FEDERAL UNIVERSITY OF MINAS GERAIS
Institute of Biological Sciences
Graduate Program in Bioinformatics

Alexandre Victor Fassio

Prioritizing promising compounds in virtual screening campaigns

Belo Horizonte
2019

Alexandre Victor Fassio

Prioritizing promising compounds in virtual screening campaigns

Final version

Dissertation presented to the Graduate Program in Bioinformatics of the Federal University of Minas Gerais in partial fulfillment of the requirements for the degree of Doctor in Bioinformatics.

Advisor: Raquel Cardoso de Melo Minardi
Co-Advisor: Rafaela Salgado Ferreira

Belo Horizonte
2019

043 Fassio, Alexandre Victor.
Prioritizing promising compounds in virtual screening campaigns [manuscrito] / Alexandre Victor Fassio. – 2019.

147 f. : il. ; 29,5 cm.

Orientadora: Dra. Raquel Cardoso de Melo Minardi. Coorientadora: Dra. Rafaela Salgado Ferreira.
Tese (doutorado) – Universidade Federal de Minas Gerais, Instituto de Ciências Biológicas. Programa de Pós-Graduação em Bioinformática.

1. Biologia computacional. 2. Reconhecimento molecular. 3. Fármacos. 4. Análise visual. 5. Receptores de Dopamina D4. I. Minardi, Raquel Cardoso de Melo. II. Ferreira, Rafaela Salgado. III. Universidade Federal de Minas Gerais. Instituto de Ciências Biológicas. IV. Título.

CDU: 573:004

Ficha catalográfica elaborada pela bibliotecária Sônia M S Moraes– CRB 6 - 1357

Universidade Federal de Minas Gerais
Instituto de Ciências Biológicas
Programa Interunidades de Pós-Graduação em Bioinformática da UFMG

ATA DA DEFESA DE TESE

Alexandre Victor Fassio

112/2019
entrada
2°/2015
CPF:
103.088.996-10

Às nove horas do dia **07 de novembro de 2019**, reuniu-se, no Instituto de Ciências Biológicas da UFMG, a Comissão Examinadora de Tese, indicada pelo Colegiado do Programa, para julgar, em exame final, o trabalho intitulado: **"Prioritizing Promising Compounds In Virtual Screening Campaigns"**, requisito para obtenção do grau de Doutor em **Bioinformática.** Abrindo a sessão, a Presidente da Comissão, **Dra. Raquel Cardoso de Melo Minardi**, após dar a conhecer aos presentes o teor das Normas Regulamentares do Trabalho Final, passou a palavra ao candidato, para apresentação de seu trabalho. Seguiu-se a arguição pelos Examinadores, com a respectiva defesa do candidato. Logo após, a Comissão se reuniu, sem a presença do candidato e do público, para julgamento e expedição de resultado final. Foram atribuídas as seguintes indicações:

Prof./Pesq.	Instituição	CPF	Indicação
Dra. Raquel Cardoso de Melo Minardi	UFMG	046.455.366-51	aprovado
Dra. Rafaela Salgado Ferreira	UFMG	055.420.566-10	APROVADO
Dr. João Paulo Ataide Martins	UFMG	802.519.713-49	APROVADO
Dr. Lucas Bleicher	UFMG	313.528.557-20	APROVADO
Dr. Rafael Victorio Carvalho Guido	IFSC/USP		
Dr. Leonardo Henrique Franca de Lima	UFSJ	048.74.[illegible]	Aprovado

Pelas indicações, o candidato foi considerado: aprovado
O resultado final foi comunicado publicamente ao candidato pela Presidente da Comissão. Nada mais havendo a tratar, a Presidente encerrou a reunião e lavrou a presente ATA, que será assinada por todos os membros participantes da Comissão Examinadora.
Belo Horizonte, 07 de novembro de 2019.

Dra. Raquel Cardoso de Melo Minardi - Orientadora _____
Dra. Rafaela Salgado Ferreira - Coorientadora Rafaela S. Ferreira
Dr. João Paulo Ataide Martins João Paulo ataide Martins
Dr. Lucas Bleicher _____
Dr. Rafael Victorio Carvalho Guido (PARTICIPAÇÃO A DISTÂNCIA)
Dr. Leonardo Henrique Franca de Lima Leonardo Henrique F. de Lima

Acknowledgments

First of all, I thank God for everything I achieved, for being who I am, for my health, perseverance, intelligence, opportunities, family, professors, and friends.

I thank my parents, who made this moment possible thanks to their encouragement to study, who gave our family the conditions to make it through, who were teachers of life when they taught us good virtues, showing us the right path to take. I also thank my brother and sister for their friendship and support.

I thank my precious, beloved, and friend, Luanna, who is my inspiration. For always supporting and encouraging me even in the ups and downs. To you, which is always with me whatever the situation, putting the pieces together in our puzzle.

I am grateful for the friendships that were established throughout my academic experience at the Federal University of Minas Gerais (UFMG) and the University of California, San Francisco (UCSF).

I thank Raquel for contributing directly to my training and for what I learned during the entire period of my Master's and PhD. Besides, I am entirely grateful for all your investment in me during this period and for making it possible for me to attend events around the world. Among these opportunities, I shall highlight especially the possibility to become a visiting scholar at UCSF.

I thank Rafaela for all her contribution to this work, consultancies, and for having accepted to be my co-advisor during the PhD. Long journey! I really had a good time discussing Chemistry with you. Finally, I am very grateful for you to be the bridge that connected me to Prof. Michael Keiser (UCSF).

I thank Michael, my abroad advisor, for having received me as a visiting scholar at UCSF. Each moment I had during this visiting program strongly contributed to my career and personal experience. I tried to make the most of my opportunity by attending lectures and international events and establishing a strong collaborative network. Thus, today, I am already reaping the harvest of this investment supported by Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES) through the process 23038.004007/2014-82, whose project was contemplated in edict 51/2013 - Computational biology.

Finally, but not least, I also thank all professors at the Federal University of Minas Gerais, who, during these 4 years, contributed to my training and knowledge. I also thank Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq) and Fundação de Amparo à Pesquisa do Estado de Minas Gerais (FAPEMIG).

After all, quoting Isaac Newton: "If I have seen further, it is by standing upon the shoulders of giants."

"*I'm a Great Believer in Luck. The Harder I Work, the More Luck I Have.*"
(Coleman Cox)

Resumo

A triagem virtual baseada na estrutura (SBVS) contribui significativamente para as etapas iniciais da descoberta de fármacos. No entanto, o SBVS geralmente depende de um processo bastante trabalhoso que consiste na seleção manual de compostos *hits*. Apesar da existência de vários trabalhos semiautomáticos que propõem métodos de *rescoring* e filtragem, a priorização e seleção automática de compostos promissores ainda é um problema em aberto. Portanto, neste trabalho, abordamos esse problema a partir de duas perspectivas. Primeiramente, no aspecto descritivo, propomos o nAPOLI (Analysis of Protein-Ligand Interactions), um servidor web que combina análises em larga escala de interações conservadas em complexos proteína-ligante a nível atômico, representações visuais interativas e relatórios detalhados sobre resíduos/átomos que interagem com a proteína. No aspecto preditivo, propomos LUNA e Functional Interaction FingerPrint (FIFP). LUNA é uma nova biblioteca Python para a descoberta de fármacos que permite a análise de qualquer complexo molecular e reúne várias funções para filtrar e visualizar interações. Por sua vez, FIFP é um novo *fingerprint* de interação do tipo *hash* que codifica complexos moleculares e suas interações como *fingerprints* binários ou de contagem. FIFP também fornece vários recursos interativos e visuais para simplificar a análise de informações dos *fingerprints*. Para validar e ilustrar a aplicabilidade do FIFP, primeiro apresentamos uma avaliação exploratória de seus parâmetros, seguida de um estudo de caso onde treinamos diferentes modelos de aprendizado de máquina para reproduzir as pontuações de *docking* observadas em um conjunto de dados composto por 86.641 moléculas em complexo com o receptor Dopamina D4. Em seguida, comparamos os modelos obtidos com quatro *fingerprints* concorrentes (ECFP, FCFP, SILIRID e PLEC). FIFP superou as abordagens concorrentes com um R^2 médio de 0,55. Portanto, vislumbramos LUNA e FIFP como estratégias promissoras para campanhas de SBVS e aprendizado de máquina.

Palavras-chave: Padrões de interação proteína-ligante, padrões de reconhecimento molecular, análise visual de interações, priorização de compostos promissores

Abstract

Structural-based virtual screening (SBVS) contributes significantly to early-stage drug discovery. However, SBVS commonly depends on a thorough manual process of hit selection. Despite the existence of several semi-automatic works that propose rescoring and filtering methods, the prioritization and automatic selection of promising compounds are still an open problem. Therefore, we tackle this problem from two perspectives. First, in the descriptive aspect, we propose nAPOLI (Analysis of PrOtein-Ligand Interactions), a web server that combines large-scale analysis of conserved interactions in protein-ligand complexes at the atomic level, interactive visual representations, and comprehensive reports of the interacting residues/atoms to detect and explore conserved non-covalent interactions. In the predictive aspect, we propose LUNA and Functional Interaction FingerPrint (FIFP). LUNA is a novel Python library for drug design that permits the analysis of any molecular complex and brings together several functions for filtering and visualizing interactions. In its turn, FIFP is a novel hashed interaction fingerprint inspired by ECFP that encodes molecular complexes and their interactions either as a binary or count fingerprint. FIFP also provides several interactive and visual features to simplify fingerprint information analysis. To validate and illustrate the applicability of FIFP, we first present an exploratory evaluation of its parameters, followed by a case study where we trained different machine learning models to reproduce the observed docking scores in a data set consisting of 86,641 molecules docked against Dopamine D4. We then compared the obtained models to four competing fingerprints (ECFP, FCFP, SILIRID, and PLEC). FIFP outperformed the competing approaches with an R^2 of 0.55. Therefore, we envision LUNA and FIFP as promising strategies for SBVS campaigns and machine learning.

Keywords: Protein-ligand interaction patterns, molecular recognition patterns, interaction visual analysis, prioritization of promising compounds

List of Figures

List of Tables

Contents

Chapter 1

Introduction

1.1 Molecular recognition and protein function

Proteins are highly abundant, versatile, complex, and vital macromolecules that play a diverse spectrum of functionalities in living beings. Countless functioning and biological processes can be pointed out, such as cellular communication, defense, metabolism, molecular recognition, movement, structural, and transport [172]. Thus, due to their importance and versatility, proteins are the focus of countless biological research, and their applicability ranges from areas such as medicine, agriculture, and biotechnology [125] to even warfare purposes [105].

Numerous biological functions performed by proteins depend on highly specific interactions with other molecules, as occurs between hormones and receptors or enzymes and substrates [125]. Thus, *molecular recognition* refers to how two or more molecules interact with each other.

To provide the means for the recognition to arise, it is required specificity and complementarity both chemical and geometric between the receptor protein and the ligand molecule (or just ligand). Such recognition takes place at regions called binding sites and is primarily controlled by noncovalent interactions [92].

An extent class of molecules can be considered a ligand: ions, carbohydrates, RNAs, DNAs, or other proteins. However, according to [5], in Chemistry, the term "ligand" is used to describe atoms and small molecules that bind to a receptor. Therefore, from now on, the term "protein-ligand interaction" will be used to refer to this particular type of interaction.

Given the crucial importance of molecular recognition in mind, several scientific research was proposed to comprehend how the recognition between two molecules occurs and which forces are involved in such a process. Understanding molecular recognition allows us to answer puzzling questions like 'how are proteins able to recognize specifically and efficiently one or more ligands?' or 'how can a ligand interact with several different proteins?'.

Although Grunenberg (2011) [98] states that these questions present a computational challenge even for simpler biological systems, computational tools are valuable and promising due to the complexity and volume of available data concerning molecular recognition.

1.2 Noncovalent interactions

Molecular recognition is a phenomenon that requires high specificity and complementarity both geometric and chemical between two molecules. The latter is usually driven by intermolecular interactions known as noncovalent interactions.

The preponderance of noncovalent interactions is explained by their weaker characteristic when compared to covalent interactions, which makes them susceptible to reversibility. This aspect is fundamental to life as it enables organisms to respond quickly to changes in the environment.

Therefore, in this section, we depict several noncovalent interactions that contribute to molecular recognition, such as aromatic stacking, electrostatic interaction, halogen bond, hydrogen bond, and hydrophobic interaction, among others.

1.2.1 Aromatic interactions

Aromatic interactions, as the very name indicates, are interactions involving aromatic rings. These interactions present an electrostatic component, which arises from the resonant double bonds in the aromatic ring, the electrons from the σ orbital within the ring plane, and π orbitals above and below the plane. The latter generates partial negative charges above and below the plane, and a partial positive charge in the ring plane. Consequently, these partial charges create a quadrupole momentum that enables the rings to interact with other aromatic rings and other systems through an electrostatic interaction [114, 125, 236].

1.2.1.1 Aromatic stacking or π-π interactions

Aromatic stackings or π-π interactions consist of electrostatic, hydrophobic, and van der Waals forces involving two aromatic rings [114]. Three typical stacking conformations are possible, namely, edge-to-face or T-shaped, parallel displaced or offset stacked interactions, and face-to-face (see Figure 1.1). Note however that the last stacking is unfavorable as the partial negative charges from the rings will be in contact.

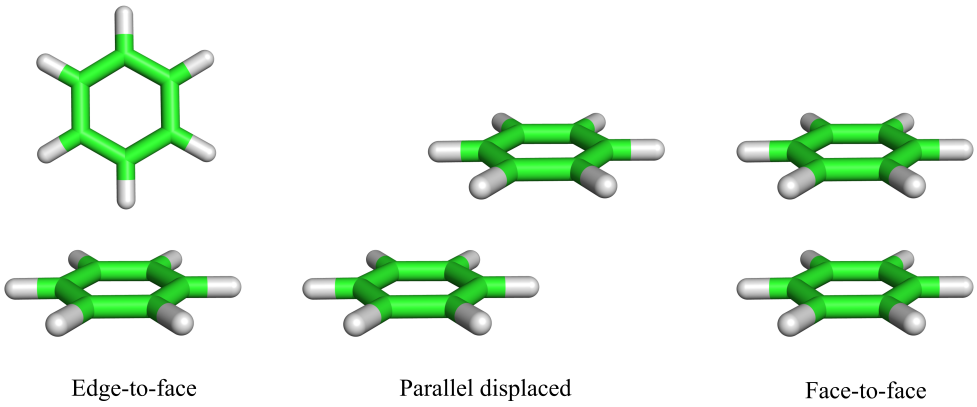


Figure 1.1: Typical geometrical arrangement of aromatic stackings.

1.2.1.2 Cation–π interactions

Cation-π interactions are electrostatic interactions involving an aromatic ring and a nearby cation. That can happen due to the partial negative charges above and below the ring plane, which produces an attraction with a positively charged atom [220].

In proteins, this type of interaction is observed between cationic side chains (ARG, HIS, and LYS) and aromatic side chains (HIS, PHE, TRP and TYR) [89]. Cationic ligands other than amino acid side chains as, for instance, acetylcholine [72] and metal cations are also examples of molecules and ions capable of establishing cation-π interactions.

1.2.1.3 Amide-π interactions

Similar to aromatic stackings, amide-π is an interaction in which aromatic ring stacks against the π orbitals of the amide (Figure 1.2), characterizing, therefore, an example of a dipole-quadrupole attraction [101, 229].

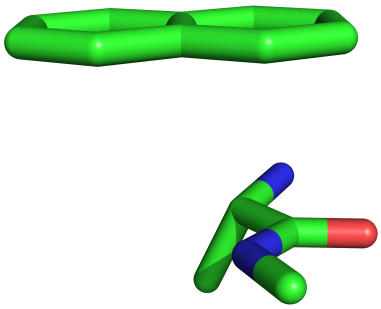


Figure 1.2: Example of an amide-π stacking.

1.2.2 Hydrophobic interactions or nonpolar interactions

The inclusion of a nonpolar substance in a polar solvent, like water, produces an entropically unfavorable system. In order to compensate for the loss in entropy, apolar substances tend to aggregate and interact with each other to reduce the contact with the solvent, the so-called hydrophobic effect [75, 125, 145, 231].

These interactions involving apolar substances are commonly referred to as hydrophobic interactions. However, Privalov and Gill (1988) [188] drew attention to the fact that hydrophobic interactions are indirect consequences of the hydrophobic effect, and, therefore, they are not considered classical atom-atom interactions.

1.2.3 Hydrogen bonds

Hydrogen bonds, one of the most important and studied interactions, arise due to the electronegativity difference between an electronegative atom (typically fluorine,

nitrogen, or oxygen) and a hydrogen covalently bound to it. This difference causes the electron cloud from the hydrogen to displace toward its partner (hydrogen donor), which leaves a partial negative charge on the donor and a partial positive charge on the hydrogen. An attractive force, namely hydrogen bond, is then formed between the hydrogen and an electronegative atom containing a partial negative charge in the vicinity and free electron pairs (hydrogen acceptor).

Hydrogens bonds' strength varies according to the atoms involved in the interaction, the bond length between the hydrogen and the acceptor, and the angle formed between the donor, the hydrogen, and the acceptor atom. Stronger hydrogen bonds were demonstrated to be approximately linear ($\approx 180°$) and to have a short bond length between the hydrogen and the acceptor (1.2 to 1.5 Å) [112, 226].

1.2.4 Weak hydrogen bonds

Although typical hydrogen bonds involve highly electronegative atoms, some non-conventional and weaker hydrogen bonds can also comprehend carbons and aromatic rings [25, 43, 62, 65, 230, 251]. In those circumstances, carbons act as hydrogen donors and aromatic rings act as acceptors, which happens due to its electron-rich cloud above and below the ring.

1.2.5 Halogen bonds

Halogen bonds are interactions with an electrostatic component that are considered similar to hydrogen bonds as they also involve an electron acceptor and donor moieties. They are frequently represented as $B\cdots X—R$, where B is a Lewis base (halogen acceptor), X is a halogen (usually chlorine, bromine, or iodine), and R is an atom covalently bound to it. By this terminology, it is important to mention that the term Lewis base comprehends any electron-rich species, including, therefore, aromatic rings.

In opposition to the commonly expected behavior of halogen atoms, when covalently bound to other atoms, they act as electron acceptors (electrophiles) rather than electron donors (nucleophiles). A well-established explanation for this unusual behavior is that the electron density becomes anisotropically distributed (not uniform) due to the covalent bond (Figure 1.3). Analysis of electrostatic potential shows that the electron

cloud is displaced toward the covalent bond, which causes this region to obtain a negative potential [186]. Consequently, a cap on the elongation of the covalent bond arises with a positive potential, the so-called σ-hole. As this region is depleted of electrons, it can be electrostatically attracted by a nucleophile moiety, giving rise to the halogen bond.

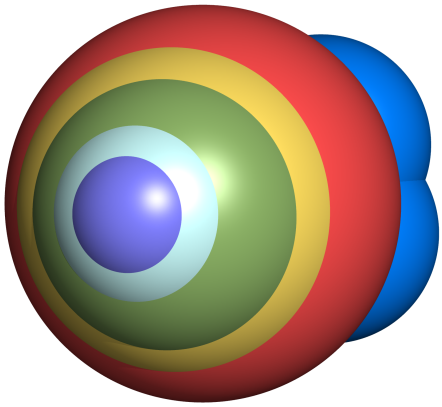


Figure 1.3: A representation of the unusual electron density in a halogen atom covalently bound to carbon. The color scale varies from blue, the most positive surface potential, including the σ-hole, to red, the most negative surface potential.

An interesting feature of halogen bonds that arises from the σ-hole is that the interactions tend to be highly directional [163, 164, 165, 166] whereby stronger ones tend to be shorter and linear (ranging from 140° to 180°).

Of all halogen atoms, fluorine is the only exception not to be pointed out as a participant of typical halogen bonds. The reason is that fluorine does not frequently form σ-holes due to its high electronegativity.

1.2.6 Chalcogen bonds

Chalcogen bonds are interactions identical to halogen bonds in their origins as they also arise from σ-holes, a cap on the elongation of the covalent bonds with positive potential. The main difference between these bonds, as the very name indicates, is that they involve chalcogen atoms, namely oxygen, sulfur, selenium, and tellurium. These interactions can be represented as $B\cdots Y—R$, where B is a Lewis base (chalcogen acceptor), Y is a chalcogen, and R is an atom covalently bound to it (usually carbon and sulfur). However, as σ-holes and the strength of chalcogen bonds increase as the atom polarizability increases, bonds formed by oxygen atoms are the weakest and are relatively rare

[18, 171]. It is also important to highlight that the term Lewis base comprehends any electron-rich species, including, therefore, aromatic rings.

Another common feature between chalcogen and halogen bonds is their high directionality [154]. However, unlike its sister interaction, chalcogen bonds are typically divalent [154], which means that two *sigma*-holes are present in the elongation of each covalent bond. As a consequence, two chalcogen bonds may be formed (Figure 1.4).

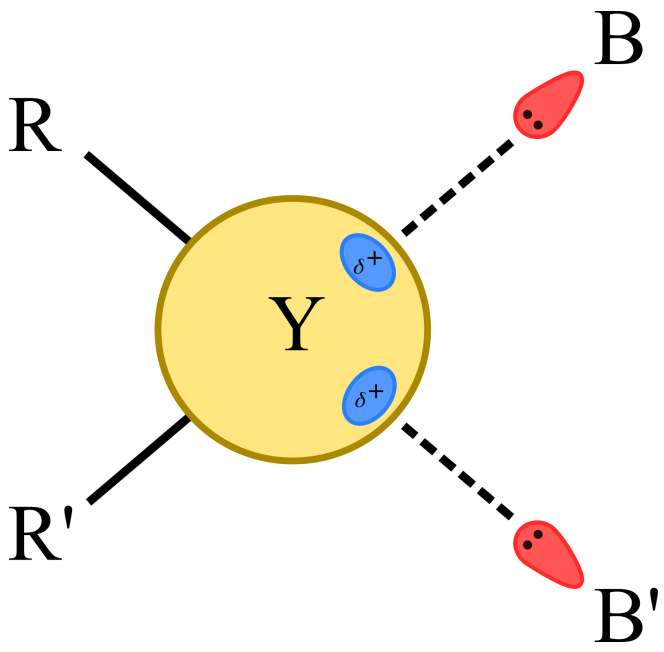


Figure 1.4: A representation of a divalent chalcogen atom (Y) establishing two chalcogen bonds with its Lewis base (B and B') partners. R and R' are atoms covalently bound to Y, usually carbon and sulfur. In blue, it highlights the *sigma*-hole.

1.2.7 Tetrel and Pnictogen bonds

Tetrel and pnictogen bonds are also interactions from the family of halogen and chalcogen bonds. These interactions, in its turn, involve groups 14 and 15 elements from the periodic table, respectively. However, to the best of our knowledge, the application and exploration of these interactions regarding their biological implications are still limited and are typically restricted to theoretical and chemical crystal structure studies [17, 18, 143, 212].

1.2.8 Van der Waals interactions

Van der Waals interactions arise from the distribution of the electronic charge on an atom at a given moment. Since electrons continuously move around the atom nuclei, its charge distribution becomes asymmetric, and a momentary dipole is formed. Although this polarization is momentary, it is enough for inducing opposite charges in another atom, forming the so-called induced dipole. As a consequence, the atoms, which momentarily are with opposite charges, are attracted to each other by a force called London force or dispersion force. As the atoms approach each other, their electrons start to collide and repeal each other, which is in accordance with Pauli's exclusion principle. The combination of attraction and repulsive forces is called van der Waals interactions. However, as pointed out by [125], the term "van der Waals" is typically used only in the description of the attractive forces.

1.2.9 Electrostatic interactions

Electrostatic interaction is a broad term used to coin all interactions involving partially or fully charged atoms or groups of atoms [118]. When both atoms have the same charge, the interaction is repulsive and unfavorable; otherwise, it is attractive and favorable.

Some examples of such interactions include ionic, dipole-dipole, and dipole-ion [32, 231]. Other interactions also contain an electrostatic component, including aromatic interactions, hydrogen bonds, and van der Waals. That said, in this section, we focus on the three first examples.

Ionic, dipole-dipole, and dipole-ion can be described by basic Physics concepts, namely Coulomb's law. Thus, the strength of an electrostatic force is inversely proportional to the distance between two charged entities. Other interactions, such as hydrogen bonds, that contain other components should also be described considering these other characteristics [125].

1.2.9.1 Ionic interactions

Ion-Ion or ionic, the strongest noncovalent interaction, arise when fully charged atoms attract or repeal each other.

1.2.9.2 Salt bridges

Salt bridges are a special type of interaction that involves simultaneously a hydrogen bond and ionic interaction, which means that the acceptor and donor atoms participating in the hydrogen bond must be fully and oppositely charged [183]. In proteins, these interactions can be established by positively charged residues (ARG, HIS or LYS) and negatively charged residues (ASP or GLU) [15]

1.2.9.3 Dipole-dipole interactions

Dipole-dipole or multipolar interactions [181] are attractive/repulsive interactions comprising partially charged atoms.

1.2.9.4 Ion-dipole interactions

As the very name indicates, ion-dipole forces arise from the attraction/repulsion between fully (ion) and partially (dipole) charged atoms.

1.3 Contacts and interaction calculation

A typical step preceding the identification of protein-ligand interactions is contact mapping. The term *contact* is often confused with *interaction* or used interchangeably. According to [55], *contact* refers to the spatial distribution of an atom and which atoms comprise its vicinity. In its turn, *interactions* consist of all forces mediating the interaction between atoms.

In this section, we first present two conventional approaches for detecting contacts, followed by interaction representations, and finally, we present a series of tools currently available in the literature.

1.3.1 Contacts calculation

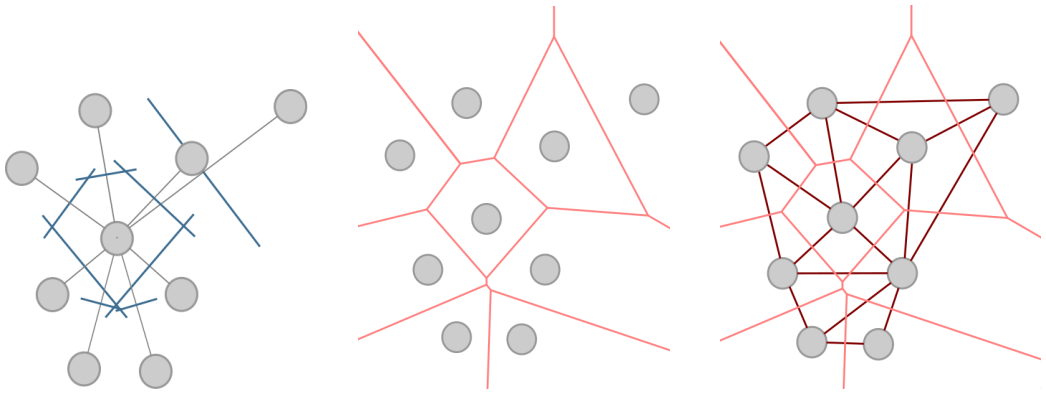
Contacts can be calculated by using a cutoff-dependent or a cutoff-free strategy. The former approach depends on a threshold value (cutoff) to determine the atom vicinity. Thus, the vicinity of an atom X is determined by identifying which atoms are inside the sphere of radius R (the cutoff) centered in X. However, the choice of this threshold is not trivial [55].

Another drawback of such an approach is that it tends to identify "false contacts" or occluded contacts, which occur when two atoms A and B are establishing a contact but there is a third atom between them. From a physical point of view, a potential interaction involving A and B are weakened and impaired by the third atom [70].

A straightforward solution for occlusions that occur due to covalently bound atoms is the calculation of angles formed by the atoms in contact and their bound neighbors. Such a method is typically taken into account during the interaction calculation step. In HBPlus [161], for instance, the angle between the donor, the acceptor, and an atom bound to the acceptor ensures that no atom is within the limits of a hydrogen bond. Besides that, occlusions can also be reduced through shorter thresholds, which are usually applied for some specific interactions.

Although these solutions reduce the number of occlusions significantly, some may still be found. Cutoff-free approaches, on the other hand, did not experience the aforementioned problems [55] since they do not require a threshold value and do not detect occluded contacts. An example of a geometric and cutoff-free approach widely used is the Delaunay tesselation (DT) that can be obtained as follows: firstly, the algorithm partitions the atoms in geometric regions called Voronoi cells whose centroid is the atom. Each

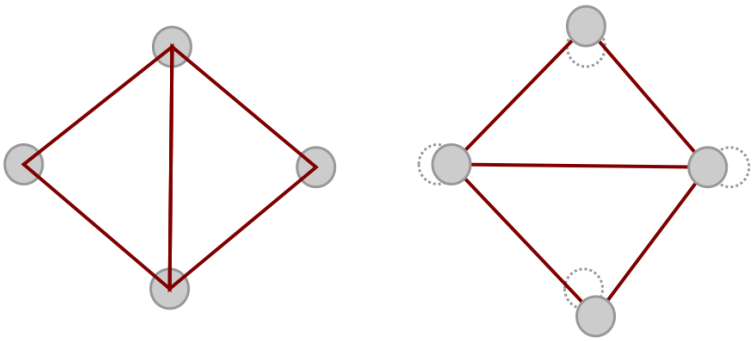
cell is defined as the smallest polyhedron formed by the linkage between a centroid and all other centroids (Figure 1.5a), and the set of all cells form a Voronoi diagram (Figure 1.5b); next, a DT is generated by connecting the centroids of adjacent cells (Figure 1.5c). Consequently, connected centroids form a graph in which edges represent atoms (vertices) in contact.



(a) (b) (c)

Figure 1.5: Voronoi diagram and Delaunay tessellation (DT). (a) shows an example of the construction of a Voronoi cell. (b) shows a set of Voronoi cells, the so-called Voronoi diagram. (c) shows a DT (in light red) obtained through a Voronoi diagram (in dark red) by connecting the centroid of adjacent Voronoi cells.

However, a limitation of DT is that it is sensible to minimal changes in atomic positions as observed, for instance, between two models of an NMR structure [187], leading to substantially different Delaunay tessellations (Figure 1.6).



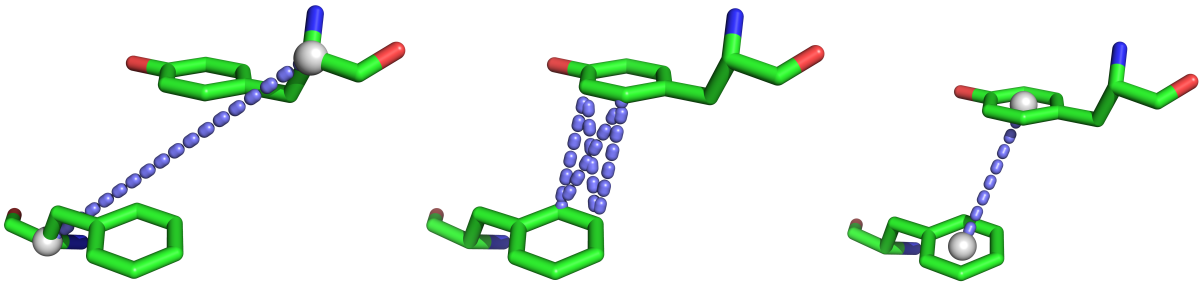
(a) (b)

Figure 1.6: Sensitivity to minimal changes in Delaunay tessellation (DT). In (a), four atoms are shown in their initial position and in (b), a small perturbation in atom positions leads to substantially different Delaunay tessellations.

1.3.1.1 Contact and interaction representation

Independently of the method chosen for contact prediction, contacts and, subsequently, interactions can be categorized by the granularity of the points [55, 133]. In coarse-grained models (Figure 1.7a) [126, 246], contacts/interactions between molecules are calculated departing from their representative point, which could be, for instance, alpha carbons (CA) or their geometric center. Such modeling is especially preferred for protein-protein analysis and molecular simulations of large biomolecules, which demand high computational resources [10, 133].

Fine-grained models (Figure 1.7b) represent contacts/interactions at an atomic-level, which are by far the most used representation. Examples of works that apply this model include [138, 155, 208, 245, 249]. In the first phase of our work (Section 2.1), we also opted for the atomic-level model [83]. However, in the second phase, we decided to use a hybrid method (discussion below), which is more approximate to how chemical interactions occur.



(a) (b) (c)

Figure 1.7: Different representations for aromatic stacking according to coarse-grained (a), fine-grained (b), and hybrid models (c). In (a) and (c), the spheres represent the alpha carbon and ring centroids, respectively.

Finally, in hybrid models (Figure 1.7c), both atoms and groups of atoms are considered during the contact/interaction perception. We believe this strategy to better model some interactions that are established due to the contribution of multiple atoms. The classic example is the interaction between two aromatic rings, where the stacking arises from the contribution of all atoms in the ring. Thus, the ring centroid can be used as the representative point, and interactions can be perceived using it as a reference. An additional example is carboxylic acids whose anionic characteristics could be attributed to all of its atoms. As a consequence, the geometric center of this chemical group could be used as the representative point. In the second phase of our work, we decided to model interactions using such a hybrid method (Section 2.3). Other examples of tools that apply this modeling include PLIP [207] and Arpeggio [121].

1.3.2 Tools for calculating protein-ligand interactions

Elucidating the mechanisms involved in molecular recognition and which forces contribute to this phenomenon is a central problem in biology [123]. Since noncovalent interactions are the primary contributors to the recognition between two molecules, investigating the mechanisms involved in protein-ligand interactions significantly contributes to the understanding of how molecular recognition occurs, and to ligand prediction, target identification, lead discovery, and drug design [90, 185].

Databases like the Protein Data Bank (PDB) [21], which comprehends more than 91,240 protein-ligand complexes (as of October 2019), provide us with essential knowledge about protein-ligand interactions. Not surprisingly, several tools and databases were proposed as an effort to investigate and depict protein-ligand interactions through such large-scale data set [3, 36, 49, 54, 60, 77, 88, 103, 107, 121, 123, 215, 155, 157, 64, 192, 138, 161, 207, 208, 214, 222, 225, 247, 233, 238, 242, 245]. These tools are widely employed within the scientific community, being of great value in structure-based drug design projects.

However, to the best of our knowledge, there is not an all-in-one tool that presents a visual and interactive large-scale automated analysis of conserved interactions, and a comprehensive report about them. Altogether, these features would further improve the understanding of molecular recognition mechanisms.

In general, the tools currently available allow an investigator to analyze only one complex at a time, and the comparison between multiple complexes is toilsome since it must be performed manually. Even Ligplot+ [138], which permits the comparison of multiple structures, has some limitations, such as the small number of simultaneous diagrams that a user can visually analyze. Some other tools [107, 157, 192, 242] were conceived to assist analysis of docking and virtual screening results. In these tools, noncovalent interactions are calculated both by angle or distance criteria [157, 192, 242] and by energy-based criteria [107]. Results are presented as matrices of interaction fingerprints (IFP), which allow the analysis of multiple complexes at once. Nonetheless, as these tools focus on analyzing docking results, they do not provide an automated feature to detect and visualize conserved protein-ligand interactions through structures available in the PDB. Determining these conserved interactions can provide evidence and suggestions to improve our comprehension of the critical factors in molecular recognition, and to guide a docking and virtual screening study.

Bearing this in mind, we proposed nAPOLI[1] (Analysis of PrOtein-Ligand Interactions), a web server that brings together an automated analysis of conserved interactions

[1] http://bioinfo.dcc.ufmg.br/napoli/

across large data sets of protein-ligand complexes, interactive visualizations, and comprehensive reports of the interacting residues/atoms to explore and make sense of conserved noncovalent interactions that work as key factors in molecular recognition.

This work was initially presented as a Master's thesis [84], but during the present work, nAPOLI was expanded with novel features, and its methods were improved, including the ones for protein-ligand interaction calculation [83]. The improved methods and novel features are presented in Section 2 and 3, while a brief overview of nAPOLI is presented in the next section.

1.3.3 nAPOLI

The core of nAPOLI is divided into two sections: *Dataset submission* and *Dataset analysis*. In *Dataset submission*, users are able to start a new project and submit their own data set. At the time nAPOLI was first presented [84], users could only submit projects by composing their data set in an exploratory way (see Section 3.1.1). However, two new submitting options were included in the current version of nAPOLI (see Section 3.1.1) [83].

After the project submission, nAPOLI calculates the protein-ligand interactions and generates a series of statistics, which will become available in *Dataset analysis*. To do so, we devised a strategy to detect conserved atomic-level interactions in protein-ligand interfaces modeled as bipartite graphs in which nodes are atoms from protein or ligand, and edges are the interactions among them. Nodes and edges are labeled with the physicochemical properties of atoms and interactions, respectively.

Some of the major questions nAPOLI aims to answer are: what are the possible interactions that each ligand can establish with the protein? What is the frequency of each type of interacting atom in the ligands data set? What is the frequency of each type of interaction? Which residues interact with the ligands? At what frequency each residue interacts with the ligands? Are there clusters of similar ligands? nAPOLI was also developed to permit analysis of the whole data set or by choosing a cluster of similar ligands for every feature.

The following sections present the visual strategies that support analyzing protein-ligand interactions.

1.3.3.1 Dataset summary

Provides an interactive table that contains the summary of the user data set (Figure 1.8a). A variety of interactive options were designed to help users to analyze their data, including information about each type of atom or interaction; images of ligand chemical structure or superposition of all ligands of a cluster, which helps to identify structural similarities; search options and table sorting; just to list a few. nAPOLI also provides an *Interaction viewer* (Figure 1.8b), where users can visualize the complex and its interactions in an interactive and straightforward way. Both 3D and 2D-view of the complex are displayed. 3D-view is a molecular viewer (3Dmol.js library [197]), while the 2D-view is the bipartite labeled graph whose edges are protein-ligand interactions. As nAPOLI can define different interactions with a pair of atoms, multiple edges are included in the representation.

1.3.3.2 Interactions by residues

The *Interactions by residues* section (Figure 1.8c) report which residues frequently interact with a set of ligands and which types of interactions they establish.

We proposed a color coded table where users can analyze the whole data set or a cluster of similar ligands at a time. This table has three main columns: *Atom (the atom name)*, *Type of interaction* and *# Ligands with which it interacts (frequency)*. The secondary headers show information about the alignment position - or the residue name if it could not be aligned to any residue of the template structure - and the total frequency with which such position/residue was found interacting. Such secondary headers group the lines of the previous three columns and are coded by a heat-based color system that varies from blue (cold lower frequencies) to orange (hot higher frequencies). Thus, through this table, it is possible to detect the total frequency of a position/residue as well as the frequency for each atom. Users can discover which residues were aligned to a certain position and which PDB structures have such residues.

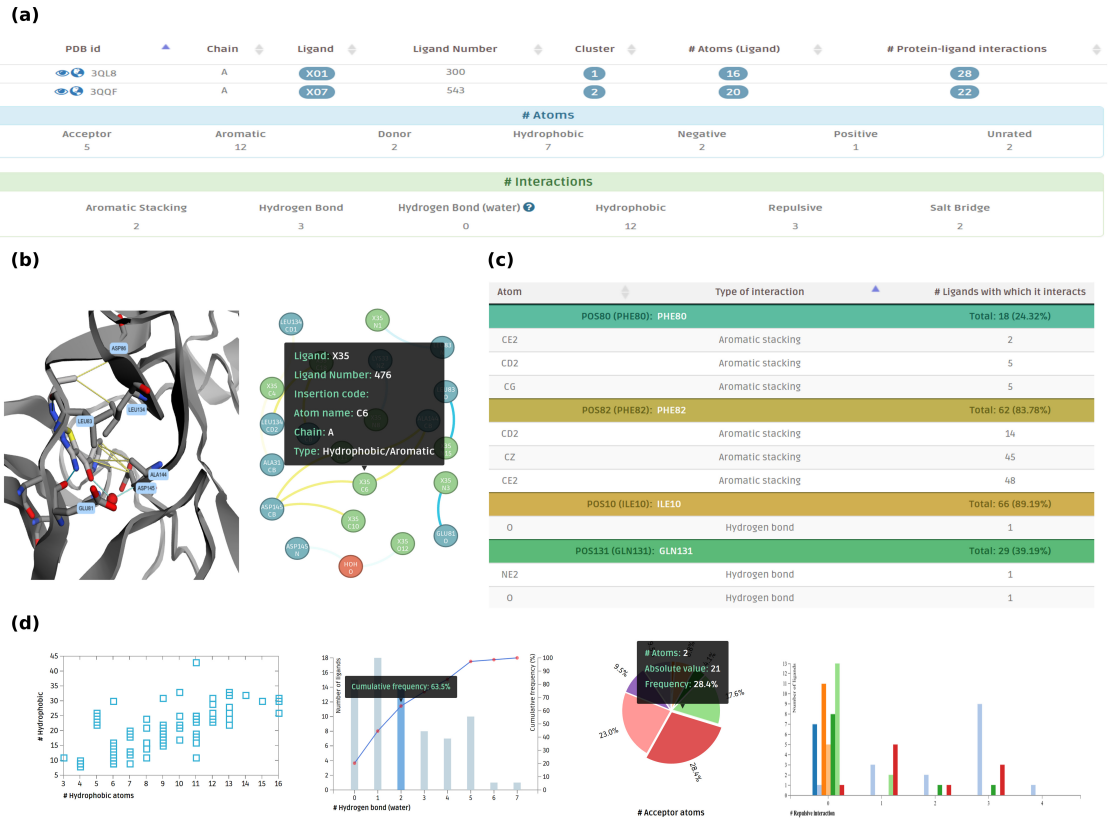


Figure 1.8: Visual strategies to analyze protein-ligand interactions. (a) *Dataset summary* table. (b) *Interaction viewer*. In this viewer, nodes represent the protein (blue) and the ligand (green) atoms or a water molecule (red), and edges represent the established interactions, which are color coded. For example, blue and yellow edges represent hydrogen bonds and hydrophobic interactions respectively. (c) Color coded table from the *Interactions by residues* section. (d) Available charts at the *Graphical analysis* section. From left to right are shown a scatter, Pareto, pie and grouped chart. Slices in pie charts can be colored based on the frequency or the number of atoms/interactions. In grouped charts, each cluster is represented by a column.

1.3.3.3 Graphical analysis

In *Graphical analysis* section (Figure 1.8d), users have a statistical report using Pareto charts, pies, grouped bars or scatter plots.

1.3.3.4 Interactions by ligands

All the protein-ligand interactions are presented and listed in a condensed and concise table where each line contains the interactions established by each ligand. Users

can also visualize the protein-ligand interactions in 3D and 2D-view such as described in Section 1.3.3.1.

1.4 Molecular and interaction fingerprints

Molecular fingerprints (MFP) are descriptors that symbolically encode chemical information, such as chemical formula, boiling and melting points, hydrophobicity, molecular weight and size, number of rotatable bonds, and polarity, among others [202, 228], typically through a binary or counting (frequency of a feature) sequence. MFPs are widely used in screening and similarity comparison between a set of molecules, which permits the discrimination of molecules according to a set of goals [202]. Similarly, interaction fingerprints (IFP) encode contacts or interactions between atoms as a means to describe a protein-ligand complex. IFPs are broadly applied in virtual screening as a post-processing step to select compounds according to their similarity to a known complex of reference.

Frequently, the similarity between two molecules or complexes is given by the Tanimoto coefficient, which varies from 0 to 1, and whose equation is presented below:

$$T(F1, F2) = \frac{c}{a + b - c} \tag{1.1}$$

Where a and b are the number of bits *on* (equal to 1) in the fingerprint F1 and F2, respectively, and c is the number of common bits *on* between the fingerprints.

There are two major types of molecular fingerprints, namely structural and hashed fingerprints. IFPs are often described as a specific fingerprint type, but due to their common characteristics with the mentioned MFPs, we will describe IFPs as falling into these categories.

In the following sections, we discuss the two major types of fingerprints. Therefore, for a broader overview of other types of fingerprints refer to [12].

1.4.1 Structural fingerprint

In structural fingerprints, each bit in a binary sequence implies the presence or absence of a predefined chemical/interaction feature. In the context of an MFP, for example, the first bit could indicate whether a guanidine exists or not in a particular

molecule. On the other hand, the first position in an IFP could indicate if the ligand establishes a hydrogen bond with a particular residue R. For IFPs, one common approach [61] is the separation of a block with N bits (one for each available interaction type) for each residue in the binding site.

The number of features (fingerprint length) and which ones will be depicted in the fingerprint directly impact the performance of algorithms that employ screening and similarity comparison. Therefore, as Leach and Gillet (2007) [142] points out, the substructures selection strongly depends on the ligands data set. Moreover, in IFPs, the choice of which information to encode limits fingerprint usage in a multi-protein context because a predefined set of bits from one fingerprint does not necessarily represent the same bits in another fingerprint. That can happen because the binding site of two different proteins may have different residues and, therefore, the information encoded in the fingerprint will not be the same. A possible workaround for this problem is to perform a structural alignment and define the bit information according to the aligned residues.

One of the most well-known MFP is the MACCS [76], although other examples can be cited, namely Avalon [91], PubChem fingerprint[2], E-state [100], BCI [16], and FP3 and FP4 from Open Babel [177].

Examples of IFPs that fall into this category are SIFt [61], Marcou's IFP [157], APIF [191], Pharm-IF [210], PyPLIF [192], TIFP [63], SILIRID [48], LORD_FP [237], Arpeggio's IFP [121], and PLIF [97].

1.4.2 Hashed fingerprint

Hashed fingerprints or topological fingerprints, differently from structural fingerprints, do not contain a predefined sequence of features. Instead, the features are perceived during the processing of molecules/complexes where unique substructures and patterns are recognized and mapped to a bit position through a hashing function - hence the name "hashed fingerprint". However, as the number of features is not available a priori and as the fingerprint length is finite, it may occur the so-called bit collision, which happens when two unrelated features are mapped to the same bit position. As shorter the fingerprint length is the more collisions may occur, which, as pointed out by Rogers and Hahn (2010) [203] and Sastry et al. (2010) [209], causes the loss of information and adds noises to the analysis. On the other hand, with a large enough fingerprint, the collision rate becomes minimal [203].

The major advantage of hashed fingerprints over structural fingerprints is the gen-

[2] ftp://ftp.ncbi.nlm.nih.gov/pubchem/specifications/pubchem_fingerprints.txt

erality, i.e., it does not require the definition of a set of features a priori. As a consequence, the fingerprint can explore the feature space more efficiently and permits the comparison between two fingerprints in multiple contexts, not being limited to a specific scenario and data set. Additionally, given such a large number of predefined features, structural fingerprints tend to be much more sparse and slow for substructure searching.

However, the lack of a predefined set of features is also a double-edged sword as it is not possible to map from a bit position back to the substructure that set the bit on, which reduces its direct interpretability [142].

Examples of hashed MFPs include Atom pairs [37], Topological torsion [173], MOLPRINT 2D [19], Daylight fingerprint [59], ECFP [203], and E3FP [9]. Regarding IFP approaches, to the best of our knowledge, SPLIF [53] and PLEC [245] are the only fingerprints that fall into this category.

1.5 Drug discovery and virtual screening

The discovery and development of new lead compounds is a highly expensive and time-consuming endeavor that takes up to 10-15 years [113]. Moreover, given the theoretical number of chemical molecules that can be considered in a study (10^{60} to 10^{100}) [148], it is infeasible for current technologies to comprise this massive volume of compounds.

Thus, computational techniques like virtual screening (VS) and molecular docking are becoming more and more popular as they contribute significantly to early-stage drug discovery. The great acceptance and advantage of VS and docking are justified by a large number of compounds (thousands of compounds) that are evaluated against a target protein *in silico*, which sharply reduces costs and narrows the lead discovery [139]. It is worthwhile to mention that these methodologies benefit from the immense amount of available data in protein and chemical databases, such as the Protein Data Bank (PDB) and ZINC, which encompass around 131,000 protein structures and 95,614,358 compounds, respectively [21, 116].

VS techniques can be classified into ligand-based (LBVS) and structure-based (SBVS) methods based on the type of data available [152, 220].

LBVS is useful when structural information of the receptor (generally, a protein) is not available, but a set of active ligand molecules for the desired receptor is known, which is, therefore, the starting point to identify candidate compounds for experimental evaluation. Some LBVS methods include similarity and substructure searching, quantitative structure-activity relationships (QSAR), pharmacophore matching, and three-dimensional shape matching [139]. In contrast, SBVS is useful when the receptor structure is known,

which supports the use of docking methods to predict the protein-ligand conformation. Drwal and Griffith (2013) [73] further add that when the structure of the receptor is available, both methods can be synergistically integrated to improve the drug design process. In our work, we focus on the SBVS strategies and, therefore, from now on next sections discuss the SBVS phases, namely the data preparation, the docking, and post-analysis.

1.5.1 Ligand and receptor preparation

The virtual screening success depends on many factors during each stage. In the initial phase, the appropriate preparation of both ligand and receptor molecules is critically important to obtain satisfactory outcomes.

The ligand preparation starts by selecting a set of ligands from chemical databases like ZINC [116] or ChemDB [46]. The chemical space covered by current databases is on the order of thousands to millions, and in a VS campaign would be a waste of time to screen all compounds available. Thus, filtering ligands for docking is highly common and reasonable. A usual strategy is to filter ligand databases by using drug-like physicochemical properties based on Lipinski Rule of Five [147], as well as filtering potentially reactive and toxic compounds. According to Klebe (2006) [132], ligand filtering can also be improved by using the property profiles from the receptor binding site, such as the pharmacophores.

Furthermore, in most cases, only a 2D representation (SMILES) of ligands is available, and the proper and realistic ligand conformer (the 3D representation) should be generated. In this procedure, several considerations must be taken into account, namely the correct assignment of ionization and tautomeric states and specification of enantiomers arising from chiral centers in the molecules [152, 234]. Lastly, sometimes, it is also necessary to assign partial charges to the compounds as a requirement of some docking tools [152].

Regarding receptor preparation, it is important to check for structural integrity, for instance, verifying if there are missing residues, especially in the binding site. Additionally, assign appropriate ionization states of residues in the binding site, and the correct tautomer for histidines must be taken into account. Usually, hydrogen atoms are added to the protein and geometry refinement is employed for optimizing both protein-ligand complex and also added hydrogen atoms. Finally, it is also recommended to maintain structural waters whenever such molecules are essential to the protein-ligand interaction [152, 234].

1.5.2 Molecular docking

Molecular docking, or just docking, is a computational strategy envisioned to predict the likely binding mode of a small molecule at a particular receptor (generally, a protein), the so-called ligand pose prediction. Since this work focus on protein-ligand complexes, from now on, we will refer to the receptor as being a protein molecule.

Docking consists of two major phases: pose prediction and pose scoring. An accurate pose prediction relies on the degree of ligand and receptor flexibility. Early docking methods considered both molecules as rigid bodies, taking into account the lock-and-key model proposed by Fischer (1894) [86]. However, currently, several different approaches already consider the flexibility of ligands and receptors to some extent in order to contemplate the induced fit and the conformational selection model [30].

Finally, the correct and precise assignment of a scoring function comprises the touchstone to rank compounds and distinguish ligands from non-ligands. Therefore, it is well-established that scoring is still the Achilles' heel of docking methodologies since it is very challenging to accomplish small processing timescale, precision, and complexity. In other words, a docking algorithm should be fast as the screening process comprehends thousands of complexes, but the correct evaluation of a pose must also be accurate, which usually involves complex calculations that demand more computational processing [131, 234].

1.5.3 Post-analysis

As already discussed, it is well-established that scoring functions have several drawbacks that trace back to their various assumptions and simplifications in the evaluation of modeled complexes. Consequently, non-ligands may be prioritized first over true ligands, which is not desirable [234].

Therefore, it is very wise to perform post-analysis procedures in order to minimize the number of false positives in the selection list and to propagate the true hits to the top of the list [152]. In this section, we will discuss some strategies commonly employed to obtain better screening outcomes.

1.5.3.1 Evaluation of the virtual screening performance

In order to evaluate the protocol employed in both ligand and receptor preparation and the docking protocol, it is crucial to validate the obtained poses. The first simple approach to access and validate the docking results is through redocking, which involves taking the crystallographic ligand from the receptor-ligand complex and then docking the same ligand into the original receptor coordinates by using a docking program. This experiment is commonly used to evaluate the accuracy of a scoring function and a docking program regarding their pose reproduction ability.

Another assessment of a docking program pose reproduction is cross-docking [130, 234], in which docking is performed in a protein structure co-crystallized with a different ligand. This type of experiment is interesting since docking algorithms use rigid proteins and therefore, a different receptor conformation may allow the sampling of different ligand conformations. Additionally, cross-docking resembles the typical scenario in virtual screening. Usually, predicted poses are compared to the crystallographic position based on the root-mean-square deviation (RMSD) between them. In [1], a docking is considered successful if the top-scoring pose was within 2.0 Å RMSD from the crystallographic position.

Another way to evaluate the performance of docking protocols is through enrichment studies. These comprise rank ordering a compound library containing a set of known ligands (actives) among a large number of non-ligands (decoys) [1], which can be obtained from the Directory of Useful Decoys (DUD) database [109]. The expectation is that known actives would rank higher than non-actives by the docking program and protocol. In [130], the authors discuss several different enrichment descriptors.

Additionally, outcomes can be evaluated by the receiver operating characteristic (ROC) curves that plot the true positive rate (sensitivity) against the false positive rate $(1 - specificity)$, based on their total area under the curve (AUC). Smaller subsets of the docking database can also be assessed; such that early enrichment is reported. A good early enrichment in a VS shows that active compounds were ranked at the very top of the database, which is crucial since only a small number (in most cases a few dozen) of compounds are usually selected for experimental testing.

1.5.3.2 Consensus scoring and rescoring

Bearing in mind the limitations of scoring functions, one strategy is the use of consensus scoring, in which more than one scoring function is used to select top-ranked compounds that are common to each function. Another strategy is to use a rescore function, such as more sophisticated methods that take into account a more appropriate description of interactions and incorporate the solvation effect. Examples of such methods are the Molecular Mechanics-Poisson–Boltzmann Surface Area (MM-PBSA) and Molecular Mechanics-Generalized Born Surface Area (MM-GBSA) [152, 234].

Some other works [53, 61, 63, 157] propose rescoring docking poses through accessing protein-ligand interactions profiles. Different approaches exist and broadly speaking, they are based on interaction fingerprints, which are obtained by converting 3D structural binding information into a one-dimensional (1D) binary string. In such strategies, docking poses are evaluated by comparing the interactions established in the complex of reference, and their similarity is measured with the Tanimoto coefficient. Thus, highly similar fingerprints are ranked first. In [61, 157], compounds were further clustered by using a hierarchical agglomerative algorithm. The critical drawback of such methods is that they do not provide their source code. Moreover, these methods depend on the reference structure to rank the compounds, which is not the case of iGEMDOCK [107], which is a fingerprint-based method where the authors presented a new score function based on interactions conservation. Nonetheless, a disadvantage of iGEMDOCK is that users are required to use the author's docking tool, the GEMDOCK [250].

Similarly, AuPosSOM [27] is a fingerprint-based approach that introduces a new scoring function and proposes the use of self-organizing maps (SOM) to cluster the compounds. A disadvantage of such a method is that it requires a list of known active compounds to train the neural network. Thus, in the following work, Mantsyzov et al. (2012) [156] proposed an improvement in the method by considering interactions conservation as an effort to remove the dependence on prior knowledge of the compounds.

In contrast, in [69], the score function is based on an atom-atom contact matrix, which means that their methodology only verifies an atom vicinity, and does not classify the contacts into noncovalent interactions. In [14], the authors proposed a score function based on footprints, which are interaction signatures whose profile corresponds to decompositions of electrostatic, steric, and hydrogen bonding interactions.

DiSCuS [242] is also an interesting tool that allows users to submit and compare different score functions, as well as perform analysis using combinations of scores. This tool also provides a filtering module for selecting compounds based on interaction fingerprints and a searching feature for finding compounds with a similar binding mode. DiSCus also provides a molecular viewer where users can analyze the 3D structure.

Finally, machine learning (ML) techniques, especially deep learning methods, have become increasingly prominent in the quest for active compound identification in recent years [44, 79, 104, 140, 141]. Some of the reasons that made it possible include the massive amount of biological data available nowadays and improvements in computational power thanks to graphics processing units (GPUs) [44, 141]. The applications of ML in drug discovery cover the task of predicting actives and inactive compounds, and binding affinity [51, 144, 160, 162, 219, 245].

1.5.3.3 Geometric analysis

An additional but widespread and fundamental strategy consists of a thorough analysis of the docking poses to select and filter hit molecules and distinguish poor poses. This procedure is useful because it involves the researcher's expertise and the use of the literature as a source of knowledge [131, 152]. In the latter case, if previous studies have already described critical residues to protein activity and which interactions are commonly established in its binding site, it is reasonable and wise to use such knowledge to select docked compounds based on the available information. Thus, tools presented in Section 1.3.2 are advantageous as they permit to calculation and to analyze protein-ligand interactions in docked structures.

On the other hand, such a "cherry-pick" procedure is remarkably toilsome and involves the manual analysis of 100-1000 top-ranked compounds through a meticulous inspection using molecular graphics programs. Furthermore, this filtering process relies on previous literature works that could not exist and user expertise.

In two recent works [99, 136], for instance, the authors declared to have selected the top hits by manual inspection. Therefore, the identification, prioritization, and automatic selection of a small number of promising compounds (hits) is still an open problem in VS field.

1.6 Motivation

Proteins are essential macromolecules to all organisms as a whole, and countless diseases are associated with their proper functioning. Not surprisingly, there is a particular interest in producing new drugs (ligands) able to modulate these macromolecules.

Accordingly, computational techniques like SBVS and molecular docking are powerful tools that contribute significantly to early-stage drug discovery.

A typical SBVS campaign consists of three major phases, namely data preparation, docking, and post-analysis. Commonly, a researcher starts with more than 20,000 compounds, and after running a protocol of docking, 100-1000 candidate molecules remain for post-analysis. The latter is an essential procedure since scoring functions have several drawbacks and non-ligands might be prioritized first over true ligands, which is not desirable. Thus, the final step in SBVS strategies is a thorough manual process of hit selection, in which binding modes of hundreds of top-scoring compounds are inspected in molecular graphics programs. In this hit selection process, researchers have the opportunity to incorporate previous knowledge of the system, such as prioritizing ligands which interact with key residues of the target protein.

In recent years, several semi-automatic works proposed new rescore functions as alternatives to docking scores in order to obtain superior enrichments. However, these tools present certain limitations as following described. Some of them: do not present user interactivity as users cannot provide their knowledge for the compound selection procedure; proposed a model to rank compounds, but unfortunately, the source code is not provided; rank compounds based on the similarity with the reference structure, or interactions conservation, but never both; do not take advantage of other protein-ligand complexes with similar binding sites available in the PDB; do not give many details about why a compound was ranked first, and it is up to the user to check for frequent interactions or interaction patterns.

Therefore, the identification, prioritization, and automatic selection of promising hits is still an open problem in VS field.

1.7 Objectives

1.7.1 General Objective

The primary objective of this work is to develop metrics, models, and algorithms for the identification, prioritization, and automatic selection of a small number of promising compounds (hits) in a structural-based virtual screening campaign.

1.7.2 Specific Objectives

Specific objectives are listed below:

- Address our major object through a descriptive and predictive perspective;
- Include new features in nAPOLI;
- Improve and expand the methods to calculate protein-ligand interactions;
- Develop an open-source library for molecular interaction analysis using the expanded methods;
- Propose a novel hashed interaction fingerprint;
- Evaluate the fingerprint parameters and their effect on the similarity between different complexes;
- Evaluate the applicability of the new fingerprint using machine learning techniques.

Chapter 2

Methods

In this section, we present the methods and algorithms proposed and implemented in this work, and describe the two perspectives we take to address the identification, prioritization, and automatic selection of a small number of promising compounds (hits) in a structural-based virtual screening campaign.

In the first aspect, we approach this subject more descriptively by providing ways to characterize and analyze protein-ligand interaction patterns across large data sets of protein-ligand complexes, as well as select and filter compounds through an interactive, visual, and analytical manner. In the second aspect, we address the problem from a predictive point of view.

This section is organized as follows. First, we discuss the descriptive aspect of this work and related methods. Then, we present novel algorithms and methods for calculating interactions, and we, finally, present the predictive aspect and its methods.

2.1 Descriptive aspect

As we mentioned in Section 1.3.2, we firstly proposed nAPOLI[1] (Analysis of PrOtein-Ligand Interactions) as a Master's thesis [84], and since then, we have improved its methods and included new functionalities. nAPOLI was conceived as a web server that brings together an automated analysis of conserved interactions across large data sets of protein-ligand complexes. Thus, the descriptive aspect of our work is especially attributed to nAPOLI, which comprises interactive visualizations and comprehensive reports of the interacting residues/atoms to explore and make sense of conserved noncovalent interactions that work as crucial factors in molecular recognition.

The improvements of nAPOLI's methods are presented in the following subsections, while novel functionalities are presented in Section 3. For a complete and detailed description of the methods refer to [83].

[1] http://bioinfo.dcc.ufmg.br/napoli/

2.1.1 Data set validation and PDB files filtering in nAPOLI

To avoid errors while running nAPOLI, we apply the following validation tests: remove *nAPOLI entries* that are not in the format presented at Section 3.1.1; remove entries whose chain was not found in the PDB file or does not have at least one of the 20 standard amino acids; remove entries whose *ligand name* and *ligand number* were not found in the PDB file. Moreover, to standardize all PDB files, we remove hydrogen atoms, keep only the first model when multiple models are available, and keep only atoms with the highest occupancy for residues with multiple conformations.

2.1.2 Physicochemical properties of atoms in nAPOLI

We classified atoms according to their physicochemical properties into one or more of the following types: *acceptor*, *aromatic*, *donor*, *hydrophobic*, *negative*, or *positive*. Atoms that do not match any type are called *unrated* atoms. These properties are assigned considering a neutral environment (pH 7).

Properties of residue atoms were manually predefined, while ligand atoms are classified automatically on the fly, i.e., during the processing of a user project. In the next subsections, we detail the rules considered to classify atoms in accordance with their physicochemical properties.

2.1.2.1 Residue atoms

Residue atoms were manually classified based on [24, 31, 161, 214, 218]. The classification of all residue atoms is shown in Table A.1.

There are three considerations to bear in mind regarding the rules presented in Table A.1: (i) all carbon atoms were labeled as *hydrophobic*, except those bound to a nitrogen or oxygen atom that remained *unrated* [218]; (ii) guanidine carbon of arginine was classified as *positive*; (iii) as histidine structure can be found in three forms depending on its protonation state and tautomeric form [146], its imidazole nitrogens were classified as *acceptor*, *donor* and *positive* to contemplate all these possibilities; (iv) finally, atoms that are not in the table remained *unrated*.

2.1.2.2 Ligand atoms

We developed an automatic classification method that consists of a preprocessing of the ligand file, and the assignment of labels to its atoms using Pmapper from ChemAxon[2].

In the preprocessing of ligand files, we first generate a new PDB file by extracting the target ligand atoms and any ligand covalently bound to them from the original file. Covalent bonds between ligands are frequently found in PDB files containing oligosaccharides that are usually represented by their small monosaccharides units. An illustrative example is found in PDB 2ZID, where a trisaccharide compound, known as isomaltriose, was represented as three glucose molecules. The impact of covalent bonds in the atoms classification can be illustrated by taking into account a hydroxyl group of glucose as an isolate species, in which its oxygens would be labeled as *acceptor* and *donor*. However, in isomaltriose example, oxygen is covalently bound to a carbon atom from another glucose molecule. In such circumstances, the oxygen is classified just as *acceptor* since it does not have any hydrogen to donate. Finally, hydrogen atoms are added to the extracted ligands and the obtained PDB file is then converted into Mol file format using Open Babel [177]. In this process, a neutral environment was considered (pH 7).

Following, Pmapper (PMapper 16.5.2, 2016) perceive physicochemical properties through a set of pharmacophore rules, which consists of an XML configuration file. ChemAxon provides two default configuration files: a calculation-based and a fragment-based one. In the former, pharmacophoric properties are obtained through chemical feature calculations as charges, partial charges or pK_as. While in the latter, such properties are acquired through the definition of functional groups.

As a first experiment, we classified the 20 most commonly found amino acids in living beings with each of these files. The objective was to evaluate which file was able to classify ligand atoms in higher agreement with our manual classification. As we compared amino acids as free ligands and residues, main chain nitrogens and oxygens had to be treated differently.

We observed that some atoms were correctly classified in one method but not in the other and vice versa. For example, the fragment-based file was not able to classify both oxygens from the carboxylate as being acceptor and negative. While the calculation-based file failed in classifying the amine and guanidine nitrogens from the arginine as positive. Therefore, we constructed a hybrid model joining the different strengths of both configuration files.

Afterwards, we evaluated and improved this new hybrid model by using different types of molecules. In order to evaluate a wide spectrum of functional groups and com-

[2]http://www.chemaxon.com

Table 2.1: List of ligand ids used to improve the hybrid model.

Ligand ids
004, 00G, 00S, 010, 01F, 020, 02Y, 03L, 03V, 055, 056, 069, 06B, 06U, 07M, 07Z, 087, 0A0, 0A9, 0AH, 0CK, 0E1, 0E5, 0EA, 0FR, 0KV, 0L1, 0N9, 0PY, 0QA, 0R1, 0U7, 0UA, 0UC, 0VD, 0WV, 10H, 11E, 11M, 12O, 13X, 14J, 172, 1AC, 1AN, 1DH, 1DJ, 1DU, 1DW, 1H2, 1H3, 1HP, 1JZ, 1KA, 1MH, 1U8, 1VL, 1XA, 23N, 2DI, 2UC, 2UE, 39R, 3BF, 412, 4AO, 4GI, 4NC, 4PO, 4SX, 6HN, 6PC, 7I2, 7NI, CDG, DMF, IMT, MAG, MIS, PCA, PYB, RIO, TEO, X01, X73

pounds, we searched PDB for ligands containing one or more of the following structures: alcohol, aldehyde, amide, amidine, amine, aminium ion, aniline, benzisoxazole, benzothiophene, carbonate, carbonyl, carboxylic acid, diazonium, ester, ether, furan, guanidine, hydroxamic acid, hydroxyl, imidazole, indazole, indole, ketone, lactam, nitrile, nitro, nitrone, nitroso, oxadiazole, phenol, phenyl, phosphoryl, piperazine, piperidine, purine, pyrazole, pyridine, pyrimidine, pyrrole, quinazoline, sulfanyl, sulfonamide, sulfonic acid, sulfonyl, thiazol, triazine, triazole. We obtained 85 compounds (Table 2.1) that were used in a cyclic process involving executing Pmapper, performing a rigorous manual analysis of the atom types of each ligand and refining the configuration file until it matched our quality criteria.

The pharmacophoric rules are defined by using SMARTS language and are presented in Section A.2.

2.1.3 Protein-ligand interactions in nAPOLI

In nAPOLI, we model protein-ligand interfaces as bipartite graphs, where nodes are atoms from the protein or ligand, and edges are the interactions among them. Nodes are labeled with physicochemical properties of atoms as explained in the previous subsection, while interactions are characterized as shown in Figure 2.1 and described in detail below.

First, we compute contacts at the atomic-level by using a cutoff-free and geometric approach called Delaunay tessellation (DT) (see Section 1.3.1), which in our work is performed by the CGAL library [40].

Next, atoms are classified according to their physicochemical properties. Finally, for each pair of atoms in contact, we define potential interactions by using physicochemical properties, distance, and angle criteria. nAPOLI identifies the following interactions: *aromatic stacking*, *hydrogen bond*, *hydrogen bond mediated by water*, *hydrophobic*, *attractive electrostatic*, and *repulsive electrostatic*. The default values are shown in Table 2.2,

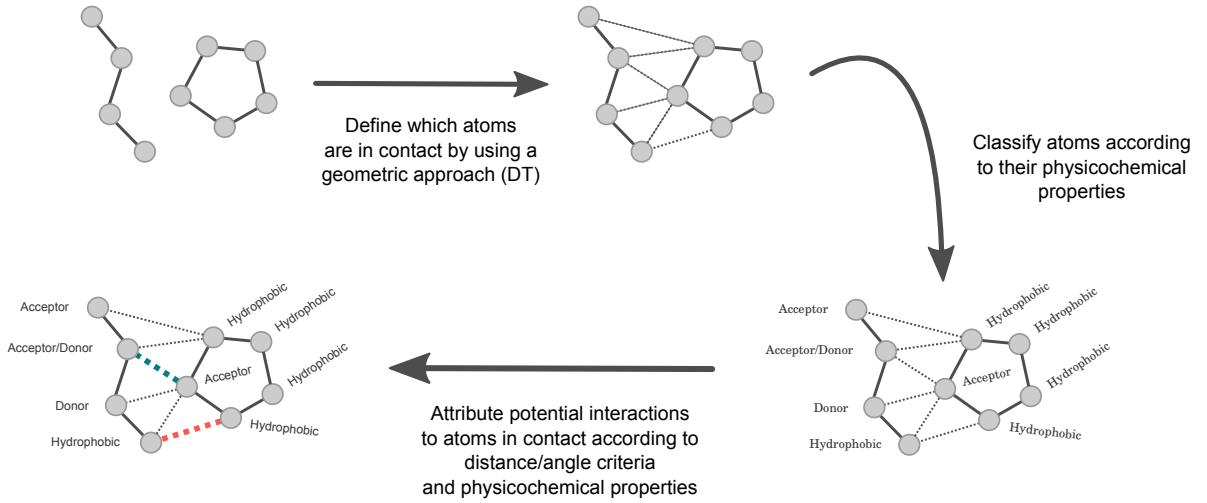


Figure 2.1: Protein-ligand interaction computation diagram. Straight lines are covalent interactions, gray dashed lines are contacts and thicker dashed lines (green/red) are interactions.

however, users can change them according to their needs.

Table 2.2: Default criteria to define interactions in nAPOLI.

Interaction type	Distance in Å	Angle in Degrees	Reference
Aromatic stacking	$2.0 \leq \overrightarrow{\|rr\|} \leq 4.0$		[157]
Hydrogen bond	$\overrightarrow{\|Ha\|} \leq 2.5$ and $\overrightarrow{\|da\|} \leq 3.9$	$d\hat{H}a \geq 120$	[13, 112, 161]
Hydrophobic	$2.0 \leq \overrightarrow{\|hh\|} \leq 4.5$		[157]
Repulsive electrostatic	$2.0 \leq \overrightarrow{\|sc\|} \leq 6.0$		[24, 155]
Attractive electrostatic	$2.0 \leq \overrightarrow{\|oc\|} \leq 6.0$		[24, 155]

Aromatic atom (r), hydrogen atom (H), acceptor atom (a), donor atom (d), hydrophobic atom (h), similarly charged atoms (sc), oppositely charged atoms (oc).
Lower limits equal to 2.0 Å avoids the inclusion of covalently bonded atoms [218].

For each nAPOLI entry, our tool only identifies interactions involving a ligand and residues from the same chain. If a user has informed the nAPOLI entry '3QL8:A:X01:300', all computed interactions will be between the residues and the ligand X01 from chain A.

Hydrogen bonds are computed using HBPlus [161]. To inform HBPlus which atoms are hydrogen bond acceptors or donors, we use the HBAdd software available with LIGPLOT [233]. After detecting all hydrogen bonds, nAPOLI filters out interactions that do not involve the target chain and the target ligand. It also removes all hydrogen bonds whose pair of involved atoms is not composed of an acceptor and a donor atom according to our method. Finally, nAPOLI searches for hydrogen bonds that are intermediated by water. After identifying such interactions, nAPOLI labels both protein-water and water-ligand hydrogen bonds as *hydrogen bond (water)*.

2.2 Descriptive case studies

As an effort to validate and illustrate the applicability of nAPOLI, especially the novel features presented in Section 3.1, we used two datasets related to ricin and human nuclear receptor subfamily 3 (hNR3). Ricin is a type 2 ribosome-inactivating protein (RIP) found in castor beans. Due to its high toxicity and ease of production [45], as well as its promising application, for instance, in immunotoxin treatment [248], ricin is an important and interesting target to be analyzed. Finally, hNR3 belongs to a family of ligand-regulated transcription factors that play crucial roles in many physiological processes. Thus, given its therapeutic importance, we chose the hNR3 as an example to show the applicability of nAPOLI in the study of conserved protein-ligand interactions along a functionally conserved binding site of a protein family. The two data sets employed in our analysis are available in the nAPOLI web server as examples in the *Data set analysis* section:

- Ricin data set: 26 complexes obtained from a literature review;
- Human nuclear receptor subfamily 3 data set: 198 complexes comprising 6 different proteins from hNR3.

In both case studies, we used the default configuration to detect interactions. For additional case studies demonstrating the applicability of nAPOLI refer to [83].

2.3 Improvement and expansion of methods for calculating interactions

In previous sections, we briefly introduced nAPOLI (Section 1.3.3) and presented some of its methods (Section 2.1). Also, we discussed how nAPOLI was specially conceived for the analysis of protein-ligand interactions in a descriptive manner.

Nonetheless, some aspects of nAPOLI took us to redesign our models and propose a new library to be accessible to more people, generic, customizable, and completely open-source. They are:

- *Protein-ligand interaction at the atomic-level*: in nAPOLI, protein-ligand interactions are modeled at the atomic-level for all interactions, however, as discussed in

Section 1.3.1.1, hybrid models better represent interactions that are established due to the contribution of multiple atoms;

- *Protein-ligands complexes only*: nAPOLI in its current version works only with protein-ligand interactions; thus, a generic tool able to deal with any molecular complex is more promising;
- *Proprietary license*: some third-party software we use, namely Pmapper, GenerateMD, and Ward, require a Chemaxon license. Although for some purposes the licenses are free, we believe a completely free-of-license tool can reach a higher number of users;
- *System developed in Perl*: nowadays, Python is one the most used languages and, not surprisingly, a large number of libraries for biological and scientific purposes are written in this language. Herein, we mainly highlight the data science libraries such as Pytorch, Keras, Tensorflow, scikit-learn, Skorch, NumPy, Pandas, Seaborn, RDKit, and others. Therefore, we believe Python offers more advantages than Perl in this context.

Taking the items above as our touchstones, we propose LUNA[3], a new Python library completely based on open-source code. The library and its functionalities are thoroughly presented in Section 3.2, while its methods are presented in the following subsections.

It is noteworthy that, although we opted for developing a new library from scratch, we envision nAPOLI and LUNA working together, where the descriptive aspect is provided by nAPOLI, while LUNA is responsible for calculating interactions and generating the data that will feed nAPOLI, as well as providing the predictive perspective for the prioritization of compounds in an SBVS campaign.

2.3.1 Dataset validation and PDB files filtering

We apply the following validation tests:

- Remove entries that are not a valid format, which is defined as follows:
  - PDB id (4 characters) or a filename: mandatory;
  - Model number (1 character): optional and required only for structures containing more than one model. By default, the first model is used;

[3]https://github.com/keiserlab/LUNA

  - Chain (1 character): mandatory. Note that if only the chain information is provided, it will calculate interactions considering the whole chain;
  - Compound name (1-3 characters): optional and required only for computing interactions involving a specific compound (residue, ligand, nucleotide, or any other molecule);
  - Compound number (valid integer): optional and required only for computing interactions involving a specific compound (residue, ligand, nucleotide, or any other molecule);
  - Insertion code (1 character): optional and required only for computing interactions involving a specific compound (residue, ligand, nucleotide, or any other molecule).

- Remove entries whose chain was not found in the PDB file;
- Remove entries whose compound was not found in the PDB file (if defined);

Moreover, to standardize all PDB files containing compounds with multiple conformations, we keep only atoms defined as the first occupancy flag, which is usually 'A' or '1'.

2.3.2 Physicochemical properties of atoms and atom groups

The recognition between two molecules is a crucial and challenging process that depends on several variables, such as the polarity and electronegativity of atoms and functional groups, solvation, hydrophobicity, environmental pH, and charge.

Based on these chemical characteristics, and departing from our previous work [83] and a thorough revision of the literature [4, 20, 22, 25, 34, 35, 50, 56, 66, 94, 95, 115, 120, 121, 174, 135, 146, 154, 181, 195, 207, 224, 227, 243, 252], we classify atoms and groups of atoms according to their physicochemical properties into one or more of the following types: *acceptor*, *amide*, *aromatic*, *atom*, *chalcogen donor*, *donor*, *electrophile*, *halogen acceptor*, *halogen donor*, *hydrophobe*, *hydrophobic*, *metal*, *negatively ionizable*, *nucleophile*, *positively ionizable*, *weak acceptor*, and *weak donor*. In LUNA, an atom group can represent both chemical functional groups or simply an arrangement of atoms as in *hydrophobes*. The latter is an optional property that represents a group of hydrophobic atoms, which better mimics how the hydrophobic effect occurs, i.e., a favorable contact between two hydrophobic surfaces. It is also important to highlight that although atoms may belong to a group, they all have their own physicochemical properties.

The terms *negatively ionizable* and *positively ionizable* were chosen to indicate that a specific atom or group of atoms may be ionized. However, one can set a different pH in order to alter the resultant classification of an atom or group.

Regarding the property *amide*, it identifies amide groups and is employed during the calculation of amide-stackings. In its turn, the feature *atom* simply identifies a heavy atom and is used to calculate atom-atom interactions, such as *covalent bond* and *van der Waals* interactions.

All these physicochemical properties are identified on the fly, i.e., during the processing of a user project. The only exception happens for protein residues to which a precomputed list of properties considering the default pH (7.0) is already available to reduce computational processing. By default, this list contains only the 20 standard amino acids and water molecules, but it can be expanded as necessary or even disabled whether a different pH is to be considered.

In the next subsections, we detail the algorithm to classify atoms in accordance with their physicochemical properties.

2.3.2.1 Physicochemical feature assignment

The feature assignment is performed on the fly for all molecules within a certain distance (in Å) of the defined target, which can be a chain, a compound, or a list of compounds. Thus, the target and the recovered molecules around it define the binding site scope for the interaction analysis. For each one of these compounds, LUNA applies three main procedures.

First, it verifies if the current compound already has a precomputed property available in an internal configuration file. If so, the tool will use this information to reduce computational processing and the property perception is successfully finalized to the current compound. Otherwise, it verifies whether the current compound contains a defined molecule file, which is usually the case for docking campaigns since the ligand pose may be available as a separate molecular file. In this particular case, the third step is promptly initialized.

Secondly, if neither situations described in the first procedure occur, the tool will identify all molecules covalently bound to the current compound and convert their structures from PDB to Mol format using Open Babel [177]. This conversion is important because the PDB format does not contain chemical information like atom charge or aromaticity, which is crucial for a proper physicochemical property perception. Moreover, it is wise to convert the current compound with its bound neighbors in order to keep correct

bond orders. During this conversion, if necessary, hydrogens can also be added to the molecules according to the specified pH.

There is another optional step that consists of validating the converted molecules and amending simple problems related to valence and charge. These problems sometimes may occur with the current versions of Open Babel when PDB structures are converted to Mol. It may happen precisely because the PDB format does not have sufficient chemical information, which can induce Open Babel to incorrectly perceive the bond order, aromaticity, valence, or charge of an atom. Our implemented solutions cover simple problems that are more recurrent during these conversions. Atom charges are amended only when they do not match the expected charge according to our implementation of OpenEye's charge model[4]. Valences are amended only for ammonium nitrogen whose structure was not previously ionized. In this case, Open Babel may perceive such atoms as hypervalent and attribute an incorrect valence to nitrogen.

Finally, LUNA perceives physicochemical properties through a set of chemical rules specified as a SMARTS-based language string and stored in a feature definition file format (FDef) as in RDKit [195]. Our rules comprise both atoms and groups of atoms, which, as mentioned before, can represent a functional group or an arrangement of atoms. Importantly, tautomeric forms of chemical groups were also envisioned in our rules as a means to account for the biological environment dynamics. The complete set of rules and geometrical models are defined in Sections B and C.

Another optional step can be applied after the third procedure and consists of grouping hydrophobic atoms to form a hydrophobe group. See Section C.13 for more information.

2.3.3 Molecular interactions calculation

For each pair of atoms/group of atoms, molecular interactions are characterized using physicochemical properties, distance, and angle criteria. In the context of atom groups, the centroid of the group is used in the geometrical analysis.

LUNA identifies the following interactions: *amide-aromatic stacking*, *anion-electrophile*, *antiparallel multipolar*, *cation-nucleophile*, *cation-pi*, *chalcogen bond*, *chalcogen-pi*, *covalent bond*, *displaced face-to-edge pi-stacking*, *displaced face-to-face pi-stacking*, *displaced face-to-slope pi-stacking*, *edge-to-edge pi-stacking*, *edge-to-face pi-stacking*, *edge-to-slope pi-stacking*, *face-to-edge pi-stacking*, *face-to-face pi-stacking*, *face-to-slope pi-stacking*, *halogen bond*, *halogen-pi*, *hydrogen bond*, *hydrophobic*, *ionic*, *multipolar*, *orthogo-*

[4]https://docs.eyesopen.com/toolkits/python/oechemtk/valence.html

nal multipolar, *parallel multipolar*, *pi-stacking*, *repulsive*, *salt bridge*, *tilted multipolar*, *unfavorable anion-nucleophile*, *unfavorable cation-electrophile*, *unfavorable electrophile-electrophile*, *unfavorable nucleophile-nucleophile*, *van der Waals*, *water-bridged hydrogen bond*, and *weak hydrogen bond*.

In addition to these interactions, we provide more three contacts: *atom overlap*, *proximal*, and *van der Waals clash*. The former contact identifies artifacts generated by low-resolution structures and homology models, which consist of an unnatural overlap of two atoms. *Van der Waals clash* characterizes repulsion between two atoms when they become too close, and *proximal* is an optional contact which simply indicates that two atoms are close to each other by a specific threshold.

The combination of chemical features, distance, angle criteria, and geometrical models utilized to calculate these interactions are presented in Section C. It is important to highlight that although LUNA implements its own methods and criteria to calculate interactions, users can also define their own functions and cutoffs. That is possible thanks to the object-oriented style employed in our library.

2.3.4 Interaction fingerprint

In this work, we also propose a novel hashed interaction fingerprint (IFP) called FIFP (Functional InteracTion FingerPrint), inspired by ECFP [203], FCFP [203], and E3FP [9]. Our fingerprint is able to encode the binding site interactions both as binary or count fingerprint. Besides it, FIFP can encode contacts, interactions, or both, and is compatible with RDKit.

Moreover, one of the most promising features of our approach is its interpretability. Different from other hashed fingerprints that are usually black-boxes, in which one has to design its own methods to interpret what each bit represents, ours already provides several features to make the analysis straightforward and out-of-the-box (see Section 3.2.4).

On the whole, we believe FIFP to be a promising approach for structure-based virtual screening and molecular dynamics, where thousands of compounds and poses can be promptly filtered or clustered according to their interaction similarities. Not to mention the possibility of using FIFP in a machine learning context, where a data set of known complexes could be used to train a model for predicting and selecting unknown compounds given a specific goal. In Sections 2.4.2 and 3.4, we present a case study illustrating the applicability of FIFP in the mentioned context.

As ECFP and E3FP, FIFP depends on three parameters (Figure 2.2): the fingerprint length, the radius growth rate, and the number of levels. The fingerprint length, as

discussed in Section 1.4.2, controls how many features at maximum can be represented in the fingerprint. In its turn, the radius growth rate and the number of levels indirectly control how many features will be included in the fingerprint. In the following subsections, we discuss how FIFP is generated and how these two parameters influence feature discovery.

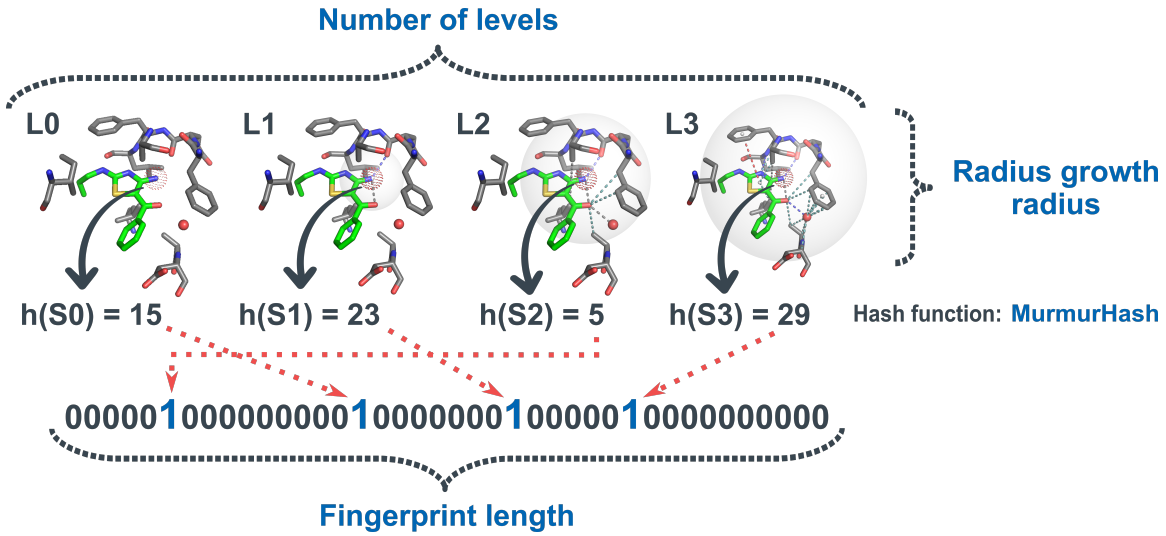


Figure 2.2: Parameters to control the FIFP creation: the fingerprint length, the radius growth rate, and the number of levels.

2.3.4.1 Generating initial identifiers

At iteration 0, initial identifiers are assigned to each atom or atom group according to a pharmacophore-based approach. In this method, only chemical features are considered during the identifier generation. For atoms, it comprehends both atom features and those inherited from the groups to which the atom belongs. For example, an aromatic carbon could have the features *Hydrophobic* and *Weak donor* as its own properties, as well as the property *Aromatic* inherited from the ring (group) that comprises this atom. In its turn, carboxylic oxygen could be considered an *Acceptor* and *Negatively ionizable*, the latter being inherited from the carboxylic group it belongs to. For atom groups, the FIFP encodes only the chemical features of the group. Therefore, this method is useful because it keeps a level of abstraction when some groups should be recognized as functionally equivalent.

After characterizing each atom (or atom group) according to their chemical information, FIFP applies a hashing function to this information to obtain their initial identifiers (a 32-bit integer). In LUNA, the hashing function we use is a Python imple-

mentation[5] of MurmurHash3 [6]. However, any hashing function can be applied as long as it generates uniform and random identifiers.

2.3.4.2 Subsequent identifiers update

After generating each initial identifier, the algorithm subsequently updates them through an iterative process that continues until it converges or it reaches the maximum number of levels. At each iteration, a sphere of size $R*L$, where R is the *radius growth* and L the current level (iteration number), is centered at each atom (atom group) and their neighborhood is characterized by capturing all interactions within the shell. A hashing function is then applied to the neighborhood, and a new identifier to the central atom (atom group) is generated.

A single iteration for a given atom or atom group is performed as follows. First, LUNA centers a sphere of size $R * L$ in the atom (atom group) and initializes an array of tuples with a pair consisting of the current level number and the identifier of the central atom (atom group) in the previous iteration. Next, it captures all atoms (atom groups) inside the shell and verifies if these entities establish any interaction with atoms/groups from the previous iteration. Note that at level 1, the list of atoms and groups from the previous iteration only contains the central atom/group. Thus, the only valid interactions in this iteration are the ones between the central atom/group and the newly discovered atoms (atom groups). Then, for each valid interaction, LUNA generates a tuple containing the interaction type and the identifier of the new atoms (atom groups) in the last iteration. These pairs are sorted and included in the array of tuples. Note that sorting the list is essential for avoiding dependence on the order of its elements. Otherwise, fingerprint generation would not be deterministic.

Finally, LUNA generates a new 32-bit integer identifier to the central atom (atom group) by hashing the sorted list. Any atom or group interacting with the central entity become part of its neighborhood and will be taken into consideration in the following levels.

This iterative process continues for each atom (atom group) until it reaches the maximum number of levels or when the algorithm converges, which happens when the shells cannot be expanded anymore. In other words, if all interactions involving the atoms (atom groups) inside each shell are already included in the last encoded neighborhood, then convergence is reached.

[5] https://pypi.org/project/mmh3/

2.4 Predictive aspect

In the predictive perspective of this work, we envision LUNA and FIFP as promising tools for the identification, prioritization, and automatic selection of compounds in a structural-based virtual screening campaign. To do so, we performed two series of experiments with different goals.

First, we evaluated the fingerprint parameters as a means to understand how they influence feature extraction and how the fingerprint could be used for selecting and filtering compounds based on similarity. While the second experiment consisted of a case study where we applied FIFP on the task of reproducing docking scores.

2.4.1 Fingerprint parametrization

As we mentioned in Section 2.3.4, FIFP depends on three parameters: the fingerprint length, the radius growth rate, and the number of levels. Together, these parameters control how features are extracted from a molecular complex and encoded into a binary or count fingerprint. Well in advance, we emphasize that, although a default combination of parameters works for most cases, some tasks or data sets may require a new parameterization for better results.

To find the best combination of parameters and to evaluate how the fingerprint behaves when varying each parameter, we analyzed their influence on the similarity of two complexes involving the same target and ligands with similar poses. If the ligands are similar and have a similar pose, it is expected they also present a similar binding mode, i.e., the interactions established with the protein would also be similar.

Bearing this in mind, we built three data sets composed by similar ligands related to the human cyclin-dependent kinase 2 (CDK2), which is a well-studied enzyme involved in cell cycle progression.

For the first data set, we manually generated different poses for the same CDK2-ligand complex (PDB 3QQF, ligand X07) by performing small transpositions on the ligand or rotating its bonds (Figure 2.3). On the whole, twelve manual poses were generated, where poses C and L represent our positive and negative references. We chose these two poses as our references because we expect that slight (pose C) and drastic (pose L) modifications to produce the highest and smallest similarities, respectively. We are aware that such manual modifications in the ligand structure could generate invalid poses, unfavorable interactions, and even clashes with protein atoms. However, herein, our goal

is only to evaluate the similarity between the fingerprints when slightly different poses are compared. In any case, these unfavorable interactions are already taken into account when calculating the molecular interactions, and, therefore, they will appear as features of the complex.

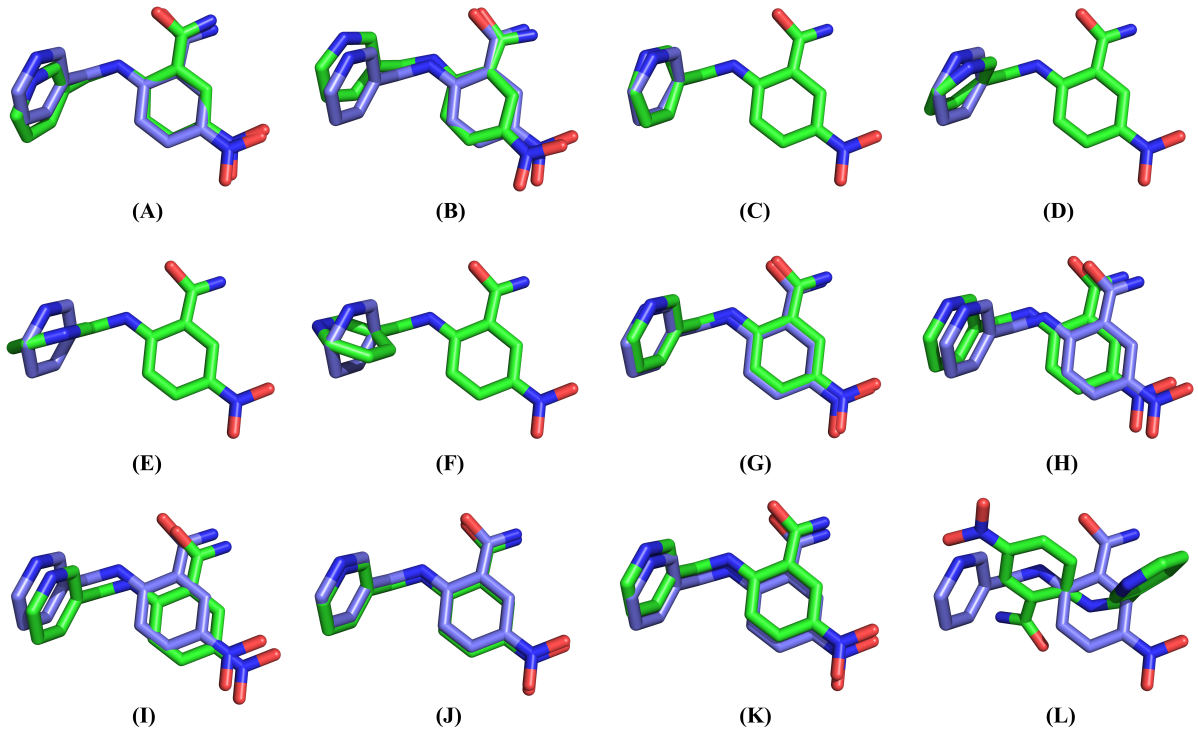


Figure 2.3: Manual poses obtained by rotating and transposing the ligand crystal pose (X07) in the CDK2 binding site (PDB 3QQF). The original pose and manually obtained poses are shown as blue and green sticks, respectively. The baseline pose is shown in the lower right corner.

For the second data set, we automatically generated a series of conformers for the ligand X02 in a complex with CDK2. To do so, we used the function EmbedMultipleConfs from RDKit with the parameter numConfs and pruneRmsThresh set to 10,000 and 0.1, respectively. The first parameter defines the number of conformers the algorithm should generate, while the second removes conformers whose distance (RMSD) to other conformers are less than 0.1. This pruning procedure is greedy, which means that the first conformation generated is retained and from then on only those that are at least pruneRmsThresh away from all retained conformations are kept. After generating the conformers, we aligned them to the ligand pose in complex with CDK2 (PDB 3QQK) and measured their distance. If the distance between the conformer and the crystal pose was less than 0.4, we retained the conformer; otherwise, we removed it. After this last pruning, we obtained 181 conformers similar to the crystal pose, which are shown in Figure 2.4.

Finally, the third data set was obtained from Schonbrunn et al. (2013) [213]. In

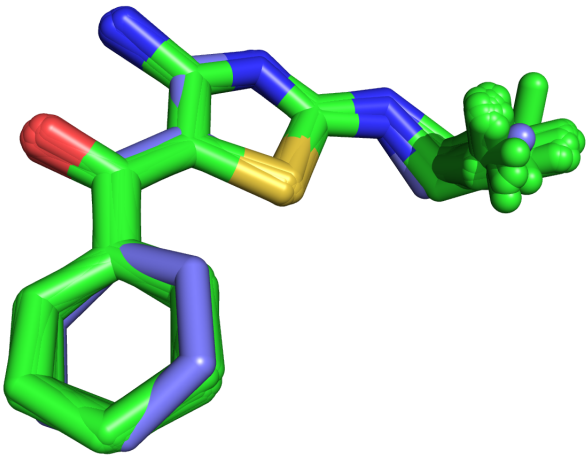


Figure 2.4: Automatically generated conformers for the ligand X02 in complex with CDK2 (PDB 3QQK). The original pose and the conformers are shown as blue and green sticks, respectively.

their work, the authors proposed 95 analogs by systematically modifying the flanking allyl and phenyl moieties of the compound 2-(allylamino)-4-aminothiazol-5-yl(phenyl) methanone (PDB id X02). Structures of 36 CDK2-ligand complexes were solved through X-ray crystallography. We also found in the PDB, an additional 38 related structures that were deposited by the authors, totalizing 74 complexes in our data set. The interesting characteristic of this data set for our study is the similarity between the ligands from a chemical perspective and their crystal poses are also similar.

In all experiments, the similarity between the fingerprints was measured using the Tanimoto coefficient (Equation 1.1), and the protein-ligand interactions were calculated with LUNA using the default parameters.

2.4.2 Predictive case study

As an effort to validate and illustrate the applicability of FIFP for prioritizing HIT compounds, we chose a large data set recently published by Lyu et al. (2019) [153]. This huge library consists of 138 million molecules docked against Dopamine D4, an important G protein-coupled receptor (GPCR) superfamily member involved in many different roles in the central nervous system. Given its relevance as a neurotransmitter, dysregulation of the Dopamine D4 signaling cascade is linked to several pathological disorders like Parkinson's disease and schizophrenia [176], which makes it an interesting target to be analyzed.

With such a large data set, FIFP could be applied to three different scenarios:

classifying molecules as active or inactive, identifying bad poses, and predicting docking score or even experimental binding affinity. Herein, we chose the third task because we are interested in evaluating if a model trained with our IFPs would be able to reproduce the Dock scores. In the future, we plan to optimize the model for predicting experimental binding affinities, which is usually referred to as fine-tuning technique and whose most straightforward application is the rescore of docking poses (see Section 1.5.3.2).

Aiming to reach this goal and starting from the whole data set, we clustered the ligands docked against Dopamine D4 based on their chemotypes. Then, to create a balanced data set, we sampled the clusters and obtained a subset composed of 86,641 samples. The hypothesis we wanted to evaluate when we decided to obtain a subset of the whole library is whether we would be able to reproduce the Dock score even with a smaller data set.

To do so, we generated several fingerprints using different combinations of parameters, namely, the fingerprint length, the radius growth rate, and the number of levels. Besides these parameters, we also evaluated variations in the methods to calculate interactions and how they impact predictive performance. To do so, we empirically explored combinations of the following options: strict or loose rules for hydrogen bond donor (*Strict H rule*); protein structure with or without hydrogens (*Struct w/ H*, pH 7.4); include or not non-covalent (excluding van der Waals) interactions (*Non-cov*); compute or not atom-atom interactions (*Atom-Atom*), which include *covalent bond*, *van der Waals*, *van der Waals clash*, and *atom overlap*; include or not proximal interactions (*Proximal*). The presence of the described labels in the experiment name (X-axis) indicates whether it was used or not during the calculation of interactions. For the best models, we also evaluated if its count fingerprint version and the inclusion of interactions in the protein side (*w/ PPI*) would improve the prediction.

As our baseline models, we chose the RDKit implementations of ECFP and FCFP. In addition to the baseline model, we also selected two other interaction fingerprints for further comparison: SILIRID [48] and PLEC [245], both available in ODDT [244]. All fingerprints were generated using their default parameters.

Three different machine learning techniques were used: Deep Neural Network (DNN), Random Forest, and XGBoost [47]. During model development, we split the data into train, validation, and test, where the proportion of each set was 60%, 20%, and 20%, respectively. The predicted and real Dock scores were evaluated in terms of the R-squared (R^2) metric.

For the random forest and XGBoost, we used their scikit-learn implementation with default parameters, except for the number of estimators that were set to 300.

Finally, the DNN models were trained using the Pytorch implementation coupled with Skorch, which is a Python library that makes it possible to use Pytorch and scikit-learn together. Besides the several helpful methods available in Skorch, one of the main

advantages of using this library is the possibility of using the methods for randomized and grid searches from scikit-learn. Thus, using the mentioned libraries, and for each fingerprint, ten different hyper-parameter combinations were sampled from the specified distributions using a randomized search (RandomizedSearchCV from scikit-learn) with five cross-validations (CV), totalizing 50 trained models. In all models, we used the Adam optimizer [129] and MSELoss criterion, which are available in Pytorch. The maximum number of epochs was set to 400; however, we allow the training process to stop earlier whether the model does not improve after six epochs (EarlyStopping class from Skorch). The best model for each hyper-parameter optimization is chosen by the average R^2 value on the five cross-validations. From the top models, we also evaluated if the performance could be improved with RAdam optimizer [149], which is a variant of Adam that was shown to outperform it in some scenarios.

Finally, we compared FIFP to the other selected fingerprints using 5-fold cross-validation. At each iteration, 80% of the full data set was used for training and the remaining for testing. Mean and standard deviation were then calculated for the R^2.

Chapter 3

Results and discussion

This section is organized as follows. First, we present the novel features included in nAPOLI, followed by the new Python library. Then, we present the experiments performed with FIFP to assess how the parameters influence the fingerprint. Finally, we present a discussion about the case study where we apply FIFP in the task of predicting Dock scores.

3.1 Novel features in nAPOLI

In this section, we describe the novel functionalities introduced in the current version of nAPOLI [83].

3.1.1 Submitting new projects

Now, users have three manners to start a new project and submit a data set composed of protein-ligand complexes, which must be in the format '<PDB id or filename (4 characters)>:<chain (1 character)>:<ligand name (1-3 characters)>:<ligand number (valid integer)>'. From now on, this representation will be referred to as a *nAPOLI entry*.

The first option is when one has a predefined list of *nAPOLI entries* and knows which complexes to analyze. In such cases, the PDB files available at the PDB are used. In the second, users can submit their own PDB files and provide a predefined list of *nAPOLI entries*. This functionality has a wide spectrum of applications in docking, virtual screening, and molecular dynamics, which are techniques that produce a huge number of structures. Finally, in the last option, which was already available in nAPOLI during the Master's thesis, users can compose their own data set in an exploratory way.

Sometimes, one has a protein of interest and wishes to build a new data set consisting of proteins containing certain similarities to this protein and whose structures are complexed with some ligand. nAPOLI performs a sequence similarity search and users can inform a *PDB id* and a *chain* of interest as well as a *sequence identity cutoff*.

After retrieving the structures, users can define three additional parameters:

- *Neighborhood cutoff*: radius to search for ligands belonging to the same protein region. The default value was empirically defined as being 5 Å.
- *Minimum number of atoms*: minimum number of heavy atoms that a ligand should have so that it is not considered as a crystallography artifact. Default is 7 atoms [185].
- *List of crystallography artifacts*: ligands considered crystallography artifacts (Table 3.1).

Table 3.1: Default list of ligand ids considered crystallography artifacts [29, 223].

Ligand ids
ACE, ACT, BME, CSD, CSW, EDO, FMT, GOL, MSE, NAG, NO3, PO4, SGM, SO4, TPO

Next, nAPOLI searches for regions containing ligands and different binding sites. As a result, a list of *nAPOLI entries* for each region is returned to users, who can then choose a list for which nAPOLI will compute interactions. Users can also remove specific ligands or remove crystallography artifacts.

3.1.2 Processing log

In *Processing log* users can access a report where any processing errors are informed. Sometimes, an error can occur when processing a particular protein-ligand complex. In this circumstance, nAPOLI stores the error message and removes the complex from the next processing steps.

3.1.3 Clusters comparison for interacting residues

The *Clusters comparison* functionality allows users to discover which residues interact exclusively in only one cluster or are common to two or more *ad hoc* combination clusters which are suitable, for instance, to propose potential residues to mutate. This comparison feature was included as part of the section *Interactions by residues* in nAPOLI (see Section 1.3.3.2).

3.1.4 Ligands filtering

This feature was conceived especially aiming at structure-based virtual screening campaigns, where the final step is typically a manual process of hit selection, in which binding modes of hundreds of top-scoring compounds are inspected in molecular graphics programs. In this hit selection process, researchers have the opportunity to incorporate previous knowledge of the system, such as prioritizing ligands that interact with key residues of the target protein.

Bearing this in mind, we built *Ligands filtering* where users can filter, in a fast and automatic way, all ligands that fit into a chosen interaction combination. Two filtering options are available: *Residues combination filtering* and *Interaction types filtering*. In the first option, a list of residue combinations is generated from a chosen list of residues. For each combination, nAPOLI searches for all ligands that interact with all residues. In the second option, users inform of a list of residues and an interaction type for each residue. nAPOLI returns all ligands that interact with the defined residues through the specified interactions.

3.1.5 Applicability of nAPOLI on two different scenarios

In this section, we present two study cases with the objective to demonstrate the applicability of nAPOLI in different scenarios and how the novel features can be employed. First, we provide a discussion on a study case involving ricin, a type 2 ribosome-inactivating protein (RIP). Then, we employ nAPOLI in a scenario comprising six different proteins from the human nuclear receptor subfamily 3 (hNR3). The corresponding data

sets are available in nAPOLI web server as examples in the *Data set analysis* section.

3.1.5.1 Ricin data set

Earlier investigations based on site-directed mutagenesis, enzymatic studies and the binding of substrate analogs to the RTA have revealed key conserved residues involved in the depurination of ribosomal RNA, namely TYR80, TYR123, GLU177, ARG180 and TRP211 [41, 58, 87, 111, 124, 127, 128, 168, 169, 170, 196, 198, 204, 205, 211, 235, 239]. Also, these studies showed that the RTA active site is highly polar and surrounded by aromatic residues. Not surprisingly, X-ray crystal structural analysis revealed essential hydrogen bonds and aromatic stackings to the catalytic process to occur.

Based on this information, several small molecules presenting a potential binding mode similar to the adenosine were evaluated against RTA [11, 38, 58, 71, 106, 167, 170, 189, 190, 206, 239, 240, 248]. From the tested compounds, 26 complexes (18 inhibitors and 8 substrate compounds) were solved and deposited in the PDB (see Table 3.2). We then submitted these complexes to nAPOLI to further analysis through the *Insert nAPOLI entries* functionality, and the results are available as an example in nAPOLI.

Analysis of the whole ricin data set revealed that the ligands are predominantly composed of polar and aromatic atoms (see Figure 3.1), which is in agreement with the literature [232]. Few of them presented negatively or positively charged atoms, while hydrophobic atoms are mostly part of aromatic rings. One can also observe that several ligands present no hydrophobic atoms. It occurred because nAPOLI does not consider carbons that are bound to polar atoms as hydrophobic (see two examples in Figure 3.2). As expected, hydrogen bonds and aromatic stackings were prevalent in the database (see Figure 3.3).

Regarding clustering, the compounds were automatically partitioned into five groups (see Table 3.2) in a similar manner as in [232]. According to the authors, these molecules can be divided into four main categories: adenine-based substrate analogues, guanine-based, pterin-based, and pyrimidine-based inhibitors. In our approach, adenine-based compounds were subdivided into a group composed of adenine and amide-based compounds (cluster 2), and a group comprising molecules similar to adenosine monophosphate (cluster 3). Guanine-based compounds are part of cluster 1, though it also includes two pterin-based compounds (PT1 and NEO). Finally, clusters 4 and 5 are formed by pyrimidine-based and pterin-base compounds, respectively. It is noteworthy that Wahome et al. grouped the compounds according to a common substructure, while our method grouped molecules based on global similarity using chemical hashed fingerprints.

Table 3.2: RTA-ligand complexes whose structures were solved and deposited in the PDB. The column *Compound type* distinguishes inhibitors from substrate molecules.

PDB	Ligand	Compound type	Cluster	Reference
1IL3	7DG	Inhibitor	1	[11]
1IL4	9DG	Inhibitor	1	[189]
1IL9	MOG	Inhibitor	1	[190]
1BR6	PT1	Inhibitor	1	[206]
1BR5	NEO	Inhibitor	1	[206]
1IFS	ADE	Substrate	2	[38]
2PJO	NMU	Substrate	2	[167]
2P8N	ADE	Substrate	2	[167]
2R3D	ACM	Substrate	2	[167]
2R2X	URE	Substrate	2	[248]
1IFU	FMC	Substrate	3	[38]
3RTI	FMP	Substrate	3	[38]
1OBT	AMP	Substrate	3	[248]
3HIO	C2X	Inhibitor	3	[240]
4Q2V	0XE	Inhibitor	4	[58]
1IL5	DDP	Inhibitor	4	[189]
3EJ5	EJ5	Inhibitor	4	[240]
3PX8	JP2	Inhibitor	5	[38]
4HUP	19M	Inhibitor	5	[71]
4HV3	19L	Inhibitor	5	[106]
4HUO	RS8	Inhibitor	5	[167]
4ESI	0RB	Inhibitor	5	[170]
4MX5	5MX	Inhibitor	5	[206]
4MX1	1MX	Inhibitor	5	[206]
4HV7	19J	Inhibitor	5	[239]
3PX9	JP3	Inhibitor	5	[239]

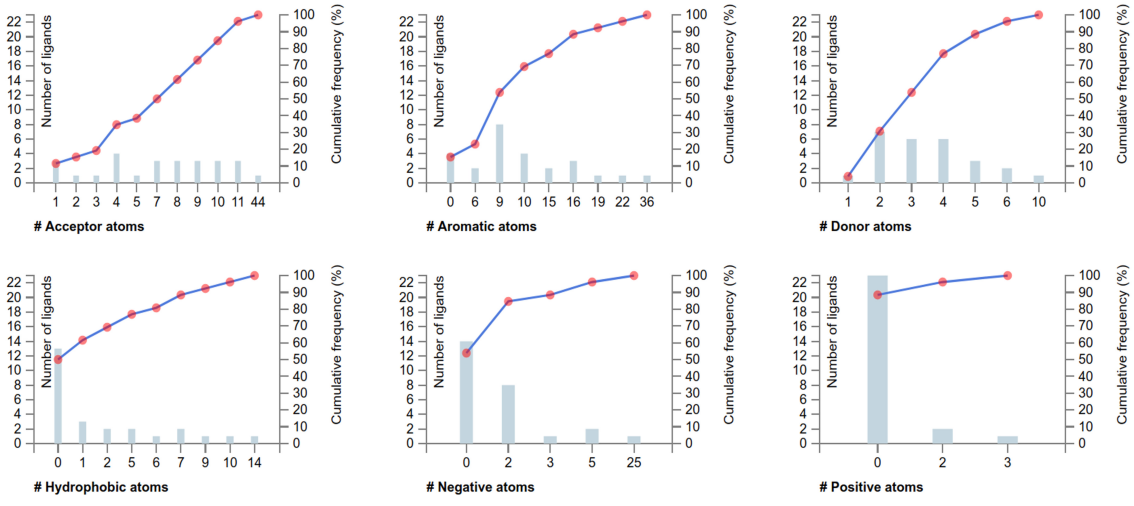
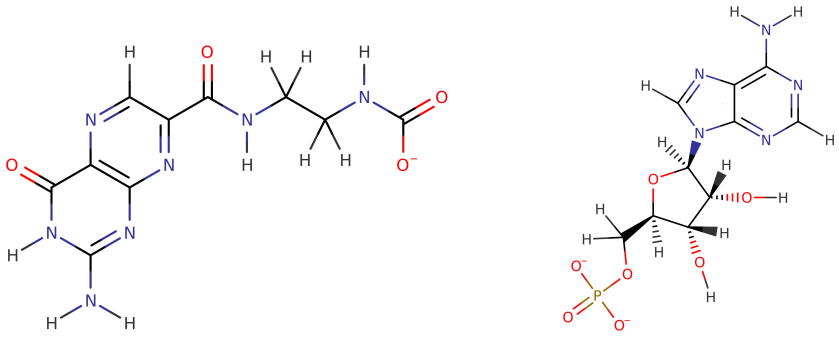


Figure 3.1: Atom type frequencies for the 26 RTA ligands. The X-axis shows the number of atoms, while the Y-axis shows the number of ligands that have a certain amount of atoms of a given type.

Consequently and not surprisingly, these two approaches can generate different cluster results [194].



(a) (b)

Figure 3.2: Examples of two ligands containing no hydrophobic atoms according to nAPOLI's method. The chemical structure from 5MX and AMP are shown in (a) and (b), respectively.

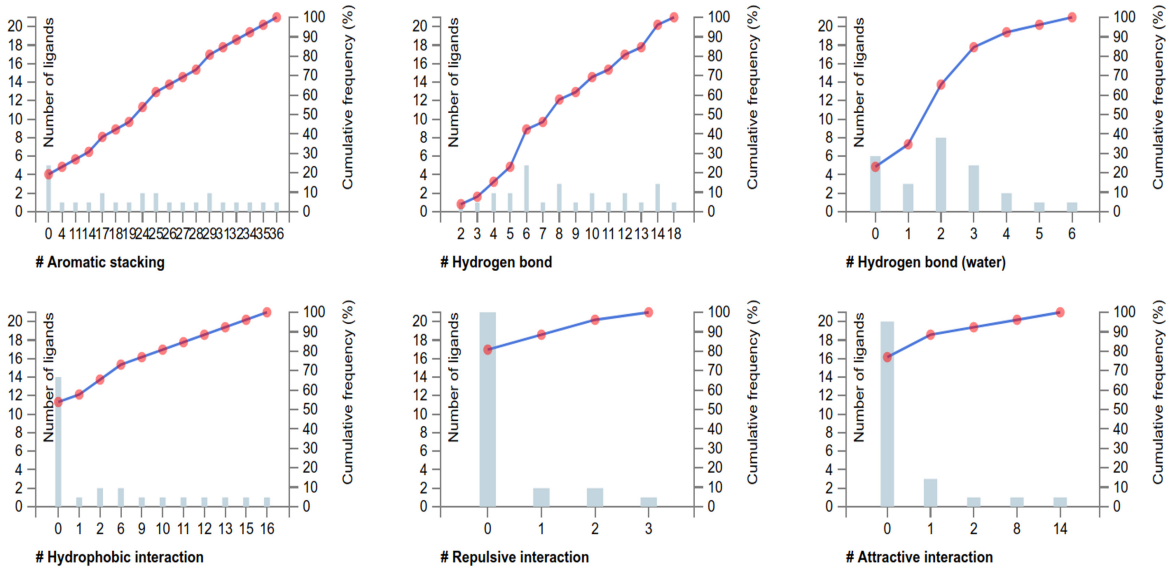


Figure 3.3: Interaction type frequencies for the RTA complexes. The X-axis shows the number of interactions, while the Y-axis shows the number of ligands that have a certain amount of interactions of a given type.

Among all clusters, the one presenting more aromatic atoms and stackings is cluster 5, which is also the group with the highest numbers of hydrogen bonds, as well as acceptor and donor atoms. Interestingly, we observed that its ligands have more acceptors than donor atoms. Considering that Ricin has evolved to depurinate a specific adenosine of the 28S RNA, the previous finding may be in accordance with recent work of Raschka et al. (2018) [193], where the authors report that proteins prefer to act as hydrogen bond donors as part of their specificity. Thus, this simple information could lead the researcher to further explore the chemical properties comprising the binding site and to guide molecular scaffold investigation [108].

Through the *Interactions by residues* table, we could also confirm literature results [41, 58, 87, 111, 124, 127, 128, 168, 169, 170, 196, 198, 204, 205, 211, 235, 239] as the key

residues TYR80 (85%), TYR123 (81%) and ARG180 (85%) presented high interacting frequencies (see Figures 3.4 and 3.5).

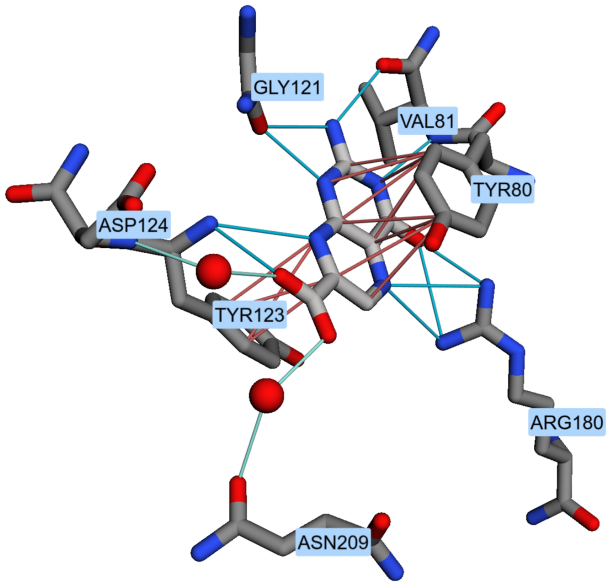


Figure 3.4: Protein-ligand interactions detected by nAPOLI to the complex Ricin and the ligand JP2:1 (PDB 3PX8). Hydrogen bonds and aromatic stackings are shown as blue and red lines. Note that all key residues TYR80, TYR123, ARG180, VAL81, and GLY121 interacted with the ligand, as well as two other residues ASP124 and ASN209.

On the other hand, GLU177 and TRP211 did not emerge as frequent interacting residues since they interacted with only 10 (38%) and 6 (23%) ligands, respectively. Regarding GLU177, the literature shows that this residue acts as a base to polarize attacking water. Indeed, from the ten ligands interacting with GLU177, eight established hydrogen bonds mediated by water. Moreover, we manually analyzed the complexes involving the ligands that did not interact with GLU177, and, in most cases, we found a water molecule placed at a favorable distance to interact with the residue but not with the ligand. Finally, TRP211 is believed to play a structural role in the binding site and is not involved in binding or catalysis [28, 201]. As a consequence, only a few ligands interacted with this residue.

Besides these key residues, VAL81 and GLY121 also presented high interacting frequencies (92% and 81%, respectively) in nAPOLI's report (Figure 3.5). Both residues were already shown to be relevant to the substrate binding [170].

Given that the ricin data set is composed of both inhibitors and substrate molecules, we also investigated if there were any differences concerning the residues interacting with these ligands. Using the *Clusters comparison* feature, we performed a comparison between all clusters (see Figure 3.6). Since clusters 2 and 3 contain only substrate molecules, except one ligand (C2X) from cluster 3 that is considered an inhibitor based on the ricin substrate, we evaluated all other clusters against these two. We first noticed that the residues ASN122, SER176, TRP211, THR216, GLU220, VAL256, and

Atom	Type of interaction	# Ligands with which it interacts
POS81 (VAL81): VAL81		Total: 24 (92.31%)
O	Hydrogen bond	24
N	Hydrogen bond	15
O	Hydrogen bond (water)	2
POS80 (TYR80): TYR80		Total: 22 (84.62%)
OH	Hydrogen bond	6
CD1	Aromatic stacking	20
CZ	Aromatic stacking	19
CG	Hydrophobic	1
OH	Hydrogen bond (water)	6
CE2	Hydrophobic	5
CD2	Hydrophobic	6
CE1	Aromatic stacking	19
CE1	Hydrophobic	3
CD1	Hydrophobic	3
CG	Aromatic stacking	20
CE2	Aromatic stacking	20
CD2	Aromatic stacking	20
POS180 (ARG180): ARG180		Total: 22 (84.62%)
NH1	Hydrogen bond	17
NH1	Hydrogen bond (water)	4
NH2	Hydrogen bond	17
NH2	Hydrogen bond (water)	3
NH2	Attractive	1
NH1	Repulsive	1
NH2	Repulsive	1
POS121 (GLY121): GLY121		Total: 21 (80.77%)
O	Hydrogen bond	21
N	Hydrogen bond (water)	1
POS123 (TYR123): TYR123		Total: 21 (80.77%)
CE2	Aromatic stacking	18
N	Hydrogen bond	15
CD2	Aromatic stacking	16
CG	Aromatic stacking	12
CD2	Hydrophobic	1
CZ	Aromatic stacking	14
CD1	Aromatic stacking	1
CE2	Hydrophobic	3
CE1	Aromatic stacking	1

Figure 3.5: Most frequent interacting residues in the ricin data set.

CYS259 are absent in the substrate clusters. Moreover, we identified four other residues (ASN78, ASP96, ASP124, and ARG213) that are found only in the inhibitor complexes, including the C2X complex. These findings may suggest residues that likely contribute to ricin inhibition.

To further evaluate if it would also be possible to distinguish inhibitors by their

Set	Set size	Residues list	Number of residues
Common residues to the Clusters 1, 2, 3, 4	4	POS172	1
Common residues to the Clusters 1, 2, 3, 4, 5	5	POS123, POS81, POS121, POS180, POS80	5
Common residues to the Clusters 1, 2, 3, 5	4	POS208, POS177	2
Common residues to the Clusters 1, 3, 4, 5	4	POS212, POS96	2
Common residues to the Clusters 1, 3, 5	3	POS78	1
Common residues to the Clusters 1, 4, 5	3	POS211	1
Common residues to the Clusters 1, 5	2	POS176	1
Common residues to the Clusters 3, 4, 5	3	POS258, POS124, POS213	3
Common residues to the Clusters 3, 5	2	POS209	1
Exclusive residues to the Cluster 2	1	POS79	1
Exclusive residues to the Cluster 3	1	POS75	1
Exclusive residues to the Cluster 4	1	POS216, POS259, POS220	3
Exclusive residues to the Cluster 5	1	POS256, POS122	2

Figure 3.6: Comparison of interacting residues in all clusters. Results are shown as alignment position. Blue rectangles highlight the residues (ASN122, SER176, TRP211, THR216, GLU220, VAL256, and CYS259) absent in clusters 2 and 3, which contain only substrate ligands. Orange rectangles highlight residues (ASN78, ASP96, ASP124, and ARG213) that are found just in the inhibitor complexes, including the C2X ligand from cluster 3.

interacting residues composition, we also compared cluster 5, which is composed of the most potent ligands [189, 190, 206, 240], against clusters 1 and 4 (see Figure 3.7). Interestingly, we observed three residues exclusive to cluster 5: ASN122, ASN209, and VAL256. Both asparagines participate in hydrogen bonds intermediated by water (see an example in Figure 3.4), while the valine establishes hydrophobic interactions. Indeed, Ready et al. (1991) [196] showed that the ASN209 plays an additive contribution to the substrate binding, and Marsden et al. (2004) [158] showed through a mutagenesis study that the N122A mutation caused a 37.5-fold reduction in the RTA activity. Therefore, we believe the *Clusters comparison* results also have the potential to reveal common and unique interacting features in the protein-ligand complexes, which may provide valuable clues regarding the molecular recognition process.

Set	Set size	Residues list	Number of residues
Common residues to the Clusters 1, 4, 5	3	POS123, POS81, POS121, POS180, POS211, POS80, POS212, POS96	8
Common residues to the Clusters 4, 5	2	POS124, POS258, POS213	3
Common residues to the Clusters 1, 5	2	POS208, POS176, POS177, POS78	4
Common residues to the Clusters 1, 4	2	POS172	1
Exclusive residues to the Cluster 5	1	POS256, POS122, POS209	3
Exclusive residues to the Cluster 4	1	POS216, POS259, POS220	3

Figure 3.7: Comparison of interacting residues in clusters 1, 4, and 5, which are the groups containing only inhibitors. Results are shown as alignment position. Blue rectangles highlight the residues ASN122, ASN209, and VAL256 that are exclusive to the most potent inhibitors (cluster 5) [189, 190, 206, 240].

3.1.5.2 Nuclear receptors subfamily 3

Human nuclear receptors (NR) are a family of ligand-regulated transcription factors that play crucial roles in many physiological processes, including development, homeostasis, and metabolism [216]. Their wide spectrum of functions is regulated by small hydrophobic signaling molecules such as hormones and dietary compounds [67, 110].

A modular structure of five domains characterizes the NR family: a variable activation function 1 domain, a conserved DNA binding domain (DBD), a variable hinge region, a conserved ligand-binding domain (LBD) and a variable C-terminal domain [110, 216]. Among them, DBD and LBD, which are the most conserved and important domains, supported the classification of the NR family into seven subgroups based on their sequence identity [80, 93].

In order to illustrate the applicability of nAPOLI in the study of conserved protein-ligand interactions along a functionally conserved binding site of a protein family, we chose the subfamily NR3 as a target for our investigation. The NR3 subfamily includes a group composed of steroid receptors and a group formed by orphan NRs. The latter comprises three estrogen-related receptors (ERRα, ERRβ, and ERRγ), while the former consists of the following proteins: androgen receptor (AR), estrogen receptor (ERα and ERβ), glucocorticoid receptor (GR), mineralocorticoid receptor (MR), and progesterone receptor (PR) [96]. Steroid receptors are important therapeutic targets as they are implicated with a series of diseases as breast and prostate cancer, cardiovascular disease, glaucoma, hyperglycemia, hypertension, obesity, osteoporosis, and disorders of the central nervous system and immunity [7, 57, 68, 110, 122, 180]. In this work, we are only interested in receptors containing endogenous ligands. Accordingly, the orphan NRs, structures that no endogenous ligands are known to bind them, were not included in the analysis.

The NR3 data set was composed as follows. First, we performed a sequence similarity search for each human steroid receptor using a 90% sequence identity cutoff. We obtained 455 structures containing ligands. Next, we searched for regions containing ligands or different binding sites in all these structures (see Section 3.1.1). To correctly select the region corresponding to the LBD binding site, we referred to the complex ERα-estradiol (PDB 1ERE; see Figure 3.8) [33]. Then, after filtering out crystallography artifacts and ions, 842 ligands remained in the LBD pocket. However, we observed through structural analysis that 51 ligands were displaced from the target pocket. Our search in the literature revealed that these ligands bound to allosteric sites [81, 137]. For this reason, we decided to discard these molecules, the remaining 791 complexes. Further analysis revealed that more than 60% of the complexes involve the ERα protein (see Table 3.3). To avoid bias towards this receptor, we generated a random uniform subsample of the data set using the Spreadsubsample algorithm from Weka [241]. The distribution spread

parameter was set to 1 (uniform distribution). Consequently, we obtained 33 complexes for each target, totalizing 198 complexes that were then submitted to nAPOLI through the *Insert nAPOLI entries* functionality, and whose results are available as an example in the *Dataset analysis* section.

Table 3.3: Number of ligands located in the LBD binding site for each Human steroid receptor.

Protein	Number of ligands
AR	99 (13%)
ERα	502 (63%)
ERβ	68 (9%)
GR	49 (6%)
MR	40 (5%)
PR	33 (4%)

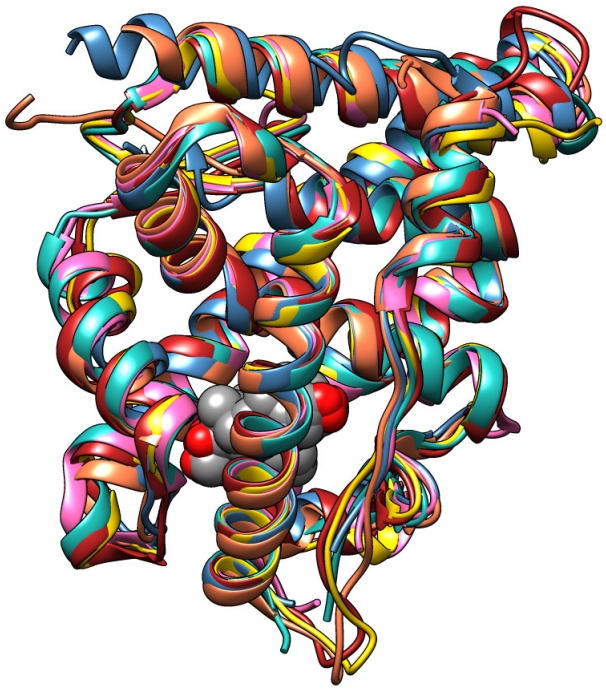


Figure 3.8: Structural alignment of the six steroid receptors. Ligands are shown as spheres and proteins are represented as cartoons. Complexes AR-dihydrotestosterone (PDB 5JJM), ERα-estradiol (PDB 1ERE), ERβ-estradiol (PDB 3OLL), GR-dexamathasone (PDB 3MNE), MR-desisobutyrylciclesonide (PDB 4UDB), and PR-progesterone (PDB 1A28) are colored red, blue, orange, pink, green, and yellow, respectively. The structural alignment and the figure were generated with LovoAlign [159] and Chimera [184], respectively.

Firstly, we assessed the ligands concerning their atom type composition and the interactions established by them. As expected, since NRs are well known for their predisposition for binding to hydrophobic molecules [80, 82], nAPOLI identified the ligands as being predominantly hydrophobic. Concerning the interactions established by these

atoms, we observed that 98% of the ligands established at least 20 hydrophobic interactions. Also, all ligands presented at least one polar atom. The literature pointed out the latter atom type as a conserved feature in the steroid family, permitting hydrogen bonds to be established in the opposite ends of the receptor pockets [110]. Indeed, we found that only three ligands did not establish any hydrogen bonds. In all these cases, the complexes (3HQ5:A:GKK:934, 3HQ5:B:GKK:2, and 3KBA:B:WOW:2) were formed between the PR and two ligands containing cyano groups (GKK and WOW). According to the HBAdd software, which we used to automatically inform HBPlus which are the acceptor and donor atoms, planar nitrogens are not pointed out as hydrogen bond acceptors. As a result, HBPlus did not identify hydrogen bonds involving the cyano nitrogen. In its turn, 133 ligands (67%) contain aromatic atoms, while 117 of them (59% of the whole data set) established aromatic interactions. Finally, charged atoms are observed in only 47 ligands (24%), and only 16 of them (8% of the whole data set) established one to three electrostatic interactions. Examples of works that discuss these two types of interactions are [74, 151, 175, 178, 199].

Regarding the frequent interacting residues, it is important to mention that although the LBD domain in the NR family is very conserved, it still presents several differences in the amino acid sequences along each receptor. These differences are both in size and physicochemical properties [78, 96]. Thus, the simple identical residues frequency count would not reveal the real importance of each residue position. Then, taking into account that different residues may be found in equivalent positions in the binding site, nAPOLI performs a structural alignment using Multiprot [217] and counts frequencies of aligned positions that interact with ligands. For example, the most frequent interacting position is the 778 (100%), in which only phenylalanines were found interacting with the set of ligands (see Figure 3.9). Another frequent position is the 759 (96%) that contains three different residues aligned to it: leucine, methionine, and serine (see Figure 3.10).

Atom	Type of interaction	# Ligands with which it interacts
POS778 (PHE778): PHE829 PHE623 PHE778 PHE311 PHE404 PHE356 PHE764 PHE629 PHE619		Total: 198 (100.00%)
CZ	Hydrophobic	69
CE2	Aromatic stacking	33
CB	Hydrophobic	6
CD2	Aromatic stacking	29
CD1	Aromatic stacking	32
CD2	Hydrophobic	91
CE1	Aromatic stacking	48
CE1	Hydrophobic	105
CE2	Hydrophobic	97
CZ	Aromatic stacking	12
CD1	Hydrophobic	102
CG	Hydrophobic	3

Figure 3.9: Alignment position containing only phenylalanines.

Despite the differences in the residue sequences, [110] demonstrated that the steroid

Atom	Type of interaction	# Ligands with which it interacts
POS759 (MET759): LEU339 LEU294 LEU387 MET759 MET610 MET600 MET604 MET745 LEU810 SER810		Total: 191 (96.46%)
CE	Hydrophobic	40
CB	Hydrophobic	181
SD	Hydrogen bond (water)	1
CD2	Hydrophobic	10
CG	Hydrophobic	40
O	Hydrogen bond (water)	131
O	Hydrogen bond	23
CD1	Hydrophobic	66

Figure 3.10: Alignment position containing three different residues: leucine, methionine, and serine.

receptors interact with their cognate ligands in a similar and conserved way. Indeed, we found 12 different positions with at least 70% of frequency, wherein each position presents on average two different residues (see Table 3.4). From these, seven positions were found interacting with the ligands in at least five of the six receptors.

Table 3.4: Most conserved interacting positions in hNR3 data set.

Interacting receptors						Alignment position	Different residues aligned to this position	Overall frequency
AR	ERα	ERβ	GR	MR	PR			
X	X	X	X	X		778	PHE	100%
X	X		X	X	X	718	LEU	98%
X	X	X	X	X	X	759	LEU, MET, SER	96%
		X	X	X		756	LEU, MET	94%
X	X	X				763	LEU, MET	94%
X	X	X	X	X	X	891	ALA, CYS, LEU, THR	92%
X		X		X		721	LEU	86%
X	X	X		X	X	890	HIS, PHE, TYR	85%
X	X	X	X	X	X	725	GLN, GLU	85%
X	X	X	X	X	X	766	ARG	80%
		X		X		755	LEU, TRP	72%
		X	X			715	HIS, LEU, MET	70%

Moreover, Huang et al. (2010) [110] also discussed conserved hydrogen bonds occurring in opposite ends of the steroid receptor pockets. All hydrogen bonds mentioned in their work were found by nAPOLI (Figure 3.11). The residues involved in these interactions were aligned in positions 719, 725, 766, 797, 890, 891, and 894 (Table 3.5). Among them, three positions (766, 735, and 719) presented more conserved participation in hydrogen bonds. Position 766, for instance, has only arginines, which are key conserved residues that play important roles in the ligand-specificity in this family [78, 102]. On the other hand, residues with a low interacting frequency could be indicative of residues that

provide target specificity, and that could be further explored. In a virtual screening campaign, for instance, one could be interested in selecting ligands that interact with specific residues from specific targets. Using the *Ligands filtering* feature from nAPOLI, a user can achieve this objective. In Figure 3.12, we show an example of a simple search for ligands establishing hydrogen bonds with the residue ASN719. This residue was pointed out by Huang et al. (2010) [110] as only establishing hydrophobic interactions, but nAPOLI automatically selected 13 complexes establishing the intended interaction.

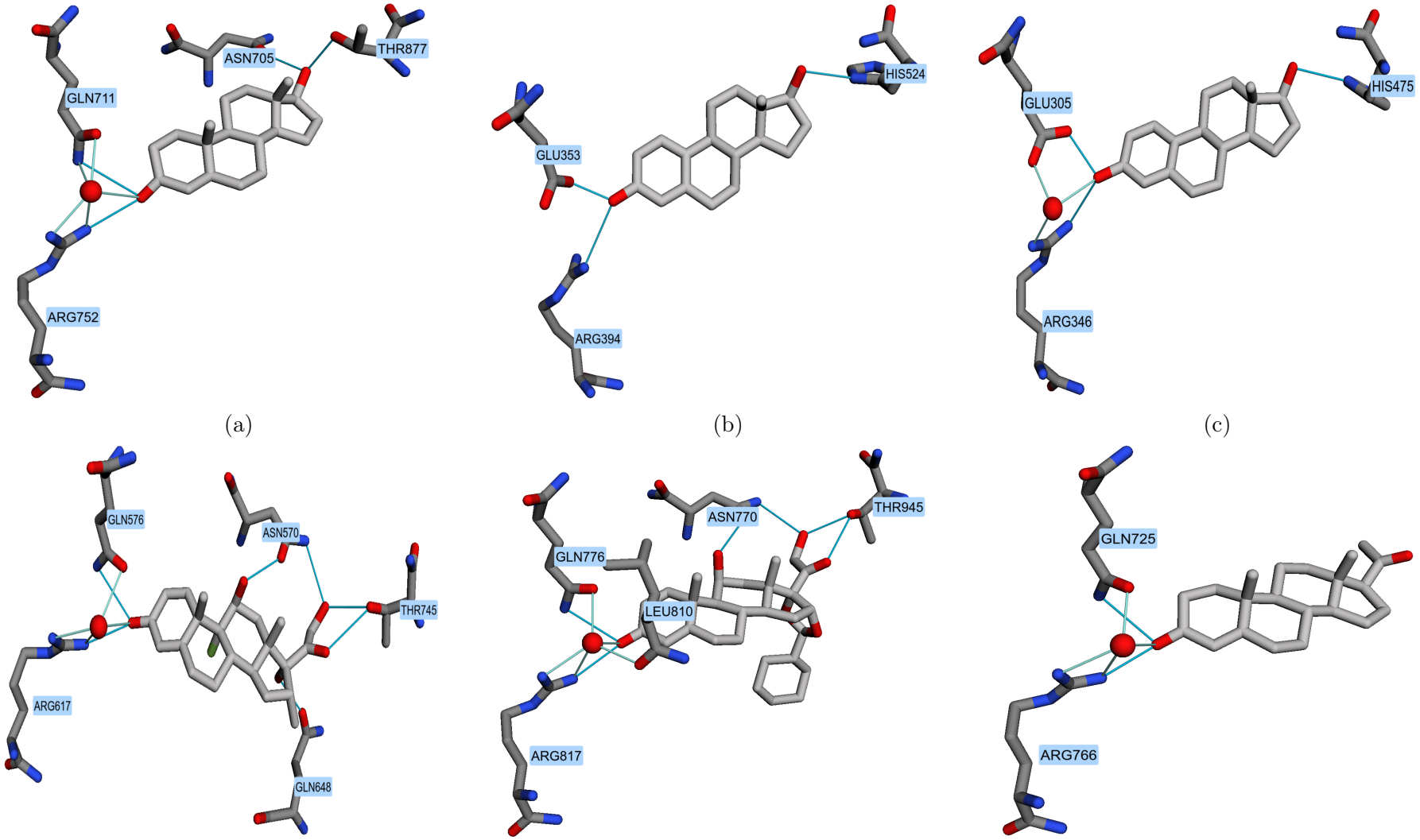
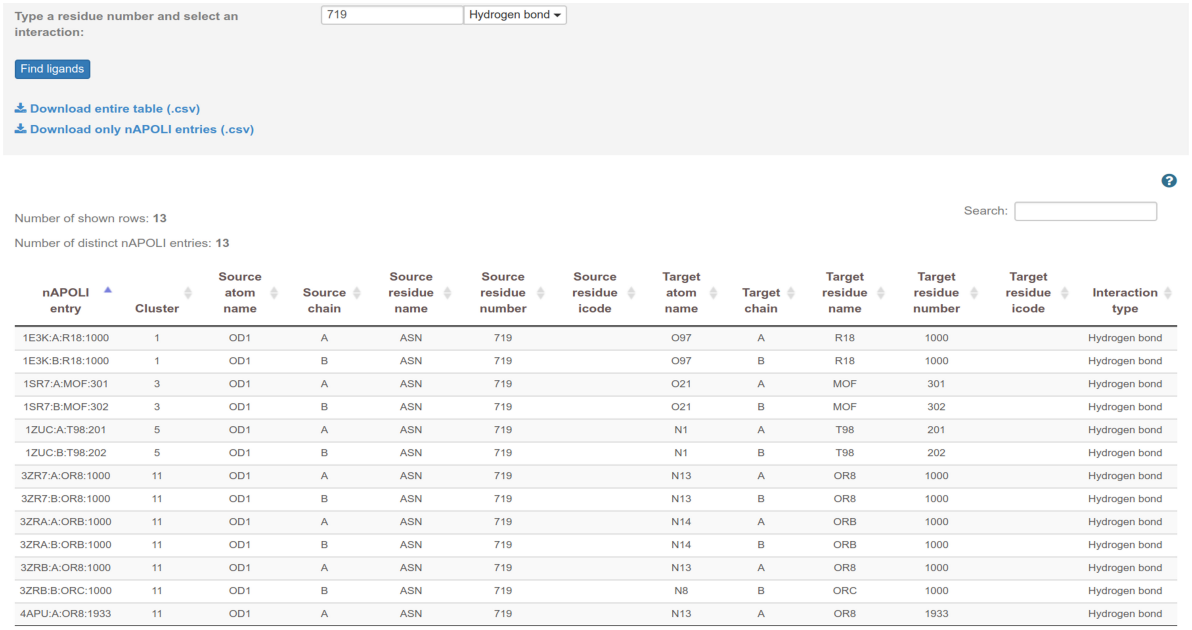


Figure 3.11: Key hydrogen bonds in the six steroid receptors are pointed out by [110] and also identified by nAPOLI. (a) AR and the ligand DHT:1001 (PDB 5JJM). (b) ERα and the ligand EST:596 (PDB 2OCF). (c) ERβ and the ligand EST:600 (PDB 3OLL). (d) GR and the ligand DEX:784 (PDB 3MNE). (e) MR and the ligand CV7:1987 (PDB 4UDB). (f) PR and the ligand STR:2 (PDB 1A28).



nAPOLI entry	Cluster	Source atom name	Source chain	Source residue name	Source residue number	Source residue icode	Target atom name	Target chain	Target residue name	Target residue number	Target residue icode	Interaction type
1E3K:A:R18:1000	1	OD1	A	ASN	719		O97	A	R18	1000		Hydrogen bond
1E3K:B:R18:1000	1	OD1	B	ASN	719		O97	B	R18	1000		Hydrogen bond
1SR7:A:MOF:301	3	OD1	A	ASN	719		O21	A	MOF	301		Hydrogen bond
1SR7:B:MOF:302	3	OD1	B	ASN	719		O21	B	MOF	302		Hydrogen bond
1ZUC:A:T98:201	5	OD1	A	ASN	719		N1	A	T98	201		Hydrogen bond
1ZUC:B:T98:202	5	OD1	B	ASN	719		N1	B	T98	202		Hydrogen bond
3ZR7:A:OR8:1000	11	OD1	A	ASN	719		N13	A	OR8	1000		Hydrogen bond
3ZR7:B:OR8:1000	11	OD1	B	ASN	719		N13	B	OR8	1000		Hydrogen bond
3ZRA:A:ORB:1000	11	OD1	A	ASN	719		N14	A	ORB	1000		Hydrogen bond
3ZRA:B:ORB:1000	11	OD1	B	ASN	719		N14	B	ORB	1000		Hydrogen bond
3ZRB:A:OR8:1000	11	OD1	A	ASN	719		N13	A	OR8	1000		Hydrogen bond
3ZRB:B:ORC:1000	11	OD1	B	ASN	719		N8	B	ORC	1000		Hydrogen bond
4APU:A:OR8:1933	11	OD1	A	ASN	719		N13	A	OR8	1933		Hydrogen bond

Figure 3.12: Filtering example in which ligands that are interacting with a specific residue (ASN719) from the PR protein were selected.

Table 3.5: Residues (grouped by their alignment position) establishing hydrogen bonds as mentioned in [110].

Receptors	Alignment positions						
	POS766	POS725	POS719	POS891	POS890	POS894	POS797
ERα	R394	E353			H524		
ERβ	R346	E305			H475		
AR	R752	Q711	N705	T877			Q642
GR	R611	Q570	N564			T739	
MR	R817	Q776	N770	C942*		T945	
PR	R766	Q725	N719*	C891*		T894*	
Overall frequency	80%	85%	64%	92%	85%	34%	63%
Fraction of HBonds	94%	88%	76%	21%	21%	46%	10%

* In [110], these residues were pointed out as only establishing hydrophobic interactions.

In Table 3.5, we also highlight four residues (CYS942, ASN719, CYS891, and THR894) presenting divergent results from [110]. In Huang et al. (2010), these residues were pointed out as involved only in hydrophobic interactions. On the other hand, nAPOLI identified these residues as also establishing hydrogen bonds, which is in agreement with literature findings [26, 150, 182].

3.2 A novel Python library for drug design

LUNA[1] was built in Python 3 and is fully object-oriented to permit users to control every processing step programmatically. The required dependencies are Biopython, mmh3, Numpy, Open Babel, Pymol, and RDKit.

The files accepted by LUNA depend on the scenario in which the library is employed. For projects containing complexes whose structures are in the same file, the options are the same as in Biopython: PDB, PDBx/mmCIF, PDBML/XML, MMTF, and bundle. Finally, for projects containing complexes in different files, as commonly happen when a docking study is performed, two separate files can be provided: one for the macromolecule and one for the ligand. The macromolecule structures can be in any format supported by Biopython as presented before. For the ligands, the accepted files are any file supported by Open Babel or RDKit.

In the rest of this section, we present some of the major LUNA features.

3.2.1 Filtering interactions

LUNA provides several filters to control how interactions are calculated. Some of them, for instance, can be applied to remove: intramolecular, protein-protein, protein-DNA/RNA, protein-small molecule, DNA/RNA-DNA/RNA, DNA/RNA-small molecule, and small molecule-small molecule interactions; and interactions involving waters and artifacts of crystallography (Table 3.1). Other filters can also be applied to calculated interactions to select those that match some geometric constraints.

3.2.2 Statistical analysis and data set characterization

LUNA provides several functions for summarizing and characterizing molecular interactions and physicochemical properties, besides the generation of statistical data with the support of Numpy, Pandas, and Seaborn. Besides it, LUNA also brings together visual strategies physicochemical properties and molecular interaction analysis. Herein,

[1] https://github.com/keiserlab/LUNA

we highlight the 2D molecular diagram (Figure 3.13) and the frequency heatmap plot that can be employed in molecular dynamics analysis to characterize and summarize the frequency in which the residues interacted with the ligands throughout the simulation (Figure 3.14).

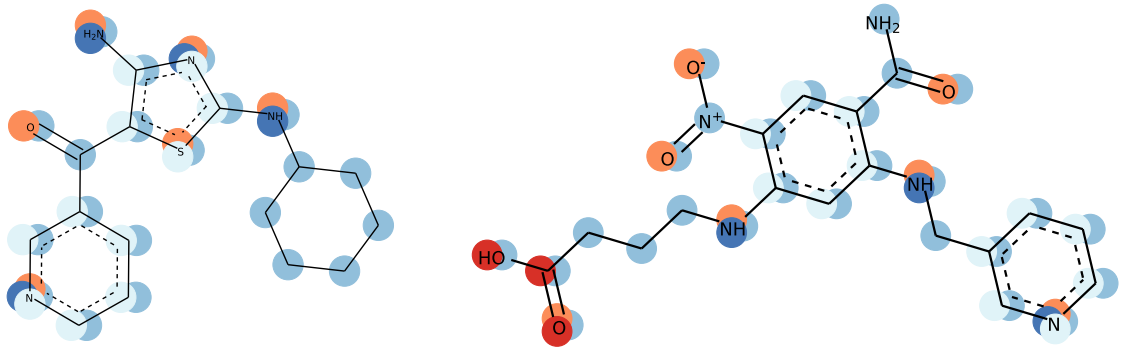


Figure 3.13: Example of a molecular 2D structure diagram with atoms colored according to their physicochemical properties.

3.2.3 Visualizing interactions

Visualization of biological data is a clearly advantageous application inasmuch as it aids the exploration, analysis, and interpretation of the data in an interactive and straightforward way. Bearing this in mind, we built a functionality where users can export their calculated interactions into a Pymol session (Figure 3.15). Thus, one can analyze the binding site using a visual approach and interact with it by using the Pymol features.

In our visualization, interactions are color-coded and grouped by type, simplifying the filtering of them. Moreover, multiple complexes can be visualized at the same time, which makes it easier to discover conservation and dissimilarities. For analyzing multiple complexes in a stacked way, it is necessary to perform a structural alignment before.

3.2.4 Visualizing fingerprint information

One drawback of most interaction and molecular hashed fingerprints refers to the fact that the structural information is lost after the generation of the fingerprint. First,

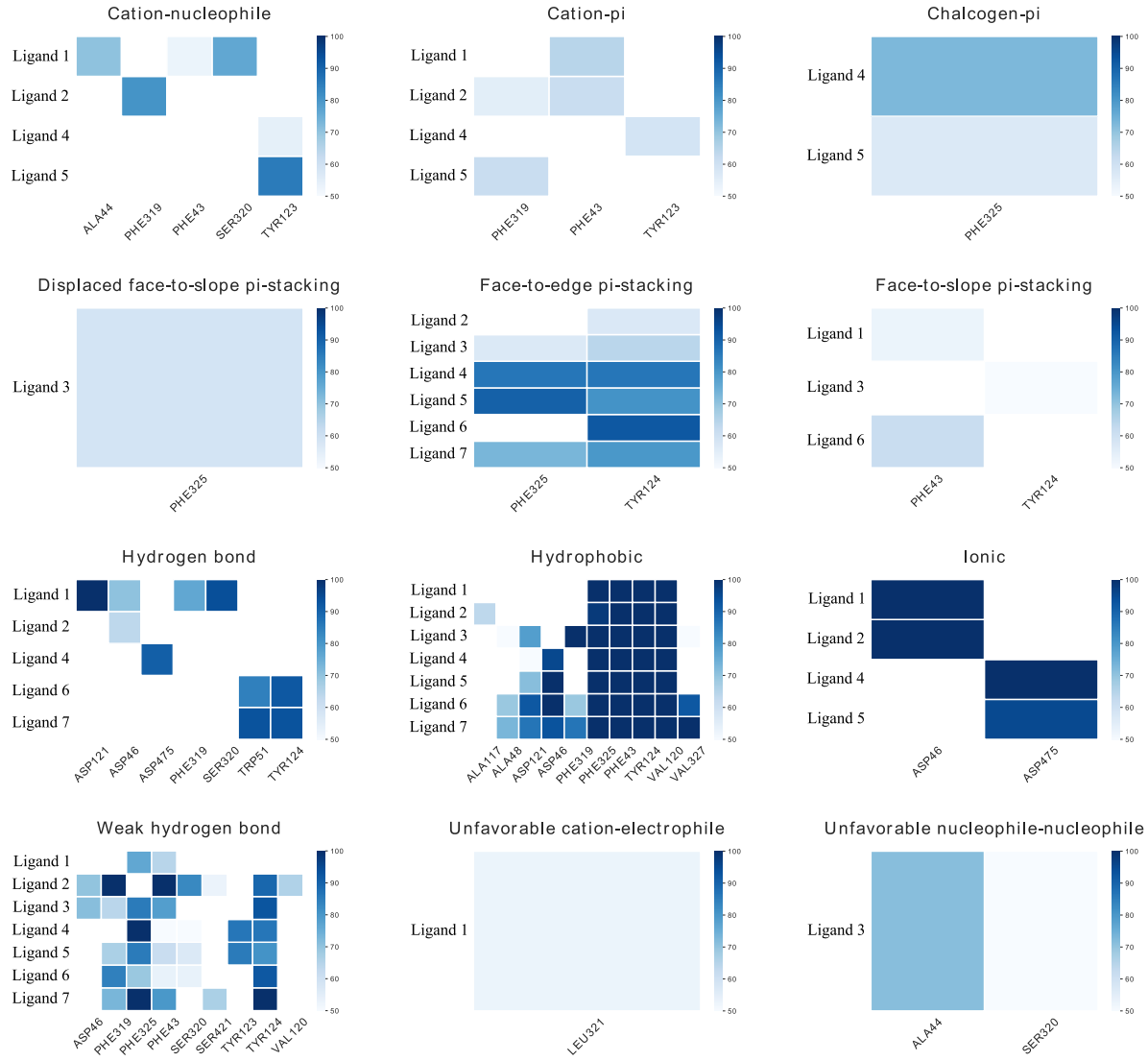
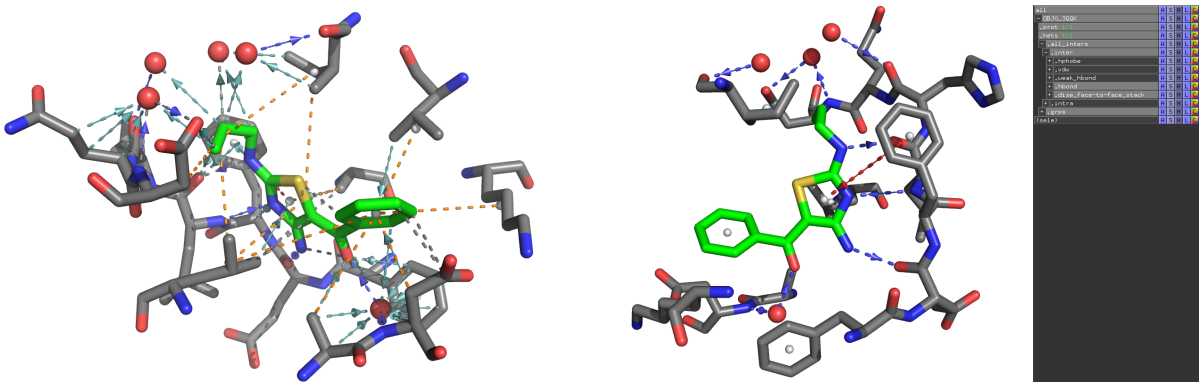


Figure 3.14: Example of an interaction heatmap summarizing the most frequent residues interacting with different ligands throughout trajectory clusters obtained from a molecular dynamics simulation.

because it is not possible to know which substructure generated the corresponding bit (1s) in the fingerprint. Second, because multiple substructures can be hashed into the same vector position, the so-called collision problem. Consequently, the hashed fingerprint becomes a black-box as no structural information can be recovered from the bits.

Bearing this in mind, we provide in LUNA a functionality where users can recover the information encoded into a bit, i.e., to reconstruct the neighborhood that generated that bit. Additionally, it allows users to export one or more recovered neighborhood into a Pymol session. Thus, one can visually analyze the atom neighborhood and interact with it using the Pymol features.

The neighborhood view generated by Pymol is similar to sessions created for molecular complexes (Section 3.2.3), but here the focus is on a central atom/group, its neighboring atoms/groups, and interactions established by them. Since multiple neighborhoods

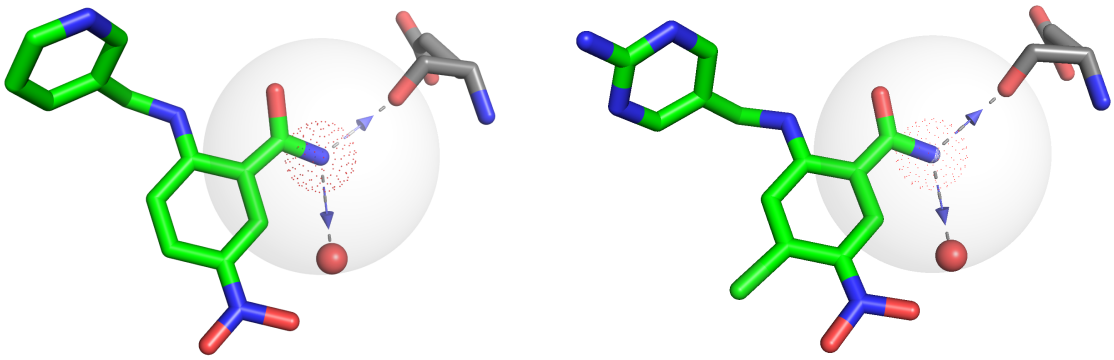


(a) (b)

Figure 3.15: A Pymol session generated by LUNA, highlighting the protein-ligand interactions between the enzyme CDK2 and the ligand X02 (PDB 3QQK). On the left, an overview of all interactions is shown. In the right, the view was rotated, and a filter was applied to the interactions to only show *hydrogen bonds* (blue) and a *displaced face-to-face stacking* (red). Arrows represent directional interactions.

can be represented in the same Pymol session, LUNA groups interactions by neighborhood.

One useful application of this feature is the possibility of visualizing similar and dissimilar bits between two fingerprints through an interactive approach. Figure 3.16 shows an example of two similar bits, i.e., two similar neighborhoods, found in two different CDK2-inhibitor complexes.



(a) (b)

Figure 3.16: A Pymol session generated by LUNA, highlighting two similar bits (neighborhood) found in the fingerprints of two CDK2-inhibitor complexes (PDBs 3QQK and 3QWJ, left and right). Blue arrows and gray dashed lines represent *hydrogen bonds* and *van der Waals* interactions. Arrows point to the interaction direction.

3.3 Fingerprint evaluation

As we do not know a priori which similarity value we should expect when two poses are compared with the proposed fingerprint, we performed an exploratory evaluation of the fingerprint parameters using three data sets composed of similar ligands related to the human cyclin-dependent kinase 2 (CDK2). These data sets were empirically designed using CDK2 inhibitors, presenting both structural and molecular similarities and low RMSD to the crystal structures. By doing so, we hypothesize that the binding mode similarity between two pairs of FIFP would be proportional to the molecular and pose similarity.

We organized this section as follows. First, we scrutinize the parameters *number of levels* and *radius growth rate*, followed by a discussion about the impact of different fingerprint lengths on the collision rate. Finally, we present an evaluation of the separability between similar from dissimilar poses using FIFP.

3.3.1 Effect of the number of levels and radius growth rate

In this section, we analyze the effect of the parameters *number of levels* and *radius growth rate* on the similarity of two complexes involving the same target and ligands with similar poses. We start with a discussion about the effects of these parameters on the similarity of manual poses and the ligand X07 (PDB id of the CDK2 complex: 3QQF). Finally, we present a similar discussion using different conformers of the ligand X02 (PDB id of the CDK2 complex: 3QQK) and an additional evaluation of the parameter effects when different CDK2 ligands are considered.

3.3.1.1 Same ligand in different manual poses

Departing from the manual generated poses, we selected pose C (see Figure 2.3), which presented the least amount of different interactions compared to the crystal pose (Figure 3.17), as our positive reference. Thus, among all manual poses, we expect it to present the highest similarity to the original pose.

On the other hand, we selected pose L as our negative reference since this pose

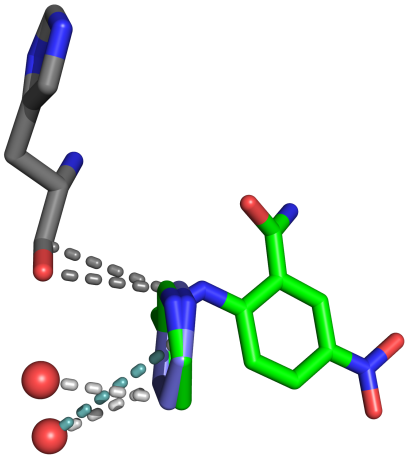


Figure 3.17: Comparison between the original crystal pose (PDB 3QQF) and pose C. Protein, original pose, and modified pose are shown as gray, blue, and green sticks. Exclusive interactions of each complex are shown as dashed lines, where light gray represents van der Waals interactions established by the original pose, and dark gray and teal represent van der Waals and weak hydrogen bonds, respectively, established by the modified ligand.

presented the highest number of dissimilar interactions in comparison to the crystal pose (Figure 3.18).

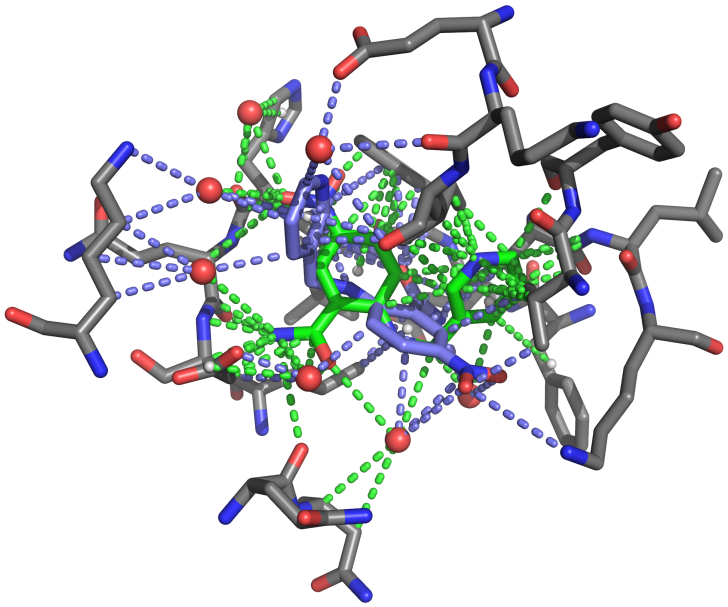


Figure 3.18: Comparison between the original crystal pose (3QQF) and pose L. Protein, original pose, and modified pose are shown as gray, blue, and green sticks. Exclusive interactions are shown as blue (original pose) and green (modified pose) dashed lines.

In Figure 3.19, we show how the variation in the number of levels influences the similarity between the manual poses and the crystal structure. Note that the positive and negative references are the ones with the highest and lowest similarities, respectively,

which is in accordance with our expectations. Moreover, as we mentioned in Section 2.3.4, the number of levels indirectly controls how many features will be included in the fingerprint. This statement is emphasized by the trend in Figure 3.19, in which the similarity decreases as we increase the number of levels. That happens because the number of levels defines how many iterations the algorithm will have to explore the feature space. Therefore, the higher the level, the more features could be found.

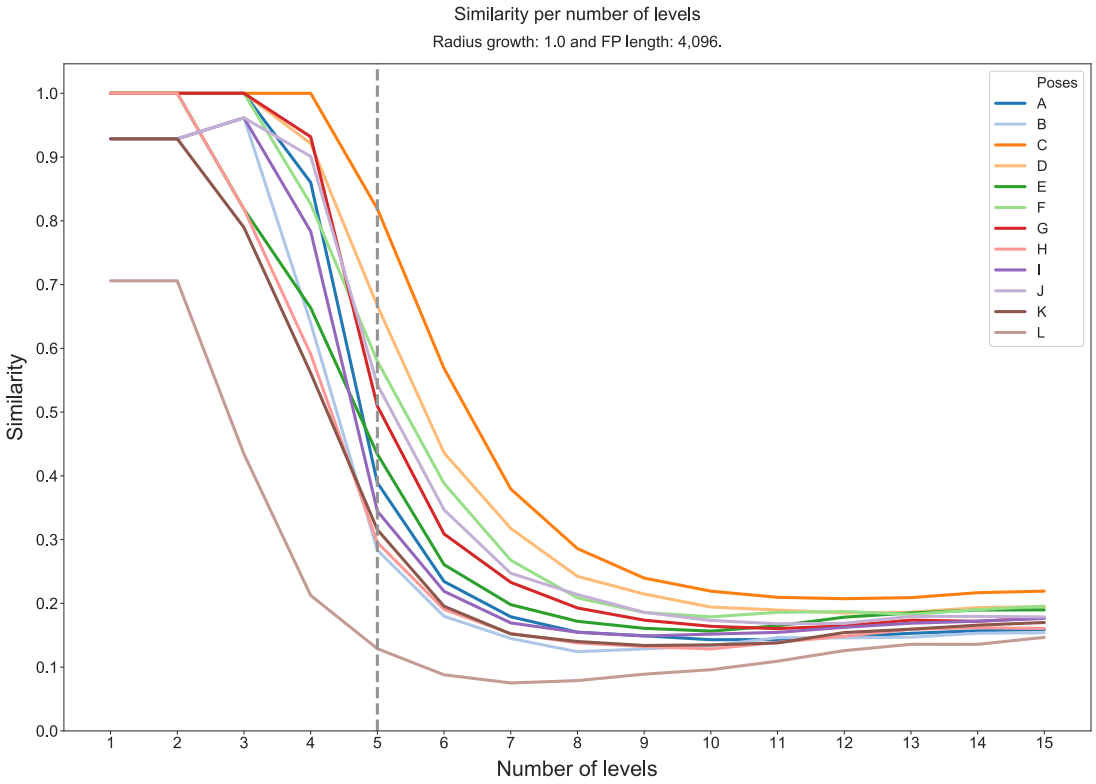


Figure 3.19: Effect of the number of levels on the similarity between manually generated poses and the crystal structure. The gray dashed line highlights the minimum number of levels for differing all poses from the original complex.

It is also noteworthy that the similarity decreases until a specific number of levels, then it slightly increases, and from then on it converges. The increase in the similarity given by higher levels is indicative of bit collisions. As we increase the number of levels, more features are generated and included in the fingerprint. However, since the fingerprint has a limited length, as higher the number of features covered by the fingerprint, the higher the chance of collisions. As a consequence, some collisions may generate false similar-bits, which artificially increases the similarity between two poses. In its turn, the convergence in higher levels indicates that the whole feature space was already covered, i.e., there is no new information to be included in the fingerprint after a specific number of levels.

Finally, the vertical gray line in Figure 3.19 highlights the minimum number of levels to obtain a separation between the conformers from the crystal pose. The same trend can be observed when varying the radius growth rate and the number of levels

(Figure 3.20). For each radius growth rate, there is a minimum number of levels that permit all manual poses to be separated from the original complex. For example, the blue line, which corresponds to the growth rate of 0.1, shows that for small steps, more levels are required to completely separate very similar poses.

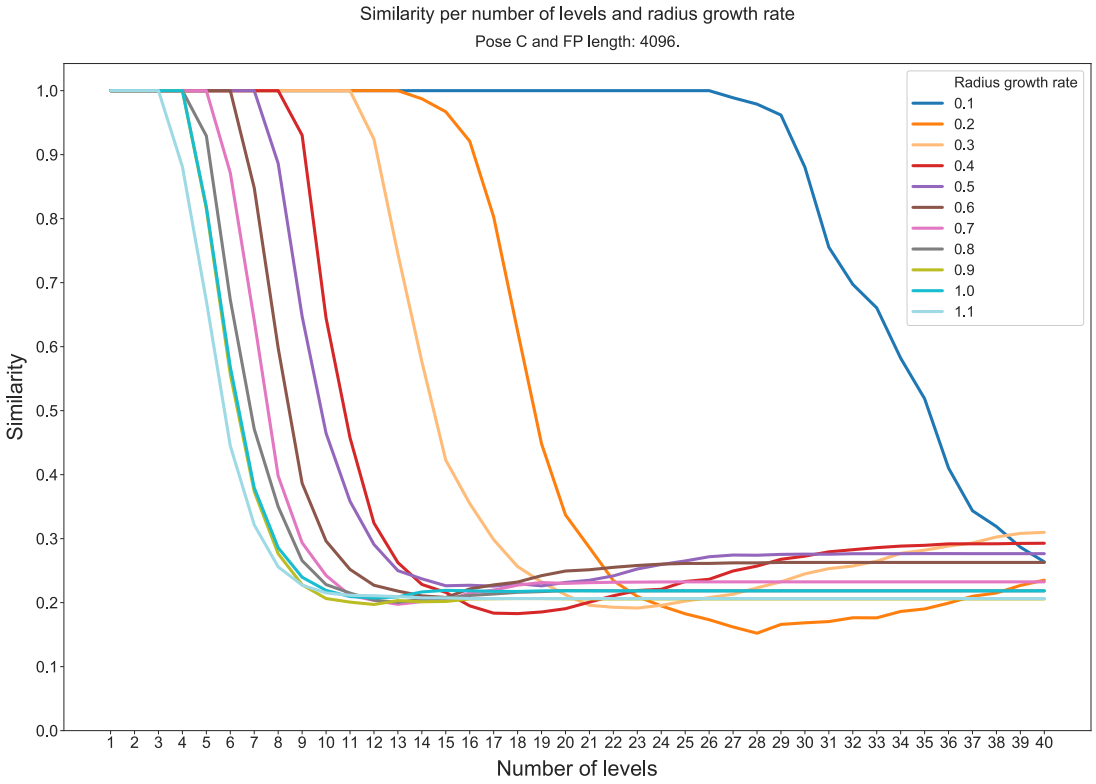


Figure 3.20: Effect on the similarity between Pose C and the original pose when varying the number of levels and radius growth rate.

3.3.1.2 Conformer analysis

In order to better assess the ability of FIFP on separating similar poses, we selected a bigger set of conformers automatically generated with RDKit and evaluated how the number of levels influences the similarity between these poses and the CDK2-ligand complex (PDB 3QQK) (Figure 3.21). As explained in Section 2.4.1, these conformers were obtained after a filtering procedure where only poses with an RMSD to the crystal less than 0.4 were kept.

Using the minimum number of levels (5) found in the previous experiment as our reference, we noticed that most of the conformers presented higher similarity to the crystal pose, which is already expected given the RMSD threshold we used for pruning conformers. As we increase the number of levels, the average similarity between the

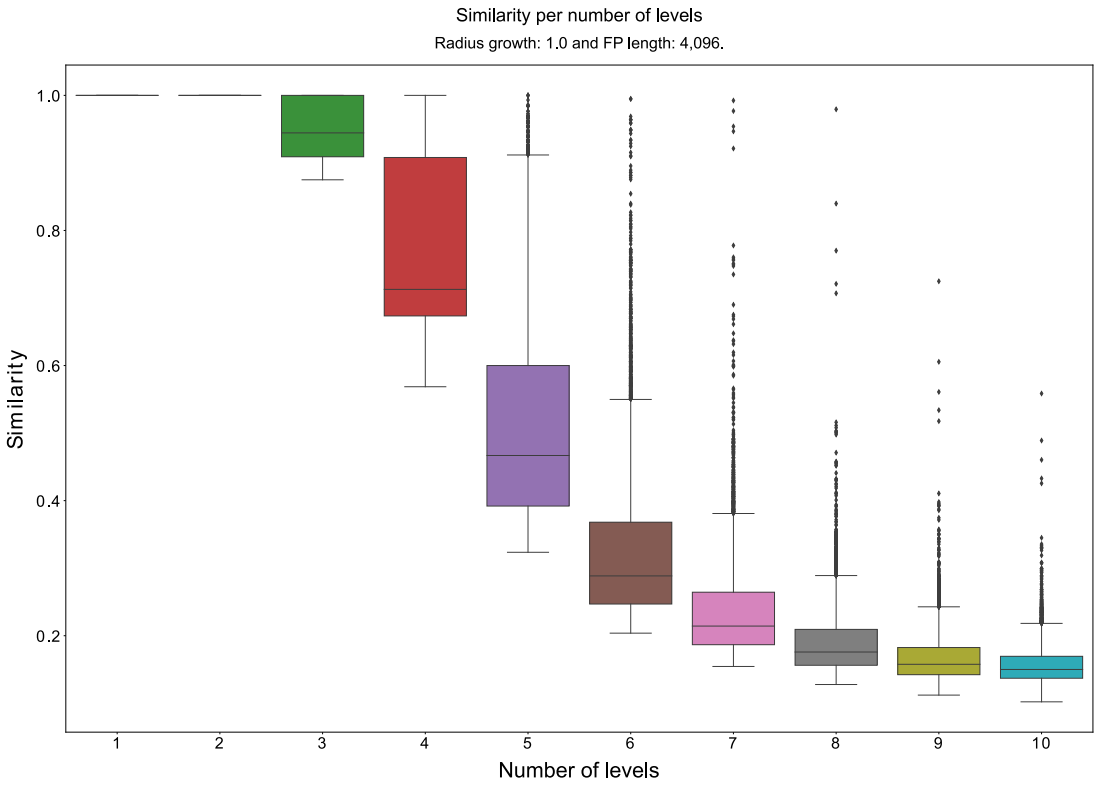


Figure 3.21: Effect on the similarity between the ligand X02 (CDK2 complex id: 3QQK) and its different conformers when varying the number of levels.

conformers and the crystal pose is usually lower than 0.5. However, it is possible to observe that some conformers still presented similarities above this value, which highlights that the fingerprint is able to separate similar from dissimilar poses.

3.3.1.3 Different CDK2 inhibitors

Besides the analysis consisting of the same ligand on different poses, we also evaluated the similarity behavior when comparing different ligands (Figure 3.22). In previous sections, we showed that the minimum number of levels for separating two very similar poses was 5. Analysing the same value for the current data set, we observed that all pairs of complexes presented a similarity lower than 0.5. Together with the previous analysis, this result indicates that more features are usually required to separate very similar poses, which could be achieved by increasing the number of levels. On the other hand, Figure 3.22 points out that the separation between dissimilar poses does not require so many features and lower levels are enough for separating two different complexes.

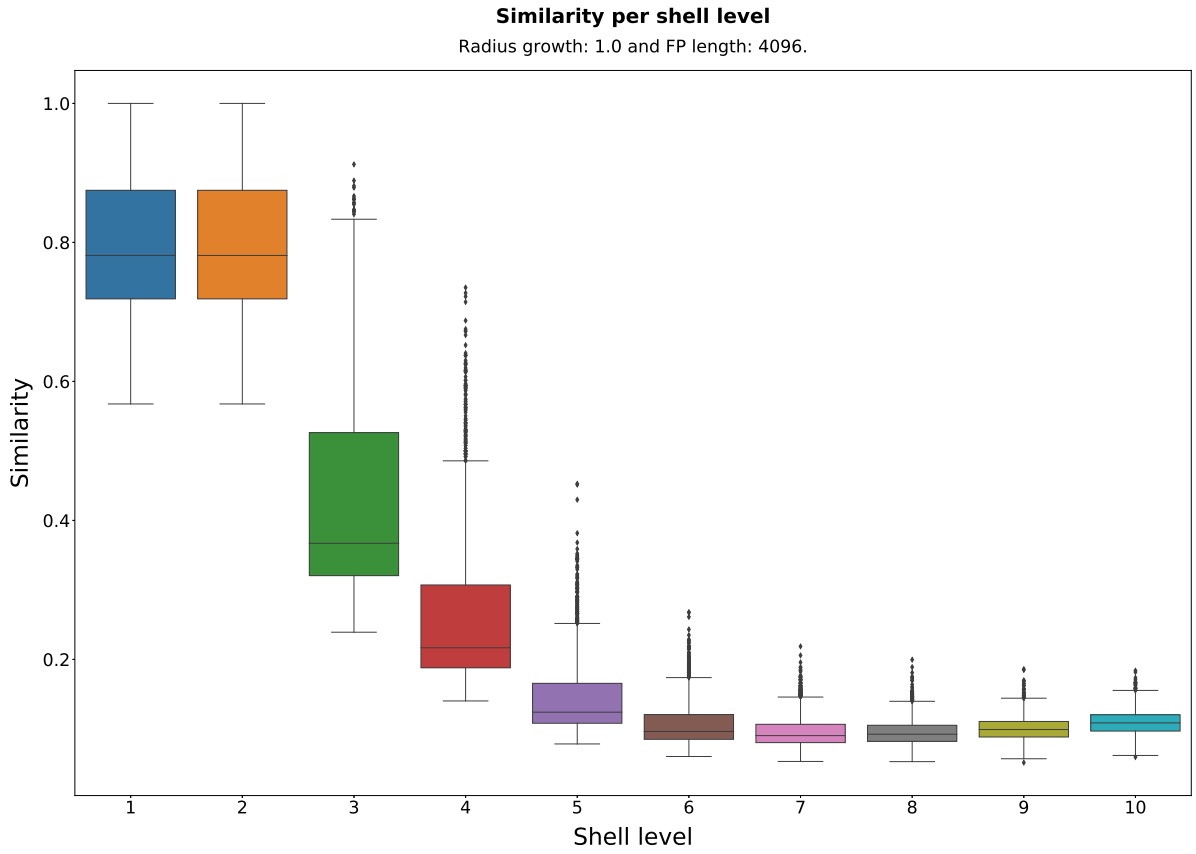


Figure 3.22: Effect of the number of levels on the similarity between pairs of CDK2 inhibitors (all against all).

3.3.2 Effect of the fingerprint length on the collision rate

For the analysis of the effect of bit collisions on fingerprint generation, we used the data set composed of 74 CDK2 inhibitors. To do so, we generated fingerprints of different sizes by fixing the number of levels and the radius growth rate to 7 and 1, respectively. Then, for each fingerprint length, we measured the number of bits on, the collision rate, and the fingerprint darkness, which is the rate of bits on in relation to the fingerprint length (Figure 3.23).

Note that the number of collisions and fingerprint darkness reduces as we increase the length of the fingerprint, which is in accordance with the expected behavior. Also, observe that even in the worst case (4,096), on average, only 4% of the bits set on presented collisions, while the maximum percentage of collisions observed was less than 8%. Thus, we conclude that the rate of collisions for the tested lengths is acceptable. However, to be more conservative, we envision the 16,384 version as more interesting given its tradeoff between collision rate and performance - computational tasks using longer fingerprints may require more processing time.

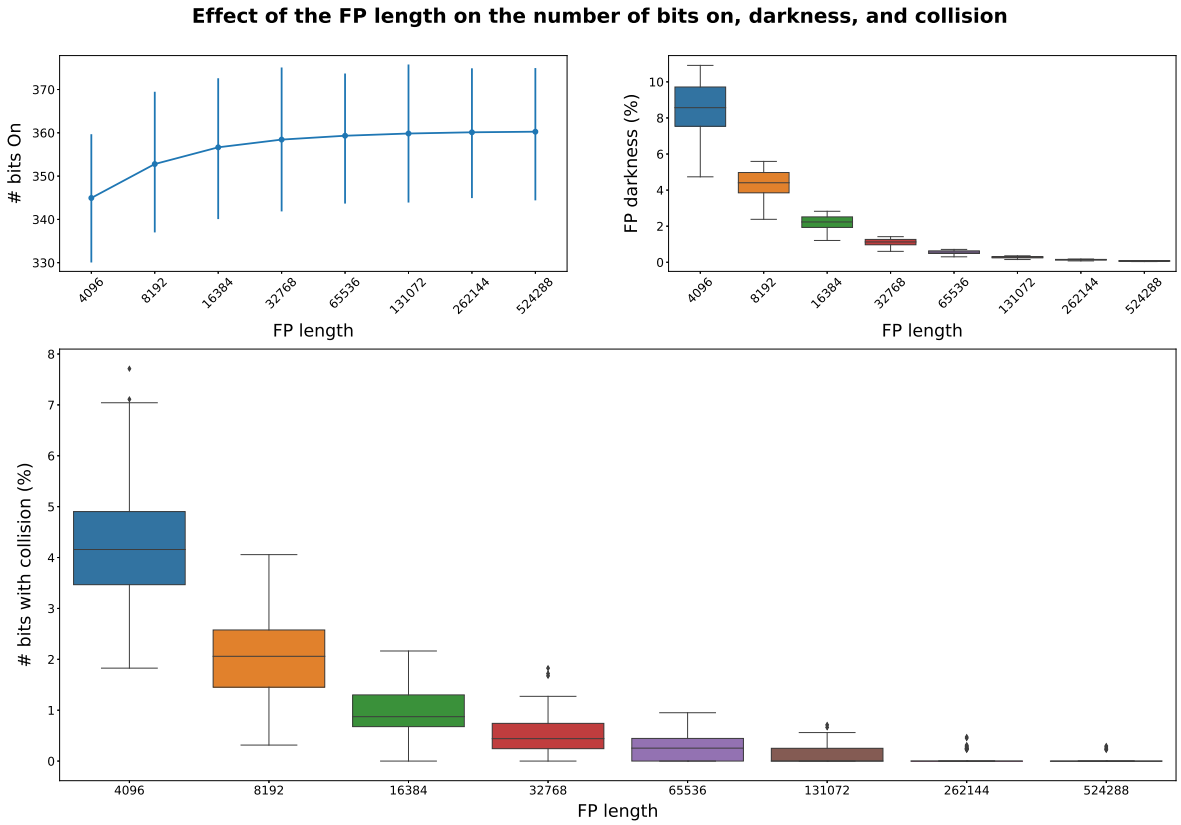


Figure 3.23: Effect of the fingerprint length on the rate of collisions and fingerprint darkness. In the top left, the point plot (a boxplot variant that better highlights the relation between two parameters) shows the number of bits for the 74 CDK2 inhibitors. In the top right and the bottom, the box plots show the variability of the fingerprint darkness and collision rate, respectively, when the fingerprint length increases.

3.3.3 Separability of similar and dissimilar binding modes

To further validate our findings, we also decided to evaluate if the fingerprint is able to separate similar from dissimilar complexes. To do so, we manually selected pairs of fingerprints from the 74 CDK2 inhibitors data set based on their molecular structure and pose similarity. For positive control, we selected the pairs of manual poses evaluated in Section 3.3.1.1, and as a negative control, we included additional pairs composed by CDK2 ligands (entries 3QL8:A:X01:300 and 3QQK:A:X02:497) and different protein-ligand complexes, namely carbonic anhydrase II (entry 1G54:A:FFB:555), ricin (entry 1BR5:A:NEO:500), thymidylate synthase (entry 1TSD:A:F89:268), and tRNA-guanine transglycosylase (entry 1K4H:A:APQ:900). By including these new pairs, we expect that the CDK2-inhibitor complexes to be highly dissimilar to the non-CDK2 complexes.

We then generated fingerprints of different lengths while fixing the number of levels and the radius growth rate to 7 and 1, respectively. In order to highlight the separability of similar and dissimilar complexes, we also grouped pairs according to their expected

similarity into the following groups: manual poses, similar CDK2 inhibitors, dissimilar CDK2 inhibitors, and different proteins and ligands (expected negative control). Poses C and L, which are our positive and negative references, respectively, remained ungrouped. The similarity between the fingerprint pairs was calculated and shown in Figure 3.24.

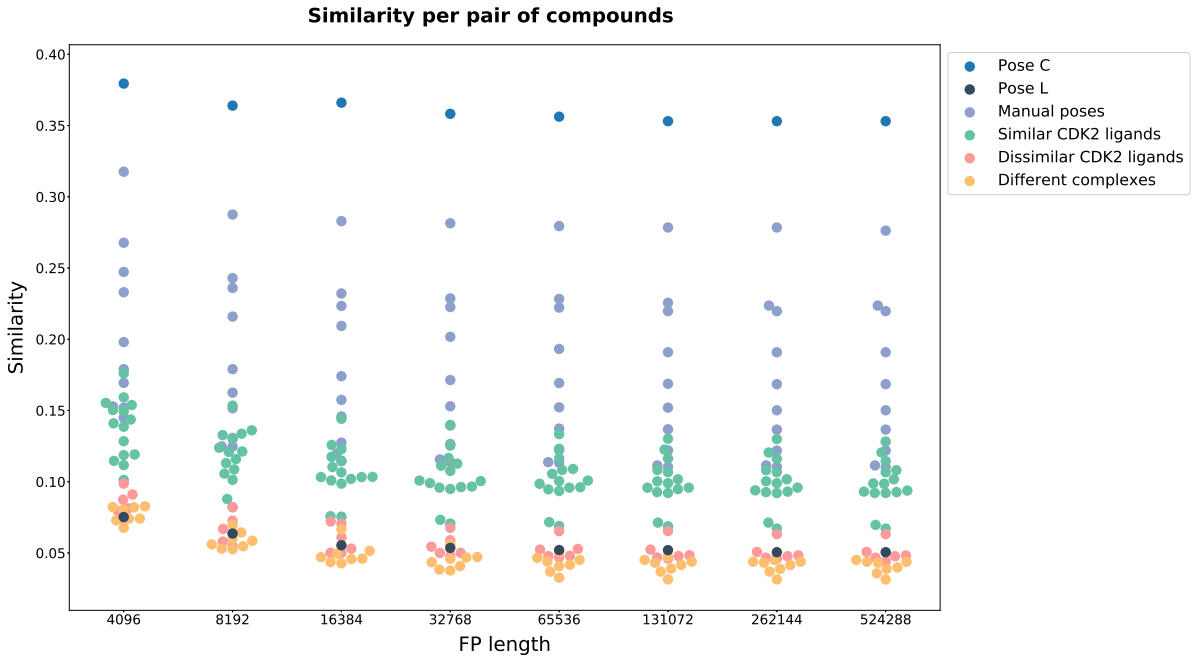


Figure 3.24: Similarity per pair of complexes for different fingerprint lengths.

Note in Figure 3.24 that pairs of compounds expected to be similar (manual poses and similar CDK2 inhibitors) presented higher binding mode similarity and are separated from pairs expected to be dissimilar (dissimilar CDK2 inhibitors and different proteins and ligands). Not surprisingly, poses C and L appear at the top and at the bottom of the plot, respectively.

Interestingly, longer fingerprints tend to accentuate the separateness of the groups, which is explained by the reduction of bit collisions. Previously, we mentioned that the collision rate artificially increases the similarity between any two poses. That happens because fingerprints contain a predefined length and so, a limit for storing novel features. Thus, the increasing the number of features, the higher will be chance to occur bit collisions. On the other hand, if the fingerprint length is extended, more features can be allocated, which, consequently, reduces the collision rate and the artificial similarity.

These remarking results potentially emphasize the applicability of FIFP on SBVS campaigns where it could be employed for selecting molecules that bind in a similar manner to known active molecules or even filtering out those ones that differ too much from active molecules.

3.4 Dock score prediction

To validate and illustrate the applicability of FIFP for prioritizing hit compounds, we decide to apply this novel fingerprint to the Dock score prediction task. To do so, we chose a large data set recently published by Lyu et al. (2019) [153] and performed several experiments to identify the best FIFP parameters followed by their comparison with the baseline (ECFP and FCFP) and two other interaction fingerprint models.

This section is organized as follows. First, we present the results and discuss the exploratory search for the best FIFP parameters. Then, we finalize with a discussion on comparing FIFP and other fingerprints.

3.4.1 Exploratory search for the best FIFP parameters

In the first experiment, we analyzed different combinations of parameters in order to select the best models for further comparison with the baseline models. Besides the fingerprint attributes (number of levels, radius growth rate, and fingerprint length) (Figures 3.26 and 3.30), we also evaluated how the methodology employed to calculate interactions influences the predictive performance. To do so, we empirically explored combinations of the following options: strict or loose rules for hydrogen bond donor (*Strict H rule*; Figure 3.25); include or not non-covalent (excluding van der Waals) interactions (*Non-cov*; Figure 3.27); compute or not atom-atom interactions (*Atom-Atom*; Figure 3.27), which include *covalent bond*, *van der Waals*, *van der Waals clash*, and *atom overlap*; include or not proximal interactions (*Proximal*; Figure 3.27); protein structure with or without hydrogens (*Struct w/ H*; Figure 3.28). The presence of the described labels in the experiment name (X-axis) indicates whether it was used or not during the calculation of interactions. For the best models, we also evaluated if its count fingerprint version and the inclusion of interactions in the protein side (*w/ PPI*) would improve the prediction (Figure 3.29).

Regarding the methodologies for hydrogen bond calculation, we observed that the models slightly improved when strict rules for hydrogen bond donors were used (Figure 3.25). That indicates that the false positive interactions identified by loose rules negatively impact the model performance. Finally, it is noteworthy in Figure 3.25 that the only exceptions are the fingerprints whose number of levels and radius rate growth are 13 and 0.5, respectively. In this particular case, the reduction in the R^2 when using strict rules are due to the artificial similarity improvement discussed in Section 3.3.1.1, i.e.,

this particular combination of levels and radius growth produces an excessive number of features, which increases the collision rate and, consequently, generates false similar-bits.

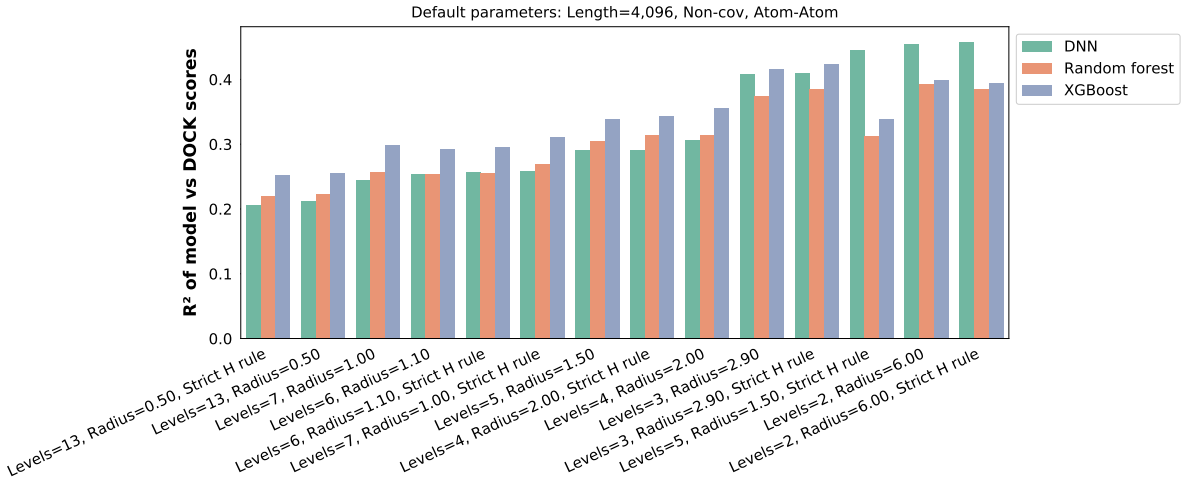


Figure 3.25: Comparison between strict and loose rules for hydrogen bonds. Default FIFP parameters are shown above the chart, and bars are ascendingly sorted from left to right according to the DNN models.

Another interesting trend we observed in the predictive performance was that fingerprints obtained by using a small number of levels produced the highest R^2 values (Figure 3.26), being the number of levels 2 (radius growth rate around 6 Å) and 3 (radius growth rate around 2.9 Å) the best options. The only exception was the combination of 5 levels and a radius growth rate equal to 1.5 that showed an R^2 close to the top scores.

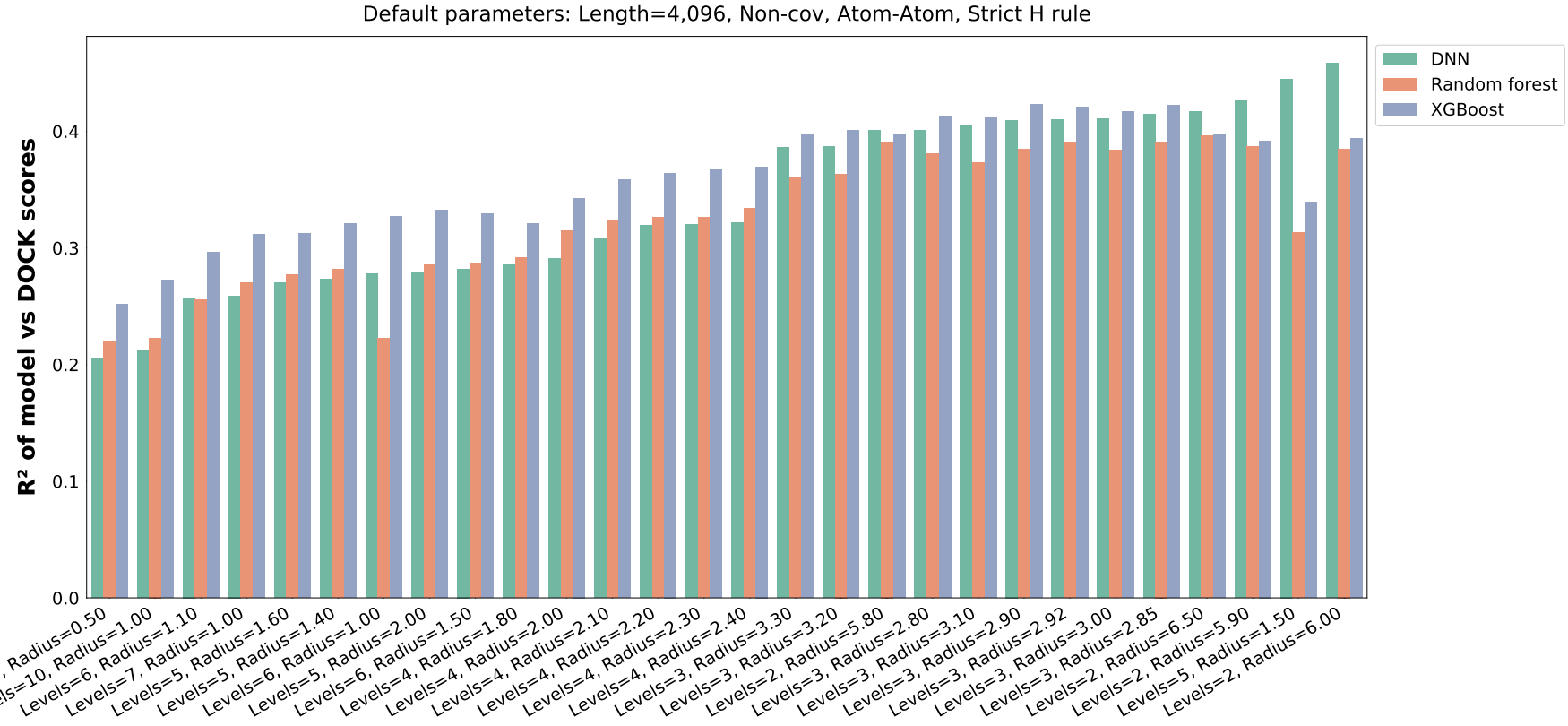


Figure 3.26: Comparison between the different number of levels and radius growth rate. Default FIFP parameters are shown above the chart, and bars are ascendingly sorted from left to right according to the DNN models.

Additionally, when it comes to the different levels of information encoded in the fingerprints (Figure 3.27), the usage of non-covalent and atom-atom interactions are encouraged as they provide a clear advantage over fingerprints with no contact information (experiment label $Levels = 1, Radius = 0.00$). However, when we also included proximal contacts, the performance of all models decreased. By default, proximal contacts are assigned to all pairs of atoms or atom groups within 6 Å, including intramolecular pairs. Consequently, an excessive number of features, especially false positive ones, are produced by such a method. For that reason (although not proved herein), we believe a short-range threshold for proximal contacts should be used, whereas intramolecular contacts should be avoided.

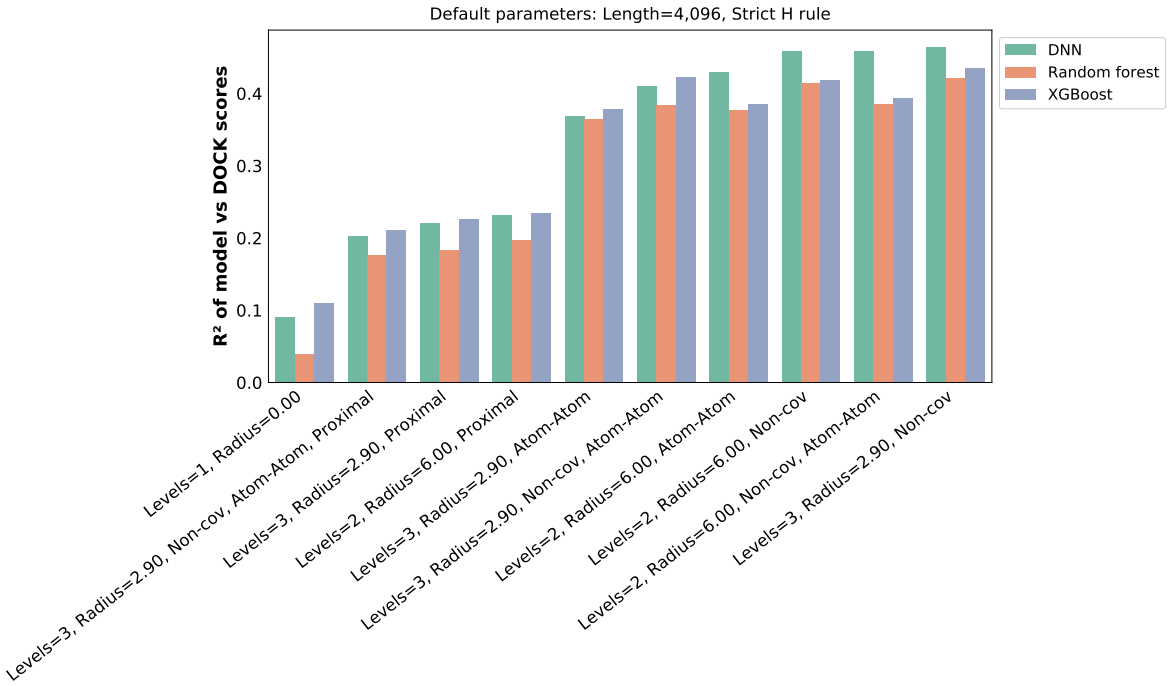


Figure 3.27: Comparison between different methodologies for computing interactions. Default FIFP parameters are shown above the chart, and bars are ascendingly sorted from left to right according to the DNN models.

Following the fine-tuning of the methodology, we then evaluated the importance of the proper addition of hydrogens to the molecules. However, since the ligand contains hydrogens, the hydrogen correction was performed only for the protein. Note in Figure 3.28 that the R^2 slightly increases when hydrogen atoms are added to the structure. So far, we have used strict rules without including hydrogens on the protein side. However, remember that strict rules require all atoms to be protonated appropriately; otherwise, (weak) hydrogen bonds are not identified. Thus, the observed improvement happened because our models were not capturing interactions in which the donor belongs to the protein side.

Concerning the count fingerprints, Figure 3.29 shows that its usage favorably contributes to the predictive task. These fingerprints, as the very name indicates, explicitly

Figure 3.28: Comparison between fingerprints with and without hydrogens added to the protein structure. Default FIFP parameters are shown above the chart, and bars are ascendingly sorted from left to right according to the DNN models.

define the frequency in which a specific feature appeared in a complex instead of only pointing out its presence or not. In other words, it means that explicitly defining how many times a feature consisting of a hydrogen bond, for instance, is more promising than only accounting for its presence. Moreover, when interactions (both non-covalent and atom-atom interactions) from the protein side are also encoded in count fingerprints (Figure 3.29), we observed a further remarking improvement in the models, where the best 4,096-bits model achieved an R^2 of 0.52.

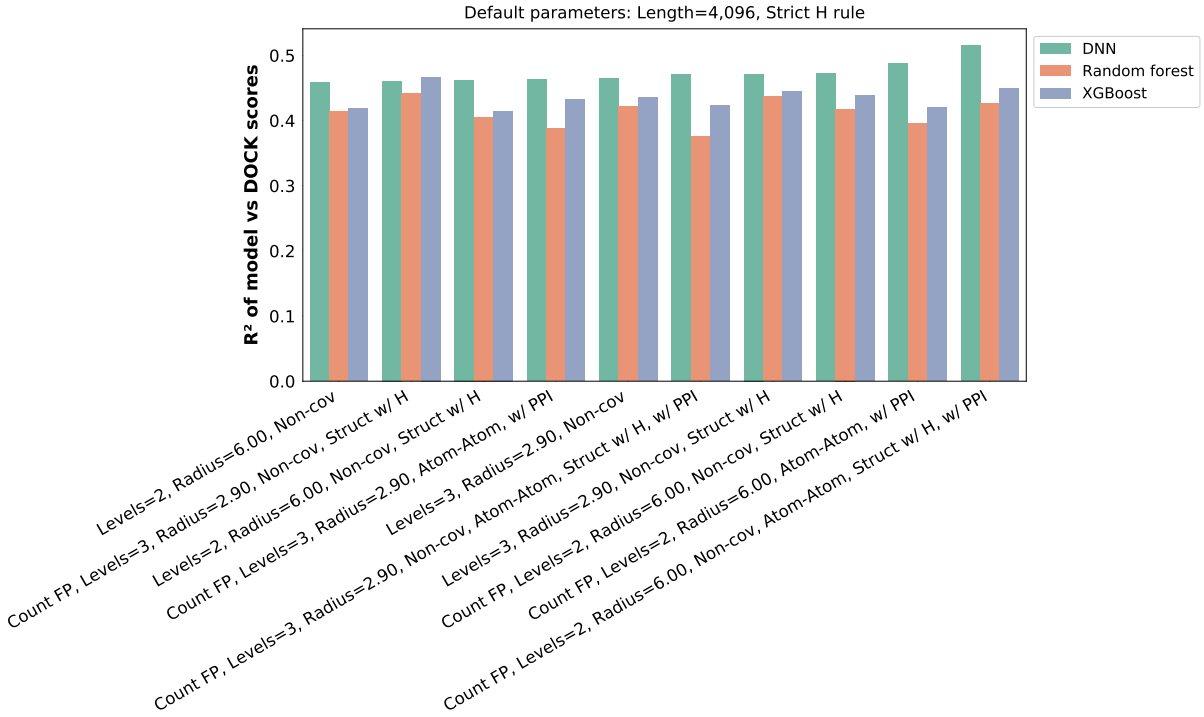


Figure 3.29: Comparison between bit and count fingerprints and contribution of interactions in the protein side. Default FIFP parameters are shown above the chart, and bars are ascendingly sorted from left to right according to the DNN models.

Finally, Figure 3.30 points out that longer fingerprints perform better than their corresponding 4,096 versions (only models with $R^2 > 0.48$). For example, the best model (most-right grouped bars) in Figure 3.30, a 16,384-bits version of the best 4,096-bits model discussed in the previous paragraph (fourth grouped bars from right to left in Figure 3.30), achieved an R^2 of 0.56 against 0.52 of its shorter version. Together with results presented in Section 3.3.2, we believe that the choice for fingerprints up to 16,384-bits may provide the best tradeoff between predictive performance, collision rate minimization, and model reliability. The latter property is mainly related to the overfitting problem, which refers to models that cannot generalize for unseen data and whose occurrence is typically higher for longer fingerprints.

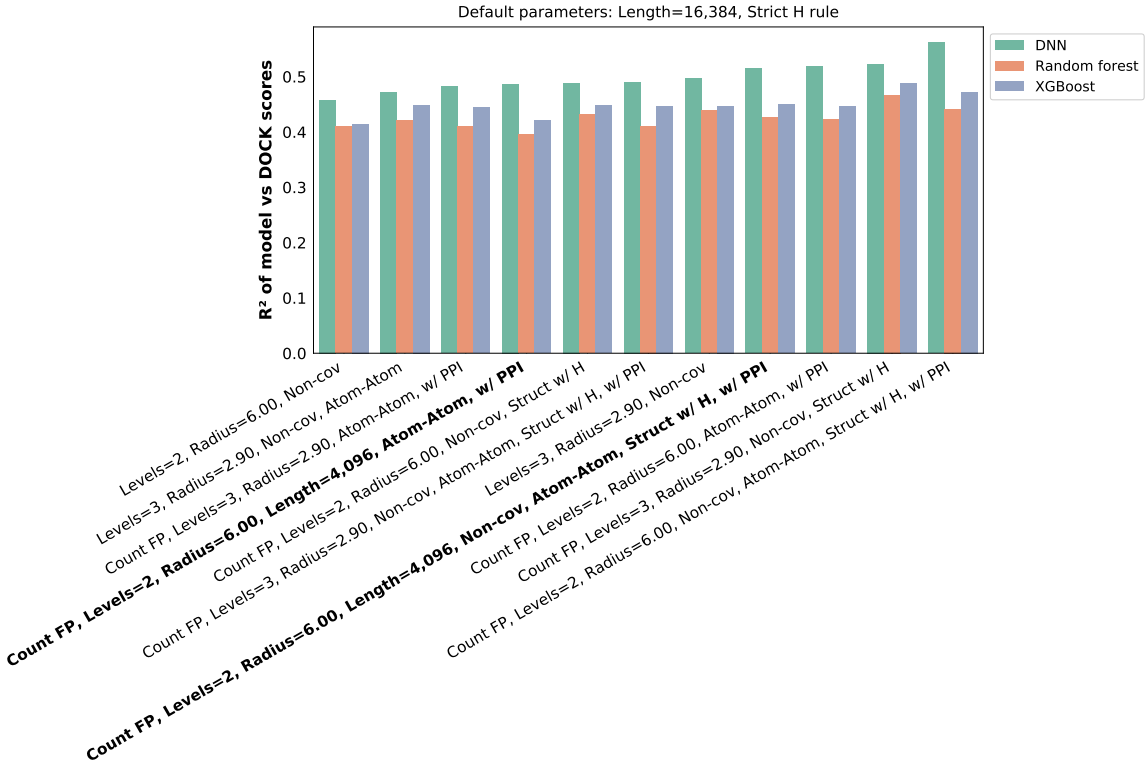


Figure 3.30: Comparison between the top 4,096-bits fingerprints ($R^2 > 0.48$; highlighted in bold) and 16,384-bits version. Default FIFP parameters are shown above the chart, and bars are ascendingly sorted from left to right according to the DNN models.

3.4.2 Baseline comparison

After the exploratory search for the best models and FIFP parameters, we selected those models whose R^2 was higher than 0.5 and compared them to the baseline (ECFP and FCFP) and two other interaction fingerprint models using a 5-fold cross-validation

strategy, whose results are shown in Figure 3.31.

Given that the molecular recognition event contains the required information for the binding affinity prediction, we hypothesized that FIFP should perform at least as well as ECFP and FCFP, which are two molecular fingerprints that only encode ligand information. Indeed, the results presented in Figure 3.31 support our hypothesis as FIFP outperformed both fingerprints, especially employing DNN strategies.

Regarding the interaction fingerprints, SILIRID showed not to be a promising approach on the Dock score prediction task (left-most grouped bars in Figure 3.31). SILIRID is a fixed-length fingerprint that summarizes protein-ligand complexes into a 168-bit vector, where each 8-bits chunk (1 bit per interaction type) is reserved for one amino acid and one additional chunk is reserved for cofactors, totalizing 21 chunks ($21 * 8 = 168$). However, we believe such a summary does not properly represent the protein-ligand interaction context for a Dock score prediction task. FOR INSTANCE, even ECFP and FCFP, which only encodes the ligand information, presented a superior performance than SILIRID. Overall, these findings emphasize how critical data representation is for success on machine learning tasks, although we highlight that SILIRID has been mainly conceived for binding site comparison only.

PLEC, on the other hand, presented the second-best performance with its 16,384-bits version, with consistent results for all machine learning algorithms (Figure 3.31). The only other model that performed better than PLEC-16,384 was the DNN model trained with FIFP count fingerprints using interactions in the protein side, number of levels and radius rate growth equal to 2 and 6, respectively. Other FIFP fingerprints presented slightly smaller results than PLEC-16,384, but the differences were not significant.

In summary, PLEC is a recently-published interaction fingerprint based on molecular topology information from ECFP. However, unlike ECFP, PLEC only encodes the environment of protein and ligand atoms in contact, which is defined by default as all atoms within 4.5 Å. Thus, it is noteworthy that PLEC achieved remarking R^2 scores by only encoding contact information coupled with topology information.

Although our approach is based on the same circular neighborhood expansion from ECFP, the current version of FIFP more closely resembles FCFP. The latter is a variation of ECFP and encodes pharmacophore properties instead of the so-called atomic invariants (the atomic number, isotope, number of neighboring heavy atoms, number of hydrogens, formal charge, ring membership, and aromaticity). Nonetheless, it has already been shown that ECFP usually performs better than FCFP [200] and, therefore, we believe that the definition of atomic invariants further contributes to the prediction task inasmuch it accounts for the specificity of each atom. On the other hand, the pharmacophore version considers that atoms with the same pharmacophore property are functionally equivalent, which seems not to add enough specificity for such a prediction problem. Given these findings, we plan to build up a FIFP version where atomic invariants are

explicitly considered and evaluate whether our models can further improve with such an update.

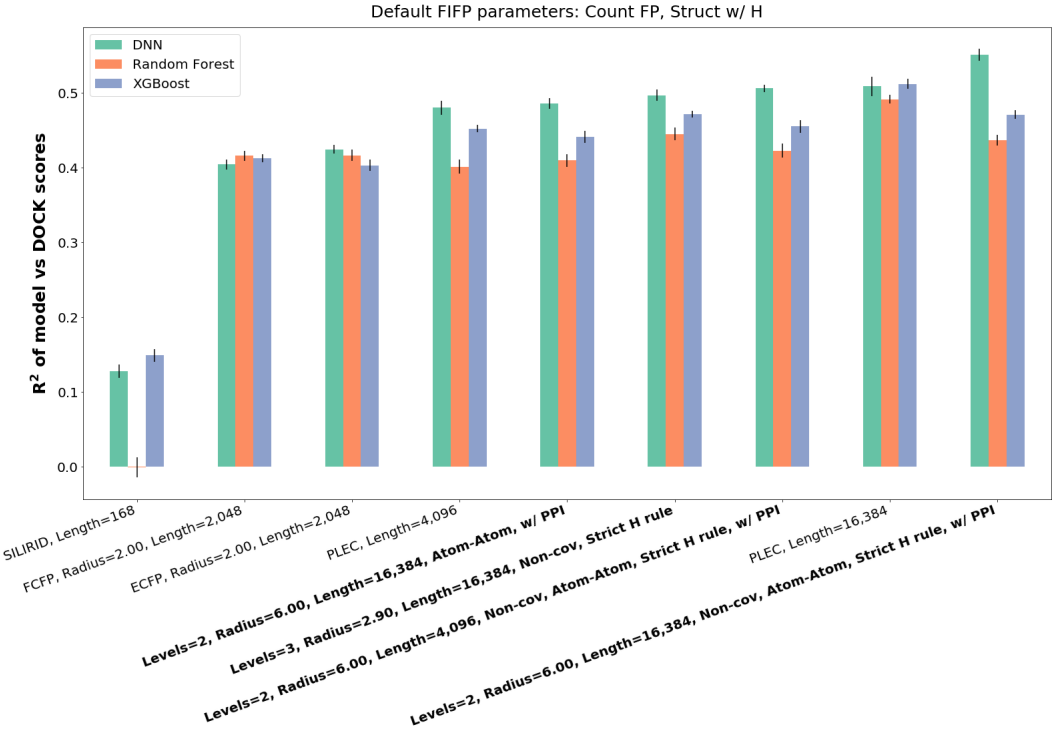


Figure 3.31: Comparison between the top FIFP models ($R^2 > 0.5$; highlighted in bold) against the baseline (ECFP and FCFP) and two other interaction fingerprint models using 5-fold cross-validation. Default FIFP parameters are shown above the chart, and bars are ascendingly sorted from left to right according to the DNN models.

Chapter 4

Conclusion

In this work, we undertook to address the problem of identification, prioritization, and automatic selection of a small number of promising compounds (HITs) in a structural-based virtual screening campaign through a descriptive and predictive perspective.

For the first aspect, we propose nAPOLI (Analysis of PrOtein-Ligand Interactions), a webserver to perform a large-scale analysis of protein-ligand interactions, consisting of a set of algorithms and interactive visual interfaces to analyze and explore comprehensive reports on conserved interactions in complexes. nAPOLI allows domain specialists to detect conserved atomic-level interactions in labeled bipartite graphs representing protein-ligand interfaces using visual strategies and statistical analysis. Furthermore, users can analyze different structural data sets by creating several projects either by submitting their local structures in PDB file format or by using the structures available at the PDB. By providing ways to characterize and analyze protein-ligand interaction patterns across large data sets of protein-ligand complexes, nAPOLI supports the understanding of processes and patterns involved in molecular recognition and permits users to select and filter compounds in a virtual screening campaign an interactive, visual, and analytical manner.

For the predictive aspect, we propose LUNA, a novel Python library for drug design that permits the analysis of multiple molecular complex types, including protein-protein, protein-DNA/RNA, protein-small molecule, and others. Moreover, the tool brings together several functions for filtering and visualizing interactions, generating statistical data, and characterizing a data set. LUNA also provides connectors to MySQL databases, RDKit, Open Babel, and Pymol. Another remarkable contribution of this work is the comprehensive expansion of methods for calculating interactions and physicochemical assignment rules.

Finally, we also propose a novel hashed interaction fingerprint called FIFP (Functional InteracTion FingerPrint), which was inspired by ECFP [203] and E3FP [9]. Different from other hashed fingerprints that are usually black-boxes, FIFP also provides several features to make fingerprint information analysis straightforward and out-of-the-box. To validate and illustrate the applicability of FIFP, we first presented an exploratory evaluation of the fingerprint parameters to identify the best combination of parameters and understand the fingerprint behavior when varying each parameter. Afterwards, we

presented a case study that applied FIFP to the Dock score prediction task. To do so, we built a data set composed of 86,641 molecules docked against Dopamine D4 departing from [153]. Following, we trained several DNN, Random forest, and XGBoost models using different combinations of FIFP fingerprints to identify the set of parameters that provides the best prediction performance. We then compared the best results with two baseline models (ECFP and FCFP) and two other interaction fingerprint models (SILIRID and PLEC). As a result, we showed that FIFP outperformed the competing approaches with an R^2 of 0.56.

Therefore, we envision LUNA and FIFP as remarking approaches for structure-based virtual screening and molecular dynamics campaigns. Additionally, they show promising applicability in machine learning tasks like classifying molecules as active or inactive, identifying bad poses, and predicting docking scores and experimental binding affinity.

Lastly, as future works, we plan to include novel features in LUNA and additional molecular interactions, namely anion-π, disulfide bond, agostic bond, hydrogen bonds with metals, metal complex, and aromatic stackings with arrays of hydrogen bonds. Regarding FIFP, we plan to propose a new FIFP flavor where atomic invariants are explicitly considered as in ECFP. Last but not least, we also plan to evaluate FIFP on different scenarios as, for instance, the classification of ligands into two (active or inactive) or more classes.

Bibliography

[1] William J. Allen, Trent E. Balius, Sudipto Mukherjee, Scott R. Brozell, Demetri T. Moustakas, P. Therese Lang, David A. Case, Irwin D. Kuntz, and Robert C. Rizzo. DOCK 6: Impact of New Features and Current Docking Performance. *Journal of computational chemistry*, 36(15):1132–1156, June 2015. ISSN 0192-8651. doi: 10.1002/jcc.23905. URL http://www.ncbi.nlm.nih.gov/pmc/articles/PMC4469538/.

[2] Mark R. Ams, Nils Trapp, Anatol Schwab, Jovana V. Milić, and François Diederich. Chalcogen Bonding "2s–2n Squares" versus Competing Interactions: Exploring the Recognition Properties of Sulfur. *Chemistry – A European Journal*, December 2018. ISSN 0947-6539, 1521-3765. doi: 10.1002/chem.201804261. URL https://onlinelibrary.wiley.com/doi/abs/10.1002/chem.201804261.

[3] Praveen Anand, Deepesh Nagarajan, Sumanta Mukherjee, and Nagasuma Chandra. PLIC: protein-ligand interaction clusters. *Database: The Journal of Biological Databases and Curation*, 2014(0):bau029, 2014. ISSN 1758-0463. doi: 10.1093/database/bau029.

[4] Rosaleen J. Anderson, Paul W. Groundwater, and Adam Todd, editors. *Antibacterial agents: chemistry, mode of action, mechanisms of resistance, and clinical applications*. John Wiley & Sons, Chichester, West Sussex, 2012. ISBN 978-0-470-97244-1 978-0-470-97245-8.

[5] Munazah Andrabi, Chioko Nagao, Kenji Mizuguchi, and Shandar Ahmad. Bioinformatics Approaches for Analysis of Protein–Ligand Interactions. In Konstantin V. Balakin, editor, *Pharmaceutical Data Mining*, pages 267–299. John Wiley & Sons, Inc., 2009. ISBN 978-0-470-56762-3. URL http://onlinelibrary.wiley.com/doi/10.1002/9780470567623.ch9/summary. DOI: 10.1002/9780470567623.ch9.

[6] Austin Appleby. MurmurHash3, September 2016. URL https://github.com/aappleby/smhasher. [Online]. Available: https://github.com/aappleby/smhasher/. Accessed: 2019-09-01.

[7] Paolo Ascenzi, Alessio Bocedi, and Maria Marino. Structure-function relationship of estrogen receptor alpha and beta: impact on human health. *Molecular Aspects of Medicine*, 27(4):299–402, August 2006. ISSN 0098-2997. doi: 10.1016/j.mam.2006.07.001.

[8] Pascal Auffinger, Franklin A. Hays, Eric Westhof, and P. Shing Ho. Halogen bonds in biological molecules. *Proceedings of the National Academy of Sciences of the United States of America*, 101(48):16789–16794, November 2004. ISSN 0027-8424. doi: 10.1073/pnas.0407607101. URL http://www.ncbi.nlm.nih.gov/pmc/articles/PMC529416/.

[9] Seth D. Axen, Xi-Ping Huang, Elena L. Cáceres, Leo Gendelev, Bryan L. Roth, and Michael J. Keiser. A Simple Representation of Three-Dimensional Molecular Structure. *Journal of Medicinal Chemistry*, 60(17):7393–7409, September 2017. ISSN 0022-2623. doi: 10.1021/acs.jmedchem.7b00696. URL https://doi.org/10.1021/acs.jmedchem.7b00696.

[10] Marc Baaden and Siewert J Marrink. Coarse-grain modelling of protein–protein interactions. *Current Opinion in Structural Biology*, 23(6):878–886, December 2013. ISSN 0959-440X. doi: 10.1016/j.sbi.2013.09.004. URL http://www.sciencedirect.com/science/article/pii/S0959440X13001735.

[11] Yan Bai, Arthur F. Monzingo, and Jon D. Robertus. The X-ray structure of ricin A chain with a novel inhibitor. *Archives of Biochemistry and Biophysics*, 483(1):23–28, March 2009. ISSN 1096-0384. doi: 10.1016/j.abb.2008.12.013.

[12] D. Bajusz, A. Rácz, and K. Héberger. 3.14 - Chemical Data Formats, Fingerprints, and Other Molecular Descriptions for Database Analysis and Searching. In Samuel Chackalamannil, David Rotella, and Simon E. Ward, editors, *Comprehensive Medicinal Chemistry III*, pages 329–378. Elsevier, Oxford, January 2017. ISBN 978-0-12-803201-5. doi: 10.1016/B978-0-12-409547-2.12345-5. URL http://www.sciencedirect.com/science/article/pii/B9780124095472123455.

[13] E. N. Baker and R. E. Hubbard. Hydrogen bonding in globular proteins. *Progress in Biophysics and Molecular Biology*, 44(2):97–179, 1984. ISSN 0079-6107.

[14] Trent E. Balius, Sudipto Mukherjee, and Robert C. Rizzo. Implementation and Evaluation of a Docking-Rescoring Method using Molecular Footprint Comparisons. *Journal of computational chemistry*, 32(10):2273–2289, July 2011. ISSN 0192-8651. doi: 10.1002/jcc.21814. URL http://www.ncbi.nlm.nih.gov/pmc/articles/PMC3181325/.

[15] D. J. Barlow and J. M. Thornton. Ion-pairs in proteins. *Journal of Molecular Biology*, 168(4):867–885, August 1983. ISSN 0022-2836.

[16] John M. Barnard and Geoff M. Downs. Chemical Fragment Generation and Clustering Software. *Journal of Chemical Information and Computer Sciences*,

37(1):141–142, January 1997. ISSN 0095-2338. doi: 10.1021/ci960090k. URL https://doi.org/10.1021/ci960090k.

[17] Antonio Bauzá, David Quiñonero, Pere M. Deyà, and Antonio Frontera. Pnicogen–π complexes: theoretical study and biological implications. *Physical Chemistry Chemical Physics*, 14(40):14061–14066, September 2012. ISSN 1463-9084. doi: 10.1039/C2CP42672B. URL https://pubs.rsc.org/en/content/articlelanding/2012/cp/c2cp42672b.

[18] Antonio Bauzá, Tiddo J. Mooibroek, and Antonio Frontera. The Bright Future of Unconventional σ/π-Hole Interactions. *ChemPhysChem*, 16(12):2496–2517, August 2015. ISSN 14394235. doi: 10.1002/cphc.201500314. URL http://doi.wiley.com/10.1002/cphc.201500314.

[19] Andreas Bender, Hamse Y. Mussa, Robert C. Glen, and Stephan Reiling. Similarity searching of chemical databases using atom environment descriptors (MOLPRINT 2d): evaluation of performance. *Journal of Chemical Information and Computer Sciences*, 44(5):1708–1718, October 2004. ISSN 0095-2338. doi: 10.1021/ci0498719.

[20] Brett R. Beno, Kap-Sun Yeung, Michael D. Bartberger, Lewis D. Pennington, and Nicholas A. Meanwell. A Survey of the Role of Noncovalent Sulfur Interactions in Drug Design. *Journal of Medicinal Chemistry*, 58(11):4383–4438, June 2015. ISSN 0022-2623. doi: 10.1021/jm501853m. URL https://doi.org/10.1021/jm501853m.

[21] H. M. Berman, J. Westbrook, Z. Feng, G. Gilliland, T. N. Bhat, H. Weissig, I. N. Shindyalov, and P. E. Bourne. The Protein Data Bank. *Nucleic Acids Research*, 28 (1):235–242, January 2000. ISSN 0305-1048.

[22] Samuel Bertrand, Jean-Jacques Hélesbeux, Gérald Larcher, and Olivier Duval. Hydroxamate, a key pharmacophore exhibiting a wide range of biological activities. *Mini Reviews in Medicinal Chemistry*, 13(9):1311–1326, July 2013. ISSN 1875-5607.

[23] Rajasri Bhattacharyya, Rudra Prasad Saha, Uttamkumar Samanta, and Pinak Chakrabarti. Geometry of interaction of the histidine ring with other planar and basic residues. *Journal of Proteome Research*, 2(3):255–263, June 2003. ISSN 1535-3893.

[24] George R. Bickerton, Alicia P. Higueruelo, and Tom L. Blundell. Comprehensive, atomic-level characterization of structurally characterized protein-protein interactions: the PICCOLO database. *BMC bioinformatics*, 12:313, July 2011. ISSN 1471-2105. doi: 10.1186/1471-2105-12-313.

[25] Caterina Bissantz, Bernd Kuhn, and Martin Stahl. A medicinal chemist's guide to molecular interactions. *Journal of Medicinal Chemistry*, 53(14):5061–5084, July 2010. ISSN 1520-4804. doi: 10.1021/jm100112j.

[26] Randy K. Bledsoe, Kevin P. Madauss, Jason A. Holt, Christopher J. Apolito, Millard H. Lambert, Kenneth H. Pearce, Thomas B. Stanley, Eugene L. Stewart, Ryan P. Trump, Timothy M. Willson, and Shawn P. Williams. A Ligand-mediated Hydrogen Bond Network Required for the Activation of the Mineralocorticoid Receptor. *Journal of Biological Chemistry*, 280(35):31283–31293, September 2005. ISSN 0021-9258, 1083-351X. doi: 10.1074/jbc.M504098200. URL http://www.jbc.org/lookup/doi/10.1074/jbc.M504098200.

[27] Guillaume Bouvier, Nathalie Evrard-Todeschi, Jean-Pierre Girault, and Gildas Bertho. Automatic clustering of docking poses in virtual screening process using self-organizing map. *Bioinformatics (Oxford, England)*, 26(1):53–60, January 2010. ISSN 1367-4811. doi: 10.1093/bioinformatics/btp623.

[28] Janice L. Bradley and Peter M. McGUIRE. Site-directed mutagenesis of ricin A chain Trp 211 to Phe. *International Journal of Peptide and Protein Research*, 35 (4):365–366, April 1990. ISSN 1399-3011. doi: 10.1111/j.1399-3011.1990.tb00062.x.

[29] G. Bricogne, E. Blanc, M. Brandl, C. Flensburg, P. Keller, W. Paciorek, P Roversi, A. Sharff, O.S. Smart, C. Vonrhein, and Womack T.O. BUSTER, 2011. Versão 2.10.1. Cambridge, United Kingdom: Global Phasing Ltd. https://www.globalphasing.com/buster/wiki/index.cgi?. Acessado em: 13/06/2015.

[30] Natasja Brooijmans and Irwin D. Kuntz. Molecular recognition and docking algorithms. *Annual Review of Biophysics and Biomolecular Structure*, 32:335–373, 2003. ISSN 1056-8700. doi: 10.1146/annurev.biophys.32.110601.142532.

[31] B.R. Brooks, C.L. Brooks, A.D. MacKerell, L. Nilsson, R.J. Petrella, B. Roux, Y. Won, G. Archontis, C. Bartels, S. Boresch, A. Caflisch, L. Caves, Q. Cui, A.R. Dinner, M. Feig, S. Fischer, J. Gao, M. Hodoscek, W. Im, K. Kuczera, T. Lazaridis, J. Ma, V. Ovchinnikov, E. Paci, R.W. Pastor, C.B. Post, J.Z. Pu, M. Schaefer, B. Tidor, R. M. Venable, H. L. Woodcock, X. Wu, W. Yang, D.M. York, and M. Karplus. CHARMM: The Biomolecular Simulation Program. *Journal of computational chemistry*, 30(10):1545–1614, July 2009. ISSN 0192-8651. doi: 10.1002/jcc.21287. URL http://www.ncbi.nlm.nih.gov/pmc/articles/PMC2810661/.

[32] Theodore L. Brown. *Chemistry: the central science*. Pearson, Boston, thirteen edition edition, 2015. ISBN 978-0-321-91041-7 978-0-321-96239-3.

[33] A. M. Brzozowski, A. C. Pike, Z. Dauter, R. E. Hubbard, T. Bonn, O. Engström, L. Ohman, G. L. Greene, J. A. Gustafsson, and M. Carlquist. Molecular basis of agonism and antagonism in the oestrogen receptor. *Nature*, 389(6652):753–758, October 1997. ISSN 0028-0836. doi: 10.1038/39645.

[34] Hans-Joachim Böhm, Stefan Brode, Ute Hesse, and Gerhard Klebe. Oxygen and Nitrogen in Competitive Situations: Which is the Hydrogen-Bond Acceptor? *Chemistry – A European Journal*, 2(12):1509–1513, 1996. ISSN 1521-3765. doi: 10.1002/chem.19960021206. URL https://onlinelibrary.wiley.com/doi/abs/10.1002/chem.19960021206.

[35] Stanislav Böhm and Otto Exner. Acidity of hydroxamic acids and amides. *Organic & Biomolecular Chemistry*, 1(7):1176–1180, April 2003. ISSN 1477-0520.

[36] Ségolène Caboche. LeView: automatic and interactive generation of 2d diagrams for biomacromolecule/ligand interactions. *Journal of Cheminformatics*, 5(1):40, August 2013. ISSN 1758-2946. doi: 10.1186/1758-2946-5-40.

[37] Raymond E. Carhart, Dennis H. Smith, and R. Venkataraghavan. Atom pairs as molecular features in structure-activity studies: definition and applications. *Journal of Chemical Information and Computer Sciences*, 25(2):64–73, May 1985. ISSN 0095-2338. doi: 10.1021/ci00046a002. URL https://pubs.acs.org/doi/abs/10.1021/ci00046a002.

[38] John H. Carra, Colleen A. McHugh, Sheila Mulligan, LeeAnn M. Machiesky, Alexei S. Soares, and Charles B. Millard. Fragment-based identification of determinants of conformational and spectroscopic change at the ricin active site. *BMC Structural Biology*, 7(1):72, November 2007. ISSN 1472-6807. doi: 10.1186/1472-6807-7-72. URL https://doi.org/10.1186/1472-6807-7-72.

[39] Gabriella Cavallo, Pierangelo Metrangolo, Roberto Milani, Tullio Pilati, Arri Priimagi, Giuseppe Resnati, and Giancarlo Terraneo. The Halogen Bond. *Chemical Reviews*, 116(4):2478–2601, February 2016. ISSN 0009-2665. doi: 10.1021/acs.chemrev.5b00484. URL http://dx.doi.org/10.1021/acs.chemrev.5b00484.

[40] CGAL. CGAL, Computational Geometry Algorithms Library, 1995. [Online]. Available: http://www.cgal.org/. Accessed: 2019-09-01.

[41] J. A. Chaddock and L. M. Roberts. Mutagenesis and kinetic analysis of the active site Glu177 of ricin A-chain. *Protein Engineering*, 6(4):425–431, June 1993. ISSN 0269-2139.

[42] Pinak Chakrabarti and Rajasri Bhattacharyya. Geometry of nonbonded interactions involving planar groups in proteins. *Progress in Biophysics and Molecular Biology*,

95(1-3):83–137, November 2007. ISSN 0079-6107. doi: 10.1016/j.pbiomolbio.2007.03.016.

[43] Pinak Chakrabarti and Sarmistha Chakrabarti. C—H··· O hydrogen bond involving proline residues in α-helices. *Journal of molecular biology*, 284(4): 867–873, 1998. URL http://www.sciencedirect.com/science/article/pii/S0022283698921994.

[44] Hongming Chen, Ola Engkvist, Yinhai Wang, Marcus Olivecrona, and Thomas Blaschke. The rise of deep learning in drug discovery. *Drug Discovery Today*, 23(6): 1241–1250, June 2018. ISSN 1359-6446. doi: 10.1016/j.drudis.2018.01.039. URL http://www.sciencedirect.com/science/article/pii/S1359644617303598.

[45] Hsiao Ying Chen, Ling Yann Foo, and Weng Keong Loke. Ricin and abrin: A comprehensive review of their toxicity, diagnosis, and treatment. In P. Gopalakrishnakone, editor, *Toxinology*, pages 1–20. Springer Netherlands, 2014. ISBN 978-94-007-6645-7. doi: 10.1007/978-94-007-6645-7_1-1.

[46] Jonathan H. Chen, Erik Linstead, S. Joshua Swamidass, Dennis Wang, and Pierre Baldi. ChemDB update–full-text search and virtual chemical space. *Bioinformatics (Oxford, England)*, 23(17):2348–2351, September 2007. ISSN 1367-4811. doi: 10.1093/bioinformatics/btm341.

[47] Tianqi Chen and Carlos Guestrin. XGBoost: A Scalable Tree Boosting System. In *Proceedings of the 22Nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '16, pages 785–794, New York, NY, USA, 2016. ACM. ISBN 978-1-4503-4232-2. doi: 10.1145/2939672.2939785. URL http://doi.acm.org/10.1145/2939672.2939785. event-place: San Francisco, California, USA.

[48] Vladimir Chupakhin, Gilles Marcou, Helena Gaspar, and Alexandre Varnek. Simple Ligand-Receptor Interaction Descriptor (SILIRID) for alignment-free binding site comparison. *Computational and Structural Biotechnology Journal*, 10(16):33–37, June 2014. ISSN 2001-0370. doi: 10.1016/j.csbj.2014.05.004.

[49] Alex M. Clark and Paul Labute. 2d Depiction of Protein-Ligand Complexes. *Journal of Chemical Information and Modeling*, 47(5):1933–1944, September 2007. ISSN 1549-9596. doi: 10.1021/ci7001473. URL http://dx.doi.org/10.1021/ci7001473.

[50] Timothy Clark. σ-Holes. *Wiley Interdisciplinary Reviews: Computational Molecular Science*, 3(1):13–20, January 2013. ISSN 17590876. doi: 10.1002/wcms.1113. URL http://doi.wiley.com/10.1002/wcms.1113.

[51] Isidro Cortés-Ciriano, Nicholas C. Firth, Andreas Bender, and Oliver Watson. Discovering Highly Potent Molecules from an Initial Set of Inactives Using Iterative Screening. *Journal of Chemical Information and Modeling*, 58(9):2000–2014, September 2018. ISSN 1549-9596. doi: 10.1021/acs.jcim.8b00376. URL https://doi.org/10.1021/acs.jcim.8b00376.

[52] Simona Cotesta and Martin Stahl. The environment of amide groups in protein–ligand complexes: H-bonds and beyond. *Journal of Molecular Modeling*, 12 (4):436–444, March 2006. ISSN 0948-5023. doi: 10.1007/s00894-005-0067-x. URL https://doi.org/10.1007/s00894-005-0067-x.

[53] C. Da and D. Kireev. Structural Protein–Ligand Interaction Fingerprints (SPLIF) for Structure-Based Virtual Screening: Method and Benchmark Study. *Journal of Chemical Information and Modeling*, 54(9):2555–2561, September 2014. ISSN 1549-9596. doi: 10.1021/ci500319f. URL http://dx.doi.org/10.1021/ci500319f.

[54] Franck Da Silva, Jeremy Desaphy, and Didier Rognan. IChem: A Versatile Toolkit for Detecting, Comparing, and Predicting Protein–Ligand Interactions. *ChemMedChem*, 13(6):507–510, 2018. ISSN 1860-7187. doi: 10.1002/cmdc.201700505. URL https://onlinelibrary.wiley.com/doi/abs/10.1002/cmdc.201700505.

[55] Carlos H. da Silveira, Douglas E. V. Pires, Raquel C. Minardi, Cristina Ribeiro, Caio J. M. Veloso, Julio C. D. Lopes, Wagner Meira, Goran Neshich, Carlos H. I. Ramos, Raul Habesch, and Marcelo M. Santoro. Protein cutoff scanning: A comparative analysis of cutoff dependent and cutoff free methods for prospecting contacts in proteins. *Proteins: Structure, Function, and Bioinformatics*, 74(3):727–743, February 2009. ISSN 08873585, 10970134. doi: 10.1002/prot.22187. URL http://doi.wiley.com/10.1002/prot.22187.

[56] Claudio Dalvit, Christian Invernizzi, and Anna Vulpetti. Fluorine as a Hydrogen-Bond Acceptor: Experimental Evidence and Computational Calculations. *Chemistry - A European Journal*, 20(35):11058–11068, August 2014. ISSN 09476539. doi: 10.1002/chem.201402858. URL http://doi.wiley.com/10.1002/chem.201402858.

[57] Andrea R Daniel, Christy R Hagan, and Carol A Lange. Progesterone receptor action: defining a role in breast cancer. *Expert review of endocrinology & metabolism*, 6(3):359–369, May 2011. ISSN 1744-6651. doi: 10.1586/eem.11.25. URL https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3156468/.

[58] P. J. Day, S. R. Ernst, A. E. Frankel, A. F. Monzingo, J. M. Pascal, M. C. Molina-Svinth, and J. D. Robertus. Structure and activity of an active site substitution

of ricin A chain. *Biochemistry*, 35(34):11098–11103, August 1996. ISSN 0006-2960. doi: 10.1021/bi960880n.

[59] Daylight. Daylight theory manual, 2011. Disponível em: http://www.daylight.com/dayhtml/doc/theory/. Acessado em: 15/06/17.

[60] Tjaart A. P. de Beer, Karel Berka, Janet M. Thornton, and Roman A. Laskowski. PDBsum additions. *Nucleic Acids Research*, 42(Database issue):D292–D296, January 2014. ISSN 0305-1048. doi: 10.1093/nar/gkt940. URL http://www.ncbi.nlm.nih.gov/pmc/articles/PMC3965036/.

[61] Zhan Deng, Claudio Chuaqui, and Juswinder Singh. Structural interaction fingerprint (SIFt): a novel method for analyzing three-dimensional protein-ligand binding interactions. *Journal of Medicinal Chemistry*, 47(2):337–344, January 2004. ISSN 0022-2623. doi: 10.1021/jm030331x.

[62] Z. S. Derewenda, L. Lee, and U. Derewenda. The occurrence of C-H...O hydrogen bonds in proteins. *Journal of Molecular Biology*, 252(2):248–262, September 1995. ISSN 0022-2836.

[63] Jérémy Desaphy, Eric Raimbaud, Pierre Ducrot, and Didier Rognan. Encoding protein-ligand interaction patterns in fingerprints and graphs. *Journal of Chemical Information and Modeling*, 53(3):623–637, March 2013. ISSN 1549-960X. doi: 10.1021/ci300566n.

[64] Jérémy Desaphy, Guillaume Bret, Didier Rognan, and Esther Kellenberger. sc-PDB: a 3d-database of ligandable binding sites—10 years on. *Nucleic Acids Research*, 43 (Database issue):D399–D404, January 2015. ISSN 0305-1048. doi: 10.1093/nar/gku928. URL http://www.ncbi.nlm.nih.gov/pmc/articles/PMC4384012/.

[65] Gautam R. Desiraju. C-H...O and other weak hydrogen bonds. From crystal engineering to virtual screening. *Chemical Communications (Cambridge, England)*, (24):2995–3001, June 2005. ISSN 1359-7345. doi: 10.1039/b504372g.

[66] Gautam R. Desiraju and Thomas Steiner. *The weak hydrogen bond: in structural chemistry and biology*. Number 9 in International Union of Crystallography monographs on crystallography. Oxford University Press, Oxford, first publ. in paperback edition, 2001. ISBN 978-0-19-850970-7. OCLC: 248364161.

[67] Vineet K. Dhiman, Michael J. Bolt, and Kevin P. White. Nuclear receptors in cancer — uncovering new and evolving roles through genomic analysis. *Nature Reviews Genetics*, 19(3):160–174, December 2017. ISSN 1471-0056, 1471-0064. doi: 10.1038/nrg.2017.102. URL http://www.nature.com/doifinder/10.1038/nrg.2017.102.

[68] Caroline H. Diep, Andrea R. Daniel, Laura J. Mauro, Todd P. Knutson, and Carol A. Lange. Progesterone action in breast, uterine, and ovarian cancers. *Journal of molecular endocrinology*, 54(2):R31–R53, April 2015. ISSN 0952-5041. doi: 10.1530/JME-14-0252. URL https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4336822/.

[69] Yun Ding, Ye Fang, Juana Moreno, J. Ramanujam, Mark Jarrell, and Michal Brylinski. Assessing the similarity of ligand binding conformations with the Contact Mode Score. *Computational Biology and Chemistry*, 64:403–413, 2016. ISSN 1476-928X. doi: 10.1016/j.compbiolchem.2016.08.007.

[70] Nikolay V Dokholyan. *Computational modeling of biological systems: from molecules to pathways*. Springer, New York, 2012. ISBN 978-1-4614-2145-0 978-1-4614-2146-7. OCLC: 883719271.

[71] Jing Dong, Yong Zhang, Yutao Chen, Xiaodi Niu, Yu Zhang, Rui Li, Cheng Yang, Quan Wang, Xuemei Li, and Xuming Deng. Baicalin inhibits the lethality of ricin in mice by inducing protein oligomerization. *The Journal of Biological Chemistry*, 290(20):12899–12907, May 2015. ISSN 1083-351X. doi: 10.1074/jbc.M114.632828.

[72] D. A. Dougherty and D. A. Stauffer. Acetylcholine binding by a synthetic receptor: implications for biological recognition. *Science (New York, N.Y.)*, 250(4987):1558–1560, December 1990. ISSN 0036-8075.

[73] Malgorzata N. Drwal and Renate Griffith. Combination of ligand- and structure-based methods in virtual screening. *Drug Discovery Today. Technologies*, 10(3): e395–401, September 2013. ISSN 1740-6749. doi: 10.1016/j.ddtec.2013.02.002.

[74] Charles B. Duke, Amanda Jones, Casey E. Bohl, James T. Dalton, and Duane D. Miller. Unexpected Binding Orientation of Bulky-B-Ring Anti-Androgens and Implications for Future Drug Targets. *Journal of Medicinal Chemistry*, 54(11):3973–3976, June 2011. ISSN 0022-2623, 1520-4804. doi: 10.1021/jm2000097. URL http://pubs.acs.org/doi/abs/10.1021/jm2000097.

[75] Michael F Dunn. Protein-Ligand Interactions: General Description. In John Wiley & Sons, Ltd, editor, *Encyclopedia of Life Sciences*. John Wiley & Sons, Ltd, Chichester, UK, April 2010. ISBN 978-0-470-01617-6 978-0-470-01590-2. URL http://doi.wiley.com/10.1002/9780470015902.a0001340.pub2. DOI: 10.1002/9780470015902.a0001340.pub2.

[76] Joseph L. Durant, Burton A. Leland, Douglas R. Henry, and James G. Nourse. Reoptimization of MDL Keys for Use in Drug Discovery. *Journal of Chemical Infor-*

mation and Computer Sciences, 42(6):1273–1280, November 2002. ISSN 0095-2338. doi: 10.1021/ci010132r. URL https://pubs.acs.org/doi/10.1021/ci010132r.

[77] Jacob D. Durrant and J. Andrew McCammon. BINANA: a novel algorithm for ligand-binding characterization. *Journal of Molecular Graphics & Modelling*, 29(6): 888–893, April 2011. ISSN 1873-4243. doi: 10.1016/j.jmgm.2011.01.004.

[78] Geeta N. Eick, Jennifer K. Colucci, Michael J. Harms, Eric A. Ortlund, and Joseph W. Thornton. Evolution of Minimal Specificity and Promiscuity in Steroid Hormone Receptors. *PLOS Genetics*, 8(11):e1003072, November 2012. ISSN 1553-7404. doi: 10.1371/journal.pgen.1003072. URL https://journals.plos.org/plosgenetics/article?id=10.1371/journal.pgen.1003072.

[79] Daniel C. Elton, Zois Boukouvalas, Mark D. Fuge, and Peter W. Chung. Deep learning for molecular design—a review of the state of the art. *Molecular Systems Design & Engineering*, May 2019. ISSN 2058-9689. doi: 10.1039/C9ME00039A. URL https://pubs.rsc.org/en/content/articlelanding/2019/me/c9me00039a.

[80] Hector Escriva, Franck Delaunay, and Vincent Laudet. Ligand binding and nuclear receptor evolution. *BioEssays*, 22(8):717–727, August 2000. ISSN 1521-1878. doi: 10.1002/1521-1878(200008)22:8<717::AID-BIES5>3.0.CO;2-I. URL https://onlinelibrary.wiley.com/doi/abs/10.1002/1521-1878%28200008%2922%3A8%3C717%3A%3AAID-BIES5%3E3.0.CO%3B2-I.

[81] E. Estebanez-Perpina, L. A. Arnold, P. Nguyen, E. D. Rodrigues, E. Mar, R. Bateman, P. Pallai, K. M. Shokat, J. D. Baxter, R. K. Guy, P. Webb, and R. J. Fletterick. A surface on the androgen receptor that allosterically regulates coactivator binding. *Proceedings of the National Academy of Sciences*, 104(41):16074–16079, October 2007. ISSN 0027-8424, 1091-6490. doi: 10.1073/pnas.0708036104. URL http://www.pnas.org/cgi/doi/10.1073/pnas.0708036104.

[82] Ronald M. Evans and David J. Mangelsdorf. Nuclear Receptors, RXR, and the Big Bang. *Cell*, 157(1):255–266, March 2014. ISSN 00928674. doi: 10.1016/j.cell.2014.03.012. URL https://linkinghub.elsevier.com/retrieve/pii/S0092867414003468.

[83] A. V. Fassio, L. H. Santos, S. A. Silveira, R. S. Ferreira, and R. C. de Melo-Minardi. nAPOLI: a graph-based strategy to detect and visualize conserved protein-ligand interactions in large-scale. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, pages 1–1, 2019. ISSN 1545-5963. doi: 10.1109/TCBB.2019.2892099.

[84] Alexandre Victor Fassio. napoli: uma ferramenta web para análise de interações proteína-ligante. Master's thesis, Universidade Federal de Minas Gerais, Universidade Federal de Minas Gerais, Belo Horizonte, 2015.

[85] Renato Ferreira de Freitas and Matthieu Schapira. A systematic analysis of atomic protein–ligand interactions in the PDB. *MedChemComm*, 8(10):1970–1981, 2017. ISSN 2040-2503, 2040-2511. doi: 10.1039/C7MD00381A. URL http://xlink.rsc.org/?DOI=C7MD00381A.

[86] Emil Fischer. Einfluss der Configuration auf die Wirkung der Enzyme. *Berichte der deutschen chemischen Gesellschaft*, 27(3):2985–2993, October 1894. ISSN 1099-0682. doi: 10.1002/cber.18940270364. URL http://onlinelibrary.wiley.com/doi/10.1002/cber.18940270364/abstract.

[87] A. Frankel, P. Welsh, J. Richardson, and J. D. Robertus. Role of arginine 180 and glutamic acid 177 of ricin toxin A chain in enzymatic inactivation of ribosomes. *Molecular and Cellular Biology*, 10(12):6257–6263, December 1990. ISSN 0270-7306.

[88] Anna Maria Gallina, Paola Bisignano, Maurizio Bergamino, and Domenico Bordo. PLI: a web-based tool for the comparison of protein-ligand interactions observed on PDB structures. *Bioinformatics (Oxford, England)*, 29(3):395–397, February 2013. ISSN 1367-4811. doi: 10.1093/bioinformatics/bts691.

[89] Justin P. Gallivan and Dennis A. Dougherty. Cation-π interactions in structural biology. *Proceedings of the National Academy of Sciences*, 96(17):9459–9464, August 1999. ISSN 0027-8424, 1091-6490. doi: 10.1073/pnas.96.17.9459. URL http://www.pnas.org/content/96/17/9459.

[90] Mu Gao and Jeffrey Skolnick. A Comprehensive Survey of Small-Molecule Binding Pockets in Proteins. *PLOS Computational Biology*, 9(10):e1003302, October 2013. ISSN 1553-7358. doi: 10.1371/journal.pcbi.1003302.

[91] Peter Gedeck, Bernhard Rohde, and Christian Bartels. QSAR - How Good Is It in Practice? Comparison of Descriptor Sets on an Unbiased Cross Section of Corporate Data Sets. *Journal of Chemical Information and Modeling*, 46(5):1924–1936, September 2006. ISSN 1549-9596. doi: 10.1021/ci050413p. URL https://doi.org/10.1021/ci050413p.

[92] Samuel H. Gellman. Introduction: Molecular Recognition. *Chemical Reviews*, 97 (5):1231–1232, August 1997. ISSN 0009-2665. doi: 10.1021/cr970328j. URL http://dx.doi.org/10.1021/cr970328j.

[93] P. Germain, B. Staels, C. Dacquet, M. Spedding, and V. Laudet. Overview of Nomenclature of Nuclear Receptors. *Pharmacological Reviews*, 58(4):685–704, December 2006. ISSN 0031-6997. doi: 10.1124/pr.58.4.2. URL http://pharmrev.aspetjournals.org/cgi/doi/10.1124/pr.58.4.2.

[94] V. J. Gillet, P. Willett, and J. Bradshaw. Identification of biological activity profiles using substructural analysis and genetic algorithms. *Journal of Chemical Information and Computer Sciences*, 38(2):165–179, April 1998. ISSN 0095-2338.

[95] Andrew C. Good and Tudor I. Oprea. Optimization of CAMD techniques 3. Virtual screening enrichment studies: a help or hindrance in tool selection? *Journal of Computer-Aided Molecular Design*, 22(3-4):169–178, March 2008. ISSN 0920-654X, 1573-4951. doi: 10.1007/s10822-007-9167-2. URL http://link.springer.com/10.1007/s10822-007-9167-2.

[96] Alexander Griekspoor, Wilbert Zwart, Jacques Neefjes, and Rob Michalides. Visualizing the action of steroid hormone receptors in living cells. *Nuclear Receptor Signaling*, 5(1):nrs.05003, January 2007. ISSN 1550-7629, 1550-7629. doi: 10.1621/nrs.05003. URL http://journals.sagepub.com/doi/10.1621/nrs.05003.

[97] Chemical Computing Group. Molecular Operating Environment, 2019. https://www.chemcomp.com/index.htm.

[98] Jörg Grunenberg. Complexity in molecular recognition. *Physical Chemistry Chemical Physics*, 13(21):10136, 2011. ISSN 1463-9076, 1463-9084. doi: 10.1039/c1cp20097f. URL http://xlink.rsc.org/?DOI=c1cp20097f.

[99] Adelaide U. P. Hain, Alexia S. Miller, Jelena Levitskaya, and Jürgen Bosch. Virtual Screening and Experimental Validation Identify Novel Inhibitors of the Plasmodium falciparum Atg8–Atg3 Protein–Protein Interaction. *ChemMedChem*, 11 (8):900–910, April 2016. ISSN 1860-7187. doi: 10.1002/cmdc.201500515. URL http://onlinelibrary.wiley.com/doi/10.1002/cmdc.201500515/abstract.

[100] Lowell H. Hall and Lemont B. Kier. Electrotopological State Indices for Atom Types: A Novel Combination of Electronic, Topological, and Valence State Information. *Journal of Chemical Information and Modeling*, 35(6):1039–1045, November 1995. ISSN 1549-9596. doi: 10.1021/ci00028a014. URL http://pubs.acs.org/doi/abs/10.1021/ci00028a014.

[101] Michael Harder, Bernd Kuhn, and François Diederich. Efficient Stacking on Protein Amide Fragments. *ChemMedChem*, 8(3):397–404, 2013. ISSN 1860-7187. doi: 10.1002/cmdc.201200512. URL https://onlinelibrary.wiley.com/doi/abs/10.1002/cmdc.201200512.

[102] Michael J. Harms, Geeta N. Eick, Devrishi Goswami, Jennifer K. Colucci, Patrick R. Griffin, Eric A. Ortlund, and Joseph W. Thornton. Biophysical mechanisms for large-effect mutations in the evolution of steroid hormone receptors. *Proceedings of the National Academy of Sciences*, 110(28):11475–11480, July 2013. ISSN 0027-8424, 1091-6490. doi: 10.1073/pnas.1303930110. URL http://www.pnas.org/content/110/28/11475.

[103] Manfred Hendlich, Andreas Bergner, Judith Günther, and Gerhard Klebe. Relibase: design and development of a database for comprehensive analysis of protein-ligand interactions. *Journal of Molecular Biology*, 326(2):607–620, February 2003. ISSN 0022-2836.

[104] Gerhard Hessler and Karl-Heinz Baringhaus. Artificial Intelligence in Drug Design. *Molecules*, 23(10):2520, October 2018. ISSN 1420-3049. doi: 10.3390/molecules23102520. URL http://www.mdpi.com/1420-3049/23/10/2520.

[105] Rickey P. Hicks, Mark G. Hartell, Daniel A. Nichols, Apurba K. Bhattacharjee, John E. van Hamont, and Donald R. Skillman. The medicinal chemistry of botulinum, ricin and anthrax toxins. *Current Medicinal Chemistry*, 12(6):667–690, 2005. ISSN 0929-8673.

[106] Meng-Chiao Ho, Matthew B. Sturm, Steven C. Almo, and Vern L. Schramm. Transition state analogues in structures of ricin and saporin ribosome-inactivating proteins. *Proceedings of the National Academy of Sciences*, 106(48):20276–20281, December 2009. ISSN 0027-8424, 1091-6490. doi: 10.1073/pnas.0911606106. URL http://www.pnas.org/content/106/48/20276.

[107] Kai-Cheng Hsu, Yen-Fu Chen, Shen-Rong Lin, and Jinn-Moon Yang. iGEMDOCK: a graphical environment of enhancing GEMDOCK using pharmacological interactions and post-screening analysis. *BMC bioinformatics*, 12(1): S33, 2011. URL https://bmcbioinformatics.biomedcentral.com/articles/10.1186/1471-2105-12-S1-S33.

[108] Ye Hu, Dagmar Stumpfe, and Jürgen Bajorath. Computational Exploration of Molecular Scaffolds in Medicinal Chemistry. *Journal of Medicinal Chemistry*, 59 (9):4062–4076, May 2016. ISSN 0022-2623. doi: 10.1021/acs.jmedchem.5b01746. URL https://doi.org/10.1021/acs.jmedchem.5b01746.

[109] Niu Huang, Brian K. Shoichet, and John J. Irwin. Benchmarking sets for molecular docking. *Journal of Medicinal Chemistry*, 49(23):6789–6801, November 2006. ISSN 0022-2623. doi: 10.1021/jm0608356.

[110] Pengxiang Huang, Vikas Chandra, and Fraydoon Rastinejad. Structural Overview of the Nuclear Receptor Superfamily: Insights into Physiology and Therapeutics. *Annual review of physiology*, 72:247–272, 2010. ISSN 0066-4278. doi: 10.1146/annurev-physiol-021909-135917. URL https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3677810/.

[111] Q. Huang, S. Liu, Y. Tang, S. Jin, and Y. Wang. Studies on crystal structures, active-centre geometry and depurinating mechanism of two ribosome-inactivating proteins. *The Biochemical Journal*, 309 (Pt 1):285–298, July 1995. ISSN 0264-6021.

[112] Roderick E Hubbard and Muhammad Kamran Haider. Hydrogen Bonds in Proteins: Role and Strength. In John Wiley & Sons, Ltd, editor, *Encyclopedia of Life Sciences*. John Wiley & Sons, Ltd, Chichester, UK, February 2010. ISBN 978-0-470-01617-6 978-0-470-01590-2. URL http://doi.wiley.com/10.1002/9780470015902.a0003011.pub2. DOI: 10.1002/9780470015902.a0003011.pub2.

[113] JP Hughes, S Rees, SB Kalindjian, and KL Philpott. Principles of early drug discovery. *British Journal of Pharmacology*, 162(6):1239–1249, March 2011. ISSN 0007-1188. doi: 10.1111/j.1476-5381.2010.01127.x. URL http://www.ncbi.nlm.nih.gov/pmc/articles/PMC3058157/.

[114] Christopher A. Hunter, Kevin R. Lawson, Julie Perkins, and Christopher J. Urch. Aromatic interactions. *Journal of the Chemical Society, Perkin Transactions 2*, (5):651–669, January 2001. ISSN 1364-5471. doi: 10.1039/B008495F. URL http://pubs.rsc.org/en/content/articlelanding/2001/p2/b008495f.

[115] Yumi N. Imai, Yoshihisa Inoue, Isao Nakanishi, and Kazuo Kitaura. Amide–π interactions between formamide and benzene. *Journal of Computational Chemistry*, 30(14):2267–2276, 2009. ISSN 1096-987X. doi: 10.1002/jcc.21212. URL https://www.onlinelibrary.wiley.com/doi/abs/10.1002/jcc.21212.

[116] John J. Irwin, Teague Sterling, Michael M. Mysinger, Erin S. Bolstad, and Ryan G. Coleman. ZINC: A Free Tool to Discover Chemistry for Biology. *Journal of Chemical Information and Modeling*, 52(7):1757–1768, July 2012. ISSN 1549-9596. doi: 10.1021/ci3001277. URL http://dx.doi.org/10.1021/ci3001277.

[117] Michio Iwaoka and Natsuki Babe. Mining and Structural Characterization of S···X Chalcogen Bonds in Protein Database. *Phosphorus, Sulfur, and Silicon and the Related Elements*, 190(8):1257–1264, August 2015. ISSN 1042-6507, 1563-5325. doi: 10.1080/10426507.2014.1002612. URL http://www.tandfonline.com/doi/full/10.1080/10426507.2014.1002612.

[118] Ilian Jelesarov and Andrey Karshikoff. Defining the role of salt bridges in protein stability. *Methods in Molecular Biology (Clifton, N.J.)*, 490:227–260, 2009. ISSN 1064-3745. doi: 10.1007/978-1-59745-367-7_10.

[119] Shuiqin Jiang, Lujia Zhang, Dongbin Cui, Zhiqiang Yao, Bei Gao, Jinping Lin, and Dongzhi Wei. The Important Role of Halogen Bond in Substrate Selectivity of Enzymatic Catalysis. *Scientific Reports*, 6:34750, October 2016. ISSN 2045-2322. doi: 10.1038/srep34750. URL https://www.nature.com/articles/srep34750.

[120] J. A. Joule and K. Mills. *Heterocyclic chemistry*. Wiley, Hoboken, N.J, 5th ed edition, 2009. ISBN 978-1-4051-9365-8 978-1-4051-3300-5.

[121] Harry C. Jubb, Alicia P. Higueruelo, Bernardo Ochoa-Montaño, Will R. Pitt, David B. Ascher, and Tom L. Blundell. Arpeggio: A Web Server for Calculating and Visualising Interatomic Interactions in Protein Structures. *Journal of Molecular Biology*, 429(3):365–371, February 2017. ISSN 1089-8638. doi: 10.1016/j.jmb.2016.12.004.

[122] Mahita Kadmiel and John A. Cidlowski. Glucocorticoid receptor signaling in health and disease. *Trends in Pharmacological Sciences*, 34(9):518–530, September 2013. ISSN 01656147. doi: 10.1016/j.tips.2013.07.003. URL https://linkinghub.elsevier.com/retrieve/pii/S016561471300120X.

[123] Kota Kasahara and Kengo Kinoshita. GIANT: pattern analysis of molecular interactions in 3d structures of protein-small ligand complexes. *BMC bioinformatics*, 15: 12, January 2014. ISSN 1471-2105. doi: 10.1186/1471-2105-15-12.

[124] B. J. Katzin, E. J. Collins, and J. D. Robertus. Structure of ricin A-chain at 2.5 A. *Proteins*, 10(3):251–259, 1991. ISSN 0887-3585. doi: 10.1002/prot.340100309.

[125] Amit Kessel and Nir Ben-Tal. *Introduction to proteins: structure, function, and motion*. Chapman & Hall/CRC mathematical and computational biology series. CRC Press, Taylor & Francis Group, Boca Raton London New York, second edition edition, 2018. ISBN 978-1-4987-4717-2 978-1-315-11387-6. OCLC: 1044746110.

[126] R. Khashan, W. Zheng, and A. Tropsha. Scoring protein interaction decoys using exposed residues (SPIDER): A novel multibody interaction scoring function based on frequent geometric patterns of interfacial residues. *Proteins: Structure, Function and Bioinformatics*, 80(9):2207–2217, 2012. doi: 10.1002/prot.24110.

[127] Y. Kim and J. D. Robertus. Analysis of several key active site residues of ricin A chain by mutagenesis and X-ray crystallography. *Protein Engineering*, 5(8):775–779, December 1992. ISSN 0269-2139.

[128] Y. Kim, D. Mlsna, A. F. Monzingo, M. P. Ready, A. Frankel, and J. D. Robertus. Structure of a ricin mutant showing rescue of activity by a noncatalytic residue. *Biochemistry*, 31(12):3294–3296, March 1992. ISSN 0006-2960.

[129] Diederik P. Kingma and Jimmy Ba. Adam: A Method for Stochastic Optimization. *arXiv:1412.6980 [cs]*, December 2014. URL http://arxiv.org/abs/1412.6980. arXiv: 1412.6980.

[130] Johannes Kirchmair, Patrick Markt, Simona Distinto, Gerhard Wolber, and Thierry Langer. Evaluation of the performance of 3d virtual screening protocols: RMSD comparisons, enrichment assessments, and decoy selection—What can we learn from earlier mistakes? *Journal of Computer-Aided Molecular Design*, 22(3-4):213–228, March 2008. ISSN 0920-654X, 1573-4951. doi: 10.1007/s10822-007-9163-6. URL https://link.springer.com/article/10.1007/s10822-007-9163-6.

[131] Douglas B. Kitchen, Hélène Decornez, John R. Furr, and Jürgen Bajorath. Docking and scoring in virtual screening for drug discovery: methods and applications. *Nature Reviews Drug Discovery*, 3(11):935–949, November 2004. ISSN 1474-1776. doi: 10.1038/nrd1549. URL http://www.nature.com/nrd/journal/v3/n11/abs/nrd1549.html.

[132] Gerhard Klebe. Virtual ligand screening: strategies, perspectives and limitations. *Drug Discovery Today*, 11(13-14):580–594, July 2006. ISSN 1359-6446. doi: 10.1016/j.drudis.2006.05.012.

[133] Sebastian Kmiecik, Dominik Gront, Michal Kolinski, Lukasz Wieteska, Aleksandra Elzbieta Dawid, and Andrzej Kolinski. Coarse-Grained Protein Models and Their Applications. *Chemical Reviews*, 116(14):7898–7936, July 2016. ISSN 0009-2665. doi: 10.1021/acs.chemrev.6b00163. URL https://doi.org/10.1021/acs.chemrev.6b00163.

[134] Mathew R. Koebel, Aaron Cooper, Grant Schmadeke, Soyoung Jeon, Mahesh Narayan, and Suman Sirimulla. S···O and S···N Sulfur Bonding Interactions in Protein–Ligand Complexes: Empirical Considerations and Scoring Function. *Journal of Chemical Information and Modeling*, 56(12):2298–2309, December 2016. ISSN 1549-9596, 1549-960X. doi: 10.1021/acs.jcim.6b00236. URL http://pubs.acs.org/doi/10.1021/acs.jcim.6b00236.

[135] Hugo Kubinyi. Hydrogen Bonding: The Last Mystery in Drug Design? In Bernard Testa, Han van de Waterbeemd, Gerd Folkers, and Richard Guy, editors, *Pharmacokinetic Optimization in Drug Research*, pages 513–524. Verlag Helvetica Chimica Acta, Zürich, February 2001. ISBN 978-3-906390-43-7 978-3-906390-

22-2. doi: 10.1002/9783906390437.ch28. URL http://doi.wiley.com/10.1002/9783906390437.ch28.

[136] Bernd Kuhn, Michal Tichý, Lingle Wang, Shaughnessy Robinson, Rainer E. Martin, Andreas Kuglstatter, Jörg Benz, Maude Giroud, Tanja Schirmeister, Robert Abel, François Diederich, and Jérôme Hert. Prospective Evaluation of Free Energy Calculations for the Prioritization of Cathepsin L Inhibitors. *Journal of Medicinal Chemistry*, 60(6):2485–2497, March 2017. ISSN 0022-2623. doi: 10.1021/acs.jmedchem.6b01881. URL http://dx.doi.org/10.1021/acs.jmedchem.6b01881.

[137] Nathan A. Lack, Peter Axerio-Cilies, Peyman Tavassoli, Frank Q. Han, Ka Hong Chan, Clementine Feau, Eric LeBlanc, Emma Tomlinson Guns, R. Kiplin Guy, Paul S. Rennie, and Artem Cherkasov. Targeting the binding function 3 (BF3) site of the human androgen receptor through virtual screening. *Journal of Medicinal Chemistry*, 54(24):8563–8573, December 2011. ISSN 1520-4804. doi: 10.1021/jm201098n.

[138] Roman A. Laskowski and Mark B. Swindells. LigPlot+: multiple ligand-protein interaction diagrams for drug discovery. *Journal of Chemical Information and Modeling*, 51(10):2778–2786, October 2011. ISSN 1549-960X. doi: 10.1021/ci200227u.

[139] A. Lavecchia and C. Di Giovanni. Virtual Screening Strategies in Drug Discovery: A Critical Review. *Current Medicinal Chemistry*, 20(23):2839–2860, August 2013.

[140] Antonio Lavecchia. Machine-learning approaches in drug discovery: methods and applications. *Drug Discovery Today*, 20(3):318–331, March 2015. ISSN 1359-6446. doi: 10.1016/j.drudis.2014.10.012. URL http://www.sciencedirect.com/science/article/pii/S1359644614004176.

[141] Antonio Lavecchia. Deep learning in drug discovery: Opportunities, challenges and future prospects. *Drug Discovery Today*, August 2019. ISSN 1359-6446. doi: 10.1016/j.drudis.2019.07.006. URL http://www.sciencedirect.com/science/article/pii/S135964461930282X.

[142] Andrew R. Leach and Valerie J. Gillet. Molecular Descriptors. In Andrew R. Leach and Valerie J. Gillet, editors, *An Introduction To Chemoinformatics*, pages 53–74. Springer Netherlands, Dordrecht, 2007. ISBN 978-1-4020-6291-9. doi: 10.1007/978-1-4020-6291-9_3. URL https://doi.org/10.1007/978-1-4020-6291-9_3.

[143] Anthony C. Legon. Tetrel, pnictogen and chalcogen bonds identified in the gas phase before they had names: a systematic look at non-covalent interactions. *Physical Chemistry Chemical Physics*, 19(23):14884–14896, June 2017. ISSN 1463-9084. doi: 10.1039/C7CP02518A. URL https://pubs.rsc.org/en/content/articlelanding/2017/cp/c7cp02518a.

[144] Eelke B. Lenselink, Niels ten Dijke, Brandon Bongers, George Papadatos, Herman W. T. van Vlijmen, Wojtek Kowalczyk, Adriaan P. IJzerman, and Gerard J. P. van Westen. Beyond the hype: deep neural networks outperform established methods using a ChEMBL bioactivity benchmark set. *Journal of Cheminformatics*, 9(1): 45, August 2017. ISSN 1758-2946. doi: 10.1186/s13321-017-0232-0. URL https://doi.org/10.1186/s13321-017-0232-0.

[145] Yaakov Levy and José N. Onuchic. Water mediation in protein folding and molecular recognition. *Annual Review of Biophysics and Biomolecular Structure*, 35:389–415, 2006. ISSN 1056-8700. doi: 10.1146/annurev.biophys.35.040405.102134.

[146] Shenhui Li and Mei Hong. Protonation, Tautomerization, and Rotameric Structure of Histidine: A Comprehensive Study by Magic-Angle-Spinning Solid-State NMR. *Journal of the American Chemical Society*, 133(5):1534–1544, February 2011. ISSN 0002-7863, 1520-5126. doi: 10.1021/ja108943n. URL http://pubs.acs.org/doi/abs/10.1021/ja108943n.

[147] C. A. Lipinski, F. Lombardo, B. W. Dominy, and P. J. Feeney. Experimental and computational approaches to estimate solubility and permeability in drug discovery and development settings. *Advanced Drug Delivery Reviews*, 46(1-3):3–26, March 2001. ISSN 0169-409X. doi: 10.1016/s0169-409x(00)00129-0.

[148] Christopher Lipinski and Andrew Hopkins. Navigating chemical space for biology and medicine. *Nature*, 432(7019):855–861, December 2004. ISSN 1476-4687. doi: 10.1038/nature03193.

[149] Liyuan Liu, Haoming Jiang, Pengcheng He, Weizhu Chen, Xiaodong Liu, Jianfeng Gao, and Jiawei Han. On the Variance of the Adaptive Learning Rate and Beyond. *arXiv:1908.03265 [cs, stat]*, August 2019. URL http://arxiv.org/abs/1908.03265. arXiv: 1908.03265.

[150] Scott J. Lusher, Hans C. A. Raaijmakers, Diep Vu-Pham, Koen Dechering, Tsang Wai Lam, Angus R. Brown, Niall M. Hamilton, Olaf Nimz, Rolien Bosch, Ross McGuire, Arthur Oubrie, and Jacob de Vlieg. Structural Basis for Agonism and Antagonism for a Set of Chemically Related Progesterone Receptor Modulators. *Journal of Biological Chemistry*, 286(40):35079–35086, October 2011. ISSN 0021-9258, 1083-351X. doi: 10.1074/jbc.M111.273029. URL http://www.jbc.org/lookup/doi/10.1074/jbc.M111.273029.

[151] Scott J. Lusher, Hans C. A. Raaijmakers, Diep Vu-Pham, Bert Kazemier, Rolien Bosch, Ross McGuire, Rita Azevedo, Hans Hamersma, Koen Dechering, Arthur Oubrie, Marcel van Duin, and Jacob de Vlieg. X-ray Structures of Progesterone

Receptor Ligand Binding Domain in Its Agonist State Reveal Differing Mechanisms for Mixed Profiles of 11β-Substituted Steroids. *Journal of Biological Chemistry*, 287 (24):20333–20343, June 2012. ISSN 0021-9258, 1083-351X. doi: 10.1074/jbc.M111.308403. URL http://www.jbc.org/lookup/doi/10.1074/jbc.M111.308403.

[152] Paul D Lyne. Structure-based virtual screening: an overview. *Drug Discovery Today*, 7(20):1047–1055, October 2002. ISSN 1359-6446. doi: 10.1016/S1359-6446(02)02483-2. URL http://www.sciencedirect.com/science/article/pii/S1359644602024832.

[153] Jiankun Lyu, Sheng Wang, Trent E. Balius, Isha Singh, Anat Levit, Yurii S. Moroz, Matthew J. O'Meara, Tao Che, Enkhjargal Algaa, Kateryna Tolmachova, Andrey A. Tolmachev, Brian K. Shoichet, Bryan L. Roth, and John J. Irwin. Ultra-large library docking for discovering new chemotypes. *Nature*, 566(7743):224–229, February 2019. ISSN 0028-0836, 1476-4687. doi: 10.1038/s41586-019-0917-9. URL http://www.nature.com/articles/s41586-019-0917-9.

[154] Kamran T. Mahmudov, Maximilian N. Kopylovich, M. Fátima C. Guedes da Silva, and Armando J. L. Pombeiro. Chalcogen bonding in synthesis, catalysis and design of materials. *Dalton Transactions*, 46(31):10121–10138, 2017. ISSN 1477-9226, 1477-9234. doi: 10.1039/C7DT01685A. URL http://xlink.rsc.org/?DOI=C7DT01685A.

[155] Adauto L Mancini, Roberto H Higa, A Oliveira, Fabiana Dominiquini, Paula R Kuser, Michel EB Yamagishi, Roberto C Togawa, and Goran Neshich. STING Contacts: a web-based application for identification and analysis of amino acid contacts within protein structure and across protein interfaces. *Bioinformatics*, 20 (13):2145–2147, 2004.

[156] Alexey B Mantsyzov, Guillaume Bouvier, Nathalie Evrard-Todeschi, and Gildas Bertho. Contact-based ligand-clustering approach for the identification of active compounds in virtual screening. *Advances and Applications in Bioinformatics and Chemistry : AABC*, 5:61–79, September 2012. ISSN 1178-6949. doi: 10.2147/AABC.S30881. URL http://www.ncbi.nlm.nih.gov/pmc/articles/PMC3459543/.

[157] Gilles Marcou and Didier Rognan. Optimizing Fragment and Scaffold Docking by Use of Molecular Interaction Fingerprints. *Journal of Chemical Information and Modeling*, 47(1):195–207, January 2007. ISSN 1549-9596. doi: 10.1021/ci600342e. URL https://doi.org/10.1021/ci600342e.

[158] Catherine J. Marsden, Vilmos Fülöp, Philip J. Day, and J. Michael Lord. The effect of mutations surrounding and within the active site on the catalytic activity of ricin

A chain. *European Journal of Biochemistry*, 271(1):153–162, January 2004. ISSN 0014-2956.

[159] Leandro Martínez, Roberto Andreani, and José Mario Martínez. Convergent algorithms for protein structural alignment. *BMC Bioinformatics*, 8:306, 2007. ISSN 1471-2105. doi: 10.1186/1471-2105-8-306. URL http://dx.doi.org/10.1186/1471-2105-8-306.

[160] Andreas Mayr, Günter Klambauer, Thomas Unterthiner, Marvin Steijaert, Jörg K. Wegner, Hugo Ceulemans, Djork-Arné Clevert, and Sepp Hochreiter. Large-scale comparison of machine learning methods for drug target prediction on ChEMBL. *Chemical Science*, 9(24):5441–5451, June 2018. ISSN 2041-6539. doi: 10.1039/C8SC00148K. URL https://pubs.rsc.org/en/content/articlelanding/2018/sc/c8sc00148k.

[161] I. K. McDonald and J. M. Thornton. Satisfying hydrogen bonding potential in proteins. *Journal of Molecular Biology*, 238(5):777–793, May 1994. ISSN 0022-2836. doi: 10.1006/jmbi.1994.1334.

[162] Benjamin Merget, Samo Turk, Sameh Eid, Friedrich Rippmann, and Simone Fulle. Profiling Prediction of Kinase Inhibitors: Toward the Virtual Assay. *Journal of Medicinal Chemistry*, 60(1):474–485, January 2017. ISSN 0022-2623. doi: 10.1021/acs.jmedchem.6b01611. URL https://doi.org/10.1021/acs.jmedchem.6b01611.

[163] Pierangelo Metrangolo and Giuseppe Resnati. Halogen Bonding: A Paradigm in Supramolecular Chemistry. *Chemistry – A European Journal*, 7(12):2511–2519, 2001. ISSN 1521-3765. doi: 10.1002/1521-3765(20010618)7:12<2511::AID-CHEM25110>3.0.CO;2-T. URL https://onlinelibrary.wiley.com/doi/abs/10.1002/1521-3765%2820010618%297%3A12%3C2511%3A%3AAID-CHEM25110%3E3.0.CO%3B2-T.

[164] Pierangelo Metrangolo, Tullio Pilati, Giuseppe Resnati, and Andrea Stevenazzi. Metric engineering of perfluorocarbon–hydrocarbon layered solids driven by the halogen bonding. *Chemical Communications*, (13):1492–1493, June 2004. ISSN 1364-548X. doi: 10.1039/B402305F. URL https://pubs.rsc.org/en/content/articlelanding/2004/cc/b402305f.

[165] Pierangelo Metrangolo, Hannes Neukirch, Tullio Pilati, and Giuseppe Resnati. Halogen Bonding Based Recognition Processes: A World Parallel to Hydrogen Bonding. *Accounts of Chemical Research*, 38(5):386–395, May 2005. ISSN 0001-4842. doi: 10.1021/ar0400995. URL https://doi.org/10.1021/ar0400995.

[166] Pierangelo Metrangolo, Franck Meyer, Tullio Pilati, Giuseppe Resnati, and Giancarlo Terraneo. Halogen Bonding in Supramolecular Chemistry. *Angewandte Chemie International Edition*, 47(33):6114–6127, August 2008. ISSN 1433-7851. doi: 10.1002/anie.200800128. URL https://onlinelibrary.wiley.com/doi/full/10.1002/anie.200800128.

[167] Darcie J. Miller, Kabyadi Ravikumar, Huafeng Shen, Jung-Keun Suh, Sean M. Kerwin, and Jon D. Robertus. Structure-based design and characterization of novel platforms for ricin and shiga toxin inhibition. *Journal of Medicinal Chemistry*, 45 (1):90–98, January 2002. ISSN 0022-2623.

[168] D. Mlsna, A. F. Monzingo, B. J. Katzin, S. Ernst, and J. D. Robertus. Structure of recombinant ricin A chain at 2.3 A. *Protein Science: A Publication of the Protein Society*, 2(3):429–435, March 1993. ISSN 0961-8368. doi: 10.1002/pro.5560020315.

[169] W. Montfort, J. E. Villafranca, A. F. Monzingo, S. R. Ernst, B. Katzin, E. Rutenber, N. H. Xuong, R. Hamlin, and J. D. Robertus. The three-dimensional structure of ricin at 2.8 A. *The Journal of Biological Chemistry*, 262(11):5398–5403, April 1987. ISSN 0021-9258.

[170] A. F. Monzingo and J. D. Robertus. X-ray analysis of substrate analogs in the ricin A-chain active site. *Journal of Molecular Biology*, 227(4):1136–1145, October 1992. ISSN 0022-2836.

[171] Jane S. Murray, Pat Lane, and Peter Politzer. Simultaneous σ-hole and hydrogen bonding by sulfur- and selenium-containing heterocycles. *International Journal of Quantum Chemistry*, 108(15):2770–2781, 2008. ISSN 1097-461X. doi: 10.1002/qua.21753. URL https://onlinelibrary.wiley.com/doi/abs/10.1002/qua.21753.

[172] David Lee Nelson and Michael M. Cox. *Lehninger principles of biochemistry*. W.H. Freeman and Company, New York, 6th ed edition, 2013. ISBN 978-1-4641-0962-1 978-1-4292-3414-6. OCLC: ocn824794893.

[173] Ramaswamy Nilakantan, Norman Bauman, J. Scott Dixon, and R. Venkataraghavan. Topological torsion: a new molecular descriptor for SAR applications. Comparison with other descriptors. *Journal of Chemical Information and Computer Sciences*, 27(2):82–85, May 1987. ISSN 0095-2338. doi: 10.1021/ci00054a008. URL https://pubs.acs.org/doi/abs/10.1021/ci00054a008.

[174] I. Nobeli, S. L. Price, J. P. M. Lommerse, and R. Taylor. Hydrogen bonding properties of oxygen and nitrogen acceptors in aromatic heterocycles. *Journal of Computational Chemistry*, 18(16):2060–2074, 1997. ISSN 1096-987X. doi: 10.1002/(SICI)1096-987X(199712)18:16<2060::AID-JCC10>3.0.

CO;2-S. URL https://onlinelibrary.wiley.com/doi/abs/10.1002/%28SICI%291096-987X%28199712%2918%3A16%3C2060%3A%3AAID-JCC10%3E3.0.CO%3B2-S.

[175] Jerome C. Nwachukwu, Sathish Srinivasan, Nelson E. Bruno, Jason Nowak, Nicholas J. Wright, Filippo Minutolo, Erumbi S. Rangarajan, Tina Izard, Xin-Qui Yao, Barry J. Grant, Douglas J. Kojetin, Olivier Elemento, John A. Katzenellenbogen, and Kendall W. Nettles. Systems Structural Biology Analysis of Ligand Effects on ERα Predicts Cellular Response to Environmental Estrogens and Anti-hormone Therapies. *Cell Chemical Biology*, 24(1):35–45, January 2017. ISSN 24519456. doi: 10.1016/j.chembiol.2016.11.014. URL https://linkinghub.elsevier.com/retrieve/pii/S2451945616304378.

[176] James N Oak, John Oldenhof, and Hubert H.M Van Tol. The dopamine D4 receptor: one decade of research. *European Journal of Pharmacology*, 405(1-3):303–327, September 2000. ISSN 00142999. doi: 10.1016/S0014-2999(00)00562-8. URL https://linkinghub.elsevier.com/retrieve/pii/S0014299900005628.

[177] Noel M. O'Boyle, Michael Banck, Craig A. James, Chris Morley, Tim Vandermeersch, and Geoffrey R. Hutchison. Open Babel: An open chemical toolbox. *Journal of Cheminformatics*, 3(1):33, October 2011. ISSN 1758-2946. doi: 10.1186/1758-2946-3-33. URL https://doi.org/10.1186/1758-2946-3-33.

[178] Jacek Ostrowski, Joyce E. Kuhns, John A. Lupisella, Mark C. Manfredi, Blake C. Beehler, Stanley R. Krystek, Yingzhi Bi, Chongqing Sun, Ramakrishna Seethala, Rajasree Golla, Paul G. Sleph, Aberra Fura, Yongmi An, Kevin F. Kish, John S. Sack, Kasim A. Mookhtiar, Gary J. Grover, and Lawrence G. Hamann. Pharmacological and X-Ray Structural Characterization of a Novel Selective Androgen Receptor Modulator: Potent Hyperanabolic Stimulation of Skeletal Muscle with Hypostimulation of Prostate in Rats. *Endocrinology*, 148(1):4–12, January 2007. ISSN 0013-7227, 1945-7170. doi: 10.1210/en.2006-0843. URL https://academic.oup.com/endo/article-lookup/doi/10.1210/en.2006-0843.

[179] Sunil K. Panigrahi and Gautam R. Desiraju. Strong and weak hydrogen bonds in the protein-ligand interface. *Proteins: Structure, Function, and Bioinformatics*, 67 (1):128–141, January 2007. ISSN 08873585. doi: 10.1002/prot.21253. URL http://doi.wiley.com/10.1002/prot.21253.

[180] Laurent Pascual-Le Tallec and Marc Lombès. The Mineralocorticoid Receptor: A Journey Exploring Its Diversity and Specificity of Action. *Molecular Endocrinology*, 19(9):2211–2221, September 2005. ISSN 0888-8809. doi: 10.1210/me.2005-0089. URL https://academic.oup.com/mend/article/19/9/2211/2737822.

[181] Ralph Paulini, Klaus Müller, and François Diederich. Orthogonal multipolar interactions in structural chemistry and biology. *Angewandte Chemie (International Ed. in English)*, 44(12):1788–1805, March 2005. ISSN 1433-7851. doi: 10.1002/anie.200462213.

[182] I. Petit-Topin, M. Fay, M. Resche-Rigon, A. Ulmann, E. Gainer, M.-E. Rafestin-Oblin, and J. Fagart. Molecular determinants of the recognition of ulipristal acetate by oxo-steroid receptors. *The Journal of Steroid Biochemistry and Molecular Biology*, 144:427–435, October 2014. ISSN 09600760. doi: 10.1016/j.jsbmb.2014.08.008. URL https://linkinghub.elsevier.com/retrieve/pii/S0960076014001873.

[183] Gregory A. Petsko and Dagmar Ringe. *Protein structure and function*. Primers in biology. New Science Press ; Sinauer Associates ; Blackwell Pub, London : Sunderland, MA : Oxford, 2004. ISBN 978-0-87893-663-2 978-0-9539181-4-0 978-1-4051-1922-1. OCLC: ocm53181467.

[184] Eric F. Pettersen, Thomas D. Goddard, Conrad C. Huang, Gregory S. Couch, Daniel M. Greenblatt, Elaine C. Meng, and Thomas E. Ferrin. UCSF Chimera—A visualization system for exploratory research and analysis. *Journal of Computational Chemistry*, 25(13):1605–1612, October 2004. ISSN 1096-987X. doi: 10.1002/jcc.20084. URL http://onlinelibrary.wiley.com/doi/10.1002/jcc.20084/abstract.

[185] Douglas E. V. Pires, Raquel C. de Melo-Minardi, Carlos H. da Silveira, Frederico F. Campos, and Wagner Meira. aCSM: noise-free graph-based signatures to large-scale receptor-based ligand prediction. *Bioinformatics (Oxford, England)*, 29(7):855–861, April 2013. ISSN 1367-4811. doi: 10.1093/bioinformatics/btt058.

[186] Peter Politzer, Pat Lane, Monica C. Concha, Yuguang Ma, and Jane S. Murray. An overview of halogen bonding. *Journal of Molecular Modeling*, 13(2):305–311, February 2007. ISSN 0948-5023. doi: 10.1007/s00894-006-0154-7.

[187] Anne Poupon. Voronoi and Voronoi-related tessellations in studies of protein structure and interaction. *Current Opinion in Structural Biology*, 14(2):233–241, April 2004. ISSN 0959440X. doi: 10.1016/j.sbi.2004.03.010. URL http://linkinghub.elsevier.com/retrieve/pii/S0959440X04000442.

[188] P. L. Privalov and S. J. Gill. Stability of protein structure and hydrophobic interaction. *Advances in Protein Chemistry*, 39:191–234, 1988. ISSN 0065-3233.

[189] Jeff M. Pruet, Karl R. Jasheway, Lawrence A. Manzano, Yan Bai, Eric V. Anslyn, and Jon D. Robertus. 7-Substituted pterins provide a new direction for ricin A chain

inhibitors. *European Journal of Medicinal Chemistry*, 46(9):3608–3615, September 2011. ISSN 1768-3254. doi: 10.1016/j.ejmech.2011.05.025.

[190] Jeff M. Pruet, Ryota Saito, Lawrence A. Manzano, Karl R. Jasheway, Paul A. Wiget, Ishan Kamat, Eric V. Anslyn, and Jon D. Robertus. Optimized 5-membered heterocycle-linked pterins for the inhibition of Ricin Toxin A. *ACS medicinal chemistry letters*, 3(7):588–591, July 2012. ISSN 1948-5875. doi: 10.1021/ml300099t.

[191] Violeta I. Pérez-Nueno, Obdulia Rabal, José I. Borrell, and Jordi Teixidó. APIF: A New Interaction Fingerprint Based on Atom Pairs and Its Application to Virtual Screening. *Journal of Chemical Information and Modeling*, 49(5):1245–1260, May 2009. ISSN 1549-9596. doi: 10.1021/ci900043r. URL https://doi.org/10.1021/ci900043r.

[192] Muhammad Radifar, Nunung Yuniarti, and Enade Perdana Istyastono. PyPLIF: Python-based Protein-Ligand Interaction Fingerprinting. *Bioinformation*, 9(6):325–328, March 2013. ISSN 0973-8894. doi: 10.6026/97320630009325.

[193] Sebastian Raschka, Alex J. Wolf, Joseph Bemister-Buffington, and Leslie A. Kuhn. Protein-ligand interfaces are polarized: discovery of a strong trend for intermolecular hydrogen bonds to favor donors on the protein side with implications for predicting and designing ligand complexes. *Journal of Computer-Aided Molecular Design*, 32 (4):511–528, April 2018. ISSN 1573-4951. doi: 10.1007/s10822-018-0105-2.

[194] John W. Raymond, C. John Blankley, and Peter Willett. Comparison of chemical clustering methods using graph- and fingerprint-based similarity measures. *Journal of Molecular Graphics & Modelling*, 21(5):421–433, March 2003. ISSN 1093-3263.

[195] RDKit. RDKit: Open-source cheminformatics, 2006. [Online]. Available: https://www.rdkit.org/. Accessed: 2019-09-01.

[196] M. P. Ready, Y. Kim, and J. D. Robertus. Site-directed mutagenesis of ricin A-chain and implications for the mechanism of action. *Proteins*, 10(3):270–278, 1991. ISSN 0887-3585. doi: 10.1002/prot.340100311.

[197] N. Rego and D. Koes. 3dmol.js: molecular visualization with WebGL. *Bioinformatics*, 31(8):1322–1324, April 2015. ISSN 1367-4803, 1460-2059. doi: 10.1093/bioinformatics/btu829. URL https://academic.oup.com/bioinformatics/article-lookup/doi/10.1093/bioinformatics/btu829.

[198] J. Ren, Y. Wang, Y. Dong, and D. I. Stuart. The N-glycosidase mechanism of ribosome-inactivating proteins implied by crystal structures of alpha-momorcharin. *Structure (London, England: 1993)*, 2(1):7–16, January 1994. ISSN 0969-2126.

[199] Johanne Renaud, Serge François Bischoff, Thomas Buhl, Philipp Floersheim, Brigitte Fournier, Martin Geiser, Christine Halleux, Joerg Kallen, Hansjoerg Keller, and Paul Ramage. Selective Estrogen Receptor Modulators with Conformationally Restricted Side Chains. Synthesis and Structure-Activity Relationship of ERα-Selective Tetrahydroisoquinoline Ligands. *Journal of Medicinal Chemistry*, 48(2): 364–379, January 2005. ISSN 0022-2623, 1520-4804. doi: 10.1021/jm040858p. URL http://pubs.acs.org/doi/abs/10.1021/jm040858p.

[200] Sereina Riniker and Gregory A Landrum. Open-source platform to benchmark fingerprints for ligand-based virtual screening. *Journal of Cheminformatics*, 5:26, May 2013. ISSN 1758-2946. doi: 10.1186/1758-2946-5-26. URL https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3686626/.

[201] Jon D. Robertus and Arthur F. Monzingo. The Structure and Action of Ribosome-inactivating Proteins. In Fiorenzo Stirpe and Douglas A. Lappi, editors, *Ribosome-inactivating Proteins*, pages 111–133. John Wiley & Sons, Ltd., Oxford, April 2014. ISBN 978-1-118-84723-7 978-1-118-12565-6. doi: 10.1002/9781118847237.ch8.

[202] Barry Robson and Andy Vaithiligam. Drug Gold and Data Dragons: Myths and Realities of Data Mining in the Pharmaceutical Industry. In Konstantin V. Balakin, editor, *Pharmaceutical Data Mining*, pages 25–85. John Wiley & Sons, Inc., 2009. ISBN 978-0-470-56762-3. URL http://onlinelibrary.wiley.com/doi/10.1002/9780470567623.ch2/summary. DOI: 10.1002/9780470567623.ch2.

[203] David Rogers and Mathew Hahn. Extended-Connectivity Fingerprints. *Journal of Chemical Information and Modeling*, 50(5):742–754, May 2010. ISSN 1549-9596, 1549-960X. doi: 10.1021/ci100050t. URL https://pubs.acs.org/doi/10.1021/ci100050t.

[204] E. Rutenber and J. D. Robertus. Structure of ricin B-chain at 2.5 A resolution. *Proteins*, 10(3):260–269, 1991. ISSN 0887-3585. doi: 10.1002/prot.340100310.

[205] E. Rutenber, B. J. Katzin, S. Ernst, E. J. Collins, D. Mlsna, M. P. Ready, and J. D. Robertus. Crystallographic refinement of ricin to 2.5 A. *Proteins*, 10(3):240–250, 1991. ISSN 0887-3585. doi: 10.1002/prot.340100308.

[206] Ryota Saito, Jeff M. Pruet, Lawrence A. Manzano, Karl Jasheway, Arthur F. Monzingo, Paul A. Wiget, Ishan Kamat, Eric V. Anslyn, and Jon D. Robertus. Peptide-conjugated pterins as inhibitors of ricin toxin A. *Journal of Medicinal Chemistry*, 56(1):320–329, January 2013. ISSN 1520-4804. doi: 10.1021/jm3016393.

[207] Sebastian Salentin, Sven Schreiber, V. Joachim Haupt, Melissa F. Adasme, and Michael Schroeder. PLIP: fully automated protein–ligand interaction profiler. *Nu-*

cleic Acids Research, 43(Web Server issue):W443–W447, July 2015. ISSN 0305-1048. doi: 10.1093/nar/gkv315. URL http://www.ncbi.nlm.nih.gov/pmc/articles/PMC4489249/.

[208] C. A. Santana, F. R. Cerqueira, C. H. d Silveira, A. V. Fassio, R. C. d Melo-Minardi, and S. d A. Silveira. GReMLIN: A Graph Mining Strategy to Infer Protein-Ligand Interaction Patterns. In *2016 IEEE 16th International Conference on Bioinformatics and Bioengineering (BIBE)*, pages 28–35, October 2016. doi: 10.1109/BIBE.2016.48.

[209] Madhavi Sastry, Jeffrey F. Lowrie, Steven L. Dixon, and Woody Sherman. Large-Scale Systematic Analysis of 2d Fingerprint Methods and Parameters to Improve Virtual Screening Enrichments. *Journal of Chemical Information and Modeling*, 50 (5):771–784, May 2010. ISSN 1549-9596. doi: 10.1021/ci100062n. URL https://doi.org/10.1021/ci100062n.

[210] Tomohiro Sato, Teruki Honma, and Shigeyuki Yokoyama. Combining Machine Learning and Pharmacophore-Based Interaction Fingerprint for in Silico Screening. *Journal of Chemical Information and Modeling*, 50(1):170–185, January 2010. ISSN 1549-9596, 1549-960X. doi: 10.1021/ci900382e. URL https://pubs.acs.org/doi/10.1021/ci900382e.

[211] D Schlossman, D Withers, P Welsh, A Alexander, J Robertus, and A Frankel. Role of glutamic acid 177 of the ricin toxin A chain in enzymatic inactivation of ribosomes. *Molecular and Cellular Biology*, 9(11):5012–5021, November 1989. ISSN 0270-7306. URL https://www.ncbi.nlm.nih.gov/pmc/articles/PMC363653/.

[212] J. Schmauck and M. Breugst. The potential of pnicogen bonding for catalysis – a computational study. *Organic & Biomolecular Chemistry*, 15(38):8037–8045, October 2017. ISSN 1477-0539. doi: 10.1039/C7OB01599B. URL https://pubs.rsc.org/en/content/articlelanding/2017/ob/c7ob01599b.

[213] Ernst Schonbrunn, Stephane Betzi, Riazul Alam, Mathew P Martin, Andreas Becker, Huijong Han, Rawle Francis, Ramappa Chakrasali, Sudhakar Jakkaraj, Aslamuzzaman Kazi, et al. Development of highly potent and selective diaminothiazole inhibitors of cyclin-dependent kinases. *Journal of medicinal chemistry*, 56 (10):3768–3782, 2013.

[214] A. M. Schreyer and T. L. Blundell. CREDO: a structural interactomics database for drug discovery. *Database*, 2013(0):bat049–bat049, July 2013. ISSN 1758-0463. doi: 10.1093/database/bat049. URL https://academic.oup.com/database/article-lookup/doi/10.1093/database/bat049.

[215] Schrödinger. Schrödinger release 2016-3: Maestro, version 10.7, 2016. https://www.schrodinger.com/.

[216] Richard Sever and Christopher K. Glass. Signaling by Nuclear Receptors. *Cold Spring Harbor Perspectives in Biology*, 5(3), March 2013. ISSN 1943-0264. doi: 10.1101/cshperspect.a016709. URL https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3578364/.

[217] Maxim Shatsky, Ruth Nussinov, and Haim J. Wolfson. A method for simultaneous alignment of multiple protein structures. *Proteins: Structure, Function, and Bioinformatics*, 56(1):143–156, 2004.

[218] Vladimir Sobolev, Anatoli Sorokine, Jaime Prilusky, Enrique E. Abola, and Marvin Edelman. Automated analysis of interatomic contacts in proteins. *Bioinformatics*, 15(4):327–332, 1999. URL http://bioinformatics.oxfordjournals.org/content/15/4/327.short.

[219] Frieda A. Sorgenfrei, Simone Fulle, and Benjamin Merget. Kinome-Wide Profiling Prediction of Small Molecules. *ChemMedChem*, 13(6):495–499, March 2018. ISSN 1860-7187. doi: 10.1002/cmdc.201700180. URL https://onlinelibrary.wiley.com/doi/abs/10.1002/cmdc.201700180.

[220] Christoph Sotriffer, editor. *Virtual screening: principles, challenges, and practical guidelines*. Number v. 48 in Methods and principles in medicinal chemistry. Wiley-VCH, Weinheim, Germany, 2011. ISBN 978-3-527-32636-5. OCLC: ocn659246406.

[221] Thomas Steiner and Gertraud Koellner. Hydrogen bonds with π-acceptors in proteins: frequencies and role in stabilizing local 3d structures. *Journal of Molecular Biology*, 305(3):535–557, January 2001. ISSN 00222836. doi: 10.1006/jmbi.2000.4301. URL http://linkinghub.elsevier.com/retrieve/pii/S0022283600943018.

[222] Katrin Stierand and Matthias Rarey. PoseView – molecular interaction patterns at a glance. *Journal of Cheminformatics*, 2(Suppl 1):P50, May 2010. ISSN 1758-2946. doi: 10.1186/1758-2946-2-S1-P50. URL http://www.ncbi.nlm.nih.gov/pmc/articles/PMC2867186/.

[223] Helena Strömbergsson and Gerard J. Kleywegt. A chemogenomics view on protein-ligand spaces. *BMC bioinformatics*, 10 Suppl 6:S13, June 2009. ISSN 1471-2105. doi: 10.1186/1471-2105-10-S6-S13.

[224] James L. Sudmeier, Elizabeth M. Bradshaw, Kristin E. Coffman Haddad, Regina M. Day, Craig J. Thalhauser, Peter A. Bullock, and William W. Bachovchin. Identification of Histidine Tautomers in Proteins by 2d $^{1}H/^{13}C^{\delta 2}$ One-Bond Correlated NMR. *Journal of the American Chemical Society*, 125(28):8430–8431,

July 2003. ISSN 0002-7863, 1520-5126. doi: 10.1021/ja034072c. URL http://pubs.acs.org/doi/abs/10.1021/ja034072c.

[225] N. Taylor. Proasis2–A Web-Based Protein Structure Database and Visualization System Linking Crystallography and Medicinal Chemistry Research, 2004.

[226] Robin Taylor. Progress in the Understanding of Traditional and Non-traditional Molecular Interactions. In *Comprehensive Medicinal Chemistry III*, pages 67–100. Elsevier, 2017. ISBN 978-0-12-803201-5. doi: 10.1016/B978-0-12-409547-2.12340-6. URL https://linkinghub.elsevier.com/retrieve/pii/B9780124095472123406.

[227] Sajesh P. Thomas, K. Satheeshkumar, Govindasamy Mugesh, and T. N. Guru Row. Unusually short chalcogen bonds involving organoselenium: insights into the Se-N bond cleavage mechanism of the antioxidant ebselen and analogues. *Chemistry (Weinheim an Der Bergstrasse, Germany)*, 21(18):6793–6800, April 2015. ISSN 1521-3765. doi: 10.1002/chem.201405998.

[228] Roberto Todeschini and Viviana Consonni. *Handbook of molecular descriptors*. Wiley-VCH, Weinheim; New York, 2000. ISBN 978-3-527-61311-3. URL http://search.ebscohost.com/login.aspx?direct=true&scope=site&db=nlebk&db=nlabk&AN=241193. OCLC: 839298309.

[229] Gergely Tóth, Charles R. Watts, Richard F. Murphy, and Sándor Lovas. Significance of aromatic-backbone amide interactions in protein structure. *Proteins: Structure, Function, and Bioinformatics*, 43(4):373–381, 2001. ISSN 1097-0134. doi: 10.1002/prot.1050. URL https://onlinelibrary.wiley.com/doi/abs/10.1002/prot.1050.

[230] Gergely Tóth, Simeon G. Bowers, Anh P. Truong, and Gary Probst. The role and significance of unconventional hydrogen bonds in small molecule recognition by biological receptors of pharmaceutical relevance. *Current Pharmaceutical Design*, 13(34):3476–3493, 2007. ISSN 1873-4286.

[231] Donald Voet and Judith G Voet. *Biochemistry*. J. Wiley & Sons, Hoboken, NJ, 2011. ISBN 978-1-118-13993-6. OCLC: 962025871.

[232] Paul G. Wahome, Jon D. Robertus, and Nicholas J. Mantis. Small-Molecule Inhibitors of Ricin and Shiga Toxins. In Nicholas Mantis, editor, *Ricin and Shiga Toxins*, volume 357, pages 179–207. Springer Berlin Heidelberg, Berlin, Heidelberg, 2011. ISBN 978-3-642-27469-5 978-3-642-27470-1. doi: 10.1007/82_2011_177.

[233] A. C. Wallace, R. A. Laskowski, and J. M. Thornton. LIGPLOT: a program to generate schematic diagrams of protein-ligand interactions. *Protein Engineering*, 8 (2):127–134, February 1995. ISSN 0269-2139.

[234] Bohdan Waszkowycz, David E. Clark, and Emanuela Gancia. Outstanding challenges in protein–ligand docking and structure-based virtual screening. *Wiley Interdisciplinary Reviews: Computational Molecular Science*, 1(2):229–259, March 2011. ISSN 1759-0884. doi: 10.1002/wcms.18. URL http://onlinelibrary.wiley.com/doi/10.1002/wcms.18/abstract.

[235] K. Watanabe, H. Dansako, N. Asada, M. Sakai, and G. Funatsu. Effects of chemical modification of arginine residues outside the active site cleft of ricin A-chain on its RNA N-glycosidase activity for ribosomes. *Bioscience, Biotechnology, and Biochemistry*, 58(4):716–721, April 1994. ISSN 0916-8451. doi: 10.1271/bbb.58.716.

[236] Marcey L. Waters. Aromatic interactions in model systems. *Current opinion in chemical biology*, 6(6):736–741, 2002. URL http://www.sciencedirect.com/science/article/pii/S1367593102003599.

[237] Julia Weber, Janosch Achenbach, Daniel Moser, and Ewgenij Proschak. VAMMPIRE-LORD: a web server for straightforward lead optimization using matched molecular pairs. *Journal of Chemical Information and Modeling*, 55(2): 207–213, February 2015. ISSN 1549-960X. doi: 10.1021/ci5005256.

[238] Martin Weisel, Hans-Marcus Bitter, François Diederich, W. Venus So, and Rama Kondru. PROLIX: rapid mining of protein-ligand interactions in large crystal structure databases. *Journal of Chemical Information and Modeling*, 52(6):1450–1461, June 2012. ISSN 1549-960X. doi: 10.1021/ci300034x.

[239] S. A. Weston, A. D. Tucker, D. R. Thatcher, D. J. Derbyshire, and R. A. Pauptit. X-ray structure of recombinant ricin A-chain at 1.8 A resolution. *Journal of molecular biology*, 244(4):410–422, December 1994. ISSN 0022-2836. doi: 10.1006/jmbi.1994.1739. URL http://europepmc.org/abstract/med/7990130.

[240] Paul A. Wiget, Lawrence A. Manzano, Jeff M. Pruet, Grace Gao, Ryota Saito, Arthur F. Monzingo, Karl R. Jasheway, Jon D. Robertus, and Eric V. Anslyn. Sulfur incorporation generally improves Ricin inhibition in pterin-appended glycine-phenylalanine dipeptide mimics. *Bioorganic & Medicinal Chemistry Letters*, 23(24): 6799–6804, December 2013. ISSN 0960-894X. doi: 10.1016/j.bmcl.2013.10.017. URL http://www.sciencedirect.com/science/article/pii/S0960894X13012146.

[241] Ian H. Witten, Eibe Frank, Mark A. Hall, and Christopher J. Pal. The weka workbench. In *Data Mining: Practical Machine Learning Tools and Techniques*,

pages 553 – 571. Morgan Kaufmann, fourth edition edition, 2017. ISBN 978-0-12-804291-5. doi: https://doi.org/10.1016/B978-0-12-804291-5.00024-6.

[242] Maciej Wójcikowski, Piotr Zielenkiewicz, and Paweł Siedlecki. DiSCuS: an open platform for (not only) virtual screening results management. *Journal of Chemical Information and Modeling*, 54(1):347–354, January 2014. ISSN 1549-960X. doi: 10.1021/ci400587f.

[243] Gerhard Wolber and Robert Kosara. Pharmacophores from Macromolecular Complexes with LigandScout. In *Pharmacophores and Pharmacophore Searches*, pages 131–150. John Wiley & Sons, Ltd, 2006. ISBN 978-3-527-60916-1. doi: 10.1002/3527609164.ch6. URL https://onlinelibrary.wiley.com/doi/abs/10.1002/3527609164.ch6.

[244] Maciej Wójcikowski, Piotr Zielenkiewicz, and Pawel Siedlecki. Open Drug Discovery Toolkit (ODDT): a new open-source player in the drug discovery field. *Journal of Cheminformatics*, 7(1):26, June 2015. ISSN 1758-2946. doi: 10.1186/s13321-015-0078-2. URL https://doi.org/10.1186/s13321-015-0078-2.

[245] Maciej Wójcikowski, Michał Kukiełka, Marta M Stepniewska-Dziubinska, and Pawel Siedlecki. Development of a protein–ligand extended connectivity (PLEC) fingerprint and its application for binding affinity predictions. *Bioinformatics*, 35(8):1334–1341, April 2019. ISSN 1367-4803. doi: 10.1093/bioinformatics/bty757. URL https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6477977/.

[246] Zhong-Ru Xie, Jiawen Chen, and Yinghao Wu. Predicting Protein–protein Association Rates using Coarse-grained Simulation and Machine Learning. *Scientific Reports*, 7(1):46622, May 2017. ISSN 2045-2322. doi: 10.1038/srep46622. URL http://www.nature.com/articles/srep46622.

[247] Akihiro Yamaguchi, Kei Iida, Nobuaki Matsui, Shirou Tomoda, Kei Yura, and Mitiko Go. Het-PDB Navi.: a database for protein-small molecule interactions. *Journal of Biochemistry*, 135(1):79–84, January 2004. ISSN 0021-924X.

[248] X. Yan, T. Hollis, M. Svinth, P. Day, A. F. Monzingo, G. W. Milne, and J. D. Robertus. Structure-based identification of a ricin inhibitor. *Journal of Molecular Biology*, 266(5):1043–1049, March 1997. ISSN 0022-2836. doi: 10.1006/jmbi.1996.0865.

[249] Jianyi Yang, Ambrish Roy, and Yang Zhang. BioLiP: a semi-manually curated database for biologically relevant ligand–protein interactions. *Nucleic Acids Research*, 41(Database issue):D1096–D1103, January 2013. ISSN 0305-1048.

doi: 10.1093/nar/gks966. URL http://www.ncbi.nlm.nih.gov/pmc/articles/PMC3531193/.

[250] Jinn-Moon Yang and Chun-Chen Chen. GEMDOCK: a generic evolutionary method for molecular docking. *Proteins*, 55(2):288–304, May 2004. ISSN 1097-0134. doi: 10.1002/prot.20035.

[251] Peng Zhou, Feifei Tian, Fenglin Lv, and Zhicai Shang. Geometric characteristics of hydrogen bonds involving sulfur atoms in proteins. *Proteins: Structure, Function, and Bioinformatics*, 76(1):151–163, July 2009. ISSN 08873585, 10970134. doi: 10.1002/prot.22327. URL http://doi.wiley.com/10.1002/prot.22327.

[252] Fabio Zuccotto. Pharmacophore Features Distributions in Different Classes of Compounds. *Journal of Chemical Information and Computer Sciences*, 43(5): 1542–1552, September 2003. ISSN 0095-2338. doi: 10.1021/ci034068k. URL https://doi.org/10.1021/ci034068k.

Appendix A

Physicochemical property definitions in nAPOLI

A.1 Residue atoms

Table A.1: Residue atom types.

Atom type	Atoms
Acceptor	ALA (O), ARG (O), ASN (O), ASN (OD1), ASP (O), ASP (OD1), ASP (OD2), CYS (O), CYS (SG), GLN (O), GLN (OE1), GLU (O), GLU (OE1), GLU (OE2), GLY (O), HIS (O), HIS (ND1), HIS (NE2), ILE (O), LEU (O), LYS (O), MET (O), MET (SD), PHE (O), PRO (O), SER (O), SER (OG), THR (O), THR (OG1), TRP (O), TYR (O), TYR (OH), VAL (O)
Aromatic	HIS (CG), HIS (ND1), HIS (CD2), HIS (CE1), HIS (NE2), PHE (CG), PHE (CD1), PHE (CD2), PHE (CE1), PHE (CE2), PHE (CZ), TRP (CG), TRP (CD1), TRP (CD2), TRP (NE1), TRP (CE2), TRP (CE3), TRP (CZ2), TRP (CZ3), TRP (CH2), TYR (CG), TYR (CD1), TYR (CD2), TYR (CE1), TYR (CE2), TYR (CZ)
Donor	ALA (N), ARG (N), ARG (NE), ARG (NH1), ARG (NH2), ASN (N), ASN (ND2), ASP (N), CYS (N), CYS (SG), GLN (N), GLN (NE2), GLU (N), GLY (N), HIS (N), HIS (ND1), HIS (NE2), ILE (N), LEU (N), LYS (N), LYS (NZ), MET (N), PHE (N), SER (N), SER (OG), THR (N), THR (OG1), TRP (N), TRP (NE1), TYR (N), TYR (OH), VAL (N)
Hydrophobic	ALA (CB), ARG (CB), ARG (CG), ASN (CB), ASP (CB), CYS (CB), GLN (CB), GLN (CG), GLU (CB), GLU (CG), HIS (CB), ILE (CB), ILE (CG1), ILE (CG2), ILE (CD1), LEU (CB), LEU (CG), LEU (CD1), LEU (CD2), LYS (CB), LYS (CG), LYS (CD), MET (CB), MET (CG), MET (CE), PHE (CB), PHE (CG), PHE (CD1), PHE (CD2), PHE (CE1), PHE (CE2), PHE (CZ), PRO (CB), PRO (CG), THR (CG2), TRP (CB), TRP (CG), TRP (CD2), TRP (CE3), TRP (CZ2), TRP (CZ3), TRP (CH2), TYR (CB), TYR (CG), TYR (CD1), TYR (CD2), TYR (CE1), TYR (CE2), VAL (CB), VAL (CG1), VAL (CG2)
Negative	ASP (OD1), ASP (OD2), GLU (OE1), GLU (OE2)
Positive	ARG (NE), ARG (CZ), ARG (NH1), ARG (NH2), HIS (ND1), HIS (NE2), LYS (NZ)

A.2 Ligand rules

Aromatic labels are assigned to aromatic atoms (‘`[*;a]`’).

Positive labels are assigned to non-negative atoms (‘[!-]’) that fulfill at least one of the following rules:

1. It has a formal charge greater than 0 or its partial charge is greater than 0.4;
2. It is a positive atom (‘[*+]’);
3. It is an N of a nitro-like group (‘[O]∼[N]=[O]’);
4. It is an N that:
   a) belongs to an amine-like (‘C[N]’) or hydrazine-like (‘NN’) or amidine-like (‘[#7][C,P,S]=[N]’) group;
   b) does not belong to a tertiary amine-like (‘C[N](C)C’) or amide-like (‘[#7][C,P,S]=O’) or aniline-like (‘c[N]’) or phenylhydrazine-like (‘c[N][N]’).

Negative labels are assigned to non-positive atoms (‘[!+]’) that fulfill at least one of the following rules:

1. It has a formal charge that is less than 0 or its partial charge is less than -0.4;
2. It is a negative atom (‘[*-]’);
3. It is an O of a nitro-like (‘[O]∼[N]=[O]’) or carboxylic acid-like (‘[H][O][C]=[O]’) or carboxylate ion-like (‘[O-][C]=[O]’) or sulfonic acid-like (‘[H][O][S](=[O])=[O]’) or phosphonic acid-like (‘[H][O][P]([O])=[O]’) group;
4. It is an atom of a sulfonate-like (‘[O-][S](=[O])=[O]’) or phosphonate-like group (‘[O][P]([O-])=[O]’).

Donor labels are assigned to atoms that fulfill at least one of the following rules:

1. It is a donor atom according to the HBDAPlugin from Chemaxon;
2. It is any atom other than carbon that is bound to hydrogen (‘[!#1!#6][H]’).

Acceptor labels are assigned to atoms that fulfill at least one of the following rules:

1. It is an acceptor atom according to the HBDAPlugin from Chemaxon;
2. It is a N/O/S atom that is not:
   a) a *positive* atom;
   b) a N of a tertiary amine-like (‘C[N](C)C’) or aniline-like (‘c[N:1]’) or nitro-like (‘[O]∼[N]=[O]’) group;
   c) an aromatic N with three total bonds (‘[nX3]’);
   d) a N/S of a amide-like (‘[#7][C,P,S]=O’) group;

e) a S of a sulfonic acid-like (‘[H][O][S](=[O])=[O]’) or sulfonate-like (‘[O-][S](=[O])=[O]’);

f) a S with bond order equal to 6 (‘[Sv6]’).

Hydrophobic labels are assigned to atoms that fulfill the following rules:

1. It is a C/F/Cl/Br/I atom that is not:

   a) a carbon bound to an O or N atoms [218];

   b) an acceptor atom;

   c) a donor atom;

   d) a negative atom;

   e) a positive atom.

Appendix B

Physicochemical property definitions in LUNA

Hydrogen donor labels are assigned to atoms that fulfill the following rules:

1. It is a tertiary amine N (‘`$([Nv3](-C)(-C)-C)`’) that is not:
   a) an amide-like N (‘`$([#7][C,P,S]=O)`’);
   b) an amidine-like N (‘`[N;![$([#7][C,P,S]=O)];$(N=[CX3][N;![$([#7][C,P,S]=O)]]),$(N[CX3]=[N;![$([#7][C,P,S]=O)]])]`’).
2. It is a tautomeric aromatic N (‘`[$(n[n;H1]),$(nc[n;H1])]`’) not in a tetrazole (‘`[nR1r5;$(n:n:n:n:c),$(n:n:n:c:n)]`’);
3. It is a tautomeric N in a guanidine-like group (‘`[$([NX2HO]=[CHOX3](N)N)]`’);
4. It is a N/O/S atom bound to a hydrogen atom (‘`[N!HOv3,N!HO+v4,nH+0,OH2v2,OH+0,SH+0]`’), where the N is not acidic (‘`[$([NH,NHO-1](S(=O)(=O))(C(=O))),$([NH1,NHO-1;R](C(=O))(C(=[O,S])))]`’) and does not belong to a tetrazole (‘`[nR1r5;$(n:n:n:n:c),$(n:n:n:c:n)]`’), while O and S are also not acidic (‘`[$([O][C,S,P](=[O,S]))]`’);
5. It is a tautomeric O in a ketene acetal-like group (‘`[$([O;H1,HO&-1]-[#6;X3]-,:[#8])]`’).

Halogen donor labels are assigned to atoms that fulfill the following rules:

1. It is a Cl/Br/I bound to C/S (‘`[$([Cl,Br,I;X1]-[#6])]`’).

Chalcogen donor labels are assigned to atoms that fulfill the following rules:

1. It is a divalent S/Se/Te bound to C/S (‘`[$([#16,#34,#52;v2;H0](-,:[#6,#16])-,:[#6,#16]),$([#16,#34,#52;v2;H1][#6,#16])]`’) or in an isothiazole-like group (‘`[$([#16,#34,#52;v2;H0;a](n)c)]`’).

Hydrogen/Halogen/Chalcogen acceptor labels are assigned to atoms that fulfill the following rules:

1. It is a tautomeric aromatic N (‘`[$([n;H1]n),$([n;H1]cn)]`’) not in a tetrazole (‘`[nR1r5;$(n:n:n:n:c),$(n:n:n:c:n)]`’);
2. It is an aromatic N with a double bond and no hydrogen (‘`[n;+0;H0;!X3]`’);
3. It is an N that is not:
   a) a positive N (‘`[*;+1,+2,+3]`’);
   b) an aniline-like N (‘`[$(N[a])]`’);
   c) an amide-like N (‘`[$([#7][C,P,S]=O)]`’);
   d) an amidine-like N (‘`[N;![$([#7][C,P,S]=O)];$(N=[CX3][N;![$([#7][C,P,S]=O)]]),$(N[CX3]=[N;![$([#7][C,P,S]=O)]])]`’);
   e) a nitro-like N (‘`[$([N+]-[O-])]`’);
   f) a basic N (‘`[$([N;H2&+0][CX4]),$([N;H1&+0]([CX4])[CX4]),[$([N;H0&+0]([CX4])([CX4])[CX4])]`’).
4. It is a chalcogen (‘`[O,$([S;!v4;!v6])]`’).

Weak hydrogen donor labels are assigned to atoms that fulfill the following rules:

1. It is a C bound to at least one hydrogen (’`[#6;!H0]`’).

Weak hydrogen acceptor labels are assigned to atoms that fulfill the following rules:

1. It is a neutral aromatic O or S (‘`[o,s;+0]`’);
2. It is a F bound to C (‘`[F;$(F-[#6]);!$(FC[F,Cl,Br,I])]`’).

Positively ionizable labels are assigned to atom and atom groups that fulfill the following rules:

1. It is a basic N (‘`[$([N;H2&+0][CX4]),$([N;H1&+0]([CX4])[CX4]),[$([N;H0&+0]([CX4])([CX4])[CX4])]`’);
2. It is a guanidine-like group (‘`N[CH0X3](=N)N`’);
3. It is an amidine-like group (‘`[N;![$([#7][C,P,S]=O)]]=[CX3][N;![$([#7][C,P,S]=O)]]`’);
4. It is a 4-aminopyridine (‘`Nc1cc[nH0]cc1`’) or 2-aminopyridine (‘`Nc1cccc[nH0]1`’);
5. It is an imidazole-like group (‘`[n;R1]1[c;R1][n;R1][c;R1][c;R1]1`’);
6. It is positive atom not bound to a negative atom (‘`[[*;+1,+2,+3];!$(*~[*;-1,-2,-3])]`’).

Negatively ionizable labels are assigned to atom and atom groups that fulfill the following rules:

1. It is a carboxylic-like group (‘C(=[O,S])-[O;H1,HO&-1]’);
2. It is a ketene acetal-like group (‘[O;H1,HO&-1]-[#6;X3]-,:[#8]’);
3. It is a tetrazole (‘c1nnn[nH,n-1]1’);
4. It is a barbiturate-like group (‘O=C1CC(=O)[NH1,NHO-1;R]C(=O)[NH1,NHO-1;R]1’);
5. It is a thiazolidinedione-like group (‘O=C1[NH1,NHO-1;R]C(=[O,S])[SX2HOR]C1’);
6. It is a diformamide-like group (‘[NH1,NHO-1;R](C(=O))(C(=O))’);
7. It is one of the hydroxamic acid forms: O anion (‘C(=O)[NX3]-[O;H1,HO&-1]’), N anion (‘C(=O)[N-1]-[OH1]’) or its resonance form (‘C(-[O-1])=N-[OH1]’);
8. It is a sulfuric-like (‘S(=[O,S])(=O)(-O)-[O;H1,HO&-1]’) or sulfonic-like (‘S(=[O,S])(=O)-[O;H1,HO&-1]’) or sulfinic-like (‘S(=[O,S])-[O;H1,HO&-1]’) acid group;
9. It is a acyl sulfonamide-like (‘[NH,NHO-1](S(=O)(=O))(C(=O))’) or a sulfonamide-like (‘[N;!HO,HO&-1]S(=O)(=O)’) group;
10. It is a phosphoric-like (‘P(=[O,S])(-O)(-O)-[O;H1,HO&-1]’) or phosphonic-like (‘P(=[O,S])(-O)-[O;H1,HO&-1]’) or phosphinic-like (‘P(=[O,S])-[O;H1,HO&-1]’) acid groups;
11. It is a negative atom not bound to a positive atom (‘[[*;-1,-2,-3];!$(*~[*;+1,+2,+3])]’).

Nucleophile labels are assigned to atom groups that fulfill the following rules:

1. It is a halogen from a haloalkane (‘[F,Cl,Br,I;X1][#6]’);
2. It is an O in a carbonyl but not in a carboxylic-like group (‘[O;!$(O=C[O;H1,HO&-1])]=[C;!$(C([O;H1,HO&-1])=O)’);
3. It is an O in alcohol but not in a carboxylic-like group (‘[O;v2;H1;!$(OC=[O,S])][#6;!$(C(=[O,S])[O;v2;H1])]’);
4. It is a cyano-like N (‘N#C’);
5. It is an O in a water molecule (‘[O;v2;H2]’);
6. It is a sulfonyl-like O (‘[$(O=[S;v4,v6]([#6])[#6])]=[$([S;v4,v6]([#6])([#6])=O)]’);
7. It is ketene acetal-like O (‘[$([O;H1,HO&-1]-[#6;X3]-,:[#8])][$([#6;X3](-,:[#8])[O;H1,HO&-1])]’);
8. It is a nitro-like O (‘[$(O=[N;D3;+][O-]),$([O-][N;D3;+]=O)]~[$([N;D3;+](=O)[O-])]’);
9. It is an ether-like O (‘[$([#8;v2]([#6])[#6])][$([#6][#8;v2][#6])]’);

10. It is a thioether-like S ('`[$([#16;v2]([#6])[#6])][$([#6][#16;v2][#6])]`');
11. It is a hydroxylamine-like O ('`[$(O-[#7;H0;X3]([#7,#6])[#7,#6])][$([#7;H0;X3]([#7,#6])([#7,#6])O)]`').

Electrophile labels are assigned to atom groups that fulfill the following rules:

1. It is a haloalkane C ('`[#6][F,Cl,Br,I;X1]`');
2. It is a C in a carbonyl but not in a carboxylic-like group ('`[#6;!$(C([O;H1,H0&-1])=[O,S])]=[O;!$([O;H1,H0&-1]C=[O,S])]`');
3. It is a cyano-like C ('`C#N`');
4. It is a nitro-like N ('`[$([N;D3;+](=O)[O-])]~[$(O=[N;D3;+][O-]),$([O-][N;D3;+]=O)]`').

Aromatic labels are assigned to atom groups that fulfill the following rules:

1. 4-membered rings ('`[a;r4,!R1&r3]1:[a;r4,!R1&r3]:[a;r4,!R1&r3]:[a;r4,!R1&r3]:1`');
2. 5-membered rings ('`[a;r5,!R1&r4,!R1&r3]1:[a;r5,!R1&r4,!R1&r3]:[a;r5,!R1&r4,!R1&r3]:[a;r5,!R1&r4,!R1&r3]:[a;r5,!R1&r4,!R1&r3]:1`');
3. 6-membered rings ('`[a;r6,!R1&r5,!R1&r4,!R1&r3]1:[a;r6,!R1&r5,!R1&r4,!R1&r3]:[a;r6,!R1&r5,!R1&r4,!R1&r3]:[a;r6,!R1&r5,!R1&r4,!R1&r3]:[a;r6,!R1&r5,!R1&r4,!R1&r3]:[a;r6,!R1&r5,!R1&r4,!R1&r3]:1`');
4. 7-membered rings ('`[a;r7,!R1&r6,!R1&r5,!R1&r4,!R1&r3]1:[a;r7,!R1&r6,!R1&r5,!R1&r4,!R1&r3]:[a;r7,!R1&r6,!R1&r5,!R1&r4,!R1&r3]:[a;r7,!R1&r6,!R1&r5,!R1&r4,!R1&r3]:[a;r7,!R1&r6,!R1&r5,!R1&r4,!R1&r3]:[a;r7,!R1&r6,!R1&r5,!R1&r4,!R1&r3]:[a;r7,!R1&r6,!R1&r5,!R1&r4,!R1&r3]:1`');
5. 8-membered rings ('`[a;r8,!R1&r7,!R1&r6,!R1&r5,!R1&r4,!R1&r3]1:[a;r8,!R1&r7,!R1&r6,!R1&r5,!R1&r4,!R1&r3]:[a;r8,!R1&r7,!R1&r6,!R1&r5,!R1&r4,!R1&r3]:[a;r8,!R1&r7,!R1&r6,!R1&r5,!R1&r4,!R1&r3]:[a;r8,!R1&r7,!R1&r6,!R1&r5,!R1&r4,!R1&r3]:[a;r8,!R1&r7,!R1&r6,!R1&r5,!R1&r4,!R1&r3]:[a;r8,!R1&r7,!R1&r6,!R1&r5,!R1&r4,!R1&r3]:[a;r8,!R1&r7,!R1&r6,!R1&r5,!R1&r4,!R1&r3]:1`');
6. 9-membered rings ('`[a;r9,!R1&r8,!R1&r7,!R1&r6,!R1&r5,!R1&r4,!R1&r3]1:[a;r9,!R1&r8,!R1&r7,!R1&r6,!R1&r5,!R1&r4,!R1&r3]:[a;r9,!R1&r8,!R1&r7,!R1&r6,!R1&r5,!R1&r4,!R1&r3]:[a;r9,!R1&r8,!R1&r7,!R1&r6,!R1&r5,!R1&r4,!R1&r3]:[a;r9,!R1&r8,!R1&r7,!R1&r6,!R1&r5,!R1&r4,!R1&r3]:[a;r9,!R1&r8,!R1&r7,!R1&r6,!R1&r5,!R1&r4,!R1&r3]:[a;r9,!R1&r8,!R1&r7,!R1&r6,!R1&r5,!R1&r4,!R1&r3]:[a;r9,!R1&r8,!R1&r7,!R1&r6,!R1&r5,!R1&r4,!R1&r3]:[a;r9,!R1&r8,!R1&r7,!R1&r6,!R1&r5,!R1&r4,!R1&r3]:1`').

Hydrophobic labels are assigned to atoms that fulfill the following rules:

1. It is a divalent sulfur, halogen (except fluorine), or a carbon not bound to an electronegative atom (‘`[s,S&H0&v2,Br,I,Cl,At,[#6;+0;![#6;$([#6]~[#7,#8,#9])]];+0]`’).

Amide labels are assigned to atom groups that fulfill the following rules:

1. It is an amide-like group (‘`[NX3][CX3](=[OX1])`’).

Atom labels are assigned to atoms that fulfill the following rules:

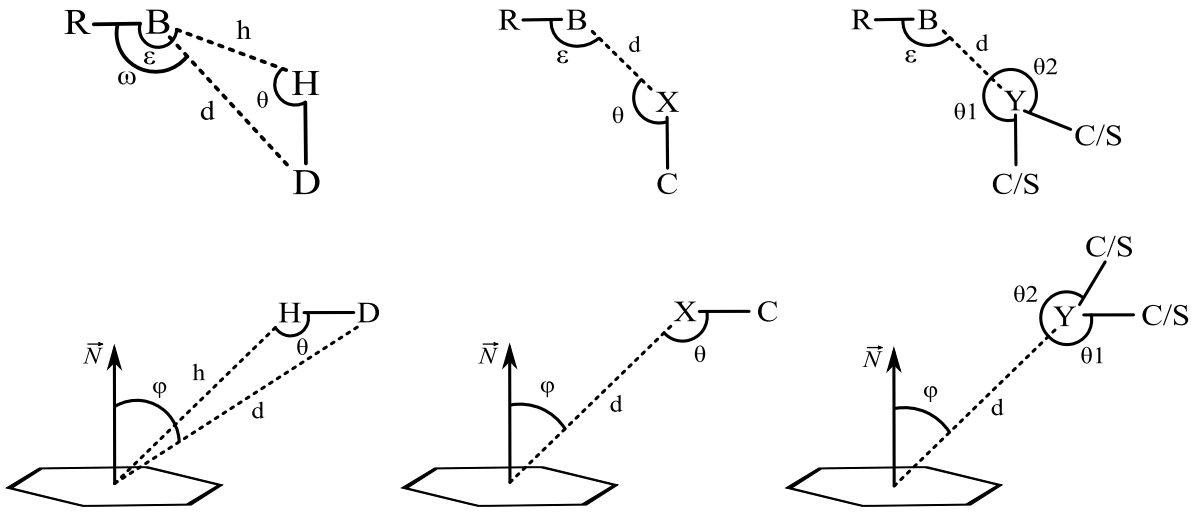
1. It is any heavy atom (‘`[!#1])`’).

Appendix C

Geometrical criteria for computing molecular interactions

In this section, we present the geometrical criteria and the models employed in the calculation of molecular interactions in LUNA. Figures C.1, C.2, and C.3 show the geometrical models for most of the interactions. Not all interactions are shown in the diagrams because they require only the evaluation of Euclidean distances between two atoms (atom groups). For these cases, the methods are discussed directly in their respective section.

It is important to mention that all geometrical criteria and models presented in this section consist of the default model of LUNA, but all of them are customizable.



Hydrogen bond

Weak hydrogen bond

Halogen bond

Chalcogen bond

Figure C.1: Models for calculating hydrogen, weak hydrogen, halogen, and chalcogen bonds. The definitions for each letter and angle depicted in the diagrams are explained in their respective interaction section.

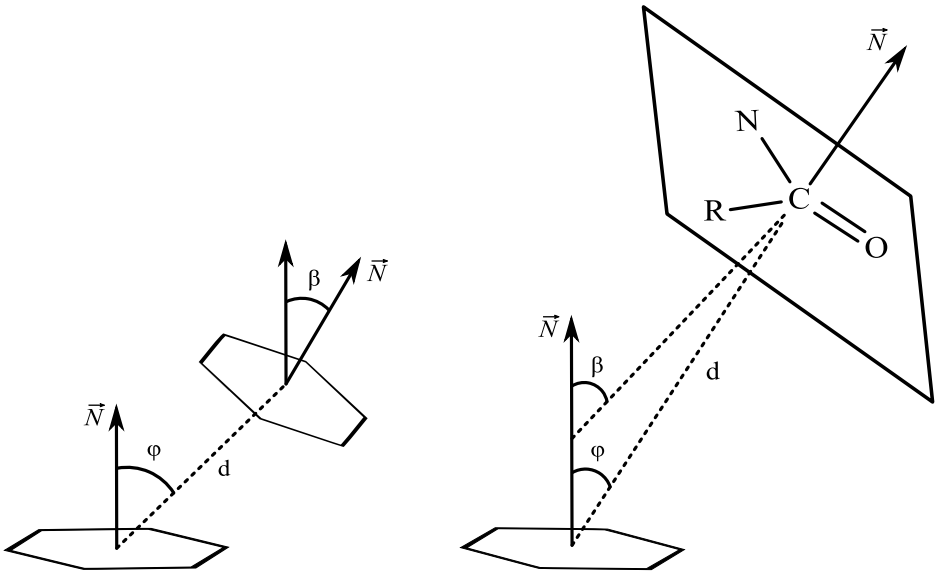


Figure C.2: Models for calculating aromatic stackings and amide-π stackings. The definitions for each letter and angle depicted in the diagrams are explained in their respective interaction section.

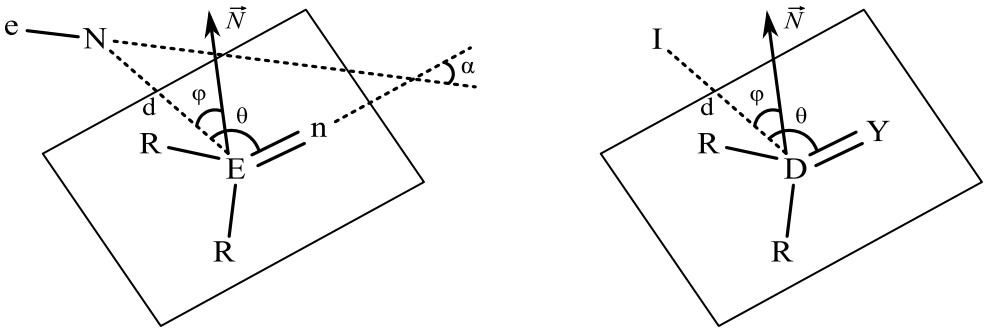


Figure C.3: Models for calculating dipole-dipole (multipolar interactions) and ion-dipole interactions. The definitions for each letter and angle depicted in the diagrams are explained in their respective interaction section.

C.1 Hydrogen bonds

Hydrogen bonds are identified according to [161] and its calculation depends on five parameters as shown in Figure C.1, where D, H, B, and R are the donor, hydrogen, a Lewis base (acceptor), and an atom covalently bound to the acceptor, respectively. Below we provide the default values extracted from literature [13, 83, 112, 161] for each of the parameters presented in Figure C.1.

- Default distance thresholds: $d \leq 3.9$Å and $h \leq 2.8$Å, respectively;

- Default angle thresholds: $\theta, \varepsilon, \omega \geq 90°$.

In Figure C.1, the model depicts only one hydrogen for the donor and one neighbor for the acceptor. However, if the donor contains two or more hydrogens, each one of them is evaluated against an acceptor, which can result in multiple hydrogen bonds. Similarly, when acceptors contain more than one neighbor, the hydrogen bond is only accepted if the angles involving these atoms match the angle criteria.

Lastly, the algorithm has two modes for applying the above criteria, a strict and loose one. In the strict mode, donor atoms must have their hydrogens explicitly defined as the algorithm requires the hydrogens' coordinates. However, not all PDB structures contain hydrogens, and it also may happen that a potential donor is not in its ionized or tautomeric form, which would impede the algorithm to detect the hydrogen bonds. Also, consider the dynamic of water molecules. In Open Babel, for instance, when hydrogens are included in the structure, usually water molecules can have their hydrogens added in different manners. As a consequence, if the exact position of hydrogens is taken into account, several results would be possible for the same system. Therefore, we also provide a loose mode bearing in mind the dynamic of the system. When the loose mode is activated, hydrogen bonds are identified despite the presence of hydrogens. Furthermore, with this mode, all hydrogen bonds involving solvents are loosely validated even if they contain hydrogens explicitly defined, which is a measure acknowledging their dynamic and multiple possible hydrogen placement.

The loose mode works similarly to the method proposed by [161]: angles involving the hydrogens are ignored and the distance between the hydrogen and the acceptor is calculated as if the hydrogen was placed 1Å away from the donor in the direction of the acceptor. As a mathematical expression it can be defined as follows:

$$h = d - 1 \leq 2.8\text{Å} \tag{C.1}$$

It is noteworthy that this method may generate more interactions than the strict one, and some of the identified interactions may be false positives.

C.2 Weak hydrogen bonds

Weak hydrogen bonds are identified through two models depending on the acceptor atom. For single atoms (Lewis bases), the model is the same as presented for hydrogen bonds. However, herein, D is a weak hydrogen donor, i.e., a carbon with an attached hydrogen bond; while, B can be an acceptor or a weak acceptor, which are defined as

any aromatic oxygen/sulfur or fluorine [34, 66, 174, 135]. In its turn, the second model comprises aromatic rings as acceptors. Below we provide the default values extracted from literature [66, 179, 221] for each of the parameters presented in Figure C.1.

- Conventional weak hydrogen bonds:
  - Default distance thresholds: $d \leq 4$Å and $h \leq 3$Å, respectively;
  - Default angle thresholds: $\theta \geq 110°$ and $\varepsilon, \omega \geq 90°$;
- Weak hydrogen bonds involving aromatic rings:
  - Default distance thresholds: $d \leq 4.5$Å and $h \leq 3.5$Å, respectively. The distances are calculated in relation to the ring centroid;
  - Default angle thresholds: $\theta \geq 120°$ and $\varphi \leq 40°$, where φ (displacement angle) is the angle formed by the ring normal ($\overrightarrow{N}$) and the vector between the ring centroid and the donor atom.

Similarly to hydrogen bonds, multiple weak hydrogen bonds can also be identified whether more than one hydrogen is covalently bound to the weak hydrogen donor. It also evaluates all possible angles with the acceptor's neighbor.

Also, as presented for hydrogen bonds, weak hydrogen bonds can also work on either strict or loose modes. Thus, similarly to hydrogen bonds, the distance expression in the loose mode is defined as follows:

$$h = d - 1 \leq 3\text{Å} \tag{C.2}$$

C.3 Water-bridged hydrogen bond

Water-bridged hydrogen bonds are identified by directly performing a search for pairs of compound-water hydrogen bonds, where both bonds involve the same water molecule. Departing from a valid pair of hydrogen bonds, a new interaction connecting the involved compounds is created and labeled *water-bridged hydrogen bond*.

These special hydrogen bonds are not identified by default because the interactions required for its computation already account for its existence implicitly. However, if necessary this interaction can be easily turned on through a flag in the method for calculating interactions.

C.4 Halogen bond

Halogen bonds are identified according to the model presented in Figure C.1, where X, C, B, R are a halogen, a carbon bound to the halogen, a Lewis base (acceptor), and an atom covalently bound to the acceptor, respectively. Below we provide the default values extracted from literature [8, 119] for each of the parameters presented in Figure C.1.

- Conventional halogen bonds:
  - Default distance threshold: $d \leq 4$Å;
  - Default angle thresholds: $\theta \geq 120°$ and $\varepsilon \geq 80°$;
- Halogen bond involving aromatic rings:
  - Default distance threshold: $d \leq 4.5$Å, where d is calculated in relation to the ring centroid;
  - Default angle thresholds: $\theta \geq 120°$ and $\varphi \leq 60°$, where φ (displacement angle) is the angle formed by the ring normal ($\overrightarrow{N}$) and the vector between the ring centroid and the halogen.

At pointed out by [39], multiple carbons may be bound to the halogen, which, in its turn, may result in multiple halogen bonds with the same acceptor. Moreover, when the acceptor contains more than one neighbor, the tool will evaluate all possible angles formed with these atoms, and the interaction is only accepted if all angles match the criteria.

C.5 Chalcogen bond

Chalcogen bonds are identified according to the model presented in Figure C.1, where Y, C/S, B, R are a chalcogen, a carbon/sulfur bound to the chalcogen, a Lewis base (acceptor), and an atom covalently bound to the acceptor, respectively. Below we provide the default values extracted from literature [2, 117, 134] for each of the parameters presented in Figure C.1.

- Conventional chalcogen bonds:

  - Default distance threshold: $d \leq 4$Å;
  - Default angle thresholds: $\theta \geq 120°$ and $\varepsilon \geq 80°$;

- Chalcogen bond involving aromatic rings:
  - Default distance threshold: $d \leq 4.5$Å, where d is calculated in relation to the ring centroid;
  - Default angle thresholds: $\theta \geq 120°$ and $\varphi \leq 60°$, where φ (displacement angle) is the angle formed by the ring normal ($\overrightarrow{N}$) and the vector between the ring centroid and the chalcogen.

In hydrogen and halogen bonds, we mentioned that the donor atoms might contain more than one neighbor covalently bound to it. However, note that this information is only explicitly depicted in chalcogen bonds. That is because chalcogen bonds are mainly established by divalent chalcogens. Consequently, chalcogen bonds could also establish multiple chalcogen bonds to the acceptor.

Finally, when multiple neighbors are bound to the acceptor atom, all angles formed with them should match the criteria; otherwise, the interaction is not accepted.

C.6 Aromatic stacking

Aromatic stackings are identified as presented in Figure C.2 and are classified according to their geometrical arrangements into nine different stackings (Figure C.4) as in [23]. Below we provide the default values extracted from literature [23, 42] for each of the parameters presented in Figure C.2.

- Default distance threshold: $d \leq 6$Å, where d is calculated in relation to the ring centroids;
- Default angle thresholds: each specific stacking depends on a combination of φ and β angles, as shown in Figure C.4, where φ (displacement angle) is the angle formed by the ring normal ($\overrightarrow{N}$) and the vector between the two ring centroids; while β (dihedral angle) is the angle between the two ring planes, which is calculated by the angle formed by the ring normals. The displacement angle is calculated using both rings as references, and the smallest angle is chosen for defining the stacking type.

If the user decides not to use the angle criteria, the interaction will be labeled as a general *aromatic stacking*.

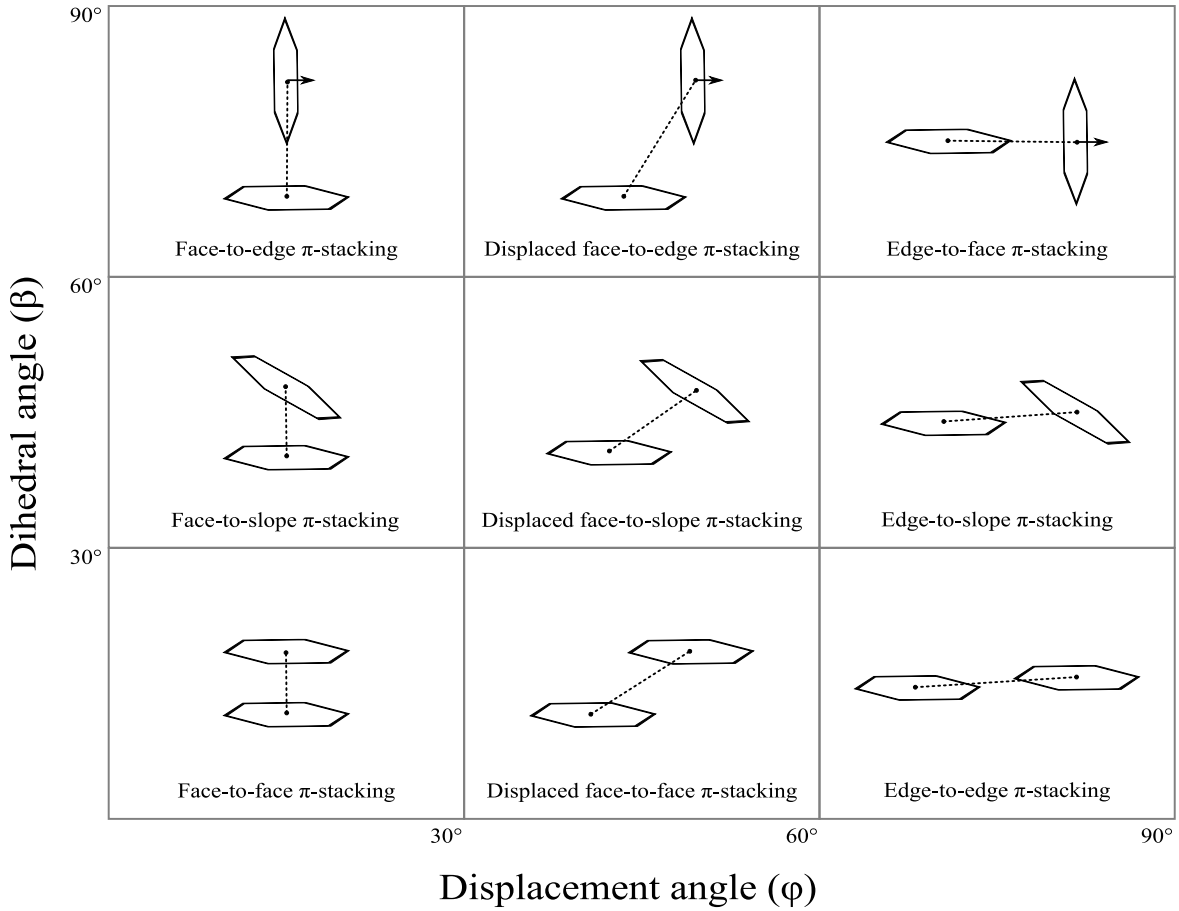


Figure C.4: Classification of aromatic stackings according to the angles φ (displacement angle) and β (dihedral angle) given two aromatic rings.

C.7 Amide-π stacking

Amide-π stackings are identified according to the model presented in Figure C.2. Below we provide the default values extracted from literature [52, 85, 101, 221] for each of the parameters presented in Figure C.2.

- Default distance threshold: $d \leq 4.5$Å, where d is calculated in relation to the ring and amide centroids;
- Default angle thresholds: $\varphi, \beta \leq 30°$, where φ (displacement angle) is the angle formed by the ring normal ($\overrightarrow{N}$) and the vector between the ring and amide centroids; while β (dihedral angle) is the angle between the ring and amide planes, which is calculated by the angle formed by their normals.

C.8 Dipole-dipole or multipolar interactions

Multipolar interactions are identified according to the model presented in Figure C.3, where E═n and N—e are the dipoles containing the interacting electrophile (E) and nucleophile (N); and R are the atoms covalently bound to the electrophile.

There are four possible arrangements for favorable multipolar interactions [181]: *parallel multipolar*, *antiparallel multipolar*, *orthogonal multipolar*, and *tilted multipolar*. Below we provide the default values extracted from literature [181] for each possible arrangement and the parameters presented in Figure C.3.

- Default distance threshold: $d \leq 4$Å, where d is calculated in relation to the nucleophilic and electrophilic atoms;
- Default angle thresholds: $70° \leq \theta \leq 110°$ and $\varphi \leq 40°$, where φ (displacement angle) is the angle formed by the electrophile normal ($\overrightarrow{N}$) and the vector connecting the interacting electrophile and nucleophile. In its turn, α is the angle formed by the dipole vectors and determines the multipolar arrangements as follows:
  - Parallel multipolar: $\alpha \leq 25°$;
  - Antiparallel multipolar: $\alpha \geq 155°$;
  - Orthogonal multipolar: $70° \leq \alpha \leq 110°$;
  - Tilted multipolar: any α not comprised by the above criteria.

The algorithm also detects unfavorable dipole-dipole interactions, which are classified either as *unfavorable nucleophile-nucleophile* or *unfavorable electrophile-electrophile*. However, the arrangements presented above are not employed for unfavorable interactions and, therefore, the tool only evaluates the distance d and angles θ and φ in the same way presented for the favorable interactions.

Also, it is noteworthy that although the diagram depicts the classic dipole-dipole interaction involving a carbonyl-like structure in the electrophile side of the interaction, other non-planar substructures are also accepted.

Lastly, in cases where the coordinate of the nucleophile neighbor (e) cannot be determined (e.g., hydrogens in water molecules), the angle α is not available and, therefore, the interaction is classified as a general *multipolar interaction* since it is not possible to define the proper dipole arrangement. In its turn, when the coordinate of the electrophile partner (n) is not available, it is only possible to calculate the distance between E and N. Thus, we also opted for classifying the interaction as a general *multipolar interaction*. However, the latter case is unlikely to occur as all default pharmacophore rules for electrophiles comprehend heavy atoms bound to the electrophilic atom (see Section B).

C.9 Ion-dipole interaction

Ion-dipole interactions are identified similarly to multipolar interactions (Figure C.3), where I is the ion centroid and D═Y is the dipole, having D as the electrophile when Y is the nucleophile and vice versa. Given the possible combinations of ions and dipoles, there are two favorable (*cation-nucleophile* and *anion-electrophile*) and two unfavorable interactions (*unfavorable anion-nucleophile* and *unfavorable cation-electrophile*).

Below we provide the default values extracted from literature [181] for each of the parameters presented in Figure C.3.

- Default distance threshold: $d \leq 4.5$Å, where d is calculated in relation to the nucleophilic/electrophilic atom (D) and the ion centroid (I);
- Default angle thresholds: $\theta \geq 60°$ and $\varphi \leq 40°$, where φ (displacement angle) is the angle formed by the dipole normal ($\overrightarrow{N}$) and the vector connecting the ion centroid and the interacting electrophile/nucleophile.

Similarly to dipole-dipole interactions, both planar and non-planar dipoles are accepted. Also, in cases where the coordinate of the atom Y cannot be determined, only the distance between D and I is evaluated.

C.10 Ionic and repulsive interactions

Interactions involving ions are classified as either *ionic* or *repulsive* when the ions are oppositely or similarly charged, respectively. The only parameter evaluated in these interactions is the distance between the ion centroids whose upper-limit threshold is 6Å [15, 24, 83, 155].

C.11 Salt bridge

Since a salt bridge consists of a hydrogen bond and an ionic interaction occurring simultaneously between the same interacting partners, salt bridges are identified by directly performing a search for pairs of hydrogen bonds and ionic interactions that match

the mentioned requirement. However, as hydrogen bonds are modeled as an atom-atom interaction and ionic interactions as a group-group interaction, this requirement is fulfilled when the acceptor belongs to one ionic group and the donor to the other.

Salt bridges are not identified by default because the interactions required for their computation already account for their existence implicitly. However, if necessary this interaction can be easily turned on through a flag in the method for calculating interactions.

C.12 Cation-π interaction

Cation-π interactions are identified when the cation and aromatic ring centroids are up to 6Å apart [89].

C.13 Hydrophobic interaction

Hydrophobic interactions can be modeled as atom-atom interactions or as surface-surface contacts, which is the default approach in our tool. The former is identified when any two hydrophobic atoms are up to 4.5Å apart [83, 121, 157].

Surface-surface contacts, in its turn, are identified as follows. Firstly, the tool computes hydrophobic interactions between atoms as explained in the atom-atom model. Then, it identifies all hydrophobic atoms covalently bound to each other and merges them to form a hydrophobic cluster/island, called *hydrophobe*. Finally, the tool converts each atom-atom interaction to its hydrophobe-hydrophobe form by identifying the hydrophobic clusters comprehending each interacting atom and attributing this interaction to them. However, it may be possible that not all hydrophobic atoms in a cluster participate in surface-surface contact. For that reason, each interaction contains the hydrophobic cluster information as a whole and keeps track of which of their specific atoms are in contact.

C.14 Covalent interaction

Covalent bonds are automatically obtained from Open Babel or RDKit. However, when the precomputed set of properties for atoms is used during the physicochemical feature assignment (see Section 2.3.2.1), it is necessary to manually compute covalent bonds as these intermediary tools are not used. To do so, we implemented the Open Babel model for covalent bonds detection, whose expression is defined below:

$$0.4 \leq d \leq A_{cov} + B_{cov} + 0.45 \tag{C.3}$$

Where d is the Euclidean distance between two atoms A and B, while A_{cov} and B_{cov} are their covalent radii, which are derived from Open Babel.

C.15 Atom overlap

Atom overlap identifies artifacts generated by low-resolution structures and homology models, which consist of the unnatural overlap of two atoms. An overlap is defined as two atoms not covalently bound separated from each other by less than or equal to the sum of their covalent radii [121].

C.16 Van der Waals clash

Van der Waals clashes are identified as in [184], which describes a van der Waals clash by the following expression:

$$A_{vdw} + B_{vdw} - d \geq 0.6 \tag{C.4}$$

Where d is the Euclidean distance between two atoms A and B, A_{vdw} and B_{vdw} are their van der Waals radii, and 0.6 is the threshold for van der Waals clashes. Van der Waals radii are derived from Open Babel.

C.17 Van der Waals interaction

Van der Waals interactions are identified as in [121], which describes a van der Waals by the following expression:

$$d \leq A_{vdw} + B_{vdw} + 0.1 \tag{C.5}$$

Where d is the Euclidean distance between two atoms A and B, A_{vdw} and B_{vdw} are their van der Waals radii, and 0.1 is a margin of error. Van der Waals radii are derived from Open Babel.

C.18 Proximal interactions

Proximal interactions are defined as any two atoms separated from each other by at least 2Å and at most 6Å.

C.19 Intramolecular interactions

Interactions involving atoms or atom groups from the same molecule are calculated using the specific interaction methods described in previous sections. However, there is an additional criterion for intramolecular interactions that consists of evaluating how many bonds separate the interacting atoms or atom groups. Note that two different molecules covalently bound to each other also fall into the rules discussed in this section.

For van der Waals clashes, the interaction is only accepted if the atoms are separated by more than 4 bonds. This threshold was defined to avoid invalid clashes typically found in structures like the one shown in Figure C.5.

For the other interactions, the number of bonds separating two atoms must be higher than 3 [184]; otherwise, the interaction is ignored. The only exception is the hydrophobic interaction which is not considered for intramolecular interactions.

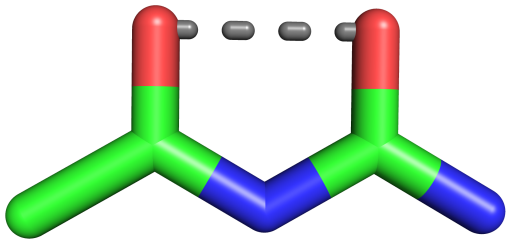


Figure C.5: Example of an invalid van der Waals clash (gray dashed line) between two atoms separated by four bonds.