Universidade de Lisboa

Faculdade de Farmácia

ChemSlotProtein: Expanding the druggable universe to unlock chemists´creativity

Ismael Rufino de Carvalho

Dissertation supervised by Professor: Dr.ª Natália Luísa de Moura Aniceto
andco-supervised by Professor . Dr.ª Rita Alexandra do Nascimento Cardoso Guedes

Mestrado em Química Medicinal e Biofarmacêutica

2023

Universidade de Lisboa

Faculdade de Farmácia

ChemSlotProtein: Expanding the druggable universe to unlock chemists´creativity

Ismael Rufino de Carvalho

Dissertation supervised by Professor: Dr.ª Natália Luísa de Moura Aniceto and co-supervised by Professor . Dr.ª Rita Alexandra do Nascimento Cardoso Guedes

Mestrado em Química Medicinal e Biofarmacêutica

2023

Acknowledgements

We acknowledge Fundação para a Ciência e a Tecnologia (FCT) for the financial support for Computational Medicinal Chemistry EXPL/QUI-OUT/1288/2021, CPCA/A2/6972/2020, and UIDB/04138/2020, and UIDP/04138/2020 to iMed.ULisboa

I would like to extend my heartfelt thanks to several individuals who played pivotal roles in my academic journey and my personal growth.
First and foremost, I would like to thank Professor Rita Guedes. Her guidance and support have been invaluable throughout this journey, helping me to grow both in the academic world and as an individual. I also want to express my deep gratitude to Professor Natália Aniceto, who has always been available to advise me on all my questions, guiding me with wisdom and patience. My mother, Joelma Cristina, and my grandmother, Maria Cardoso da Silva, deserve special thanks. Their unwavering support was a beacon in challenging moments, and I am forever grateful for everything they have done for me. I cannot fail to mention my laboratory friends, Bruno Gomes, Carlota Silva, Filipe Estrada, Nuno Martinho, Patrícia Serra, Pedro Fernandes, and Silvestre Isidoro. We shared significant learning experiences and unforgettable moments together, and I am thankful for your contributions to my growth. I also want to thank my friends Hellen Cristina, Leonardo Lima, José Jacinto, and Vitória Maria for their support and friendship over the years. Last but not least, my brothers, Samuel Rufino and Gabriel Rufino, deserve my profound gratitude. Samuel, you have always been by my side, encouraging and supporting me at every step. Gabriel, your profound reflections have helped me grow as a person, and I am grateful for that. To all of you, my gratitude is eternal.

Declaro ter desenvolvido e elaborado o presente trabalho em consonância com o Código de Conduta e de Boas Práticas da Universidade de Lisboa. Mais concretamente, afirmo não ter incorrido em qualquer das variedades de fraude académica, que aqui declaro conhecer, e que atendi à exigida referenciação de frases, extratos, imagens e outras formas de trabalho intelectual, assumindo na íntegra as responsabilidades da autoria

Abstract

Recurrent research in the discovery of new drugs has predominantly focused on a limited subset of the human proteome, accounting for just 15%. This emphasis arises from researchers' focus on a small set of genes, to evade unexplored research paths. Consequently, a significant gap exists in the validation process for new pharmaceutical compounds, with an estimated 85% of disease-associated targets remaining underexplored. To address this critical issue, this thesis undertakes a series of computational studies with a primary objective of detecting and analyzing protein binding sites to facilitate the discovery of novel drug targets. In pursuit of these goals, this project categorizes proteins into three distinct groups, drawing from the work by Oprea and collaborators, as well as data from the UniProt, ChEMBL, and Protein Data Bank (PDB) databases. The first group contains targets that have either approved drugs or those currently undergoing clinical trials, classifying them as Well-Known Targets (WKT). The second group contains targets for which no active compounds have yet been identified, designated as Yet NOT druggable targets (YNOTs). Finally, the last group contains targets posing modeling challenges, which we classify as Difficult to Obtain Pharmacological Effect (DOPE). The analytical approach involves performing pocket detection for each PDB structure associated with protein targets. To achieve this, we employed the CAVIAR software, enabling the identification of cavities within each protein, including their binding sites. Subsequently, we conducted a search for similarities among these pockets. The most promising targets identified in the YNOTs group were the crystal structure of isobutyryl-CoA dehydrogenase complexed with substrate and the crystal structure of Human UDP-glucuronic acid decarboxylase. For the DOPE group, the most promising target discovered was the phosphodiesterase 6 delta subunit (PDE6δ). Furthermore, we compiled a database featuring 178 targets that also exhibit promise as potential pharmacological targets.

Keywords: Binding site; Pocket Detection; computational methods

RESUMO

A descoberta de novos medicamentos tem-se centrado predominantemente num subconjunto limitado de alvos do proteoma humano, representando apenas 15% do total de proteínas. Isto acontece porque os cientistas apenas se focam num pequeno conjunto de proteínas, de forma a evitar alvos e vias de sinalização para as quais ainda pouco se conhece. Consequentemente, existem profundas lacunas de identificação e validação de novos alvos farmacêuticos, o que leva a que cerca de 85% dos alvos associados a doenças permaneçam inexplorados. De forma a contribuir para melhorar o conhecimento nesta área, esta tese realiza uma série de estudos computacionais com o principal objetivo de identificar e analisar sítios de ligação de proteínas de forma a facilitar a descoberta de novos alvos cuja função possa ser modulada. Para atingir estes objetivos, este projeto categoriza as proteínas em três grupos distintos, com base no trabalho de Oprea e colaboradores, bem como em dados dos bancos de dados UniProt, ChEMBL e Protein Data Bank (PDB). O primeiro grupo contém proteínas para as quais existem medicamentos já aprovados ou que se encontram em ensaios clínicos, classificando-os como "Alvos Bem Conhecidos" (WKT). O segundo grupo contém proteínas para as quais ainda não foram identificados compostos ativos, designados como "Alvos Yet NOT druggable" (YNOTs). Por fim, o último grupo contém alvos que apresentam desafios para modelação, que classificamos como de "Difícil Obtenção de Efeito Farmacológico" (DOPE). A abordagem analítica utilizada envolve a identificação de cavidades para cada um dos PDB's associados aos três grupos. Para isso, utilizámos o software CAVIAR, que possibilita a identificação de cavidades nas proteínas, incluindo os seus sítios de ligação. Posteriormente, realizámos uma procura por semelhança de cavidades com o grupo WKT. Os alvos mais promissores encontrados no grupo dos YNOTs foram, a estrutura cristalina da isobutiril-CoA desidrogenase complexada com o seu substrato e a estrutura cristalina da decarboxilase do ácido UDP-glucurónico humano. Para o grupo DOPE, o alvo mais promissor encontrado foi a subunidade delta da fosfodiesterase 6 (PDE6δ). Além disso, obtivemos um conjunto de 178 proteínas que consideramos como promissores alvos farmacológicos.

Palavra-chave: Sítio de ligação; Deteção de pocket; métodos computacionais; similaridade de cavidade

INDEX

1 Chapter 1 Introduction: From Human Proteome to Binding Site. 11
1.1 The Human Proteome in Drug Discovery 11
1.2 *In-silico* Study in Drug Discovery 12
1.3 Computational Binding Sites: Insights and Analysis 13
1.3.1 The main strategies for the studies of binding sites 16
1.4 Pocket Detection 17
1.5 Pocket Similarity 20
1.6 The objective of this thesis 21
2 Chapter 2 Behind the Data, software and large-scale analysis 23
2.1 Datasets, algorithms. 23
2.1.1 UniProt database 23
2.1.2 ChEMBL database 24
2.1.3 Protein Data Bank (PDB) 24
2.1.4 PDBbind database 25
2.2 Softwares 25
2.2.1 Automated Ligand Identification 26
2.2.2 A Comprehensive Comparative Study of CAVIAR vs. Multiple Sofware Alternatives in pocket detection 26
2.2.3 Molecular Operating Environment (MOE) 27
2.2.4 PocketMatch 27
2.2.5 PyMOL 29
3 Chapter 3 Methodology 31
3.2 Division of Analyse groups 31
3.2.3 Criteria of Division of YNOTs groups 31
3.2.4 Criteria of Division of Well-Knowns groups 33
3.2.5 Criteria of Division of DOPE groups 35
3.2.6 Determining the similarity between targets. 36
4 Results and Discussion 39
4.2 YNOTs groups 39
4.3 Well-Known Groups 43
4.4 DOPE Groups 46
4.5 Analysis of Similarities Between YNOTs and WKT 46
4.6 Analysis of Similarities Between DOPE and WKT 48
5 Conclusion 51

Figure Index

Figure 1: Fisher projection model *projection model* 13
Figure 2: Relationship between the pocket druggability of proteins 14
Figure 3: p53 strucutre and compounds 16
Figure 4 : Diagram divided by method analysis and software related to pocket protein detection studies. 18
Figure 5: Representative formula for determining Caviar's score 26
Figure 6: Relation between the volumes in Å³ detected by the software POCASA, CAVIAR, DOGSITE, MOE, and Cb-DOCK 27
Figure 7: formula for PocketMatch PM score function 28
Figure 8 : YNOTs characterization workflow 33
Figure 9: WKT characterization process 35
Figure 10: DOPE characterization process 36
Figure 11: Representation of number of compound per targets of Ynots Goup 39
Figure 12: Graphic representation of the pockets of the Ynots groups 42
Figure 13: Representation of number of compounds per targets of WKT Groups 43
Figure 14: Graphic representation of the pockets of the WKT groups 45
Figure 15: Comparison between the physical chemical properties of pockets, WKT and YNOTs 47
Figure 16 : Analyses of YNOTs group after the structures underwent evaluation with PocketMatch 47
Figure 17: Comparison between the physical chemical properties of pockets, WKT and DOPE 49
Figure 18: Analyses of YNOTs group after the structures underwent evaluation with PocketMatch 49
Figure 19 : Pocket alignment between proteins 5ml3 and 5lm4 50
Figure 20: Sequence alignment of total proteins between 5ML3 and 5LM4 50
Figure 21: Sequence alignment of proteins pocket between 5ML3 and 5LM4 51

Table Index

Table 1: Process of exclusion of proteins from the YNTs groups 40
Table 2 : Relationship of the chains of the YNOts groups 41
Table 3 CAVIAR Cavity detection results for the structure PDB 1buh 41
Table 4: PDBS relationships and Uniprots codes between the WKT groups and the Ynots 44

Acronyms and Abbreviations

ChEMBL	Database of bioactive molecules with drug-like
DOPE	Difficult to Obtain Pharmacological Effect
EAAT1	Excitatory amino acid transporter 1
PDB	Protein Data Bank
UniProt	Universal Protein Knowledgebase
WKT	Well-Known Targets
YNOTs	Yet not a druggable targets
IC_{50}	Half maximal inhibitory concentration
InChI	International Chemical Identifier
InChIKeys	Fixed-length format directly derived from InChI
IUPAC	International Union of Pure and Applied
K_i	Constant of binding affinity of a ligand to a biomolecule
LigExtract	Automated Ligand Identification and Extraction from PDB Structure
MOE	Molecular Operating Environment
NCBI	National Center for Biotechnology Information
NMR	Nuclear Magnetic Resonance
PDBbind	Collection of binding affinities for the protein-ligand complexes
SMILES	Simplified Molecular-Input Line-Entry System
UniMes	UniProt Metagenomic and Environmental Sequences
UniParc	UniProt Archive
UniProtKB	UniProt Knowledgebase
UniRef	UniProt Reference Clusters

Chapter 1

Introduction: From the Human Proteome to Binding Sites

1 Chapter 1 Introduction: From Human Proteome to Binding Site.

1.1 The Human Proteome in Drug Discovery.

The exploration of new pharmacological targets is currently focused on a limited portion of the extensive human proteome. As noted by Tudor I. Oprea (2018), the pursuit of innovative therapeutic agents predominantly centers around a minor subset of the human proteome, approximately 15%(Oprea et al., 2018). This tendency often arises because researchers focus their efforts on a small set of genes, thereby sidestepping unexplored research paths. This leaves a gap concerning the validation process of new drugs. It is estimated that a significant 85% of disease-associated targets remain underexplored and have yet to be adequately investigated (Oprea et al., 2018).

In this way, organizations that fund the discovery of novel medicines tend to search for more substantiated and meticulously researched data (Morales-Navarro et al., 2019). This approach has both positive and negative aspects. On the positive side, it promotes precision in research focus. Conversely, it generates a potential downside akin to a “snowball effect”, wherein researchers predominantly focus their research projects towards areas of study already endowed with a large amount of data and analyses. Consequently, a multitude of unexplored opportunities for investigating new therapeutic targets remain underemphasized.

With the continued evolution of computational techniques, the pursuit of these novel targets is becoming increasingly attainable. However, there is a lackin established methodologies and strategies, largely fueled by the issue outlined in the preceding paragraph. This underscores the imperative need for the development of innovative approaches and the integration of techniques to surmount current challenges and propel the field of target discovery.

Certainly, in light of the rapid progress in algorithms rooted in structural biology, the development of diverse in-silico screening methodologies has gained momentum. These methodologies enable large-scale analyses of the human proteome, leveraging experimental data to enable the parametrization of various characteristics associated with therapeutic targets and its intricate interactions of small molecules within their respective binding sites.

1.2 *In-silico* Study in Drug Discovery

The concept of “in silico” originates from an expression used in the realm of computer simulations and associated domains to occurrences that transpire "in or through a computer simulation". This nomenclature was coined from the Latin phrases “in vivo” and “in vitro”. Computational modeling has witnessed significant impetus, thanks to funding initiatives from both private and government agencies, which have facilitated the establishment of numerous laboratories dedicated to in silico analysis. Such analyses have enabled great advances in the discovery of novel drugs.

Computational methods have continuously evolved and advanced, enabling the analysis of large databases for the identification of potential therapeutic targets and molecules for treating diseases (Jean-Quartier et al., 2018). The discovery of novel targets stands as the main mechanisms for current pharmacological research, alongside the validation of these methods, which in turn facilitates the screening of new compounds with respect to their binding site interactions.

Computational analyses offer several advantages in contrast to experimental investigations, such cost-effectiveness, speed, scalability, and flexibility. Nevertheless, they also have some disadvantages, including complexity in validation and inherent limitations of available data. In silico methods constitute valuable tools for addressing diseases, but their application should be used in a critically conscious way, serving as a complementary approach rather than a replacement for experimental studies. (Jean-Quartier et al., 2018).

For instance, these computational methods can be used to elucidate the underlying mechanisms driving disease progression and to evaluate potential therapeutic targets across different treatment modalities, among other possible analyses that can be carried out.

Many of these prediction methods are based on similarity principles, encompassing both target similarity and small molecule similarity. Additionally, there are methods based on molecular structures that exhibit analogous interactions with a given target (Jean-Quartier et al., 2018). These two fields are interrelated because most of these studies contribute to an enhanced understanding of the binding sites of diverse targets (Jean-Quartier et al., 2018)

1.3 Computational Binding Sites: Insights and Analysis

As far back as 1894, Fischer introduced one of the first models elucidating protein-ligand binding (Cramer, 1995). He employed the analogy of a lock for a rigid protein binding pocket and a key for a specific ligand, illustrating the interaction between a protein and its ligands (Cramer, 1995). Protein binding pockets are instrumental to understanding the role and function of proteins, as well as the type of molecules with which they bind or are bound to. Characteristics of binding regions, including their shape, size, physical properties, binding modes, and the mechanisms of action of therapeutic agents on disease-related proteins, are of paramount importance. The diversity of target types arises from the various binding sites, leading to different mechanisms for targeting them with small molecules. Amino acids assume a critical role in shaping binding sites, facilitating interactions with specific substrates or ligands, inducing changes in their chemical structures (Flower, 2006).

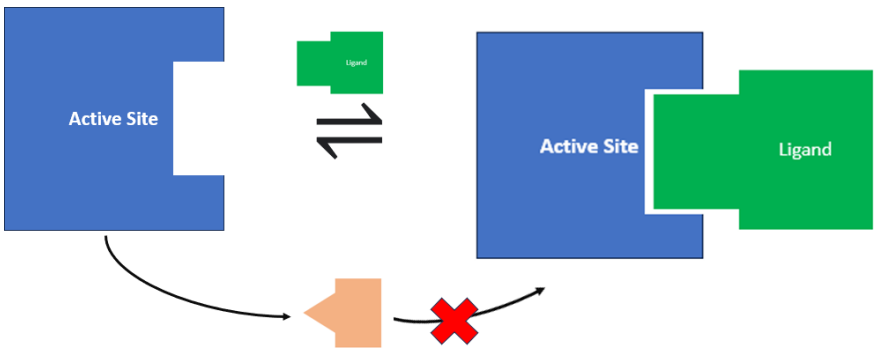


Figure 1: Fisher projection model *projection model*

The protein cavities exhibit a broad range of sizes, ranging from the compact cavities found in enzymatic interactions to the large ones associated with antibody binding. Consequently, computational studies into protein binding sites serve as valuable tools for their identification and characterization through modeling and simulation techniques. These methodologies empower researchers to predict the three-dimensional (3D) structure of proteins, thereby enhancing comprehension of their interactions with other molecules, the development of therapeutic interventions, and the design of pharmaceutical agents (Kahraman and Thornton, 2008).

Protein cavities exhibit multiple variables, including surface area, spatial location, configurations of alpha-beta carbons, and the presence of heavier atoms, among other distinctive characteristics.

Two concepts of binding sites exist: the first is knowledge-based, wherein if the location of a binding site in one protein is known, it can be transposed or inferred onto another protein. The second approach involves a priori prediction of a binding site based on either sequence or structural information. In contrast, the fold of a protein dictates the more general aspects of the binding site, including its shape and size.

At the genomic level or within a population of genomes, the nature of the protein-binding site is not predetermined. Consequently, identifying and analyzing these binding sites remains a fundamental challenge, engaging both in vitro and in silico approaches. (Flower, 2006).

The study of binding sites has witnessed significant advancements. These developments encompass various computational techniques such as binding site prediction, molecular docking, molecular dynamics, among other methods, including the recent progress in machine learning algorithms, These resources offer expanded capabilities for the analysis of diverse types of targets, representing a substantial enhancement compared with traditional methodologies (Jean-Quartier et al., 2018).

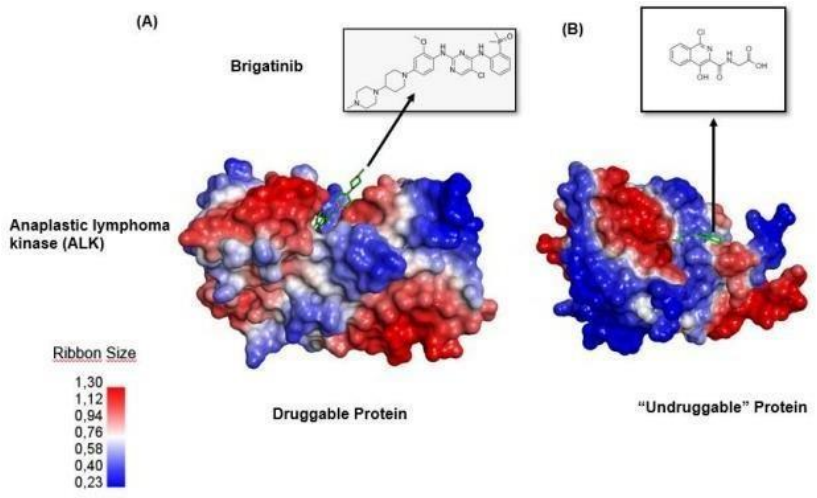


Figure 2: Relationship between the pocket druggability of proteins

The Figure 2 illustrates a comparison between two proteins, where (a) presents an instance of a druggable protein, Anaplastic Lymphoma Kinase (ALK), targeted by a new second-generation ALK Tyrosine Kinase Inhibitor (TKI) called Brigatinib, which demonstrates preclinical activity against a wide range of ALK mutations. In contrast, (b)

serves as an example of a "difficult to drug" protein, CDK9, which functions as the kinase for the positive transcription elongation factor b and facilitates the transition of paused RNA polymerase II to processive transcription elongation.

The concept of a binding site is intrinsically linked to protein druggability, a term associated with the protein's ability to bind small molecules with high affinity. Consequently, some proteins may possess more easier "druggable" binding sites than others. An “undruggable” protein lacks specific characteristics such as cavity size, conformation, that readily permit drug binding, although this does not imply their inability to interact with small molecules. Rather, it implies that identifying specific drugs for these protein structures may be more challenging. The term “undruggable” has lost favor in recent times because many targets once deemed undruggable have since become promising candidates for drug development. This transformation is primarily attributed to advancements in technology rather than inherent properties of the proteins themselves.

In recent years, advancements in drug discovery techniques, such as the development of stapled peptides and small molecules that can reactivate mutant p53, have made it possible to target p53 effectively (Figure 3). These innovative approaches have transformed p53 from an undruggable protein to a promising druggable target for cancer therapy, and several compounds are currently in clinical development for this purpose.

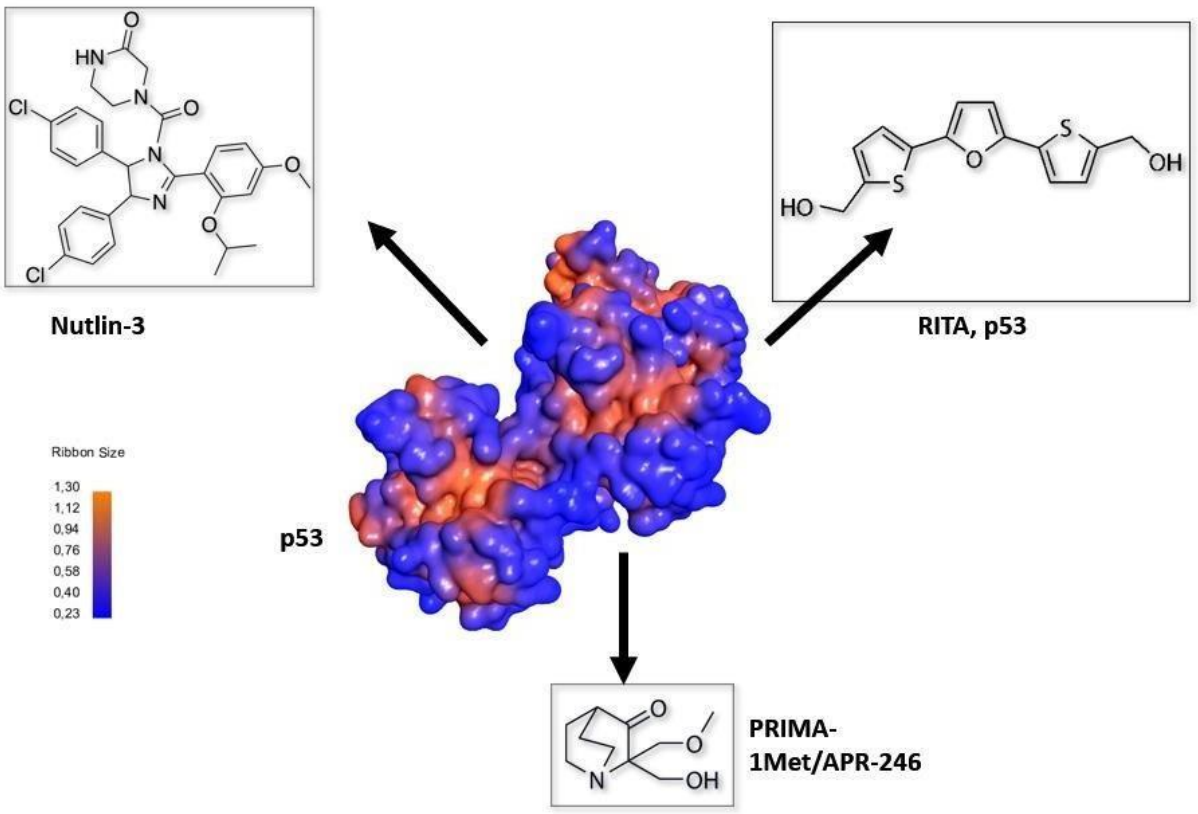


Figure 3: p53 strucutre and compounds

1.3.1 The main strategies for the studies of binding sites

Numerous in silico strategies are available for the exploration of protein's binding sites. These approaches can be tailored to specific protein families, individual target, or large-scale analyses. The selection of analytical techniques is typically dependent on the research focus (Marchand et al., 2021).

For instance, molecular docking serves as a valuable tool for probing interactions involving multiple ligands. This method provides a detailed understanding of the amino acid residues within the binding site, critical for modulating the activation or inhibition of a particular protein.

Large-scale analyses, on the other hand, are less constrained in scope but offer a broader parameterization of all available structural data. This broad approach allowsthe classification and discovery of various targets and mechanisms that may not be accessible through individual protein examinations. An illustrative example is the study conducted by Shen et al. in 2017, which entailed a proteome-scale investigation of protein allosteric

regulation perturbed by somatic mutations across 7,000 cancer genomes. In summary, they conclude that some findings clarify the role of allosteric binding site regulation during tumorigenesis and provide a useful tool for the timely development of targeted cancer therapies (Shen et al. in 2017).

1.4 Pocket Detection

Structural binding site analysis typically commences with data derived from the Protein Data Bank (PDB) (Burley et al., 2017). which contains an extensive repository of experimental information pertaining to enzymes, proteins, and other types of macromolecules. In the year 2020, the PDB recorded 13,998 macromolecules and a total of 172,880 registered entities. These structures provide invaluable insights for drug designers.

Understanding that the folded protein cavities are fundamental in functions, such as signal transduction and enzymatic activity, several software tools have been developed to detect, compare, and discover new binding site cavities. (Marchand et al., 2021).

Pocket detection software serves the vital role of assessing the nature and significance of the cavities found in a protein. These software tools are instrumental in pinpointing regions within the protein that may hold crucial importance for its biological function, such as potential binding sites for drugs, enzymes and antibodies. Some of the primary tools for pocket detection include CAVIAR, LIGSITE, PocketPicker, DoGSite, FPocket, among others software applications.

To gauge the features of these pocket cavities, it is fundamental to use a scoring system. This scoring system assigns numerical values to each cavity identified by the program, and it is computed based on various cavity characteristics, including size, shape, depth, location within the protein and other pertinent parameters (Marchand et al., 2021). Some programs also take additional factors into account, such as the presence of functional groups of amino acids in proximity to the cavity or the distance to the protein's surface.

It is noteworthy that these scores are used for ranking the detected cavities generated by the programs, thereby enabling the selection of those with a higher likelihood of playing a significant role in the protein's function. Lower cavity scores are typically regarded as more promising and merit closer investigation, though it is important to exercise caution in interpretation because failures such as inaccurate detection may occur, requiring different types of analyzes to determine the results.

The scoring system serves as a tool designed to aid in the identification of potentially significant protein cavities. Thus, it should not be considered as an absolute measure of a cavity's importance. To corroborate the significance of a cavity for a protein, it is advisable to employ additional analytical methods, as suggested by Marchand et al.(2021).

Methods for predicting cavities using a structure-based approach can be categorized into three main groups: Geometry-based, energy-based and Evolutionary base algorithms. Figure 4 show how these methods are organized.

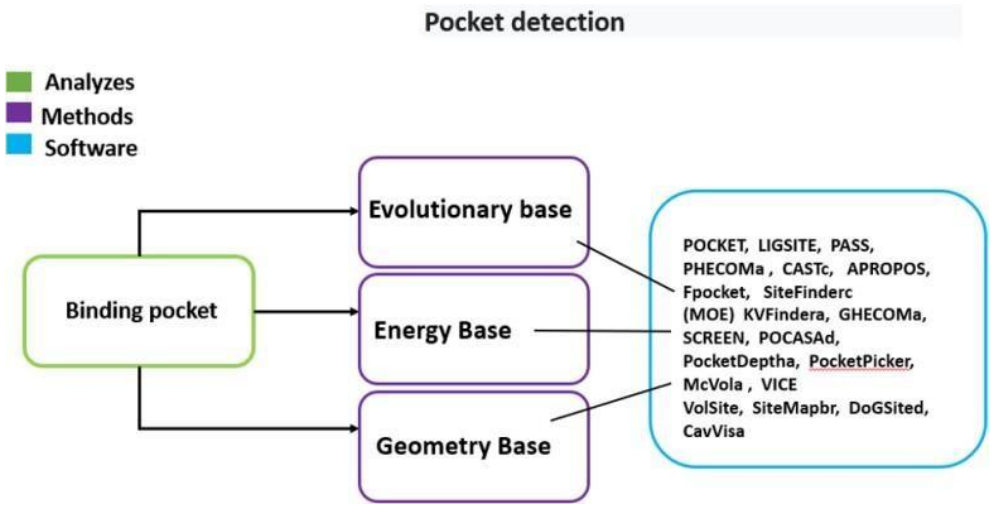


Figure 4 : Diagram divided by method analysis and software related to pocket protein detection studies**.**

Geometry-based algorithms require fewer computational resources and demonstrate greater stability with respect to side-chain position. They offer scalability and automation, making them suitable for large scale applications. These algorithms identify cavities primarily based on their geometric shapes, which can be further enhanced with additional features such as scaffolds (Marchand et al., 2021).

Geometry Based: use the calculation ofthe distance between points, the calculation of the intersection angle between two lines, and the calculation of the area of a pocket to detect pockets from its geometry.

These methods utilize principles from molecular mechanics or molecular dynamics to estimate the binding affinity or stability of a ligand (small molecule) within a specific region of the protein.

Energy-based methods typically involve the calculation of various types of energy terms, such as van der Waals forces, electrostatic interactions, hydrogen bonding, and solvation energy, among others. By evaluating these energy components, the methods aim to predict whether a particular region of the protein can accommodate and interact favorably with a ligand, indicating the likelihood of it being a functional binding site.
These approaches are valuable for identifying potential ligand-binding pockets, as they take into account the energetics of molecular interactions, which play a crucial role in determining the feasibility and strength of ligand-protein binding.

In addition to the previously mentioned algorithms, there are more recent approaches in protein detection. One such approach is the Evolutionary Based, which uses evolutionary algorithms to detect pockets. These algorithms are based on genetic algorithms and are designed to identify the characteristics of a pocket. Another noteworthy approach is the Combination Approach, which combines multiple algorithms and methods to enhance pocket detection accuracy while mitigating false positives (Marchand et al., 2021). Several calculation models can be used for these predictions. For instance, in the Combination Approach, a form of multiple regression can be used to combine various algorithms and methods for pocket detection (Yu et al., 2009). Alternatively, linear regression methods can be employed to estimate the contribution of each feature in the detection of pockets (Yu et al., 2009).

Moreover, other techniques such as discriminant analysis, logistic regression, decision trees, and neural networks can also be applied. In addition, we can use edge detection techniques to detect surface edges and contours (Liu et al., 2020).

1.5 Pocket Similarity

As previously mentioned, the function of a protein can be elucidated by analyzing its 3D structure, specifically by inferring and querying its active sites through comparisons based on pocket similarities. Alignment and comparison of protein and ligand structures have become fundamental in drug design, as they encompass the intricate 3D configurations of their surfaces, which play a pivotal role in various aspects of molecular recognition. Over time, comparison methods have evolved to offer greater precision and ease of selection. This advancement in comparison techniques has brought significant and pertinent implications for drug design (Eguida and Rognan, 2022).

Pocket Similarity uses various comparison systems, and the selection of a specific system depends on the objectives and reference criteria defined for each particular model. Protein cavities exhibit multiple variables, including surface area, spatial location, configurations of alpha-beta carbons, and the presence of heavier atoms, among other distinctive characteristics. The choice of a particular model is based upon the intrinsic characteristics established within each model for example, alignment model, structural model, among others (Eguida and Rognan, 2022).

Comparing specific protein pockets presents a significant challenge involving the delicate balance between minimizing inaccuracies, such as false positive results, and preserving potentially similar data(Eguida and Rognan, 2022). The utilization of subpocket comparisons offers a more viable approach as it addresses situations characterized by high conformational variability within an entire cavity. It is crucial to emphasize that the decision to analyze an entire cavity or not depends on the specific objectives of the ongoing studies. In general, pocket similarity software falls into four categories: graph matching, geometry matching, fingerprint histogram, and machine learning (Eguida and Rognan, 2022). The Graph Matching model facilitates comparison between two or more sets of data by applying nonlinear patterns to the relationships between variables. It involves mapping the information within the data and the selected structural object. As a result, a pair of two elements can be linked, with the first set represented as node and the second as edges in the graph-based analysis(Eguida and Rognan, 2022).

In summary, these various models, each employing different software for protein pocket comparisons, provide estimates of pockets similarities with a certain precision. Nevertheless, it is crucial to emphasize that the human protein complex is established by different criteria, so for each specific disease there are specific variations of similarity. For example, PocketMatch uses comparison between a pair of sites of an average size of 50 atoms. The algorithm has been used for large scale database searches and all-vs-all comparisons. others software like SiteHopper uses Surface atoms by Gaussian shapes matching to compare the similarities.

1.6 The objective of this thesis

This work seeks to elucidate several applications within the field of protein binding sites. As we exploredin this topic, in-silico analyses involve an expansive field that has the potential to lead to the discovery of novel molecules. Thus, we aim to demonstrate how detection methods and the assessment of protein pockets similarities can facilitate innovative discoveries. To address this critical issue, this thesis undertakes a series of computational studies with a primary objective of detecting and analyzing protein binding sites to facilitate the discovery of novel drug targets.

Chapter 2

Behind the Data, Software and Large-Scale Analysis.

2 Chapter 2 Behind the Data, software and large-scale analysis.

2.1 Datasets, algorithms.

This chapter will focus on presenting the various Datasets and algorithms that were used in the preparation of this work, while also providing the primary justifications for employing both the software and algorithms, as well as detailing the databases involved. It is important to note that during this work, a diverse array of databases, Python scripts, and software were employed, all aimed at challenging classifications and analyzing various types of protein.

2.1.1 UniProt database

The Universal Protein Database, commonly referred to as UniProt, is an accessible and free database that compiles diverse information related to protein sequences and their functions. Many of its entries stem from genome sequencing projects and encompass a wealth of data regarding various biological functions derived from the scientific literature. UniProt's primary collaborations are associated with Swiss-Prot, TrEMBL (both components of UniProtKB), UniParc, UniRef, and UniMes (Bateman et al., 2017). This database encompass more than 227 million protein sequences, all of which have been carefully selected and reviewed by multiple experts. The UniProt databases empower the research community to explore the life’s diversity by investigating the proteins expressed by each organism. The UniRef databases cluster sequence sets at different levels of sequence identity, while the UniProt Archive provides a comprehensive collection of known unique sequences, including historical obsolete ones. The uncurated sections of UniProt are designated as UniProtKB/TrEMBL. This sector exhibits the most rapid growth in terms of data. Although the data in this section is not manually reviewed, UniProt automatically supplements annotations to organize and mimic them based on experimental data. With the inclusion of this information and these resources, UniProt offers mechanistic insights into how variations, as well as different sequences and mutations, can contribute to the development of various types of diseases. Therefore, in this thesis, various parameters will be explored, with a primary focus on UniProtKB data as a source for analyzing novel pharmacological targets.

2.1.2 ChEMBL database

ChEMBL is an openly accessible database that offers extensive bioactivity data. Within ChEMBL, multiple sources provide access to a wealth of data related to small molecules at various stages of research. The data within ChEMBL undergoes a comprehensive manual review process and is sourced from substances or compounds subjected to bioactivity testing. Thus, this data is extracted from submitted documents containing lists of compound records and associated assays categorized by their activity, complete with actual parameters, values, and units, all conforming to ChEBI and InChIKeys standards. When data is obtained from external resources, original identifiers are also preserved (e.g., SIDs and AIDs for PubChem substances and assays, HET codes for PDBe ligands). PubMed identifiers or Digital Object Identifiers (DOIs) are maintained for reference documents.

The compounds featured in the ChEMBL database are categorized based on their structures and their respective pharmacological targets. This approach enables ChEMBL to generate non-redundant data entries within a structured dictionary. The structural representation in this database is founded on IUPAC (InChI) standards, facilitating the classification of identical compounds under new identifiers. Protein targets are represented by primary accessions in the UniProt protein database, while organism targets receive NCBI taxonomy IDs and names.

2.1.3 Protein Data Bank (PDB)

The Protein Data Bank (PDB) is one of the largest databases of protein crystallographic structures available to date. As of 2023, the number of crystallographic structures for Homo sapiens has exceeded 60,000. The vast amount of data associated with these structures empowers a diverse range of computational analyses aimed at elucidating their functions within this database.

Within the PDB, three-dimensional structures are stored, offering insights into the spatial arrangements of atoms and the chemical bonds present in amino acids, small molecules, ions, peptides, and various other molecules types. In summary, the stored information results from analysis of X-ray crystallography of protein, nuclear magnetic resonance (NMR) techniques. These registered structures are identified by a code consisting of four alphanumeric characters, always commencing with a number.

The PDB not only furnishes a plethora of valuable data regarding the cataloged proteins but also interfaces with several other databases to augment data accessibility and facilitate in-depth analyses. One such database in PDBbind, which is intricately to the PDB and which will be discussed in section 2.4.1.

Furthermore, the parameters associated with the presented structures are of utmost importance for understanding and selection purposes. The resolution of PDB structures constitutes qualitative data that facilitates the differentiation of atoms within the protein structure. Lower resolution values correspond to greater precision in atom discrimination. According to the literature, structures with resolutions below 1 are deemed highly precise, while those with values exceeding 3 do not offer substantial precision regarding their structural details.

2.1.4 PDBbind database

PDBbind is a database that provides a large collection of affinity data concerning protein and ligand complexes. In accomplishes this by utilizing the PDB and conducting searches for ligand references, thereby elucidating the relationship between structure and ligand through activities. The primary analysis parameters employed are K_d, K_i and IC_{50}. The most recent release, version 2020, is derived from the contents of the PDB officially released during the first week in 2020. This release provides binding affinity data for a total of 23,496 biomolecular complexes within the PDB. These complexes include protein-ligand interactions (19,443), protein-protein interactions (2,852), protein-nucleic acid interactions (1,052), and nucleic acid-ligand complexes (149). In comparison to the previous release (v.2019), the binding data included in this version have grown by approximately 10%. All binding data are meticulously curated by our group from approximately 40,500 original references.

2.2 Software

2.2.1Automated Ligand Identification

LigExtract, a user-friendly tool developed by our research group, excels at precise localization and extraction of ligands from proteins. It employs a well-defined set of criteria and references to extract ligands from PDB structures by utilizing the UniProt code.

The advantages over other tools of its kind are that LigExtract allows for a more automated extraction, requiring only a few steps for the ligand extraction procedures. Being a great tool for large-scale analysis of PDBs structure.

2.2.2A Comprehensive Comparative Study of CAVIAR vs. Multiple Sofware Alternatives in pocket detection

CAVIAR is an adaptable open-source tool designed for the identification and visualization of protein cavities, with machine learning. CAVIAR allows the location of cavities based on grid points It provides both a command-line interface and a graphical user interface, serving as the primary programming language used in its development. Remarkably, CAVIAR does not require any specialized knowledge of ligands. It incorporates a robust sub-cavity segmentation algorithm that effectively dissects protein cavities into meaningful components (Marchand et al., 2021).

$$score_{cavity} = \frac{size * median * q}{100}$$

Figure 5: Representative formula for determining Caviar's score**.**

In this context, 'size' refers to the dimensionality of the cavity measured in grid points, 'median' denotes the central value of buriedness, and 'q' represents the 8th quantile of buriedness.

A work carried out by our group compared the performance of the Caviar software with other similar programs and demonstrated that Caviar significantly outperforms in terms of accuracy in ligand volume comparisons compared to other software of its kind, as was the case in the analysis of the structure 6GFA shown in Figure 6.

As we can see, the lines in the graphic represent the different ligand volumes of ATP ligand. The Vsurf_D1 represents the Interaction field in the surface area of the ligand, the Vol in yellow represents the hydrophobic volume and the vdw_vol represents the van der Waals volume (Å3) calculated using a connection table approximation. The bar represents the different volumes detection by the software.

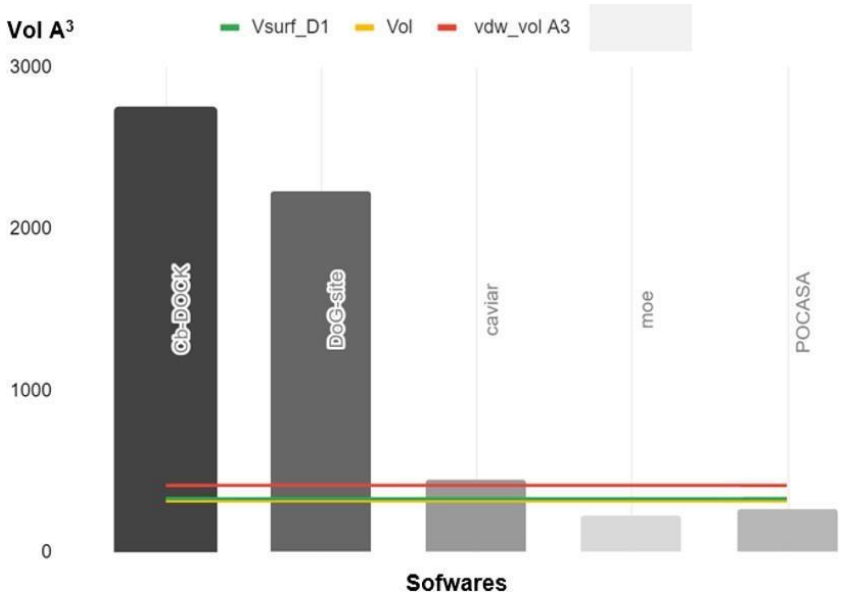


Figure 6: Relation between the volumes in Å³ detected by the software POCASA, CAVIAR, DOGSITE, MOE, and Cb-DOCK

One of the remarkable features of CAVIAR is its ability to rapidly and accurately localize protein cavities within specific regions of a protein structure, including individual chains. Additionally, CAVIAR permits cavity decomposition (Marchand et al., 2021).

2.2.3 Molecular Operating Environment (MOE).

In the fields of computational chemistry and molecular modeling, scientists and researchers frequently utilize the software platform known as the Molecular Operating Environment (MOE) (Chemical Computing Group ULC, 2022) This platform offers a diverse array of functions and tools for the investigation and analysis of molecular structures, prediction of molecular properties, and the execution of various simulations and computations relevant tothe fields of chemistry and biology.

MOE has been developed to assist researchers in various tasks including, encompassing protein modeling, ligand-receptor interactions, drug development, and chemical informatics. Users can engage in molecular docking studies to ascertain how chemicals bind to proteins, examine and modify molecular structures. Serving as a comprehensive tool for researchers in the biological sciences and pharmaceutical sectors, MOE is routinely employed for a wide range of scientific endeavors.

2.2.4 PocketMatch

PocketMatch is a valuable tool employed or the comparative analysis of binding sites within protein structures, enabling the establishment of connections and commonalities between these protein molecules. PocketMatch excels in the rapid comparison of multiple binding sites, thanks to its algorithm's scalability for large-scale binding site comparisons. This tool finds extensive utility in the examination of both the structural and functional attributes of proteins.

In PocketMatch, each binding site is meticulously represented by 90 sorted distance lists, providing an accurate reflect the site's geometric and chemical properties. Through an incremental alignment procedure, these ordered matrices are aligned and scored, resulting in the derivation of PM (PocketMatch) scores for site pairings (Yeturu and Chandra, 2008). Researchers working with proteins can significantly benefit from the PocketMatch tool, as it enables them to swiftly analyze multiple association sites and identify similarities among the binding sites of different proteins.

$$PMScore = \frac{\sum_{i=1}^{90} Count_i}{maximum(|S_1|, |S_2|)}$$

Figure 7: Formula for PocketMatch PM score function

The alignment score between two sites is determined by calculating the net average of matching distance elements within 90 lists, expressed as a fraction of the total number of distance elements in the larger set, using a chosen threshold τ. This similarity measure, referred to as PMScore, serves as the default scoring scheme in the study. Additionally, a variant known as PMScore min was investigated. In this variant, the denominator is determined by the minimum between the cardinalities of the two sets (|S1|, |S2|). This modification places emphasis on local structural similarity while disregarding the relative size mismatch between the compared sites.

2.2.5 PyMOL

For scientists and researchers working in the domains of computational chemistry, structural biology, and biochemistry, PyMOL stands as a crucial piece of software. This program offers a robust platform for the visualization and analysis of three-dimensional structures of molecules, including proteins, nucleic acids, ligands, and a diverse array of chemical compounds.

The widespread adoption of PyMOL can be largely attributed to its ability to deliver precise and in-depth three-dimensional representations of molecules, enabling scientists to intuitively explore their shapes, conformations, and interactions. Additionally, PyMOL excels at producing high-quality images and animations, making it an ideal choice for presentations and scientific publications. Among its notable attributes are analysis techniques that facilitate the understanding of geometry, chemical bonding, and surface properties.

Chapter 3

Methodology

3 Chapter 3 Methodology.

3.1 Division of Analyse groups

Following a detailed analysis and discussion of research challenges in earlier sections, we identified three distinct target groups:

1. DOPE targets: These are targets that are notoriously challenging to modulate. Despite evaluating several compounds against them, none have exhibited target-associated activity.
2. YNOTs targets: The second and most significant group for this work, encompasses proteins that remain untested against any molecules and lack any known active compounds.
3. Well-Known targets (WKT): Serving as a benchmark or reference, the WKTs are characterized by their association with both approved drugs and molecules that demonstrate significant biological activities relevant to the target.

3.1.1 Criteria of Division of YNOTs groups

The The YNOTs group was curated through a series of methodological steps, further elaborated in this section. The initial step involved querying the UniProt database for proteins from the entire human proteome, using the reference taxonomy "Homo sapiens (Human/Man) [9606]. "

With data related to all human UniProt entries, we assessed the bioactivities of these protein targets. Our primary focus was on targets devoid of bioactivity information associated with small molecules. We utilized the ChEMBL database (version 31) to filter targets that had no active compounds and had been tested with fewer than 10 compounds in total. To facilitate this filtration process, a Python script was developed through collaborative effort within our research group. This script was specifically designed to meet the needs of YNOTs filtration, along with analyses of other groups.

Following the target selection procedure outlined above, we employed another in-house software tool developed by our research group, named LigExtract. This tool was instrumental in identifying PDB structures associated with the UniProt codes of the YNOTs, ensuring they had a resolution finer than 3 angstroms. Additionally, LigExtract

also facilitated the separation of individual chains within each crystallographic structure and extracting all ligands.

To ensure the YNOTs did not include any WKT, we utilized the PDBbind database. A Python script was developed to query the bioactivity associated with the targets of each structure, aiming to eliminate proteins that exhibited activities beyond a threshold of 5 angstroms.

Additionally, we instituted a selection criterion based on amino acid length. Structures with fewer than 100 amino acids were omitted since they typically lack sufficiently sized binding site cavities. UniProt codes that did not meet all the criteria set forth were likewise removed.

Upon initiating the YNOTS' PDB structure selection, we undertook a structure preparation step. This pivotal step ensured uniformity across all structures, organizing them sequentially for comparison against other analysis groups. Initially, the crystallographic structures underwent a cleansing process, which entailed the removal of ligands and non-standard residues that did not conform to essential binding site analysis standards.

Next, these streamlined structures underwent protonation. Thus, a Python script, tailored with PyMOL functionalities, was crafted to protonate the YNOTs' PDB structures. This process added specific hydrogen atoms to amino acid residues that were suitable for accepting hydrogen ions, aiming to mimic physiological conditions of pH. Protonation is crucial as it enables a higher precision of interactions, making it possible to compare protein pockets. It is important to note that the protonation performed was simpler and less robust due to the number of structures involved compared to methods that involve small-scale analysis.

Subsequently, an additional criterion of analysis was employed to refine our selection of individual PDB structures. For each unique UniProt code, the structure with the largest chain was chosen. This procedure streamlined the data set, optimizing calculation time (i.e., reduce the number of structures to minimize) and conserving computational resources, resulting in a single structure for each target.

Upon finalizing this selection, structure minimization was carried out using the MOE software, according to the following conditions. Post-minimization, the structures were submitted to the CAVIAR software to compute all potential pockets of structures. For the identified pockets, a detailed characterization process was carried out to determine their physical and chemical characteristics. The CAVIAR output underwent clustering analysis

to assess factors such as hydrophobicity, ligandability, score function, and the pocket count in each processed structure with the aim of obtaining a list of characteristics of the pockets and also to continue with the similarity of the pockets.

Next, we extracted cavities from the CAVIAR output which provides its results in three formats: PDB structures showcasing associated cavities, a ".txt" file detailing the results, and another PDB file with subcavities. Leveraging the ID data for each cavity, the coordinates of these cavities were selected and then the distance within 4 angstroms from the cavity coordinates was checked. These results were extracted from the PDB structures, thus obtaining pockets that ranked highest based on the prior script. Figure 8 shows the YNOTs characterization workflow.

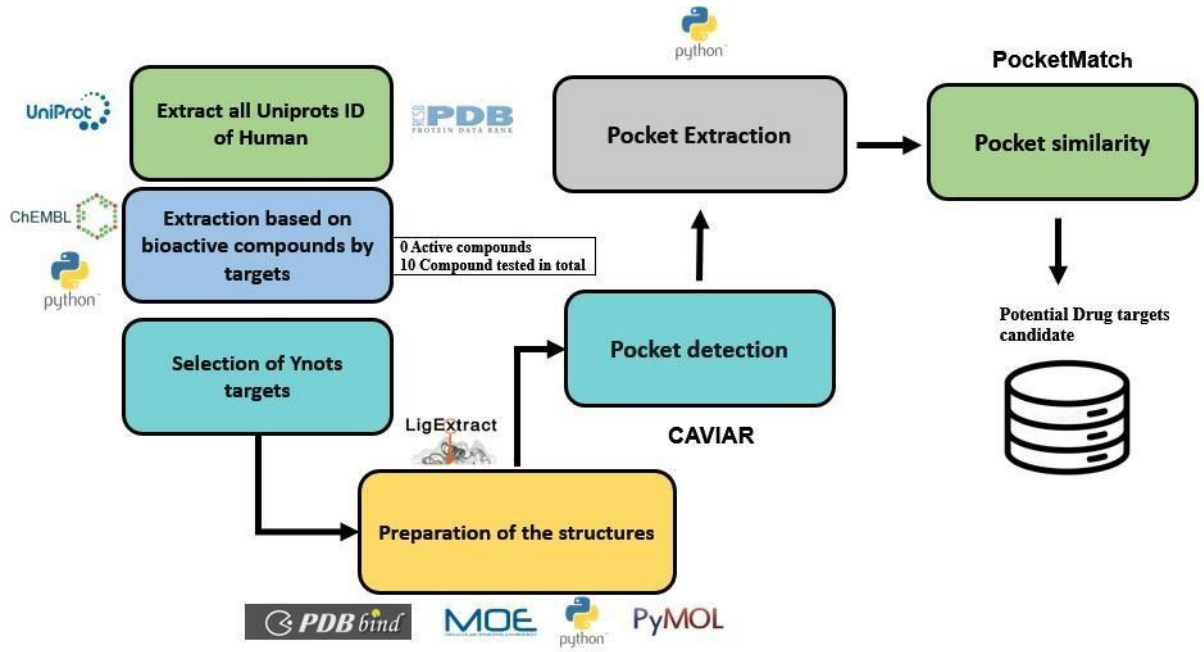


Figure 8 : YNOTs characterization workflow

3.1.2 Criteria of Division of Well-Knowns groups

The The identification of WKT groups followed a series of referenced analyses,ensuring that the obtained structures played a crucial role in treating specific diseases andwere indeed well-known structures. To achieve this, proteins referenced in the Oprea et al. publication within the Tchem category that have at least one ChEMBL compound with an activity cutoff of < 30 nM, and Tclin category targets that have at least one approved drug,

were considered. This resulted in a compilation list containing 840 UniProt codes associated with well-known proteins..

Subsequently, a search was conducted within ChEMBL (version 31) to identify targets possessing WKT. In the initial phase, the chemical information files were processed, and certain transformations standardized the chemical structures. This facilitated the computation of unique InChIKeys for each chemical structure linked to the UniProt codes. Activity values were normalized to a specific unit (nanomolar) following their original units, and filtered based on standard maximum allowable activity value relationships. Duplicative entries were then removed.

Canonical SMILES data were extracted using the following criterion: compounds with activity exceeding 10 nM were considered active. Thus, a list was compiled containing data extracted from Oprea et al., pairing activities associated with each target based on the aforementioned criteria. Subsequently, the search for crystallographic structures linked to the respective UniProt codes was carried out.

Regarding the PDB filtering process, the same criteria and procedures as those used for other groups were applied. Ligextract was employed to filter out PDB structures smaller than 3 angstroms, followed by the removal and extraction of all ligands associated with the chains. Subsequently, the relationship of each known ligand with its respective targets was determined. Next, a process to exclude proteins smaller than 100 angstroms was carried out, thereby excluding proteins that lacked sufficient binding sites.

After this separation process of Well-Knowns, the steps for structure preparation were initiated. The crystal structures were cleaned by removing non-conventional amino acid residues to avoid issues in the calculation. Following this, the structures underwent protonation using the scripts. Subsequently, the selection of unique structures associated with their respective targets was performed. This selection aimed to obtain structures with the longest chains for subsequent calculations. The process of structure minimization was then carried out using the MOE software.

Following this procedure, pocket detection was carried out using the CAVIAR software. Comprehensive analyses of the Well-Knowns' pockets were conducted, evaluating ligandability, hydrophobicity, and the scoring of each selected target.

With these results in hand, the validation process of the methods employed commenced. This involved a meticulous analysis of the Well-Knowns' components to verify if the selected structures genuinely belonged to recognized structures. These analyses included

identifying each chain within the structures and matching them to their respective UniProt codes, confirming the correlation of the structure with activity, as well as the ligand associated with the structure.

After this phase, the Well-Knowns were prepared for analysis in PocketMath, details of which will be elaborated upon in the concluding paragraphs of the methodology. Figure 9 shows the WKT characterization workflow.

.

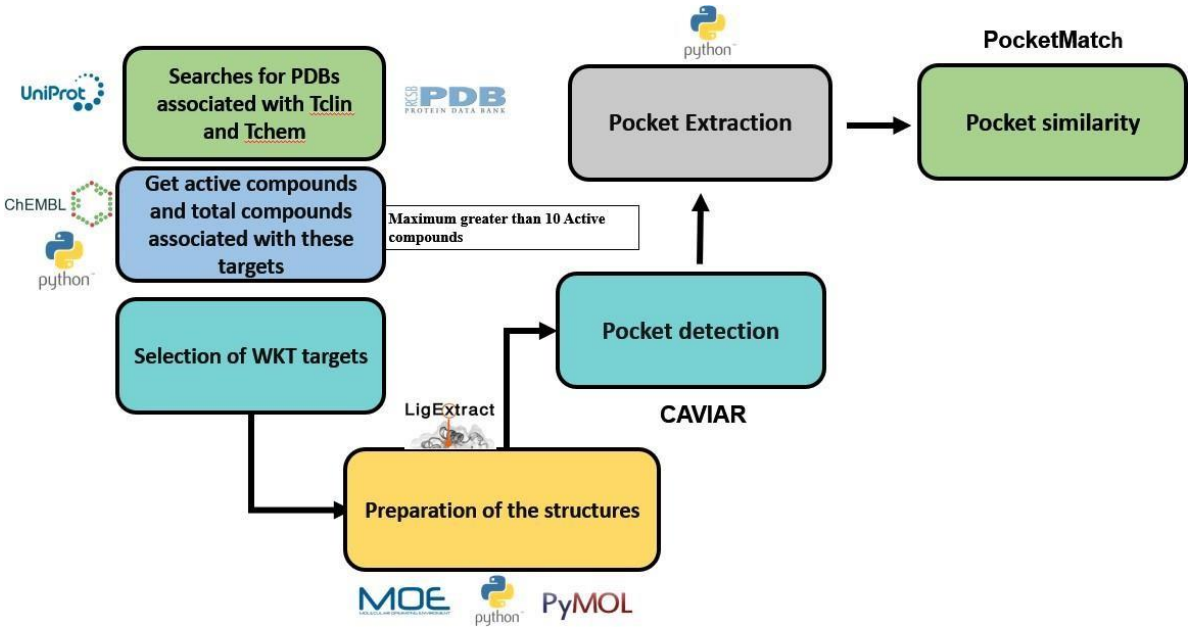


Figure 9: WKT characterization process

3.1.3 Criteria of Division of DOPE groups

As delineated earlier, the DOPE Group targets are those for which numerous small molecule compounds have been tested, but no active compounds have been identified. The selection criteria involved identifying targets with a plethora of tested compounds but with none or just one active compound. This selection involved gathering UniProt codes for the human proteome, using the taxonomic reference "Homo sapiens (Human/Man) [9606]."

Next, chemical data was extracted, followed by certain transformations like to standardize the chemical structures. This enabled the computation of unique InChIKeys for each chemical structure associated with the UniProt codes. Activity values were normalized to a specific unit (nanomolar) and adjusted according to their original units. The data was

then filtered based on standard relationships of maximum allowed activity values. Subsequently, duplicate entries were removed. Canonical SMILES data was extracted using the criteria that there should be more than 3 active compounds and more than 100 compounds tested in total.

Post this, the same steps as mentioned in the earlier sections were repeated. These steps included employing Ligextract to select structures smaller than 3 angstroms for exclusion of smaller chains, selecting unique PDBs based on the largest chain, protonation of structures, and their minimization via MOE software. Finally, CAVIAR software was employed to detect pockets in the protein structures.

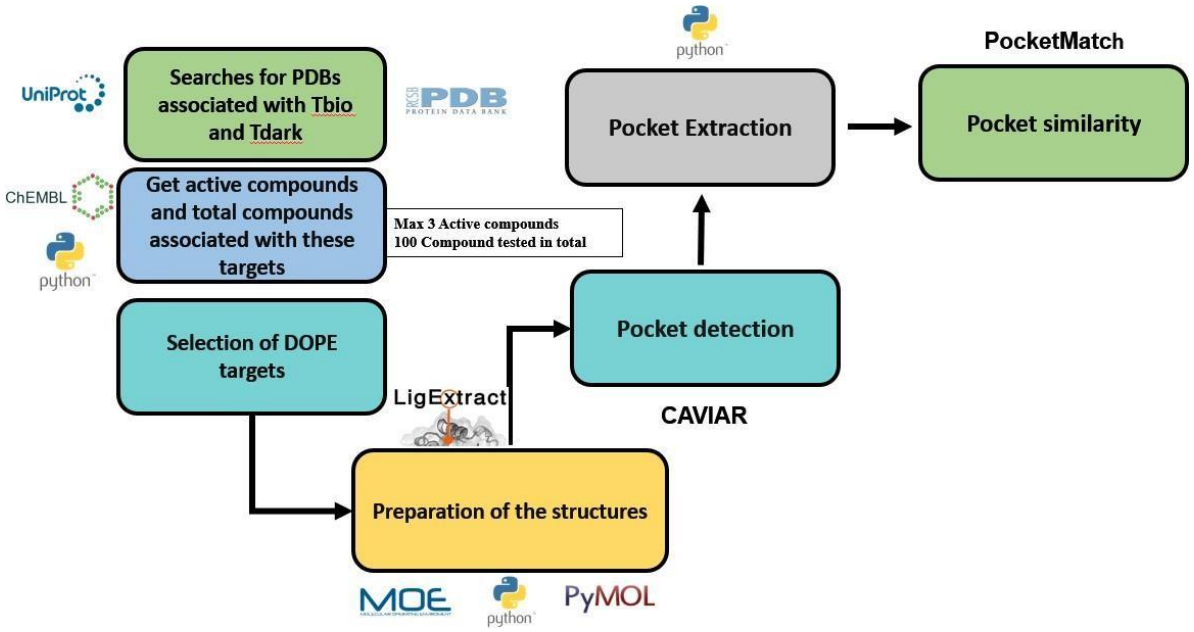


Figure 10: DOPE characterization process

3.1.4 Determining the similarity between targets.

After segregating all the structures, an analysis was conducted on the shared characteristics of protein pockets using thePocketMatch program. This tool emerged as the most suitable for the in-depth analysis required, enabling the comparison of pocket collections from DOPE and YNOTs with the well-known ones. In this way, pockets with similar attributes were identified among the groups.

Following this procedure, analyzes were carried out on the targets that obtained the best scores. These analyzes involved a visual comparison of the pockets as well as data processing and verification of the values obtained for each group, with in-depth

comparisons being carried out. This approach, which encompasses the identification of comparable grants and the meticulous analysis and refinement of the resulting data, was a vital and basic step in the current research work. This thorough approach allowed the acquisition of significant results, elaborated in the subsequent sections...

Chapter 4

Results and Discussion.

4 Results and Discussion

4.1 YNOTs groups

We initially filtered the YNOTs group and obtained approximately 20,427 UniProt codes related to the entire human proteome. Using a Python script developed by our research group, we were able to assess the correlation between the tested compounds and active compounds, since our objective was to acquire the target-compound relationship. The distribution of these 20,427 codes, in relation to their respective compounds, is illustrated in the graph below.

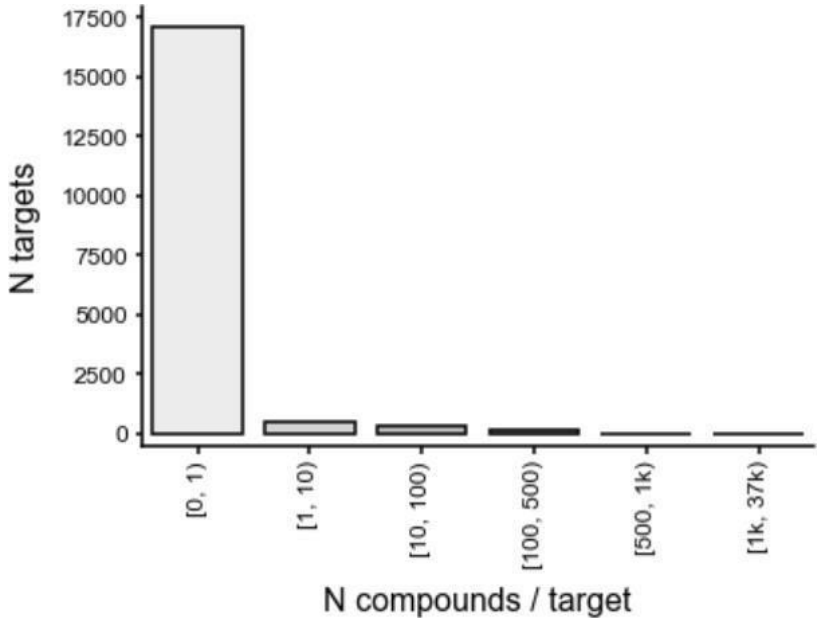
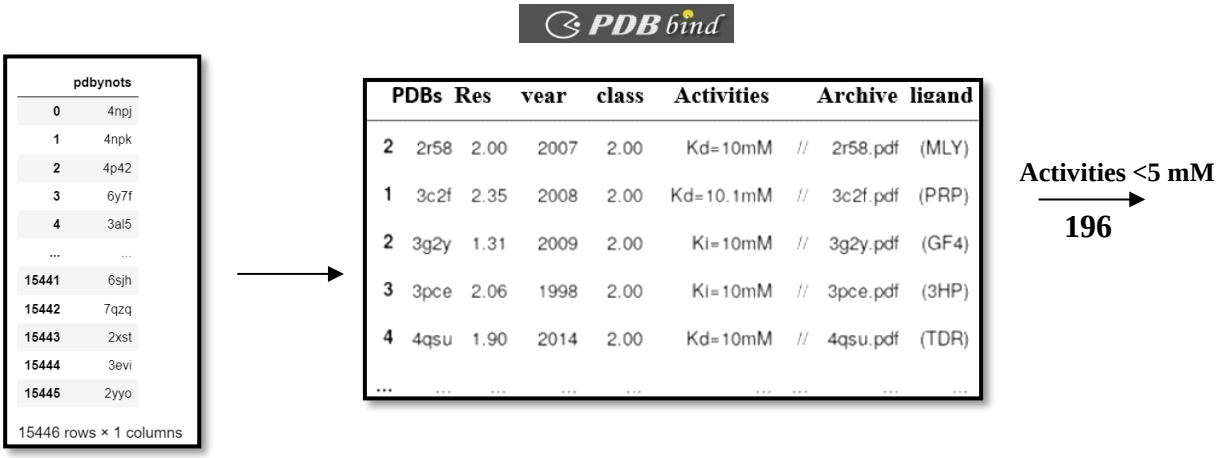


Figure 11: Representation of number of compound per targets of Ynots Goup

Following these procedures, we utilized another Python script to verify the relationships among these crystallographic structures. We successfully managed to compile a list of 3,700 UniProt codes that corresponded to crystallographic structures with resolutions less than 3 angstroms and lengths exceeding 100 amino acid residues. However, the majority of the UniProt codes still did not have associated crystallographic structures. The YNOTs list was then used to input the UniProt codes into Ligextract to

retrieve all the crystallographic structures linked to each UniProt code. Consequently, a total number of 15,446 PDB structures were associated with these 3,710 targets.

Subsequently, to ensure the YNOTs dataset did not contain known structures, we used the PDBbind database to remove structures of these ligands presented at Ki, Kd values less than 5 angstroms. The table below provides an overview of the data frame regarding ligand activities associated with PDB structures.



	pdbynots
0	4npj
1	4npk
2	4p42
3	6y7f
4	3al5
...	...
15441	6sjh
15442	7qzq
15443	2xst
15444	3evi
15445	2yyo

15446 rows × 1 columns

PDB bind

	PDBs	Res	year	class	Activities		Archive	ligand
2	2r58	2.00	2007	2.00	Kd=10mM	//	2r58.pdf	(MLY)
1	3c2f	2.35	2008	2.00	Kd=10.1mM	//	3c2f.pdf	(PRP)
2	3g2y	1.31	2009	2.00	Ki=10mM	//	3g2y.pdf	(GF4)
3	3pce	2.06	1998	2.00	Ki=10mM	//	3pce.pdf	(3HP)
4	4qsu	1.90	2014	2.00	Kd=10mM	//	4qsu.pdf	(TDR)
...	...	...	...	...	...	...	...	...

Activities <5 mM

196

Table 1: Process of exclusion of proteins from the YNTs groups

The value 196, as represented in the diagram, corresponds to the UniProt codes with activity values less than 5 micromolar corresponding to the value that was removed. This resulted in a final count of 3,514 codes. Using this data, along with the established relationships between PDBs, we pursued the process of obtaining the unique structure. This was based on the relationship between the chains, selecting the PDB structures containing the longest chains for their corresponding UniProt codes. The correlation between the structures and chains is detailed in table 2.

	PDBs	PDBs_lig	Chains
0	4p42	4p42_chain-A_lig-EGC.pdb	A;B
1	6y7f	6y7f_chain-A_lig-37X.pdb	A
2	3al6	3al6_chain-B_lig-AKG.pdb	B
3	6f4r	6f4r_chain-A_lig-OGA.pdb	A;B
4	6f4s	6f4s_chain-A_lig-AKG.pdb	A;B
...	...	...	...
2607	6skj	6skj_chain-B_lig-COA.pdb	B;D
2608	7bu5	7bu5_chain-A_lig-GLY.pdb	A
2609	2b9e	2b9e_chain-A_lig-SAM.pdb	A
2610	7ckf	7ckf_chain-A_lig-AF3.pdb	A
2611	7e5a	7e5a_chain-A_lig-AF3.pdb	A

Table 2 : Relationship of the chains of the YNOts groups

Upon concluding the cleaning steps of the crystallographic structures and completing the structure preparation in the MOE software, we proceeded to the analysis of the YNOT structures using the CAVIAR tool. This step revealed crucial information such as score, and the cavity of each target, regarding the identification and characterization of structural pocketsin protein-ligand complexes.

The detection and characterization of these structural pockets are paramount, because as they provide insights on the topology and nature of cavities within the proteins. table 3 illustrates the representation of the output files generated from CAVIAR results, highlighting a specific example associated with structure “1buh”:

PDB_chain	CavID	Ligab	Score	Size	Hydrophob	InterCh	AltLoc	Miss	Subcavs
1buh_A	1	0.8	4.6	346	45%	0	0	1	4
1buh_A	2	0.4	1.8	97	61%	0	0	1	1
1buh_AB	3	0.6	1.6	159	52%	1	0	1	3
1buh_A	4	1.0	0.5	41	73%	0	0	1	1
1buh_A	5	0.4	0.5	50	54%	0	0	1	1
PDB_chain	CavID	SubCavID	Size	Hydrophob.	Polar	Neg	Pos	Other	
1buh_A	1	1	21	24%	24%	33%	5%	14%	
1buh_A	1	2	96	31%	51%	12%	5%	0%	
1buh_A	1	3	126	40%	32%	20%	9%	0%	
1buh_A	1	4	103	68%	17%	12%	3%	1%	
1buh_A	2	1	97	61%	32%	7%	0%	0%	
1buh_AB	3	1	73	49%	1%	23%	26%	0%	
1buh_AB	3	2	33	39%	0%	15%	0%	45%	
1buh_AB	3	3	53	64%	21%	11%	2%	2%	
1buh_A	4	1	41	73%	15%	12%	0%	0%	
1buh_A	5	1	50	54%	44%	0%	0%	2%	

Table 3 CAVIAR Cavity detection results for the structure PDB 1buh

As illustrated in table 3, the output from CAVIAR reveals comprehensive details about the characteristics of individual pockets of YNOts groups. Additionally, CAVIAR parameterized the cavities based on their smoothness, which allowed greater accuracy in binding sites analyses. Among these results, we were particularly interested in the score values, as they provided a probabilistic insight into the likelihood of a specific site serving as a binding site.

Therefore, Figure 12 offers an analysis of all the PDB structures based on the CAVIAR results, aiming to understand the characteristics of these pockets that have the highest probability of being binding sites for the yet-to-be-identified proteins. The structures presented here correspond to the cavities of 2070 YNOTs, where each point represents a cavity obtained with the highest score values detected for each structure in CAVIAR.

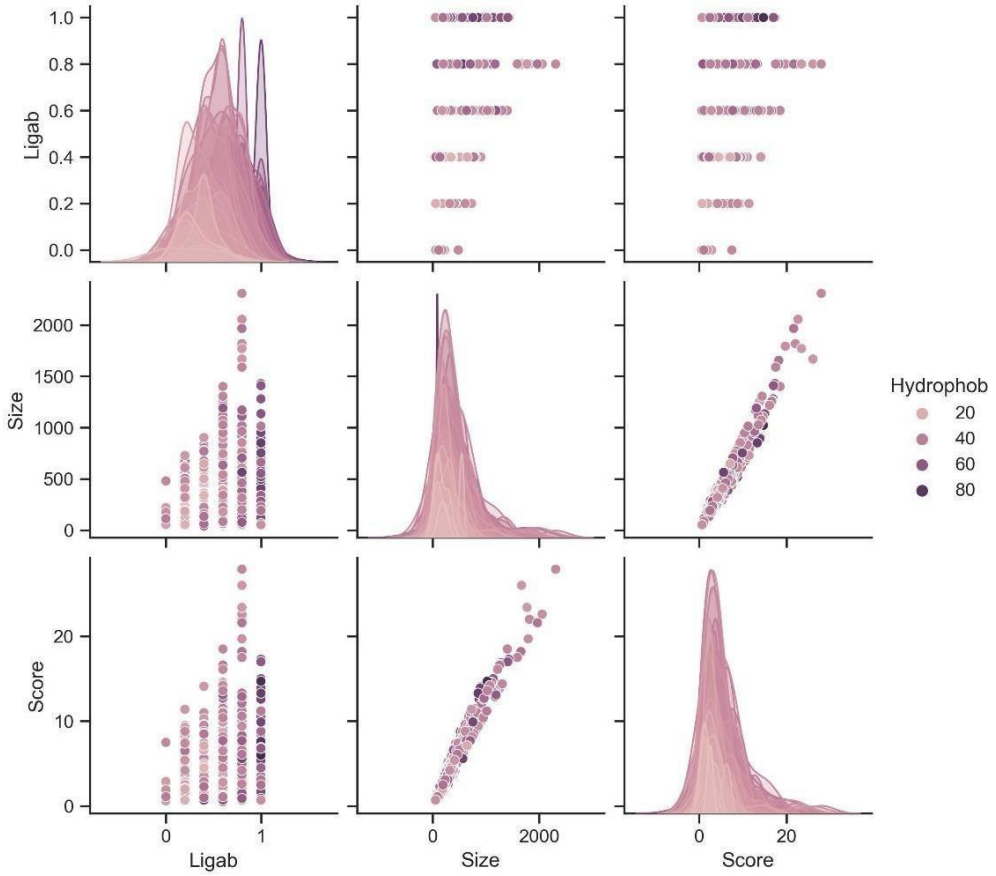


Figure 12: Graphic representation of the pockets of the Ynots groups

It is essential to pay attention to the subtleties of the data and its underlying links to improve our analysis. The graph presented intriguing insights, indicating that pockets with larger volumes stood out by showing higher scores when compared to those with lower

hydrophobic values. This discovery prompts us to think carefully about the variables that might influence these outcomes.

4.2 Well-Known Groups

The investigation of Well-known proteins commenced with a rigorous selection process. This procedure was guided by the guidelines established by Oprea et al., resulting in a comprehensive list of 2002 UniProt codes that represent these widely recognized within the scientific community. The assessment of the relationships between these proteins and their corresponding compounds utilized the extensive ChemBl database as a source of information.

These analyses for the well-known proteins were conducted using the same methodology we previously applied to the YNOTs proteins. Consequently, Figure 11 illustrates the relationship of well-known proteins with their respective ligand-activity profiles.

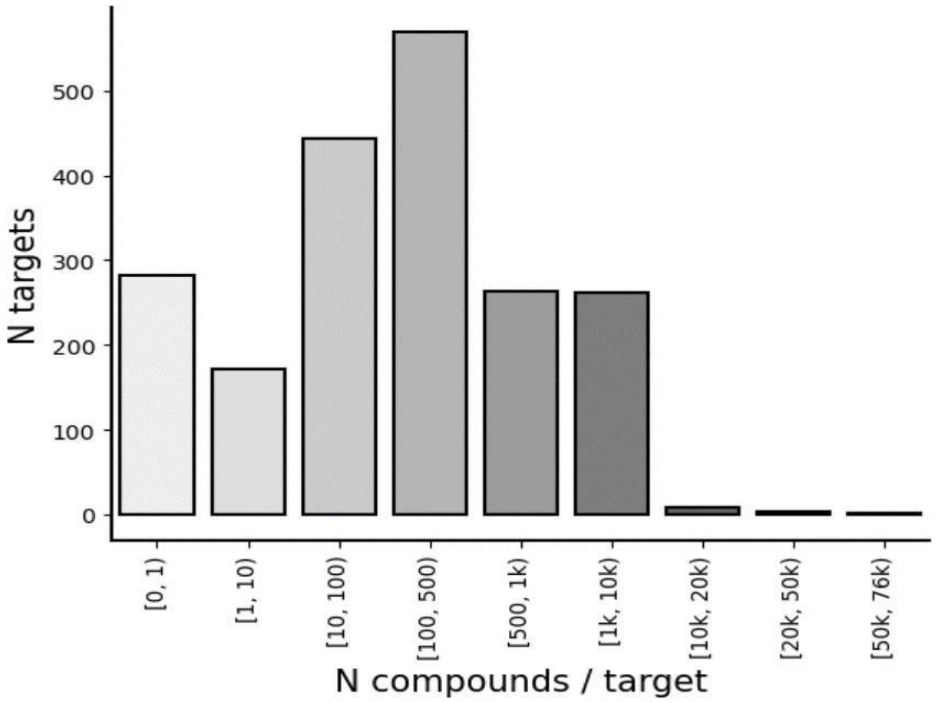


Figure 13: Representation of number of compounds per targets of WKT Groups

The analysis of ChEMBL data for the WKT reveals a noticeable and anticipated variation in the quantity of compounds attributed to different targets. Targets associated with WKT typically have a marked abundance of assigned compounds, in stark contrast to other target groups. Notably, a significant proportion of targets have either no registered compounds or just

one. Such observation can be attributed to two main reasons: the potential gaps of complete information in ChEMBL for certain targets and the existence of targets that, having already been well-established over time, no longer require new ligands registration. This heterogeneity in data reflects the diverse and ever-evolving nature of medicinal chemistry research, highlighting the need for consistent curation and updates of the ChEMBL database.

Table 4 demonstrates the differences between the number of structures associated with the WKT compared to the YNOTs Groups.

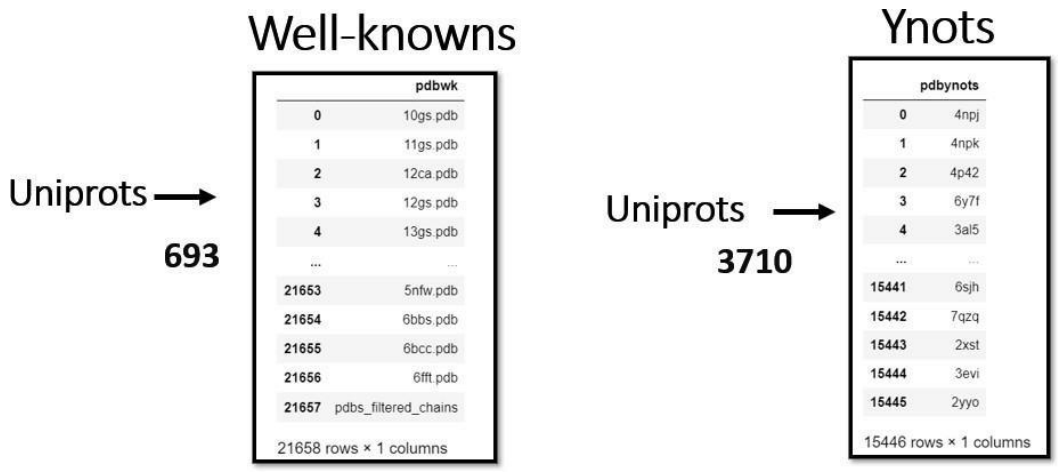


Table 4: PDBS relationships and Uniprots codes between the WKT groups and the Ynots

The comparative analysis revealed a notable discrepancy in the number of PDB structures available for protein-ligand relationship between WKT and YNOT complexes. For the WKT complexes, we identified a total of 21,658 structures associated with 693 distinct UniProt codes. In contrast, the Ynot complexes showed a more limited dataset with 15,446 structures linked to 3,710 UniProt codes.

Once these results were obtained, we proceeded with a series of data preprocessing steps for the WKT structures, as detailed in the methodology sections. These steps encompassed crucial tasks such as analysis using LigExtract, partitioning of structures based on chains, structure preparation through cleaning and protonation processes, and the use of the MOE software for additional optimization.

Subsequently, we subjected the WKT structures to detailed analyses of cavities using CAVIAR. Additionally, we performed a cluster analysis to identify significant groupings within the dataset. Figure 14 clearly illustrates the relationships between individual pocket and their corresponding peak score values obtained from the analysis.

.

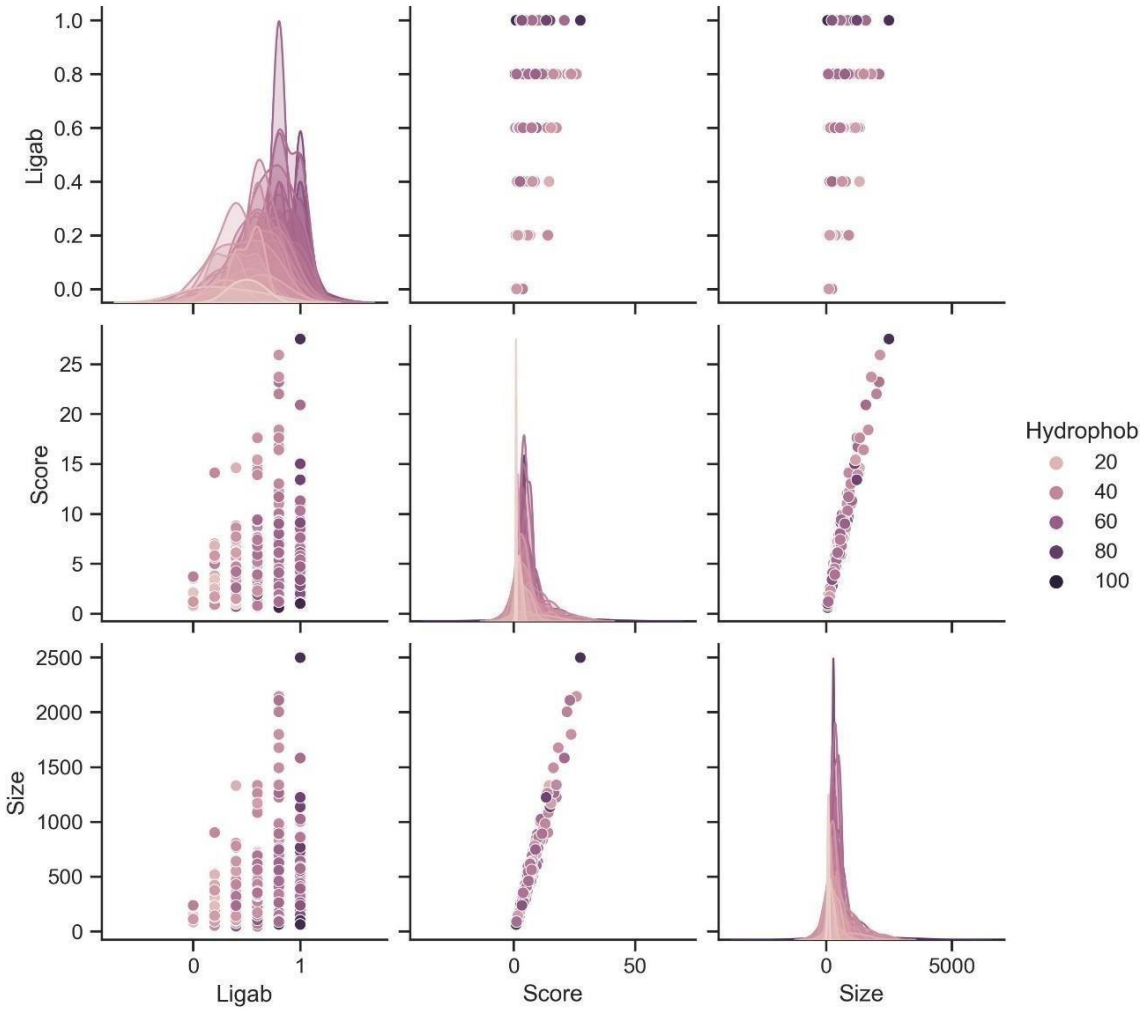


Figure 14: Graphic representation of the pockets of the WKT groups

Figure 14 demonstrates that pockets with greater volumes also exhibit elevated score values, accompanied by a substantial increase in hydrophobicity. This allows us to infer the presence of a significant relationship between the pocket's volume and its potential to act as a binding site The structures presented here correspond to the cavities of 693 WKT, where each point represents a cavity obtained with the highest score values detected for each structure in CAVIAR.

These procedures of pocket detection provided a more in-depth insight into the structural characteristics and binding properties of WKT complexes, highlighting the rich diversity of available structures. These results will serve as a foundation for further investigations, aiming to elucidate the interplay between structure and function in WKT protein-ligand complexes.

4.3 DOPE Groups

Contrary to the parameters of previous analyses, this approach resulted in a significantly reduced selection of DOPE targets, totaling only 12. This number was in line with our initial expectations and further facilitated a comprehensive manual examination during the detection and comparison stages between the sites. The following table lists the targets identified upon implementing the criteria described in the methodology:

Uniprot	PDB
P17174	3II0
Q15758	5LM4
P63096	1KJY
Q16828	1MKP
Q9C000	4IFP
O00764	2AJP
P63000	1MH1
Q9GZT4	3L6B
Q92499	4XW3
P09471	8E9X
P00442	2PB7

Table 5 : Dope Goups

4.4 Analysis of Similarities Between YNOTs and WKT

Initially, we carried out a comparative analysis of the physicochemical properties of the pockets present in the evaluated structures, based on the results from CAVIAR. This analysis is fundamental to understanding the diversity and nature of pockets across the various structures under investigation.

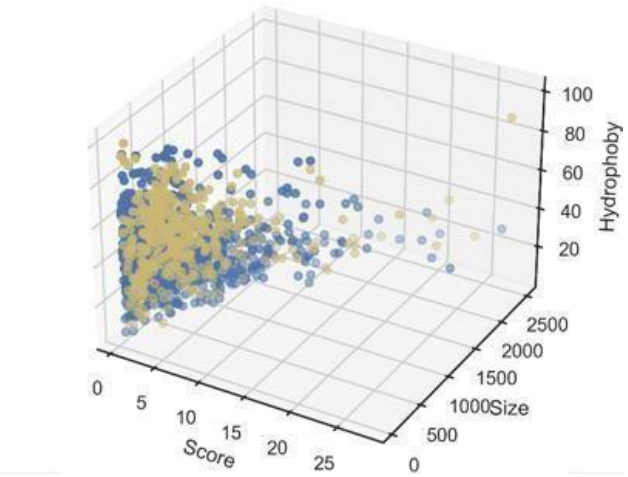


Figure 15: Comparison between the physical chemical properties of pockets, WKT and YNOTs

Notably, in the three-dimensional graph, the majority of the pockets are situated farther to the right, suggesting a convergence in the physicochemical characteristics between YNOts and WKT. This overlap might result from shared structural or functional characteristics of the proteins analyzed in both groups.

The last stage was to identify within the YNOTs group that have characteristics similar to the binding sites of WKT proteins. Figure 16 below presents the results of the analyses after the structures underwent evaluation with PocketMatch:

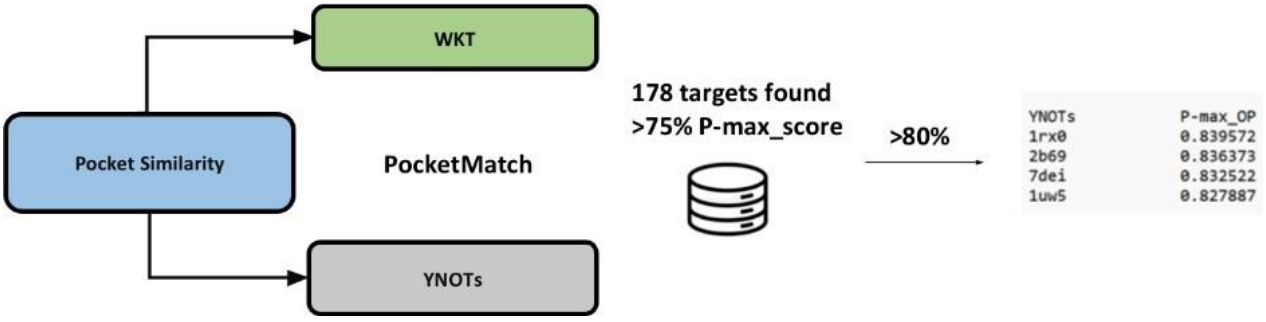


Figure 16 : Analyses of YNOTs group after the structures underwent evaluation with PocketMatch

A total of 410,240 combinations were generated between the "YNOTs" and the "WKT" groups. Within this dataset, we sought to identify targets with a "P-max" value exceeding 0.75, signifying a similarity of 75% or higher.

We identified 178 targets that exhibited similarities surpassing 75%. Moreover, 4 targets identified displayed similarities above 80%. These results suggest the potential for creating a new

database comprising structures that remain unexplored and might serve as potential targets for small molecules.

The main targets identified with similarities above 80% are as follows:

- PDB 1rx0: Presents the crystal structure of isobutyryl-CoA dehydrogenase complexed with substrate. The UniProt reference code for this target is Q9UKU7.

- PDB: Crystal structure of Human UDP-glucuronic acid decarboxylase. The UniProt reference code for this target is Q8NBZ7.

- Structure: Human ORP3 ORD domain in complex with PI(4)P. The UniProt reference code for this target is Q9H4L5.

- Structure: PITP-alpha complexed to phosphatidylinositol. The UniProt reference code for this target is Q9H4L5.

- These targets may represent interesting opportunities for future investigations and the development of therapeutic molecules.

4.5 Analysis of Similarities Between DOPE and WKT

The same analysis system that we applied to the YNOTs group was also employed for the Dope group. Therefore, the figure 17 presents the 12 structures associated with the WKT group, illustrating the relationships between the corresponding.

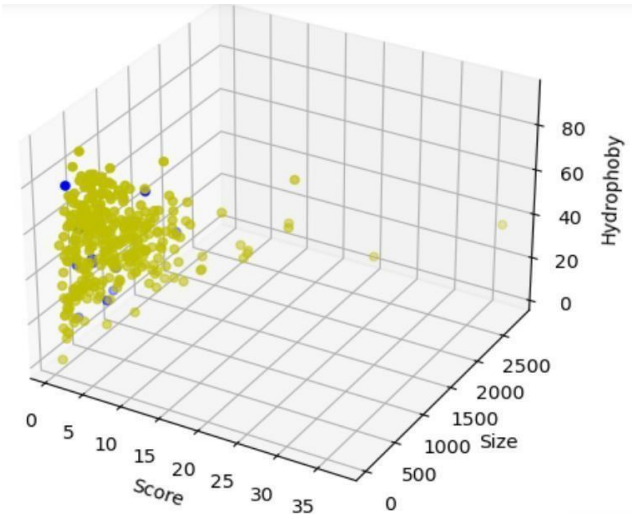


Figure 17: Comparison between the physical chemical properties of pockets, WKT and DOPE

The concluding step was the search for pockets within the DOPE group that presented characteristics similar to the binding sites of proteins in the WKT group. These results findings are presented in figure 18, which displays the results obtained through PocketMatch.

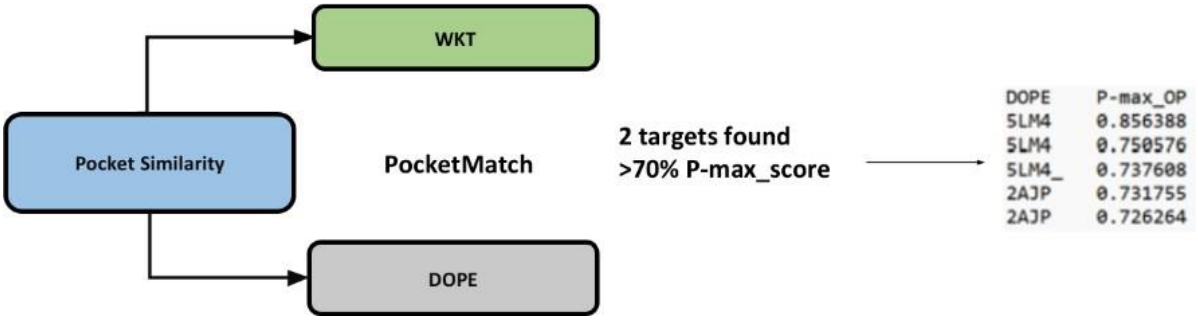


Figure 18: Analyses of YNOTs group after the structures underwent evaluation with PocketMatch

Using our established platform, the results revealed that the cavity corresponding to the allosteric center of Excitatory Amino Acid Transporter 1 (EAAT1) exhibited a good similarity score 94% with Phosphodiesterase 6 Delta Subunit (PDE6δ). This positioned it as the prime target candidate for further studies. Figure 19 illustrates the overlay of the pockets from both structures.

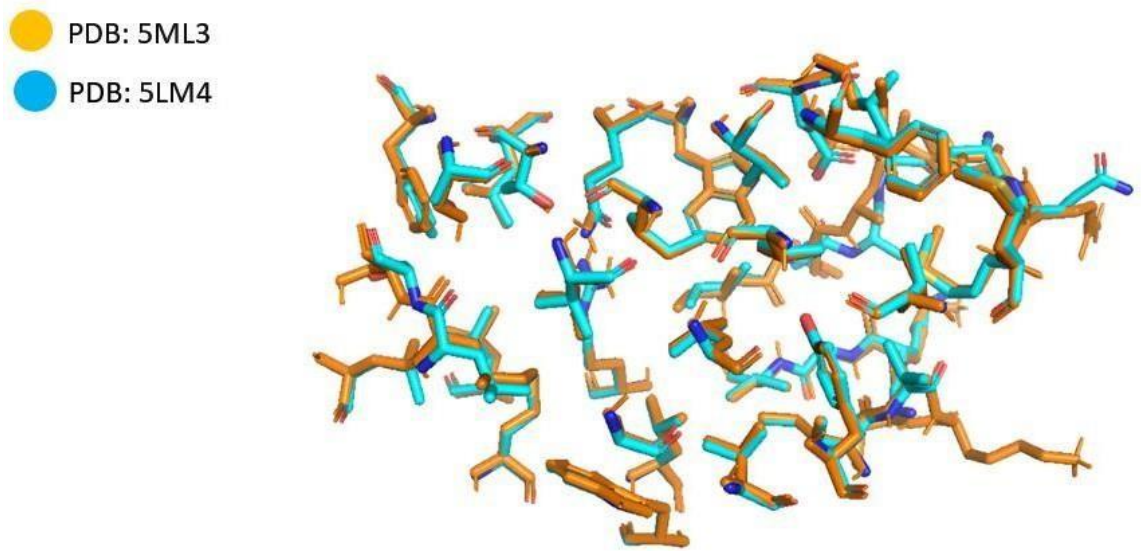


Figure 19 : Pocket alignment between proteins 5ml3 and 5lm4

As we can observe in figure 19, there was a very similar conformation between these structures, with some differences in certain amino acid residues that may contribute to the binding of different ligands within the pocket's structure. The figure 20 illustrates the relationship between the amino acids of these two proteins, as well as the comparison of amino acid sequences within the pockets.

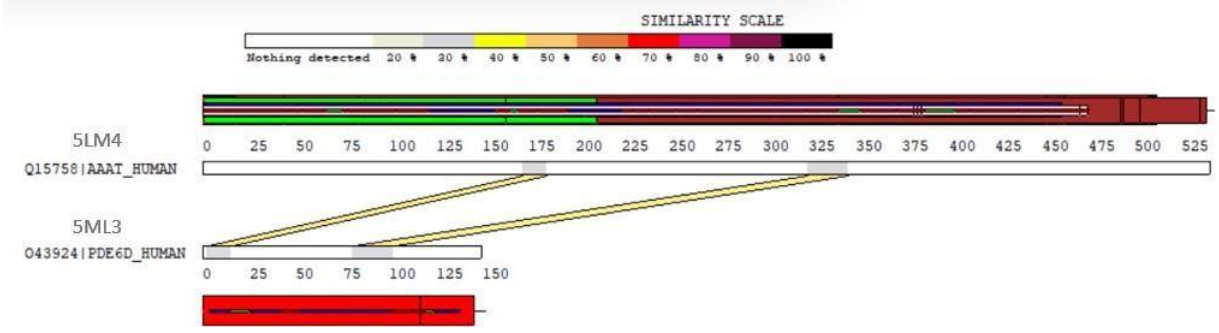


Figure 20: Sequence alignment of total proteins between 5ML3 and 5LM4

Regarding the amino acid sequences, these two PDB structures are totally different from each other, with 5LM4 being much longer in terms of amino acid length. However, when the focus is reduced to the pockets, it becomes evident that their binding sites share significant similarities, as we can see in figure 21, concluding that they are almost identical in their entirety.

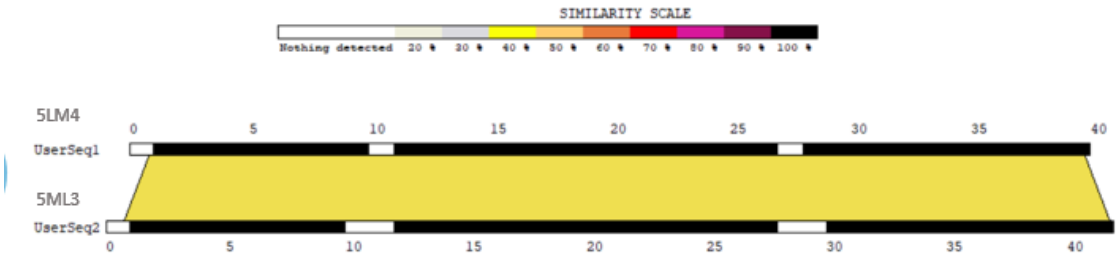


Figure 21: Sequence alignment of proteins pocket between 5ML3 and 5LM4

5 Conclusion

This project allowed for the in-depth analysis of protein binding sites, which is directly implicated in the discovery of novel therapeutics. This was achieved through a combination of data analysis techniques and the utilization of Python-based tools and scripts, which augmented our understanding and facilitated various discoveries.

In this way, for each analysis step conducted here, distinct insights emerged, emphasizing the vast field of the human proteome yet to be thoroughly explored. Consequently, this work allowed for the identification of these gaps and demonstrated many targets that can be explored in the field of drug discovery. Notably, within the YNOTs group, the crystal structures of isobutyryl-CoA dehydrogenase complexed with its substrate and the Human UDP-glucuronic acid decarboxylase stood out as the most promising targets. In the DOPE group, the phosphodiesterase 6 delta subunit (PDE6δ) emerged as a key target. Furthermore, we have identified a list of 178 potential pharmacological targets. Looking ahead, we intend to compile these findings into a comprehensive database.

REFERENCES

Burley, S.K., Berman, H.M., Kleywegt, G.J., Markley, J.L., Nakamura, H., Velankar, S., 2017. Protein Data Bank (PDB): The single global macromolecular structure archive, in: Methods in Molecular Biology. Humana Press Inc., pp. 627–641. https://doi.org/10.1007/978-1-4939-7000-1_26

Bairoch, A., Apweiler, R., Wu, C.H., Barker, W.C., Boeckmann, B., Ferro, S., Gasteiger, E., Huang, H., Lopez, R., Magrane, M., Martin, M.J., Natale, D.A., O'Donovan, C., Redaschi, N., Yeh, L.S.L., 2005. The Universal Protein Resource (UniProt). Nucleic Acids Res. 33, D154–D159. https://doi.org/10.1093/nar/gki070

Boutet, E., Lieberherr, D., Tognolli, M., Schneider, M., Bairoch, A., 2007. UniProtKB/Swiss-Prot, in: Plant Bioinformatics. Humana Press, pp. 89–112. https://doi.org/10.1007/978-1- 59745-535-0_4

Cramer, F., 1995. PHARMACEUTICA ACTA HEIJETIAE, Pharmaceutics Acta Helvetiae.

Eguida, M., Rognan, D., 2022. Estimating the Similarity between Protein Pockets. Int J Mol Sci. https://doi.org/10.3390/ijms232012462

Chemical Computing Group ULC, 2022. Molecular Operating Environment (MOE)

DeLano, W.L., 2002. The PyMOL molecular graphics system.

Ehrt, C., Brinkjost, T., Koch, O., 2019. Binding site characterization-similarity, promiscuity, and druggability. Medchemcomm 10, 1145–1159. https://doi.org/10.1039/c9md00102f

Garavelli, J.S., Huang, H., Laiho, K., McGarvey, P., Natale, D.A., Ross, K., Vinayaka, C.R., Wang, Q., Wang, Y., Yeh, L.S., Zhang, J., Ruch, P., Teodoro,

D., 2021. UniProt: the universal protein knowledgebase in 2021. Nucleic Acids Res. 49, D480–D489. https://doi.org/10.1093/nar/gkaa1100

Gaulton, A., Hersey, A., Nowotka, M.L., Patricia Bento, A., Chambers, J., Mendez, D., Mutowo, P., Atkinson, F., Bellis, L.J., Cibrian-Uhalte, E., Davies, M., Dedman, N., Karlsson, A., Magarinos, M.P., Overington, J.P., Papadatos, G., Smit, I., Leach, A.R., 2017. The ChEMBL database in 2017. Nucleic Acids Res. 45, D945–D954. https://doi.org/10.1093/nar/gkw1074

Jean-Quartier, C., Jeanquartier, F., Jurisica, I., Holzinger, A., 2018. In silico cancer research towards 3R. BMC Cancer 18. https://doi.org/10.1186/s12885-018-4302-0

Labute, P., 2009. Protonate3D: Assignment of ionization states and hydrogen coordinates to macromolecular structures. Proteins Struct. Funct. Bioinforma. 75, 187–205. https://doi.org/10.1002/prot.22234

Lee, A., Kim, D., 2019. CRDS: Consensus Reverse Docking System for target fishing. Bioinformatics 36, 959–960. https://doi.org/10.1093/bioinformatics/btz656

Lee, A., Lee, K., Kim, D., 2016. Using reverse docking for target identification and itsapplications for drug discovery. Expert Opin. Drug Discov. https://doi.org/10.1080/17460441.2016.1190706

Leelananda, S.P., Lindert, S., 2016. Computational methods in drug discovery.

Morales-Navarro, S., Prent-Peñaloza, L., Núñez, Y.A.R., Sánchez-Aros, L., Forero-Doria, O., González, W., Campilllo, N.E., Reyes-Parada, M., Martínez, A., Ramírez, D., 2019. Theoretical and experimental approaches aimed at drug design targeting neurodegenerative diseases. Processes. https://doi.org/10.3390/PR7120940

Oprea, T.I., Bologa, C.G., Brunak, S., Campbell, A., Gan, G.N., Gaulton, A., Gomez, S.M., Guha, R., Hersey, A., Holmes, J., Jadhav, A., Jensen, L.J., Johnson, G.L., Karlson, A., Leach, A.R., Ma'ayan, A., Malovannaya, A., Mani, S., Mathias, S.L., McManus, M.T., Meehan, T.F., Von Mering, C., Muthas, D., Nguyen,

D.T., Overington, J.P., Papadatos, G., Qin, J., Reich, C., Roth, B.L., Schürer, S.C., Simeonov, A., Sklar, L.A., Southall, N., Tomita, S., Tudose, I., Ursu, O., Vidović, D., Waller, A., Westergaard, D., Yang, J.J., Zahoránszky-Köhalmi, G., 2018. Unexplored therapeutic opportunities in the human genome. Nat Rev Drug Discov. https://doi.org/10.1038/nrd.2018.14

Papadatos, G., Overington, J.P., 2014. The ChEMBL database: A taster for medicinal chemists. Future Med. Chem. https://doi.org/10.4155/fmc.14.8

Shen, Q., Cheng, F., Song, H., Lu, W., Zhao, J., An, X., Liu, M., Chen, G., Zhao, Z., Zhang, J., 2017. Proteome-Scale Investigation of Protein Allosteric Regulation Perturbed by Somatic Mutations in 7,000 Cancer Genomes. Am J Hum Genet 100, 5–20. https://doi.org/10.1016/j.ajhg.2016.09.020

Soga, S., Shirai, H., Koborv, M., Hirayama, N., 2007. Use of amino acid composition to predict ligand-binding sites. J. Chem. Inf. Model. 47, 400–406. https://doi.org/10.1021/ci6002202

Sneader, W., 2005. Drug Discovery, Drug Discovery: A History. Wiley. https://doi.org/10.1002/0470015535

Sydow, D., Rodríguez-Guerra, J., Kimber, T.B., Schaller, D., Taylor, C.J., Chen, Y., Leja, M., Misra, S., Wichmann, M., Ariamajd, A., Volkamer, A., 2022. TeachOpenCADD 2022: open source and FAIR Python pipelines to assist in structural bioinformatics and cheminformatics research. Nucleic Acids Res. 50, W753–W760. https://doi.org/10.1093/nar/gkac267

Simonovsky, M., Meyers, J., 2020. Deeply Tough: Learning Structural Comparison of Protein Binding Sites. J. Chem. Inf. Model. 60, 2356–2366. https://doi.org/10.1021/acs.jcim.9b00554

Terstappen, G.C., Schlüpen, C., Raggiaschi, R., Gaviraghi, G., 2007. Target deconvolution strategies in drug discovery. Nat. Rev. Drug Discov. https://doi.org/10.1038/nrd2410

Volkamer, A., Kuhn, D., Rippmann, F., Rarey, M., 2012. DoGSiteScorer: a web server for automatic binding site prediction, analysis and druggability assessm

Yeturu, K., Chandra, N., 2008. PocketMatch: A new algorithm to compare binding sites in protein structures. BMC Bioinformatics 9. https://doi.org/10.1186/1471-2105-9-543

Yu, J., Zhou, Y., Tanaka, I., Yao, M., 2009. Roll: A new algorithm for the detection of protein pockets and cavities with a rolling probe sphere. Bioinformatics 26, 46–52. https://doi.org/10.1093/bioinformatics/btp599