

3D structural prediction, analysis and validation of Sars-Cov-2 protein molecules

Clarice de Souza, João Alfredo Bessa, Rosiane de Freitas

Optimization, Algorithms and Computational Complexity CNPq research group

Institute of Computing, Federal University of Amazonas,

Manaus-AM, Brazil

{clarice, joao.bessa, rosiane}@icomp.ufam.edu.br

and

Micael Davi de Oliveira, Jonathas da Silva, Kelson Mota de Oliveira

Laboratory of Theoretical and Computational Chemistry

Chemistry Department - ICE, Federal University of Amazonas,

Manaus-AM, Brazil

{michaeloliveira, jonathaswater, kelsonmota}@ufam.edu.br

Abstract

Knowing the structure of a protein is of enormous importance but presents great theoretical and technological challenges. The COVID-19 pandemic showed how important it was to be able to determine the structure and form of SARS-COV-2 to better understand its functioning and to be able to develop vaccines and combat drugs. This article presents mathematical-computational and physical-chemical aspects involved in the reconstruction and validation of the three dimensional molecular conformation of SARS-CoV-2 virus proteins, including the variant discovered in patients from Brazil in 2021, the lineage *B.1.1.28/P.1*. The methodology used is based on the sequencing of the virus protein through the incorporation of new *in silico* mutations in already known structures, the result is then submitted to computational reconstruction using an enumerative feasibility algorithm validated by the Ramachandran diagram and alignment structural. After the structural reconstruction of the virus, a stability study is performed with the protein generated through classical molecular dynamics.

Keywords: algorithms, Branch-and-Prune method, COVID-19 pandemic, Nuclear Magnetic Resonance, protein folding, SARS-CoV-2, molecular dynamics.

1 Introduction

It is known that the spatial conformation of the structure is decisive in the function performed by proteins and it is directly linked to their three-dimensional folding, constituting the main factor that governs biological interactions [1]. Protein folding is critical in many diseases, including type II diabetes, as well as neurodegenerative diseases like Alzheimer's and Parkinson's. Despite major advances in this field, it is still not possible to accurately predict the three-dimensional structures of [2] proteins.

Thus, research in this field has essential value in the discovery of new drugs, in the understanding of metabolic processes and the mechanism of various diseases. In 1968, Cyrus Levinthal realized that although there are a large number of possible conformations to a protein, its folding to the most stable structure occurs with extreme precision and very quickly. Even today, the rules by which nature carries out the conformational search remain uncertain [2].

In the year 2020, the importance of predicting the three-dimensional structure of proteins in a faster and simpler way was evidenced with the appearance of a new virus of the *Coronaviridae* family that spread in a few weeks around the world, giving rise to a pandemic of a magnitude not seen since the Spanish flu in 1918.

The pandemic called COVID-19 (Coronavirus Disease, identified in late 2019) is caused by the SARS-CoV-2 virus, largely similar to the SARS-CoV virus [3].

Although the first cases were officially detected in December 2019, in the city of Wuhan, China, being associated with the city's seafood market, their origin is still unknown. Although the virus apparently first appeared in a bat, there was possibly an unknown intermediate host. In addition, it is not proven how and where the transmission of SARS-CoV-2 to humans occurred. Within a few weeks the disease spread to China, reaching other countries and continents, taking on the proportions of a pandemic, as enacted by the World Health Organization (WHO) on March 11, 2020 [4].

At August of 2021, it had over 206 million confirmed cases globally, with more than 4.3 million deaths and about 185 million recovered cases. In Brazil, at the end of 2021, there were already more than 20.3 million confirmed cases and almost 567 thousand deaths [5]. The first case of COVID-19 in Brazil and Latin America (identified as BR1), had a molecular diagnosis confirmed by the Adolfo Lutz Institute, on February 26, 2020 (despite studies that indicate its presence since the end of January). The complete genome of this virus was sequenced and made available only 2 days later, on February 28, 2020, by researchers from the Adolfo Lutz Institute, in conjunction with the Institute of Tropical Medicine of the Faculty of Medicine of the University of São Paulo and the University of Oxford [6].

In January of 2021, Brazilian researchers identified a novel variant of the SARS-CoV-2, called P.1 (20J/501Y.V3), during a period of unprecedented increase in hospitalizations and incidence of cases of reinfection in the city of Manaus, Amazonas state in the Northeast Region of the Brazil, where it was shown to be correlated with the emergence of a new lineage. From the genetic sequencing it was perceived to consist of multiple mutations, among the most worrying are 3 (three) that affected the RBD region of the Spike glycoprotein: (K417T, E484K and N501Y) [7]. The P.1 strain has many similarities to the discovery in the UK B.1.1.7 (VOC-202012/01). It was found that there was an increase in transmissibility by 1.4-2.2 times, in addition to the greater probability of lethality by 1.1-1.8 times [8, 9]. In the Figure 1 there is an illustration of the 4 (four) most worrying variants of SARS-CoV-2, where we present their respective impacts on transmissibility, lethality and antigenicity.

The latest Variant-Of-Concern (VOC) named Omicron (B.1.1.529) has raised awareness of possible increased transmissibility over previous variants of SARS-CoV-2. The Omicron variant was designated by the WHO (World Health Organization) on November 26, 2021. This variant has an unprecedented 32 mutations in the Spike protein. This variant was first reported in South Africa and poses serious risks for increased viral transmissibility or detection difficulties in the RT-PCR diagnosis [13]. The first experimental results indicated that existing vaccines may be less effective against the Omicron variant, but with evidence that the booster dose may be essential to minimize the impact [14]. In this way, the serum induced by the vaccine when interacting with Omicron was reduced, and therefore T-cell mediated immunity is essential to minimize the more serious effects of COVID-19. A suggestion made in the same research was the need for specific vaccines for the Omicron variant [14].

Sequencing the virus genome and reconstructing parts or completely its molecular structure provides a greater understanding of its behavior in the host organism, which helps in the development of drugs and vaccines to combat it. On the other hand, the process of reconstructing the 3D structure of complex molecules brings interesting technological and mathematical-computational challenges from a theoretical and practical point of view. Thus, this work focuses on proteins from the main strains of the SARS-CoV-2 virus, analyzing algorithms to determine and validate a three-dimensional structure of the virus.

The remainder of this article is structured as follows. In Section 2, preliminary concepts about the biological and physicochemical aspects of the application scenario are presented. In Section 3, the modeling of the problem from a mathematical-computational point of view is presented, as well as the main solution methods. In Section 4, the framework designed for protein acquisition, 3D structure prediction, and validation via Ramachandran diagram is presented. In Section 5, the results are presented and discussed. Finally, in Section 6, the final considerations are made.

In molecular biology, much research focuses on proteins and their properties. Proteins are fundamental structures of biological systems and consist of a linear chain of amino acids that determine their three-dimensional structure [15]. The spatial conformation and folding of the protein arising from the linkage of the amino acid sequence are fundamental in the function performed by the protein [16]. Hence, methods that accurately describe the spatial conformation of proteins have used computational and experimental techniques that avoid unrealistic representations of protein structure. Furthermore, despite the existence of trillions of proteins, the human being is aware of the three-dimensional structure of only a few thousand. The function performed by proteins is directly related to their three-dimensional conformation, constituting the main factor governing biological interactions [1]. Despite enormous advances in algorithms for structural prediction, such as AlphaFold2 and RoseTTAFold, unfortunately problems of a physicochemical nature persist since they usually only use a computational approach [17]. Furthermore, the Branch-and-Prune

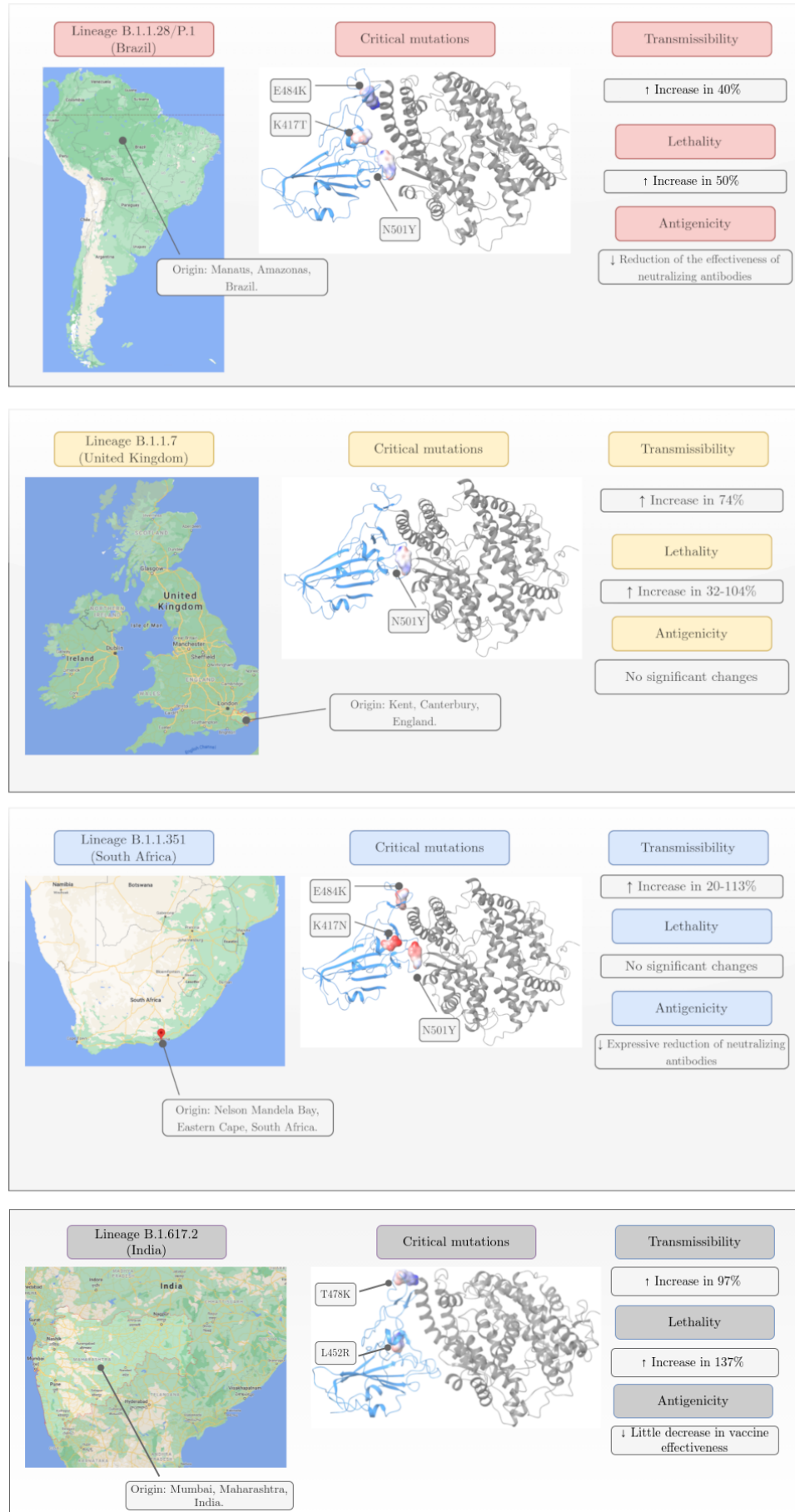


Figure 1: Illustration of the cities where the most worrying variants of the novel coronavirus emerged, as well as the positions that affected the crystallographic structure in the ACE2-RBD interaction (PDB ID: 6M0J). The transmissibility, lethality and antigenicity data for the variants were obtained from [8,10–12].

optimization approach is not so recurrent in these algorithms [17]. Inevitably the inclusion of structural minimization by molecular dynamics considerably increases the computational cost, but it certainly increases the probability of similarity to the real structure.

Protein energy landscapes are surfaces with many local minima. Thus, the exploration of these landscapes requires efficient sampling methods. Consequently, optimization methods are essential in this research field. Among the existing optimization methods, we have the molecular dynamics simulations where conformational sampling is described by Newton's laws as a potential energy function of the system. Finally, the native state is always the one with the lowest free energy and potential energy [17].

The three-dimensional structure can be obtained theoretically by Quantum and Classical Molecular Mechanics, experimentally by X-ray Crystallography, Nuclear Magnetic Resonance (NMR) and in a Hybrid way it is possible to use Homology Modeling from sequencing. In the X-ray method, the atomic coordinates are obtained with high accuracy from electron diffraction on a crystallized sample of the protein, allowing, from a rigorous mathematical refinement, the three-dimensional structure to be obtained [18]. The NMR method in turn is applied to structures that cannot be crystallized, but the NMR signal provides only a small subset of the distances between atoms in a molecule, because it only obtains distances between pairs of atoms that are close in the range of 5 to 6 Å [19].

Regardless of the structural elucidation method, the problem is to estimate the complete structure of the molecule, determining the correct position in space of all its atoms, also known as the Molecular Distance Geometry Problem (MDGP) and traditionally formulated as a continuous optimization problem [16, 19, 20]. Determining the three-dimensional structure when all distances between atoms are known is a polynomial problem, but starting from an incomplete subset of distances (as in NMR data), becomes a non-polynomial-hard problem [16]. The study of the problem under the Geometry of Distances approach can contribute decisively in identifying the most feasible protein structure, from a set of experimentally obtained distances.

2 Reconstruction of the Three-Dimensional Structure of Proteins

In this section, the main methods and tools that have been used to describe and determine the three-dimensional structure of a protein are evaluated, focusing on experimental, theoretical/computational and hybrid methods.

2.1 Molecular Modeling of Complex Molecules

Complex molecules are composed of carbon bonded to other elements, especially oxygen, hydrogen and nitrogen. They are molecules similar to polymers, like proteins, due to the immense possible variety of these polymers, they are complex [21]. The term Molecular Modeling refers to theoretical methods and computational techniques to model or mimic the behavior of molecules.

Proteins are organic compounds of high molecular weight, constructed from amino acids arranged in several specific sequences and linked together through peptide bonds. They are considered the most important macromolecules in cells and make up almost 50% of their mass [22].

Analyzes of various types of proteins show that they are all formed from a standard set of twenty amino acids, called α -amino acids. The α -amino acids are formed by a central carbon (C_α) bonded to four groups: amine group (NH_2), carboxyl group ($COOH$), hydrogen and side chain (R) which is specific for each amino acid. The Figure 2 shows the general structure of the α -amino acids [23].

In Figure 2 the different regions that constitute an amino acid are presented (the planar structure was built in the academic version of MarvinSketch 20.18). The three-dimensional structure was modeled in the software GaussView 6.0.16 with a previous optimization in Gaussian 09W at DFT level with the functional hybrid B3LYP and Gaussian bases 6-311G++ (2d,2p). The α -amino acids are joined by **peptide bond**. When the protein is being synthesized the carboxyl radical of an amino acid loses a group $-OH$ (hydroxyl) and the amine radical of another amino acid loses a hydrogen atom $-H$. The amino acids then join and the OH binds to the H , yielding a water molecule as a product, peptide bond occurs between the carbon (C) of one amino acid and the nitrogen (N) of the next amino acid, classified as a covalent bond ($C-N$). [23].

2.2 Methods Used in the Structural Determination of Proteins

Methods that accurately describe the spatial conformation of proteins have used experimental and theoretical techniques. Experimentally, the three-dimensional structure can be obtained by two methods: X-ray crystallography and Nuclear Magnetic Resonance (NMR).

In the X-ray method, the atomic coordinates are obtained with great precision from the diffraction of electrons on a crystallized sample of the protein, generally in the presence of water molecules, allowing, from a rigorous mathematical refinement, to obtain the structure three-dimensional [18].

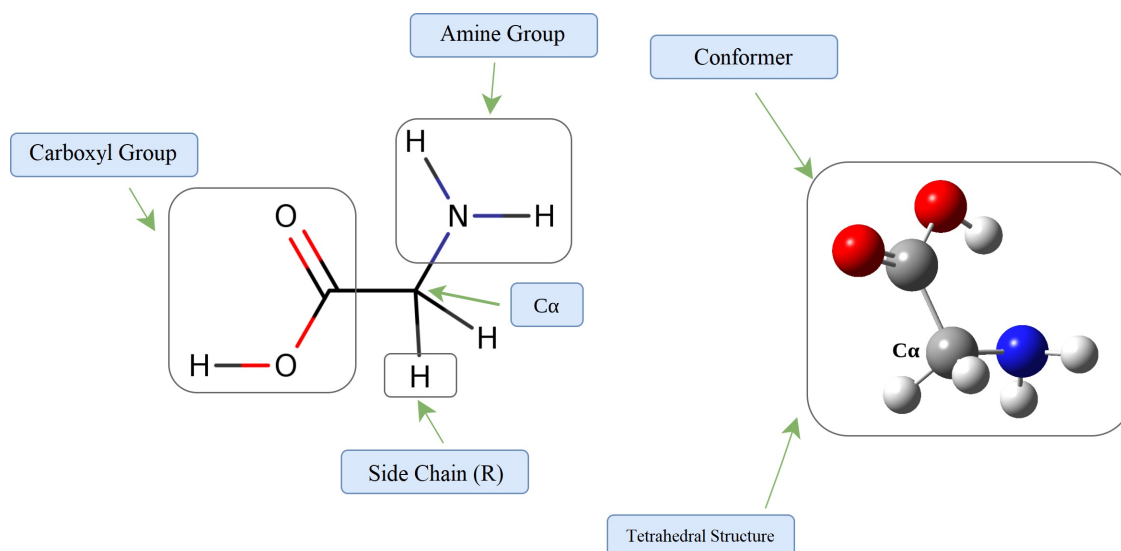


Figure 2: Structure of the amino acid Glycine (Gly), the least complex of all amino acids, containing only a single hydrogen as a side chain.

The NMR method in turn is applied to structures that cannot be crystallized. The techniques for obtaining structures are varied and rely heavily on computational aid for the construction or inference of a viable model [24, 25]. In general, the atomic coordinates between sets of pairs of atoms are modeled from the magnetic resonance coupling signal of atomic nuclei, in general of hydrogen atoms [26]. The NMR signal only gives a small subset of distances between atoms in a molecule, as they only get distances between pairs of atoms that are close together in the range of 5 to 6 Å [19].

The theoretical methods of building protein structures are very varied, in general there are two basic approaches, classical mechanics and quantum mechanics. In classical, only the interactions of classical force fields, such as Coulomb interactions, Van der Waals forces, among others, are applied to the atoms that make up the structure.

In quantum mechanics, the system is constructed taking into account the quantum properties of matter. Classical modeling is computationally cheap in terms of processing, in quantum the costs are prohibitive when it comes to high molecular mass proteins (which is common to the vast majority of proteins). Between these two methods there is a profusion of hybrid methods that mix experimental data and theoretical approaches, whether classical, quantum, mixed [27, 28] or computationally complex [29].

In any method used to build and validate the structure of a protein, the problem will be basically the same: estimating the complete structure of the molecule, determining the correct position in space of all the atoms that compose it, also known as the Molecular Distance Geometry Problem (*Molecular Distance Geometry Problem* - MDGP) and traditionally formulated as a continuous optimization problem [16].

2.3 Molecular Distance Geometry Problem - MDGP

Given a molecule formed by n atoms $a_1, a_2, \dots, a_n$ of which a set of distances d_{ij} between pairs of atoms a_i and a_j are known. The Molecular Distance Geometry Problem (MDGP) can be defined as obtaining a three-dimensional configuration $x_1, x_2, \dots, x_n$ for the molecule respecting the known Euclidean distances [16, 30].

The coordinates $x_1, x_2, \dots, x_n$ can be obtained from the distances d_{ij} by solving the system of distance calculation equations:

$$\|x_i - x_j\| = d_{ij} \quad \forall (i, j) \in S, \quad (1)$$

where S is the set of pairs of atoms whose distance d_{ij} is known, where $x_i = (v_i, w_i, z_i)^T$ a vector of coordinates with v_i , w_i and z_i being the first, second and third coordinates of the atom i and $\|\cdot\|$ being the Euclidean norm.

The structure is achieved by embedding the atoms in $\mathbb{R}^3$. Considering the position of the point as $p = (v, w, z)$, embedding a point means finding the values of the dimensions v , w and z for the point.

The problem of determining the coordinates of the atoms of a molecule from distances between pairs of atoms can be investigated by placing the atoms in any metric space respecting the defined distances [31]. In this work, the problem will be studied only in the 3D Euclidean space. Depending on the set of distances provided and the metric space, the MDGP may have one solution, several solutions or not have a valid

solution. In practice, the set of distances may contain errors, as the distances are obtained from experiments, such as NMR and crystallography, or theoretical estimates [32]. Therefore, a more practical way to define the problem is to use lower and upper limits in the set of distances, in this case the restriction of the problem is:

$$l_{ij} \leq \|x_i - x_j\| \leq u_{ij} \quad \forall (i, j) \in S. \quad (2)$$

MDGP can be classified depending on the set of distances in two forms [33,34]:

- **Complete set of distances** - all the distances between any pairs of atoms are known, Dong and Wu presented a polynomial algorithm that solves this version of the problem [15];
- **Arbitrary set of distances** - only some distances among atoms in the molecule are known, this version is NP-complete for embedding in one dimension $MDGP_1$ and NP-hard for embedding in dimensions greater than 1 $MDGP_k$ for $k > 1$ [35], [16].

There are several algorithms to solve the MDGP, the vast majority of them solve the classical continuous version of the problem, such as *Geometric Build-up* [36]. However Lavor et al. proposed a discrete model, the *Discretizable Molecular Distance Geometry Problem* (DMDGP) for the arbitrary set of distances [16,29].

The Discrete Molecular Distance Geometry Problem is defined as given a weighted simple graph not directed $G = (V, E, d)$ and an order given to the vertices $v_1, \dots, v_n$ V called *backbone ordering* that meets the following requirements:

1. E contains all 4-clicks of consecutive vertices:
 $\forall i \in \{4, \dots, n\} \quad \forall j, k \in \{i-3, \dots, i\} \quad (\{j, k\} \in E)$
2. The strict triangular inequality $\forall i \in \{4, \dots, n\} \quad d_{i-3,i-1} < d_{i-3,i-2} + d_{i-2,i-1}$ it's true,

there is a valid embedding $x : V \rightarrow \mathbb{R}^3$ such that $\|x_u - x_v\| = d_{uv}$ for all $\{u, v\} \in E$?

The first requirement demands that the lengths between atoms separated by three consecutive bonds are known. The second requirement forces the connection angles not to be multiples of π . And that means that the embedding of the v atom, denoted by x_v , is at the intersection of the three spheres S_3 , S_2 and S_1 centered on x_{v-3} , x_{v-2} and x_{v-1} of radius length equal to $d_{v-3,v}$, $d_{v-2,v}$ and $d_{v-1,v}$ as seen in Figure 3. Discretization requires that all consecutive vertex 4-clicks are subgraphs of G , each 3-subclicks $k_v^3 = \{v-3, v-2, v-1\}$ is used to test the inequality of triangles. If the two assumptions are true then each atom will have only two possible positions.

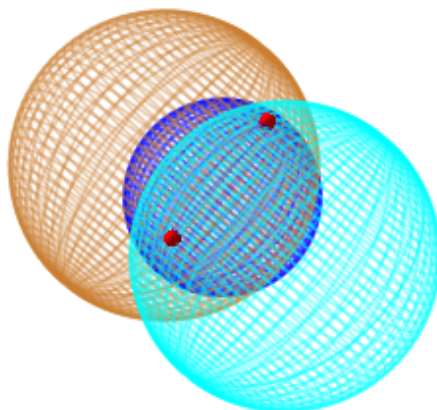


Figure 3: The figure shows the two intersecting points of three spheres, these points represent the two possibilities of positions of each vertex in the plane during the execution of the algorithm. For each atom represented by a vertex, BP tries to find its position using the known distances of it for three atoms already embedded, each possibility found opens a branch of the solution search tree.

For this version of the problem (DMDGP) the main way of solving is the *Branch and Prune* (BP) which is an implicit enumeration method, which enumerates all possible positions of the atoms and discards invalid ones. In this work, the BP algorithm was used for the structural reconstruction of the viral protein mutations.

2.3.1 Branch and Prune Algorithm

The algorithm assigns an order to the atoms of the molecule, and for each atom $v \in V$ that will be embedded in $\mathbb{R}^3$ two assumptions are made for resolution [16,37,38]:

- Valid embedding are known for all the atoms that precede v ;
- The edges $\{v-3, v\}$, $\{v-2, v\}$ and $\{v-1, v\} \in E$. The distances $d_{v-3,v-1}$, $d_{v-3,v-2}$ and $d_{v-2,v-1}$ are known, and the strict triangular inequality $d_{v-3,v-1} < d_{v-3,v-2} + d_{v-2,v-1}$ is true.

The Branch-and-Prune algorithm shown in Algorithm 1 has five input arguments:

- graph G ;
- vertex v that will be embedding in $\mathbb{R}^3$;
- subset $U \subseteq N(v)$, where $N(v)$ is the adjacent set of v with $|U| = 3$;
- a valid embedding x' for the subgraph $G[U]$;
- a set X of valid embedding of G already found.

Algorithm 1: BranchAndPrune(G, v, U, x', X)

```

1 P=intersection of the 3 spheres  $S^2(x'_u, d_{u,v})$  for  $u \in U$ 
2 forall  $p \in P$  do
3    $x = (x'_u, p)$ 
4   if  $x$  is a valid embedding then
5     if  $x$  is the last atom then
6       It adds  $x$  to  $X$ 
7     else
8        $U' = (U - \min(U)) \cup \{v\}$  BranchAndPrune( $G, v+1, U', x, X$ )

```

The process of recursion starts with a BranchAndPrune($G, 4, \{1, 2, 3\}, y, \emptyset$), where y is a valid embedding for the $\{1, 2, 3\}$ atoms. The algorithm builds a binary tree in which each level v represents possible spatial positions p for the vertex v . At the end of the run, the set X contains all valid positions of G extending x'

Pruning can be done using the *Direct Distance Feasibility (DDF)* method which considers the distances from v to the vertices of the subset $\bar{U} = \{u \in V \mid u < v \wedge u \in N(v) \wedge u \notin U\}$, that is, vertices that have known distances to v and have not been used to determine your position. If $\|x_u - x_v\| \neq d_{u,v}$ is not a valid embedding, so any subtree below it can be pruned, as shown in Figure 4

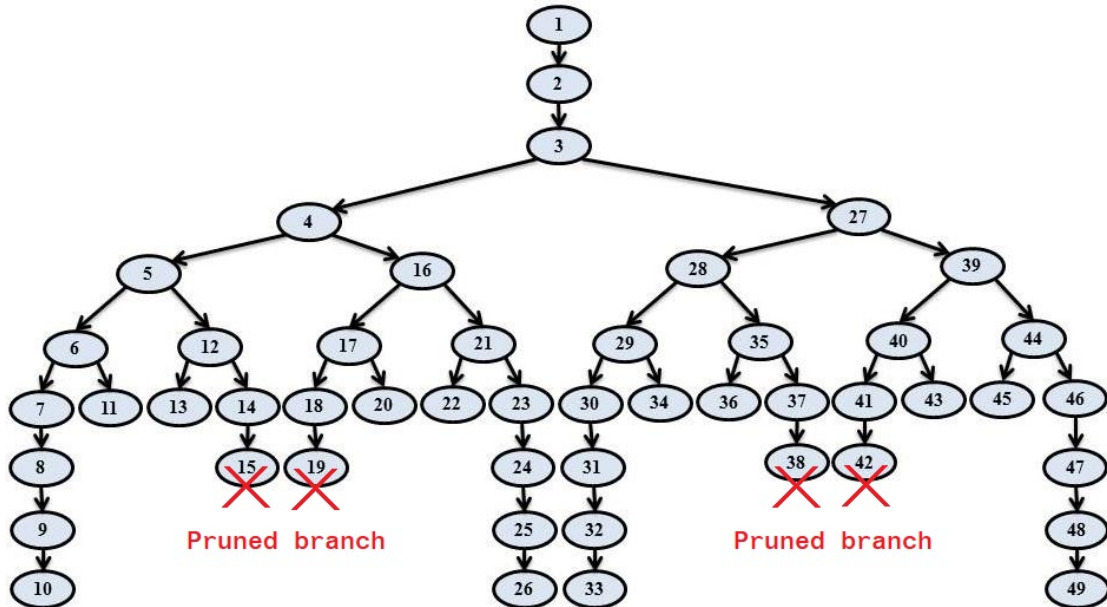


Figure 4: Binary tree resulting from the execution of the Branch-and-Prune algorithm. After fixing the first three base atoms in the Cartesian plane, for each atom of the protein the BP generates two possibilities of mathematically valid positions using the intersection of three spheres. To test whether the new branch generated is valid, the algorithm searches in the input instance for a known distance between the current vertex and an already embedded vertex that was not part of determining its position, so an pruning test is made using the DDF method, if the method invalidates the embedding, then the entire subtree is pruned.

2.4 Biochemical Validation by Ramachandra Graph

The Ramachandran plot describes the $\phi - \psi$ torsion angles of the protein backbone, providing an overview of the conformation of a protein. The $\phi - \psi$ angles are grouped into distinct regions on the Ramachandran graph, where each region corresponds to a specific secondary structure. There are four basic types of Ramachandran graphs, depending on the stereochemistry of the amino acid: generic (which refers to the 18 non-glycine, non-proline amino acids), glycine, proline, and pre-proline (which refers to residues preceding a proline).

The torsion angles ϕ (phi) and ψ (psi) are defined for each of the amino acid residues. These are angles that define rotation, with ϕ defining rotation around the $C_\alpha - N$ bond of the residue, and ψ defining rotation around the $C_\alpha - C$ bond of the same residue [39]. In principle, the dihedral angles ϕ and ψ in the amino acids can have any value between $+180^\circ$ and -180° , but several values are forbidden due to steric hindrance between the backbone atoms and the amino acid side chains, and only a few values reasonably accurately reproduce the spatial conformation of the protein, and results that prove to be forbidden or unrealistic must be discarded.

Theoretically to validate whether the conformation of proteins is the best possible, the use of the Ramachandran diagram has been found to be very useful as it tests the quality of three-dimensional structures [22]. A Ramachandran plot model where there are contour lines highlighted in green for areas allowed at dead angles, and in white as areas that cause collision is showed in Figure 5.

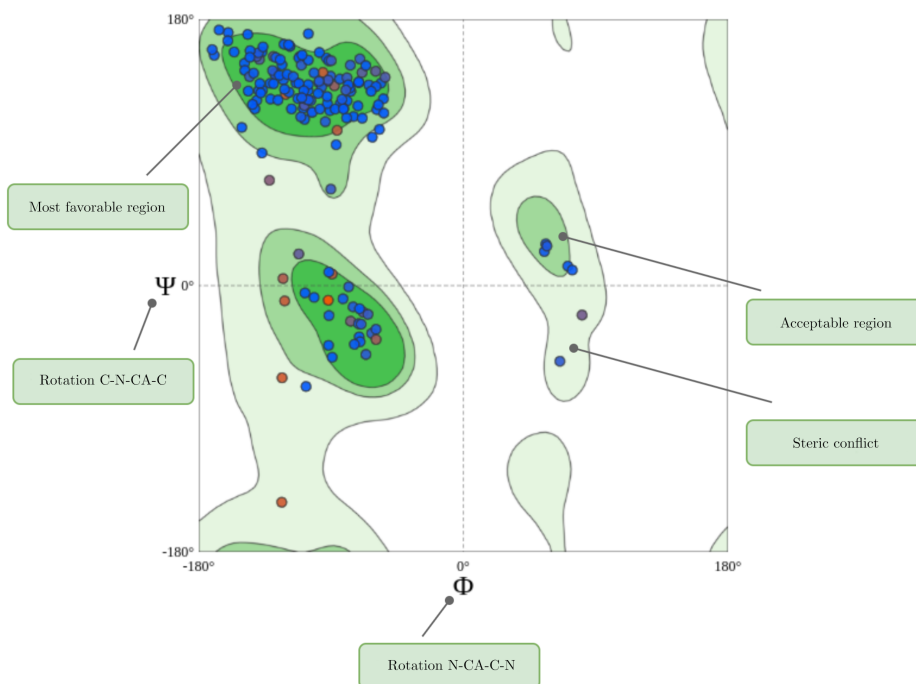


Figure 5: Regions of the Ramachandran plot having as an example the validation for an ORF8 structure (PDB ID: 7JX6) obtained with the "Structure Assessment" module on the MolProbity platform [40].

3 Methodology Proposed for the Structural Calculation of Variants

Due to the unavailability of more accurate information about the structural properties, such as the crystallographic structures that contain the mutations that make up the P.1 lineage in the state of Amazonas, the methodology shown in Figure 6 was adopted for the structural calculation of variants [41].

First, a search of the original file without mutation of the virus protein is carried out, in the RCSB protein database PDB [42]. The *in silico* mutation is then applied using the aid of the PyMol 2.3 software [43] with the "Mutagenesis" module, with the rotamer having the lowest steric tension (automatically from the software). In the third stage, instances of tests for the problem were generated by simulating an arbitrary set of distance, with distances less than 6 angstrom, thus resembling the restrictions present in the characterization by NMR. To strengthen the tests, instances of tests were made using only the atoms of the backbone and some using the backbone and side chain.

The structural reconstruction stage was subdivided into structural calculation and structural validation. The *Branch and Prune* algorithm was used for the structural calculation of virus variations, as the struc-

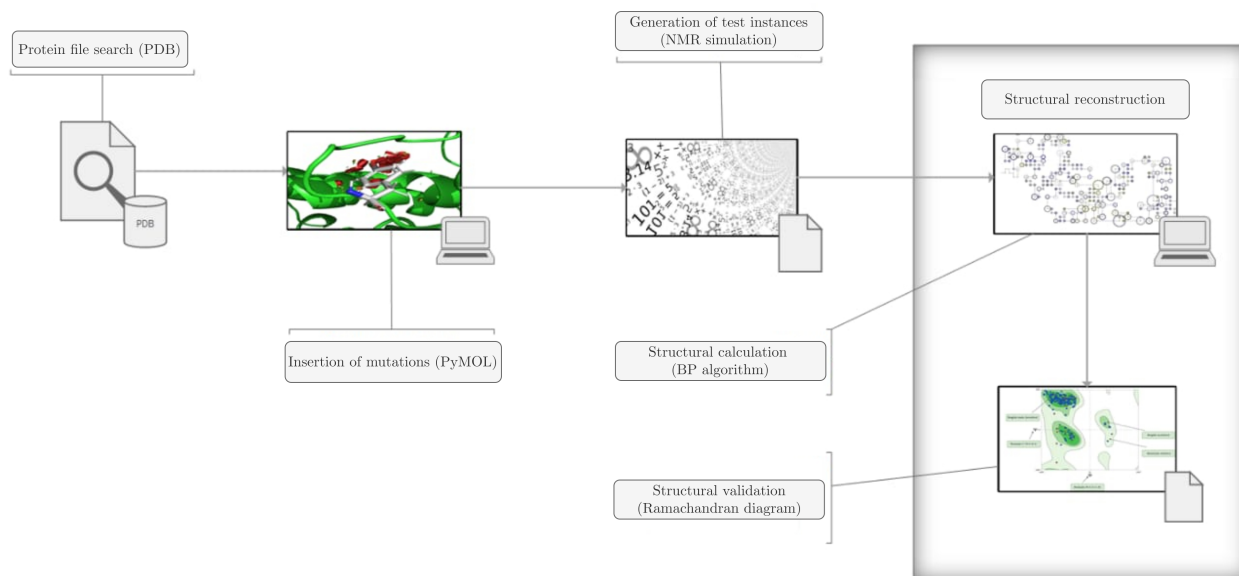


Figure 6: Flow of structural calculation of SARS-CoV-2 strain proteins.

tural prediction algorithm can generate several mathematically valid structures as a solution but does not guarantee the chemical validation of each generated structure, since it uses only as information on distances between atoms for embedding of the vertices in the plane, a chemical validation step was elaborated. For this structural validation, the Ramachandran Graph was used, which verifies the conformation of proteins (see Figures 7 and 8).

The platform adopted to generate the Ramachandran diagrams was MolProbity (<https://swissmodel.expasy.org/assess>) [44] with the “Evaluation Module” implemented on the SWISS-Model [45] by which it is possible to obtain the percentage of amino acids belonging to the favorable regions and with steric restrictions. The reconstructed structures were aligned with the one obtained through crystallography with the TM-Align <https://zhanglab.ccmb.med.umich.edu/TM-align/> [46]. The software used for the subsequent visualization was Schrödinger Maestro 2020-4 with the module ‘Superposition ’ between C- α atoms.

3.1 Molecular dynamics parameters

Molecular dynamics simulations were also performed for the ACE2-RBD complex (PDB ID: 6M0J) [47] against the B.1.1.28/P.1 (E484K, K417T, N501Y) strain in the 50 ns interval. All proteins were previously prepared with the software in its version for academic purposes Schrödinger Maestro 2020-4 using the “Protein Preparation” module. The generation of all input and configuration files was done using the QwikMD plugin 1.3 [48] implemented in the VMD graphical interface 1.9.4.48a [49]. It was immersed in a solvation box with cubic geometry and PBC (periodic boundary conditions) containing water molecules described by the TIP3P model, as well as the addition of Na^+ and Cl^- ions to neutralize the system at a molar physiological concentration of $0.15mol \cdot L^{-1}$. Meanwhile, all topology files have been generated with the CHARMM36 [50] force field. The system was minimized using 1000 steps with a conjugate gradient approach. Shortly thereafter, there was a gradual heating of 60 – 300 K under the NPT ensemble and then the system was balanced in NPT with algorithm in steepest descent over 1 ns .

Finally, the trajectory calculation an NPT thermodynamic cycle with an approximate time of 50 ns for ACE2-RBD. The Langevin dynamics stochastic piston was used to control pressure under 1 atm analogous to a barostat, while a Nosé-Hoover thermostat was also adopted to maintain the temperature at 300.0 K . Finally, the NAMD3 (Nanoscale Molecular Dynamics) algorithm [51] was used to run all simulations with acceleration of the 2 GB Nvidia GTX 1050 GPU with 640 CUDA cores in a 2 fs time step. At the end, conformational changes in the macromolecule as a result of the mutations were measured through the temporal evolution RMSD and RMSF of the atomic displacement in relation to the C α of the ACE2-RBD complex where frame 0 was adopted as a reference. The trajectory analysis was possible with the MDAnalysis [52] library.

4 Computational Results

In this section, the 3D structures generated from some experiments carried out with proteins from the new coronavirus will be shown. In order to carry out the tests, test instances were created that obey the same restrictions present in the characterization by NMR.

The SARS-CoV-2 structures chosen for testing in this work are shown in Table 1.

Table 1: Information on the structures studied in this work [47, 53–59].

Protein	Mass	Atom	Residue	Resolution	Region
6W37	7,54	525	66	2,9 Å	ORF7a
6YWK	95,51	6935	861	2,2 Å	Nsp3
6WTT	105,02	7430	906	2,15 Å	Main protease
6XS6	417,78	18294	2357	3,7 Å	Spike-ACE2
6XDC	64,33	3150	386	2,9 Å	ORF3a
7JX6	24,00	1822	192	1,61 Å	ORF8
6M0J	97,14	6571	832	2,45 Å	ACE2-RBD
7BWJ	71.63	4820	662	2,85 Å	Antigen-antibody

4.1 Study of the Amazonian Lineage B.1.1.28 of Clade P.1

All the structure possibilities of the generation tree that the BP algorithm builds were calculated and analyzed, all the generated structures respected the mathematical constraints of the problem, but some of them were considered chemically invalid when they hurt the physicochemical constraints of the proteins when the Ramachandran graph was generated.

The first test performed was for the ACE2-RBD complex instance and generated two three-dimensional structures as solutions (structure X,Y,Z), which respect the mathematical distance constraints. However, when Ramachandran graphs were generated for the two solutions, it was found that the first one was totally invalidated because it generated a graph with almost all points in wrong regions. The second however, generated a Ramachandran plot that respected the valid regions, thus being an acceptable structure, as shown in Figure 7.

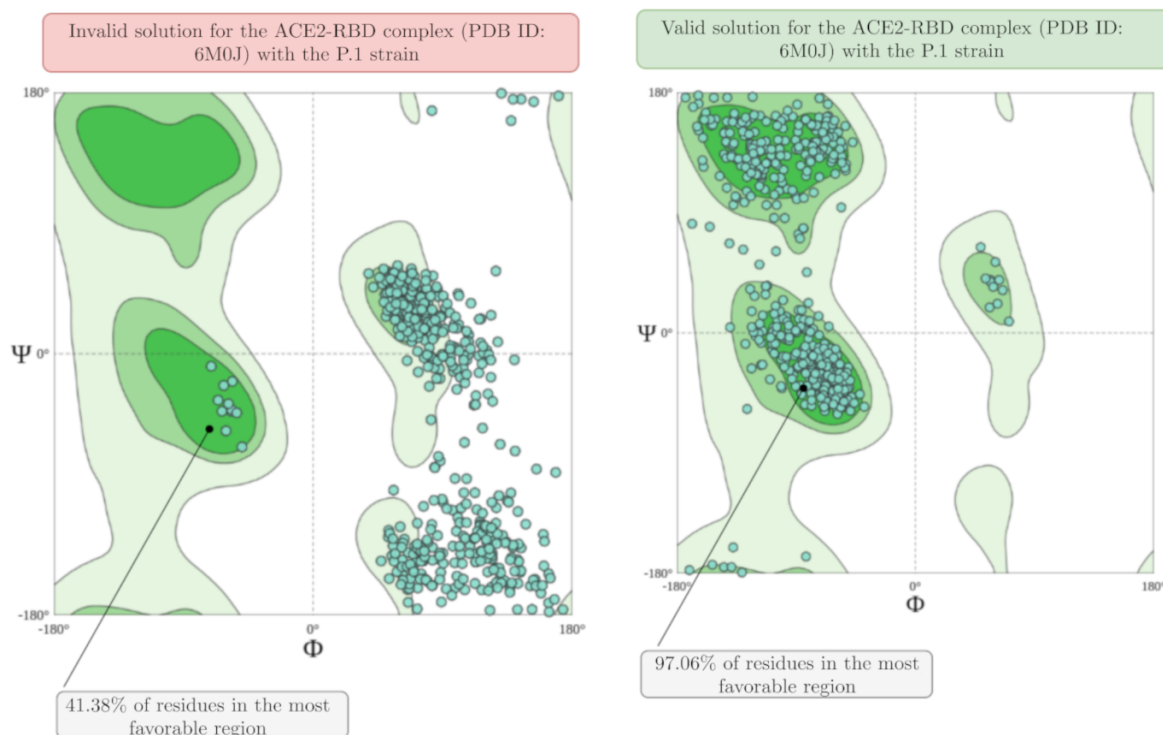


Figure 7: Ramachandran diagrams for the solutions obtained in the algorithm implemented for strain P.1. The reference structure where mutagenesis and the respective reconstruction were applied was the ACE2-RBD (PDB ID: 6M0J). All images were constructed with the aid of the MolProbity platform.

The structure for the ACE2-RBD complex (PDB ID: 6M0J) containing the mutations of the P.1 strain, considered chemically valid showed a total of 97.06% of the amino acids in the region without steric hindrance. In addition the most consistent solution showed 2.92% of the torsion angles under marginal conditions while 0.38% under conditions of total steric hindrance. Meanwhile, the reference structure for the P1 variant of the ACE2-RBD complex (PDB ID: 7NXC) presented 95.0% of the amino acids in the region without steric hindrances. Therefore an important indication of the great consistency of the developed algorithm in this work.

The test for the antibody-antigen (PDB ID: 7BWJ) followed the pattern of the first one, they generated two 3D structures as a solution, however after validation one of them proved to be inconsistent. The reconstructed structure showed 93.97% of the amino acids in a region without steric hindrance, as shown in Figure 8. Thus, even though the algorithm is not able to remove the steric conflicts induced by mutagenesis, the structural reconstruction was satisfactorily close to the crystallographic structure.

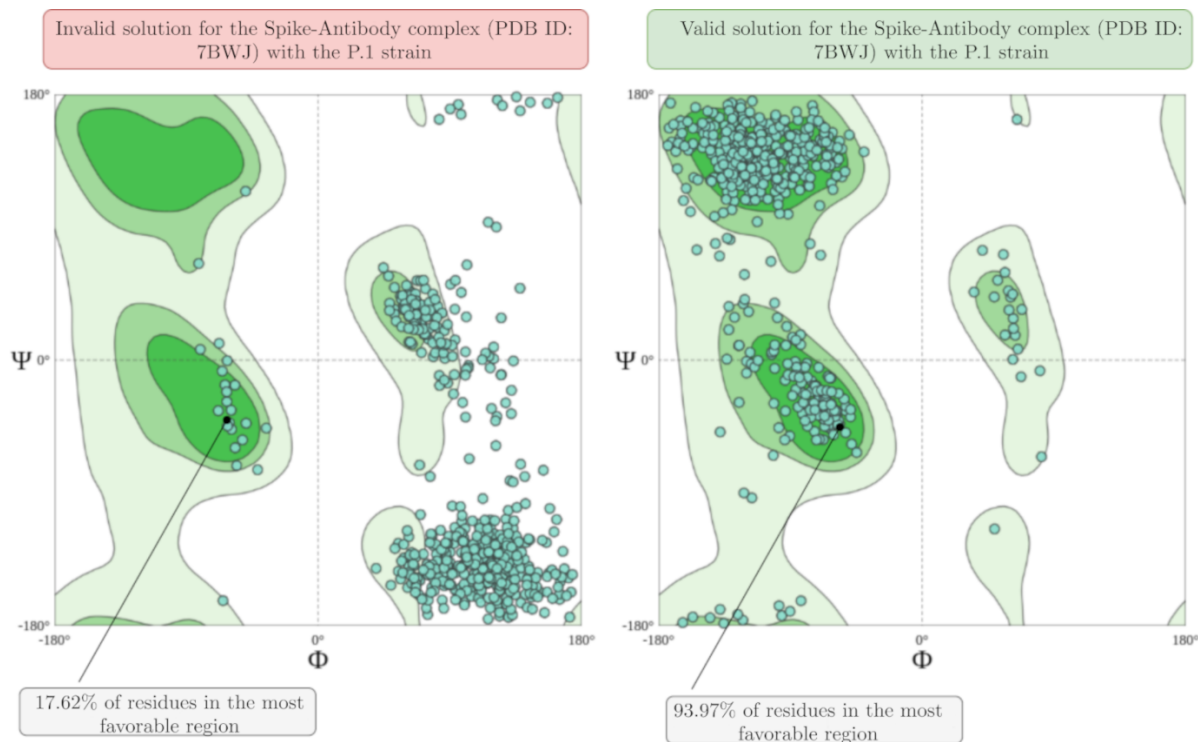


Figure 8: Ramachandran diagrams for the solutions obtained in the algorithm implemented for strain P.1. The reference structure where mutagenesis and the respective reconstruction were applied was the Spike-Antibody (PDB ID: 7BWJ). All images were constructed with the aid of the MolProbity platform.

In Figure 9 we see that the most consistent solution reconstructed by the algorithm showed an RMSD of 0.483Å for the ACE2-RBD complex containing P.1 when aligned with the PDB ID: 7NXC crystallographic structure. On the other hand, the invalid solution generated a higher RMSD of 7.49Å. We can notice that the Branch-and-Prune implementation presented the second best RMSD result from a total of 6 (six) algorithms with 0.483Å, only behind the SwissModel algorithm with 0.480Å. Nevertheless, we should note that the apparently promising results of the BP implementation may result from mere chance. This is because the insertion of mutations that constitute the P.1 variant did not reflect significant conformational changes. Therefore, as the algorithm is based on the Spike protein without mutations (PDB ID: 6M0J) it coincidentally presented promising results and very similar to the reference structure with the P.1 variant (PDB ID: 7NXC).

From the analysis of the average results of molecular dynamics (see Table 2), we noticed lower values of RMSF and lower exposure to the solvent measured by SASA and large formation of Hydrogen bonds. Therefore, in general, it can be seen that mutations of the P.1 variant have stabilized the ACE2-RBD structure, although in some central residues an increase has been noted structural flexibility according to the RMSF analysis. While only 2 (two) analyzes reflect the hypothesis of greater instability, these being: Higher radius of gyration (Rg) and lower native contacts and relatively higher RMSD values.

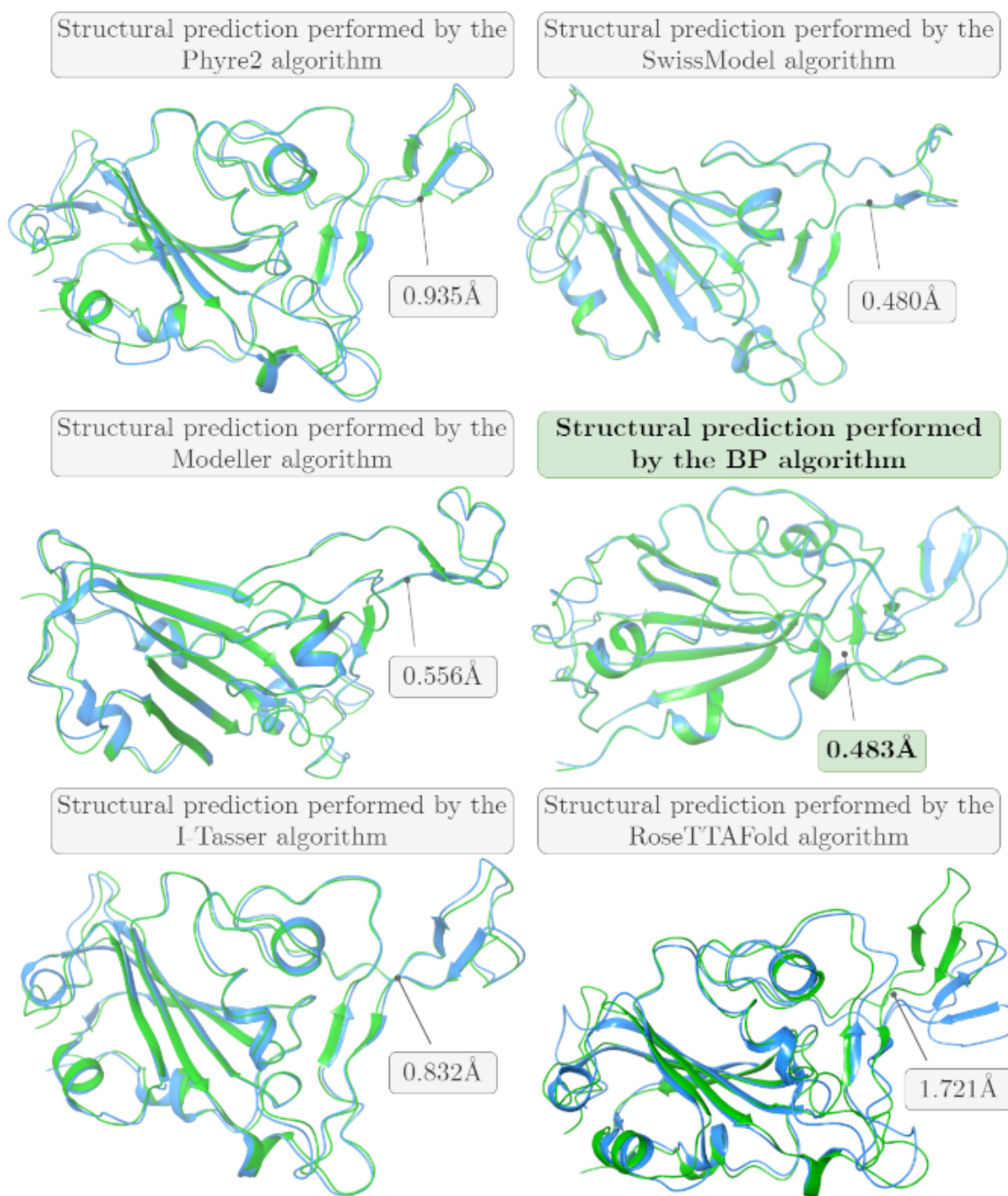


Figure 9: Benchmarking where different algorithms are compared regarding the structural alignment between the prediction performed for the ACE2-RBD complex and the crystallography containing the P.1 variant (PDB ID: 7NXC). The visualization and RMSD values for the alignment were calculated in the Schrödinger Maestro 2021-2 software. In green color, we have the reference structure (PDB ID: 7NXC). While in blue color, we have the structure predicted through algorithm.

The greater stabilization in the ACE2-RBD complex as a result of certain mutations may be an explanation of why the virus has followed a convergent evolution as already reported in some experimental studies [60–63], although the causes until then are unknown. Through the simulations of this work, we conclude that there is a tendency towards greater structural stability in the most frequent mutations. In other words, mutations that have been repeated in several strains, such as E484K and N501Y, tend to have greater thermodynamic stability. Finally, we must remember that the impacts of lineage P.1 will become clearer as we repeat the simulations or increase the time interval, which would generate more conclusive results, convergence of fluctuations and greater reproducibility.

Table 2: Comparison between the P.1 variant in relation to the average values of some parameters resulting from the molecular dynamics in the range of 50ns for ACE2-RBD (PDB ID: 6M0J) that quantify structural changes in the Spike RBD region in the interaction with ACE2. All simulations were performed in NAMD3 algorithm [51].

Analysis	Without mutations	P.1 lineage
Average RMSD	$(1.97 \pm 0.36)\text{\AA}$	$(2.01 \pm 0.29)\text{\AA}$
Average RMSF	$(1.28 \pm 0.51)\text{\AA}$	$(1.13 \pm 0.46)\text{\AA}$
$H \cdots$ bonds	197 ± 11	201 ± 11
Average SASA	$(380.41 \pm 3.47)\text{nm}^2$	$(377.64 \pm 3.66)\text{nm}^2$
Native contacts	0.9903 ± 0.0017	0.9892 ± 0.0016
Average Rg	$(31.557 \pm 0.296)\text{\AA}$	$(31.573 \pm 0.183)\text{\AA}$

4.2 Comparative Analysis Between Different Methods of Structural Minimization Based on Structural Reconstruction of the P.1 Variant

The structural reconstruction of the P.1 variant was very consistent as the RMSD in alignment with crystallography was only 0.483 Å (see Figure 9). After the structural reconstruction using BP algorithm, we performed a minimization using a conjugate gradient, the potential energy (see Table 3) was -33060.39 kcal/mol while 97.4% of the amino acids were in the most favorable region of the Ramachandran diagram.

Table 3: Comparison among different structural minimization methods for the ACE2-RBD complex. In all softwares, the best solution of the BP algorithm was used as input for minimization.

Minimization Methods	Potential energy (kcal/mol)	Favourable residues (%)
Without minimization	-25768.23	97.1%
Conjugate gradient (Prime)	-33060.39	97.4%
Simulated annealing (Desmond)	-23739.09	72.7%
LBFGS (Desmond)	-32523.4	94.1%
Simulated annealing (NAMD)	-20998.0	92.9%

Therefore, by minimizing the structure, it became even more consistent than that obtained recently in the Protein Data Bank (PDB ID: 7NXC) with 95.0%.

4.3 Structural Reconstruction of Amazonian G196V and L84S Mutations

Another important prediction of mutations that affected the state of Amazonas was the folding of the ORF3a protein with the G196V variant as well as the L84S mutation belonging to ORF8.

When performing structural alignment on ORF3a, differences were relatively small with RMSD of 3.29Å. Regarding the ORF8 structure, there was a displacement with RMSD of 0.275Å compared to the variant. Although the changes may only have been due to the stochastic nature of molecular dynamics, the mutation may actually have induced subtle conformational changes seen after structural minimization.

The most likely solution for the ORF8 structure (PDB ID: 7JX6) resulted in a total of 91.07% of the residues in the most favorable region, as shown in Figure 10. The crystallographic structure showed exactly the same number of amino acids in this region, although some twist angles in the prediction are still not consistent with crystallography. The structural alignment provided an RMSD of 0 Å, indicating that the algorithm could correctly reconstruct the input structure.

Regarding the ORF3a structure (PDB ID: 6XDC), the most consistent prediction showed 95.93% of the residues in the most favored region. And as in the previous tests, the structure obtained experimentally presented the same percentage in the region without steric hindrances in the Ramachandran diagram and whose RMSD of alignment was 0 Å. The ramachandra graphs made for the structure are shown in Figure 10

4.4 Tests Performed with Other Virus Variants to Validate the Methodology

The other tests performed with the test instances followed the pattern of the previous ones. Each test generated two or more 3D structures mathematically valid solutions, however, after validation with Ramachandran Plot, one of them has always been inconsistent, thus being chemically invalid. A summary of the results is shown in Table 4.

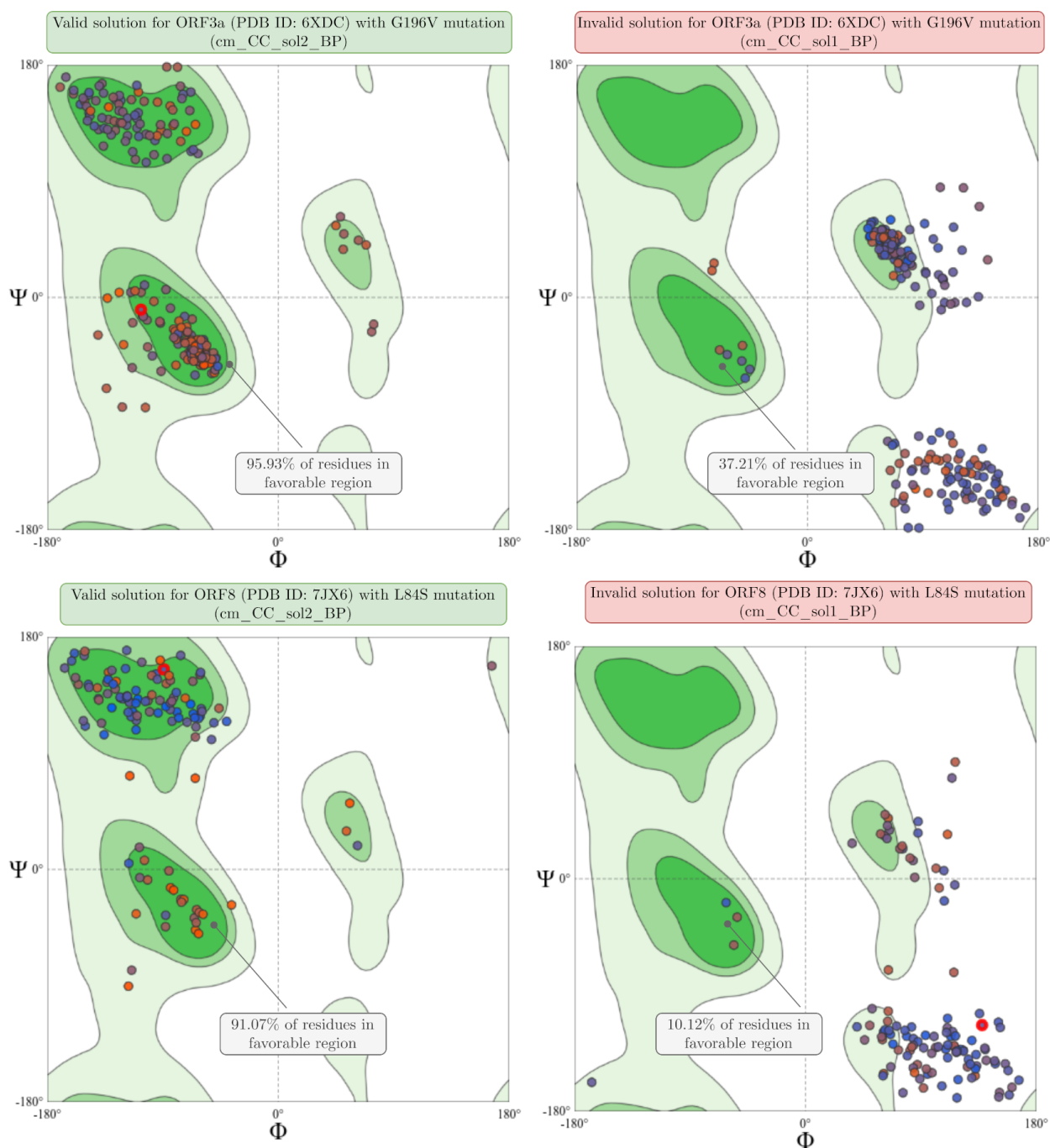


Figure 10: Ramachandran diagrams for the solutions obtained in the algorithm implemented for G196V and L84S mutations. The reference structures where mutagenesis and the respective reconstruction were applied were the ORF3a (PDB ID: 6XDC) and ORF8 (PDB ID: 7JX6) complexes. All images were constructed with the aid of the MolProbity platform.

The Ramachandra graphs presented in Figure 11 show the comparison between the crystallographic structure referring to the PDB ID: 6WTT file and the two structures reconstructed in the test. The structure considered valid obtained a percentage of residues in the favorable region equal to that of the crystallographic structure of 94.78%, while the invalid solution presented a very low percentage of 26.33%, which invalidated the structure.

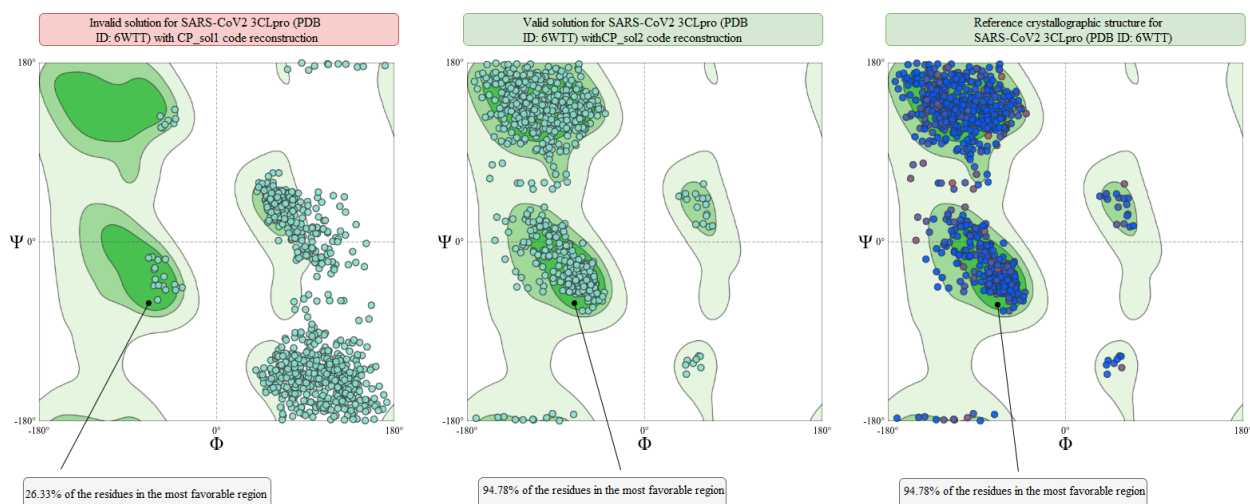


Figure 11: Ramachandran diagrams referring to the solutions provided by the algorithm for PDB ID: 6WTT. All images were constructed with the aid of the MolProbity platform.

For the structure of the variant PDB ID: 6XS6, the crystallographic structure presented a percentage of 93.88% of the residues in the favorable region. The most likely solution reconstructed by BP for this structure entailed a slightly lower percentage, a total of 93.83% of the residues in the favorable region, as shown in Figure 12. The structure invalidated by the Ramachandra plot generated only 24.32% of the residues in the correct region.

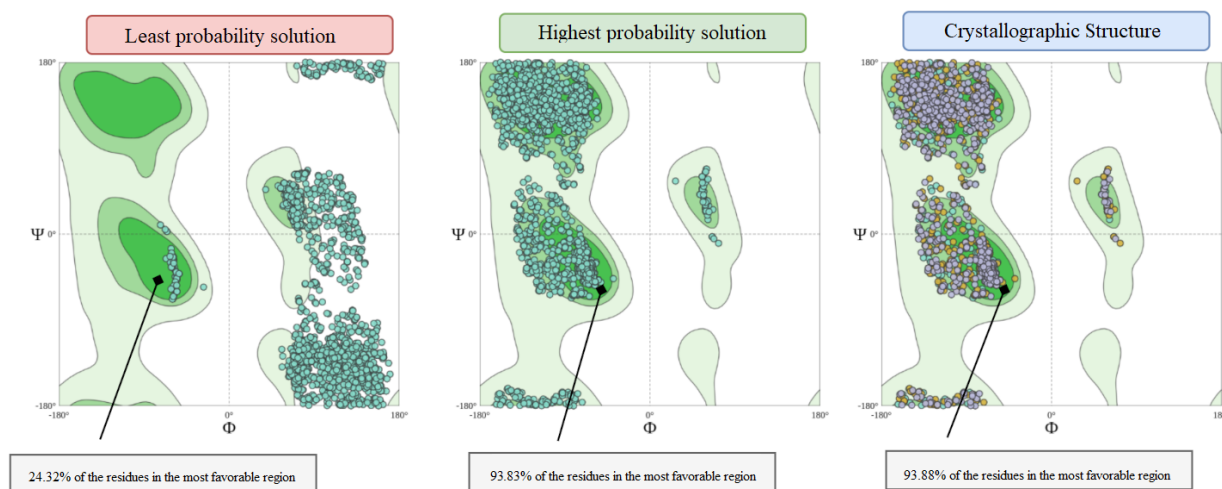


Figure 12: Ramachandran diagrams referring to the solutions provided by the algorithm for PDB ID: 6XS6. All images were constructed with the aid of the MolProbity platform.

The tests for the reference file PDB ID: 6YWK are shown in Figure 13, the solution considered chemically invalid presented a percentage of 39.37% in the valid region, while the one chemically validated by the Ramachandra graph obtained 99, 65% of waste in favorable region. The percentage of the valid structure was equal to that of the crystallographic structure, which once again showed the efficiency of the Branch-and-Prune algorithm in obtaining at least one fully valid solution, very close to the real one.

The other tests performed with the file PDB ID: 6W37 followed the pattern of the previous ones, generating two 3D structures as a solution, however after validation one of them was inconsistent, as shown in Figure 14. In the structural overlap between the valid solution for the obtained prediction and the respective crystallographic structure, an RMSD tending to zero was obtained. On the other hand, the solution with the most inconsistencies in the Ramachandran diagram was precisely the one that had low alignment with an RMSD of 3.06 Å.

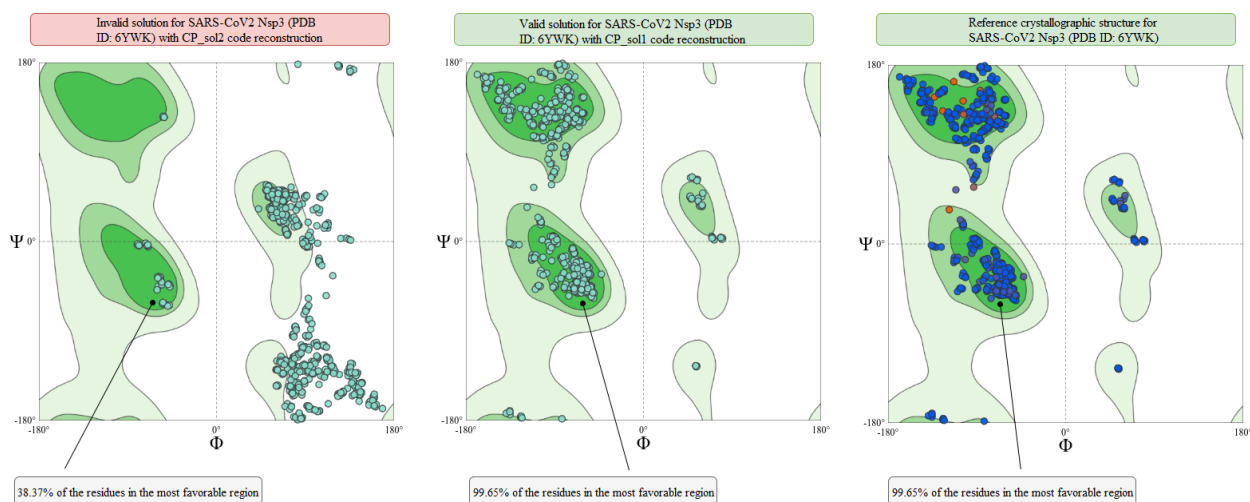


Figure 13: Ramachandran diagrams referring to the solutions provided by the algorithm for PDB ID: 6YWK. All images were constructed with the aid of the MolProbity platform.

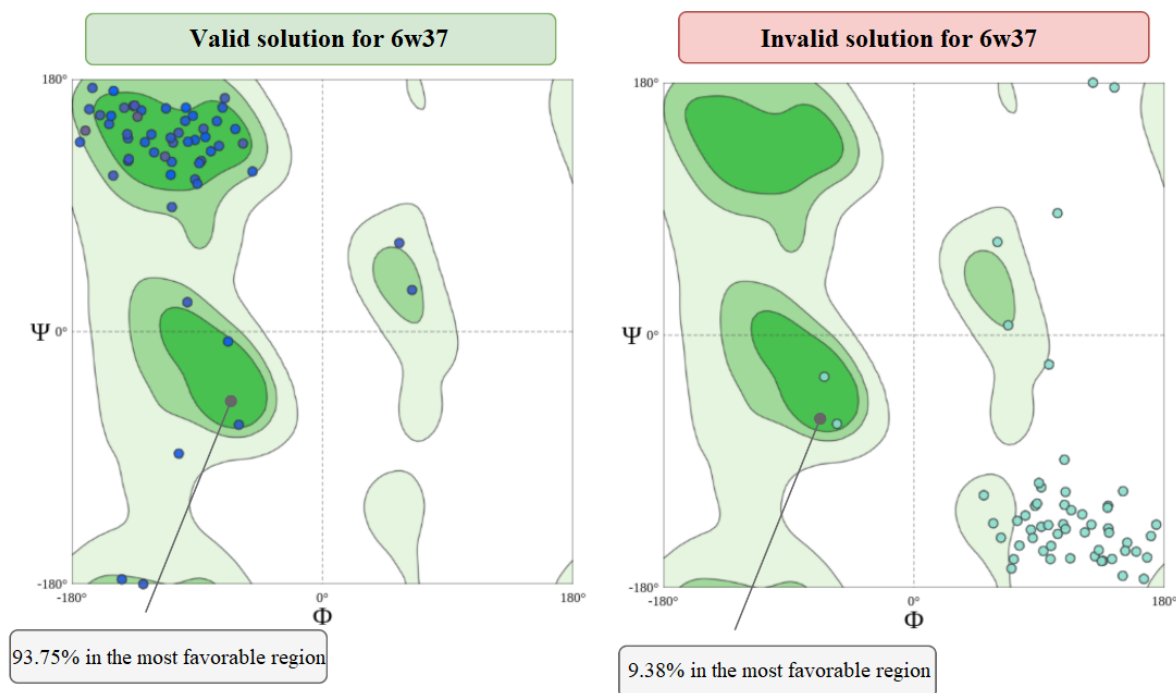


Figure 14: Ramachandran diagrams referring to the solutions provided by the algorithm for PDB ID: 6W37. All images were constructed with the aid of the MolProbity platform.

5 Concluding Remarks

The methodology developed in this work proved to be effective in reconstructing proteins of the P.1 variant of the SARS-CoV-2 virus emerged in Amazonas state, Brazil, from mutated data of the wild-type proteins. Applying the BP algorithm with Ramachandran validations, we found the great consistency of structural reconstructions based on the same restrictions of experimental characterization by NMR. Thus, it is increasingly perceived that the challenge of reconstructing proteins has been solved with the great help of computer science, as the physical-chemical principles that govern folding are not yet fully understood.

An interesting point identified when testing the instances with the backbone and the side chain was that although Branch-and-Prune was designed to find the backbone structure, it was possible to reconstruct the entire tested structure, one of its solutions being valid according to mathematical and physical-chemical

Table 4: Summaries of tests performed to validate the methodology used. The percentages of residues in the favorable regions of the Ramachandran graph are presented for both solutions validated and invalidated by the tests.

Protein		% of Residues in Favorable Region	
PDB ID	Region	Valid Solution	Invalid Solution
6W37	ORF7a	93,75%	9,38%
6YWK	Nsp3	99,65%	39,37%
6WTT	Main protease	94,78%	26,33%
6XS6	Spike-ACE2	93,83%	24,32%
6XDC	ORF3a	95,93%	37,21%
7JX6	ORF8	91,07%	10,12%
6M0J	ACE2-RBD	97,06%	41,38%
7BWJ	Antigen-antibody	93,97%	17,62%

restrictions. However, more tests are being carried out to validate this observation. Among the future perspectives, we intend to implement a heuristic that receives the sequence of amino acids as input, and that despite the high computational complexity, can predict its tertiary structure. In addition, we intend to add an objective function based on a classic force field to create physical-chemical restrictions throughout the protein reconstruction.

Acknowledgement This work is the result of a multidisciplinary project, being partially supported by the Brazilian agencies CNPq (National Council for Scientific and Technological Development), CAPES (Coordination for the Improvement of Higher Education Personnel), and FAPEAM (Foundation Support of Research the State of Amazonas).

References

- [1] Y. Levy, P. G. Wolynes, and J. N. Onuchic, “Protein topology determines binding mechanism,” *Proceedings of the National Academy of Sciences*, vol. 101, no. 2, pp. 511–516, 2004. [Online]. Available: <https://doi.org/10.1073/pnas.2534828100>
- [2] K. A. Dill and J. L. MacCallum, “The protein-folding problem, 50 years on,” *Science*, vol. 338, no. 6110, pp. 1042–1046, 2012.
- [3] P. Zhou, X.-L. Yang, X.-G. Wang, B. Hu, L. Zhang, W. Zhang *et al.*, “A pneumonia outbreak associated with a new coronavirus of probable bat origin,” *Nature*, vol. 579, no. 7798, pp. 507 – 513, 2020.
- [4] M. F. Boni, P. Lemey, and X. Jiang, “Evolutionary origins of the SARS-CoV-2 sarbecovirus lineage responsible for the COVID-19 pandemic,” *Nature Microbiology*, vol. 5, no. 11, pp. 1408–1417, 2020.
- [5] SUS. (2021) Coronavírus Brasil. Accessed in August 14, 2021. [Online]. Available: <https://covid.saude.gov.br/>
- [6] FMUSP, “Genoma do SARS-CoV-2 do primeiro caso de COVID-19 da América Latina sequenciado em 48 horas no Instituto Adolfo Lutz,” 2020. [Online]. Available: <https://www.fm.usp.br/fmusp/noticias/-genoma-do-sars-cov-2-do-primeiro-caso-de-covid-19-da-america-latina-sequenciado-em-48-horas-no-instituto-adolfo-lutz>
- [7] F. Naveca, V. Nascimento, V. Souza, A. Corado, F. Nascimento, G. Silva, Ágatha Costa *et al.*, “Phylogenetic relationship of SARS-CoV-2 sequences from Amazonas with emerging Brazilian variants harboring mutations E484K and N501Y in the Spike protein,” 2021. [Online]. Available: <https://virological.org/t/phylogenetic-relationship-of-sars-cov-2-sequences-from-amazonas-with-emerging-brazilian-variants-harboring-mutations-e484k-and-n501y-in-the-spike-protein/585>

- [8] N. R. Faria, T. A. Mellan, C. Whittaker, I. M. Claro, D. d. S. Candido, S. Mishra, Crispim *et al.*, “Genomics and epidemiology of the p.1 sars-cov-2 lineage in manaus, brazil,” *Science*, 2021.
- [9] R. M. Coutinho, F. M. D. Marquitti, L. S. Ferreira, M. E. Borges, R. L. P. da Silva, O. Canton, T. P. Portella, S. P. Lyra, C. Franco, A. A. M. da Silva, R. A. Kraenkel, M. A. d. S. M. Veras, and P. I. Prado, “Model-based estimation of transmissibility and reinfection of sars-cov-2 p.1 variant,” *medRxiv*, 2021.
- [10] E. Volz, S. Mishra, M. Chand, J. C. Barrett, R. Johnson, L. Geidelberg *et al.*, “Transmission of sars-cov-2 lineage b.1.1.7 in england: Insights from linking epidemiological and genetic data,” *medRxiv*, 2021.
- [11] D. Planas, T. Bruel, L. Grzelak, F. Guivel-Benhassine, I. Staropoli, F. Porrot *et al.*, “Sensitivity of infectious sars-cov-2 b.1.1.7 and b.1.351 variants to neutralizing antibodies,” *Nature Medicine*, 03 2021.
- [12] D. N. Fisman and A. R. Tuite, “Progressive increase in virulence of novel sars-cov-2 variants in ontario, canada,” *medRxiv*, 2021.
- [13] “Sars-cov-2 variants of concern and variants under investigation: Technical briefing 29,” accessed in 06 December, 2021. [Online]. Available: https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/1036501/Technical_Briefing_29_published_26_November_2021.pdf
- [14] A. Wilhelm, M. Widera, K. Grikscheit, T. Toptan, B. Schenk, C. Pallas, M. Metzler, N. Kohmer, S. Hoehl, F. A. Helfritz, T. Wolf, U. Goetsch, and S. Ciesek, “Reduced neutralization of sars-cov-2 omicron variant by vaccine sera and monoclonal antibodies,” *medRxiv*, 2021.
- [15] Q. Dong and Z. Wu, “A linear-time algorithm for solving the molecular distance geometry problem with exact inter-atomic distances,” *Journal of Global Optimization*, vol. 22, pp. 365–375, 2002.
- [16] C. Lavor, L. Liberti, N. Maculan, and A. Mucherino, “The discretizable molecular distance geometry problem,” *Computational Optimization and Applications*, 2011.
- [17] B. Kuhlman and P. Bradley, “Advances in protein structure prediction and design,” *Nature Reviews Molecular Cell Biology*, vol. 20, no. 11, pp. 681–697, 2019.
- [18] J. Newman, “A review of techniques for maximizing diffraction from a protein crystal *in stilla*,” *Acta Crystallographica Section D*, vol. 62, no. 1, pp. 27–31, 2006.
- [19] L. Liberti, C. Lavor, and N. Maculan, “A branch-and-prune algorithm for the molecular distance geometry problem,” *International Transactions in Operational Research*, vol. 15, pp. 1–17, 2008.
- [20] A. Mucherino, C. Lavor, L. Liberti, and N. Maculan, *Distance Geometry: Theory, Methods and Applications*. Springer Science & Business Media, 2013.
- [21] K. Kobayashi, “Complex organic molecules,” in *Encyclopedia of Astrobiology*, M. Gargaud, R. Amils, J. C. Quintanilla, H. J. J. Cleaves, W. M. Irvine, D. L. Pinti, and M. Viso, Eds. Berlin, Heidelberg: Springer Berlin Heidelberg, 2011, pp. 352–352. [Online]. Available: https://doi.org/10.1007/978-3-642-11274-4_337
- [22] D. L. Nelson and M. M. Cox, *Lehninger Principles of Biochemistry*, 7th ed. New York: Macmillan Higher Education, 2017.

- [23] D. VOET, J. G. VOET, and C. W. Pratt, *Fundamentals of Biochemistry*, USA, 2006.
- [24] J. A. Purslow, B. Khatiwada, M. J. Bayro, and V. Venditti, “NMR Methods for Structural Characterization of Protein-Protein Complexes,” *Frontiers in Molecular Biosciences*, vol. 7, p. 9, 2020.
- [25] P. Guerry and T. Herrmann, “Advances in automated nmr protein structure determination,” *Quarterly Reviews of Biophysics*, vol. 44, no. 3, p. 257–309, 2011.
- [26] A. Cavalli, X. Salvatella, C. M. Dobson, and M. Vendruscolo, “Protein structure determination from nmr chemical shifts,” *Proceedings of the National Academy of Sciences*, vol. 104, no. 23, pp. 9615–9620, 2007.
- [27] K.-C. Chow and C.-T. Zhang, “Prediction of protein structural classes,” *Critical Reviews in Biochemistry and Molecular Biology*, vol. 30, no. 4, pp. 275–349, 2008.
- [28] M. Elstner, T. Frauenheim, and S. Suhai, “An approximate DFT method for QM/MM simulations of biological structures and processes,” *Journal of Molecular Structure: THEOCHEM*, vol. 632, no. 1, pp. 29–41, 2003.
- [29] C. Lavor, L. Liberti, N. Maculan, and A. Mucherino, “Recent advances on the discretizable molecular distance geometry problem,” *European Journal of Operational Research*, 2011.
- [30] C. de Souza and R. de Freitas, “Análise do problema de geometria de distâncias moleculares,” *2013 XLV Simpósio Brasileiro de Pesquisa Operacional (XLV SBPO)*, 2013.
- [31] L. L. Jorge Alencar, Carlile Lavor, “Realizing euclidean distance matrices by sphere intersection,” *Discrete Applied Mathematics*, 2019.
- [32] A. Sit, “Solving distance geometry problems for protein structure determination,” Ph.D. dissertation, Iowa State University, 2010.
- [33] C. de Souza and R. de Freitas, “Constraint programming algorithms to determine the position of atoms in a molecule,” *The International Conference on Operations Research (OR 2013 Conference)*, 2013.
- [34] —, “Enumeration algorithms for determining the 3d structure of proteins,” *SEIO 2013 XXXIV Congreso nacional de estadística e investigación operativa. VII Jornadas de Estadística Pública. Libro de abstracts*, 2013.
- [35] L. Liberti, C. Lavor, B. Masson, and A. Mucherino, “Polynomial cases of the discretizable molecular distance geometry problem,” 2011.
- [36] Q. Dong and Z. Wu, “A geometric build-up algorithm for solving the molecular distance geometry problem with sparse distance data,” *Journal of Global Optimization*, vol. 26, pp. 321–333, 2003.
- [37] C. de Souza, R. de Freitas, and K. Mota, “Uma proposta de adaptação do algoritmo branch-and-prune usando a interseção de quatro esferas para o problema de geometria de distâncias moleculares,” *XXXIII Congresso da Sociedade Brasileira de Computação/VII e-Science workshop (CSBC/e-Science)*, 2013.
- [38] C. de Souza, R. de Freitas, and M. Salvatierra, “Sphere intersection algorithms for molecular distance geometry problem,” *2013 XXXIX Latin American Computing Conference (CLEI)*, pp. 1–4, 2013.
- [39] J. M. Berg, J. L. Tymoczko, L. Stryer, and G. J. Gatto, *Biochemistry*, New York, 2012.

- [40] C. J. Williams, J. J. Headd, N. W. Moriarty, M. G. Prisant, L. L. Videau, L. N. Deis, V. Verma *et al.*, “Structure of mers-cov spike receptor-binding domain complexed with human receptor dpp4,” *Protein science*, vol. 27, no. 1, pp. 293–315, 2018.
- [41] C. de Souza, J. Bessa, R. de Freitas, M. Oliveira, and K. Mota, “Estratégia algorítmica para a reconstrução e validação da estrutura molecular de variantes do sars-cov-2,” *Anais do XV Brazilian e-Science Workshop*, pp. 65–72, 2021.
- [42] H. M. Berman, J. Westbrook, Z. Feng, G. Gilliland, T. N. Bhat, H. Weissig *et al.*, “The Protein Data Bank,” *Nucleic acids research*, vol. 28, no. 1, pp. 235–242, 2000. [Online]. Available: <https://doi.org/10.1093/nar/28.1.235>
- [43] Schrödinger, LLC, “The PyMOL molecular graphics system, version 2.3.0,” November 2015.
- [44] V. B. Chen, W. B. Arendall, III, J. J. Headd, D. A. Keedy, R. M. Immormino, G. J. Kapral, L. W. Murray, J. S. Richardson, and D. C. Richardson, “MolProbity: all-atom structure validation for macromolecular crystallography,” *Acta Crystallographica Section D*, vol. 66, no. 1, pp. 12–21, 2010.
- [45] A. Waterhouse, M. Bertoni, S. Bienert, G. Studer, G. Tauriello, R. Gumienny, F. T. Heer, T. A. de Beer, C. Rempfer, L. Bordoli, R. Lepore, and T. Schwede, “SWISS-MODEL: homology modelling of protein structures and complexes,” *Nucleic Acids Research*, vol. 46, no. 1, pp. 296–303, 2018.
- [46] Y. Zhang and J. Skolnick, “TM-align: a protein structure alignment algorithm based on the TM-score,” *Nucleic Acids Research*, vol. 33, no. 7, pp. 2302–2309, 2005.
- [47] J. Lan, J. Ge, J. Yu, S. Shan, H. Zhou, S. Fan *et al.*, “Structure of the SARS-CoV-2 spike receptor-binding domain bound to the ACE2 receptor,” *Nature*, vol. 581, no. 7807, pp. 215–220, 2020. [Online]. Available: <https://doi.org/10.1038/s41586-020-2180-5>
- [48] J. V. Ribeiro, R. C. Bernardi, T. Rudack, J. E. Stone, J. C. Phillips, P. L. Freddolino, and K. Schulten, “Qwikmd-integrative molecular dynamics toolkit for novices and experts,” *Scientific Reports*, vol. 6, no. 1, 2016.
- [49] W. Humphrey, A. Dalke, and K. Schulten, “Vmd: Visual molecular dynamics,” *Journal of Molecular Graphics*, vol. 14, no. 1, pp. 33–38, 1996.
- [50] J. Huang and A. D. MacKerell Jr, “Charmm36 all-atom additive protein force field: Validation based on comparison to nmr data,” *Journal of Computational Chemistry*, vol. 34, no. 25, pp. 2135–2145, 2013.
- [51] J. C. Phillips, D. J. Hardy, J. D. C. Maia, J. E. Stone, J. V. Ribeiro, R. C. Bernardi, R. Buch *et al.*, “Scalable molecular dynamics on CPU and GPU architectures with NAMD,” *The Journal of Chemical Physics*, vol. 153, no. 4, p. 044130, 2020.
- [52] N. Michaud-Agrawal, E. J. Denning, T. B. Woolf, and O. Beckstein, “MDAnalysis: A toolkit for the analysis of molecular dynamics simulations,” *Journal of Computational Chemistry*, vol. 32, no. 10, pp. 2319–2327, 2011.
- [53] C. Nelson, G. Minasov, L. Shuvalova, and D. Fremont, “Structure of the SARS-CoV-2 ORF7a encoded accessory protein,” 2020. [Online]. Available: <https://www.rcsb.org/structure/6w37>
- [54] X. Ni, M. Schroeder, V. Olieric, E. Sharpe, J. Wojdyla, M. Wang, S. Knapp, and A. Chaikuad, “Crystal structure of the sars-cov-2 orf8 protein,” 2020. [Online]. Available: <https://www.rcsb.org/structure/6ywk>

- [55] C. Ma, M. D. Sacco, B. Hurst, J. A. Townsend, Y. Hu, T. Szeto, X. Zhang, B. Tarbet, M. T. Marty, Y. Chen, and J. Wang, “Boceprevir, gc-376, and calpain inhibitors ii, xii inhibit sars-cov-2 viral replication by targeting the viral main protease,” *bioRxiv*, 2020.
- [56] L. Yurkovetskiy, X. Wang, K. E. Pascal, C. Tomkins-Tinch, T. P. Nyalile, and Y. Wang, “Structural and functional analysis of the d614g sars-cov-2 spike protein variant,” *Cell*, vol. 183, no. 3, pp. 739–751, 2020.
- [57] D. M. Kern, B. Sorum, C. M. Hoel, S. Sridharan, J. P. Remis, D. B. Toso, and S. G. Brohawn, “Cryo-em structure of the sars-cov-2 3a ion channel in lipid nanodiscs,” *bioRxiv*, 2020.
- [58] C. Nelson, P. Hall, and D. Fremont, “Crystal structure of the sars-cov-2 orf8 protein,” 2020. [Online]. Available: <https://www.rcsb.org/structure/7jx6>
- [59] B. Ju, Q. Zhang, J. Ge, R. Wang, J. Sun, X. Ge, J. Yu *et al.*, “Human neutralizing antibodies elicited by sars-cov-2 infection,” *Nature*, vol. 584, no. 7819, pp. 115–119, 2020.
- [60] T. T.-Y. Lam, N. Jia, Y.-W. Zhang, M. H.-H. Shum, J.-F. Jiang, H.-C. Zhu, Yi-Gan *et al.*, “Identifying sars-cov-2-related coronaviruses in malayan pangolins,” *Nature*, vol. 583, no. 7815, pp. 282–285, 2020.
- [61] L.-M. Bobay, A. C. O’Donnell, and H. Ochman, “Recombination events are concentrated in the spike protein region of betacoronaviruses,” *PLOS Genetics*, vol. 16, no. 12, pp. 1–14, 12 2020.
- [62] E. B. Hodcroft, D. B. Domman, D. J. Snyder, K. Y. Oguntuyo, M. Van Diest, K. H. Densmore, K. C. Schwalm, J. Femling, J. L. Carroll, R. S. Scott, M. M. Whyte, M. W. Edwards, N. C. Hull, C. G. Kevil, J. A. Vanchiere, B. Lee, D. L. Dinwiddie, V. S. Cooper, and J. P. Kamil, “Emergence in late 2020 of multiple lineages of sars-cov-2 spike protein variants affecting amino acid position 677,” *medRxiv*, 2021.
- [63] S. Kemp, W. Harvey, S. Lytras, A. Carabelli, D. Robertson, and R. Gupta, “Recurrent emergence and transmission of a sars-cov-2 spike deletion h69/v70,” *bioRxiv*, 2021.