ISABELA DE SOUZA GOMES

PROTEIN BIOINFORMATICS: OVERVIEW AND COMPUTATIONAL STRATEGIES TO DETERMINE PROTEIN INTERACTION PATTERNS TO ASSIST IN DISEASE CONTROL

Dissertation presented to the Universidade Federal de Viçosa, in partial requirement of Graduate Program in Computer Science, to obtain the degree of *Magister Scientiae*.

Advisor: Sabrina de Azevedo Silveira

VIÇOSA - MINAS GERAIS
2022

Ficha catalográfica elaborada pela Biblioteca Central da Universidade Federal de Viçosa - Campus Viçosa

T
G633p
2022

Gomes, Isabela de Souza, 1995-
Protein bioinformatics: overview and computational strategies to determine protein interaction patterns to assist in disease control / Isabela de Souza Gomes. – Viçosa, MG, 2022.
1 dissertação eletrônica (156 f.): il. (algumas color.).

Texto em inglês.
Inclui apêndices.
Orientador: Sabrina de Azevedo Silveira.
Dissertação (mestrado) - Universidade Federal de Viçosa, Departamento de Informática, 2022.
Referências bibliográficas: f. 111-134.
DOI: https://doi.org/10.47328/ufvbbt.2022.550
Modo de acesso: World Wide Web.

1. Bioinformática. 2. Interação proteína-proteína - Processamento de dados. 3. Peptídeos. 4. Aprendizado do computador. 5. Dinâmica molecular. 6. Proteínas. I. Silveira, Sabrina de Azevedo, 1983-. II. Universidade Federal de Viçosa. Departamento de Informática. Programa de Pós-Graduação em Ciência da Computação. III. Título.

CDD 22. ed. 570.285

Bibliotecário(a) responsável: Bruna Silva CRB-6/2552

ISABELA DE SOUZA GOMES

PROTEIN BIOINFORMATICS: OVERVIEW AND COMPUTATIONAL STRATEGIES TO DETERMINE PROTEIN INTERACTION PATTERNS TO ASSIST IN DISEASE CONTROL

Dissertation presented to the Universidade Federal de Viçosa, in partial requirement of Graduate Program in Computer Science, to obtain the degree of *Magister Scientiae*

APPROVED: September 1, 2022.

Assent:

Isabela de Souza Gomes
Author

Sabrina de Azevedo Silveira
Advisor

I dedicate this work to my family, that never abandoned me on this journey, and for always believing in me.

Acknowledgments

First, I thank God for allowing my entire academic education at the Federal University of Viçosa. These were many years of great learning and growth, both professional and personal.

I thank my family for always supporting and motivating me during the entire master's program, which was particularly difficult because it happened during the COVID-19 pandemic. The support given by my mother Eneida, my aunt Elisangela, and my brother Leonardo, who closely followed my entire trajectory, was fundamental. A special thanks to my sister Daniela who gave me great support in the research. Without her, I would not have been able to get where I am now. I thank Carlos, Charles, Vagner, Felippe, Vinícius, and Luciana for their contributions to my work and Amanda, Nicholas, and Marco for their support during their scientific initiation.

I thank all the professors of the graduate program in Computer Science for their teachings, tips, and orientations. I thank Professor Sabrina for the trust, advisory, and for always being willing to help. I also thank the partner professors in the research group for contributing so much to the work developed.

I thank my friends, especially Rubens, Daniel, Kayo, and Westerley, for their strength and support.

Finally, I thank CAPES for the investment. This work was carried out with the support of the Coordenação de Aperfeiçoamento de Pessoal de Nível Superior - Brazil (CAPES) - Funding Code 001.

"I can't be as confident about computer science as I can about biology. Biology easily has 500 years of exciting problems to work on. It's at that level."
(Donald Knuth)

Abstract

GOMES, Isabela de Souza, M.Sc., Universidade Federal de Viçosa, September, 2022. **Protein bioinformatics: overview and computational strategies to determine protein interaction patterns to assist in disease control**. Advisor: Sabrina de Azevedo Silveira.

Proteins are fundamental biomolecules for the metabolism of living beings and have several biotechnological uses. The computational study of this class of macromolecules allows the expansion of knowledge and speed in research applications, such as catalytic processes, three-dimensional conformations, enzyme inhibition, molecular engineering, among others. In this dissertation, we present a set of papers that purposes computational strategies to study three-dimensional structures of proteins. In the first work, we combine an in-house developed machine learning strategy with docking, MM-PBSA, and metadynamics simulations to detect potential inhibitors for SARS-COV-2 main protease. Computational strategies can help to speed up the process of drug discovery, reducing the time and cost of wet-lab experiments because they will be focused on fewer molecules. Our work points out six ligands that have a good interaction with our target in its active pocket, indicating an inhibitor behavior. We highlighted the strongest interaction of our experiments, M^{pro}-mirabegron complex, which was used as input for subsequent *in vitro* assays to validate the inhibition potential suggested by *in silico* experiments. In the second paper, we present a literature review of several bioinformatics tools for the study of proteins. The article is a very detailed material to support the choice of students and professionals for the most appropriate tool for a particular application. In the third work, we introduced the Propedia database for protein-peptide identification, which comprises over 19,000 high-resolution structures from the Protein Data Bank. Protein-peptide interactions can be useful for predicting, classifying, and scoring complexes or for designing new molecules. The main advantage of Propedia over other peptide databases is that it allows a more comprehensive analysis of similarity and redundancy. The papers presented here provide an overview of the diversity of protein bioinformatics study and some of its applications in biological problems.

Keywords: Bioinformatics. Protein-protein interaction. Peptide-protein interaction. Machine learning. Docking. Molecular dynamics. Proteins.

Resumo

GOMES, Isabela de Souza, M.Sc., Universidade Federal de Viçosa, setembro de 2021. **Bioinformática de proteínas: panorama e estratégias computacionais para determinação de padrões de interações entre proteínas para auxiliar no controle de doenças**. Orientador: Sabrina de Azevedo Silveira.

Proteínas são biomoléculas fundamentais para o metabolismo dos seres vivos e possuem diversas aplicações biotecnológicas. O estudo computacional dessas macromoléculas permite expandir o conhecimento e dar velocidade em pesquisas em diversas áreas, como processos catalíticos, conformações tridimensionais, inibição enzimática, engenharia molecular, entre outras. Nesta dissertação, apresentamos um conjunto de artigos que fazem a proposta de estratégias computacionais para o estudo de estruturas tridimensionais de proteínas. No primeiro trabalho, combinamos uma estratégia de aprendizagem de máquinas, desenvolvida pelo nosso grupo de pesquisa, com simulações de docking, MM-PBSA, e metadinâmica para detectar potenciais inibidores da *main protease* de SARS-COV-2. Estratégias computacionais podem acelerar o processo de descoberta de fármacos, reduzindo o tempo e o custo dos experimentos em bancada, visto que estes se concentrarão em menos moléculas. Nós evidenciamos seis ligantes que têm uma boa interação com a proteína alvo no seu sítio ativo, indicando um comportamento inibidor promissor. A interação mais forte detectada, o complexo M^{pro}-mirabegron, foi utilizado como insumo para ensaios *in vitro* subsequentes a fim de validar o potencial de inibição sugerido pelos ensaios *in silico*. No segundo trabalho, apresentamos o trabalho de revisão de literatura sobre diversas ferramentas de bioinformática para o estudo de proteínas. O artigo é um material bastante abrangente para embasar a escolha de estudantes e profissionais pela ferramenta mais adequada para determinada aplicação. No terceiro trabalho, apresentamos a base de dados Propedia para consulta de interações de peptídeo-proteína, que compreende mais de 19.000 estruturas de alta resolução do Protein Data Bank. As interações proteína-peptídeo podem ser úteis para prever, classificar e pontuar complexos ou para a concepção de novas moléculas. A principal vantagem do Propedia sobre outras bases de dados de peptídeos é que permite uma análise mais abrangente da similaridade e redundância. Os artigos aqui apresentados trazem uma visão geral da diversidade do estudo de bioinformática de proteínas e algumas de suas aplicações em problemas biológicos.

Palavras-chave: Bioinformática. Interação proteína-proteína. Interação peptídeo-proteína. Aprendizado de máquina. Ancoragem molecular. Dinâmica molecular.Proteínas.

List of Figures

List of Tables

Contents

Introduction

Bioinformatics is an area of applied computing that is very broad and of great importance in various areas of knowledge, such as biology, pharmacy, medicine, and agriculture, among others (PAIVA et al., 2022; LUSCOMBE; GREENBAUM; GERSTEIN, 2001; BILOTTA; TRADIGO; VELTRI, 2019). One of the most prominent branches is structural bioinformatics, which deals with the structural conformation of biomolecules and how these structural arrangements interact with each other. Intermolecular interactions are fundamental to macromolecules such as nucleic acids and proteins, and these interactions determine how the chains are organized. These two classes of biomolecules are key components of metabolism, and for this reason, they are the subject of many bioinformatics research projects (CAZALS; DREYFUS, 2016; MEDEMA, 2021).

One focus of our research group is the calculation of interactions between protein and other molecules (ligands, peptides, and proteins). However, cases of COVID-19 began to appear in Brazil and worldwide in the early 2020s, which lead the scientific community to study ways to mitigate the proliferation of SARS-COV-2 and its social consequences. The spread of a new disease is a common trigger for virtual screening studies (GALINDEZ et al., 2021). This technique aims to raise many ligands for a given protein target and refine this result through machine learning algorithms. Once this initial filtering is done, it is possible to apply more techniques of computer simulations, such as docking and molecular dynamics, to further reduce the number of possible molecules that interact with the target. The principal goal in this strategy is to direct the essential wet-lab studies for drug and vaccine discovery, reducing the number of possibilities for only the molecules that have already demonstrated interaction potential computationally (WU et al., 2020; AHMED; QUADEER; MCKAY, 2020; JIN et al., 2020; ZHANG et al., 2020).

Thus, in this work we design, implement, and evaluate a computational strategy that can respond to this important demand of society, scientifically and socially. Also, we work on the design and development of a database that allows to cluster protein-peptide interfaces according to varied criteria. Last, we work on the development of Protein Structural bioinformatics Overview (PreStO), an interactive visualization strategy created to support a description and main topics of protein structural bioinformatics.

This dissertation is structured as a collection of papers, standardized by the Graduate Technical Council of the Universidade Federal de Viçosa (UFV, Conselho Técnico de Pós Graduação da Universidade Federal de Viçosa, 2018). This format is composed by: gen-

eral introduction, papers and manuscripts (Chapters 2, 1, 3), conclusion, and appendixes (Appendixes A, B, C).

The Chapter 1 presents the first paper, *Protein structural bioinformatics: an overview* (PAIVA et al., 2022) which was published in August, 2022 in Computers in Biology and Medicine (IF 6.698, 2022). In this work, we describe several tools used in protein structural bioinformatics to guide researchers to choose a suitable solution for the problem. We propose PreStO (Protein Structural bioinformatics Overview) an interactive visualization strategy to present and summarize the structural bioinformatics tools to be accessible for the scientific community.

Chapter 2 presents the second paper, *Computational prediction of potential inhibitors for SARS-COV-2 main protease based on machine learning, docking, MM-PBSA calculations, and metadynamics* (GOMES et al., 2022), which was published in April 2022 in PLOS One journal (IF 3.24, 2020). It proposes a machine learning strategy to perform virtual screening, supporting on the prediction of potential ligands to an important enzyme of SARS-COV-2 metabolism, using machine learning and structural bioinformatics techniques.

The Chapter 3 presents the third paper, *Propedia: a database for protein-peptide identification based on a hybrid clustering algorithm* (MARTINS et al., 2021). This work was published in January, 2021 in BMC Bioinformatics (IF 3.169, 2020). We develop a protein-peptide interaction database clustered by sequences, interfaces, and binding sites.

The last section is the general conclusion about all the works developed during the master's degree and the opportunities derived from them. The Appendixes are related to the Supplementary Information of each paper and related to the patent request to the Universidade Federal de Viçosa regarding the results presented in Chapter 2.

Besides the papers presented in this dissertation, I work in the project about the webserver PRIorI (PRotein-PRotein InteractiOn gRaph Isomorfism). It has been developed to explore and calculate protein-protein interaction in all structures deposited in Protein Data Bank (PDB) based on ppiGReMLIN strategy (QUEIROZ et al., 2020) developed in our group. This work is not in the scope of this dissertation because we are currently reviewing the manuscript for submission.

Chapter 1

Protein structural bioinformatics: an overview

Proteins play a crucial role in organisms in nature. They can perform structural, catalytic, transport, and defense functions in cells, among others. We understand that a variety of resources do exist to work with protein structural bioinformatics, which performs tasks such as protein modeling, protein docking, protein molecular dynamics, protein interaction, active, and binding site prediction, and mutation analysis. Nonetheless, they are generally spread all over different online repositories. For the students or professionals interested in working with protein structural bioinformatics, it may not be trivial to know what resources he/she should learn/use or where these could be accessed. Here, the main subareas in the field of protein structural bioinformatics are introduced with a brief description, and we point to and discuss several online resources, such as methods, databases, and tools, to give an overview of this research field. Furthermore, we developed Protein Structural bioinformatics Overview (PreStO), a web tool available at <http://bioinfo.dcc.ufmg.br/presto/>, to organize and make it possible to retrieve these online resources based on a search term. We believe that this paper can be a starting point for potential bioinformaticians to trace a path that can be followed to build competencies and achieve knowledge milestones in the context of protein structural bioinformatics.

1.1 Introduction

Bioinformatics is a research field that aims to search for knowledge on biological data, more specifically biomolecules, through models and algorithms from Computer Science. It covers the collection, storage, retrieval, manipulation, and modeling of data for analysis, visualization, or prediction through the development of computational algorithms and strategies (Nature, 2021). It also employs knowledge of Physics, Chemistry, Statistics, and Mathematics in the solution of biological problems, thus promoting the development of the various areas involved (LUSCOMBE; GREENBAUM; GERSTEIN, 2001; BILOTTA; TRADIGO; VELTRI, 2019). A variety of biological data, such as nucleotide sequence, gene expression, protein sequence, and protein structure, which give rise to omics data, have been generated at a fast pace, and demand integration, organization, and the development of reliable and accurate computational

strategies to improve the understanding of the relationship between structure and function of biomolecules (CAZALS; DREYFUS, 2016; MEDEMA, 2021).

A natural question that arises is how to prepare human resources to work in such a broad field as bioinformatics. The International Society for Computational Biology (ISCB) has set a Curriculum Task Force that has published a set of reports and papers focused on the definition and refinement of curricular guidelines for training in bioinformatics. This work can be followed in the papers (WELCH et al., 2014; WELCH et al., 2016) and (MULDER et al., 2018). In a general manner, the idea was to define the core competencies, different types of users (profiles), and how these competencies match each type of user. This was performed in an iterative process involving not only ISCB but also other education venues, including the Global Organisation for Bioinformatics Learning, Education, and Training (GOBLET).

Structural bioinformatics comprises data resources, algorithms, and tools for investigating, analyzing, predicting, and interpreting biomacromolecular structures. More specifically, we are interested in protein structural bioinformatics. Proteins are large and complex molecules that perform a myriad of functions in organisms. They are formed by the union of amino acids and can take on different sizes and shapes. Proteins perform much of the work in cells and are necessary for the structure, function, and regulation of tissues and organs in organisms (KESKIN et al., 2008). Currently, there is a growing number of protein structures and sequences deposited in specialized databases, mainly provided by the advance in genomic sequencing technologies and methods for determining structures (AKCAPINAR; SEZERMAN, 2017).

Initially, information about proteins was available in terms of their amino acid sequence. 2021 marks 50 years of the Protein Data Bank (PDB) (BERMAN et al., 2000a), a catalog of all macromolecule structures we know to date. In 1971, the PDB was established at Brookhaven National Laboratory, led by Walter Hamilton, which comprises seven structures. The genome sequencing of organisms and the rapid increase in the number of three-dimensional macromolecular structures available have given rise to structural bioinformatics. Structural Bioinformatics has two major goals: the creation of general-purpose methods for manipulating information about biological macromolecules and the application of these methods to solve problems in biology, to generate new knowledge (ALTMAN; DUGAN, 2009).

Since the early years of PDB, much progress has been made in protein structure prediction, with emphasis on AlphaFold2 (JUMPER et al., 2021) in the CASP14 competition, which was able to predict protein domain structures with accuracy close to that of experimental methods (KRYSHTAFOVYCH et al., 2021). AlphaFold2 was released with over 300k protein models and is scheduled to cover over 100 million proteins, which demands structural biology tools that can be applied on a proteome-wide scale thus posing new challenges and opportunities for these tools (AKDEL et al., 2021).

In recent years, the growing volumes of biological data have demanded scalable and data-driven bioinformatics models, algorithms, and tools. Thus, the need for life science scien-

tists to develop basic bioinformatics skills has increased. Even experienced bioinformaticians need to update their knowledge and skills, so new algorithms and tools can be developed in response to the advances in science. However, basic data science, bioinformatics, and programming can be still relatively rare in life science curricula, and most biologists receive little or no formal training for the computational aspects of their field (ATTWOOD et al., 2019). Another important aspect of bioinformatics is that finding appropriate material can be challenging as it is often scattered on the internet or hidden in its home institution.

To tackle this challenge, the scientific community proposed a set of rules named FAIR (Findable, Accessible, Interoperable, and Reusable) to support potential users in easily finding these digital objects (GARCIA et al., 2020). Click2Drug is an online repository that provides a comprehensive list of protein structure tools and databases. On its website [1], users can find hundreds of works from the most diverse applications, mainly focused on drug design research. Zhang Lab [2] brings a series of tools, publications, and databases related to structural bioinformatics. Protein structure prediction, protein-protein interactions, and ligand-docking are some examples of the work carried out by the group. In bio.tools [3], users find an extensive record of software, databases, and services related to bioinformatics. These resources can be found by searching for expressions or keywords and are described using accurate semantics and syntax. Datasets2Tools (TORRE et al., 2018) consists of an online repository with thousands of analyses, databases, and computational tools. The website [4] analyzes the resources according to some features, such as accessibility, interoperability, and reusability.

Nonetheless, despite the relevant contributions of the mentioned works, repositories that are focused on a specific topic in a subarea or bring a list of tools, without delving into them to some degree, do not provide a basic understanding of the field and its subareas and do not mention positive and negative aspects of each tool. Hence the potential user has no support in choosing an appropriate tool for a task of interest. In addition, some repositories are outdated and present resources that are no longer available, such as LISE (XIE et al., 2013), Pocketome (KUFAREVA; ILATOVSKIY; ABAGYAN, 2012), and OpenAstexViewer (OPENASTEXVIEWER,), which were not accessible as of the date of writing.

To overcome these challenges, in this work we present an overview of the main subareas of protein structure bioinformatics. In the next sections, we discuss docking, molecular dynamics, molecular visualization, structure prediction, mutation analysis, catalytic and binding site prediction, and databases. For each subarea, we present a brief introduction and definition followed by a set of widely used tools to address common tasks in the subarea. To bring as many examples of tools as possible, in addition to the works presented here, several other examples are described in the Supplementary Material[5]. For each tool, in turn, we try to give an idea of

[1] <https://www.click2drug.org/>
[2] <https://zhanggroup.org/>
[3] <https://bio.tools/>
[4] <https://maayanlab.cloud/datasets2tools>
[5] The supplementary material of this paper is not in the content of this dissertation

how it works, with the intuition behind the proposed algorithm (when it is the case), as well as positive aspects and limitations, in addition to mentioning which other methods/tools it was compared with. It is important to point out that this paper does not present an exhaustive list of methods/tools in the field of structural bioinformatics of proteins, as it would be infeasible since there is a huge number of methods and tools available. We consider this set of tools a relevant starting point for those that are beginning in the field.

1.1.1 PreStO

We developed Protein Structural bioinformatics Overview (PreStO), an interactive visualization tool that organizes hierarchically or in a tabular manner all the resources presented in this paper, pointing to their papers' DOI and URL and providing a search tool that allows users to retrieve these resources. This tool is shown in Figure S1[6], with a tree graph of the subareas and a wordle (cloud of words) created based on frequent words from titles and abstracts of each online resource described in the manuscript. Moreover, a table is available, which lists the title, abstract, DOI, and tool URL of all resources. The user may select a keyword from the wordle or submit a search term of interest. The web tool uses the Term Frequency–Inverse Document Frequency (TF–IDF) (JONES, 1972) to measure and rank the importance of a search term for each online resource, based on their titles or abstracts. Finally, the user can download the search result in BibTeX or CSV format.

1.2 Docking

The elucidation of the interactions between proteins and ligands, or between proteins, is very relevant to various fields of knowledge, such as the development of new drugs and elucidation of molecular recognition (SOUSA et al., 2013; WENG et al., 2020). This stimulated the emergence of the technique of molecular docking, which is now widespread. It consists of calculating the best orientation that a molecule assumes to form a stable complex with a receptor. It is composed of two fundamental components: the sampling algorithm and the scoring function (WANG et al., 2016).

The sampling algorithm intends to generate different poses of the ligand arrangement inside the delimited simulation box. Such algorithms can essentially be classified into two categories: systematic and stochastic searches. The first seeks to exhaustively explore all the space allowed for simulation and has a high degree of freedom as a consequence. For such reason, the systematic search is a typical method for blind docking, whose main objective is to find the protein binding site. The second one, on the other hand, is a refinement search, through genetic algorithms, for example, and may not explore all the simulation space allowed. It is often used

[6] The supplementary material of this paper is not in the content of this dissertation

as local docking, when the binding site is already precisely known or after a blind docking step (WANG et al., 2016; WENG et al., 2020). The pose generation is heavily affected by the simulation space and the flexibility of the molecules involved. With the evolution of graphics boards, the determination of poses using much more flexible molecules has become faster. Therefore, the docking efficiency in predicting the interaction poses has increased (SOUSA et al., 2013).

The scoring function, the other fundamental component of docking technique, aims to rank the conformations obtained by the sampling algorithm and predict the binding affinity between the two components. It is an equation that can estimate the thermodynamic properties of the complex arrangement and rank instability order. Currently, this aspect of the docking methodology is the bottleneck to define the best fit between two molecules, since it is a challenge to accurately calculate the affinity of several molecular groups (WANG et al., 2016). The scoring functions can be of three sorts: empirical, force field and knowledge-based. Empirical functions estimate the affinity by adding important contact terms (H-bonds, for example) whose values have been previously determined. Force field ones estimate the affinity by calculating the parameters with a specific force field. The knowledge-based apply machine learning to predict the affinity (SOUSA et al., 2013).

Here, we present some tools that are commonly used for docking, summarized in Table 1. For protein-ligand, we describe GOLD, AutoDock Vina, and SwissDock, for protein-protein, we present ClusPro, pepATTRACT, HDOCK, and ZDOCK, and the hybrid tool HADDOCK (it can be used to perform protein-ligand and protein-protein docking). The distinction between the tools is commonly observed, due to the discrepancy in the complexity of the structures involved. The flexibility of the system components is the most important factor in this matter. Ligands, which are small molecules, have more freedom of translation, rotation and torsion than proteins, which have their movements restricted by a well-defined tertiary structure. Under this difference in the structures involved in docking, the scoring functions for affinity prediction and ranking are developed differently to best suit each scenario.

The molecular docking technique is widely used to predict molecular interactions. Several tools have been developed to perform this type of calculation, with certain specificities when it comes to a ligand or a peptide/protein. Based on the tools listed, we realize that the most classical tools for protein-ligand docking mostly run locally, since ligands are small molecules, which facilitates the positioning calculation. As for protein-protein docking, the tools developed tend to be web servers, which allows simulations to run more efficiently on more robust machines. In both types, there is a constant evolution of the scoring functions so that the energy description of the conformations is increasingly accurate.

It is important to emphasize that docking assays are usually used as an initial step in more complex simulations, such as molecular dynamics and metadynamics. Since it is a simulation performed in a vacuum and the molecular motions are restricted, its results may not represent reality with precision. However, these simulations are highly dependent on well-defined

Table 1 – Summary of docking tools.

Name	URL	Features	Limitations	Reference
GOLD	<https://www.ccdc.cam.ac.uk/solutions/csd-discovery/Components/Gold/>	partial flexibility close to protein active site frequently improvements internal H-bonds as parameter	only supports protein-ligand docking commercially available	(JONES et al., 1997)
AutoDock Vina	<http://vina.scripps.edu/download.html>	open source gradient optimization method quicker and more accurate than AutoDock 4 calculate grid maps automatically can use multiple CPU cores	no GUI it uses pdbqt format dependence of AutoDock Tools only supports protein-ligand docking	(TROTT; OLSON, 2009)
SwissDock	<http://www.swissdock.ch>	friendly user interface web service blind and local docking	only supports protein-ligand docking limited parameters control	(GROSDIDIER; ZOETE; MICHIELIN, 2011)
ClusPro	<https://cluspro.org>	web-server frequently updates CAPRI validation	only supports protein-protein docking only supports PDB format limited scoring function	(KOZAKOV et al., 2017)
pepATTRACT	<https://bioserv.rpbs.univ-paris-diderot.fr/services/pepATTRACT/>	web service does not require the protein binding site precise results	only performs peptide-protein docking 18h of processing it does not run on GPU	(VRIES et al., 2017)
HDOCK	<http://hdock.phys.hust.edu.cn/>	it can receive aminoacid sequence as input user friendly interface allows the incorporation of experimental information online molecular visualization allows protein-RNA/DNA docking	dependence of high-quality homologous complex templates	(YAN et al., 2020)
ZDOCK	<http://zdock.umassmed.edu>	user-friendly interface editable scoring function display input and output visualization with Jmol about 12 minutes of runtime	no clustering no post processing analysis included	(PIERCE et al., 2014)
HADDOCK	<https://wenmr.science.uu.nl/haddock2.4/>	web server allows all kind of docking allows docking with more then 2 molecules clustering and post processing analysis	limited control for new users	(HONORATO et al., 2021; van Zundert et al., 2016)

three-dimensional structures, mainly obtained by crystallographic methods. For this reason, it is possible to perform minimizations and quick molecular dynamics simulations prior to the docking step, to ensure the quality of the structures that were not defined with enough resolution (IGLESIAS et al., 2018).

It is important to mention that the quality of methods has increased over time due to initiatives as Critical Assessment of PRedicted Interactions (CAPRI), which conducted an independent assessment of docking techniques. Its main goal is to verify the quality of protein-protein docking of experimentally determined complexes to predict three-dimensional structures (JANIN et al., 2003).

1.2.1 GOLD

Genetic Optimisation for Ligand Docking (GOLD) was one of the first pieces of software developed for automated protein-ligand docking that explores the full flexibility of the ligand and partial for the protein close to its active site. It uses a genetic algorithm to fit the ligand into the protein binding site, modifying its position, orientation, and conformation. GOLD passed through a lot of improvements during the years, one of them is the scoring function that, nowadays is composed of four ligand energy parameters: internal and external van der Waals, external H-bond, and internal torsion. Another component, internal H-bond, may be added if necessary (JONES et al., 1997). Currently, GOLD is commercially available by The Cambridge Crystallographic Data Centre (CCDC) at <https://www.ccdc.cam.ac.uk/solutions/csd-discovery/Components/Gold/>.

1.2.2 AutoDock Vina

AutoDock Vina is a widely used docking program freely available for academic purposes, designed to be quicker and more accurate in binding prediction than its ancestor AutoDock4. It also has the advantage of calculating the grid maps automatically and generating a clean output for the user. A grid map is a set of regularly spaced points, used to segment the region of interest in the docking calculation. Vina can use multiple CPU cores, which is one of the reasons behind its speed-up. The positioning algorithm is based on iterated local search global optimizer, which is a combination of stochastic global and local optimization approaches. There is a gradient optimization method that uses the derivatives of the empirical scoring function among the ligand position, orientation, torsion tree and rotatable bonds. These parameters are set on an additional software, AutoDock Tools, previously developed to manage AutoDock files (TROTT; OLSON, 2009). Vina has the limitation of lacking graphical user interface. Besides, it uses a very specific file format (pdbqt), which makes it difficult to visualize the results. AutoDock Vina can be downloaded at <http://vina.scripps.edu/download.html>.

1.2.3 SwissDock

SwissDock is a web service developed by the Swiss Institute of Bioinformatics, which is part of the tools designed in the SwissDrugDesign project. It is a protein-ligand docking platform available at <http://www.swissdock.ch> with a friendly user interface, which makes the tool more accessible for non-experts in molecular simulations and programming. Among the advantages of a web service, we can highlight that all docking processing and the visualization of the results occur on the server-side, which does not require great computational resources from the user. SwissDock works with the EADock DSS engine, which allows performing blind and local docking in four steps, involving evolutionary algorithm and energy calculation with CHARMM force field. The limitation of the tool is the restricted control over the simulation parameters. (GROSDIDIER; ZOETE; MICHIELIN, 2011). An example of the results page of the tool is presented in Figure S2[7].

1.2.4 ClusPro

ClusPro is a web server used for performing protein-protein docking, developed in 2004, but with constant updates to the present day. It is available at <https://cluspro.org>. The system only requires two protein structures in PDB format, and the user may or may not change the default settings of the simulation, which is completed within four hours. The docking calculation is divided into three steps. The first step consists of rigid-body docking based on Fast Fourier Transform (FFT) correlation to generate more than one billion possible conformations.

[7] The supplementary material of this paper is not in the content of this dissertation

From this set, the one thousand most stable poses are selected for the second clustering step to identify the most likely conformations. From there, thirty poses are selected to go through the last energy refinement step, using the CHARMM force field in the energy parameterization. Finally, the algorithm returns ten models that represent the best fit between the proteins. The scoring function used in the FFT correlation is based on structure interactions, in addition to energy terms, such as electrostatics and desolvation (KOZAKOV et al., 2017). ClusPro is a recognized software by CAPRI, but its scoring function based on rigid body method is not as accurate as that of flexible methods (DESTA et al., 2020).

1.2.5 pepATTRACT

pepATTRACT is a blind peptide-protein docking web service that is available at <https://bioserv.rpbs.univ-paris-diderot.fr/services/pepATTRACT/>. It does not require any information about the protein binding site and the peptide structure. pepATTRACT brings together several algorithms already developed by the research group. It works, initially, by performing a rigid docking, with the three most probable conformations of the peptide, for the putative determination of the binding site, using the ATTRACT algorithm. Then, a refinement with local docking is performed in two steps, which allows great peptide flexibility to explore as many conformations as possible. First, iATTRACT is used, and then a molecular dynamics stage is executed with AMBER. This refinement takes approximately 18 hours to run, and the software does not integrate with GPU, even when it is possible, which we consider a limitation (VRIES et al., 2017).

1.2.6 HDOCK

The HDOCK is a web server designed to perform protein-protein docking based on template modeling and structure prediction. The great advantage of this tool is that it can receive the three-dimensional structure in PDB format or the amino acid sequence as input. If the sequence is provided, HDOCK predicts the structure using MODELLER. From there, the hybrid docking strategy predicts the interaction between the chains, and experimental information can be incorporated to calculate the interaction and the scoring function. In the end, the top 100 poses are available for download, and the top 10 can be visualized on the web page. HDOCK is one of the first docking platforms to allow protein–RNA/DNA docking, which has its intrinsic scoring function. High dependency of high-quality homologous complex templates to execute hybrid docking is an issue faced by the platform (YAN et al., 2020). It can be accessed on <http://hdock.phys.hust.edu.cn/>.

1.2.7 ZDOCK

ZDOCK Server is a user-friendly protein-protein docking web platform that allows editing the scoring function parameters and displaying input and output visualization. The process is divided into three steps: (1) the input, which can be carried out from an upload of the structure or the PDB id; (2) selection of contacting or blocking residues; and (3) calculation and results for visualization with JMol. However, the platform allows the user to download the results and use them in other analysis software. ZDOCK runtime is about 11.5 min and succeeds with CAPRI 'Acceptable' criteria. Eventually, the authors of the tool intend to develop clustering algorithms for the results and post-processing analysis on the webserver. ZDOCK is available on <http://zdock.umassmed.edu> (PIERCE et al., 2014).

1.2.8 HADDOCK

HADDOCK (High Ambiguity Driven biomolecular DOCKing) is an open web server available on <https://wenmr.science.uu.nl/haddock2.4/> in which the user can perform protein-protein and protein-ligand docking and use a variety of molecules as input, including cyclic peptides and glycans. The platform allows docking calculating with two molecules up to 20. It has many interfaces, varying from Easy to Guru tier, that allow different levels of user control of the simulation parameters. The Easy tier is the most basic, which explains its limitations. The user can upload two structures, define active/inactive residues and submit to the platform by applying the default parameters for docking calculation. The full access and control can be obtained by the Guru tier, in which the user can modify over 500 parameters (HONORATO et al., 2021; van Zundert et al., 2016). The result page, as observed on Figure S3[8], brings the summary of the complex pose clusters in several energy scores and the graphical comparison between clusters, using Bokeh Python library.

1.3 Molecular dynamics

Molecular dynamics is a technique of computational calculation in which Newton's classical laws of motion are extrapolated to molecular systems. To do so, the behavior of molecules is described by force fields that prescribe the spatial and energetic pattern temporally. In addition to this approximation, a cutoff is also established to determine the maximum interaction distance between two atoms to reduce the number of operations that the algorithm needs to perform (GOEL et al., 2015). These considerations make molecular dynamics a very efficient method for many applications, such as protein folding prediction, complex system stability, and inhibition potential of proteins with small ligands.

[8] The supplementary material of this paper is not in the content of this dissertation

There are three types of molecular dynamics simulation: the conventional, which is used for several studies, such as protein folding, complex stability analysis, and structure refinement; the QM/MM (quantum mechanics/molecular mechanics) simulations, which can be used to simulate chemical reactions; and metadynamics, in which the free energy of systems can be predicted (CASE et al., 2005b). In all simulation cases, we employ algorithms for calculating the integration of the motion and energy equations, to obtain the vectors describing velocity, position, and force for all atoms in the system. These vectors are defined at small time intervals throughout the simulation, which allow obtaining the system's physical properties at the time interval considered in the full simulation (VLACHAKIS et al., 2014). Before any of these simulations, there is a preparation of the molecules, in which minimization steps are applied to ensure the quality and stability of the structures.

A critical factor for these simulations is the setting of equations of motion and energy, which are directly related to the force field selected for its parameterization. Several force fields have been developed with different resolutions, either expliciting all atoms or not or even working with entire groups. The choice for one force field or another should take into account the properties to be measured during the simulation, the level of detail of the interactions, atomic or interchain scale, for example, and the class of biomolecules. Neglecting to select the correct force field may result in low accuracy and long delays, as problems in the simulation are usually detected only when the simulation is finished, which can take up to months (RINIKER, 2018; DAUBER-OSGUTHORPE; HAGLER, 2018).

A force field is defined by the Equation 1.1, which represents the bonded, nonbonded interactions and specific parameters of the force field. Bonded interactions include intramolecular bonds, valence angles, rotation, and dihedrals. Nonbonded interactions involve intermolecular bonds, such as electrostatic, dispersion, and Pauli exclusion (GUVENCH; MACKERELL, 2008).

$$\begin{aligned} E_{total} = & \sum_{bonds} k_b(b-b_0)^2 + \sum_{angles} k_\theta(\theta-\theta_0)^2 + \sum_{dihedrals} k_\chi[1+cos(n\chi-\sigma)] + \\ & + \sum_{nonbonded} \left(\varepsilon_{ij} \left[\left(\frac{R_{min,ij}}{r_{ij}} \right)^{12} - 2 * \left(\frac{R_{min,ij}}{r_{ij}} \right)^{6} \right] + \frac{q_i q_j}{r_{ij}} \right) + E_{forcefield} \end{aligned} \tag{1.1}$$

CHARMM, Amber, GROMOS, and OPLS-AA, all of them developed by academic research groups, are the most common force fields used in molecular dynamics simulations. They are all present in the major molecular dynamics software, except for GROMOS, which was specially developed for GROMACS. Even though they were designed focusing on proteins, some modifications were added, and other classes can be described as well, such as nucleic acids, lipids, carbohydrates, and small molecules. The main differences between these force fields are related to nonbonded interaction and the "improper" dihedrals (this characteristic is

specific for proteins, but it is estimated for the other biomolecular classes). For this reason, these parameters should be the criteria employed to decide which force field to use (GUVENCH; MACKERELL, 2008; MONTICELLI; TIELEMAN, 2012).

We discuss the main tools used in academia for performing molecular dynamics simulations, which are summarized in Table 2. In most of them, there is a great diversity of force fields available, and they allow more complex simulations, such as metadynamics, for example. Molecular dynamics is a technique that has been widely used to better understand the formation of complexes between macromolecules and ligands, as well as broader phenomena on a cellular scale. Both the software that allows this type of calculation and the force fields used are constantly evolving to bring the simulations closer to reality. Another aspect that has been incremented in software is the diversification of the simulations that can be performed, such as metadynamics and QM/MM, which expands the events that can be studied computationally. In addition, molecular dynamics simulations are computationally expensive. Therefore, powerful GPU processors are used to accelerate and parallelizing calculations of complex systems, and recent versions of this kind of software include GPU setup (YANG; WANG; CHEN, 2007; KONDRATYUK et al., 2021).

1.3.1 Amber

Amber (CASE et al., 2005a; SALOMON-FERRER; CASE; WALKER, 2013) is a tool for performing molecular dynamics simulations created in the 1970s, which has been maintained by the development community to this day. This tool is a set of several programs that work together to configure, execute and analyze molecular dynamics simulations. To execute the simulation, Amber has several molecular dynamics force fields and simulation software packages, including source code and demo files.

Currently, Amber consists of three different molecular dynamics simulation tools: Sander, Pmemd, and Pmemd.cuda. In summary, Sander has been the primary platform for Amber's computing and development, to explore new features. Pmemd and Pmemd.cuda are more focused on maximizing performance, thus working to implement Sander features on high-performance architectures. Several force fields are supported by Amber, including pairwise amino and nucleic acid variants, fixed-charge protein force fields, CHARMM force fields, Glycam series of force fields, polarizable force fields, and AMOEBA force field.

The tool is available free of charge via the following web address: <https://ambermd.org/>. The web page contains a lot of important information about Amber, such as reference manuals, a description of force fields, and several tutorials, including installation and running information. It is also possible to find demonstration materials for educators on the website, where several examples and commented codes are made available, to provide an overview of relevant algorithms.

Table 2 – Summary of molecular dynamics tools.

Name	URL	Features	Limitations	Reference
Amber	<https://ambermd.org/>	Free of charge Standalone software Usually applied in biomolecular simulations Several force fields available Many available tutorials	Only available for Linux	(CASE et al., 2005a; SALOMON-FERRER; CASE; WALKER, 2013)
LAMMPS	<https://www.lammps.org/>	Free of charge Standalone software Available for Windows, Linux and macOS Run in parallel or single processor MPI and OpenMP parallel support	Does not allow interactively visualization Lacks output data plotting resources	(PLIMPTON, 1995)
GROMACS	<https://www.gromacs.org/>	Free of charge Standalone software Several force fields available GROMOS force field Many available tutorials	Only available for Linux	(SPOEL et al., 2005a; ABRAHAM et al., 2015)
CHARMM	<https://www.charmm.org/>	Free of charge Standalone software Available for Linux and macOS MPI and OpenMP parallel support Usually applied in biomolecular simulations	Requires licence for commercial use	(BROOKS et al., 2009)
NAMD	<https://www.ks.uiuc.edu/Research/namd/>	Free of charge Standalone software Available for Windows, Linux and macOS Many available tutorials Extensive documentation and training section	Does not support non-equilibrium MD Does not suppoort rigid bond between heavy atoms	(PHILLIPS et al., 2020)
Desmond	<https://www.schrodinger.com/products/desmond>	Commercial tool Standalone software Source code available for non-commercial use State-of-the-art GPU acceleration technology Focused on scalability	Only available for Linux	(BOWERS et al., 2006)
OpenMD	<https://openmd.org/>	Free of charge Open source project Standalone software Available for Windows, Linux and macOS MPI parallel support	-	(LOUDEN et al., 2017)
ORAC	<http://www1.chim.unifi.it/orac/>	Free of charge Standalone software Scaling parallel support via OpenMP Focused on simulations at atomistic level complex systems	Only available for UNIX operating systems	(PROCACCI, 2016; MARSILI et al., 2010; PROCACCI et al., 1997)
AMMP VE	<https://www.ddl.unimi.it/cms/index.php?Software_projects:AMMP_VE>	Free of charge Standalone software Available for Windows and Linux Can be embedded in other programs	Lacks of support information and documentation	(PROGRAM, 2012)
ACEMD	<https://www.acellera.com/products/molecular-dynamics-software-gpu-acemd/>	Free of charge Standalone software Parallel support Designed for GPUs	Needs a licence to use all its resources Only available for Linux	(HARVEY; GIUPPONI; FABRITIIS, 2009)
Tinker	<https://dasher.wustl.edu/tinker/>	Free of charge Standalone software Available for Windows, Linux, macOS and Android Support very large molecular systems	Requires licence for commercial use Does not focus on free energy calculations	(RACKERS et al., 2018)
Chemsol	<https://laetro.usc.edu/doc/chemsol/cs-1.0/index.html>	Free of charge Webserver Works with aqueous solutions	Lacks of support information and documentation	(FLORIÁN; WARSHEL, 1999)
Abalone	<http://www.biomolecular-modeling.com/Abalone/index.html>	Free of charge Standalone software Several force fields available Education section available	Only available for Windows	(SOFTWARE,)
YASARA Dynamics	<http://www.yasara.org/products.htm>	Commercial tool Standalone software Available for Windows, Linux, macOS and Android Run simulations in aqueos solutions Shows simulations in real-time Allows screen interactions during simulations	-	(KRIEGER; VRIEND, 2014a; KRIEGER; VRIEND, 2015)

1.3.2 GROMACS

GROMACS, a software commonly used in the chemical area, is another example of a molecular dynamics simulation tool (SPOEL et al., 2005a; ABRAHAM et al., 2015). It was designed to work with biochemical molecules, such as protein, nucleic acids, and lipids, but it has also been used in non-biological systems, such as polymers, due to its efficiency in calculating non-bonded interactions, which generally account for most of the simulation. GROMACS comes with a variety of built-in force fields for molecules, including GROMOS53a6, OPLS, Encad, OPLS-AA/L, Amber99SB-ILDN, CHARMM27, among others. This tool was created to provide efficient modeling and has been developed through open-source and free software development, with a codebase generated through the sharing of infrastructure and contributions from several researchers.

The ease of use of the tool is a strength of GROMACS. First of all, it does not use

any scripting language, just provides a simple interface. The simulations generated by the tool can be monitored as they are carried out, and the remaining time to finish the simulation is provided. In addition, GROMACS allows the users to select the accuracy and comprises a large set of tools for trajectory analysis, using lossy compression that can offer a compact way to store trajectory data. GROMACS is a free-of-charge tool that is available on its website through the link: <https://www.gromacs.org/>. At this address, a variety of information is provided, including features of the tool, funding, and support information. The website contains a link to extensive documentation of GROMACS, which supplies download addresses and release notes for various of its versions.

1.3.3 CHARMM

CHARMM (BROOKS et al., 2009), an acronym for Chemistry at Harvard Macromolecular Mechanics, is a tool for molecular simulation with wide application in multi-particle systems with a set of energy functions, improved sampling methods, and implicit solvent models. It was developed for the study of biomolecules, such as peptides, proteins, carbohydrates, nucleic acids, small molecule ligands, and lipids, which can be used to solve problems of multi-particle systems.

The software provides computational apparatus, including molecular minimization, path sampling methods, dynamics, and free energy estimators. The calculations performed with the tool can be used for various functions and energy models. CHARMM contains a variety of analysis and model creation tools and can deliver high performance on a large group of platforms, including those using parallel computing and GPU.

The tool is available for free for academic use. Information about CHARMM is available at <https://www.charmm.org/>. The website provides a forum for discussion of various topics involving the tool, such as installation issues, discussion of parameters, and questions about molecular dynamics and chemistry. In addition, there is a section dedicated to the academic use of CHARMM, which covers aspects of the tool along with external packages that can be used. The user also has access to the CHARMM documentation, which contains a complete list and description of its modules.

1.3.4 NAMD

NAMD is another example of a molecular dynamics simulation program (PHILLIPS et al., 2020). This tool features an efficient parallel implementation and offers scalable system performance, running on up to hundreds of thousands of processing cores. It is used to simulate large systems with millions of atoms, whose purpose is to enable biomedical research through practical supercomputing. NAMD is a tool that aggregates several state-of-the-art algorithms to perform simulations in thermodynamic sets, including CHARMM, AMBER, OPLS, and

GROMOS biomolecular force fields. Like many other molecular dynamics simulation tools, NAMD is available as open-source, with no commercial use.

NAMD supports classical molecular dynamics simulations in explicit solvent, with periodic boundary conditions, in addition to coarse-grained models, in implicit solvent, with unrelated periodic boundary conditions (non-periodic and semi-periodic). Some characteristics can also be considered to attach external forces in the molecular simulations, similar to the Colvars module (where the user can set in calculations some variables as control parameters), flexible adaptation of structures to electron density maps, and methods for sampling acceleration.

Users can find NAMD at <https://www.ks.uiuc.edu/Research/namd/>, a website that contains information about the tool, including an overview, news, and announcements. In addition, it is possible to find a training session, with several development workshops. Besides, together with the documentation section, the user can obtain support for using the tool.

1.3.5 OpenMD

LOUDEN et al. (2017) proposed OpenMD, an open-source molecular dynamics engine that allows working with simulations of proteins, lipids, nanoparticles, transition metals, and several other systems, through the use of atoms with orientational degrees of freedom. OpenMD was implemented with a parallel computing approach, the Message Passing Interface (MPI), which makes it an efficient method to develop molecular dynamics simulations. The tool provides several trajectory analyses and utility programs. In addition, OpenMD supports various force fields and even allows users to define their own, using empirical energy functions.

OpenMD is available for free through the website <https://openmd.org/>, with plenty of information about the tool, including news and examples, as well as its related source code. Users can find a Documentation section where a manual and documentation code of the OpenMD are available. In addition, the website has a Community section, with discussions on various subjects involving OpenMD and molecular dynamics, both for common users and developers. OpenMD is available for Linux, Windows, and macOS operating systems.

1.3.6 ORAC

ORAC (PROCACCI, 2016; MARSILI et al., 2010; PROCACCI et al., 1997) is an open-access molecular dynamics simulation method that works with several complex molecular systems at the atomic level. Its main feature is the use of state-of-the-art molecular dynamics algorithms along with the flexibility to manipulate the most diverse types and sizes of molecules. In addition, ORAC has algorithms for biological systems with periodic boundary conditions, can perform simulations for computing electrostatic interactions, and allows the simulation of molecular systems in different thermodynamic ensembles.

ORAC was mainly developed using the Fortran language and is compatible with most compilers currently available. However, it is only available for UNIX-based systems. The recent versions present an implementation of parallel architectures, using MPI (Message Passing Interface) and OpenMP, which guarantees a faster and more efficient execution. In addition, some features were implemented, such as support for the fast switching double annihilation method (FS-DAM) (NERATTINI; CHELLI; PROCACCI, 2016; PROCACCI; CARDELLI, 2014), for the calculation of binding free energies in drug-receptor systems.

At the web address <http://www1.chim.unifi.it/orac/>, users can find information about ORAC, including links and instructions for downloading the current version as well as older releases. An extensive manual, with information about ORAC inputs, interactions, and simulations, is available to users to provide support for using the tool. In addition, from the ORAC website, there is a link to PrimaDORAC (PROCACCI, 2017), a web interface that generates parameter files and topology for molecular dynamics from organic molecule coordinates.

1.3.7 Tinker

(RACKERS et al., 2018) proposed the Tinker method, a molecular mechanics and dynamics package for molecular modeling. Several parameter sets are found in the tool, including Amber, CHARMM, OPLS, and AMOEBA, among other force fields. Tinker has a series of algorithms for the most diverse applications, such as free energy calculations, continuum solvation treatments, vibrational analysis, reaction field treatment of long-range electrostatics, fitting charge, multipole, and polarization models, among many others.

The method, implemented in Fortran, uses a module concept and, together with dynamic memory allocation, allows working with very large molecular systems. To further improve its performance, Tinker uses a parallel approach, through OpenMP, in various aspects of its code. A new package, Tinker9 (WANG, 2021), was recently developed to provide performance on single NVIDIA GPU-based systems.

Tinker is available free of charge for academic and non-profit use through the website <https://dasher.wustl.edu/tinker/>. Before use, the tool must be downloaded and installed locally, on any machine with Windows, Linux, or macOS operating systems. Users will find many resources on the website, such as user and installation guides, Force Field Parameter Sets, and the source code of the tool.

1.3.8 YASARA Dynamics

YASARA Dynamics (KRIEGER; VRIEND, 2014a; KRIEGER; VRIEND, 2015) is software for molecular dynamics simulations, which is part of the YASARA tool package. This package consists of four stages: View (stage 1), Model (stage 2), Dynamics (stage 3), and Struc-

ture (stage 4). Each stage contains all the features of the previous stages, plus specific additional functions of a given application.

Stage 1 of the YASARA package (View) consists of a set of features for exploring macromolecular structures iteratively. Stage 2, YASARA Model, provides resources to explore, analyze and model macromolecules in a production environment. Stage 3, YASARA Dynamics, contains all the functionality of the previous stages and also has support for molecular simulations. Some features of YASARA Dynamics can be mentioned, such as treatment of long-range electrostatic interactions, parallel execution of the simulations, and calculation of energies and binding energies. Furthermore, an interesting feature of the tool is that it allows users to view the molecular simulation in real-time, unlike some methods that work like a black box. YASARA Dynamics makes available the use of its force fields, NOVA and YAMBER, besides allowing users to employ other known force fields such as AMBER. Stage 4, YASARA Structure, in addition to the molecular dynamics package (and all previous stages), contains functionalities for the identification and validation of macromolecular structures.

YASARA Dynamics, along with all other stages of the package, can be found via the web address: <http://www.yasara.org/products.htm>. The tool can be used on Linux, macOS, Android, and Windows operating systems. Only stage 1, YASARA View, is provided free of charge. It is necessary to purchase a license to use the other stages, including YASARA Dynamics. On the website, users can also find much more information about the YASARA package, as well as news and user support resources.

1.4 Molecular visualization

The advance of new experimental techniques and structure determination methods for obtaining atomic positions in three-dimensional space has contributed significantly to studies related to the functions of proteins. An increased number of 3D structures deposited in the PDB (Protein Data Bank) database enabled new studies (CONSORTIUM, 2018). Many of the so-called molecule viewers are freely available, and open-source enables open science aiming to learn about structures and physical-chemical phenomenons in proteins.

Molecule viewers are versatile and can be integrated with web platforms, mobile or stand-alone systems in studies involving structural similarities, mutations, drug design, and molecular docking, among others. In addition, graphical user interface (GUI) and simulations are critical for students entering the field of bioinformatics and computational biology (BROWN; BEVAN, 2017). The understanding of molecular structures is becoming more democratized as a result of the many molecule visualization software systems made available. Currently, in addition to high-quality images, the developers of such tools are also concerned with usability and offer several new features.

This constant evolution brings significant gains, especially when it comes to studies involving non-static and complex systems with several components. The concern with the graphical interface also allows the extrapolation of structural data interpretation for researchers who do not have as much contact with lines of code, thus greatly enriching research discussions. Nowadays, many systems are available freely or under a commercial license. In the next subsections, we present a relevant set of 3D molecule viewers largely used by the scientific community. A summary of the 3D viewer tools described in this section is presented in Table 3.

Table 3 – Summary of the 3D viewer tools.

Name	URL	Features	Limitations	Reference
PyMOL	<http://www.pymol.org/>	open-source creating movies cross-platform support different input file formats use command lines high quality images	not on easy-of-use for new users programming-knowledge for scripts no official support	(Schrödinger, LLC, 2015)
VMD	<https://www.ks.uiuc.edu/Research/vmd/>	open-source cross-platform user support support different input file formats GPU acceleration creating movies	does not work well representations of cyclic proteins bug in the old "cartoon" representation	(HUMPHREY; DALKE; SCHULTEN, 1996a)
Chimera	<https://www.cgl.ucsf.edu/chimera/>	no-commercial free cross-platform creating movies support other tools	no longer under active development no official support	(EF et al., 2004)
NGL Viewer	<http://proteinformatics.charite.de/ngl>	open-source web application allows the use of plugins	support PDB, mmCIF, and GRO input file default viewer PDB not create movies	(ROSE et al., 2018)
3Dmol.js	<https://3dmol.csb.pitt.edu/>	open-source web application JavaScript language support different input file formats	used in web project not create movies	(REGO; KOES, 2014)
EzMol	<http://www.sbg.bio.ic.ac.uk/ezmol/>	no-commercial free web application high quality images easy-of-use for new users	support PDB file only not create movies	(REYNOLDS; ISLAM; STERNBERG, 2018)
Schrödinger	<https://www.schrodinger.com>	commercial integrated computer platform version free for academic purposes support other tools user support high quality images	general purpose viewer focused on drug design	(WANG et al., 2015)
MOE	<https://www.chemcomp.com/Products.htm>	commercial cross-platform structure database associated GPU acceleration high quality images user support	general purpose viewer	((MOE), 2019)
Jmol	<http://jmol.sourceforge.net/>	open-source cross-platform web and desktop application support different input file formats	Java language specific	(HERRÁEZ, 2006; HANSON et al., 2008)
GLmol	<https://www.glmol.com/>	open-source web application JavaScript and GLSL languages	used in web project not create movies	(NAKANE, 2014)
DS Visualizer	<https://www.3ds.com/products-services/biovia/>	no-commercial free integrated computer platform cross-platform	general purpose viewer	(VISUALIZER, 2020))
Crystal Studio	<http://www.crystal0studio.com/products.php>	commercial high quality images structure database associated support other tools	general purpose viewer	(RAKOVAN, 2018)
CueMol	<http://www.cuemol.org>	open-source web and desktop application cross-platform high quality images support other tools	no official support	(ISHITANI; NAKANE, 2014)
YASARA View	<http://www.yasara.org/products.htm#view>	no-commercial free integrated computer platform cross-platform high quality images GPU acceleration support different input file formats	only standard ASCII characters	(KRIEGER; VRIEND, 2014b)
TextMol	<https://cvcweb.oden.utexas.edu/cvcwp/software/texmol/>	open-source cross-platform GPU acceleration high quality images	PDB and PQR formats only no official support	(BAJAJ et al., 2004)

1.4.1 PyMOL

PyMOL is the most classical software when it comes to biomolecule visualization. It is open-source software for educational and commercial purposes that is distributed by Schrödinger, Inc. Written in C, C++, and Python, PyMOL can be installed on several operating systems, such as Unix and Windows, and through the tool's official website (<http://www.pymol.org/>), the user can obtain its license (Schrödinger, LLC, 2015). PyMOL can be used to produce both images and movies of biomolecular systems in various types of representations, as described in Figure S4[9]. In addition, several plugins have been developed that can be installed and add several more features to the application, including protein and ligand modeling, PL docking, Gromacs-based molecular dynamics simulations, Dehydron (plugin that displays the hydrogen interactions of the backbone that are unprotected from water attacks, which compromises protein folding) and even QM/MM calculation (YUAN; CHAN; HU, 2017).

The software consists, basically, of three interfaces: (a) Graphic User Interface (GUI) with menu bars, which allows the user to achieve several features of the visualization; (b) mouse controls that allow precise selections of molecule regions and movements for better structure positioning; and (c) command lines that can be used with no need of the two other options. The last one is a powerful way to modify the structure and achieve additional function combinations (PAN; ALLER, 2015). Due to the many uses of PyMOL, a support community has been developed, the PyMOLWiki (<http://www.pymolwiki.org>). There are tutorials, descriptions of features and plugins, image-making protocols, and answers to users' most frequently asked questions. The page is constantly updated with new information and features of the software.

1.4.2 VMD

VMD (Visual Molecular Dynamics) (HUMPHREY; DALKE; SCHULTEN, 1996a) is a software developed by the Theoretical and Computational Biophysics Group from the University of Illinois to allow visualization, modeling, and analysis of biomolecules, both static and in a simulated trajectory, using a built-in scripting system and three-dimensional graphics. A sample of the VMD interface is represented in Figure S5[10]. We can mention some technical features of the tool, such as support for various computing platforms (MacOS X, Unix, or Windows), efficient management of memory usage, use of multicore processing, and GPU acceleration. Furthermore, the VMD has no limit on the number of atoms, residues, molecules, or trajectory frames, being limited only to the amount of memory available in the machine. The visualization features of the tool include multiple coloring and rendering methods, support for over 60 different file formats, support for using other visualization and simulation tools such, as NAMD (PHILLIPS et al., 2005a), as well as extensions capable of allowing users to create

[9] The supplementary material of this paper is not in the content of this dissertation
[10] The supplementary material of this paper is not in the content of this dissertation

their routines for molecular analysis.

VMD is distributed free of charge through its website <https://www.ks.uiuc.edu/Research/vmd/>, which also includes its source code. A range of information about the tool is available on the website, including news, announcements, development topics, and related publications. There is also a user support section, where the VMD documentation is available, with user guides, bug lists, and FAQs. A series of tutorials and manuals are available to users and include a very detailed walkthrough with descriptions and images of various problems in visualizing molecular structures. Examples of the use of VMD are also found in conjunction with other chemical or biological tools, such as AMBER (CASE et al., 2005a; SALOMON-FERRER; CASE; WALKER, 2013), which works with molecular dynamics, and APBS (BAKER et al., 2001), which performs electrostatic calculations.

1.4.3 Chimera

Chimera (EF et al., 2004) is a molecular visualization tool written in Python, which is designed to be simple to use on several platforms and capable to produce classical graphics. The tool can be divided into two parts: core and extensions. The first one consists of a series of basic services and molecular visualizations, in addition to ensuring that the extensions can run robustly. Extensions allow for higher-level functionality of Chimera, are loaded via the tool's menu, and used when users access it, on-demand. We can mention some of the extensions used by Chimera such as Multiscale (useful to explore large molecular assemblies), Multalign Viewer (useful to align structures), ViewDock (integrated to DOCK, it helps to perform the screening of ligands), Movie (which allows visualization of molecular dynamics trajectories) and also Volume Viewer (to present three-dimensional information).

The authors developed a new version of the tool, ChimeraX (PETTERSEN et al., 2021), to work with larger structures while maintaining high performance. This new version contains some new features and advantages, such as high-performance manipulation and rendering of a large number of atoms, interactive ambient-occlusion lighting, and new windows, panels, and bars that allow users to navigate between the features of the tool more simply and clearly, plus a platform for virtual reality, made for Steam VR systems.

The tool is available for free through the link <https://www.cgl.ucsf.edu/chimera/>. On the website, users find a wealth of information about using Chimera, including its documentation, a download section, publications, datasets, and related software. In the documentation section, users have at their disposal guides and tutorials about the tool, visualization, and analysis videos, release notes, and a series of commands that can be used in Chimera. In addition, a gallery section is available, with an extensive collection of images and animations that can be made from the software, including RNA base models, protein interfaces, binding footprints, and wobble motion.

1.4.4 NGL Viewer

Proposed by (ROSE et al., 2018), the NGL Viewer is an open-source web application that uses WebGL to allow online molecular manipulation and visualization. The tool is one of the default viewers for Protein Data Bank (PDB) 3D view [11] (BERMAN et al., 2000a), which can be integrated into Jupyter Notebooks. The NGL Viewer was developed focused on memory efficiency, which was solved by parsing the input files quicker using the binary Macromolecular Transmission Format (MMTF). This advantage of the tool determined its option to be used in the PDB and has registered from small structures to representations of entire viruses, which demands a lot of memory space (ROSE et al., 2016).

The NGL Viewer can load and display the main structural files type (PDB, mmCIF, and GRO) in a diverse way, such as ball & stick, cartoon, and surfaces. Moreover, it also loads molecular dynamics trajectories for analysis (ROSE et al., 2018; ROSE; HILDEBRAND, 2015). On the tool's web page (<http://proteinformatics.charite.de/ngl>), users can use various features of the viewer in their browser, without the need to install local software. The tool has menus and icons that allow users to view and manipulate structures easily and make use of plugins. It is also possible to find a documentation section, where users find information about the use of the tool and descriptions of objects, components, and representations used in the visualization of molecules with NGL.

1.4.5 3Dmol.js

Proposed by (REGO; KOES, 2014), 3Dmol.js is a tool for molecular visualization in an online environment, which is an object-oriented library, WebGL based on JavaScript language. Among the features offered by the tool, we can highlight the different ways of representing surfaces, atoms, and chains, as well as the detail on demand when users hover the mouse over some part of the structure. We can also mention some other relevant features, such as support for various file formats (such as pdb, mol2, xyz, sdf, and cube), molecular surface parallel computation, as well as various structure visualization styles, such as cartoon, stick, line, and sphere. Besides the implementation in JavaScript, 3Dmol was also implemented in Python with some function reductions and can be used in Jupyter Notebook.

3Dmol.js can be accessed at <https://3dmol.csb.pitt.edu/>, where users can also download the tool's code. On the website, users can find examples of interactive visualization of the tool, such as protein structures and atomic characteristics. A notable facility concerning 3Dmol.js is that all the visualizations (including its several variations regarding style, color, and other characteristics) can be presented in the web browser itself, with no need to install local software. It is also possible to use library features through the Jupyter notebook, in addition to finding information about the development of web applications using a featured API. 3Dmol.js

[11] <https://www.rcsb.org/3d-view/>

provides documentation, FAQ, and contact section, besides a teaching section, which allows students to learn about molecular structures through an active learning environment.

1.4.6 Jmol

JMol (HERRÁEZ, 2006; HANSON et al., 2008) is one of the oldest open-source viewers for chemical structures in 3D used by students and researchers around the world. It started being developed in the Open-Science project by Dan Gezelter and is based on Java language, which turns it compatible with different operating systems, such as Windows, Linux, and Mac. Its modules have different purposes: Jmol application (desktop), JSmol (web pages), and JmolViewer (Java applications). Jmol supports a wide variety of file formats for input, automatically identified: mol, cif, pdb, xyz, and some specific ones, such as ADF (Amsterdam Density Functional), Gaussian, and SDF (V2000 and V3000). For output, the user can download the generated image in jpeg, png, or ppm or export it into a pdf file.

Jmol can represent the whole periodic table and it accepts a variety of atoms representations, such as van der Waals radius or absolute size, ionic radius, and dots (spheres, stars, tetrahedra, or octahedra). For biological macromolecules, the tool allows the summarization of the structure in classic representation models, such as backbone, trace, ribbons, strands, and cartoon. Jmol is available on <http://jmol.sourceforge.net/>, where we can also find the software documentation.

1.4.7 DS Visualizer

Discovery Studio Visualizer or DS Visualizer is part of a suite of programs able to simulate and share analyses of small and macromolecules. It is developed and distributed by Dassault Systemes BIOVIA (<https://www.3ds.com/products-services/biovia/>). It has many collaborations from the academic community, which makes it relevant for scientific research.

It is free software and provides a comprehensive collection of features to capture the specific nuances of research. It supports scientific research and uses known software and algorithms developed originally by the scientific community, such as CHARMM (molecular dynamic), MODELLER (3D modelling), DELPHI (electrostatic potential), ZDOCK (molecular docking), and DMol3 (electronic property).

In addition, the program suite is used for simulations and calculations, Ligand Design, Pharmacophore modeling, Structure-based Design, Macromolecule engineering, QSAR (quantitative structure-activity relationship), ADME (proprieties pharmacokinetics and pharmacology for "absorption, distribution, metabolism, and excretion) and more. For more details of features, access (VISUALIZER, 2020).

1.4.8 YASARA View

YASARA View is a potent and free 3D viewer (<http://www.yasara.org/products.htm#view>) used to explore macromolecular structure interactively. Its architecture has an innovative engine for high-efficiency graphics and computation on modern GPUs enabling many structures to be loaded simultaneously. Therefore, it is possible to produce publication-quality ray-traced images including labels. Besides, it is possible to integrate programs or scripts in Python language to run the viewer.

Its important features include a cross-platform and easy installation (Windows, Linux, and macOS in the same directory, run directly from a USB stick), support for over 70 molecular file formats, download of PDB files from the RCSB, measurement (distances, angles, dihedrals), alignment of multiple proteins based on their structure or sequence, using a variety of methods, and Parallel (orthographic) and perspective projection (KRIEGER; VRIEND, 2014b). For more details, access <http://www.yasara.org>.

1.5 Structure prediction

The number of known protein sequences has been increasing exponentially in recent years, and about 150 million entries were deposited in the UniProtKB database (BATEMAN et al., 2020). However, sequence information alone is not enough to understand the protein function, and protein structure information is crucial for this task. There is a gap between the growth rates of the sequential and structural information due to intrinsic difficulties and the costly nature of the experimental determination. For example, the Protein Data Bank (BERMAN et al., 2000b), a database of protein structures, contains about 170 thousand entries. Moreover, predicting a protein structure using the amino acid sequence as a starting point remains an unsolved problem in Bioinformatics.

Due to the importance of the theme, initiatives such as the Critical Assessment of Protein Structure Prediction (CASP) (MOULT et al., 2018), have promoted the development of the protein prediction methods and provided standard evaluation and definitions. The core of protein structure prediction is the assumption that the protein's native state is the one with the lowest free energy (ANFINSEN, 1972). Thus, the methods of protein structure prediction combine sampling of possible conformations with a ranking of these conformations through energy functions, aiming to find the lowest energy state.

Methods to predict protein structure can be labeled into two approaches: template-based and template-free (KUHLMAN; BRADLEY, 2019). When predicting a structure of a protein amino acid sequence, if there are related structures previously determined in PDB, template-based methods can use these structures as examples to model the target sequence. Otherwise, if related proteins are not found, template-free methods try to predict the protein structure directly

from the protein sequence, using energy functions combined with conformational sampling. Table 4 summarizes the structure prediction tools described in this section.

Table 4 – Summary of the structure prediction tools.

Name	URL	Features	Limitations	Reference
SWISS-MODEL	<https://swissmodel.expasy.org/>	Free of charge Webserver or standalone version User-friendly interface Build models at different levels of complexity	Only return a single result	(ARNOLD et al., 2006; WATERHOUSE et al., 2018)
Modeller	<https://salilab.org/modeller/>	Free for academic use Standalone program Can perform de novo modeling of loops Several models optimization	No GUI available	(WEBB; SALI, 2016; WEBB; SALI, 2017)
I-TASSER	<https://zhanggroup.org/I-TASSER/>	Free of charge (non-commercial use) Webserver Uses a multiple thread approach Give functional annotation information	Slower predictions compared to other methods	(ZHANG, 2008; YANG et al., 2015)
AlphaFold	<https://alphafold.ebi.ac.uk/>	Free for academic use Standalone program Machine Learning-based model	May not provide fold prediction in context-free scenarios	(JUMPER et al., 2021; ALQURAISHI, 2019)
LOMETS	<https://zhanggroup.org//LOMETS/>	Free of charge Webserver Give functional annotation information Focused on give quicker response	Does not refine threading models	(WU; ZHANG, 2007; ZHENG et al., 2019)
RaptorX	<http://raptorx.uchicago.edu/>	Free of charge Webserver or standalone version Deep learning-based method Inter-residue/inter-atom distance and orientation probability distribuition	Only provides contacts, not distances	(KÄLLBERG et al., 2012)
MPACK	<http://curie.utmb.edu/mpack/>	Free of charge Standalone program Multiple-templates used for modelling Support for MOLMOL visualization	Do not allow gaps in the secondary structure region	(SOMAN et al., 2001; SANNER et al., 1989; SCHAUMANN; BRAUN; WÜTHRICH, 1990)
CABS	<http://biocomp.chem.uw.edu.pl/CABSfold/>	Free of charge Webserver Outputs a corse grained trajectory of conformations Use Jmol representation	Small set of best scoring models Predictions can take up to 12 hours	(KOLIŃSKI et al., 2004)
Rosetta	<https://www.rosettacommons.org/software>	Free for academic use Webserver or standalone version Deep learning-based method Can model multi-chain complexes	No GUI available Not recommended to use on Windows	(LEMAN et al., 2020)
QUARK	<https://zhanggroup.org/QUARK/>	Free of charge (non-commercial use) Webserver Suitable for proteins that do not have homologous templates	Can not submit multiple jobs at once	(XU; ZHANG, 2012)
ModPipe	<https://salilab.org/modpipe/>	Free of charge Open-source software Standalone program Almost no manual intervention is needed	Do not calculate profile-profile alignments Only builds a single model per alingment	(ESWAR et al., 2003)

Beyond the protein structure, the design of ligands that interact with these proteins is crucial for predicting binding on active and allosteric regions. Therefore, some software has been developed to generate the best structural conformation of small molecules. These systems take molecular formulas, in the case of simple molecules, or more detailed linear representations, such as SMILES, and arrange the atoms three-dimensionally, respecting the geometry of the molecular orbitals. Some examples of this kind of tool are LigPrep (RELEASE, 2019), ChemSketch (ACD/CHEMSKETCH ADVANCED CHEMISTRY DEVELOPMENT, 2008), Avogadro (HANWELL et al., 2012), SCIGRESS (STEWART, 2009) and ChemDraw (PERKINELMER, 2012). Despite its relevance, this type of tool is still underdeveloped compared to tools specifically focused on biomolecules (HANWELL et al., 2012). We discuss these tools in the Supplementary Material[12].

1.5.1 Template-based modeling

This class of methods predicts 3D protein models using examples that share high sequence identity with a target protein. This approach is effective when the query shares at least 30% sequence identity with the protein examples. In general, the steps of standard template-based modeling include the selection of a suitable structural template, alignment of the query

[12] The supplementary material of this paper is not in the content of this dissertation

sequence to the template, and assembling the 3D protein model according to the target-template alignment (BORDOLI et al., 2009).

1.5.1.1 SWISS-MODEL

SWISS-MODEL (ARNOLD et al., 2006; WATERHOUSE et al., 2018) is a web-based modeling system used to build a protein structure model based on homology modeling. This method aims to build a 3D protein structure model using experimentally determined structures of templates that share an ancestry with the target sequence. First, a target protein is used as input, and its sequence serves as a query to find evolutionarily related proteins contained in the SWISS-MODEL template library SMTL (BIASINI et al., 2014). Then, templates are aligned against the target sequence to verify whether templates can represent conformational states or cover different locations of the query protein. Finally, a 3D protein model is generated for each selected template, transferring conserved atom coordinates according to the target-template alignment (WATERHOUSE et al., 2018). Figure S6[13] shows the SWISS-MODEL result page. Different template structures can be analyzed using the 3D viewer.

1.5.1.2 Modeller

Modeller also uses structures that share some sequence identity to model protein structures. The input to the Modeller is an alignment file of the query sequence with its respective templates, the atomic coordinates of the templates, and a script file (WEBB; SALI, 2016). Modeller uses satisfaction of spatial restraints to perform comparative protein structure modeling. These spatial limitations include stereochemical restraints, such as bond length and bond angles, restraints of distance and dihedral angles in the query sequence, and statistical knowledge for angle and inter-atomic distances, obtained from experimental protein structures. The spatial restraints are combined into an objective function that is optimized (WEBB; SALI, 2017).

1.5.1.3 I-TASSER

Another successful modeling method, I-TASSER (ZHANG, 2008; YANG et al., 2015), constructs 3D protein models by interactive treading assembly simulations. The target protein is compared against a set of representative protein structures aiming to find possible folds. The conformation in I-TASSER is represented by a feature of Cα atoms and side-chain centers of mass. The reassembling process is conducted by replica-exchange Monte Carlo simulations. The energy function elements of this method include information about secondary structure properties, backbone hydrogen bonds, and correlations based on the structural statistics from the PDB library. The lowest free-energy conformations undergo a refinement step to remove steric clashes and refine global topology (YANG et al., 2015).

[13] The supplementary material of this paper is not in the content of this dissertation

1.5.1.4 AlphaFold

In the recent CASP13, the participating methods achieve remarkable progress due to the adoption of deep learning and the exploration of co-evolutionary information in protein structure prediction. AlphaFold is a co-evolution method that achieves impressive results by CASP standards (ALQURAISHI, 2019). Co-evolution methods work by constructing multiple sequence alignments of proteins homologous to the target protein. Such a class of methods infers spatial proximity between residues by detecting mutations that occurred in the same evolutionary timeframes in response to other mutations. Another feature of AlphaFold came from applying convolutional networks, showing the potential of deep learning for protein structure prediction. In the coming years, the trend of improvements in template-based modeling is expected to continue, due to the increasing amount of available structures. More recently, AlphaFold2 (JUMPER et al., 2021) was developed, which brought improvements using deep learning architectures. AlphaFold2 achieved the highest accuracy in CASP14.

1.5.1.5 RaptorX

Proposed by (KÄLLBERG et al., 2012), RaptorX is a tool for predicting the structure and function of proteins, freely available for non-commercial use at <http://raptorx.uchicago.edu/>. The method was developed taking into account cases in which the target sequence is distant from the related protein templates. For such, a profile-entropy scoring method was used, which analyzes the number of non-redundant homologs and template structures. Conditional random fields (CRFs), which incorporate various biological signals, are also used. Finally, RaptorX implements a multi-template threading procedure to use multiple templates to model a single target sequence. The results obtained by RaptorX in CASP9 indicated values similar or even superior to those of other tools in the area.

1.5.2 Template-free modeling

It is not always possible to find experimental structures that share similarities with the target sequence. In this scenario, methods *ab initio* and *de novo* are used where there is no information about the target sequence. Due to the lack of a structural template, these methods require conformational sampling and ranking criteria by which near-native conformations can be chosen as candidates.

The basis for template-free approaches is the Anfinsen's thermodynamic hypothesis (ANFINSEN, 1972), which claims that the native state is the one with the lowest free energy. Thus, protein structure prediction methods combine sampling of alternative conformations and scoring functions to rank the sampled conformations and identify the state with the lowest energy. However, the size of conformational space that must be searched grows exponentially,

making the exhaustive search infeasible. Therefore, the exploration of the conformational space is guided by search algorithms that navigate through the energy landscape towards near-native conformations (KUHLMAN; BRADLEY, 2019).

1.5.2.1 Rosetta

Rosetta is an *ab initio* approach that generates protein models by assembling small fragments taken from the PDB library. For conformational search, multiple rounds of Monte Carlo minimization are carried out, where each move is evaluated by a scoring function, and the move is accepted based on the Metropolis criterion, according to the energy difference between the original and the new conformation (LEMAN et al., 2020). Rosetta's scoring function is a linear combination of terms, including physically based and statistically derived, which describes elements such as non-covalent interactions, residue solvation, and backbone torsion angles. Rosetta is a consolidated tool for protein model prediction that has been very successful for the free modeling targets in CASP experiments.

1.5.2.2 QUARK

QUARK is an *ab initio* protein structure prediction pipeline based on continuous fragment assembly that uses both physics-based energy and knowledge terms (XU; ZHANG, 2012). The target sequence is threaded through non-redundant high-resolution PDB structures, and at each residue position, structural fragments (1 to 20 residues) are generated (ZHANG et al., 2018). The scoring function of the threading is composed of torsion angle, solvent accessibility, and secondary structure matches. Replica-exchange Monte Carlo simulations are performed to assemble the fragments into complete models.

1.5.2.3 ModPipe

ModPipe (ESWAR et al., 2003) is an automated pipeline software used to calculate protein structure models from sequences. The modeling performed by the tool is based on four steps: fold assignment, sequence-structure alignment, model building, and model evaluation. First, for a protein sequence given as input, potential templates are found, and then the alignment between templates and the input sequence is calculated. ModPipe is then executed, which generates results for all templates found. Although the standard ModPipe protocol is a template-free approach, it is possible to add a wrapper to the tool, making it run in template-based model. ModPipe is available for free at <https://salilab.org/modpipe/>.

1.6 Mutation analysis

Missense mutations are a specific type of genetic mutation characterized by single base-pair substitution that results in the change of one amino acid by another. It is commonly observed in nature and can modify important properties in proteins, such as conformation, stability, flexibility, drug resistance and protein-small molecule or antibody-antigen affinities. These variants (mutations) often contribute to the emergence of serious diseases, such as cancer, which corroborates the relevance of their analysis for the emergence of new therapies.

The identification or prediction of their effects is an important ally in the prevention or treatment of these disorders (CELNIKER et al., 2013). Analysis of changes in the properties of the substituted amino acids contribute to the identification of potentially critical mutations that becomes a problem of enormous relevance. Furthermore, many researches aim to evaluate the effects of a point mutation with varied purposes, including the identification of new drugs or treatments.

The technological advance has enabled many computational approaches to be combined with research for the prediction and analysis of the effects of missense mutations on proteins. Several systems, using different approaches to variant analysis, are currently available and can be used by researchers around the world. In this section, we listed the main computational tools to help to understand mutations in various aspects.

A summary of the tools and methods described in this section is presented in Table 5.

1.6.1 EVmutation

EVmutation combines prediction and data visualization techniques to explore mutational effects on proteins. According to (HOPF et al., 2017), it is an unsupervised statistical method that models protein residue interaction, providing quantitative data on the effects of mutations. For this, it uses a "global probability" approach to analyze the interactions between all pairs of residues. In the work conducted by the authors, the results computed by the tool were compared with indicators from 34 experiments involving genetic variations, showing itself as a viable method that can be applied in different contexts and species of interest. However, the work also describes some limitations, such as trends coming from more recent families and with low diversity.

EVmutation presents itself as an interesting alternative for studies related to the analysis of mutations, making available for free through a web platform (<http://evmutation.org/>) a set of predictions for human proteins. Currently, 9,935 entries are provided, which can be accessed individually or distributed in the following datasets: mutation effects, evolutionary couplings and sequence alignments.

Table 5 – Summary of mutation analysis tools.

Name	URL	Features	Limitations	Reference
EVmutation	<http://evmutation.org/>	Free of charge Webserver available Unsupervised statistical method	User guide not available Predefined input data Trends coming from more recent families and with low diversity	(HOPF et al., 2017)
DynaMut	<http://biosig.unimelb.edu.au/dynamut2/>	Free of charge Webserver available Normal Mode Analysis (NMA) and machine learning method (Randon Forest)	Performance: lower Pearson coefficient than methods such as I-Mutant2	(RODRIGUES; PIRES; ASCHER, 2021)
PMut	<http://mmb.irbbarcelona.org/PMut>	Free of charge Webserver available Machine learning method (Randon Forest)	Performance: lower accuracy than methods such as PON-P2	(FERRER-COSTA et al., 2005)
SNPnexus	<http://www.snp-nexus.org/>	Free of charge Webserver available Functional annotation	Its use depends on filling out a form, so that the data is processed and the results sent to the user	(ULLAH et al., 2018)
Interactome INSIDER	<http://interactomeinsider.yulab.org/>	Free of charge Webserver available Machine learning method (Randon Forest)	User guide not available	(MEYER et al., 2018)
Metadome	<https://stuart.radboudumc.nl/metadome/>	Free of charge Webserver available Meta-domain based	Predefined input data - only those available in the own database	(WIEL et al., 2019)
mutfunc	<http://www.mutfunc.com/>	Free of charge Webserver available Provides an extensive database with pre-computed forecasts	Does not allow considering other types of mutation, such as variations in the number of copies Does not include many organisms with frequently studied mutations	(WAGIH et al., 2018)
ActiveDriverDB	<https://www.ActiveDriverDB.org>	Open-source Webserver available Based on information about post-translational modifications (PTMs)	The related study does not present comparisons with other tools with the same purpose	(KRASSOWSKI et al., 2017)
Condel	<http://bg.upf.edu/condel>	Free of charge Webserver available Consensus tool	Requires the user to login to access the features	(GONZALEZ-PEREZ; LÓPEZ-BIGAS, 2011)
PredictSNP	<https://loschmidt.chemi.muni.cz/predictsnp/>	Free of charge Webserver available Consensus tool	Allows you to use only sequences as input (FASTA format)	(BENDL et al., 2014)
I-Mutant2.0	<https://folding.biofold.org/i-mutant/i-mutant2.0.html>	Free of charge Webserver available Machine learning method (SVM)	User experience: does not provide a modern interface	(CAPRIOTTI; FARISELLI; CASADIO, 2005)
FATHMM	<http://fathmm.biocompute.org.uk/>	Free of charge Webserver available Markov Models-based	Little interactive features for presenting results	(SHIHAB et al., 2013)
SIFT	<https://sift.bii.a-star.edu.sg/>	Free of charge Webserver available Evolutionary conservation-based	User experience: does not provide a modern and intuitive interface	(NG; HENIKOFF, 2003)
Polyphen	<http://genetics.bwh.harvard.edu/pph2/>	Free of charge Webserver available Evolutionary conservation and structure-based	The related study does not present comparisons with other tools with the same purpose User experience: does not provide a modern interface	(Ramensky; Bork; Sunyaev, 2002)
AUTO-MUTE	<http://proteins.gmu.edu/automute>	Free of charge Webserver available Stand alone (version 2.0) Statistical methods of classification and regression	User experience: does not provide a modern and intuitive interface For the local version (2.0), you need to install additional packages	(MASSO; VAISMAN, 2010)
MDPPM	-	Molecular dynamics and classification methods (Randon Forest and KNN)	Performance: increased processing time due to intensive computational simulations	(MCCOY et al., 2021)

1.6.2 DynaMut2

DynaMut is a web tool that uses Normal Mode Analysis (NMA) to assess mutational effects on protein stability and dynamics. In (RODRIGUES; PIRES; ASCHER, 2021), NMA is described as a computational approach capable of exploring harmonic motions and providing information about structure-function relationships. Its application can also produce higher performance, considering the use of simplified structural representations that contribute to the reduction of computational cost. The DynaMut prediction model uses a Random Forest classifier with 10-fold cross-validation.

The experimental results obtained from DynaMut were compared with those of methods already consolidated for the study of mutations: I-Mutant, Maestro, DUET, SDM2, mCSM, ENCoM and FoldX. Although the study suggests the tool as a more suitable approach for predicting "destabilizing and stabilizing" mutations, the results demonstrate a lower Pearson correlation in relation to the I-Mutant2, DUET and mCSM methods.

DynaMut is freely available (<http://biosig.unimelb.edu.au/dynamut2/>) and offers a simple interface, providing users with three analysis options: single mutation (for processing a specific mutation), multiple mutations (for batch processing of a list of mutations) and NMA. The tool was developed using the Bootstrap framework version 3.3.7 (front-end) and the Python

programming language, through the Flask framework version 0.12.2 (back-end).

1.6.3 PMut

PMut is a web tool for the annotation of pathological variants in proteins. Its first version was released in 2005, as a neural network-based classifier trained with a dataset extracted from SwissProt (<https://www.uniprot.org>). In (FERRER-COSTA et al., 2005), the authors present a new version of the tool, including the PyMut software package, through which users can prepare their own predictors for specific families of proteins. This creates an advantage for those who want to integrate these prediction features into their applications. Access to PMut is freely available at <http://mmb.irbbarcelona.org/PMut>, which also includes a tutorial with usage guidelines.

The new version of PMut (PMut2017) uses a Random Forest classifier, and its evaluation was performed using different approaches: 10-fold cross-validation, blind validations with SwissVar and ClinVar entries, and comparisons with specific genes. In a blind validation with SwissVar inputs, for example, PMut obtained greater accuracy than other methods such as SIFT and PROVEAN, but is inferior to LRT and PON-P2 methods.

The PyMut package included in this new version was developed as a Python 3 module and is based on consolidated libraries such as NumPy (for numerical computation), Pandas (for data management), and Scikit-learn (for machine learning). This module can be downloaded and installed locally. Its source code is available at <https://github.com/inab/pymut> and also at the official Python package repository at <https://pypi.python.org/pypi/pymut>.

1.6.4 SNPnexus

SNPnexus (ULLAH et al., 2018) is a web tool for the analysis of genetic sequence variation. It uses a Perl pipeline and a MySQL database to perform instant functional annotations. An advantage of the tool is that it provides information about annotations already available, as well as different examples that can help the assembly of new sequences. The portal includes a section with guidance on each available annotation category.

The SNPnexus architecture is structured in a request and response scheme that comprises two layers: access layer (with different search options and output formats to support the study of variants) and storage layer (composed of data sets for annotations and auxiliary data sets). SNPnexus is freely available at <http://www.snp-nexus.org>. Its use depends on filling out a form so that the data is processed and the results sent to the user.

1.6.5 Interactome INSIDER

INtegrated Structural Interactome and Genomic Data browser (Interactome INSIDER) is a web tool for the analysis and enrichment of mutations specifically in human diseases. It allows user to explore mutations of diseases already known and available in different databases, as well as mutations included by other users. It provides a diverse set of pre-computed mutations that can be accessed through the tool portal available at <http://interactomeinsider.yulab.org>.

As a prediction mechanism, Interactome INSIDER uses a framework called Ensemble Classifier Learning Algorithm to predict Interface Residues (ECLAIR), which combines 8 independent classifiers based on the Random Forest classification algorithm from the scikit-learn library. ECLAIR was compared with other prediction methods (PIER, PINUP, SPPIDER, CPORT and PRISE) obtaining a performance as good or even a little better in some cases, such as accuracy and recall metrics, for example (MEYER et al., 2018). The tool does not optionally include a user guide.

1.6.6 I-Mutant2.0

I-Mutant2.0 is a web tool to predict changes in protein stability after single point mutations (CAPRIOTTI; FARISELLI; CASADIO, 2005). This tool is based on a support vector machine (SVM) and can receive protein structure or sequence data as input.

The I-Mutant2.0 classifier trained and tested the input using a cross-validation procedure. The dataset taken from the Thermodynamic Database for Proteins and Mutants (ProTherm) were able to correctly predict 80% for structures and 77% for sequences of the dataset.

I-Mutant2.0 is available through a web interface at <https://folding.biofold.org/i-mutant/i-mutant2.0.html>. For both inputs (structures and sequences), the tool gives the main result the value of the free energy change or only its sign. However, considering a better user experience, the web tool does not offer a modern interface.

1.6.7 SIFT

In (NG; HENIKOFF, 2003; Kumar; Henikoff; Ng, 2009), the authors present the SIFT (Sorting Intolerant From Tolerant) method, a tool for analyzing mutations in proteins through sequence homology. Its functioning is based on the premise that important amino acids will be conserved, and alterations in positions with a high degree of conservation tend to be intolerant of substitutions.

To predict the impacts of these substitutions on protein function, SIFT considers the position where the change occurred and the type of change. Thus, the probability that an amino acid is tolerated in one position is calculated from a pre-established sequence alignment. Sub-

stitution is given as deleterious if the normalized value of this probability is less than a cutoff point. For (Kumar; Henikoff; Ng, 2009), a limitation of the method is that it does not apply structural data to assess the effects of substitutions.

SIFT provides a toolbox with six features for data entry: two batch tools, which provides predictions for multiple proteins and their substitutions, and another four tools that provide detailed predictions for a single protein, considering all substitutions or only selected ones. SIFT is currently available through a web platform at <https://sift.bii.a-star.edu.sg/>. Considering a better user experience, it does not provide a modern and intuitive interface.

1.6.8 Polyphen

Polyphen (Ramensky; Bork; Sunyaev, 2002) is a web-based tool for predicting the effects of non-synonymous coding SNPs (nsSNPs) on protein structure and function. Polyphen may also be relevant in discovering the structural basis of mutations in diseases, enabling an understanding of their molecular cause.

It is a server dedicated to the automatic functional annotation of encoding nsSNPs, which receives an amino acid sequence as input and, from which it performs a series of actions (figure process). The PolyPhen server was applied to annotate all SNPs deposited in the HGVbase database. For the authors, the availability of this collection of annotated data could be useful in selecting nsSNPs for association studies based on candidate genes. However, the study does not present comparisons of the tool with others with the same purpose. The tool is available in its second version (PolyPhen-2) at <http://genetics.bwh.harvard.edu/pph2/>. An example of the results page of the tool is represented in Figure S7[14]. Considering a better user experience, the tool does not provide a modern interface.

1.7 Interactions at atomic/residue level

Studies involving information about protein structures and their interactions with different types of molecules have significantly grown in recent years, but determining aspects associated with such interactions still face challenges (PAN et al., 2019). Several experimental and computational techniques have been used to study interactions involving proteins, such as determining, evaluating, and understanding structures and protein interactions, this has a fundamental importance in molecular biology. Many proteins associate with other molecules to perform their functions and form essential complexes in a large number of cellular functions, such as cell signaling, proliferation, DNA repair, and immunity (BICKERTON; HIGUERUELO; BLUNDELL, 2011). Three of the main interactions involving proteins will be described here, along

[14] The supplementary material of this paper is not in the content of this dissertation

with computational tools involving them: protein-protein, protein-ligand, and protein-peptide interactions.

Protein-protein interactions (PPI) are physical inter-chain contacts that lead to the forming of protein agglomerates as a result of a biochemical process in a cell. According to (RIVAS; FONTANILLO, 2010), the definition of PPI has to take into account two aspects that involve the interaction interface. The first one says that the interaction interface should be deliberate and not accidental, and the second aspect says that the interaction interface must be non-generic. Protein-protein interactions deal with a wide variety of biological processes, including cellular and metabolic interactions. In addition, PPIs play an important role in predicting protein function and the druggability of molecules (RAO et al., 2014).

Biological macromolecules, such as proteins, interact with a large number of molecules with a high degree of specificity and high affinity, called ligands, which can be defined as any molecule capable of binding to a protein that does not belong to the protein class. (OLSSON et al., 2008). The interactions between these ligands and proteins are called protein-ligand interactions (PLI), the second type of protein interaction discussed here. Such interactions play a key role in enzyme catalysis, signal transduction, and several other biochemical processes (SCHREYER; BLUNDELL, 2009). Thus, understanding the aspects involving PLIs helps to understand protein functions, furthermore, can be linked to the development of new drugs, for instance (DU et al., 2016).

The last interaction mentioned here is the protein-peptide. These interactions are present in several cells of living beings and play an important role in the protein-protein interaction network, in addition to participating in signaling and regulation (LONDON; RAVEH; SCHUELER-FURMAN, 2011). Structural analysis of protein-peptide interactions indicates that most peptides, when binding, do not perform conformational changes in the protein, minimizing the entropy cost of the binding (LONDON; MOVSHOVITZ-ATTIAS; SCHUELER-FURMAN, 2010). Peptides usually bind in the largest pockets available in proteins, being completely located in cavities, bound in pockets or form, at the surface of a protein, beta-strand interactions (STANFIELD; WILSON, 1995). In the next subsections, we will describe some tools that work with protein-protein, protein-ligand, and protein-peptide interactions. Table 6 summarizes the tools presented in this topic.

1.7.1 Protein-protein interaction methods

1.7.1.1 PrePPI

(ZHANG et al., 2012) present a method for protein-protein interactions (PPI) prediction based on three-dimensional structural information. The developed tool, named PrePPI (predicting protein-protein interactions), combines structural information along with other functional

Table 6 – Summary of the protein interaction tools.

Name	URL	Features	Limitations	Reference
PrePPI	<https://bhapp.c2b2.columbia.edu/PrePPI/>	Free of charge (PPI database) Based on structural information of proteins	Only predicted interactions are available, not the method	(ZHANG et al., 2012)
ppiGReMLIN	<https://ppigremlin.github.io/pages/files.html>	Free of charge Standalone program Based on structural information of proteins Graph-based method Only requires python environment	Does not calculate water mediated interactions	(QUEIROZ et al., 2020)
mCSM-PPI2	<http://biosig.unimelb.edu.au/mcsm_ppi2/>	Free of charge Webserver Graph-based method	-	(RODRIGUES et al., 2019)
PLIP	<https://plip-tool.biotec.tu-dresden.de/plip-web/plip/index>	Free of charge Webserver Offers publication ready images	Outputs limited to binary interaction fingerprints Few statistical analysis available	(SALENTIN et al., 2015)
LIGPLOT	<https://www.ebi.ac.uk/thornton-srv/software/LIGPLOT/>	Free for academic use Standalone program Available for Windows, Linux and macOS	Requires license to use	(WALLACE; LASKOWSKI; THORNTON, 1995; LASKOWSKI; SWINDELLS, 2011)
nAPOLI	<http://bioinfo.dcc.ufmg.br/napoli/>	Free of charge Webserver Works in large-scale scenarios Graph-based method	Not designed to detect interactions in complexes from virtual screening	(FASSIO et al., 2019)
LeView	<http://www.pegase-biosciences.com/leview-ligand-environment-viewer/>	Free of charge Standalone software Available for Windows, Linux and macOS	Generates 2D images only	(CABOCHE, 2013)
BINANA	<https://durrantlab.pitt.edu/binana/>	Free of charge Webserver or standalone program Only requires python environment	-	(DURRANT; MCCAMMON, 2011)
PoseView	<https://proteins.plus/2ozr#poseview>	Free of charge Webserver Generate structure diagrams from scratch Describe moties at atomic detail following IUPAC conventions	Lacks of statistical data	(STIERAND; MAASS; RAREY, 2006)
GalaxyPepDock	<http://galaxy.seoklab.org/pepdock>	Free of charge Webserver or standalone program Template-based docking approach Uses energy-based optimization	Does not perform predictions of peptides in isolation	(LEE et al., 2015)

aspects and it is available free of charge at <https://bhapp.c2b2.columbia.edu/PrePPI/>. The tool involves a few steps, when a pair of proteins is given as input, a sequence alignment is first used to find structural representatives. Afterward, a structural alignment is performed to find close and remote structural neighbors. When two pairs of neighbors of structural representatives form a pair reported in the PDB, this is defined as a template for modeling the interaction between the input proteins.

For the evaluation of the created protein-protein interaction model, some metrics are used (more specifically, five empirical scores) to analyze properties from individual monomer alignments to their templates. PrePPI uses a Bayesian network to combine structural and non-structural aspects, bringing more reliability in PPI predictions, as well as identifying more interactions compared to using a single source of information. The results obtained by the tool allow the use of homology models that can be used to study the close and remote geometric relationship between proteins. The tool facilitates the generation of experimentally testable hypotheses and allows the creation of a structural model for PPIs that play an important role in biological systems.

1.7.1.2 ppiGReMLIN

(QUEIROZ et al., 2020) proposed ppiGReMLIN, a tool for analyzing protein-protein interactions with a fine level of granularity at both the residual and atomic levels. The method is freely available at <https://ppigremlin.github.io/pages/files.html> and uses a graph-based strategy, where the nodes represent the atoms of the proteins (labeled according to their physicochemical properties) and the edges the non-covalent interactions (based on the physicochemical information of their proteins atoms and distance criteria). For each PDB entry, connected components are calculated and serve as the basis for clustering analyses. This step is performed with the Spectral Clustering algorithm (NG; JORDAN; WEISS, 2002) and is performed to find

similar graphs for the next step of frequent subgraph mining, which is executed to find arrays of conserved structures.

To demonstrate the tool's ability to describe and find structural arrangements, at the atomic level, at protein-protein interfaces, the authors used two databases containing protein-protein complexes: a serine protease dataset and a BCL-2 dataset (details, including graphical analyses about these datasets can be found on the tool's webpage). Thus, it was possible to deduce that ppiGReMLIN is capable of detecting substructures at protein-protein interfaces on a large scale, which were compared with relevant residues and interactions found in the literature. Tests performed with the tool showed that ppiGReMLIN can find conserved structures automatically and the results for each of the datasets used ranged from 69% to 100% regarding the accuracy, with 100% of recall.

1.7.1.3 mCSM-PPI2

(RODRIGUES et al., 2019) proposed the development of the mCSM-PPI2 tool that predicts the effects of mutation in protein-protein interactions. Although we have already discussed aspects of mutation analysis in Section 6, we think it is pertinent to bring this tool to exemplify protein-protein interactions, as this is the context in which mutations are studied in (RODRIGUES et al., 2019)'s work. To create the optimized predictor, the tool uses a graph-based approach to be able to model the effects of variations, including several aspects, like inter-residue interaction network, complex network metrics, information about evolutionary and energetic terms. The mCSM-PPI2 models geometric and physicochemical properties of protein-protein interactions and can be applied in studies involving small molecules and protein structures.

The main component of the tool is the use of graph-based structural signatures (mCSM), which represent the context of wild-type residues. In this model, nodes represent atoms, and edges represent the interactions between nodes. The physicochemical information is coded according to the residual properties of the amino acids and the distance between atoms is described by their properties, defined in signatures. The tool also uses a machine learning technique (which uses six new features in the training stage) along with the graph-based signatures approach to explore the effects of mutations on protein-protein bonds.

The authors also implemented a web server (available at <http://biosig.unimelb.edu.au/mcsm_ppi2/>) to host mCSM-PPI2. The tool offers two services on its website: the first one is used to analyze the effects of user-specified mutations, and the second one performs the prediction of mutation effects in the protein-protein context. As a result, the website shows the whole protein binding environment, together with an interactive 3-dimensional viewer. Furthermore, a 2D viewer that shows non-covalent interactions of wild-type and mutant structures is provided.

1.7.2 Protein-ligand interaction methods

1.7.2.1 PLIP

The PLIP (protein-ligand interaction profiler) (ADASME et al., 2021) is a web service (available at <https://plip-tool.biotec.tu-dresden.de/plip-web/plip/index>) used for identification and visualization of non-covalent contacts in the context of protein-ligand interactions, working with 3D structures and allowing in-depth analysis involving patterns of such interactions. The tool, which is free and open-source, offers automated high-quality images, and session files PyMOL to create custom images in addition to results files for data processing.

Identification and report of protein-ligand interactions are performed by the PLIP algorithm in four different steps: preparation, functional characterization, rule-based matching, and filtering of interactions. In the first stage, preparation, the structure is hydrogenated and its ligands and binding sites are identified. The functional characterization step includes the detection of hydrophobic atoms and acceptors/donors for halogen and hydrogen bonds, in addition to a search for aromatic rings and charge centers. In the next step, putative interacting groups are combined by applying geometric criteria. Finally, in the last step, filtering interactions are performed to eliminate redundant or overlapping interactions.

It is possible to apply the information brought by the web service in docking result evaluation, drug development, and repositioning and binding site similarity evaluation. As input, the webserver accepts a protein-ligand complex in PDB format and outputs 2D and 3D diagrams, visualization files, and various details involving interaction patterns for each binding site.

1.7.2.2 LIGPLOT

LIGPLOT is an algorithm for plotting protein-ligand interactions developed by (WALLACE; LASKOWSKI; THORNTON, 1995). The tool allows users to create two-dimensional representations of protein-ligand complexes from PDB input data files. The functioning of the algorithm can be summarized in a few steps. First, connectivity is calculated from 3D coordinates and hydrogen bond information, then rotatable bonds are identified and all ring groups are flattened. Finally, the structure is unrolled and cleaned up and a graphical representation is created through a postscript file. LIGPLOT is available at <https://www.ebi.ac.uk/thornton-srv/software/LIGPLOT/> and it has an extensive operating manual session, where the users can learn about the tool's operation, as well as links to references and FAQ.

To improve LIGPLOT, (LASKOWSKI; SWINDELLS, 2011) proposed the development of LigPlot+, a tool alleged to be a successor to the original algorithm and which is available on the following website: <https://www.ebi.ac.uk/thornton-srv/software/LigPlus/>. This new version also generates two-dimensional representations of protein-ligand interactions, like the first one, but with improvements. The first one is a new interface, developed using Java lan-

guage, which offers diagram editing through click-and-drag mouse operations in an improved plotting environment. LigPlot+ also allows superposition of related diagrams, making it easier to visualize similar protein-ligand complexes and gives the users the option to visualize the representations in the PyMOL and RasMol tools.

1.7.2.3 nAPOLI

nAPOLI (Analysis of PrOtein-Ligand Interactions), proposed by (FASSIO et al., 2019), is a tool for analyzing protein-ligand interactions. The method works with the analysis of conserved interactions of protein-ligand complex datasets, along with interactive visualizations and reports of residues and atoms interactions. The approach brought by the authors consists of using bipartite graphs to model the protein-ligand interactions so that the atoms are represented by the nodes of the graph (characterized by their physicochemical properties) and the edges indicate the interactions between the atoms. Then, similar ligands are grouped to elucidate conserved interactions in groups of ligands. For this, the clustering algorithms GenerateMD (D. Szisz, 2021) and Ward (KELLEY; GARDNER; SUTCLIFFE, 1996), both from ChemAxon, are used. Finally, superposition is used to find equivalences of residues and conserved interactions in various complexes.

The tool is available free of charge at <http://bioinfo.dcc.ufmg.br/napoli/>. On this web page, users are faced with two main menus: dataset submission and dataset analysis. In the first one, nAPOLI entries are inserted with information on the PDB ID of the proteins, chain, and information about the ligands, in addition to aspects about structural alignment. Still, in the dataset submission menu, users can even compose a new dataset and also define several parameters of the tool, which involve aromatic stacking, hydrogen bonds, and also hydrophobic, repulsive, and attractive interactions. In the dataset analysis menu, through the link generated when a project was previously submitted, users have access to various information about protein-ligand interactions. It is possible to graphically visualize analysis of interactions (Figure S8[15]), perform filtering of ligands, and several other submission details. In addition, the website has a help section, where users can find detailed materials on all aspects of using nAPOLI, as well as descriptions of graphical analysis and interaction summary.

1.7.3 Protein-peptide interaction methods

1.7.3.1 GalaxyPepDock

GalaxyPepDock (<http://galaxy.seoklab.org/pepdock>)(LEE et al., 2015) is a web server capable of generating high-resolution complex structure models of protein-peptide binding by a similarity-based method. These models are built based on templates from experimen-

[15] The supplementary material of this paper is not in the content of this dissertation

tally known protein-peptide interaction structures and GALAXY (HEO; PARK; SEOK, 2013) energy-based optimization that improves structural flexibility on the template-based search.

The method proposed by the authors consists firstly in selecting templates from the PepBind (DAS et al., 2013) database, according to similarity measures of protein structures and their interactions. For the construction of the model, for each template, 50 models of complex structures are generated using the GalaxyTBM (KO et al., 2012; KO; PARK; SEOK, 2012) method, using both protein structure and peptide sequence alignment. For model optimization, distance controls between interactions are inserted in GALAXY energy, and finally, 10 structures with the best energy values are selected for each template and then refined through GalaxyRefine (HEO; PARK; SEOK, 2013).

The website can present the generated structures using PyMOL and offers a download option. Along with the models, information about the binding sites and their estimated accuracy is also available. The average accuracy of GalaxyPepDock binding site residues identification is 75.4%.

1.8 Catalytic and binding site prediction

Currently, there is an increase in the number of structures and protein sequences deposited in specialized databases, being provided mainly by the advancement of genomic sequencing technologies and methods to determine their structure. Among the challenges involving proteins, we can highlight the prediction of their function, which increases the need for automatic and reliable approaches that are capable of performing such prediction. (AKCAPINAR; SEZERMAN, 2017).

Despite efforts to define and annotate the functions of proteins, the Pfam (EL-GEBALI et al., 2019), a database that catalogs protein families, contains about 22% (3961) of all entries marked as proteins of unknown function. As another example, the Uniprot database, responsible for cataloging protein sequences, contains over 120 million entries. The amount of proteins that have been analyzed by experts is only less than a million (CONSORTIUM, 2019).

A large number of biological processes, including cellular defenses, catalysis of enzymes, and signal transmission, depending on the interaction between proteins and small molecules. An enzyme can be defined as a protein molecule that acts as a catalyst in chemical reactions, regulating and synchronizing them. Therefore, it is crucial to develop methods that can be capable of identifying protein binding sites, contributing to studies on protein functions and new functional roles (SANTANA et al., 2020a).

The function of a protein can be predicted by searching for enzyme binding sites. Enzyme binding sites are areas on the surface of an enzyme designed to interact with other molecules and they can be divided into two different parts: the substrate-binding site and the

catalytic site. The first one recognizes the molecule on which the enzyme performs and the second one is a collection of two to six amino acids that performs the catalytic role (SCHWEDE; PEITSCH, 2008).

There are some experimental methods performed in web labs that can identify catalytic and binding sites in proteins, however, they still face difficulties to be executed, due to problems related to cost, time, and automation of processes. These issues create the demand for computational techniques to grow even more, making efforts to create and improve prediction methods. These created techniques have been able to search and predict catalytic and binding sites, as well as their geometry, function, and various other information that can help different aspects of research in the area (MALLICK; VIDYARTHI et al., 2011).

In the literature, there are a large number of computational methods that have been proposed to predict the binding and catalytic sites of a protein, providing a reduction in costs and time compared to experimental procedures. These methods can be grouped into three main categories: algorithms based on sequence, structure, or hybrid techniques. Table 7 summarizes catalytic and binding site prediction tools described in this section.

Table 7 – Summary of the catalytic and binding site prediction tools.

Name	URL	Features	Limitations	Reference
GASS	<https://gass.unifei.edu.br/>	Free of charge Webserver Based on structural information of proteins Genetic Algorithm-based method	Dependent on precisely oriented residues	(IZIDORO; MELO-MINARDI; PAPPA, 2015; IZIDORO; LACERDA; PAPPA, 2015; MORAES et al., 2017)
PINGU	-	Free of charge Webserver Based on sequence information of proteins SVM-based method	Predicts specific residue functions Data processing takes long time	(PAI; RANJANI; MONDAL, 2015)
iCataly-PseAAC	<http://www.jci-bioinfo.cn/iCataly-PseAAC>	Free of charge Webserver Based on sequence information of proteins Fuzzy KNN-based method	-	(XUAN et al., 2015)
Chien and Huang method	-	Based on sequence and structural information of proteins SVM-based method	Method is not available via webserver or standalone program	(CHIEN; HUANG, 2013)
GRaSP	<https://grasp.ufv.br/>	Free of charge Webserver Based on structural information of proteins Graph-based method	Binary results may be uninformative	(SANTANA et al., 2020a)
FunFOLD	<http://www.reading.ac.uk/bioinf/FunFOLD>	Free of charge Webserver or standalone program Based on structural information of proteins Uses TM-Align method	Limited quality assessment features	(ROCHE; TETCHNER; MCGUFFIN, 2011)
DeepSite	<https://www.playmolecule.com/deepsite/>	Free of charge Webserver Based on structural information of proteins CNN-based method	No customize protein options Undocumented web API	(JIMÉNEZ et al., 2017)
TRAPP	<https://trapp.h-its.org/>	Free of charge Webserver Based on structural information of proteins	Identifies only specific pockets Slow running (Molecular Dynamics approach)	(KOKH et al., 2013; STANK et al., 2017)
CAVER	<https://loschmidt.chemi.muni.cz/caverweb/>	Free of charge Webserver Based on sequence and structural information of proteins Uses TM-Align method	Tunnel length is sometimes shortened Works only with static structures	(STOURAC et al., 2019)
MED-SuMo	<http://www.medit.fr/>	Commercial software Standlaone program Based on structural information of proteins	GUI available only for Windows	(DOPPELT-AZEROUAL et al., 2009)
eFindSite	<http://www.brylinski.org/efindsite>	Free of charge Standalone program Based on structural information of proteins Perform ligand-based virtual screening against identified pockets	Unbalanced base of templates	(BRYLINSKI; FEINSTEIN, 2013)
POVME	<https://github.com/POVME/POVME>	Free of charge Standalone program Based on structural information of proteins Integrate results into large data workflows	Limited when working with poorly understood proteins	(DURRANT; OLIVEIRA; MCCAMMON, 2011)
SiteEngine	<http://bioinfo3d.cs.tau.ac.il/SiteEngine/SiteEngine.html>	Free of charge Standalone program Based on structural information of proteins Also uses physicochemical properties Can handle large protein structures quickly	Limited quality of biological predictions No implicit treatment of electrostatic potentials	(SHULMAN-PELEG; NUSSINOV; WOLFSON, 2004)
SVILP_ligand	<http://www.sbg.bio.ic.ac.uk/svilp_ligand/>	Free of charge Standalone program Based on structural information of proteins SVM-based method	Method not tested on unbound structures Poor information available on the web address	(KELLEY et al., 2009)

1.8.1 Methods for catalytic site prediction

1.8.1.1 GASS

As an example of methods for predicting catalytic sites, (IZIDORO; MELO-MINARDI; PAPPA, 2015; IZIDORO; LACERDA; PAPPA, 2015; MORAES et al., 2017) propose GASS, a genetic algorithm that searches for active sites based on templates, which uses only distance information between residues. The search problem treated by GASS can be described as given a collection of N amino acids that make up the active site A of an enzyme of known function (template), and a protein B with M amino acids of unknown function, the method looks for the pattern A in B.

GASS defines an individual, used in his genetic algorithm, as a group of amino acids that make up candidate solutions to the problem. An individual is defined as a vector where each index contains information about a single amino acid. The method involves a population of candidate active sites in which genetic operators such as crossover and mutation act. The fitness function used by GASS is based on the Root Mean Square Deviation (RMSD) and calculates the distance between a template and an active candidate site. The web server can be found at <https://gass.unifei.edu.br/>. Figure S9[16] shows the GASS result page, with found active sites and matched templates.

The authors compared GASS with 17 other methods submitted in CASP 10 (CASSARINO; BORDOLI; SCHWEDE, 2014), for protein function prediction, where it was ranked fourth compared to the other methods, according to Matthew Correlation Coefficient (MCC) values. Other results also showed GASS effectiveness in identifying catalytic sites cataloged by the CSA with accuracy greater than 90% in several cases.

1.8.1.2 PINGU

An SVM-based method was proposed by (PAI; RANJANI; MONDAL, 2015), the algorithm named PINGU (PredIction of eNzyme catalytic residues using seqUence information), classifies sets of catalytic and non-catalytic residues, then filters these results through a post-processing method, leaving just the most probable catalytic ones. The train and the independent test data sets were selected from the CSA and the PDB, 650 enzymes for the training and 200 enzymes for the test dataset. The SVM model was chosen over Logistic Regression and Radial Basis Function Network as a better performance was presented using this type of data and inputs.

The PINGU algorithm outperformed other similar methods by using specific physico-chemical residue attributes, together with an evolutionary conservation index. During the SVM step, the algorithm selects as many catalytic residues as possible, even at the cost of a high

[16] The supplementary material of this paper is not in the content of this dissertation

false-positive rate. This result is then filtered by the post-processing through S-SITE, a ligand-binding site predictor that works based on template recognition. This step allows for minimizing the false positives and leaves almost all of the catalytic residues. A boost of 16% in precision and 0.138 in MCC is marked after filtering.

1.8.1.3 iCataly-PseAAC

When using residue's physicochemical information, (XUAN et al., 2015) proposes an interesting method called iCataly-PseAAC. The classification is made upon a preprocessed 21x20 pseudo amino acid composition matrix, where each line represents the amino acid position in a peptide, and each column represents the amino acid type. The gray-PSSM encoding is used in this case, to better represent the evolutionary conservation score data of each peptide and each residue within it. But the encoding itself does not have classification powers, thus the authors apply a Fuzzy K-NN to the prepared dataset to perform a nonparametric classification.

To assess the quality of the method, the authors compared the iCataly-PseAAC with two other predictors, CRpred (ZHANG et al., 2008) and EXIA (CHIEN; HUANG, 2012), where the first is based on protein sequence information, and the second on protein structure and sequence conservation. The results were measured according to the accuracy in prediction, and the iCataly-PseAAC scores were greater than 87% in all three benchmark datasets, making the proposed method superior to the other two also evaluated.

The tool is available at the following web address: <http://www.jci-bioinfo.cn/iCataly-PseAAC>. On this page, the user can enter the query protein sequence in FASTA format or upload a batch prediction file. Authors report that predictions generally take about 20 minutes to run for each protein, and thus it is possible to receive prediction results via the user's email. Furthermore, users can find additional information about iCataly-PseAAC on the webpage, such as support data and a read me section.

1.8.2 Methods for binding site prediction

1.8.2.1 GRaSP

GRaSP (SANTANA et al., 2020a) is a computational strategy that represents the residue environment (encoded by 2 shells of neighboring residues) through graphs to predict protein-ligand binding sites. The method uses machine learning to model, at the atomic level, the residue environment in the form of graphs. For each residue in a protein, some atom features used include topological and physicochemical properties. The interaction between them is represented as a graph, which is encoded as a vector. GRaSP is available for free at <https://grasp.ufv.br/>. In Figure S10[17], GRaSP presents the suggested ligands of the binding site, along with a 3D

[17] The supplementary material of this paper is not in the content of this dissertation

viewer.

The problem of binding site prediction was defined as follows: a characteristic vector is created for every single residue of a protein, based on 14 descriptors. These descriptors can be grouped into several levels, such as residue, atom, and interaction. Then, GRaSP builds a matrix that represents the entire set of proteins. In the experimental evaluation, six different datasets were used, as well as tests were performed comparing GRaSP to several state-of-the-art methods.

To evaluate the method, the authors used six different datasets, and GRaSP presented compatible or superior results. To illustrate, it ought to mention that GRaSP outperformed the other six methods that predict binding site residues. Also, the method ranked seventh among 17 CASP 10 methods (there was no statistical difference between the first 10 methods) and outperformed RaptorX-Binding, which is a method from CAMEO independent assessment similar to GRaSP. The method ranked second when compared to methods that predict pockets (that can potentially be binding sites) and it takes 10-20 seconds to calculate a binding site while the state-of-the-art method takes 2-5 hours.

1.8.2.2 FunFOLD

The FunFOLD algorithm, proposed by (ROCHE; TETCHNER; MCGUFFIN, 2011), works with an automatic approach for the selection of residues and cluster identification. The method calculates binding sites, their residues, and other information about a target protein and its ligands. FunFOLD is based on the concept that templates that contain ligands, coming from the PDB, with the same fold as 3D models of the target protein, probably contain similar binding sites. FunFOLD uses the TM-Align method to superpose each of the structure templates that contain relevant ligands onto the 3d protein model.

The developed tool can be combined with existing fold recognition servers, using as input a list of templates and a 3D model of proteins. The authors also proposed the FunFOLD2 server (ROCHE; BUENAVISTA; MCGUFFIN, 2013), which can perform prediction of proteins from their sequences via structure. It uses connection sites prediction from FunFOLD, and quality assessment protocols from FunFOLDQA (ROCHE; BUENAVISTA; MCGUFFIN, 2012). FunFOLD2 is available for use at <http://www.reading.ac.uk/bioinf/FunFOLD>, where users can download both FunFOLD and FunFOLDQA.

1.8.2.3 DeepSite

(JIMÉNEZ et al., 2017) propose a method called DeepSite for binding site prediction based on convolutional neural networks. The approach works by mapping protein structures from a computer vision perspective, in 3D images discretized into voxels. These voxels consist of a group of pharmacophoric properties at the atomic level and their occupations are defined

according to the atoms of the protein, taking into account their excluded volume and some atomic properties, such as aromatic, hydrophobic, and metallic aspects. The tool is available at <https://www.playmolecule.com/deepsite/>, where the user can perform the prediction through a protein structure in PDB format and view the results through the WebGL viewer.

DeepSite makes use of a Deep Convolutional Neural Network (DCNN) composed of four convolutional layers, and every two of these layers, max-pooling, and dropout are used. At the end of the DCNN, there is a fully connected layer. All layers, except the last one, use the Exponential Linear Unit (ELU) activation function and the network output uses the sigmoidal activation function. For training and evaluation of the DCNN, a dataset formed by 7,622 proteins from the scPDB database was used, where filters involving similarities of binding sites were applied to the structures.

To evaluate the proposed method, the authors used two different criteria. The first one evaluates the performance of predictions based on two metrics: distance to the center of the binding site (DCC) and discretized volumetric overlap (DVO). Both metrics are also used in two other prediction methods, fPocket and Concavity, as shown in (CHEN et al., 2011) work, to compare them with DeepSite. In addition, the second criterion evaluates the method on different structural groups of proteins. Tests were performed using the SCOPe (FOX; BRENNER; CHANDONIA, 2014) dataset, which provides curated information on PDB structures.

1.8.2.4 TRAPP

The TRAnsient Pockets in Proteins (TRAPP) (KOKH et al., 2013; STANK et al., 2017) is a tool for detecting pockets and sub-pockets that may have been generated by internal motion in proteins. Moreover, this tool provides resources for exploring the dynamics of protein binding sites. TRAPP is not a method focused on identifying all binding pockets in proteins but works to analyze physicochemical properties and spatial changes in specific pockets, which may have been generated from protein motion.

Three different modules make up the webserver: TRAPP Structure, TRAPP Analysis, and TRAPP Pocket. After defining parameters and user inputs, the three modules are executed in sequence. In the first module, TRAPP Structure, protein structures are generated, representing characteristics of the binding pocket, such as the diversity conformation. The second module, TRAPP Analysis, comes next and is responsible for providing tools for comparing binding pockets in protein structures. Finally, the TRAPP Pocket module locates pockets and identifies transient regions in protein structures.

TRAPP is available free of charge at <https://trapp.h-its.org/> and can be run via a web application. Users can also download a desktop version of the tool, which comprises a command-line version of TRAPP but only for Linux systems. Although it is necessary to obtain a license to use TRAPP, there is currently no charge to acquire it, simply fill out a form with

user information. Several tools for exploring binding sites can also be found on the webserver, as well as resources for analyzing the dynamics of binding sites. In addition, other information can be found on the TRAPP website, such as examples of work done by the tool, as well as documentation to help users.

1.8.2.5 CAVER

(STOURAC et al., 2019) proposed Caver Web, an interactive tool for the identification of protein tunnels and channels, in addition to allowing the analysis of ligand transport. The server is built using two different methods. For the detection of tunnels, the Caver 3.02 method (CHOVANCOVA et al., 2012) is used, a tool that provides high-quality results with fast calculations, and is based on the Voronoi diagram representation of protein structures. Ligand transport analysis is performed using the CaverDock 1.0 tool (FILIPOVIČ et al., 2019), which is based on molecular docking iteration along the tunnel and performs the analysis quickly and accurately.

Caver Web is available for free through its website <https://loschmidt.chemi.muni.cz/caverweb/>, which, in addition to the tool, also brings examples of using the method, as well as other tools of the group. The identifications and analyzes carried out by Caver are made entirely via the web, without the need to install software locally by the user. For the method execution, some steps must be followed. First, the user must select the protein structure to work on (PDB format) and also select the biological unit. In the second step, the starting point selection for tunnel detection is defined, according to several different modes, such as a pocket, catalytic pocket, ligands, sequence, and manual tweaking. Finally, in the third step, Caver settings and parameters are selected, such as shell radius and depth, clustering threshold, minimum probe radius, and also maximal distance.

After following all the steps described above, Caver Web is executed, calculating the protein tunnels. A results page is shown containing various analyses, including information about the executed job and tunnels. About the tunnels, several characteristics and details are described, such as bottleneck radius, length and curvature of the tunnel, list of all centerline spheres, and also the residues and atoms around the tunnel. In addition, the results page allows analysis of ligands transportation through the tunnel.

1.8.2.6 POVME

Proposed by (DURRANT; OLIVEIRA; MCCAMMON, 2011), POVME (POcket Volume MEasurer) is a tool for ligand-binding pocket analysis in proteins. The method works with a grid-based representation of the binding pockets through voxels and performs the analysis through descriptions of the flexibility and shape of the cavity. Its latest version, POVME 3.0 (WAGNER et al., 2017), adds several features, such as methods for clustering and principal

component analysis, identification of features through chemical coloring scheme, as well as new features for analyzing and comparing binding pockets.

Four steps are required to run POVME. First, the user must select the region of the binding pocket (using spheres and prisms). In the second step, an encompassing region is defined and a volume-grid file with equispaced points is also generated. Then, volume-grid points close to the protein atoms are deleted, leaving only the points that are in the binding pocket. Finally, in the last step, the volume of the binding pocket is calculated, according to the number of the remaining points. POVME is an open-source tool, developed in Python language and is only available in a standalone version, available for installation at <https://github.com/POVME/POVME>.

1.9 Databases

Data related to bioinformatics have been generated at a fast pace by researchers from all over the world. This increase in data has made it essential to use and interconnect (cross-references) databases for storage and analysis. Several types of databases differ in their structure (e.g. flat-file format and relational form), as well as in the management structures (e.g. object-oriented databases, data warehouses, and distributed databases). In addition, it is necessary to use a DBMS (database management system) to control the database (ZVELEBIL et al., 2008).

Due to a large number of databases in bioinformatics, there is a lot of redundant and replicated data, making it difficult to search for information. The journal Nucleic Acids Research (NAR) can help with this search task. Every year, NAR produces a special issue reporting new and updated databases. In addition, there is a list of the URLs of databases that have been reported in these annual issues of NAR, called the Molecular Biology Database Collection (available from the NAR home page), where can be found data from gene expression and its regulation, genome structure, protein domains, and protein-protein interactions. Also, there are database centers, such as EBI (<http://www.ebi.ac.uk>) and NCBI (<http://www.ncbi.nlm.nih.gov/>) that have links to several databases involving different data such as sequence, structure and genome (ZVELEBIL et al., 2008; GALPERIN; FERNáNDEZ-SUáREZ; RIGDEN, 2016).

Bioinformatics databases can be classified according to the type of data they contain (e.g. sequence data, experimental data, structure data, and protein interaction data) (ZVELEBIL et al., 2008). However, the databases do not store the data itself. For example, an enzyme database may also contain information about catalytic residues (position in the sequence) and links to other databases with structural information. Next, some characteristics from important databases frequently used in bioinformatics will be presented. Table 8 summarizes the databases described in this section.

Table 8 – Summary of the database.

Name	URL	Features	Limitations	Reference
PDB	<http://www.wwpdb.org/>	About 180,000 annotated structures (nucleic acids, proteins, and large macromolecular complexes) using NMR, X-ray crystallography and electron microscopy techniques. It has several tools to help students and researchers (e.g. Mol* 3D Viewer, Protein Feature View, Pair Structure Alignment, Protein Symmetry, Statistics).	Need for pre-processing because in some old structures entries (X-ray crystallography) some atoms may be missing.	(BERMAN et al., 2000a; BERMAN; HENRICK; NAKAMURA, 2003; BERMAN et al., 2006)
M-CSA	<https://www.ebi.ac.uk/thornton-srv/m-csa/>	Database with information about enzyme catalytic residues manually curated and annotation about enzyme mechanisms.	The homologous proteins section can show wrong catalytic residues.	(BARTLETT et al., 2002; PORTER; BARTLETT; THORNTON, 2004) (HOLLIDAY et al., 2005; HOLLIDAY et al., 2006; HOLLIDAY et al., 2012) (FURNHAM et al., 2013; RIBEIRO et al., 2017)
MetalPDB	<https://metalpdb.cerm.unifi.it/>	Information on metal-binding sites detected in three-dimensional structures. The last release includes new contents and tools, statistical analyses involving protein families.	Limitations in the structure preview window.	(PUTIGNANO et al., 2017)
BioLiP	<https://zhanglab.dcmb.med.umich.edu/BioLiP/>	Database semi-manually curated of ligand-protein interactions. Focusing on the biological relevance of the residues, contains 529,047 entries and some resources (COACH, Search, Browse and Download).	Some search options do not work. Unstable web server.	(YANG; ROY; ZHANG, 2012)
UniProt	<https://www.uniprot.org/>	The most complete compendium of all known protein sequence data (experimentally verified, or computationally predicted) and functional annotation. The databases are distributed in several formats.	-	(BATEMAN et al., 2021)
SWISS-MODEL	<https://swissmodel.expasy.org>	Repository of automatically generated 3D models. Contains more than 400,000 high-quality models.	-	(BIENERT et al., 2017)
ProThermDB	<https://web.iitm.ac.in/bioinfo2/prothermdb/>	It contains protein information, structural information, experimental conditions, literature information and experimental thermodynamic data. method	More information about upload and download options is missing. More details and information are missing from the tutorial.	(NIKAM et al., 2020)
MEROPS	<http://www.ebi.ac.uk/merops/>	Manually curated peptidases information database integrated with their substrates and inhibitors. Includes a manually curated bibliography for an extensive number of entries (peptidases, clan, family, inhibitor and substrate).	-	(RAWLINGS et al., 2017)
SCOP2	<https://scop2.mrc-lmb.cam.ac.uk/>	It is an evolution of database SCOP, with the purpose of organizing and categorizing proteins according to their structural and evolutionary relationships.	-	(ANDREEVA et al., 2019; ANDREEVA et al., 2013)
LigBase	<https://modbase.compbio.ucsf.edu/ligbase/>	Database of ligand-binding sites of known structures aligned with related protein sequences.	The last update of the base was in 2002 when it contained approximately 50,000 ligand binding sites for small molecules found in PDB.	(STUART; ILYIN; SALI, 2002)

1.9.1 PDB

Established at Brookhaven National Laboratories (BNL) in 1971, The Protein Data Bank (PDB) is a database of biological macromolecular structures. In October 1998, the management of the PDB became the responsibility of the Research Collaboratory for Structural Bioinformatics (RCSB) (BERMAN et al., 2000a). In 2003, three organizations have established a collaboration to oversee the newly formed worldwide Protein Data Bank (wwPDB; <http://www.wwpdb.org/>): the Protein Data Bank Japan (PDBj) at the Institute for Protein Research in Osaka University, the Macromolecular Structure Database (MSD) at the European Bioinformatics Institute (EBI) and the Research Collaboratory for Structural Bioinformatics (RCSB). The goal was to maintain wwPDB publicly available to the global community (BERMAN; HENRICK; NAKAMURA, 2003; BERMAN et al., 2006).

Currently, PDB (<https://www.rcsb.org/>) has about 180,000 annotated structures, including nucleic acids, proteins, and large macromolecular complexes that have been determined using NMR, X-ray crystallography, and electron microscopy techniques. The PDB also provides several tools to help students and researchers. In addition to the search section, with

various filters and features, there are also visualization tools (e.g. Mol* 3D Viewer, Protein Feature View), analysis methods (e.g. Pair Structure Alignment, Protein Symmetry, Statistics), and downloads tools. The PDB-101 section provides a series of information to help students and researchers explore the proteins and nucleic acid structures. There are videos, interactive animations, and a guide to understanding PDB data.

1.9.2 M-CSA

M-CSA (Mechanism and Catalytic Site Atlas) (RIBEIRO et al., 2017) is a new database formed by merging of databases CSA (Catalytic Site Atlas) (BARTLETT et al., 2002; PORTER; BARTLETT; THORNTON, 2004; FURNHAM et al., 2013) and MACiE (Mechanism, Annotation and Classification in Enzymes) (HOLLIDAY et al., 2005; HOLLIDAY et al., 2006; HOLLIDAY et al., 2012). The CSA database contained information about enzyme catalytic residues manually curated. Each entry in CSA was formed by a reference PDB entry, a list of catalytic residues with their chemical functions, the literature evidence, and the overall reaction. For each entry, there was also a list of homologous PDB structures. The database MACiE contained annotations about enzyme mechanisms. Its annotations also included PDB links, UniProtKB, CATH, and other databases. In addition, there were the roles of catalytic residues and cofactors.

The new M-CSA contains 538 entries (manually curated) where catalytic residues have been identified, but the complete mechanism is unknown, and 423 entries (manually curated) with detailed reaction mechanisms. The annotation was extended to 51,993 homologous PDB structures and over 5 million homologous sequences using the UniProtKB reference dataset.

The M-CSA home page (<https://www.ebi.ac.uk/thornton-srv/m-csa/>) contains a brief qualitative description of the main statistics, citation information, and additional navigation links. The browse page perhaps is the most catching resource once it gives the user a complete overall view of the database. This page shows a table with all the entries that can be sorted by any of the PDB, UniProtKB, CATH, and EC identifiers. Residues and cofactors also can be used to refine the search. When selecting a specific entry, the user can, among other information, access homologous proteins and their respective catalytic residues, as well as visualize the structure using Litemol (SEHNAL et al., 2017). Figure S11[18] shows part of the results of a search with the homologous of the 3NOS protein.

1.9.3 MetalPDB

Updated in 2018, MetalPDB (PUTIGNANO et al., 2017) is a database providing information on metal-binding sites detected in three-dimensional structures. The current release in-

[18] The supplementary material of this paper is not in the content of this dissertation

cludes new contents and tools, statistical analyses involving protein families, as well as 287,122 sites from 50,797 structures (there was a growth of 64% in the last 6 years).

In MetalPDB, the metal sites are stored as Minimal Functional Sites (MFSs), where each MFS is a set of atoms of the metal cofactor, the metal ligands, and any other residue within 5 Å from a ligand. In general, each MFS has at least one associated function among structural, electron transfer, catalytic, substrate, protection, regulatory, and transport. Beyond providing information about the functions or mechanisms of action of a metalloprotein, MFSs can be useful for predicting the role of 3D structures in the absence of experimental biochemical data.

In addition to offering several improvements to the web interface, the new version of MetalPDB (<https://metalpdb.cerm.unifi.it/>) includes some tools like MetalS 3, which is designed to search similar sites with a query site. Figure S12[19] shows part of the results of a search done in MetalPDB.

1.9.4 BioLip

The BioLiP (YANG; ROY; ZHANG, 2012) is a database that lists ligand-protein interactions, just as the PDB. BioLip site residues, however, often differ from the ones on PDB, since it's semi-manually curated and focuses on the biological relevance of the residues. The majority of ligand-binding site prediction methods use templates, generated by the alignment of the same function proteins. This method is very effective on the statistical and computational side, but not all ligands are biologically relevant.

The BioLiP database (<https://zhanglab.dcmb.med.umich.edu/BioLiP/>) contains the following basic resources: COACH, Search, Browse and Download. BioLiP contains 529,047 entries, these being: 109,998 PDB proteins; 57,059 DNA/RNA ligands; 25,960 peptide ligands; 146,969 metal ligands; 299,051 regular ligands; and 23,492 entries with binding affinity data. Due to its relevance, BioLiP is used and cited by many other methods and articles (e.g., GRASP, MionSITE and I-TASSER) (SANTANA et al., 2020a; QIAO; XIE, 2019; MCCULLOUGH et al., 2021; JAMASB et al., 2021).

1.9.5 UniProt

The recently updated UniProt (BATEMAN et al., 2021) databases are a great source for biological, biomedical, and bioinformatics fields research. It aims at providing the most complete compendium of all known protein sequence data (experimentally verified, or computationally predicted) and functional annotation. All the reviewed and curated Swiss-Prot entries

[19] The supplementary material of this paper is not in the content of this dissertation

are combined by UniProt Knowledgebase (UniProtKB), with the unreviewed TrEMBL entries generated by in silico methods.

All UniProt databases are some of the most accessible, being distributed in numerous formats for download, such as XML, RDF, plain text, GFF, Excel tables, Tab-separated, and FASTA directly from the website (<https://www.uniprot.org/>). In addition, all data can be retrieved from an FTP interface (<ftp://ftp.uniprot.org>) and also through a public RESTful API capable of performing complex queries by the SPARQL API endpoint.

This extensible accessibility is proof that the UniProtKB fully supports the FAIR (Findable, Accessible, Interoperable, and Reusable) data principles. This summed to the scale of the databases, approximately 190 million (despite some sequence redundancy), and the fact that it is updated every eight weeks, makes UniProt a database of ultimate importance.

1.9.6 SWISS-MODEL

SWISS-MODEL Repository (SMR) (BIENERT et al., 2017) is an example of a database of automatically generated 3D models. The structural prediction of protein is based on its sequence, using homology modeling. These structure assessments are made using a repository of annotated protein structures and their sequences as templates, with a sequence identity of at least 30% with the input sequence.

The models built by this pipeline are added to the SMR every 15 minutes, and it tries to avoid redundancy by prioritizing the inclusion of longer sequences and the highest model quality (measured by QMEANS (BENKERT; BIASINI; SCHWEDE, 2011)). With almost 20% of SwissProt/UniProtKB entries, SMR contains more than 400,000 high-quality models.

SMR functionalities and the established models are accessible through the web page <https://swissmodel.expasy.org/repository> featuring a simple but complete search box. Queries to the repository or the pipeline can also be made programmatically through a RESTful API.

1.9.7 SCOP2

The original purpose of the Structural Classification of Proteins (SCOP) (ANDREEVA et al., 2019) database model was to organize and categorize proteins according to their structural and evolutionary relationships. It has been a reference database for structure annotation, but its tree-based model became unsuccessful in annotating some outlier proteins and some previously unknown relationships. So it was redesigned into the SCOP2 database (ANDREEVA et al., 2013).

SCOP2 (<http://scop2.mrc-lmb.cam.ac.uk/>) uses an acyclic graph-based categorization that allows different ramifications to the evolutionary and structural relationships. It is

also an expert-curated database and uses the PDB as a source of protein structures. There are four main categories: protein relationships (which include structural, evolutionary, and 'Other'), evolutionary events, protein types, and structural classes. SCOP2 also counts with the classical evolutionary levels from the original SCOP, Species, Protein, Family, and Superfamily.

1.10 Conclusion

Many resources such as methods, tools, and databases are available to perform main tasks in the context of protein structure bioinformatics. However, they are commonly scattered across different online repositories, making it not straightforward which topics should be learned/used and where these topics could be accessed. This task can be time-consuming, especially for those beginning in the field of bioinformatics.

In this paper, we covered the main subareas of protein structure bioinformatics, presenting, for each subarea, a succinct definition and a set of tools frequently used to address tasks in the subarea. Each tool is described in terms of the main ideas behind how it works or its algorithm, positive aspects, and drawbacks. As a complementary resource to the paper, we developed a website that allows users to retrieve the online resources described here by selecting a keyword from the wordle visualization or submitting a search term of interest. Our tool applies TF–IDF to measure and rank the importance of a search term for each resource covered in this paper, based on their title and abstract. The user can download the results in BibTeX or CSV format.

In future work, we intend to add to the proposed website the datasets, algorithms, tools, and online courses that we have been developing since 2014 in the context of the Structural Bioinformatics of Proteins (Babel) project, which involves 6 universities in the northeast, southeast and south of Brazil and was financed by Brazilian public funding institutions. This material was proposed to create a path in the area of structural bioinformatics of proteins to train our students during the project and also as a result of the research developed. We believe that making this resource available for free would be a significant contribution to society.

Chapter 2

Computational prediction of potential inhibitors for SARS-COV-2 main protease based on machine learning, docking, MM-PBSA calculations, and metadynamics

The development of new drugs is a very complex and time-consuming process, and for this reason, researchers have been resorting heavily to drug repurposing techniques as an alternative for the treatment of various diseases. This approach is especially interesting when it comes to emerging diseases with high rates of infection, because the lack of a quickly cure brings many human losses until the mitigation of the epidemic, as is the case of COVID-19. In this work, we combine an in-house developed machine learning strategy with docking, MM-PBSA calculations, and metadynamics to detect potential inhibitors for SARS-COV-2 main protease among FDA approved compounds. To assess the ability of our machine learning strategy to retrieve potential compounds we calculated the Enrichment Factor of compound datasets for three well known protein targets: HIV-1 reverse transcriptase (PDB 4B3P), 5-HT2A serotonin receptor (PDB 6A94), and H1 histamine receptor (PDB 3RZE). The Enrichment Factor for each target was, respectively, 102.5, 12.4, 10.6, which are considered significant values. Regarding the identification of molecules that can potentially inhibit the main protease of SARS-COV-2, compounds output by the machine learning step went through a docking experiment against SARS-COV-2 M^{pro}. The best scored poses were the input for MM-PBSA calculations and metadynamics using CHARMM and AMBER force fields to predict the binding energy for each complex. Our work points out six molecules, highlighting the strong interaction obtained for M^{pro}-mirabegron complex. Among these six, to the best of our knowledge, ambenonium has not yet been described in the literature as a candidate inhibitor for the SARS-COV-2 main protease in its active pocket.

2.1 Introduction

During an outbreak, it is necessary to quickly respond to the unknown pathogen to avoid the uncontrolled spread of the disease. Just like what happened with the novel COVID-19 disease, caused by the Severe Acute Respiratory Syndrome Coronavirus 2 (SARS-COV-2), that appeared for the first time in Wuhan, China at the end of 2019, spreading quickly all over the globe. In approximately one year, the infection reached more than 170 million cumulative confirmed cases and caused over than 3.5 million deaths [1] (as of Jun 2021). Considering that the development of new drugs is expensive and time-consuming, in this scenario computational strategies can potentially speed up the process of drug discovery. The repurposing of existing drugs to treat new diseases can accelerate the approval process, giving a quick response against unknown pathogens (GALINDEZ et al., 2021). The use of the antiviral Remdesivir is an example, as this drug was initially indicated for the treatment of Ebola, and in October 2020, FDA approved it for use against COVID-19 (VEKLURY, 2021). However, the mortality rate for patients treated with Remdesivir is still quite high and does not differ significantly from placebo treatment in clinical trials (BEIGEL et al., 2020; SPINNER et al., 2020). This shows that treatment with this antiviral alone is still not enough and further research to identify other promising drugs should continue.

Drug development often starts with the identification of key molecules, generally proteins, also called targets, which are crucial for the treatment of a specific disease. High-throughput screening (HTS) experiments are performed aiming to identify compounds that interact with a target protein to achieve a biological purpose, and these compounds can be used as drug candidates. However, designing HTS experiments is expensive, which requires time and resources such as advanced laboratories having chemical and biological libraries (RIFAIOGLU et al., 2019; GIMENO et al., 2019). The problem increases when considering the drug development as a whole since the cost of all process is upwards of US$2.8 billion (EKINS et al., 2019). Moreover, there is a lack of correlation between *in vivo* and *in vitro* assays, since HTS is not able to predict clinical failures, such as side effects and toxicity problems. To address these challenges, computational methods, such as virtual screening (VS), have been developed to mitigate costs and improve productivity in drug development (WU et al., 2020; AHMED; QUADEER; MCKAY, 2020; JIN et al., 2020; ZHANG et al., 2020).

VS is a computational approach in drug discovery, which aims to predict drug-like small molecules that can bind a drug target, generally, a protein (GONCZAREK et al., 2018). The VS pipeline has the potential to highly reduce the cost and time required for HTS experiments, discarding unlikely compound-target pairs and, as result, potential active combinations are selected (RIFAIOGLU et al., 2019). The binding affinity between a protein target and a ligand candidate is assessed by scoring functions associated with the method, in which potential ligands are

[1] <https://covid19.who.int/>

ranked according to their binding capability. There are two main approaches of VS, structure-based and ligand-based. These can also be combined, generating a hybrid approach(LUO et al., 2020).

The structure-based virtual screening (SBVS) uses structural information of a protein target, such as biding pockets, to dock a ligand candidate to obtain energy predictions. It is highly dependent on the three-dimensional structure of the protein, which can be considered as a limitation because there is a considerable gap between the availability of protein sequences and protein structures. According to The UniProt Consortium, there are 189 million sequence records in Uniprot, a catalogue of all known protein sequences. On the other hand, in Protein Data Bank (PDB), the catalogue of all known structures, there are 175,000 macromolecules (BURLEY et al., 2021). Therefore, depending on the target, it is necessary to create a protein structure model first, which can potentially impact on the quality of the final result (CHENG et al., 2007; ANDERSON, 2003). Furthermore, docking techniques may disregard dynamic features of the protein and, thus, the three-dimensional structure may not represent the most favorable conformation for binding a ligand that would be highly favorable in another thermodynamic state. For these reasons, structure-based methods are more commonly used for optimization, along with some other techniques, such as molecular dynamics (ŚLEDŹ; CAFLISCH, 2018). The primary characteristic that distinguishes SBVS tools is how spatial conformations are ranked, according to the score function. There is a current trend to replace classical score functions with more sophisticated methods involving machine learning, to increase the accuracy of the tools (LI et al., 2021). Thus, recent methods look for improvements in the calculation of score functions, such as AtomNet (WALLACH; DZAMBA; HEIFETS, 2015), which uses deep convolutional neural network, and SIEVE (YASUO; SEKIJIMA, 2019), which uses features of intermolecular binding energy.

Another popular class of methods to perform virtual screening is the ligand-based, which explores chemical and structural features among small molecules to identify compounds having a similar pharmacological activity to the protein. Therefore, it requires a large volume of known data for the specific target, opening space for the employment of a variety of machine learning methods (ACHARYA et al., 2011). The main algorithms used for this purpose are support vector machines, k-nearest neighbors, random forests, genetic algorithms, and artificial neural networks, which can be combined (BANEGAS-LUNA; CERÓN-CARRASCO; PÉREZ-SÁNCHEZ, 2018; GIMENO et al., 2019; MENDOLIA et al., 2020). There are some data structures usually employed to represent molecules in ligand-based virtual screening (LBVS) algorithms: graphs, based on the molecular structure (LEŠNIK et al., 2015; GARCIA-HERNANDEZ; FERNANDEZ; SERRATOSA, 2019), fingerprint vectors, which characterize the ligand in terms of its physicochemical and structural properties (CERETO-MASSAGUé et al., 2015; SHANG et al., 2017; MENDOLIA et al., 2020), and ACFs (Atom-Centered Fragments) (KÜHNE; EBERT; SCHÜÜRMANN, 2009; ZHENG et al., 2014). We can highlight some examples involving each of the structures. LiSiCA (LEŠNIK et al., 2015) is a software that applies the click algorithm

to find and rank similarities between pairs of two or three-dimensional molecules, represented in the shape of graphs. The LBVS (ZHENG et al., 2014) is a web server that uses BindingDB (GILSON et al., 2016) and ChEMBL (DAVIES et al., 2015) as data sources for the Naive Bayes classifier, whose features are ACFs (Atom-Centered Fragments). Another web server is HybridSim-VS which performs virtual screening ligand prediction based on 2D and 3D fingerprints to calculate the similarity between molecules (SHANG et al., 2017).

SARS-COV-2 is an enveloped RNA positive genome virus with about 30kb that encodes two polyproteins, which are cleaved to constitute the 16 non-structural proteins and four structural proteins, essential to virus cycle (WU et al., 2020; MIRZA; FROEYEN, 2020). The polyproteins cleaveage is mostly done by the main-protease (M^{pro}) which makes this protein a potential drug target to inhibit virus replication (JIN et al., 2020).

In this work, we combine an in-house developed machine learning strategy, docking, MM-PBSA calculations, and metadynamics to identify molecules approved by the Food and Drug Administration (FDA) that can potentially act as inhibitors of the enzyme activity of the main protease of SARS-COV-2, which is relevant for interrupting the viral cycle. Our computational strategy is hybrid as it combines LBVS and SBVS: (i) We devised a machine learning strategy that couples different molecule fingerprints to perform a first step of LBVS. (ii) Next, the resulting molecules went through SBVS steps. These molecules are docked against the target protein (SARS-COV-2 M^{pro}) using Autodock Vina. The poses with lowest vina scores show the preliminary favorable fit of the ligand inside the M^{pro} active site. Finally, we selected the best-scored poses to perform Molecular Mechanics Poisson-Boltzmann Surface Area (MM-PBSA) and metadynamic simulations using two distinct force fields, CHARMM and AMBER. These simulations intended to predict the binding energy for each complex and inspect the impact of the force field on this kind of simulation.

This work aims to identify, through computational techniques, promising molecules to interact with the target protein. The results obtained here serve as input for subsequent in vitro assays to validate the inhibition potential suggested by in silico experiments. All the source code and input datasets are available at <https://github.com/IsabelaGomes/Prediction_SARSCOV2_inhibitors>.

2.2 Materials and methods

This section details our computational strategy to predict candidate inhibitors for SARS-COV-2 M^{pro}. (Fig 1) presents the workflow that outlines the process. We explain data collection and preprocessing, prediction of molecules through the in-house developed supervised learning strategy, and refinement of these molecules through docking and molecular dynamics.

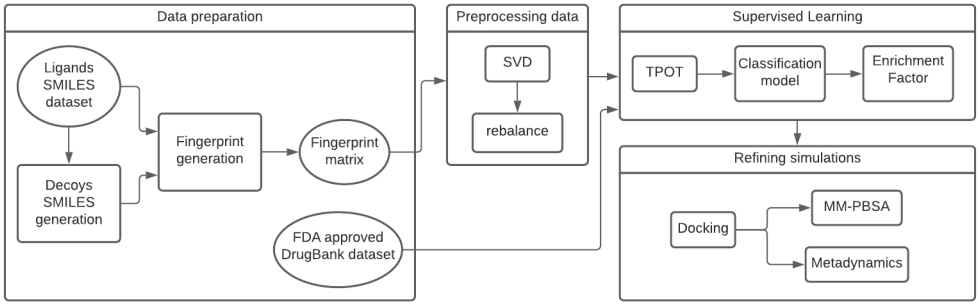


Figure 1 – The workflow of our strategy. It is composed of four blocks: Data preparation; Preprocessing data; Supervised learning; and Refining simulations. Rectangles indicate processing steps and ellipsoids denote datasets.

2.2.1 Data preparation

Our strategy intends to predict small molecules to favorable interact with a protein in an active pocket. As a first step, it is necessary to collect robust data to feed the machine learning strategy.

Only positive samples are insufficient to build a robust predictive model and, to make a model that correctly filters non-active compounds, is important to have inactive molecules in the dataset. One challenge in molecule screening modeling is how to label compounds as inactive since the range of inactive molecules is very wide, so we use decoys to fulfill this task. We generated 50 molecular decoys for each active ligand using the "Database of Useful Decoys: Enhanced" server (DUD-E) (MYSINGER et al., 2012).

An important step in machine learning is feature engineering, where the raw data is converted into a more informative value, called a feature. In the drug repurposing context, we aim to convert the chemical data into an informative numerical feature understandable for predictive models. This is achieved by computational processing for molecular fingerprint generation, which are numerical features extracted from chemical structures that can be used to assess the similarity of two compounds.

The structural similarity of two molecules is commonly evaluated by computing the Tanimoto coefficient on their chemical fingerprints (CERETO-MASSAGUÉ et al., 2015). High values indicate two compounds are similar but do not provide information about specific chemical groups they share. This is because the compounds are compared one by one, resulting in a local point of view, unlike when compounds are analyzed together, achieving a global view. In this scenario, we decide to use techniques to increase the overall quality of the data and to help the machine learning strategy better capture the relationships between features, beyond the Tanimoto coefficient.

The active ligands and the decoys composes our dataset and we described the molecule features through a combination of two molecular fingerprints, implemented by RDKit (THE. . . ,

2021). The first of them is a topological-based fingerprint inspired by Daylight fingerprint that extracts chemical patterns from the chemical graph, and the other one is a structure keys-based fingerprint implementation of the public MACCS keys, indicating the presence or absence in the compounds of particular substructures or features from a given list of structural keys (CERETO-MASSAGUÉ et al., 2015). Thus, we built a data matrix containing numerical information about active molecules and decoys on the rows and fingerprint features on the columns. The resulting matrix is composed of 2214 columns and used as input to build the supervised machine learning model.

2.2.2 Preprocessing data

The next step concerns processing the data in order to improve the predictive performance. Usually, the matrix data generated by fingerprints is sparse, because molecules have different sizes and functional groups. This results in many cells filled with zeros. Moreover, the assumption that similar molecules share similar activities might not always be valid, since minor modification on functional groups causes an abrupt change in activity (LO et al., 2018). Small changes of descriptor values can lead to considerable changes in molecular properties. Thus, before using raw fingerprints as input for a machine learning strategy, we perform singular value decomposition (SVD) that allows us to obtain lower-dimensional projections of the data (ZAKI; MEIRA JR, 2014). The SVD reduces the dimensionality while preserving essential properties of the raw data matrix, detecting meaningful patterns while discarding weak signals and noises (WALL; RECHTSTEINER; LM., 2003).

As 50 decoys were generated for each active molecule, there is an imbalance between actives and decoys, that is, positive and negative classes. Imbalanced data refers to a dataset in which one or more classes have much grater number of examples than others (HAIXIANG et al., 2017). In drug discovery, the prevalent class, called majority class, comprises inactive molecules, while the rarest class, or minority class, is composed by active compounds. Generally, imbalanced data significantly challenges traditional machine learning models. To address this issue, a rebalance is performed before building the learning model, aiming to attain a more balanced input data and improve the model prediction capacity. We use a resampling technique to rebalance the sample space in order to relieve the effect of the skewed class distribution in the learning process. Given an imbalanced dataset, we split the majority class set into subsets similar in size to positive class. We combine each negative subset with the positive one, and create many smaller balanced subdatasets that will be used as input to an ensemble of classifiers, similar to the strategy implemented in (SANTANA et al., 2020b).

2.2.3 Supervised learning

The matrix of active compounds and their respective decoys for each target after pre-processing data served as training data to build a prediction model. To select an appropriated learning model and optimize its parameters for each dataset, we used TPOT (LE; FU; MOORE, 2020), an automated machine learning system written in Python. The pipelines were trained and evaluated under 5-fold cross validation using scikit-learn library in Python (PEDREGOSA et al., 2011).

Nearly 2,000 approved molecules were obtained from Drugbank (WISHART et al., 2018) and screened against each target using the best pipeline according TPOT. This set of approved drugs was used as a validation set, aiming to verify if the best machine learning pipeline is able to retrieve active compounds from Drugbank FDA approved drugs. After screening, Drugbank approved compounds were ranked according to the class probabilities returned by the respective pipeline. Best ranked compounds can be selected as candidates for future analysis in more robust experiments, such as molecular docking and molecular dynamics. The metric used to assess the performance of the VS strategy is the Enrichment Factor (EF), which is a measure of how much the compound dataset is enriched with actives after applying the VS strategy (GIMENO et al., 2019). Usually, a sample of one percent (top-1%) of the original data, after labeling probabilities, containing the best classified compounds is used to calculate the EF. This metric is computed as the ratio between the proportion of actives after and before applying the VS strategy.

$$EF = \frac{\frac{a2}{a2+d2}}{\frac{a1}{a1+d1}} \tag{2.1}$$

in which a1 = actives before the VS step, a2 = actives after the VS step, d1 = decoys before the VS step, and d2 = decoys after the VS step.

For our search for potential SARS-COV-2 M^{pro} inhibitors, we collected a benchmark on April 24, 2020 in PDB (BERMAN et al., 2000a) consisting of 74 SARS-COV-2 main protease structures in complex with diverse ligands. The best machine learning algorithm calculated by TPOT under a 5-fold cross validation was the MLP Classifier, with parameters: alpha=0.0001, learning_rate_init=0.1, max_iter=500, hidden_layer_sizes=(100,75,50,25,).

2.2.4 Docking

After using our VS strategy on Drugbank FDA approved drugs, we selected 17 compounds candidates to interact against SARS-COV-2 M^{pro}, excluding those that act on the central nervous system or are illegal – allowed to medical use under specific conditions. We executed the local docking focusing on the binding site: H41, S46, M49, Y54, F140, L141, N142, G143,

C145, H164, M165, E166, L167, P168, H172, A173, F185, D187, Q189, T190, A191, and Q192 (JIN et al., 2020). The grid box was elaborated using Autogrid from AutoDock Tool (MORRIS et al., 2009). We added polar hydrogen to the macromolecule, intended to correct any deprotonations of the structure, and rotation was allowed to all rotatable bonds of the ligands. Every molecule was saved in pdbqt format so they could be used as input for AutoDock Vina (TROTT; OLSON, 2010). We executed three simulations replica and we set the tool to generate seven poses for each drug in each replica. The best scored complexes were analyzed by Biovia Discovery Studio 2020 (BIOVIA, 2017) to describe the contacts profile.

Docking indicates preferable areas where the ligands interact with the protein. However, this is a static result that can be improved with molecular dynamics approximating the system to real conditions. Moreover, the docking score function does not reflect the the binding energy, which can be estimated by molecular dynamics methods such as MM-PBSA and metadynamics.

2.2.5 MM-PBSA simulations

We selected the six best-scored poses from docking for molecular dynamics simulations, MM-PBSA and metadynamic, using two distinct force fields: CHARMM (BJELKMAR et al., 2010) and AMBER (SORIN; PANDE, 2005). These simulations intended to predict the binding energy for each complex and inspect the impact of the force field on this kind of simulation. Using two different ways to estimate the interaction power of the two molecules allows us to view the complex binding from more than one perspective. The use of distinct force fields allows us to see if even with different parameters governing the molecular trajectories, the tendency of the interaction force is maintained.

We performed the MM-PBSA on GROMACS 5.1.4 (SPOEL et al., 2005b). We followed the SwissParam (ZOETE et al., 2011) protocol to generate drugs topologies with the CHARMM force field (BJELKMAR et al., 2010) and antechamber (WANG et al., 2006) to generate drugs topologies with the AMBER force field (SORIN; PANDE, 2005). Each system was centered on a dodecahedral box with a distance of 14 Å between the complex and the edges, solvated with TIP3P water, and neutralized with 0.15 mol.L^{-1} sodium chloride.

For energy minimization, we applied the steepest descent (MORSE; FESHBACH, 1953) and conjugate gradient (HESTENES; STIEFEL, 1952) algorithms with convergence criteria of 0.24 kcal.mol^{-1} in both cases. Subsequently, we equilibrated the system in 300 K and 1.0 atm using seven simulations stages with restriction forces being gradually removed. The first stage was in canonical coupling and, the others, in isotherm-isobaric coupling. The simulation protocol consisted of the Verlet Leap-frog integration algorithm (GUSTEREN; F.; BERENDSEN, 1990), with an integration step of 2.0 fs. Also, v-rescale algorithms (BERENDSEN et al., 1984; BUSSI; DONADIO; PARRINELLO, 2007) with a tau_t = 0.1 ps and Parrinello-Rahman (PARRINELLO; RAHMAN, 1981) with a tau_p = 2.0 ps were used for the temperature and pressure

couplings, respectively. For the CHARMM force field, we applied the settings suggested by GROMACS for van der Waals parameters on the molecular dynamics parameters file.

Next, we submitted the complexes to 100 ns of molecular dynamics each, under the same parameters. From the trajectories obtained, we compressed it into 500 snapshots due to MM-PBSA calculation using g_mmpbsa tool (KUMARI; KUMAR; LYNN, 2014) and default parameters.

Since the sample data set is small to compare the energies calculated with both force fields, only six points, we used Spearman's rank correlation to assess the concordance of the results.

2.2.6 Metadynamics

Metadynamics is an appropriate method to determine the binding energy for protein-ligand complexes, mainly when the active site is large and the ligands very small (MARTINS et al., 2021; BRANDT et al., 2016b) and it is a quicker simulation than MM-PBSA. This simulation also allows us to establish different sets of collective variables, which govern the simulation in specific ways. The combination of several of these sets of collective variables describes more clearly the unbinding behavior of the ligand from the active site of the protein.

We used the same six complexes to perform metadynamics simulations with the same force fields used previously, CHARMM, and AMBER. First, each system was solvated using the Solvate v1.7 plugin for Visual Molecular Dynamics software (HUMPHREY; DALKE; SCHULTEN, 1996b) with TIP3P water in a cubic box with 14 Å padding. Next, we neutralized the systems with 0.15 mol.L^{-1} sodium chloride using Autoionize v1.7 plugin for VMD software (HUMPHREY; DALKE; SCHULTEN, 1996b).

We used the NAMD 2.14 software (PHILLIPS et al., 2005a) to perform the simulations with a time-step of 2.0 fs, in an NPT ensemble with Langevin thermostat and barostat devices set at 300K and 1 atm. We set periodic boundary conditions and a cutoff of 12 Å for the non-bonded interactions, and we calculated the long-range electrostatic interactions by the Particle mesh Ewald (PME) method. The systems were minimized for 1000 steps by minimization with conjugate gradient, then a protocol of relaxation with harmonic restrains was performed. This protocol consists of:

1. 500 ps MD with harmonic restrains in the M^{pro} and the ligand;
2. 500 ps MD with harmonic restrains in the backbone of the M^{pro} and ligand;
3. 500 ps MD with harmonic restrains in the M^{pro};
4. 500 ps MD without harmonic restrains;

5. 8 ns MD without harmonic restrains, restarting the velocities;
6. 500 ps MD with harmonic restrains in the backbone of the M^{pro};
7. 500 ps MD with harmonic restrains in the M^{pro};
8. 1 ns MD with harmonic restrains in the M^{pro}, restarting the velocities.

For the actual metadynamics, we restrained the M^{pro} harmonically, and we kept the ligands free for all the systems during the 7 ns. We executed three simulation replica for each ligand with each force field. We set the height of the Gaussians to 0.02 kcal.mol^{-1}, added every 1.0 ps with a width of 1.77. We chose two groups of collective variables (CV):

- **Group 1:** The first CV were the distance between the center of mass of M^{pro} C145 and the center of mass of the closest atoms of the ligand to M^{pro} C145, varying between 0 Å and 30 Å with an amplitude fluctuation of 2 Å. For the second CV, we established the angle between the center of mass of M^{pro} C145, the center of mass of the closest atoms of the ligand to M^{pro} C145 and the center of mass of the entire ligand, varying from 0º to 180º with an amplitude fluctuation of 10º.
- **Group 2:** The first CV were the same distance for group 1. For the second CV, we established specific angles for each ligand, considering their internal coordinate variations, varying from 0º to 180º with an amplitude fluctuation of 10º. The atoms involved in each angle are described in Table 9. The label of each ligand is available on Supporting Information.

Table 9 – Set of atoms for each angle component for the selected ligands

Ligand	**Atoms involved in the CV**		
	Set 1	**Set 2**	**Set 3**
Ambenonium (DB01122)	C145	O,O1,N1,N2,C8,C9, C10,C11	ligand
Plerixafor (DB06809)	C145	N,N1,N2,N3,C7,C8, C9,C10,C11,C12, C13,C14,C15,C16	N4,N5,N6,N7,C18, C19,C20,C21,C22, C23,C24,C25,C26,C27
Revefenacin (DB11855)	C145	H41,H04,H05,O3,N4, C31,C34	ligand
Mirabegron (DB08893)	H1,H2,H11,N1,N2, C8,C9,C10,S	ligand	H19,H20,H21,H22, H23,C15,C16,C17, C18,C19,C20
Diloxanide furoate (DB14638)	beta carbon of C145	O,O1,O2,C,C1,C2	O3,C4,C5,C6,C12
Vorinostat (DB02546)	C145	H1,H2,H14,H15,H16, H17,H18,H19,O1, O2,N1,C10,C11, C12,C13	ligand

We chose this pair of collective variables based on the principal movements of the ligands inside the protein on the equilibrium molecular dynamics stage. These CVs showed better sampling and convergence about the configurations on each minimum before the recrossing.

Despite operating the metadynamics by varying two collective variables, distance (CV_{dist}), and a specific angle (CV_{ang}), the unbinding process is mainly described by pull off the ligand from the protein. For this reason, we can estimate the metadynamics free energy onto the CV_{dist} according to the (Eq 2.2) (BRANDT et al., 2016b).

$$-\beta G(CV_{dist}) = ln\frac{\int e^{-\beta G(CV_{dist},CV_{ang})} dCV_{ang}}{\int e^{-\beta G(CV_{dist},CV_{ang})} dCV_{ang} dCV_{dist}} \tag{2.2}$$

in which β = 1/k_bT, where k_b is the Boltzmann constant 1.9858 x 10^{-3} kcal.mol^{-1}·K^{-1}, T = 300K, and G(CV_{dist}, CV_{ang}) is the free energy value for each distance and angle pair described in the PMF (Potentials of Mean Force) map.

Since Mpro is an enzyme with a shallow and exposed active site, the 7.0 ns of simulation maps conformations of the ligand onto the active site, the bulk, and even allows the molecule to interact with the protein again in other regions. Thus, it is critical to analyze the distance profiles and potential and electrostatic energies over time to select the most appropriate PMF files, since they are constructed by overlapping Gaussians during the simulation. These files describe the energies within the catalytic site and on the bulk to reconstruct the free energy landscape of the system before the ligand explores states beyond the two intended to avoid overestimating the energies in these two states. We calculated the distance in each frame using CPPTRAJ (ROE; CHEATHAM, 2013) and nonbonded energies using NamdEnergy. We analyzed the results with VMD software (HUMPHREY; DALKE; SCHULTEN, 1996b) and Python in-house scripts.

The binding free energy from each ligand with both force fields was determined by the subtraction of the minimum energy of the ligand inside the protein binding site (G^{in}) and the minimum energy of the ligand in the water (G^{out}), without the protein influence according to (Eq 2.3).

$$\Delta G_{bind} = G^{in}_{CV_{ang}}(CV_{dist}) - G^{out}_{CV_{ang}}(CV_{dist}) \tag{2.3}$$

This influence could be noticed from the distance profile, in which we established a cutoff of 20 Å to consider the ligand inside the protein, concomitantly with energy profiles. When the complex energy is null, it indicates no influence of the M^{pro} on the ligand.

For each ligand, we then performed six simulations with each force field. We calculated the final free energy by weighting the free energy of each simulation using the Boltzmann average, which highlights more favorable states, which are the more negative ones. The Boltzmann average applied for our systems can be written as the summation in (Eq 2.4). Thus with it was done for the MM-PBSA essay, we measured the concordance of the two force fields with

Spearman rank correlation.

$$\Delta G_{bind} = -kTln(\sum_{i=1}^{3}\sum_{j=1}^{2}\frac{e^{-G_{R_iCV_j}}}{kT}) \tag{2.4}$$

in which R_iCV_j is the i^{th} replica with the j CV set.

2.3 Results and discussion

This section discusses detailed results for each step of our method. First, we show that our supervised learning strategy is able to enrich compound datasets for three well known targets. Then our strategy is applied to point out potential molecules to inhibit SARS-COV-2 M^{pro} that will be refined through docking and molecular dynamics. Next, molecules that went through docking simulation and their respective scores are presented. Finally, the binding energies calculated through MM-PBSA and metadynamics simulations for two force fiels are shown.

2.3.1 Supervised learning

To execute the methodology and validate it, we used the data of three targets that have a large amount of FDA approved drugs available: HIV-1 reverse transcriptase, 5-HT2A serotonin receptor, and H1 histamine receptor. We searched for specific ligands on BindingDB (GILSON et al., 2016) and removed from the set all drugs deposited on DrugBank, used as method validation. Thus, we established a set of 1,492 ligands for HIV-1 reverse transcriptase (PDB 4B3P), 1,199 ligands for H1 histamine receptor (PDB 3RZE), and 1,974 ligands for 5-HT2A serotonin receptors (PDB 6A94). Based on these data, we followed the data preparation described and generated the matrix for the training set of the predictive model.

Table 10 shows the results achieved by our VS strategy. For each target, we used the TPOT to select the appropriated learning model and ran it on the Drugbank FDA approved data (U..., 2021) to retrieve drugs already used in the treatment of diseases related to these targets. The results were satisfactory, highlighting the HIV-1 target, where an enrichment above a hundred was obtained, in other words, at least a hundred times better than random picking.

Table 10 – Values of enrichment after apply the virtual screening strategy in three different datasets.

Target	Best Algorithm	Drugbank EF
HIV-1 reverse transcriptase	K-Neighbors Classifier	102.5
5-HT2A serotonin receptor	MLP Classifier	12.4
H1 histamine receptor	MLP Classifier	10.6

The enrichment factor (EF) measures how well the VS strategy filtered out positive ligands and is not directly related to the model used. For the HIV target assay, the model returned only three results above the established cutoff line, and for this reason, the EF was significantly high. The results for the other targets were not as good, but are still favorable since they indicate that the strategy selects better results for molecules that interact with the protein than for ligands not yet classified for this purpose. In this case, these other molecules that stood out in the model could be potential study targets to identify if they interact well with the protein.

Observing the results, the next step was modeling and screening compounds from Drugbank FDA approved drugs for the SARS-COV-2 M^{pro}. Ligands in complex with the published X-ray crystal structure for this protein (JIN et al., 2020) can be used as examples in the training model, in order to find other molecules that share similar features, since that compounds with similar structures are expected to have similar biological activities (TROPSHA, 2010). To obtain these ligands we collected 74 SARS-COV-2 M^{pro} complexes from PDB (BERMAN et al., 2000a) and executed our VS, which returned 28 ligands whose probability of being M^{pro} inhibitors is superior to 0.9. Among these, we selected to proceed to the docking step the 17 best-ranked ones, excluding those that act on the central nervous system or are illegal.

2.3.2 Docking

The compounds selected for docking simulation are displayed in Table 11 with their respective vina and our LBVS scores.

The poses with lowest vina scores show the preliminar favorable fit of the ligand inside the M^{pro} active site. The 2D structure of the six best ranked ligands through the docking assay are shown in (Fig 2). We can observe that there is diversity among the ligands and the interactions that they establish (Fig 3). However, some characteristics prevail, such as the aromatic character and the electronegative content due to nitrogen, oxygen, and chlorine atoms. This indicates the predominance of favorable interactions: hydrogen bonds, attractive electrostatic interactions, hydrophobic interactions, and some involving π orbitals. For M^{pro}-vorinostat, there is only one unfavorable donor-donor interaction involving histidine 41. For M^{pro}-ambenonium, there is one unfavorable electrical interaction between histidine 41 and one of the charged nitrogens. Besides, there is another unfavorable donor-donor involving glycine 143. For M^{pro}-plerixafor there are two unfavorable donor-donor interactions, involving serine 46 and glutamic acid 166.

2.3.3 MM-PBSA simulations

We chose the six vina best-scored poses to perform MM-PBSA (Molecular Mechanics Poisson-Boltzmann Surface Area) and (Fig 4) shows the binding energy calculated with the two distinct force fields.

Table 11 – Top-scored ligands for SARS-COV-2 M^{pro} selected for docking calculation in Autodock vina.

Ligand	DrugBank ID	Vina score	LBVS score
Plerixafor	DB06809	-8.2	0.92
Revefenacin	DB11855	-7.6	0.90
Mirabegron	DB08893	-6.7	0.92
Ambenonium	DB01122	-6.1	0.96
Diloxanide furoate	DB14638	-5.9	0.94
Vorinostat	DB02546	-5.8	0.95
Acetarsol	DB13268	-5.4	0.90
Lacosamide	DB06218	-5.3	0.98
Procainamide	DB01035	-5.3	0.98
Alverine	DB01616	-5.0	0.91
Phenacemide	DB01121	-4.9	0.98
Acetaminophen	DB00316	-4.7	0.96
Isoniazid	DB00951	-4.6	0.91
Mephentermine	DB01365	-4.4	0.97
Levmetamfetamine	DB09571	-4.3	0.97
Phentermine	DB00191	-4.1	0.95
Pargyline	DB01626	-4.0	0.98

The choice for a given force field impacts how the atoms interact with one another and how we calculate the dihedrals, which can lead the trajectories in distinct ways, even under the same conditions. This had a considerable influence on complexes' behavior. The M^{pro}-plerixafor complex was the one in which the energy contrasted the most considerably when varying the force field. This may have been due to its structure, which is composed of two wide loops, with nitrogens between carbons, attached to an aromatic ring. Because it is not an usual arrangement, the parameterization varied considerably with the force field and influenced the simulations in a wide range.

The other complexes also presented certain variations, but they were within the expected for MM-PBSA simulations, considering the peculiarities of each force field. It can be seen that the ligands ambenonium, vorinostat, and revefenacin were the least divergent because they are composed of predominantly open chains and without the presence of rarer groups. In the case of ligands mirabegron and dioxanide furoate, the influence on the parameterization that generated the energy difference may be due to the presence of an atom other than carbon in the middle of the aromatic ring, and each force field treats this differently.

The binding energies calculated with the MM-PBSA technique provided the estimated

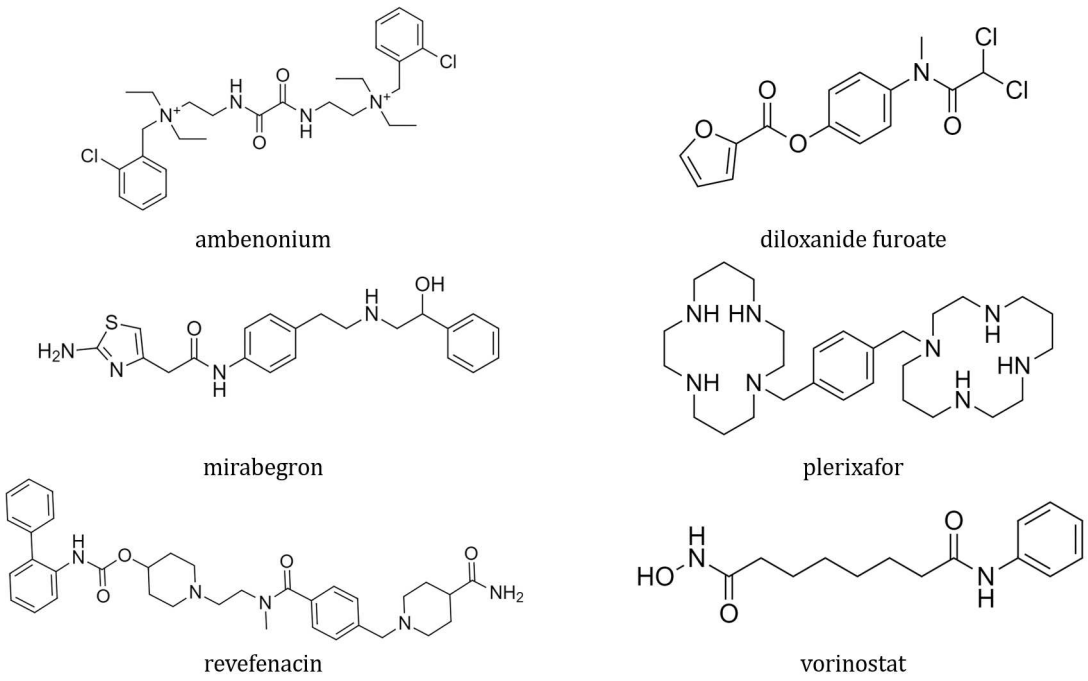


Figure 2 – Compounds highlighted by docking assay. The 2D structure of the best scored ligands in docking experiment.

values with high oscillations, and it was very sensitive to the change of force field. We can see this from the Spearman rank correlation coefficient between the energies calculated based on both force fields, which is 0.143 with p-value 0.787. The detailed energy contributions are available on Supporting Information. Intending to have another perspective, we also calculated the binding energy with metadynamics using both force fields.

2.3.4 Metadynamics

To calculate the binding energy, we choose among the PMF maps generated during each simulation the first one that described both the ligand inside and outside of the active site. This process is necessary because the simulation occurs with Gaussian potentials stacking according to the movements the complex assumes. Then, if we pick a PMF map in a simulation point much later than both events occurred, the energy may not be precise, because more Gaussian potentials will be considered (MARTINS et al., 2021).

Analyzing the distance and energy profiles together to calculate the binding free energy we can expect that the ligands did not have the same cutoff distance to indicate whether they were in the protein site or not, even though they were the same ligand but in different replicas. Considering the energies in this process makes the calculation more accurate, despite this cutoff variation. An example of these profiles of what was obtained for all simulations is available on (Fig 5). It represents the energy behavior along CV_{dist} for M^{pro}-mirabegron. The other profiles are available in the Supporting Information.

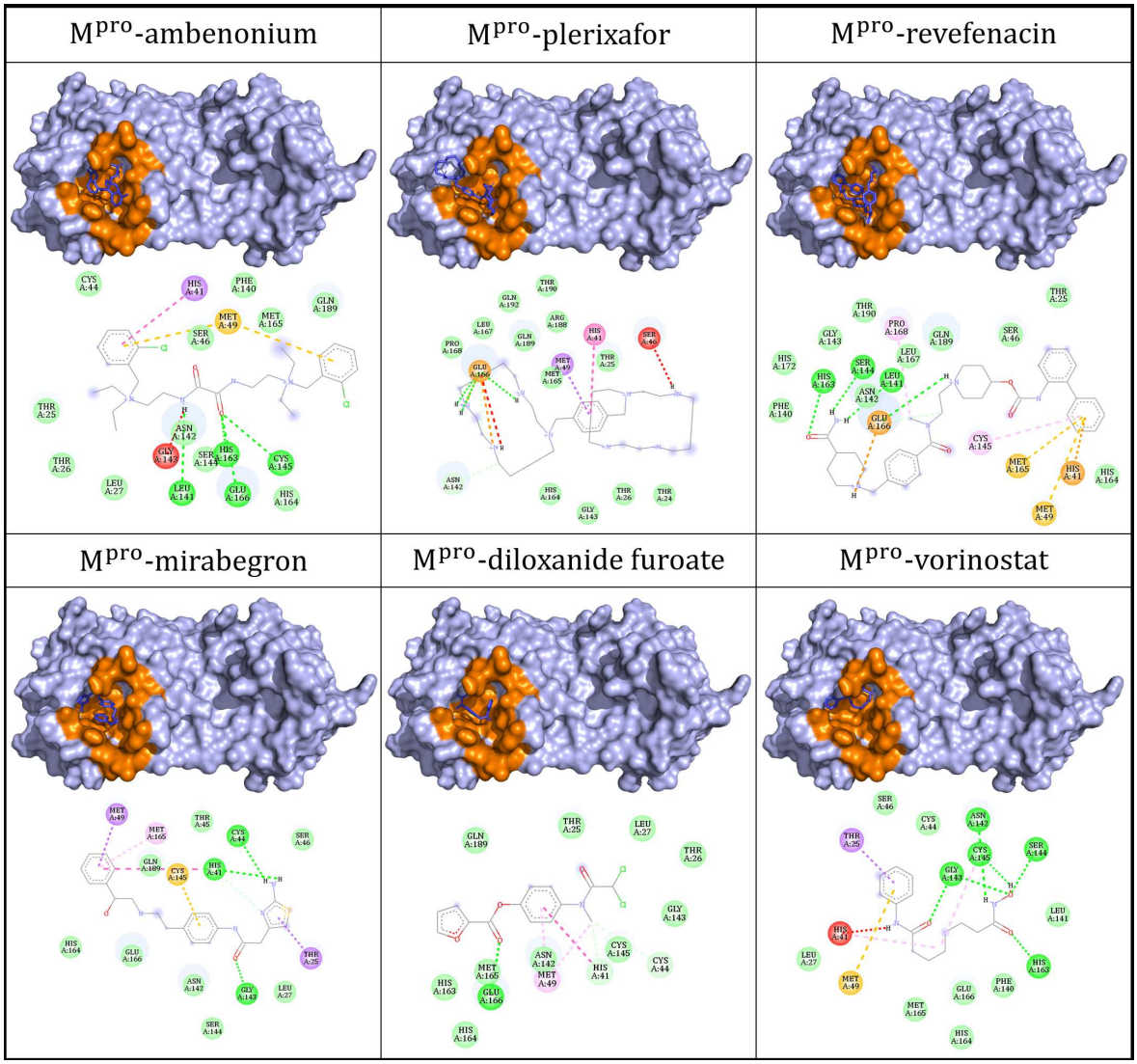
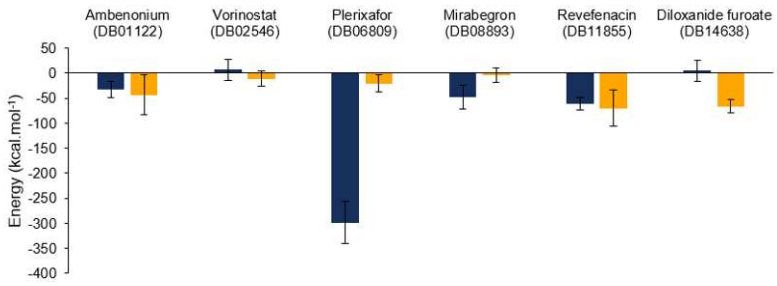


Figure 3 – Top-scored complexes calculated by Autodock vina. The ligands (in blue) are well fitted in the M^{pro} active site (in orange). Hydrogen bonds are represented by green dashed lines; attractive electrostatic interactions, as orange dashed lines; orbital π interactions, as pink, yellow and purple dashed lines; unfavorable interactions, as red. Residues involved in hydrophobic interactions are represented by light green circles.



	CHARMM		AMBER	
	energy	std error	energy	std error
DB01122	-32	0.70	-42	1.8
DB02546	6.7	0.91	-11	0.70
DB06809	-298	1.9	-21	0.78
DB08893	-48	1.0	-3.9	0.65
DB11855	-61	0.56	-70	1.6
DB14638	5.1	0.92	-66	0.60

Figure 4 – Binding energy on MM-PBSA calculation for the complexes. CHARMM is represented in blue and AMBER, in yellow. The error bars represents the standard error.

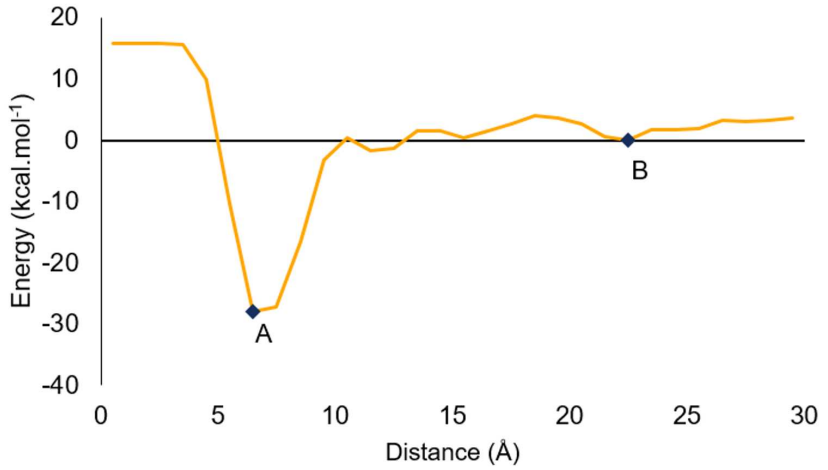


Figure 5 – Energy profile for M^{pro}-mirabegron in its first replica with CHARMM and the first CV set. The points A and B represents, respectively the minimum energy inside the active site and in the water.

Performing the calculation based on (Eq 2.4), we considered different modes of unbinding the ligand from the protein, some less energetic than others, but the influence of each is considered proportionally. It is important to emphasize that it is not possible to mix simulations executed with different force fields, because the parametrization influences translational and rotational movements of the molecules during the trajectory.

This possibility of different behavior according to the force field applied is the reason why the calculated energy, in each case, have a divergency, as already shown in MM-PBSA experiment. Table 12 explores the relation of the calculated energies with CHARMM and AMBER.

Table 12 – Relation between calculated energies with CHARMM and AMBER. The energies described are in kcal.mol^{-1}.

Ligand	**CHARMM**		**AMBER**	
	Energy	**std error**	**Energy**	**std error**
Ambenonium (DB01122)	-10	2.0	-17	3.0
Vorinostat (DB02546)	-15	1.2	-10	2.0
Plerixafor (DB06809)	-20	3.7	-20	3.4
Mirabegron (DB08893)	-35	8.7	-26	6.3
Revefenacin (DB11855)	-19	5.9	-17	4.4
Diloxanide furoate (DB14638)	-8.2	0.66	-13	1.6

Metadynamics had a more concordant behavior between the force fields compared to the MM-PBSA experiment. The Spearman rank correlation coefficient is 0.714 (p-value 0.111) when all points are considered. However, by removing the outlier which is the point corresponding to the complex M^{pro}-vorinostat, this value reaches 0.9 (p-value 0.037) indicating that CHARMM and AMBER energy ranks are well correlated.

For CHARMM and AMBER, compound DB08893 (mirabegron) stands out from the

others, and seems to be promising. Next, we notice DB06809 (plerixafor), which formed the lowest energy complex in MM-PBSA using CHARMM and the second-best ranked in the docking score. The other molecules do not follow a definite pattern for both force fields.

Another compound that we can highlight is DB01122 (ambenonium), a molecule that, to the best of our knowledge, had not yet been identified in any other study involving possible inhibitors for SARS-COV-2 M^{pro} in its active pocket. Although it was not the top-ranked compound in any of the docking and molecular dynamics simulations, it showed favorable interaction in all simulations. For these reasons, this ligand also ranks among the promising ones for further assays to prove inhibition activity.

Comparing all the steps performed, shown in Table 13, we can notice that the ligands end up changing their ranking during the virtual screening. The main difference occurred in the classifications of DB06809 (plerixafor) and DB11855 (revefenacin), which had the lowest LBVS scores of the set of the six best ligands but stood out in the other SBVS simulations. The opposite occurred with DB02546 (vorinostat) and DB14638 (diloxanide furoate). This variation demonstrates the relevance of combining ligand and structure based strategies for efficient inhibitor prediction, considering several positive aspects of both techniques.

Table 13 – Ranking of compounds for each step of our method.

Supervised learning	**Docking**	**MM-PBSA CHARMM**	**MM-PBSA AMBER**	**Metadynamic CHARMM**	**Metadynamic AMBER**
DB01122	DB06809	DB06809	DB11855	DB08893	DB08893
DB02546	DB11855	DB11855	DB14638	DB06809	DB06809
DB14638	DB08893	DB08893	DB01122	DB11855	DB01122
DB08893	DB01122	DB01122	DB06809	DB02546	DB11855
DB06809	DB14638	DB14638	DB02546	DB01122	DB14638
DB11855	DB02546	DB02546	DB08893	DB14638	DB02546

2.4 Conclusion

This work brings together an in-house developed machine learning strategy, docking, MM-PBSA calculations and metadynamics to identify FDA approved molecules that can potentially inhibit the main protease of SARS-COV-2. First, we devised a machine learning strategy that couples different molecule fingerprints to perform a first step of LBVS. Next, the resulting molecules go through SBVS steps, which consists of docking these molecules against the target protein (SARS-COV-2 M^{pro}) using Autodock Vina and then selecting the poses with lowest vina scores. Finally, we selected the best-scored poses to perform MM-PBSA calculations and metadynamic simulations using CHARMM and AMBER force fields to predict the binding energy for each complex.

To assess the ability of our in-house developed machine learning strategy to retrieve potential candidates for the molecular docking and molecular dynamics, the Enrichment Factor was used, which is a measure of how much the compound dataset is enriched with active molecules after applying the screening strategy. Data for three well-studied targets, HIV-1 reverse transcriptase (PDB 4B3P), 5-HT2A serotonin receptor (PDB 6A94), and H1 histamine receptor (PDB 3RZE) went through screening using the proposed machine learning strategy and the EF was, respectively, 102.45, 12.4, 10.6, which are considered significant values. Regarding the identification of molecules approved by the FDA that can potentially act as inhibitors of the enzyme activity of the main protease of SARS-COV-2, this work points out six molecules, highlighting the strong interaction obtained for M^{pro}-mirabegron complex. Among these six molecules, to the best of our knowledge, ambenonium has not yet been described in the literature as a candidate inhibitor for the SARS-COV-2 main protease in its active pocket.

The computational prediction of molecules that can potentially inhibit SARS-COV-2 main protease does not give the certainty that the top-ranked compounds inhibit the target, as it is an entirely computational experiment. Thus, as future work we envision to select the promising candidates for in vitro and in vivo studies that can show the inhibitory action of the proposed molecules for the target of interest.

Chapter 3

Propedia: a database for protein-peptide identification based on a hybrid clustering algorithm

Background: Protein-peptide interactions play a fundamental role in a wide variety of biological processes, such as cell signaling, regulatory networks, immune responses, and enzyme inhibition. Peptides are characterized by low toxicity and small interface areas; therefore, they are good targets for therapeutic strategies, rational drug planning and protein inhibition. Approximately 10% of the ethical pharmaceutical market is protein/peptide-based. Furthermore, it is estimated that 40% of protein interactions are mediated by peptides. Despite the fast increase in the volume of biological data, particularly on sequences and structures, there remains a lack of broad and comprehensive protein-peptide databases and tools that allow the retrieval, characterization and understanding of protein-peptide recognition and consequently support peptide design. **Results:** We introduce Propedia, a comprehensive and up-to-date database with a web interface that permits clustering, searching and visualizing of protein-peptide complexes according to varied criteria. Propedia comprises over 19,000 high-resolution structures from the Protein Data Bank including structural and sequence information from protein-peptide complexes. The main advantage of Propedia over other peptide databases is that it allows a more comprehensive analysis of similarity and redundancy. It was constructed based on a hybrid clustering algorithm that compares and groups peptides by sequences, interface structures and binding sites. Propedia is available through a graphical, user-friendly and functional interface where users can retrieve, and analyze complexes and download each search data set. We performed case studies and verified that the utility of Propedia scores to rank promissing interacting peptides. In a study involving predicting peptides to inhibit SARS-COV-2 main protease, we showed that Propedia scores related to similarity between different peptide complexes with SARS-COV-2 main protease are in agreement with molecular dynamics free energy calculation. **Conclusions:** Propedia is a database and tool to support structure-based rational design of peptides for special purposes. Protein-peptide interactions can be useful to predict, classifying and scoring complexes or for designing new molecules as well. Propedia is up-to-date as a ready-to-use webserver with a friendly and resourceful interface and is available

at: <http://bioinfo.dcc.ufmg.br//propedia>

3.1 Background

Peptides are short chains of amino acid residues connected by peptide bonds that act in cell signaling and as immune modulators, among other important functions. It is estimated that between 15% to 40% of all protein-protein interactions in cells are mediated by these molecules (NEDUVA et al., 2005). Additionally, peptides are structurally diverse, versatile, induce low resistance with limited nontarget activity and can be modulated to interact with specific cellular targets, making them good therapeutic agents (LIU et al., 2019). However, their short half-life and poor oral bioavailability has discouraged the search for peptides as therapeutics in the past (LAU; DUNN, 2018).

With the recent emergence of new synthetic approaches that permit changes in the biophysical and biochemical properties of peptides, these molecules are once again being considered as drug candidates (ANGELOVA et al., 2019; LEE et al., 2019; VINOGRADOV; YIN; SUGA, 2019). In fact, over 60 peptide drugs have been approved in major pharmaceutical markets and hundreds of others are in active clinical development at the moment (LAU; DUNN, 2018). Peptide-like inhibitors are used as well to treat cancer, diabetes, and autoimmune diseases and have high success rates in commercial development (PANT et al., 2020). Multiple next-generation drug candidates (derived from exenatide, a synthetic form of a natural 39-amino acid peptide secreted by *Heloderma suspectum*), have been proposed as therapeutic agents for type 2 diabetes mellitus (LAU; DUNN, 2018).

Understanding the structure and recognition of protein-peptide complexes may aid the design of novel peptides and peptide-based compounds for drug development or biotechnological purposes. Databases of protein-peptide complexes can pave the way for the analysis and comprehension of the mechanisms of protein-peptide recognition. There are several peptide databases, with varied purposes, as databases of bioactive peptides (WANG et al., 2018), antimicrobials (WANG; LI; WANG, 2016), cell penetrating peptides (GAUTAM et al., 2012), hemolytic peptides (GAUTAM et al., 2014), etc. (WANG et al., 2018). Here, we briefly review some representative examples of protein-peptide databases.

London and colleagues (LONDON; MOVSHOVITZ-ATTIAS; SCHUELER-FURMAN, 2010) in 2010 proposed PeptiDB, comprising 103 high-resolution peptide-protein complex structures. It was proposed as a nonredundant set of high resolution complexes to investigate the structural bases of interactions between proteins and peptides and to improve understanding binding strategies for short peptides (5 - 15 residues).

Also in 2010, Vanhee Vanhee *et al.* (VANHEE et al., 2010) devised PepX, comprising protein-peptide complexes clustered based on binding interfaces. It was updated in 2014 for

the last time (505 unique protein-peptide interface clusters from 1431 complexes) and is not available anymore.

Das *et al.*, in turn, proposed PepBind (DAS et al., 2013) in 2013 as a curated set of 3100 protein-peptide complexes clustered according to structure determination methods and manually curated for cellular activity of complexes. The authors mentioned that there was a web interface but it seems to no longer be available.

More recently, in 2018, Frappier *et al.* (FRAPPIER; DURAN; KEATING, 2018) presented PixelDB a database that comprises 1966 non-redundant high-resolution complexes. Entries are clustered based on structural similarities of receptors and then on binding modes. The authors claim to identify conserved peptide core structural motifs. We found a version of this database on GitHub updated 3 years ago.

Wen *et al.* (WEN et al., 2019) released PepBDB also in 2018 and this database is available through a web interface and for download. It contains 13,299 complexes and was last updated in March 2020. The web interface presents the whole list and an individual interactive visualization of the 3D interface and a 2D plot of hydrogen bonds and hydrophobic interactions using LigPlot (WALLACE; LASKOWSKI; THORNTON, 1995). Protein-peptide complexes can be filtered considering sequence features, structure resolution and experimental method.

At the end of 2019, Xu *et al.* (XU; ZOU, 2020) proposed PepPro, a nonredundant benchmarking tool for testing peptide–protein docking algorithms composed of only 89 complexes. For 58 complexes, the unbound protein structures are available, which is useful for evaluating to what extent docking algorithms can accommodate binding-related protein conformational changes.

In summary, a variety of databases have been proposed to explore and increase the understanding of protein-peptide interactions. Nevertheless, despite their relevant contributions when released, most of them are obsolete and/or no longer supported. Among those mentioned, PepBDB is the most comprehensive, as it contains approximately 13,000 complexes. In addition, it is the only one that provides features for binding mode analysis.

To fill these gaps, aiming at automatically collecting a broad and up-to-date data-set of protein-peptide complex structures as a useful resource for diverse peptide studies, we propose Propedia. This database is a comprehensive, general purpose and up-to-date protein-peptide resource that contains over 19,000 high-resolution structures from the Protein Data Bank (PDB) segmented in clusters to reduce redundancy if desired. Structures of complexes have been organized, facilitating search and visualization by different criteria such as PDB id, sequence similarity, peptide classification, source organism, binding area, molecular weight, aromaticity, instability index, isoelectric point, and hydrophobicity, among other computed data. These clusters not only help accommodate redundancy in the database but also allow comparisons among sequences, interfaces, interactions and functions. Therefore, Propedia is a comprehensive and powerful tool for structural studies of protein-peptide recognition, support for construction of

training and test data sets for docking and scoring approaches, and facilitation of peptide rational design.

Propedia was inspired by our previous work on defense of plants against insects and pathogens. Soybean, when injured by the caterpillar *Anticarsia gemmatalis* Hübner, produces the Kunitz trypsin inhibitor (KTI) and the Bowman–Birk inhibitor (BBI), which impede protease-catalyzed degradation in the insect gut (PILON et al., 2017; PATARROYO-VARGAS et al., 2017). Based on these inhibitors that are naturally produced by soybean, we are interested in proposing peptide or mimetic peptide molecules to inhibit the proteases of the caterpillar gut. We believe these molecules have the potential to be used in the ecological control of this pest insect. We formerly designed peptides manually, with the support of certain bioinformatic tools. Now we are investing in the development of automatic tools to support this process, such as ppiGReMLIN (QUEIROZ et al., 2020). In this context, Propedia aims to deliver a comprehensive data set of experimental protein-peptide complexes organized in three types of clusters based on : (i) sequence similarity; (ii) interface structure; and (iii) protein-peptide binding site. It permits analysis of structures under different perspectives, supporting the detection of potential peptides for interacting with a target of interest, for example, peptides that are likely to inhibit proteases of the caterpillar gut. It is important to note that our database is not specific to soybean and its insect pest *Anticarsia gemmatalis* Hübner and can be applied to other data sets involving protein-peptide complexes.

3.2 Construction and content

In this section, we detail the project decisions and the design process followed to build Propedia as well as the contents of the database.

3.2.1 Database construction

We used the following criteria to retrieve PDB entries: (1) structures composed by two or more chains, (2) one chain with at least 2 and no more than 50 residues (for peptides), and (3) structures solved by NMR or X-ray crystallography with resolution below 2.5 Å. The present release is composed of 19,813 complexes (May 02, 2020). We developed in-house Python scripts and the Biopython library (COCK et al., 2009) to extract PDB data and populate the database. Each file was filtered to remove hydrogen atoms, water molecules, alternative positions (HAMELRYCK; MANDERICK, 2003) and crystallographic artifacts (FASSIO et al., 2019).

We identified protein-peptide complexes from the remaining files. Chains with lengths of 2 - 50 residues were classified as "peptides". The reason for this choice is to keep Propedia comprehensive comprising the ranges used by the existing databases. Chains with more than 60

residues were classified as "receptors". This decision was empirical since, by allowing smaller receptors, we observed complexes involving two peptides (or small unstructured proteins).

The protein-peptide interfaces were computed as follows: if there was at least one peptide atom at a distance of 6Å from any receptor atom and the protein-peptide complex had an interface area (greater than 0), then the protein-peptide complex was included in the database. We used the method of Lee and Richards (LEE; RICHARDS, 1971) to compute the interface area (IA) and the accessible surface area (ASA). This algorithm returns the surface area of a protein in $Å^2$ and was computed by NACCESS (HUBBARD; THORNTON, 1993) software. The software receives a PDB file as input and returns the ASA of each atom. The IA was calculated using the following equation:

$$IA = (ASA(A) + ASA(B)) - ASA(AB) \tag{3.1}$$

where $ASA(A)$ and $ASA(B)$ are the ASA of the protein (A) and peptide (B), respectively, while $ASA(AB)$ is the protein-peptide complex (AB) ASA. Then, IA is assumed to be the set of atoms that gained solvent accessibility.

With this procedure, we identified 19,813 complexes, including 19,177 from X-ray structures and 636 by NMR. There were peptides missing residues or containing nonstandard amino acid residues or binding with multiple chains, described in Table 14. Peptides bound with multiple receptors may affect both it is structural conformation and those interface residues. Therefore, we removed these complexes, obtaining 5971 protein-peptide complexes and, from now on, we refer to them as the Clusterable Complex Dataset (CCD).

Table 14 – Summary of the number of complexes identified, by complexes with only standard amino acid residues peptides and binding with multiple receptors chains.

# of receptors bound with peptide	**# of complexes**	**# of complexes with only standard amino acid residues peptides**
1	8990	5971
2	7040	4232
3	2205	1449
4	1204	656
5	290	50
6	84	84
Total	19813	12442

Data collected in previous steps and computed clusters were stored in a MySQL database. The entity-relationship model is depicted in Figure 6. We have the following entities: pdb, complex, peptide, receptor, organism, cluster (three types: sequence, interface, binding site) and alignment (clustal (peptide sequence), mustang, probis). The group table contains keywords derived from the pdb classification. For example, the Coronavirus Main Proteinase (3CLpro)

(PDB id: 1p9u) is classified as 'Hydrolase/Hydrolase Inhibitor' and was labeled so as to be included in the groups: 'Hydrolase' and 'Inhibitor'. Alignment tables store data from the results of molecular pair alignment, according to the type of clustering, and therefore have double foreign keys (id_complex1, id_complex2) corresponding with the complex table due to efficiency requirements.

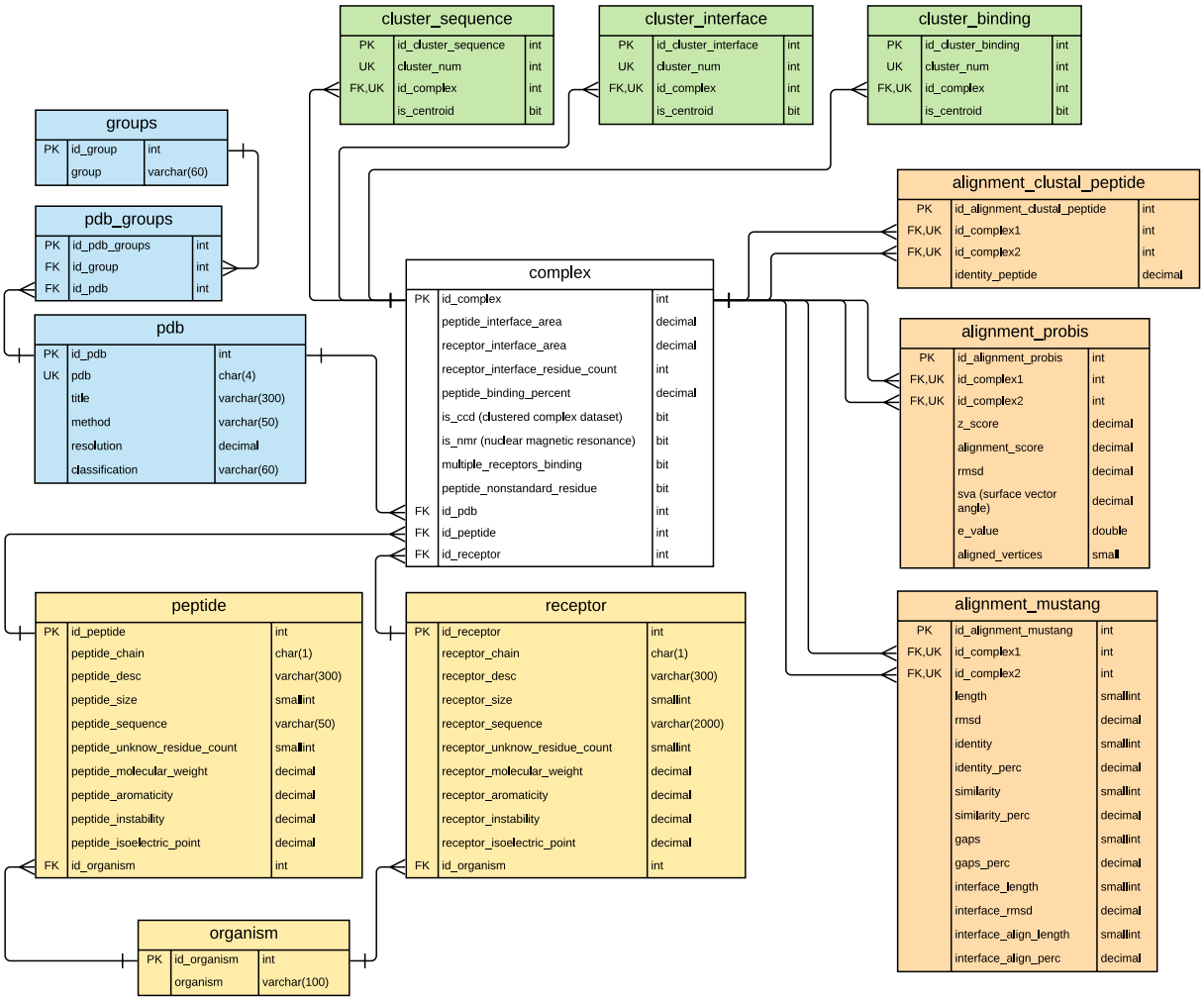


Figure 6 – Propedia database schema, presenting the tables, fields and relationships. The complex table (white) is the core of the database and interconnects all the data; pdb entities (blue) including group, pdb_groups and pdb tables; peptide/receptor and organism tables (yellow); cluster tables (green); and alignment tables (orange).

3.2.2 Clustering

3.2.2.1 Sequences

We classified peptide sequences using the tool Hammock (1.2.0) (KREJCI et al., 2016). It uses hidden Markov model profiles for peptide sequence clustering and three external tools for multiple alignments, similarity search, and HMM-HMM comparison: Clustal Omega (SIEVERS et al., 2011; SIEVERS; HIGGINS, 2014), HMMER 3.0 (FINN; CLEMENTS; EDDY,

2011), and HHSuite (SÖDING, 2005). We ran Hammock using mode 'full' with default parameters with the exception of '–min_conserved_positions', which was set to 3, and '–count_threshold', which was set to 300. These values were set empirically. Sequence labels were assigned using Python in-house scripts. CCD was used as input and after the filtering step, Hammock returned 3,495 unique sequences and classified them into 771 clusters and 1074 unique clusters (singletons), totaling 1845 peptide sequence clusters. For each cluster (non singletons) a consensus sequence was generated using the WebLogo tool (CROOKS et al., 2004), and the sequence alignment was determined using Clustal Omega (SIEVERS et al., 2011; SIEVERS; HIGGINS, 2014) to store the sequence identity among the peptides of each cluster. Centroids were identified as the peptides having the same sequence as the main sequence of each cluster.

3.2.2.2 Interface

Protein-peptide interfaces were aligned with MUSTANG (KONAGURTHU et al., 2006), a multiple protein structural alignment tool that superposes structures using the distances of the C-α coordinates of residues. A pairwise structural alignment was performed using only the protein structures of the CCD. To avoid unfavorable pairwise alignments, we considered only pairs of receptors sharing over 50% sequence identity. A total of 353,545 alignments were performed in parallel in a multicore processor, and the interface RMSD (iRMSD) was calculated from the results. The protein-peptide interface was considered to be all residues within 6 Å of a peptide (BICKERTON; HIGUERUELO; BLUNDELL, 2011; PLAXCO; SIMONS; BAKER, 1998). In-house Python scripts were developed to create an undirected graph network using the NetworkX (version 1.11) Python library (HAGBERG; SWART; CHULT, 2008). Nodes representing receptors and the edges (with the iRMSD between them) were added if 75% of the residues that composed the interfaces were aligned and had Cα distance less than or equal to 2 Å. This threshold is the same for PepX (VANHEE et al., 2010). Each connected subgraph from the undirected graph was considered a cluster. Altogether, 535 clusters were formed, plus 1356 singletons, for a total of 1891 non redundant protein-peptide interfaces. Each centroid was defined as the receptor node with the highest degree, and in the case of a tie, the one with the lowest sum of iRMSDs.

3.2.2.3 Binding sites

We used the ProBiS algorithm (KONC; JANEŽIČ, 2010) to identify similar protein-peptide binding sites. ProBis is a local alignment algorithm that aligns similar binding sites in proteins with dissimilar folds through 3D patterns of physicochemical properties of their surfaces, considering geometrical and functional groups. Functional groups are specific groups of atoms in residues with particular physicochemical properties, which include hydrogen bond acceptors, hydrogen bond donors, acceptor/donors, aromatics and aliphatic groups(SCHMITT; KUHN; KLEBE, 2002). ProBiS returns an alignment score for each pairwise alignment. The

higher the alignment score, the more similar the binding sites are. It also computes a Z-score, a statistical measure, based on alignment scores of the population. This parameter is calculated using the Karlin-Altschul equation (KARLIN; ALTSCHUL, 1990). The input we supplied to Propedia was the CCD and we extracted surface structural patches of each receptor at a distance of 6 Å from the corresponding peptide. Then, we performed a pairwise alignment using ProBiS.

$$Z\_score = \frac{alignment\_score - \mu}{\sigma} \tag{3.2}$$

The population mean (μ) and population standard deviation (σ) were computed from pairwise alignment scores in the CCD, where μ and σ are 1.488 and 4.951, respectively.

We used a similar method to define the clusters based on interfaces. An edge with alignment score, as weight, between two nodes (receptors) was created if the Z-score between them was greater than 1.5. This value was estimated as the point at which the number of clusters starts to increase exponentially. Connected subgraphs defined each cluster, and centroids were selected in the same way we described for previous clusters. Finally, 521 clusters and 945 singletons were generated, totaling 1466 distinct binding sites.

3.2.3 Propedia webserver

The Propedia database can be accessed through an interactive webserver implemented in the CodeIgniter PHP framework. Graph visualizations were implemented with the D3.js library [1]. Protein-peptide three-dimensional structure visualizations were generated using the 3Dmol.js library (REGO; KOES, 2015). The receptor/peptide sequence search mechanism is based on the blastp tool from NCBI-BLAST+ suite (ALTSCHUL et al., 1990; CAMACHO et al., 2009) and for the binding site search we use the ProBiS algorithm (KONC; JANEŽIČ, 2010).

3.3 Utility and discussion

The Propedia interface (bioinfo.dcc.ufmg.br/propedia) is user-friendly, visual and interactive. It allows database searches with several options (Figure 7A). Each entry in Propedia represents a protein-peptide complex. The web tool allows access to entries through PDB id, which can be followed (optionally) by protein chain id and peptide chain id.

Propedia's interface allows searching by pdb, complex id, organism, group (classification keyword), peptide and protein sizes, resolution, protein and peptide sequences (using BLAST), protein binding site (using ProBiS), and similar complexes using different clustering methods. When the user selects a particular complex to analyze, the web page presents the pdb, complex id, resolution, protein/peptide description and organism, and their data, includes chain,

[1] <https://d3js.org/>

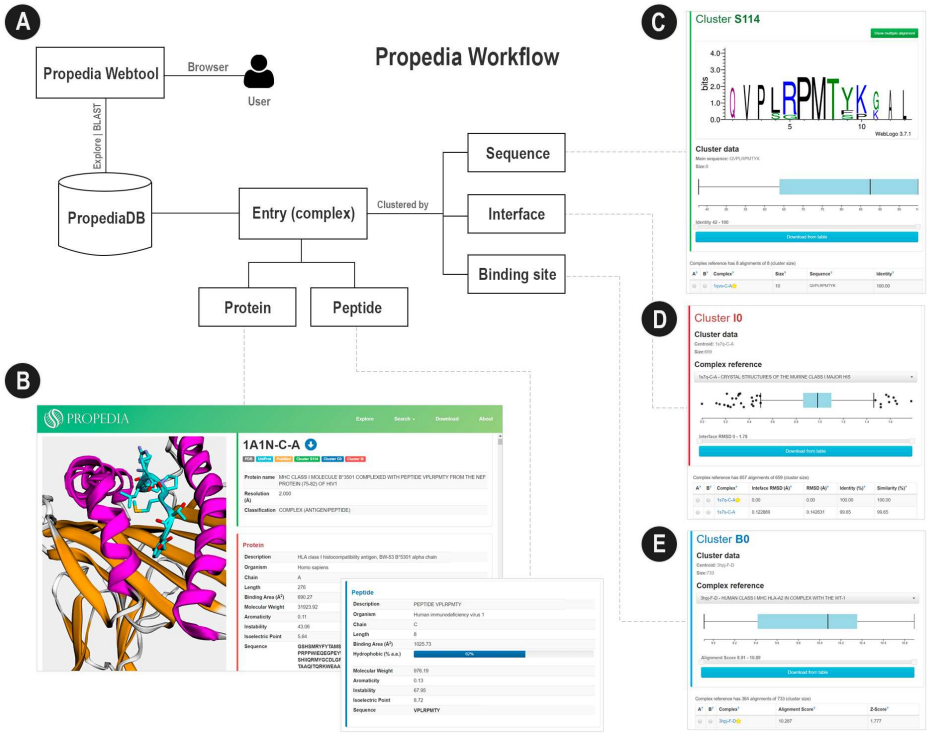


Figure 7 – (A) Propedia scheme. The user accesses Propedia through a browser. Propedia presents each protein-peptide as a complex. Each complex can be associated with a cluster based on sequence, interface or binding site. (B) Propedia interface. Three-dimensional structure visualization of a complex. Protein is shown as a cartoon (alpha-helix in magenta and beta-strands in orange). The peptide is shown as a cartoon with cyan sticks. Complex information includes receptor features, peptide features, clustering classification and similar complexes. (C, D, E) Sequence, interface and binding site, cluster pages. Sequence cluster containing the sequence WebLogo (consensus) and main sequence. Each cluster page has a distribution chart (boxplot), used to filter complexes, according to the attributes used for clustering: sequence identity, iRMSD and alignment score.

length, binding area (Å^2), molecular weight, hydrophobic percent (peptide only), aromaticity, instability index, isoelectric point, and sequence. We enriched Propedia with other relevant information from multiple databases such as UniProt and PubMed: protein chain, length, binding area and sequence information (Figure 7B).

The Propedia database was built based on a hybrid clustering approach that segments the set of complexes by the following: (1) sequence similarity; (2) interface structure; and (3) protein-peptide binding site. Due to this organization in clusters, users find similar protein-peptides complexes not only by traditional sequence and/or structure conservation but by interactions as well. Interactions, in fact, are essential for molecular recognition. A user can choose among these three different approaches to eliminate redundancy of the data set, if needed.

3.3.1 Comparison with other peptide databases

There are several peptide databases available. Table 15 compares some of their features. Each existing database contributes mainly to a specific piece of biological information. PepX (VANHEE et al., 2010) is a protein-peptide interaction database clustered by binding interfaces. It has 1,431 complexes with peptide sizes between 5 and 35 amino acids. PepBind (DAS et al., 2013) compiles structures, sequences and experimental information for protein-peptide complexes with peptides up to 35 amino acids. PeptiDB (LONDON; MOVSHOVITZ-ATTIAS; SCHUELER-FURMAN, 2010) comprises only 103 high-resolution complexes with peptides ranging in size from 5 to 15 amino acids. Some of these databases are not being updated and, for others, data are not even available.

PepPro (XU; ZOU, 2020), PixelDB (FRAPPIER; DURAN; KEATING, 2018) and PepBDB (WEN et al., 2019), on the other hand, are more recent and up-to-date efforts. They aggregate structural data from peptides up to 50 amino acids. PepPro (XU; ZOU, 2020) is a benchmark database built specifically for evaluation of protein-peptide docking algorithms. It contains 89 nonredundant complex structures retrieved from 1,198 high-resolution PDB entries with peptide size ranging from 5 to 30 residues. PixelDB (FRAPPIER; DURAN; KEATING, 2018) contains 1,966 nonredundant protein-peptide structures organized into clusters to provide structural conservation data for peptide binding modes. Finally, PepBDB (WEN et al., 2019) comprises 12,241 protein-peptide complex structures and their interaction information and is useful for analyzing and benchmarking docking algorithms and scoring functions.

Propedia is a more recent and fully automated database and webserver that will be updated quarterly. It is broader (comprises the entire PDB data-set) and general purpose protein-peptide analysis tool that is ready to collect, filter, clean, and compute several features and cluster data automatically always providing a comprehensive and up-to-date resource. For instance, researchers can already retrieve SARS-COV-2 proteins along with peptides in Propedia.

Table 15 – Comparison between Propedia and other protein-peptide complex databases.

Name	# of complexes	Peptide length (aa)	Resolution (Å)	Type	Availability
Propedia	19,813	2 – 50	< 2.5	Web server	✓
PepX	1,431	5 – 35	< 2.5	Web server	N.A.*
PeptiDB	103	5 – 15	< 2.0	PDB IDs' list	✓
PepBind	5,314	≤ 35	N.A*	Web server	N.A.*
PixelDB	1,966	5 – 50	< 2.5	GitHub	✓
PepBDB	12,241	< 50	N.A*	Web server	✓
PepPro	1,198	5 – 30	< 2.5	PDB IDs' list	✓

* N.A.: not available.

3.3.2 Case studies

We designed three case studies using Propedia's varied features to exemplify possible use cases of the tool and the adjoining database.

3.3.2.1 Estrogen receptors in complexes with different peptides (2JF9 and 4IV2)

We performed a case study with the estrogen receptor alpha LBD in complex with a tamoxifen-specific peptide antagonist (PDB id: 2JF9; peptide: chain Q; protein: chain B). This is a *Homo sapiens* protein classified in the PDB in the transcription category. The main objective of this case study was to test if Propedia would be able to find structures with similar binding sites but with different peptide sequences.

We compared the estrogen complex with the crystal structure of the estrogen receptor alpha ligand-binding domain in complex with dynamic way-derivative (PDB id: 4IV2; peptide: chain C; protein: chain A). Although these complexes were classified in the same cluster (B1) considering their interactions, the clusters for sequence was different (Table 16).

Table 16 – Comparison between protein and peptide characteristics of 2JF9 and 4IV2.

		4IV2–C–A	**2JF9–Q–B**
Protein	**Chain**	A	B
	Length	232	210
	Binding Area (Å^2)	484.85	519.30
Peptide	**Chain**	C	Q
	Length	10	13
	Binding Area (Å^2)	559.07	547.74
	Hydrophobic (% a.a.)	40%	30%
	Molecular Weight	1272.50	1539.71
	Aromaticity	0.00	0.15
	Instability	95.31	34.72
	Isoelectric Point	8.76	5.79
	Sequence	HKILHRLLQD	SPGSREWFKDMLS
Clusters	**Sequence cluster**	S 0	S 1024
	Interface cluster	I 1	I 1
	Binding cluster	B 1	B 1

Thus, we aligned the PDB files using the PyMol tool (DELANO, 2002) and compared the results manually (Figure 8). We observed that although the peptide primary structures were different, the peptide α-helix folding remained the same. In addition, Propedia was able to de-

tect similar contacts in the protein-peptide interactions, suggesting conservation in the mechanism of recognition. Additionally, our analysis showed that the protein residues were conserved, but the residues of the peptides were not. However, the interaction patterns were maintained.

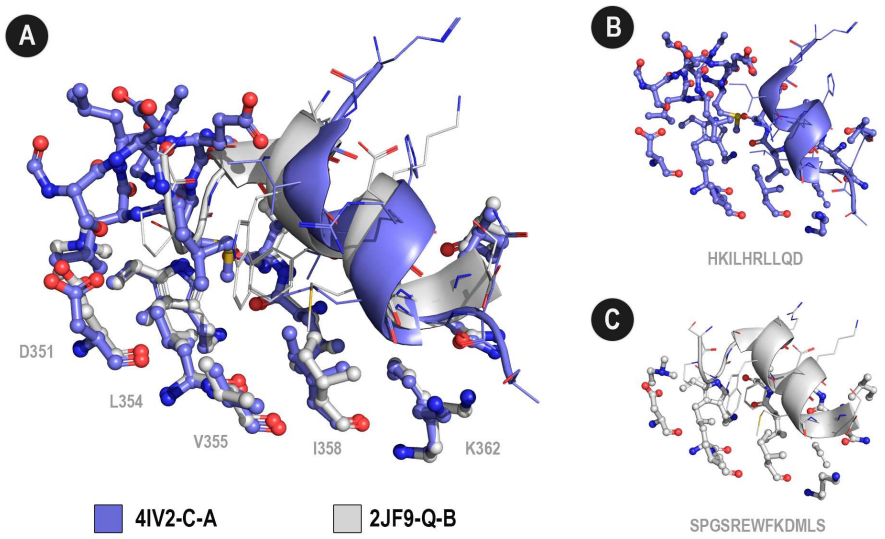


Figure 8 – (A) Structural alignment between 2JF9 and 4IV2. The protein residues were conserved, but the peptide residues were not. (B) Estrogen receptor alpha LBD in complex with a tamoxifen-specific peptide antagonist (PDB ID: 2jf9; peptide: chain Q; protein: chain B). (C) Estrogen receptor alpha ligand-binding domain in complex with dynamic way-derivative (PDB ID: 4IV2; peptide: chain C; protein: chain A).

This study case highlights the potential of Propedia to find similar binding patterns between proteins and peptides, even when peptide primary structure is not conserved.

3.3.2.2 SARS-COV-2 main protease interactions with peptides (6LU7)

From a SARS-COV-2 main protease structure (PDB id: 6lu7) we performed a case study to find peptides that can potentially recognize the binding site and inhibit it competitively. We submitted the 6lu7 structure to the Propedia database webserver using CCD searching scope, setting the chain (A) and binding site residues (residues within 6 Å of the N3 inhibitor). We searched for complexes with similar binding sites and the best results were ranked by alignment score (result value from the ProBiS). The top 10 results were retrieved.

From these 10 peptides retrieved in complex with similar binding sites, we obtained the peptides presented in Table 17. We verified that both of 1lvb peptides (chains C and D) have the same sequence and structural conformation. Therefore, only the complex with 1lvb-D peptide was kept for the next analyses, which has a better bound receptor (chain B), based on it is alingment score and RMSD from Propedia query results.

Propedia was able to retrieve 2 proteases from SARS-COV (previous coronavirus infecting human beings) and other viral proteases along with peptides that could be useful for the design of antiviral peptides capable of inhibiting the SARS-COV-2 main protease. Consequently,

Table 17 – List of retrieved peptides for SARS-COV-2 main protease case study. *Sequences omitted due to their long length.

PDB id	Description	Protein Chain	Peptide Chain	Peptide AA Sequence
2q6g	SARS-COV main protease H41A mutant	A	C	-TSAVLQSGFRK
1uk4	SARS-COV main proteinase	B	H	--NSTLQ----
1lvm	Thermotoga maritima methyltransferase	B	D	-ENLYFQ----
1lvm	Thermotoga maritima methyltransferase	A	C	-ENLYFQ----
3mmg	Tobacco vein mottling virus protease	A	C	-ETVRFQS---
1lvb	Tobacco etch virus protease	B	D	TENLYFQSGT--
1lvb	Tobacco etch virus protease	A	C	TENLYFQSGT--
5om5	Human alpha1-antichymotrypsin	A	B	-TSAVLQSGFR-
6hgj	SARS-COV main protease variant NewBG-III	A	B	*
3caa	Cleaved antichymotrypsin A347R	A	B	*

we performed molecular docking experiments with the peptides returned by the search on SARS-COV-2 protease. We used the Rosetta FlexPepDock docking protocol (RAVEH; LONDON; SCHUELER-FURMAN, 2010). It computes high-resolution complex structures from an approximate model of a peptide within a receptor binding site, allowing full flexibility of the peptide backbone and all side chains. To provide the initial structure of each complex, we superposed the SARS-COV-2 protease (PDB id: 6lu7) with the Propedia retrieved complex and removed the protein retaining SARS-COV-2 and the peptide. This procedure was successful for 8 complexes, in which the SARS-COV-2 protease was properly superposed with the model receptors (manual inspection and $RMSD \leq 3$ Å). For two peptides whose receptors did not align properly due to structural high dissimilarity (PDB id: 3caa:B and 6hgj:B), we performed a global blind docking using HADDOCK (ZUNDERT et al., 2016). Then, we selected the best model from the best cluster (most negative HADDOCK score) and submitted it as thanbe initial structure to Rosetta FlexPepDock protocol as we did with the previous 8 peptides. We had to discard both peptides (PDB id: 3caa:B and 6hgj:B) because FlexDock accommodates only peptides shorter than 30 residues. Consequently, we obtained 8 docked models, and all of them exhibited considerable affinity to the SARS-COV-2 main protease (Table 18, column "Rosetta score"). In addition to acceptable scores, we verified apparently adequate poses (Figure 10) of each peptide for cleavage by the site considering the proximity (α-C of P1) to CYS145's sulfur atom (Table 18, column "RosettaCYS distance").

In fact, according to the MEROPS database of proteolytic enzymes (RAWLINGS; BARRETT; BATEMAN, 2010; GOETZ et al., 2007), SARS coronavirus main proteases show preference for substrates of the general form: P4=V/T/A/S P3=V/W/K P2=L P1=H/Q. These positions are depicted in shades of blue in Figure 9. According to these previous works, the site P1 is well conserved but the other sites are very mutable. The peptides identified using Propedia have residues highlighted in yellow and it can be viewed in the webserver [2]. Notice that this set of peptides is generally consistent with peptides known to inhibit SARS coronavirus proteases.

[2] <http://bioinfo.dcc.ufmg.br/propedia/search/binding/covid/>

Table 18 – RMSDs for SARS-COV-2 main protease and superposition of receptors identified by Propedia

PDB id	chain	Propedia Alig. score	Propedia site RMSD	Rosetta receptors RMSD	Rosetta score	P1-CYS145 distance
2q6g	A	10.36	0.34	0.997	-542.507	3.6
1uk4	B	9.47	0.44	0.659	-525.999	3.6
1lvm	B	5.69	0.84	2.175	-525.907	4.0
1lvm	A	5.26	1.53	2.085	-528.398	5.5
3mmg	A	4.67	0.47	1.785	-530.833	3.7
1lvb	B	4.54	1.21	2.521	-530.517	3.6
5om5	A	3.18	1.63	6.665	-538.985	3.7
6hgj	A	3.38	2.20	11.156	-	-
3caa	A	3.25	2.16	9.943	-	-

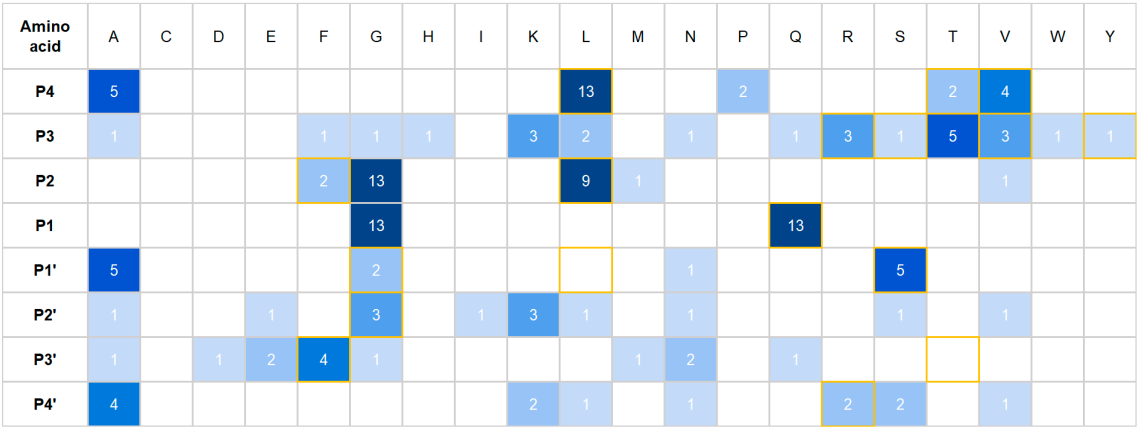


Figure 9 – MEROPS specificity matrix in shades of blue and residues from Propedia suggested peptides highlighted in yellow.

3.3.2.3 Metadynamics estimated ΔG_{bind} correlates with the major Propedia scores for the SARS-COV-2 M^{pro}

The free energy landscape (FEL) for the respective triplicates of the unbinding metadynamics (MetaD) for the M^{pro}: peptides complexes with the PDB id: 2q6g (chain C), 1uk4 (chain H), 1lvm (chain D), and 1lvb (chain D) are shown on Figure 30 from the Additional file B (maps 1-3, 4-6, 7-9 and 10-12, respectively). At each system, the minima inside the protein (A) and at the aqueous environment (B) could be characterized with enough accuracy in order to estimate the binding free energy (ΔG_{bind}) according the described on equations (B.1, B.2 and B.3) from the Additional file B.

It could be obtained a significant convergence for the MetaD recovered ΔG_{bind} values for each system in our protocol, with maximal standard deviation of 1.57 kcal·mol^{-1} for the systems 1lvm (chain D), 1lvb (chain D) and 2q6g (chain C) and a relatively higher deviation of 4.32 kcal·mol^{-1} just for 1uk4. In fact, such convergence is not surprising, once the already

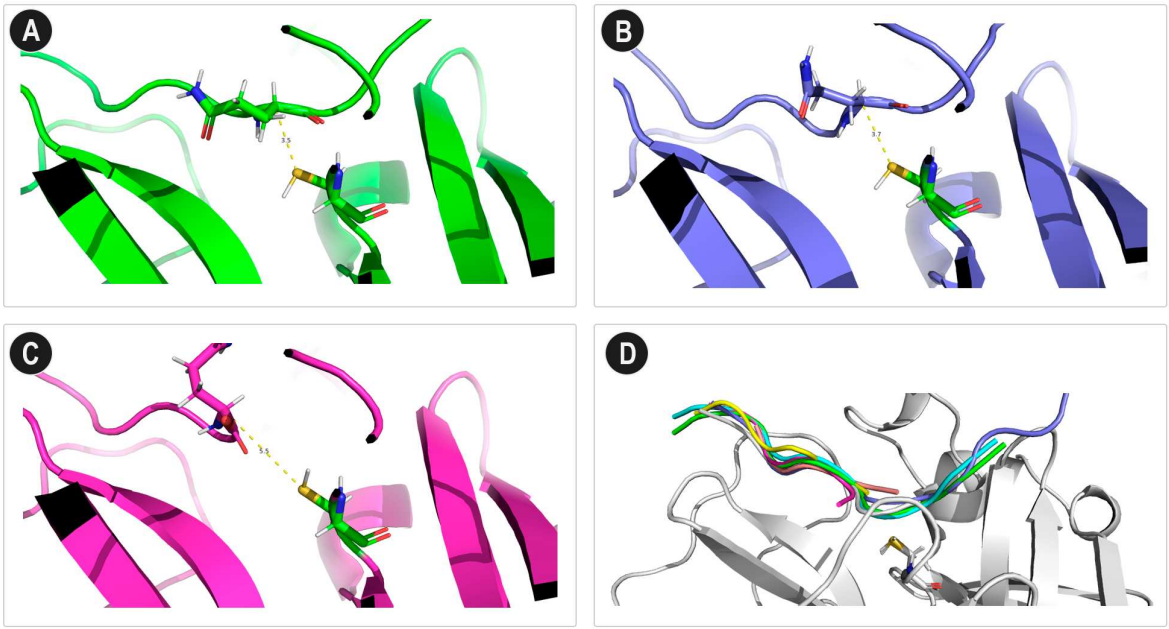


Figure 10 – (A) PDB ID: 1lvb; peptide: chain C; protein: chain A; Rosetta score: -538.306; Distance: 3.5 (B) PDB ID: 5om5; peptide: chain B; protein: chain A; Rosetta score: -538.985; Distance: 3.7 (C) PDB ID: 1lvm; peptide: chain C; protein: chain A; Rosetta score: -528.398; Distance: 5.5 (D) the whole set of evaluated peptides.

consolidated situation of the metadynamics technique as an accurate computational tool to estimate the binding free energy for usual ligands and peptides, being a powerful method on drug screening procedures (CAVALLI et al., 2015; SÖLDNER; HORN; STICHT, 2019; BRANDT et al., 2016a). The accuracy of this technique, in this way, makes it a providential instrument to validate the Propedia methodology at the screening of peptides with differential affinities for the SARS-COV-2 M^{pro}, given the still sparse availability of experimental data for peptide affinity at this new and important target. In this way, the correlation between the ΔG_{bind} recovered by the metadynamics higher performance method and the Propedia recovered scores was carried aiming the validation of this computational tool. It is notorious at Figure 11 (Additional file B: Table 23), the significant negative correlation of the MetaD ΔG_{bind} with the Propedia recovered alignment score (R^2 of 0.98) and the positive correlation with the Propedia recovered RMSD in Å at the active site alignment procedure (R^2 of 0.96). Even, it is noteworthy that both the Propedia scores as the MetaD recovered ΔG_{bind} values put the known MPro specific substrate (PDB:2q6g) and the substrate-analogous MPro inhibitor (1uk4) at the top of the affinity ranking with this protein. In this way, both the significant correlation with the results from the high performance metadynamics method, as well the self-consistence with known functional data can be taken together as an indicative of validation for our new software, as well its applicability at the screening for functional peptides for this and other important targets.

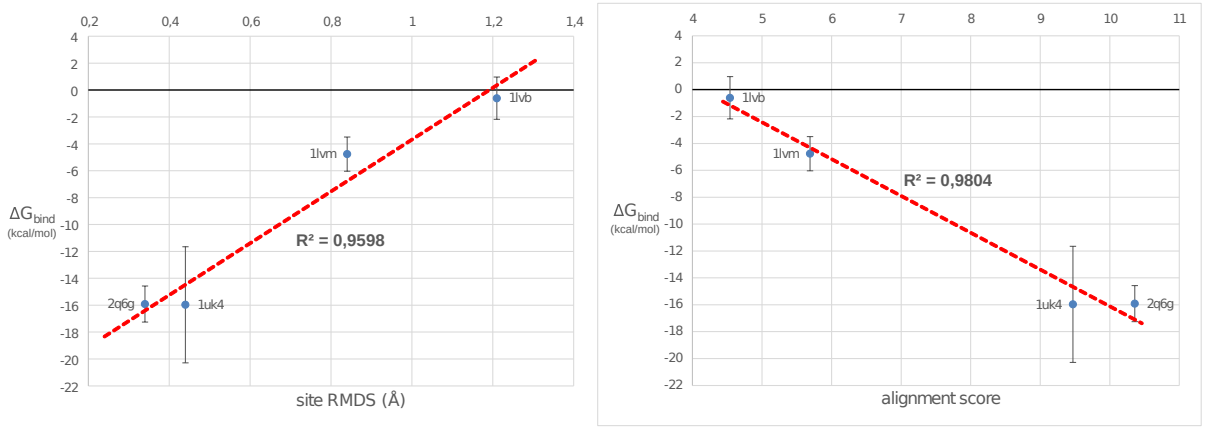


Figure 11 – Correlation of MetaD ΔG_{bind} with site RMSD (left) and alignment score (right) from the SARS-COV-2 M^{pro} with peptide complexes from the PDB id: 2q6g (chain C), 1uk4 (chain H), 1lvm (chain D), and 1lvb (chain D)

3.3.2.4 Anticarsia gemmatalis protease

The velvetbean caterpillar, *Anticarsia gemmatalis* (AG) Hübner (Lepidoptera: Noctuidae) is one of the primary defoliating pests in the Americas, affecting mainly soybean crops, and a major cause of economic losses in agriculture (VIANNA et al., 2011; MOSCARDI et al., 2012). In recent years, alternative approaches towards pest control, such as the development of biopesticides, have been explored. For instance, the use of protease inhibtors is high regarded in insect pest management, as it affects the bioavailability of essential amino acids, which ultimately hinders larvae growth and the development of insects for several species, as has been shown in (MOREIRA et al., 2011; PILON et al., 2018).

In this case study, we used the sequence of a trypsin-like serine protease extracted from the AG's gut, sequenced by our research group and deposited at GenBank(BENSON et al., 2005) (accession JX898746.1 (GENBANK. . . , 1982)). Additionally, a 3D model was produced using the I-TASSER server(YANG et al., 2015). We performed a structural alignment of the model with the highest ranked templates from the modeling step in order to identify the highly conserved residues from the catalytic triad in the protease (PERONA; CRAIK, 1995). These residues were identified as HIS6, ASP56, and SER143 in the model. Additional file B: Figure 31) shows the superposition of the structures where the triad residues are highlighted.

We queried the Propedia database server using the protease sequence and the residues from the catalytic triad as binding site residues (along with the 3D model) in two separate experiments, and the top 10 results in each of them were selected according to their alignment scores. The results are shown in Tables 19 and 20 respectively. Then, we performed molecular docking experiments of the peptides retrieved to the model of the AG's protease using only HADDOCK, since a considerable number of the peptides retrieved contained non-standard residues, which is not supported by PepFlexDock. Also, for the binding site query experiment, two peptides entries from the Propedia results were not used (which are not listed in Table 20): 1p11-P, due to

its format not being supported by HADDOCK and 3kf2-D, due to the high sequence similarity to peptide 3kf2-C in the same PDB structure.

Table 19 – List of retrieved peptides for the AG protease case study using sequence query.

PDB id	Description	Protein Chain	Peptide Chain
1ekb	Bovine enteropeptidase	B	A
1ekb	Bovine enteropeptidase	B	C
2stb	Salmon trypsin	I	E
2sta	Salmon trypsin	I	E
3qgn	Human thrombin	A	B
2zdv	Human thrombin	L	H
1ca8	Human thrombin	A	B
1ca8	Human thrombin	A	B
4dii	Human thrombin	L	H
4dih	Human thrombin	L	H
4lz1	Human thrombin	B	A

Table 20 – List of retrieved peptides for the AG protease case study using binding site query.

PDB id	Description	Protein Chain	Peptide Chain
3qgj	Lysobacter enzymogenes protease	D	C
1p11	Lysobacter enzymogenes protease	I	E
2obq	Hepacivirus NS3-4A protease	B	C
2oin	Hepacivirus NS3-4A protease R155K	C	A
2o8m	Hepacivirus NS3-4A protease S139A	D	B
3kn2	Hepacivirus NS3-4A protease	B	C
3kf2	Hepacivirus NS3-4A protease	C	A
3sga	Streptomyces griseus protease	P	E
6rw2	Human Ephrin type-A receptor 2	B	A
4a1t	Hepacivirus NS3-4A protease	D	B

For the sequence based dataset, we set the residues from the catalytic triad as active residues for the docking procedure, as well as the complete chain of each peptide. We selected the best resulting structures primarily according to the HADDOCK score (most negative) and then, according to the RMSD ($\leq$ 3 angstroms) of each structure relative to the overall lowest energy model. Table 21 summarize the results. Finally, peptide poses in the protease were analysed for the top 5 scored models according to the HADDOCK score, for which we identified the closest residues to the SER143 residue at the S1 site, considering the distance between C-α atoms. The closest residues found were cysteine residues located in models 3qgn-A (3.9 Å) , 4dii-L (4.4 Å) and 1ca8-A (5.1 Å). The presence of cysteine residues close to the serine in the catalytic tryad indicate a potential use of the peptide as an inhibitor since substrates with these residues at position P1 are not usually cleaved by trypsin-like serine proteases (PAGE; CERA, 2008) . Figure 12 shows models 3qgn-A and 4dii-L, where the distance between residues is highlighted.

Table 21 – HADDOCK score and RMSD for the selected models for each peptide chain in the sequence based experiment

PDB id	Chain	HADDOCK iRMSD	HADDOCK score	S1 closest residue
1ekb	C	2.564	-49.202	-
1ekb	A	2.499	-63.477	-
2stb	I	0.000	-86.803	-
2sta	I	4.275	-81.387	-
3qgn	A	0.000	-97.979	CYS (3.9 Å)
2zdv	L	0.000	-100.560	GLU (7.4 Å)
1ca8	A	1.530	-102.975	CYS (5.1 Å)
1ca8	C	1.158	-75.138	-
4dii	L	2.598	-95.836	CYS (4.4 Å)
4dih	L	1.508	-95.317	ARG (5.4 Å)

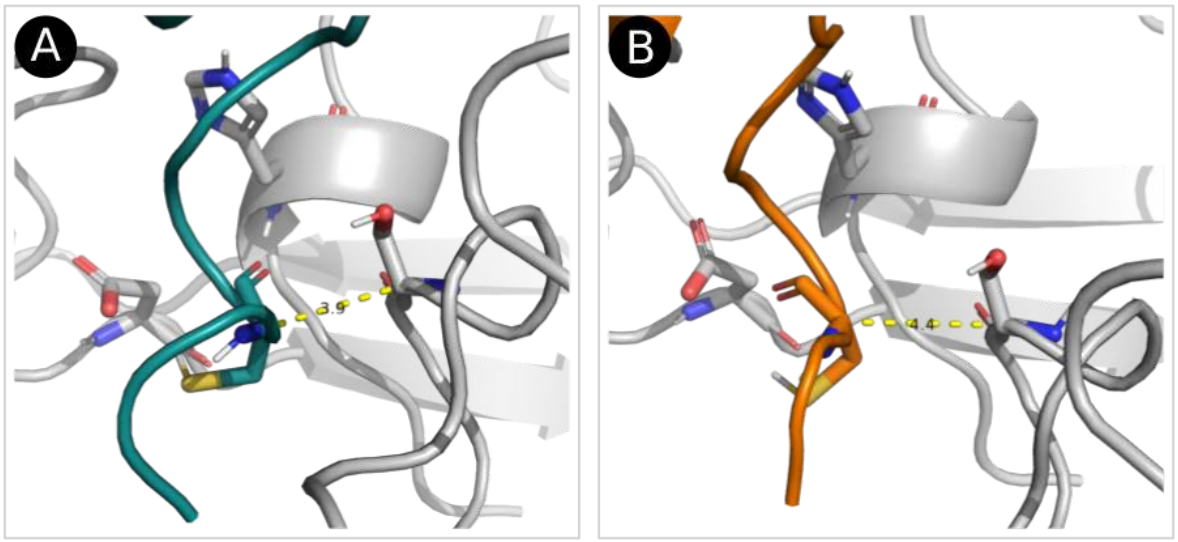


Figure 12 – AG's Protease model, in gray, coupled with peptides 3qgn-A (A) and 4dii-L (B). The distance between the SER143 residue from the S1 site in the protease to the cysteine residues in the peptides are 3.9 Å and 4.4 Å respectively

Similar to the sequence based dataset, we performed the docking for the binding site dataset using the residues from the catalytic triad, as well as complete peptide chains as active residues. A binding site signature is how a protein interacts with its ligand, and which amino acids are essential to keep the complex stable. A proper metric to verify the similarity of binding sites is the fraction of common contacts (FCC). The FCC_{AB} is the ratio of contacts between structures *A* and *B* to all contacts in *A*, whose value ranges from zero, when the chains share no contacts, to a maximum of one, when all contacts of chain A are with chain B (RODRIGUES et al., 2012). Therefore, for the binding site docking experiments, a higher average value of FCC in a cluster indicates higher similarity of the interactions between different peptide poses and the protease model, which also means that the binding site is more conserved.

For each peptide, we selected the cluster with the highest FCC score (relative to its

lowest energy models produced by HADDOCK), from which sets we chose the best models according to their HADDOCK scores. FCC values and HADDOCK scores are shown in Table 22 for all peptides. The best 4 models for each of the top 3 scored clusters are shown in Figure 13. In all models, contacts are centered in the catalytic triad (highlighted in red), while the remaining contact areas bind to different ligands, where neighboring residues on the protease side have great relevance by establishing hydrophobic and hydrogen bonds. The complete interaction map of each model is available in Additional file B: Figure 32. This emphasizes the importance of using FCC as a suitable metric for binding site analysis rather than RMSD, and also demonstrates Propedia's accuracy in determining binding site patterns in regard to the ligand specificity.

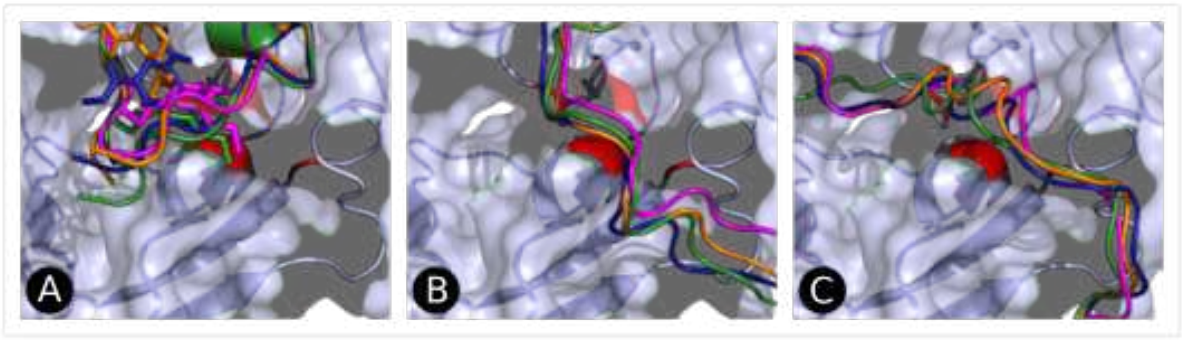


Figure 13 – AG's Protease model, in gray, coupled with the 4 top scored poses of peptides 6rw2-B (A), 3kn2-B (B) and 2obq-B (C). Residues in red represent the catalytic residues from the catalytic triad.

Table 22 – HADDOCK score and FCC for the selected models for each peptide chain in the binding site experiment

PDB id	Chain	Cluster FCC	Lowest HADDOCK Score
3qgj	D	0.409	-34.823
1p11	I	0.621	-49.383
2obq	B	0.833	-80.258
2oin	C	0.696	-85.060
2o8m	D	0.196	-54.616
3kn2	B	0.840	-78.998
3kf2	A	0.236	-55.432
3sga	P	0.650	-61.368
6rw2	A	0.883	-75.354
4a1t	D	0.648	-87.944

3.4 Conclusions

As far as we know, Propedia is the broadest and most comprehensive set of protein-peptide complexes. At the moment of publication of this paper, it comprises approximately

20,000 complexes. Furthermore, we developed hybrid clustering strategies that organized data into 1845 clusters based on sequences, 1891 clusters based on interface structures similarity and 1466 clusters based on binding sites. These groups may be used for detecting either nonredundant or similar complexes with several purposes going from peptide docking and scoring function benchmarking, design of biotechnological peptides and even peptide-based rational drug design. Finally, Propedia is available through a web interface, searches and analysis can be performed by a user-friendly interface and all the data are available to download.

Conclusion

In this dissertation, we present some papers that propose computational strategies based on machine learning to characterize and predict protein intermolecular interactions and to summarize protein structural bioinfomatics.

In our first contribution, we work on was the literature review paper, where we describe relevant tools for eight fields of protein bioinformatics: docking, molecular dynamics, molecular visualization, structure prediction, mutation analysis, interactions at atomic/residue level, catalytic and binding site prediction and databases. We use the research material to build the web-server PreStO (Protein Structural bioinformatics Overview). The web-server is an orientation to students and researchers to help them choose the most suitable tools for their application. PreStO lists all the tools described in the paper and their respective reference. For this project, we can pursue future work involving studies with humans to qualitatively evaluate the user experience of the tools raised. This would be a step further to assist the user in selecting the most suitable and intuitive tool for the research purpose.

In our second work, we develop a machine learning strategy that couples different FDA-approved molecule fingerprints to perform LBVS over the main protease of SARS-COV-2. With the results of this step, we follow with SBVS techniques: docking, MM-PBSA, and metadynamics simulations to filter the small molecules and predict the binding energy between each ligand and our target. Six molecules have been identified as potential inhibitors of the SARS-COV-2 main protease. As ambenonium was not yet described in the literature interacting with this protein in its active pocket. For this reason, we made a patent request for the use of this drug for coronavirus proteases inhibition.

This subject can be approached from other perspectives in future work to enrich the LBVS methodology and broaden its application. We can evaluate the impact of rebalancing steps, dimensionality reduction, segregation of training and validation data that are input to TPOT, as well as the selection of other classification algorithms and the use of stratified cross-validation. Another point of improvement would be to use model explainability techniques so that the user can better understand what was considered in predicting a molecule with inhibition potential. To enrich the data set and identify molecules beyond those already approved by the FDA stored in the DrugBank, we could consider in future work the use of broader databases such as ZINC. In this context, we would have a deeper insight into the chemical composition of the ligands, evaluating the toxicity and drug-like characteristics of the constituent functional groups.

Furthermore, we also propose a machine learning based strategy to cluster protein-peptide interfaces according to three criteria and structured the interactions database, Propedia. It was developed considering hybrid clustering strategies to group the registers in the database. As we show in our case studies, the tool can be applied to identify interactions patterns to molecular recognition and modeling. Here it could also conduct studies with humans to validate the user experience with the tool. Furthermore, we could consider new versions of Propedia with new features, such as the inclusion of non-ribosomal peptides and with post-translational changes in the database, as well as allowing binding prediction. For this last feature, the degree of prediction reliability could be incorporated into Propedia to indicate how well the user can consider the prediction consistent with the way the structures were obtained.

Our contributions presented in this dissertation explores many fields of structural bioinformatics and how they can be applied in the studies of proteins. These applications are of great relevance to support the investigation of relevant research topics and emerging problems, such as COVID-19 outbreak.

Bibliography

ABRAHAM, M. J. et al. GROMACS: High performance molecular simulations through multi-level parallelism from laptops to supercomputers. **SoftwareX**, Elsevier, v. 1, p. 19–25, 2015.

ACD/CHEMSKETCH ADVANCED CHEMISTRY DEVELOPMENT, I. Acd/labs release: 12.00 product version: 12.00. **Build**, v. 29305, 2008.

ACHARYA, C. et al. Recent advances in ligand-based drug design: Relevance and utility of the conformationally sampled pharmacophore approach. **Current Computer Aided-Drug Design**, Bentham Science Publishers Ltd., v. 7, n. 1, p. 10–22, mar. 2011. Disponível em: <https://doi.org/10.2174/157340911793743547>.

ADASME, M. F. et al. Plip 2021: Expanding the scope of the protein-ligand interaction profiler to dna and rna. **Nucleic Acids Res**, 2021.

AHMED, S. F.; QUADEER, A. A.; MCKAY, M. R. Preliminary identification of potential vaccine targets for the covid-19 coronavirus (sars-cov-2) based on sars-cov immunological studies. **Viruses**, MDPI AG, v. 12, n. 3, p. 254, Feb 2020. ISSN 1999-4915. Disponível em: <http://dx.doi.org/10.3390/v12030254>.

AKCAPINAR, G. B.; SEZERMAN, O. U. Computational approaches for de novo design and redesign of metal-binding sites on proteins. **Bioscience reports**, Portland Press, v. 37, n. 2, 2017.

AKDEL, M. et al. A structural biology community assessment of AlphaFold 2 applications. **bioRxiv**, Cold Spring Harbor Laboratory, 2021.

ALQURAISHI, M. AlphaFold at CASP13. **Bioinformatics**, Oxford University Press, v. 35, n. 22, p. 4862–4865, 2019.

ALTMAN, R. B.; DUGAN, J. M. **Structural Bioinformatics**. 2. ed. [S.l.]: Wiley-Blackwell, 2009.

ALTSCHUL, S. F. et al. Basic local alignment search tool. **Journal of molecular biology**, Elsevier, v. 215, n. 3, p. 403–410, 1990.

ANDERSON, A. C. The process of structure-based drug design. **Chemistry & Biology**, Elsevier BV, v. 10, n. 9, p. 787–797, set. 2003. Disponível em: <https://doi.org/10.1016/j.chembiol.2003.09.002>.

ANDREEVA, A. et al. SCOP2 prototype: a new approach to protein structure mining. **Nucleic Acids Research**, v. 42, n. D1, p. D310–D314, 11 2013. ISSN 0305-1048.

ANDREEVA, A. et al. The SCOP database in 2020: expanded classification of representative family and superfamily domains of known protein structures. **Nucleic Acids Research**, v. 48, n. D1, p. D376–D382, 11 2019. ISSN 0305-1048.

ANFINSEN, C. B. The formation and stabilization of protein structure. **Biochemical journal**, Portland Press Ltd, v. 128, n. 4, p. 737, 1972.

ANGELOVA, A. et al. Pep-lipid cubosomes and vesicles compartmentalized by micelles from self-assembly of multiple neuroprotective building blocks including a large peptide hormone pacap-dha. **ChemNanoMat**, Wiley Online Library, v. 5, n. 11, p. 1381–1389, 2019.

ARNOLD, K. et al. The SWISS-MODEL workspace: a web-based environment for protein structure homology modelling. **Bioinformatics**, Oxford University Press, v. 22, n. 2, p. 195–201, 2006.

ATTWOOD, T. K. et al. A global perspective on evolving bioinformatics and data science training needs. **Briefings in Bioinformatics**, Oxford University Press, v. 20, n. 2, p. 398–404, 2019.

BAJAJ, C. et al. TexMol: interactive visual exploration of large flexible multi-component molecular complexes. In: **IEEE Visualization 2004**. [S.l.: s.n.], 2004. p. 243–250.

BAKER, N. A. et al. Electrostatics of nanosystems: application to microtubules and the ribosome. **Proceedings of the National Academy of Sciences**, National Acad Sciences, v. 98, n. 18, p. 10037–10041, 2001.

BANEGAS-LUNA, A.-J.; CERÓN-CARRASCO, J. P.; PÉREZ-SÁNCHEZ, H. A review of ligand-based virtual screening web tools and screening algorithms in large molecular databases in the age of big data. **Future Medicinal Chemistry**, Future Science Ltd, v. 10, n. 22, p. 2641–2658, nov. 2018. Disponível em: <https://doi.org/10.4155/fmc-2018-0076>.

BARDUCCI, A.; BONOMI, M.; PARRINELLO, M. Metadynamics. **Wiley Interdisciplinary Reviews: Computational Molecular Science**, Wiley Online Library, v. 1, n. 5, p. 826–843, 2011.

BARTLETT, G. J. et al. Analysis of Catalytic Residues in Enzyme Active Sites. **journal of Molecular Biology**, v. 324, n. 1, p. 105–121, 2002. ISSN 0022-2836.

BATEMAN, A. et al. UniProt: the universal protein knowledgebase in 2021. **Nucleic Acids Research**, 2020.

BATEMAN, A. et al. UniProt: the universal protein knowledgebase in 2021. **Nucleic Acids Res**, v. 49, n. D1, p. D480–D489, 01 2021.

BEIGEL, J. H. et al. Remdesivir for the treatment of COVID-19 — final report. **New England Journal of Medicine**, Massachusetts Medical Society, v. 383, n. 19, p. 1813–1826, nov. 2020. Disponível em: <https://doi.org/10.1056/nejmoa2007764>.

BENDL, J. et al. PredictSNP: Robust and Accurate Consensus Classifier for Prediction of Disease-Related Mutations. **PLoS Computational Biology**, v. 10, 2014.

BENKERT, P.; BIASINI, M.; SCHWEDE, T. Toward the estimation of the absolute quality of individual protein structure models. **Bioinformatics**, v. 27, n. 3, p. 343–350, Feb 2011.

BENSON, D. A. et al. Genbank. **Nucleic acids research**, Oxford University Press, v. 33, n. suppl_1, p. D34–D38, 2005.

BERENDSEN, H. J. C. et al. Molecular dynamics with coupling to an external bath. **The Journal of Chemical Physics, v.**, v. 81, p. 3684–3690, 1984.

BERMAN, H.; HENRICK, K.; NAKAMURA, H. Announcing the worldwide Protein Data Bank. **Nature Structural & Molecular Biology**, v. 10, n. 12, p. 980–980, Dec 2003. ISSN 1545-9985.

BERMAN, H. et al. The worldwide Protein Data Bank (wwPDB): ensuring a single, uniform archive of PDB data. **Nucleic Acids Research**, v. 35, n. suppl_1, p. D301–D303, 11 2006. ISSN 0305-1048.

BERMAN, H. M. et al. The Protein Data Bank. **Nucleic Acids Research**, v. 28, n. 1, p. 235–242, 01 2000. ISSN 0305-1048.

BERMAN, H. M. et al. The protein data bank. **Nucleic acids research**, Oxford University Press, v. 28, n. 1, p. 235–242, 2000.

BEST, R. B. et al. Optimization of the additive charmm all-atom protein force field targeting improved sampling of the backbone ϕ, ψ and side-chain $\chi 1$ and $\chi 2$ dihedral angles. **Journal of chemical theory and computation**, ACS Publications, v. 8, n. 9, p. 3257–3273, 2012.

BIASINI, M. et al. SWISS-MODEL: modelling protein tertiary and quaternary structure using evolutionary information. **Nucleic acids research**, Oxford University Press, v. 42, n. W1, p. W252–W258, 2014.

BICKERTON, G. R.; HIGUERUELO, A. P.; BLUNDELL, T. L. Comprehensive, atomic-level characterization of structurally characterized protein-protein interactions: the piccolo database. **BMC bioinformatics**, Springer, v. 12, n. 1, p. 313, 2011.

BIENERT, S. et al. The SWISS-MODEL Repository-new features and functionality. **Nucleic Acids Res**, v. 45, n. D1, p. D313–D319, 01 2017.

BILOTTA, M.; TRADIGO, G.; VELTRI, P. Bioinformatics Data Models, Representation and Storage. In: **Encyclopedia of Bioinformatics and Computational Biology**. [S.l.]: Elsevier, 2019. p. 110–116.

BIOVIA, D. S. **Discovery studio modeling environment**. [S.l.]: Release, 2017.

BJELKMAR, P. et al. Implementation of the charmm force field in gromacs: Analysis of protein stability effects from correction maps, virtual interaction sites, and water models. **Journal of Chemical Theory and Computation**, American Chemical Society, v. 6, n. 2, p. 459–466, Feb 2010. ISSN 1549-9618. Disponível em: <https://doi.org/10.1021/ct900549r>.

BORDOLI, L. et al. Protein structure homology modeling using SWISS-MODEL workspace. **Nature protocols**, Nature Publishing Group, v. 4, n. 1, p. 1, 2009.

BOWERS, K. J. et al. Scalable Algorithms for Molecular Dynamics Simulations on Commodity Clusters. In: **SC '06: Proceedings of the 2006 ACM/IEEE Conference on Supercomputing**. [S.l.: s.n.], 2006. p. 43–43.

BRANDT, A. M. et al. Exploring the unbinding of l eishmania (l.) amazonensis cpb derived-epitopes from h 2 mhc class i proteins. **Proteins: Structure, Function, and Bioinformatics**, Wiley Online Library, v. 84, n. 4, p. 473–487, 2016.

BRANDT, A. M. L. et al. Exploring the unbinding of leishmania (l.) amazonensis cpb derived-epitopes from h2 mhc class i proteins. **Proteins: Structure, Function, and Bioinformatics**, v. 84, n. 4, p. 473–487, 2016. Disponível em: <https://onlinelibrary.wiley.com/doi/abs/10.1002/prot.24994>.

BROOKS, B. R. et al. CHARMM: the biomolecular simulation program. **journal of computational chemistry**, Wiley Online Library, v. 30, n. 10, p. 1545–1614, 2009.

BROWN, A. M.; BEVAN, D. R. Introducing Protein 3-D Visualization Software to Freshman Undergraduate Students: Making Connections and Building Skills. In: **Proceedings of the Practice and Experience in Advanced Research Computing 2017 on Sustainability, Success and Impact**. [S.l.]: Association for Computing Machinery, 2017. p. 1–6.

BRYLINSKI, M.; FEINSTEIN, W. P. e findsite: Improved prediction of ligand binding sites in protein models using meta-threading, machine learning and auxiliary ligands. **Journal of computer-aided molecular design**, Springer, v. 27, n. 6, p. 551–567, 2013.

BURLEY, S. K. et al. Rcsb protein data bank: powerful new tools for exploring 3d structures of biological macromolecules for basic and applied research and education in fundamental biology, biomedicine, biotechnology, bioengineering and energy sciences. **Nucleic acids research**, Oxford University Press, v. 49, n. D1, p. D437–D451, 2021.

BUSSI, G.; DONADIO, D.; PARRINELLO, M. Canonical sampling through velocity rescaling. **The Journal of Chemical Physics, v.**, v. 126, 2007.

BUSSI, G.; LAIO, A.; PARRINELLO, M. Equilibrium free energies from nonequilibrium metadynamics. **Physical review letters**, APS, v. 96, n. 9, p. 090601, 2006.

CABOCHE, S. LeView: automatic and interactive generation of 2D diagrams for biomacromolecule/ligand interactions. **journal of cheminformatics**, BioMed Central, v. 5, n. 1, p. 1–7, 2013.

CAMACHO, C. et al. Blast+: architecture and applications. **BMC bioinformatics**, Springer, v. 10, n. 1, p. 421, 2009.

CAPRIOTTI, E.; FARISELLI, P.; CASADIO, R. I-Mutant2.0: predicting stability changes upon mutation from the protein sequence or structure. **Nucleic Acids Research**, v. 33, n. suppl_2, p. W306–W310, 07 2005. ISSN 0305-1048.

CASE, D. A. et al. The Amber biomolecular simulation programs. **journal of Computational Chemistry**, Wiley, v. 26, n. 16, p. 1668–1688, 2005.

CASE, D. A. et al. The Amber biomolecular simulation programs. **journal of Computational Chemistry**, v. 26, n. 16, p. 1668–1688, 2005.

CASSARINO, T. G.; BORDOLI, L.; SCHWEDE, T. Assessment of ligand binding site predictions in CASP10. **Proteins: Structure, Function, and Bioinformatics**, Wiley Online Library, v. 82, p. 154–163, 2014.

CAVALLI, A. et al. Investigating drug–target association and dissociation mechanisms using metadynamics-based algorithms. **Accounts of chemical research**, ACS Publications, v. 48, n. 2, p. 277–285, 2015.

CAZALS, F.; DREYFUS, T. The structural bioinformatics library: modeling in biomolecular science and beyond. **Bioinformatics**, v. 33, n. 7, p. 997–1004, 12 2016. ISSN 1367-4803.

CELNIKER, G. et al. ConSurf: using evolutionary data to raise testable hypotheses about protein function. **Israel journal of Chemistry**, Wiley Online Library, v. 53, n. 3-4, p. 199–206, 2013.

CERETO-MASSAGUÉ, A. et al. Molecular fingerprint similarity search in virtual screening. **Methods**, Elsevier, v. 71, p. 58–63, 2015.

CERETO-MASSAGUé, A. et al. Molecular fingerprint similarity search in virtual screening. **Methods**, v. 71, p. 58–63, 2015. ISSN 1046-2023. Virtual Screening. Disponível em: <https://www.sciencedirect.com/science/article/pii/S1046202314002631>.

CHEN, K. et al. A critical comparative assessment of predictions of protein-binding sites for biologically relevant organic compounds. **Structure**, Elsevier, v. 19, n. 5, p. 613–621, 2011.

CHENG, A. C. et al. Structure-based maximal affinity model predicts small-molecule druggability. **Nature Biotechnology**, Springer Science and Business Media LLC, v. 25, n. 1, p. 71–75, jan. 2007. Disponível em: <https://doi.org/10.1038/nbt1273>.

CHIEN, Y.-T.; HUANG, S.-W. Accurate prediction of protein catalytic residues by side chain orientation and residue contact density. **PloS one**, Public Library of Science San Francisco, USA, v. 7, n. 10, p. e47951, 2012.

CHIEN, Y.-T.; HUANG, S.-W. On the Structural Context and Identification of Enzyme Catalytic Residues. **BioMed research international**, v. 2013, p. 802945, 02 2013.

CHOVANCOVA, E. et al. CAVER 3.0: A Tool for the Analysis of Transport Pathways in Dynamic Protein Structures. **PLoS Computational Biology**, Public Library of Science (PLoS), v. 8, n. 10, p. e1002708, out. 2012.

COCK, P. J. et al. Biopython: freely available python tools for computational molecular biology and bioinformatics. **Bioinformatics**, Oxford University Press, v. 25, n. 11, p. 1422–1423, 2009.

CONSORTIUM, U. UniProt: a worldwide hub of protein knowledge. **Nucleic acids research**, Oxford University Press, v. 47, n. D1, p. D506–D515, 2019.

CONSORTIUM wwPDB. Protein Data Bank: the single global archive for 3D macromolecular structure data. **Nucleic Acids Research**, v. 47, n. D1, p. D520–D528, 10 2018. ISSN 0305-1048.

CROOKS, G. E. et al. Weblogo: a sequence logo generator. **Genome research**, Cold Spring Harbor Lab, v. 14, n. 6, p. 1188–1190, 2004.

D. Szisz. **Chemical Hashed Fingerprint**. 2021. <https://docs.chemaxon.com/display/docs/chemical-hashed-fingerprint.md>. [Online; access 26 August 2021].

DAS, A. A. et al. Pepbind: a comprehensive database and computational tool for analysis of protein–peptide interactions. **Genomics, proteomics & bioinformatics**, Elsevier, v. 11, n. 4, p. 241–246, 2013.

DAUBER-OSGUTHORPE, P.; HAGLER, A. Biomolecular force fields: where have we been, where are we now, where do we need to go and how do we get there? **journal of Computer-Aided Molecular Design**, v. 33, p. 133–203, 2018.

DAVIES, M. et al. ChEMBL web services: streamlining access to drug discovery data and utilities. **Nucleic Acids Research**, Oxford University Press (OUP), v. 43, n. W1, p. W612–W620, abr. 2015. Disponível em: <https://doi.org/10.1093/nar/gkv352>.

DELANO, W. L. **PyMOL**. 2002.

DESTA, I. T. et al. Performance and Its Limits in Rigid Body Protein-Protein Docking. **Structure**, Elsevier BV, v. 28, n. 9, p. 1071–1081.e3, set. 2020.

DOPPELT-AZEROUAL, O. et al. A Review of MED-SuMo Applications. **Infectious Disorders - Drug Targets**, v. 9, n. 3, p. 344–357, 2009. ISSN 1871-5265/2212-3989.

DU, X. et al. Insights into protein–ligand interactions: mechanisms, models, and methods. **International journal of molecular sciences**, Multidisciplinary Digital Publishing Institute, v. 17, n. 2, p. 144, 2016.

DURRANT, J. D.; MCCAMMON, J. A. BINANA: a novel algorithm for ligand-binding characterization. **journal of Molecular Graphics and Modelling**, Elsevier, v. 29, n. 6, p. 888–893, 2011.

DURRANT, J. D.; OLIVEIRA, C. A. F. de; MCCAMMON, J. A. Povme: an algorithm for measuring binding-pocket volumes. **Journal of Molecular Graphics and Modelling**, Elsevier, v. 29, n. 5, p. 773–776, 2011.

EF, P. et al. Ucsf Chimera–a visualization system for exploratory research and analysis. **J Comput Chem**, v. 25, n. 13, p. 1605–12, Oct 2004.

EKINS, S. et al. Exploiting machine learning for end-to-end drug discovery and development. **Nature materials**, Nature Publishing Group, v. 18, n. 5, p. 435, 2019.

EL-GEBALI, S. et al. The Pfam protein families database in 2019. **Nucleic acids research**, Oxford University Press, v. 47, n. D1, p. D427–D432, 2019.

ESWAR, N. et al. Tools for comparative protein structure modeling and analysis. **Nucleic acids research**, Oxford University Press, v. 31, n. 13, p. 3375–3380, 2003.

FASSIO, A. V. et al. napoli: a graph-based strategy to detect and visualize conserved protein-ligand interactions in large-scale. **IEEE/ACM transactions on computational biology and bioinformatics**, IEEE, 2019.

FERRER-COSTA, C. et al. PMUT: a web-based tool for the annotation of pathological mutations on proteins. **Bioinformatics**, v. 21, n. 14, p. 3176–3178, 05 2005.

FILIPOVIČ, J. et al. CaverDock: a novel method for the fast analysis of ligand transport. **IEEE/ACM transactions on computational biology and bioinformatics**, IEEE, v. 17, n. 5, p. 1625–1638, 2019.

FINN, R. D.; CLEMENTS, J.; EDDY, S. R. Hmmer web server: interactive sequence similarity searching. **Nucleic acids research**, Oxford University Press, v. 39, n. suppl_2, p. W29–W37, 2011.

FLORIÁN, J.; WARSHEL, A. **ChemSol, version 2.1**. [S.l.]: University of Southern California: Los Angeles, 1999.

FOX, N. K.; BRENNER, S. E.; CHANDONIA, J.-M. SCOPe: Structural Classification of Proteins—extended, integrating SCOP and ASTRAL data and classification of new structures. **Nucleic acids research**, Oxford University Press, v. 42, n. D1, p. D304–D309, 2014.

FRAPPIER, V.; DURAN, M.; KEATING, A. E. Pixeldb: Protein–peptide complexes annotated with structural conservation of the peptide binding mode. **Protein Science**, Wiley Online Library, v. 27, n. 1, p. 276–285, 2018.

FURNHAM, N. et al. The Catalytic Site Atlas 2.0: cataloging catalytic sites and residues identified in enzymes. **Nucleic Acids Research**, v. 42, n. D1, p. D485–D489, 12 2013. ISSN 0305-1048.

GALINDEZ, G. et al. Lessons from the covid-19 pandemic for advancing computational drug repurposing strategies. **Nature Computational Science**, Nature Publishing Group, v. 1, n. 1, p. 33–41, 2021.

GALPERIN, M. Y.; FERNáNDEZ-SUáREZ, X. M.; RIGDEN, D. J. The 24th annual Nucleic Acids Research database issue: a look back and upcoming changes. **Nucleic Acids Research**, v. 45, n. D1, p. D1–D11, 12 2016. ISSN 0305-1048.

GARCIA-HERNANDEZ, C.; FERNANDEZ, A.; SERRATOSA, F. Ligand-based virtual screening using graph edit distance as molecular similarity measure. **Journal of chemical information and modeling**, ACS Publications, v. 59, n. 4, p. 1410–1421, 2019.

GARCIA, L. et al. **Ten simple rules for making training materials FAIR**. [S.l.]: Public Library of Science San Francisco, CA USA, 2020. e1007854 p.

GAUTAM, A. et al. Hemolytik: a database of experimentally determined hemolytic and non-hemolytic peptides. **Nucleic acids research**, Oxford University Press, v. 42, n. D1, p. D444–D449, 2014.

GAUTAM, A. et al. Cppsite: a curated database of cell penetrating peptides. **Database**, Narnia, v. 2012, 2012.

GENBANK Internet, Bethesda MD. National Library of Medicine (US), National Center for Biotechnology Information, 1982. Disponível em: <https://www.ncbi.nlm.nih.gov/nuccore/JX898746.1>.

GILSON, M. K. et al. Bindingdb in 2015: A public database for medicinal chemistry, computational chemistry and systems pharmacology. **Nucleic acids research**, Oxford University Press, v. 44, n. D1, p. D1045–D1053, Jan 2016. ISSN 1362-4962. Gkv1072[PII].

GIMENO, A. et al. The light and dark sides of virtual screening: What is there to know? **International Journal of Molecular Sciences**, MDPI AG, v. 20, n. 6, p. 1375, mar. 2019. Disponível em: <https://doi.org/10.3390/ijms20061375>.

GOEL, S. et al. Diamond machining of silicon: a review of advances in molecular dynamics simulation. **International journal of Machine Tools and Manufacture**, v. 88, p. 131–164, jan 2015. ISSN 0890-6955.

GOETZ, D. et al. Substrate specificity profiling and identification of a new class of inhibitor for the major protease of the sars coronavirus. **Biochemistry**, ACS Publications, v. 46, n. 30, p. 8744–8752, 2007.

GOMES, I. d. S. et al. Computational prediction of potential inhibitors for sars-cov-2 main protease based on machine learning, docking, mm-pbsa calculations, and metadynamics. **PLOS ONE**, Public Library of Science, v. 17, n. 4, p. 1–20, 04 2022. Disponível em: <https://doi.org/10.1371/journal.pone.0267471>.

GONCZAREK, A. et al. Interaction prediction in structure-based virtual screening using deep learning. **Computers in biology and medicine**, Elsevier, v. 100, p. 253–258, 2018.

GONZALEZ-PEREZ, A.; LÓPEZ-BIGAS, N. Improving the assessment of the outcome of nonsynonymous SNVs with a consensus deleteriousness score, Condel. **American journal of human genetics**, v. 88 4, p. 440–9, 2011.

GROSDIDIER, A.; ZOETE, V.; MICHIELIN, O. SwissDock, a protein-small molecule docking web service based on EADock DSS. **Nucleic Acids Research**, Oxford University Press (OUP), v. 39, n. suppl, p. W270–W277, maio 2011.

GUSTEREN, V.; F., W.; BERENDSEN, H. J. C. Computer simulation of molecular dynamics: Methodology, applications, and perspectives in chemistry. **Angewandte Chemie International Edition in English, v.**, v. 29, p. 992–1023, 1990.

GUVENCH, O.; MACKERELL, A. D. Comparison of Protein Force Fields for Molecular Dynamics Simulations. In: **Methods in Molecular Biology**. [S.l.]: Humana Press, 2008. p. 63–88.

HAGBERG, A.; SWART, P.; CHULT, D. S. **Exploring network structure, dynamics, and function using NetworkX**. [S.l.], 2008.

HAIXIANG, G. et al. Learning from class-imbalanced data: Review of methods and applications. **Expert Systems with Applications**, Elsevier, v. 73, p. 220–239, 2017.

HAMELRYCK, T.; MANDERICK, B. Pdb file parser and structure class implemented in python. **Bioinformatics**, Oxford University Press, v. 19, n. 17, p. 2308–2310, 2003.

HANSON, B. et al. Jmol: an open-source Java viewer for chemical structures in 3D. **URL: www. jmol. sourceforgenet. net.–2008**, 2008.

HANWELL, M. D. et al. Avogadro: an advanced semantic chemical editor, visualization, and analysis platform. **journal of Cheminformatics**, Springer Science and Business Media LLC, v. 4, n. 1, ago. 2012.

HARVEY, M. J.; GIUPPONI, G.; FABRITIIS, G. D. ACEMD: accelerating biomolecular dynamics in the microsecond time scale. **journal of chemical theory and computation**, ACS Publications, v. 5, n. 6, p. 1632–1639, 2009.

HEO, L.; PARK, H.; SEOK, C. GalaxyRefine: Protein structure refinement driven by side-chain repacking. **Nucleic Acids Res**, v. 41, n. Web Server issue, p. W384–388, Jul 2013.

HERRÁEZ, A. Biomolecules in the computer: Jmol to the rescue. **Biochemistry and Molecular Biology Education**, Wiley, v. 34, n. 4, p. 255–261, jul. 2006.

HESTENES, M. R.; STIEFEL, E. Methods of conjugate gradients for solving linear systems. **Journal of Research of the National Bureau of Standards, v.**, v. 49, p. 409–436, 1952.

HOLLIDAY, G. L. et al. MACiE (Mechanism, Annotation and Classification in Enzymes): novel tools for searching catalytic mechanisms. **Nucleic Acids Research**, v. 35, n. suppl_1, p. D515–D520, 11 2006. ISSN 0305-1048.

HOLLIDAY, G. L. et al. MACiE: exploring the diversity of biochemical reactions. **Nucleic acids research**, Oxford University Press, v. 40, n. Database issue, p. D783–D789, Jan 2012. ISSN 1362-4962.

HOLLIDAY, G. L. et al. MACiE: a database of enzyme reaction mechanisms. **Bioinformatics**, v. 21, n. 23, p. 4315–4316, 09 2005. ISSN 1367-4803.

HONORATO, R. V. et al. Structural Biology in the Clouds: The WeNMR-EOSC Ecosystem. **Frontiers in Molecular Biosciences**, v. 8, p. 708, 2021. ISSN 2296-889X.

HOPF, T. A. et al. Mutation effects predicted from sequence co-variation. **Nature Biotechnology**, v. 35, p. W128–135, 2017.

HUANG, J.; JR, A. D. M. Charmm36 all-atom additive protein force field: Validation based on comparison to nmr data. **Journal of computational chemistry**, Wiley Online Library, v. 34, n. 25, p. 2135–2145, 2013.

HUBBARD, S. J.; THORNTON, J. M. Naccess. **Computer Program, Department of Biochemistry and Molecular Biology, University College London**, v. 2, n. 1, 1993.

HUMPHREY, W.; DALKE, A.; SCHULTEN, K. VMD – Visual Molecular Dynamics. **journal of Molecular Graphics**, v. 14, p. 33–38, 1996.

HUMPHREY, W.; DALKE, A.; SCHULTEN, K. Vmd: Visual molecular dynamics. **Journal of Molecular Graphics**, v. 14, n. 1, p. 33 – 38, 1996. ISSN 0263-7855. Disponível em: <http://www.sciencedirect.com/science/article/pii/0263785596000185>.

HUMPHREY, W. et al. Vmd: visual molecular dynamics. **Journal of molecular graphics**, Guildford: Butterworth Scientific Limited, c1983-c1996., v. 14, n. 1, p. 33–38, 1996.

IGLESIAS, J. et al. Computational structure-based drug design: Predicting target flexibility. **WIREs Computational Molecular Science**, Wiley, v. 8, n. 5, abr. 2018.

ISHITANI, R.; NAKANE, T. **CueMol: molecular visualization framework**. 2014.

IZIDORO, S.; LACERDA, A. M.; PAPPA, G. L. MeGASS: Multi-Objective Genetic Active Site Search. In: **Proceedings of the Companion Publication of the 2015 Annual Conference on Genetic and Evolutionary Computation**. [S.l.: s.n.], 2015. p. 905–910.

IZIDORO, S. C.; MELO-MINARDI, R. C. de; PAPPA, G. L. GASS: identifying enzyme active sites with genetic algorithms. **Bioinformatics**, Oxford University Press, v. 31, n. 6, p. 864–870, 2015.

JAMASB, A. R. et al. Deep for Protein–Protein Interaction Site Prediction. In: **Methods in Molecular Biology**. [S.l.]: Springer US, 2021. p. 263–288.

JANIN, J. et al. CAPRI: A Critical Assessment of PRedicted Interactions. **Proteins: Structure, Function, and Genetics**, Wiley, v. 52, n. 1, p. 2–9, maio 2003.

JIMÉNEZ, J. et al. DeepSite: protein-binding site predictor using 3D-convolutional neural networks. **Bioinformatics**, Oxford University Press, v. 33, n. 19, p. 3036–3042, 2017.

JIN, Z. et al. Structure of mpro from sars-cov-2 and discovery of its inhibitors. **Nature**, v. 582, n. 7811, p. 289–293, Jun 2020. ISSN 1476-4687. Disponível em: <https://doi.org/10.1038/s41586-020-2223-y>.

JONES, G. et al. Development and validation of a genetic algorithm for flexible docking 1 1Edited by F. E. Cohen. **journal of Molecular Biology**, Elsevier BV, v. 267, n. 3, p. 727–748, abr. 1997.

JONES, K. S. A statistical interpretation of term specificity and its application in retrieval. **journal of documentation**, MCB UP Ltd, 1972.

JUMPER, J. et al. Highly accurate protein structure prediction with alphafold. **Nature**, Nature Publishing Group, v. 596, n. 7873, p. 583–589, 2021.

KÄLLBERG, M. et al. Template-based protein structure modeling using the raptorx web server. **Nature protocols**, Nature Publishing Group, v. 7, n. 8, p. 1511–1522, 2012.

KARLIN, S.; ALTSCHUL, S. F. Methods for assessing the statistical significance of molecular sequence features by using general scoring schemes. **Proceedings of the National Academy of Sciences**, National Acad Sciences, v. 87, n. 6, p. 2264–2268, 1990.

KELLEY, L. A.; GARDNER, S. P.; SUTCLIFFE, M. J. An automated approach for clustering an ensemble of NMR-derived protein structures into conformationally related subfamilies. **Protein Engineering, Design and Selection**, Oxford University Press, v. 9, n. 11, p. 1063–1065, 1996.

KELLEY, L. A. et al. Discovering rules for protein–ligand specificity using support vector inductive logic programming. **Protein Engineering, Design & Selection**, Oxford University Press, v. 22, n. 9, p. 561–567, 2009.

KESKIN, O. et al. Principles of protein- protein interactions: what are the preferred ways for proteins to interact? **Chemical reviews**, ACS Publications, v. 108, n. 4, p. 1225–1244, 2008.

KO, J. et al. GalaxyWEB server for protein structure prediction and refinement. **Nucleic acids research**, Oxford University Press, v. 40, n. W1, p. W294–W297, 2012.

KO, J.; PARK, H.; SEOK, C. GalaxyTBM: template-based modeling by building a reliable core and refining unreliable local regions. **BMC bioinformatics**, Springer, v. 13, n. 1, p. 1–8, 2012.

KOKH, D. B. et al. **TRAPP: A Tool for Analysis of Tra nsient Binding P ockets in P roteins**. [S.l.]: ACS Publications, 2013.

KOLIŃSKI, A. et al. Protein modeling and structure prediction with a reduced representation. **Acta Biochimica Polonica**, v. 51, 2004.

KONAGURTHU, A. S. et al. Mustang: a multiple structural alignment algorithm. **Proteins: Structure, Function, and Bioinformatics**, Wiley Online Library, v. 64, n. 3, p. 559–574, 2006.

KONC, J.; JANEŽIČ, D. Probis algorithm for detection of structurally similar protein binding sites by local structural alignment. **Bioinformatics**, Oxford University Press, v. 26, n. 9, p. 1160–1168, 2010.

KONDRATYUK, N. et al. GPU-accelerated molecular dynamics: State-of-art software performance and porting from Nvidia CUDA to AMD HIP. **The International journal of High Performance Computing Applications**, v. 35, n. 4, p. 312–324, 2021.

KOZAKOV, D. et al. The ClusPro web server for protein–protein docking. **Nature Protocols**, Springer Science and Business Media LLC, v. 12, n. 2, p. 255–278, jan. 2017.

KRASSOWSKI, M. et al. ActiveDriverDB: human disease mutations and genome variation in post-translational modification sites of proteins. **Nucleic Acids Research**, v. 46, n. D1, p. D901–D910, 11 2017. ISSN 0305-1048.

KREJCI, A. et al. Hammock: a hidden markov model-based peptide clustering algorithm to identify protein-interaction consensus motifs in large datasets. **Bioinformatics**, Oxford University Press, v. 32, n. 1, p. 9–16, 2016.

KRIEGER, E.; VRIEND, G. YASARA View—molecular graphics for all devices—from smartphones to workstations. **Bioinformatics**, Oxford University Press, v. 30, n. 20, p. 2981–2982, 2014.

KRIEGER, E.; VRIEND, G. YASARA View—molecular graphics for all devices—from smartphones to workstations. **Bioinformatics**, Oxford University Press, v. 30, n. 20, p. 2981–2982, 2014.

KRIEGER, E.; VRIEND, G. New ways to boost molecular dynamics simulations. **journal of computational chemistry**, Wiley Online Library, v. 36, n. 13, p. 996–1007, 2015.

KRYSHTAFOVYCH, A. et al. Critical assessment of methods of protein structure prediction (CASP)—Round XIV. **Proteins: Structure, Function, and Bioinformatics**, Wiley Online Library, v. 89, n. 12, p. 1607–1617, 2021.

KUFAREVA, I.; ILATOVSKIY, A. V.; ABAGYAN, R. Pocketome: an encyclopedia of small-molecule binding sites in 4d. **Nucleic acids research**, Oxford University Press, v. 40, n. D1, p. D535–D540, 2012.

KUHLMAN, B.; BRADLEY, P. Advances in protein structure prediction and design. **Nature Reviews Molecular Cell Biology**, Nature Publishing Group, v. 20, n. 11, p. 681–697, 2019.

KÜHNE, R.; EBERT, R.-U.; SCHÜÜRMANN, G. Chemical domain of qsar models from atom-centered fragments. **Journal of Chemical Information and Modeling**, American Chemical Society, v. 49, n. 12, p. 2660–2669, Dec 2009. ISSN 1549-9596. Disponível em: <https://doi.org/10.1021/ci900313u>.

Kumar, P.; Henikoff, S.; Ng, P. C. Predicting the effects of coding non-synonymous variants on protein function using the SIFT algorithm. **Nature Protocols**, v. 4, n. 7, p. 1073–1081, 2009.

KUMARI, R.; KUMAR, R.; LYNN, A. g_mmpbsa—a gromacs tool for high-throughput mm-pbsa calculations. **Journal of Chemical Information and Modeling**, American Chemical Society, v. 54, n. 7, p. 1951–1962, Jul 2014. ISSN 1549-9596. Disponível em: <https://doi.org/10.1021/ci500020m>.

LASKOWSKI, R. A.; SWINDELLS, M. B. **LigPlot+: multiple ligand–protein interaction diagrams for drug discovery**. [S.l.]: ACS Publications, 2011.

LAU, J. L.; DUNN, M. K. Therapeutic peptides: Historical perspectives, current development trends, and future directions. **Bioorganic & medicinal chemistry**, Elsevier, v. 26, n. 10, p. 2700–2707, 2018.

LE, T. T.; FU, W.; MOORE, J. H. Scaling tree-based automated machine learning to biomedical big data with a feature set selector. **Bioinformatics**, Oxford University Press, v. 36, n. 1, p. 250–256, 2020.

LEE, A. C.-L. et al. A comprehensive review on current advances in peptide drug development and design. **International journal of molecular sciences**, Multidisciplinary Digital Publishing Institute, v. 20, n. 10, p. 2383, 2019.

LEE, B.; RICHARDS, F. M. The interpretation of protein structures: estimation of static accessibility. **Journal of molecular biology**, Academic Press, v. 55, n. 3, p. 379–IN4, 1971.

LEE, H. et al. GalaxyPepDock: a protein–peptide docking tool based on interaction similarity and energy optimization. **Nucleic acids research**, Oxford University Press, v. 43, n. W1, p. W431–W435, 2015.

LEMAN, J. K. et al. Macromolecular modeling and design in Rosetta: recent methods and frameworks. **Nature methods**, Nature Publishing Group, v. 17, n. 7, p. 665–680, 2020.

LEŠNIK, S. et al. Lisica: A software for ligand-based virtual screening and its application for the discovery of butyrylcholinesterase inhibitors. **Journal of Chemical Information and Modeling**, American Chemical Society, v. 55, n. 8, p. 1521–1528, Aug 2015. ISSN 1549-9596. Disponível em: <https://doi.org/10.1021/acs.jcim.5b00136>.

LI, H. et al. Machine-learning scoring functions for structure-based virtual screening. **WIREs Computational Molecular Science**, v. 11, n. 1, p. e1478, 2021. Disponível em: <https://onlinelibrary.wiley.com/doi/abs/10.1002/wcms.1478>.

LIU, D. et al. Self-assembly of mitochondria-specific peptide amphiphiles amplifying lung cancer cell death through targeting the vdac1–hexokinase-ii complex. **Journal of Materials Chemistry B**, Royal Society of Chemistry, v. 7, n. 30, p. 4706–4716, 2019.

LO, Y.-C. et al. Machine learning in chemoinformatics and drug discovery. **Drug discovery today**, Elsevier, v. 23, n. 8, p. 1538–1546, 2018.

LONDON, N.; MOVSHOVITZ-ATTIAS, D.; SCHUELER-FURMAN, O. The structural basis of peptide-protein binding strategies. **Structure**, Elsevier, v. 18, n. 2, p. 188–199, 2010.

LONDON, N.; RAVEH, B.; SCHUELER-FURMAN, O. Modeling peptide–protein interactions. In: **Homology Modeling**. [S.l.]: Springer, 2011. p. 375–398.

LOUDEN, P. et al. OPENMD-2.5: molecular dynamics in the open. **Department of Chemistry and Biochemistry, University of Notre Dame, Notre Dame, Indiana 46556**, 2017.

LUO, H. et al. Biomedical data and computational models for drug repositioning: a comprehensive review. **Briefings in Bioinformatics**, Oxford University Press (OUP), fev. 2020. Disponível em: <https://doi.org/10.1093/bib/bbz176>.

LUSCOMBE, N.; GREENBAUM, D.; GERSTEIN, M. What is bioinformatics? A proposed definition and overview of the field. **Methods of information in medicine**, FK SCHATTAUER VERLAGSGESELLSCHAFT MBH, v. 40, n. 4, p. 346–358, 2001.

MALLICK, M.; VIDYARTHI, A. S. et al. Tools for predicting metal binding sites in protein: a review. **Current Bioinformatics**, Bentham Science Publishers, v. 6, n. 4, p. 444–449, 2011.

MARSILI, S. et al. ORAC: A molecular dynamics simulation program to explore free energy surfaces in biomolecular systems at the atomistic level. **journal of computational chemistry**, Wiley Online Library, v. 31, n. 5, p. 1106–1116, 2010.

MARTINS, P. M. et al. Propedia: a database for protein–peptide identification based on a hybrid clustering algorithm. **BMC Bioinformatics**, Springer Science and Business Media LLC, v. 22, n. 1, jan. 2021. Disponível em: <https://doi.org/10.1186/s12859-020-03881-z>.

MASSO, M.; VAISMAN, I. I. AUTO-MUTE: web-based tools for predicting stability changes in proteins due to single amino acid replacements. **Protein Engineering Design and Selection**, Oxford University Press (OUP), v. 23, n. 8, p. 683–687, jun. 2010.

MCCOY, M. D. et al. Predicting Genetic Variation Severity Using Machine Learning to Interpret Molecular Simulations. **Biophysical journal**, v. 120, n. 2, p. 189–204, 2021. ISSN 0006-3495.

MCCULLOUGH, J. et al. **Annotating Putative D. discoideum Proteins Using I-TASSER**. 2021.

MEDEMA, M. H. The year 2020 in natural product bioinformatics: an overview of the latest tools and databases. **Natural Product Reports**, Royal Society of Chemistry (RSC), v. 38, n. 2, p. 301–306, 2021.

MENDOLIA, I. et al. Convolutional architectures for virtual screening. **BMC Bioinformatics**, Springer Science and Business Media LLC, v. 21, n. S8, set. 2020. Disponível em: <https://doi.org/10.1186/s12859-020-03645-9>.

MEYER, M. J. et al. Interactome INSIDER: a structural interactome browser for genomic studies. **Nature Methods**, v. 15, p. W107–114, 2018.

MIRZA, M. U.; FROEYEN, M. Structural elucidation of sars-cov-2 vital proteins: Computational methods reveal potential drug candidates against main protease, nsp12 polymerase and nsp13 helicase. **Journal of Pharmaceutical Analysis**, 2020. ISSN 2095-1779. Disponível em: <http://www.sciencedirect.com/science/article/pii/S2095177920303865>.

(MOE), M. O. E. 01; chemical computing group ulc, 1010 sherbooke st. **West, Suite**, v. 7, p. 2021, 2019.

MONTICELLI, L.; TIELEMAN, D. P. Force Fields for Classical Molecular Dynamics. In: **Methods in Molecular Biology**. [S.l.]: Humana Press, 2012. p. 197–213.

MORAES, J. P. et al. GASS-WEB: a web server for identifying enzyme active sites based on genetic algorithms. **Nucleic acids research**, Oxford University Press, v. 45, n. W1, p. W315–W319, 2017.

MOREIRA, L. et al. Survival and developmental impairment induced by the trypsin inhibitor bis-benzamidine in the velvetbean caterpillar (anticarsia gemmatalis). **Crop Protection**, Elsevier, v. 30, n. 10, p. 1285–1290, 2011.

MORRIS, G. M. et al. Autodock4 and autodocktools4: Automated docking with selective receptor flexibility. **Journal of computational chemistry**, v. 30, n. 16, p. 2785–2791, Dec 2009. ISSN 1096-987X. 19399780[pmid].

MORSE, P.; FESHBACH, H. **Methods of theoretical physics. Part I: chapters 1 to 8**. [S.l.]: McGraw Hill Book Co, 1953.

MOSCARDI, F. et al. Capítulo 4-artrópodes que atacam as folhas da soja. **Soja-manejo integrado de insetos e outros artrópodes-praga. Brasília: Embrapa**, p. 213–334, 2012.

MOULT, J. et al. Critical assessment of methods of protein structure prediction (CASP)—Round XII. **Proteins: Structure, Function, and Bioinformatics**, Wiley Online Library, v. 86, p. 7–15, 2018.

MULDER, N. et al. The development and application of bioinformatics core competencies to improve bioinformatics training and education. **PLoS computational biology**, Public Library of Science San Francisco, CA USA, v. 14, n. 2, p. e1005772, 2018.

MYSINGER, M. M. et al. Directory of useful decoys, enhanced (dud-e): better ligands and decoys for better benchmarking. **Journal of medicinal chemistry**, ACS Publications, v. 55, n. 14, p. 6582–6594, 2012.

NAKANE, T. GLmol-Molecular Viewer on WebGL. **Javascript. webglmol. sourceforge. jp**, 2014.

Nature. **Bioinformatics**. 2021. <https://www.nature.com/subjects/bioinformatics>. [Online; access 10 August 2021].

NEDUVA, V. et al. Systematic discovery of new recognition peptides mediating protein interaction networks. **PLoS biology**, Public Library of Science, v. 3, n. 12, 2005.

NERATTINI, F.; CHELLI, R.; PROCACCI, P. II. Dissociation free energies in drug–receptor systems via nonequilibrium alchemical simulations: application to the FK506-related immunophilin ligands. **Physical Chemistry Chemical Physics**, Royal Society of Chemistry, v. 18, n. 22, p. 15005–15018, 2016.

NG, A. Y.; JORDAN, M. I.; WEISS, Y. On spectral clustering: Analysis and an algorithm. In: **Advances in neural information processing systems**. [S.l.: s.n.], 2002. p. 849–856.

NG, P. C.; HENIKOFF, S. SIFT: predicting amino acid changes that affect protein function. **Nucleic Acids Research**, v. 31, n. 13, p. 3812–3814, 07 2003. ISSN 0305-1048.

NIKAM, R. et al. ProThermDB: thermodynamic database for proteins and mutants revisited after 15 years. **Nucleic Acids Research**, v. 49, n. D1, p. D420–D424, 11 2020. ISSN 0305-1048.

OLSSON, T. S. et al. The thermodynamics of protein–ligand interaction and solvation: insights for ligand design. **journal of molecular biology**, Elsevier, v. 384, n. 4, p. 1002–1017, 2008.

OPENASTEXVIEWER. **OpenAstexViewer**. <http://www.openastexviewer.net/>. [Online; accessed 18-April-2022].

PAGE, M. J.; CERA, E. D. Serine peptidases: classification, structure and function. **Cellular and Molecular Life Sciences**, Springer, v. 65, n. 7-8, p. 1220–1236, 2008.

PAI, P. P.; RANJANI, S. S. S.; MONDAL, S. PINGU: PredIction of eNzyme catalytic residues usinG seqUence information. **PloS one**, Public Library of Science, v. 10, n. 8, p. e0135122–e0135122, Aug 2015. ISSN 1932-6203.

PAIVA, V. de A. et al. Protein structural bioinformatics: An overview. **Computers in Biology and Medicine**, p. 105695, 2022. ISSN 0010-4825. Disponível em: <https://www.sciencedirect.com/science/article/pii/S0010482522004784>.

PAN, A. C. et al. Atomic-level characterization of protein–protein association. **Proceedings of the National Academy of Sciences**, National Acad Sciences, v. 116, n. 10, p. 4244–4249, 2019.

PAN, L.; ALLER, S. Tools and procedures for visualization of proteins and other biomolecules. **Current Protocols in Molecular Biology**, v. 2015, p. 19.12.1–19.12.47, 2015.

PANT, S. et al. Peptide-like and small-molecule inhibitors against covid-19. **Journal of Biomolecular Structure and Dynamics**, Taylor & Francis, v. 39, n. 8, p. 1–15, 2020.

PARRINELLO, M.; RAHMAN, A. Polymorphic transitions in single crystals: A new molecular dynamics method. **Journal of Applied Physics, v.**, v. 52, 1981.

PATARROYO-VARGAS, A. M. et al. Kinetic characterization of anticarsia gemmatalis digestive serine-proteases and the inhibitory effect of synthetic peptides. **Protein and peptide letters**, Bentham Science Publishers, v. 24, n. 11, p. 1040–1047, 2017.

PEDREGOSA, F. et al. Scikit-learn: Machine learning in Python. **Journal of Machine Learning Research**, v. 12, p. 2825–2830, 2011.

PERKINELMER. **ChemBioDraw**. [S.l.]: CambridgeSoft Waltham, MA, USA, 2012.

PERONA, J. J.; CRAIK, C. S. Structural basis of substrate specificity in the serine proteases. **Protein Science**, Wiley Online Library, v. 4, n. 3, p. 337–360, 1995.

PETTERSEN, E. F. et al. UCSF ChimeraX: Structure visualization for researchers, educators, and developers. **Protein Science**, Wiley Online Library, v. 30, n. 1, p. 70–82, 2021.

PHILLIPS, J. C. et al. Scalable molecular dynamics with namd. **Journal of Computational Chemistry**, v. 26, n. 16, p. 1781–1802, 2005. Disponível em: <https://onlinelibrary.wiley.com/doi/abs/10.1002/jcc.20289>.

PHILLIPS, J. C. et al. Scalable molecular dynamics with namd. **Journal of computational chemistry**, Wiley Online Library, v. 26, n. 16, p. 1781–1802, 2005.

PHILLIPS, J. C. et al. Scalable molecular dynamics on CPU and GPU architectures with NAMD. **The journal of chemical physics**, AIP Publishing LLC, v. 153, n. 4, p. 044130, 2020.

PIERCE, B. G. et al. ZDOCK server: interactive docking prediction of protein-protein complexes and symmetric multimers. **Bioinformatics**, Oxford University Press (OUP), v. 30, n. 12, p. 1771–1773, fev. 2014.

PILON, A. M. et al. Protease inhibitory, insecticidal and deterrent effects of the trypsin-inhibitor benzamidine on the velvetbean caterpillar in soybean. **Anais da Academia Brasileira de Ciências**, SciELO Brasil, v. 90, n. 4, p. 3475–3482, 2018.

PILON, F. M. et al. Purification and characterization of trypsin produced by gut bacteria from anticarsia gemmatalis. **Archives of insect biochemistry and physiology**, Wiley Online Library, v. 96, n. 2, p. e21407, 2017.

PLAXCO, K. W.; SIMONS, K. T.; BAKER, D. Contact order, transition state placement and the refolding rates of single domain proteins. **Journal of molecular biology**, Elsevier, v. 277, n. 4, p. 985–994, 1998.

PLIMPTON, S. Fast parallel algorithms for short-range molecular dynamics. **journal of computational physics**, Elsevier, v. 117, n. 1, p. 1–19, 1995.

PORTER, C. T.; BARTLETT, G. J.; THORNTON, J. M. The Catalytic Site Atlas: a resource of catalytic sites and residues identified in enzymes using structural data. **Nucleic Acids Research**, v. 32, n. suppl_1, p. D129–D133, 01 2004. ISSN 0305-1048.

PRICE, D. J.; III, C. L. B. A modified tip3p water potential for simulation with ewald summation. **The Journal of chemical physics**, American Institute of Physics, v. 121, n. 20, p. 10096–10103, 2004.

PROCACCI, P. **Hybrid MPI/OpenMP implementation of the ORAC molecular dynamics program for generalized ensemble and fast switching alchemical simulations**. [S.l.]: ACS Publications, 2016.

PROCACCI, P. Primadorac: A free web interface for the assignment of partial charges, chemical topology, and bonded parameters in organic or drug molecules. **journal of chemical information and modeling**, ACS Publications, v. 57, n. 6, p. 1240–1245, 2017.

PROCACCI, P.; CARDELLI, C. Fast switching alchemical transformations in molecular dynamics simulations. **journal of chemical theory and computation**, ACS Publications, v. 10, n. 7, p. 2813–2823, 2014.

PROCACCI, P. et al. ORAC: A molecular dynamics program to simulate complex molecular systems with realistic electrostatic interactions. **journal of computational chemistry**, Wiley Online Library, v. 18, n. 15, p. 1848–1862, 1997.

PROGRAM, A. **Another Molecular Mechanics Program VE**. 2012. <https://www.ddl.unimi.it/cms/index.php?Software_projects:AMMP_VE>. [Online; accessed 9-December-2021].

PUTIGNANO, V. et al. MetalPDB in 2018: a database of metal sites in biological macromolecular structures. **Nucleic Acids Research**, v. 46, n. D1, p. D459–D464, 10 2017. ISSN 0305-1048.

QIAO, L.; XIE, D. MIonSite: Ligand-specific prediction of metal ion-binding sites via enhanced AdaBoost algorithm with protein sequence information. **Analytical Biochemistry**, v. 566, p. 75–88, 2019. ISSN 0003-2697.

QUEIROZ, F. C. et al. ppiGReMLIN: a graph mining based detection of conserved structural arrangements in protein-protein interfaces. **BMC bioinformatics**, Springer, v. 21, p. 1–25, 2020.

RACKERS, J. A. et al. Tinker 8: software tools for molecular design. **journal of chemical theory and computation**, ACS Publications, v. 14, n. 10, p. 5273–5289, 2018.

RAITERI, P. et al. Efficient reconstruction of complex free energy landscapes by multiple walkers metadynamics. **The journal of physical chemistry B**, ACS Publications, v. 110, n. 8, p. 3533–3539, 2006.

RAKOVAN, J. Computer Programs for Drawing Crystal Shapes and Atomic Structures. **Rocks & Minerals**, Taylor & Francis, v. 93, n. 1, p. 60–64, 2018.

Ramensky, V.; Bork, P.; Sunyaev, S. Human non-synonymous SNPs: server and survey. **Nucleic Acids Research**, v. 30, n. 17, p. 3894–3900, 2002.

RAO, V. S. et al. Protein-protein interaction detection: methods and analysis. **International journal of proteomics**, Hindawi, v. 2014, 2014.

RAVEH, B.; LONDON, N.; SCHUELER-FURMAN, O. Sub-angstrom modeling of complexes between flexible peptides and globular proteins. **Proteins: Structure, Function, and Bioinformatics**, Wiley Online Library, v. 78, n. 9, p. 2029–2040, 2010.

RAWLINGS, N. D.; BARRETT, A. J.; BATEMAN, A. Merops: the peptidase database. **Nucleic acids research**, Oxford University Press, v. 38, n. suppl_1, p. D227–D233, 2010.

RAWLINGS, N. D. et al. The MEROPS database of proteolytic enzymes, their substrates and inhibitors in 2017 and a comparison with peptidases in the PANTHER database. **Nucleic Acids Research**, v. 46, n. D1, p. D624–D632, 11 2017. ISSN 0305-1048.

REGO, N.; KOES, D. 3Dmol.js: Molecular Visualization with WebGL. **Bioinformatics (Oxford, England)**, v. 31, 12 2014.

REGO, N.; KOES, D. 3dmol. js: molecular visualization with webgl. **Bioinformatics**, Oxford University Press, v. 31, n. 8, p. 1322–1324, 2015.

RELEASE, S. 4: Ligprep. **Schrödinger, LLC, New York, NY**, 2019.

REYNOLDS, C.; ISLAM, S.; STERNBERG, M. EzMol: A Web Server Wizard for the Rapid Visualization and Image Production of Protein and Nucleic Acid Structures. **journal of Molecular Biology**, v. 430, n. 15, p. 2244–2248, 2018. Cited By 48.

RIBEIRO, A. J. M. et al. Mechanism and Catalytic Site Atlas (M-CSA): a database of enzyme reaction mechanisms and active sites. **Nucleic Acids Research**, v. 46, n. D1, p. D618–D623, 11 2017. ISSN 0305-1048.

RIFAIOGLU, A. S. et al. Recent applications of deep learning and machine intelligence on in silico drug discovery: methods, tools and databases. **Briefings in bioinformatics**, Oxford University Press, v. 20, n. 5, p. 1878–1912, 2019.

RINIKER, S. Fixed-Charge Atomistic Force Fields for Molecular Dynamics Simulations in the Condensed Phase: An Overview. **journal of Chemical Information and Modeling**, American Chemical Society (ACS), v. 58, n. 3, p. 565–578, mar. 2018.

RIVAS, J. D. L.; FONTANILLO, C. Protein–protein interactions essentials: key concepts to building and analyzing interactome networks. **PLoS Comput Biol**, Public Library of Science, v. 6, n. 6, p. e1000807, 2010.

ROCHE, D. B.; BUENAVISTA, M. T.; MCGUFFIN, L. J. FunFOLDQA: a quality assessment tool for protein-ligand binding site residue predictions. **PloS one**, Public Library of Science, v. 7, n. 5, p. e38219, 2012.

ROCHE, D. B.; BUENAVISTA, M. T.; MCGUFFIN, L. J. The FunFOLD2 server for the prediction of protein–ligand interactions. **Nucleic acids research**, Oxford University Press, v. 41, n. W1, p. W303–W307, 2013.

ROCHE, D. B.; TETCHNER, S. J.; MCGUFFIN, L. J. FunFOLD: an improved automated method for the prediction of ligand binding residues using 3D models of proteins. **BMC bioinformatics**, BioMed Central, v. 12, n. 1, p. 1–20, 2011.

RODRIGUES, C. H. et al. mCSM-PPI2: predicting the effects of mutations on protein–protein interactions. **Nucleic acids research**, Oxford University Press, v. 47, n. W1, p. W338–W344, 2019.

RODRIGUES, C. H.; PIRES, D. E.; ASCHER, D. B. DynaMut2: Assessing changes in stability and flexibility upon single and multiple point missense mutations. **Protein Science**, Wiley Online Library, v. 30, n. 1, p. 60–69, 2021.

RODRIGUES, J. P. et al. Clustering biomolecular complexes by residue contacts similarity. **Proteins: Structure, Function, and Bioinformatics**, Wiley Online Library, v. 80, n. 7, p. 1810–1817, 2012.

ROE, D. R.; CHEATHAM, T. E. PTRAJ and CPPTRAJ: Software for processing and analysis of molecular dynamics trajectory data. **Journal of Chemical Theory and Computation**, American Chemical Society (ACS), v. 9, n. 7, p. 3084–3095, jun. 2013. Disponível em: <https://doi.org/10.1021/ct400341p>.

ROSE, A. S. et al. NGL viewer: web-based molecular graphics for large complexes. **Bioinformatics**, v. 34, n. 21, p. 3755–3758, 05 2018. ISSN 1367-4803.

ROSE, A. S.; HILDEBRAND, P. W. NGL Viewer: a web application for molecular visualization. **Nucleic Acids Research**, v. 43, n. W1, p. W576–W579, 04 2015. ISSN 0305-1048.

ROSE, P. W. et al. The RCSB protein data bank: integrative view of protein, gene and 3D structural information. **Nucleic Acids Research**, v. 45, n. D1, p. D271–D281, 10 2016. ISSN 0305-1048.

SALENTIN, S. et al. PLIP: fully automated protein–ligand interaction profiler. **Nucleic acids research**, Oxford University Press, v. 43, n. W1, p. W443–W447, 2015.

SALOMON-FERRER, R.; CASE, D. A.; WALKER, R. C. An overview of the Amber biomolecular simulation package. **Wiley Interdisciplinary Reviews: Computational Molecular Science**, Wiley Online Library, v. 3, n. 2, p. 198–210, 2013.

SANNER, M. et al. Geom: a new tool for molecular modelling based on distance geometry calculations with nmr data. **Journal of computer-aided molecular design**, Springer, v. 3, n. 3, p. 195–210, 1989.

SANTANA, C. A. et al. GRaSP: a graph-based residue neighborhood strategy to predict binding sites. **Bioinformatics**, Oxford University Press, v. 36, n. Supplement_2, p. i726–i734, 2020.

SANTANA, C. A. et al. GRaSP: a graph-based residue neighborhood strategy to predict binding sites. **Bioinformatics**, Oxford University Press (OUP), v. 36, n. Supplement_2, p. i726–i734, dez. 2020. Disponível em: <https://doi.org/10.1093/bioinformatics/btaa805>.

SCHAUMANN, T.; BRAUN, W.; WÜTHRICH, K. The program fantom for energy refinement of polypeptides and proteins using a newton–raphson minimizer in torsion angle space. **Biopolymers: Original Research on Biomolecules**, Wiley Online Library, v. 29, n. 4-5, p. 679–694, 1990.

SCHMITT, S.; KUHN, D.; KLEBE, G. A new method to detect related function among proteins independent of sequence and fold homology. **Journal of molecular biology**, Elsevier, v. 323, n. 2, p. 387–406, 2002.

SCHREYER, A.; BLUNDELL, T. CREDO: a protein–ligand interaction database for drug discovery. **Chemical biology & drug design**, Wiley Online Library, v. 73, n. 2, p. 157–167, 2009.

Schrödinger, LLC. The PyMOL molecular graphics system, version 1.8. PyMOL The PyMOL Molecular Graphics System, Version 1.8, Schrödinger, LLC. 2015.

SCHWEDE, T.; PEITSCH, M. C. **Computational structural biology: Methods and applications**. [S.l.]: World scientific, 2008.

SEEBER, M. et al. Wordom: a program for efficient analysis of molecular dynamics simulations. **Bioinformatics**, Oxford University Press, v. 23, n. 19, p. 2625–2627, 2007.

SEHNAL, D. et al. LiteMol suite: interactive web-based visualization of large-scale macromolecular structure data. **Nature Methods**, v. 14, n. 12, p. 1121–1122, Dec 2017. ISSN 1548-7105.

SHANG, J. et al. HybridSim-VS: a web server for large-scale ligand-based virtual screening using hybrid similarity recognition techniques. **Bioinformatics**, Oxford University Press (OUP), v. 33, n. 21, p. 3480–3481, jun. 2017. Disponível em: <https://doi.org/10.1093/bioinformatics/btx418>.

SHIHAB, H. A. et al. Predicting the Functional, Molecular, and Phenotypic Consequences of Amino Acid Substitutions using Hidden Markov Models. **Human Mutation**, v. 34, p. 57 – 65, 2013.

SHULMAN-PELEG, A.; NUSSINOV, R.; WOLFSON, H. J. Recognition of functional sites in protein structures. **Journal of molecular biology**, Elsevier, v. 339, n. 3, p. 607–633, 2004.

SIEVERS, F.; HIGGINS, D. G. Clustal omega. **Current protocols in bioinformatics**, Wiley Online Library, v. 48, n. 1, p. 3–13, 2014.

SIEVERS, F. et al. Fast, scalable generation of high-quality protein multiple sequence alignments using clustal omega. **Molecular systems biology**, John Wiley & Sons, Ltd, v. 7, n. 1, 2011.

ŚLEDŹ, P.; CAFLISCH, A. Protein structure-based drug design: from docking to molecular dynamics. **Current Opinion in Structural Biology**, Elsevier BV, v. 48, p. 93–102, fev. 2018. Disponível em: <https://doi.org/10.1016/j.sbi.2017.10.010>.

SÖDING, J. Protein homology detection by hmm–hmm comparison. **Bioinformatics**, Oxford University Press, v. 21, n. 7, p. 951–960, 2005.

SOFTWARE, A. **Abalone Molecular Simulations**. <http://www.biomolecular-modeling.com/Abalone/>. [Online; accessed 9-December-2021].

SÖLDNER, C. A.; HORN, A. H.; STICHT, H. A metadynamics-based protocol for the determination of gpcr-ligand binding modes. **International journal of molecular sciences**, Multidisciplinary Digital Publishing Institute, v. 20, n. 8, p. 1970, 2019.

SOMAN, K. V. et al. Homology modeling and simulations of nuclease structures. **Nuclease Methods and Protocols**, Springer, p. 263–286, 2001.

SORIN, E. J.; PANDE, V. S. Exploring the helix-coil transition via all-atom equilibrium ensemble simulations. **Biophysical journal**, Biophysical Society, v. 88, n. 4, p. 2472–2493, Apr 2005. ISSN 0006-3495.

SOUSA, S. et al. Protein-Ligand Docking in the New Millennium – A Retrospective of 10 years in the Field. **Current Medicinal Chemistry**, Bentham Science Publishers Ltd., v. 20, n. 18, p. 2296–2314, abr. 2013.

SPINNER, C. D. et al. Effect of remdesivir vs standard care on clinical status at 11 days in patients with moderate COVID-19. **JAMA**, American Medical Association (AMA), v. 324, n. 11, p. 1048, set. 2020. Disponível em: <https://doi.org/10.1001/jama.2020.16349>.

SPOEL, D. V. D. et al. GROMACS: Fast, flexible, and free. **journal of Computational Chemistry**, Wiley, v. 26, n. 16, p. 1701–1718, 2005. Disponível em: <https://doi.org/10.1002/jcc.20291>.

SPOEL, D. V. D. et al. Gromacs: Fast, flexible, and free. **Journal of Computational Chemistry**, v. 26, n. 16, p. 1701–1718, 2005. Disponível em: <https://onlinelibrary.wiley.com/doi/abs/10.1002/jcc.20291>.

STANFIELD, R. L.; WILSON, I. A. Protein-peptide interactions. **Current opinion in structural biology**, Elsevier, v. 5, n. 1, p. 103–113, 1995.

STANK, A. et al. TRAPP webserver: predicting protein binding site flexibility and detecting transient binding pockets. **Nucleic acids research**, Oxford University Press, v. 45, n. W1, p. W325–W330, 2017.

STEWART, J. Scigress, version 2.9. 0. **Fujitsu Limited: United States**, 2009.

STIERAND, K.; MAASS, P. C.; RAREY, M. Molecular complexes at a glance: automated generation of two-dimensional complex diagrams. **Bioinformatics**, Oxford University Press, v. 22, n. 14, p. 1710–1716, 2006.

STOURAC, J. et al. Caver Web 1.0: identification of tunnels and channels in proteins and analysis of ligand transport. **Nucleic acids research**, Oxford University Press, v. 47, n. W1, p. W414–W422, 2019.

STUART, A. C.; ILYIN, V. A.; SALI, A. LigBase: a database of families of aligned ligand binding sites in known protein sequences and structures . **Bioinformatics**, v. 18, n. 1, p. 200–201, 01 2002. ISSN 1367-4803.

THE RDKit: Open-Source Cheminformatics Software. 2021. <https://www.rdkit.org/>. Accessed June 7, 2021.

TORRE, D. et al. Datasets2tools, repository and search engine for bioinformatics datasets, tools and canned analyses. **Scientific data**, Nature Publishing Group, v. 5, n. 1, p. 1–10, 2018.

TROPSHA, A. Best practices for qsar model development, validation, and exploitation. **Molecular informatics**, Wiley Online Library, v. 29, n. 6-7, p. 476–488, 2010.

TROTT, O.; OLSON, A. J. AutoDock Vina: Improving the speed and accuracy of docking with a new scoring function, efficient optimization, and multithreading. **journal of Computational Chemistry**, Wiley, p. NA–NA, 2009.

TROTT, O.; OLSON, A. J. Autodock vina: improving the speed and accuracy of docking with a new scoring function, efficient optimization, and multithreading. **Journal of computational chemistry**, v. 31, n. 2, p. 455–461, Jan 2010. ISSN 1096-987X. Disponível em: <https://pubmed.ncbi.nlm.nih.gov/19499576>.

UFV, Conselho Técnico de Pós Graduação da Universidade Federal de Viçosa. **Normas de Redação de Teses e Dissertações**. 2018. <http://www.dpi.ufv.br/arquivos/ppgcc/doc/PPG-2015-normascorrigidas.pdf>. Accessed: 2018-06-14.

ULLAH, A. Z. D. et al. SNPnexus: assessing the functional relevance of genetic variation to facilitate the promise of precision medicine. **Nucleic Acids Research**, v. 46, n. W1, p. W109–W113, 05 2018. ISSN 0305-1048.

U.S. Food and Drug Administration website. 2021. <https://www.fda.gov/drugs>. Revised June 2021. Accessed June 22, 2021.

van Zundert, G. et al. The HADDOCK2.2 Web Server: User-Friendly Integrative Modeling of Biomolecular Complexes. **journal of Molecular Biology**, v. 428, n. 4, p. 720–725, 2016. ISSN 0022-2836. Computation Resources for Molecular Biology.

VANHEE, P. et al. Pepx: a structural database of non-redundant protein–peptide complexes. **Nucleic acids research**, Oxford University Press, v. 38, n. suppl_1, p. D545–D551, 2010.

VEKLURY. **(remdesivir). U.S. Food and Drug Administration website**. 2021. <https://www.accessdata.fda.gov/drugsatfda_docs/label/2020/214787Orig1s000lbl.pdf>. Revised October 2020. Accessed June 7, 2021.

VIANNA, U. et al. Espécies e/ou linhagens de trichogramma spp. (hymenoptera: Trochogrammatidae) para o controle de anticarsia gemmatalis (lepidoptera: Noctuidae). **Arquivos do Instituto Biológico**, v. 71, p. 81–87, 03 2011.

VINOGRADOV, A. A.; YIN, Y.; SUGA, H. Macrocyclic peptides as drug candidates: recent progress and remaining challenges. **Journal of the American Chemical Society**, ACS Publications, v. 141, n. 10, p. 4167–4181, 2019.

VISUALIZER, D. S. Dassault Systèmes BIOVIA. **San Diego, nd**, 2020.

VLACHAKIS, D. et al. Current State-of-the-Art Molecular Dynamics Methods and Applications. In: **Advances in Protein Chemistry and Structural Biology**. [S.l.]: Elsevier, 2014. p. 269–313.

VRIES, S. J. de et al. The pepATTRACT web server for blind, large-scale peptide–protein docking. **Nucleic Acids Research**, Oxford University Press (OUP), v. 45, n. W1, p. W361–W364, abr. 2017.

WAGIH, O. et al. A resource of variant effect predictions of single nucleotide variants in model organisms. **Molecular Systems Biology**, EMBOpress, v. 14, 2018.

WAGNER, J. R. et al. Povme 3.0: software for mapping binding pocket flexibility. **Journal of chemical theory and computation**, ACS Publications, v. 13, n. 9, p. 4584–4592, 2017.

WALL, M. E.; RECHTSTEINER, A.; LM., R. Singular value decomposition and principal component analysis. In: DUBITZKY, W.; GRANZOW, M. (Ed.). **Berrar DP**. A Practical Approach to Microarray Data Analysis. Norwell, MA: Kluwer, 2003.

WALLACE, A. C.; LASKOWSKI, R. A.; THORNTON, J. M. Ligplot: a program to generate schematic diagrams of protein-ligand interactions. **Protein engineering, design and selection**, Oxford University Press, v. 8, n. 2, p. 127–134, 1995.

WALLACH, I.; DZAMBA, M.; HEIFETS, A. Atomnet: A deep convolutional neural network for bioactivity prediction in structure-based drug discovery. **arXiv preprint arXiv:1510.02855**, 10 2015.

WANG, G.; LI, X.; WANG, Z. Apd3: the antimicrobial peptide database as a tool for research and education. **Nucleic acids research**, Oxford University Press, v. 44, n. D1, p. D1087–D1093, 2016.

WANG, J. et al. Automatic atom type and bond type perception in molecular mechanical calculations. **Journal of Molecular Graphics and Modelling**, v. 25, n. 2, p. 247 – 260, 2006. ISSN 1093-3263. Disponível em: <http://www.sciencedirect.com/science/article/pii/S1093326305001737>.

WANG, J. et al. Strapep: a structure database of bioactive peptides. **Database**, Narnia, v. 2018, 2018.

WANG, L. et al. Accurate and Reliable Prediction of Relative Ligand Binding Potency in Prospective Drug Discovery by Way of a Modern Free-Energy Calculation Protocol and Force Field. **journal of the American Chemical Society**, American Chemical Society (ACS), v. 137, n. 7, p. 2695–2703, fev. 2015.

WANG, Z. **Tinker9: Next Generation of Tinker with GPU Support**. 2021. Washington University in St. Louis, accessed: Dec. 6, 2021. Disponível em: <https://github.com/TinkerTools/tinker9>.

WANG, Z. et al. Comprehensive evaluation of ten docking programs on a diverse set of protein–ligand complexes: the prediction accuracy of sampling power and scoring power. **Phys. Chem. Chem. Phys.**, The Royal Society of Chemistry, v. 18, p. 12964–12975, 2016.

WATERHOUSE, A. et al. SWISS-MODEL: homology modelling of protein structures and complexes. **Nucleic acids research**, Oxford University Press, v. 46, n. W1, p. W296–W303, 2018.

WEBB, B.; SALI, A. Comparative protein structure modeling using MODELLER. **Current protocols in bioinformatics**, Wiley Online Library, v. 54, n. 1, p. 5–6, 2016.

WEBB, B.; SALI, A. Protein structure modeling with MODELLER. In: **Functional genomics**. [S.l.]: Springer, 2017. p. 39–54.

WELCH, L. et al. Applying, evaluating and refining bioinformatics core competencies (an update from the curriculum task force of ISCB's Education Committee). **PLoS computational biology**, Public Library of Science San Francisco, CA USA, v. 12, n. 5, p. e1004943, 2016.

WELCH, L. et al. Bioinformatics curriculum guidelines: toward a definition of core competencies. **PLOS computational biology**, Public Library of Science San Francisco, USA, v. 10, n. 3, p. e1003496, 2014.

WEN, Z. et al. Pepbdb: a comprehensive structural database of biological peptide–protein interactions. **Bioinformatics**, Oxford University Press, v. 35, n. 1, p. 175–177, 2019.

WENG, G. et al. Comprehensive Evaluation of Fourteen Docking Programs on Protein–Peptide Complexes. **journal of Chemical Theory and Computation**, American Chemical Society (ACS), v. 16, n. 6, p. 3959–3969, abr. 2020.

WIEL, L. et al. MetaDome: Pathogenicity analysis of genetic variants through aggregation of homologous human protein domains. **Human Mutation**, v. 40, p. W1030–1038, 2019.

WISHART, D. S. et al. Drugbank 5.0: a major update to the drugbank database for 2018. **Nucleic acids research**, Oxford University Press, v. 46, n. D1, p. D1074–D1082, 2018.

WU, C. et al. Analysis of therapeutic targets for sars-cov-2 and discovery of potential drugs by computational methods. **Acta Pharmaceutica Sinica B**, v. 10, n. 5, p. 766 – 788, 2020. ISSN 2211-3835. Disponível em: <http://www.sciencedirect.com/science/article/pii/S2211383520302999>.

WU, S.; ZHANG, Y. Lomets: a local meta-threading-server for protein structure prediction. **Nucleic acids research**, Oxford University Press, v. 35, n. 10, p. 3375–3382, 2007.

XIE, Z.-R. et al. LISE: a server using ligand-interacting and site-enriched protein triangles for prediction of ligand-binding sites. **Nucleic acids research**, Oxford University Press, v. 41, n. W1, p. W292–W296, 2013.

XU, D.; ZHANG, Y. Ab initio protein structure assembly using continuous structure fragments and optimized knowledge-based force field. **Proteins: Structure, Function, and Bioinformatics**, Wiley Online Library, v. 80, n. 7, p. 1715–1735, 2012.

XU, X.; ZOU, X. Peppro: A nonredundant structure data set for benchmarking peptide–protein computational docking. **Journal of Computational Chemistry**, Wiley Online Library, 2020.

XUAN, X. et al. iCataly-PseAAC: Identification of Enzymes Catalytic Sites Using Sequence Evolution Information with Grey Model GM (2,1). **The journal of Membrane Biology**, v. 248, 06 2015.

YAN, Y. et al. The HDOCK server for integrated protein–protein docking. **Nature Protocols**, Springer Science and Business Media LLC, v. 15, n. 5, p. 1829–1852, abr. 2020.

YANG, J.; ROY, A.; ZHANG, Y. BioLiP: a semi-manually curated database for biologically relevant ligand–protein interactions. **Nucleic Acids Research**, v. 41, n. D1, p. D1096–D1103, 10 2012. ISSN 0305-1048.

YANG, J.; WANG, Y.; CHEN, Y. GPU accelerated molecular dynamics simulation of thermal conductivities. **journal of Computational Physics**, v. 221, n. 2, p. 799–804, 2007. ISSN 0021-9991.

YANG, J. et al. The i-tasser suite: protein structure and function prediction. **Nature methods**, Nature Publishing Group, v. 12, n. 1, p. 7, 2015.

YASUO, N.; SEKIJIMA, M. Improved method of structure-based virtual screening via interaction-energy-based learning. **Journal of Chemical Information and Modeling**, American Chemical Society, v. 59, n. 3, p. 1050–1061, Mar 2019. ISSN 1549-9596. Disponível em: <https://doi.org/10.1021/acs.jcim.8b00673>.

YUAN, S.; CHAN, H. S.; HU, Z. Using PyMOL as a platform for computational drug design. **WIREs Computational Molecular Science**, v. 7, n. 2, p. e1298, 2017.

ZAKI, M.; MEIRA JR, W. **Data Mining and Analysis: Fundamental Concepts and Algorithms**. Cambridge: Cambridge University Press, 2014.

ZHANG, C. et al. Template-based and free modeling of I-TASSER and QUARK pipelines using predicted contact maps in CASP12. **Proteins: Structure, Function, and Bioinformatics**, Wiley Online Library, v. 86, p. 136–151, 2018.

ZHANG, L. et al. Crystal structure of sars-cov-2 main protease provides a basis for design of improved α-ketoamide inhibitors. **Science**, American Association for the Advancement of Science, v. 368, n. 6489, p. 409–412, 2020. ISSN 0036-8075. Disponível em: <https://science.sciencemag.org/content/368/6489/409>.

ZHANG, Q. C. et al. Structure-based prediction of protein–protein interactions on a genome-wide scale. **Nature**, Nature Publishing Group, v. 490, n. 7421, p. 556–560, 2012.

ZHANG, T. et al. Accurate sequence-based prediction of catalytic residues. **Bioinformatics**, Oxford University Press, v. 24, n. 20, p. 2329–2338, 2008.

ZHANG, Y. I-TASSER server for protein 3D structure prediction. **BMC bioinformatics**, Springer, v. 9, n. 1, p. 1–8, 2008.

ZHENG, M. et al. LBVS: an online platform for ligand-based virtual screening using publicly accessible databases. **Molecular Diversity**, Springer Science and Business Media LLC, v. 18, n. 4, p. 829–840, set. 2014. Disponível em: <https://doi.org/10.1007/s11030-014-9545-3>.

ZHENG, W. et al. Lomets2: improved meta-threading server for fold-recognition and structure-based function annotation for distant-homology proteins. **Nucleic acids research**, Oxford University Press, v. 47, n. W1, p. W429–W436, 2019.

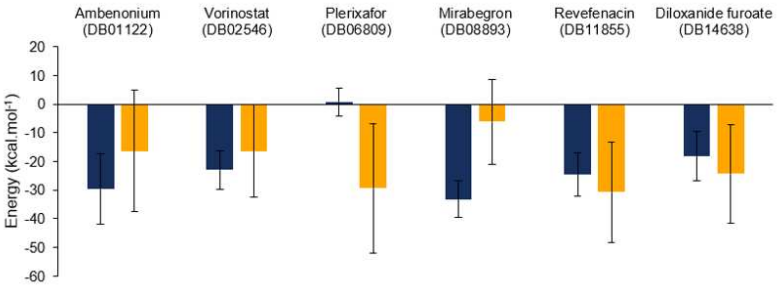
ZOETE, V. et al. Swissparam: A fast force field generation tool for small organic molecules. **Journal of Computational Chemistry**, v. 32, n. 11, p. 2359–2368, 2011. Disponível em: <https://onlinelibrary.wiley.com/doi/abs/10.1002/jcc.21816>.

ZUNDERT, G. V. et al. The haddock2. 2 web server: user-friendly integrative modeling of biomolecular complexes. **Journal of molecular biology**, Elsevier, v. 428, n. 4, p. 720–725, 2016.

ZVELEBIL, M. et al. **Understanding Bioinformatics**. [S.l.]: Garland Science, 2008. ISBN 9780815340249.

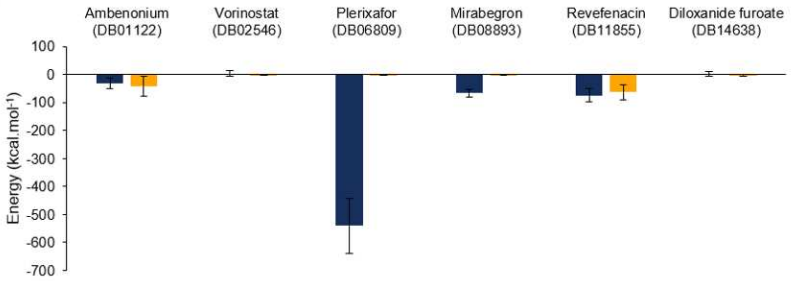
Appendix A

Supplementary information: Chapter 2



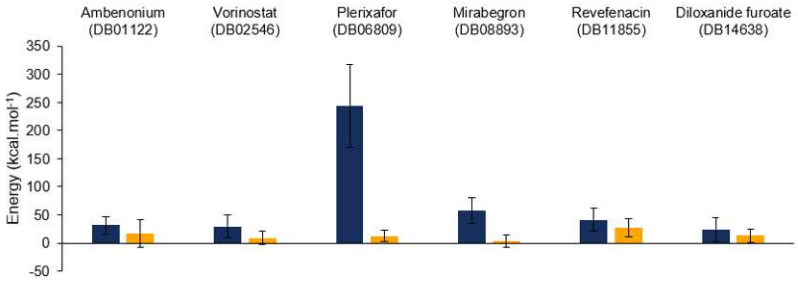
	CHARMM		AMBER	
	energy	std error	energy	std error
DB01122	-30	0.55	-16	0.95
DB02546	-23	0.30	-16	0.73
DB06809	0.77	0.21	-29	1.0
DB08893	-33	0.29	-6.1	0.66
DB11855	-24	0.34	-31	0.79
DB14638	-18	0.39	-24	0.77

Figure 14 – van der Waals energy on MM-PBSA calculation for the complexes. CHARMM is represented in blue and AMBER, in yellow. The error bars represent the standard error.



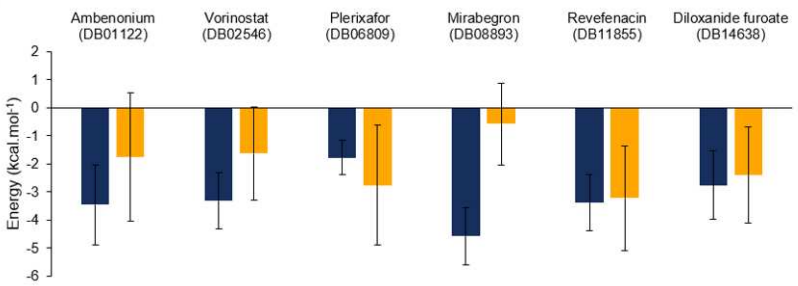
	CHARMM		AMBER	
	energy	std error	energy	std error
DB01122	-31	0.87	-41	1.6
DB02546	3.6	0.42	-1.4	0.080
DB06809	-541	4.4	-1.2	0.058
DB08893	-67	0.65	-0.52	0.064
DB11855	-74	1.0	-63	1.2
DB14638	2.2	0.34	-2.3	0.13

Figure 15 – Electrostatic energy on MM-PBSA calculation for the complexes. CHARMM is represented in blue and AMBER, in yellow. The error bars represent the standard error.



	CHARMM		AMBER	
	energy	std error	energy	std error
DB01122	32	0.65	17	1.1
DB02546	29	0.93	8.6	0.53
DB06809	244	3.3	13	0.47
DB08893	57	1.0	3.2	0.52
DB11855	41	0.89	27	0.71
DB14638	24	0.97	13	0.54

Figure 16 – Polar solvation energy on MM-PBSA calculation for the complexes. CHARMM is represented in blue and AMBER, in yellow. The error bars represent the standard error.



	CHARMM		AMBER	
	energy	std error	energy	std error
DB01122	-3.5	0.064	-1.8	0.10
DB02546	-3.3	0.044	-1.6	0.074
DB06809	-1.8	0.027	-2.8	0.095
DB08893	-4.6	0.045	-0.58	0.065
DB11855	-3.4	0.045	-3.2	0.083
DB14638	-2.8	0.055	-2.4	0.077

Figure 17 – SASA energy on MM-PBSA calculation for the complexes. CHARMM is represented in blue and AMBER, in yellow. The error bars represent the standard error.

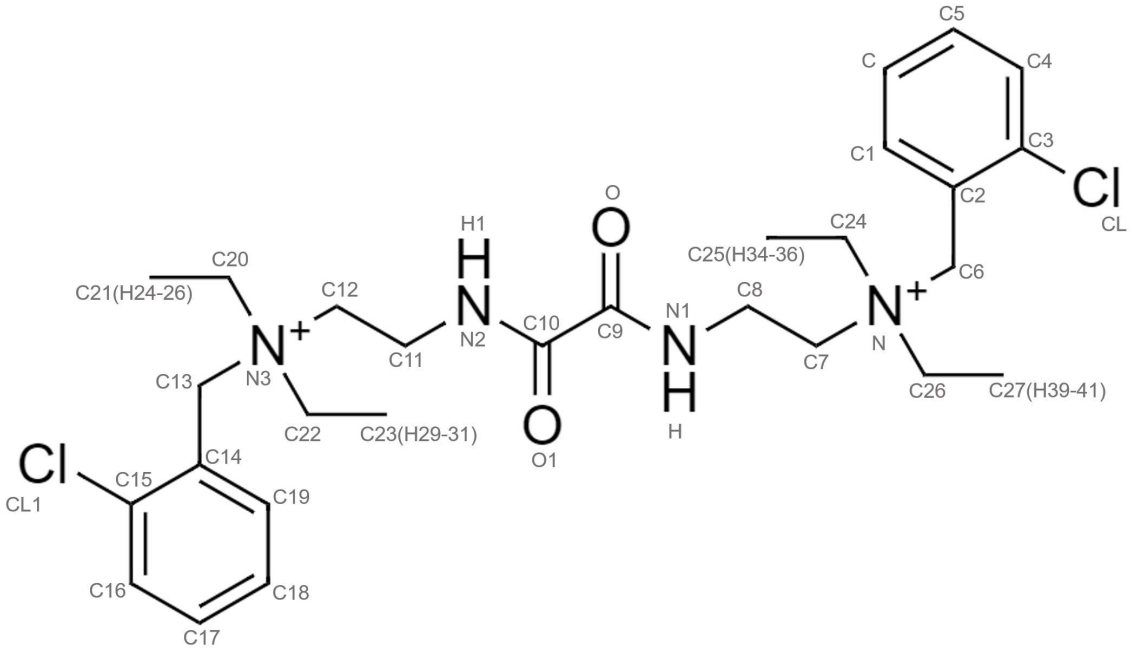


Figure 18 – Labels for ambenonium atoms.

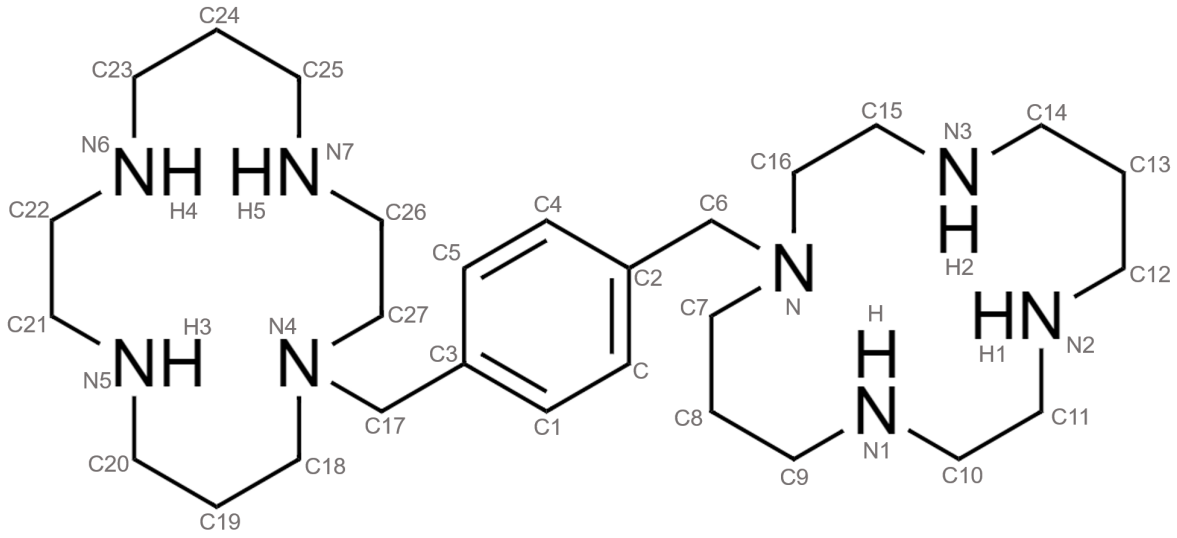


Figure 19 – Labels for plerixafor atoms.

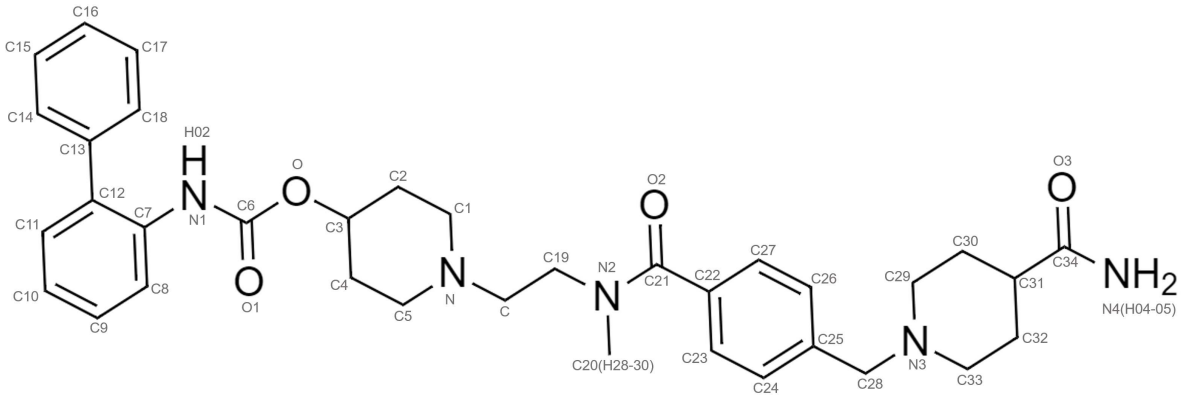


Figure 20 – Labels for revefenacin atoms.

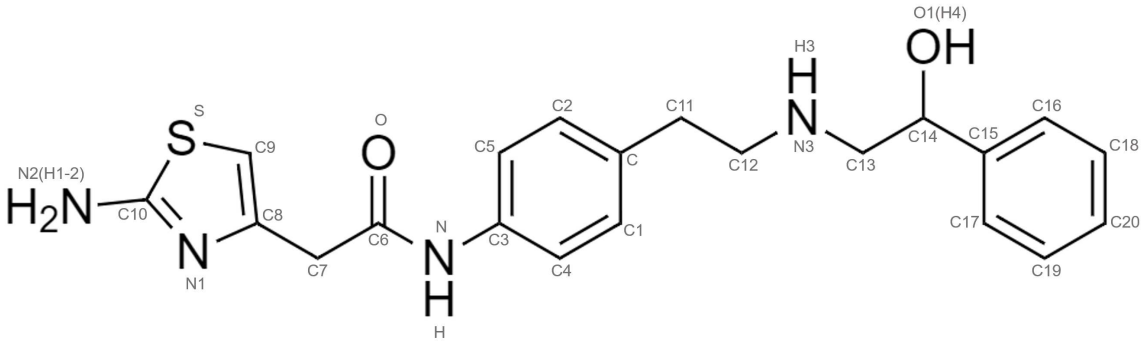


Figure 21 – Labels for mirabegron atoms.

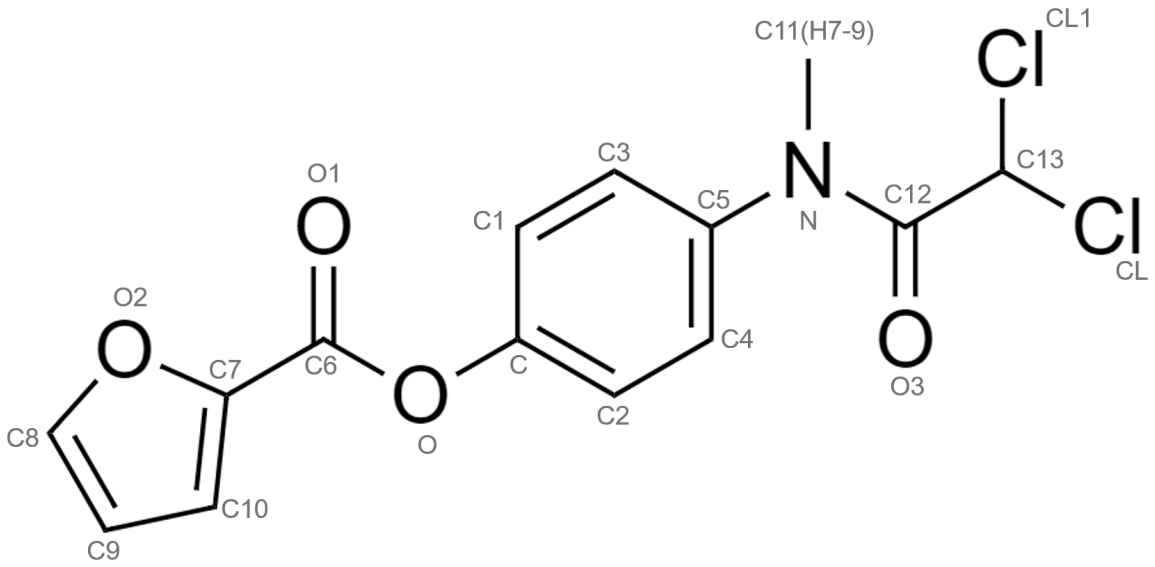


Figure 22 – Labels for diloxanide furoate atoms.

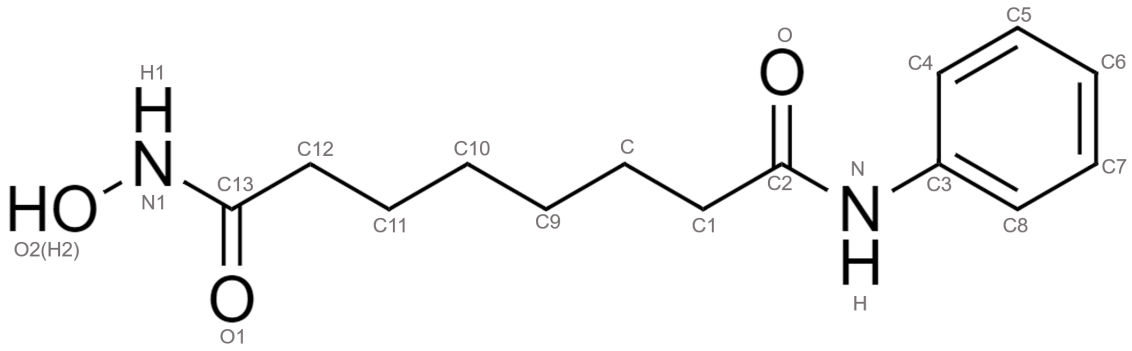


Figure 23 – Labels for vorinostat atoms.

Figure 24 – Energy profile for M^{pro}-ambenonium. Energy profile in all replicas (represented by the R1, R2 or R3) with AMBER (represented by A) and CHARMM (represented by C) and the two CV sets (represented by 1 or 2 preceding the force field label). The highlighted points represents, respectively the minimum energy inside the active site and in the water. The vertical axis represents the energy in kcal.mol^{-1} and the horizontal axis represents the distance in Å.

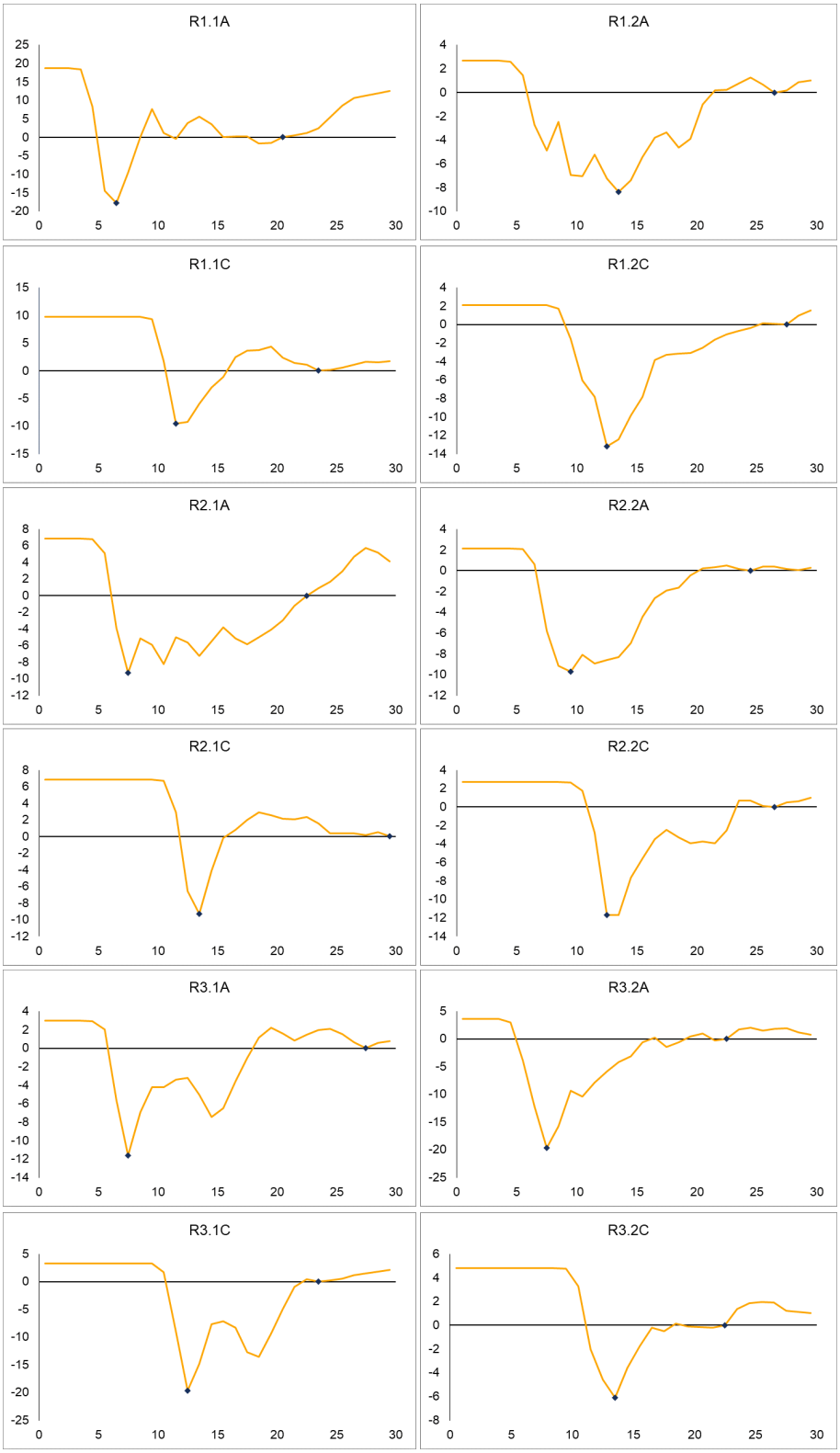


Figure 25 – Energy profile for M^{pro}-plerixafor. Energy profile in all replicas (represented by the R1, R2 or R3) with AMBER (represented by A) and CHARMM (represented by C) and the two CV sets (represented by 1 or 2 preceding the force field label). The highlighted points represents, respectively the minimum energy inside the active site and in the water. The vertical axis represents the energy in kcal.mol^{-1} and the horizontal axis represents the distance in Å.

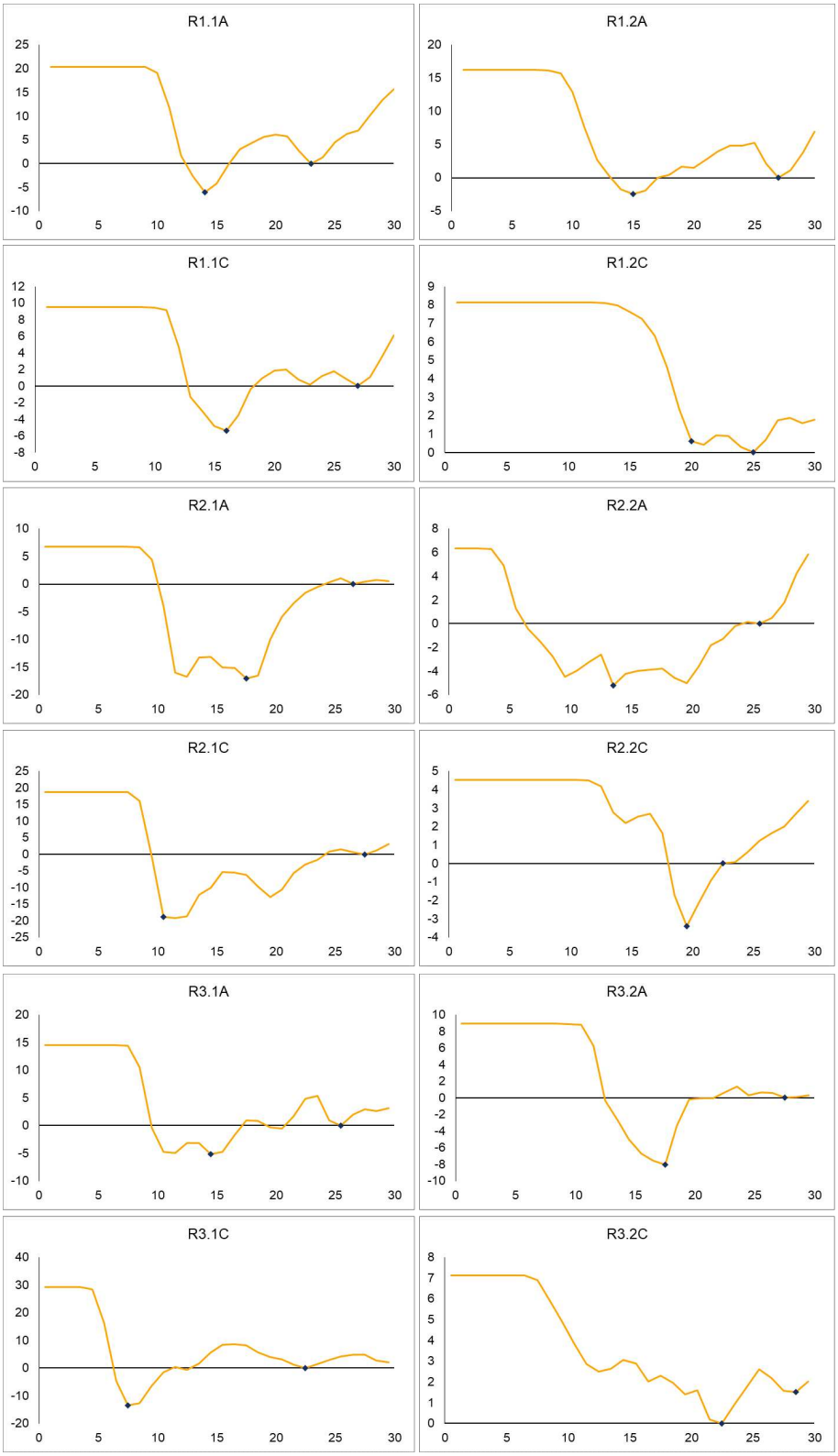


Figure 26 – Energy profile for M^{pro}-revefenacin. Energy profile in all replicas (represented by the R1, R2 or R3) with AMBER (represented by A) and CHARMM (represented by C) and the two CV sets (represented by 1 or 2 preceding the force field label). The highlighted points represents, respectively the minimum energy inside the active site and in the water. The vertical axis represents the energy in kcal.mol^{-1} and the horizontal axis represents the distance in Å.

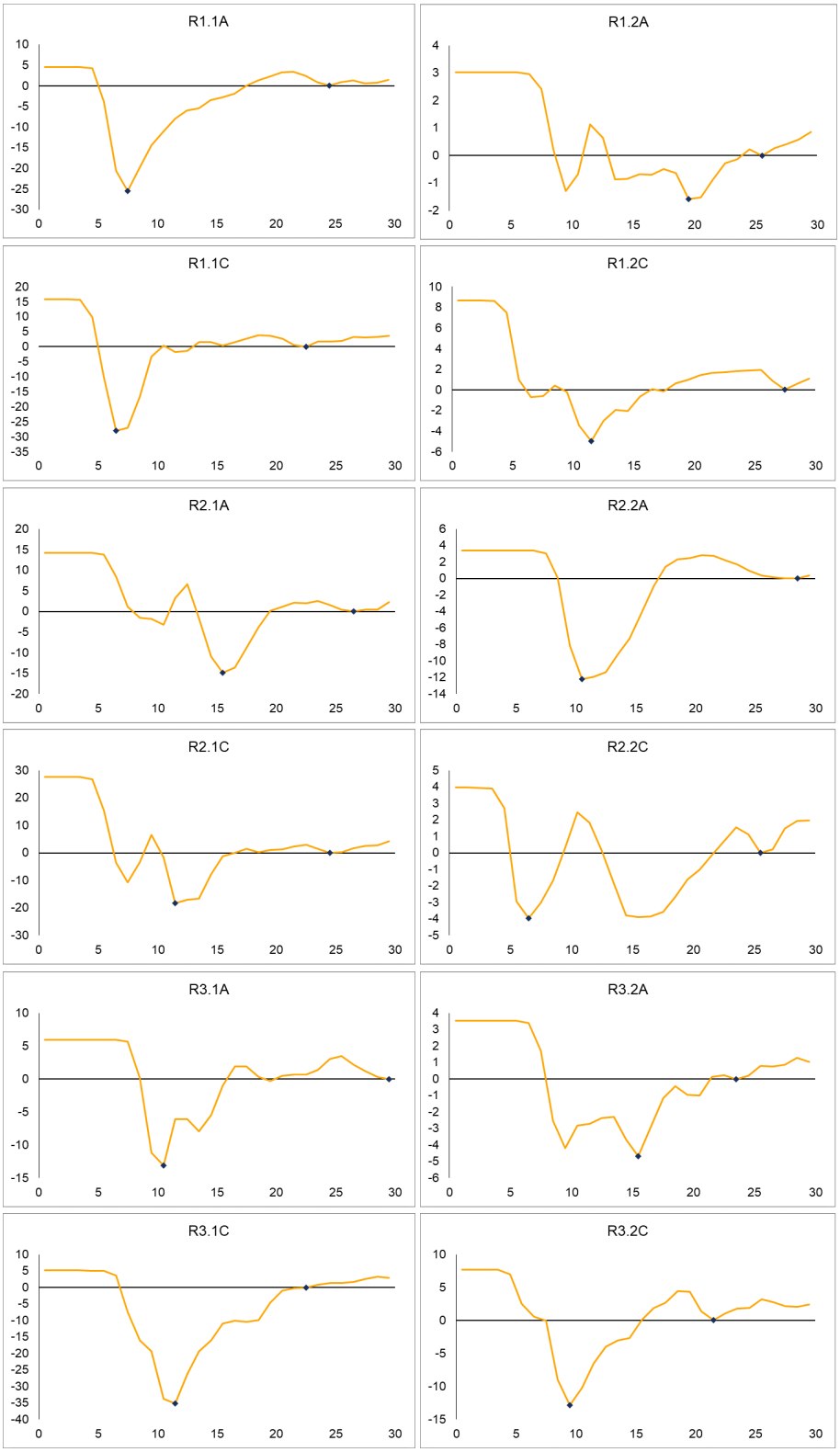


Figure 27 – Energy profile for M^{pro}-mirabegron. Energy profile in all replicas (represented by the R1, R2 or R3) with AMBER (represented by A) and CHARMM (represented by C) and the two CV sets (represented by 1 or 2 preceding the force field label). The highlighted points represents, respectively the minimum energy inside the active site and in the water. The vertical axis represents the energy in kcal.mol^{-1} and the horizontal axis represents the distance in Å.

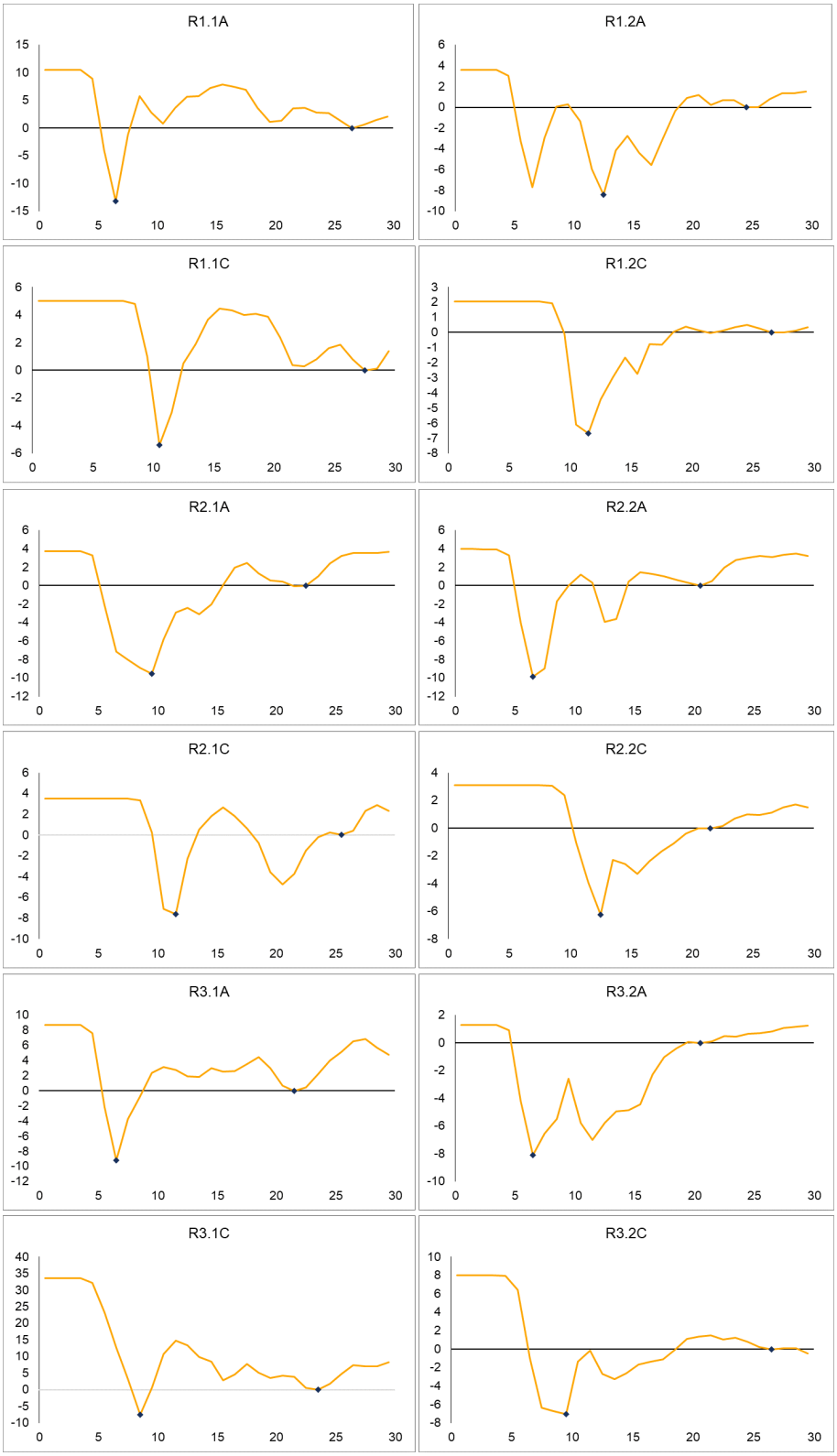


Figure 28 – Energy profile for M^{pro}-diloxanide furoate. Energy profile in all replicas (represented by the R1, R2 or R3) with AMBER (represented by A) and CHARMM (represented by C) and the two CV sets (represented by 1 or 2 preceding the force field label). The highlighted points represents, respectively the minimum energy inside the active site and in the water. The vertical axis represents the energy in kcal.mol^{-1} and the horizontal axis represents the distance in Å.

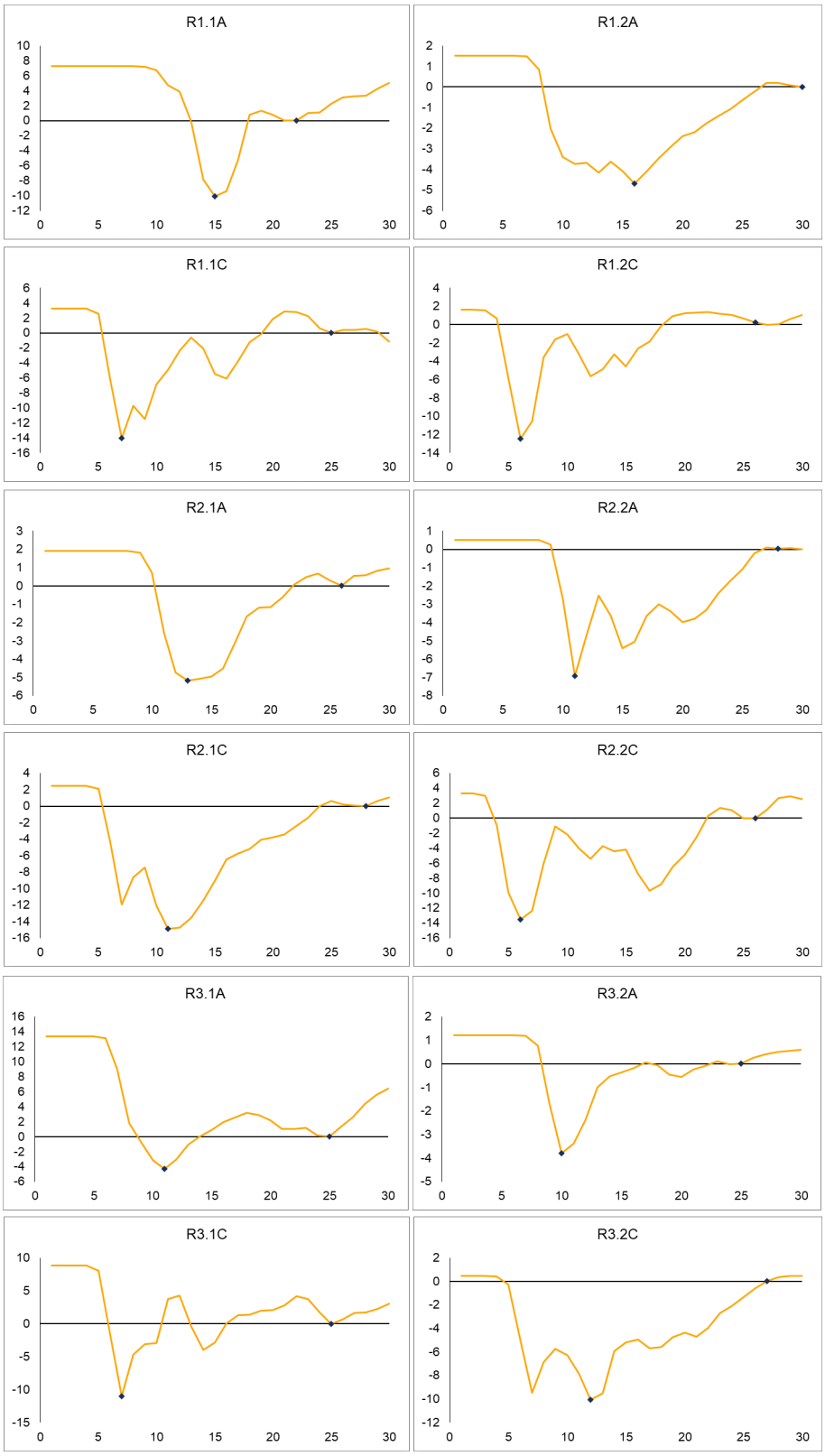


Figure 29 – Energy profile for M^{pro}-vorinostat. Energy profile in all replicas (represented by the R1, R2 or R3) with AMBER (represented by A) and CHARMM (represented by C) and the two CV sets (represented by 1 or 2 preceding the force field label). The highlighted points represents, respectively the minimum energy inside the active site and in the water. The vertical axis represents the energy in kcal.mol^{-1} and the horizontal axis represents the distance in Å.

Appendix B

Supplementary information: Chapter 3

In this document we present additional details and figures about Propedia case studies. The document is organized in sections that are correspondent to details of metadynamics and simulations performed.

B.1 Peptide Selection for Metadynamics Validation and System Setup

Four representative peptide-protein complexes were selected from the Propedia output for the crystallographic structure of the SARS-COV-2 main protease (the M^{pro} with PDB:ID 6lu7) to posterior binding free energy (ΔG_{bind}) calculation by metadynamics simulation (BARDUCCI; BONOMI; PARRINELLO, 2011; BUSSI; LAIO; PARRINELLO, 2006; BRANDT et al., 2016a), aiming the validation of the Propedia scores. The peptides were selected looking for a maximal coverage of the Propedia ranking based both on the alignment score as at the RMSD considering the active site (Table 23).

For the four selected complexes, the Rosetta pose presenting the best score and preserving the previously attributed protonation states was carried to a procedure of topology assembly, solvation on a water box with a 12 Å padding, and Na^+/Cl^- addition until the system neutralizing and ionic strength of 0.150 M. For these procedures, the respective psfgen, solvate and ionize tools from the VMD/NAMD packages were used (HUMPHREY et al., 1996; PHILLIPS et al., 2005b). The CHARMM36 force field (HUANG; JR, 2013; BEST et al., 2012) was used both for the protein, peptide, water and ions, as well the TIP3P model for the water molecules (PRICE; III, 2004).

B.2 Simulation Procedures

All the simulations were carried at the NAMD 2.13 package (PHILLIPS et al., 2005b), at the NPT ensemble, using Langevin thermostat and barostat devices, respectively set to 300 K and 1 atm. Periodic boundary conditions were used, as well particle mesh Ewald (PME)

for the calculations of the long range electrostatic forces, with a 12 Å cutoff for the nonbond interactions and a 2 fs time-step. The hydrogen atoms dynamics was constrained and estimated by the SETTLE algorithm according implemented in NAMD 2.13 (PHILLIPS et al., 2005b).

Before the metadynamics procedures themselves, a meticulous minimization/relaxation/equilibration protocol was carried for each system. First, each system was minimized for 10,000 steps by conjugate gradient algorithm according implemented in NAMD 2.13 (PHILLIPS et al., 2005b). In sequence, a 10 steps relaxation/equilibration molecular dynamics (MD) protocol, at the previously listed simulation conditions and with gradual adaptation of harmonic restraints, was carried as below:

- 500 ps MD with harmonic restrains for all the atoms of the receptor and the ligand.
- 500 ps MD with harmonic restrains just for the backbone atoms of the receptor and the ligand.
- 500 ps MD with harmonic restraints just for the backbone atoms of the receptor.
- 500 ps MD without harmonic restraints.
- 8 ns MD without harmonic restraints and with previous reboot of the velocities according a 300 K and 1 atm NPT ensemble.
- 500 ps MD reintroducing the harmonic restraints at the backbone atoms of the receptor and the ligand.
- 500 ps MD reintroducing the harmonic restraints at all atoms of the receptor and the ligand.
- 300 ps MD with removal of the harmonic restraints from the ligand side chains.
- 300 ps MD with removal of the harmonic restraints from the ligand all atoms.
- 1 ns MD at the aforementioned conditions and rebooting the velocities to a 300 K and 1 atm compatible NPT ensemble.

The last 5 steps were carried in order to prepare the system for the restraints conditions used along the metadynamics procedure (i.e., harmonic restraints for the entire receptor and complete freedom just for the ligand (see below)).

After relaxation, the last frame for each system was taken to three respective and independent 10 ns metadynamics procedures (i.e., three metadynamics simulations per system) of ligand unbind from the active site according a similar protocol to the described in (BRANDT et al., 2016a). Basically, all the atoms of the M^{pro} were maintained harmonically restrained, while complete dynamics freedom was given to the peptide, water molecules and ions. Two collective

variables (CVs) were used to describe the unbinding process of the peptide from the catalytic pocket. The first, CV_{dist} was defined as the distance in Å between the respective centers of mass of the catalytic C145 in M^{pro} and of the closer residue (at the starting pose) of the peptide. The respective residues for each one of the four analyzed peptides were: S7 for the PDB:2q6g peptide; Q5 for the PDB:1uk4 peptide; Q307 both for PDB:1lvm as PDB:1lvb peptides (see Results and Discussions). The second CV, CV_{ang}, was defined as the angle in degrees (º) determined by the center of mass of the M^{pro} C145, the residue afore mentioned for each peptide and the center of mass of the peptide as a whole. The height of the Gaussians for the metadynamics was set to 0.02 Kcal/mol and added every 2ps with a width of 1.77. The CV_{dist}. ranged between 0 and 30 Å with an amplitude fluctuation of 2 Å, while the CV_{ang} has varied between 0º and 180º with an amplitude fluctuation of 10º. The potential of mean force (PMF) landscapes were saved every 1 ps. Results were analyzed with VMD software (HUMPHREY et al., 1996), in-house R and Python scripts, as well the Wordom package (SEEBER et al., 2007).

B.3 Free Energy Maps, Projections of the Metadynamics Energies Along the CV_{dist} dimension and Estimation of ΔG_{bind}

To select the number of frames to be considered along the PMF reconstruction, we have used both the molecular mechanics nonbond interaction energies between the peptide and the protein, as the distance associated to CV_{dist}. (i.e., the distance between the respective C145 and the closer residue mass centers) as a metric to observe: 1) the access and the filling of the energy minimum A (i.e., the energy minimum for the peptide inside the protein); 2) the access and the filling of the energy minimum B (i.e., the energy minimum for the peptide outside the protein, at the aqueous environment); 3) the re-crossing event (i.e., the simulation phase in which, once the system has reached and completely filled both minima, the peptide gain higher dynamics freedom and turn to visit both minima repeatedly). Following the suggested in literature, we have selected the PMF maps saved until the simulation step immediately before the re-crossing event to reconstruct the free energy landscape (FEL) along the unbinding event (BRANDT et al., 2016a; RAITERI et al., 2006). This is made in order to avoid the over-filling of the energy minima by the metadynamics Gaussian potentials and a loss of accuracy along such FEL reconstruction.

The CV most directly related to the unbinding process is, naturally, the one that describes the distance variation between the peptide and the active site (CV_{dist}). In this way, the projections of the metadynamics free energy onto the CV_{dist}. was calculated similarly to (BRANDT et al., 2016a) as following equation:

$$-\beta G_{CV_{ang}}(CV_{dist}) = ln\frac{\int e^{-\beta G(CV_{dist},\, CV_{ang})} dCV_{ang}}{\int e^{-\beta G(CV_{dist},\, CV_{ang})} dCV_{ang} dCV_{dist}} \tag{B.1}$$

where β = 1/kbT, being k_b the Boltzmann constant (1,9858 x 10^{-3}·kcal·mol^{-1}·K^{-1}); T = 300 K and G(CV_{dist}, CV_{ang}) accounts for the free energy value at the position (CV_{dist}, CV_{ang}) position on the PMF map.

For the estimation of ΔG_{bind} from each metadynamics replica, the minima value of $G_{CV_{ang}}(CV_{dist})$ at a CV_{dist} compatible with the complete independence of the peptide environment from the protein influence (i.e., $CV_{dist} \geq 25$ Å) was diminished from the minima value of this measure inside a distance compatible with the peptide bound (specifically or not) to the protein active site (i.e., $CV_{dist} \leq 20$ Å) according equation:

$$\Delta G_{bind} = G'_{CV_{ang}}(CV_{dist}) - G''_{CV_{ang}}(CV_{dist}) \tag{B.2}$$

where $G'_{CV_{ang}}(CV_{dist})$ is the minimum value inside, while $G''_{CV_{ang}}(CV_{dist})$ is the minimum outside the protein. For the cases in which two or more minima with similar favorability were found inside the protein, both minima were equally weighted at a global value of $G'_{CV_{ang}}(CV_{dist})$ (i.e.,$G_{CV_{ang}}(CV_{dist})^{Min}{}_{inside}$) according equation:

$$-\beta G_{CV_{Ang.}}(CV_{Dist.})^{Min.}_{inside} = ln\frac{\sum_{i=1}^{n} e^{-\beta G'_{CV_{Ang.}}(CV_{Dist.})}}{\int e^{-\beta G_{Ang.}(CV_{dist})} dCV_{Dist.}} \tag{B.3}$$

where i = 1, 2,..n is the number of equivalent minima at diferent CV_{dist}. values inside the protein. Finally, the accuracy of Propedia was probed by the metadynamics analysis by measuring the correlation of the ΔG_{bind} values estimated according equations (B.1, B.2, B.3) and the respective values of the alignment score and the site RMSD recovered by our tool for each one of the four M^{pro}:peptide complexes choosen for validation.

B.4 Tables

Table 23 – Correlation between the metadynamics estimated binding free energy (MetaD ΔG_{bind} and it is standard deviations(σ)) and the Propedia recovered alignment score (Align. Score) and site RMSD. At the last two columns the respective negative and positive correlation coefficients of the ΔG_{bind} with each Propedia parameter are depicted.

	Propedia		**MetaD**		**R^2 ΔG_{bind}**	
PDB id	**Align. score**	**site RMSD (Å)**	**ΔG_{bind} (kcal/mol)**	**σ**	**Align. score**	**site RMSD**
2q6g	10.36	0.34	-15.92	±1.34	(-) 0.98	(+) 0.96
1uk4	9.47	0.44	-15.97	±4.32		
1lvm	5.69	0.84	-4.76	±1.27		
1lvb	4.54	1.21	-0.61	±1.57		

B.5 Figures

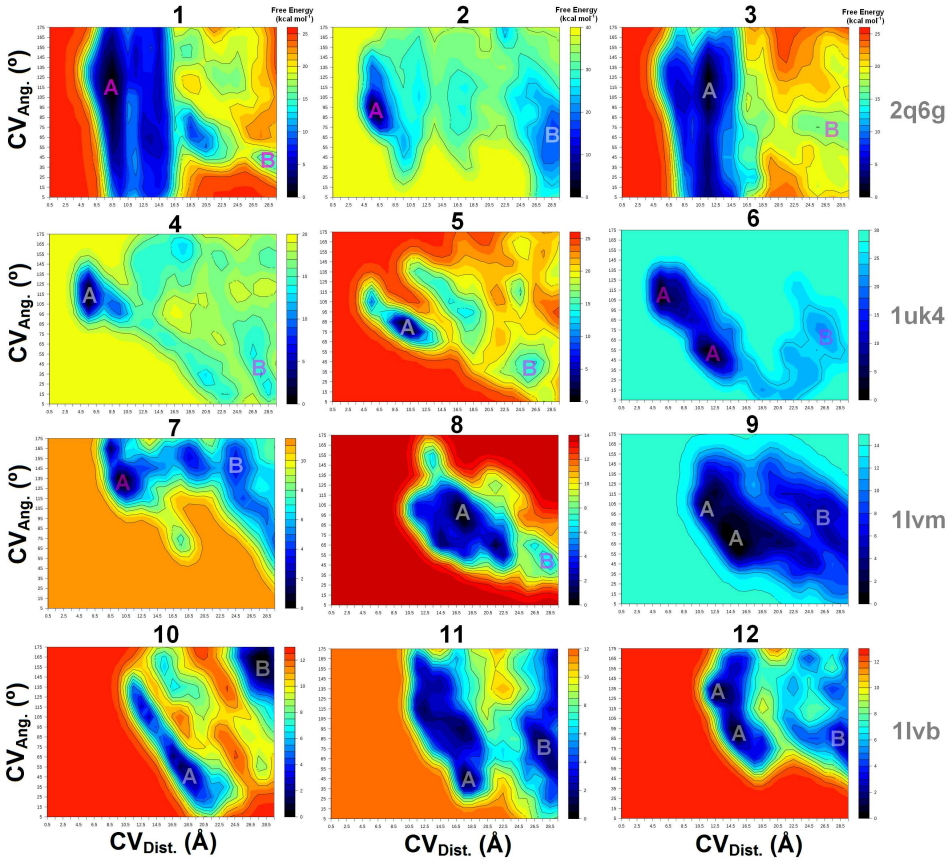


Figure 30 – The free energy landscape for the respective triplicates of the unbinding metadynamics for the SARS-COV-2 M^{pro}: peptide complexes with the PDB id: 2q6g(chain C), 1uk4 (chain H), 1lvm (chain D), and 1lvb (chain D)

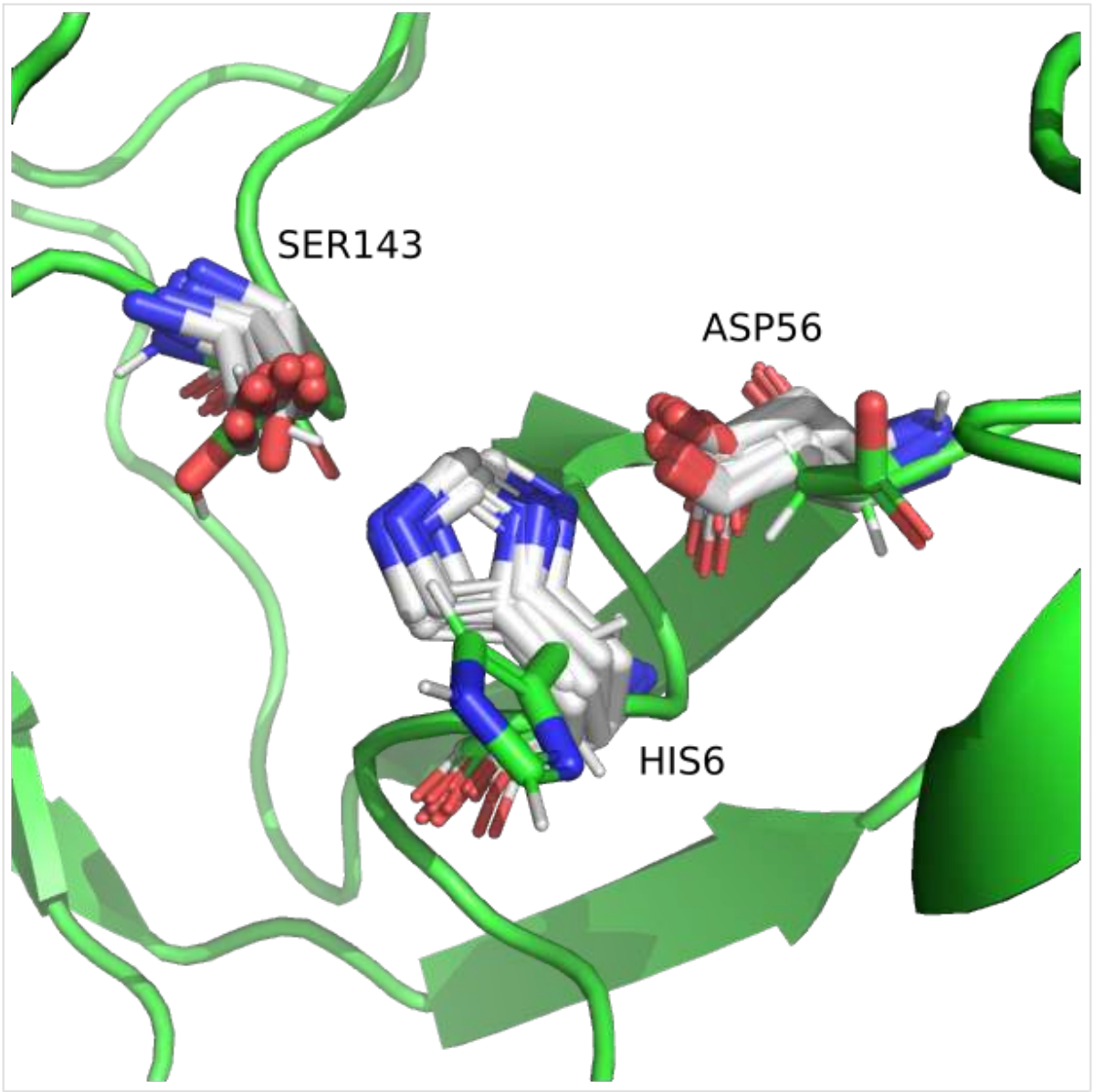
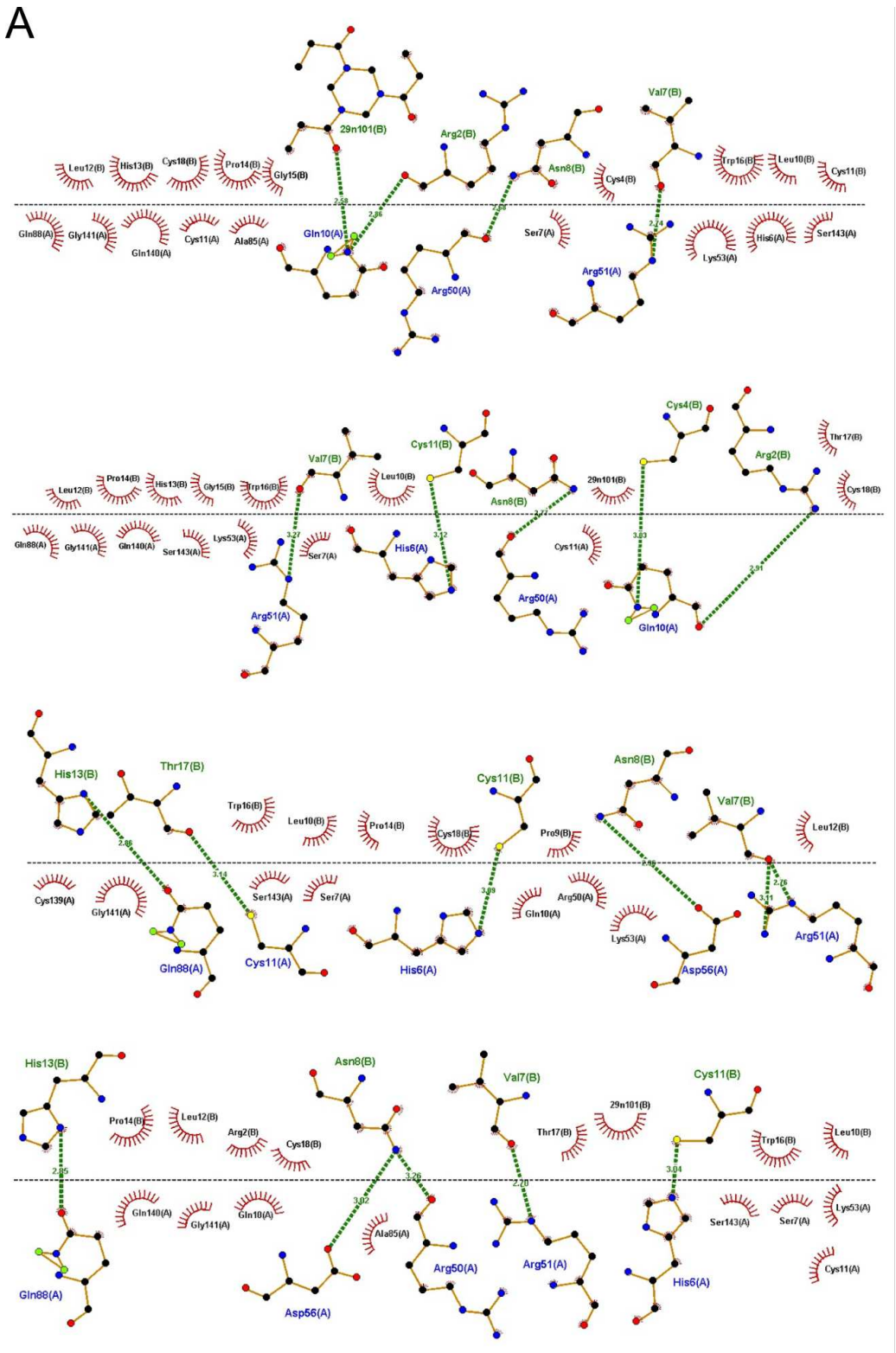
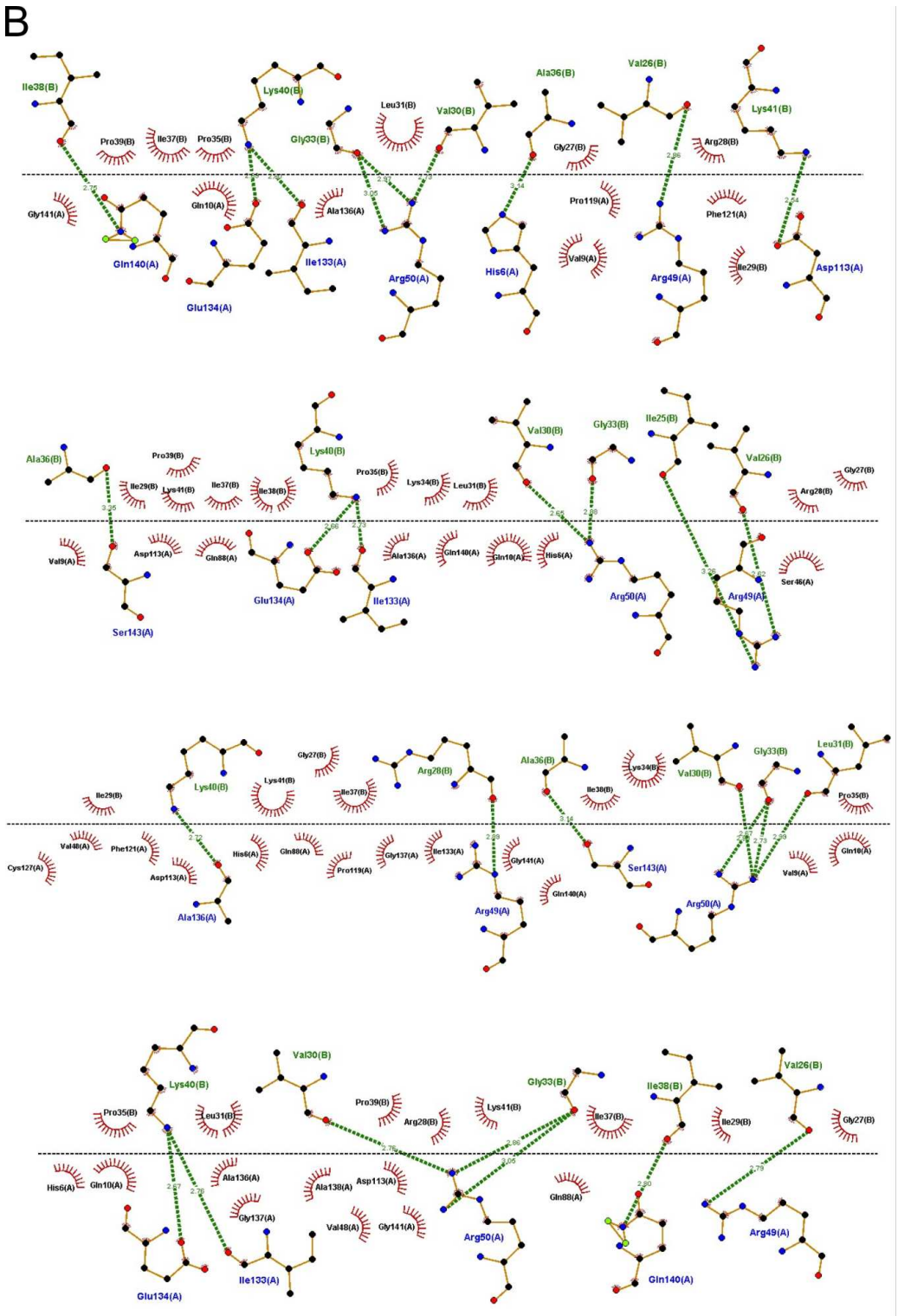


Figure 31 – Structural alignment of AG's protease model (in green), and the highest ranked templates used in the modeling procedure. Residues in gray are the template's residues corresponding to the catalytic triad


A
29n101(B)
Arg2(B)
Asn8(B)
Val7(B)
Leu12(B)
His13(B)
Cys18(B)
Pro14(B)
Gly15(B)
Cys4(B)
Trp16(B)
Leu10(B)
Cys11(B)
Gln88(A)
Gly141(A)
Gln140(A)
Cys11(A)
Ala85(A)
Gln10(A)
Ser7(A)
Lys53(A)
His6(A)
Ser143(A)
Arg51(A)
Arg50(A)
Thr17(B)
Ser7(A)
Cys139(A)
Asp56(A)


B
Ile38(B)
Pro39(B)
Ile37(B)
Pro35(B)
Lys40(B)
Gly33(B)
Leu31(B)
Val30(B)
Ala36(B)
Gly27(B)
Val26(B)
Arg28(B)
Lys41(B)
Gly141(A)
Gln140(A)
Gln10(A)
Glu134(A)
Ala136(A)
Ile133(A)
Arg50(A)
His6(A)
Pro119(A)
Val9(A)
Arg49(A)
Phe121(A)
Ile29(B)
Asp113(A)
Ala36(B)
Pro39(B)
Ile29(B)
Lys41(B)
Ile37(B)
Ile38(B)
Lys40(B)
Pro35(B)
Lys34(B)
Leu31(B)
Val30(B)
Gly33(B)
Ile25(B)
Val26(B)
Gly27(B)
Arg28(B)
Val9(A)
Asp113(A)
Gln88(A)
Ala136(A)
Gln140(A)
Gln10(A)
His6(A)
Ser46(A)
Glu134(A)
Ile133(A)
Arg50(A)
Arg49(A)
Ser143(A)
Gly27(B)
Lys40(B)
Lys41(B)
Ile37(B)
Arg28(B)
Ala36(B)
Ile38(B)
Lys34(B)
Val30(B)
Gly33(B)
Leu31(B)
Pro35(B)
Ile29(B)
Val48(A)
Phe121(A)
Cys127(A)
Asp113(A)
Ala136(A)
His6(A)
Gln88(A)
Pro119(A)
Gly137(A)
Ile133(A)
Gly141(A)
Arg49(A)
Ser143(A)
Gln140(A)
Arg50(A)
Val9(A)
Gln10(A)
Lys40(B)
Pro35(B)
Leu31(B)
Val30(B)
Pro39(B)
Arg28(B)
Lys41(B)
Gly33(B)
Ile37(B)
Ile38(B)
Ile29(B)
Val26(B)
Gly27(B)
His6(A)
Gln10(A)
Ala136(A)
Ala138(A)
Asp113(A)
Gln88(A)
Gly137(A)
Val48(A)
Gly141(A)
Arg50(A)
Arg49(A)
Glu134(A)
Ile133(A)
Gln140(A)

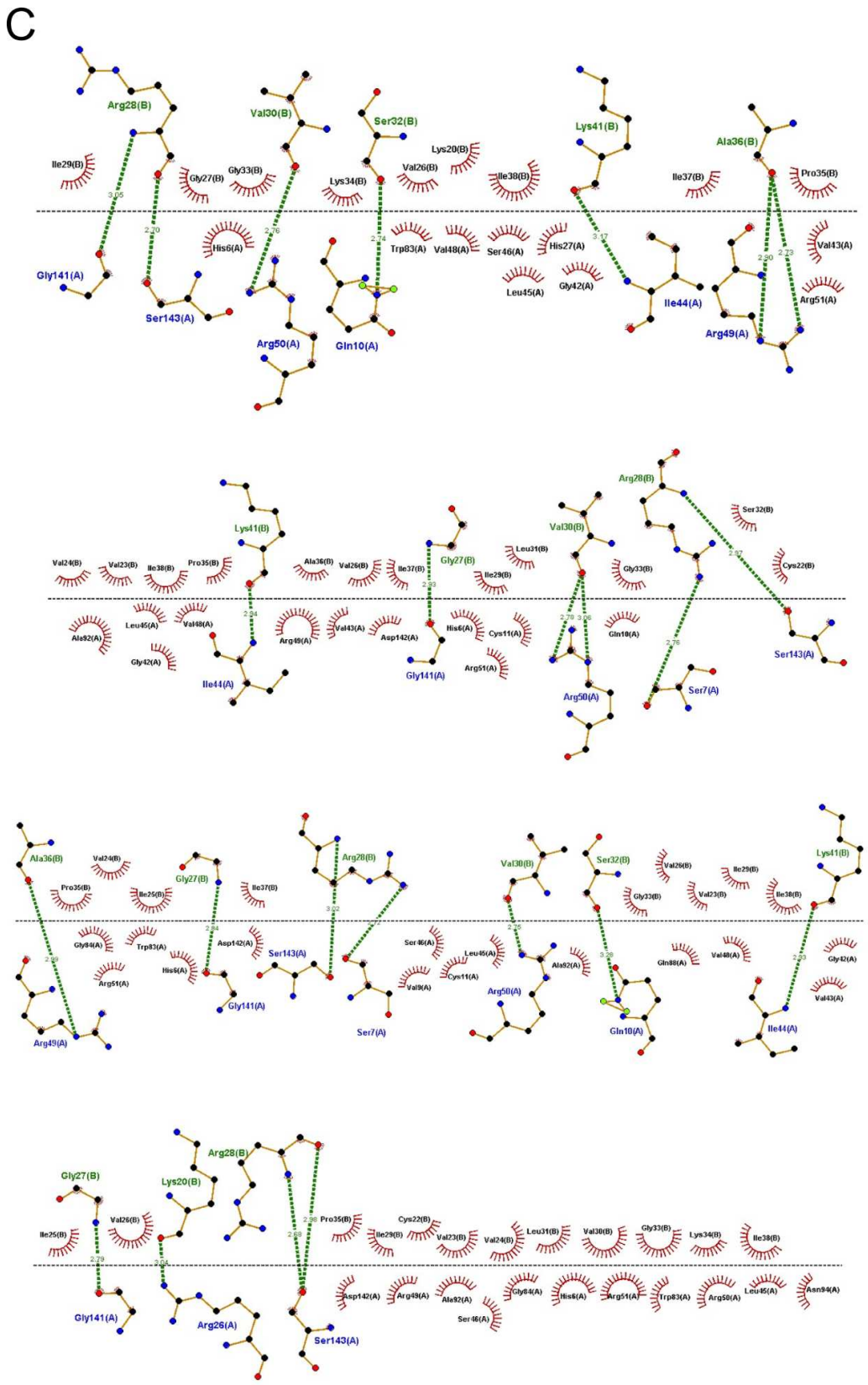


Figure 32 – Interaction maps for the four best poses of the complexes 6rw2_B (A), 3kn2_B (B) and 2obq_B (C). Protein residues are labeled in blue and peptides in green. H-bonds are highlighted in dash green lines and hydrophobic interactions are indicated by red arcs.

Appendix C

Patent request: Uso do ambenônio para inibição de proteases de coronavírus, com aplicação no combate a coronaviroses

For the security of the authors' personal data, only the order data pages are shown.

INSTITUTO NACIONAL DA PROPRIEDADE INDUSTRIAL

25/08/2021 870210078494 17:13

29409161938079166

Pedido nacional de Invenção, Modelo de Utilidade, Certificado de Adição de Invenção e entrada na fase nacional do PCT

Número do Processo: BR 10 2021 016873 0

Dados do Depositante (71)

Depositante 1 de 1

Nome ou Razão Social: UNIVERSIDADE FEDERAL DE VIÇOSA
Tipo de Pessoa: Pessoa Jurídica
CPF/CNPJ: 25944455000196
Nacionalidade: Brasileira
Qualificação Jurídica: Instituição de Ensino e Pesquisa
Endereço: Campus UFV, Pró-Reitoria de Pesquisa e Pós Graduação, sala 04.
Cidade: Vicosa
Estado: MG
CEP: 36570-900
País: Brasil
Telefone: (31) 3612 2334
Fax:
Email: propriedadeintelectual@ufv.br

PETICIONAMENTO ELETRÔNICO

Esta solicitação foi enviada pelo sistema Peticionamento Eletrônico em 25/08/2021 às 17:13, Petição 870210078494

Dados do Pedido

Natureza Patente:	10 - Patente de Invenção (PI)
Título da Invenção ou Modelo de Utilidade (54):	Uso do ambenônio para inibição de proteases de coronavírus, com aplicação no combate a coronaviroses
Resumo:	A presente invenção refere-se ao uso do medicamento ambenônio (DB01122) como potencial inibidor de proteases de coronavírus, macromolécula essencial no ciclo reprodutivo da família de vírus causadores de infecções, principalmente respiratórias. O ambenônio é usado, atualmente, no tratamento da doença muscular myasthenia gravis, por ser um inibidor de acetilcolinesterase, principalmente receptores muscarínico e nicotínico. A proposta do atual pedido de patente foi baseada em estudos computacionais de uma representante do grupo das main proteases de coronavírus, a 3CLpro de SARS-COV-2. Foi verificado que existe potencial de inibição do fármaco em relação à atividade da 3CLpro viral, e consequentemente seu uso no tratamento da covid-19, coronaviroses e infecções respiratórias causadas pelo coronavírus.
Figura a publicar:	1