Author

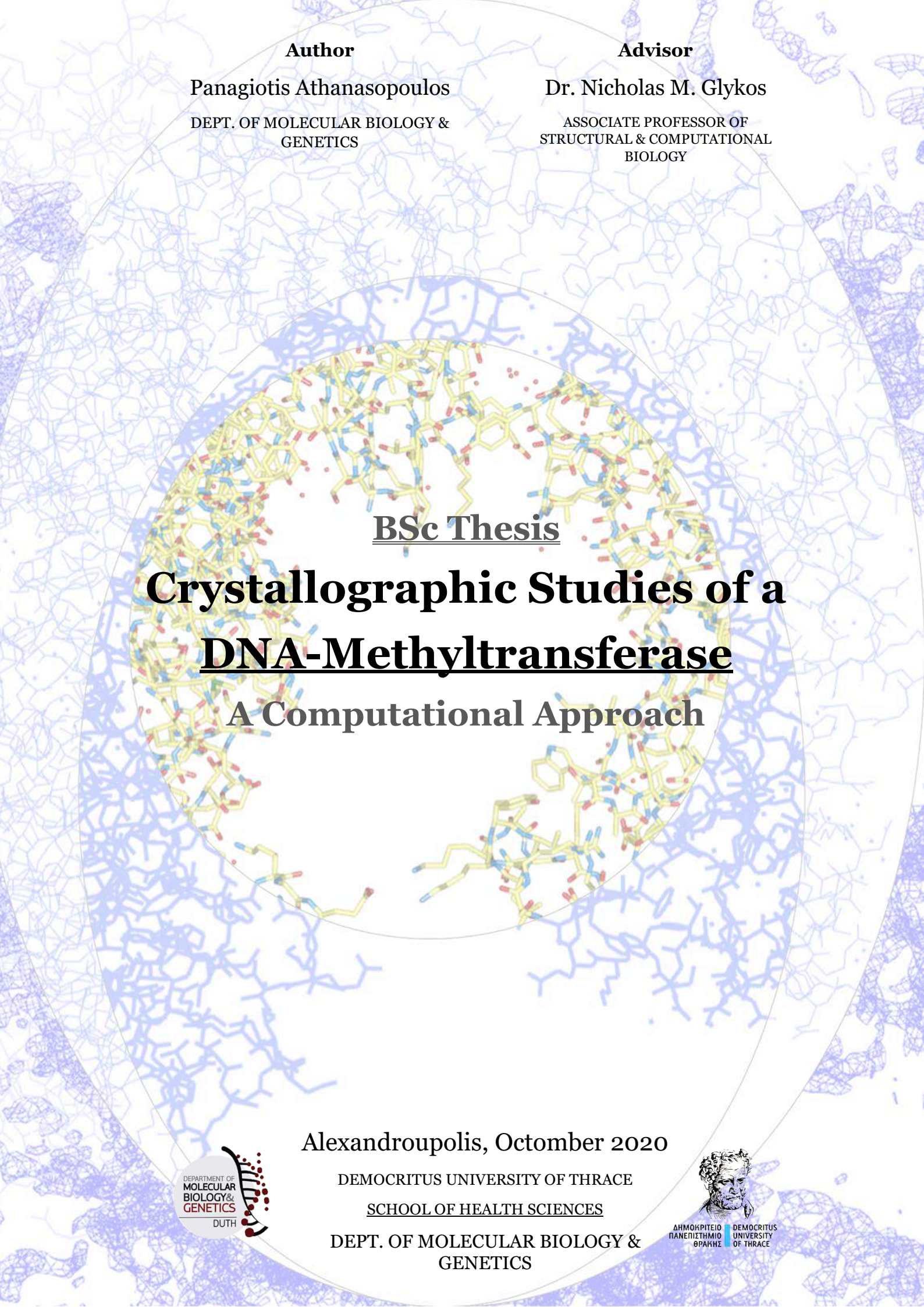
Panagiotis Athanasopoulos

DEPT. OF MOLECULAR BIOLOGY & GENETICS

Advisor

Dr. Nicholas M. Glykos

ASSOCIATE PROFESSOR OF STRUCTURAL & COMPUTATIONAL BIOLOGY



BSc Thesis

Crystallographic Studies of a DNA-Methyltransferase

A Computational Approach

Alexandroupolis, October 2020

DEMOCRITUS UNIVERSITY OF THRACE

SCHOOL OF HEALTH SCIENCES

DEPT. OF MOLECULAR BIOLOGY & GENETICS

ΔΗΜΟΚΡΙΤΕΙΟ ΠΑΝΕΠΙΣΤΗΜΙΟ ΘΡΑΚΗΣ
ΣΧΟΛΗ ΕΠΙΣΤΗΜΩΝ ΥΓΕΙΑΣ
ΤΜΗΜΑ ΜΟΡΙΑΚΗΣ ΒΙΟΛΟΓΙΑΣ & ΓΕΝΕΤΙΚΗΣ

Διπλωματική Εργασία

“ Κρυσταλλογραφικές Μελέτες μίας DNA-Μεθυλοτρανσφεράσης:
Μια Υπολογιστική Προσέγγιση”

Παναγιώτης Αθανασόπουλος
(Α.Ε.Μ. 1385)

Επιβλέπων Καθηγητής:
Δρ. Νικόλαος Μ. Γλυκός
ΑΝΑΠΛΗΡΩΤΗΣ ΚΑΘΗΓΗΤΗΣ ΔΟΜΙΚΗΣ
ΚΑΙ ΥΠΟΛΟΓΙΣΤΙΚΗΣ ΒΙΟΛΟΓΙΑΣ

ΑΛΕΞΑΝΔΡΟΥΠΟΛΗ
ΟΚΤΩΒΡΙΟΣ, 2020

Acknowledgments

The "Acknowledgments" Section is dedicated, not only to those who supported me during the preparation and completion of my bachelor thesis, but mainly to those who accompanied me through my long and extended journey of my Bachelor studies in Alexandroupolis, Greece.

First of all, I would like to thank **my supervisor and mentor** Nicholas M. Glykos. Usually a student member of the NMG group is quite grateful for our supervisor's patience. I'm *more than* thankful to him for being *more than* patient and supportive through our collaboration towards the completion of this thesis. Most importantly, I am entirely grateful to him for sharing with me his experience and academic wisdom, which I feel today as parts of my identity as a developing scientist. I feel quite honored to call him *my mentor*.

I'm also grateful to **my *academic family***, the NMG group, and more specifically, to Fillina, Giannis, Dionysia, Alex, Penny, Dora, Dimitris, and to the newest members I was able to develop a close relationship with, Magda and Evaggelos. Each one of you played a significant role to me feeling the Dpt. Of Molecular Biology and Genetics, and our field by extension, like home.

I am more than grateful to my **close friends** Eva, Alexandra, Maria, and Artemis, along with my most recent *encounters*; Magda, Nikos and Efstathia. Thank you all for being part of what I like to mention as my *chosen family*.

I would like to thank **the Athanasopoulos family**; my father Andreas, my mother Vassiliki, and my sister and brother, Eugenia and Kostantinos, for supporting me through my Bachelor years in any way they could possibly do. I'm especially grateful to my brother Kostas, for allowing me, putting aside our kinship, to feel him as part of my *chosen family* in my new life in Alexandroupolis, and also for the honor of being his *colleague*.

I am grateful to my **friends and colleagues** Despina Kiousi, Christina Katzastra, Dimitris Mitsikas, Vasso Garoufalaki, Michalis Spathakis, Aggelos Mannolias and Athina Gavanozi(s) for their support and patience, and for being close to me to the most difficult parts of that journey.

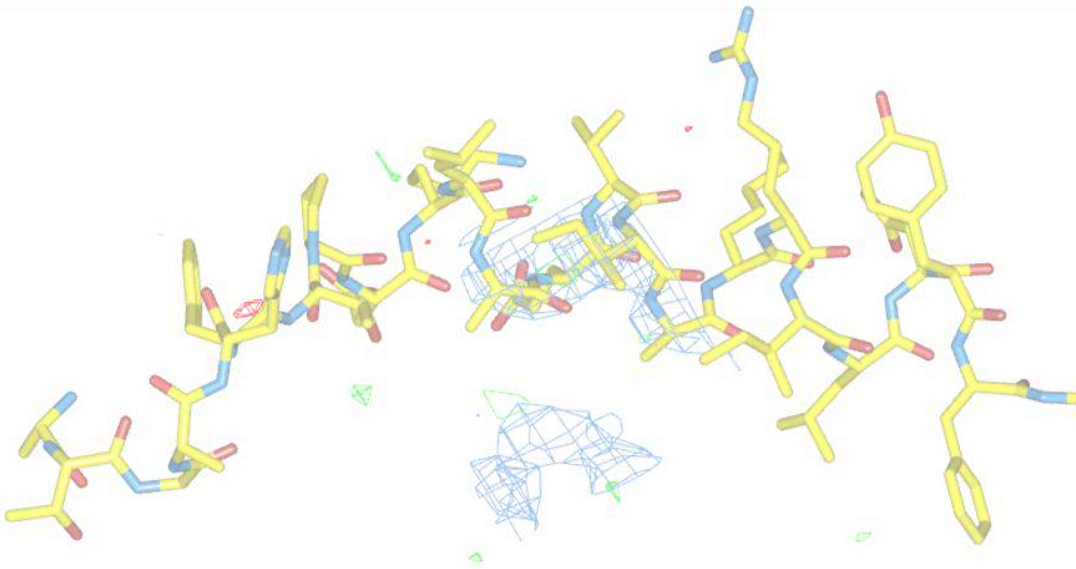
Since the COVID-19 pandemic was included in my Bachelor experience, I cannot thank enough my ***quarantine squad***; Domna, Anastasia, Elissavet, Panagiotis and Archondis, for supporting each other during this ever-changing and exasperating period.

To honor our **previous relationship**, I would like to thank Dionysis and Penny, but also Anastasia and Michalis Sv. Even though our *paths* were separated, I treasure the experiences and feelings we shared.

Last, but not least, I am grateful to three persons- ***landmarks*** to my life: Vassilis Blitsas, for being another important mentoring figure and revealing to me the importance of feeling accepted. My childhood friend, Vassilis Gekas, for reminding me each time that our relationship is *not bound from time and space*. And my dear, good friend Dora Boukoura, for being my closest companion in my self-developing journey, and for choosing me each time as a companion for hers.

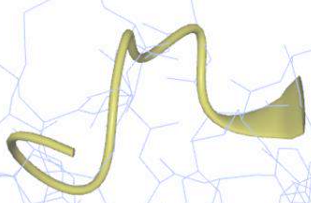
I hold a very special place in my heart for each and every one of you.

Table of Contents…



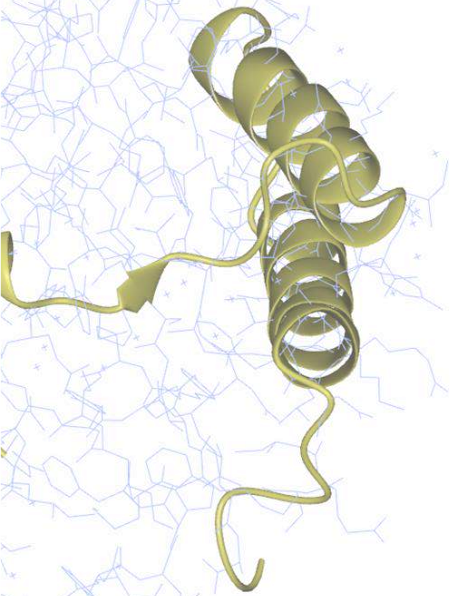
Abstract

X-ray Protein Crystallography is the most known method for the determination of a protein's three-dimensional structure. The scattering of an X-ray beam from the protein's crystallized solution, and the analysis of the diffracted waves' electromagnetic properties, reveal the exact position of the molecule's atoms. All those positions collectively comprise the *electron density map,* which can be studied and modelled through computational means. An X-ray Crystallographic experiment is usually divided in two main experimental procedures; A. the Extraction and Analysis of the X-ray data, and B. the Computational Building and Optimization of the three-dimensional Protein Model. In this thesis, the focus was the Computational Optimization of BsecI DNA-Methyltransferase 3D model, and the building of some *loop* regions, in which the electron density indications were almost non-existent. For this reason, the model *underwent* repeated *Fitting* and *Refinement* cycles, the *Simulated Annealing* trial, and the *Automated Model Validating Algorithm of the PDB-REDO Server*. This procedure provides a deeper look on how the conserved region-specific methylation mechanism of BsecI works, and its importance to the evolution of species. The thorough studies of a protein's crystallographic structure reveal important information about the molecule's function, and how the molecular world works in general.

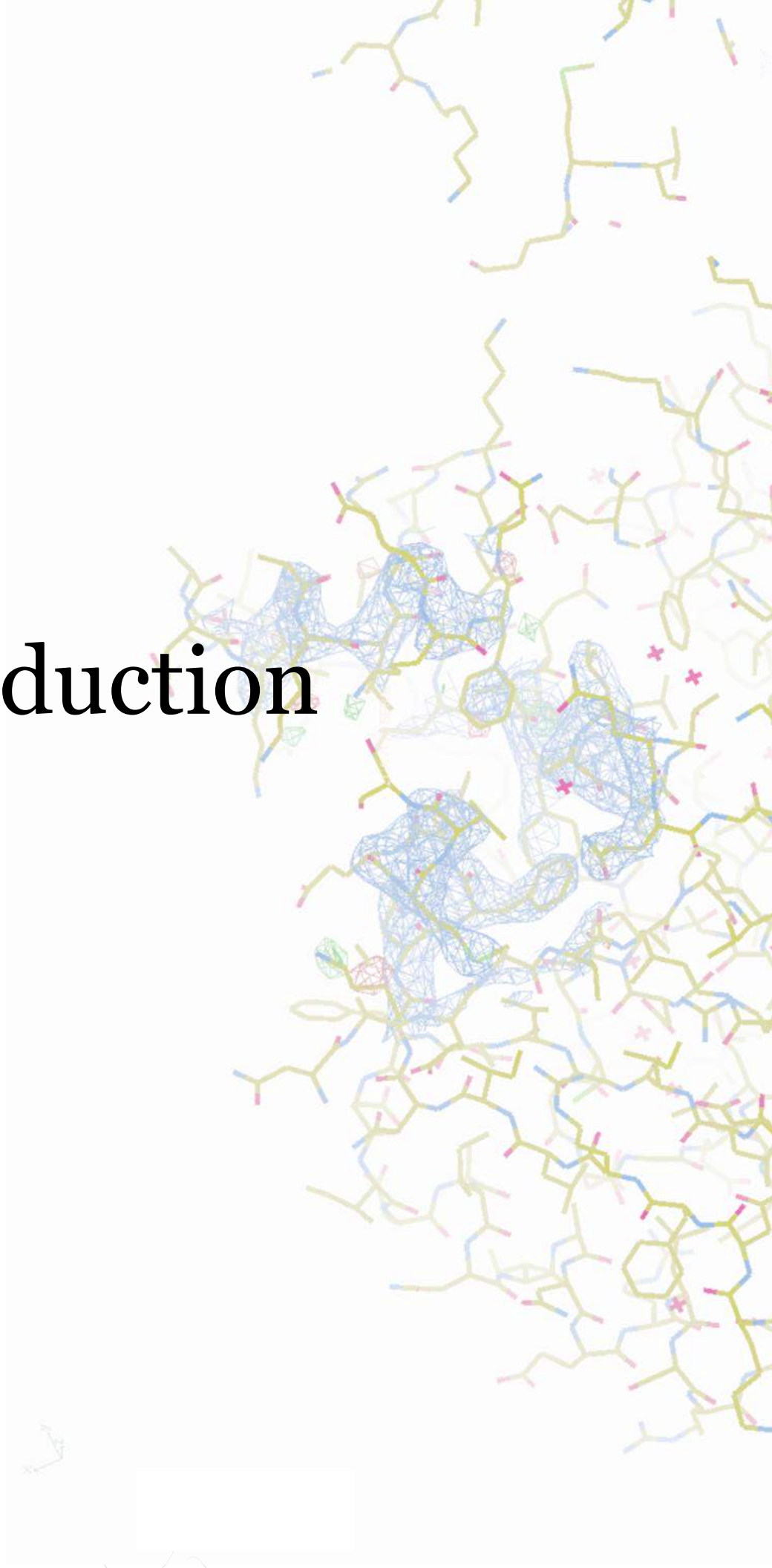


Περίληψη

Η Κρυσταλλογραφία Ακτίνων-Χ είναι η πιο διαδεδομένη τεχνική για την εύρεση της τρισδιάστατης δομής μίας πρωτεΐνης. Η περίθλαση μίας δεσμίδα ακτίνων-Χ από το κρυσταλλικό πρωτεϊνικό διάλυμα, καθώς και η ανάλυση των ηλεκτρομαγνητικών ιδιοτήτων των σκεδαζόμενων κυμμάτων, δίνει πληροφορία για την ακριβή θέση των ατόμων του μορίου στον χώρο. Το άθροισμα των διαφορετικών θέσεων όλων των ατόμων αποτελούν τον *χάρτη ηλεκτρονιακής πυκνότητας*, στον οποίο βασίζεται η μελέτη και κατασκευή του τρισδιάστατου μοντέλου, η οποία πραγματοποιείται με τη βοήθεια υπολογιστικών μέσων. Ένα Κρυσταλλογραφικό Πείραμα Ακτίνων-Χ αποτελείται από δύο βασικά μέρη: Α. την Εξαγωγή και Ανάλυση δεδομένων από την σκέδαση των Ακτίνων-Χ από τον κρύσταλλο και Β. την Υπολογιστική Κατασκευή και Βελτιστοποίηση του τρισδιάστατου μοντέλου της Πρωτεΐνης. Η παρούσα διπλωματική εργασία εστιάζεται στην Υπολογιστική Βελτιστοποίηση του 3D μοντέλου της BsecI DNA-Μεθυλοτρανσφεράσης, και στην κατασκευή κάποιων *βρόγχων* , για τους οποίους τα κρυσταλλογραφικά δεδομένα ήταν χαμηλής διακριτικότητας. Για τον λόγο αυτό, η δομή υποβλήθηκε σε επαναλαμβανόμενες δοκιμές *Fit & Refinement, σ*τη δοκιμασία του *Simulated Annealing* και στον *Αλγόριθμο Αυτοματοποιημένης Αξιολόγησης και Βελτιστοποίησης του PDB-REDO*. Μέσα από αυτήν τη διαδικασία δίνεται η δυνατότητα για περεταίρω εμβάθυνση στον μηχανισμό με τον οποίο η Μεθυλοτρανσφεράση μεθυλιώθει μία συγκεκριμένη περιοχή στη διπλή έλικα του DNA, καθώς και στη σημασία αυτού του συντηρημένου εξελικτικά μηχανισμού. Οι ενδελεχείς μελέτες της κρυσταλλογραφικής δομής μίας πρωτεΐνης αποκαλύπτει σημαντικές πληροφορίες για τη λειτουργία του μορίου και, ευρύτερα, για το *πώς δουλεύει ο κόσμος των μορίων*.



I. Introduction



Main Idea…

As you may already noticed, this thesis' title is "Crystallographic Studies of a DNA-Methyltransferase". My main goal for this project is to study and optimize the already solved three-dimensional (3D) structure of a specific DNA-Methyltransferase, using Crystallographic Computational Methods.

A Computational Crystallographic project is mainly an *empirical* procedure. Beyond a basic knowledge background on the Proteins' Structural Properties and Computational Crystallographic methods, an empirical view on the molecules' structures is mainly needed during the whole study. So, my thesis is basically an *Empirical Diary* on my journey of studying this DNA-methyltransferase's structure and becoming more and more acquainted with the *molecular world*.

Through the Introduction section I will set the basic knowledge background that was needed for this project.

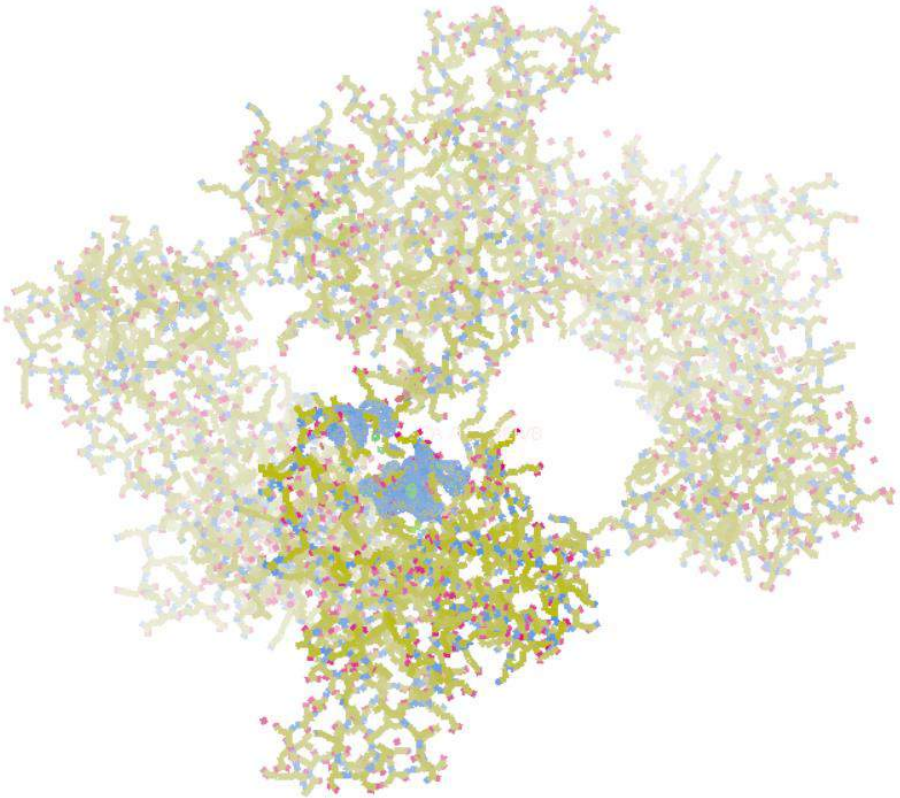


Figure 1| *A Figure I like to call "A Normal Day in Computational Crystallography Lab".*

But before all that, let us begin with some questions that may -if not already, they will- occur:

1. Proteins…

…are large biomolecules, or macromolecules, consisting of one or more long chains of amino acid residues. Proteins perform a vast array of functions within organisms, inside the "molecular world", including catalyzing metabolic reactions, DNA replication and transcription, responding to stimuli, and transporting molecules from one location to another. In other, more poetic, words, *"proteins build life"*.

A linear chain of amino acid residues is called a polypeptide. A protein contains at least one long polypeptide. Short polypeptides, containing less than 20-30 residues, are rarely considered to be proteins, and are commonly called peptides, or sometimes oligopeptides.

Proteins differ from one another primarily in their sequence of amino acids, which is dictated by the nucleotide sequence of the genes they are encoded from, and which usually results in protein folding, a specific three dimensional (3D) structure that determines a protein's activity. (Perret, 2007)

Maybe it' s already becoming a bit clear how important studying the 3D structure of a protein is. The structure is connected to the molecule's properties and function. Studying the protein structures is like *"a diver's leap into the vast and wonderful abyss of the molecular world"*.

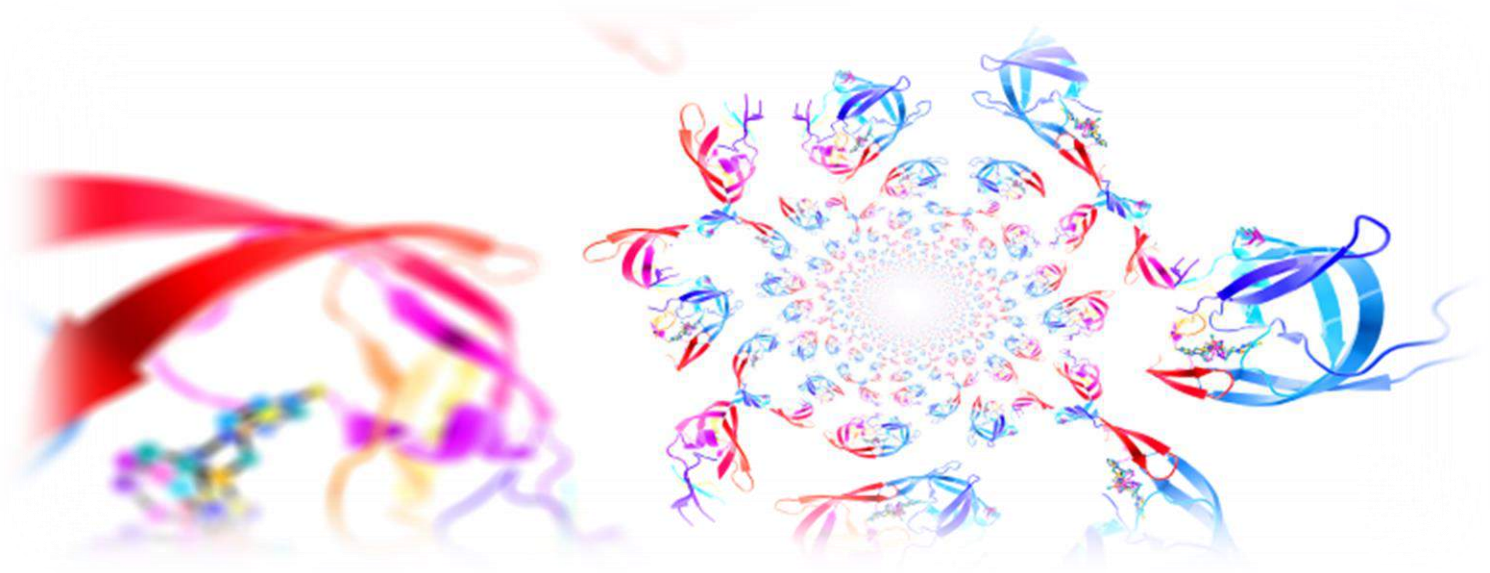


Figure 2| *"The wonderful world of Proteins" (Perrin, 2018)*

1.a. The Central Dogma of Molecular Biology

Let' s connect the basic terms of Molecular Biology to the Proteins' functions and Structural Studies. First, proteins are products of cellular processes, which begin from the genetic code. These processes are explained in Molecular Biology through the Central Dogma of Molecular Biology.

The Central Dogma is an explanation of the flow of genetic information within a biological system. It is often stated as "DNA makes RNA, and RNA makes protein" (Leavitt, 2010), although this is not its original meaning. It was first stated by Francis Crick in 1957, then published in 1958:

> *"The Central Dogma. This states that once "information" has passed into protein it cannot get out again. In more detail, the transfer of information from nucleic acid to nucleic acid, or from nucleic acid to protein may be possible, but transfer from protein to protein, or from protein to nucleic acid is impossible. Information means here the precise determination of sequence, either of bases in the nucleic acid or of amino acid residues in the protein."* (Crick, F., 1958, pp. 138–163)

and re-stated in a Nature paper:

> *"The Central Dogma of molecular biology deals with the detailed residue-by-residue transfer of sequential information. It states that such information cannot be transferred back from protein to either protein or nucleic acid."* (Crick, F., 1970)

There are not many changes in the Central Dogma since the publications of Crick. The dogma is a framework for understanding the transfer of sequence information between information-carrying biopolymers, molecules with repeating regions in their structure. These biopolymers are found in the most common or general case, in living organisms. There are 3 major classes of such biopolymers; DNA and RNA (both chemically known as nucleic acids) and Protein. There are 3 × 3 = 9 conceivable direct transfers of information that can occur between these. The dogma classes these into 3 groups of 3; Three general transfers -believed to occur normally in most cells- , three special transfers -known to occur, but only under specific conditions in case of some viruses or in a laboratory- and three unknown transfers -believed never to occur. (**Figure 3**)

The general transfers describe the normal flow of biological information; DNA can be copied to DNA, also known as DNA replication. DNA information can be copied into the types of RNA, also known as transcription. And Proteins can be synthesized using the information in mRNA, one of the RNA types, as a template, also known as the process of translation.

The special transfers describe; RNA being copied from RNA, also known as RNA replication. DNA being synthesized using an RNA template, also known as reverse

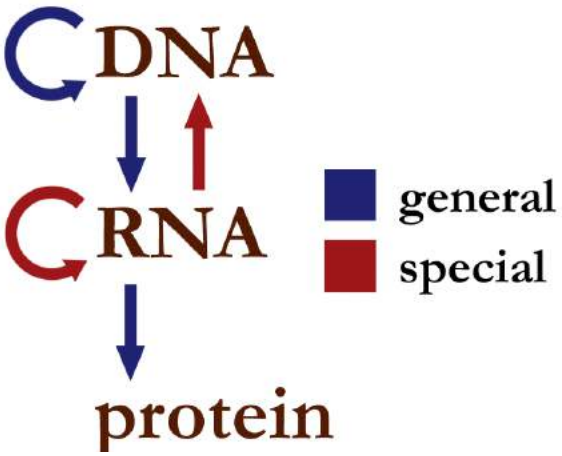


Figure 3| *Information flow in biological systems.*

transcription. And Proteins being synthesized directly from a DNA template without the use of mRNA.

The unknown transfers describe; A Protein being copied from a Protein. Synthesis of RNA using the primary structure (See **Figure 3**) of a protein as a template. And DNA synthesis using the primary structure of a protein as a template. These cases are not thought to occur naturally. (Crick, 1970)

The biopolymers that comprise DNA, RNA and (poly)peptides are linear polymers -i.e.: each monomer is connected to at most two other monomers. The sequence of their monomers effectively encodes information. The monomers of a protein are amino acids.

Regarding this project, it is quite important to have both a knowledge base, and an empirical one, built around the properties of amino acids.

1.b. Amino acids- "The building blocks of Proteins"

The amino acids are the *building blocks* of a protein. They are organic compounds that contain amine (-NH2) and carboxyl (-COOH) functional groups, along with a side chain (*R* group) specific to each amino acid (Nelson, D. L., Cox, M. M., 2005). The key elements of an amino acid are Carbon (C), Hydrogen (H), Oxygen (O) and Nitrogen (N), although other elements are found in the side chains of certain amino acids. The generic formula H2NCHRCOOH in most cases, where R is an organic substituent known as a "side chain" (Clark, 2007) (**Figure 4.A**). About 500 naturally occurring amino acids are known -though only 20 emerge from the transcription and translation of the genetic code- and can be classified in many ways (Wagner, Musso, 1983).

They can be classified according to the core structural functional group's locations as alpha- (α-), beta- (β-), gamma- (γ-) or delta- (δ-) amino acids; other categories relate to polarity, pH level and side chain group type -aliphatic, acyclic, aromatic, containing hydroxyl or sulfur, etc. Beyond their role as residues in proteins, amino acids also participate in a number of processes such as neurotransmitter transport and biosynthesis.

In biochemistry, amino acids having both the amine and the carboxylic acid groups attached to the first (alpha-) Carbon atom have particular importance. They are known as 2-, alpha-, or *α*-amino acids and often the term "amino acid" is used to refer specifically to these. They include the 22 proteinogenic –"protein building"- amino acids, which combine into peptide chains -polypeptides- to form the building blocks of a vast array of proteins. These are all L-stereisomers, or "left-handed" isomers, although a few D-amino acids –"right-handed" isomers- occur in bacterial envelope. (**Figure 4.B**)

Twenty of the proteinogenic amino acids are encoded directly by triplet codons in the genetic code and are known as "standards" amino acids. The other two "non-standard" or "non-canonical" are Selenocysteine -present in many prokaryotes as well as most eukaryotes, but not coded directly by DNA- and Pyrrolysine -found only in some archaea and one bacterium. (Jakubke, Sewald, 2008. p.20)

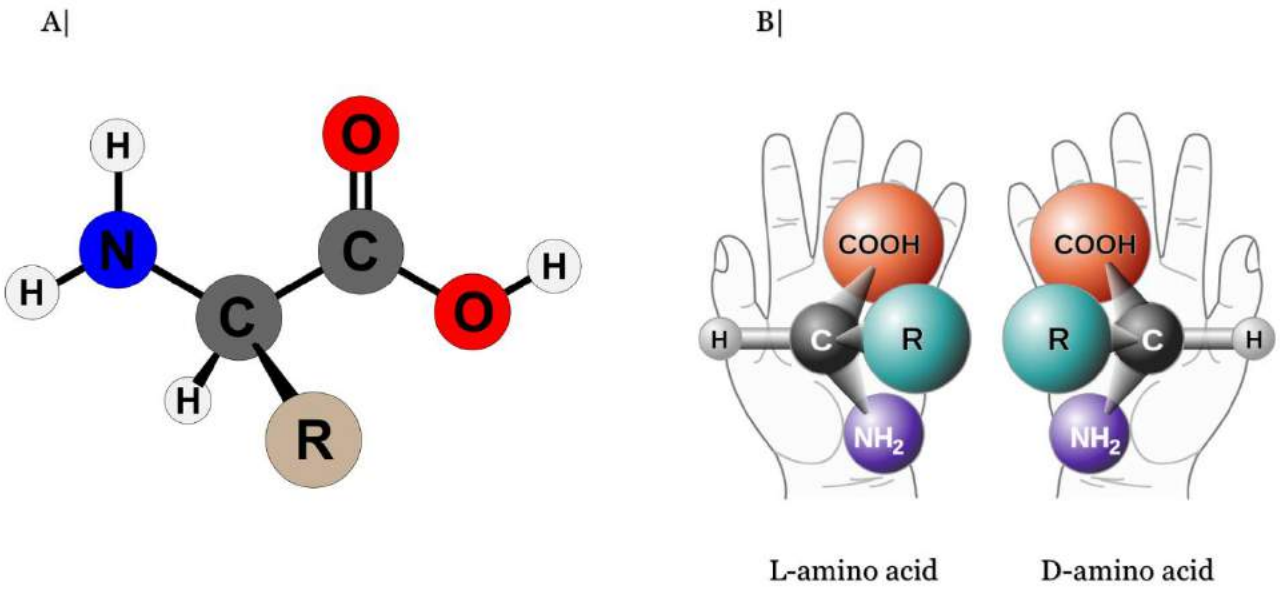


Figure 4| A. *The chemical structure of an amino-acid and* **B.** *The L- (left handed) & D- (right handed) stereoisomers of amino acids.*

As mentioned above, the sequence of amino acid residues in a protein is defined by the sequence of a gene, which is encoded in the genetic code and *translated* into amino acid residues through the processes of *transcription* and *translation*.

The amino acids are usually classified by the properties of their side chain into four groups. The side chain can make an amino acid a weak acid or a weak base, and a hydrophile, if the side chain is polar, or a hydrophobe, if it is non-polar (Creighton, 1993).

In a peptide chain all amino acids interact with each other, based on their side chain properties and environmental conditions, enabling in that way the process of *protein folding* and thus the protein's distinct 3D structure. (**Figure 5**)

Figure 5| *The Crystal Structure of Collagen Model Peptide (POG)3-PRG-(POG)4 (Okuyama, et. al., 2014)*

In **Figure 6.A-C** I present the 20 amino acids, which are usually found in proteins.

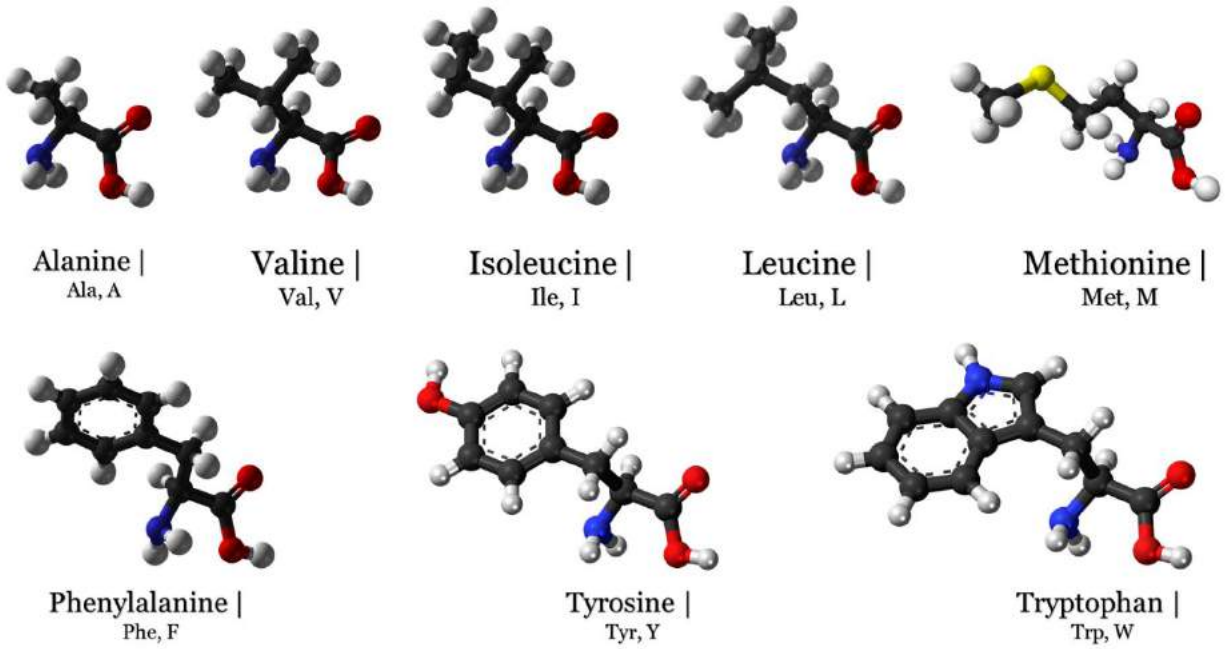


Figure 6.A| *Amino Acids with* ***Hydrophobic*** *Side Chains (****black*** *are* ***Carbon*** *atoms,* ***red*** *are* ***Oxygen*** *atoms, white are Hydrogen atoms,* ***blue*** *are* ***Nitrogen*** *atoms and* ***yellow*** *are* ***Sulfur*** *atoms).*

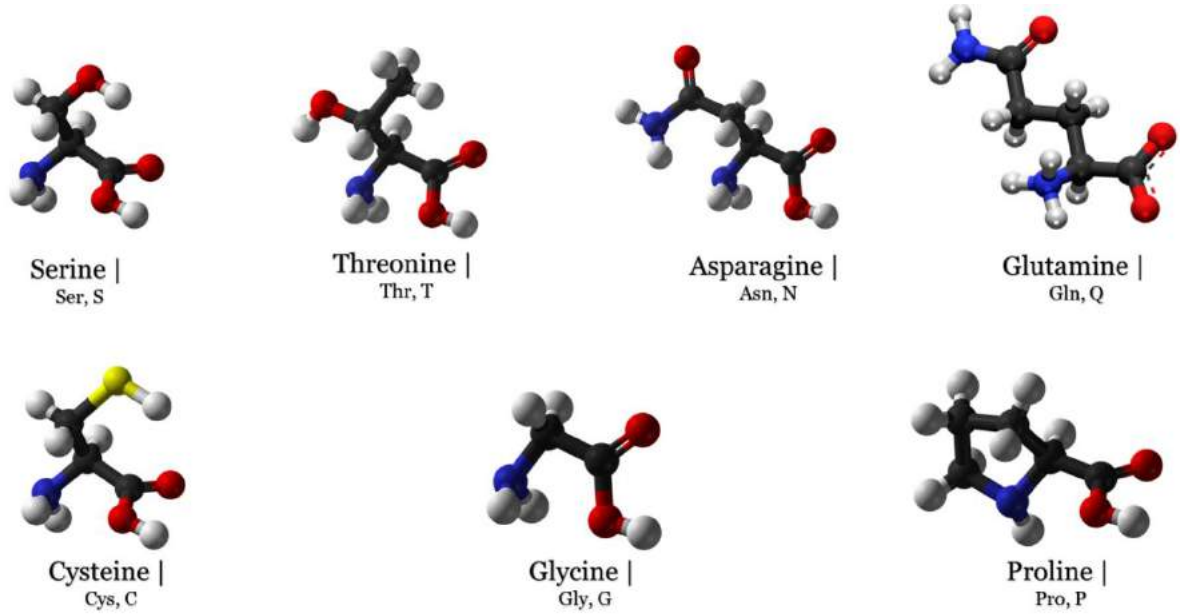


Figure 6.B| Amino Acids with **Polar Uncharged** Side Chains (above) & some **Special Cases** (below).

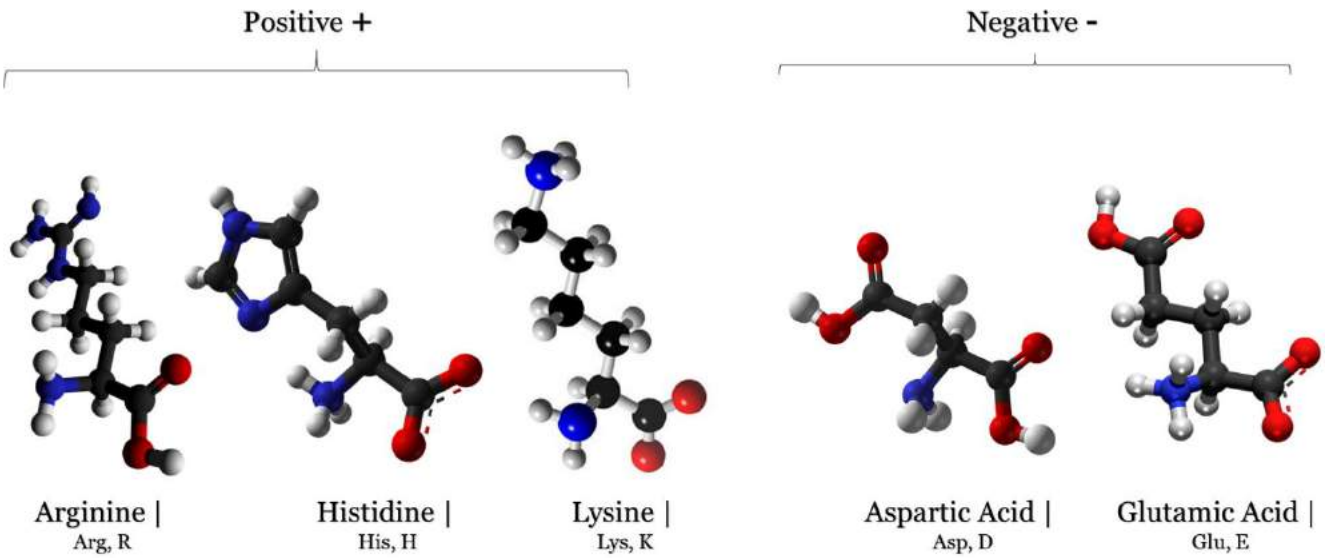


Figure 6.C| *Amino Acids with* ***Electrically Charged*** *Side Chains.*

But how, in the end, the final Structure is defined?

1.c. Structural Characteristics & Protein Folding

The individual amino acid residues are bonded together with peptide bonds and adjacent amino acid residues, forming in that way a peptide chain. A peptide bond is an amide type of covalent chemical bond linking two consecutive alpha-amino acids from C1 (carbon number one) of one alpha-amino acid and N2 (Nitrogen number 2) of another, along a peptide or protein chain (IUPAC-IUB, 1984). It is a type of condensation reaction: the products of this reaction are a molecule of water (H2O) and two amino acids joined by a peptide bond (-CO-NH-). (**Figure 7**)

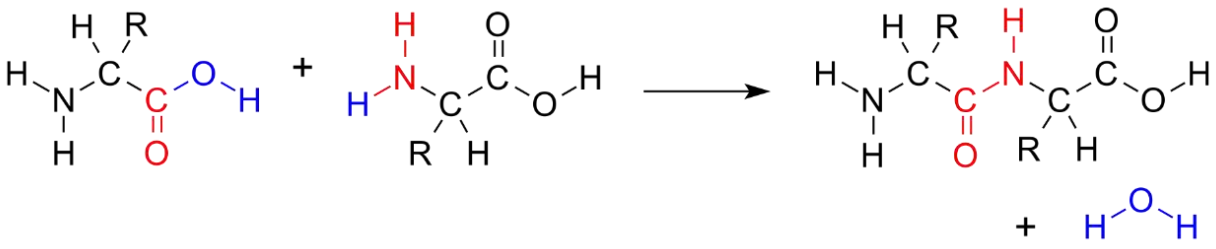


Figure 7| *The chemical diagram of the formation of a Peptide Bond.*

The amino acids bond together forming a linear peptide chain. This is called Primary Structure of a protein. Once the amino acid sequence of a protein -and so its Primary Structure- is complete, the protein starts to *fold* into its native conformation, which is the protein's final three-dimensional structure. The native conformation is usually biologically functional, in a reproducible manner. So, the physical process, by which a polypeptide folds into its characteristic and functional 3D structure from a random coil, is called Protein Folding. (Alberts, et al., 2002) (**Figure 8**)

Folding begins to occur even during translation of the polypeptide chain. As I mentioned before, amino acid residues interact with each other under the physicochemical properties that their side chains confer on them, the structural characteristic of the main chain -which is also called backbone-, and the physical conditions of the polypeptide chain's environment, to produce a well-defined 3D structure. According to the above, it is quite clear that the amino acid sequence or primary structure defines the 3D native state -conformation- of the protein, something that Anfinsen has described in his dogma. (1972)

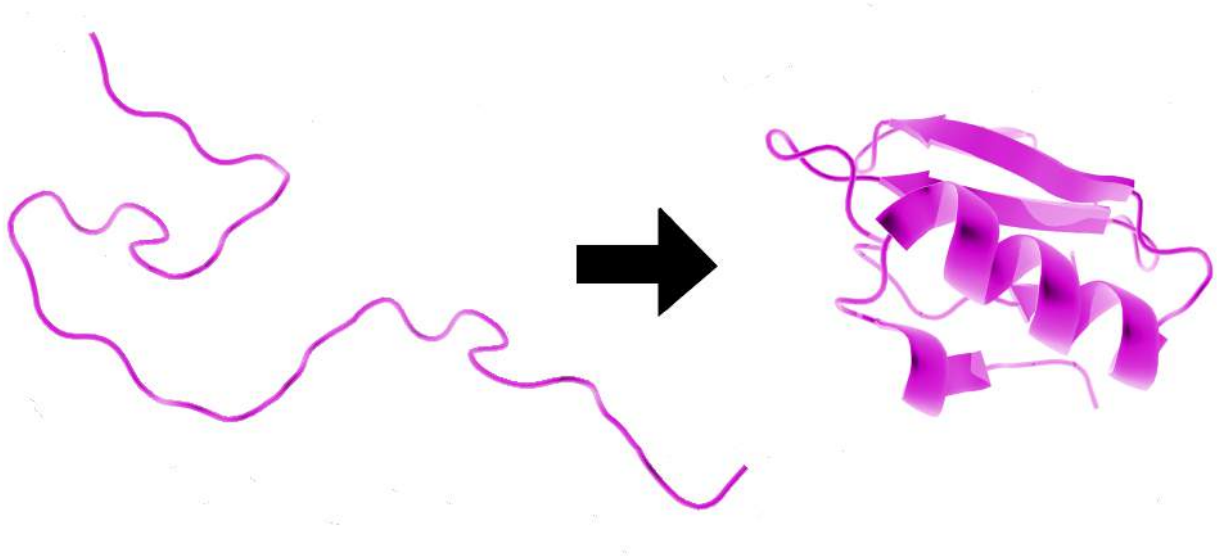


Figure 8| *The folding of Chymotrypsin Inhibitor 2 from an unfolded state to the native – presented as lines and ribbons.*

I mentioned above that beyond the physical properties of each amino acid residue in a polypeptide chain and the many possible interactions with its environment and other residues, a critical role to the Folding process plays the structural characteristics of the main chain.

In a protein chain there are three different structural factors, known as dihedral angles. In geometry, a dihedral angle is formed between two intersecting planes (**Figure 9.A**). In Structural Biology, these are also known as torsion angles. The two torsion angles of the polypeptide chain, also called Ramachandran angles, describe the rotations of the polypeptide backbone around the bonds between N-C*alpha*, called phi (φ) and C*alpha*-C, called Psi, (ψ) (**Figure 9.B**). The Ramachandran plot, which is a way to visualize dihedral angles ψ against φ of amino acid residues, provides an easy way to view the distribution of torsion angles of a protein structure (**Figure 9.C**). It also provides an overview of allowed and disallowed regions of torsion

angle values, serving as an important factor in the assessment of the quality of proteins' three dimensional (3D) structures.

Torsion angles are among the most important local structural parameters that control protein folding. Essentially, if there was a way to predict the Ramachandran angles for a particular protein it would be possible to predict its 3D structure. The reason is that these angles provide the flexibility required for folding of the polypeptide backbone, since the third possible torsion angle within the protein backbone, called omega (ω), is essentially flat and fixed to 180 degrees (180°). This is due to the partial double-bond character of the peptide bond, which restricts rotation around the C-N bond, placing two successive alpha-carbons and C, O, N and H between them in one plane. Thus, rotation of the main chain – backbone- of a protein can be described as the rotation of the peptide bond planes relative to each other. (Al-Karadaghi, 2019)

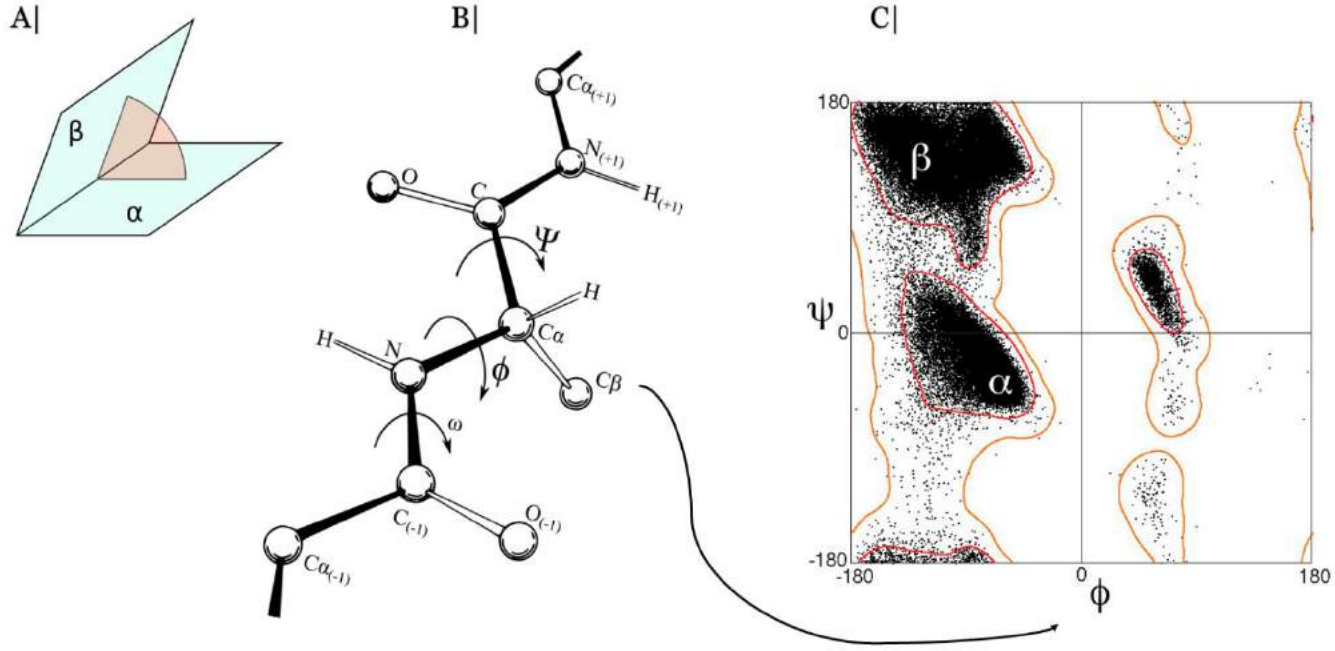


Figure 9| A. *A diagram of how Dihedral Angles are defined geometrically,* **B.** *"The Torsion angles in a protein chain" and how they are translated in* **C.** *the Ramachandran plot with psi (**ψ**) against phi (**φ**) and 100,000 protein data points. (inspired by Marz, E. ,2018)*

So, to sum up, there are three dihedral angles in a protein's main chain defined as φ (phi), ψ (psi) and ω (omega). The planarity of the peptide bond usually restricts ω to be 180°- the typical *trans* case- or 0° -rare *cis* case. So, the main factors that contribute to the protein's main chain flexibility are the φ and ψ torsion angles, also known as Ramachandran angles (Singh, et al., 2018).

Keep in mind, that the *cis-trans* isomerism I mentioned above is a term used in organic chemistry and they are prefixes from Latin, meaning "this side of" and "the other side of", respectively. In the context of organic chemistry *cis* indicates that *the functional groups* are on the same side of carbon side chain, and *trans* indicates that *the functional groups* are on the other side of the carbon chain (Lewis, Short, 1879). The distance between Cα in the *trans* and *cis* isomers is approximately 3.8 and 2.9 Angstrom (Å), respectively. The vast majority of the peptide bonds in proteins are *trans*, though the bond between the nitrogen of proline and another residue has an increased prevalence of *cis* compared to other amino acid pairs. (Singh, et al., 2018)

For this project, also keep in mind that it is common to represent polymers backbones, notably proteins, in internal coordinates, which is a list of consecutive dihedral angles and bond lengths. However, some types of computational chemists use cartesian coordinates instead. In computational structure optimization, like the main idea of this project, some programs need to flip back and forth between these representations during their iterations. This task can dominate the calculation time. For processes with many iterations or with long chains, it can also introduce cumulative numerical inaccuracy. While all conversion algorithms produce mathematically identical results, they differ in speed and numerical accuracy. (Parson, et. al., 2005)

In conclusion, folding is a spontaneous process that it depends on different factors, both biological and physicochemical. Some of them are hydrophobic interactions, formation of intramolecular hydrogen bonds, van der Waals forces and it is opposed by conformational entropy. The process of folding often begins co-translationally, so that the N-terminus of the protein begins to fold, while the C-terminal portion of the protein is still being synthesized by the ribosome, the complex of molecules responsible for the protein synthesis. However, a protein molecule may fold spontaneously during or after the process of synthesis. While these macromolecules may be regarded as "*folding themselves*", the process also depends on the solvent -water or lipid bilayer-, the concentration of salts, the pH, the temperature, the possible presence of cofactors and of molecular chaperones, which are molecules-providers of the special conditions for a protein's folding. (Berg, John, Stryer, 2007)

Of course, the limitations on protein folding are defined by the restricted bending angles or conformations that are possible. The torsion angles, which play a critical role in protein folding, as mentioned above, are presented in the two-dimensional Ramachandran plot, depicting ψ with φ angles of allowable rotation around the axis, which is defined by the backbone.

1.d. Protein Structures

Now that a basic knowledge background on proteins' Structural Characteristics is set, it is time to focus on the different types of Protein Structures, which are present during the Folding process and at the end of it, where it is folded in the *native conformation*.

As I already mentioned, once the translation of mRNA to amino acid sequence is complete, the product could be related to a line of amino acid residues -linear peptide chain-, which is called the *Primary Structure* of protein molecule. The specific amino acid residues and their position in the polypeptide chain are the determining factors for which portions of the protein fold closely together and form its 3D conformation. The amino acid composition is not as important as the sequence. The essential fact of folding, however, remains that the amino acid sequence of each protein contains the information that specifies both the native structure and the pathway to attain that state. This is not to state that nearly identical amino acid sequences always fold similarly. Conformations differ based on environmental factors as well. (Anfinsen, 1973)

Formation of a *Secondary Structure* is the first step in the folding process that a protein *takes,* to assume its native structure. Characteristic of Secondary are the structures known as alpha-helices and beta-sheets, that fold rapidly, because they are stabilized by intramolecular hydrogen bonds, as was first characterized by Linus Pauling. Formation of intramolecular hydrogen bonds provides another important contribution to protein stability. *α*-helices are formed by hydrogen bonding of the backbone and form a spiral shape.

The amino acids in an *α*-helix are arranged in a right handed helical structure where each amino acid residue corresponds to a 100° turn in the helix (3.6 residues per turn in the helix), and a translation of 1.5 Å, which means 0.15 nm (0.5 Å = atomic radius of Hydrogen) along the helical axis. Sometimes, short pieces of left-handed helix occur with a large content of achiral - not distinguishable from its mirror image- glycine amino acids, but usually they are unfavorable for the other L-amino acids. The hydrogen bonds, which stabilize an alpha-helix structure occur between the N-H group of an amino acid with the C=O group of the amino acid four residues earlier; also described as repeated $i + 4 \rightarrow i$ hydrogen bonding (**Figure 10.A**). In terms of φ, ψ torsion angles, there is a rule of 6.3 degrees of repeating angles along the backbone chain. Similar structures include the 3_{10} helix ($i + 3 \rightarrow i$ hydrogen bonding) and the π-helix ($i + 5 \rightarrow i$ hydrogen bonding). (Berg, John, Stryer, 2007) (**Figure 10.B**)

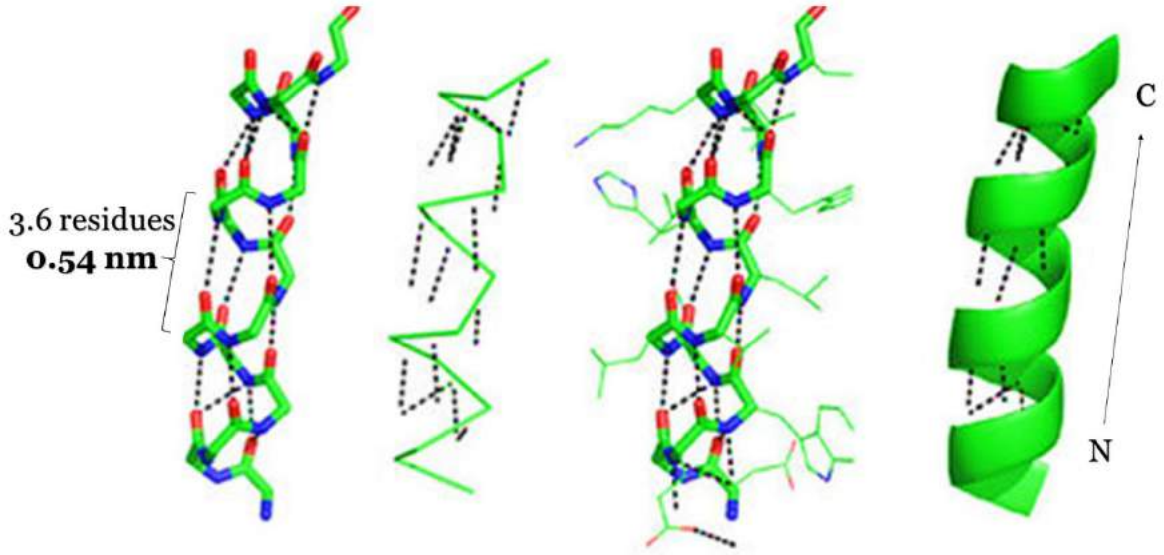


Figure 10. A| *The structural properties of α-helix: (from left to right) backbone, linear representation, plus side chains, ribbons representation. The dotted lines represent the Hydrogen Bonds.*

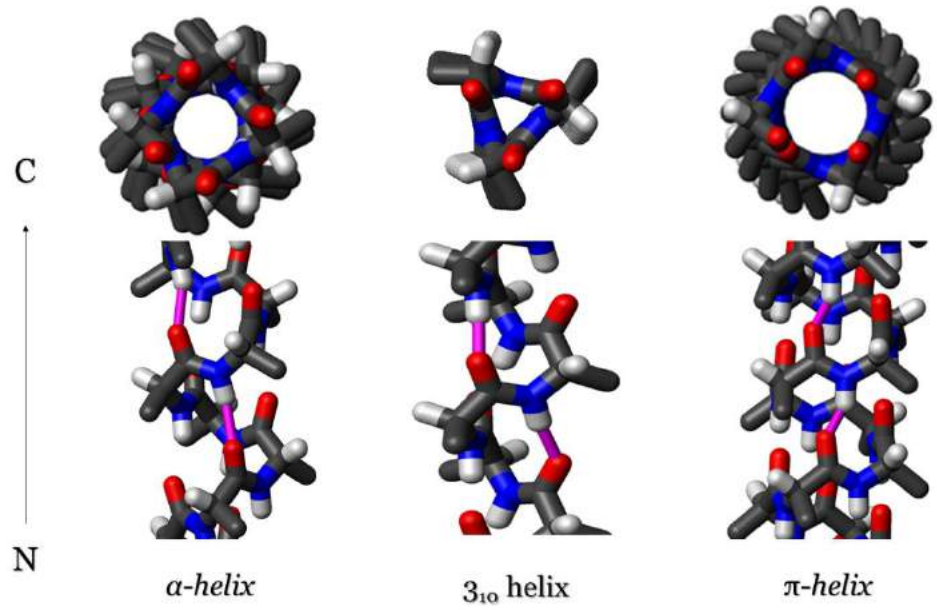


Figure 10.B| *The differences between the three main types of helices, which can be found in a protein structure- backbone view (below) & vertical view (above). From the N- to the C-terminus (arrow).*

The *β*- pleated sheet is a structure which is formed with the backbone bending over itself to form the hydrogen bonds. The hydrogen bonds are between the amide hydrogen and carbonyl oxygen of the peptide bond. There exists anti-parallel *β*-pleated sheets and parallel *β*-pleated sheets, where the stability of the hydrogen bond is stronger in the anti-parallel *β*-sheet as it bonds with the ideal 180 degree angle compared to the slanted hydrogen bonds formed by parallel sheets. (**Figure 11.A & B**)

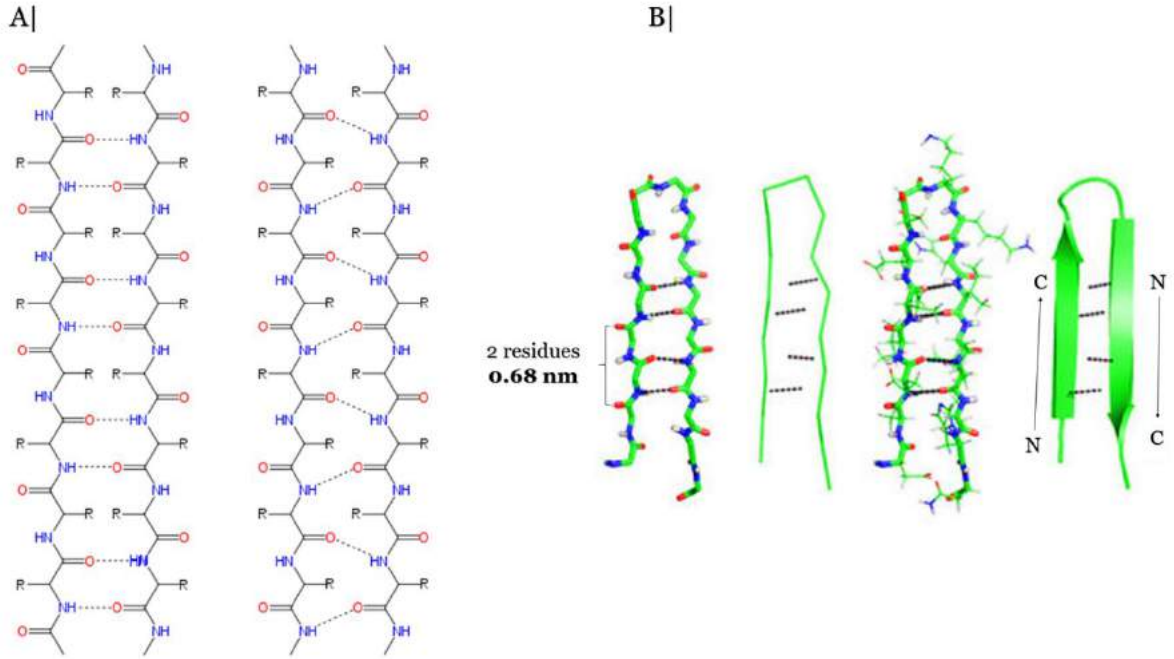


Figure 71| A. *The parallel and anti-parallel beta-sheets,* **B.** *Structural properties of beta-sheets.*

In computational structural studies, the secondary structures can be identified with many methods, one of which is Define Secondary Structure of Protein -also known as DSSP. (Kabsch, Sander,1983)

So, from a linear peptide chain in the *Primary Structure* (**Figure 12.A**) emerges through folding the *Secondary Structures*. The helices and β-sheets can be amphipathic in nature (**Figure 12.B**). They contain a hydrophilic portion and a hydrophobic portion. This property of Secondary Structures aids in the *Tertiary Structure* of a protein in which the folding occurs so that the hydrophilic sides are facing the aqueous environment surrounding the protein and the hydrophobic sides are facing the hydrophobic core of the protein. Secondary Structure leads to the Tertiary Structure formation. Once the protein's Tertiary's Structure is formed, it is stabilized by the hydrophobic interactions and possible other types of interactions, such as disulfide bridges, which are formed between two Cysteine amino acid residues. (**Figure 12.C**)

Tertiary structure of a protein involves a single polypeptide chain. However, additional interactions of folded polypeptide chains give rise to *Quaternary Structure* formation. In that stage, multiple folded polypeptide chains interact to form a fully functional Quaternary protein molecule. (**Figure 12.D**) (Berg, John, Stryer, 2007)

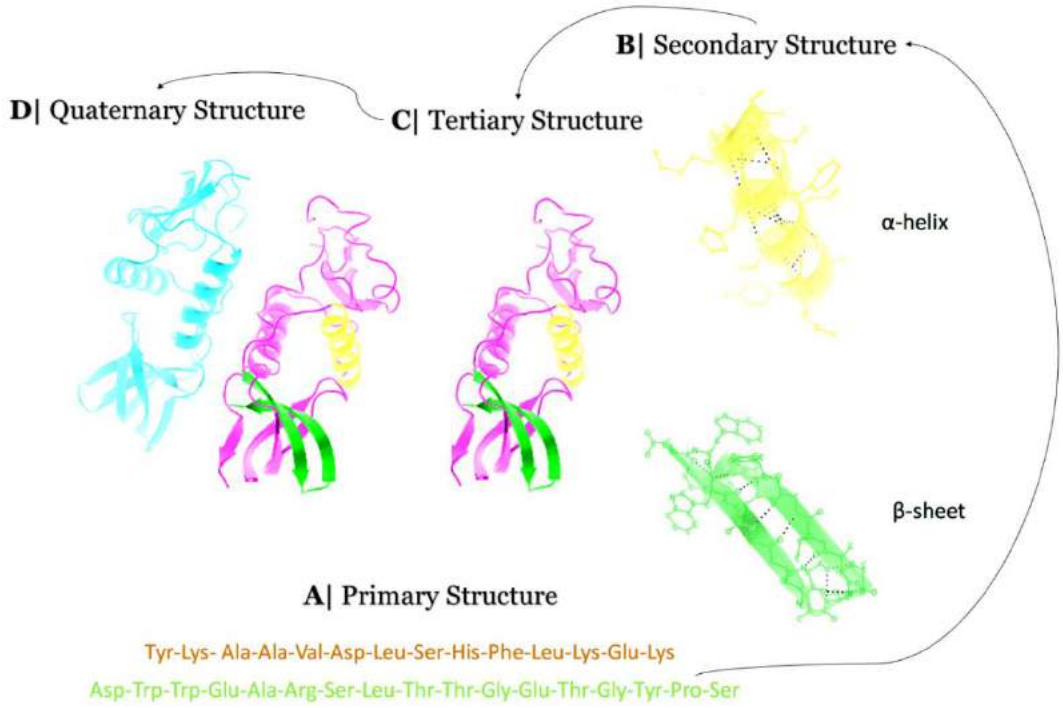


Figure 12| A. *The Primary Structure as a three-letter code of amino acid sequence,* **B.** *The Secondary Structure with the distinct alpha-helices and beta-sheets,* **C.** *The Tertiary Structure at the end of the Folding and* **D.** *The Quaternary Structure when the protein (Copyrighted work)*

1.d. Loops & Turns…

There are two other important states of Secondary Structure, known as Turns and Loops. Both structure types connect two stabilized secondary structures, like an α-helix or a β-plane.

The basic difference between Turns and Loops is, that Turns are stabilized with interactions -f.e. through hydrogen bonding -between the amino acid residues they have, and Loops have more flexible and non-specific structure without necessarily being stabilized with hydrogen bonding between the residues, which participate in the formation of a Loop structure.

Turns have been classified in multiple ways before. One of the most extended and accurate classifications that has been described is from the work of J. M. Thornton and her partners in 1988. In their paper they describe:

> *"There are four types of connecting loops: aa, βa, αβ and ββ (That is, an aa loop connects two a-helices, etc.)[…]"* (p. 63)

Turns can be classified by their backbone dihedral angles, so they can be described with the Ramachandran plot. They are small peptides -usually 1 to 5- and are defined as intersecting regions between two Ca atoms of two different residues with a distance of less than 7 Angstrom (Å). (**Figure 13.A**)

A|

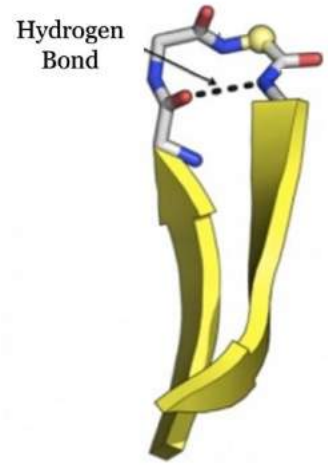


Figure 13.A| *A beta turn connecting two beta-planes.*

Loops, on the other hand, -also called ω loops- are *patternless* and quite flexible regions. Usually they play an important role in the protein's function; by interacting with the protein's substrate, providing kinetic flexibility to the molecule, etc. Some loop regions are one of my main points of interest for this project. (**Figure 13.B**)

B|

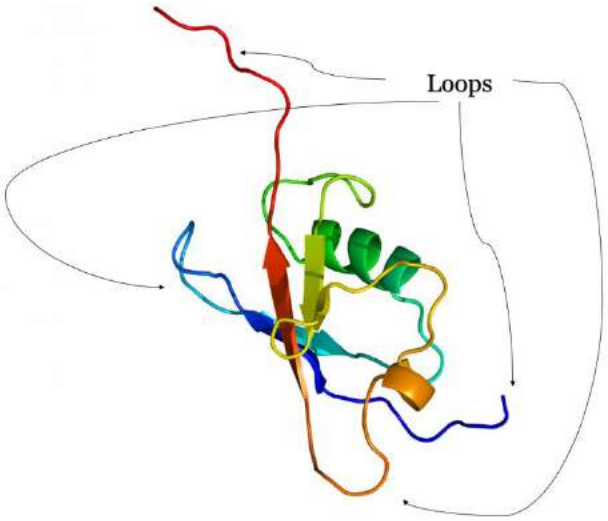


Figure 13.B.| *Loops in the structure HERPUD1.*

Maybe, reaching the end of the Protein Structural section, it is a bit or more clear that studying the structure of loops is both important and quite difficult. Many techniques advanced for the determination of a protein's structure are sometimes incapable of producing sufficient data for the studying of these regions. Mainly because of the un-stabilized and flexible structure they present. But at the same time, the way these loops participate in a protein's function make them more than worth studying.

This problem usually can be solved through the means of Computational Methods.

2. Computational X-ray Crystallography

Computational Crystallography is the field of Crystallography, which focuses on the definition and studies of molecules' structure via computational methods. It is one of the most famous *in silico* -on the computer- applications of research processes in Molecular Biology.

But before all that, what is Crystallography and more specifically, what is X-ray Crystallography?

2.a. X-ray Crystallography- *"Breaking down the terms"*

X-ray Crystallography is the experimental science determining the atomic and molecular structure of a crystal, in which the crystalline structure causes a beam of incident X-rays to diffract into many specific directions. The crystalline structure can be that of a nucleic acid -usually DNA-, drug molecules, vitamins and proteins. Regarding this project, as always, the focus is mainly on Protein X-ray Crystallography.

So why X-rays to study proteins' structures? To answer that question, it is necessary to refer to the meaning of *observation in an experimental procedure*. The observation of an object during an experimental procedure is based on the collection of *information*, and the type of that information by extension, from the *observer*. In order the observer to collect that information, *an intermediate factor* is needed. That factor, which is chosen by the observer, interacts with the object of interest and defines the type of the collected information. For the use of such factor, an *observatory tool* is needed. For example, in Microbiology, to *observe* a microorganism -*the object of interest*-, a microbiologist may use a microscope. The microscope is the *observatory tool*. The *intermediate factor* is *light* -a specific electromagnetic radiation- produced by a source in the microscope, interacting with a sample of that microorganism and scattered by it, refracted by the microscope's lenses, and reaching the observer's eye. (**Figure 14**)

Figure 14| *The path of light in a simple light microscope.*

The information about the microorganism is *carried* through the light to the observer's eye. The image is the main element for a microscopical observation. To observe a protein's structure, it is needed to produce an image -or more specifically, a *model*- of the structure. For imaging methods, like the determination of protein's structure, the meaning of *resolution* is quite important.

Resolution is the measurement of *detail/specificity* of an object's image. To understand that, there is a quite simple example; Imagine you are watching an image of two objects one next to each other. Then take a distance from that image. The longer distance you take from the image, the closer the *forms* of the objects get together, merging in one indistinct *form*. The higher the resolution, the longer

distance is needed, in order for these two distinct *forms* to merge in one. (Glykos, 2015, p. 12-15) (**Figure 15**)

Figure 15|. *High resolution (left) and low resolution (right). As resolution gets lower, it seems like independent points from the image tend to merge together (Glykos, 2015, p. 14).*

Another important factor in imaging is the properties of electromagnetic (EM) radiation. In physics, EM radiation refers to the wave of the electromagnetic field propagating -radiating- through space, *carrying* electromagnetic energy in packages, known as quanta and photons. The spectrum of electromagnetic radius includes radio waves, microwaves, infrared, the visible light -what humans can see- ultraviolet, X-rays and gamma rays and they are characterized according to the radius wavelength (m) and frequency (Hz).

Classically, electromagnetic radiation consists of electromagnetic waves, which are synchronized oscillations of electric and magnetic fields. In a vacuum -space with no air- electromagnetic waves travel at the speed of light. In homogeneous, isotropic media -like a sample of microorganisms- the oscillations of the two fields are perpendicular to each other and perpendicular to the direction of energy and wave propagation, forming a transverse wave. (Reitz, Milford, Christy, 1992) (**Figure 16.A**)

So, talking about electromagnetic radius, I am talking about electromagnetic waves. Waves have properties/characteristics, which are *phase*, *length*, *amplitude*, and *direction*. When an electromagnetic wave interacts with a sample, it is *scattered*. The diffracted wave can be captured by a special sensor, and the characteristics of the wave can be studied. The characterization of the diffracted wave *gives* information for the object that caused its diffraction.

Of course, when the diffracted electromagnetic wave belongs to the spectrum of visible light, there is no reason to analyze its characteristics, to see the object that caused the diffraction. Once it is diffracted, the only thing that needs to be done is to use *special lenses*, in order to *magnify* it and redirect it to your eye. The human eye can break down that information and *perceive* it as an image. But when the electromagnetic wave, that needs to be studied, cannot be *seen* -such as X-rays- then an analytical study of its characterizations is needed.

There is a mathematical model that can be used to study the properties of a wave and it is called *Fourier analysis*. Through that, somebody can represent the refractions that happens to a wave once it interacts with a lens. So, in a hypothetical experiment, a beam of not visible radiation *runs* through a sample. Multiple waves diffract from it. The only thing needed to be done is to capture the *scattered waves* from the sample, to measure the *properties* of each wave -direction, wavelength, wave amplitude and phase- and study the information from these waves with the lens's equation -Fourier analysis. (Glykos, 2015, p. 27-31)

A|

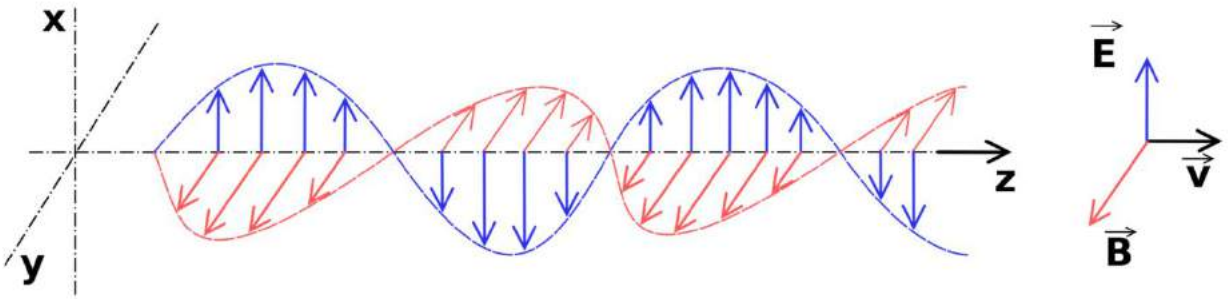


Figure 16.A| *A linearly polarized sinusoidal electromagnetic wave, propagating in the direction +**z** through a homogeneous isotropic, dissipationless medium, such as vacuum. The electric field -blue- oscillates in the +/- **x** direction, and the orthogonal magnetic field -red- oscillates in phase with the electric field, but in the +/- **y** direction.*

To connect the electromagnetic properties of the waves with the meaning of resolution: to produce the image of an object -especially the image of a protein's structure- in high resolution, it is important to be aware of which electromagnetic radiation is needed for the study. The *smaller* the wavelength of the electromagnetic radiation that is used, the *higher* the resolution of the image will be.

In Structural Biology the measure of resolution is Ångström (Å), which as mentioned above it is comparable to the atomic diameter of a Hydrogen atom, which is close to 1 Å. So, to observe the atoms of a protein's structure, a wavelength close to 1 Å is needed. In the electromagnetic spectrum, X-rays are in that wavelength. Hence, X-ray Crystallography. (Glykos, 2015, p. 13-16) (**Figure 16.B**)

B|

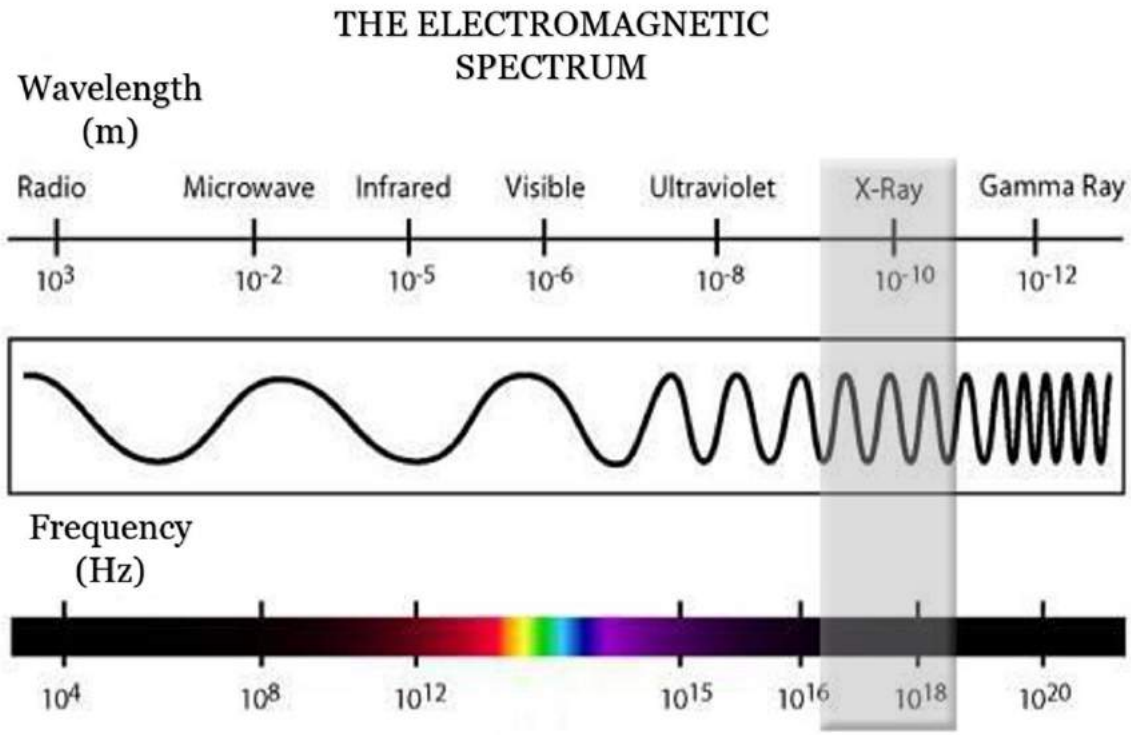


Figure 16.B| *The Electromagnetic Spectrum and the characteristics of X-rays (grey bar).*

In relation to the example of microscopy, a crystallographic experiment would look like **Figure 17**; X-rays would interact with a single isolated molecule of the targeted protein. The X-rays would scatter from the sample. The scattered beams would be refracted by a lens and the inverted and magnified image of the protein molecule would be produced and ready for observation.

Unfortunately, it is not that simple and there are many issues in the diagram of the *Ideal Crystallographic Experiment*. First, there is no available material in nature with properties that can *bend* X-rays and change their direction. In other words, there is no such thing as "X-ray lenses".

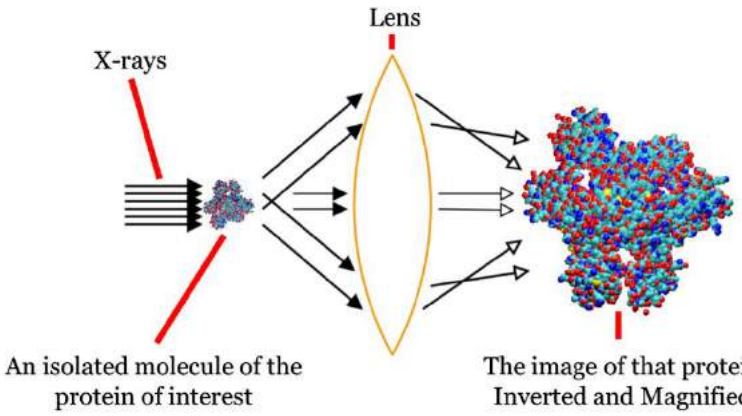


Figure 17| *"The ideal crystallographic experiment" (Glykos, 2015, p.17)*

Second, even if there were such lenses, because of the extremely high frequencies the X-rays present, it is almost impossible to determine these waves' phases. And third, a single copy of the target-protein's molecule is usually not enough for a complete data collection.

The third issue is solved using crystallized proteins.

2.b. Why Crystals?

Crystal is a solid material, characterized by the three-dimensional periodic conformation of the atoms and molecules, which constitute the crystal. Crystals are made from the repetition and simple *relocations* -of these repetitions, in 3D space- of a main motif. This motif is usually in cuboid shape and it is called crystal's Unit Cell. So, the Unit Cell is the *structural unit* of a crystal.

A crystal is usually characterized by the geometry and symmetry of the unit cell. In a protein crystal, in the unit cell may be contained at least one copy of the protein's molecule. The overall symmetry or non-symmetry of a crystallized protein structure is determined by the above factors and it is quite fundamental for the further study of the protein's structure. (**Figure 18**)

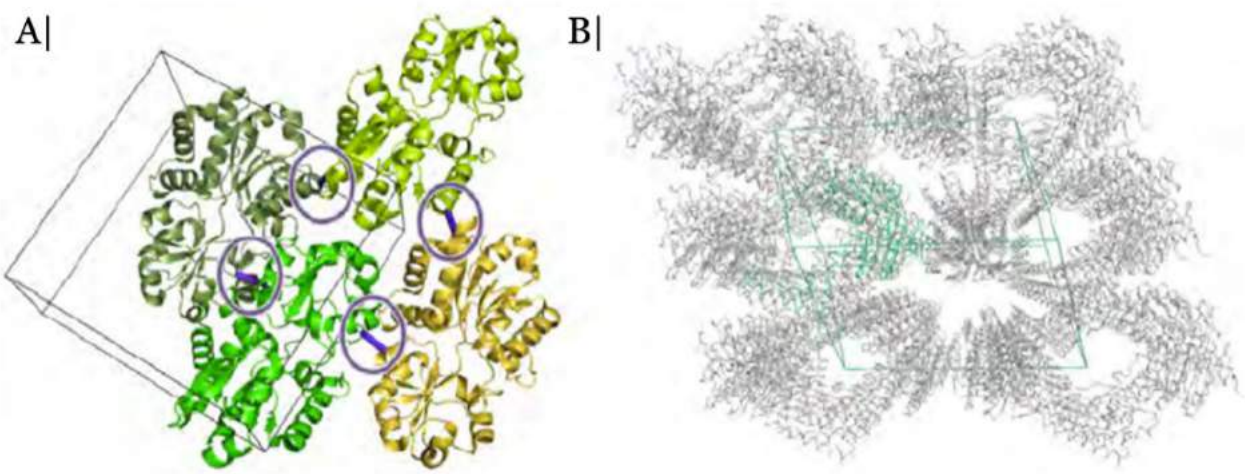


Figure 18| A. *Example of protein packing in a Unit Cell.* **B.** *Three-dimensional crystal packing of a different protein molecule. (Abts, Schmitt, et al., 2012)*

The creation of a protein crystal is a quite laborious and time-consuming procedure. Especially because of the many different factors that have to be accounted for the protein molecule of interest. But there are several methods to create a protein crystal. One of them is known as the "Hanging Drop" method. (Glykos, 2015, p. 19-23)

In order to create a protein crystal, first, a reliable source of the protein of interest must be available, together with a *high-yield* purification/concentration protocol for the for the protein's molecule copies. Then, the main idea is to put a solution containing multiple molecule

copies of the targeted protein in conditions, where the concentration of the protein molecules is higher than the solubility equilibrium for these specific conditions.

As Smyth and Martin explain in their review (2000):

"The principle of crystallisation, whether of macromolecules or salts (unfortunately!) is to take a solution of the sample at high concentration and induce it to come out of solution; if this happens too fast then precipitation will occur, but under the correct conditions crystals will grow." (p.9) (**Figure 19**)

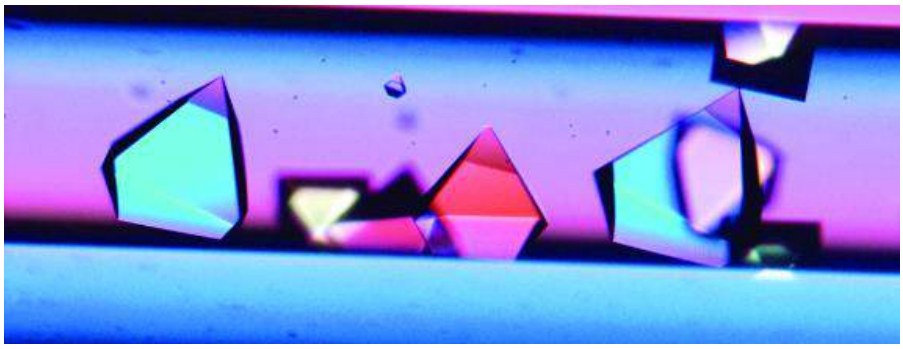


Figure 19| *Protein Crystals*

The use of crystals for the defining of a protein's structure solves lots of issues. It is a sample easier to handle during the experimenting procedure and because of its solid phase, it is easy to rotate the sample and have information about the protein from multiple sides of the crystal. In that way there is information for a 3D *model building* of the molecule.

Inside the crystal are contained multiple copies of the protein of interest, which are all symmetrical to each other, as it is defined by the Unit Cell's symmetrical properties. This means that there is information about the structure from every copy inside the crystalized structure, and because of the symmetry, the diffracted X-ray waves interfere with each other, amplifying the signal and reducing the noise. That explains why we need crystals in "X-ray Crystallography".

Another important part about the crystals that is needed to be mentioned is the Crystal Symmetrical Properties. Crystal structure is described in terms of geometry of arrangement of particles in the unit cell. The unit cell is defined as the smallest repeating unit having the full symmetry of the crystal structure, just as mentioned above.

The geometry of the unit cell is defined as a parallelepiped, providing six lattice parameters taken as the lengths of the cell edges (*a, b, c*) and the angles between them (α, β, γ). The positions of particles inside the unit cell are described by the fractional coordinated (*xi, yi, zi*) along the cell edges, measured from a reference point. Usually, it is only necessary to report the coordinates of a smallest asymmetric subset of particles. This group of particles may be chosen so that it occupies the smallest physical space, which means that not all particles need to be physically located inside the boundaries given by the lattice parameters. All other particles of the unit cell are generated by the *symmetry operations* that characterize the symmetry of the unit cell.

The collection of symmetry operations of the unit cell is expressed formally as the *Space Group* of the crystal structure.

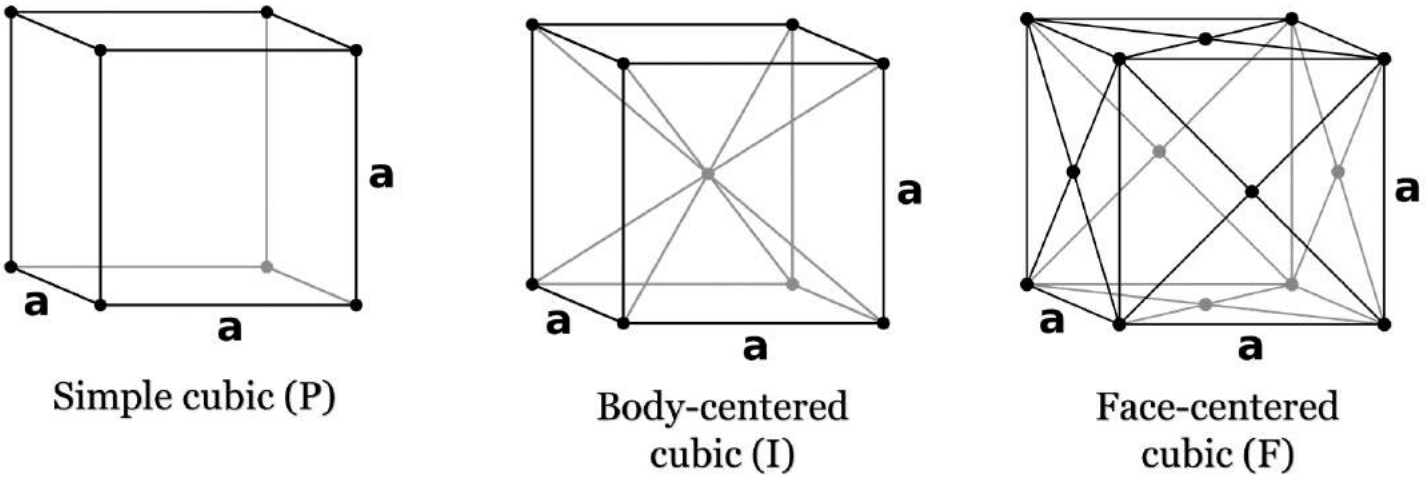


Figure 20| *The main classes of a unit cell cubic symmetry.*

In mathematics, physics and chemistry, a space group is the symmetry group of a configuration in space, usually in three dimensions. In crystallography, space groups are called crystallographic or Fedorov groups and represent a description of the symmetry of the crystal. A definitive source regarding 3-dimensional space groups is the International Tables for Crystallography. (Hahn, 2002)

For the classification of space groups symmetry there are multiple parameters. Some of the most basic are the *Bravais lattices*, also referred to as space lattices. They describe the geometric arrangement of the lattice points and therefore the translational symmetry of the crystal. The three dimensions of space afford 14 distinct Bravais lattices describing the translational symmetry. All crystalline materials recognized today, not including the quasicrystal -an exception of crystals which are ordered but not periodic-, fit in one of these arrangements.

The crystal structure consists of the same group of atoms, the basis, positioned around each lattice point. This group of atoms therefore repeats indefinitely in three dimensions according to the arrangement of one of the Bravais lattices. (**Figure 21.A & B**)

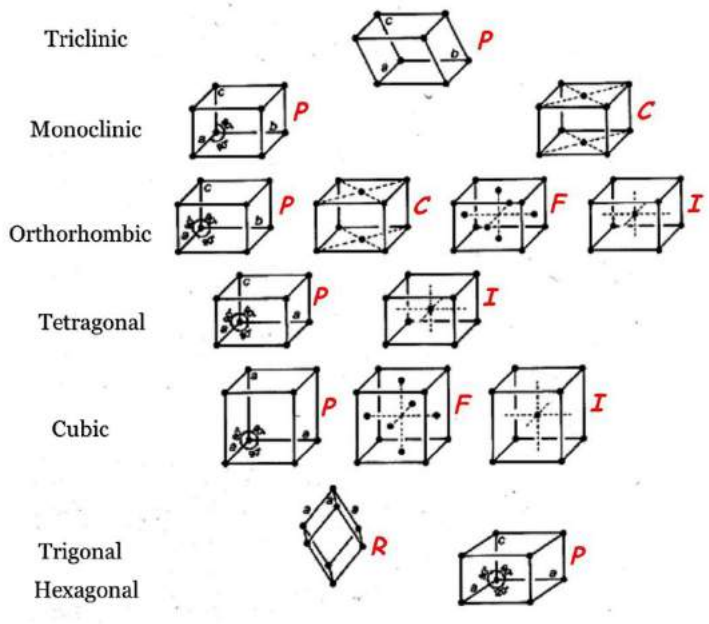


Figure 21.A| *The 14 Bravais Lattices*

Figure 21.B| *The 230 Space Group List Project (Copyrights reserved to Frank Hoffmann) – Zoom In for clarity.*

To sum up, once a crystal is obtained, data can be collected using a beam of X-ray radiation. Although many universities that engage in Crystallographic research have their own X-ray producing equipment, synchrotrons are often used as X-ray sources, because of the more precise and complete patterns such sources can generate. Synchrotron sources also have a much higher intensity of X-ray beams, so data collection takes a fraction of the time. (**Figure 22.A**)

The study of the symmetry of the Unit Cell is important for all the steps of the process of finding the 3D structure of a protein. Mainly because the shape and symmetry of the cell define the directions of the diffracted beams and the locations of all atoms in the cell define their intensities on the *diffraction pattern* -the result of an X-ray experiment.

The larger the unit cell, the more diffracted beams, also called *reflections*, can be observed. Moreover, the positions of each atom in the crystal structure influences the intensities of all the reflections and, conversely, the intensity of each individual reflection depends on the positions of all atoms in the unit cell. (Wlodawer, Minor, Dauter, Jaskolski, 2008)

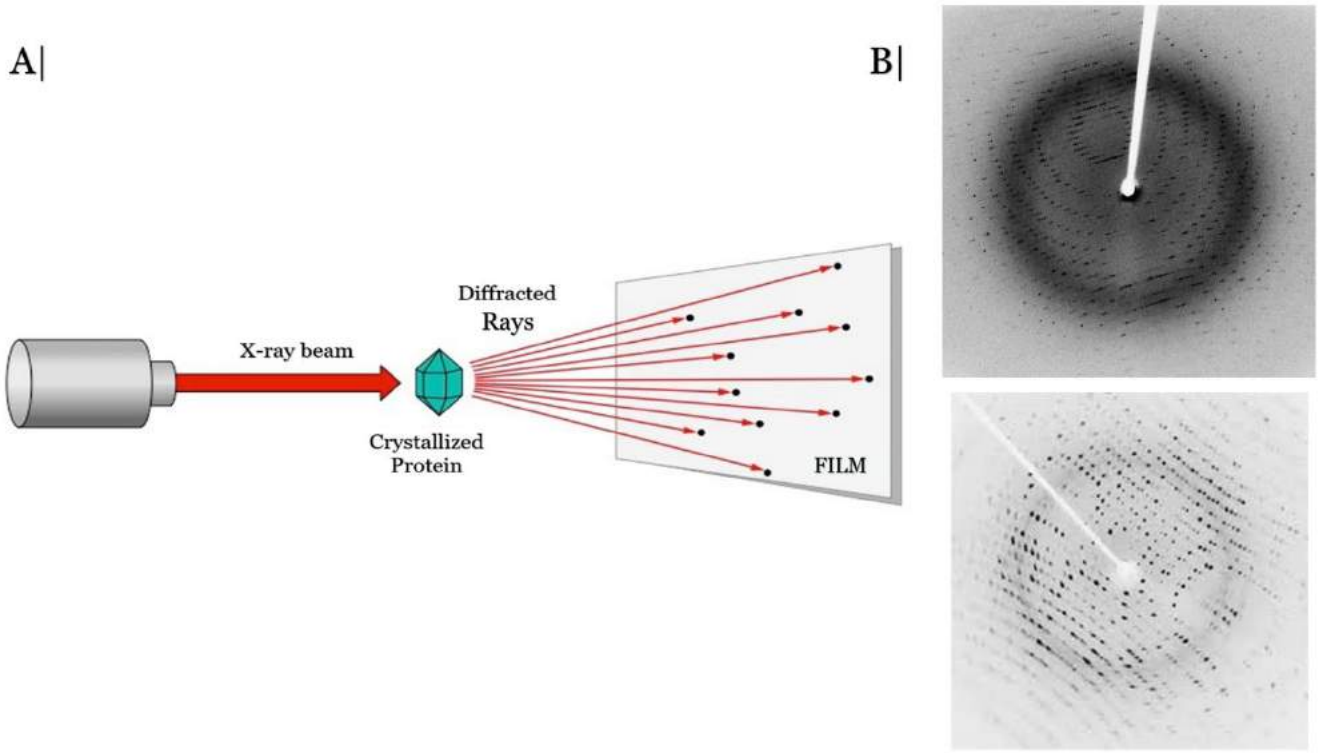


Figure 22| A. *The X-ray Crystallographic Experiment* **B.** *Diffraction patterns (above & below); The black dots indicate the reflections of the crystal.*

Once the diffraction pattern is generated, the crystal is rotated, and the X-ray experiments starts again. In that way, will be generated diffraction patterns from multiple sides of the crystal, and they will all contribute to the final model building of the protein's structure.

The diffraction patterns can be used to characterize the reflections. Each *dot* represents a reflection. Each reflection can be allocated in a three-dimensional grid -with three-dimensional coordinates- defined by each diffraction pattern (**Figure 22.B**). The three coordinates used for the characterization of each reflection in the three-dimensional space are integral numbers and are known as *Miller indices*. They are symbolized as *h*, *k* and *l*. (Glykos, 2015, p. 50-57)

The Miller indices are used, as shown in **Figure 23**:

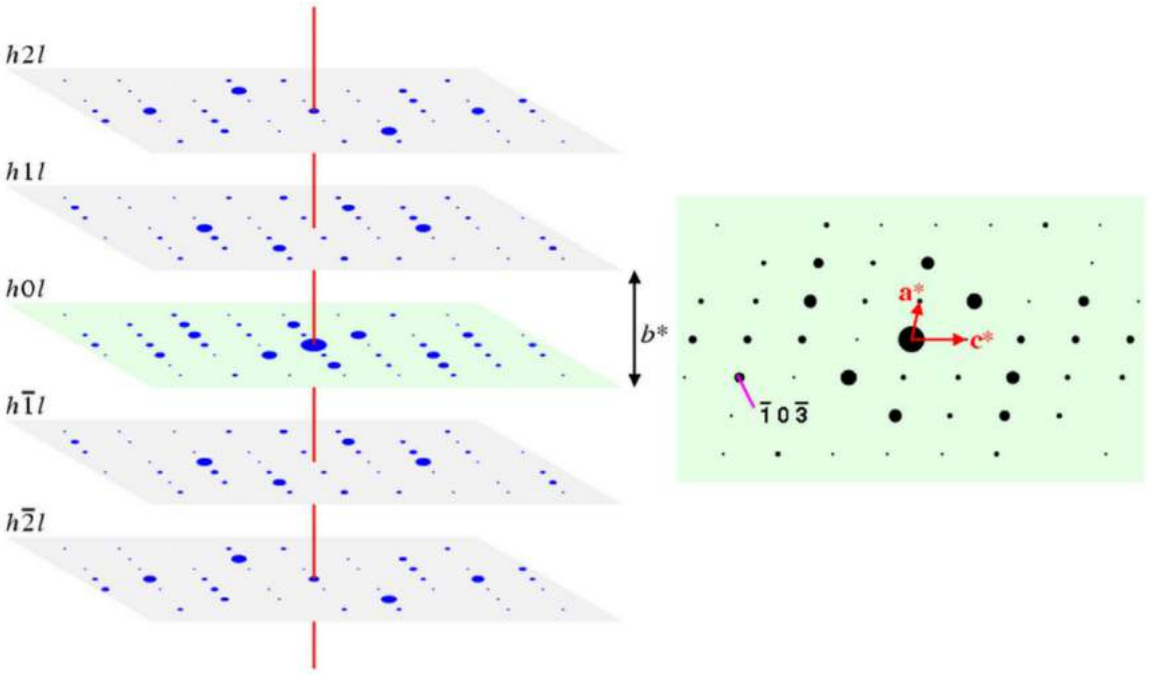


Figure 23| *Different diffraction patterns in a three-dimensional grid. The* l *index is used for the characterization of the plane. The a*, b* and c* are vectors, which define the reference system of the axes. (Glykos, 2015, p.51)*

Most of the crystallographic issues are solved by using crystals. But there is still one great issue in Crystallography…

2.c. Phase Problem…

The Phase Problem in Crystallography arises from the fact that the frequencies of the X-ray radius are quite *high*, which means that the phase of each diffracted wave is nearly impossible to be calculated directly. That means that the analysis Fourier cannot be used, because of incomplete data. At the same time, the part of the information related to the phases of the waves are significantly more important than the part of the information from other wave properties, like the wave amplitude. (**Figure 24.A & B**) (Glykos, 2015, p. 34)

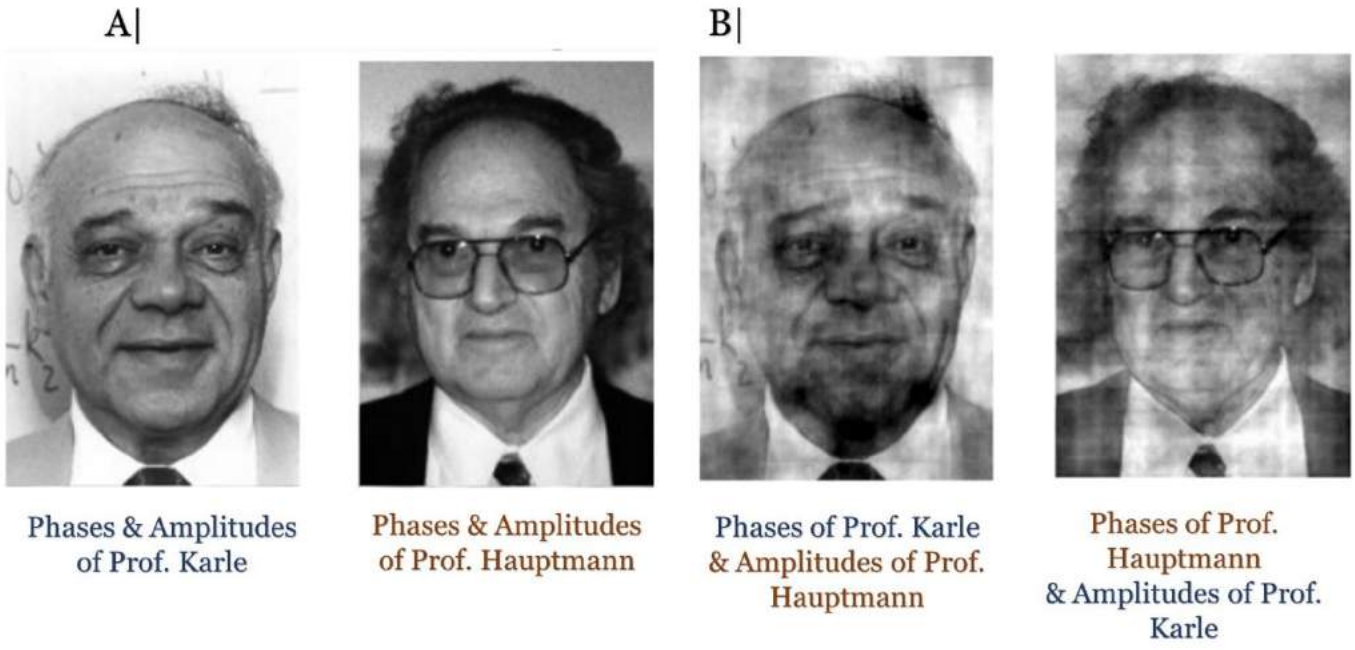


Figure 24| *In* **A.** *there are the photos of two Nobel winners, with both phases & amplitudes from the images and in* **B.** *they have the same phases, but different amplitudes. (Glykos, 2015, p.35)*

Although the Phase Problem is the real issue in Protein X-ray Crystallography, there are multiple ways that have been developed that approach a solution to that Problem. One of them is known as *Molecular Replacement* (MR); where a similar molecule's already-known phases are grafted onto the intensities of the molecule at hand, which are observationally determined. These phases can be obtained experimentally from a homologous molecule, or if the phases are known for the same molecule but in a different crystal, by simulating the molecule's packing in the crystal and obtaining theoretical phases.

Other famous methods -and most common approaches to solving the Phase Problem- are the *Single isomorphous replacement* (SIR) and the *Multiple isomorphous replacement* (MIR). SIR is conducted by soaking the crystal of a sample to be analyzed with a *heavy atom* solution or co-crystallization with the heavy atom. The addition of the heavy atom to the structure should not affect the crystal formation or unit cell dimensions in comparison to its native form, hence, they should be *isomorphic*.

Data sets from the native and heavy-atom derivative of the sample are first collected. Then the interpretation of the Patterson -which is another, more mathematical, approach to phase problem solving- difference map reveals the heavy atom's location in the unit cell. This allows both the *amplitude* and the *phase* of the *heavy-atom contribution* to be determined.

To study and compare mathematically the wave properties of each reflection, a *vector* is used to represent a reflection -diffracted wave from the crystal- known as the *Structure Factor* ($\boldsymbol{F}_{hkl}$). This algebraic expression relates the amplitudes and phase of the beam diffracted by the planes of the crystal -which are characterized by the Miller indices- to that produced by a single scattering unit at the vertices of the primitive unit cell. It can also be mentioned as a *vector presentation* of a diffracted wave by a crystal. In a two-dimensional cartesian coordinate system, the *length* of the vector is defined by the wave amplitude and the *angle* from the phase of the

reflection presented. That vector can be presented in a diagram, known as *Argand diagram*. (**Figure 25.A**)

Back to the *Isomorphous Replacement*, the *Structure Factor* of the heavy-atom derivative (***F*ph**) (***p*** for *native protein* crystal and ***h*** for *heavy-atom*) of the crystal is the vector sum of the lone heavy atom (***F*h**) and the native crystal (***F*p**), then the phase of the native ***F*p** and ***F*ph** vectors can be solved geometrically, for example using the Argand diagram. This algebraic approach ends up with at least two possible solutions (**Figure 25.B**). The Multiple isomorphous replacement -which is the same procedure with more heavy atoms- solves the problem of the multiple possible solutions, by combining the answers to conclude on *one possible solution*.

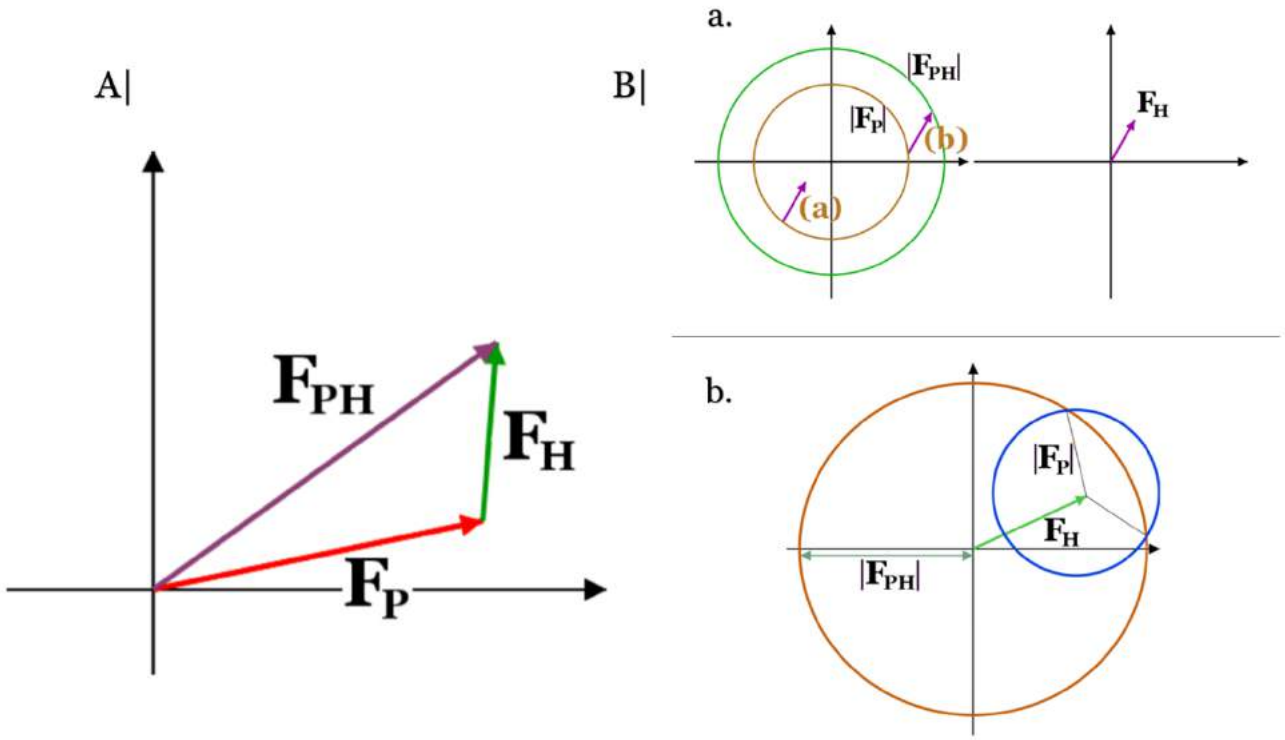


Figure 25|A. *The Argand diagram and vector representations of the Structural Factors of heavy atoms and native crystal phases as the sum* **F*p*+ F*h*= F*ph***. **B. a & b.** *Searching solution in the phase problem using the Argand diagram. (Glykos, 2015, p. 87-92)*

Another known method is using the effect known as *Anomalous Scattering*. These methods are known as *Single-wavelength anomalous diffraction* (SAD) and *Multi-wavelength anomalous diffraction* (MAD). In MAD, atoms' inner electrons absorb X-rays of particular wavelengths and reemit the X-rays after a delay, inducing a phase shift in all of the reflections, known as the *anomalous dispersion* effect. Analysis of this phase shift results in a solution for the phases. In a similar way works the SAD method, but for single wavelength shifts.

Last, but definitely not least, the *Patterson function* is a mathematical approach to the solution of the Phase Problem in X-ray Crystallography. It is essentially the Fourier transform of the intensifies rather than the structure factors:

$$P(u,v,w) = \sum_{hkl} |F_{hkl}|^2 \; e^{-2\pi i(hu+kv+lw)}.$$

Furthermore, a Patterson map of *N* points will have *N (N-1) peaks*, excluding the central -origin- peak and any overlap. The peaks positions in the Patterson function are the interatomic distance vectors and the peak heights are proportional to the product of the number of electrons in the atoms concerned. (Glykos, 2015)

From a knowledge of the amplitude and the phase we can determine the S*tructure Factor*, from which the arrangement of the atoms in the unit cell can be calculated. The phase angle cannot normally be determined directly in the case of protein crystals and so must be found in an *indirect* way. The two most frequently used methods are isomorphous replacement and molecular replacement. (Smyth, Martin, 2000)

After solving the phase problem, the next step is the *Fourier analysis* and the calculation of a map, known as *electron density map*. Once the amplitudes and phases are calculated, the Structure Factors -the vectors representing the waves and the information from the crystal structure- can also be calculated. From the calculation of amplitudes and wave phases to the *building* of a protein model, the procedure becomes *strictly computational*.

For the rest of the crystallographic study, there are some things that need to be accounted for, regarding the quality of the data collection. One of them is the Crystallographic Symmetry, which in the current stage, that term is referring to the amount of Symmetry present in the crystal system and Space group.

Another is the Non-Crystallographic Symmetry (NCS); the amount of symmetry present in an asymmetric Unit cell. The ways of amplitudes and phases' determination. And the upper resolution limit required, which is determined by the crystallization procedure usually. With this information the study of maps and model building become more specific.

Based on these, the three-dimensional coordinates for each atom in the target-protein's molecule can be determined via the crystal's symmetrical and structural properties, enhancing in that way the quality of the model building and the *final model* by extension.

2.d. Maps & Computational Studies

An electron density map is like a 3D *grid* with indications on *how to build* the protein model. More specifically, it indicates where the atoms' electrons of the molecule are in the three-dimensional space. (**Figure 26**)

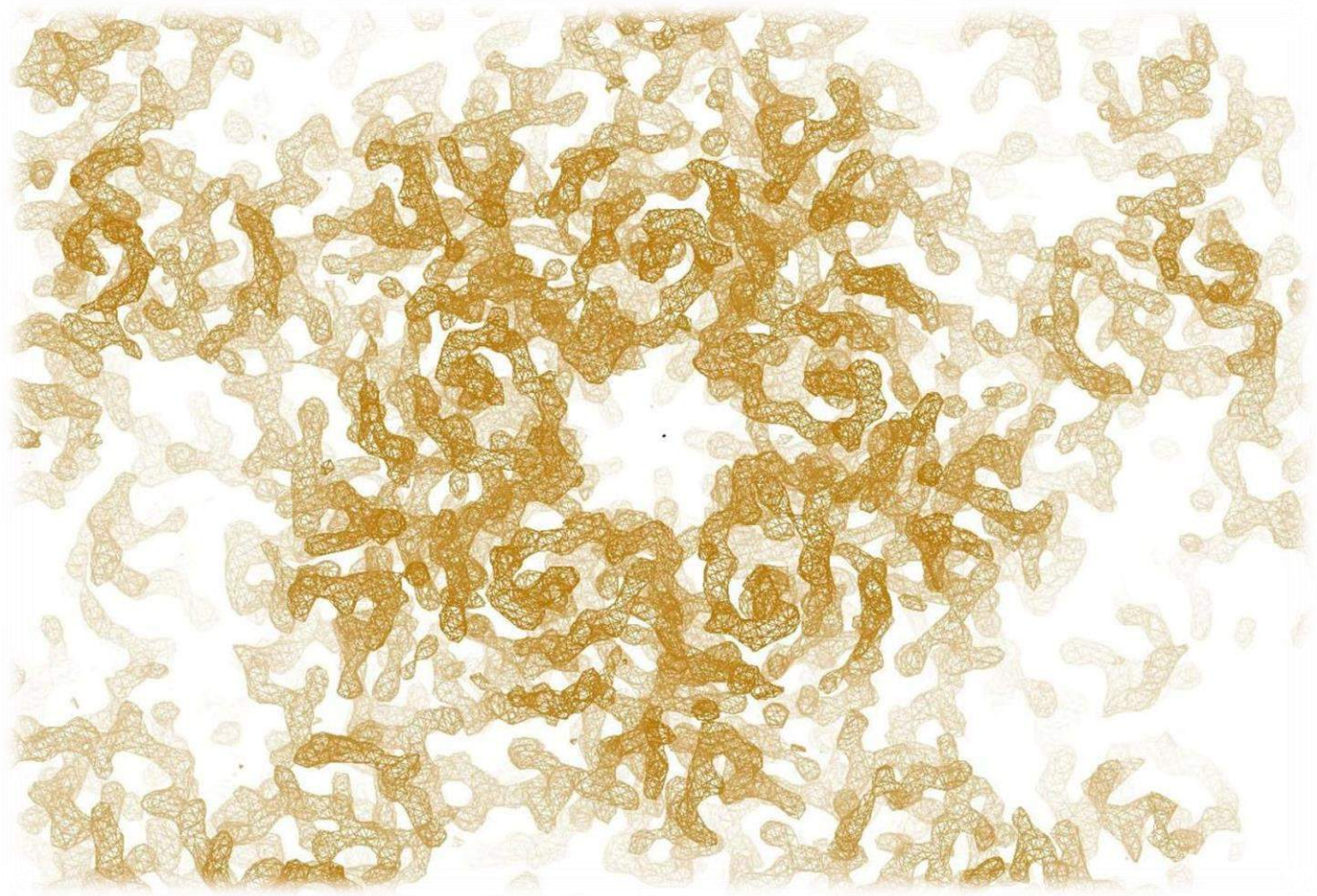


Figure 26| *An electron density map.*

The information of the electron density map is saved in an MTZ *format file*. As mentioned above, from now on the work on the protein structure is strictly computational -unless there are changes in the protein crystal, from which the data was collected.

According to the CCP4 -Collaborative Computational Program No. 4- Program Suite the MTZ format file is used for the storage of reflection data. The file contains the data and a header of metadata -data that provides information for other data. The data section is held as a table with rows presenting reflections and columns representing different quantities for each reflection. The metadata section aims to make the file *self-contained* by including all necessary information, such as symmetry operations, unit cell dimensions etc. The MTZ file is a *flat-file* representation -with rows and columns- of a particular data model and, as mentioned before, the beginning of model building.

As Smyth and Martin stated (2000):

> *"The resulting electron density map will form the three-dimensional contours into which the protein structure will be built. Each of the unit cell edges is divided into spacing of a few Ångstroms. The spacing determines the quality of the map detail and the speed of calculation. This creates a three-dimensional grid within the unit cell and the electron density is calculated at each of the grid points. This computation may be greatly accelerated by applying limits to the size of the grid, according to where the protein molecule is located within the Unit Cell."* (p. 13)

Once the -first- electron density map is fully calculated it is needed to be tested for its *quality*. The meaning of Resolution now makes greater sense; The higher the resolution of the density map, the easier the three-dimensional conformation of the peptide chain can be defined. (**Figure 27.A & B**)

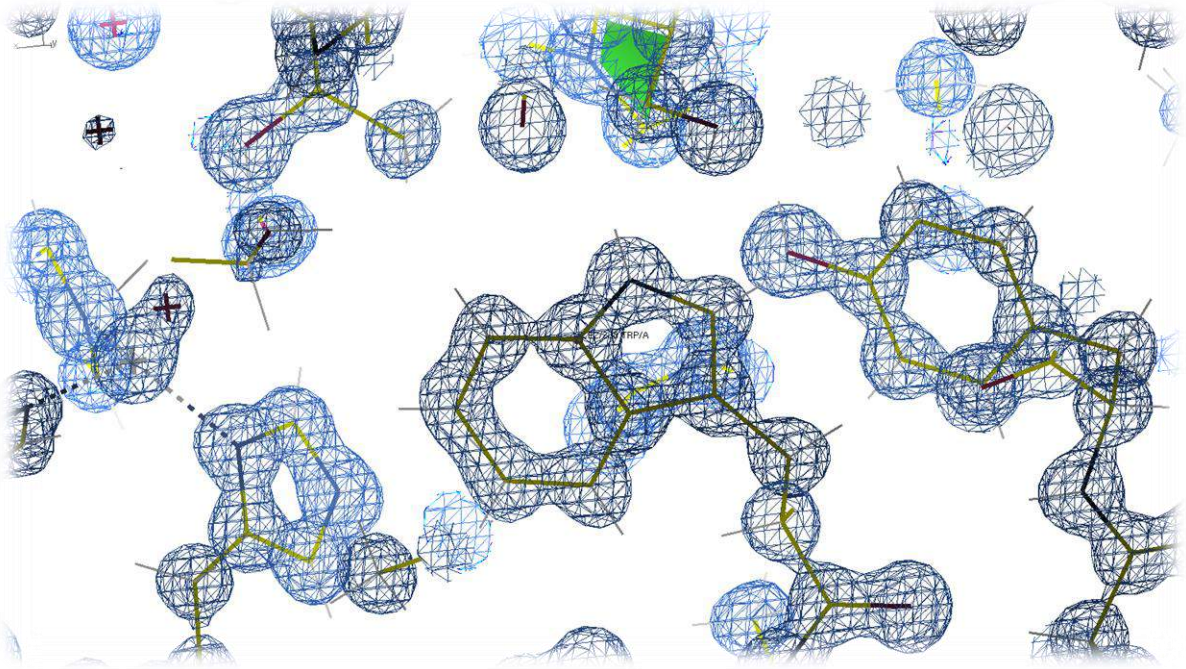


Figure 27.A| *A model with high resolution density map (dark blue grid). At the center of the Figure is the conformation of a Tryptophan residue.*

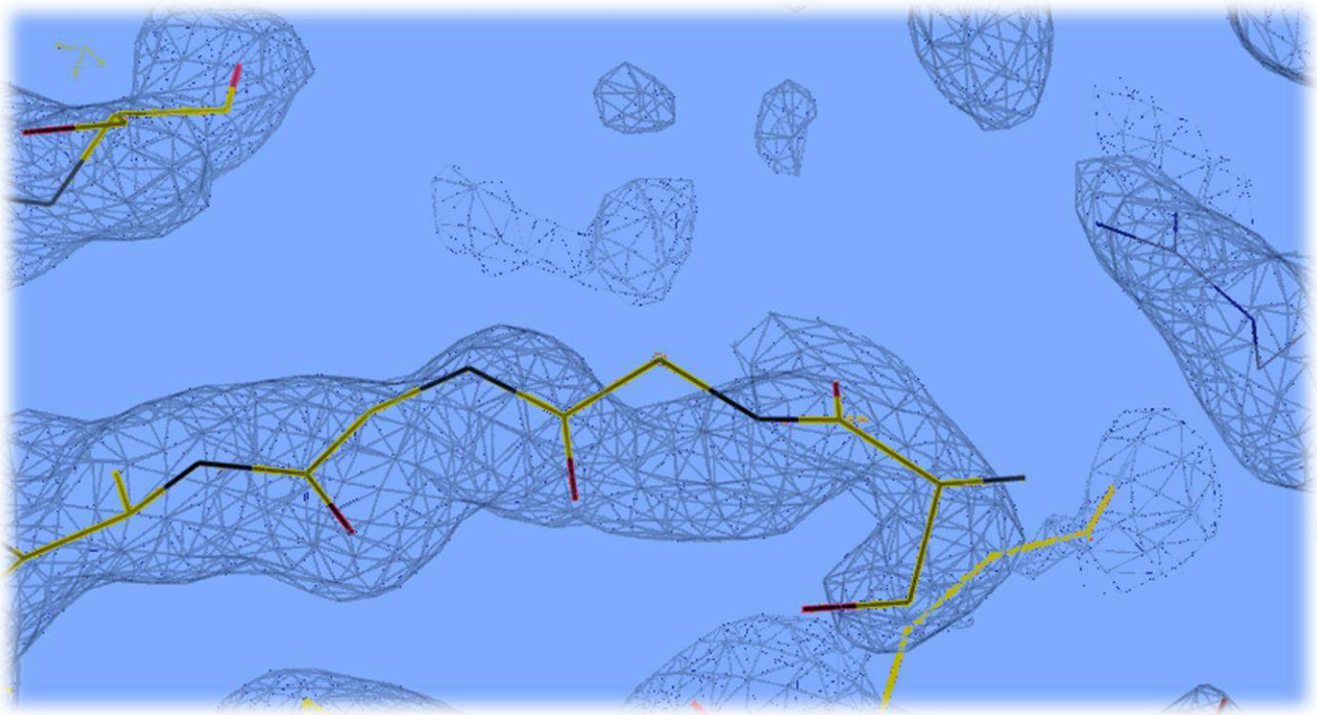


Figure 27.B| *A tail in lower resolution density map. The conformation of the mode's tail is not that clearly defined by the electron density map.*

There are many types of electron density maps that can be used in Computational Crystallographic studies and the building of a protein's model. The maps, which are calculated from the amplitudes and phases from the crystal diffraction or a built model, are known as *Direct Maps*. The Direct maps calculated using the *observed* Structure Factors and the *calculated* phases of the atomic model are symbolized as ***Fo**-maps* (*o* for *observed)*. The maps calculated from Structure Factors and phases from the atomic model are symbolized as ***Fc**-maps* (*c* for *calculated)*.

Another important type of density maps are known as *Difference maps*. Difference maps provide a comparison between a model and the diffraction data, with positive and negative indications, according to where are less or few electrons -which structurally indicate parts of the peptide chain- in a protein's model. These indications may usually look like atomic orbits and are usually depicted as a 3D grid, just like the Direct maps. They are classically symbolized as ***Fo** -**Fc***. (Tronrud, 2015) (**Figure 28**)

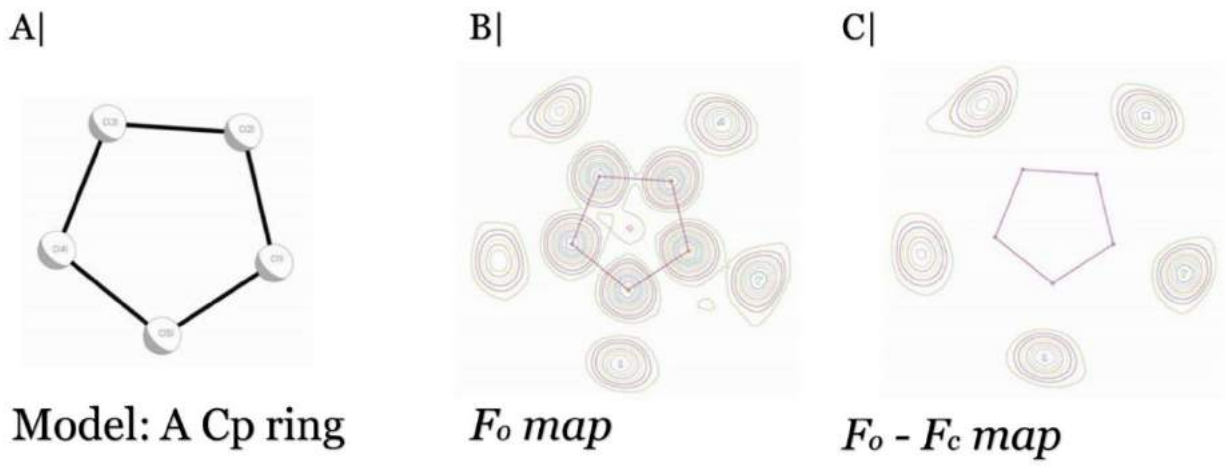


Figure 28| A. *Atomic model of a Cp* ring ligand* **B.** *An **Fo-map*** of that model and **C.** The ***Fo-Fc map*** of the same model. (MIT Open Course Ware, 2008)

Difference maps are combined with the Direct maps to provide better indications for the model building. (**Figure 29**)

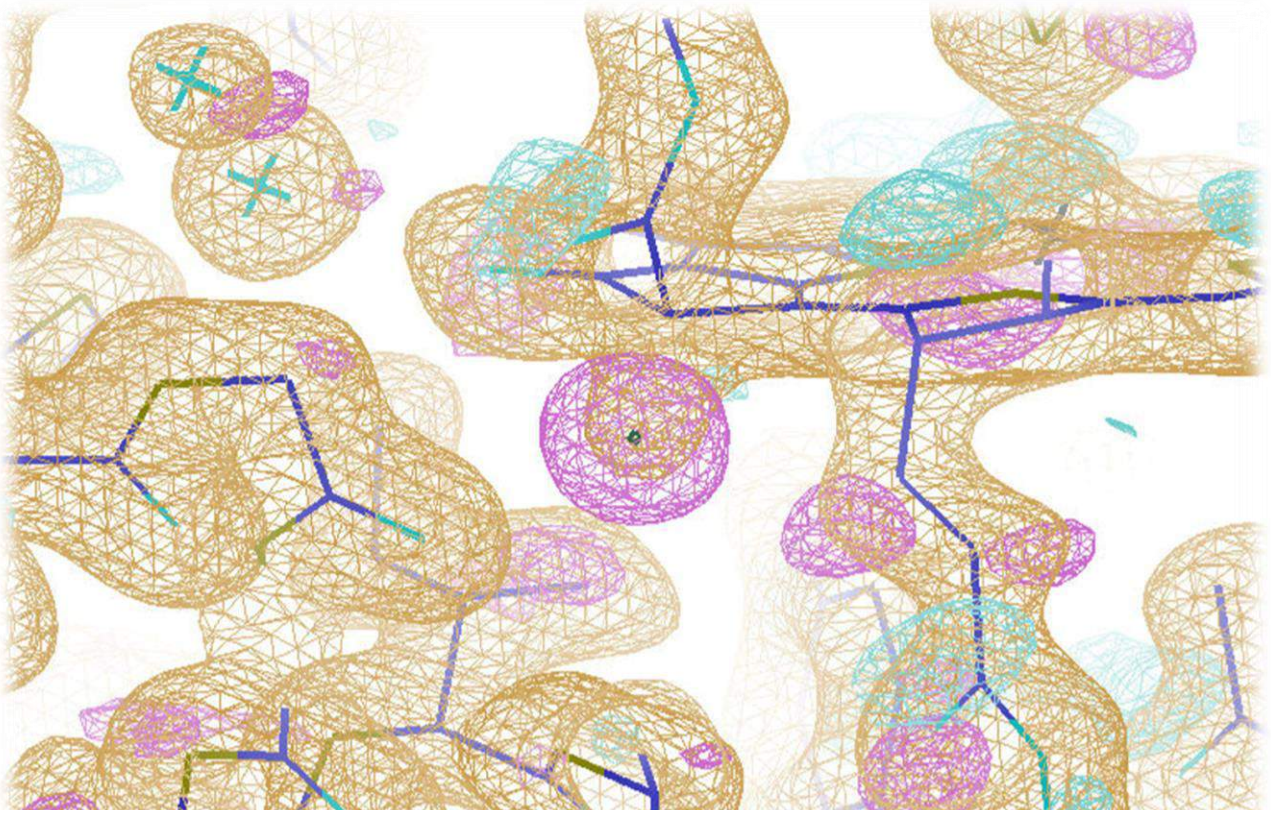


Figure 29| *An atomic model (**blue**) with a Direct map (**yellow**), combined with a Difference Map's positive (**green**) and negative (**pink**) indications.*

Another type of Density map is the *Composite Map*, which is *preferable* combination of Direct and Difference maps and it is usually used for the next steps of the computational methods. Also, the *Demonstration maps* are usually used to emphasize on some points of the model building.

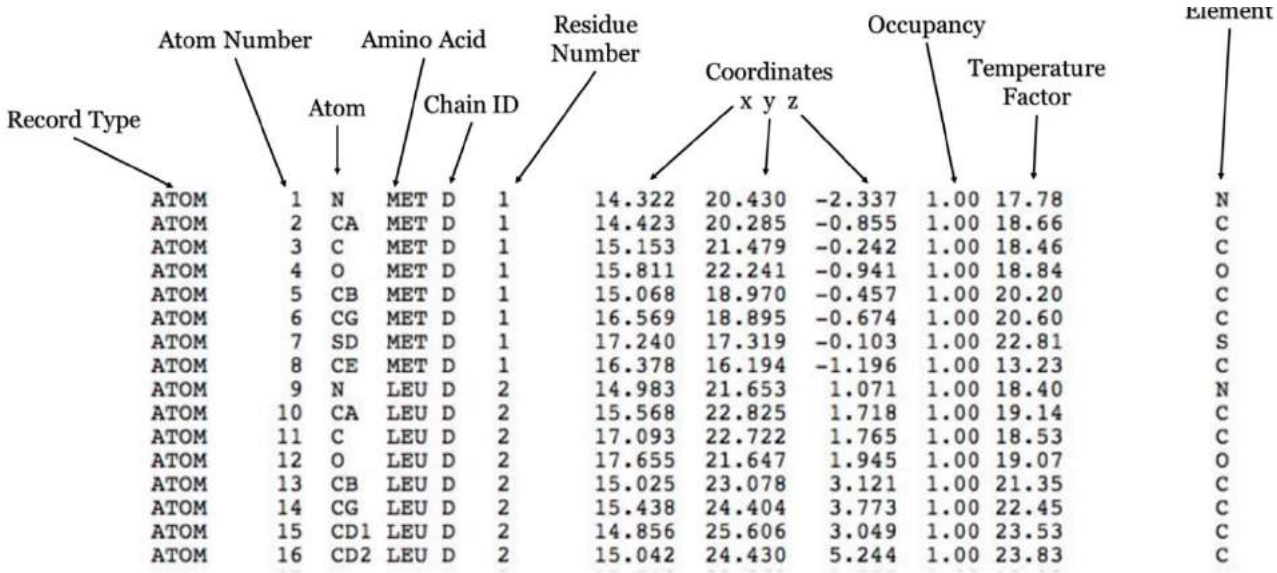
It is extremely important for all the stages of model building to test which types of electron density map are suitable for the study of the structure. And that's why, once an analyst publishes the protein's model, it is important to report the exact conditions of the calculations of the maps, which were used -and not mislead the readers of the publication. (Tonrud, 2015)

For the next step, for the *improvement* of the calculated map's quality, a first model of the 3D protein's structure must be created. This step takes place using a Molecular Graphics Modification program.

One of the most used programs for Molecular Graphics Modifications for Structural Studies is COOT (Emsley, Lohkamp, Scott, Cowtan, 2010). COOT can read the information of the MTZ file and present it as a 3D grid, representing the electron density map in the three-dimensional space -just like the Figures above. Apart from viewing, there is also the potential of building a model according to the indications of the electron density map. So, based on the 3D grid, a model backbone can be built and saved as a file. The file format for models in Molecular Biology is called PDB.

PDB format is found in the published molecules' structures in the Protein Data Bank database, hence the name of the format. That format provides a standard representation for macromolecular structure data derived from X-ray diffraction and NMR studies. This representation was created in the 1970's and is used by many Structural programs. From viewing, to many types of modifications and studying.

Table 1| *The format of a PDB file.*



ATOM	1	N	MET	D	1	14.322	20.430	-2.337	1.00	17.78	N
ATOM	2	CA	MET	D	1	14.423	20.285	-0.855	1.00	18.66	C
ATOM	3	C	MET	D	1	15.153	21.479	-0.242	1.00	18.46	C
ATOM	4	O	MET	D	1	15.811	22.241	-0.941	1.00	18.84	O
ATOM	5	CB	MET	D	1	15.068	18.970	-0.457	1.00	20.20	C
ATOM	6	CG	MET	D	1	16.569	18.895	-0.674	1.00	20.60	C
ATOM	7	SD	MET	D	1	17.240	17.319	-0.103	1.00	22.81	S
ATOM	8	CE	MET	D	1	16.378	16.194	-1.196	1.00	13.23	C
ATOM	9	N	LEU	D	2	14.983	21.653	1.071	1.00	18.40	N
ATOM	10	CA	LEU	D	2	15.568	22.825	1.718	1.00	19.14	C
ATOM	11	C	LEU	D	2	17.093	22.722	1.765	1.00	18.53	C
ATOM	12	O	LEU	D	2	17.655	21.647	1.945	1.00	19.07	O
ATOM	13	CB	LEU	D	2	15.025	23.078	3.121	1.00	21.35	C
ATOM	14	CG	LEU	D	2	15.438	24.404	3.773	1.00	22.45	C
ATOM	15	CD1	LEU	D	2	14.856	25.606	3.049	1.00	23.53	C
ATOM	16	CD2	LEU	D	2	15.042	24.430	5.244	1.00	23.83	C

A PDB file is flat-file type, like the MTZ. There is a *Header*, explaining the details of the experimental procedure of the data extraction and the *data* of the model, categorized in columns as shown in **Table 1**.

Now that there is an MTZ file with the data referring to the electron density map and a PDB file with the data referring to the protein's backbone model, it is time for the improvement of the density map's quality, using the method of *Refinement*.

Refinement is a general term that refers to almost all the operations needed to develop a trial model into one that best represents the observed data. In Crystallographic terms it is a way of optimizing the model & map parameters.

As Watkin (2008) stated, referring to Refinement:

> *"Just as there is not a well-defined mathematical technique for extracting valid phases from the observed intensities, so also there is no single well defined path from the trial model to the completed structure -if there were, it would have been programmed long ago."* (p. 1)

Beyond a mathematical and computational procedure, Refinement is needed to be used based on the empirical background of each analyst. *"Refinement is a step-wise procedure, with increasingly subtle features being introduced",* in order to develop a protein's model. (Watkin, 2008)

But how Refinement in Computational Crystallography works?

In order Refinement to be succeded, the protein's structure model must be varied, to achieve the best agreement between the observed reflection amplitudes (***Fobs)*** and those calculated from the model (***Fcalc)***. This agreement is judged by the residual or crystallographic R-factor (***R***). (Wlodawer, et al., 2008)

The R-factor in crystallography is a measure of the agreement between the crystallographic model and the experimental X-ray diffraction data. In other words, it is a measure of how well the refined structure predicts the observed data. In mathematical context, the value is also sometimes called the discrepancy index, as it mathematically describes the difference between the experimental observations and the ideal calculated values. It is defined by the following equation:

$$R = \frac{\sum \left| |F_{\text{obs}}| - |F_{\text{calc}}| \right|}{\sum |F_{\text{obs}}|}$$

The minimum ideal value is *zero* (0.0), indicating perfect agreement between experimental observations and the structure factors predicted from the model. There is no *theoretical maximum*, but in practice, values are considerably less than *one* (1.0) even for *poor models*. Random experimental errors contribute to ***R*** even for a perfect model and these have more leverage when the data is *weak* or not enough -for example low resolution data set. Model inadequacies such as incorrect or missing parts and model's disorder are the main contributors to ***R***, making it useful to assess the progress and result of a crystallographic model refinement. For large molecules, the R-factor usually ranges between 0.6 (when computed for a random model and against an experimental data set) and 0.2 (a well-refined macromolecular model at a resolution of 2.5 Å). Small molecules usually form better-ordered crystals than large molecules and thus it is possible to attain lower R-factors.

Crystallographers also use the Free R-Factor (***Rfree***) to assess possible issues in the model of the data. ***Rfree*** is computed according to the same formula given above, but on a small, random sample of data, which are set aside for that purpose and never included in the refinement. ***Rfree*** will always be greater than ***R***, because the model is fitted to the reflections that contribute ***Rfree***, but the two statistical values should be similar because a correct model should predict all the data with accuracy. If the two values agree, the observed and calculated data from the model are probably close to the protein's native conformation. If the two values differ significantly then that indicates the model has been over-parameterized, so that to some extent it predicts not the ideal error-free data for the correct model, but rather the error-afflicted data, which is actually observed.

The quantities ***Rsym*** and **Rmerge** are similarly used to describe the internal agreement of measurements in a crystallographic data set. (Brunger, 1992)

Once Refinement is complete, the phases are *corrected*, in order the observed and calculated values to agree in a greater grade. So, a new density map – MTZ file- and model -PDB file- are calculated, probably with *higher precision*. From the new electron density map, more accurate atomic positions can be derived, which lead to even better phase angles, and so forth. In every such cycle, adjustments to the atomic model are made -atom types are changed, missing atoms are introduced, etc. That process is called *Structure Refinement*.

Two methods are widely used in refinement: *Maximum Likelihood* and *Simulated Annealing*. Both methods use restraints to how an atomic model is built based on the bond distances, angles and torsions and *temperature factors*, usually referred to as B-*factors*.

Regarding the importance of the temperature factor; in crystallography, uncertainty in the positions of atom increases with disorder in the protein crystal. Resolution usually represents the average uncertainty for all atoms. The temperature value, B *factor*, quantitates the uncertainty for each atom. At typical resolutions for protein crystals, a high temperature factor reflects a low empirical electron density for the atom. Generally, a temperature value of less than 30 Å^2 signifies confidence in its position, while temperature value of

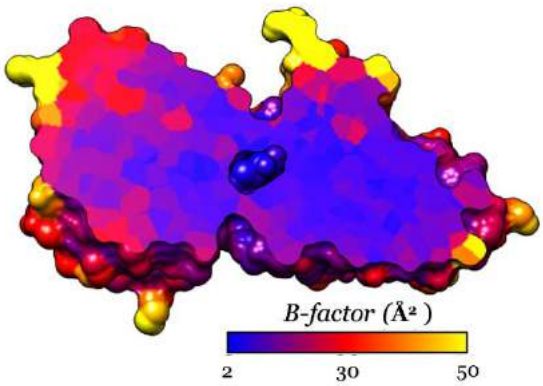


Figure 30| *B-factor representation of Galactose/Glucose Binding Protein (PDB entry: 2gbp) with Chimera Program.*

greater than 60 Å² signifies disorder. (Wlodawer, et al., 2008) B-*factor* can be depicted in molecular viewing programs, like Coot. (**Figure 30**)

Back to the refinement methods, in Maximum Likelihood the phases are adjusted to minimize the ***R***-factor. Maximum likelihood estimation (MLE) is a method of estimating the parameters of probability distribution by maximizing a likelihood function, so that under the assumed statistical model, the observed data is most probable. (Rossi, 2018)

In Simulated Annealing the structure is "*heated*" -to add randomness- and slowly *cooled* and refined. The randomness reduces the probability of falling into a wrong *local minimum*. The products of Simulated Annealing calculations are known as Omit maps, which are more precise electron density maps. (**Figure 31.A & B**)

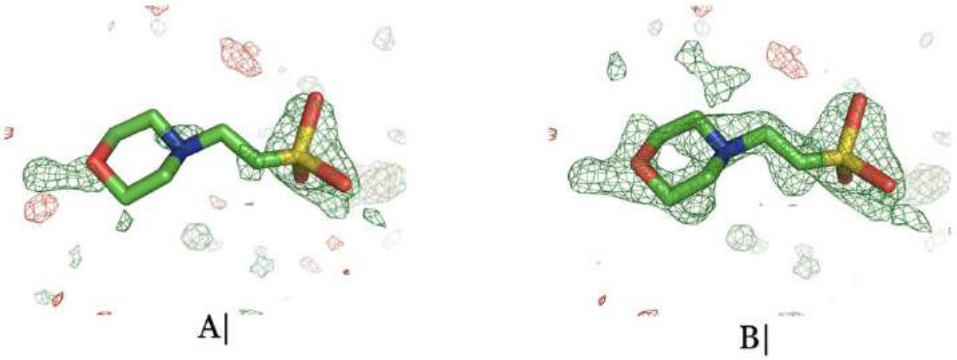


Figure 31| A. *A model with low resolution density map &* **B.** *The same model with an omit map.*

One of the most commonly used programs for the Refinement of a model is the Refmac5 of the CCP4 Program Suite (Vagin, et al., 2004). Refmac5 *utilizes* as input the MTZ and PDB file and creates as output the new *refined* map and model, in new MTZ and PDB files. With a Refinement program it is easier to calculate and observe in *real-time* the optimization procedures and the error-related data of the model. The user has also the option of changing the parameters of the Refinement procedure.

One of these parameters, which usually are used to improve the quality of map, are the *Non-crystallographic symmetry restraints* (NCS), which are used to produce an *average* from the different phases in symmetrical parts -different molecules- of the *asymmetric* unit. One other is the *Density Modification* (DM), which aims to adjust the density to the expectations of how it should generally look like. The solvent does not diffract normally and the electron density should therefore be *zero* in the solvent region.

Also, in proteins often the temperature factors are averaged in all dimensions -isotropic- instead of individual -anisotropic. TLS refinement -*Translation*, *Libration* (small movements) and *Screw-rotation* of a group of atoms- can give a good approximation of anisotropy with much fewer parameters. The TLS group of atoms can be an entire molecule or a domain. All these parameters are available in Refinement Programs, like Refmac5. (**Figure 32**)

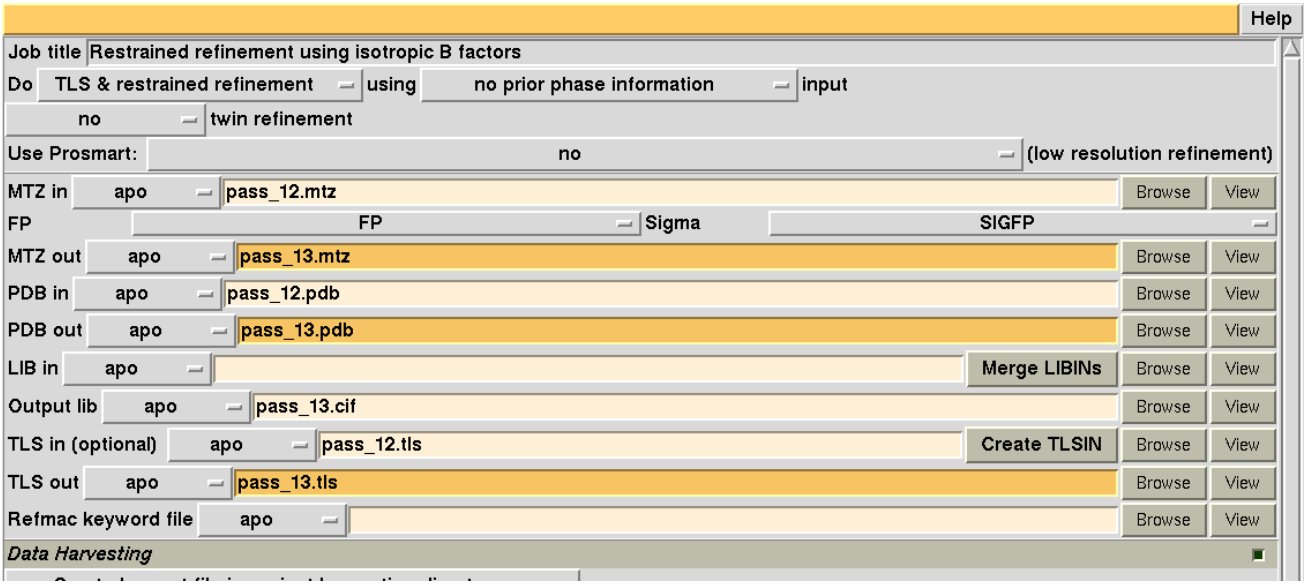


Figure 32| *The start settings of a Refmac5 Run.*

Once the first refined maps are calculated, a full protein's model can be built with the protein's sequence, almost automatically. Each amino acid needs to be fitted accordingly to the density maps' indications and the residues' biochemical properties. This is one of the most empirical parts of the whole process and it is called "*Model Fitting*".

The structural analyst has to take in consideration most of the Structural and biochemical properties of each amino acid on the peptide chain and validate the already calculated model and map. Each time a *Model Fitting* stage ends, comes the *Refinement* process of the corrected model and back again to *Model Fitting*. Further computational studies may take place if there are issues, as for example regions with high mobility -like *big loops*- or other properties of the molecules. At the end of these *Fitting-Refinement cycles* a most precise model of a protein's structure is almost complete.

The last step before the *publication* of a protein's structure in the Protein Data Bank is the step of *Validation*. The main reason the step of validation is needed, is because *a model stays a model* even if it is *built* at high resolution and if *fits* the electron density *well*. Even after all these processes, there are still many sources of error -experimental or due to wrong interpretation- during the model building. Validation methods *detect* inconsistencies in the final model based on information that was not used during the refinement process.

At the same time validation may be seen as an additional step of Refinement. When something really does not seem right, through Validation there is the possibility to go back and check that region. One of the most common Validating tools, before publishing in the Protein Data Bank, is the PDB-REDO. (Joosten P., Joosten K., Murshudovb, Perrakis, 2012) (**Figure 33**)

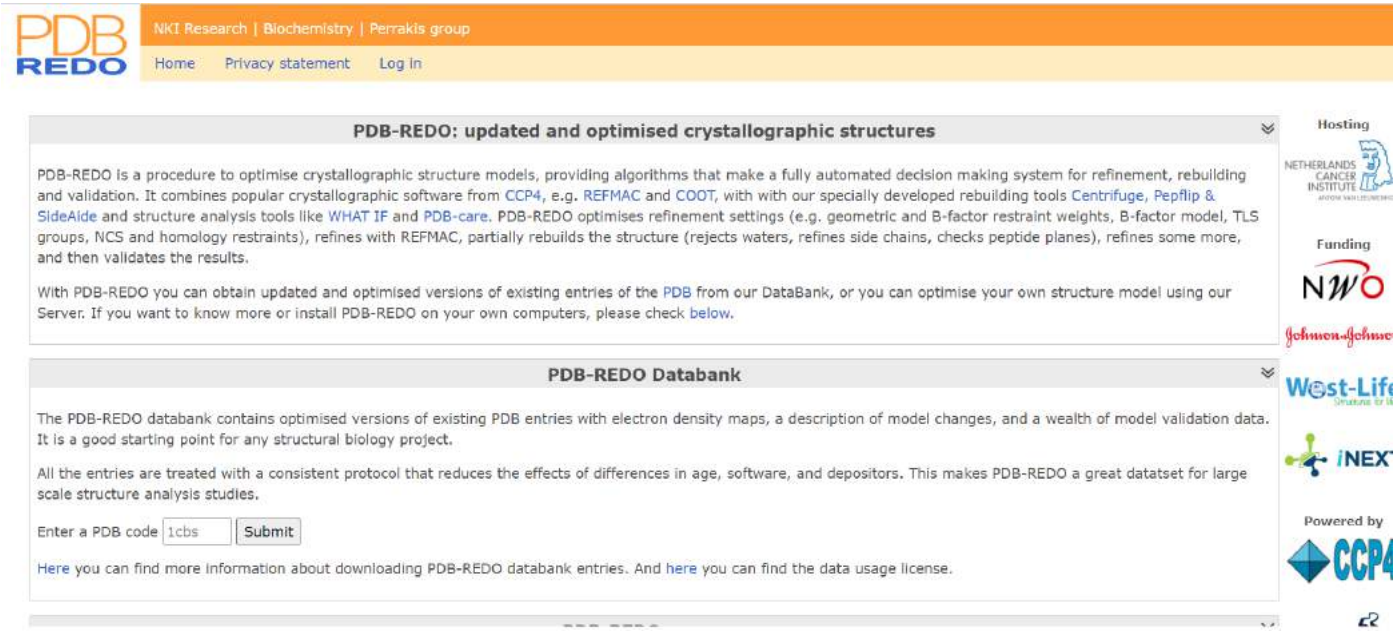


Figure 33| *The PDB-REDO interface.*

To sum up the main points for X-ray Protein Crystallography, producing an image from a *diffraction pattern* requires sophisticated mathematics and often an iterative process of *modelling* and *refinement*. In this process, the mathematically predicted diffraction patterns of a hypothesized structure -or "model", hence modelling- are compared to the actual pattern generated by the crystalline sample. Ideally, researchers make several initial guesses, which through the process of refinement all converge on the same answer.

Models are refined until their predicted patterns match to as great a degree as can be achieved without radical revision of the model. This is an agonizing process, made a lot easier today by computers. (**Figure 34**)

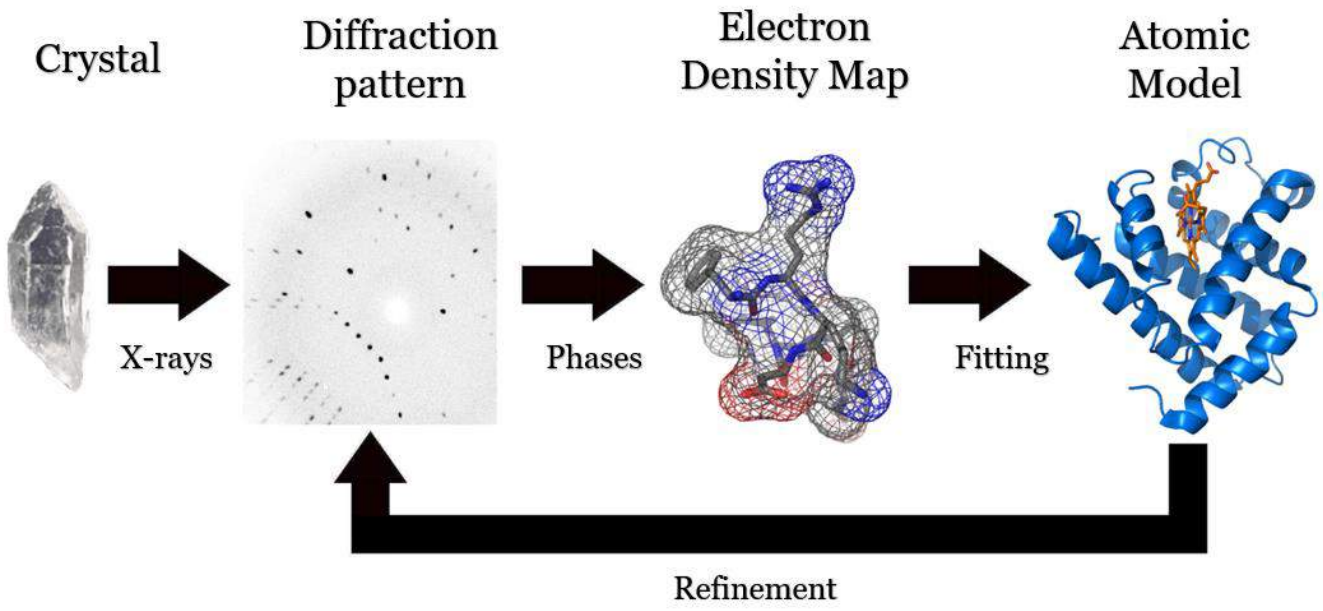


Figure 34| *The whole procedure to the definition of a 3D model of a protein's structure from a protein Crystal.*

3. DNA-methyltransferase…

From all the proteins that have been discovered and from those which have not, my interest regarding this project is one, belonging in the famous protein family, known as DNA-methyltransferases. These types of proteins are categorized, firstly, as *enzymes*.

Enzymes are proteins that act as biological catalysts. Catalysts accelerate chemical reactions. The molecules, upon which enzymes may act, are called *substrates* and the enzyme converts the substrates into different molecules, known as *products*. Almost all metabolic processes in the cell need enzyme catalysis, to occur at rates fast enough to sustain life. (Berg, John, Stryer, 2007)

3.a. Function & Classification

DNA-methyltransferases are enzymes that catalyze the transfer of *a methyl group* to DNA. DNA methylation serves a wide variety of biological functions. Basically, through methylation can change the activity of a DNA segment without changing the sequence. When located in a gene promoter, DNA methylation typically acts to repress gene transcription.

Two of DNA's four bases can be methylated; Cytosine and Adenine. Cytosine methylation is widespread in both eukaryotes and prokaryotes, even though the rate of Cytosine DNA methylation can differ greatly between species. Adenine methylation has been observed in bacterial, plant and recently mammalian DNA, but has received considerably less attention. All the known DNA methyltransferases use *S-adenosyl methionine* (SAM) as the methyl donor. (**Figure 35.A & B**)

In mammals, DNA methylation is essential for normal development and is associated with several key processes, such as aging and carcinogenesis.

In many bacteria, Adenine or Cytosine methylation is part of their restriction modification system, in which specific DNA sequences are methylated periodically throughout the genome. A methylase is the enzyme that *recognizes a specific sequence* and *methylases* one of the bases in or near that sequence. Foreign DNAs -which are not methylated- that are introduced into the cell are degraded by sequence-specific restriction enzymes and cleaved. Bacterial genomic DNA is not recognized by these restriction enzymes. The methylation of native DNA acts as a sort of primitive immune system, allowing the bacteria to protect themselves from infection by bacteriophage -bacteria viruses.

According to which nucleobase, and in which place of that nucleobase the catalysis of methylation occurs, the methyltransferases are classified into three groups. The m6A (those that produce *N6-methyladenine*), the m4C (those that generate *N4-methylcytosine*) and the m5C (those that generate *C5-methylcytosine*). (Rina, Markaki, Bouriotis, 1994)

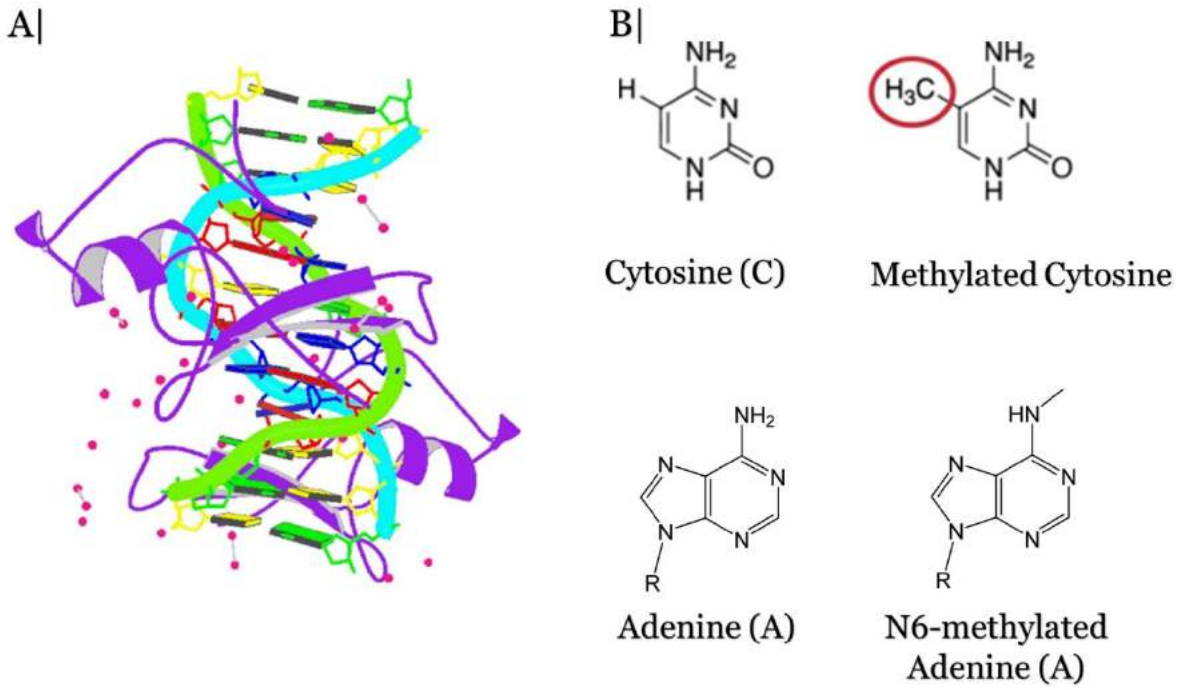


Figure 35| A. *The 3D structure of a DNA-methyltransferase with a DNA complex and* **B.** *Some methylations of Cytosine (above) and Adine (below).*

3.b. Base Flipping

The DNA-methyltransferases present multiple ways of binding on the DNA molecule and catalyze the methylation, according -of course- to the special characteristics that each one has, based on its class. But there is one characteristic methylation mechanism, quite known as *Base Flipping*.

Base flipping became known from the studies of the structure of the enzyme HhaI, which is a DNA Cytosine-5-Methyltransferase (it catalyzes the methylation of the 5th Carbon of a Cytosine in DNA). More specifically when M. HhaI interacts with its substrate DNA, the *target cytosine* is *flipped* completely *out* of the DNA helix and *into* a *cavity* in the enzyme where the chemistry of catalysis takes place (**Figure 36.A & B**). As Roberts (1995) states in his review "On Base Flipping":

> *"This dramatic but elegant distortion in the DNA structure contrasts sharply with the kinks and bends induced by other proteins upon binding. Remarkably, there is no external energy supply required during this base flipping process since it takes place in the presence of only protein, DNA, and the cofactor S-adenosylmethionine."* (p. 9)

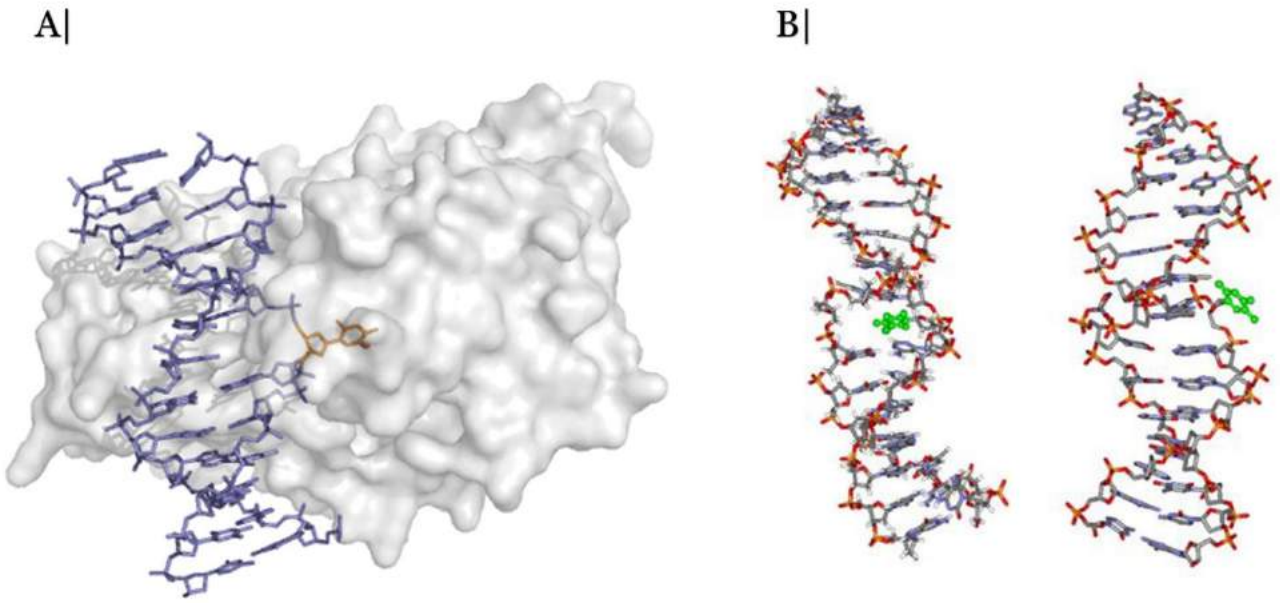


Figure 36| A. *The M. HhaI base flipping* **B.** *The change of the DNA conformation during Base Flipping. (Copyrights reserved to Dr. Olaf Wiest, University of Notre Damme- chemistry.nd.eu)*

DNA nucleotides are held together with hydrogen bonds, which are relatively weak and can be easily broken. Base flipping occurs on a millisecond timescale by breaking the hydrogen bonds between bases and unstacking the base from its neighbors. The base is *rotated* out of the double helix by 180 degrees -typically via the major groove of the double helix- and into the active site of the enzyme. This opening leads to *small conformational changes* in the DNA backbone, which are quickly stabilized by the increased *enzyme-DNA interactions*. Studies looking at the *free energy* (*ΔG*) profiles of base-flipping have shown that the free energy barrier to flipping can be lowered by 17 kcal/mol for M. HhaI in the closed conformation.

More studies have shown that DNA base flipping is used by many different enzymes in a variety of biological processes, such as DNA methylation, various DNA repair mechanism, RNA transcription and DNA replication. (Huang, Banavali, MacKerell, 2002)

Roberts (1995) and his colleagues, already from the first crystals of M. HhaI in 1994, had connected this mechanism to evolutionary advantages in the molecular processes of the organisms it is present, and foreseen the importance of the studies of such mechanism. In his review's conclusion he stated:

> "More examples [of enzymes using the Base Flipping mechanism] lie in the wings, and there are good reasons to think that base flipping may be quite widespread. Just 2 years ago, this proposal would have been greeted with great skepticism, perhaps even laughter. It seems much more plausible today. Of course, that is at once the great challenge and the great joy of molecular biology, where much remains to be discovered." (p.12)

Today, Base Flipping is one of the most studied mechanisms in the Methyltransferases' *world*. And it is connected to multiple functions. That is the main reason of the Structural study of this particular Methyltransferase. The DNA-Methyltransferase I was working on, presents this quite interesting mechanism to catalyze the methylation of Adenine.

The whole Introduction Section -up until now- was a prologue to the presentation of this DNA-Methyltransferase's structure, which -as mentioned above- is the continuity of the molecular studies of the biological mechanism, known as Base Flipping.

So, *without further ado*, I present to you…

3.c. Modification Methylase BsecI…

…also known as Adenine-specific methyltransferase BsecI, or BsecIM for short (Rina, Bouriotis, 1993).

BsecIM is a methyltransferase, encoded by the *bseCIM* gene, isolated by the *Geobacillus stearothermophilus* (*Bacillus stearothermophilus*) bacterium.

Geobacillus stearothermophilus -previously *Bacillus stearothermophilus*- is a rod-shaped, Gram-positive bacterium. It is a *thermophile* -it is developed in relatively high temperatures- and is widely distributed in soil, hot springs, ocean sediment, and is a cause of spoilage in food products. It will grow within a temperature range of 30 to 75°C. Some strains are capable of oxidizing carbon monoxide aerobically. It is commonly used as a challenge organism for sterilization validation studies and periodic check of sterilization cycles. (Donk, 1920) (**Figure 37**)

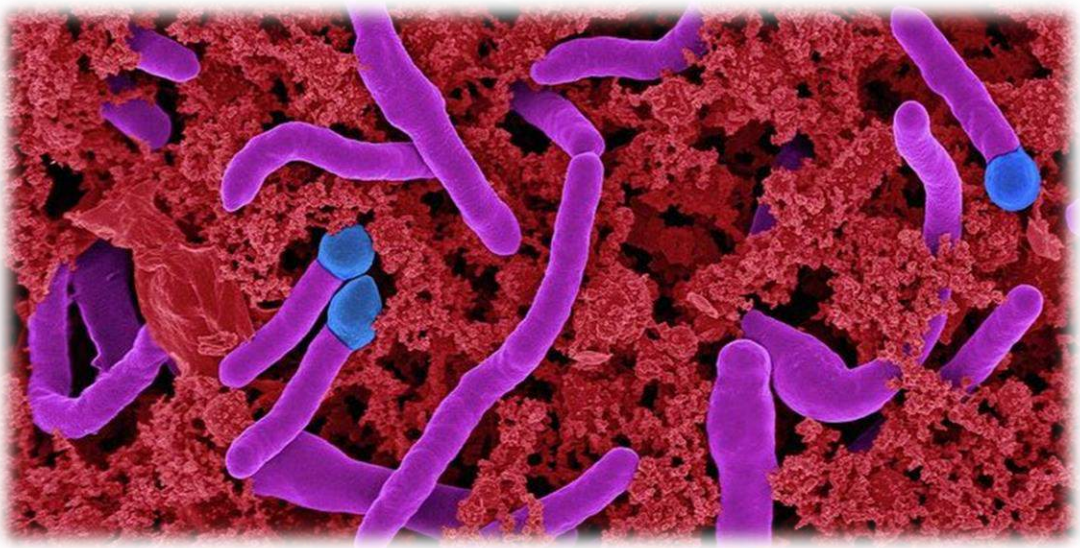


Figure 37| *G. stearothermophilus picture. (Copyrights to Dennis Kunkel)*

BsecIM is connected with defensive mechanisms for the bacterial DNA against virus intrusion and DNA cleavage mechanisms. From studies of its function, this methylase recognizes the double-stranded sequence ATCGAT, causes specific methylation on the 5th Adenine on both strands, and protects the DNA from cleavage by the BanIII *endonuclease*. (**Figure 38.A**)

A|

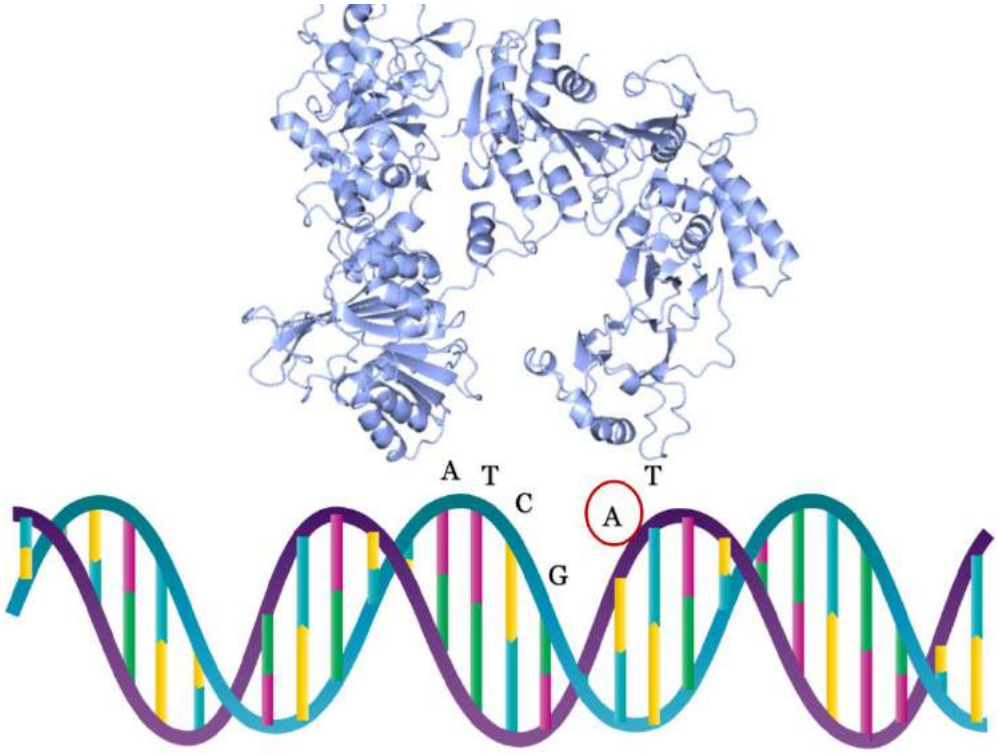


Figure 38. A| *The BsecIM recognition and catalysis region.*

Once it forms the complex with the bacterial DNA double helix, BsecIM *flips* the 5th Adenosine of the recognition sequence and the catalysis of methylation begins. The *ingredients* for the catalysis are the 5th Adenosine of the recognition sequence and an S-adenosyl-L-methionine, as a donor of the methyl-group. The *products* of this reaction are an N6-methyl-2'-deoxyadenosine in DNA, an H+ and an S-adenosyl-L-homocysteine. (**Figure 38.B**)

B|

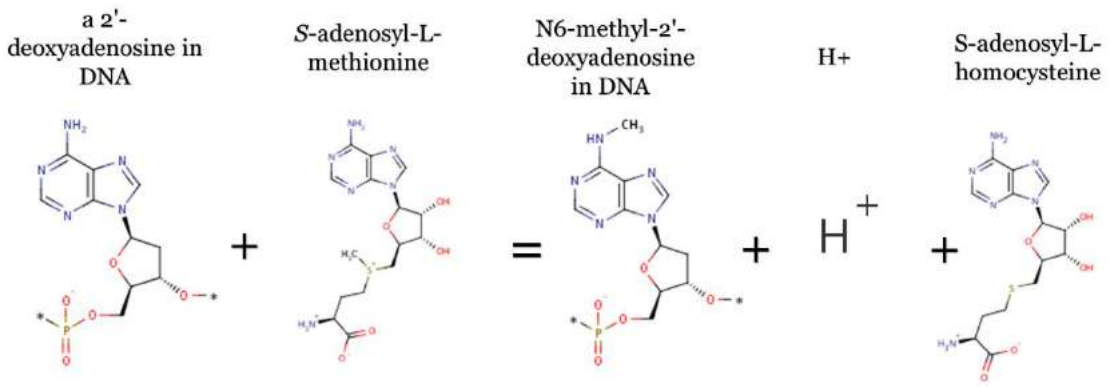


Figure 8.B| *The chemical reaction of the BsecIM methylation.*

From a Structural point of view, BsecIM is a *protein dimer*. More speficially, the protein's Quaternary Structure is formed by two identical peptide chains folding and interacting with each other, producing the final structure of BsecIM.

This is quite an important information for the crystallographic studies of a protein. The fact that two same peptide chains form one molecule can be used in multiple ways for the model building of the protein. More accurate data from one chain can be very useful for the studies of *poor electron density* regions on the other chain.

Of course, BsecIM is respectively a *large* molecule. Each peptide chain consists of 579 amino acids. So, there are many regions, which are more flexible -such as loops- and makes even more difficult the whole model building and refinement process. To build an accurate model of this protein, multiple strategies and approaches were needed to be followed.

That is the main goal for this project; the Determination and Optimization of Modification Methylase BsecI's structure, through computational crystallographic studies. As mentioned above, that is a main step into the further analysis and determination of the enzyme's function, which may be followed by further structural studies on the BsecIM- DNA complex. But regarding this project, the optimization of the final 3D model of BsecIM which its function is connected to the Base Flipping mechanism, completes the first step in the enzyme's analysis and study, and *makes way for* the rest steps to follow.

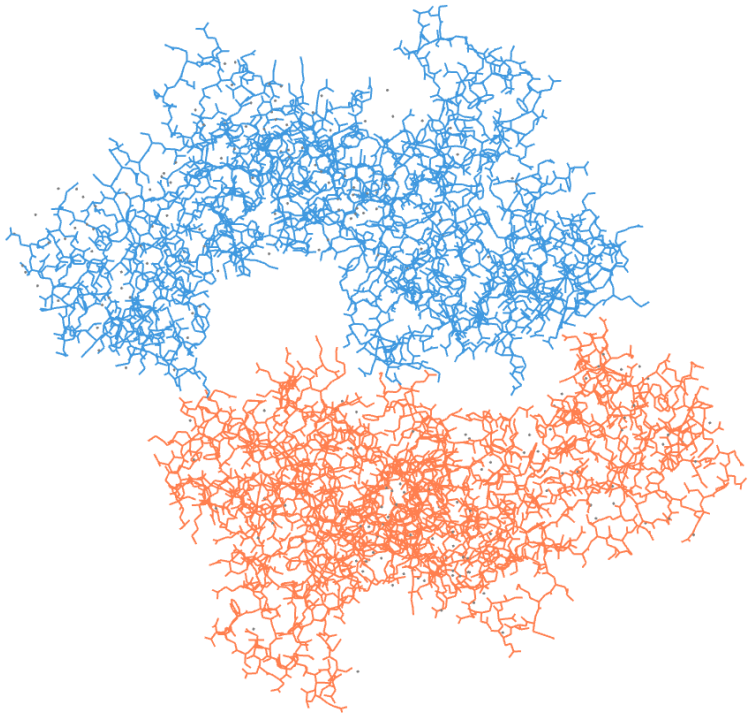
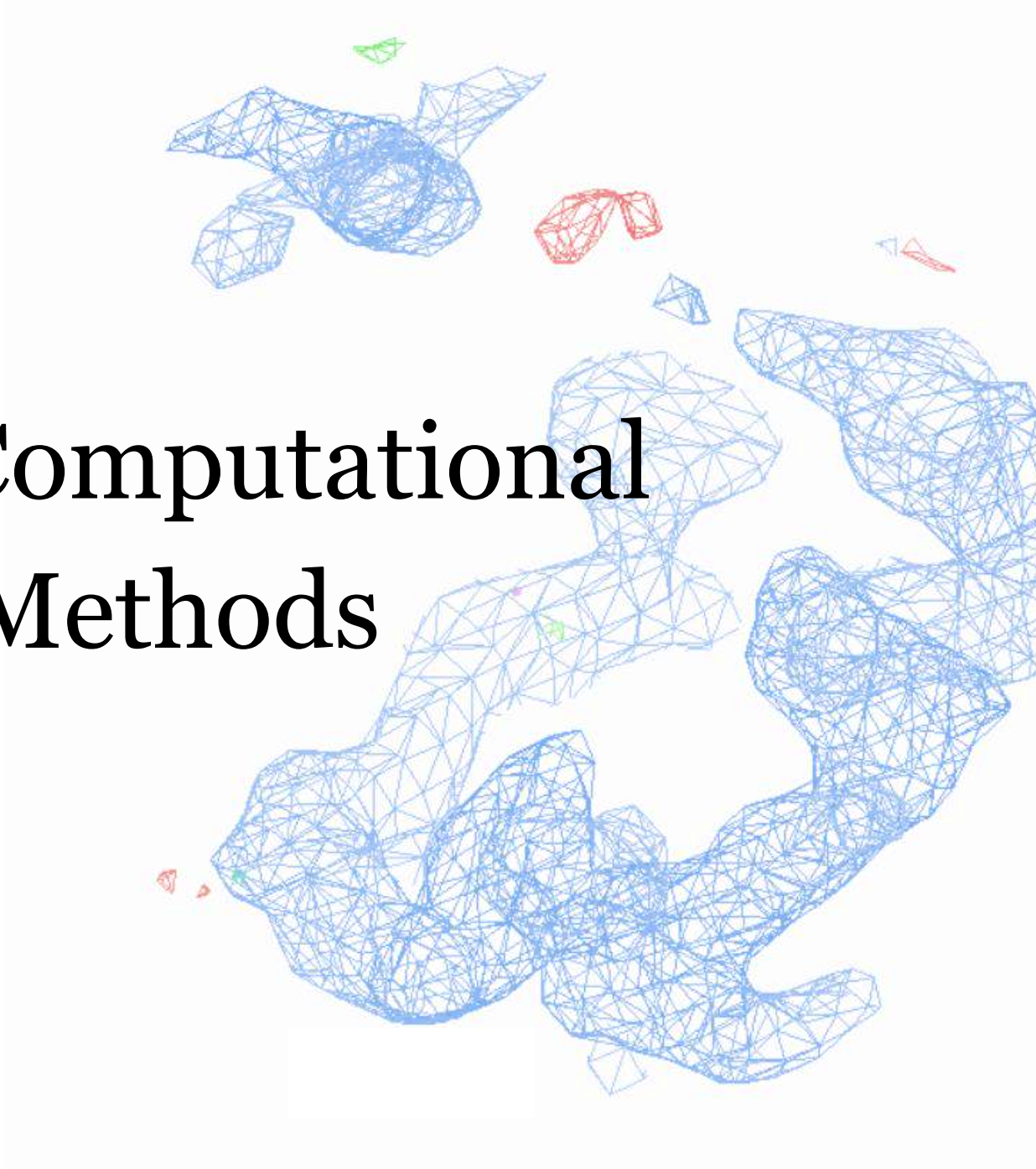


Figure 39| *The whole BsecIM model. (Each chain is colored differently).*

II. Computational Methods



Main goal…

The Computational Crystallographic Studies of BsecIM started with the already *solved* phases for the molecule's X-ray crystallographic data and an already *built first* model for the enzyme. So, you may ask: "What is the meaning of this thesis?".

For the Completion and Optimization of the molecule's 3D model, first, some *unmodeled* loop regions needed to be built. As mentioned in the Introduction Section, loop regions *tend to be quite flexible* in proteins' structures, which probably means that they are low electron density indications, and therefore it is quite difficult to determine a specific structure for these regions.

Second, to optimize the model, every residue needed to be in agreement with the protein's sequence and electron density map's indications. And third, some missing solvent/water molecules needed to be added to the final BescIM's model, representing the molecule's interactions with the environmental conditions.

So, the main goal for this project was the optimization of the protein's model through model fitting and missing regions building processes, and through multiple Refinement cycles. In that Section I will present the different Computational Methods I followed, to succeed a better model's optimization.

0. Starting Point…

This project started with a PDB, an MTZ and a TLS file for the protein's already built and refined first model. The resolution for this protein's model was at 2.5 Ångström (**Å**), with regions with higher and lower electron density indications, as described in the model's data. Higher electron density indications were usually connected to Secondary Structures, and lower to the high mobility missing loop regions.

0.i. Starting Files & Data

The files I used for this project were named numerically, according to the stage of model optimization and study. In that way it is easier to manage, track and present the results from each stage. The starting files are named **file_0.pdb**, **.mtz** and **.tls** accordingly.

The starting model was already refined with Refmac5 Program of the CCP4 Suite, using TLS (*Translation, Liberation* and *Srew-Rotation*) parameters, hence the TLS file format, which contains the corresponding data for these parameters.

The MTZ file, containing the electron density data, was produced by merging two different maps; A Direct Map for the molecule and a Difference ***2mFo-DFc*** Map, with multiple positive and negative indications. Information about the Crystal Symmetry and Resolution characteristics is also included in the MTZ file.

Table 2 | *General data included in the* ***file_0.mtz*** .

MTZ Type:	Merged (Direct Map & Difference Map ***Fo –Fc)***
Space Group:	P 1 21 1
Unit Cell:	53.7 85.7 151.8 90 95.1 90
Number of Lattices:	**1**
Number of Reflections:	**47,621**
Low Resolution:	**24.71**
High Resolution:	**2.5**

In **Table 2** is presented the general data included in the **file_0.mtz**. The BsecIM's crystal Space Group, which is P 1 21 1, the Unit Cell dimensions, and the Number of Lattices is an important information for the protein's model optimization. The way that the protein's model is contained in that specific Unit Cell is shown in **Figure 40**.

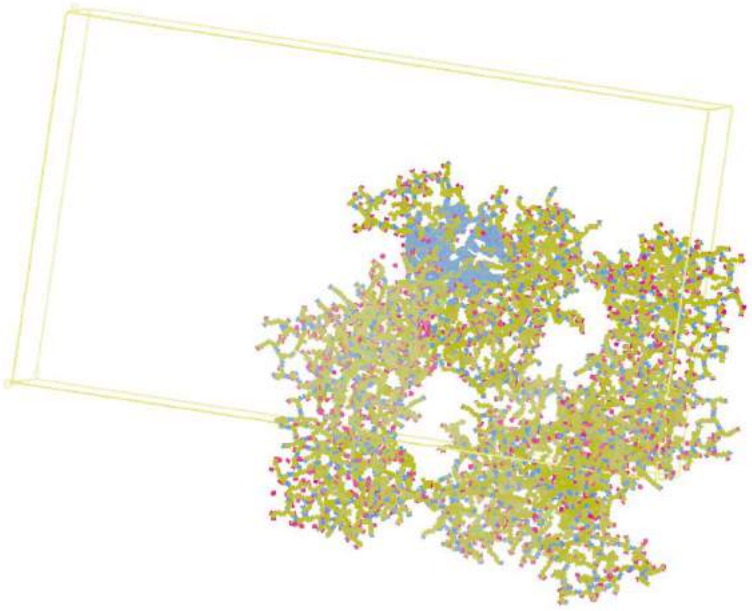


Figure 40| *The Unit Cell containing a whole BsecIM's molecule. (image captured with COOT)*

The data presented in the PDB file for the atoms' coordinates are separated in 4 Chains. Chain A and B refer to the different peptide chains of the protein, and Chain C and D to the different water molecules present near some residues on each peptide chain, respectively. (**Figure 41**)

Chain A|

```
ATOM      1  N   THR A  22       1.772  58.007 104.899  1.00102.69           N
ATOM      2  CA  THR A  22       2.823  58.805 104.189  1.00102.91           C
ATOM      3  CB  THR A  22       3.237  60.062 105.004  1.00101.79           C
```

Chain B|

```
ATOM   4450  N   THR B  22      27.613 -16.984 122.689  1.00 97.71           N
ATOM   4451  CA  THR B  22      27.132 -18.354 122.889  1.00 94.51           C
ATOM   4452  CB  THR B  22      26.635 -18.579 124.332  1.00 94.37           C
```

Chain C|

```
HETATM 8875  O   HOH C   1      -0.920   9.189 144.577  1.00 37.34           O
HETATM 8876  O   HOH C   2     -41.315  27.187  74.409  1.00 24.97           O
HETATM 8877  O   HOH C   3     -18.226  22.090  79.620  1.00 34.42           O
```

Chain D|

```
HETATM 8986  O   HOH D   1      -5.196   9.102 154.629  1.00 59.41           O
HETATM 8987  O   HOH D   2     -10.362   1.831 146.107  1.00 62.06           O
HETATM 8988  O   HOH D   3     -16.633   2.279 155.464  1.00 64.77           O
```

Figure 41| *The first three atoms and their coordinates for each one of the four chains in the* ***file_0.pdb*** .

As a reference for the protein's model, I used the uploaded sequence in the UniProt Data Base (Uniprot Consortium, 2018), which was determined from protein sequencing experimental procedures. (Rina, Bouriotis, 1994) (**Figure 42.A.**)

There are some main differences between the UniProt and the model's sequences. The UniProt sequence has **579 amino acid residues** -for each chain- and the **file_0.pdb**'s sequence has **40 residues less** for each chain. Also, the model's sequence's numbering starts a unit further compared to the UniProt's sequence numbering – for example, if *i* is the numerical position of a residue in UniProt BsecIM's main sequence, then *i+1* is the same residue's position in the model's sequence. (**Figure 42.B.**)

These differences are due to the fact, that the model's sequence is based on the way the protein's model was built. The 40 missing residues, from the **file_0.pdb**'s sequence, are not built within the model's backbone, so they are not present in the model's sequence.

It may be already clear how useful the UniProt sequence can be for the remaining model building and final optimization.

A|

```
        10         20         30         40         50
MMSVQKANTV SRQKATGAHF TPDKLAEVIA KRILDYFKGE KNRVIRVLDP
        60         70         80         90        100
ACGDGELLLA INKVAQSMNI QLELIGVDFD IDAINIANER LSRSGHKNFR
       110        120        130        140        150
LINKDFLEMV SEGDNYDLFN IEELEPVDII IANPPYVRTQ ILGAEKAQKL
       160        170        180        190        200
REKFNLKGRV DLYQAFLVAM TQQLKSNGII GVITSNRYLT TKGGGSTRKF
       210        220        230        240        250
LVSNFNILEI MDLGDSKFFE AAVLPAIFFG EKKNKEYQKE NSNVPKFFKI
       260        270        280        290        300
YEQSDIEASS SVNSEFNSLI ELLEVNKSGL YSVEDKTYSI SLGKIISPEN
       310        320        330        340        350
YKEPWILATE DEYEWFMKVN QNAYGFIEDF AHVKVGIKTT ADSVFIRSDW
       360        370        380        390        400
GELPEEQIPE DKLLRPIISA DQANKWSVSL VGNNKKVLYT HEIRDGQIKA
       410        420        430        440        450
INLEEFPRAK NYLESHKERL ASRKYVLKAN RNWYEIWVPH DPSLWDKPKI
       460        470        480        490        500
IFPDTSPEPK FFYEDKGSVV DGNCYWIIPK KENSNDILFL IMGICNSKFM
       510        520        530        540        550
SKYHDIAFQN KLYAGRRRYL TQYVNKYPIP DPESIYSKEI ISLVRELVNN
       560        570
KKETQDINEI ENRIEKLILR AFDIESLKY
```

B|

```
       10+1       20+1       30+1       40+1       50+1
_____ TPDKLAEVIA KRILDYFKGE KNRVIRVLDP
       60+1       70+1       80+1       90+1      100+1
ACGDGELLLA INKVAQSMNI QLELIGVDFD IDAINIANER LSRSGHKNFR
      110+1      120+1      130+1      140+1      150+1
LINKDFLEMV _____ ELEPVDII IANPPYVRTQ ILGAEKAQKL
      160+1      170+1      180+1      190+1      200+1
REKFNLKGRV DLYQAFLVAM TQQLKSNGII GVITSNRYLT TKGGGSTRKF
      210+1      220+1      230+1      240+1      250+1
LVSNFNILEI MDLGDSKFFE AAVLPAIFFG EKKNKE_____ SNVPKFFKI
      260+1      270+1      280+1      290+1      300+1
YEQSDIEASS SVNSEFNSLI ELLEVNKSGL YSVEDKTYSI SLGKIISPEN
      310+1      320+1      330+1      340+1      350+1
YKEPWILATE DEYEWFMKVN QNAYGFIEDF AHVKVGIKTT ADSVFIRSDW
      360+1      370+1      380+1      390+1      400+1
GELPEEQIPE DKLLRPIISA DQANKWSVSL VGNNKKVLYT HEIRDGQIKA
      410+1      420+1      430+1      440+1      450+1
INLEEFPRAK NYLESHKERL ASRKYVLKAN RNWYEIWVPH DPSLWDKPKI
      460+1      470+1      480+1      490+1      500+1
IFPDTSPEPK FFYEDKGSVV DGNCYWIIPK KENSNDILFL IMGICNSKFM
      510+1      520+1      530+1      540+1      550+1
SKYHDIAFQN KLYAGRRRYL TQYVNKYPIP DPESIYSKEI ISLVRELVNN
      560+1      570+1
KKETQDINEI ENRIEKLILR AFDIESL_____
```

Figure 42| A. *The UniProt's Protein Sequence of BsecIM compared to* **B.** *the differences from the model in* ***file_0.pdb*** .

0.ii. Programs & Suites

For the building of the missing regions, and the Optimization and Validation of the model, multiple computational tools were needed. Beyond the most common programs in Computational Crystallography, I also used some other tools to test how much they could be of use for the Optimization of BsecIM's final model, to compare the different optimized results, and to test their validity for future work on computational crystallographic studies.

0.ii.A. COOT

Coot -*Crystallographic Object-Oriented Toolkit*- (version 0.8.9.2) is a Molecular Graphics Modification Program created by the MRC Lab.

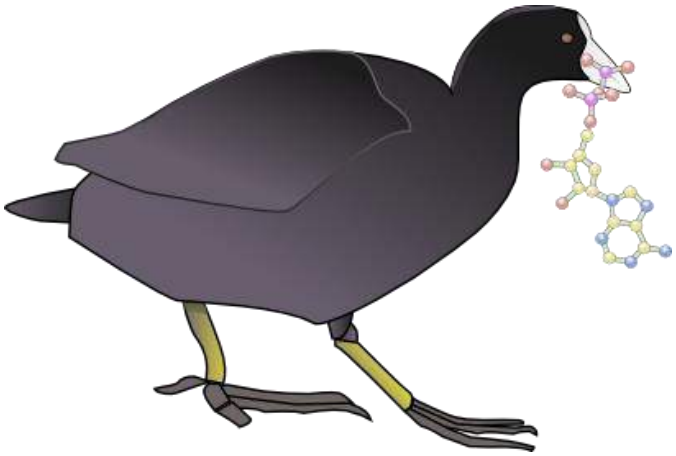


Figure 43| *The COOT logo.*

More specifically, COOT is for macromolecular model building, model completion and validation. It is particularly suitable for protein modelling using X-ray data. It displays electron density maps, such as MTZ format files and molecules' models, from PDB files.

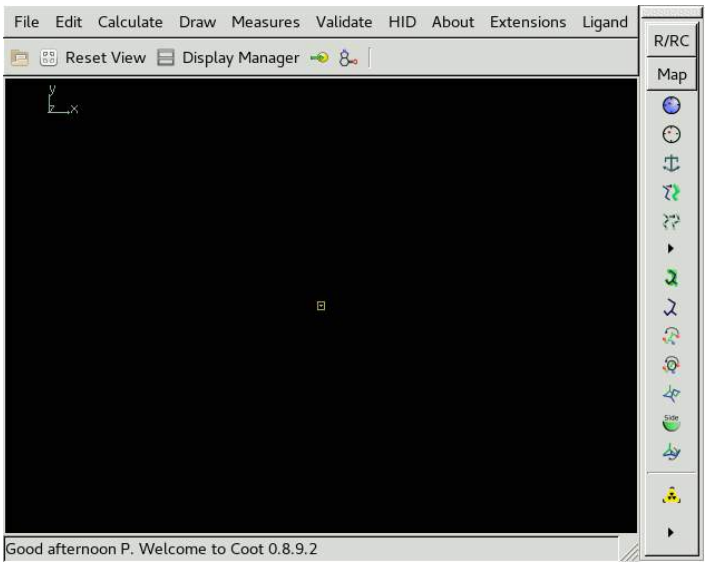


Figure 44| *Coot starting Interface.*

Some of the main COOT tools that I used for the model building are:

Manual Rotation & Navigation- Coot is a molecular graphics program. It is quite easy to navigate through a model's chain and *jump* from one residue to another with just one *click*. Using the mouse, to zoom In & Out in the 3D model or label residues, and the keyboard to change the focus point and navigate through the different residues on the model's backbone. Its graphic display makes it easier to use it for model building.

Real Space Refinement- A tool, which is used to refine a specific amino acid residue, based on its stereochemical properties and electron density maps' indications (both Direct and Difference maps, according to what maps are used). There is also the possibility to *run* Real Space Refinement for the whole model.

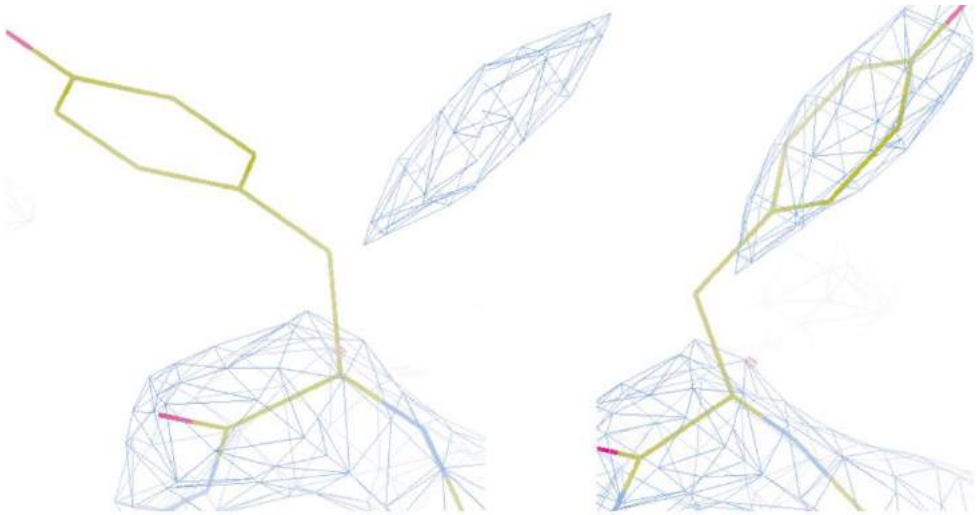


Figure 45| *Real Space Refinement of a Tyrosine.*

Mutation- Mutation is a tool used to change one residue to another – for example an Alanine *being mutated* to Lysine.

Add Terminal Residue- With *Add Residue*, I could add one by one amino acid residues, from the C-terminus, filling the model's missing regions. The residue that is added each time is an Alanine. After the addition, with the *Mutation* tool I could turn Alanine into the amino acid residue of interest.

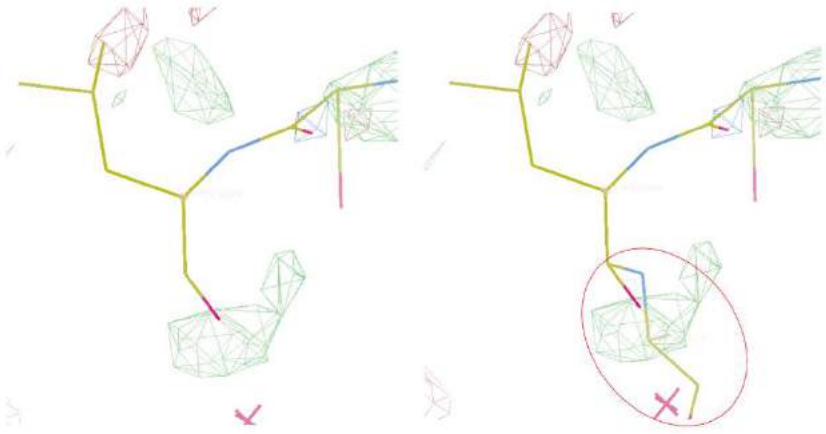


Figure 46| *The addition of an Alanine with Add Residue.*

Some other COOT Tools, which could be combined with the above are:

Rotamers… and ***Auto fit Rotamers***, which shows a list of distinct rotamer conformations for a residue and can be used in combination with *Real Space Refinement* for the best fitting conformation of a residue according to the map's indications.

Change Residue's Phi and Psi, which shows the possible angles represented by the Ramachandran plot and allows the user to change the φ and ψ torsion angles of a residue, accordingly.

Flip Peptide, which allows the user to flip a residue by 180 degrees on the backbone.

COOT is one of the most important programs I used for this project. The many editing possibilities, that COOT provides, made the completion of the model building of BsecIM much easier, putting aside the validation stage of the study. (Emsley, Lohkamp, Scott, Cowtan, 2010)

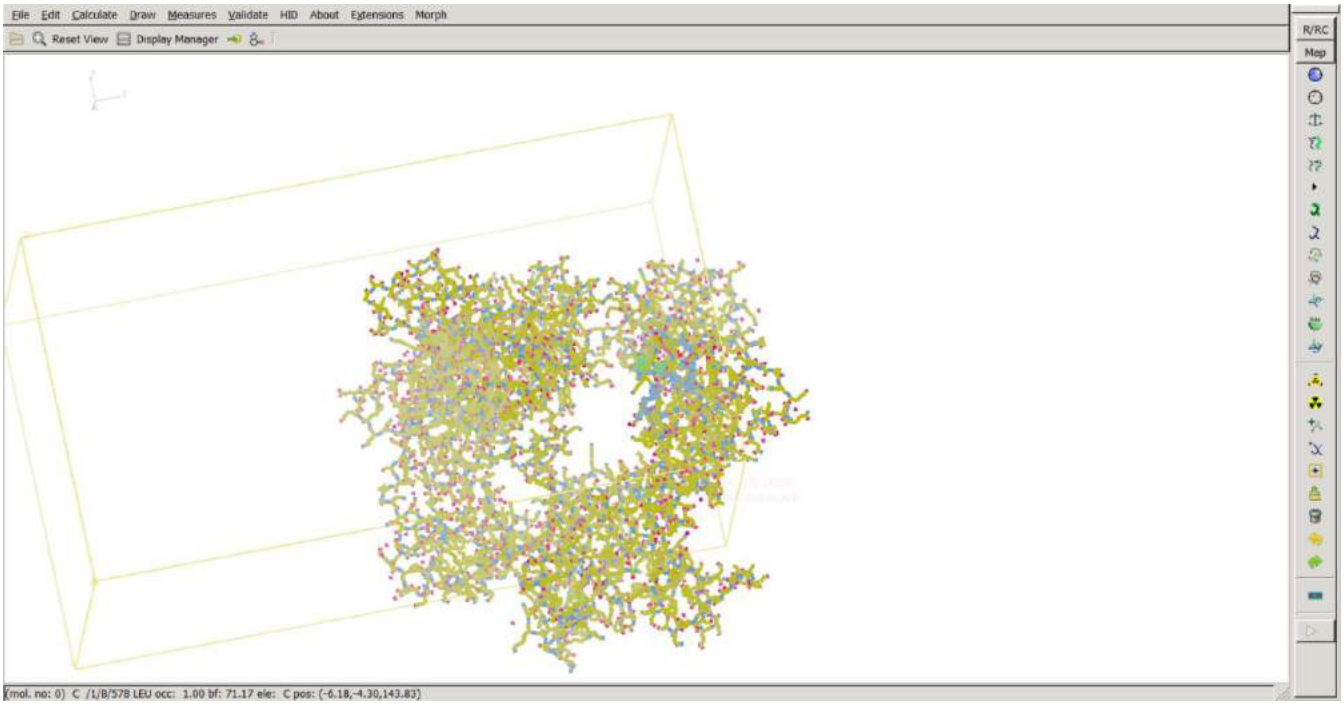


Figure 47| *A typical day with COOT.*

0.ii.B. CCP4-Refmac5

Refmac5 is one of the most basic Macromolecular Crystallography Refinement Program. It is provided by the CCP4 Suite, just as mentioned in the Introduction Section.

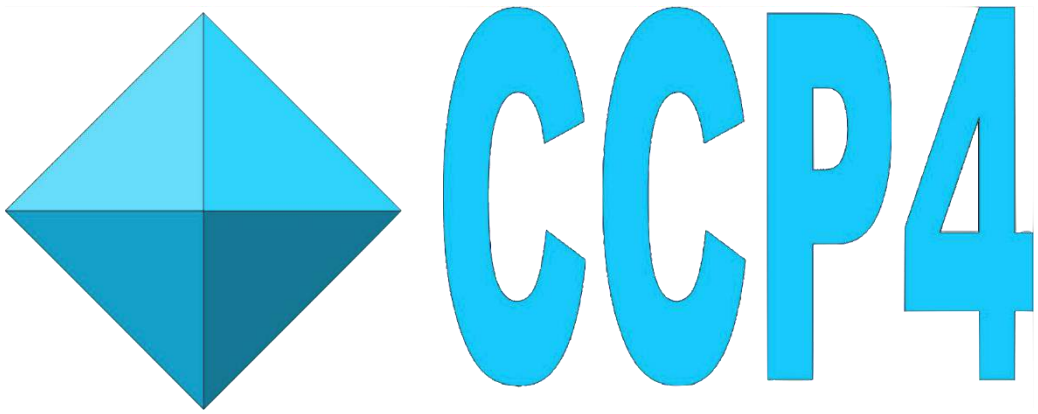


Figure 48| *The CCP4 Suite Logo.*

So, it is quite clear that this program was mostly used for the Computational Refinement process, just like COOT was the one that I used the most for the Model Building procedure.

To keep the Refinement results comparable to the previous studies on the BsecIM's structure, the same restraints and parameters were needed to be used in this study, and therefore the same Algorithmic Setup. As shown in **Figure 49.A.** , running a *Job* with Refmac5 starts with the *Job's Title* (**1**). For the Refinement cycles that I *ran*, the different jobs had the same name; "Restrained refinement using isotropic B factor", which summarize this Refinement's main parameters.

The "TLS & restrained parameter" (**2**) was the second basic setting of a Refmac5 run, with "no prior phase information" and "no twin refinement" -which is usually used when there are two crystal structures with similar orientation.

To complete the main Setup, I needed each time to import the PDB, MTZ and TLS files, which needed to be refined, and to state the names and file paths for each OUTPUT file of the same formats. If, for example, the INPUT files were **file_0_editted.pdb**, **.mtz** and **.tls** , the OUTPUT files would be **file_1.pdb**, **.mtz** and **.tls,** accordingly.

In the *Data Harvesting* section (**3**), the same project was set for all *Job Runs*.

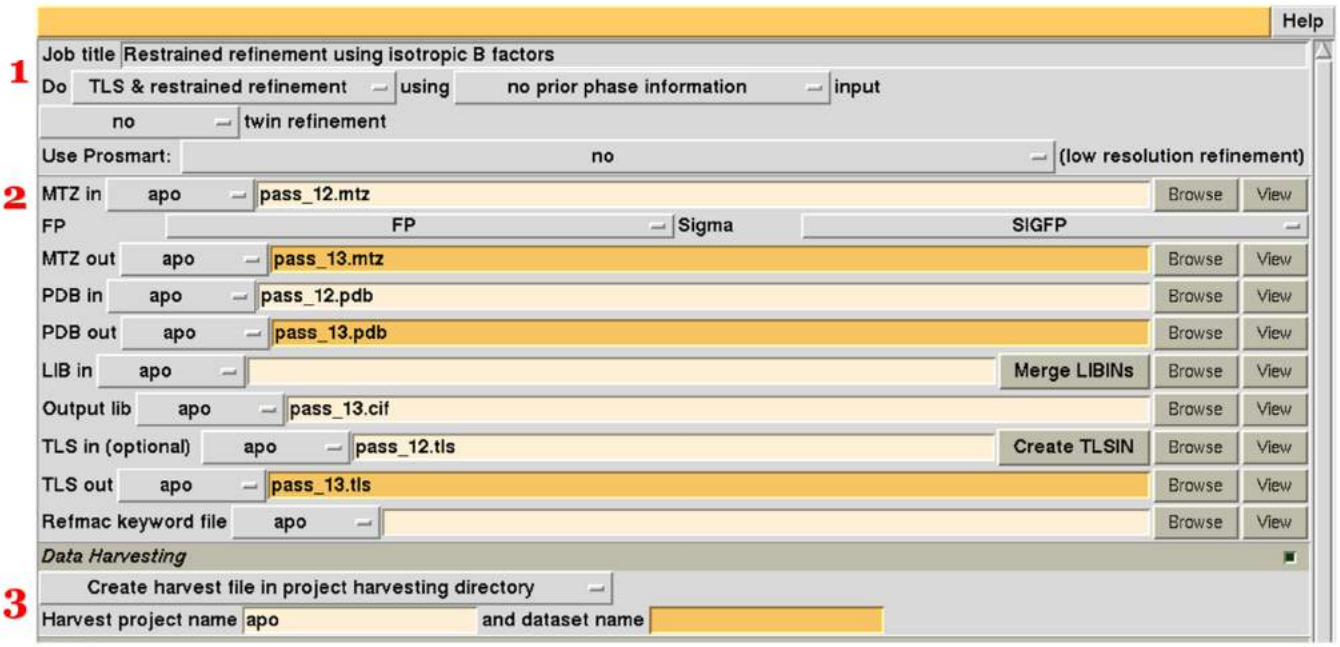


Figure 49.A. | *The main Setup of Refmac5 Job Run.*

The next Step is shown in **Figure 49.B**. Once the Main Setup is complete, it's time for the Parameters' Setup. The *TLS Parameters* are checked in each *job* I *ran* (**1**), and set by default to perform 10 Refinement cycles.

In the *Refinement Parameters* (**2**) section the number of cycles is also set in 10, for *maximum lilelihood restrained after TLS refinement*, and it is important to be set in the same, or close enough, number of cycles as TLS Refinement. Instead of *automatic weighting*, the use of *experimental sigmas to weight Xray terms*, restrains the Refinement data according to the crystallographic experimental data.

The *Setup* of *Geometric Restraints* (**3**) are set by default in the *cis-peptide* and *links between symmetry related atoms* according to the INPUT files for the Refinement cycles. The *Setup* of *Non-Crystallographic Symmetry (NCS) Restrains* is also important for this particular crystallized structure, because of its position in the Unit Cell, and set to be generated automatically (**4**).

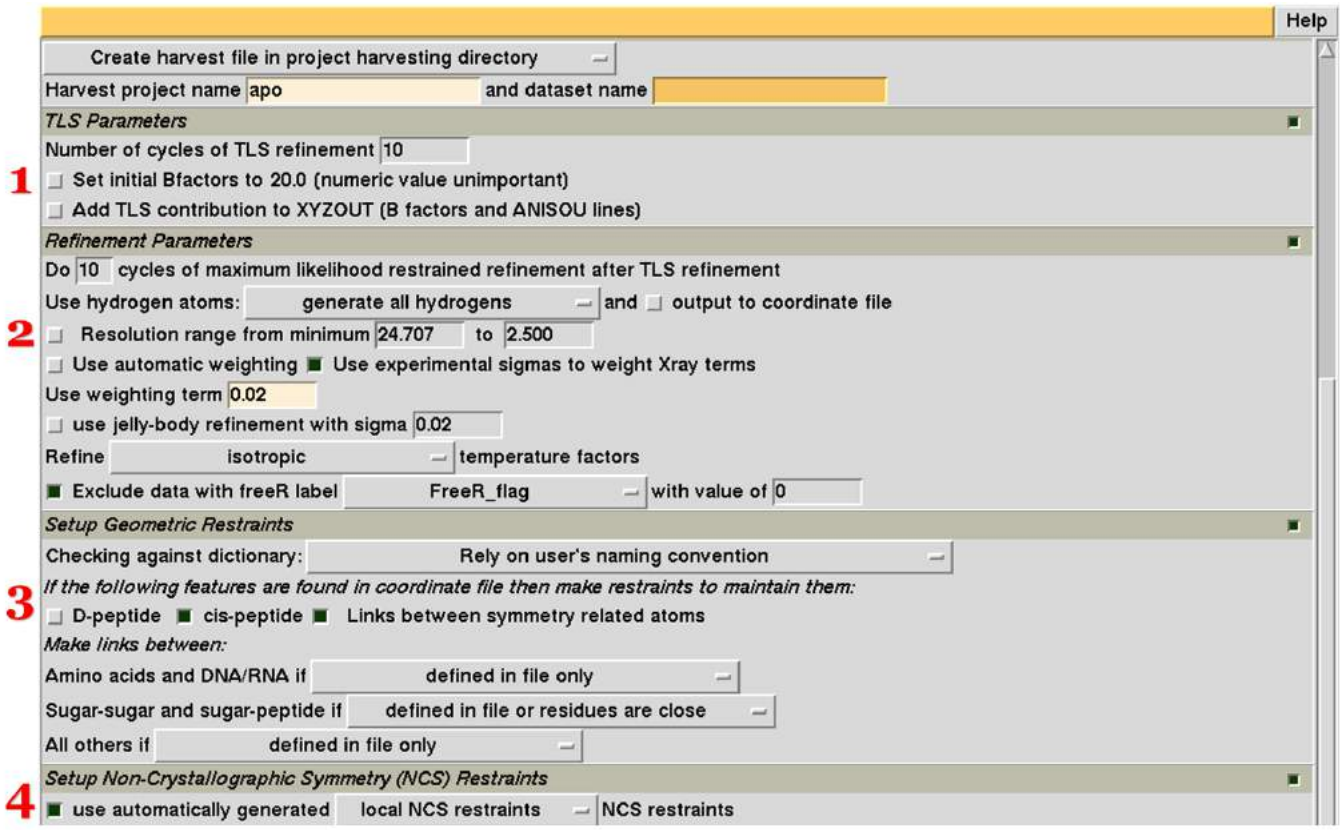


Figure 49.B| *The Setup of* ***1.*** *TLS and* ***2.*** *Refinement Parameters,* ***3.*** *Geometric and* ***4.*** *NCS Restraints in Refmac5 Runs for BsecIM crystallographic studies.*

The rest Setup, as shown in **Figure 49.C**, is about other type of Restraints, like *External Restraints* (**1**), for weighting the 3D parameters of the model and map, and (**2**) the *Monitoring and Output Options*, which are same as the INPUT files' characteristics.

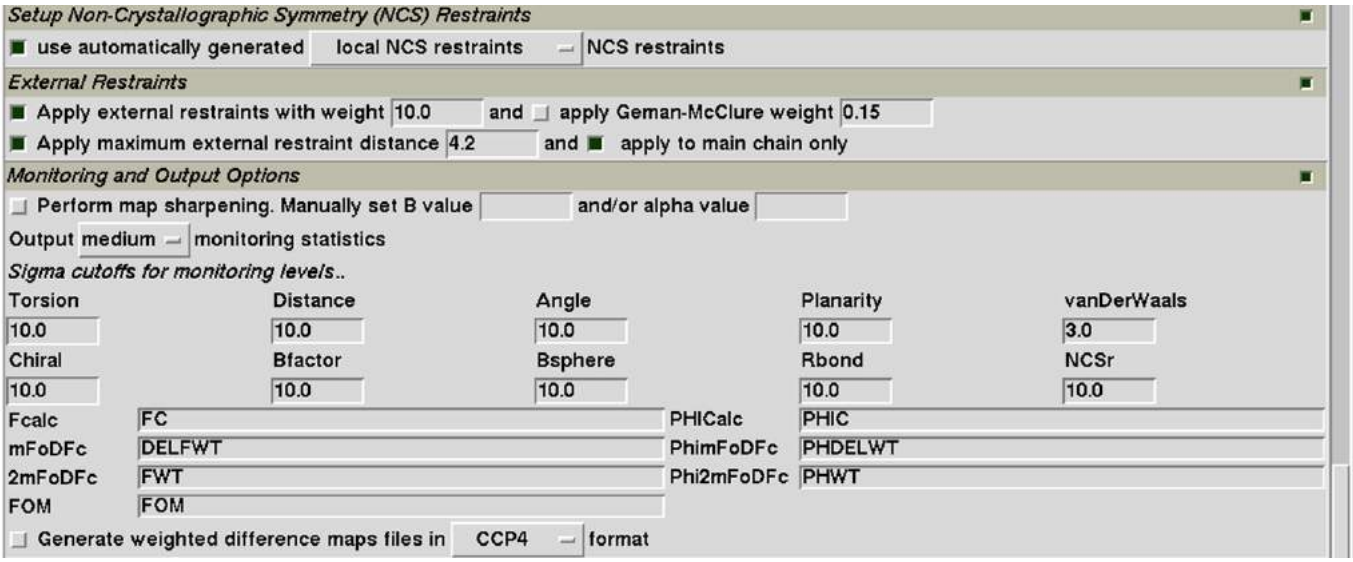


Figure 49.C.1-2| *The External Restrains and Output Options of my Refmac5 Run.*

For the *Scaling* (3) section, the calculation of the contribution from the solvent region is needed, and it is referred to the contribution of the water molecules in the model.

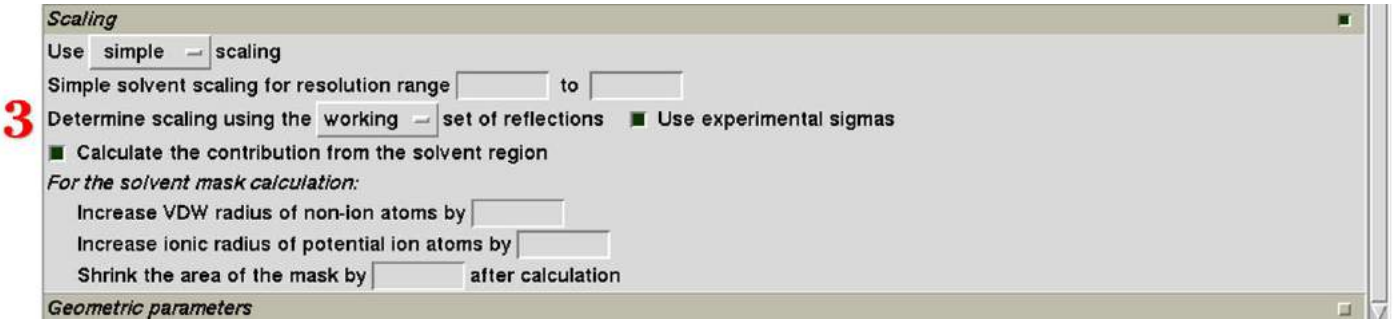


Figure 49.C.3| *The Scaling Section.*

And last, but not least, with no extra restraining contributions from the Geometric Parameters (**Figure 49.D.1**), the Refmac5 *Job is set and ready for Run* (2).

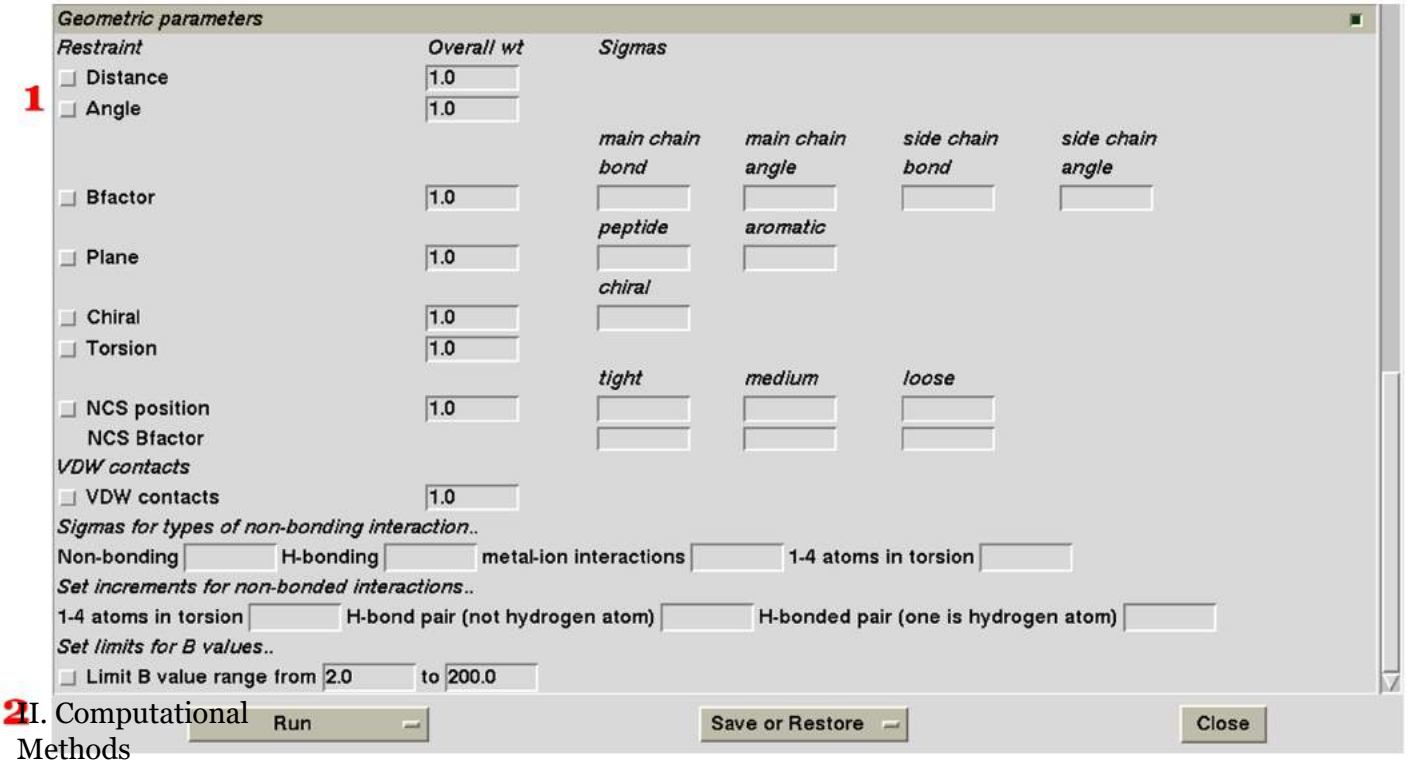


Figure 49.D| *The Final Step of Refmac5 Setup for the Refinement of BsecIM model.*

Once the Setup is ready, the Refinement cycles start. Each cycle is perfomed with the same restraints and every time R and R_{free} factors are calculated for every cycle. At the end of a Refinement cycle the data from the previous cycle works as *starting point* for the next, producing

in that way a two-dimensional diagram of the ***R*** factors for each cycle against the electron density Resolution (**Å**) for each refined map. (**Figure 50**) (Winn, M. D., et al., 2011)

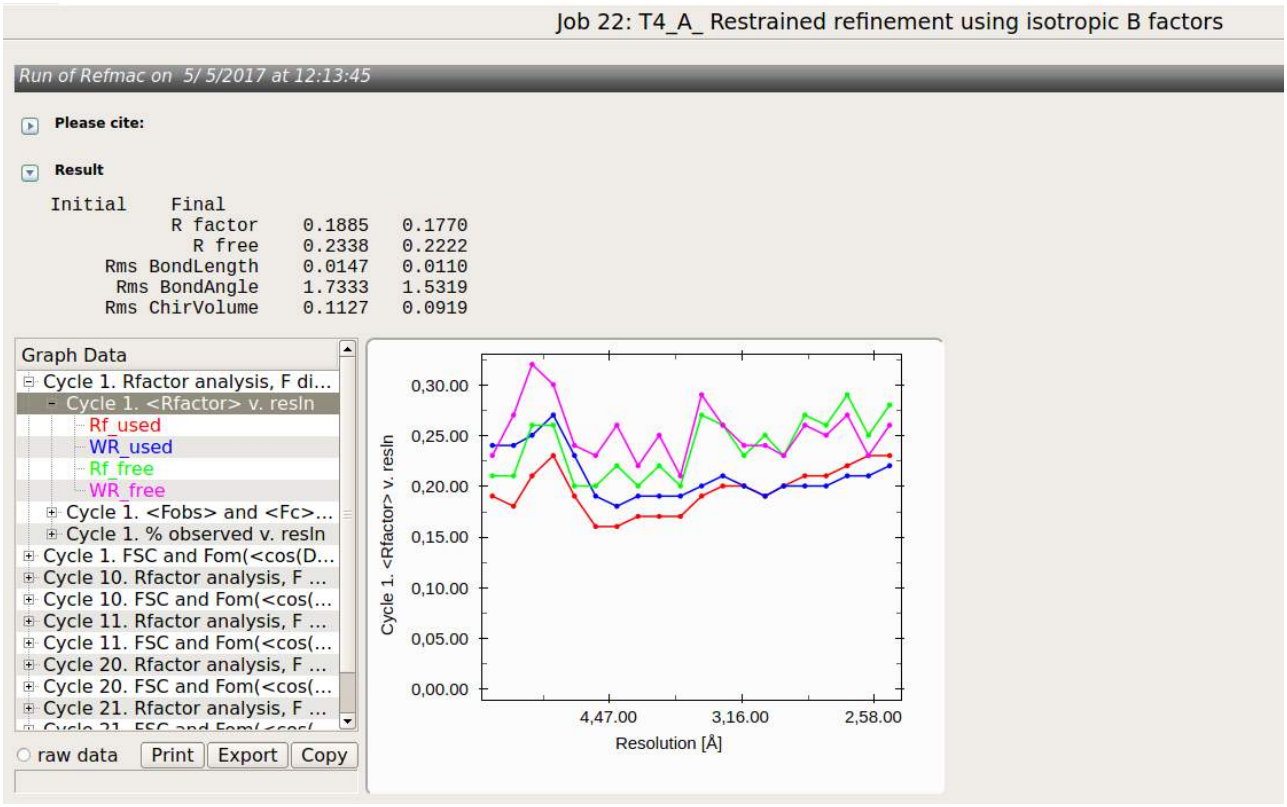


Figure 50| *The final diagram and data after the completion of the Refinement cycles of a Refmac5 Run.*

As mentioned above, the closest the ***R*** is to 0.0, the greater the agreement between experimental observations and the structure factors predicted from the model. Of course, some contributing parameters to the decrease or increase of the R factor are the molecule's size and structural complexity, and also the resolution of the crystallographic data.

Decreasing the R factor as much as closer to 0.0 isn't necessarily the wanted result, especially when the molecule's structural complexity and the data's resolution have not been taken into consideration. Nevertheless, the significant increase or decrease of the R and R-free factors work as great indications for the process of model buiding and fitting.

0.ii.C. Phenix- Simulated Annealing & Omit maps

The model building for the missing loop regions was quite difficult because of the missing electron density indications for these regions. To build some well-defined models for these loops, I needed to test some alternative methods. One of these was the Simulated Annealing process, which is usually used for producing higher resolution data for crystallographic experiments.

One of the Program Suites that provided a *region-specific* Simulated Annealing was the Phenix Suite.

Figure 51| *The Phenix Suite Logo.*

The Simulated Annealing process for Protein Crystallographic Structures works like a *simulation*, in which the structure's *environmental temperature* is increased and directly decreased, to add randomness in the structure's formation process and reduce the probability of falling into the wrong structural *local minimum*. The *products* of such procedures are called Omit Maps.

The Omit are maps with probably enhanced resolution for the whole molecule or the model's regions of interest. The Phenix Suite Simulated Annealing algorithm provides a user-friendlier environment, a fast-algorithmic process and a feature focusing on regions of interest on the model. In that way, the enhancement of the crystallographic data for the high-mobility loop regions in BsecIM's model could be of greater help for the definition of these regions' 3D conformation. (Terwilliger, T.C., et al., 2008)

Of course, changing the use of tools from the CCP4 Suite to Phenix, creates some issues around the Validation processes of the produced maps and model. Something that will be discussed further in the Computational Methods Section.

o.ii.D. PDB-REDO

The PDB-REDO is an automated Validation and Optimization Procedure for Protein Structures. Many protein structures, before or after their upload in the Protein Data Bank, undergo the Validating process of PDB-REDO.

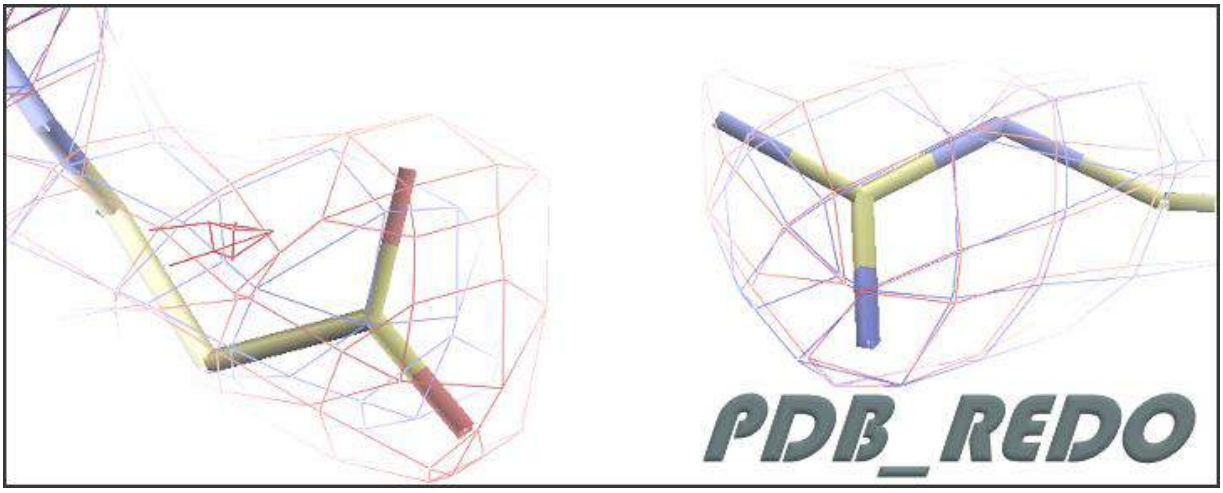


Figure 52| *The PDB-REDO Logo.*

In the platform of PDB-REDO are embedded most of the CCP4 Suite programs, like Refmac and COOT. Once a protein model and electron density map are uploaded on the PDB-REDO server, multiple optimization settings begin.

First, the refinement settings for the model are optimized -NCS restraints, TLS and geometric parameters, etc.- and some first Refinement cycles are carried out. Once the reconstruction of some of the model's regions is complete, these structural changes run through validating processes repeatedly, until the whole process reaches a threshold.

Once the validation and optimization are complete, the validated model is *fetched back* to the user with the Refinement results uploaded. (Joosten, R., Joosten, K., Murshudovb, Perrakis, 2012).

It is quite critical to use a Validating tool, like PDB-REDO, before the publication of a protein structure. But it is also important to take into consideration that an automated process like this lacks the Structural researcher's empirical point of view, and it is possible to emerge a model based mainly on the decrease of the R and R-free factors.

But even if that's the case, the results from the Validation may be used as an extremely useful *guideline* for the completion of the model.

o.ii.E. CCP4-MG

Finally, the Molecular Graphics program of CCP4 Suite is the main Molecular Presentation program I used for this project.

This program provides multiple tools for viewing macromolecular structures and generating high-resolution images of Molecules. The main interface is quite similar to common molecular graphics programs, like Rasmol and Pymol. The differences are that it provides a graphically friendlier environment for the user and the capability to combine different structural data to generate more accurate crystallographic figures -for example a peptide chain with an electron density map.

Most of the structural Figures in this thesis were generated with CCP4MG. (McNicholas, Potterton, Wilson, Noble, 2011)

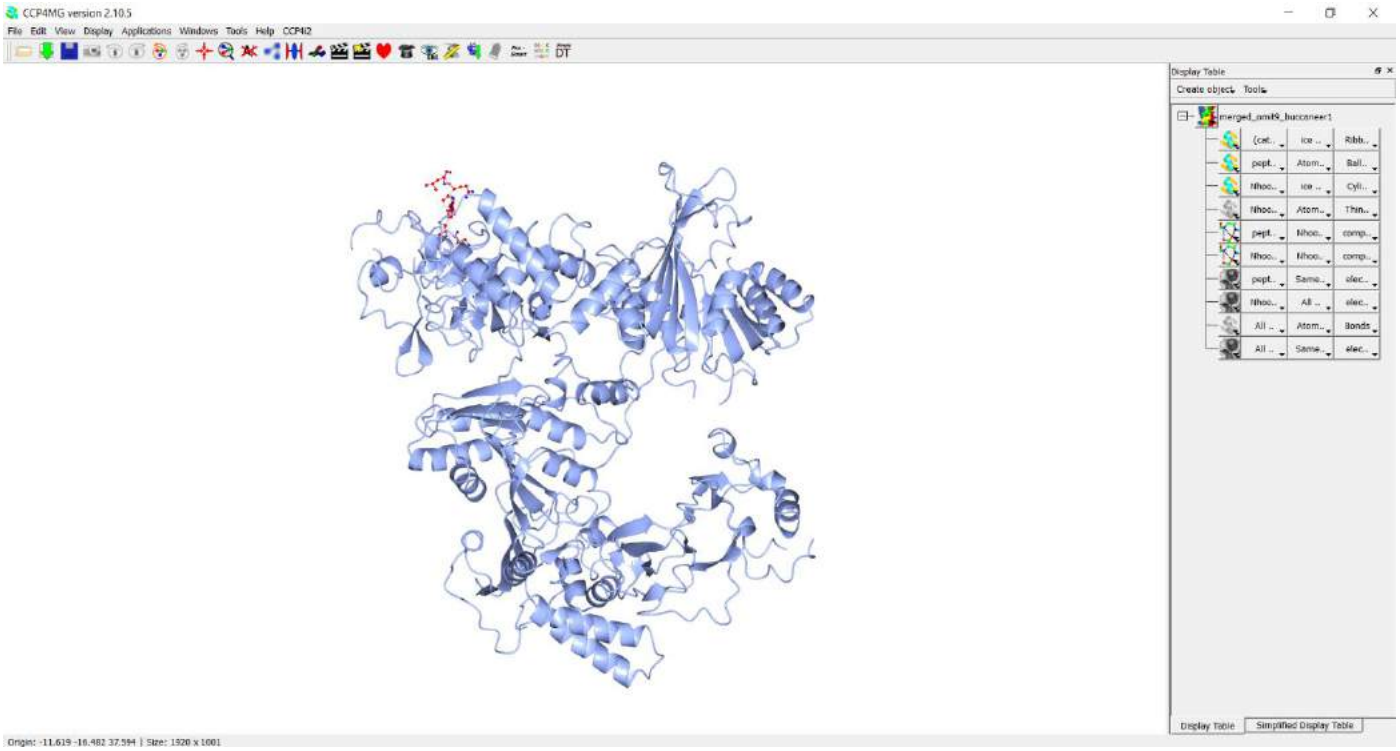


Figure 53| *The CCP4 Molecular Graphics Interface.*

0.iii. Hardware & Software

The characteristics of the Hardware and Software used for this computational experiment are the most important factors needed to be mentioned. In every computational experimental procedure, the main Setup of the project starts from choosing a satisfactory Hardware, if not the most efficient, for the *needs* of the project.

The next step is the choice of the Software, which is important to be in congruity to the algorithms' versions and requirements. In other words, it is futile to setup a Software, in which some -if not all- algorithms *do not run*. Also, the *stability* of the Software's characteristics and the *adaptability* the Software's environment may provide to the algorithms' properties are the main reasons, somebody needs to consider before the beginning of the study. For example, if I wanted to make a change to a program's function, without making a lot of changes to the main program, I would need an environment, which would make it easier to change these algorithmic properties.

Without describing further, for this study I used *two different computer systems*, one *main* and one for *back-up use*, each with different characteristics:

- The **Main Computer**'s **Hardware's** characteristics are; a **Processor CPU G440 @1.60 GHx** -*which is quite efficient for the runs of programs like Refmac5*-, and a **2nd Generation Core Processor Family Integrated Graphics card** from Intel Corp. -*which is of pivotal importance for this project, because of the high graphic requirements, most of the programs had.*
  The **Software's** characteristics are **Linux-based**, the Ubuntu 17.10 version, which is compatible with all the CCP4 Suite and Phenix Suite Programs and all the new versions and updates for almost all programs.

- The **Back-up Computer**'s **Hardware** system contains; a **Processor Core i5-5200U CPU @2.20 GHx** from Intel Corp. -*quite efficient processor for a laptop 64bit machine*-, and an **HD Graphics 5500 Graphics card** from Intel Corp. -*which is also quite satisfactory for all the Molecular Graphics Programs, like Coot*-.
  The **Software's** characteristics are Microsoft **Windows-based**, and more specifically Windows 10 the 2004 version. All the programs I was using had a Windows compatible version, like for Example WinCoot, CCP4_win and Phenix 1.10.1.

Every main characteristic for each System is more than satisfactory for the requirements of this computational study. But there is a great issue. It is not recommended -*almost forbidden*- to use in a Computational Study two entirely different Computational Systems. Note that this is pointed out mainly for the different Software and not the Hardware.

The Software-*swapping* between Linux and Windows is never preferable. Something I learned almost at the end of this project.

The programs used in both Software systems are some CCP4 Suite tools and COOT. The Phenix Suite's Simulated Annealing was mainly *run* in Ubuntu 17.10. The interesting fact is that

both CCP4 Suite and COOT are updated to the latest version for Linux and Windows. In both Systems, there is a CCP4 Suite 7.1.002 version and a COOT 0.8.9.2 version. Another interesting point is that, even though the versions for each system are the same, there are some *troubling* differences in the results.

At this point, this Software-*swapping* became an interesting testing on how these two different Software systems can or cannot be used in combination to perform future similar studies, and if not for the whole studies, maybe for some parts. To fix the issue mentioned above, I tested which system performs more efficiently in which part of the study. This was not a detailed testing though, but more of an empirical one, to isolate the technical issue and *handle* it in a much more convenient way, without creating difficulties for the rest of the study.

In the end, for editing I preferred using WinCoot; it has more consistent geometrical parameters and the editing tools usually ended up with a well-based conformation for a residue, than the Ubuntu COOT. In **Figure 54** is shown the region 23-26 of Chain B from the protein. For the optimization of this region the same COOT tools were used; *Rotamer* change + *Real-Space Refinement* with the same electron density map. The results are different. The conformation of the Linux COOT is quite unstable, forming non-canonical hydrogen bonding (**54.A**). With the same conditions the WinCOOT performed much better, according to the geometric parameters of the backbone and the side chains' structural characteristics (**54.B**).

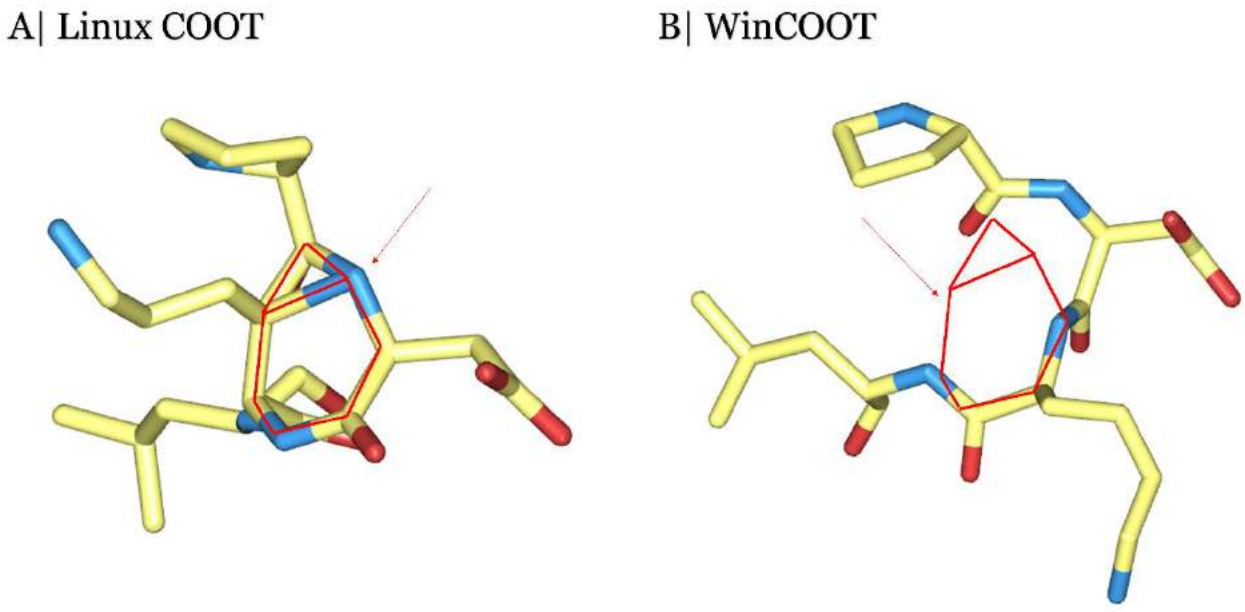


Figure 54| A. *The conformation of 23/PRO/B, 24/ASN/B, 25/LYS/B and 26/VAL/B with Rotamer change and Real-Space Refinement with Linux-based COOT.* **B.** *The same region's conformation, with the same tools and requirements in WinCOOT, but with better performance than Linux-based COOT.*

For the Refmac5 runs I preferred using the Linux-based CCP4 Suite, because it seemed more consistent on its results. The results from the same three runs on Refmac5 are shown in **Table 3**. In each Run the same INPUT files, with the same edited model from WinCOOT, are used for both Linux and Windows Refmac5. The OUTPUT files from each Run are edited in the same way with WinCOOT and work as INPUT files for the next Run.

The Linux version' s statistics are significantly better (**Table 3.A.**) than Windows' runs, which seem to run with slightly more *strict* automated parameters. (**Table 3.B.**)

Table 3| A. *The Linux-based Refmac5 for three*

A| Linux Refmac5

1st RUN	Initial	Final
R-factor	0.1767	0.1763
R-free	0.2235	0.2236

2nd RUN	Initial	Final
R-factor	0.1885	0.1770
R-free	0.2338	0.2222

3rd RUN	Initial	Final
R-factor	0.1919	0.1762
R-free	0.2438	0.2217

B| Windows Refmac5

1st RUN	Initial	Final
R-factor	0.1776	0.1826
R-free	0.2256	0.2264

2nd RUN	Initial	Final
R-factor	0.1892	0.1842
R-free	0.2352	0.2248

3rd RUN	Initial	Final
R-factor	0.1952	0.1835
R-free	0.2454	0.2209

As mentioned above, the recurring alternation from one Software environment to an entirely different one -in this case from Ubuntu to Windows, and back again- is usually called a *first-timers' accident*. But this *accident* allowed me to consider if the combination of WinCOOT and Linux-based Refmac5 would create further issues on the project, leave the conditions of the study almost unaffected or it would lead to some interesting results for future use.

More about the results of the study will be described in the *Results* Section.

1. Step 1: First Optimization & Loop-Building

The focus of Step 1 is to optimize the BsecIM' s model described in **file_0.pdb**, based on the electron density map of **file_0.mtz** by:

A. Checking **residue-by-residue** the model, to detect the model's issues and correct the conformational errors on both chains' backbone, and …

B. **Building the missing regions 111-124** and **237-242** for each chain, even though the density's indications are almost non-existent for these regions.

1.A. *Residue-by-Residue Check…*

To check the whole model I used as INPUT files the **file_0.pdb** & **.mtz** in **WinCOOT** and started from the N-terminal (*first*) amino acid residue on each chain of BsecIM's modelled structure.

The **first amino acid** for **Chain A** was **Threonine/22** and for **Chain B**, **Histidine/20**. The **last amino acid** (in the C-terminus) was **Leucine/578** for both Chains. Every residue was refined into a better conformation with the Real-Space Refinement tool of COOT.

In every editing step of the model building or model reconstruction, it was important to take in account the electron density maps' contour level. The contour level helped me to find easier the conformation of a model's region based on the map. If the indications for some regions are quite *noisy,* the user can lower the contour level, or do exactly the opposite if the resolution of the map for some regions is low. The sigma (***σ***) factor defines the contour levels of an electron density map.

The corrections for the main changes in the residue's conformations on Chain A and B are shown in **Table 4. A-C**. The model of the **file_0** residues' conformations are presented with **yellow**-colored *skeleton* model and the edited results are presented with **red**-colored *skeleton*.

The **sigma factor** for the Direct map's blue *grid* is **~0.2**, for the Negative Difference map's red *grid* is **~ -0.1** and for the Positive Difference map's green grid is **~ 0.1**.

In **Table 4.A** and **B** are shown the changes made for **23 amino acid residues** with some indications I recognized as errors in their conformation on **Chain A**. In **Table 4.C**, there are **6 edited amino acid residues' conformations** for **Chain B**.

In every edited residue there is a description of the COOT tools that were used and the goal of the editing process for this particular residue.

Table 4. A| *The conformational corrections of **12 amino acid residues in Chain A** of the BsecIM' s model in **file_0.pdb**, using the **file_0.mtz** Direct and Difference map. Under every residue's image there is a short description of the editing's tools and main goal. The memo for the Table's reading is on the top of the table.*

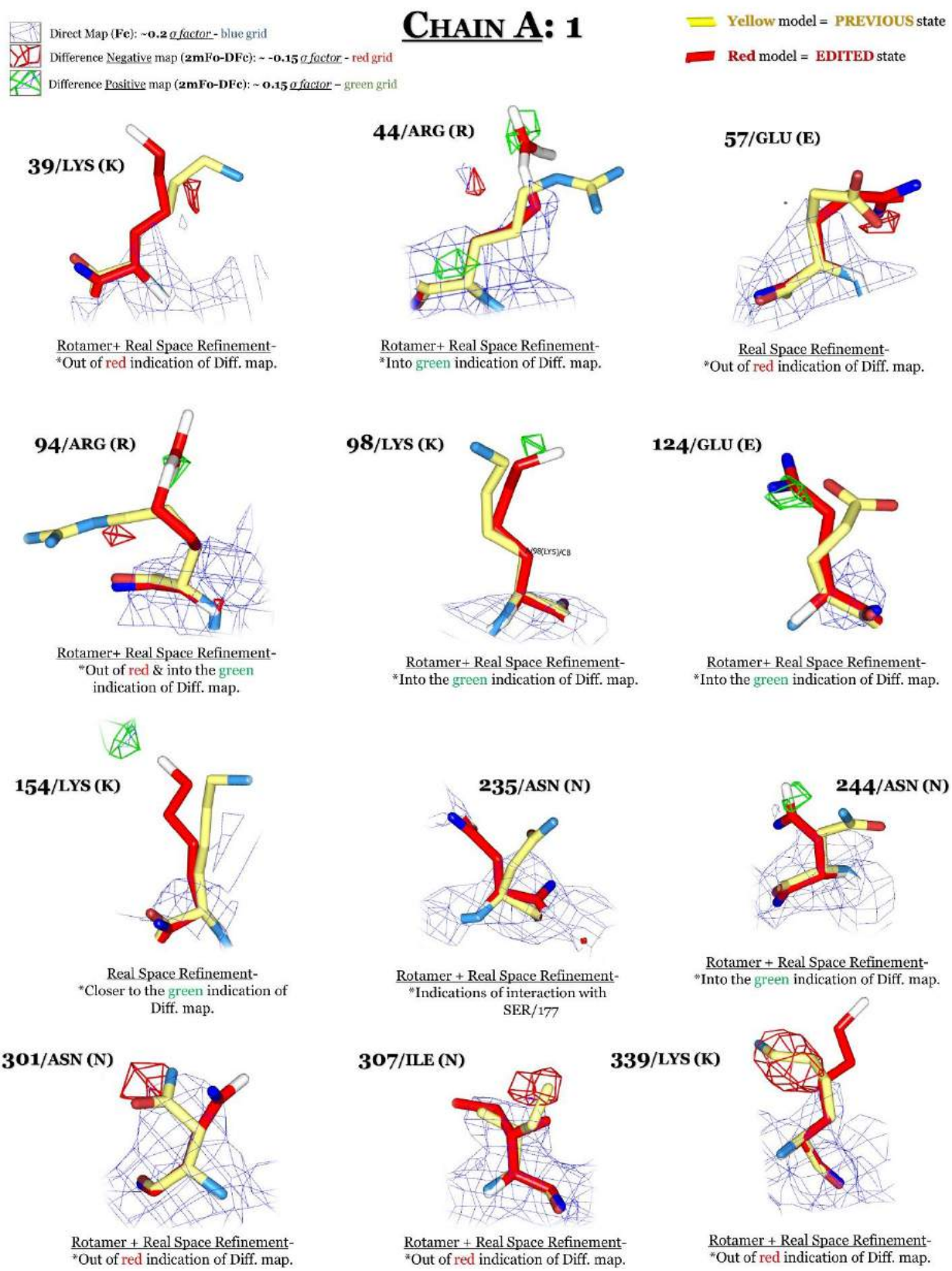


Table 4. B| *The conformational corrections of the rest **11 amino acid residues in Chain A** of the BsecIM' s model in **file_0.pdb**, using the **file_0.mtz** Direct and Difference map. Under every residue's image there is a short description of the editing's tools and main goal. The memo for the Table's reading is on the top of the table.*

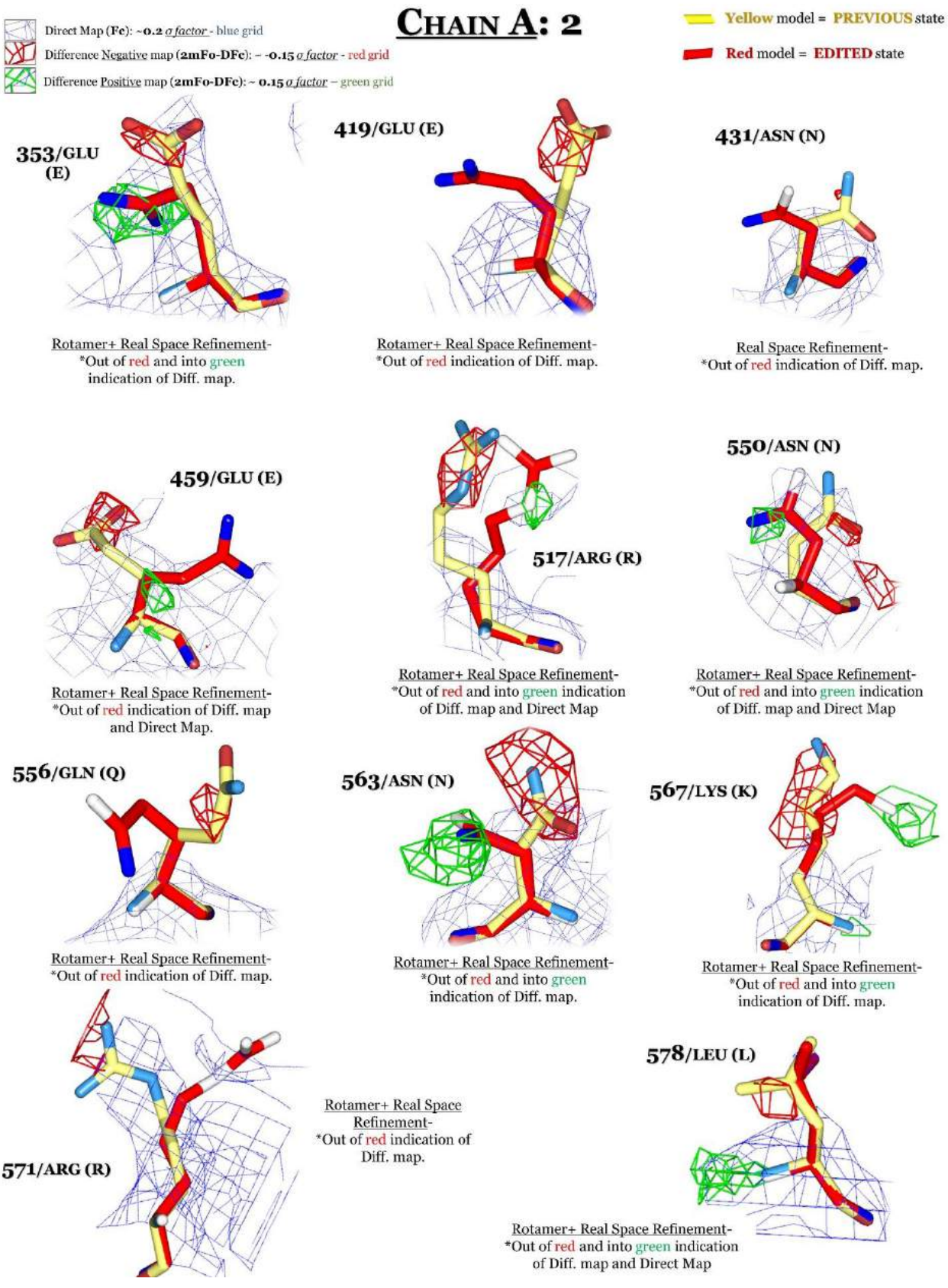
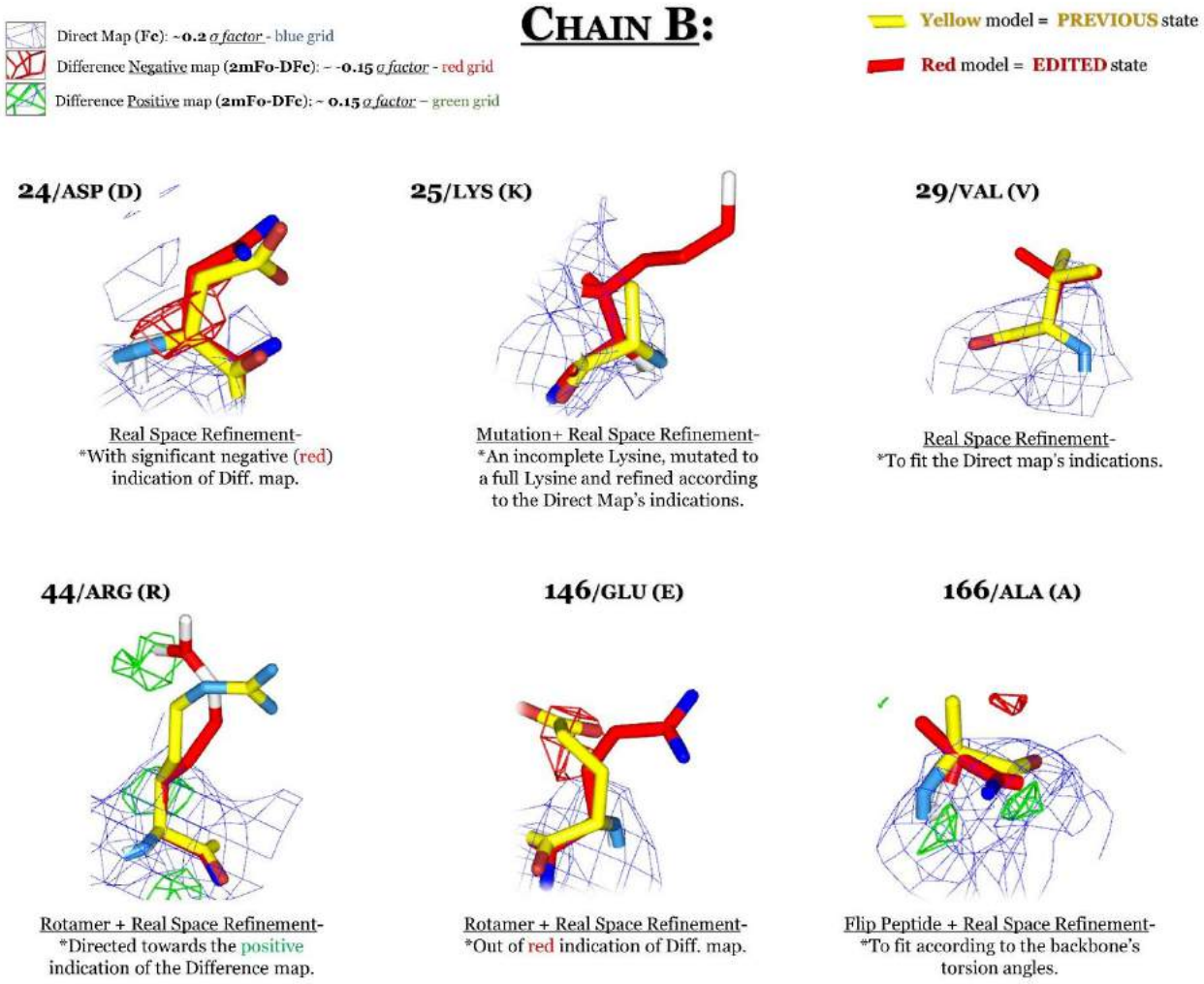


Table 4. C| *The conformational corrections of the* ***6 amino acid residues in Chain B*** *of the BsecIM' s model in* ***file_0.pdb****, using the* ***file_0.mtz*** *Direct and Difference map. Under every residue's image there is a short description of the editing's tools and main goal. The memo for the Table's reading is on the top of the table.*



Once the *residue-by-residue check* was complete and all residues in both chains *Real-Space refined*, it was the time for the building of the missing regions.

1.B. *Filling the gaps…*

In both chains there were the same missing regions, with low resolution electron density map. (**Figure 55.A & B**)

112-122 *(Big Loop)*

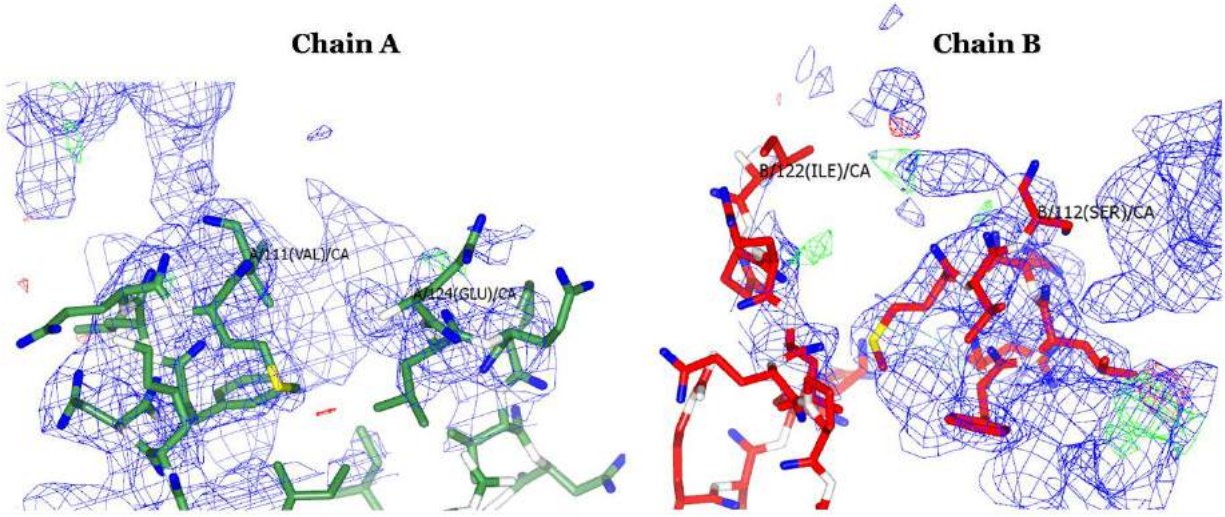


Figure55. A| *The electron density map at the beginning for the **112-122** region in both Chains. The low resolution is presented as little to none indications for the rest of the loop. The sigma (σ) factor for these regions is **0.14** for the **blue** Direct Map and **0.18** for the **red** and **green** Difference Map.*

238-243 *(Small Loop)*

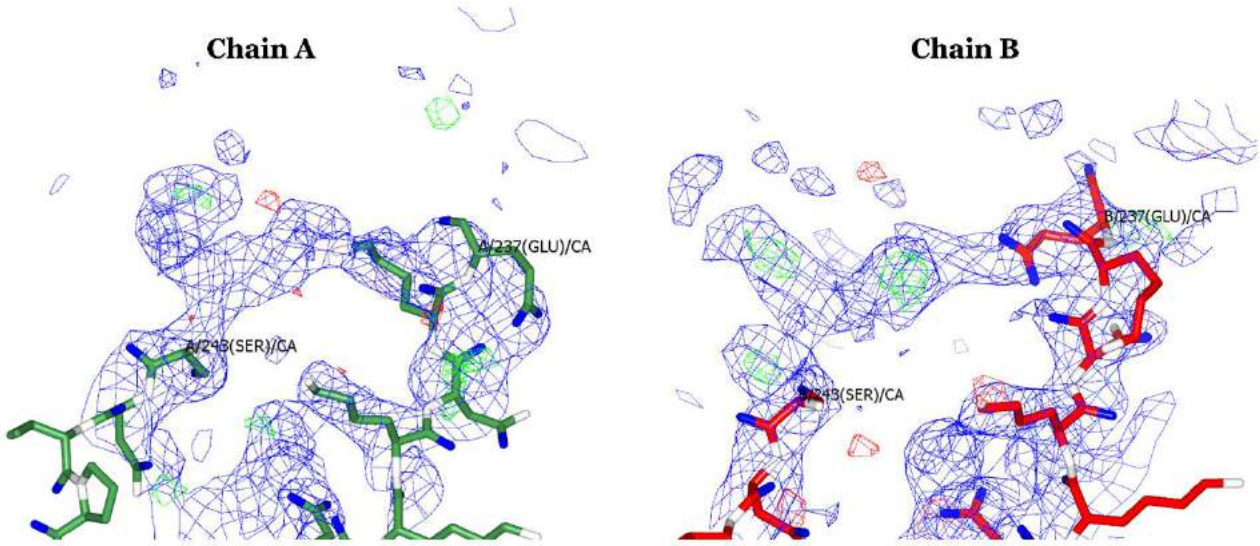


Figure 55. B| *The electron density map at the beginning for the **238-243** region in both Chains. The indications of the map are much better than the Big Loop, due to its smaller size. The sigma (σ) factor for these regions is **0.14** for the **blue** Direct Map and **0.18** for the **red** and **green** Difference Map.*

First, to start building these unmodelled regions, I needed to check the protein sequence and add these residues after that to check in which chain there would be slightly more evident the indications of the electron density map.

The missing regions from the model are shown in **Figure 56.A-C** with one-letter code representation for each residue. In **Figure 56.A** there is a diagrammatical representation of the N-terminus missing region of the model, also known as the *N-terminus Tail,* from **residue-1 to residue-21;**

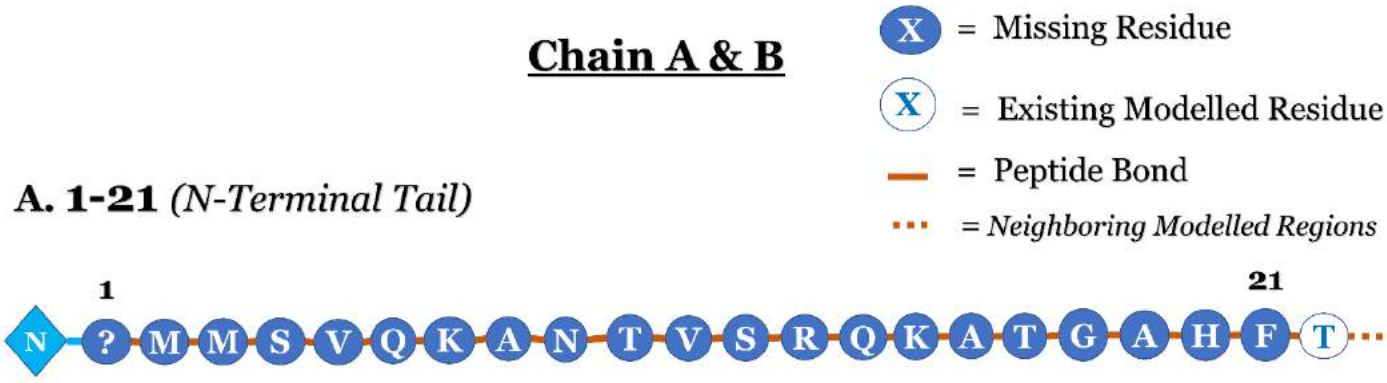


Figure 56.A| *The **N-Terminal missing region** with one-letter code representation for every missing residue from that region of the model.*

In **Figure 56.B & C** are shown the regions **112-122**;

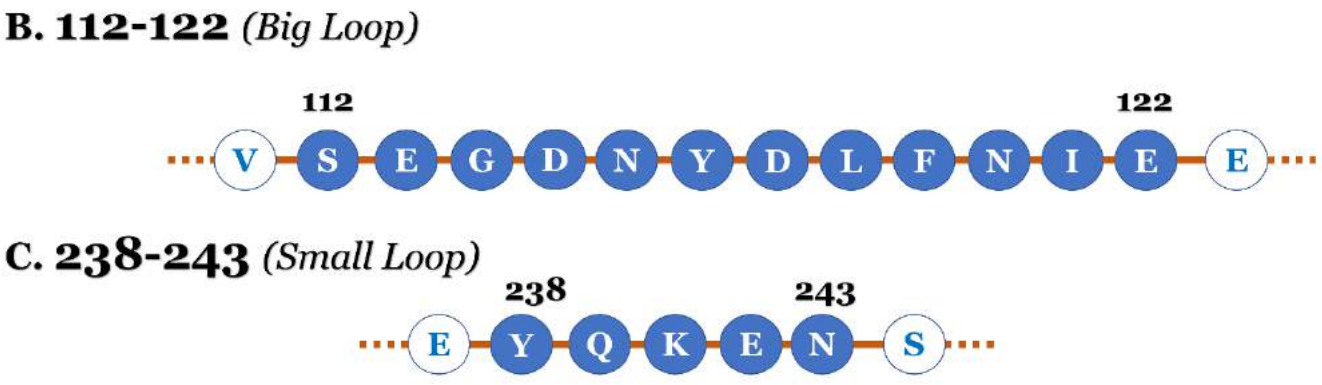


Figure 56.B & C| *The **112-122** and **238-243** intervening missing regions, also known as The Big & Small Loop.*

The C-terminal region was incomplete for only three missing residues, just as shown in **Figure 56.D**;

Figure 56.D| *The **C-terminal** missing regions.*

For the building of these regions I used the following **tools from COOT**;

Add Residue, to fill in a place in the loop regions with an Alanine residue.

Mutate, to turn the Alanine residue into the amino acid of interest, always based on the protein sequence.

Rotamers… *and* ***AutoFit Rotamer***, to find the best suited conformation of the residue according to its position and auto-fit it based on the geometric parameters of the backbone and the neighboring side-chains.

Change Phi and Psi Angles, to find the best conformation for the angles of the peptide bond that was formed with the already existing previous residue and re-direct the conformation of the backbone for the other residues.

Real Space Refinement, to refine all the geometric parameters and changes that I performed on the *newly-added* residue.

Each time a residue was added and edited based on the steps and tools mentioned above, I added the next residue, according to the protein's sequence, and performed the same steps until it was best fitted in the developing region (**Figure 57. A-D**). Note, that sometimes the conformation of the existing residues, located right before and after the missing regions, needed to be *edited* too, in order to redirect the backbone and allow the building of the missing regions.

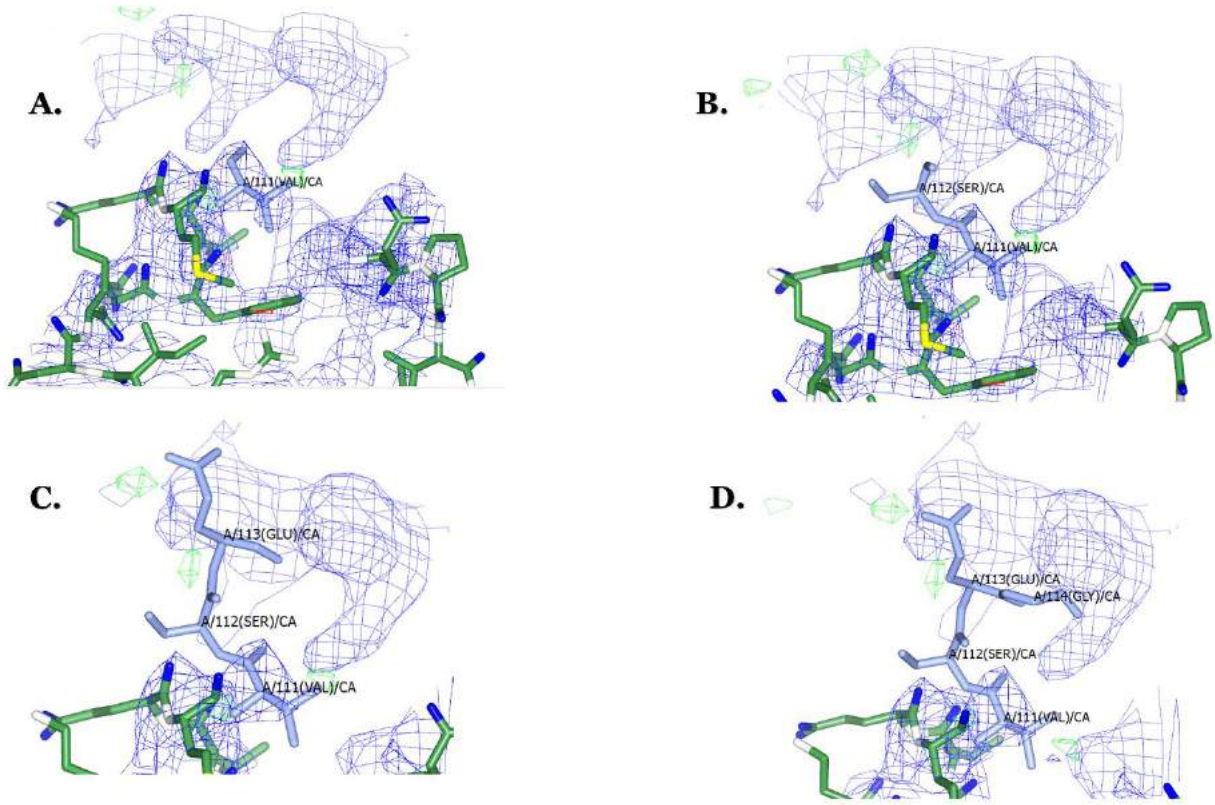


Figure 57| A. *Editing and re-adding the VAL/111 in the backbone of the Chain A, beginning in that way the development of the 112-122 missing loop in Chain A.* **B.** *Adding SER/112,* **C.** *GLU/113 and* **D.** *GLY/114, and continuing the building accordingly. The sigma (σ) factor for the maps is;* ***0.14*** *for the Direct Map and* ***0.18*** *for the Difference.*

In that point it is important to note that in Chain B the electron density indications are much clearer than in Chain A, something that can be explained from the difference of B factor that these chains present. In **Figure 58.1-2** is shown the first complete structure for the 112-122 region for both chains. The difference in the electron density map indications is quite evident; the Difference Map indications are much more clear in Chain B than Chain A, on sigma factor 0.18 .

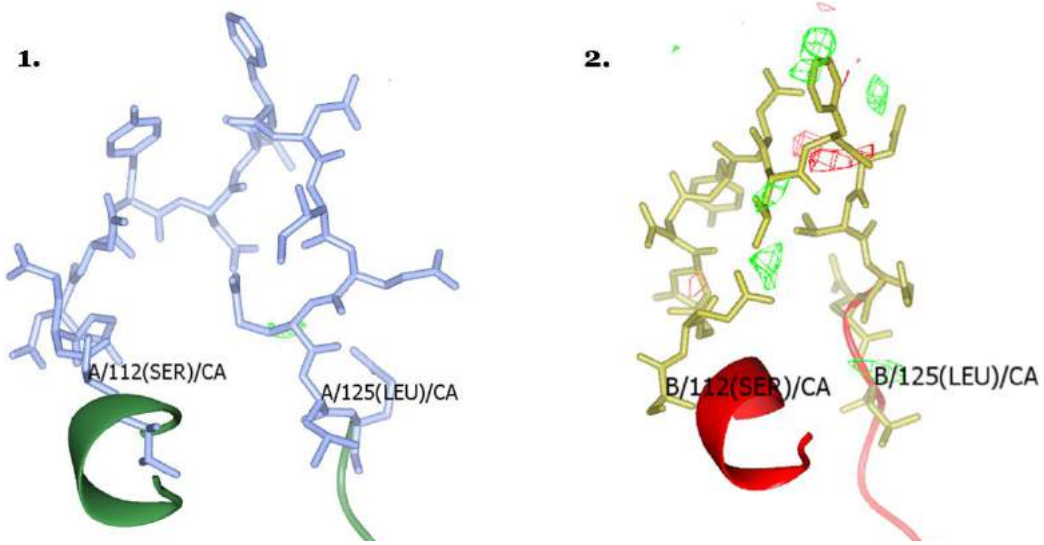


Figure 58| 1. *Chain A 112-122 region in grey-blue with less Difference map indications than.* **2.** *Chain B for the same region, in yellow. The previous and next regions are presented in "Ribbons" displaying format. The sigma (σ) factor for the Difference map is* ***0.18****.*

A complete representation of the complete missing 112-122 region is shown in **Figure 59. A-F**.

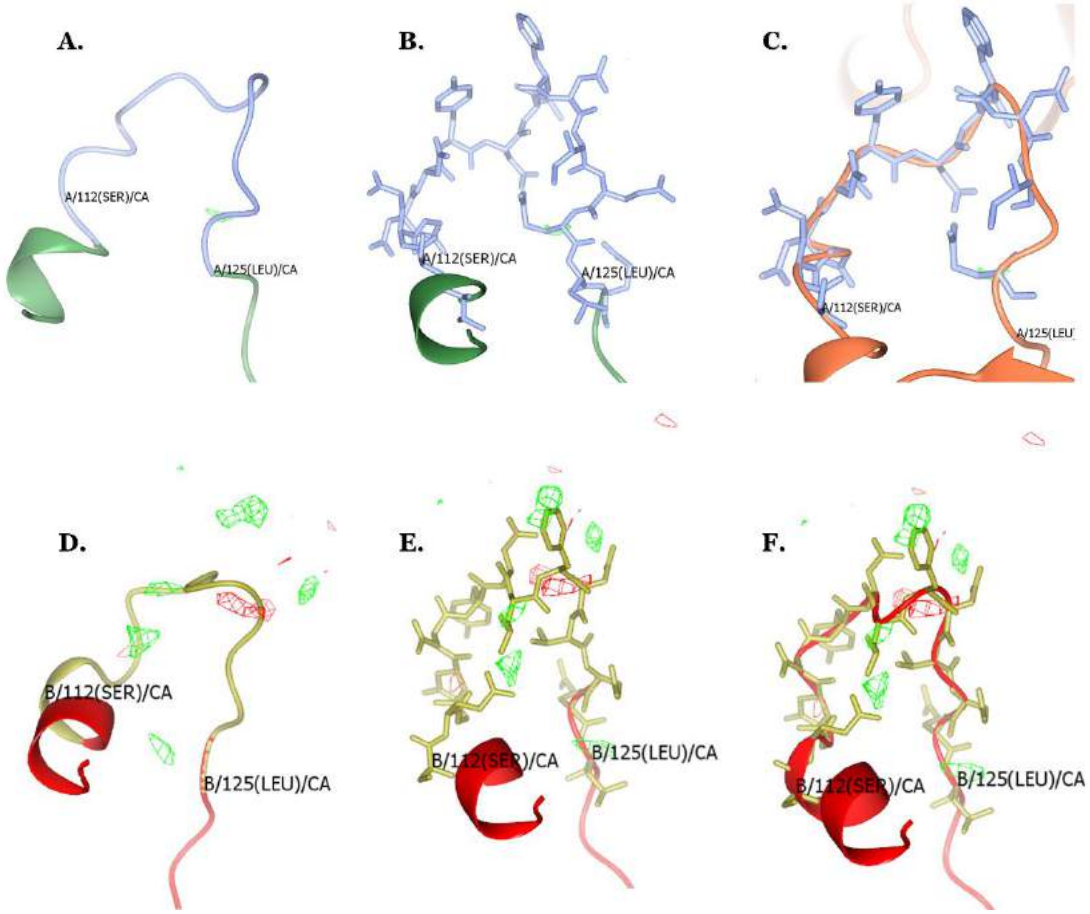


Figure 59| A-C. *Chain A 112-122 region in "Ribbon", "Bonds" and "Both" display, accordingly .* ***D-F.*** *Chain B with the same representation. The previous and next regions are presented in "Ribbons" displaying format. The sigma (σ) factor for the Difference map is* ***0.18****.*

The complete **238-243** turn is shown in **Figure 60.1-2,** below ;

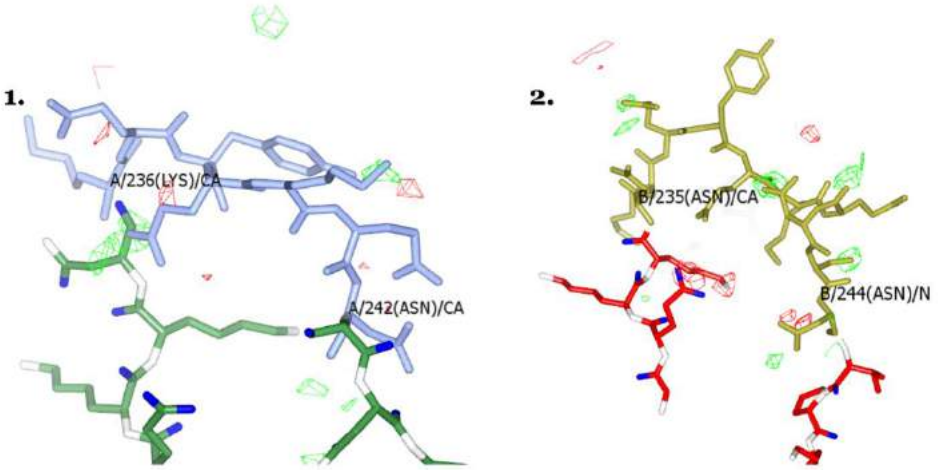


Figure 60| 1. *Chain A 238-243 region in grey-blue with* ***slightly*** *less Difference map indications than.* **2.** *Chain B for the same region, in yellow. The previous and next regions are presented in "Ribbons" displaying format. The sigma (σ) factor for the Difference map is* ***0.18****.*

The much clearer electron density data present in Chain B indicates that the model in built for that Chain can be used as a reference for Chain A in the following editing steps.

The focus of that step was to build *a* model for the loops and turns, even if there wasn't enough data or indications on how to position the missing residues. During Refinement there is a possibility that the electron density map will be more specific for these regions, because of the almost complete model.

The PDB file created from that step was named **file-0-editted.pdb**.

2. Step 2: Refinement & Omit Maps

In that Step I explain the cycles of Model Fitting-Refinement and how I chose to perform the Simulated Annealing procedure, to create Omit Maps and use them for the Optimization of the BsecIM model.

2.A. Fitting-Refinement Cycles

Once the building of the missing 112-122 and 238-243 regions was complete for both Chains I *run* Refmac5 with input files the **file-0-editted.pdb** and **file-0.mtz**, **.tls**, with the Set-Up mentioned above, performing the first Refinement. The result was an increasing in both R and R-free factors, in comparison to the results from the Refinement of **file-0.pdb**. It is time for the fitting processes of the OUTPUT model from the **file-0-editted** Refinement; the refined **file-1.pdb** model, using the **file-1.mtz** as the model's refined map

The focus now is to fit the already built loop regions, based on the refined maps. Once the fitting is complete, I would *run* once again Refmac5, using as a model the edited one and repeat fitting processes on the refined model (**Figure 61.A**). That process of *Fitting-Refinement* was repeated four times in general, until there will be generated as OUTPUT files from Refmac5 the **file-4.pdb, .mtz, .tls** . (**Figure 61.B**)

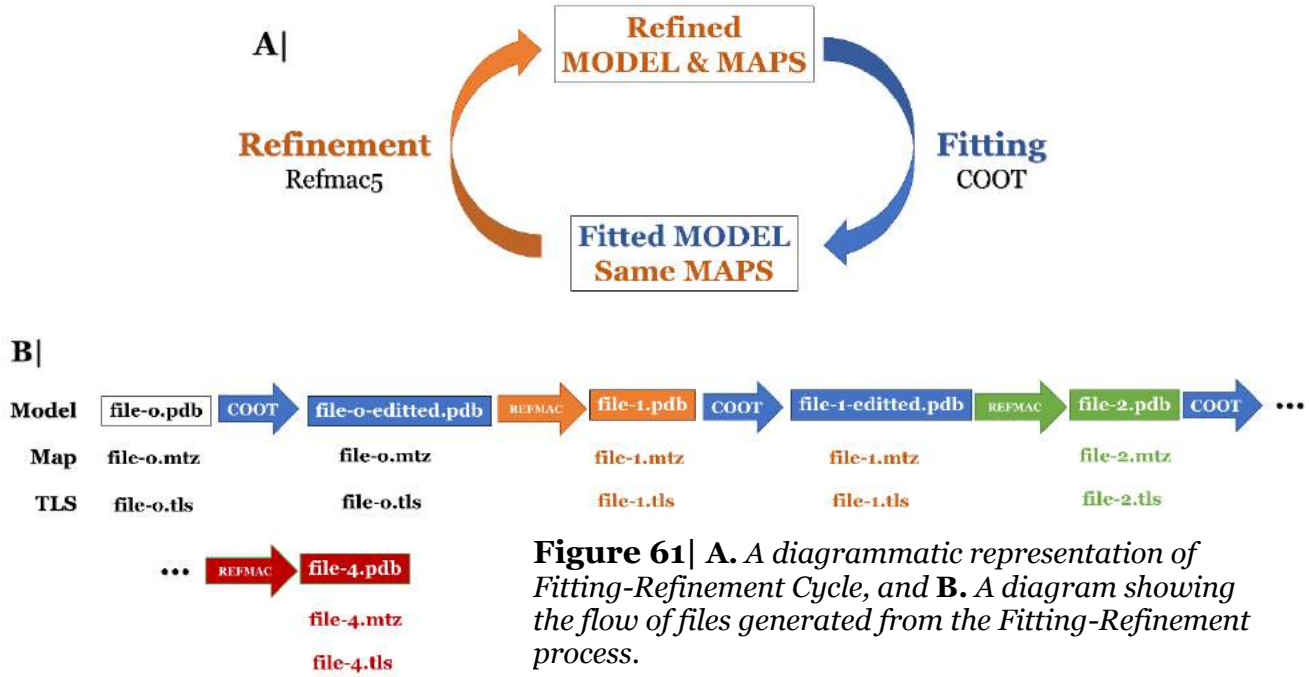


Figure 61| A. *A diagrammatic representation of Fitting-Refinement Cycle, and* **B.** *A diagram showing the flow of files generated from the Fitting-Refinement process.*

2.B. An Alternative Process: Simulated Annealing & Omit Maps

The Fitting-Refinement Cycles is usually a standard procedure in Model Building and Optimization (Smyth & Martin, 2000). But in certain cases, and especially in low resolution regions, there is a need for the use of alternative procedures, such as Simulated Annealing.

Up until this step, the 238-243 region seem quite *well fitted* according to the maps indications for both chains. Although, *well-fitted* is not quite valid as a characterization for the 112-122 region; there is still not enough evidence from the data for even a possible ***backbone orientation*** for that specific loop. The *Fitting-Refinement Cycles* did not seem to give more refined results for that specific region, no matter how many cycles were performed. (**Figure 62**)

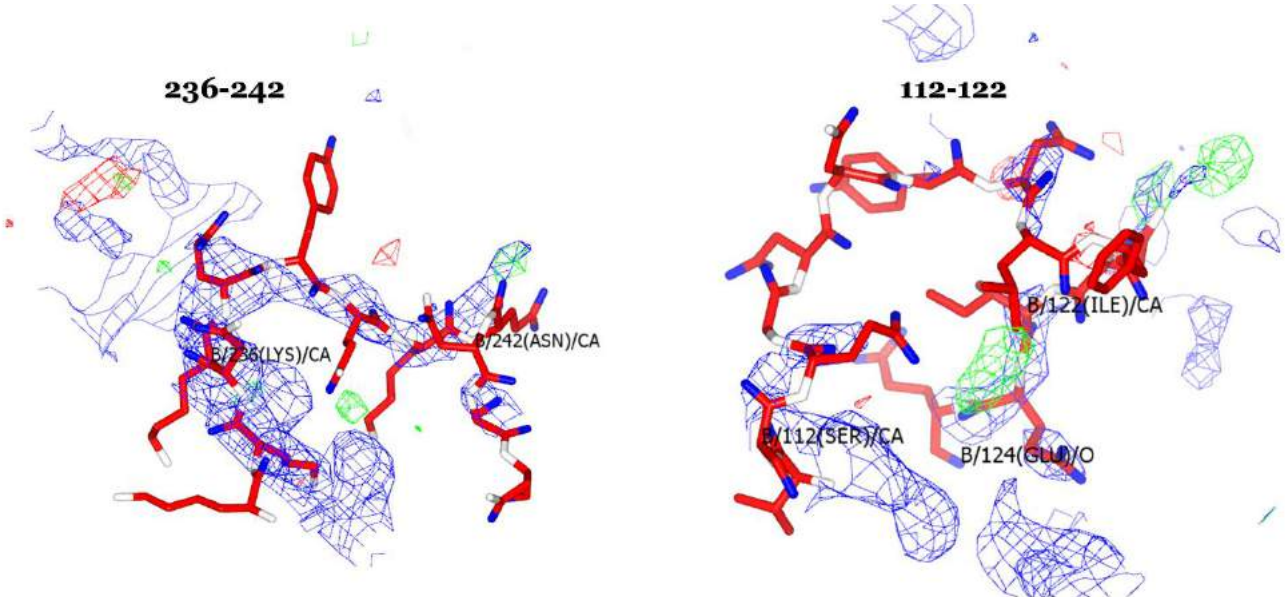


Figure 62| *In Chain B, the backbone orientation, and even some side-chains, of **236-242** is quite clear (**left**). In **112-122** (**right**), there is no evidence of any possible conformation. The sigma (σ) factor is at **0.17** for the Direct Map and at **0.14** for the Difference Map.*

To solve the problem mentioned above I used Phenix's *Optional Simulated Annealing*, also known as "Composite Omit Map" tool, to generate Omit Maps specified for the 112-122 region. *Composite Omit Map* has a very simple Set-Up procedure, which is described below;

First, I needed to set as INPUT files for the algorithm the model described in **file-4.pdb** and the map **file-4.mtz**, to use as a reference for the composition of the Omit. (**Figure 63.A**)

Second, to set the Annealing process of Omit Map Composition, I needed to set it in the corresponding setting field, in **Map Options** section (**Figure 63.B.1**). To specify the Simulated Annealing, in the **Atom selection** (**Figure 63.B.2**) I described to perform the Omit map composition specified for the residue range 112-122 for both chains, in the coding "resseq 112:122". And before the *run* I needed to set the temperature parameters in the **More options…** section (**Figure 63.B.3**)

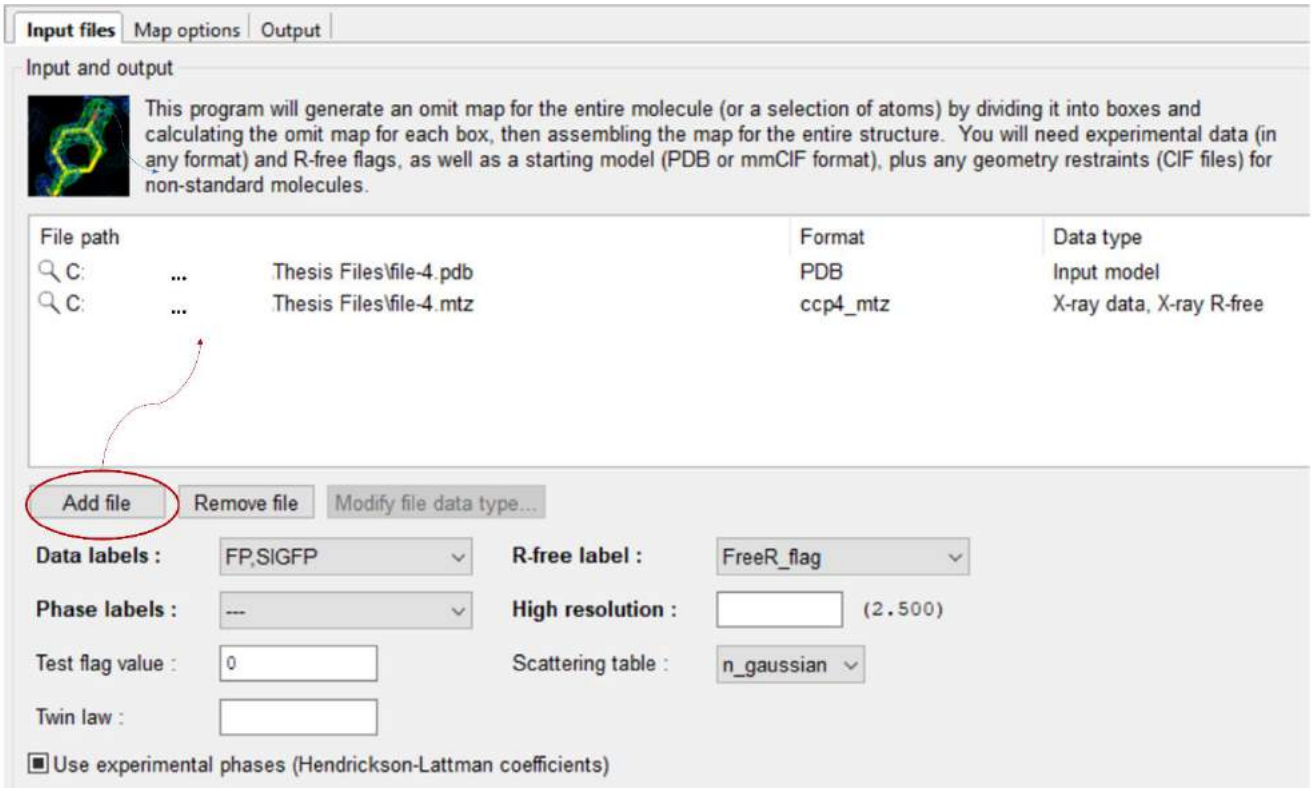


Figure 63.A| *The **Input files** interface of the "Composite Omit Map" tool in Phenix Suite.*

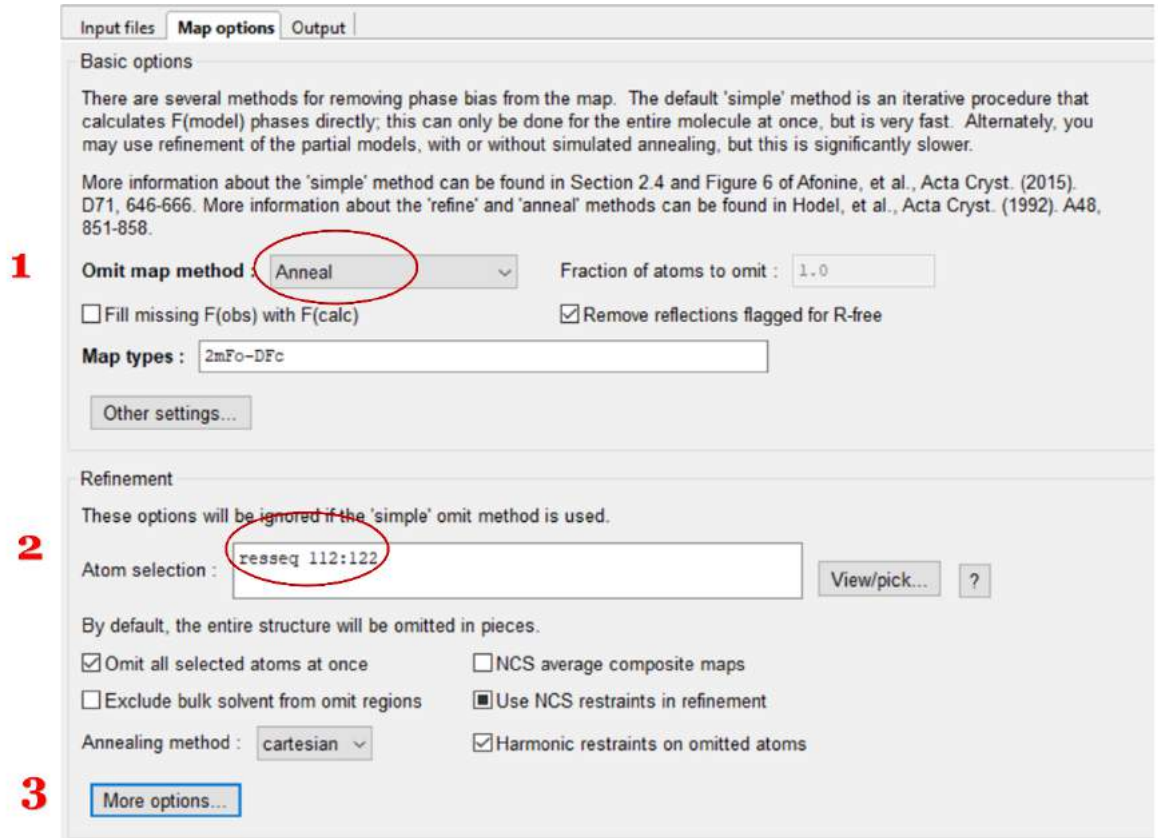


Figure 63.B| *The **Map Options** interface of the "Composite Omit Map" tool in Phenix Suite. **1)** Setting the Omit map method to **Anneal**, **2)** Set the regions of interest for both chains, and **3)** Press the **More Options...** .*

The **Annealing Temperature** defines the specificity of the Omit. I noticed that each time I lowered the Annealing temperature, the Omit was more and more specific for the region of interest. Starting from **2000 Kelvin**, I lowered it to **150 K**, which was the lowest point of temperature I reached for the Annealing.

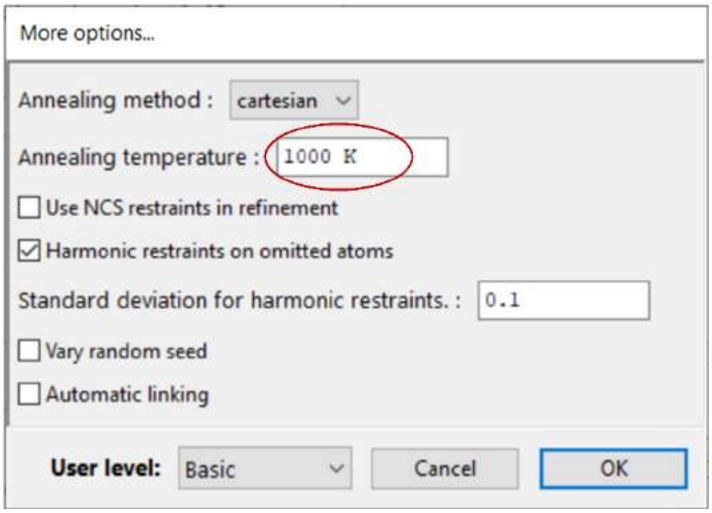


Figure 63.C| *The **More options...** section, including the Annealing Temperature setting.*

From the intermediate temperatures I used the map, which was generated at **250 K** Annealing Temperature. It is important to note that the "Composite Omit Map" would not work specified for the 112-122 region if that region was not already built in the model. Hence, the use of the **file-4.pdb** model.

The problem with the Composition of Omit maps in low Annealing temperatures is that, the lowest the temperature, the most possible is that the in the Map will be present false-positive indications. To solve this problem, I needed to perform a **Map Validation** process. There are at least two *Map Validation* tools; the Phenix Suite's ***Real Space CC*** and the CCP4 Suite's ***EDSTAT***.

The problem with both tools and the Omit Maps, which were generated with Phenix, was that the **.mtz** files of the Omit were missing some important information for the algorithms' settings. A possible solution to that problem would be to merge the **250k-omit.mtz** with the **file-4.mtz**, using CCP4'S ***CAD*** -a map merging tool- to generate an Omit file with the missing information and be able to perform the map's validation. But, the map merging is always risky to confuse the different MTZ information.

Putting aside the Validation process, the Omit could indeed work combined with the **file-4.mtz** as a more specific reference for the 112-122; In **Figure 64** the combination of both maps provides more indications for the loop.

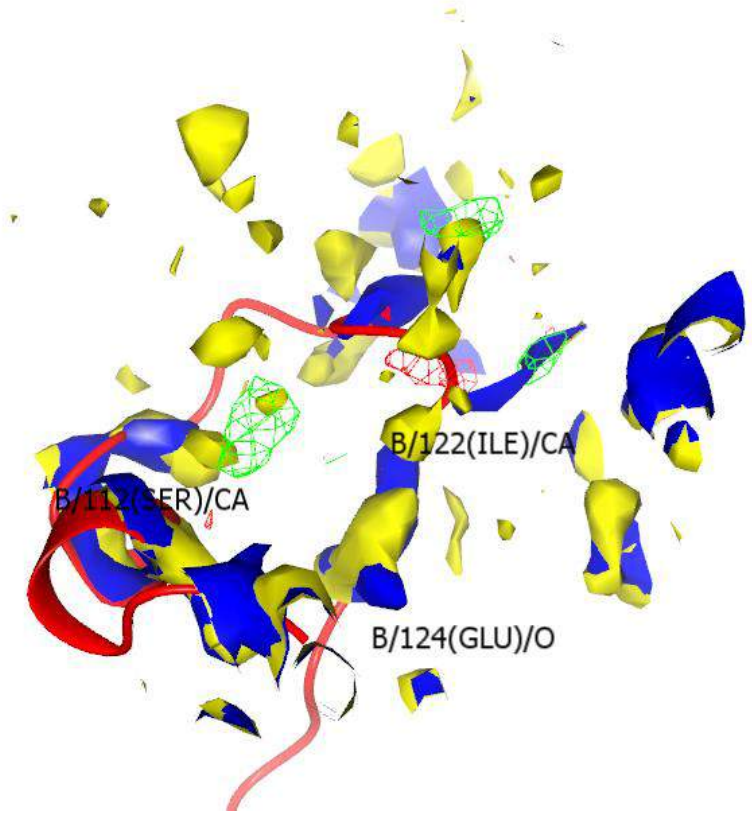


Figure 64| *The **112-122 region** in **Chain B**; The **file-4.mtz** Direct map is displayed with **solid blue surface** and the **250k-omit.mtz** Direct is displayed with **solid yellow surface**. Both maps are at **0.16 σ**.*

Another important evidence that the generated Omit Map could be proven useful was the fact that the COOT's function ***fit_gap*** could work when the **250k-omit.mtz** was the reference model's map.

The ***fit_gap*** is a **Python scripted function** in COOT, which performs the building and fitting of unmodelled regions in a molecule's model. To work, the model's map's resolution is necessary to be as high as possible for the region of interest. If it is high enough, some possible conformations for the region of interest will be built and fitted accordingly. If the resolution is quite low, the function will not *run*. With **file-4.mtz** the ***fit_gap*** could not run, but the **250k-omit.mtz** probably provided enough data for the function's *run*.

To run the ***fit_gap*** function in COOT I needed to type in the ***Coot Python Scripting*** the Python command, accordingly; ***calling*** the function, setting the ***reference map*** -the Omit Map- , specialized for **Chain B**, ***first*** and ***second*** residue for the region of interest, and the ***sequence of interest***. (**Figure 65**)

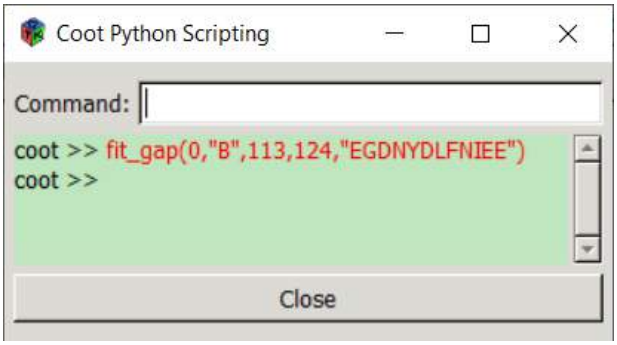


Figure 65|*Python Scripting the **fit_gap** function in COOT for the 112-122 loop in Chain B.*

Once the function is *set*, COOT starts building and fitting as many possible conformations for the region of interest. With the completion of the calculations, all the possible conformations are presented in a *Selection Box*. From these, I chose the structure, which was in most agreement with the possible amino acid interactions of the loop, and the geometric properties of the backbone orientation. (**Figure 66**)

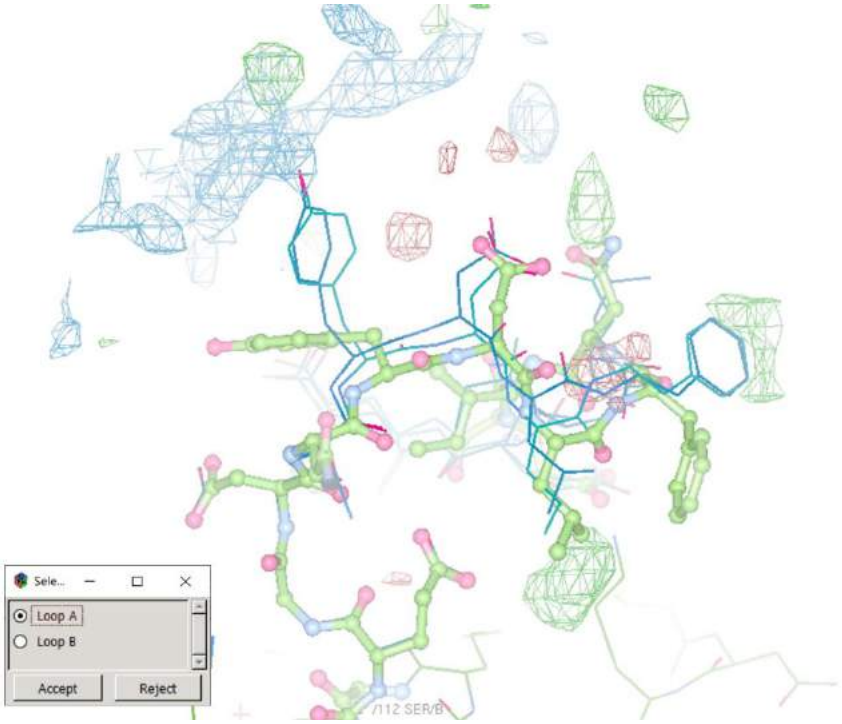


Figure 66| *The **112-122 region** in **Chain B*** after the ***fit-gap command*** *run, with two suggested possible structures for the loop. (the previously built loop is represented with **Sticks & Balls**, and the suggested structures with **Bonds** representation, in different colors)*

Each time the function was running, the previous model was used as a reference for possible corrections. I repeated the process until the ***fit_gap*** calculations started repeating the same structures, and the algorithm was *stuck in a loop*.

Once the final Chain "B" structure for the loop was ready, I used the COOT ***copy_residue_range*** Python scripted command to transfer that structure in Chain "A", in which the electron density data was significantly of lower resolution than Chain "B". For that function's *run* I used the following scripting setting;

copy_residue_range_from_ncs_master_to_others('*imol*',"B",113,114)

Where *imol* was the **file-4.mtz** map as a reference. The above configuration performs a *copying* process of the loop structure in Chain "B", to Chain "A". To fit the copied structure on the backbone of Chain "A", the region was directly refined with COOT's **Real Space Refinement**. (**Figure 67**)

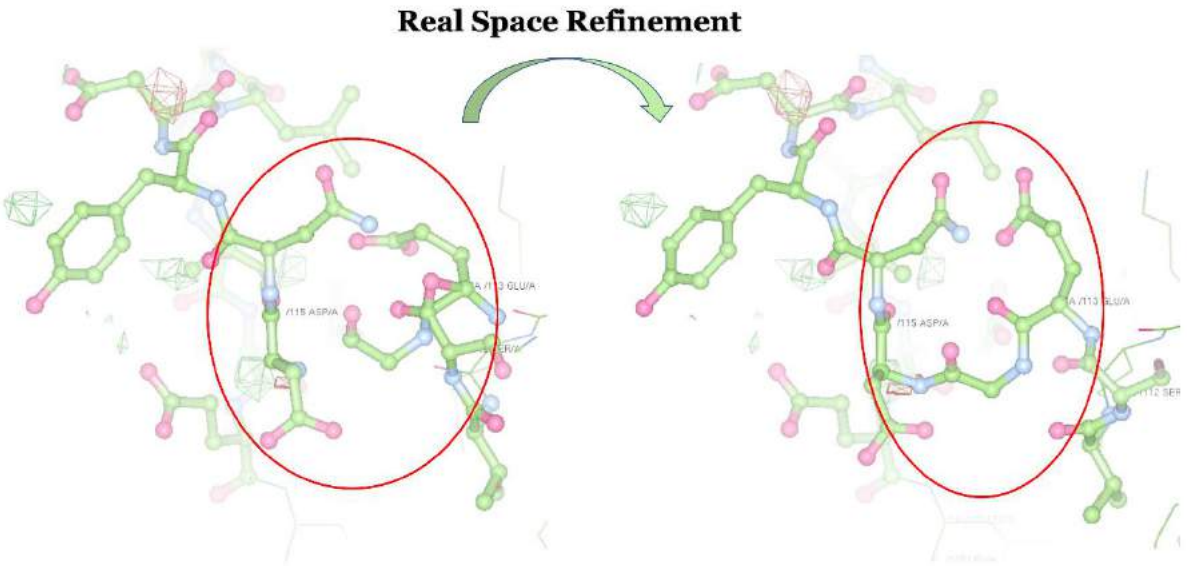


Figure 67| *The **copied 112-122 region** in **Chain A*** after the ***copy_residue_range command*** *before (**left**) and after (**right**) the use of **Real Space Refinement** in COOT.*

All the changes in the model were saved in the PDB file, named **file-4-editted** and refined with **Refmac5**. The OUTPUT files of the Refinement were named **file-5.pdb, .mtz, .tls**.

The reasons I trusted the **fit-gap** and **copy-residue-range** functions was due to the fact, that the structures, which were generated from these calculations, were seemingly more consistent to the peptides' geometric properties than the previous structures I built for these regions empirically. Even though the calculated Refinement factors before and after the Simulated Annealing process were quite close at rating with each other, the electron density indications for the loop regions after the Refinement of file-4-editted seemed to be slightly more specific.

In the next step, the focus is on the building of the N- and C-terminus.

3. Step 3: Building the Tails

At the beginning of the N- and C-terminal tails building step, there was no trace of the backbone orientation from the electron density data. In N-terminus, the model was beginning from **Thr/22**, and in C-terminal the **Lys/579** and **Tyr/580** final residues were missing. With COOT's **Add residue** tool it was not possible to *add* the missing residues and start the development of the missing tails, because of how low the resolution was.

Moving towards to the solution of that problem, I **placed some water molecules** at the electron density ***blobs***, in places where I thought that the backbone of the tails was supposed to be (**Figure 68.A**). These changes were saved in **file-5-editted-pseudo.pdb**. **After Refinement** the **file-6** series was generated; the map was slightly more evident for the backbone orientation in N-terminal of Chain "B", and also the missing **Lys/579** residue. A process I like to call; "**Pseudo Water Placement**".

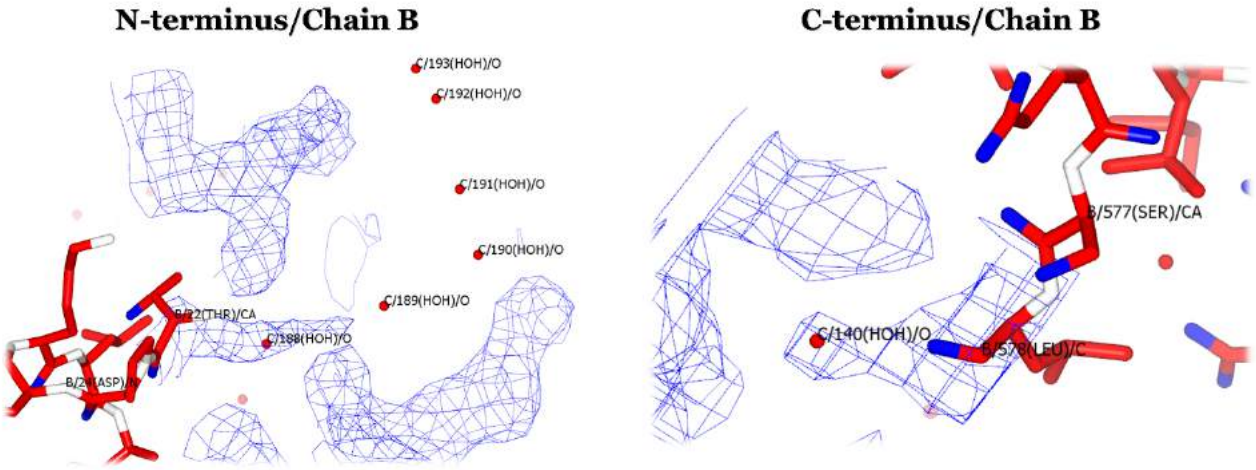


Figure 68| *The Chain B **N-terminus (left)** and **C-terminus (right)**. The **red dots** represent the water molecules placed accordingly to the "Pseudo Water Placement" process. The sigma in that Figure is at **0.14**.*

Once the map indications became much *clearer*, the placement of the final residues in the C-terminal tail in Chain "B" was finally *allowed*, and the completion of the region finished (**Figure 69**).

In the N-terminus, the electron density indicated a need for an *adjustment* in the already existing residue range **22-26** (**Figure 70**). Once the fitting of that region was complete, I started the building of the remaining N-terminus from **residue 23 to 1**.

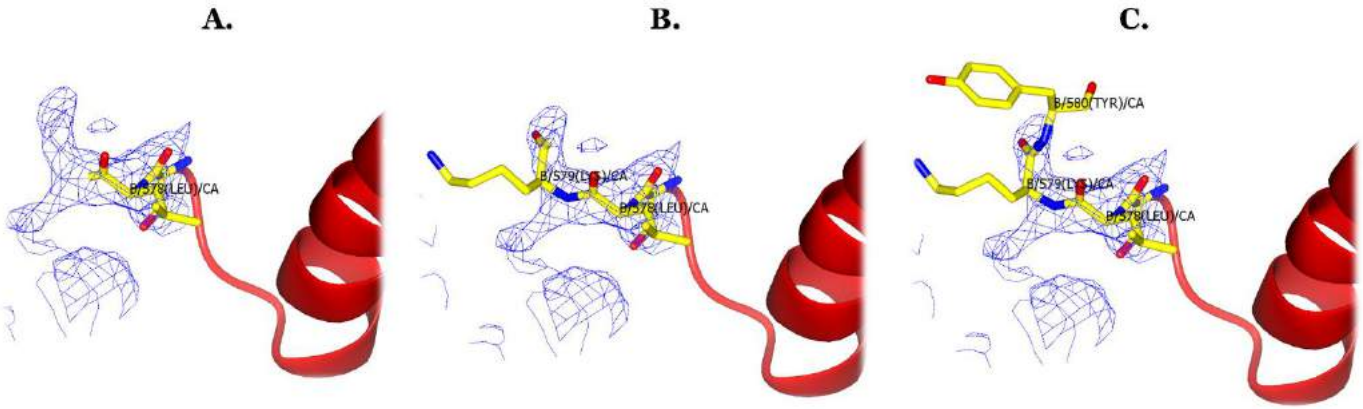


Figure 69| *The completion of **Chain "B" C-terminus**; The fitting of **A. Leu/578**, **B. Lys/579** and **C. Tyr/580**. Note: Sigma factor of Direct Map is at **0.14**.*

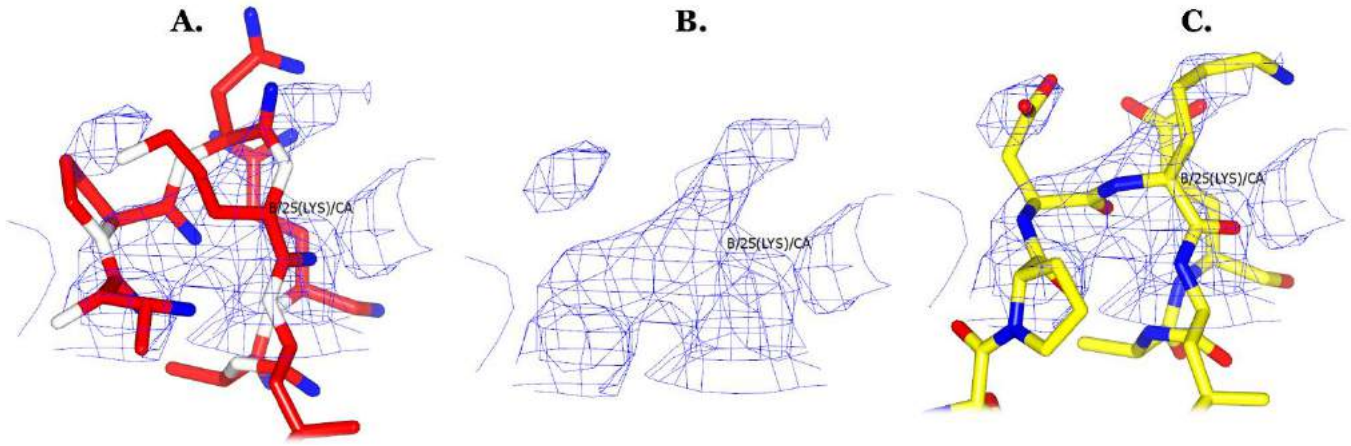


Figure 70| *The adjustment of the **22-26** region in **Chain "B" N-terminus**; **A.** Previous conformation, **B.** The Direct Map's indications after the "Pseudo Water" Refinement, **C.** The well-fitted region.*

To complete the N-terminal tail, each time new residues were added on it, a Refinement was performed. After each Refinement, I noticed that the Density map was *sharpened* more, reporting the next residue's place and, in that way, the continuity of the tail. The process would be complete until there was no further indication of the next residue to be added.

The process stopped in **Ala/16**. According to the UniProt sequence of BsecIM, from the missing residues, **6 were added** and **4 adjusted** to fit, as shown in the following one-letter diagram in **red** and **green** respectively;

? M M S V Q K A N T V S R Q K A T G A H F T P D K ...

5 10 15 20 25

Every **building stage** of the N-terminus in Chain "B" and the corresponding files are shown in **Figure 71**;

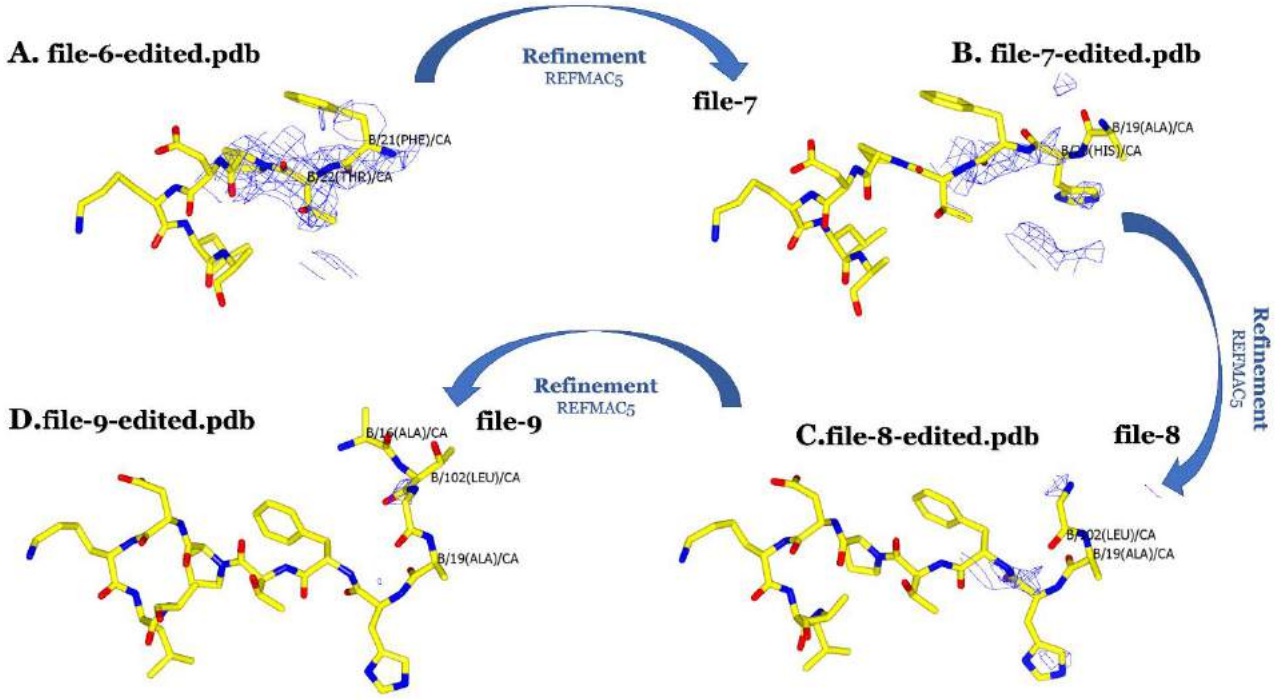


Figure 71| *The process of **Build-Fit-Refine cycles** in the backward building of **N-terminal tail** in **Chain B**; **A.** Adding **Phe/21** and run Refmac5 for Refinement, **B.** Adding **His/20** and **Ala/19** and Refine, **C.** Adding **Leu/18** and Refine, and finally **D.** Adding **Ala/16** and perform one last Refinement, generating the **file-10** file series.*

At the end of that Step, the N-terminal and C-terminal tail were **copied** with **copy_residue_range_from_ncs_masters** Python command script, from Chain "A" to Chain "B". Not just the copied regions in Chain "A", but the whole model was refined with **Real Space Refine All** COOT's tool. The changes were saved in **file-10-edited.pdb**.

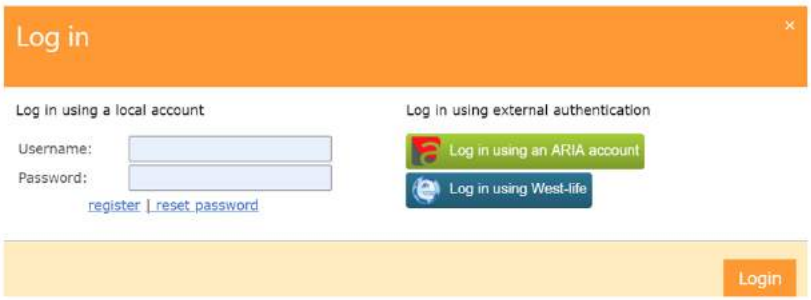
The last **Refinement** of the study was performed, generating the **file-11.pdb, file-11.mtz and file-11.tls**, which contain the data for the complete BsecIM crystallographic model.

The last step for that study is, Model Validation with PDB-REDO.

4. Step 4: Final Optimization; PDB-REDO

For the BsecIM's model last validation process, I used the PDB-REDO Server. To submit a model for Validation in the PDB-REDO Server, I needed to create an academic account (**Figure 72.A**). Through my account I could submit my MTZ and PDB files for a PDB-REDO run (**Figure 72.B**)

A.



B.

Figure 72| ***A.*** *The Login area in PDB-REDO and* ***B.*** *The* ***Submit new PDB-REDO run*** *section; the possible files that can be added are MTZ, PDB, mmCIF and FastA for sequence.*

Once the *job* is complete, the Server provides diagrams and tables of the changes on the model, how well fitted it ended being to the electron density map, along with the factors from the Refinement cycles it ran. (**Figure 72.C**) (Joosten, Long, Murshudov, Perrakis,2014)

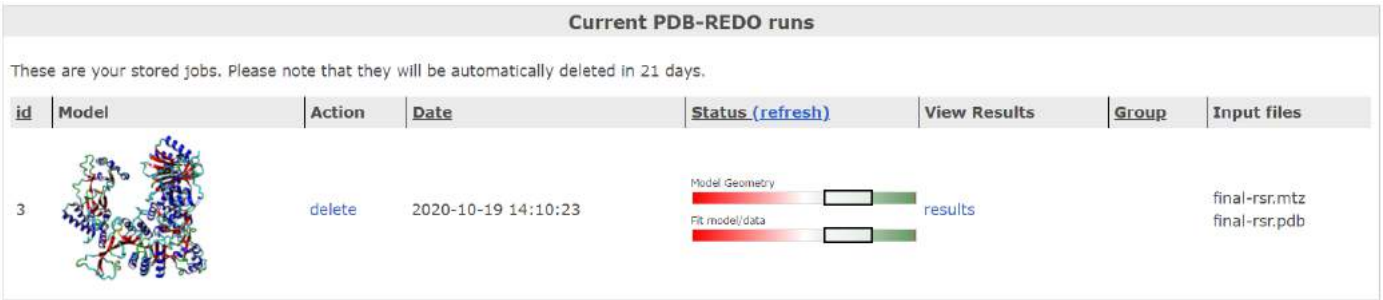


Figure 72.C| *The Job's stored results in PDB-REDO personal account.*

For this project I submitted **two PDB-REDO runs**;

The **first *run*** was with the **file-11.pdb** and **file-11.mtz**, to validate and optimize the model I studied.

The **second *run*** was with the starting files **file-0.pdb** and **file-0.mtz**, along with the BsecIM's UniProt sequence in FASTA file format, to study the efficiency of the PDB-REDO Server in comparison with the Computational Methods mentioned above in the section.

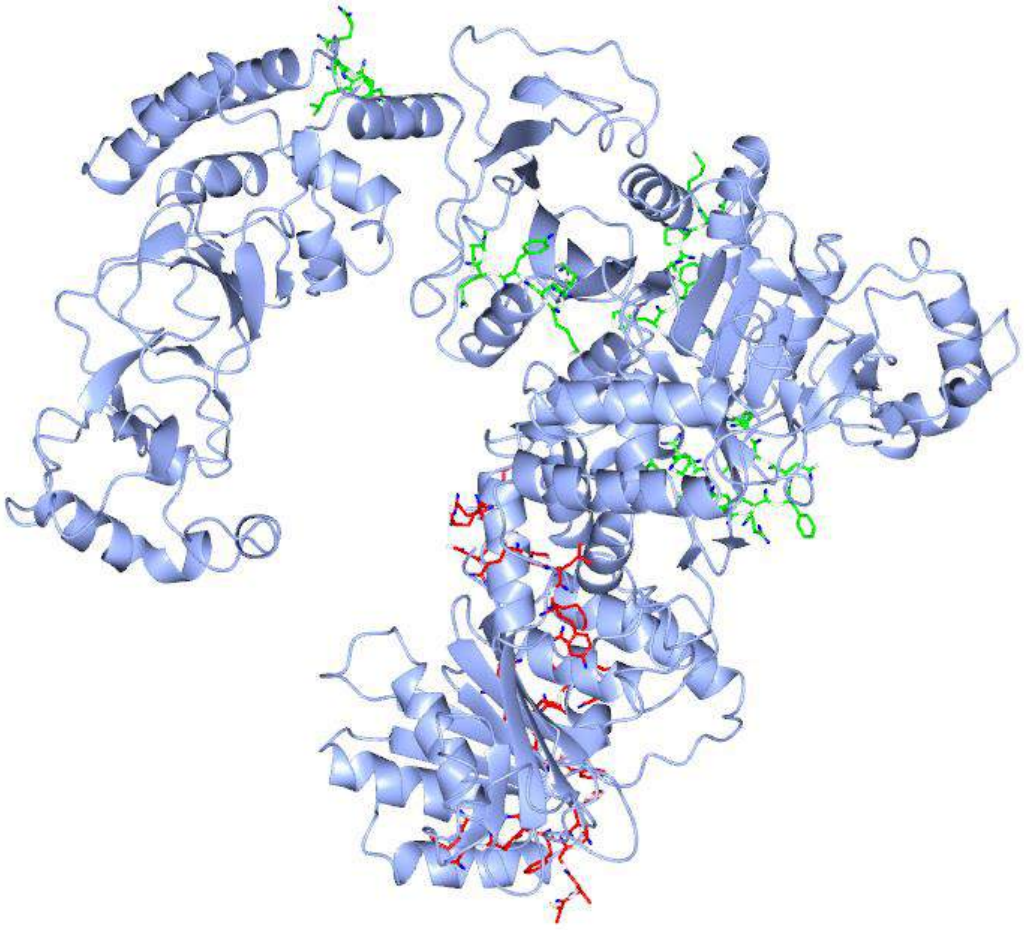
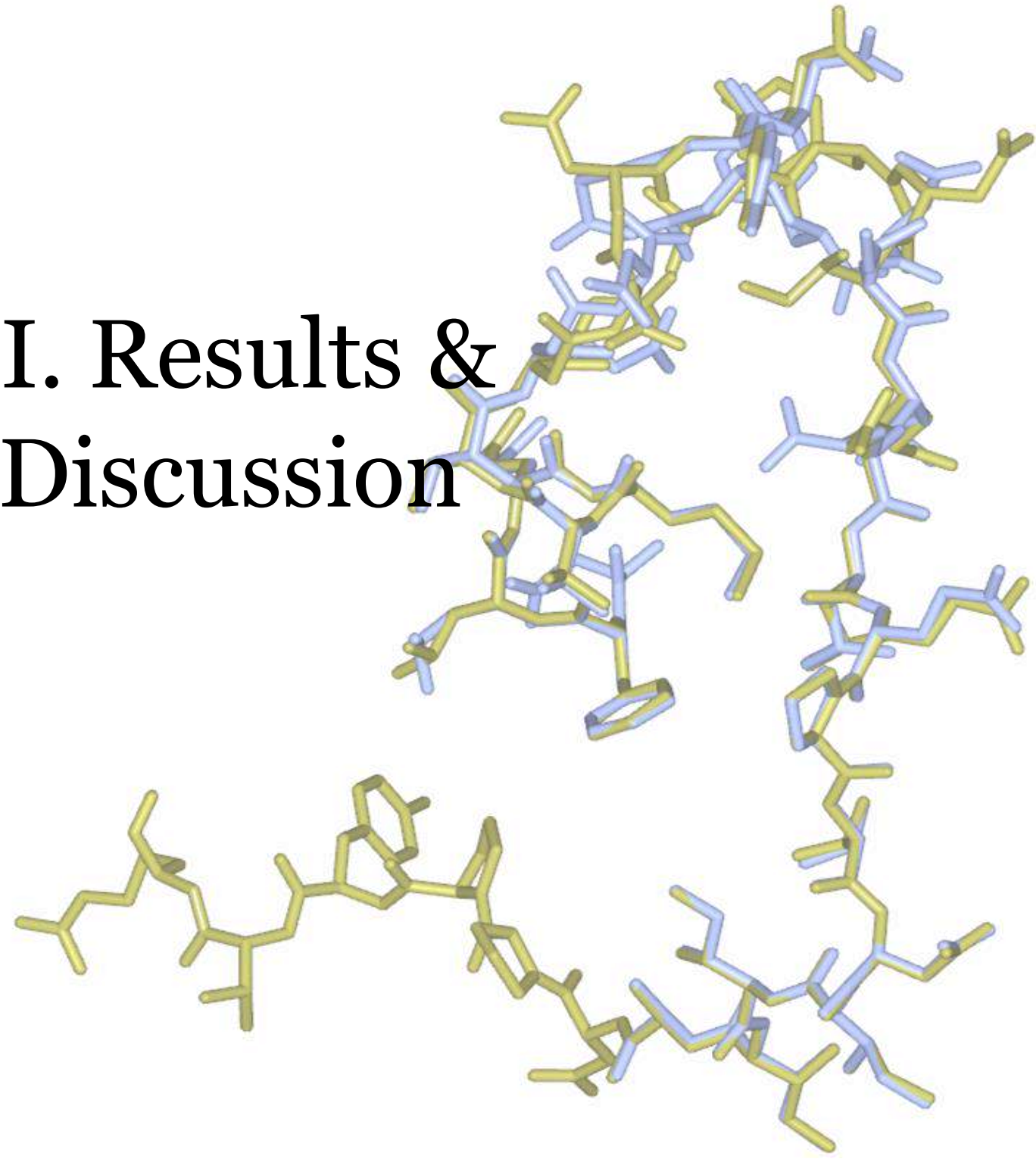


Figure 73| *The optimized BsecIM's model. The loops and tails regions mentioned in the study are represented in **green** and **red** color, in **Chain A** and **Chain B**, respectively.*

III. Results & Discussion



In this Section I will present the Results from every step of the BsecIM's Model Building & Optimization and at the end I will compare them with PDB-REDO Server's performance on the process of "Automated Model Building and Optimization". The performance of the process, in comparison with the results from the PDB-REDO Server, not only shows a lot about the effectiveness of these methods for the *construction* of a three-dimensional crystallographic structure, like BsecIM's, but also the possible intramolecular properties of the protein.

1. Refinement Results

In **Table 5** I present the *Initial* and *Final* **R-factor** and **R-free**, calculated from the **Refmac5 Refinement cycles**, for **file-0** and **file-1** refined structures. (Murshudov, Skubak, Lebedev, Pannu, Winn, 2011).

Table 5| *The Refmac5 results from **file-0.pdb, .mtz, .tls** files in comparison with the Refinement for **file-1.pdb, .mtz, .tls** files.*

OUTPUT: file-0	Initial	Final
R-factor	0.1848	0.1709
R-free	0.2180	0.2133

OUTPUT: file-1	Initial	Final
R-factor	0.1878	0.1708
R-free	0.2260	0.2200

The *Final* **R-factors** from the reference model and the first optimization are at the same rating. But there is an increasing in the *Initials*, and a significant increasing in the **R-free**. Seems like, once the first building of the loop regions and the *residue-by-residue* check was complete, I was quite focused on the *Fitting* process, that I fall in the *trap* of *overfitting* the structure to the electron density map, and paid less attention to the residues' biochemicals and geometrical properties.

After the results from the first Refinement, my main focus was on the loops and turns' conformation in the model.

In **Table 6.A & B** I present all the *Initial* and *Final* **R-factors** and **R-free factors** from every step of the Model Refinement, including the PDB-REDO Refinement factors results.

I excluded from the results the **file-6** file series because they contain the Pseudo Water molecules placement in the N- and C-terminal tails.

Table 6| *The Refmac5 Runs results from **A.** Step 1 & 2 and **B.** Step 3 & 4 of the Computational Methods, with the corresponding OUTPUT files, from each Refinement cycles set, mentioned.*

A| Step 1 & 2:

OUTPUT: file-1	Initial	Final
R-factor	0.1878	0.1708
R-free	0.2260	0.2200
OUTPUT: file-2	**Initial**	**Final**
R-factor	0.1882	0.1720
R-free	0.2286	0.2224
OUTPUT: file-3	**Initial**	**Final**
R-factor	0.1814	0.1721
R-free	0.2213	0.2212
OUTPUT: file-4	**Initial**	**Final**
R-factor	0.1881	0.1720
R-free	0.2285	0.2225

Model First Optimization & Loop Building-Refinement

OUTPUT: file-5	Initial	Final
R-factor	0.1809	0.1721
R-free	0.2217	0.2211

Simulated Annealing

B| Step 3 & 4:

OUTPUT: file-7	Initial	Final
R-factor	0.1809	0.1727
R-free	0.2203	0.2208
OUTPUT: file-8	**Initial**	**Final**
R-factor	0.1760	0.1721
R-free	0.2218	0.2216
OUTPUT: file-9	**Initial**	**Final**
R-factor	0.1762	0.1724
R-free	0.2215	0.2241
OUTPUT: file-10	**Initial**	**Final**
R-factor	0.1763	0.1729
R-free	0.2241	0.2259

Building Tails

OUTPUT: file-11	Initial	Final
R-factor	0.1836	0.1735
R-free	0.2251	0.2243

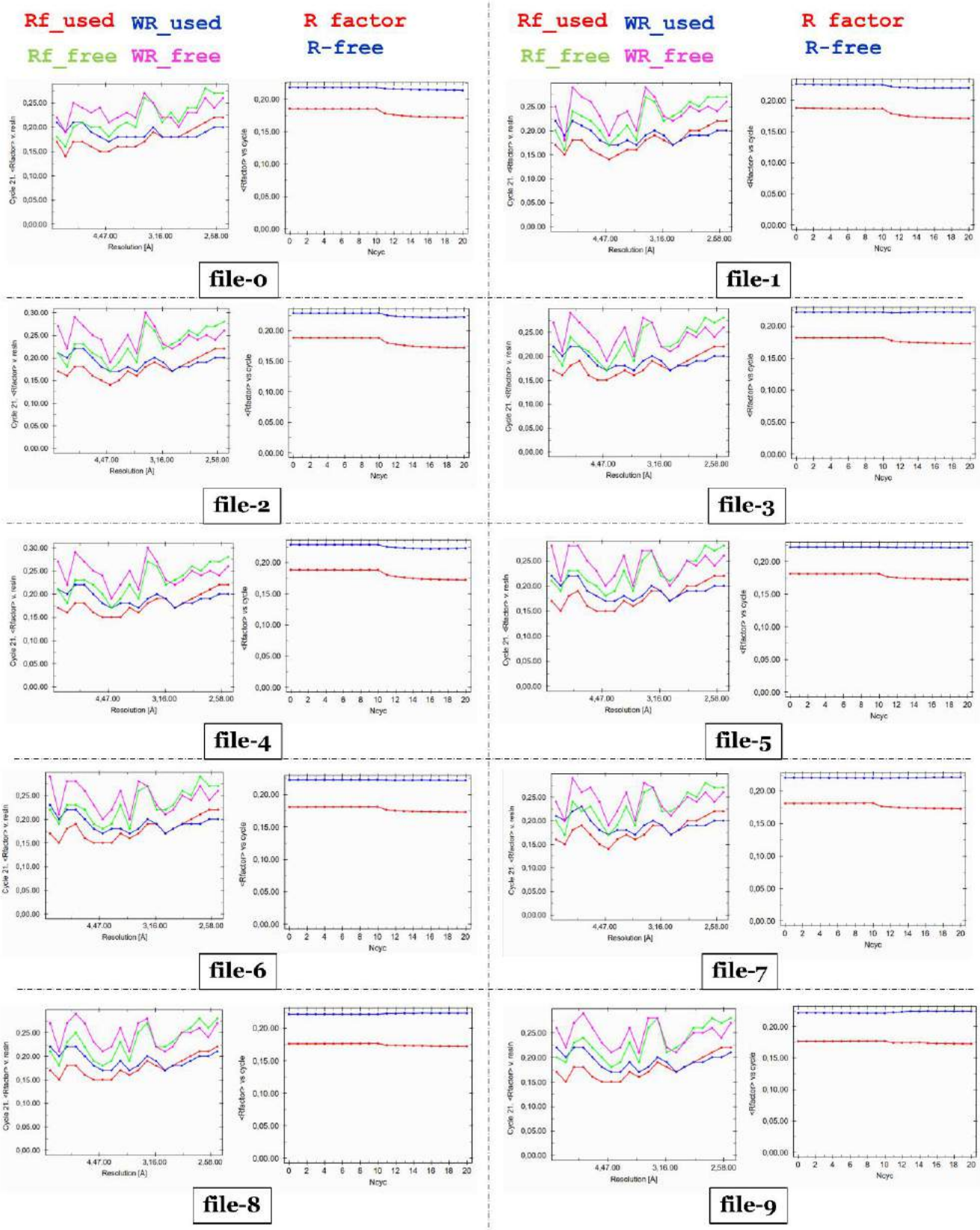
Final Optimization

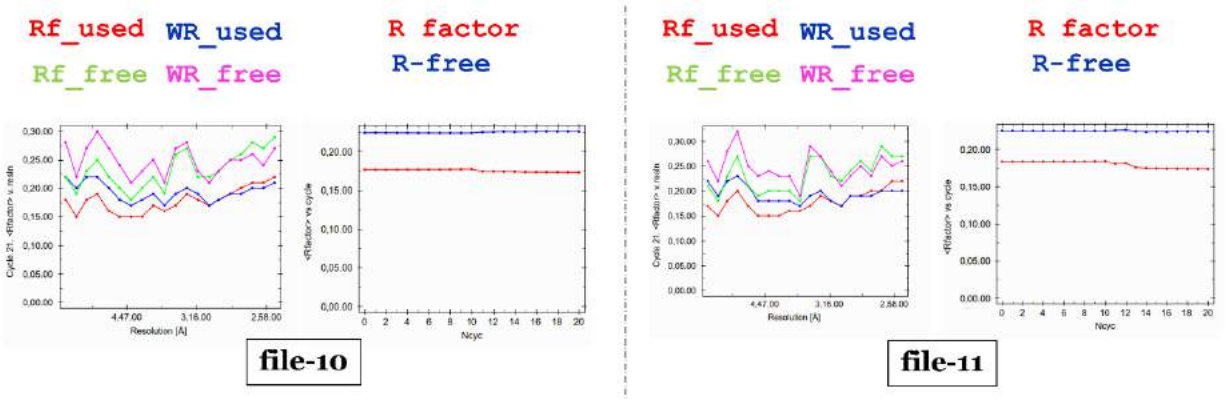
PDB-REDO		Final
R-factor		0.1590
R-free		0.2057

PDB-REDO

From the decreasing of **R-factor** there is evidence, that the protein's model reached a level of agreement with the molecule's crystallographic data. The significant decrease of both ***R*** and ***R*free** in the PDB-REDO Validation stage may point out that the Server's restraining settings were much more *loose* for the model's optimization, and therefore a further study on the Server's validation processes for molecules like BsecIM is probably needed.

Table 7| *The Refmac5 graphs for each Run: on the* ***right*** *of each file series the graph shows the distribution of* ***R*** *and* ***R*****free** *,along with the corresponding* ***weights****, on the last Refinement cycle, based on the resolution of the data, and* ***left*** *the fluctuation of* ***R*** *and* ***R*****free** *in the different Refinement cycles of a Refmac5 Run.*

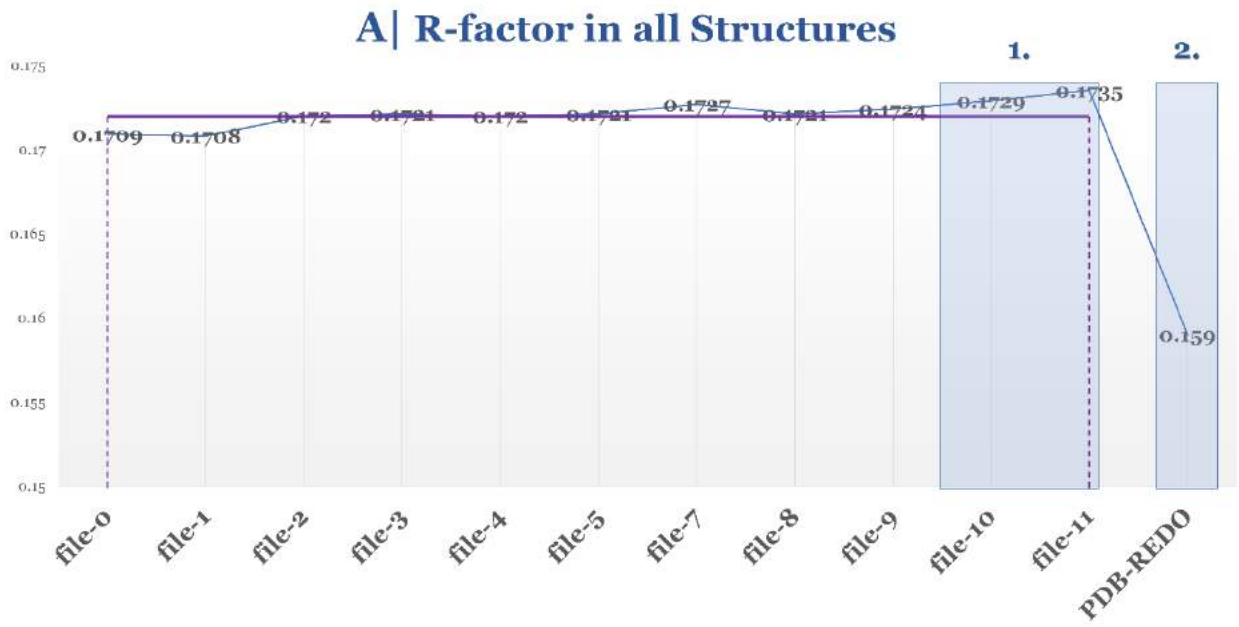




In **Table 7**, I present the analytical diagrams from the Refinement cycles of Refmac5 for the same optimization stages I presented in Table 6. For each file series the **left diagram** shows the **R factor vs resolution** distribution for the final Refinement cycle of that *run*, and the **right diagram** shows the overall ***R* and *Rfree* vs the Number of the Refinement cycles** in Refmac5.

As mentioned above, the difference between the *initial* and *final rating* of the Refinement factors was decreasing along the Model Optimization process. However, the *final* ***R* and *Rfree*** reveal how well the model and the map were refined. (**Table 8.A & B**)

Table 8| A. *The **Final R factors** calculated from the Refmac5 Refinement cycles, for each refinement stage of the model's optimization. The **purple** straight line indicates the **R-factors Average** for all the file series, **except the PDB-REDO R-factor,** and the blue **frames 1 and 2** show the points of smaller or greater deviation from that Average.* **B.** *The same for the **Final R-free**, where the straight line indicates the **R-free Average** -the PDB-REDO excluded-, and **frames 1 and 2** the deviated points.*



In the diagram of **Table 8.A** is shown the distribution of the *final* R-factors from the Refinement cycles, and the **Average**, which is at **0.1720**. In **Table 8.B**, the ***Rfree*** distribution is show, with an **Average** rating at **0.2220**.

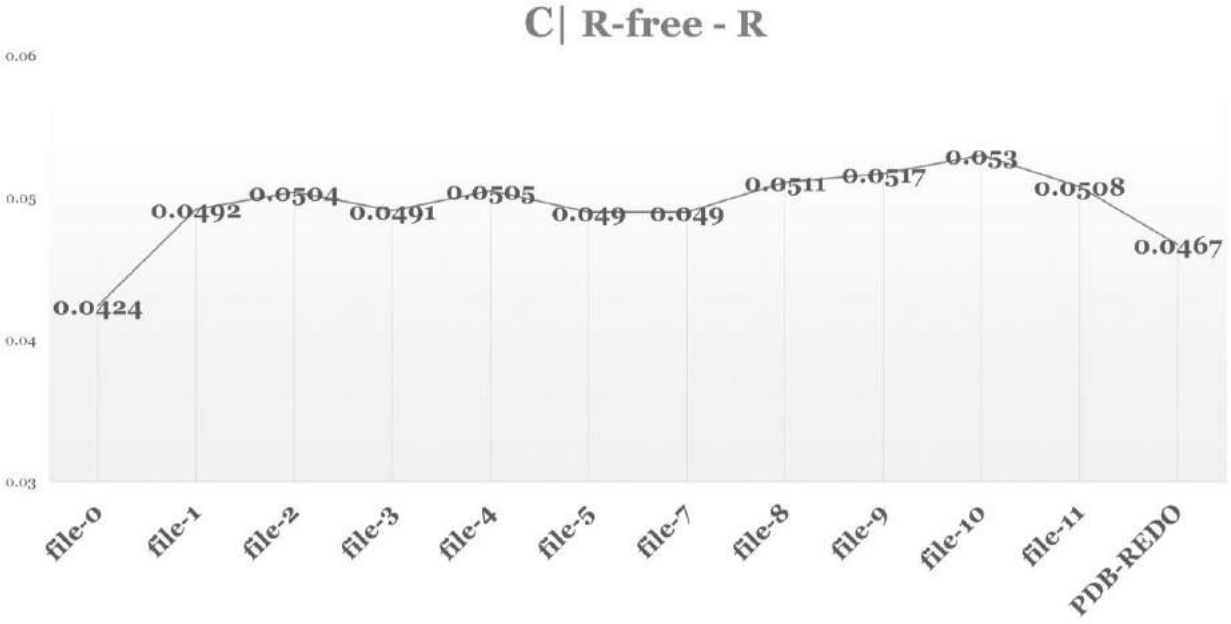
From **file-0** till the beginning of the repeated *Fit & Refinement* process it seems that both ***R*** and ***Rfree*** were increased and remained increased till the *Final Refinement* -before the PDB-REDO *run*. An increasing of ***Rfree*** occurred on the *Tail-building* step.

The increasing of the Refinement factors was almost inevitable, because of the focus on fitting and *gap-filling* the model, and especially in regions of *high mobility*, like the loop regions. The continuous increasing of the R-free on the stage where I built and fit the N- and C-terminal tails for the model, is probably due to the fact that **the tails were built in *noisy regions***. On the *final Refinement*, ***Rfree*** was decreased by **0.0008**.

On the final Refinement, after *copying* the missing regions from Chain "B" to Chain "A" and perform *Real Space Refinement* for the whole molecule, ***R*** was increased by **0.0015** from the ***R-factors'* Average**. The increasing was probably due to the *copying* of regions from one Chain to the other, and the fact that the copied regions were not in full agreement with the small -yet existing- indications of the electron density map on the Chain A.

The difference between ***R*** and ***Rfree*** show how effectively the performance of the Refmac5 Refinement process was. The closer these two factors tend to agree, the more the *refined* model agrees with the *refined* corresponding crystallographic data. In the graph of **Table 8.C** is clear that there is not a great divergence between the two factors for the whole process.

Table 8.C| *The graph of Difference between* ***Rfree*** and ***R***.



The increasing at the starting point is once again explained due to the *overfitting* procedures. Also, the more I was trying to model *noisy regions*, the greater the difference of the two factors was, but approximately at **0.05**.

In the case of other models, these increases and decreases of both ***R*** and ***Rfree***, and their difference, would probably be treated as small and insignificant changes. But for the sake of this study, is not. Even though the Refinement factors don't change significantly along the process, the generated electron density maps from each Refinement step are indeed more specified for these regions. (**Figure 74**)

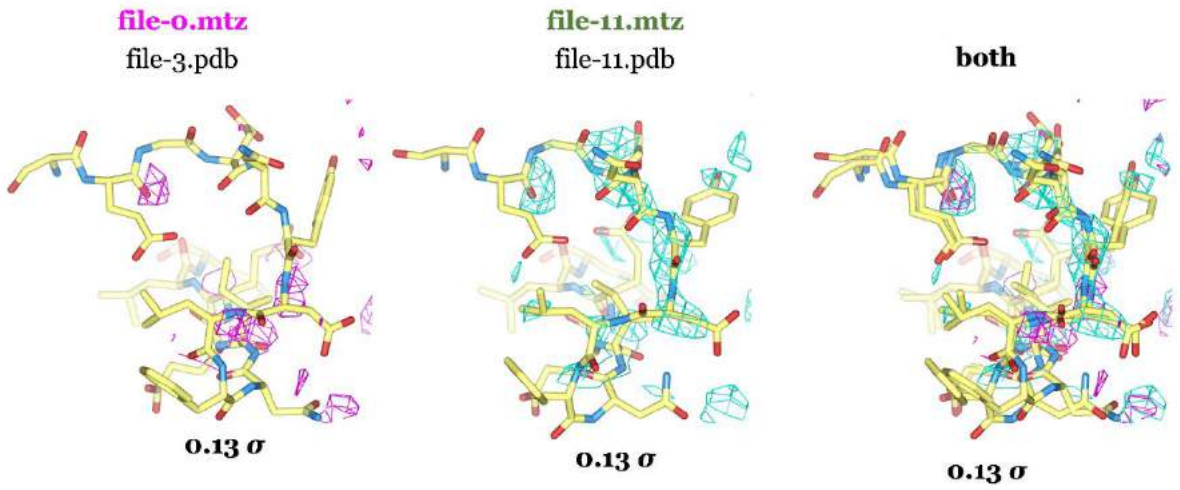


Figure 74| *The 112-122/Chain "B" region;* ***(left to right)*** *The* ***file-0.mtz*** *density map with the* ***file-3.pdb*** *loop, the* ***file-11.mtz and .pdb*** *for the same region and both, with clear indications for the backbone orientation. Both maps are* ***0.13 σ.***

And more importantly, the **Table 8 graphs** show that the PDB-REDO results are significantly lower than the Average of the Refinement factors. This probably means that the Automated Validation *corrected* the model in all ways possible, but also raises questions that the Server's *algorithmic process* was *quite strict* with the parameterization of the Validation.

2. Empirical vs Automated Optimization

As mentioned in the computational methods section, there were two PDB-REDO submissions; the **file-11** Validation (***Job 1***), and the **file-0** as control submission (***Job 2***). Both *jobs* showed a significant decrease for both ***R*** and ***Rfree*** (**Table 9**) (Joosten, Long, Murshudov, Perrakis, 2014)

Table 9| *The PDB-REDO results for both **file-0 (Job 2)** and **file-11 (Job 1)**; **Rfree** and **.R**.*

	File-0	File-11
R-factor	0.1600	0.1590
R-free	0.2068	0.2057

The analytical reports for ***Job 1*** and ***Job 2*** are shown in **Table 10** and **Table 11 respectively**;

JOB 1: file-11

(Already complete and empirically optimized structure)

Table 10.A| *The PDB-REDO **validation metrics report** for **file-11 (Job 1)***

Validation metrics from PDB-REDO	PDB	PDB-REDO
Crystallographic refinement		
R	0,1828	0,1590
R-free	0,2199	0,2057
Bond length RMS Z-score	0,560	0,427
Bond angle RMS Z-score	0,789	0,663
Model quality (raw scores | percentiles)		
Ramachandran plot appearance	24	30
Rotamer normality	41	49
Coarse packing	38	52
Fine packing	24	43
Bump severity	51	53
Hydrogen bond satisfaction	23	33

Table 10.B| *The PDB-REDO **reported model changes** for **file-11 (Job 1)***

Significant model changes	
Description	**Count**
Rotamers changed	27
Side chains flipped	1
Waters removed	87
Peptides flipped	16
Chiralities fixed	0
Residues fitting density better	31
Residues fitting density worse	2

Table 10.C| *The PDB-REDO* ***Model quality compared to resolution neighbors Graphs*** *for* ***R-free,*** *the* ***Ramachandra plot angles*** *and the quality of the* ***Rotamers****. For* ***file-11 (Job 1)***

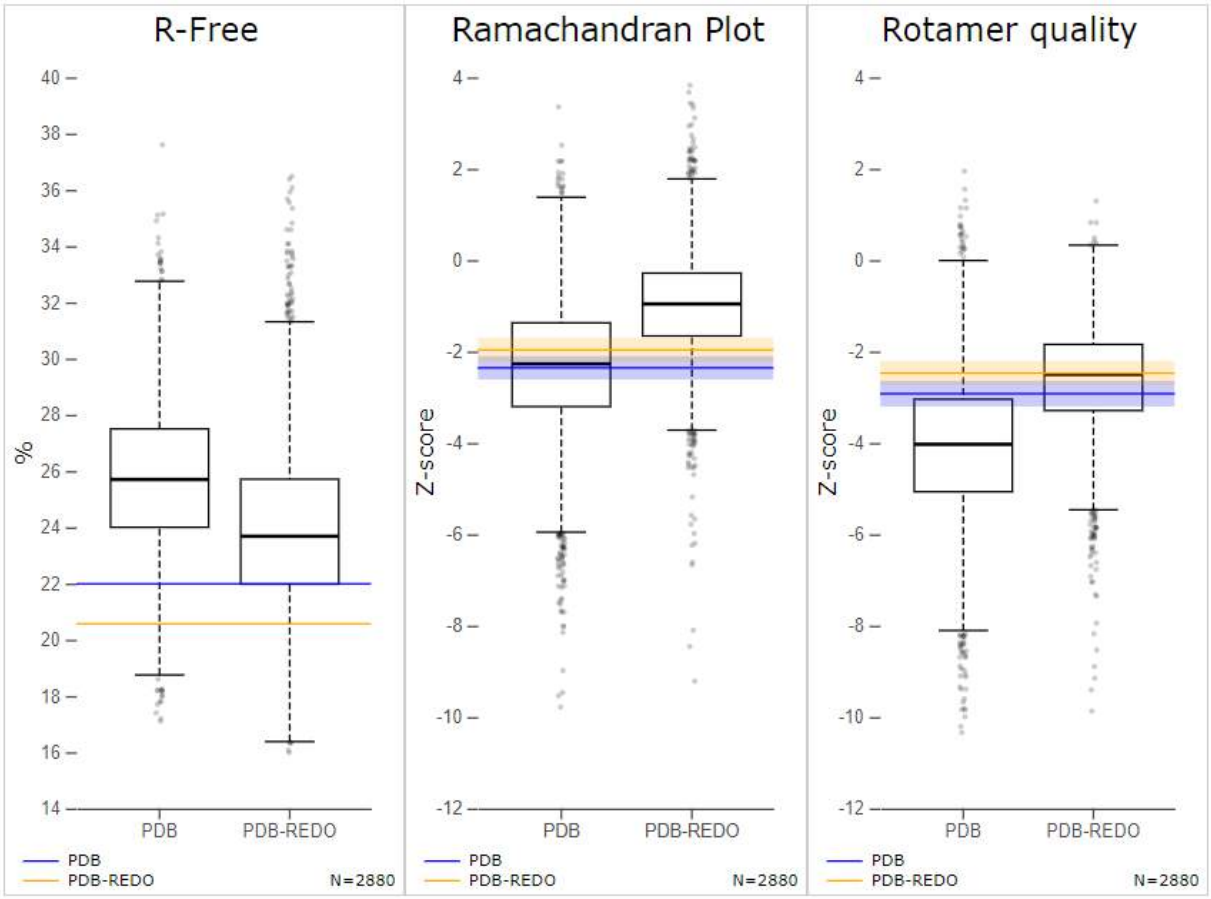


Table 10.D.i| *The PDB-REDO* ***report of Changes in density map fit (RSCC)*** for all the **previously unmodelled regions** in **Chain "A"**. *For* ***file-11*** *(****Job 1****).*

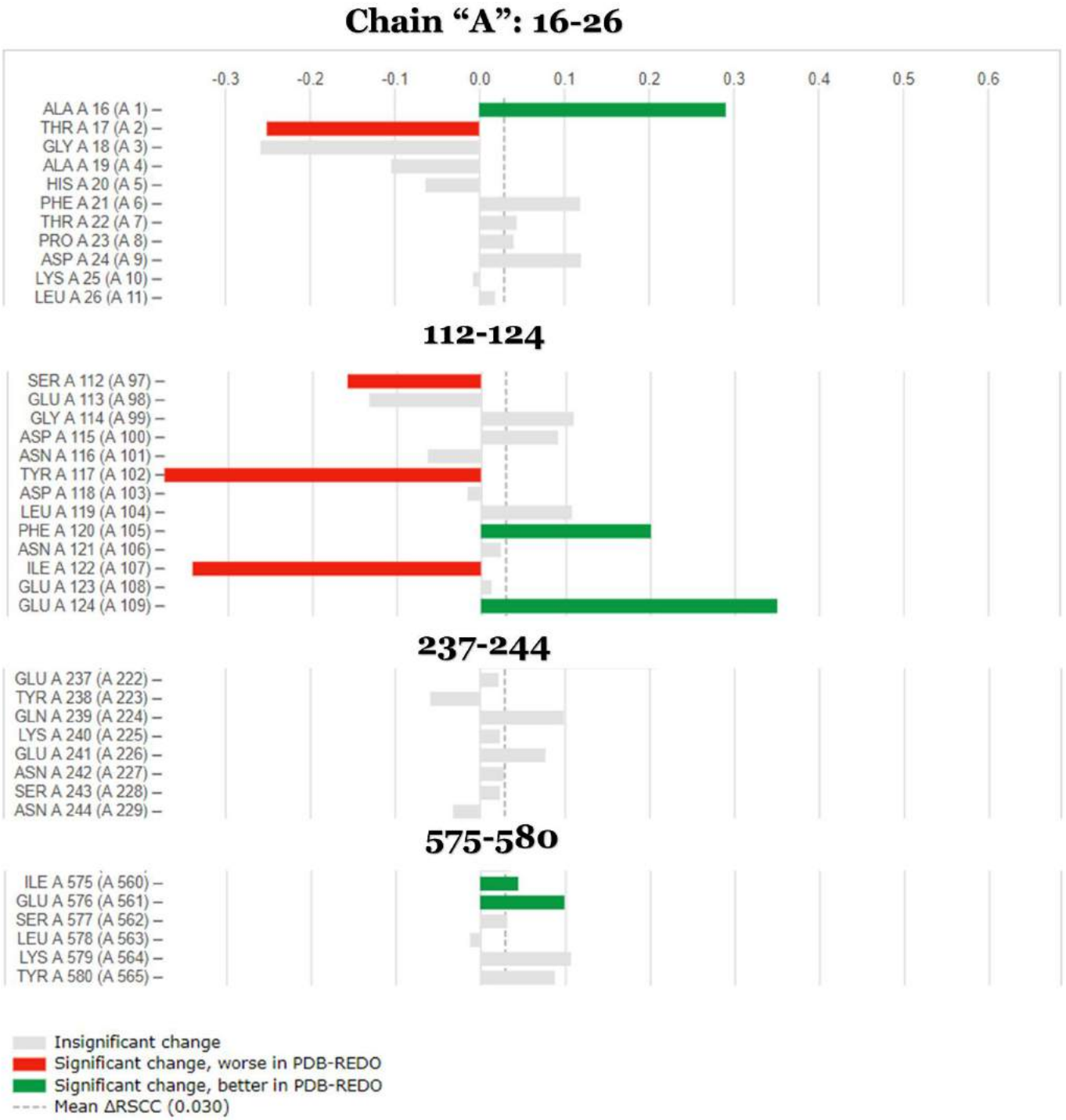
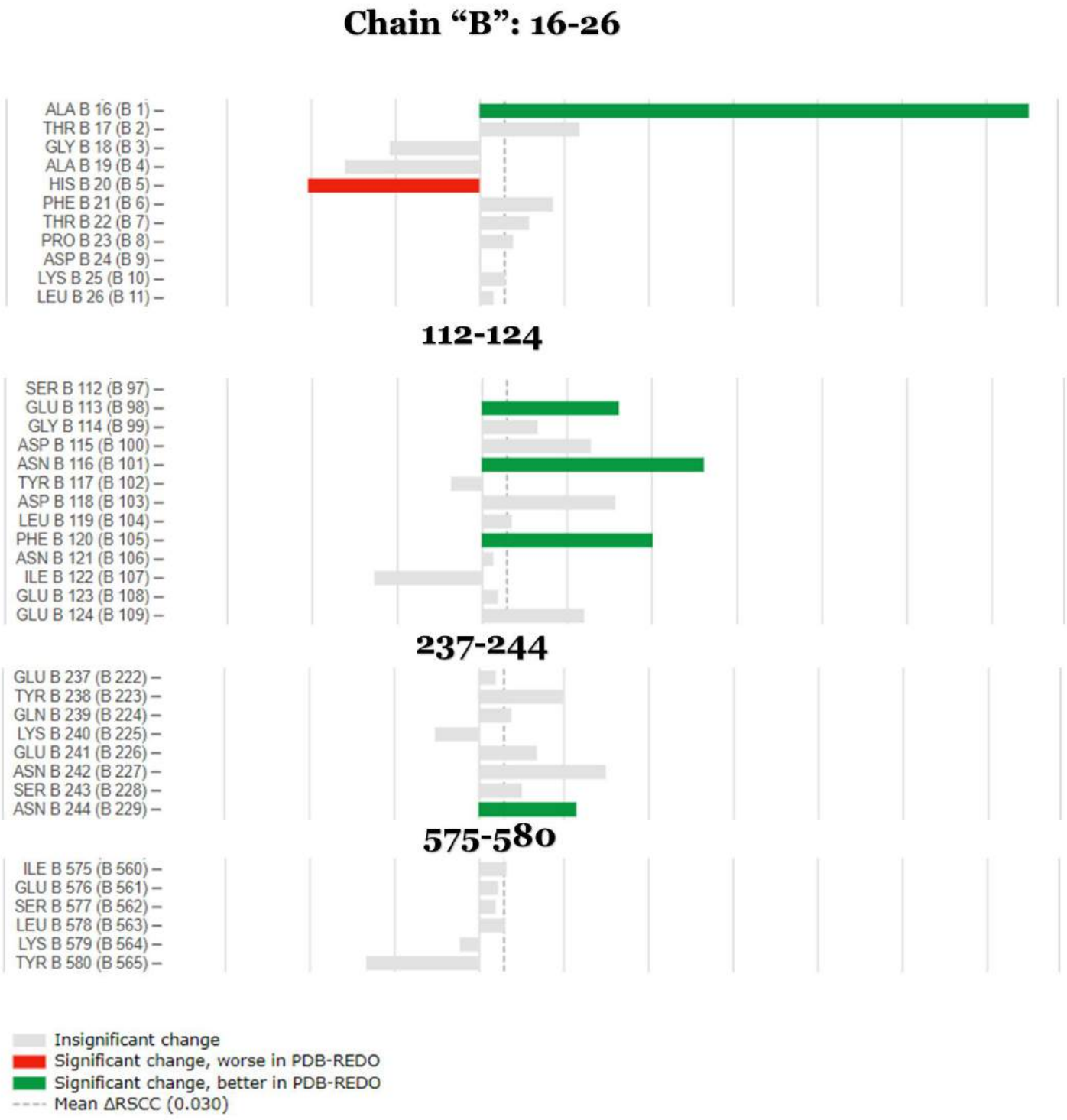


Table 10.D.ii| *The PDB-REDO* ***report of Changes in density map fit (RSCC)*** for all the **previously unmodelled regions** in **Chain "B"**. *For* ***file-11 (Job 1)***.



The statistics from the PDB-REDO Validation show some significant changes and corrections of the model. Even the regions I studied, which were quite low resolution regions, the algorithm seemed to be able to refine and fit them, with some significant fitting changes, as they are described in **Table 10.D.i-ii**.

The **RMSD** between the **file-11** model and the **PDB-REDO** validated one is at **0.27**. In **Figure 80.A and B** I show the superposed file-11 and PDB-REDO regions on **Chain "A" and "B"**. The main differences between the two models in these regions are the changing of some rotamers -but still quite similar- and some torsion angles in extent.

Since these regions tend to be *quite flexible* in the molecule, the changed conformations in the **PDB-REDO** model are not in a significant level different from the **file-11** model.

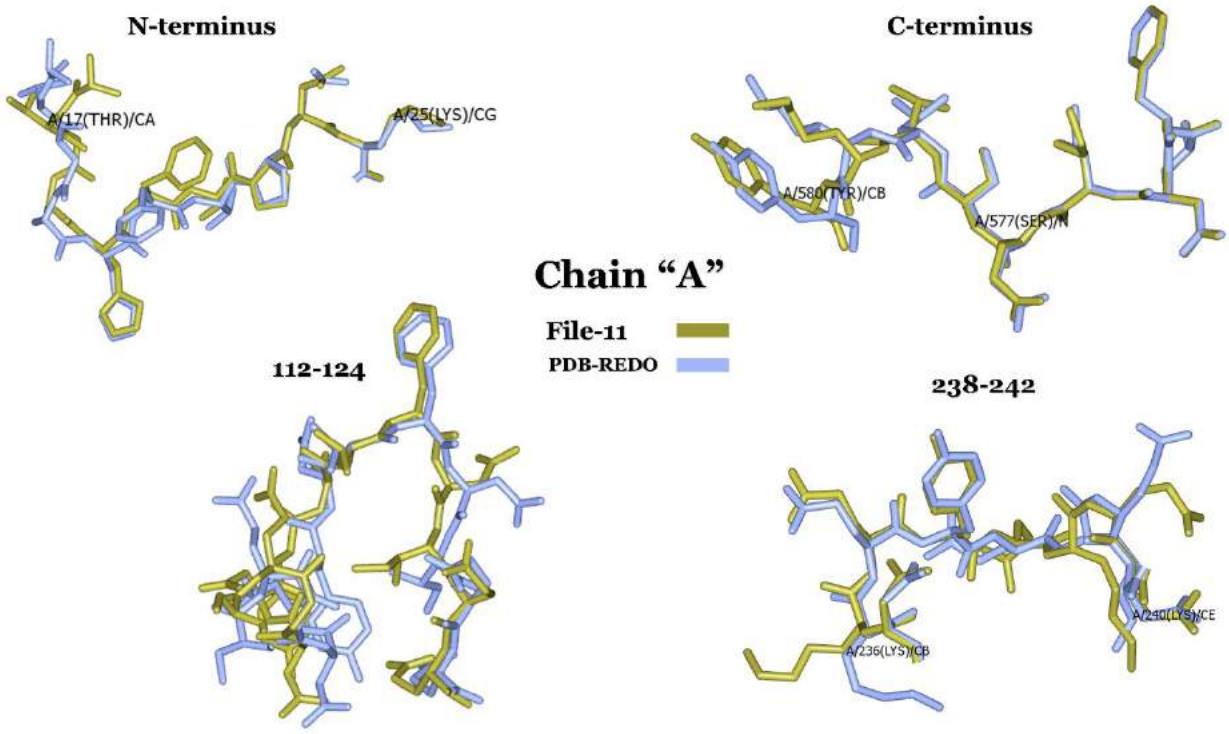


Figure 80.A| *The regions of interest for this thesis of* ***file-11*** *superposed with the* ***PDB-REDO*** *validated model for Chain "A".*

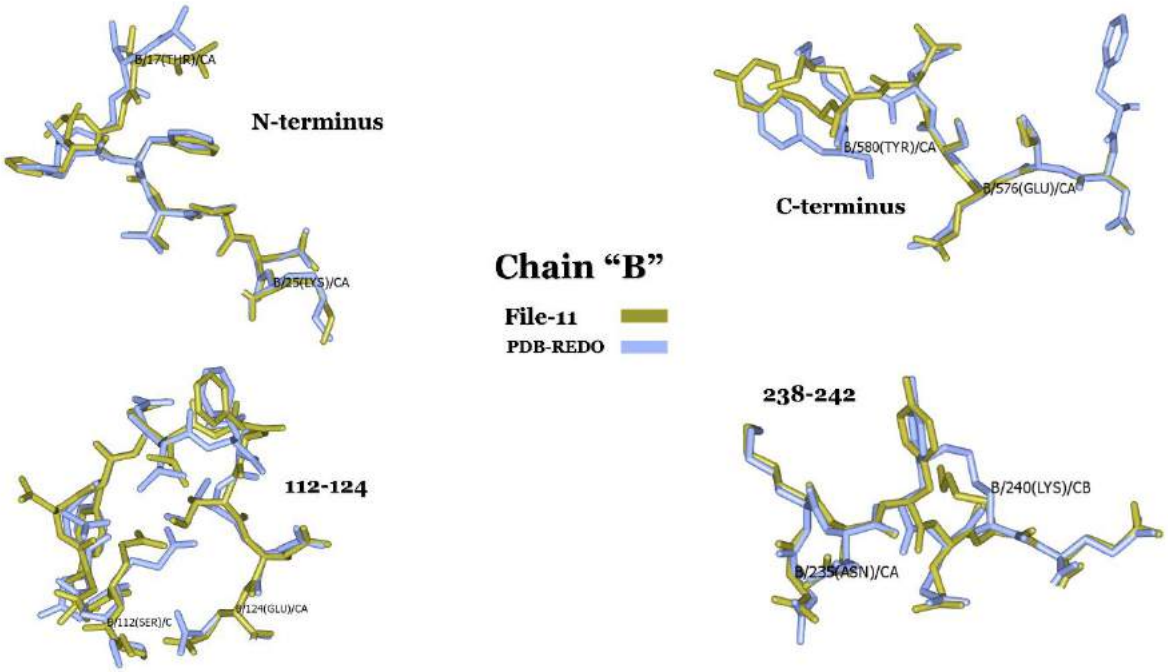


Figure 80.B| *The regions of interest for this thesis of **file-11** superposed with the **PDB-REDO** validated model for Chain "B".*

The other *Job* was also quite important, because it works as *control* for the comparison between the *Empirical Optimization*, that I studied in this thesis, and the *Automated* one, which is performed with PDB-REDO. For the Optimization of **file-0** model with the Automated Process of PDB-REDO, the **file-0.mtz and .pdb** were needed for submission, and the protein's **sequence** in **FASTA file format**. Using the sequence as reference, the PDB-REDO allows the building of the missing regions, if the data is clear enough. The whole process is described as ***Job 2***.

JOB 2: file-0

(The control *Job* for the effectiveness and performance of PDB-REDO)

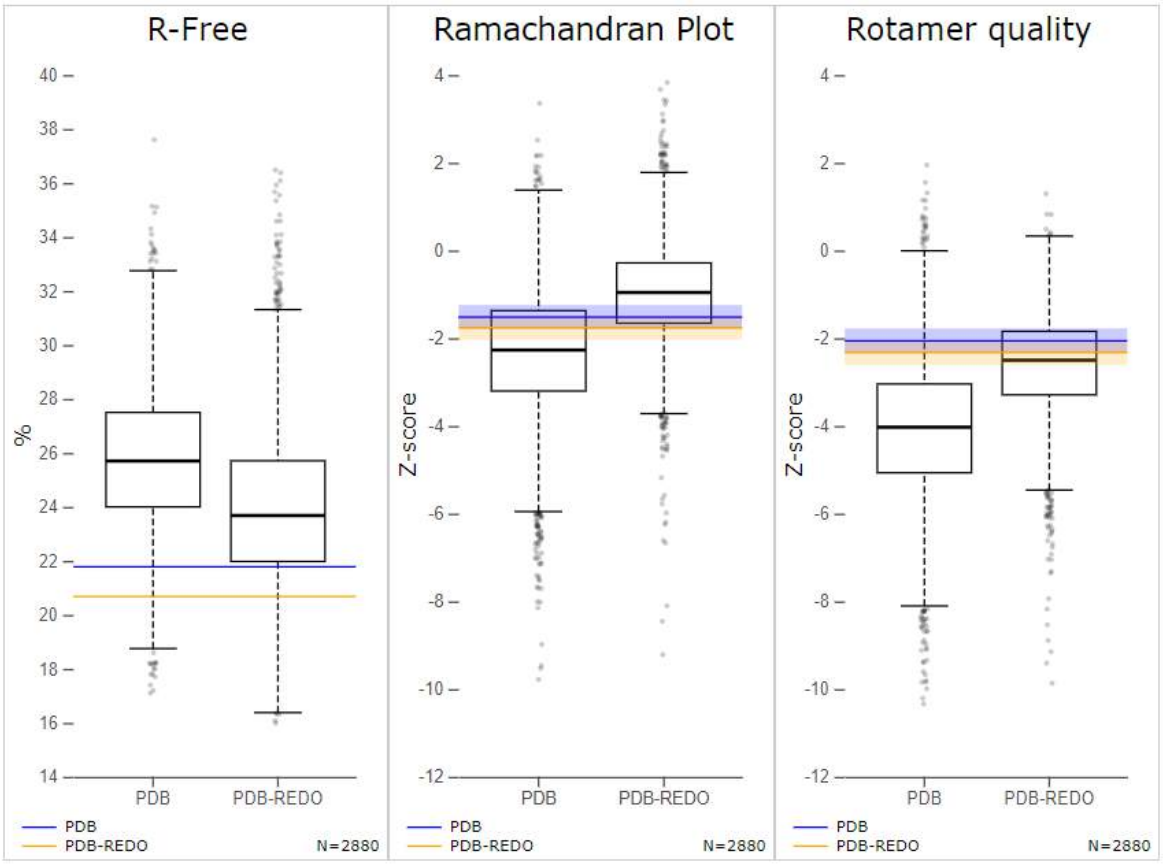
Table 11.A| *The PDB-REDO* ***validation metrics report*** *for* ***file-0 (Job 2)***

Validation metrics from PDB-REDO		
	PDB	**PDB-REDO**
Crystallographic refinement		
R	0,1848	0,1600
R-free	0,2178	0,2068
Bond length RMS Z-score	0,298	0,411
Bond angle RMS Z-score	0,531	0,659
Model quality (raw scores | percentiles)		
Ramachandran plot appearance	40	34
Rotamer normality	57	52
Coarse packing	49	59
Fine packing	34	50
Bump severity	58	49
Hydrogen bond satisfaction	25	32
WHAT_CHECK	Report	Report

Table 11.B| *The PDB-REDO* ***reported model changes*** *for* ***file-0 (Job 2)***

Significant model changes	
Description	**Count**
Rotamers changed	23
Side chains flipped	3
Waters removed	71
Peptides flipped	10
Chiralities fixed	0
Residues fitting density better	26
Residues fitting density worse	1

Table 11.C| *The PDB-REDO* ***Model quality compared to resolution neighbors Graphs*** *for* ***R-free,*** *the* ***Ramachandra plot angles*** *and the quality of the* ***Rotamers****. For* ***file-o*** *(****Job 2****)*



In *Job 2* a significant decreasing of both **R and R-free factors** is also present, and the rating of these factors is almost equal to those from *Job 1*. All statistical reports show important modifications to the model. However, PDB-REDO Server **was not able to build any of the missing regions** from **file-o** based on the sequence only, probably due to the fact there is not enough data for these regions, to build the loops.

The algorithm treated these regions as non-existent and continued the optimization for the rest of the model.

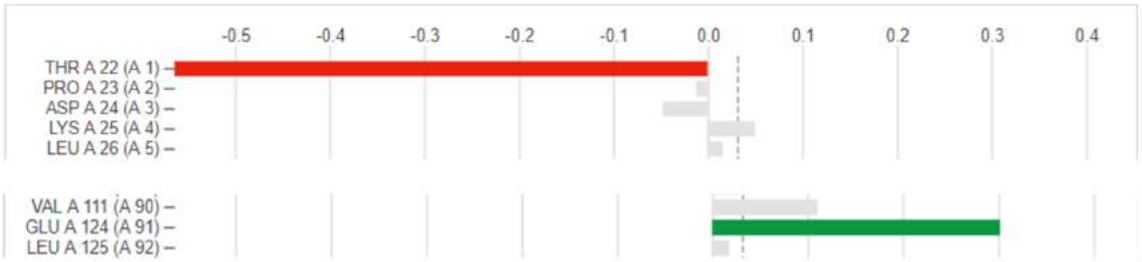


Figure 81| *The fitting graph of PDB-REDO, where the missing regions are not present.*

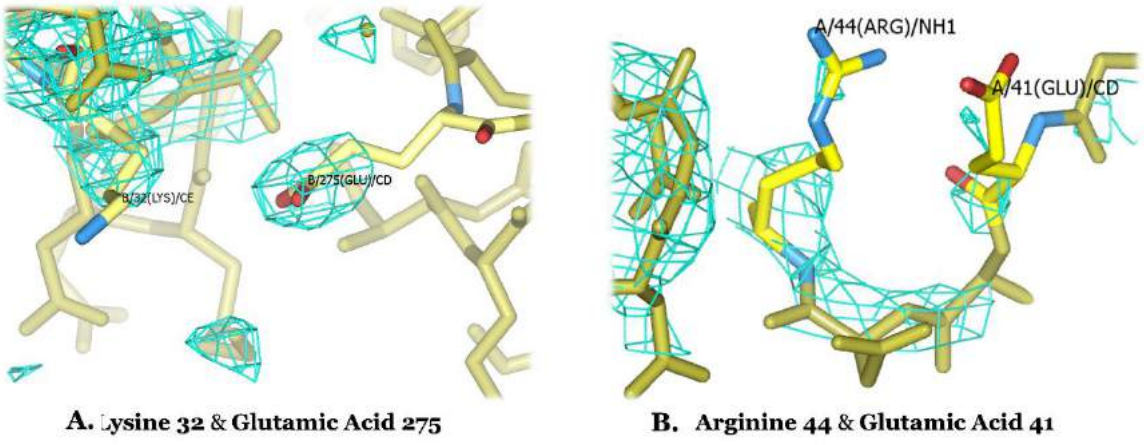
3. In Conclusion…

The Refinement factors are an important element for the quantification of the Model Building and Optimization effectiveness. However, it is clear that for a Protein Model, like the one for BsecI-Methyltransferase, it is not the determining factor for the validation of the model, something that became clear from the PDB-REDO results; in the control *Job* run for **file-o**, the **R factor was significantly decreased, but the model was still incomplete**. So, the decreasing of the Refinement factors does not necessarily mean that the model is optimized.

An empirical view and thorough computational studies are important for the **modelling of high-mobility regions**. With that being stated, a Validation process of a Protein model cannot be complete, if both the Empirical and the Automated Optimization processes are not combined for the *Fitting and Refinement* of the final structure.

Furthermore, the error I mentioned at the beginning of the Computational Methods as the "*software swapping",* describing the use of two entirely different computational environments such as *Windows* and *Linux*, didn't seem to affect the continuity of the study. Before that, I pointed out that an error like this needs to be avoided. But it seemed that a program like COOT performed better in *Windows* environment, and CCP4 Refmac5 in Linux. So, with a great reservation, I note that the process of Validation seemed like, not only wasn't obstructed by that error, but quite possibly it was aided to be completed in the best possible way, based on the conditions of the study.

Last and not least, the computational study of BsecI-Methyltransferase's three-dimensional structure reveals possible intramolecular interactions between the protein's residues, and molecular and structural properties of the enzyme's methylation mechanism. Some of the possible amino acid interactions I noted, are shown in **Figure 75.A-C**.



A. Lysine 32 & Glutamic Acid 275 **B.** Arginine 44 & Glutamic Acid 41

Figure 75| *Possible interactions between* ***A. Lys/32/B and Glu/275/B****, and* ***B. Arg/44/A and Glu/41/A.***

As mentioned in the Introduction section, the residue's interactions between them and their environmental conditions define the final conformation of the protein. Hydrogen bonded *bridges* between amino acid residues are usually detected in regions with some connection to the molecule's function. That is why, regarding this project, I report some examples of possible interactions, that I noticed. The Lysine residues are well-known for their *flexibility*. Even though multiple *Refinement* processes usually tend to change possible residues' conformations, for Lysine 32 in Chain "B" the picture was a bit different. The final conformation and the electron density indications may reveal a possible interaction with Glutamic Acid 275, also in Chain "B". The fact that these two residues are in between two main Secondary structures raises the question if they are part of a process of BsecIM's function. (**Figure 75.A**) For similar reasons, the same question is raised for the Chain "A" residues Arginine 44 and Glutamic Acid 41. These residues are part of a *turn* between two main Secondary structures also. It is possible that the interaction between them stabilizes the turn's backbone. (**Figure 75.B**)

The studying and modelling of high-mobility regions in the model is also important, because of the *high flexibility* they present. The loops and tails of the structure are detected inside the molecule's *pocket*. It is possible that they are connected to the *Flipping Base* mechanism of BsecI. **Future studies** on the enzyme's complex with the DNA double helix may reveal the importance of these structures.

Nevertheless, the final model of the BsecI DNA-Methyltransferase is the first step to the *unravelling* of the enzyme's properties and role in the biological systems. (**Figure 76**)

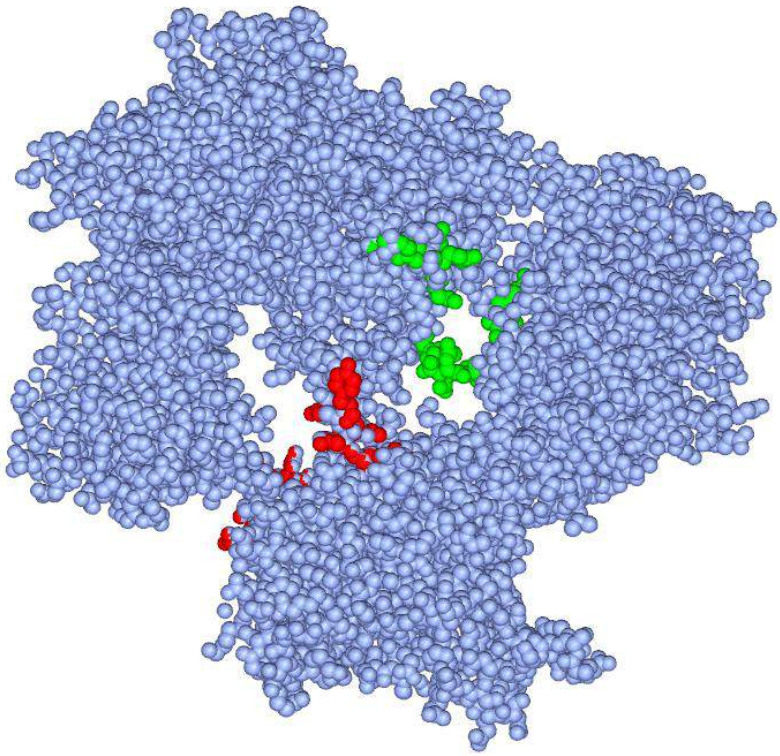


Figure 76| *The optimized model of the BsecI DNA-methyltransferase with the position of the loops and tails regions colored in* ***red*** *and* ***green****.*

References

Abts, A., Schwarz, C., Tschapek, B., Smits, S. J., Schmitt, L. (2012). Rational and Irrational Approaches to Convince a Protein to Crystallize. in *Modern Aspects of Bulk Crystal and Thin Film Preparation-Chapter 22*. January. doi: 10.5772/1348. isbn: 978-953-307-610-2.

Anfinsen, C. B. (1972). The formation and stabilization of protein structure. *The Biochemical Journal*. 128 (4): 737–49. doi:10.1042/bj1280737. July.

Anfinsen, C. B. (1973). Principles that govern the folding of protein chains. *Science*. 181 (4096): 223–30. July.

Alberts B., Johnson A., Lewis J., Raff M., Roberts K., Walters P. (2002). The Shape and Structure of Proteins. *Molecular Biology of the Cell; Fourth Edition*. New York and London: Garland Science.

Al-Karadaghi, S. *Torsion Angles and the Ramachandran Plot*. Retrieved May 28, 2020 by http://proteinstructures.com/Structure/Structure/Ramachandran-plot.html

Berg, J. M., John, T. L., Stryer, L. (2007). *Chapter 2: Protein Composition and Structure*. Biochemistry. New York: W. H. Freeman.

Brunger, A. T. (1992). Free R value: a novel statistical quantity for assessing the accuracy of crystal structures. *Nature*. 355 (6359): 472–475. January.

Clark, J. (2007). An introduction to amino acids. *Chemguide*. August. Retrieved 4 July 2017.

Creighton, T. H.(1993).*Chapter 1- Proteins: structures and molecular properties*. San Francisco: W. H. Freeman.

Crick, F. H. (1958). *On Protein Synthesis*. In F. K. Sanders (ed.). Symposia of the Society for Experimental Biology, Number XII: The Biological Replication of Macromolecules. Cambridge University Press.

Crick, F. H. (1970). Central dogma of molecular biology. *Nature*. 227 (5258): 561–3. August.

Diederichs, K., & Karplus, P. A. (1997). Improved R-factors for diffraction data analysis in macromolecular crystallography. *Nature Structural Biology,* 4, 269-75.

Donk, P.J. (1920). A highly resistant thermophilic organism. *Journal of Bacteriology*. 192, 5, 373- 374.

Emsley, P., Lohkamp, B., Scott, W., Cowtan, K. (2010). Features and Development of Coot. *Acta Crystallographica Section D-BIOLOGICAL CRYSTALLOGRAPHY*, 66, 486-501.

Glykos, N. M. (2015). *A Non-Mathematical Introduction to Protein Crystallography*. Greece, Alexandroupolis, Retrieved November 2,2015, from utopia.duth.gr/ glykos/pdf/Protein_crystallography.pdf .

Hahn, Th. (2002). Volume A: Space Group Symmetry. *International Tables for Crystallography*. 5th ed. Berlin. New York.

Huang, N., Banavali, N. K., MacKerell, A. D. (2002). Protein-facilitated base flipping in DNA by cytosine-5-methyltransferase. *Proceedings of the National Academy of Sciences*. 100 (1): 68–73. December.

IUPAC-IUB Joint Commission on Biochemical Nomenclature. (1984). "Nomenclature and Symbolism for Amino Acids and Peptides. Recommendations 1983". *European Journal of Biochemistry*. 138 (1): 9–37. doi:10.1111/j.1432-1033.1984.tb07877.x. ISSN 0014-295

Jakubke, H., Sewald, N.(2008). *"Amino acids"- Peptides from A to Z: A Concise Encyclopedia*. Germany: Wiley-VCH. p. 20. ISBN 9783527621170 – via Google Books.

Joosten, R. P., Joosten, K., Murshudovb, G. N., Perrakis, A. (2012). PDB_REDO: constructive validation, more than just looking for errors. *Acta Crystallographica Section D: Biological Crystallography*. p. 484–496.

Joosten, R. P., Long, F., Murshudov, G. N., Perrakis, A. (2014). The PDB_REDO server for macromolecular structure model optimization. IUCr Journal. 1:213-20.

Kabsch, W., Sander, C.(1983). Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features. *Biopolymers*. 22 (12): 2577–637. December.

Kovalevskiy, O., Nicholls, R. A., Long, F., Murshudov, G. N. (2018). Overview of refinement procedures within REFMAC5: utilizing data from different sources. *Acta Crystallographica*, D74, 492-505

Leavitt, S. A. (2010). *Deciphering the Genetic Code: Marshall Nirenberg*. June.Office of NIH History.

Lewis, C., Short, C. (1879). A Latin Dictionary. *Clarendon Press*. (Entry for *cis*)

Liebschner, D., Afonine, P. V., Baker, M. L., Bunkóczi, G., Chen, V. B., Croll, T. I., ... Adams, P. D. (2019). Macromolecular structure determination using X-rays, neutrons and electrons: Recent developments in Phenix. Acta Crystallographica Section D-BIOLOGICAL CRYSTALLOGRAPHY, 75, 861-877.

Marz, E. (2018). *The Ramachandran Principle- Phi (φ) and Psi (ψ) Angles in Proteins*. Retrieved May 20, 2020 by http://bioinformatics.org/molvis/phipsi/

McNicholas, S., Potterton, E., Wilson, K. S., Noble, M. M. (2011). Presenting your structures: the CCP4mg molecular-graphics software. Acta Crystallographica. D67, 386-394.

MIT OpenCourse Ware. (2008). *5.069 Crystal Structure Analysis*. Spring. Retrieved May 14, 2018 by http://ocw.mit.edu .

Murshudov, G. N., Skubak, P., Lebedev, A. A., Pannu, N. S., Steiner, R. A., Nicholls, R. A., Winn, M. D., Long F., Vagin, A. A. (2011). REFMAC5 for the refinement of macromolecular crystal structures. *Acta Crystallographica*, D67, 355-367

Murshudov, G. N., Vagin, A. A., Dodson, E. J. (1997). Refinement of Macromolecular Structures by the Maximum-Likelihood Method. *Acta Crystallographica*, D53, 240-255, EU Validation contract: BIO2CT-92-0524

Nelson, D. L., Cox, M. M. (2005), *Principles of Biochemistry* (4th ed.), New York: W. H. Freeman.

Nicholls, R. A., Tykac, M., Kovalevskiy, O., Murshudov, G. N. (2018). Current approaches for the fitting and refinement of atomic models into cryo-EM maps using CCP-EM. *Acta Crystallogrica*, D74, 215-227.

Okuyama, K., Haga, M., Noguchi, K., Tanaka, T.(2014). Preferred side-chain conformation of arginine residues in a triple-helical structure. *Biopolymers 101*: 1000-1009. doi: 10.1002/bip.22478.

Parsons, J., Holmes, J. B., Rojas, J. M., Tsai, J., Strauss, C. E. (2005). Practical conversion from torsion space to cartesian space for in silico protein synthesis. *Journal of Computational Chemistry*, 26 (10): 1063–1068, doi:10.1002/jcc.20237

Perrett, D. (August, 2007). From 'protein' to the beginnings of clinical proteomics. *Proteomics: Clinical Applications*. 1 (8): 720–38. doi:10.1002/prca.200700525. PMID 21136729.

Perrin, H. (2018). *A guide for protein structure prediction methods and software*. July. Retrieved May 30, 2020 by https://medium.com/@HeleneOMICtools/

Reitz, J., Milford, F., Christy, R. (1992). *Foundations of Electromagnetic Theory*. Addison Wesley. 4th edition.

Roberts, J. R. (1995). On Base Flipping. *Cell*. 82, p. 9-12. July.

Rina, M., Bouriotis, V. (1994). Cloning, Purification and characterization of the BseCI DNA methyltransferase from *Bacillus stearothermophilus*. *Gene*. Elsevier Science Publishers. 133, p. 91-94.

Rina, M., Markaki, M., Bouriotis, V.(1994). Sequence of the cloned bseC134 gene: MSBseCI reveals high homology to MaBanIII. *Gene*. Elsevier Science Publishers. 150, p. 71-73.

Rossi, R. J. (2018). *Mathematical Statistics : An Introduction to Likelihood Based Inference*. New York: John Wiley & Sons. p. 227.

Singh, J., Hanson, J., Heffernan, R., Paliwal, K., Yang, Y., Zhou, Y., (2018). Detecting Proline and Non-Proline Cis Isomers in Protein Structures from Sequences Using Deep Residual Ensemble Learning. *Journal of Chemical Information and Modeling*. August, 58, (9): 2033–2042. doi:10.1021/acs.jcim.8b00442

Smyth, M. J., Martin, J. H. J. (2000). X ray Crystallography. *Mol Pathol*. Clin Pathol. 53, p. 8-14.

Terwilliger, T.C., Grosse-Kunstleve, R.W., Afonine, P.V., Moriarty, N.W., Adams, P.D., Read, R.J., Zwart, P.H., Hung, L.-W. (2008). Iterative-build OMIT maps: map improvement by iterative model building and refinement without model bias. Acta Crystallographica. D64, 515-524.

The UniProt Consortium. (2018). UniProt: The Universal Protein Knowledgebase. *Nucleic Acids Res.,* 46: 2699.

Thornton, J. M., Sibanda, B. L., Edwards, M. S., Barlow, D. J. (1988). Analysis, Design and Modification of Loop Regions in Protein. *BioEssays*. 8, (2), p. 63-9. February/March.

Tronrud, D. (2015). *The Wonderful World of Maps*. Retrieve 28 July, 2017 by www.daletronrud.com/crystallography

Vagin, A. A., Steiner, R. A., Lebedev, A. A., Potterton, L., McNicholas, S., Long, F., & Murshudov, G. N. (2004). REFMAC5 Dictionary: Organization of prior chemical knowledge and guidelines for its use. *Acta Crystallographica Section D-BIOLOGICAL CRYSTALLOGRAPHY*, 60, 2184-2195.

Wagner, I., Musso, H., (1983). New Naturally Occurring Amino Acids. *Angewandte Chemie*, International Edition in English. November.

Watkin, D. (2008). Structure refinement: some background theory and practical strategies. *Journal of Applied Crystallography*. ISSN 0021-8898. March.

Winn, M. D., et al. (2011). Overview of the CCP4 suite and current developments. *Acta Crystallographica Section D- BIOLOGICAL CRYSTALLOGRAPHY,* 67, 235-242.

Winn, M., Isupov, M., & Murshudov, G. N. (2000). Use of TLS parameters to model anisotropic displacements in macromolecular refinement. *Acta Crystallographica Section D- BIOLOGICAL CRYSTALLOGRAPHY*, 2001:57, 122-133.

Wlodawer, A., Minor, W., Dauter, Z., & Jaskolski, M. (2008). Protein crystallography for non-crystallographers, or how to get the best (but not more) from published macromolecular structures. *The FEBS journal*, *275*(1), 1–21. doi:10.1111/j.1742-4658.2007.06178.x