



Bioinformatics and Computational Biology Research at the Computer Science Department at UFMG

Diego Mariano  [Universidade Federal de Minas Gerais | diego@dcc.ufmg.br]


Frederico Chaves Carvalho  [Universidade Federal de Minas Gerais | fredericochaves@dcc.ufmg.br]

Luana Luiza Bastos  [Universidade Federal de Minas Gerais | luizabastos.luana9@gmail.com]

Lucas Moraes dos Santos  [Universidade Federal de Minas Gerais | lucas.santos@dcc.ufmg.br]

Vivian Moraes Paixão  [Universidade Federal de Minas Gerais | vivianmp95@ufmg.br]

Raquel C. de Melo-Minardi   [Universidade Federal de Minas Gerais | raquelcm@dcc.ufmg.br]

 *Laboratory of Bioinformatics and Systems (LBS), Department of Computer Science, Universidade Federal de Minas Gerais, Av. Pres. Antônio Carlos, 6627 - Pampulha, Belo Horizonte, MG, 31270-901, Brazil.*

Received: 5 May 2022 • Published: 16 February 2024

Abstract Bioinformatics is an emerging research field that encompasses the use of computational methods, algorithms, and tools to solve life science problems. At the Laboratory of Bioinformatics and Systems (LBS), our research lines include the use of graph-based algorithms to improve the prediction of the structure and function of macromolecules, the detection of molecular recognition patterns, the application of mathematical models and artificial intelligence techniques to assist enzyme engineering, and development of models, algorithms, and tools. Additionally, the group has played a role in scientific outreach and spreading bioinformatics in Brazil. In this article, we summarize the 20 years of Bioinformatics and Computational Biology research conducted by our group at LBS in the Department of Computer Science at the Universidade Federal de Minas Gerais (DCC-UFMG).

Keywords: Artificial Intelligence, Computational Biology, Data Mining, Data Visualization, Machine Learning, Research Group, Structural Bioinformatics

1 History

The research group in Bioinformatics and Computational Biology was created in 2003 by professors Marcelo Matos Santoro (Department of Biochemistry and Immunology at UFMG) and Wagner Meira Jr. (Department of Computer Science - DCC at UFMG). The laboratory was born at the same time as the creation of the Graduate Program in Bioinformatics at UFMG (Universidade Federal de Minas Gerais). Professor Santoro had experience in the field of Biochemistry, with an emphasis on proteins, and was studying trypsin enzymes, the thermodynamic stability of proteins, protein purification, calorimetry, and protein stability. Professor Meira works in the areas of parallel and distributed systems and data mining and its application in social networks, cybersecurity, electronic commerce, information retrieval, and bioinformatics, among others.

They started a set of weekly meetings to discuss the group's lines of research and began to mentor graduate students together. Their students were allocated to a computational research laboratory that came to be called the "Laboratory of Bioinformatics and Systems (LBS)". Among the first students advised by them are professors: Carlos Henrique da Silveira (UNIFEI - Campus de Itabira) and Raquel Cardoso de Melo-Minardi (Department of Computer Science at UFMG). Both defended their thesis in the first semester of 2008 and worked on topics related to the search for structural/functional signatures in proteins. Professor Santoro advised four other PhDs in Bioinformatics until he died in 2014. Professor Meira advised five other PhDs in Bioinformatics

and left the Bioinformatics Graduate Program in 2016. In 2010, Professor Minardi became a professor at the Department of Computer Science at UFMG and rejoined the lab, and since then has advised six PhDs and eight Masters in Bioinformatics.

A relevant milestone for the group was the approval of a project in the CAPES Computational Biology grant in 2013. We got second place (out of 39 proposals) in the call aimed at inducing training of human resources and research in Computational Biology, and we received the amount of R\$3.2 million for this project. Our proposal aimed to face two major challenges of biotechnological and pharmacological innovation: (i) mutant cellulases in producing second-generation biofuels; and (ii) ricin inhibitors, a highly lethal phytotoxin that makes castor bean agribusiness difficult.

Unfortunately, Professor Santoro passed away on March 10, 2014. The CAPES project was implemented by the team under the coordination of Professor Minardi according to the group's decision. Since then, the professor has led the research group in Bioinformatics and Computational Biology at DCC/UFMG.

2 Research lines and results

As part of the highly interdisciplinary field of Bioinformatics, our research lines are diverse. We have experience developing machine learning models and algorithms, data mining techniques, and bioinspired algorithms. Additionally, we use these tools to push the boundaries of the current knowledge about the molecular basis of open problems of biotechnolog-

ical relevance. For instance, some of our current research lines aim to solve problems related to structural computational biology by developing and applying artificial intelligence methods, particularly machine learning and deep neural networks. We have also worked on several problems and application scenarios in collaboration with experimental research groups. The word cloud in Figure 1 shows some of the most important keywords that describe our recent publications.

2.1 Structural-functional signatures and classification

The first research line created at LBS focused on the protein folding problem (PFP), more specifically, on the understanding of the sequence vs structure vs function of proteins and was headed by Professor Santoro. One of our initial goals was to raise our level of understanding of why proteins from different organisms with different sequences fold into such similar structures with identical functions. In this context, our research line on structural signatures emerged. *Structural signatures* are a set of characteristics capable of unequivocally identifying a protein folding and the nature of the interactions it can establish with other proteins and ligands. In this way, the group explored different possibilities and applications of these signatures.

During her thesis, Professor Minardi designed and implemented methods based on contact maps for the structural classification of proteins [de Melo et al., 2006]. *Contact maps* are two-dimensional matrices in which the x and y axes represent the linear sequence of amino acids in a protein chain, and the dots (x, y) indicate the existence of contacts representing possible non-covalent chemical interactions between the x and y amino acid residues. She also showed that these methods could identify molecular recognition patterns at protein-protein interfaces [de Melo et al., 2007].

Professor Silveira created the Protein Cutoff Scanning approach [Silveira et al., 2009], which would later give rise to the emergence of Cutoff Scanning Matrices (CSM). This gave rise to a very robust signature for several tasks involving structural classification, function prediction, molecular recognition, and mutation impact, among other applications. The idea arose from his observation of the packing pattern of atoms in protein families. He hypothesized a conservation pattern in the packing mode, which he formalized as vectors of frequency pairs of atoms at various distances (neighborhoods) in 3D space.

This work was continued and improved later by Douglas Pires during his PhD studies. He proposed the use of these matrices and variations of them in problems related to the classification of protein structures [Pires et al., 2011] and problems involving molecular recognition [Pires et al., 2013]. After his PhD, he has already shown numerous applications of these matrices in diverse problems, such as predicting the impact of a mutation on the stabilization/destabilization of a protein [Pires et al., 2014], predicting the affinity of binding sites with small molecules [Pires and Ascher, 2016], and predicting antibody/antigen affinity [Myung et al., 2022], to name a few examples.

More recently, in this line of research, we developed more

fundamental models for the identification of structural and functional signatures of proteins based on chemical interactions [Martins et al., 2018; Gadelha Campelo et al., 2019].

2.2 Function prediction and Molecular recognition patterns

Proteins are macromolecules that perform many of the essential functions of living beings. In this regard, they have been considered the machinery of life. They are composed of a linear sequence of amino acid residues bound covalently in peptide bonds. It is well known that protein sequences fold into specific three-dimensional structures and that protein function is highly dependent on this fold. In the folded proteins, residues participate in non-covalent interactions, herein called contacts, that will determine the three-dimensional shape of the protein and have an important role in the interaction of proteins with their ligands - small molecules, peptides, and other proteins. Our goal with this line of research is to understand the relation between sequence, structure, and function, as well as the patterns of contacts involved in molecular recognition.

In her post-doc, Prof. Minardi developed ASMC (*Active Sites Modeling and Clustering*), an unsupervised method to classify sequences using structural information of protein pockets. The method predicts functional amino acids by proposing active site SDP residues (*Specificity Determining Positions*) and active site CP residues (*Conserved Positions*) profiles. ASMC combines homology modeling of family members, structural alignment of modeled active sites, and a subsequent hierarchical conceptual classification of obtained alignments. Comparison of profiles obtained from computed clusters allows identifying the residues correlated to sub-families function divergence. This method was used to discover diverse enzymatic activities [Bastard et al., 2014] catalyzed within protein families of unknown or little-known functions. As a case study, we investigated the DUF849 Pfam family and proposed 14 potential new enzymatic activities, designating these proteins as β -keto acid cleavage enzymes.

Later in Boari de Lima et al. [2016], we addressed the protein function prediction problem by a new constrained clustering algorithm based on genetic programming. We have also proposed a method for mapping the dynamics of EC number annotations in Swiss-Prot and identifying possible errors based on supervised learning [Silveira et al., 2014a].

2.3 Enzyme engineering

A more recent research line of the group focuses on applying computational approaches to detect mutation sites of interest for bioengineering enzymes used in second-generation biofuel production. One such enzyme is the β -glucosidases, essential enzymes that act in the last step of the saccharification process. However, they are inhibited by their product: glucose. Hence, since the 1970s, several studies have searched for strategies to improve the efficiency of the degradation of lignocellulose biomass, for example, by improving β -glucosidases tolerance to inhibition by glucose. The recent advances in Bioinformatics have brought new methodologies

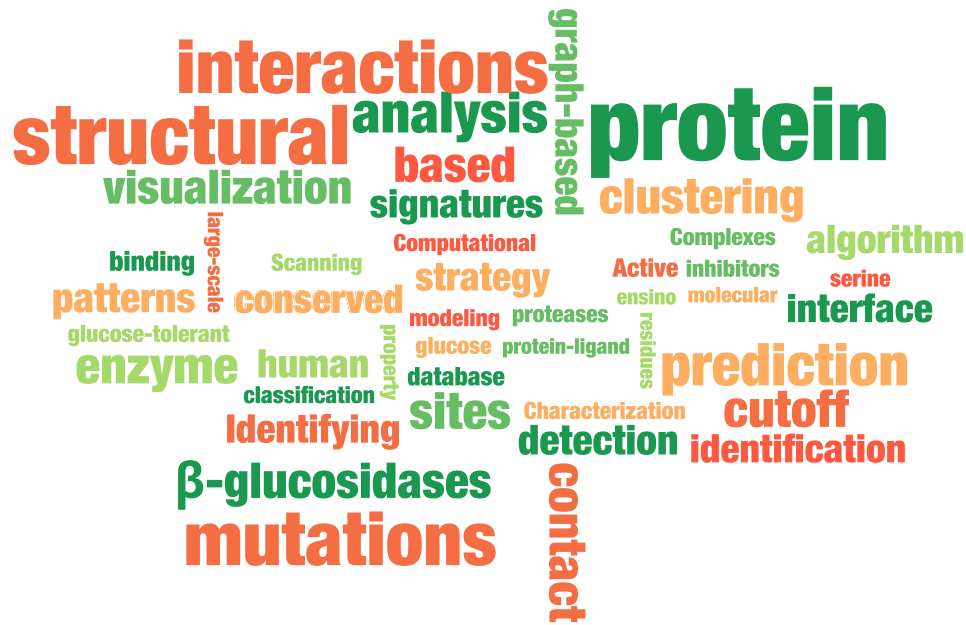


Figure 1. "Word cloud" with the most cited terms in the titles of our recent publications.

and strategies that can be applied to this problem, opening new possibilities for solving this problem.

In 2017, our group published a systematic literature review of glucose-tolerant β -glucosidases with data publicly available. During this project, we collected 23 structures of glucose-tolerant β -glucosidase enzymes previously described in the literature [Mariano *et al.*, 2017a] and analyzed the mutations described in the papers reviewed. This enabled us to propose a method based on the variation of structural signatures [Mariano *et al.*, 2019b] to identify mutation target sites of interest for increased glucose inhibition resistance.

In the aforementioned method, we modeled each protein as a graph where atoms are the vertices, and the distance of atoms pairs are the edges. We hypothesized that more glucose-tolerant mutants would have a Yesilar structural signature to the glucose-tolerant described in the literature. Thus, we used Euclidean distance to verify if induced mutations could lead to producing more efficient enzymes for biofuel production. Our results pointed out mutation target sites that could lead to glucose tolerance improvement, which agreed with previous studies in the literature [Yang *et al.*, 2015; de Giuseppe *et al.*, 2014]. Some of the mutant enzymes developed by our team have great efficiency and will be patented.

Another outcome of the literature review process was the establishment of a new methodology for performing systematic reviews specially applied to articles in bioinformatics [Mariano *et al.*, 2017b]. Our Bioinformatics systematic review strategy was based on other successful methodologies proposed for other areas, such as medical research and software development.

We also have been working on the development of models, algorithms, visualizations, and tools for the optimization of essential enzymes in the cellulose digestion process for the production of second-generation ethanol [Mariano *et al.*, 2017a, 2019b; Costa *et al.*, 2019; Mariano *et al.*, 2020b; Bar-

roso *et al.*, 2020; Lima *et al.*, 2021]. We used comparative molecular modeling to obtain three-dimensional structures of more than 3,500 β -glucosidase sequences available in public data banks, such as UniProt. We then extrapolated mutations described in the literature to these modeled enzymes. Lastly, we tested the docking of the product and substrate to evaluate the importance of some mutations in the protein-ligand interactions. Finally, we turned all available data obtained into a new public database called Glutant β ase [Mariano *et al.*, 2020b].

Using molecular dynamics, we proposed an explanation for the mechanisms of glucose withdrawal in glucose-tolerant β -glucosidases [Costa *et al.*, 2019]. Our results suggested that a "slingshot mechanism" performed by the D238 residue was responsible for removing glucose from the active site. Also, we demonstrated in a posterior study, using an accelerated molecular dynamics method, that the movements of loops located around the active site pocket entrance were related to glucose-tolerant structures [Lima *et al.*, 2021].

In 2019, we constructed a web tool called Proteus to detect stabilizing mutation pairs based on three-dimensional structures available in the Protein Data Bank [Barroso *et al.*, 2020]. Later, using our web tool VTR [Pimentel *et al.*, 2021], we demonstrated an essential difference in structural ionic interactions in the loops around the active site pocket between a glucose-tolerant β -glucosidase enzyme and a non-tolerant. Finally, we used several molecular dynamics techniques to verify the impacts of canonical single amino acid substitutions on the thermostabilizing mechanisms of GH1 β -glucosidase enzymes [Rocha *et al.*, 2023]. All these results corroborate themselves, promoting this as a successful research line for our group.

2.4 Tools, databases, and data visualization

As mentioned in the previous sessions, our group expertise also includes developing web tools, web platforms, and databases to aid in important tasks of high relevance for bioinformatics research. Over the last seven years, we developed more than ten tools and four databases that helped numerous researchers worldwide. Our goal is to provide high-quality tools and curated data that allow fellow scientists to achieve their research goals faster and more reliably by bringing more automation to the analysis process while offering useful visualizations and refined results that help them extract the most value from their data. Our tools also help lower the entry barriers in bioinformatics by reducing the need for advanced computational knowledge of the users.

The web tools and web platforms developed by our group in the recent years include: (1) PDBest [Goncalves *et al.*, 2015], a platform for searching, collecting and standardizing protein structures and their ligands; (2) VERMONT [Fassio *et al.*, 2017], a platform that allows us understand mutations by interpreting position-specific structural and physicochemical properties; (3) nAPOLI [Fassio *et al.*, 2019] which was developed to be a web server for large-scale analysis of protein-ligand interactions; (4) CALI [Medina *et al.*, 2017], that proposes a strategy based on complex network model of protein-ligand interactions to reveal frequent and relevant patterns among them; (5) MutaGraph [Rodrigues, 2017] is a computational model able to predict and evaluate the effect of the substitution of a single amino acid for another in a protein complex; (6) SSV [Mariano *et al.*, 2019b], a tool to propose mutations for enzymes used in industrial applications based on structural signatures; (7) Proteingo [Silva *et al.*, 2019], which was created as a game for analyzing contacts in proteins; (8) Proteus [Barroso *et al.*, 2020], a platform based on a database of conformations of contacts and their neighborhood, which aims to assist in the proposal of mutations in pairs of residues involved in interactions, designed primarily for use in proteins engineering; (9) VTR [Pimentel *et al.*, 2021], a recently developed that calculates and matches contacts between two proteins; (10) and finally, the recently published E-Volve [Dos Santos *et al.*, 2022], a web tool designed to model mutations in the input protein complex and is powered by Modeller [Webb and Sali, 2016].

Some of the aforementioned software designed by our group also serve as powerful visualization tools for understanding biological problems. For instance, VERMONT [Fassio *et al.*, 2017] is also useful for generating visualizations that help us to analyze mutations in protein sequences. nAPOLI's visualizations [Fassio *et al.*, 2019] are useful for detecting protein-ligand interactions, while VTR [Pimentel *et al.*, 2021] helps the user visualize matching contacts. In Santana *et al.* [2016] and Ribeiro *et al.* [2020], we propose a graph-based model for mining and interactive visualization of frequent subgraphs in protein-ligand interaction graphs. Finally, we also have some works on biological data visualization presented in different editions of the IEEE Biological Data Visualization [Silveira *et al.*, 2012, 2014b], one of them was awarded in the *Data Contest Biology Experts Pick* category in 2013.

The E-Volve tool [Dos Santos *et al.*, 2022] is our lat-

est web tool that also has visualization capabilities. In this case, the visualizations generated by E-Volve aim to assist the user in evaluating the effect of mutations in the interactions between proteins. Recently, we used E-Volve to systematically probe and visualize the impact of these mutations in several SARS-COV-2 variants of concern, which helped us understand how they affect the interactions between the SARS-COV-2 spike's protein and the human acceptor ACE2 [Dos Santos *et al.*, 2022].

In addition to this, as previously mentioned, our group developed four databases: (1) Betagdb Mariano *et al.* [2017a] contains structures and data referent to beta-glucosidase that are resistant to glucose; (2) the CAPRI database [Martins *et al.*, 2018] contains information of about 45,000 protein complexes accompanied by data regarding interactions between pairs of atoms, residues, and chains; (3) Glutantbase [Mariano *et al.*, 2020b] is, at the same time, a database, a web tool and a method to evaluate mutations for beta-glucosidases proteins used in industrial applications; (4) PROPEdia [Martins *et al.*, 2021], our latest database hosts peptide-protein complexes and has aimed to provide new insights into peptide design. Recently, PROPEdia received a major update and now has almost 50,000 structures of protein-peptide complexes [Martins *et al.*, 2023].

Lastly, we use user-friendly web tools to tell the story of bioinformatics and show the impact of collaboration networks of Brazilian authors for this field of research [Mariano *et al.*, 2020a].

2.5 Teaching and scientific outreach

In addition to conducting basic and applied research, our group has promoted actions related to teaching and scientific outreach on bioinformatics and computational biology. Our objective in this topic is helping to diffuse bioinformatics to students of all levels, discover new talents, and empower a new generation of scientists.

In 2019, we proposed a game-based strategy to detect occlusions in residue pairs interacting in protein structures [Silva *et al.*, 2019]. Our web-mobile game was called "pro-teingo" and offered a user-friendly interactive interface to allow users to evaluate if two residues are performing contact or not. The user could choose between five types of interactions: hydrogen bonds, ionic, hydrophobic, aromatic stacking, or non-contact. We evaluated proteingo performing a live experiment with hundreds of users during the "X-Meeting Conference of 2017", where the work was awarded the best poster prize.

Also, students of the group collaborated with the organizing committee of the events: the Course on Bioinformatics (2020, 2022, and 2023) of UFMG and the Online Workshop on Bioinformatics (Nov/2020). An event report and a guide with tips for organizing online and in-person events were published in [Silva and et al, 2021].

Furthermore, we published a report on the experience of teaching Python programming language to life science students [Mariano *et al.*, 2019a]. In addition, we proposed an article for young audiences showing how computers have been used to improve biofuel production [Mariano *et al.*, 2022]. We also published a series of books in the Portuguese lan-

Table 1. The tools developed by our group.

Tool	Description	Access Link
PDBBest	PDB Enhanced Structures Toolkit.	http://www.pdbbest.dcc.ufmg.br/
VERMONT	Tool for viewing mutations.	http://bioinfo.dcc.ufmg.br/vermont/
NAPOLI	Large-scale analysis of protein-ligand interactions	http://bioinfo.dcc.ufmg.br/napoli
CALI	Complex Analysis of Protein-Ligand Interactions	http://bioinfo.dcc.ufmg.br/cali
MutaGraph	Predict the effect of the substitution of a single amino acid	http://bioinfo.dcc.ufmg.br/mutagraph
SSV	Method to propose mutations for enzymes	http://bioinfo.dcc.ufmg.br/ssv/
Proteingo	Game for analyzing contacts in proteins	http://bioinfo.dcc.ufmg.br/proteingo/
Proteus	An algorithm to propose stabilizing mutations	http://proteus.dcc.ufmg.br
VTR	Calculate and match contacts in two proteins	http://bioinfo.dcc.ufmg.br/vtr
E-VOLVE	Designed to model mutations in the protein complex	http://bioinfo.dcc.ufmg.br/evolve
BETAGDB	Database of glucose-tolerant beta-glucosidase structures	http://bioinfo.dcc.ufmg.br/betagdb
Glutantbase	Database for glucose-tolerant beta-glucosidase models	http://bioinfo.dcc.ufmg.br/glutantbase
CAPRI	Database for analysis of protein-protein interaction	http://bioinfo.dcc.ufmg.br/capri
Propedia	Database of peptide-protein complexes clusterized	http://bioinfo.dcc.ufmg.br/propedia

guage to teach scientific writing [Mariano and Santos, 2021] and programming for bioinformatics using Python [Mariano and et al, 2015], Perl [Mariano and de Melo Minardi, 2016], and Web-based script languages [Mariano and de Melo Minardi, 2017; Paixão *et al.*, 2023].

In 2021, we presented an analytic-descriptive study about implementing an online extension course in Bioinformatics at UFMG [de Melo-Minardi and Bastos, 2021]. We published conceptual papers about structural alignments [Santos *et al.*, 2021], biological data banks [Liborio and Resende, 2021], and computational modeling of proteins [Xavier *et al.*, 2021]. Finally, we organized a special issue of *Frontiers in Bioinformatics* journal called "Bioinformatics in the age of data science: algorithms, methods, and tools applied from Omics to structural data" [Mariano *et al.*, 2023].

3 Students, Alumni, and Collaborators

Our group has already graduated 15 PhDs and 8 masters in Bioinformatics, published dozens of articles in the best venues in the area, and continues to be a center of attraction for students and training of scientists absorbed by several educational, science, and technology institutions in Brazil and abroad. A detailed list of our past and present members can be found in the appendix ??.

In Figure 2, we show our collaboration network, which highlights our papers published between 2003-2022 (blue nodes) and Authors (red nodes). The table 2 indicates our most recent collaborators. The complete list of corresponding papers to each ID is included in the appendix ??.

4 Perspectives and Future directions

Several challenging problems have served as research topics for master's and doctoral students in our group. Work in progress includes the study of SARS-CoV-2 since Covid-19 has become one of the most worrisome diseases in recent times, mainly because of its characteristic of generating new variants, thus making it important to understand the

best ways to fight it. For this, developing methods capable of predicting the potential of mutations in the spike target protein may be of great importance in containing the disease. There are ongoing projects that aim to develop models and algorithms for predicting the impact of mutations in the spike protein using molecular modeling, contact calculation, and calculation of structural signatures. In addition, algorithms are being developed for the rational design of peptides and peptidomimetic compounds to inhibit potential SARS-CoV-2 targets. Still, regarding SARS-CoV-2, convolutional neural networks are being used to identify conformational changes caused by mutations in specific regions of proteins from patterns found in distance maps. As a case study, the SARS-CoV-2 spike protein receptor-binding domain, as well as mutations monitored by the World Health Organization (WHO), existing in this region, are being analyzed. Finally, a study is being carried out on inhibiting the TMPRSS2 protease, which is important for SARS-CoV-2 replication. We will mainly study aprotinin. The aim is to use machine learning and protein structure data to prospect molecules with potential action against the virus [Paixão and Melo-Minardi, 2022].

Another topic of interest in several ongoing research projects is the study and prospecting of autism-causing molecules using machine learning models. Considering that the increased incidence of autism may be related to substances present in our everyday lives and that this occurs because of epigenetic modifications, it is important to identify key factors at the genesis of autism. The aim is to study the hypothesis and prospecting molecules that cause autism through machine learning. Another approach being studied in relation to autism spectrum disorders is the use of preventive therapies based on compounds of plant origin as a curative benefit and reduction of side effects of conventional treatment. In this context, the project aims to perform multi-target virtual screening and the search for pharmacophoric signatures of phytochemical compounds promising to treat autism spectrum disorder.

In addition, the development and improvement of methodologies and algorithms for computational structural biology problems continue to be part of the focus of our research projects, among which are the development of machine learn-

Table 2. Recent collaborators.

Name	Institution	Area	Country
Adriano de Paula Sabino	UFMG	Pharmaceutical Sciences	Brazil
Wagner Meira Jr.	UFMG	Computer Science	Brazil
Gisele Lobo Pappa	UFMG	Computer Science	Brazil
Rafaela Ferreira	UFMG	Biochemistry	Brazil
Lucas Bleicher	UFMG	Biochemistry	Brazil
Ronaldo Nagem	UFMG	Biochemistry	Brazil
Lúis Fernando Marins	FURG	Genetics	Brazil
Karina Machado	FURG	Computer Science	Brazil
Adriano Werhli	FURG	Computer Science	Brazil
Carlos Silveira	UNIFEI	Bioinformatics	Brazil
Sabrina Azevedo Silveira	UFV	Bioinformatics	Brazil
Leonardo Lima	UFSJ	Biophysics	Brazil
Rafael Oliveira Rocha	UFMG	Biochemistry	Brazil
Maurício Schneedorf	UNIFAL	Biochemistry	Brazil
Demetrius Araújo	UFPB	Biochemistry	Brazil
Gerd Rocha	UFPB	Computational Chemistry	Brazil
Luis Cezar Rodrigues	UFPB	Chemistry	Brazil
Lucianna Helene Santos	Institut Pasteur	Biophysics	Uruguay
François Artiguenave	<i>Commissariat à l’Energie Atomique et aux Énergies Alternatives</i>	Genomics	France

ing algorithms applied to computational structural biology problems, as well as the elaboration of a computational pipeline that combines mechanistic models, machine learning and bioinspired algorithms to generate peptides optimized for functions of medical interest. Since peptide engineering is a field that still depends primarily on in vitro experiments, a process of trial and error, the use of new methodologies can bring more efficiency, speed, and precision to the area. Another ongoing project is the development of a computational method based on structural signatures and machine learning to predict the interaction of protein-protein complexes since the identification of protein-protein interactions has several challenges in its in vitro experiments and the methodologies available for protein-protein docking has several limitations, including the size of the complexes.

Considering the growing interest in the application of bioinformatics and machine learning methodologies to problems faced by the agronomic sector and the good results already obtained by our group, we are developing a database for the rational design of insecticides based on natural products, using the techniques of protein-peptide docking, virtual screening, Machine Learning, and molecular modeling.

4.1 Artificial Intelligence and Molecular Simulations

The Big Data phenomenon has become a reality in bioinformatics. With the already large and evergrowing amount of publicly available data of biological interest, it is increasingly important to employ powerful data processing and analysis techniques to make sense of this data, uncovering patterns and revealing new insights.

We are currently implementing lines of research that focus on investigating how different artificial intelligence methods can help achieve our team’s research goals. This line of research focuses on applying machine learning and deep learn-

ing methods to understand better how to improve the rational design of proteins and peptides. In particular, we seek to develop novel ways to represent these biomolecules computationally in manners that can increase the performance of deep learning models trained on large datasets such as the Propedia database [Martins *et al.*, 2021, 2023]. To this end, we are investigating suitable data mining methods and deep learning based natural language processing techniques to extract meaningful features from sequence data.

In parallel, we are building generative strategies to propose novel sequences for peptides with desirable properties by employing two different approaches: bioinspired algorithms and deep learning-based generative models. In the former, we combine genetic algorithms with molecular simulations to optimize the binding affinity of peptides to a selected target while generating a few thousand simulation results that can be further used to train predictive and generative models. The second approach consists of using experimental and simulated data to train deep learning models capable of proposing peptides with high binding affinity for a region of interest of a particular target. In both cases, our targets are the spike protein and the main protease of the SARS-CoV2 virus.

Recent approaches based on Deep Learning have excelled in solving open problems in structural biology, such as protein folding. Applications such as TrRosetta [Yang *et al.*, 2020] and, more recently, AlphaFold2 [Senior *et al.*, 2020] have become the state of the art when it comes to *de novo* protein structure prediction, achieving significant performances in the modeling of protein structures. We are developing methods to leverage such approaches and models to build an end-to-end pipeline for the rational design of antiviral peptides.

One aspect that contributes to the performance of these applications is the use of deep learning algorithms such as Convolutional Neural Networks, where successive convolutional layers form dense neural networks capable of iden-

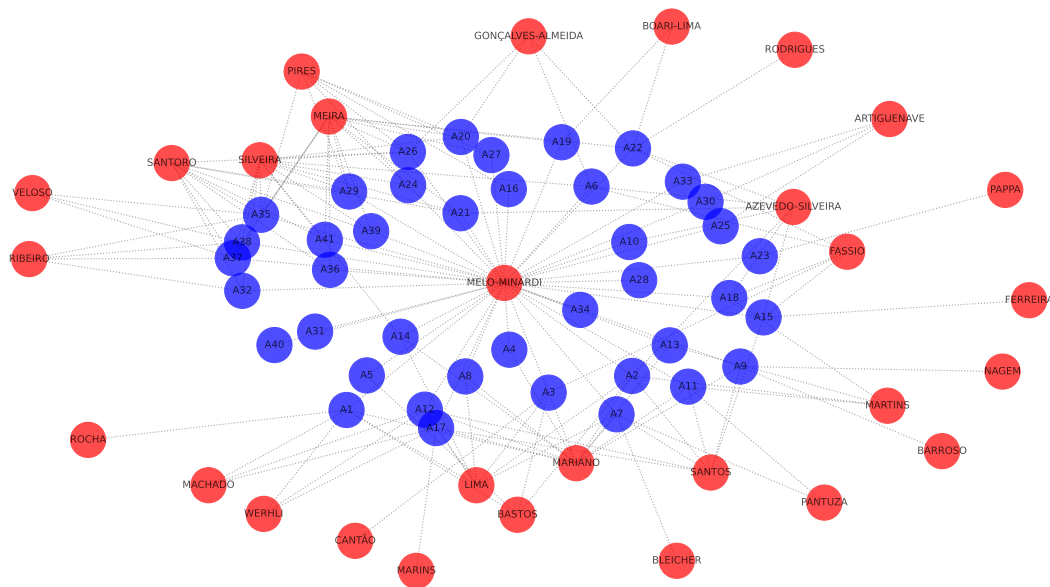


Figure 2. LBS' collaborations map

tifying complex patterns in images. Another recent work of our group aims to discover whether convolutional neural networks can identify conformational changes in proteins caused by point mutations in their structure. Conformational changes in protein structures are strongly correlated with functional changes. Some modifications may be easily noticeable, while others are more subtle. To tackle this problem, we model the protein conformation changes problem as a classification problem through its representation as images that illustrate matrices of interatomic distances, known as *distance maps*.

In a recent work, we employed CNNs to identify large-scale conformational changes in protein structures, which were represented through distance maps. Our focus was on discerning two distinct states of the SARS-CoV-2 spike glycoprotein: distant trimers, often referred to as pre-fusion (open), and the closest trimers (closed). The trained classifier successfully recognized both states of the spike, achieving an accuracy of 71.8% during testing, accompanied by an error rate of 28.2% [Santos and Melo-Minardi, 2022]. Nevertheless, we aim to enhance the evaluation of these architectures by assessing their capability to identify even more subtle conformational changes, including those induced by point mutations observed in virus variants.

Thus, we propose a model based on convolutional neural networks capable of classifying mutated (or not) protein structures only from the distance patterns in these maps. As a case study, we used molecular dynamics simulations of the RBD region of the S protein of the SARS-CoV2 virus, which contains the contact residues with the human cell receptor ACE2. This makes the region conducive to the emergence of critical mutations, with the potential to increase ligand-receptor affinity or even immune evasion.

Acknowledgements

The authors thank the funding agencies: Coordenação de Aper-

feiçoamento de Pessoal de Nível Superior (CAPES), Fundação de Amparo à Pesquisa do Estado de Minas Gerais (FAPEMIG) and Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq).

Funding

Between 2011 and 2016, we had 11 projects supported by CAPES, CNPq, FAPEMIG, and PRPq/UFMG. The main one is the CAPES Computational Biology Grant, which contributed more than R\$3 million in scholarships, equipment, and funding. It is a project submitted in 2013 with Prof. Marcelo Matos Santoro as coordinator. Most group studies were financed in part by the Coordenação de Aperfeiçoamento de Pessoal de Nível Superior - Brasil (CAPES) - Finance Code 001. Project grant number 51/2013–23038.004007/2014–82.

Authors' Contributions

DM, FCC, LLB, LMS, VMP, and RCMM wrote the manuscript. All authors read and approved the final manuscript.

Competing interests

The authors declare that they have no conflict of interest.

Availability of data and materials

For a more detailed description of ongoing research projects and members, please visit our GitHub repository. Supplementary material is available at <https://github.com/LBS-UFMG/20years>.

- **Supplementary Table S1:** Concluded and ongoing students.
- **Supplementary Table S2:** IDs and citations of group publications from 2003–2022.
- **Supplementary text 1:** Current members and research projects.

References

- Barroso, J. R. M., Mariano, D., Dias, S. R., Rocha, R. E., Santos, L. H., Nagem, R. A., and de Melo-Minardi, R. C. (2020). Proteus: an algorithm for proposing stabilizing mutation pairs based on interactions observed in known protein 3d structures. *BMC Bioinformatics*, 21(1):1–21.
- Bastard, K., Smith, A. T., Vergne-Vaxelaire, C., Perret, A., Zaparucha, A., de Melo-Minardi, R., Mariage, A., Boutard, M., Debard, A., Lechaplais, C., et al. (2014). Revealing the hidden functional diversity of an enzyme family. *Nature Chemical Biology*, 10(1):42–49.
- Boari de Lima, E., Meira, W., and de Melo-Minardi, R. C. (2016). Isofunctional protein subfamily detection using data integration and spectral clustering. *PLoS Computational biology*, 12(6):e1005001.
- Costa, L. S. C., Mariano, D. C. B., Rocha, R. E. O., Kraml, J., Silveira, C. H. d., Liedl, K. R., de Melo-Minardi, R. C., and Lima, L. H. F. d. (2019). Molecular dynamics gives new insights into the glucose tolerance and inhibition mechanisms on β -glucosidases. *Molecules*, 24(18):3215.
- de Giuseppe, P. O., Souza, T. d. A., Souza, F. H. M., Zaphorlin, L. M., Machado, C. B., Ward, R. J., Jorge, J. A., Furriel, R. d. P. M., and Murakami, M. T. (2014). Structural basis for glucose tolerance in gh1 β -glucosidases. *Acta Crystallographica Section D: Biological Crystallography*, 70(6):1631–1639.
- de Melo, R., Lopes, C., Jr., F. F., da Silveira, C., Santoro, M., Carceroni, R., Jr., W. M., and Araújo, A. (2006). A contact map matching approach to protein structure similarity analysis. *Genet. Mol. Res.*, 5(2):284–308.
- de Melo, R., Ribeiro, C., Murray, C., Veloso, C., da Silveira, S., Neshich, G., Jr., W. M., Carceroni, R., and Santoro, M. (2007). Finding protein-protein interaction patterns by contact map matching. *Genet. Mol. Res.*, 6:1–10.
- de Melo-Minardi, R. C. and Bastos, L. L. (2021). Expandindo as paredes da sala de aula: aprendizados com o ensino a distância e ensino remoto emergencial. *Revista da Universidade Federal de Minas Gerais*, 28(1):106–125.
- Dos Santos, V. P., Rodrigues, A., Dutra, G., Bastos, L., Mariano, D., Mendonça, J. G., Lobo, Y. J. G., Mendes, E., Maia, G., dos Santos Machado, K., et al. (2022). Evolve: understanding the impact of mutations in sars-cov-2 variants spike protein on antibodies and ace2 affinity through patterns of chemical interactions at protein interfaces. *PeerJ*, 10:e13099.
- Fassio, A. V., Martins, P. M., Guimarães, S. d. S., Junior, S. S., Ribeiro, V. S., de Melo-Minardi, R. C., and Silveira, S. d. A. (2017). Vermont: a multi-perspective visual interactive platform for mutational analysis. *BMC Bioinformatics*, 18(10):403.
- Fassio, A. V., Santos, L. H., Silveira, S. A., Ferreira, R. S., and de Melo-Minardi, R. C. (2019). nAPOLI: a graph-based strategy to detect and visualize conserved protein-ligand interactions in large-scale. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, 17:1317–1328.
- Gadella Campelo, J. A. F., Rodrigues Monteiro, C., da Silveira, C. H., de Azevedo Silveira, S., and Cardoso de Melo-Minardi, R. (2019). Protein structural signatures revisited: Geometric linearity of main chains are more relevant to classification performance than packing of residues. In Rojas, I., Valenzuela, O., Rojas, F., and Ortuño, F., editors, *Bioinformatics and Biomedical Engineering*, pages 391–402, Cham. Springer International Publishing.
- Goncalves, W. R., Goncalves-Almeida, V. M., Arruda, A. L., Meira Jr, W., da Silveira, C. H., Pires, D. E., and de Melo-Minardi, R. C. (2015). Pdbest: a user-friendly platform for manipulating and enhancing protein structures. *Bioinformatics*, 31(17):2894–2896.
- Liborio, L. and Resende, V. (2021). *Revista Brasileira de Bioinformática*, volume 1 of 1, chapter Introdução aos bancos de dados biológicos. Alfahelix, Lagoa Santa, first edition.
- Lima, L. H. F. d., Fernandez-Quintéro, M. L., Rocha, R. E. O., Mariano, D. C. B., de Melo-Minardi, R. C., and Liedl, K. R. (2021). Conformational flexibility correlates with glucose tolerance for point mutations in β -glucosidases—a computational study. *Journal of Biomolecular Structure and Dynamics*, 39(5):1621–1634.
- Mariano, D., Da Fonseca Júnior, N. J., Santos, L. H., and Minardi, R. C. d. M. (2023). Bioinformatics in the age of data science: algorithms, methods, and tools applied from omics to structural data. *Frontiers in Bioinformatics*, 3:1246859.
- Mariano, D. and de Melo Minardi, R. C. (2016). *Introdução à Programação Para Bioinformática Com Perl*, volume 1 of 1. CreateSpace Independent Publishing Platform, first edition.
- Mariano, D. and de Melo Minardi, R. C. (2017). *Introdução à Programação Web para Bioinformática: HTML, CSS, PHP and JavaScript.*, volume 1 of 1. CreateSpace Independent Publishing Platform, first edition.
- Mariano, D. and et al (2015). *Introdução à Programação para Bioinformática com Biopython.*, volume 1 of 1. CreateSpace Independent Publishing Platform, first edition.
- Mariano, D., Ferreira, M., Sousa, B. L., Santos, L. H., and de Melo-Minardi, R. C. (2020a). A brief history of bioinformatics told by data visualization. In *Advances in Bioinformatics and Computational Biology: 13th Brazilian Symposium on Bioinformatics, BSB 2020, São Paulo, Brazil, November 23–27, 2020, Proceedings 13*, pages 235–246. Springer.
- Mariano, D., Leite, C., Santos, L., Marins, L., Machado, K., Werhli, A., Lima, L., and de Melo-Minardi, R. (2017a). Characterization of glucose-tolerant β -glucosidases used in biofuel production under the bioinformatics perspective: A systematic review. *Genet Mol Res*, 16(3):10–4238.
- Mariano, D., Martins, P., Santos, L., and de Melo-Minardi, R. C. (2019a). Introducing programming skills for life science students. *Biochemistry and Molecular Biology Education*, 0(0).
- Mariano, D., Pantuza, N., Santos, L. H., Rocha, R. E., de Lima, L. H., Bleicher, L., and de Melo-Minardi, R. C. (2020b). Glutant β ase: a database for improving the rational design of glucose-tolerant β -glucosidases. *BMC Molecular and Cell Biology*, 21(1):1–15.

- Mariano, D. and Santos, L. (2021). *Manual de Escrita Científica: Teoria e Prática Aplicadas à Bioinformática*. Alfa-helix.
- Mariano, D., Santos, L. H., Meleiro, L. P., Henrique, L., de Lima, F., Marins, L. F., and de Melo-Minardi, R. C. (2022). Using computers to improve biofuel production. *Frontiers Young Minds*. DOI: 10.3389/frym.2022.751195.
- Mariano, D. C., Leite, C., Santos, L. H., Rocha, R. E., and de Melo-Minardi, R. C. (2017b). A guide to performing systematic literature reviews in bioinformatics.
- Mariano, D. C. B., Santos, L. H., Machado, K. d. S., Werhli, A. V., de Lima, L. H. F., and de Melo-Minardi, R. C. (2019b). A computational method to propose mutations in enzymes based on structural signature variation (ssv). *International journal of molecular sciences*, 20(2):333.
- Martins, P., Mariano, D., Carvalho, F. C., Bastos, L. L., Moraes, L., Paixão, V., and Cardoso de Melo-Minardi, R. (2023). Propedia v2. 3: A novel representation approach for the peptide-protein interaction database using graph-based structural signatures. *Frontiers in Bioinformatics*, 3:1103103. DOI: 10.3389/fbinf.2023.1103103.
- Martins, P. M., Mayrink, V. D., de A. Silveira, S., da Silveira, C. H., de Lima, L. H. F., and de Melo-Minardi, R. C. (2018). How to compute protein residue contacts more accurately? In *Proceedings of the 33rd Annual ACM Symposium on Applied Computing, SAC '18*, page 60–67, New York, NY, USA. Association for Computing Machinery. DOI: 10.1145/3167132.3167136.
- Martins, P. M., Santos, L. H., Mariano, D., Queiroz, F. C., Bastos, L. L., de S. Gomes, I., Fischer, P. H. C., Rocha, R. E. O., Silveira, S. A., de Lima, L. H. F., de Magalhães, M. T. Q., Oliveira, M. G. A., and de Melo-Minardi, R. C. (2021). Propedia: a database for protein-peptide identification based on a hybrid clustering algorithm. *BMC Bioinformatics*, 22(1). DOI: 10.1186/s12859-020-03881-z.
- Medina, S. G., Fassio, A. V., de A. Silveira, S., da Silveira, C. H., and de Melo-Minardi, R. C. (2017). CALI: A Novel Visual Model for Frequent Pattern Mining in Protein-Ligand Graphs. In *2017 IEEE 17th International Conference on Bioinformatics and Bioengineering (BIBE)*, pages 352–358. IEEE. ISSN: 2471-7819. DOI: 10.1109/BIBE.2017.00-29.
- Myung, Y., Pires, D. E., and Ascher, D. B. (2022). Csm-ab: Graph-based antibody-antigen binding affinity prediction and docking scoring function. *Bioinformatics*, 38(4):1141–1143.
- Paixão, V., Puelles, A., de Abreu, A., Bastos, L., dos Santos, L., Carvalho, F., Mariano, D., and de Melo Minardi, R. (2023). *LBS Tech - Desenvolvimento web*. LBS Tech. Alfa-helix Publicações.
- Paixão, V. M. and Melo-Minardi, R. C. (2022). Computational methodology for discovery of potential inhibitory peptides. In *Advances in Bioinformatics and Computational Biology*, pages 91–96. Springer Nature.
- Pimentel, V., Mariano, D., Cantão, L., Bastos, L., Fischer, P., de Lima, L., and Fassio, A. (2021). Melo-minardi rcd (2021) vtr: A web tool for identifying analogous contacts on protein structures and their complexes. *Front. Bioinformatics*, 1:1–10. DOI: 10.3389/fbinf.2021.730350.
- Pires, D. E. and Ascher, D. B. (2016). Csm-lig: a web server for assessing and comparing protein-small molecule affinities. *Nucleic acids research*, 44(W1):W557–W561.
- Pires, D. E., Ascher, D. B., and Blundell, T. L. (2014). mcsm: predicting the effects of mutations in proteins using graph-based signatures. *Bioinformatics*, 30(3):335–342.
- Pires, D. E., de Melo-Minardi, R. C., da Silveira, C. H., Campos, F. F., and Meira Jr, W. (2013). aCSM: noise-free graph-based signatures to large-scale receptor-based ligand prediction. *Bioinformatics*, 29(7):855–861.
- Pires, D. E., de Melo-Minardi, R. C., dos Santos, M. A., da Silveira, C. H., Santoro, M. M., and Meira, W. (2011). Cutoff Scanning Matrix (CSM): structural classification and function prediction by protein inter-residue distance patterns. In *BMC Genomics*, volume 12, page S12. Springer.
- Ribeiro, V. S., Santana, C. A., Fassio, A. V., Cerqueira, F. R., da Silveira, C. H., Romanelli, J. P., Patarroyo-Vargas, A., Oliveira, M. G., Gonçalves-Almeida, V., Izidoro, S. C., de Melo-Minardi, R. C., et al. (2020). visGREMLIN: graph mining-based detection and visualization of conserved motifs at 3D protein-ligand interface at the atomic level. *BMC Bioinformatics*, 21(2):1–12.
- Rocha, R. E. O., Mariano, D. C. B., Almeida, T. S., CorrêaCosta, L. S., Fischer, P. H. C., Santos, L. H., Cafarella, E. R., da Silveira, C. H., Lamp, L. M., Fernandez-Quintero, M. L., et al. (2023). Thermostabilizing mechanisms of canonical single amino acid substitutions at a gh1 β -glucosidase probed by multiple md and computational approaches. *Proteins: Structure, Function, and Bioinformatics*, 91(2):218–236.
- Rodrigues, L. M. (2017). *Mutagraph: modelos e algoritmos para predição na afinidade de complexos proteicos através de Graph Kernel e métricas de redes complexas*. PhD thesis, Universidade Federal de Minas Gerais.
- Santana, C. A., Cerqueira, F. R., Da Silveira, C. H., Fassio, A. V., de Melo-Minardi, R. C., and Silveira, S. d. A. (2016). GREMLIN: A graph mining strategy to infer protein-ligand interaction patterns. In *2016 IEEE 16th International Conference on Bioinformatics and Bioengineering (BIBE)*, pages 28–35, Taichung, Taiwan. IEEE, IEEE.
- Santos, L. M. and Melo-Minardi, R. C. (2022). Identifying large scale conformational changes in proteins through distance maps and convolutional networks. In *Advances in Bioinformatics and Computational Biology*, pages 56–67. Springer Nature.
- Santos, V., Martins, P., and Mariano, D. (2021). *Revista Brasileira de Bioinformática*, volume 1 of 1, chapter Alinhamentos estruturais: métodos de sobreposição de proteínas e outras moléculas. Alfa-helix, Lagoa Santa, first edition.
- Senior, A. W., Evans, R., Jumper, J., Kirkpatrick, J., Sifre, L., Green, T., Qin, C., Židek, A., Nelson, A. W., Bridgland, A., Penedones, H., Petersen, S., Simonyan, K., Crossan, S., Kohli, P., Jones, D. T., Silver, D., Kavukcuoglu, K., and Hassabis, D. (2020). Improved protein structure prediction using potentials from deep learning. *Nature*, 577:706–710. DOI: 10.1038/s41586-019-1923-7.

- Silva, A. L. D. and et al (2021). From in-person to the online world: Insights into organizing events in bioinformatics. *Frontiers in Bioinformatics*, 1.
- Silva, M. F., Martins, P. M., Mariano, D. C., Santos, L. H., Pastorini, I., Pantuza, N., Nobre, C. N., and de Melo-Minardi, R. C. (2019). Proteingo: motivation, user experience, and learning of molecular interactions in biological complexes. *Entertainment Computing*, 29:31–42.
- Silveira, C., Pires, D., de Melo, R., Habesch, R., Ribeiro, C., Veloso, C., Lopes, J., Neshich, G., Meira Jr., W., and Santoro, M. (2009). Protein cutoff scanning: a comparative analysis of cutoff dependent and cutoff free methods for prospecting contacts in proteins. *Proteins*, 74(3):727–743.
- Silveira, S., de Melo-Minardi, R., Silveira, C., Santoro, M., and Meira Jr., W. (2014a). Enzymap: Exploiting protein annotation for modeling and predicting ec number changes in uniprot/swiss-prot. *PloS One*, 9(2):e89162.
- Silveira, S., Rodrigues, A., de Melo-Minardi, R., da Silveira, C., and Meira Jr., W. (2012). ADVISE: Visualizing the dynamics of enzyme annotations in uniprot/swiss-prot. In *Biological Data Visualization (BioVis), 2012 IEEE Symposium on*, pages 49–56, Seattle. IEEE, IEEE.
- Silveira, S. A., Fassio, A. V., Gonçalves-Almeida, V. M., de Lima, E. B., Barcelos, Y. T., Aburjaile, F. F., Rodrigues, L. M., Meira Jr, W., and de Melo-Minardi, R. C. (2014b). VERMONT: Visualizing mutations and their effects on protein physicochemical and topological property conservation. In *BMC Proceedings*, volume 8, page S4. BMC Proceedings. DOI: 10.1186/1753-6561-8-S2-S4.
- Webb, B. and Sali, A. (2016). Comparative protein structure modeling using modeller. *Current protocols in bioinformatics*, 54:5.6.1–5.6.37.
- Xavier, L., Bastos, L. L., and Santos, L. H. (2021). *Revista Brasileira de Bioinformática*, volume 1 of 1, chapter Modelagem computacional de proteínas. Alfahelix, Lagoa Santa, first edition.
- Yang, J., Anishchenko, I., Park, H., Peng, Z., Ovchinnikov, S., and Baker, D. (2020). Improved protein structure prediction using predicted interresidue orientations. *Proceedings of the National Academy of Sciences*, 117(3):1496–1503.
- Yang, Y., Zhang, X., Yin, Q., Fang, W., Fang, Z., Wang, X., Zhang, X., and Xiao, Y. (2015). A mechanism of glucose tolerance and stimulation of gh1 β -glucosidases. *Scientific reports*, 5(1):1–12.