UNIVERSIDADE DE SÃO PAULO
Instituto de Física de São Carlos

MATÍAS NAHUEL ELLENA

Structural determination, by cryo-EM, of the hemocyanin from the
mollusk *Concholepas concholepas* with substantial biomedical impact

São Carlos
2024

MATÍAS NAHUEL ELLENA

Structural determination, by cryo-EM, of the hemocyanin from the mollusk *Concholepas concholepas* with substantial biomedical impact

Dissertation presented to the Graduate Program in Physics at the Instituto de Física de São Carlos, Universidade de São Paulo, to obtain the degree of Master of Science.

Concentration area: Biomolecular Physics

Advisor: Prof. Dr. Andre Luis Berteli Ambrosio

Corrected Version

(original version available on the Program Unit)

São Carlos
2024

I AUTHORIZE THE REPRODUCTION AND DISSEMINATION OF TOTAL OR PARTIAL COPIES OF THIS DOCUMENT BY CONVENTIONAL OR ELECTRONIC MEDIA FOR STUDY OR RESEARCH PURPOSES SINCE IT IS REFERENCED.

Para meus pais, minhas irmãs, minhas filhas e meu amor.
Para todos aqueles que estiveram ao meu lado,
que em algum momento me acompanharam na vida,
que me apoiaram e me ensinaram a amar,
e que tanto estimo.
E principalmente para aqueles que não estão mais comigo,
mas que sempre estarão ao meu lado.

ACKNOWLEDGMENTS

To my Godmother, Natalia Alvarez Failache, who showed me the magic of science.

To my dear friends from Pelotão, who have been with me since I was a child.

And to my partner Gabriel, who taught me that everything has its time and that it's not worth stressing over things we cannot control.

DISCLAIMER

This document was written in the English language by non-native speakers. Hence, it was revised and corrected for grammar and improved for clarity using automated tools such as Grammarly Premium (Grammarly, Inc.) and the GPT-4 model from ChatGPT (OpenAI), with versions available in October 2023.

"It's a dangerous business, Frodo, going out of your door. You step into the Road, and if you don't keep your feet, there is no telling where you might be swept off to." *Bilbo Baggins*

J.R.R. Tolkien (1892-1973)

ABSTRACT

The single-particle cryo-electron microscopy (cryo-EM) advances since the "*Resolution Revolution*" has allowed the study of a wide range of proteins that were impossible to analyze with such precision with previous techniques and resolution. The recent acquisition of a Titan Krios microscope in Brazil is a remarkable achievement that positions the country as a leader in cryo-EM in Latin America. However, training researchers across the region is essential to optimize and maximize this technology. Collaborative efforts between Chilean and Brazilian research groups can promote widespread access to this methodology, fostering regional scientific development using state-of-the-art techniques. Here, we present a synergistic approach to determine the cryo-EM structure of *Concholepas concholepas* hemocyanin (CCH), an oxygen-transporting mega-dalton oligomer found in many invertebrates. Hemocyanins are employed as natural immunostimulants with significant biomedical and clinical applications. Understanding their precise molecular mechanisms is essential for targeted improvement. Structural diversity among homologous hemocyanins, particularly regarding glycosylation patterns, might account for their varying efficiency in generating suitable adjuvant effects in immunotherapy. In preclinical studies, Chilean researchers have shown that CCH holds promise for biomedical applications due to its superior stability, solubility, and immunological response induction. However, its structure and amino acid sequence remain unexplored. This collaboration enabled the elucidation of the structure of the CCH, a heterodidecamer with 8 MDa, and D5 symmetry, presenting two distinct monomers called CCHA and CCHB distributed equally along forming a protein with 366 Å height and 316 Å diameter, for the external collar, and almost 60 glycosylation sites distributed internally and externally, that have application for biomedical properties.

Keywords: Cryo-EM. Hemocyanin. Mega-Dalton protein.

RESUMO

Os avanços da criomicroscopia eletrônica de partícula única (cryo-EM) desde a "*Revolução da Resolução*" permitiram o estudo de uma ampla gama de proteínas que eram impossíveis de analisar com tanta precisão e resolução com técnicas anteriores. A recente aquisição de um microscópio Titan Krios no Brasil é uma conquista notável que posiciona o país como líder em Cryo-EM na América Latina. Mas para otimizar e maximizar o uso desta tecnologia é de grande importância formar pesquisadores em toda a região. Os esforços colaborativos entre grupos de pesquisa chilenos e brasileiros podem promover o acesso generalizado a esta metodologia, fomentando o desenvolvimento científico regional utilizando técnicas de ponta. Aqui apresentamos uma abordagem sinérgica para determinar por Cryo-EM a estrutura da hemocianina *Concholepas concholepas* (CCH), um oligômero transportador de oxigênio da ordem do mega-dalton encontrado em muitos invertebrados. As hemocianinas são empregadas como imunoestimulantes naturais com aplicações biomédicas e clínicas significativas. Compreender seus mecanismos moleculares precisos é essencial para uma melhoria direcionada. A diversidade estrutural entre as hemocianinas homólogas, particularmente no que diz respeito aos padrões de glicosilação, pode ser responsável pela sua eficiência variável na geração de efeitos adjuvantes adequados na imunoterapia. Pesquisadores chilenos demonstraram que o CCH é promissor para aplicações biomédicas devido à sua estabilidade superior, a sua solubilidade e a sua indução de resposta imunológica em estudos pré-clínicos. No entanto, a sua estrutura e sequência de aminoácidos permanecem inexploradas. Esta colaboração possibilitou a elucidação da estrutura do CCH, um heterodidecámero, com 8 MDa, e simetria D5, apresentando dois monômeros distintos denominados CCHA e CCHB distribuídos igualmente ao longo da proteína com uma altura de 366 Å e diâmetro de 316 Å, para o colar externo, e quase 60 sítios de glicosilação distribuídos interna e externamente, que têm aplicação e propriedades biomédicas.

Palavras-chave: Cryo-EM. Hemocianinas. Mega-Dalton proteína.

LIST OF FIGURES

LIST OF ABBREVIATIONS

Cryo-EM   Single Particle Cryo-Electron Microscopy

CCH   *Concholepas concholepas* Hemocyanin

RVH   *Rapana venosa* hemocyanin

HDH   *Haliotis diversicolor* hemocyanin

KLH   "Keyhole Limpet" hemocyanin (*Megathura crenulata*)

FU   Functional unit

CONTENTS

1 INTRODUCTION

1.1 Hemocyanins and Hemoglobins

The respiratory system is crucial for all living organisms, facilitating the transport of gassesgases like O_2 and CO_2. Various cells and proteins are specialized for this purpose, their types and functions differing among species. In vertebrates, gas transport is primarily facilitated by erythrocytes, or red blood cells, which contain the protein hemoglobin. Hemoglobin, a tetrameric protein with an iron ion at each active site, binds respiratory gases. This protein, weighing approximately 64 kDa, imparts the characteristic red color to the blood due to the iron ions that oxidize from Fe^{2+} to Fe^{3+} when binding to O_2.[1]

However, hemoglobin is not the sole protein responsible for gas transport in animals. Hemocyanins play a similar role in mollusks, crustaceans, and arthropods. Unlike hemoglobin, hemocyanins comprise around ten subunits, each with two copper ions at their active sites. This composition, weighing between 400 and 9000 kDa, gives the body fluids of these animals a blue or green hue. A notable distinction between hemocyanin and hemoglobin is their location within the body. Hemocyanins are found in the hemolymph and transport nutrients and gasses throughout the body, whereas hemoglobin is housed within erythrocytes.[2]

1.2 Hemocyanins: Structure and Characteristics

Hemocyanins, beyond their role as oxygen carriers in many animals, exhibit additional functions and characteristics. They act as glycoprotein storage and carriers and play a role in the innate immune response of arthropods and mollusks. Being dissolved in the hemolymph necessitates hemocyanins to possess negatively charged surfaces and a polar nature. This characteristic prevents their association with similarly charged surfaces, such as glycocalyx and epithelial cells, and enhances their solubility. The improved solubility is also attributed to their synthesis in specialized cells before being released into the hemolymph.[3]

Hemocyanins belong to the type-3 copper oxygen-binding protein superfamily, characterized by a di-copper active site center (Figure 1). Three histidine residues coordinate each metal atom, termed Cu-A and Cu-B. These atoms reversibly bind to peroxide ions, capturing O_2 molecules, a reaction shared among hemocyanins and tyrosinases. The reversible reaction changes the deoxidized copper ions from a cuprous state (Cu I ion) to a cupric state (Cu II ion) upon oxidation. This shift transforms the color from translucent in the deoxidized state to intense blue upon oxidation, lending the protein its name: hemocyanin (hemo: blood, cyanin: cyan). [2, 4]
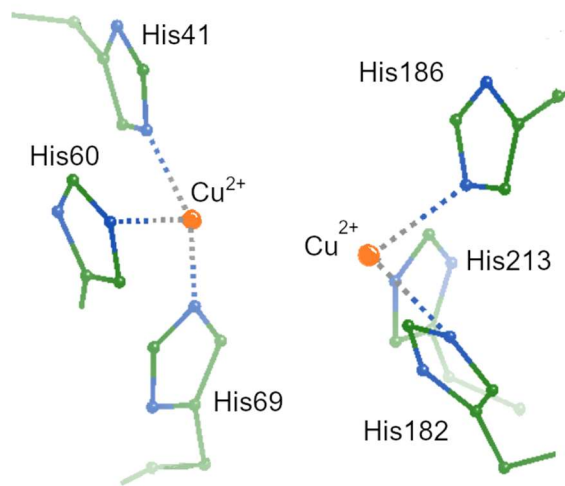
Figure 1 - **Representation of type-3 copper oxygen-binding active site on mollusks' hemocyanins.** The dinuclear copper site comprises six histidine residues, which can be separated into two groups of three linked to the copper íons. Those ions are then connected to the oxygen molecules. Every Functional Unit (FU) possesses one active site, around 7 or 8 active sites in total in each subunit, depending on the quantity of FU composing the subunit.

Source: By the author.

1.3 Molluscan hemocyanins

Hemocyanins are found in various animal groups, including mollusks, arthropods, and crustaceans, each exhibiting distinct characteristics. Molluscan hemocyanins are notable for their size, being among the largest glycoproteins studied, with molecular masses ranging from 3.3 MDa to 13.5 MDa. They are structured as cylindrical multimeric proteins comprising paralogous Functional Units (FUs), each containing approximately 420 amino acids and weighing between 45 and 50 kDa.[3]

Each FU comprises two or three globular folded domains connected by flexible peptide linkers of 10 to 20 amino acids (Figure 2). The FUs, typically ranging from FU-a to FU-h (Figure 3), are sequentially arranged on a polypeptide chain. Their domain consists of an α-helix bundle containing the di-copper center active site and a β-domain of anti-parallel β-sheets. These monomers assemble into a cylindrical decamer, which can further associate to form didecamers or tridecamers with distinct compositions, leading to significant structural variations between mollusca and arthropod hemocyanins.[3, 5]
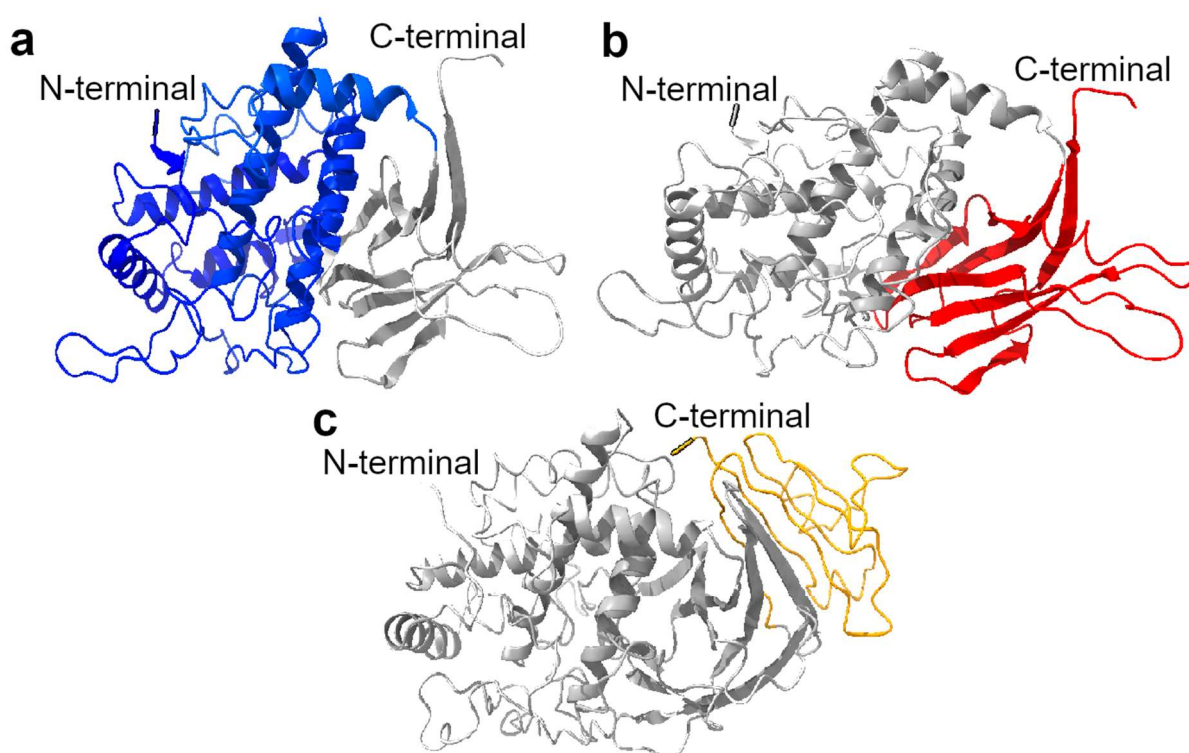
Figure 2 – Functional unit (FU) structure of molluscan hemocyanins. N-terminal and C-terminal are highlighted. It can be observed in alpha-helix (blue) and beta-sheet (red) domains, present in all functional units—FU-g with its extra domain (yellow) present on the C-terminal domain. The active site is present in the alpha-helix domain.

Source: By the author.
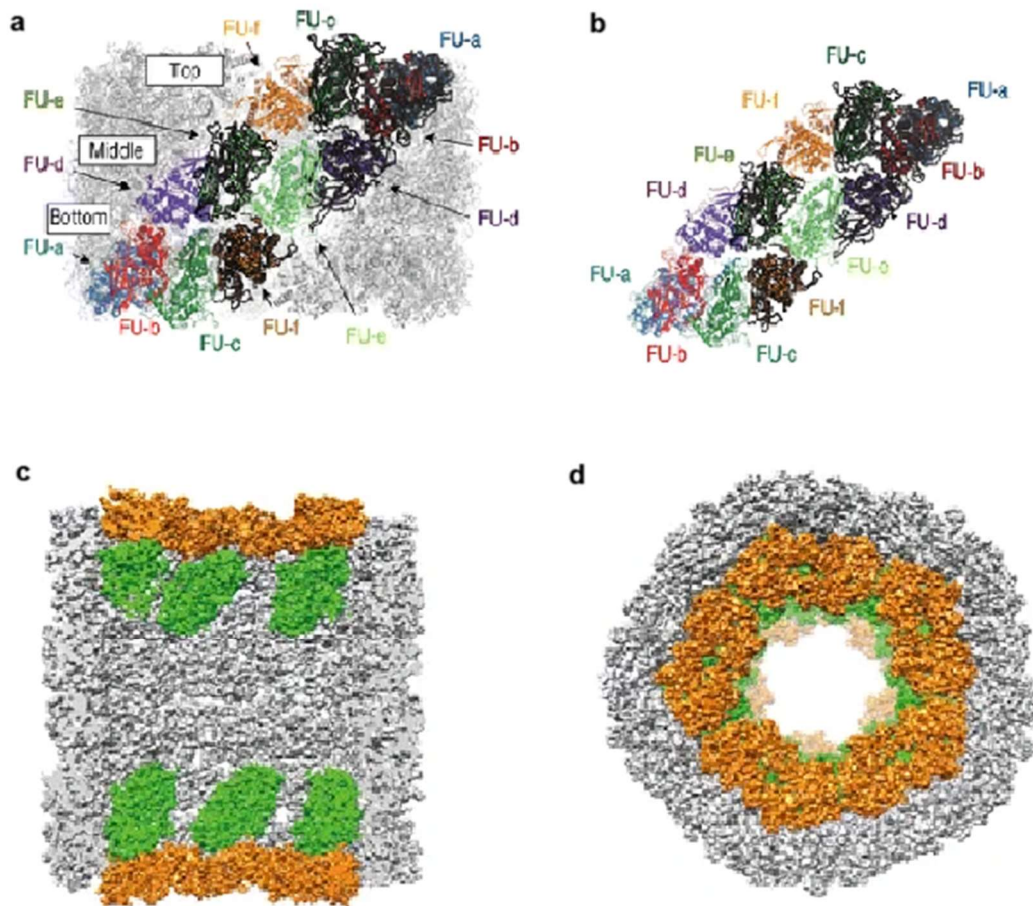
Figure 3 - Architecture of the plate-like dimer. (a) FUs in one plate-like dimer are shown in separate colors. The outline of one of the protomer is highlighted. (b) The architecture of plate-like dimer. Keyhole limpet-type hemocyanin. Conformation of keyhole limpet-type hemocyanin illustrated from lateral (c) and top views (d). Wall region, FU-g, and FU-h are shown as gray, green, and orange surfaces, respectively.

Source: Adapted from KATO *et al.* [6]

Despite the remarkable similarity between the marine hemocyanins and the composition of their subunits, significant structural differences can be responsible for different immunogenetic responses. Molluscan hemocyanins are classified into four types based on subunit quantity and composition:

Type 1: Found in keyhole limpets, it comprises eight subunits (FU-a to -h) and forms hetero or homo didecamers. The first six types form the external cylinder structure, while the last two form an internal collar, weighing approximately 8MDa (Figure 4). [7-8]

Type 2: Known as mega-hemocyanin, these tridecamers have two similar subunits as Type 1 and a third one with twelve subunits. The additional six subunits (FU-f1 to -f6) fill the cylinder's internal space, contributing to its significant molecular weight of approximately 13.5 MDa (Figure 4).[9]

Type 3: Found in nautilus, this type has seven subunits (FU-a to -g) and forms decamers. It resembles Type 1 but has a larger internal hole due to the collar being formed only by FU-g, weighing approximately 3.5 MDa (Figure 4).[8,10]

Type 4: Found in squids, this type is similar to Type 3 but features an additional FU-d* subunit. It forms decamers and weighs approximately 3.8 MDa (Figure 4).[10-11]

The Keyhole Limpet Hemocyanin (KLH) from the gastropod *Megathura crenulata* serves as the base model for Type 1 hemocyanin and is extensively studied for its medical applications, including diagnosing human schistosomiasis. Hemocyanins are gaining attention in medical research, particularly as potential anticancer agents. Their immunogenetic properties, influenced by carbohydrate content and glycosylation sites, are crucial in stimulating cellular and humoral immune responses. Interactions with macrophages, lymphocytes, and polymorphonuclear cells enhance the immune system's recognition and elimination of cancerous cells.[3-4]
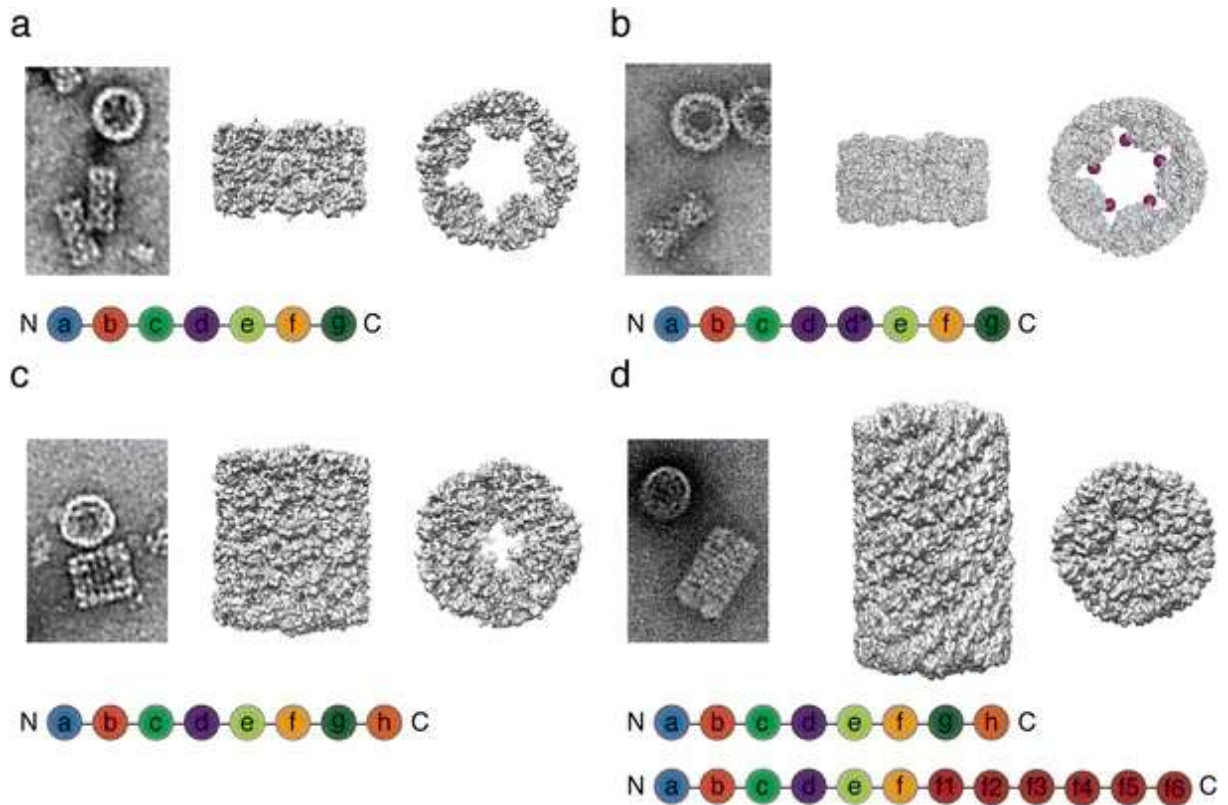
Figure 4 - **Structure of molluscan hemocyanin and its subunits.** a–d Representative area of negatively stained transmission electron microscopy (TEM) image (left), scheme of hemocyanin subunit (bottom), and three-dimensional (3D) model of four molluscan hemocyanin types (right) are shown. (a) Nautilus-type hemocyanin (TEM image: *Enterooctopus dofleini* hemocyanin; 3D structure: nautilus hemocyanin). (b) Squid-type hemocyanin (TEM image: *Todarodes pacificus* hemocyanin; 3D structure: crystal structure of *T. pacificus* hemocyanin. The Cu2O2 cluster of FU-d*, one of the inner domains, is indicated as purple spheres because the precise structure of FU-d* has not been determined yet). (c) Keyhole limpet-type hemocyanin (TEM image: *Nordotis discus hannai* hemocyanin; 3D structure: cryo-EM structure of keyhole limpet hemocyanin). (d) Mega-hemocyanin-type hemocyanin (*Melanoides tuberculata* hemocyanin). *C* C-terminus, *N* N-terminus. Scheme of hemocyanin subunits: *a, b, c, d, e, f, g, h* Functional units (FUs) of molluscan hemocyanins, *f1, f2, f3, f4, f5, f6* homologous FUs.

Source: Adapted from KATO *et al.* [6]

1.4 Concholepas concholepas Hemocyanin

Concholepas concholepas (Figure 5), a sea snail native to the west coasts of South America, particularly Chile and Peru, possesses hemocyanin (CCH) with a structure akin to KLH. CCH is synthesized in the hepatopancreas, a multifunctional organ involved in digestion, endocytosis, enzyme secretion, nutrient storage, and detoxification, among other processes.[8,12]

Figure 5 -   **Photo of the mollusk from the species *Concholepas concholepas*.** It's native to the Pacific coasts of Chile and Peru, and the source of the hemocyanins focused on this study because of its immunotherapeutic applications.

Source: Wikipedia.org

While CCH shares similarities with KLH, distinct characteristics influence the immune responses they induce. CCH consists of two subunits, CCHA (405 kDa) and CCHB (350 kDa), which intertwine to form a heterodecamer. The pairing of these subunits results in the formation of heterodidecamers. Notably, CCH exhibits enhanced stability at lower pH levels, though not at higher ones, and does not rely on Mg2+ or other divalent cations for stabilization, distinguishing it from KLH.[4,13]

CCH, a Type 1 hemocyanin, comprises eight subunits per monomer, connected by flexible peptide linkers. Each subunit's α-helix domain houses an active site with two copper ions. Despite CCHA and CCHB having similar amino acid counts and sizes, a significant difference in molecular mass exists, attributed to CCHA containing a higher proportion of "heavier" amino acids, impacting its properties.[14]

1.5 Single Particle Cryo-Electron Microscopy

Single particle Cryo-Electron Microscopy or Single Particle Analysis (SPA Cryo-EM), is a technique with the most significant future in structural biochemistry due to its fast development and acceptance in the scientific environment. The Cryo-EM technique is used in biomolecular studies to determine three-dimensional structures of proteins, nucleic acids and other biological molecules with at least some dozens of

kDa. It involves the creation of vitreous ice in thin layers by fast freezing in liquid ethane, preserving the native state of the sample without the need for chemical additions or structural modifications.

This technique, though not new, has undergone a renaissance thanks to technological advancements and computational innovations that have enhanced its precision and applicability. The first images of biological samples made from Cryo-EM were obtained in the early 1980s, but only 20 years later, in the early 2000s, with the improvement of computational methods and direct electron detectors that the technique started to gain knowledge on the scientific community and was more used to determine structures of large and complex biological molecules with a better resolution, gaining more recognition year after year until the well-known "*Resolution Revolution*". In 2017, the Nobel Prize in Chemistry was awarded to Jacques Dubochet, Joachim Frank, and Richard Henderson for high-resolution determination of biomolecules in solution with Cryo-EM, and from then on, the technique has grown exponentially, being possible to study proteins with some dozens of kDa and to achieve a resolution of 1.2 Å.[15]

As we delve deeper into the molecular world, the role of SPA Cryo-EM is becoming increasingly pivotal. It's not just about capturing images but about rendering detailed, accurate, and comprehensive 3D structures that offer insights into biological molecules' function, behavior, and interaction. Each structure, each image, is a piece of a giant puzzle that, when complete, promises to unveil the intricate interaction of molecules that define life. Because the study of structural biology is essential for understanding how macromolecular complexes work and their function on living organisms, and so the greater the details of a structure, the more can be learned from it, and that is the reason why it's so important to find, or at least understand, the limits of a technique, being it new or a developing one as cryo-EM.

While Single Particle Cryo-EM is widespread, it's worth noting the contributions of its counterparts, Cryo-Electron Tomography (Cryo-ET) and Micro-Electron Diffraction (Micro-ED). These branches, though distinct, complement Single Particle Cryo-EM, offering alternative approaches and perspectives in the study of biological molecules, both in a more native environment or as crystalline species.

But the scarcity of high-end electron microscopes, especially in the southern hemisphere, poses a significant challenge. With approximately 360 advanced Cryo-EM equipment globally, only a handful are accessible in regions like Latin America and Oceania. This disparity underscores a technological gap, highlighting the urgent need for equitable access to cutting-edge tools instrumental in advancing structural biochemistry.

In SPA Cryo-EM, the process involves rapidly freezing samples to capture them in their native state, eliminating the need for chemical stabilizers or additives. Electron beams then capture detailed images meticulously assembled into comprehensive 3D models using specialized software. This intricate process has been central to the "*Resolution Revolution*", a term coined to describe the unprecedented clarity and detail achieved in recent years. After the collection of multiple images, from hundreds of thousands to millions, a 3D model can be constructed with the help of specialized softwares like Cryosparc and Relion.[15, 16]

The resolution in Cryo-EM is described by the Fourier Correlation Shell (FCS), a concept created to evaluate the resolution of a 3D structure map created by Cryo-EM in a way that it could be compared to more standard and well-known techniques like crystallography. The FCS is a curve that shows the correlation between two half-maps of a reconstructed 3D structure as a function of spatial frequency. Once the data is obtained, it is divided into two independent groups, and then the Fourier transform of each half as well as the correlation between the corresponding frequency components and making an average of the correlation over the entire dataset, allowing a comparison of how much the halves agree with each other at different spatial frequencies. Once the FCS is ready, the resolution of the reconstructed 3D structure can be determined by the spatial frequency at which the correlation drops below a certain threshold. From this intersection, the resolution of the map can be calculated. Initially, the threshold used was 0.5, the limit where two points in the structure are barely similar, but this caused an overestimated resolution, especially if they have much noise or the structure has low contrast. Therefore, the 0.143 threshold, the most used and accepted nowadays, was proposed, a correlation equal to 1/7th of the initial value created, where a signal-to-noise ratio approaches 1, to allow a better comparison with X-ray crystallography, and a limit established to determine where the structure

cannot be reliably determined anymore, based on the fact that the signal-to-noise ratio decreases as the resolution increases.[17]

The resolution limit in X-ray crystallography is related to the maximum condition of constructive interference at which the crystal structure can be determined, and the diffraction limit of the crystal lattice measures it. Although both Cryo-EM and X-ray crystallography differ significantly, the resolution limit concept is fundamental in both techniques, in a way that the 0.143 threshold for Cryo-EM has a similar idea to the resolution limit in X-ray crystallography. Despite the initial differences between both methods and the discussion of them being mutually exclusive, it is essential to highlight that they are complementary due to their unique properties, allowing different approaches when using both to study the same structure. When combined, they exhibit a more complete understanding of a biological specimen. The main factors that can affect the resolution of Cryo-EM are the quality of the micrographs, which can be improved with better electron detectors, the signal-to-noise ratio, and some software-based procedures, like the alignment, reconstruction, and correction.

However, there are still some difficulties that surround this technique that need to be overcome to make Cryo-EM more accessible and regular, starting with the immense quantity of images of identical macromolecules that need to be collected to acquire a high-resolution map because higher the resolution, higher is the quantity of the images required to reach it.[17] With it, there is the computational processing and disk space needed since a collection of images can occupy more than a dozen terabytes in disk space, with days of continuous computational processing power to achieve a good resolution map.

And so, the future trajectory of Single Particle Cryo-EM is intricately linked to integrating artificial intelligence and machine learning. These technologies promise to expedite data processing and enhance the accuracy and reliability of the resulting 3D structures. The ongoing evolution of AI and machine learning algorithms is expected to mitigate some of the current challenges associated with the extensive computational requirements of Single Particle Cryo-EM.

In the specific context of the *Concholepas concholepas* hemocyanin (CCH), Single Particle Cryo-EM has unveiled intricate details that were previously elusive. This protein, synthesized in the hepatopancreas of the *Concholepas concholepas* sea snail,

exhibits unique characteristics, notably its stability at lower pH levels and its independence from divalent cations for stabilization. These attributes have not only distinguished CCH but have also underscored the versatility and precision of Single Particle Cryo-EM, especially since the amino acids sequence of CCH is still unknown. [8,12]

      In conclusion, SPA Cryo-EM stands at the intersection of technology, biology, and computational science. It's a testament to modern scientific exploration's collaborative and interdisciplinary nature. As technology evolves, so will our understanding of the complex and intricate world of biological molecules, heralding discoveries, innovations, and applications that will shape the future of science, medicine, and beyond. And studying proteins like CCH is not just academic but central to understanding biological processes, diseases, and potential treatments. Every molecule, every atom captured and rendered by Single Particle Cryo-EM, adds a layer of understanding, bringing us a step closer to unraveling the mysteries of life at the molecular level.

2 JUSTIFICATION

This dissertation is a product of a collaborative research project between the Crystallography Group at IFSC/USP and the Department of Biology of the Universidad de Chile, aiming to pioneer the structural study of a giant mollusk hemocyanin that holds considerable promise in the scope of immunotherapy. The research was funded by FAPESP (2019/13318-5) and the Comisión Nacional de Investigación Científica y Tecnológica (CONICYT) from Chile. Unraveling the structure of CCH will significantly enhance our comprehension of its potent immunomodulatory properties.

However, at its core, this proposal strives to cement the synergistic efforts of both institutions in bolstering the methods of single particle Cryo-EM in the Latin American structural biology community, a field still in the early stages of development in the region; the establishment of the Cryo-EM facility at LNNano represents a monumental stride in advancing local structural biology.

A multidisciplinary team of researchers underpins our endeavor, each bringing a unique yet complementary set of skills to the work. Dr. Maria Inez Becker's team boasts a rich history in the molecular biology of *C. concholepas*, having made notable strides in the biochemical characterization of CCH. This includes the development of protocols for its purification and characterization by negative staining electron microscopy. On the other hand, Dr. Victoria Guixe's and Dr. Victor Castro´s (Universidad de Chile) teams possess the requisite knowledge and experience to purify CCH in its entirety and its functional units. We are responsible for conducting the high-resolution structure determination in its entirety. Together, we work towards interpreting the biological interpretation revealed by the high-resolution structure of CCH.

3 OBJECTIVES

3.1 Major Objective

To determine the high-resolution structure of the intact hemocyanin from *Concholepas concholepas* by single particle Cryo-EM and locate potential glycosylation sites.

3.2 Specific objectives

1. **Sample preparation:** To identify the most suitable vitrification conditions for cryo-grid preparation of intact CCH by optimizing protein concentration and ice thickness.

2. **Data collection and processing:** To collect a complete high-resolution dataset in a high-end microscope, such as the Titan Krios (300 kV, ThermoFisher Scientific), and conduct the entire processing pipeline with cryoSPARC.

3. **Model building and interpretation:** To perform thorough structural analysis, detect potential glycosylation sites in the Cryo-EM maps, and compare CCH with other mollusk hemocyanins.

4 METHODS AND RESULTS

4.1 Purification and electrophoresis of CCH

The CCH samples, commercially available as inmunocyanin®, were supplied by Dr. Maria Inez Becker's group at the Laboratory of Biochemistry and Molecular Biology, University of Chile, responsible for its production. The samples were aliquoted at a concentration of 26 mg.ml^{-1} in Phosphate Buffered Saline (PBS) and stored at 4°C, with a shelf life of up to one year. Upon receipt via international courier, the purity and integrity of the samples were immediately verified through electrophoretic migration analysis under denaturing conditions (SDS PAGE) and dynamic light scattering (DLS), affirming the high sample quality (Figure 6).

Hydrodynamic Radius

Z-Average (± SD) (r.nm): 18,87 ± 1,895
Polydispersity Index: 0,010

Estimated MW (± SD) (KDa): 3250 ± 326,7
%Polydispersity: 10,0
Sample Polydispersity: Monodisperse

Distribution Results

	Mean ± SD (nm)	%Pd	Est. MW (KDa) (Mean ± SD)*	% Intensity	% Mass	Peak Polydispersity
Peak 1:	19,15 ± 2,246	11,7	3,37e+3 ± 394,7	100,0	100,0	Monodisperse
Peak 2:	0,000 ± 0,000	0	0,0 ± 0,0	0,0	0,0	
Peak 3:	0,000 ± 0,000	0	0,0 ± 0,0	0,0	0,0	
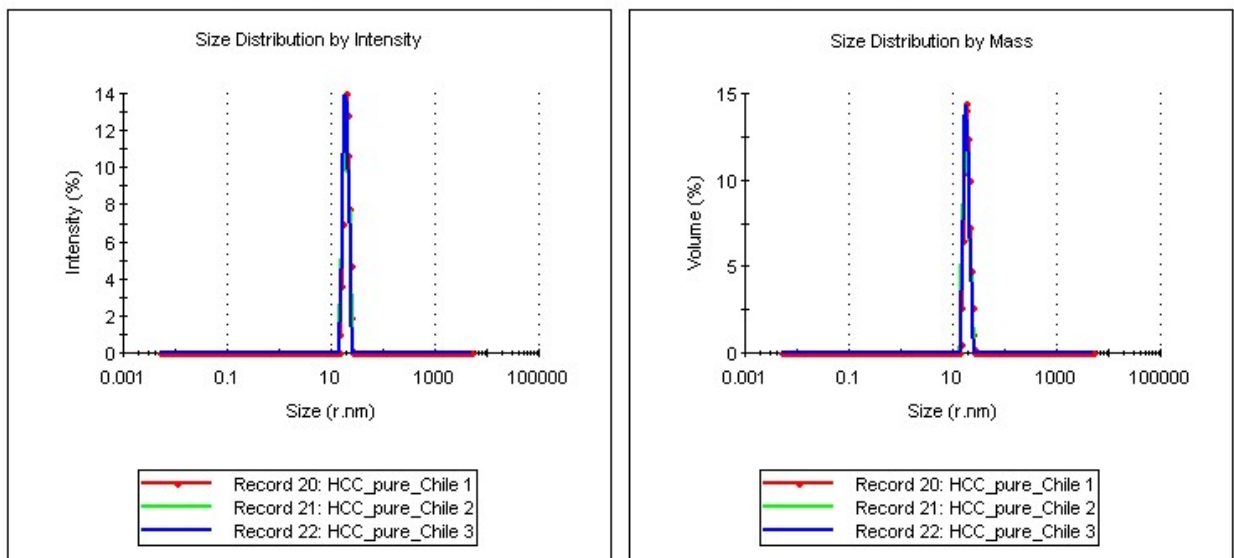
*The molecular weight reported here is only an estimate, calculated using an empirical mass vs. size calibration curve.

Figure 6 - Dynamic Light Scattering analysis of CCH, in 1x PBS, at 1 mg.ml^{-1}. Three independent measurements confirmed populational homogeneity (monomer), suggesting a hydrodynamic radius of 19 ± 2 nm and a molecular mass of 3.3 ± 0.3 MDa.

Source: By the author.

4.2 Preliminary analysis by negative staining electron microscopy

In preparation for cryo-EM studies, the visual confirmation of sample size and shape, purity, and populational homogeneity was made by transmission electron microscopy analysis of negatively stained CCH, ensuring the best quality of the samples before sending them to cryo-EM analysis. Once sample homogeneity, quality, and quantity were reliably established, the cryo sample screening began.

The negative staining experiments were conducted at LNNano/CNPEM, Campinas, and the Instrumental Chemical Analysis Center (CAQI) of the Institute of Chemistry of São Carlos (IQSC). First, the negative staining analysis at LNNano was entirely performed by the facility staff upon delivery of the CCH sample, without direct participation, because of the access restrictions applied during the pandemic of COVID-19 for external users. Access to the microscope was granted via a fast-track submission under number TEM-27734F.

At LNNano, CCH samples were diluted in 1x PBS to 0.2, 0.4, and 0.8 mg.ml^{-1} and used for grid preparation. Briefly, 3 µl of the diluted CCH sample was applied to a glow-discharged electron microscopy grid, and 3 µl of a mixture of 1:1 of 1x PBS and 5% Uranyl Acetate (in di-distilled water). After 1 minute, required for protein adsorption, the excess liquid was removed with a piece of filter paper from the side of the drop, followed by a second drop of the staining mixture and further resting and blotting. The grid was allowed to sit until thoroughly dried and stored in a grid holder.

Micrographs were collected on a JEM-1400Plus microscope (Jeol, Japan), operating at 120 kV, and equipped with a OneView CMOS 4k x 4k camera (GATAN, USA). Magnifications of 20, 50, and 100 thousand times were used, and exposure times ranged between 0.5 and 1 second, with an electron dose of 20-30 e^{-}/Å^2. Representative micrographs, which confirm the overall high quality of the samples, are shown in Figure 7.

The same experiment was performed in the local laboratory at IFSC/USP and analyzed at CAQI/IQSC/USP. The grid preparation followed the same procedure LNNano did, adapting the buffer to 5% Uranyl Formate. At CAQI, the Negative staining analysis was conducted on a JEM-2100 (Jeol, Japan) microscope, operating at 120 kV

and equipped with a Gatan CCD camera. Magnifications of 50, 150, and 400 thousand times were used, with an exposure time of 1 second and an electron dose of 20 e$^-$/Å^2. Representative micrographs can be found in Figure 7.
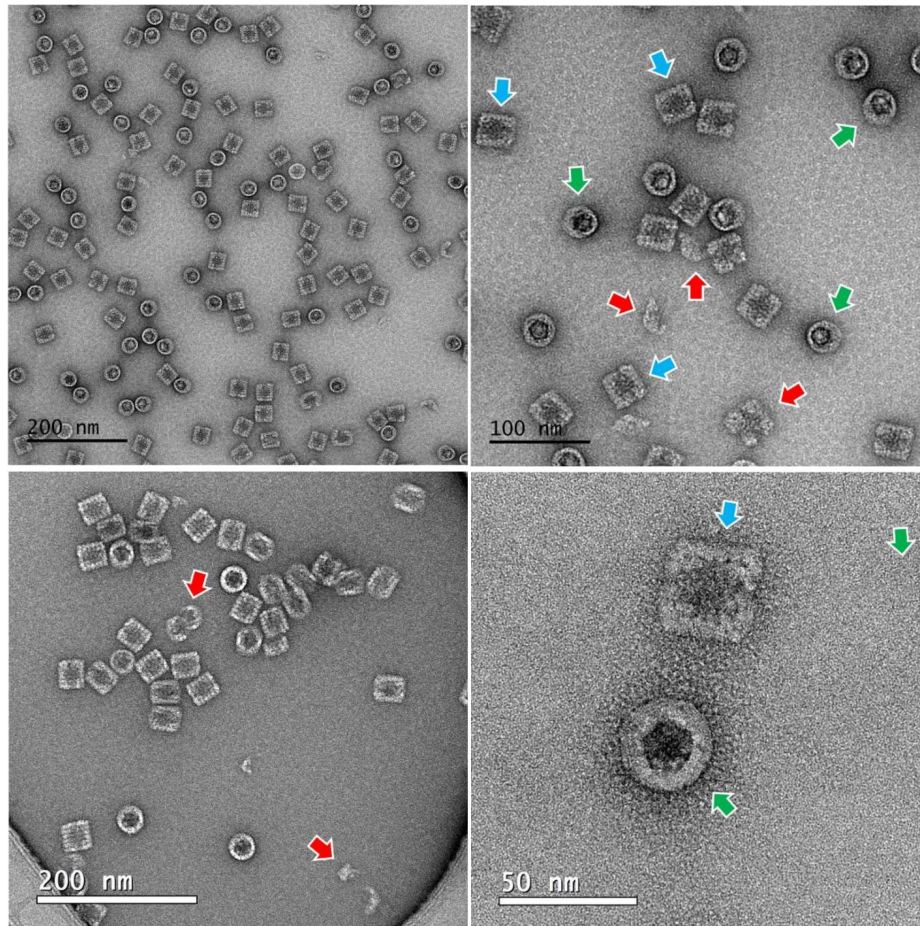
Figure 7 - Negative staining analysis of CCH and comparison between the results obtained at LNNano and CAQI. The quid preparation and parameters used were similar at both experiments. Representative micrographs confirm purity, size, and conformational homogeneity. A cylindrical shape for CCH is observed. Side (blue arrows) and top (green arrows) views are the dominating orientations. Eventually, partially degraded particles can be spotted (red arrows). Both images at the top are from LNNano, and those at the bottom are from CAQI. The concentrations used were around 4 to 5 mg.ml^{-1}

Source: By the author.

Figure 7 confirms the homogeneity and distribution of the particles over the grid, that the majority are intact with side and top views dominating due to preferential adsorption onto the grid material, with some partially degraded particles. Besides confirming high purity, the negative staining analysis also provided an estimate of the

initial concentrations to be tested for Cryo-EM. The concentrations prepared for the grids respective to Figure 7 were around 4 to 5 mg.ml^{-1}.

4.3 Vitrification and data collection

Upon confirmation that the CCH samples were ready for cryo-EM studies, we required access to the Titan Krios microscope at LNNano for data collection. However, due to the access restrictions imposed by the SARS-CoV2 during the pandemic, the grid preparation and the experiments were held on the LNNano facility by the local staff.

Protein concentration and ice thickness were defined according to the following parameters: the samples used were prepared with 2.5 µL of CCH at a concentration of 5 mg.ml^{-1} in 1x PBS deposited onto a Quantifoil R2/2 200 mesh copper grid and inserted in a Vitrobot Mark IV (Thermo Fisher Scientific) at 4 $^{\circ}$C and 100% humidity for plunge-freezing, with a blot time of 4.5 s with a relative blot force of -4. After blotting, a quick drain time of 0.5 s was applied, and the grid was rapidly plunged into liquid ethane for vitrification. Frozen grids were stored in liquid nitrogen.

The total of movies counting all the datasets was 11,010, acquired in the Titan Krios, operating at 300 kV with a pixel size of 1.1 Å squared, equipped with a Gatan K3 direct detection camera, and a varying defocus of -0.3, -0.6 and -0.9 µm. The total dose in each movie was 30 e^{-}.Å^{-1} divided into 20 fractions.

4.4 Image processing and structure determination

After acquiring the movies, they underwent processing steps to produce the final 3D potential density map. Various software tools can perform the necessary corrections and constructions to generate this map. Some tools, like Zorro, MotionCor2, Unblur, and SerialEM, specialize in realign and correcting the bean-induced motion that results in an enhancement of the initial movie's resolution. Once the images are corrected for single-particle analysis, they are processed using

reconstruction software packages such as RELION, FREALIGN, CryoSPARC, IMAGIC, or EMAN2.

CryoSPARC (Structura Biotechnology) v3.2.0 was the primary software for processing our dataset. This freely available software for Linux offers optimized algorithms and GPU acceleration throughout all stages, from pre-processing to particle picking, 2D particle classification, 3D *ab-initio* structure determination, high-resolution refinement, and heterogeneity analysis.

There is a standard workflow for achieving an optimal density map (Figure 8). While the specific steps can vary based on the software used, some general procedures are essential regardless of the chosen software. Initially, micrographs require refinement and corrections; this involves camera normalization and anisotropy magnification correction. Subsequently, motion correction aligns the movie particles to eliminate any movement from the particle or equipment. Concurrently, the CTF estimation describes the modulation of the electron microscope's imaging system, which can affect the movie recordings and the particles' surrounding environment, thereby reducing the final resolution.
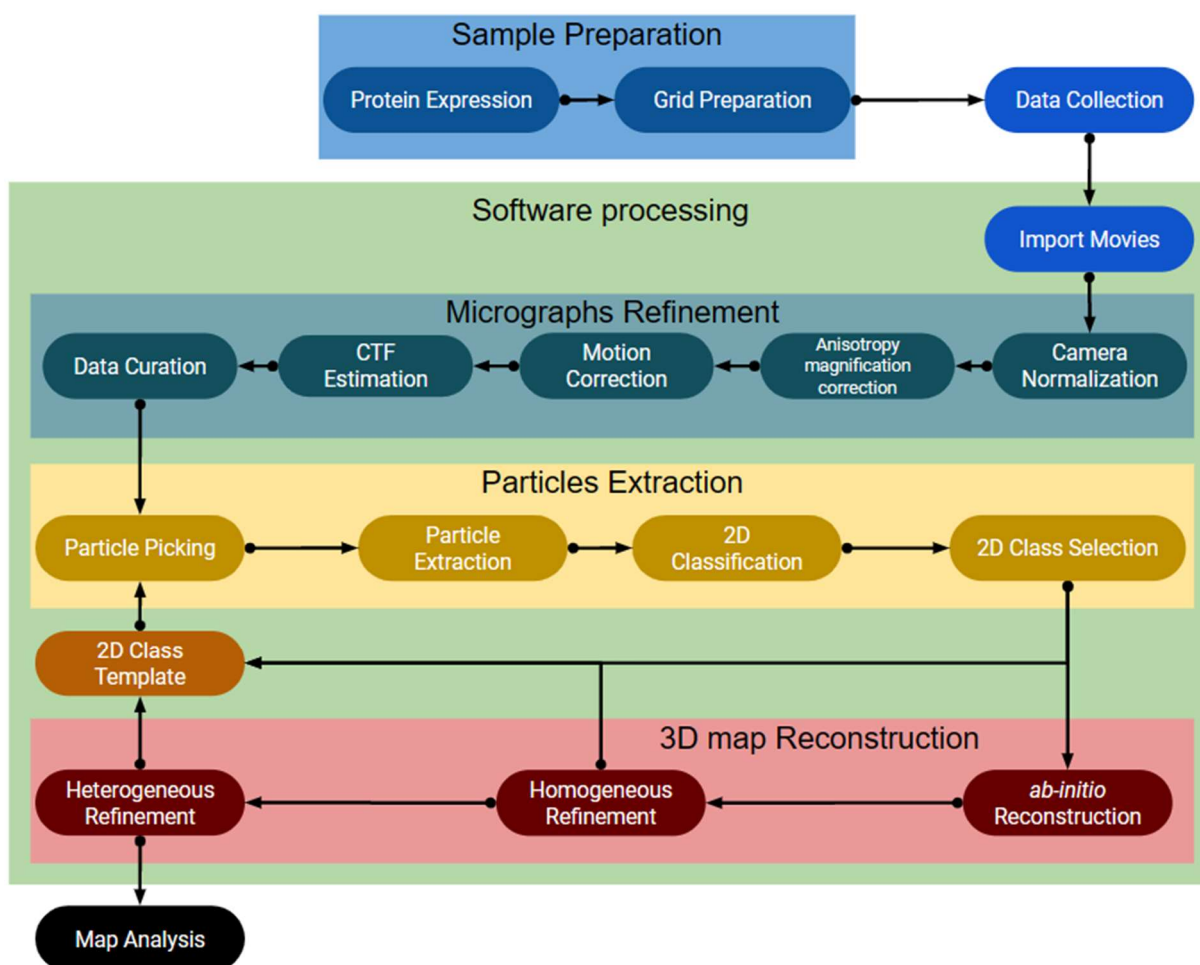
Figure 8 - Flowchart summarizing the general workflow of our Cryo-EM study.
Source: By the author.

After optimizing the image quality, an initial curation removes images that either lack sufficient resolution or do not meet the required parameters. The next step involves extracting particles from these curated images. Initially a picking is made, both manually or automatically, but then an extraction is done using specialized automatized softwares. Automation is crucial because a movie can contain hundreds of particles, and a collection of movies might comprise millions.

With the majority of particles extracted, the density map was constructed. Leveraging prior knowledge, we selected the map that best represented the studied protein. With this preliminary map, a series of refinements began, aiming to enhance its quality to its utmost potential; thus, the refinement and correction of the micrographs initiated with camera (or image) normalization. In a typical dataset, micrographs are often presumed to maintain a uniform signal-to-noise ratio. However, this assumption

can be misleading in the realm of Cryo-EM due to inherent variations in electron beam intensity and sample conditions. Camera normalization addresses this by standardizing the entire dataset to a standard grayscale. This is achieved by applying specific additive and multiplicative factors to each image, typically by subtracting the mean and dividing by the standard deviation, ensuring all images converge to a mean of zero and a standard deviation of one. Each sensor pixel has unique characteristics determined by its average density and standard deviation. These pixels exhibit varying sensitivities or "*gains*." Camera normalization is performed to address these disparities and ensure uniformity across the sensor. During this normalization process, the Fourier Ring Correlation (FRC) plays a pivotal role in evaluating the efficacy of the *a posteriori* correction. A successful correction is indicated by an FRC curve that oscillates around zero without ever reaching the 3σ threshold, either positively or negatively. It's worth noting that spurious correlations can emerge without adequate a priori flat-field correction, potentially skewing the data. The *a posteriori* normalization is designed to rectify these unintended correlations and any systematic errors stemming from the camera's readout electronics. The final result is a corrected image where undesired background-pattern correlations are effectively suppressed, ensuring the integrity and accuracy of the captured data.[18]

After camera normalization, anisotropic magnification correction was employed. This step is pivotal in rectifying micrograph distortions introduced by the microscope's electron optics.

Addressing anisotropic magnification takes precedence over Contrast Transfer Function (CTF) determination. The distortion's presence can complicate the defocus and astigmatism assessment, making early correction crucial. Conveniently, this correction is optimally performed on the movie data, negating the need for separate calibration measurements. Large Cryo-EM datasets inherently contain the water ring information, providing all the data required to diagnose and amend the anisotropy. This is done by summing the amplitude spectra of all fractions from all movies in the dataset, using the crystalline ice ring at around, and quantifying its distortion from a circle to an ellipse. The following distortion parameters were found: major and minor scaling values along two orthogonal axes of 1.001 and 0.999, distortion angle 63º, totaling about 1.8% distortion.[19]

The electron magnification is often assumed to be isotropic, but inherent electron optical aberrations can skew this assumption. These aberrations, if uncorrected, can introduce systematic astigmatism in the parameters estimated for the Contrast Transfer Function (CTF). To counteract these distortions, sophisticated Fourier space transformations are employed, ensuring the fidelity of the micrographs for subsequent refinement steps. Both these pivotal corrections were executed using the IMagic software. Anisotropic magnification, characterized by a 1.8% distortion, poses significant challenges in imaging a 350 Å particle diameter, such as CCH, via a transmission electron microscope. This distortion results in an approximate 3.15 Å displacement at the particle's edges from the average position, a deviation that is directional and thus varies in different orientations. In the realm of single-particle Cryo-EM, where averaging multiple images of proteins captured at random orientations is commonplace, such distortion undermines the coherence and resolution of the subsequent 3D reconstructions. For the subsequent steps, CryoSPARC was the software of choice.[19]

Individual frames from each movie were meticulously aligned and summed using the Patch Motion Correction protocol (Figure 9). This step is crucial to counteract stage drift and anisotropic deformation, common artifacts arising from beam-induced sample motion and thermal drifts. Segmenting the movies into distinct patches facilitated alignment and set the stage for subsequent processing, ensuring the integrity of the samples and identifying any movies that exhibited aberrations or required intensive repair.
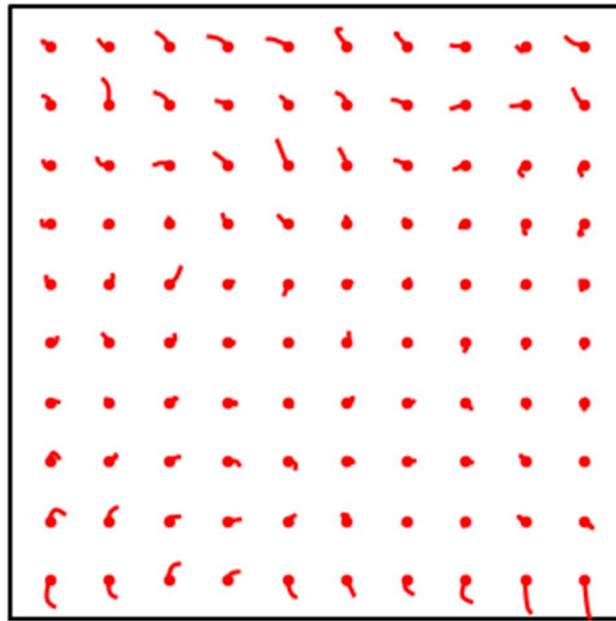
Figure 9 - Cryo-EM CCH motion correction graph. The motion graphic illustrates the path corrected from the patches of the respective micrograph since each micrograph has its own motion correction graph. It also shows the path followed for each patch on the sequential movies, red lines, as a result of the stage drift and the anisotropic sample deformation.

Source: By the author.

The next step of the data processing was the CTF estimation, an essential step to improve the resolution of the map in which the Fourier transform is used to measure the defocus degree of the microscope, allowing better differentiation of the particle from the noise. It is possible to filter the movies that present a CTF adjusted to a lower resolution, which could prejudice the 3D reconstruction resolution. In the CTF estimation of a movie, it is possible to observe the power spectrum produced by the electron stages at the moment of the collection of the micrograph (black), the calculated CTF for those stages (red), and the correlation function (cyan), in which a limit value of 0.3 is used in the example as parameter between both specters. The Patch CTF protocol performed whole micrograph CTF (Contrast Transfer Function) estimation and correction. Representative micrographs, aligned, summed, and CTF corrected, are shown in Figure 10. The CTF resolution in this example corresponds to 2.95 Å (Figure 10).
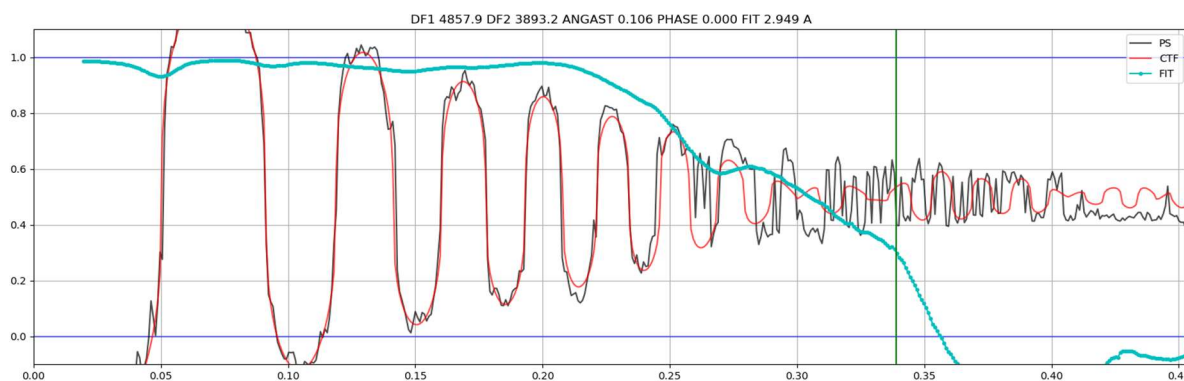
Figure 10 - CTF estimation graph. The plot presents the match between the computed CTF, red line, and the observed power spectrum of the micrograph, black line. The cross-correlation is shown in cyan, while the green line represents the frequency at which the fit drops below a respective threshold, 0.3 in this case. With this, it is possible to increase the resolution of the micrographs and estimate the possible resolution of volume to be reconstructed from them.

Source: By the author.

After these corrections, a rigorous assessment of the movie's quality was undertaken. This involved setting stringent criteria based on defocus, astigmatism, and CTF estimation resolution. Specifically, micrographs with defocus larger than -1 µm, astigmatism beyond 16,000 Å and CTF resolutions lower than 2 Å and larger than 5 Å were excluded. Of the initial 11,010 movies, a curated set of 8064 movies met these stringent quality benchmarks. A representative corrected image is displayed in Figure 11.
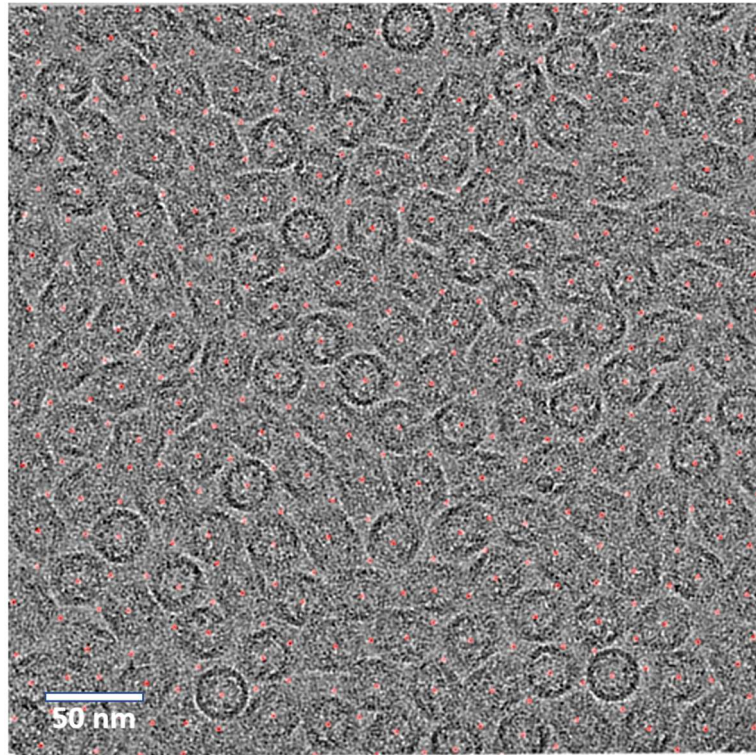
Figure 11 – Representative cryo-EM micrograph of CCH obtained with Titan Krios after camera, magnification, motion, and CTF corrections. Compared to the negative staining studies, vitreous ice allows for many different projected orientations of the CCH particle while maintaining an intact structure; these features are critical for volume reconstruction. The red dots indicate the possible particles that were extracted afterwards.

Source: By the author.

From the resulting movies, about 500 particles were selected manually from the initial 100 micrographs and then 2D classified based on their projection similarity, with a box size of 600 pixels (660 Å). Those classes were then used as a base template for the desired shape for the automatic pickers, resulting in 1,060,718 candidates. Once the selected particles were classified again, the best classes were established. This process was repeated several times until many particles were present in the desired 2D classes. This way, after some rounds of selection and 2D classification, 32 classes were selected, accounting for 626,735 particles, representing distinct projections of CCH. The 2D class averages are presented in Figure 12.
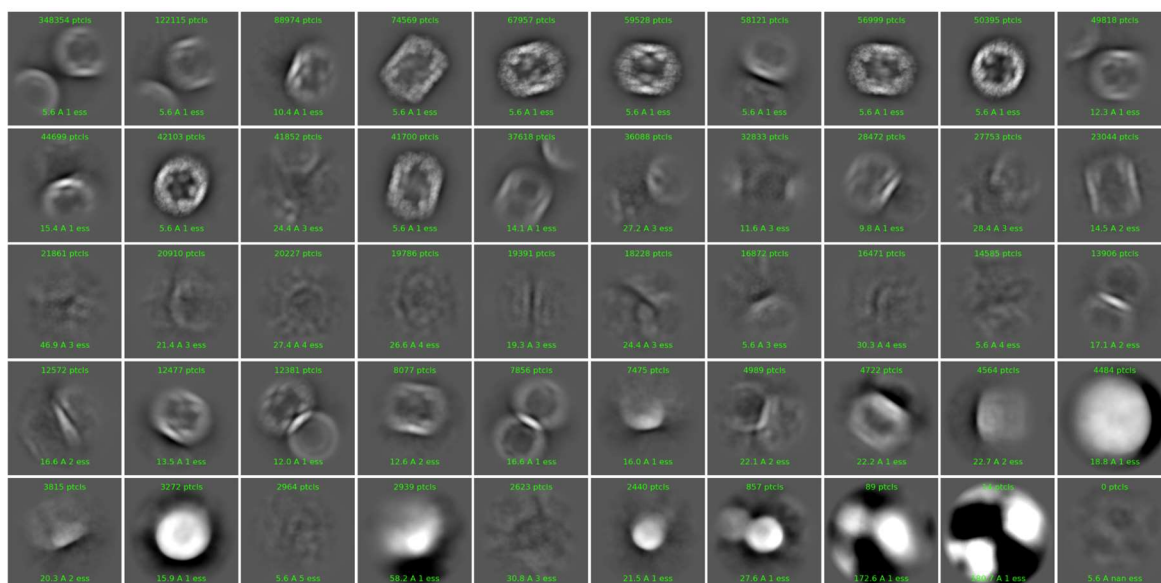
B
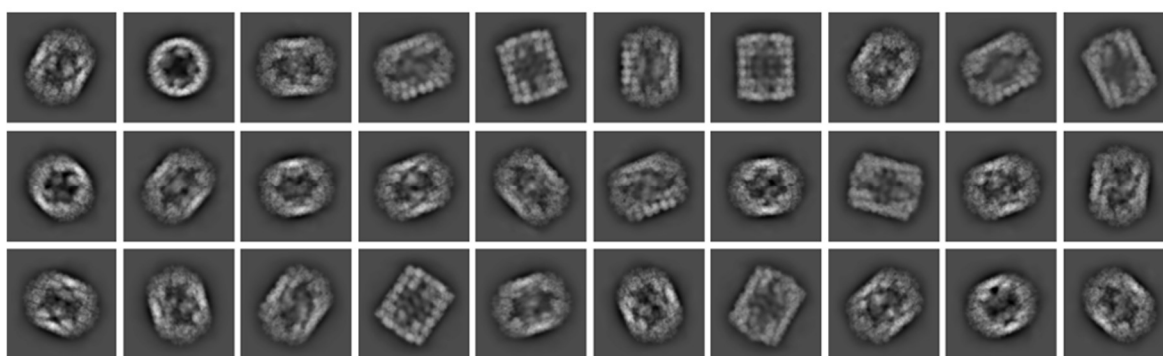
Figure 12 - **2D class averages of CCH.** Following the manual particle picking (not shown), newer particles were automatically picked based on the templates and redistributed to 50 distinct classes based on their projected orientations of CCH. A greater diversity of orientations is essential for a more precise and volume reconstruction. From all the classes presented (a), only seven were selected and reclassified as in (b), based on their orientation diversity, particle quantity, and resolution, producing 30 distinct views.

Source: By the author.

Once the particles were extracted from the micrographs, the first step of the 3D reconstruction, the *ab-Initio* 3D reconstruction, began. Here, 3D maps were reconstructed from a set of particles without the need for initial models or information, but some symmetry can be imposed on this reconstruction to assist the reconstruction. In this case, two symmetries were used for the rebuilding, the base C1 symmetry and a D5 symmetry, because, although the protein presents similarities with D5 symmetry, imposing it can inflict a decrease of resolution, and so the C1 symmetry, the lesser

one, is used as a base of comparison. For each symmetry, three maps were reconstructed, and based on what is known of the CCH, the best ones applied either C1 or D5 symmetry constraints. The number of particles used in each 3D reconstruction was 623,418 for both the C1 and the D5 maps.[19]
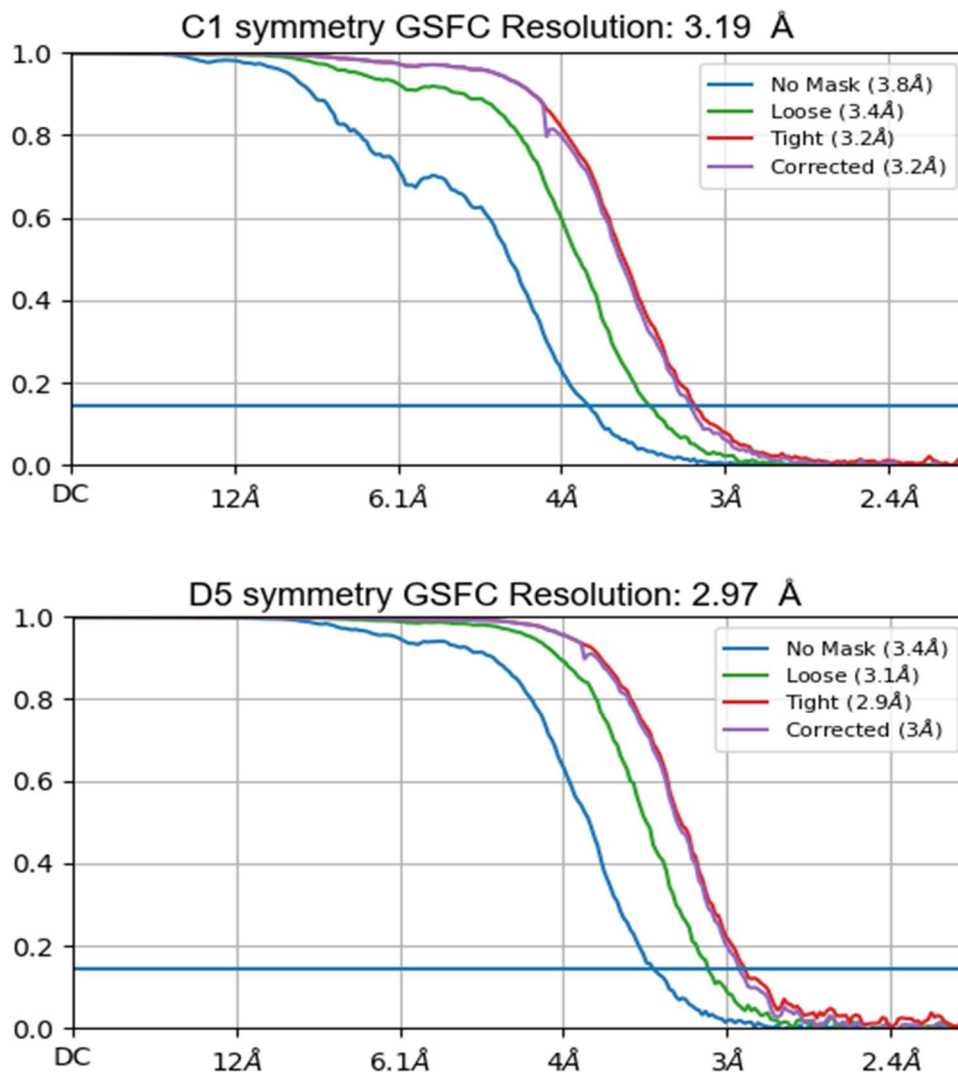
Figure 13 - Gold-Standard Fourier Shell Correlation (GSFCS) curve, with indicated global resolution of 3.19 Å and 2.97 Å, according to the 0.143 threshold (blue line) for the CCH. The curve shows the correlation in the function of the resolution. The blue curve indicates the direct correlation between the two independent half-maps, without adding a mask, and so with the structure and solvent together. The green one means the resolution after adding a loose mask over both half-maps, and the red curve applies a tight mask, reducing the expansion to differentiate the structure from the solvent. Finally, the purple curve applies a similar mask as the one presented on the red curve, but with a noise correction. This is the one used for the calculation of the final resolution, where the 0.143 threshold, blue line, crosses the purple curve, resulting in the global resolution of the electron density map.

Source: By the author.

Once *ab-initio* reconstruction was completed, rounds of 3D refinement started. It begins with a heterogeneous refinement protocol, in which the models, particles and imposed symmetry are used to reclassify the particles and allow the identification of more minor variations between models. Subsequently, the homogeneous refinement was done, distinctively from the previous one; it uses the existing particles making optical corrections, as on the defocus and higher-order aberrations. Once the refinement steps were completed, considering that they can be repeated and other steps of refinement can be done too, but usually with no resolution improvement after some repetitions, the quality of the density map is measured based on the Gold Standard Fourier Shell Correlation (GS-FSC), giving as a result a global resolution of the map. The resolutions of the C1 and D5 maps were respectively 3.19 Å and 2.97 Å, based on the 0.143 threshold in which the GS-FSC value best indicates the map's actual resolution, allowing a better comparison with X-ray crystallography. The general parameters microscope, data collection, and processing can be observed in Table 1, and a better identification of the processing steps can be kept in the flowchart in Figure 8.

With the enormous size of the CCH quaternary structure, visually consisting roughly of 160 FU, belonging to the A and B subunits, the only homologous structure available was the Keyhole Limpet Hemocyanin (KHL, PDB 4bed), from *Megathura crenulata*, solved in 2013 by Cryo-EM, at 9 Å global resolution. Thus, we estimate that full CCH will have a molecular weight of about 8 MDa or 70 thousand residues. The discrepancy observed between the calculated molecular weight from the DLS and the Cryo-EM structure (3.3 MDa vs. 8 MDa) is likely due to mathematical approximations of the Stokes equation, which assume the protein particle flowing through the liquid to be a solid sphere with homogeneous mass distribution across the volume. However, CCH is a hollow cylinder.

4.5 Structure analysis and glycosylation site identification

Due to the lack of the complete sequence information about CCHs FU, homologous models, crystallographic structures and sequence fragments of CCHs FU were used collectively, using AlphaFold, to generate models for the structures of CCHA and CCHB. FU-a, -b, -c, -e, and -f were generated using the FU subunits of RVH (RVH1

and RVH2) as templates, in combination with the structure of CCHB-g. FU-d was generated based on the sequence of CCHA-d. FU-g was generated based on the crystallographic structure of CCHB-g, obtained by the collaboration group from the University of Chile. And FU-h was generated with the combination of the KLH FU-h (PDB: 3QJO), in combination with the available sequence of CCHX2. Ultimately, the *Haliotis diversicolor* hemocyanin *(HDH)* was used as a template for the FU's organization, generating models for both CCHA and CCHB.

After the models were ready, they were aligned and docked to the CCH Cryo-EM map with the help of the software ChimeraX. Phenix and WinCoot were used to refine Cryo-EM maps, visual inspection, and manual building of loops and other missing portions, such as loops and linkers. Both maps had their correlation coefficients evaluated to determine the fitting rate of the models to the CCH cryo-EM map.

Once the structure was completed and the fitting was optimized in the map, the residues present were substituted with the actual sequence of the CCH, which was later experimentally obtained by the Chilean collaborators. Again, the new sequence was aligned, replaced, refined, and inspected using Phenix and WinCoot.

5 DISCUSSION

5.1 Symmetries Comparison analysis

Although both maps presented a good global resolution, with the D5 presenting the best between them, a local analysis was necessary to decide which presented the qualities necessary for the following analysis and structure construction, with the help of the software ChimeraX. In this analysis it was observed that the external wall of the CCH, primarily because of its stability, presented an excellent local resolution in both maps, allowing not only the identification of the majority of residues but also the identification of the possible active site, loops and some linkers (Figure 14). This shows that a general and external analysis of the maps wasn't enough to determine which of both was the best one to continue the studies.
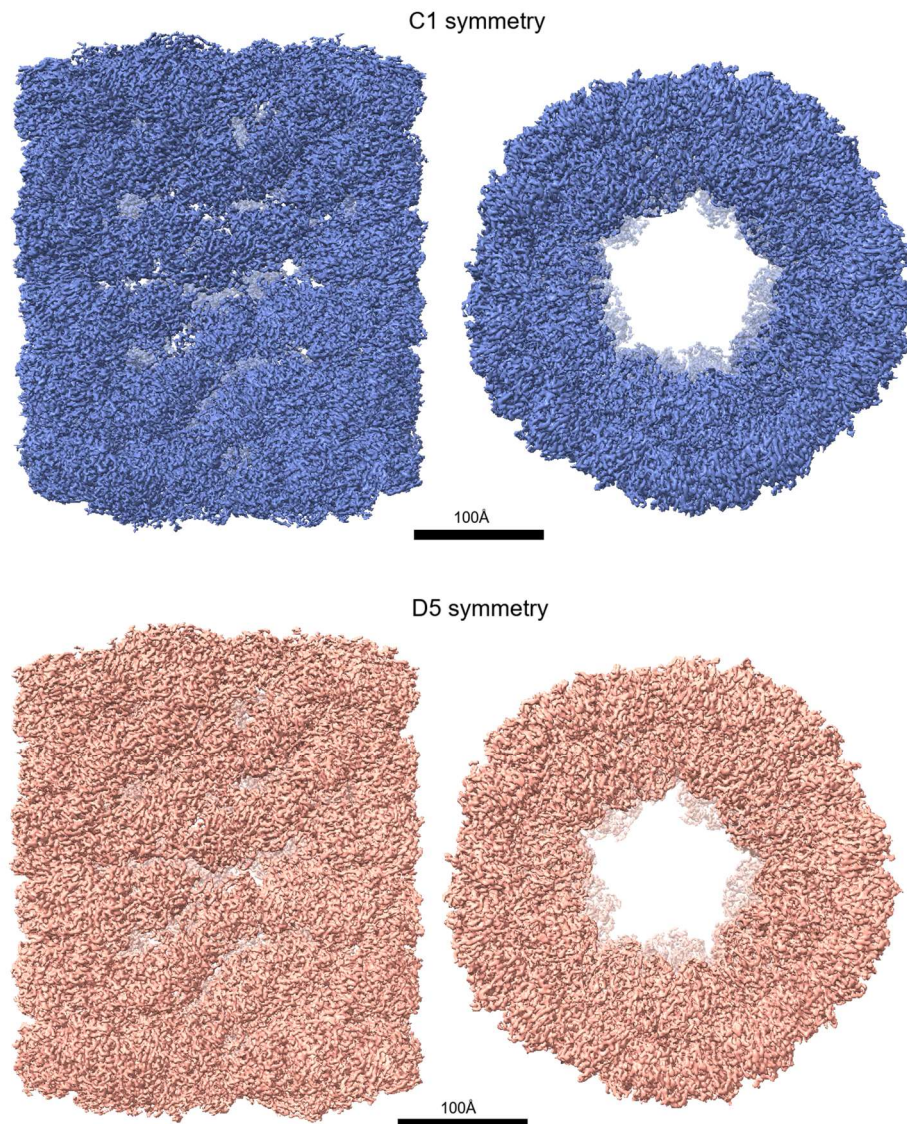
Figure 14 - Density maps with D5 and C1 symmetry. Side and top views of the resulting density maps reconstructed from the same data but with different symmetry parameters used for refinement. At the top, in blue, the C1 symmetry can be observed, while on the bottom, in peach, the D5 symmetry map can be observed. Although very similar and similar global resolutions, some specific regions present a lower resolution than the others.In this way, the C1 symmetry density map was chosen to be used in the following steps of the study. The cryo-EM maps are contoured at 5σ.

Source: By the author.

But when the extremes and the inner collars were examined, which are more flexible and more challenging to reconstruct, some crucial discrepancies were observed between the C1 and the D5 symmetry maps, as shown in Figure 15, possibly due to averaging regions that were not wholly symmetric Most loops and linkers on the top surface of the D5 symmetry map were fragmented and incomplete. In contrast, on the C1 symmetry map, most of these secondary structures were complete and with a good resolution. For this reason, the studies were continued using the density map

with the C1 symmetry. However, it doesn't indicate the lack of symmetry on the protein, just the presence of flexible regions that do not respect the protein symmetry entirely, and to allow a better study of them, is necessary the use of less rigid symmetry.
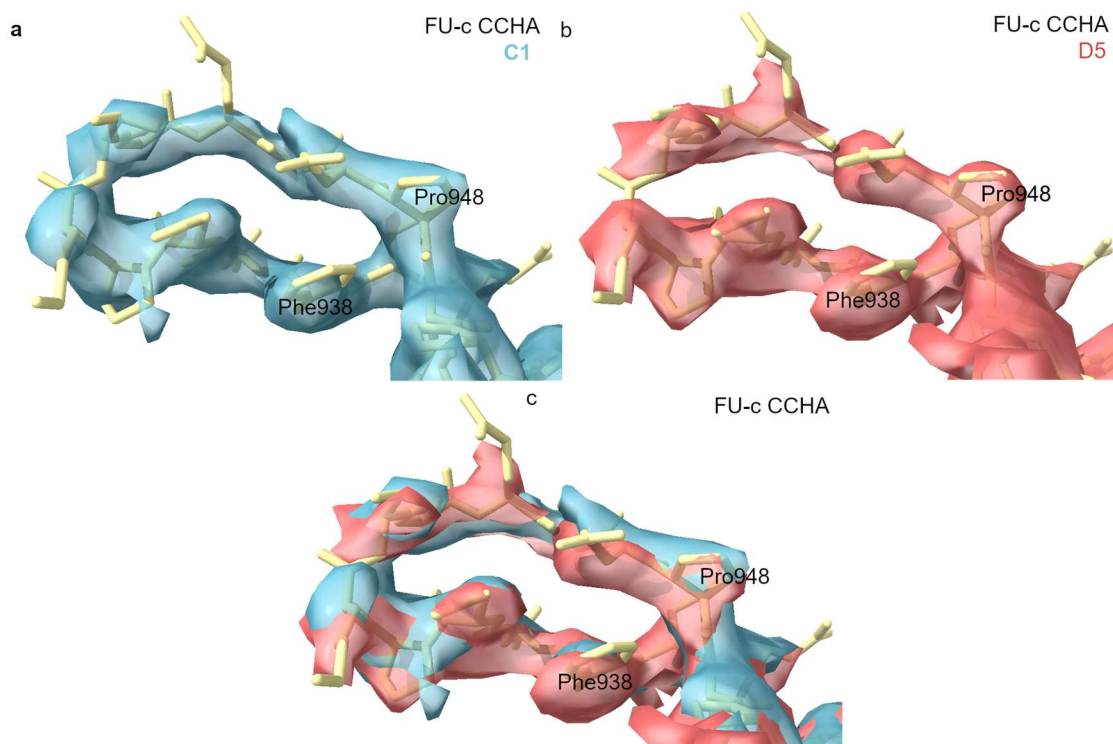
Figure 15 - Density maps comparison of D5 and C1 symmetries based on fitted structure. The fitted loop on FU-c of CCHA clearly demonstrates why the C1 symmetry was chosen over the D5 symmetry map, being present at the external ring in the top region of the molecule. The C1 symmetry map is presented in blue (a) and the D5 symmetry map in red (b), overlapped in c for a better comparison of density's presence, or lack, in this region. The D5 map presents a more fragmented map compared to the C1 map, which is more continuous and abundant. The cryo-EM maps shown are contoured at 5σ. The cryo-EM maps are contoured at 5σ.
Source: By the author.

Analyzing the local resolution map made it possible to determine the resolution distribution in the whole map. The closer to the center of the density map, the more uniform and continuous the resolution distribution is, varying between 3 and 4 Å, as seen in Figure 16. But as it gets closer to the corners, especially at the top, the resolution decreases, being around 5 and 6 Å, with some pointed regions with a better resolution, especially around the FU-a of CCHA. In the interior, it is also possible to see some regions with greater resolutions, respective to FU-g and FU-h of CCHA. This indicates that the subunit A of the CCH is more stable, allowing a better construction of its regions and a better resolution. This affected the structure because the FU

present on the regions with a greater resolution were easy to fit and visualize, while the regions at the extremes presented a lack of continuity, difficulting the in sequence fitting.
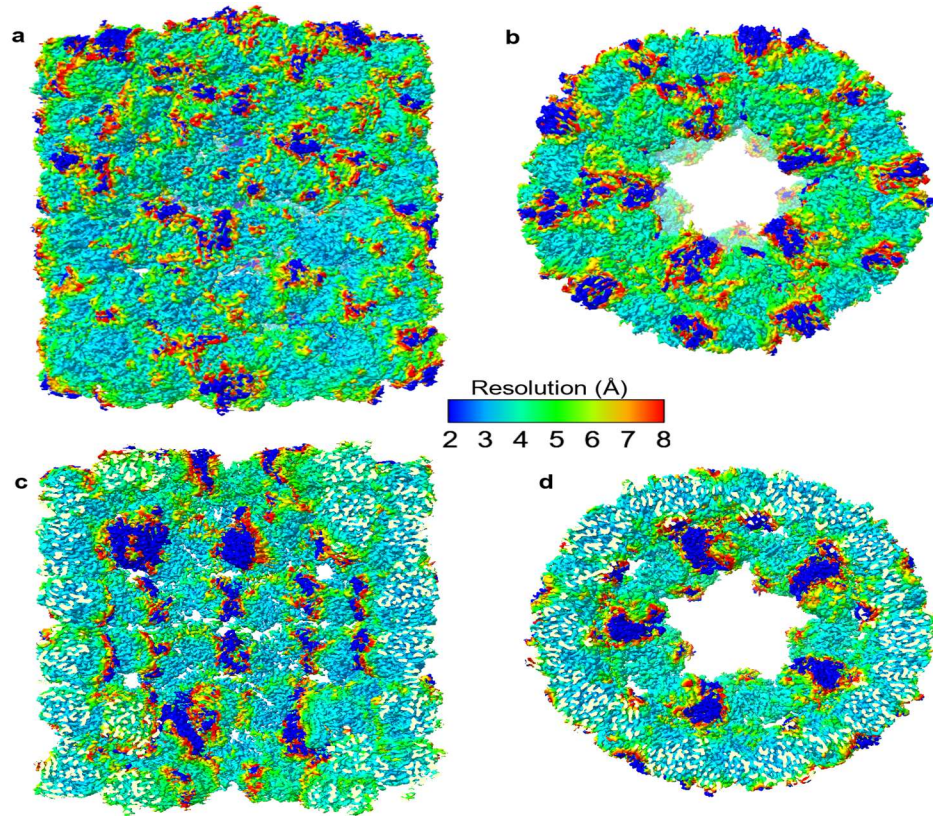
Figure 16 - Local resolution distribution on C1 symmetry map. The distribution of the local resolution around the Cq symmetry map after it was selected to be used on the study and then refined. On the side view (a) it is possible to observe that there is a continuous distribution of resolution, around 3 and 4 Å, on the body of the density map, but as it gets closer to the top view (b), the corners of the map have a decrease in resolution, being around 5 and 6 Å. The inner side view (c) and inner top view (d) show some regions with a greater resolution.
Source: By the author.

Table 1 - Information on the data collection and processing from the cryo-EM analysis of the CCH.

Microscope	Titan Krios K3
Detector	Falcon 3 (direct detection)
Voltage (kV)	300
Electron exposure (e⁻ /Å), # of frames	30, 20
Defocus range (μm)	-0.3 to -0.9

Spherical aberration (mm)	0.01
Pixel size (Å)	1.1
Symmetry	C1
Number of initial particles	2,335,087
Number of final particles	624,297
Map resolution (Å)	3.19
FCS threshold	0.143

Source: By the author.

5.2 Protein Model Building

The lack of a primary sequence information was a significant problem during the development of the CCH structure, even with a density map that allowed the correct main chain tracing and the observation of short or bulky side chains. Many techniques were tested, such as manually constructing the map, using automatic software like Modelangelo, replicating the consensus or known FU sequences in the entire structure, and refining the necessary spots. In the end, the one that presented the best result was the construction of the structure based on known sequences of FU from other mollusk hemocyanins, which presented a high degree of similarity with the pieces of the structure, along with the identified fragments of the CCH's FU. This was accomplished using the software Alphafold. For the FU-a, -b, -c, -e, and -f, both from CCHA and CCHB, the respective RVH1 and RVH2 sequences were used, using the known crystallographic structure of CCHB-g as a template. For the FU-d, the sequence of CCHA-d was used as well for the FU-h that is available, the sequence of CCHB-h, and for the FU-g, the crystallographic structure of CCHB-g was used.
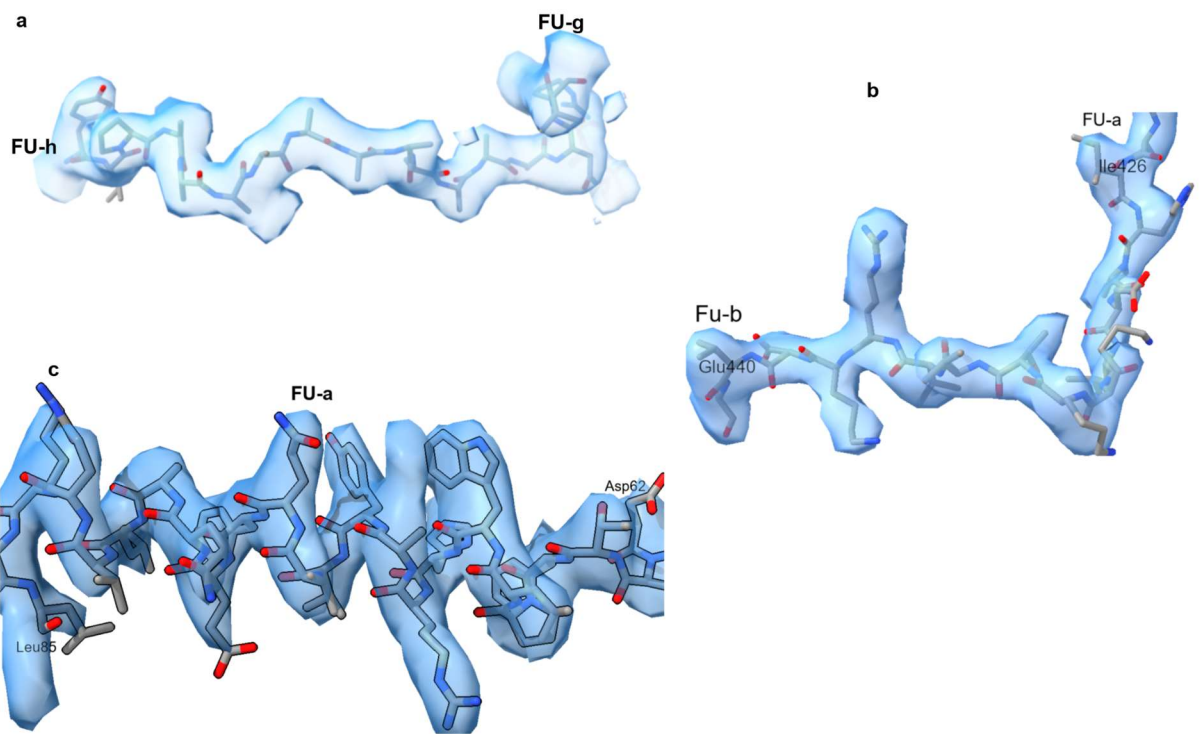
Figure 17 - Structure representation of interest regions. After the map constructions it was possible to observe and analyze more precisely some regions of interest. In **a** it's possible to observe a manually constructed linking region between functional units -g and -h of subunit B, interpreted as polyalanine, but the lack of sequence information led to a manual reconstruction of some regions, especially linking parts and loops in which the employed structure did not present a good fit, leading to a manual construction with polyalanine structures. This was a way to overcome the lack of sequence data that could limit the structural model analysis. In **b** is a fitted linker structure inside the density map of the region between functional units -a and -b of subunit B. The sequence information of this region was available, making it possible to make a great fit of the structure inside the density map, even for the more flexible regions, like the linkers regions between FU, allowing a more precise analysis of the whole structure. In **c** there is a sequenced Alpha-helix fitted inside the density map. The helix is present between amino acids 62 and 85, in FU-a of the subunit A. The map is not only able to present secondary structures, like alpha-helices and beta-sheets but also allows it to identify and allocate the residues in their corresponding regions. The Cryo-EM map is shown as a blue semi-transparent surface, contoured at 5σ.

Source: By the author.

Once there was a model for each FU, it was necessary to allocate them inside the density map. For this, determining their distribution along the dimers became essential. Based on what we know about the protein and the similarities both in shape and size with the HDH, this protein was used as a base for the FU distribution along the dimers together with the symmetry, a symmetry present in most of the protein due to its stability but less present on the inner collar that is more flexible. The result was two models of dimers, with alternated positions of the CCHA and CCHB. The model I

showed the CCHA at the center of the protein, in the more stable region and the CCHB at the up corner, and model II the opposite. In this way, it was possible to obtain a substantial part of the structure, only needing manual construction and adjustment on the loops, linking regions, and some other minor parts that did not present a good fit, but without the sequence, especially for the linkers, regions had to be completed with polyalanine.

The dimers models were used to complete the heterodidecamer structure through rigid *docking*, and then real space was refined with the help of the software Phenix and COOT. Based on the density map to validate the best structure, the known crystallographic structures of CCHB-g and -h were used as fitting parameters, since those are the only known complete fragments that belong to the CCH, decreasing the possible errors. Thus, the correlation coefficient was calculated for those FU in each model and compared, with a manual analysis of the fitting in both models.
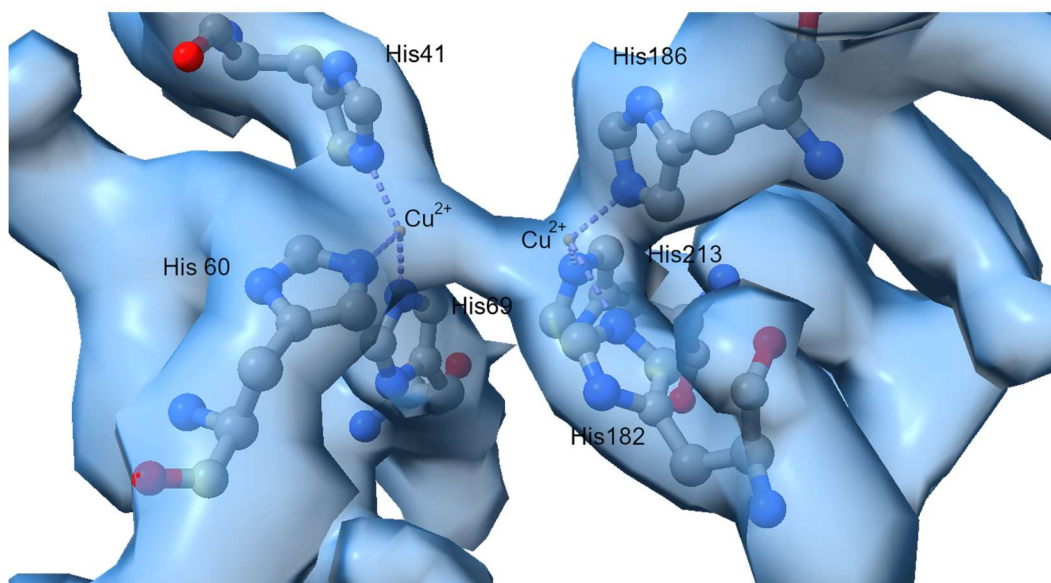
Figure 18 - **Active site of FU-a on subunit B.** The molecules presented can be separated into two groups: one with histidines 41, 60, 69 linked to one Cu^{2+} and the other with histidines 182, 186, 213 linked to another Cu^{2+}. Due to the distance and position of the molecules in the active sites, it is possible that they are not bound to molecular oxygen, but further analysis is needed to ensure this. The Cryo-EM map is shown as a blue semi-transparent surface, contoured at 5σ.

Source: By the author.

The CC factors indicate that model II is correct, ensured by the analysis of specific linkers and residues. This is more evident in the cases of the residue Arg15

from the CCHA-d, which fits in the density present in model II. Still, the same density isn't present in model I, but it accommodates a cysteine with a high degree of precision, something present in the FU-d of RVH1 and RVH2. The residues Pro50 and Ile94, suggest the same due to the lack of density in the model I, but an exact fit in the model II. If replaced with aspartic acid and phenylalanine residues present at these positions in RVH1 and RVH2, the residues fit correctly at the densities. These details support model II as the correct one.

Model I
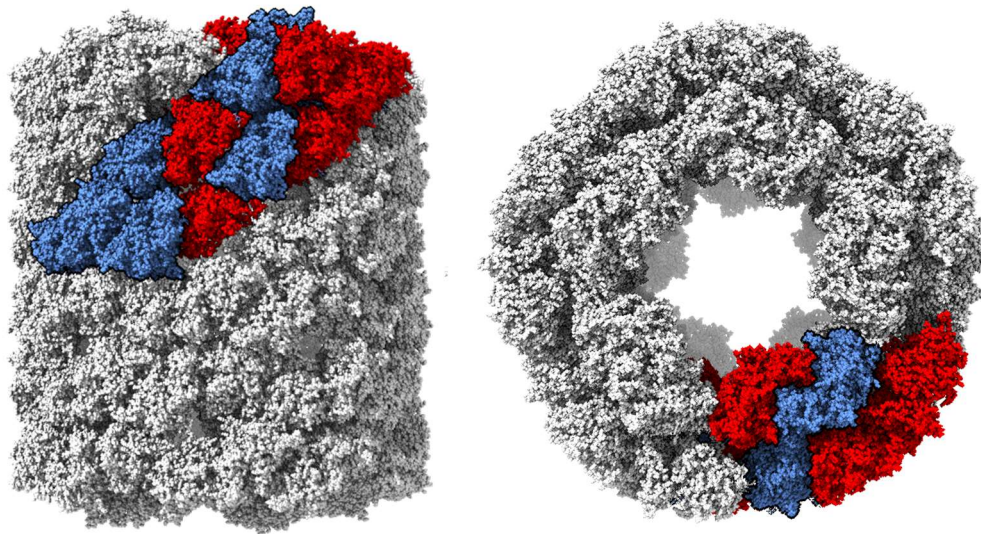
Model II
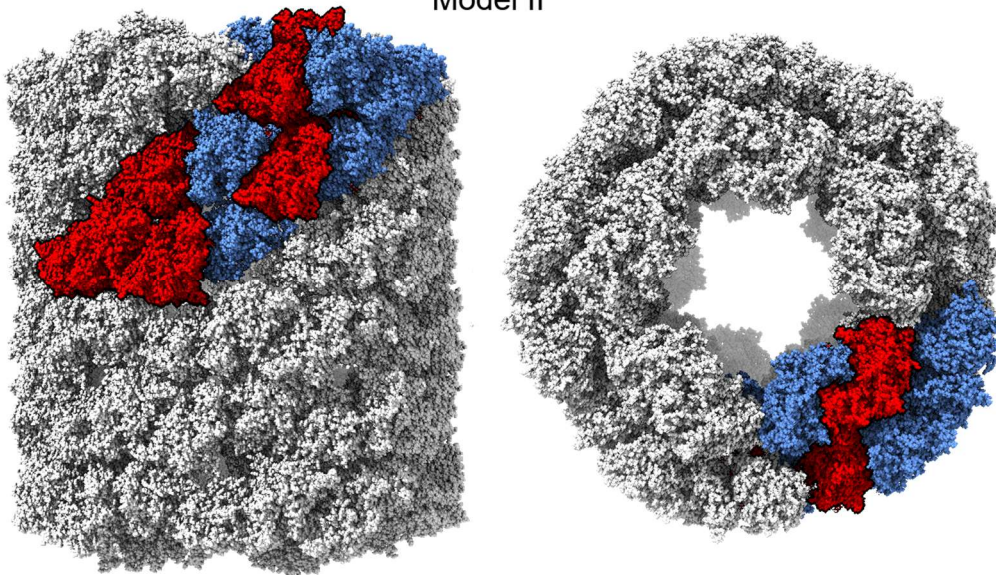
Figure 19 - Heterodidecameric models of the subunit organization of the CCH. By rigid docking, two models were created alternating the organization of the CCHA (red) and CCHB (blue) subunits in the structure, resulting in Model I and Model II. The distribution of the FU was based on the structure of HDH due to its high similarity to CCH.

Source: By the author.

With the best density map and structure available, the identification of characteristics and unique features began. It was possible to confirm that the present structure of the CCH was, as expected, a didecamer of type I, with an external diameter of 316 Å, an inner collar diameter of 218 Å, and a height of 366 Å, with each decamer measuring approximately 183 Å, similar to those presented in other hemocyanins, as the one used as a guide for the FU sequence, the HDH.

Table 2 - Correlation Coefficient (CC) values for each one of the different subunit organization models for the CCH. Model I shows the CCHA at the center of the protein, in the more stable region, and the CCHB at the up corner, while model II does the opposite.

	Model I (CC)	Model II (CC)
Didecamer Model	0.80	0.81
Refined Dimer Model	0.80	0.82
Solo Model CCHB-g and CCHB-h	0.68	0.81
Solo Model CCHA-d	0.76	0.79

Source: Adapted from MUÑOZ [20]

The high-resolution cryo-EM map facilitated the identification of six putative densities consistent with N-glycosylation in asparagine residues. Among these, four are located on the external surface of CCH, while the remaining two are positioned on the cylinder's internal surface. A representative glycosylation site is depicted in Figure 18, which features an omit difference map calculated using the Servalcat software. Thus, those residues that presented extra densities in their positions were analyzed, and the possible glycosylations were inspected, but even so, only the first saccharide of the glycosylation chain of each site was identified.

The same sites were projected over the entire protein through symmetry expansion, allowing the identification of 60 glycosylation sites all over the structure. However, a better resolution of the map density is needed to identify other sites and the subsequent residues of each pointed glycosylation.
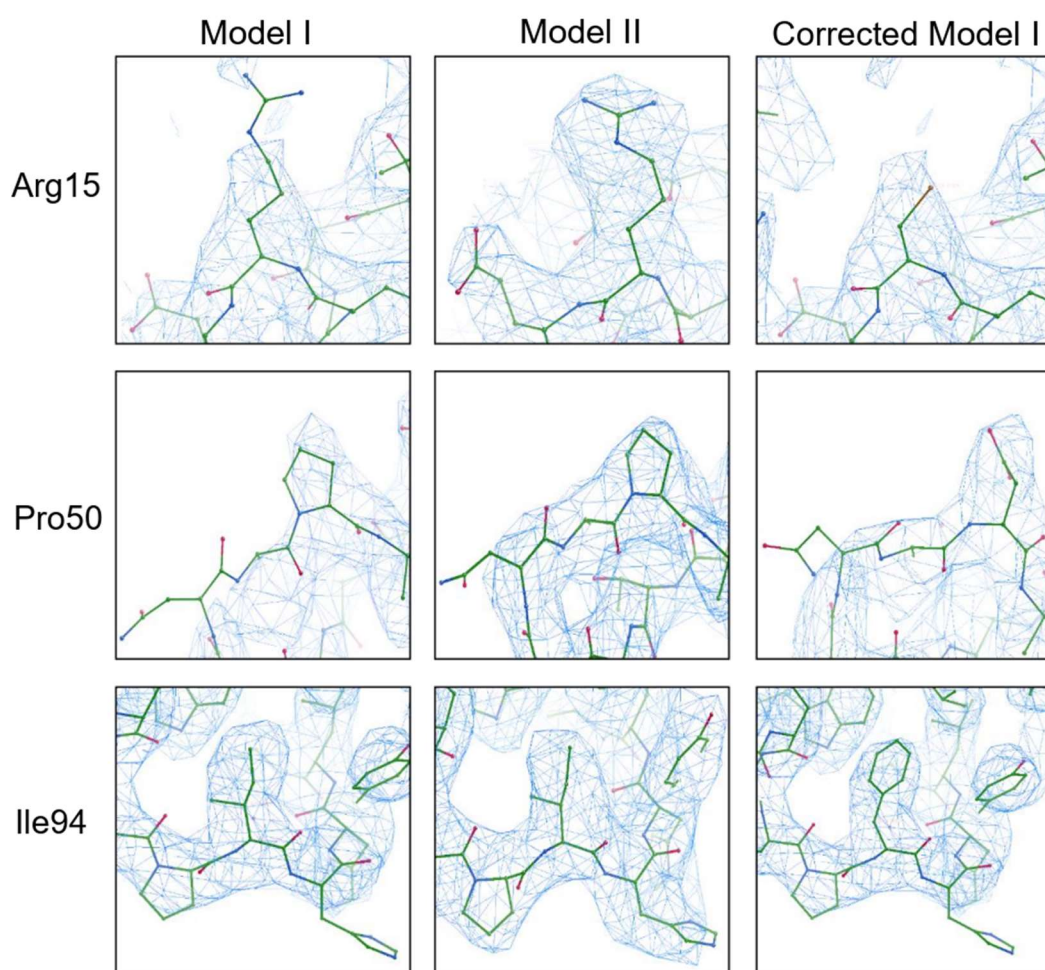
Figure 20 - Residues comparison of the CCHA-d FU between both subunit organization models and the corrected model. The first and second column shows the position of the residues Arg15, Pro50 and Ile94 at the FU CCHA-d and their fits at the corresponding density map regions at model I and II. The third column shows the same region but with their respective corrections substitutions, with a Cys, Asp and Phe, observed in the FU -f and -g of the subunits of RVH.

Source: Adapted from MUÑOZ. [20]

The final architecture of the intact CCH by single particle cryo-EM is depicted in Figure 21, representing the relative orientations of subunits A and B, which form the basic unit of the heterodidecamer. Notably, the crystal structure of CCHB-g FU was entirely used in the final cryo-EM. Subunit A starts and ends near the top region of the cylinder, while subunit B starts at the center and ends at the top of the cylinder.
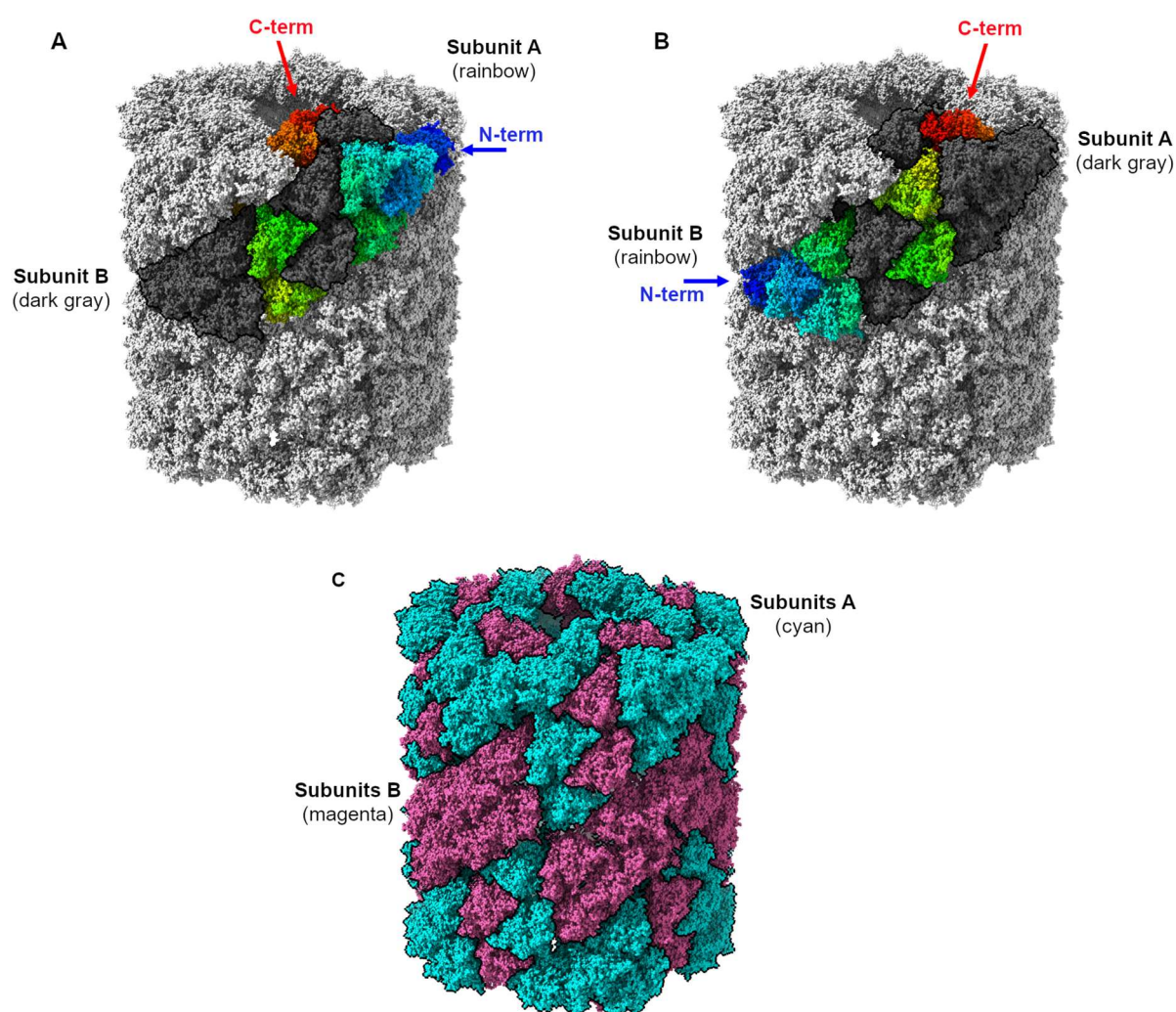
Figure 21 - The overall structure of intact CCH. Detailed representation of the comprehensive atomic model of CCH, emphasizing the relative positions and interplay of subunits A and B, which together form the foundational unit of the heterodidecamer. (A) Subunit A is illustrated using a rainbow gradient, transitioning from blue at the N-terminus to red at the C-terminus to depict the backbone's trajectory. Subunit B is represented in a contrasting dark gray, providing a spatial context for Subunit A, while the rest of the molecule is shaded in white for clarity. (C) Overall distribution of the A and B subunits (ten copies each) across the overall molecule.

Source: By the author

In summary, using Cryo-EM combined with other techniques like crystallography and negative staining allowed the elucidation of the structure of one of the biggest proteins studied with cryo-EM in Brazil. It was possible to confirm the structure of the CCH as a heterodidecamer, formed from ten dimers, each one composed of a subunit of CCHA and CCHB, entangled together and distributed along the protein with a, almost, D5 symmetry. The FU's distribution along the subunits

shows most of them at the external wall, but with the FU's -g and -h usually at the inner part of the protein, especially forming the inner collar, making them the most flexible, and thus more demanding of attention and refinement. The protein presents an external diameter of 314 Å, an inner collar diameter of 220.6 Å, and a height of 317 Å, with each decamer measuring approximately 158.5 Å, similar to other hemocyanins. Although both subunits present the same size, distribution and function, they have a weight difference of approximately 50 kDa, but the reason for this is still unknown.
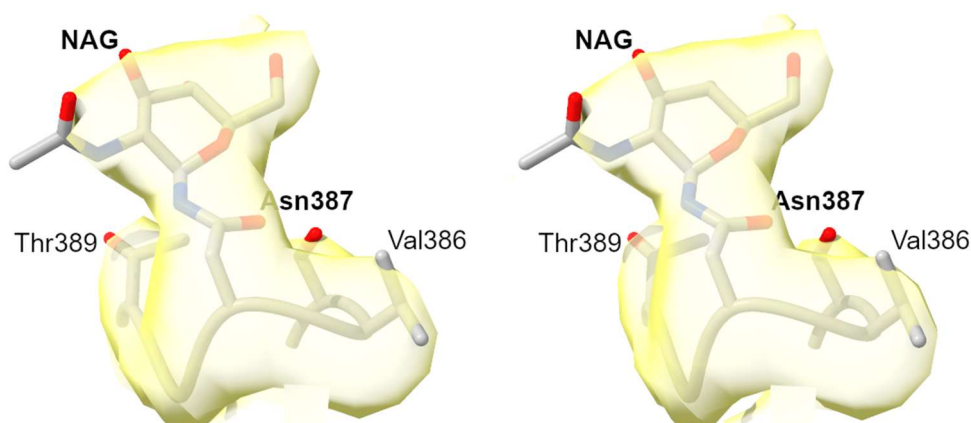
Figure 22 - Representative N-glycosylation in CCH. Wall-eyed stereoscopic representation of the Fo-Fc omit map (yellow) for visualization of the putative N-glycosylated Asp382, bound to a N-acetylglucosamine monomer, and neighboring densities, contoured at 10 σ. The difference map was calculated using the Servalcat software, which is based on Bayesian statistics derived from half-map variances, after omitting all the atoms between Ser385 and Thr389.

Source: By the author.

Despite the lack of the complete sequence, it was possible to obtain the complete structure of the protein, and from its analysis it was possible to observe the active site, linkers and loops and how the FU's are distributed and how they interact between them. But more importantly, it was possible to identify the position of 60 different glycosylation sites, allowing a better understanding of the medical applications of the protein.
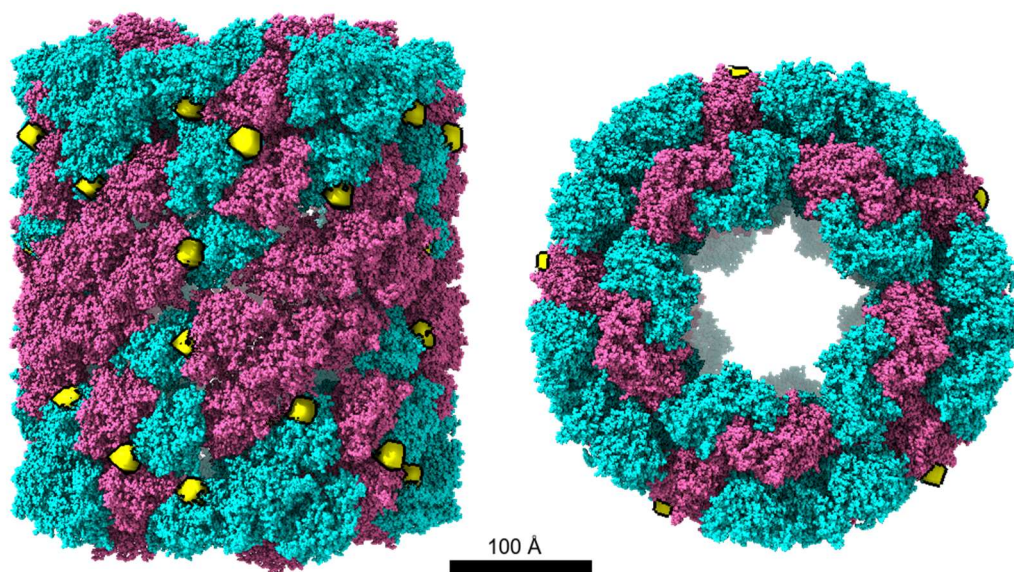
Figure 23 - N-glycosylation sites around CCH. Comprehensive depiction of all N-glycosylation positions within the complete CCH structure. The putative densities are consistent with N-glycosylation in asparagine residues (yellow spheres) spanning all the ten copies of subunits A (cyan) and subunits B (magenta). Two thirds of these spheres are located on the external surface of CCH, with the remaining one third is positioned on the cylinder's internal surface. This representation offers an enhanced understanding of the relative spatial distribution of the glycosylation sites across the entire protein surface.

Source: By the author.

The reconstruction of a 3.19 Å map of an 8 MDa protein is a significant achievement, specifically considering that the previous best resolution Cryo-EM structures of a hemocyanin is 9 Å for the KLH, one of the most studied hemocyanins and 4.5 Å for the HDH, making the CCH the best candidate for future studies and comparison to other hemocyanins. Even when there is a lack of the complete sequence of the CCH structure, its sequence is being sequenced with great success, and when most part is completed, those data will be used to increase the structure certainty for even more precise results.
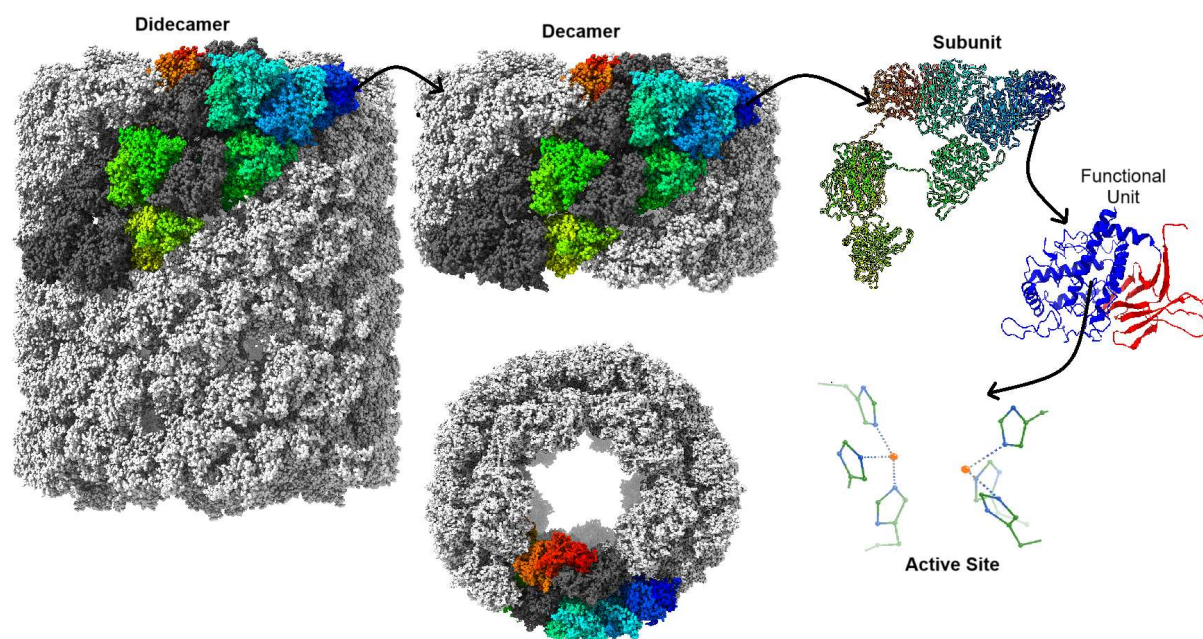
Figure 24 - **CCH configuration.** Top and side views of the didecamers, with the FU that compose the inner collar highlighted. Decamer representations with the subunits distribution, each decamer is composed of five CCHA subunits and five CCHB subunits. Subunits representation and how the FU are distributed inside it.

Source: By the author.

6 CONCLUSION

During the evolution of this study, while learning the methods of single particle Cryo-EM, it was also possible to observe the technological improvement over the years, with the advances in microscopes, the development of more efficient software and algorithms, and the association and addition of other techniques to be used together with Cryo-EM. This indicates that this is an excellent technique with much more to improve and impress, not only in studying small proteins but also in giant complexes.

In summary, the integration of Cryo-EM, crystallography and negative staining has facilitated the structural analysis of one of the largest proteins examined using Cryo-EM in Brazil. This research confirmed the structure of the CCH as a heterodidecamer, consisting of ten intertwined dimers. These dimers comprise CCHA and CCHB subunits, arranged with a D5 symmetry. The FU's distribution along the subunits present the majority of them at the external wall, but with the FU's -g and -h usually at the inner part of the protein, essentially forming the inner collar, making them the most flexible, and thus more demanding of attention and refinement. The protein presents an external diameter of 313.9 Å, an inner collar diameter of 220.6 Å, and a height of 317 Å, with each decamer measuring approximately 158.5 Å, similar to other hemocyanins. Both subunits share similar characteristics but differ in weight, in approximately 50 kDa, for reasons yet to be determined.

Despite the lack of the complete sequence at the beginning of the study, it was possible to obtain the complete structure of the protein, and from its analysis, it was possible to observe the active site, linkers, and loops and how the FU's are distributed and how they interact between them. More importantly, it was possible to identify the position of 60 different glycosylation sites, allowing a better understanding of the medical applications of the protein. Later, once the structure was substituted with the original sequence, the more fitted residues allowed a deeper structure, confirming the assumptions made when the sequence wasn't present.

Achieving a 3.19 Å map for an 8 MDa protein is significant. Previously, the best Cryo-EM resolutions for hemocyanins were 9 Å for the well-researched KLH and 4.5 Å for the HDH. CCH is a prime subject for future research and comparisons with other

hemocyanins. While the entire sequence of the CCH structure is still being determined, sequencing efforts have been promising. With the whole structure, the proper sequence and in significant resolution, further studies will allow the study of the biological and biochemical properties of the CCH as well as its function and applications in the medical field.

As future steps, we are currently teaming up with collaborators from Chile to unveil the mechanisms and functional details of CCH brought about by the high-resolution atomic model.

REFERENCES

1 COATES, C.J.; DECKER, H. Immunological properties of oxygen-transport proteins: hemoglobin, hemocyanin and hemerythrin. *Cell Molecular Life Science*, v. 74, n. 2, p. 293-317, 2016. DOI: 10.1007/s00018-016-2326-7.

2 MARKL, J. Evolution of molluscan hemocyanin structures. *Biochimica et Biophysica Acta*, v. 1834, p. 1840-1852, 2013. DOI: 10.1016/j.bbapap.2013.02.020.

3 GESHEVA, V. *et al*. Anti-cancer properties of gastropodan hemocyanins in murine model of colon carcinoma. *BMC Immunology*, v.15, p. 34, 2014. DOI: 10.1186/s12865-014-0034-3.

4 GONZÁLEZ, A. *et al*. The oxygen-binding properties of hemocyanin from the mollusk Concholepas concholepas. *Biochimica et Biophysica Acta (BBA)* - proteins and proteomics, v. 1865, p. 1746-1757, 2017. DOI: 10.1016/j.bbapap.2017.08.017.

5 DOLASHKA, P. *et al*. De novo structural determination of the oligosaccharide structure of hemocyanins from molluscs. *Biomolecules*, v. 10, n. 11, p. 1470, 2020. DOI: 10.3390/biom10111470.

6 KATO, S. *et al*. Molluscan hemocyanin: structure, evolution, and physiology. *Biophysical Reviews*, v.*10*, p. 191-202, 2017. DOI: 10.1007/s12551-017-0349-4.

7 ZHANG, Q. *et al*. Cryo-EM structure of a molluscan hemocyanin suggests its allosteric mechanism. *Structure*, v. 21, p. 604-613, 2013. DOI: 10.1016/j.str.2013.02.018.

8 GATSOGIANNIS, C.; MARKL, J. Keyhole limpet hemocyanin: 9-Å CryoEM structure and molecular model of the KLH1 didecamer reveal the interfaces and intricate topology of the 160 functional units. *Journal of Molecular Biology*, v. 385, p. 963-983, 2009. DOI: 10.1016/j.jmb.2008.10.080.

9 GATSOGIANNIS, C. *et al*. Nautilus pompilius hemocyanin: 9 Å Cryo-EM structure and molecular model reveal the subunit pathway and the interfaces between the 70 functional units. *Journal of Molecular Biology*, v. 374, p. 465-486, 2007. DOI: 10.1016/j.jmb.2007.09.036.

10 GAI, Z. *et al*. Crystal structure of the 3.8-MDa respiratory supermolecule hemocyanin at 3.0 A° resolution. *Structure*, v. 23, p. 2204-2212, 2015. DOI: 10.1016/j.str.2015.09.008.

11 TANAKA, Y. *et al*. Cryo-EM reveals the asymmetric assembly of squid hemocyanin. *IUCrJ*, v. 6, p. 426-437, 2019. DOI: 10.1107/S205225251900321X.

12 MANUBENS, A. *et al*. Concholepas hemocyanin biosynthesis takes place in the hepatopancreas, with hemocytes being involved in its metabolism. *Cell and Tissue Research*, v. 342, p. 423-435, 2010. DOI: 10.1007/s00441-010-1057-6.

13 IDAKIEVA, K. *et al*. Spectroscopic properties and conformational stability of concholepas hemocyanin. *Journal of Fluorescence*, v.18, p. 715-725, 2008. DOI: 10.1007/s10895-008-0338-x.

14 IOANNES, P. D. *et al*. Hemocyanin of the molluscan concholepas exhibits an unusual heterodecameric array of subunits. *Protein Structure and Folding*, v. 279, p. 26134-26142, 2004. DOI: 10.1074/jbc.M400903200.

15 STARK, H.; CHARI, A. Sample preparation of biological macromolecular assemblies for the determination of high-resolution structures by cryo-electron microscopy. *Microscopy*, v. 65, p. 23-34, 2016. DOI: 10.1093/jmicro/dfv367.

16 MARQUES, M. A.; PURDY, M. D.; YEAGER, M. CryoEM maps are full of potential. *Current Opinion in Structural Biology*, v. 58, p. 214-223, 2019. DOI: 10.1016/j.sbi.2019.04.006.

17 YIP, K. M. *et al*. Atomic-resolution protein structure determination by cryo-EM. *Nature*, v. 587, p. 157-161, 2020. DOI: 10.1038/s41586-020-2833-4.

18 AFANASYEV, P. *et al*. A posteriori correction of camera characteristics from large image data sets. *Scientific Reports*, v. 5, p. 10317, 2015. DOI: 10.1038/srep10317.

19 GUIMEI, Y. *et al*. An algorithm for estimation and correction of anisotropic magnification distortion of cryo-EM images without need of pre-calibration. *Journal of Structural Biology*, v. 195, n. 2, p. 207-215, 2016. DOI: 10.1016/j.jsb.2016.06.003.

20 MUÑOZ, S. M. *Estudios estructurales de la hemocianina de Concholepas concholepas:* un enfoque combinado de las técnicas de cristalografía de rayos X y Cryo-EM. 2022. Tesis (Ciencias Biológicas) - Facultad de Ciencias, Universidad de Chile, Santiago, 2022.

21 PUNJANI, A. *et al*. cryoSPARC: algorithms for rapid unsupervised cryo-EM structure determination. *Nature Methods*, v.14, p. 290-296, 2017. DOI: 10.1038/nmeth.4169.

22 DECKER, H.; TERWILLIGERT, N. Cops and robbers: putative evolution of copper oxygen-binding proteins. *Journal of Experimental Biology*, v. 203, n12, p.1777-1782, 2000.

23 GATSOGIANNIS, C. *et al*. Structure of mega-hemocyanin reveals protein origami in snails. *Structure*, v. 23, p. 93-103, 2015. DOI: 10.1016/j.str.2014.10.013.

24 HANSKE, J.; SADIAN, Y.; MÜLLER, C. W. The cryo-EM resolution revolution and transcription complexes. *Current Opinion in Structural Biology*, v. 52, p. 8-15, 2018. DOI: 10.1016/j.sbi.2018.07.002.

25 HRISTOVA, R. *et al*. Spectroscopic properties and stability of hemocyanins. *Spectrochimica Acta Part A*: molecular and biomolecular spectroscopy, v. 53, p. 471-478, 1997. DOI: 10.1016/S1386-1425(96)01837-9.

26 MÄEOTS, M. E.; ENCHEV, R. I. Structural dynamics: review of time-resolved cryo-EM. *Acta Crystallographica Section D*, v. 78, p. 927-935, 2022. DOI: 10.1107/S2059798322006155.

27 OLIVEIRA, T. M. *et al*. Cryo-EM: the resolution revolution and drug discovery. *SLAS Discovery*, v. 26, n.1, p.17-31, 2021. DOI: 10.1177/2472555220960401.

28 PERARO, M. D.; SONG, J.-J. Editorial overview: the future after the cryo-EM resolution revolution. *Current Opinion in Structural Biology*, v. 58, n. 3-4, 2019. DOI: 10.1016/j.sbi.2019.07.007.

29 PUNJANI, A.; ZHANG, H.; FLEET, D. J. Non-uniform refinement: adaptive regularization improves single-particle cryo-EM reconstruction. *Nature Methods*, v. 17, p. 1214-1221, 2020. DOI: 10.1038/s41592-020-00990-8.

30 WIGGE, C.; STEFANOVIC, A.; RADJAINIA, M. The rapidly evolving role of cryo-EM in drug design. *Drug Discovery Today*: technologies, v. 38, p. 91-102, 2020. DOI: 10.1016/j.ddtec.2020.12.003.

1. **ANNEX A - CCHA FU distribution diagrams and sequence.**

FU-a (CCHA)

SLVRKNVDSLTEDEILTLQSALRDLENDETDHGFQALAAYHGEPSQCALKDGKLIACCLHGDPTFPLWHRAYVVQFEQALVQKGLANLGVPYWDWTEPLT
1   5   10   15   20   25   30   35   40   45   50   55   60   65   70   75   80   85   90   95   100

HLPELVSHPIFRDPKGGKGQKNAWFTADIEHGDVHEHTARAVDDRLFEHVEPGHKTRLFEMILDALEQDEFCHFEAQFEVAHNHIHYLVGGRHHFSMSTL
101 105 110 115 120 125 130 135 140 145 150 155 160 165 170 175 180 185 190 195 200

EYTSYDPLFFLHHSNADRLFVIWQELQKKRGKPHDHSDCGVELFGKPMEPFNHENNPVELTKTFSKPKTDLPLGYTYDTLSLNGMTPEQLHVLLEERRGR
201 205 210 215 220 225 230 235 240 245 250 255 260 265 270 275 280 285 290 295 300

GPLRVLRRFLHPGLSANLRQICHPESEEEDHCEFSGDFFILGGANEMPWEFTVPYYFDVTETMDHMLHGIEEDHGHLEVKAEVYSVNGTLLGSQILGKPY
301 305 310 315 320 325 330 335 340 345 350 355 360 365 370 375 380 385 390 395 400

FU-b (CCHA)

Linker a-b (CCHA)

ISHRPAHGFTDRRTTTVGRKDVDVLTDEEVYHLRQALARFQNDTSIDGFQAVAEFHGLPAKCPHPDAAVRYACCIHGMPTFPHWHRLFVTEVEDELRSRG
401 405 410 415 420 425 430 435 440 445 450 455 460 465 470 475 480 485 490 495 500

LEIGIPYWDWTKPNSHVPALASEENYEHPYTHETLHNPFHDASMLLEGERTTRQIQDELSETPEFGDHTTLFDGMLLAFEQTDFCDFEVQFEVVHNAIHF
501 505 510 515 520 525 530 535 540 545 550 555 560 565 570 575 580 585 590 595 600

LVGGFGPYTMATLHYSAFDPIFFLHHSNVDRLWAIWQELQMRRGKPYKAHCAESLEKEPMKPFGFPPPYHNDEKTHDNARPTKIYDYESTLHYSYDDLQF
601 605 610 615 620 625 630 635 640 645 650 655 660 665 670 675 680 685 690 695 700

GGMTIDQLDEYLHERKQHDRVFVGIMLHNIGVSAWWTLFINRAGGDKYEVGKLAVLGGEKEMPWRFDRAYKVEITDALKTLGFSYDDDFHVHLDITDVTG
701 705 710 715 720 725 730 735 740 745 750 755 760 765 770 775 780 785 790 795 800

FU-c (CCHA)

Linker b-c (CCHA)

KKWPSDTFPHPTIIHEPARHRQE   GGHHVRSDVEHLTAEQMQNLREALLALKADSSISGYNQIASFHGQPNWCPSPDAEKKYACCAHGMSIFPHWHRLL
801 805 810 815 820   826 830 835 840 845 850 855 860 865 870 875 880 885 890 895 900

TVQAENALMAHGMTSGLPYWDWTLPIHHLPGLVSGENFDNPKTSQSEPNPWFSGEVDGHTTQRHAREDLFKQPAFGEMTPLAEKVMMAFEQDNFCDFEIQ
901 905 910 915 920 925 930 935 940 945 950 955 960 965 970 975 980 985 990 995 1000

FEIAHNFIHAVVGGTELYSMASLRYTAFDPLFYLHHSSTDRLWAIWQKLQKFRNKPYNSANCALSEMRKPLQPFAQTSATNPDPVTRDHSVPFDVFNYHN
1001 1005 1010 1015 1020 1025 1030 1035 1040 1045 1050 1055 1060 1065 1070 1075 1080 1085 1090 1095 1100

SFHYEYDNFQFNGMSIPQLQRETLKRQGLERVFAGFMLHGIKKSSLVVFNICTPDNSCTKAGEFYLLGDEFELPWEYNRLYKHEITKQLKAMGLDHHDPF
1101 1105 1110 1115 1120 1125 1130 1135 1140 1145 1150 1155 1160 1165 1170 1175 1180 1185 1190 1195 1200

FU-d (CCHA)

Linker c-d (CCHA)

DIHYNVFDLESNDLGHDVFGNASIIFKSG   VRASHIRRNIEDLTKGERESLRSALHDMEEDGGFEKIAKFHGYPGMCTHDNHPASCCVHGSPTFPHWHR
1201 1205 1210 1215 1220 1225   1251 1255 1260 1265 1270 1275 1280 1285 1290 1295 1300 1305 1310 1315

LYVENVENALLNHGSAVSMPYWDWTQPIHHLPPLIAEPTYYDSRTHHKEPNAFFRWKVPDTEEYTSRDPRPELYDSHYFLDNVLLALEQTSFCDFEVQFE

1320 1325 1330 1335 1340 1345 1350 1355 1360 1365 1370 1375 1380 1385 1390 1395 1400 1405 1410 1415

IVHNGLHSFLGGRGQYSMSTLDYSAFDPVFFLHHANVDRIWAIWQALQKFRGLPFDETDCAMNLMNIPLHPFDYDELNDCWVTHKYSRPKDVFDYSHHFD

1420 1425 1430 1435 1440 1445 1450 1455 1460 1465 1470 1475 1480 1485 1490 1495 1500 1505 1510 1515

YQYDNLNFHGLTIAQLEHLLEERKSEDRAFLNFMLKGIGTSADVEFELCDKDNHCNFAGTFAVLGGPLEMPWSFDRVFKYDVTKVFKQMQLRPDSEYHFE

1520 1525 1530 1535 1540 1545 1550 1555 1560 1565 1570 1575 1580 1585 1590 1595 1600 1605 1610 1615

Linker d-e (CCHA)

HHIRAVNGTDLNPHLLKPPSVSFVPGRKSY   HRNLVRKSVRNLSPAERRSLVLAMKSLQEDSSADGFQSLASFHAQPPLCPYPEANKRFACCVHGMATF

1620 1625 1630 1635 1640 1645 1662 1670 1675 1680 1685 1690 1695 1700 1705 1710 1715 1720 1725

FU-e (CCHA)

PEWHRLYTVQFEDALRRHGSVVGIPYWDTVVPQEDLPAFFNDETWDDPLFHANFTNPWHGADIEFNHHEVARDINMEKLGQKGPKGYDTWSWKQAIYALE

1730 1735 1740 1745 1750 1755 1760 1765 1770 1775 1780 1785 1790 1795 1800 1805 1810 1815 1820 1825

QEDYCDFEVQFEIAHNAIHAWMGGHEEYSMGHLHYASYDPVFILHHSNTDRIFALWQELQKFRGHDPNEVNCALEMMREPLKPFSFGAPYNLNPTTKEHS

1830 1835 1840 1845 1850 1855 1860 1865 1870 1875 1880 1885 1890 1895 1900 1905 1910 1915 1920 1925

KPEDTFDYKGHFHYEYDHLELQGMSVQRLNDYINQEKERDRVFVGFLLRGIGTSAHVDFSICKPDGECTHAGFFDVLGGSLETPWQFDRLYKYEITDVLE

1930 1935 1940 1945 1950 1955 1960 1965 1970 1975 1980 1985 1990 1995 2000 2005 2010 2015 2020 2025

Linker e-f (CCHA)

SKGLDVHDVFDIKVTSTTWDNEDIPTDRFPPPSVIYIPKSRDIELHEVAPNRVRNDLSKLSGRDIQNLKSAIRDLQLDESTDGYQNIASFHGLPALCHDA

030 2035 2040 2045 2050 2055 2060 2065 2070 2075 2080 2085 2090 2095 2100 2105 2110 2115 2120 2125

FU-f (CCHA)

EGHEYACCIHGNPTFPHWHRLYTVEMEDALIRHGSGVALPYWDWSAAITKLPDLFTSESYYDAWRDEVIPNPFARSYINVAKGYTVRDPQPELTKLAKDG

2130 2135 2140 2145 2150 2155 2160 2165 2170 2175 2180 2185 2190 2195 2200 2205 2210 2215 2220 2225

KHSMLFDEMLLVLEQTDYCDFEVQFEVLHNAIHFLVGGRQEYSLSTLEYSSYDPIFFVHHSFVDKVWAVWQELQKKRGLVYDRVDCAVNYINQPMHPFDW

2230 2235 2240 2245 2250 2255 2260 2265 2270 2275 2280 2285 2290 2295 2300 2305 2310 2315 2320 2325

ESLNPDVRTREHASPSSVFDYEDLGYHYDNFELGGNDLEKLAELIHEQKSHARVFAAFHLHNIGTSAAVVFEVCRTESNCVRAGGFFILGGKLEMPWSFD

2330 2335 2340 2345 2350 2355 2360 2365 2370 2375 2380 2385 2390 2395 2400 2405 2410 2415 2420 2425

Linker f-g (CCHA)

RLFKYDITSVLHDIGIEPEDVFNPEAPFFLKYEIQGVNGSSLPLSTVSPPTLVFKPAEE   AGVGVRKNINTLTDAEMENLREAMRRVQAGTGRLSYNKV

2430 2435 2440 2445 2450 2455 2460 2465 2470 2475 2480 2485 2499 2505 2510 2515 2520 2525 2530 2535

FU-g (CCHA)

AKAHGYPAQCHMKDDGKEVACCHHGMASFPQWHRVFAKQMEAALMWEGAELGLPYWDWTESFTELPRLISQEHDNPFHHFEVPDKEHHVTTRAPRPQLFK

2538 2545 2550 2555 2560 2565 2570 2575 2580 2585 2590 2595 2600 2605 2610 2615 2620 2625 2630 2635

DPEHGDESFFYRQVLIAFEQRDYCDFEVQFEVIHNAIHSWIGGTSPYGMSTLEYSAYDPVFFIHHSNVDRQFAIWQALQKYRGLDYNTANCNIHELREPL
2638 2645 2650 2655 2660 2665 2670 2675 2680 2685 2690 2695 2700 2705 2710 2715 2720 2725 2730 2735

EPFNHEDNPVHVTRAHSRAIDAFDYDELDYQYDNLNFHGLTIPQLEHLLEERKSEDRAFLNFMLKGIGTSADVEFELCDKDNHCNFAGTFAVLGGPLEMP
2738 2745 2750 2755 2760 2765 2770 2775 2780 2785 2790 2795 2800 2805 2810 2815 2820 2825 2830 2835

Linker g-h (Manubens)

WSFDRVFKYDVTKVFKQMQLRPDSEYHFEHHIRAVNGTDLNPHLLKPPSVSFESDA   EDLVRYEVSALSLAQVTNLRNALNKLMHDHGPNGYEAIAGFH
2838 2845 2850 2855 2860 2865 2870 2875 2880 2885 2890   2906 2910 2915 2920 2925 2930 2935 2940 2945

FU-h (CCHA)

GAPGLCPENAEEHYSCCQHGMPAFPHMHSLLTVQFEHALREQGALVGIPYWDWTRPSKSLPSLFTDPSENNPFKTYRMSFNNEFIKREVHEDLFNHPSEG
2948 2955 2960 2965 2970 2975 2980 2985 2990 2995 3000 3005 3010 3015 3020 3025 3030 3035 3040 3045

DPESLFHQALELLEQTTYCDFEVQYEMLHNAVHELVGGPNTYSMSTLDYSAFDPFFMVHHASIDRLWHIWQQLQKLRHKPYNYAVCAARTLYRHLEPFSY
3048 3055 3060 3065 3070 3075 3080 3085 3090 3095 3100 3105 3110 3115 3120 3125 3130 3135 3140 3145

ESVNMDPLTRNNAQPVQIFDKKKFHYHYDNLQLSGHALQELEDMVHDMQSHTRIFAGFVLHGFGTSARVQVDIVEDGHVEKVGNFFVLGGASEMPWAYER
3148 3155 3160 3165 3170 3175 3180 3185 3190 3195 3200 3205 3210 3215 3220 3225 3230 3235 3240 3245

IYKLDMTEAAEHLNMTDSCYFHFQLTVTKYDGTPLDVEFPDPVVIRKRANTEFDEVVLPLKPLNKLPPKIVVRRGTHVIFHPVEDALHGPIRELGSYTNS
3248 3255 3260 3265 3270 3275 3280 3285 3290 3295 3300 3305 3310 3315 3320 3325 3330 3335 3340 3345

VHCAIPPGEAHAYAFDESHELEPGDYYFTANDAQQCHGGARIQLTVDEE
3348 3355 3360 3365 3370 3375 3380 3385 3390 3395

The FU and linkers regions, from CCHA, are enumerated on the following order:

- FU-a: 1-407
- linker a-b: 408-421
- FU-b: 422-821
- linker b-c: 822-835
- FU-c: 836-1228
- linker c-d: 1229-1260
- FU-d: 1261-1647
- linker d-e: 1648-1672
- FU-e: 1673-2058
- linker e-f: 2059-2092
- FU-f : 2093-2487
- linker f-g: 2488-2547
- FU-g: 2548-2894
- linker g-h: 2895-2926
- FU-h: 2927-3396

2. ANNEX B - CCHB FU distribution diagrams and sequence.

FU-a (CCHA)

SLLRKNVDSLTEQEVLRFQNTLLELEKDESAHGFQALGAFHGEPSGCHMKDGRAIACCVHGEATFPMMHRAYTVQFEQLLVEKGLANLGVPYWDWTKELT
```
1    5   10   15   20   25   30   35   40   45   50   55   60   65   70   75   80   85   90   95   100
```

HLPELVQHQIFKDPDGGVGQRNAWFSAEINHGDVHEHTARAVDERLFEHPEEGGKTRLFNLILDALEQDDYCHFEAQFEVAHNHIHYLVGGRHHFSMSTL
```
101  105  110  115  120  125  130  135  140  145  150  155  160  165  170  175  180  185  190  195  200
```

EYTSYDPIFFLHHSNVDRIFVIWQEMQKRRGKPYDHADCSVELFNKPMEPFKRDDNPVELTRTYSKPRDLFSHLQLGYNYDDLHLGGMSLDELYALLEER
```
201  205  210  215  220  225  230  235  240  245  250  255  260  265  270  275  280  285  290  295  300
```

HGRDRTFATFALHGLGFSANVRVKVCDPHDDGEAQAAAASDRNDEHCEFSGDFFVLGGANEMPWEFTVPYYFDVTEAVHKLKLESGGHYGVHADVYSVNG
```
301  305  310  315  320  325  330  335  340  345  350  355  360  365  370  375  380  385  390  395  400
```

FU-b (CCHA)

Linker a-b (RVH1)

TLLGAHVLGSRPYLSHRPEGYHDPEIHPDLKKTVVRKDVETLTDEEVYYLRQALAKFQNDTSIDGYQAVAEYHGLPPKCPHPDAAVRYACCIHGMATFPH
```
401  405  410  415  420  425  430  435  440  445  450  455  460  465  470  475  480  485  490  495  500
```

WHRLFVTEVEDELKSRGLEIGIPYWDWTKPNTHVPSLAAEETYEDPHSHHQVHNPFHDAVVAFLGEKTTRDVQTELSETPAFGDHTALFDGLLLAFEQTD
```
501  505  510  515  520  525  530  535  540  545  550  555  560  565  570  575  580  585  590  595  600
```

FCDFEVQFEVVHNAIHFLVGGFESYTMATLHYSAYDPIFYLHHSNVDRLWAIWQKLQMRRGKLYKVHCAGSLTQEVMKPFGFPPPYHNDPKTHDNARPSK
```
601  605  610  615  620  625  630  635  640  645  650  655  660  665  670  675  680  685  690  695  700
```

IYDYERVLDYTYDSMQFGGMTIDQLDHYLEERQTHDRTFVGIMLHNIGISAWATLTIEMKNEEEHTVGKIAVLGGEKEMPWHFDRVYKVEITDALHKLGY
```
701  705  710  715  720  725  730  735  740  745  750  755  760  765  770  775  780  785  790  795  800
```

FU-c (CCHA)

Linker b-c (RVH1)

RYDDDYEVTLHITDVAGNEWPEDTFPHHTIIHEPASTHLEE   EEEHVRKDVDTLTPEQMNNLREALDTLKHDHSVGGFNHIAAFHGQPNWCPSMEAEHK
```
801  805  810  815  820  825  830  835  840       850  855  860  865  870  875  880  885  890  895  900
```

VACCTHGMPVFPHWHRLLTVQAENALIAHGMHSGLPYWDWTLPMTALPEFVSAAEYENPKTHQTVPNPWFSGEVDTVGHATSRMVRDELFEQPEFDKMTR
```
902  910  915  920  925  930  935  940  945  950  955  960  965  970  975  980  985  990  995  1000
```

IAEKVMLAFEQDNFCDFEIQYEIAHNHIHALVGGNKLYSMASLRFTAFDPLFFLHHSNTDRIWAIWQALQKLRGKPYNSANCAISELRKPLQPFAQTSVT
```
1002 1010 1015 1020 1025 1030 1035 1040 1045 1050 1055 1060 1065 1070 1075 1080 1085 1090 1095 1100
```

NPDPVTRDHAVPFDVFKYQENFHYHYDNLQFNGMSPPQLNRELVKRKGLDRVFAGFMLHGIQKSSLVKFNICKPDATCHPAGEFYLLGDEFEMPWEYDRL
```
1102 1110 1115 1120 1125 1130 1135 1140 1145 1150 1155 1160 1165 1170 1175 1180 1185 1190 1195 1200
```

FU-d (CCHA)

Linker c-d (RVH1)

FKYDITEQLDKFDLHPADHYDIAYTVFDLNGQNLGEDLFGNATIIVHEPGLGHMKGHEEDYRAEV   HVRKNLSSLTLGECESLRSAFHDMENDGTYESIA
```
1202 1210 1215 1220 1225 1230 1235 1240 1245 1250 1255 1260 1265      1275 1280 1285 1290 1295 1300
```

RFHGYPGLCDHNGKHVGCCVHGAPTFPHVHRLFVEQVENALLNHGSAVSVPYWDWTEDFDHLPKLLGMPTYFKVPDTRKDVNPFFRWKVTGTGEYTTRDP

| 1304 | 1310 | 1315 | 1320 | 1325 | 1330 | 1335 | 1340 | 1345 | 1350 | 1355 | 1360 | 1365 | 1370 | 1375 | 1380 | 1385 | 1390 | 1395 | 1400 |

RFDLFSADYFLENVLLALEQTSYCDFEVQFEVSHNAIHSFLGGRAKMSMSTLDYSAFDPLFFLHHANMDRIWAIWQALQKHRDLPFDETDCAVNLMSKTL

| 1404 | 1410 | 1415 | 1420 | 1425 | 1430 | 1435 | 1440 | 1445 | 1450 | 1455 | 1460 | 1465 | 1470 | 1475 | 1480 | 1485 | 1490 | 1495 | 1500 |

HPFDNKDINHFGLTYQHSRPADVWDYSRYLDYHYDDLHLHGWSIAQLDEVLKKQRSRDRMFAGFMLHHIGTSAVVEIDVCVDTGDGGRSCHHPAGKFSIL

| 1504 | 1510 | 1515 | 1520 | 1525 | 1530 | 1535 | 1540 | 1545 | 1550 | 1555 | 1560 | 1565 | 1570 | 1575 | 1580 | 1585 | 1590 | 1595 | 1600 |

Linker d-e (Meneses)

GGEYEMPFTFDRLYKHDITDAIRKLGLRLDSAADFDLEVKIHSFNGSYIDASLLHPPTIIFQPGKSVRNLSPAERRSDLVRKSVWTLSPAERRSLVLAMK

| 1604 | 1610 | 1615 | 1620 | 1625 | 1630 | 1635 | 1640 | 1645 | 1650 | 1655 | 1660 | 1665 | 1670 | 1675 | 1680 | 1685 | 1690 | 1695 | 1700 |

FU-e (CCHA)

SLQADSSADGFQSLASFHAQPPLCPYPEATKRFACCVHGMATFPAWHRLYTVQFEDALRRHGALVGIPYWDTVVPQRELPPFFNDEVWDDPLFHANFPNP

| 1704 | 1710 | 1715 | 1720 | 1725 | 1730 | 1735 | 1740 | 1745 | 1750 | 1755 | 1760 | 1765 | 1770 | 1775 | 1780 | 1785 | 1790 | 1795 | 1800 |

WHGADIEFNHQKVTRHFDMAELGKKGPKGYDTWSWKQYIYALEQEDYCDFEVQFEIAHNAIHAWWGGNEEHSMGHLHYASYDPIFILHHSNTDRIFALWQ

| 1804 | 1810 | 1815 | 1820 | 1825 | 1830 | 1835 | 1840 | 1845 | 1850 | 1855 | 1860 | 1865 | 1870 | 1875 | 1880 | 1885 | 1890 | 1895 | 1900 |

KLQKFRGHDPNNVNCALELMREPLKPFSFGAPYNLNENTHEHSKPEETLDYHTHFHYDYDSLELQGMDVHRLQDYINQQKVEDRAFAGFLLQGIGTSAHV

| 1904 | 1910 | 1915 | 1920 | 1925 | 1930 | 1935 | 1940 | 1945 | 1950 | 1955 | 1960 | 1965 | 1970 | 1975 | 1980 | 1985 | 1990 | 1995 | 2000 |

Linker e-f (Meneses)

AFSVCKDGGECTSAGFFDVLGGALETPWRFDRLYKYEITDVLKSKDLEVRDAFHIEVTITASNGSALPSSLIPEPSVIYAPKSRKLSGRDIQPNRVRHDL

| 2004 | 2010 | 2015 | 2020 | 2025 | 2030 | 2035 | 2040 | 2045 | 2050 | 2055 | 2060 | 2065 | 2070 | 2075 | 2080 | 2085 | 2090 | 2095 | 2100 |

FU-f (CCHA)

ADLTDRDAKNLKSAIRDLQNDNSKDGYQHIASFHGAPALCPDPNAATHACCIHGNPTFPHVHRLYTVEMEDALIRHGSGVALPYWDVMKPITELPELFTS

| 2104 | 2110 | 2115 | 2120 | 2125 | 2130 | 2135 | 2140 | 2145 | 2150 | 2155 | 2160 | 2165 | 2170 | 2175 | 2180 | 2185 | 2190 | 2195 | 2200 |

ESYYDPWRDEVVPNPFVRAYIKDTDGYTVRDPRPELTKLSHDGKHSILFEEVLLALEQTDYCDFEVQYEVTHNAIHYLVGGRQLYSLSSLEYSSYDPIFF

| 2204 | 2210 | 2215 | 2220 | 2225 | 2230 | 2235 | 2240 | 2245 | 2250 | 2255 | 2260 | 2265 | 2270 | 2275 | 2280 | 2285 | 2290 | 2295 | 2300 |

VHHSFIDKIWTVWQELQKRRGLVYDRADCAVNFMQHKMHPFDMDELNPDVRTREHAHPQTVFDYEDLGYHYDDFKLGGNTLEELEELIHERKSHPRVFAG

| 2304 | 2310 | 2315 | 2320 | 2325 | 2330 | 2335 | 2340 | 2345 | 2350 | 2355 | 2360 | 2365 | 2370 | 2375 | 2380 | 2385 | 2390 | 2395 | 2400 |

Linker f-g (RVH1)

FHLQSIGTSADILFHICKTEDACHHAGGIFVLGGALEMPWAYDRLYKYDITEDLHDLGIEPEDVFHAQEPFHLKYEIHAVNGTVLPPSTIHAPTLIYKPG

| 2404 | 2410 | 2415 | 2420 | 2425 | 2430 | 2435 | 2440 | 2445 | 2450 | 2455 | 2460 | 2465 | 2470 | 2475 | 2480 | 2485 | 2490 | 2495 | 2500 |

FU-g (CCHA)

EGAINTLTDAEMENVGVRKDINTLTAAETTNLRDALRRVQAGTGRMTYDFIAGAHGYPAECKMGEYDVACCQHGMASFPGWHRVFTRQMEIALSWEGAKV

| 2504 | 2510 | 2515 | 2520 | 2525 | 2530 | 2535 | 2540 | 2545 | 2550 | 2555 | 2560 | 2565 | 2570 | 2575 | 2580 | 2585 | 2590 | 2595 | 2600 |

GLPYWDWTEAFTELPTLVSQEHDNPFHHGHIPGKAENITTTRAPRPQLFKDPEHGEESFFFRQALLAFEQRDFCDFEVQFEVLHNALHSWIGGTSPYGMS
2604 2610 2615 2620 2625 2630 2635 2640 2645 2650 2655 2660 2665 2670 2675 2680 2685 2690 2695 2700

TLEYAAYDPIFFIHHSNVDRQFAIWQELQKHRGLDYNTANCHIQDLRKPLEPFNRANNPVLVTRVHSRAIDAFNYDQYGYQYDHLHFHGLTVDKLDEKLE
2704 2710 2715 2720 2725 2730 2735 2740 2745 2750 2755 2760 2765 2770 2775 2780 2785 2790 2795 2800

Linker g-h (Manubens)

KRKEQDRVFLNFMLRGIKMSADVVFDLCNAQGTCNFAGTFAILGGPLEMPWNFDRVFKYDVTKIFQQMRLRPDSNYTIPIRIRAVNGMQLDPNLLEPPSV
2804 2810 2815 2820 2825 2830 2835 2840 2845 2850 2855 2860 2865 2870 2875 2880 2885 2890 2895 2900

FU-h (CCHA)

TFVPGKK   DLTRHDLTSLSIPQISNLADALYKLQNDHGPNGFEAIASYHGAPGLCPENATEHYSCCQHGMPAFPHMHRLLTVQFEHALKEKGAVIGVPY
2904 2910   2935 2940 2945 2950 2955 2960 2965 2970 2975 2980 2985 2990 2995 3000 3005 3010 3015

WDWTRPVKGMPTLFTDNSEINPFHTYLMSFNNHHIERQVQEEMFTHSAEGDPESLFHQALELLEETNYCDFEVQYEMLHNAVHELVGGSHTYSMSTLEYS
3020 3025 3030 3035 3040 3045 3050 3055 3060 3065 3070 3075 3080 3085 3090 3095 3100 3105 3110 3115

AFDPFFMVHHSSIDRIWQIWQTLQKLRHKPFNYAVCAARSMYRPLEPFSYESINPDPLTRENSKPAQIFDSHKFHYHYDNLNLNGHSVMELNTMIHAMQA
3120 3125 3130 3135 3140 3145 3150 3155 3160 3165 3170 3175 3180 3185 3190 3195 3200 3205 3210 3215

QSRTFAGFVLSGIGTSARVHVDIFKGEETAAVGNFFVLGGSTEMPWAYERIYKLDMTEAAKKLGLSGSSHFDFKLTVTKYDGTALDAKFPDPVIVKRRTN
3220 3225 3230 3235 3240 3245 3250 3255 3260 3265 3270 3275 3280 3285 3290 3295 3300 3305 3310 3315

AEFDELILPIRKENKLPPKIVVRRGTRVLFHPTEEGIAGPIRELGSFTNSEHCAIPPGHAHAYELDVLHVLEPGDYYFVSNNLEACKAGSRVQISVDEE
3320 3325 3330 3335 3340 3345 3350 3355 3360 3365 3370 3375 3380 3385 3390 3395 3400 3405 3410 3415
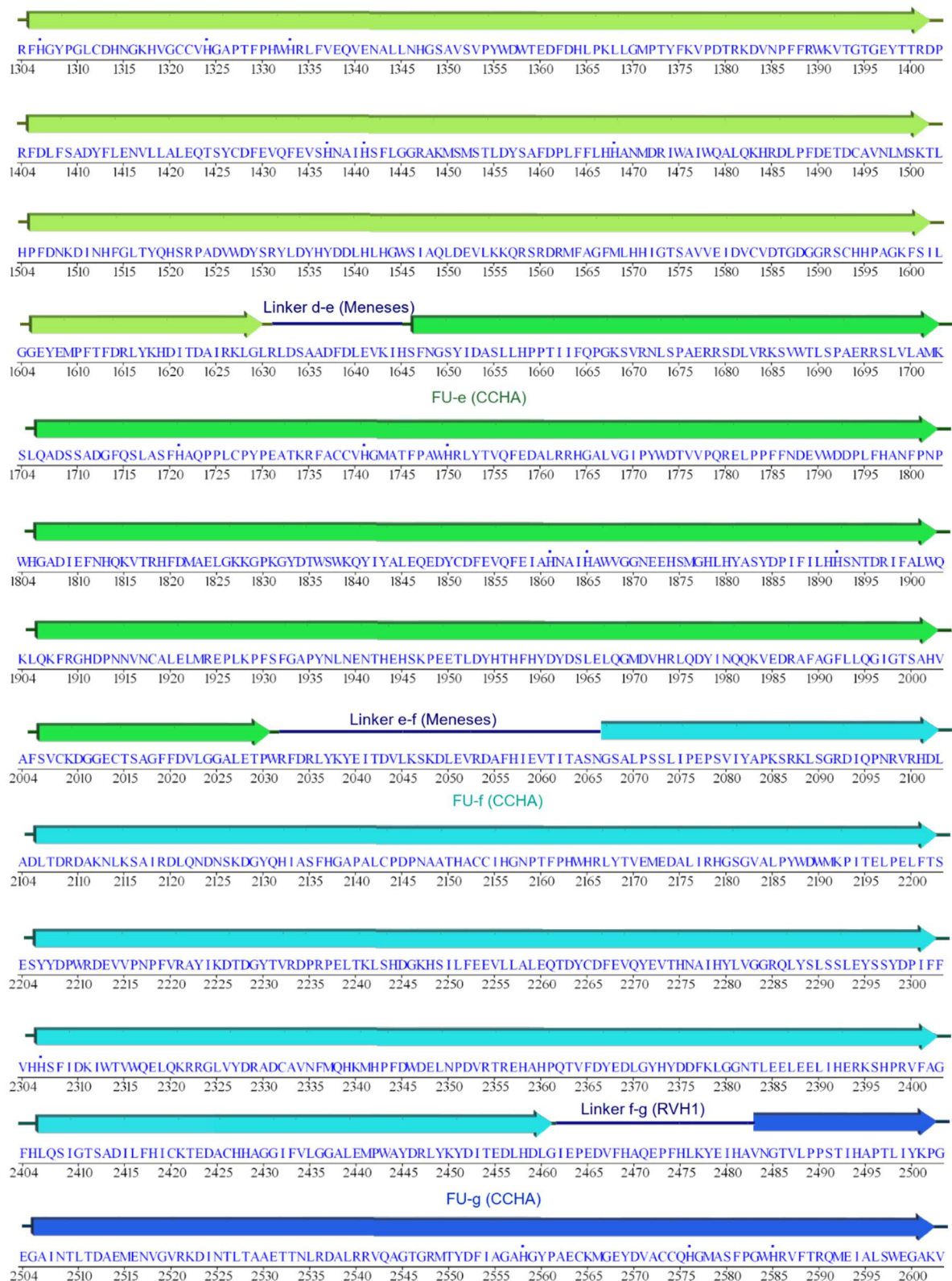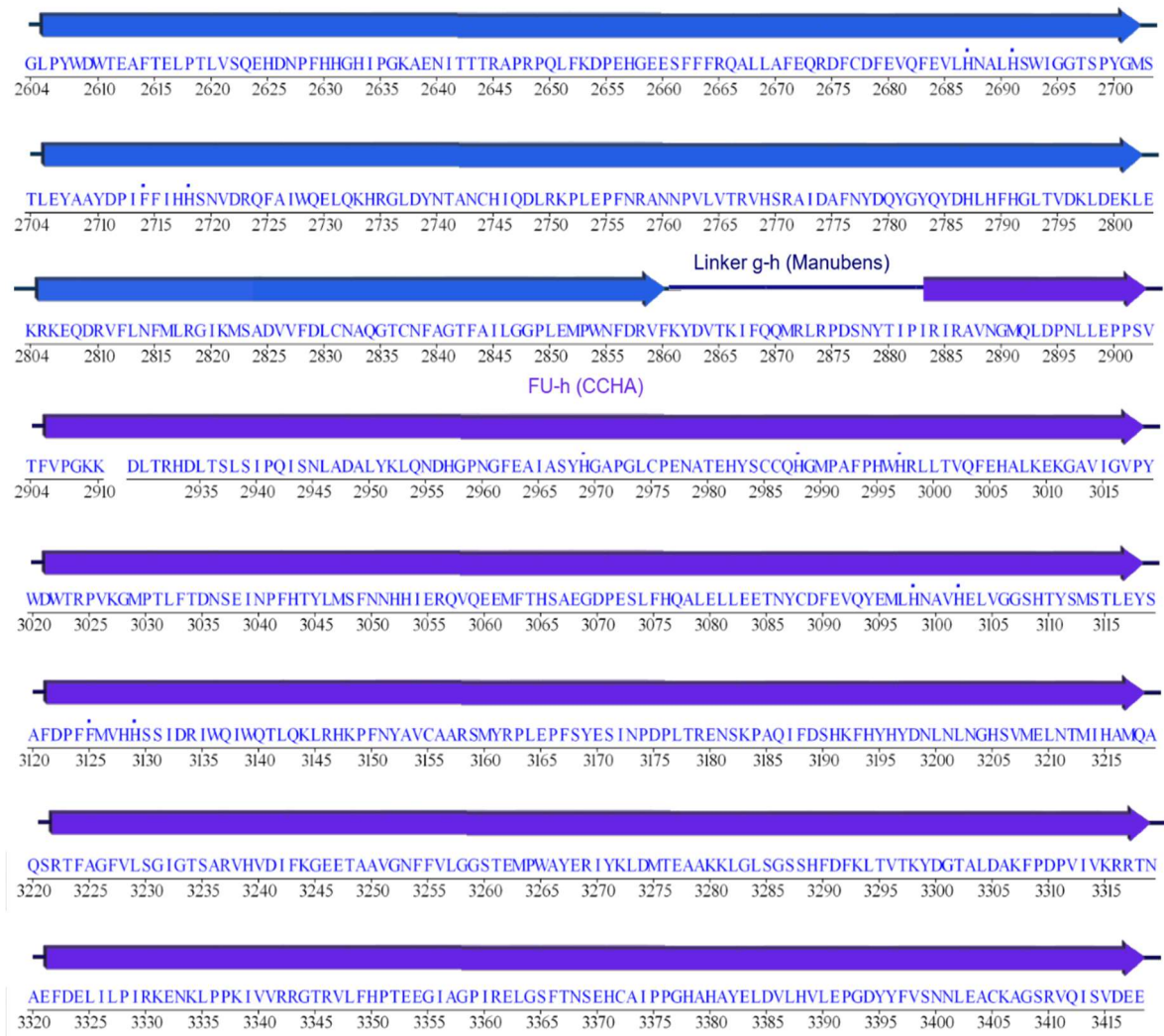
The FU and linkers regions, from CCHB, are enumerated on the following order:

- FU-a: 1-407
- linker a-b: 408-421
- FU-b: 422-821
- linker b-c: 822-835
- FU-c: 836-1229
- linker c-d: 1230-1242
- FU-d: 1243-1631
- linker d-e: 1632-1645

- FU-e: 1646-2031
- linker e-f: 2032-2066
- FU-f : 2067-2460
- linker f-g: 2461-2483
- FU-g: 2484-2860
- linker g-h: 2861-2883
- FU-h: 2884-3418

3.