U.PORTO
FACULDADE DE CIÊNCIAS
UNIVERSIDADE DO PORTO

U.PORTO
INSTITUTO DE CIÊNCIAS
BIOMÉDICAS ABEL SALAZAR
UNIVERSIDADE DO PORTO

MSc

2.º
CICLO

FCUP
ICBAS
2021

U.PORTO

Development of an antidote for Russell's viper (*Daboia russelii russelii*) venom factor X activating enzyme (RVV-X)

Juliana de Castro Amorim

FC

U.PORTO
FACULDADE DE CIÊNCIAS
UNIVERSIDADE DO PORTO

U.PORTO
INSTITUTO DE CIÊNCIAS
BIOMÉDICAS ABEL SALAZAR
UNIVERSIDADE DO PORTO

Development of an antidote for Russell's viper (*Daboia russelii russelii*) venom factor X activating enzyme (RVV-X)

Juliana de Castro Amorim

Dissertação de Mestrado em Bioquímica apresentada à Faculdade de Ciências da Universidade do Porto (FCUP) e ao Instituto de Ciências Biomédicas Abel Salazar (ICBAS)

2021

Development of an antidote for Russell's viper (*Daboia russelii russelii*) venom factor X activating enzyme (RVV-X)

Juliana de Castro Amorim

Mestrado em Bioquímica
Departamento de Química e Bioquímica
2021

Orientadora
Doutora Ana Luísa Novo de Oliveira, LAQV-REQUIMTE, Laboratório de Bioquímica Computacional, Departamento de Química e Bioquímica, Faculdade de Ciências da Universidade do Porto

Coorientador
Professor Doutor Pedro Alexandrino Fernandes, Laboratório de Bioquímica Computacional, Departamento de Química e Bioquímica, Faculdade de Ciências da Universidade do Porto

U.PORTO

FACULDADE DE CIÊNCIAS
UNIVERSIDADE DO PORTO

U.PORTO

INSTITUTO DE CIÊNCIAS
BIOMÉDICAS ABEL SALAZAR
UNIVERSIDADE DO PORTO

Todas as correções determinadas pelo júri, e só essas, foram efetuadas.
O Presidente do Júri,

Porto, ______/______/_________

Acknowledgments

In the first place, I would like to thank my supervisor, Dra. Ana Oliveira, for her guidance, support, motivational words, patience, and availability that made this whole journey even more exciting. I am deeply grateful.

I am very thankful to Prof. Pedro Fernandes, who, together with Prof. Maria João Ramos, welcomed me into the research team and provided me excellent guidance, discussion, and support.

To Óscar Passos, for his sympathy and for providing the necessary softwares and functionalities.

I am also grateful to Rui Neves, for his enormous patience and teachings about QM/MM calculations that enriched my work.

To all members of the laboratory for the sympathy as well as excellent work environment that directly or indirectly have provided me.

To my flatmates, Ana Moura, Alexandra Simões, Erika Loupy, Ani Gabelia, for all the support, for the moments of joy, for the deep conversations and above all for being great housemates during this journey. To Bianca Rodrigues, Rui Silva, Cândido Xavier and Sara Rodrigues for all their support and for keeping their friendship throughout these years. To my master colleagues, for having welcomed me and immediately integrated me when I didn't know anyone. To Afonso Oliveira, the best person I could ever have by my side during this cycle, with whom I have shared moments of deep anxiety but also of big excitement. To Carlos Oliveira, Carla Costa, Andreia Neves, Bruna Fernandes and Mariana Teixeira for being my favourite Valencianos, my best friends, for always putting a smile on my face and for being always present. Thank you all for always have supported, encouraged, and believed in me.

Finally, a special thanks to my family, my biggest fans, for being my pillars and examples of strength and dedication.

Last but not least, my cats, Brrnirr and Rita, for giving me company and relaxing me with their purr.

Abstract

Snakebite envenoming is a major cause of mortality and disability worldwide. This work focuses on the *Daboia russelii* viper, one of the most dangerous and deadliest snakes in the Indian subcontinent. Its venom contains more than 10 families of toxins, and one of the main proteolytic enzymes that contributes the most to its toxicity is Russell's viper venom factor X activator (RVV-X). RVV-X activates the coagulation factor X (FX) leading to abnormal blood clotting and consequently, death. Despite vipers' impact, to date, there are no potent antidotes in the remote areas of the Indian subcontinent and Southeast Asia, highlighting the urgent need of a deep understanding on metalloproteinase's mechanisms.

Overall, structural information of snake proteins is limited. Thus, the 3D models of RVV-X and FX were built using homology modeling. The crystallographic structure of the *D. Siamensis* RVV-X was used as a template (PDB ID: 2E3X). While the FX zymogen model was constructed using the structures of activated FX (PDB ID: 1XKA, 1IOD, 2JKH) and chymotrypsinogen (PDB ID: 2CGA). To form the complex and understand how the two molecules interact, protein-peptide docking was performed using the HPepDock webserver. The top two solutions which corresponded to opposite orientations of the substrate were selected. To evaluate which was the most suitable, the two models were subjected to molecular dynamics simulations. Results indicated that the best model corresponds to the pose where the orientation of the substrate is in accordance with an earlier proposal by Takeda et al. in 2007. In the chosen model, the backbone of the FX Ile-Arg residues is correctly oriented to the cleavage process. Finally, this model was used for the study of the RVV-X hydrolytic mechanism using the two-layered ONIOM model (B3LYP/6-31G(d):AMBER) method. The first and rate-limiting step of the reaction corresponds to the nucleophilic attack of a water molecule on the substrate carbonyl with an activation barrier of 14.4 Kcal.mol^{-1}. Follows a rearrangement of the H-bond network, proton transfer to the peptide nitrogen and C-N bond cleavage with an energy barrier of 8.0 Kcal.mol^{-1}.

Moreover, this study also aimed to find novel metalloprotease inhibitors that could in a future act as antivenom drugs. To select 50 potential metalloprotease inhibitors, two molecular screening approaches were carried out: (1) ligand-based (LBVS) and (2) structure-based (SBVS). The first method returned 338836 commercially available compounds. These compounds were characterized by molecular descriptors, such as Lipinsky's rules and clustered by chemical similarity. Of these, 20 were selected for experimental evaluation. In the SBVS method, a chemical library of 3.5 million compounds was explored and, following the same procedure, another 30 compounds were selected. All compounds will be tested in vitro at the Professor Ashis Mukherjee laboratory in the University of Tezpur, India.

Keywords: Snakebite, Russell’s viper, RVV-X, FX, Homology modelling, SBVS, LBVS, QMMM, ONIOM

Resumo

O envenenamento por mordida de cobra é uma das principais causas de morte e incapacidades em todo o mundo. Este trabalho foca-se na víbora *Daboia russelii*, uma das cobras mais perigosas e mortíferas do subcontinente indiano. O seu veneno contém mais de 10 famílias de toxinas, sendo que uma das enzimas proteolíticas principais e que mais contribui para a sua toxicidade é a enzima activadora do factor de coagulação X (RVV-X). Esta enzima activa o factor de coagulação X (FX) desencadeando alterações no funcionamento da cascata de coagulação, e, consequentemente, morte. Apesar do impacto destas víboras, até à data, ainda não se desenvolveram antídotos suficientemente potentes, e de fácil acesso em regiões remotas do subcontinente Indiano e do Sudeste Asiático, sublinhando a necessidade urgente de aprofundar o conhecimento dos mecanismos das metaloproteases. Em geral, a informação estrutural sobre as proteínas presentes no veneno de cobra é ainda muito limitada. Assim, os modelos 3D das proteínas RVV-X e FX foram construídos recorrendo à técnica de modelação por homologia. A estrutura cristalográfica da RVV-X da *D. Siamensis* foi usada como template (PDB ID: 2E3X). Enquanto que o modelo do FX foi construído usando as estruturas do FX activo (PDB ID: 1XKA, 1IOD, 2JKH) e do quimotripsinogénio (PDB ID: 2CGA). Para formar o complexo e compreender como as duas moléculas interagem entre si, realizou-se docking proteína-péptido recorrendo ao servidor online HPepDock. As duas poses mais pontuadas que correspondem a orientações opostas do substrato foram selecionadas. Para avaliar qual das conformações era mais favorável, os dois modelos foram submetidos a simulações de dinâmica molecular. Os resultados indicaram que o melhor modelo corresponde à pose onde a orientação do substrato está de acordo com a proposta por Takeda et al. em 2007. No modelo construído, o backbone dos resíduos Ile-Arg do FX encontra-se corretamente orientado para o processo de clivagem. Por fim, o modelo escolhido foi usado para o estudo do mecanismo hidrolítico da RVV-X usando o método do modelo ONIOM de duas camadas (B3LYP/6-31G(d):AMBER). A primeira etapa da reação que corresponde ao ataque nucleofílico por uma molécula de água no carbonilo do substrato, apresenta uma barreira de energia de 14.4 Kcal.mol^{-1}. Segue um rearranjo das ligações H, transferência do protão para o nitrogénio do péptido e clivagem da ligação C-N com uma barreira energética de 8.0 Kcal.mol^{-1}.

Além disso, este estudo também teve como objetivo encontrar novos inibidores de metaloprotease que poderiam, no futuro, atuar como antídotos. Para selecionar 50 potenciais inibidores da metaloprotease, levaram-se a cabo duas abordagens de triagem molecular: (1) baseado no ligando (LBVS) e (2) baseado na estrutura (SBVS). O primeiro método retornou 338836 mil compostos disponíveis comercialmente. Estes compostos foram caracterizados por descritores moleculares, tais como as regras de Lipinsky e agrupados por semelhança química. Destes, 20 foram selecionados para avaliação experimental. No método SBVS, explorou-se uma quimioteca de 3.5 milhões de compostos e, seguindo o procedimento supracitado, foram selecionados mais 30 compostos. Todos os compostos serão testados in vitro no laboratório do Professor Ashis Mukherjee da Universidade de Tezpur, Índia

Palavras-chave: Envenenamento, Russell’s viper, RVV-X, FX, Modelação por homologia, SBVS, LBVS, QMMM, ONIOM

Index

List of figures

List of tables

List of abbreviations

3FTx Three-finger toxins

aMD accelerated Molecular Dynamics

ADAM A disintegrin-like and metalloproteinase

AMBER Assisted Model Building with Energy Refinement

AP Activation Peptide

BLAST Basic Local Alignment Search Tool

CRISP Cysteine-rich Secretory Proteins

Cryo-EM Cryogenic electron microscopy

Cys-rich Cysteine-rich domain

DFT Density Functional Theory

Dis-like disintegrin-like domain

D-loop disintegrin loop

DOPE Discrete Optimized Protein Energy

DUD-E Directory of Useful Decoys–Enhanced

EF Enrichment Factor

EGF Epidermal-growth-factor

ES Reactant State

FX Coagulation Factor X

GA Genetic Algorithm

Gla γ-carboxyglutamic acid

GMQE Global Model Quality Estimation

HADDOCK High Ambiguity Driven DOCKing proteins

HBA Hydrogen Bond Acceptor

HBD Hydrogen Bond Donor

HC Heavy chain

HL High-layer

HPEPDOCK Hierarchical flexible Peptide Docking

HVR Hypervariable Region

INT Intermediate state

IRC Intrinsic Reaction Coordinate

LBVS Ligand-based Virtual Screening

LC Light Chain

LL Low-layer

M metalloproteinase domain

MCPB Metal Center Parameter Builder

MM Molecular Mechanics

MP Metalloproteinase

NMR Nuclear Magnetic Resonance

NTD Neglected Tropical Disease

ONIOM Our Own N-layer Integrated molecular Orbital molecular Mechanics
PLA$_2$ Phospholipase A$_2$
PDB Protein Data Bank
PDBe Protein Data Bank in Europe
PES Potential Energy Surface
ProSA Protein Structure Analysis
QM Quantum Mechanics
QM/MM Quantum Mechanics/Molecular mechanics
RESP Restricted Electrostatic Potential
RMSd Root-mean square deviation
RMSF Root-mean-square Fluctuation
RVV Russell's Viper Venom
RVV-X Russell's viper venom factor X activator
SANDER Simulated Annealing with NMR-Derived Energy Restraints
SBVS Structure-based Virtual Screening
sMD steered Molecular Dynamics
SMTL Swiss Model Library of Templates
SP Serine Protease
SVMP Snake Venom Metalloproteinase
SVSP Snake Venom Serine Protease
Snaclec Snake C-type lectin
TS Transition state
TIP3P Transferable Intermolecular Potencial 3P
UniProtKB AC UniProt Knowledgebase Accession Code
VICC Venom-induced Consumption Coagulopathy
VS Virtual Screening
VMD Visual Molecular Dynamics
WHO World Health Organization
X-ray X-Ray Crystallography

Chapter I - Introduction

1.1. Background

1.1.1. Snake venom as an evolutionary innovation

Since prehistory, the behaviour of venomous animals has been increasing both curiosity as well as fascination in mankind. As soon as contact between humans and these organisms began to be correlated with illness or even death, this has become an active research field [1, 2].

Venom genes evolved independently over millions of years from the Toxicofera common ancestor (includes all snakes and some lizards) [3]. These are thought to have arisen from physiological protein-encoding genes that have evolved by a myriad of processes, counting with gene duplication, recombination, alternative splicing, among others. However, recruitment of a copy into the venom gland and duplication events have been recognized as crucial steps to protein neofunctionalization in the form of a novel toxic activity [3, 4].

Many toxins are encoded by large multigene families from which they were convergently recruited in several animal lineages, implying that there are structural and functional constraints on the recruitment and evolution of animal venoms. Each family share common evolutionary origins but diverge on their pharmacological activities. Examples of those are phospholipases (PLAs), metalloproteases (MPs), serine proteases (SPs), cysteine-rich secretory proteins (CRISPs), three-finger toxins (3FTxs), among others [5, 6]. Their folding is preserved despite mutations at key surface residues which lead to different biological targets [7]. Toxin diversity as well as multifunctionality contributed as an adaptive advantage that led to the diversification of several animal lineages [1].

Across the world, thousands of venomous animal groups have been identified, including platypuses, snakes, scorpions, lizards, fishes, spiders, insects, centipedes, cone snails, and sea anemones. On a general basis, these species are capable of producing venoms with distinct complexity consistent with their ecological needs, such as predation and defense [5, 8]. The venom system is dynamic, differing both interspecifically and intraspecifically due to a diverse array of factors, including age, gender, geographic distribution, prey availability (diet), and season [1]. However, most venom research has been applied to investigate human medical targets, as is the case of snakes that are one of the major human killers worldwide [4].

Snake venom consists of a highly complex mixture of toxic and pharmacologically active enzymes and peptides, as well as some non-enzymatic compounds [5, 8, 9]. It is produced in specialized secretory glands located above the upper bone of the jaw on both sides of the head. These are connected to venom ducts that transfer the secreted venom along the inside of the fangs, which in turn inoculate it on the target through a wound infliction [10]. Upon injection, venom toxins interact with functionally relevant receptors, ion channels and enzymes disrupting physiological and biochemical processes in the envenomed prey [5, 11].

In many cases, venom represents a key innovation that evolved in conformity with changes in biotic and abiotic factors. It allowed, for instance, in snakes, to replace its mechanical strategies (constriction/suffocation), hitherto prevalent [5, 11]. Thus, the venom itself functions as a natural weapon for prey subjugation and digestion and, also as a defensive trait against predators and aggressors [5, 8, 9].

Some of the important snake venom toxins, awakened in medical-scientific community a great interest for the development of effective antivenoms and new drugs [12]. However, venom composition variations is a major challenge to make efficient snakebite treatments [13]. Therefore, snake venom complexity needs to be unravelled urgently [14].

1.1.2. Proteomic profile and its pharmacological effects

Recently, several studies of venom glands and snake venom have been carried out at the level of transcriptomics and proteomics, respectively. As a consequence, it became possible to achieve a deeper understanding about snake venoms [10].

Snake venoms are mostly made up of hydrolytic enzymes that act on biological molecules, such as proteins and phospholipids, breaking them down. In addition to their contribution to the digestion of prey, several of these enzymes exhibit pharmacological effects in the envenomed victim [8]. The scope of pathological and pathophysiological manifestations inflicted on the prey is very wide, involving local and systemic alterations that can be life-threatening [15].

There are four big families in snakes: Elapidae, Viperidae, Colubridae, and Atractaspididae. The first two are the main medically important groups of snakes, as they are responsible for the majority of the highly toxic human envenomings [16]. The venoms of the Elapids are rich in neurotoxins, like 3FTx, that exert their effects at the neuromuscular system quickly immobilizing the victim. Whereas viper venoms typically encompass numerous hemotoxins (haemorrhage and coagulopathy effects) and cytotoxins (tissue damage) to kill prey and aid in its digestion [17].

Vipers venom effects tend to occur more intensely at the site of the bite (locally), leading to necrosis, edema, and bleeding, which are mostly associated with the action of two proteases: snake venom metalloproteinases (SVMPs) and snake venom serine proteases (SVSPs). In contrast, the systemic effects, lead to coagulopathies culminating in spontaneous haemorrhage which is predominantly attributed to the synergistic action between SVMPs, SVSPs and PLA_2s (Fig. 1) [17-19].

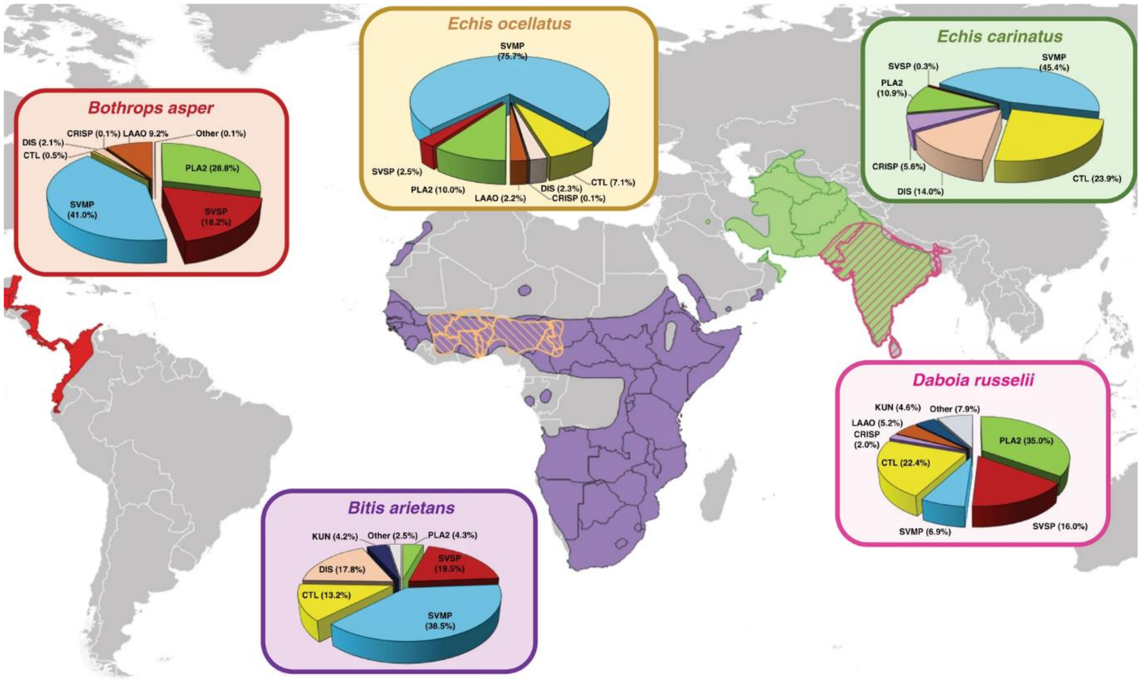


Figure 1 – Geographical distribution and venom proteomic profile of the world's deadliest vipers: *Echis ocellatus* (Nigeria), *Echis carinatus* (India), *Bothrops asper* (Costa Rica), *Bitis arietans* (Nigeria) and *Daboia russelii* (Sri Lanka). Overall dominant toxin families SVMP, SVSP, and PLA_2; Adapted from [16].

Typically, venom proteomes are dominated (> 60%) by the aforementioned enzymatic families [16]. From the proteomics analysis of viperid venoms reported to date, the most abundant enzymes are the SVMPs (*Fig. 1*), suggesting that they play a key role in pathologies [5, 10]. The effects caused by these enzymes are the main contributors to the high morbidity rates observed in snakebite envenoming [20].

1.1.3. The global snakebite envenoming burden

Snakebite envenoming represents a public health hazard that affects mostly rural populations in underdeveloped countries. Tropical and subtropical regions of the world, like Asia, sub-Saharan Africa, Latin America and parts of Oceania suffer the biggest impact (Fig. 2) [2, 17]. Despite being globally distributed, the incidence of snakebite envenoming in India is the highest when compared to any other country (*Fig. 2*), accounting for nearly half of global snakebite deaths [14]. Here, most of the snakebites, around 90%, are inflicted by the saw scaled viper (*Echis carinatus*) and Russell's viper (*Daboia russelii*). The latter, which will be further discussed, is one of main responsible for the high rates of morbidities and fatal envenomings in the Indian subcontinent, representing a serious medical threat to mankind [15].

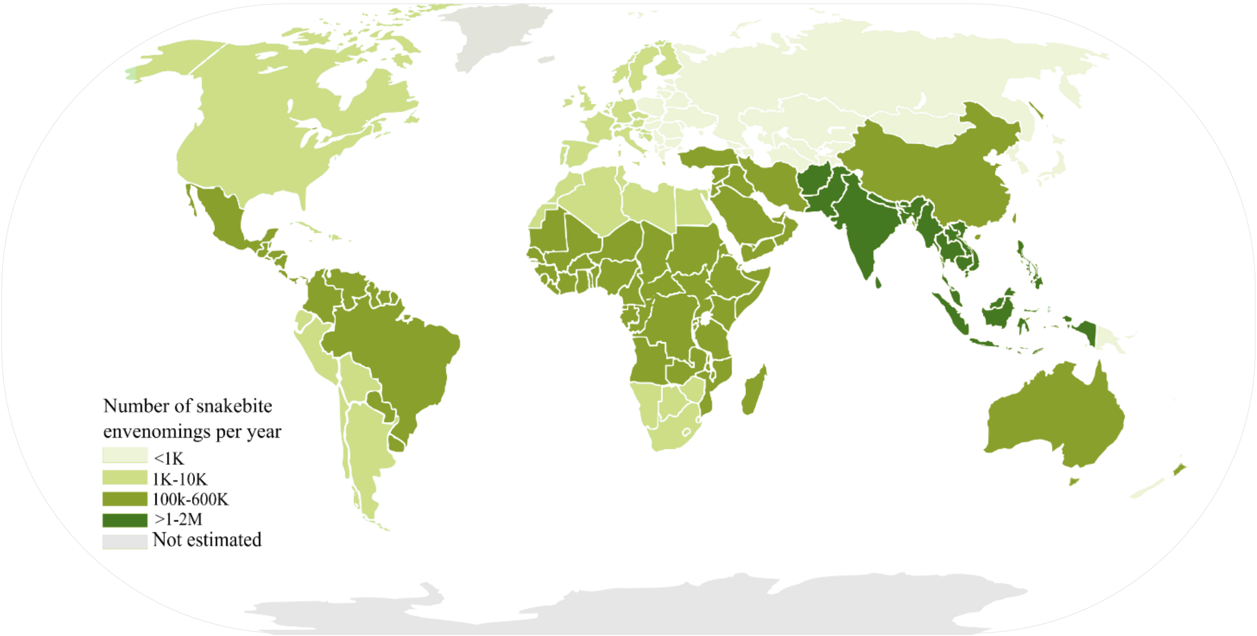


Figure 2 - Map showing an approximate number of snakebite envenomings per year at a global scale. The highest number occurs in the Indian subcontinent followed by Latin America, Oceania, and Sub-Saharan Africa [21].

The problem becomes even more serious because, in these areas, the access to health services is reduced and treatments — antibody-based antivenoms — are of low availability and affordability to those who need them. Moreover, most victims use traditional treatments at home leading to a significant, but unknown, number of morbidities and sequelae which have serious repercussions in the victim's life. Examples include severe tissue damage, loss of limbs, renal failure, visual impairment, thrombosis, and neurological damage [12, 18, 19].

The problem might be even bigger because epidemiological data on snakebite envenoming is fragmentary and presumably underestimated. This is a cause of great concern, however, despite its serious impacts, it has been neglected by global health authorities and government agencies [12, 22]. Following several requests, in 2017, the World Health Organization (WHO) recognized snakebite envenoming as a prior neglected tropical disease (NTD), the one that leads to the larger number of annual deaths (*Fig. 3*). Hereinafter, in May 2019, a work plan was launched by the WHO, outlining the goal of halving mortality and disabilities caused by envenomings by 2030 [23].

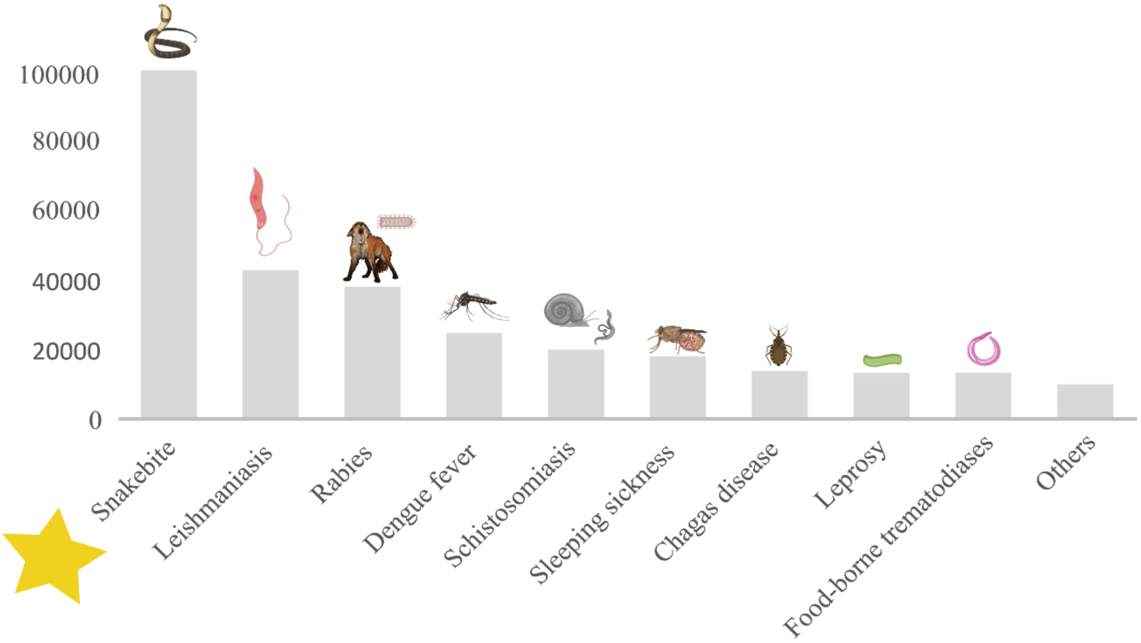


Figure 3 - Annual deaths worldwide from neglected tropical diseases (Snakebite envenoming, leishmaniasis, rabies, dengue fever, schistosomiasis, sleeping sickness, chagas disease, leprosy, food-born trematodiases, among others) [22].

To this end, the action plan consists of sensitizing communities increasing awareness, strengthening rural health care systems, and enhance the supply of affordable, safe and effective treatments. This plan promotes the research and development of small molecule inhibitors that could be ministered outside a hospital settlement [23].

Despite all the recent efforts, venom extraction is difficult, and the amount of venom obtained is very small which limits the discovery of drugs and treatments. Thus, venom research has become even more focused on snakes that are most dangerous for humans [13].

1.2. Russell’s viper (*Daboia russelii russelii*)

1.2.1. Geographic distribution and habitat

Russell's viper (Viperidae family) is a widely distributed snake of public health importance in the Indian subcontinent and Southeast Asia, including Sri Lanka, Pakistan, Myanmar, southern China and Taiwan. It is abundant in the southern, western, and eastern states of India but very rare in the Ganges Valley, Northern Bengal and Assam. It is reported as the principal cause of snakebite-induced morbidity and mortality in these regions [15, 24, 25]. Despite being widespread, to date, the knowledge on the influence of the biogeographic conditions on the venom proteome and potency is still very limited, posing great challenges regarding treatments [26].

India has more than 60 species of venomous snakes. *Daboia russelii*, *Naja naja* (Spectacled Cobra), *Bungarus caeruleus* (Common Krait) and *Echis carinatus* (Saw-scaled viper) are the four deadliest venomous snake species and thus are designated as “The Big Four”. However, Russell’s Viper is a major source of snakebites that occur on the vicinity of the agricultural lands and villages where its primary dietary preference, rats and mice, grow. This is a serious issue for rice cultivators and farmers that work in their habitat, and therefore are its major victims [12, 15, 27].

1.2.2. Morphological features

The size of this snake varies from medium to large, averaging 120 cm, however it can grow up to 180 cm. The head is flat and triangular, but distinct from the neck. The snout is blunt and rounded and the body colour is typically yellowish to brown [12]. The dorsal part of its body consists of several patterns of brownish colour, with three series of dark, round, and oblong spots with black and white edges around it that run the length of the body (*Fig. 4*). These characteristics allow the identification of these vipers easily. For example, in the west of India the lower part is white, in south-eastern India it is partially spotted and deeply spotted in the northeast [12].

Figure 4 - Russell's viper (Daboia russelii).

Russell’s viper was classified into five subspecies based on its morphological characteristics: *D. r. russelii* (India, Pakistan, Nepal, and Bangladesh), *D. r. pulchella* (Sri Lanka), *D. r. siamensis* (Thailand, Myanmar and China), *D. r. formosensis* (Taiwan), and *D. r. limitis* (Indonesia). However, phylogenetic analysis attributed to morphological characteristics and mitochondrial DNA, classified Russell’s Vipers into two distinct species: *D. russelii* (from the South of Asia) and *D. siamensis* (from southeast Asia, southern China, Indonesia, and Taiwan) [12].

1.2.3. Venom and its pharmacological effects

Russell's viper venom (RVV) is a potent cocktail of several toxic and non-toxic components that intervene in vital physiological processes leading to haemostatic disturbances [27]. The amount of venom produced by this snake is considerable, and for adult specimens it can vary from 130 to 250 mg. In mice, their Median Lethal Dose (LD_{50}) is 0.13 mg/kg, 0.40 mg/kg and 0.75 mg/kg, for intravenous, intraperitoneal and subcutaneous administration, respectively [28].

Among the major families of proteins found in RV venom there are PLA2s and SVMPs. There are also other toxins that contribute to venom-induced pathophysiological effects, namely the SVSPs, C-type lectins and L-amino oxidases that together with PLA2s and

SVMPs constitute over 90% of its proteome. These enzymes are thought to be responsible for several local and systemic clinical manifestations, including myonecrosis, edema, neurotoxicity, renal failure, and hypotension [15, 25, 28]. These clinical manifestations rely both on the qualitative and quantitative venom composition leading to variations in different geographical regions [15, 24]. For example, in some geographic regions, RV venom exhibited potent procoagulant effects, whereas in others it exhibited potent anticoagulant activities. These anticoagulant effects are mainly caused by coagulopathies such as venom-induced consumption coagulopathy (VICC). Here, blood coagulation factors are consumed due to their continuous activation by the action of the procoagulant toxins such as the SVMP RVV-X. Consequently, in serious situations, it can culminate in life-threatening haemorrhage [26].

Although rare, damage to the central nervous system has been reported in southern India, leading to ischemic stroke and anterior pituitary infarction [28] . However, the main cause of death after Russell’s viper bite is acute renal failure, whose mechanism is not yet well known [29].

This way, Russell’s viper was assigned as a category-I medically important snake in the Indian subcontinent [12, 28].

1.2.3.1. Snake venom metalloproteinases (SVMPs)

SVMPs are zinc-dependent metalloproteinases that evolved from mammalian ADAM (A disintegrin-like and metalloproteinase) proteins, before the radiation of the advanced snakes (approximately 60 million years ago) [30]. Both ADAMs and the thrombospondin motif (ADAMTs) constitute the adamalysin/reprolysin subfamily of zinc metalloproteinases, which, in turn, belong to the metzincin superfamily. They have a modular structure which is homologous to the metalloproteinase (M), disintegrin-like (Dis-like), and cysteine-rich (Cys-rich) domains of the membrane-anchored ADAMs [5, 31]. SVMPs are characterized by a catalytic consensus H139EXXH143GXXH149 sequence, which coordinate the zinc metal centre. This catalytic zinc-ion is tetrahedrally coordinated by the Nε_2 atoms of three conserved histidine residues (His139, His143 and His149) and a water molecule that plays the role of the fourth ligand anchored to Glu140. The Glu140 acts as a catalytic base at the bottom of the catalytic cleft [32]. Some residues downstream the consensus sequence, where a conserved Met folds into the so-called Met-turn (loop with methionine residue) and forms a hydrophobic base underneath the three zinc-binding imidazole rings [6, 31].

The evolution of the SVMP multigenic family resulted from several evolutionary adaptation processes. It started with the recruitment of an ADAM (ADAM7 or ADAM28), followed by the loss of C-terminal domains, gene duplication, positive selection and neofunctionalization. Such processes resulted in the formation of class III (P-III) SVMPs, followed by classes II, I, and respective subclasses, contributing to the significant level of diversity on substrate specific proteolytic activity [32].

1.2.3.1.1. Classification

SVMPs are divided into three classes (P-I, P-II, P-III) and subclasses, depending on their quaternary structure and post-translation modifications, respectively (*Fig. 5*) [6, 32, 33].

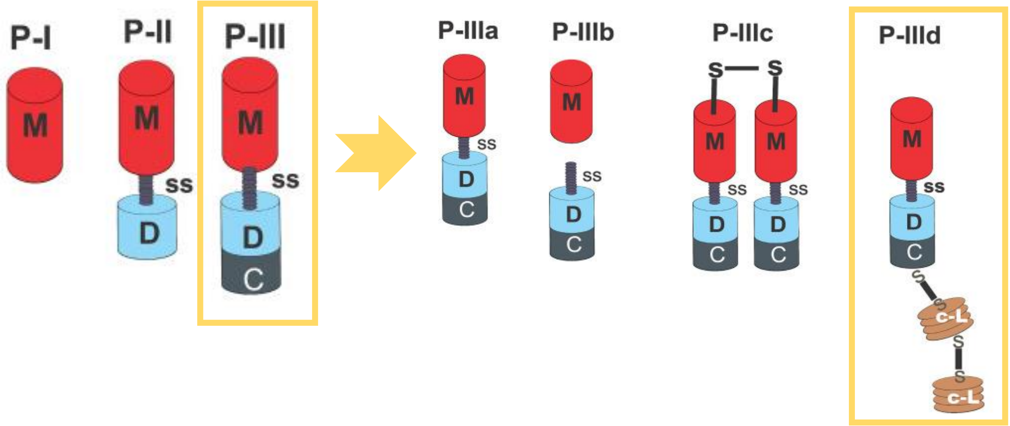


Figure 5 – The three classes of SVMPs, P-I, P-II and P-III and P-III subclasses: a, b, c and d. M: M domain; D: Dis-like domain; C: Cys-rich domain; c-L: snaclecs; P-III is subdivided in P-IIIa, the canonical structure, PIII-b resulting from a proteolytic cleavage, P-IIIc that forms homo- or heterodimers and P-IIId that forms a complex with two covalently bridged Snake C-type lectin-like proteins. Adapted from [30] and edited.

Class III (P-III) SVMPs (60–100 kDa; P-IIIa, P-IIIb, P-IIIc and P-IIId) contains a M domain followed by Dis-like and Cys-rich domains (canonical form – P-IIIa). In this class, there are three subclasses depending on post-translational modifications, such as (i) proteolytic processing of the dis-like and cys-rich domains from the M domain (P-IIIb), (ii) dimerization (P-IIIc), and (iii) complexation (P-IIId). The latter applies to RVV-X, which combines with snake venom C-type lectin-like proteins (snaclecs).

Class II (P-II) SVMPs (30–60 kDa; P-IIa, P-IIb, P-IIc, P-IId and P-IIe), which also are divided in subclasses, diverge from P-IIIs because the enzyme displays a metalloproteinase and a disintegrin domain. Finally, class I (P-I) SVMPs (20–30 kDa) represents the structurally simplest configuration and consists of a single catalytic metalloproteinase domain [5, 6, 30, 33].

1.2.3.1.2. Overall structure

SVMPs have a very typical topology and domain organization, consisting of six α-helices and five stranded β-sheets in the M domain where the zinc ion is localized. These enzymes contain structural motifs essential for catalytic activity. Among them is the above mentioned conserved coordination sequence, the Met-turn and Ca^{2+} binding sites for structural stabilization [31, 34].

The M domain has a hydrophobic pocket (S1' pocket) surrounded by the residues Val136, Ala139, Ile163 and Phe176, which are conserved among several SVMPs. The buried Asp150 is also conserved in the SVMPs and is thought to stabilize the hydrophobic basement of the active site (*Fig. 6*) [35].

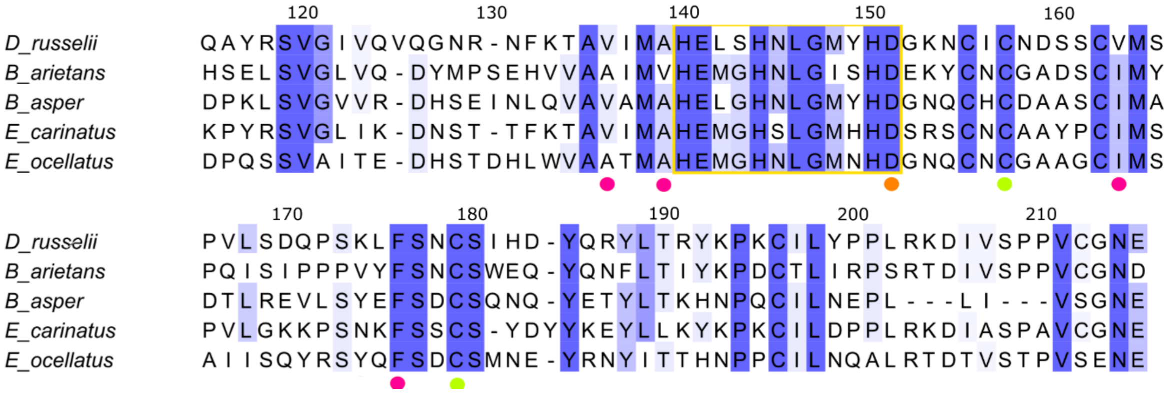

120
130
140
150
160
D_russelii QAYRSVGIVQVQGNR-NFKTAVIMAHELSHNLGMYHDGKNCICNDSSCVMS
B_arietans HSELSVGLVQ-DYMPSEHVVAAIMVHEMGHNLGISHDEKYCNCGADSCIMY
B_asper DPKLSVGVVR-DHSEINLQVAVAMAHELGHNLGMYHDGNQCHCDAASCIMA
E_carinatus KPYRSVGLIK-DNST-TFKTAVIMAHEMGHSLGMHHDSRSCNCAAYPCIMS
E_ocellatus DPQSSVAITE-DHSTDHLWVAATMAHEMGHNLGMNHDGNQCNCGAAGCIMS
170
180
190
200
210
D_russelii PVLSDQPSKLFSNCSIHD-YQRYLTRYKPKCILYPPLRKDIVSPPVCGNE
B_arietans PQISIPPPVYFSNCSWEQ-YQNFLTIYKPDCTLIRPSRTDIVSPPVCGND
B_asper DTLREVLSYEFSDCSQNQ-YETYLTKHNPQCILNEPL---LI---VSGNE
E_carinatus PVLGKKPSNKFSSCS-YDYYKEYLLKYKPKCILDPPLRKDIASPAVCGNE
E_ocellatus AIISQYRSYQFSDCSMNE-YRNYITTHNPPCILNQALRTDTVSTPVSENE

Figure 6 – Sequence alignment of several SVMP sequences: *Daboia russelii* (accession number: K9JAW0), *Bitis arietans* (P0DM97), *Bothrops asper* (Q072L5), *Echis carinatus* (E9JG28) and *Echis ocellatus* (Q14FJ4). Residues with a percentage of conservation above 30% when compared to each other are coloured in dark (the most conserved) and pale blue (the least conserved). Yellow box corresponds to the zinc structural motif; Pink circles correspond to the residues that constitute the hydrophobic pocket (Phe, Val, Ala, Ile); Orange circle corresponds to the conserved buried Asp; Green circles correspond to the disulfide-bridge. Image made using Jalview software.

However, depending on the composition of residues that line this pocket and the presence of disulfide-bridges, this pocket may suffer variations in depth (shallow, intermediate, or deep). These structural features may, in turn, influence the specificity of the cleavage, as well as the binding of possible inhibitors [35].

Non-metalloproteinase domains are also important and may enhance catalytic activity, as is the case of the dis-like and cys-rich domains [5, 32]. These are involved in target selectivity, influencing the specificity of the substrate [35].

The dis-like domain contains Ca^{2+} binding sites and a disintegrin loop (D-loop) (*Fig. 7*), whose residues are highly mobile and accessible to the solvent. The cys-rich domain, on the other hand, include a hypervariable region (HVR) that exhibits variability between SVMPs and mediates protein-protein interactions [30, 34].

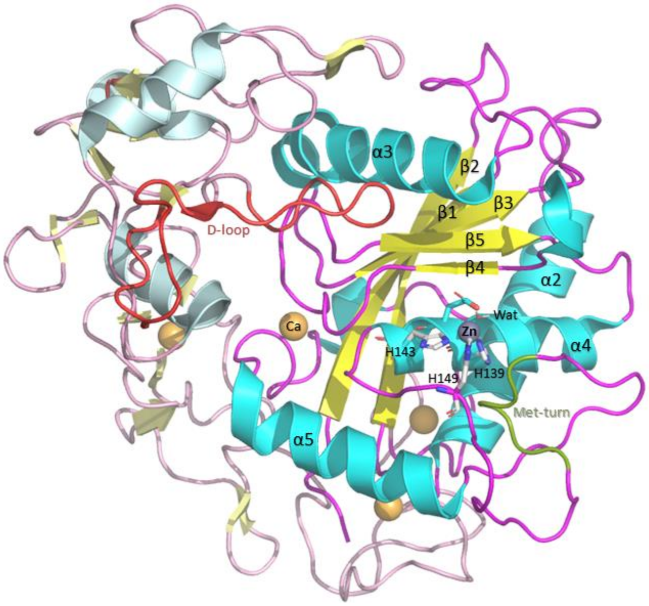
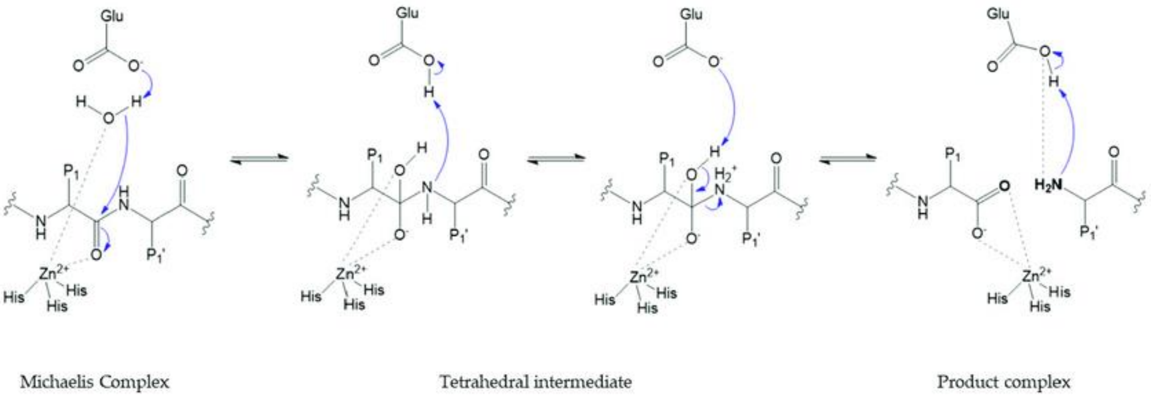


Figure 7 - Modular structure of SVMPs with its zinc ion tetrahedrally coordinated by the three histidines (grey sticks) and the water molecule (sticks). M domain is represented by intense colours, whereas dis and cys-rich domains are represented by pale colours. α-helices in blue, β-strands in yellow and loops in magenta, are represented by new cartoon. Calcium ions are represented by orange spheres. The Met-turn segment is represented by green colour and the D-loop is represented by red colour.

Although the SVMPs have a high degree of structural conservation, their specificity at the level of the substrate is highly variable, resulting in very different functions [20].

1.2.3.1.3. Mechanism of action: general acid/base

A mechanism of action has been proposed for SVMPs in general (*Fig. 8*), in which Glu140 acts as a base and polarizes the catalytic water molecule (by removing a proton). This step facilitates the nucleophilic attack by the newly formed hydroxide (bound to the catalytic zinc) at the carbonyl carbon of the scissile peptide bond of the substrate, by transferring a water proton to a carbonyl oxygen. The oxygen from the substrate carbonyl group becomes negatively charged. In a second step, the Glu140 acts as an acid and transfers the proton, taken from the water to the nitrogen group of the substrate. This leads to the cleavage of the peptide bond and release of the products. It means that mutations in these highly conserved catalytic residues could affect one or more stages of the catalytic process [32, 35].



Michaelis Complex

Tetrahedral intermediate

Product complex

Figure 8 - Schematic representation of the overall proposed metalloproteinase's catalytic mechanism. Adapted from [36].

1.2.4. Russell's viper venom factor X activator (RVV-X)

The potent procoagulant activities caused by *D. russelii* venom on human plasma are mainly due to a specific toxin of its venom, RVV-X, already mentioned [28]. RVV-X is the most potent venom coagulation activator known that causes fatal envenomation [37].

The RVV-X enzyme is a non-haemorrhagic P-IIId SVMP that cleaves a specific peptide bond between Arg194 and Ile195 of the FX , resulting in the release of a 52-residue peptide, the activation peptide (AP) (Fig. 9) [8].

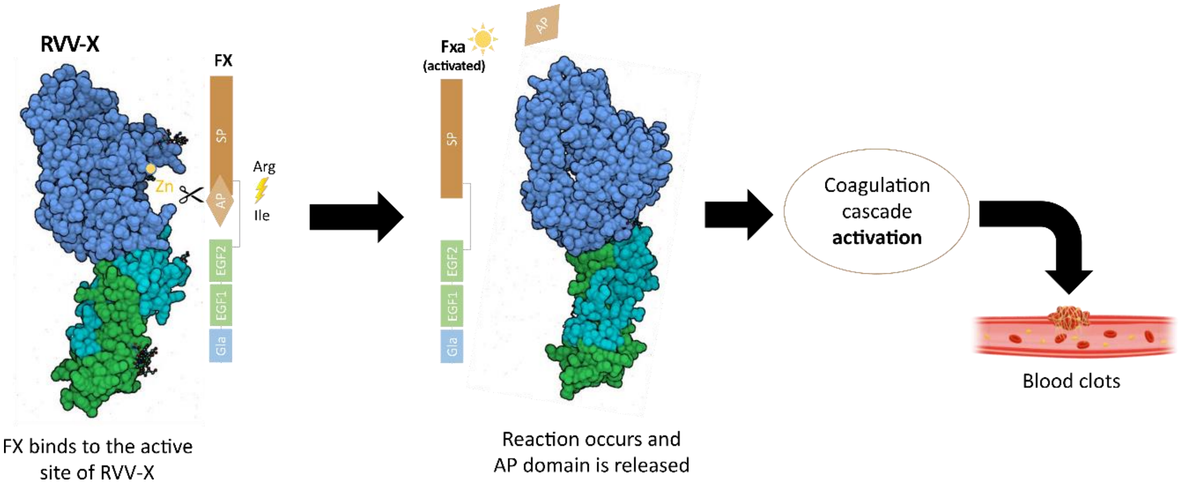

RVV-X
FX
Zn
SP
AP
EGF2
EGF1
Gla
Arg
Ile
Fxa
(activated)
AP
Coagulation cascade activation
Blood clots
FX binds to the active site of RVV-X
Reaction occurs and AP domain is released

Figure 9 - Schematic model of factor X activation by the RVV-X and its effects.

The activation of most clotting factors involves the same process that is the loss of an N-terminal heavy chain peptide. In the FX case, it leads to its conversion into Factor Xα and, consequently, in the activation of the coagulation cascade [8, 29]. Thus, just like for the other vipers, SVMPs play major roles in the pathogenesis of Russell's viper snakebites [33].

1.2.4.1. RVV-X structure-function relationship

The metalloprotease of the RV venom, RVV-X, has already been characterized. Its full crystal structure with a resolution of 2.91 Å was determined by Takeda and colleagues in 2007 [38]. Accordingly, RVV-X is a 93 kDa heterotrimeric glycoprotein composed of three disulfide-linked glycosylated polypeptide chains. One heavy chain (HC) and two light chains (LC1 and LC2), exhibiting a "hook spanner-wrench" configuration. The M/dis-like domains are the hook, whereas the rest of the molecule forms a handle. The backbone structure of the HC (α-chain, 57.6 kDa) follows the characteristics of P-III SVMPs. The HC is composed of three distinct domains: M, Dis-like and Cys-rich domains. These domains are characterized by being organized in a C-shaped configuration where the M domain interacts with the Cys-rich domain. The last interacts with the Cys133 of the LC1 via the unique cysteine residue (Cys389) found in the middle of the HVR forming a disulfide bond [15, 30, 37, 38]. Furthermore, residues from HVR and its neighbours allow the formation of hydrophobic and aromatic interactions and hydrogen bonds with LC1 residues, contributing to a greater stabilization of the HC/LC1 structure [6, 34].

The LC subunit (β- and γ-chains, 19.4 kDa and 16.4 kDa) forms a domain-swapped dimer. They share sequence identity with snake venom C-type lectin-like proteins (snaclecs). These are thought to be regulatory subunits that recognize the γ-carboxyl glutamate residues of the calcium-bound conformation of the Factor X Gla domain and contribute to its selectivity [6, 8, 33, 37, 39]. The dimeric interface formed by the two LC is a concave structure that may function as an exosite, which confers both affinity and binding specificity for the Gla domain of FX in the presence of Ca^{2+} [8, 9].

In addition to activate the factor X, RVV-X also inhibits collagen- and ADP-induced platelet aggregation through its ECD (glutamic, cystein, aspatic) motif present in the disintegrin-like domain. As a result, it leads to intravascular coagulation in the prey [30, 33].

1.2.4.2. Coagulation factor X (FX)

Blood coagulation factor X (FX), also named "Stuart factor", is a vitamin K-dependent (VKD) serine protease zymogen composed of two disulfide linked polypeptide chains (~59 kDa), a 139-residue LC and a 306-residue HC [40, 41].

FX is synthesized in the liver (hepatocytes) and undergoes co- and post-translational modifications before its secretion. These modifications include the maturation of vitamin-K-dependent y-carboxylation of 11 glutamic acid residues, glycosylation with the addition of carbohydrate moieties, among others [41]. When the process is complete, it is then secreted into the blood circulation as a precursor [39, 40].

The activation of factor X by RVV-X occurs in solution, whereas the physiological pathways are a membrane-mediated event [42]. Its activation plays a key role in the blood coagulation cascade as it leads to the final stages of haemostasis. Once activated, FXa associates with another factor (FVa) forming a macromolecular membrane complex (prothrombinase) on the phospholipid surface of the activated platelets in the presence of calcium ions. This complex is responsible for the conversion of prothrombin to thrombin, which in turn, converts fibrinogen to fibrin leading to the formation of blood clots [39, 43].

1.2.4.2.1. Overall structure

The LC is composed of a N-terminal γ-carboxyglutamic acid (Gla)-domain, consisting of 11 γ-carboxyglutamic acid residues that allow the anchorage to platelet and endothelial cell membranes, in a Ca^{2+}-dependent process. Binding to the membrane allow the redirection of the enzymes to the injury site, resulting in an orientation that favours interactions with activators, cofactors and substrates [44].

This domain is followed by a short hydrophobic amino acid stack (residues Phe40-Lys45), two epidermal-growth-factor (EGF)-like domains (EGF1 and EGF2), and a HC. The HC harbours an activation peptide (AP) and a trypsin-like serine protease (SP) domain in which the active-site catalytic triad — His236, Asp282, and Ser379 — is located [39, 43, 44].

Studies suggest that the EFG domains acquire the role of flexible spacers between the Gla domain and the SP domain, considerably distancing the active site from the biological membranes, enabling FX biological activity [39, 45].

The activation peptide (AP) consists of 52 residues (Ser143-Arg194), in an external loop, from Cys132 to Cys259, and is glycosylated with carbohydrates. It is stipulated that, carbohydrates in glycoproteins influence certain physicochemical characteristics, such as conformational stability of the zymogen, resistance to proteases mechanisms via steric effects, and ability to bind to water. So, glycosylation of the activation peptide may be needed to protect the cleavage bond from undesirable physiological and non-physiological activators [41, 46]. Also, the clustering of acidic (glutamic) residues in the AP increases disorder in the protein, interfering with its activation rate [41].

1.3. Conventional treatment and its pitfalls

Antivenoms consist of immunoglobulins (IgG) purified from the plasma/serum of animals (commonly horse or sheep) hyperimmunised with either a single venom (monospecific antivenoms), or a mixture of venoms (polyspecific antivenoms). Intravenous administration of these animal-derived antivenoms is the only treatment available after snakebite envenomation [17, 19].

Antivenoms, when used quickly and appropriately, can neutralize systemic envenoming associated with severe pathologies, such as coagulopathy, haemorrhage, neurotoxic effects and other adverse pharmacological effects [5]. However, it has its drawbacks as venom toxins may act before the antibodies neutralization, even when high amounts of antibodies are present, leading to irreversible sequelae [17, 19]. Another reason for concern is related to the safety of the antivenom. Because its administration needs large doses, it can lead to nonspecific antibodies and therefore induce a series of inflammatory responses.

Furthermore, it requires a proper transport to remote villages, specific storing conditions and delivery, which are a very big limitation in the rural communities. Furthermore, in most cases snakebite victims do not have the means to transport themselves to health facilities, as these are very far away. This poses great challenges as the immediate administration of antivenoms is practically impossible [5, 16, 18].

Unravelling the entire venom composition and pathophysiological manifestations among Russell's venom variants across the Indian subcontinent is necessary and may shed light on how to develop life-saving drugs [31]. The antivenoms used for the treatment of snakebite by the wide distributed Russell’s viper are typically raised against the “Big four” species or those of a particular geographical origin [28]. Moreover, snakes that are subject to experimental studies are usually held captive, despite venom injection being inflicted by wild free snakes [15]. Consequently, some venom toxins may not be neutralized by the antivenom due to variations in the venom composition. As refereed before, the venom composition varies with the age, location and gender of the animal, even among the same species [15, 25]. The ineffectiveness of these antivenoms, in turn, ends up slowing down the hospital management of the envenomed patients [24].

This highlights the urgency to explore and develop more safe, affordable and effective therapeutic tools that might serve as first-aid before hospitalization. It also needs better storage logistics, to overcome the defects of the conventional methods, and completely inhibit the major toxic components of snake venoms [5, 18]. For this purpose, the most commonly present enzymes in venoms which in this case are SVMPs, namely RVV-X, must be studied in detail and, after, explore their neutralization mechanisms.

To address this problem, the pre-reactive complex of RVV-X and FX, never studied before, was firstly estimated, and clarified on how both molecules are put together to favour the reactive mechanism. Besides, 50 potential inhibitory compounds were found for the RVV-X using virtual screening studies.

1.4. Aims

The goal of this work is to elucidate on how rational drug design can be applied against one of the most dangerous snakes in India. Because of the complexity of the venomics, and the lack of detailed information on snake enzymes, particularly on Russell's viper, two major goals for this study were defined:

i. better mechanistic and dynamic understanding of the principal Russell's Viper Venom Factor X activator (RVV-X).

ii. development of a generic, and accessible antidote to inhibit the enzymatic activity of RVV-X.

Chapter II - Materials and Methods

2.1. Homology Modelling

Understanding the target 3D structure is crucial for the following drug design process. Structures are mainly determined by experimental techniques such as X-Ray Crystallography (X-Ray), Nuclear Magnetic Resonance (NMR) and Cryogenic electron microscopy (Cryo-EM). The electronic map is then transformed in special atomic coordinates which are deposited in the Protein Data Bank (PDB) or in the European DB (PDBe). These databases are the resources for three-dimensional (3D) structural data of biological macromolecules. However, when the experimental 3D structure of a given protein is not available, one can resort to homology modelling [47]. The method assumes that homologous proteins (templates) with similar primary sequences exhibit similar three-dimensional structures and biological functions [48].

Furthermore, the first thing to consider when choosing the template is the quality of the sequence alignment between the template and the target (our protein of interest). Ideally, it should display at least 40% of sequence identity, with extensive aligned regions and a low number of gaps in order to produce a biologically accurate model [49].

Structural information of both FX and its activating enzyme (RVV-X) was limited. To overcome this problem, the 3D models of both proteins were built using homology modelling.

As such, once the target protein to be modelled was defined, homology modelling was carried out according to six steps: (1) identification of related sequences whose structures are experimentally (XRC or NMR) available; (2) selection of the template(s); (3) Alignment of target and template sequences; (4) building the model for the target based on the 3D structure of the template; (5) model validation; (6) structure refinement (Fig. 10) [49, 50].

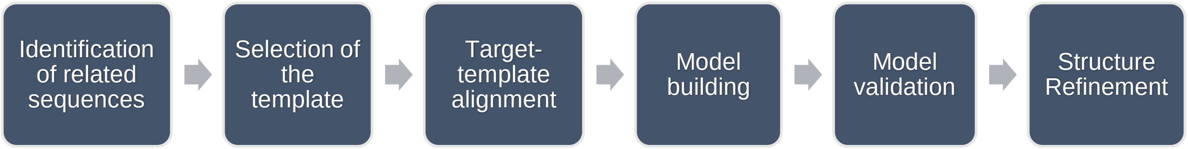

Identification of related sequences
Selection of the template
Target-template alignment
Model building
Model validation
Structure Refinement

Figure 10 - Homology modelling workflow.

2.1.1. Modelling the structure of RVV-X enzyme

2.1.1.1. Template Selection and alignments

The first step of this work was the identification and analysis of the primary structure of the protein of interest, the coagulation factor X-activating enzyme (RVV-X) from *Daboia russelii russelii* venom, which was further used as a query string for the computation of its three-dimensional model.

As mentioned in the introduction (topic 1.2.4.1), RVV-X is composed of three disulfide-linked polypeptide chains, a HC with the metalloproteinase/disintegrin-like/cysteine-rich (MDC) domains and two lectin-like subunits which constitute LC1 and LC2. For this purpose, the amino acid sequence of its heavy (UniProtKB AC: K9JAW0) and light chains (UniProtKB ACs: K9JAX2 and K9JCB2) were retrieved from the Universal Protein Resource (UniProtKB) database. The last consists of a curated database of protein sequences and functional information [51, 52]. These sequences were retrieved in FASTA format and given as input for the SWISS-MODEL web server [53]. The server's Swiss Model Library of Templates (SMTL) [54] was searched against the PDB using two search methods, BLAST [55] and HHblits [56], in order to identify experimentally determined homologous sequences (templates) and automatically align them with the target. A list of 50 suitable candidate templates was returned and ranked accordingly to their Global Model Quality Estimation (GMQE) [54] and QMEAN score, which estimated modelling errors and its accuracy on a global and *per* residue manner [53, 57].

The alignment of the target sequence with its relatives and its corresponding percentage of sequence identity, was obtained from a default automatic multiple sequence alignment (ClustalW). The 1st template on the list was the full crystal structure of RVV-X from the (also) highly venomous Indian *Daboia Siamensis* (PDB ID: 2E3X) viper, which showed 97,66% of sequence identity to the target's HC and 100% and 91,04% to LC1 and LC2, respectively.

However, to study the enzyme's catalytic activity and further search for inhibitors, the choice of an appropriate template must also consider an open conformation state of the enzyme, i.e., its holo form. Therefore, its binding pocket and active site must be intact, with its zinc cofactor coordinated by three histidine residues. According to those requirements, the most suitable candidate template in this work with a high GMQE score was indeed the RVV-X of the *D. Siamensis* viper. This structure was co-crystallized with an inhibitor, Ilomastat (GM6). Furthermore, the resolution at which the template was crystallized was 2.91 Å.

Inspection of the obtained homology model revealed that it did not contain the catalytic water, which was subsequently manually modelled.

2.1.1.2. Model building and Validation

The modelled structure of RVV-X was fully assembled using PyMOL [58] by its superimposition with the crystal structure of the selected template. In addition, Visual Molecular Dynamics (VMD) software [59] was used to align both the model and the template to calculate the Root-Mean-Square Deviation (RMSd) between them, which is a value that reveals the measure of difference between the crystal structure and the model. Both structures showed to be practically completely superimposable, ensuring homology and reliability of the model.

In addition, the backbone of the modelled structure was calculated by analysing phi (φ) and psi (ψ) torsion angles using PROCHECK (Ramachandran plot) analysis [60].

In order to further assess the accuracy and reliability of the homology model, the ProSA (Protein Structure Analysis) [61, 62] web server was used, which recognizes potential issues in structural integrity, thus allowing to check its quality. The total energy deviation of the studied structure could be measured by the *Z*-score, which basically compared it with any energy distribution derived from random conformations.

2.1.2. Modelling the structure of the FX

2.1.2.1. Template Selection and alignments

The primary sequence of human coagulation factor X was identified (UniProtKB AC: P00742) and retrieved from the UniProt database in FASTA format.

As mentioned in the introduction, this protein is composed of 4 domains, namely the serine SP domain with the AP, EGF1 and EGF2 and the GLA domains. To determine if there were homologous sequences to each domain with available X-ray structures, the FASTA sequences were used as query and submitted for a Basic Local Alignment Search Tool (BLAST) analysis with the BlastP (protein-protein blast) engine against the PDB database. Several structures corresponding to human Factor Xa (activated) with high sequence identity were selected for each domain modelling. The PDB codes of the structures used as template for each domain, as well as the resolution and percentage of sequence identity are listed in *Table 1*.

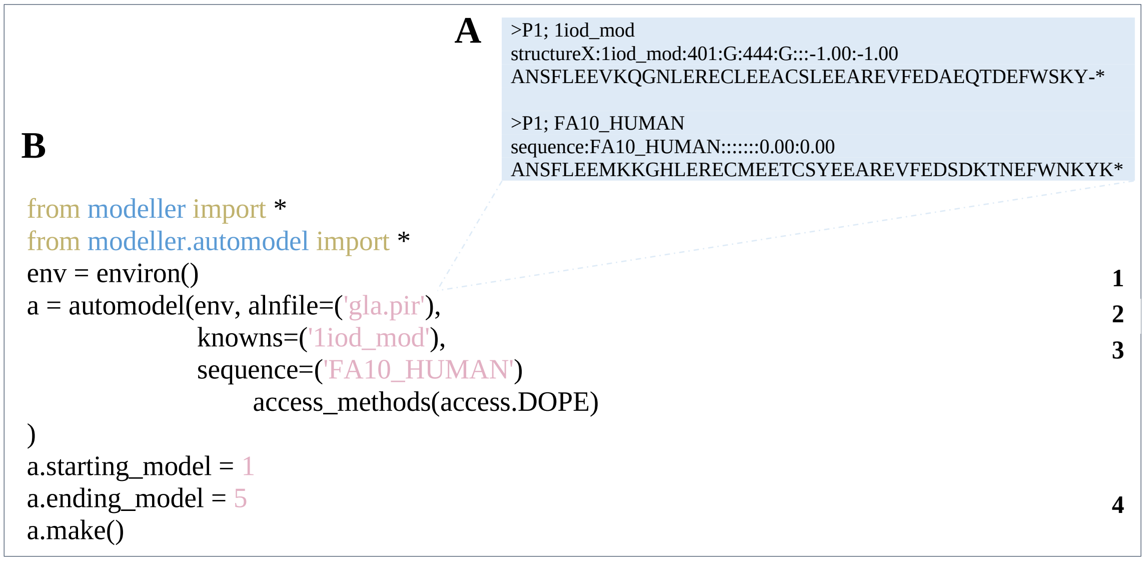
Table 1 – PDB structures selected as suitable templates for the FX model. (*) corresponds to the PDB structure that was later added to the table and that served as a template to build the inactive form (zymogen) of FX.

Domain	PDB ID	Resolution (Å)	% of sequence identity
SP	2XC0 (active) 2CGA (zymogen)*	2.05 2.35	100 -
EFG1	2JKH	1.25	100
EGF2	1XKA	2.3	100
GLA	1IOD	2.3	75

Furthermore, considering that there was no crystallographic structure of the Factor X zymogen, i.e., that possessed the AP, and knowing *a priori* that this segment consisted of 52 amino acids with high prevalence of glutamic residues (negative environment), it is suspected that the region is intrinsically disordered. Therefore it was not possible to experimentally solve its structure to date. Accordingly to Stojanovski et al. 2020, the presence of acid or basic amino acids along the activation peptide creates dipolar forces which in turn increases the intrinsic disorder propensity [46].

Neverthless, x-ray crystal structures exist for the zymogen form of several SPs, such as chymotrypsinogen [63] and proproteinase E [64]. As such, for the modelling of the zymogenic SP domain, these zymogen structures were analysed and superimposed using Pymol. The backbone atoms of fundamental catalytic residues were very similar when superimposed. For this reason, this domain was reconstructed by using the chymotrypsinogen (PDB ID: 2CGA) structure as a template (added to *Table 1* - *).

As it was not possible to find a crystallographic structure that possessed all domains with high sequential identity and since the downloaded templates did not contain all the residues necessary for the assembly of the full protein, the homology modelling of FX was carried out using VMD with a default ClustalW for sequence alignment and the MODELLER 9v11 software [65]. MODELLER 9v11 is a software that builds models accordingly to the alignment of the target sequence previously obtained in the UniProt database and known three-dimensional structures (templates). The advantage of this software compared with the SWISS MODEL, is that the program allows the use of more than one templates at the same time. This way, different regions of the protein can be selected to obtain a more accurate model. This software was used to automatically generate models considering all nonhydrogen atoms for each domain. To achieve this, each domain sequence was aligned against the corresponding templates listed in *Table 1* using VMD and this information was saved in a PIR format file (*Fig.11A*). However, the software also needs the template's spatial coordinates file [65] and so, a script (*Fig. 11B*) was created with the instructions to be followed by MODELLER, for the creation of the distinct models.



```
>P1; 1iod_mod
structureX:1iod_mod:401:G:444:G:::-1.00:-1.00
ANSFLEEVKQGNLERECLEEACSLEEAREVFEDAEQTDEFWSKY-*

>P1; FA10_HUMAN
sequence:FA10_HUMAN::::::0.00:0.00
ANSFLEEMKKGHLERECMEETCSYEEAREVFEDSDKTNEFWNKYK*
```

```
from modeller import *
from modeller.automodel import *
env = environ()
a = automodel(env, alnfile=('gla.pir'),
                    knowns=('1iod_mod'),
                    sequence=('FA10_HUMAN')
                         access_methods(access.DOPE)
)
a.starting_model = 1
a.ending_model = 5
a.make()
```

Figure 11 - An example of a PIR file and a script with instructions for the modelling of the Gla domain by MODELLER. (A) Sequences on the blue box, correspond, respectively to the template (1iod_mod) and the target sequence (FA10_HUMAN) alignment; (B) (1) the PIR file, (2) the template, (3) the target sequence, and (4) the number of models to be created. This script also creates a log file (summary file) with information about the modelling.

The quality of the generated models was ranked using DOPE (Discrete Optimized Protein Energy) assessment score [65]. DOPE is a scoring function based on atomic distance-dependent statistical potential which compares energies from different generated models considering the target sequence. More accurate models possessed lower scores. It is typically used for model prediction and assessment, as it can evaluate the structure of the modelled protein. In addition, it accesses problematic regions of the structure, generating a residue-by-residue energy profile for the resulting model. Out of the 10 resulting models, the one with the best score was chosen [65].

2.1.2.2. Model building and validation

The output models were visualized using PyMOL and its assembly was carried out by their superimposition with the corresponding crystal structures of the templates listed in *Table 1*, and manually curated.

In addition to the predictive quality parameters used by MODELLER, the accuracy and reliability of the protein model was evaluated using the Z-score calculation at the ProSa-Web server.

2.2. Structure Refinement

After building both RVV-X and FX models, its refinement was carried out by minimizing its energy with the AMBER (Assisted Model Building with Energy Refinement) package of computer programs. This process relaxes the system and leads it to its local minimum energy, allowing the correction of possible deviations from the correct protein geometry that could impair its result.

2.2.1. Setup of the initial structure and molecular simulations

In this step, a pdb4amber script was used and as its name implies, it is designed to prepare PDB files in the AMBER format. It removes irrelevant information (TITLE, REMARK, HELIX, SHEET e CONECT) and renumbers both residues and respective atoms, keeping all the information needed to start with Molecular Dynamics setup.

Both RVV-X and FX included thousands of atoms (10376 and 6285, respectively). The AMBER package allowed not only the preparation of the setup (necessary input files) to perform simulations, but also to analyse the results [66]. It contained libraries with parameter sets for the residues and respective protonation states under physiological conditions. These parameters described each residue’s type of atoms, force constants and equilibrium values for each bond, angle, dihedral, Van der Waals forces and partial charges of each atom. In summary, this force field is an “All-atom” type which provides parameters for all atoms in the system, including hydrogens. AMBER force field equation (*Equation 1*) is defined as:

$$U(\overrightarrow{R}) = \sum_{bonds} K_b \ (b - b_0)^2 + \sum_{angles} K_\Theta \ (\Theta - \Theta_0)^2 + \sum_{dihedrals} \frac{V_n}{2} \ [1 + \cos(n\emptyset - \gamma)] + \sum_{nonbonded} [\frac{A_{ij}}{R_{ij}^{12}} - \frac{B_{ij}}{R_{ij}^{6}} + \frac{q_i q_j}{\varepsilon_1 R_{ij}}]$$

(Equation 1)

Where K_b, K_Θ e V_n represent force constants related to bond connections, angles and dihedrals, respectively; b, Ө and ∅ represent the connection length, angle and dihedral, with index 0 concerning the equilibrium values for each individual term; γ corresponds to the phase angle (values between 0º and 180º); A_{ij} and B_{ij} are the terms of repulsion and attraction of dispersive interactions; q_i e q_j are the partial atomic charges of atoms i and j, while ε_1 are the dielectric constant of the surroundings.

2.2.1.1. Metal Center Parameters generation

The python-based Metal Center Parameter Builder (MCPB.py) [67] module of the AmberTools18 package has been widely used to parametrize metalloproteinase's metal centre for further MD simulations.

In this work, the bonded model was adopted to represent the interactions between the metal ion and the neighbouring residues. First, the PDB files for the standard and non-standard residues of the RVV-X were prepared. Protonation states were assigned based on the pKa value using *H*++ webserver (http://biophysics.cs.vt.edu/H++) [68, 69]. However, it does not consider either the metal ion or the catalytic water while adding hydrogen atoms, and so, the protonation state of His139 was wrong, which should be neutral. This had to be manually fixed, by deleting the extra "Hε2" atom and renaming the residue to "HID" instead of "HIE". The resulting PDB file was renumbered using pdb4amber, explained in the previous topic.

In the next step, a MCPB.py script generated three models using the ff14SB force field, a small model which was used for bond and angle parameters calculation, a standard model which contained the atom type information, and a large model for charge (RESP) calculation. The small model contained the zinc metal with the coordinating histidine's sidechains, in which hydrogen atoms were added to the truncated bonds, whereas the bigger model possessed the smaller model with entire residues and the truncated bonds were capped by NME and ACE.

Then, B3LYP/6-31G* level of theory [70, 71] was employed to perform geometry optimization and force constant calculation on the smaller model. Merz-Kollman atomic charges were calculated for the larger model with the Restricted Electrostatic Potential

(RESP) method and assigned to the corresponding atoms. These calculations were carried out with the Gaussian 09 software [72].

Finally, the Seminario method which uses the Cartesian Hessian matrix derived the metal site force field parameters for AMBER, that included the connection length, connection angle, dihedral angle, and RESP charges, and then the entire forming metal bond was integrated. MCPB.py then returned the *.frcmod* file necessary for the generation of the topology and parameters file during tLeap modelling.

2.2.2. Generation of the topology and parameter files

The *tLeap* module from the Amber program, was then used to create conditions that mimic the system in its physiological conditions.

To carry out the simulations of the metalloproteinase RVV-X, the parameters file (*.frcmod*) generated in the previous topic (*2.2.1.1.*), as well as the AMBER ff14SB library of force-fields were loaded on tLeap, whereas for the FX simulations, only the last was loaded to model all standard amino acid residues.

Although it is possible to perform simulations in the vacuum, a realistic simulation of molecular movements implies the existence of an aqueous environment, usually water, in which the protein is immersed. Therefore, it was necessary to include the solvent effect in the simulation due to hydrophobic interactions and hydrogen bonds contribution on the force field. A solvation box was used with the molecules of interest within, and its empty spaces filled with water. To do this, the systems were solvated in an octahedral box of TIP3P (Transferable Intermolecular Potencial 3P) water molecules so that the box boundaries were at least 15 Å away from any protein atom. Finally, counterions (8 Cl^- for the RVV-X model and 17 Cl^- for the FX model) were added to maintain the electroneutrality of the simulated system. Also, missing hydrogen atoms were added by the *tLeap* because the crystal structures and homology models lack them.

In the end, topology parameters (*.prmtop*) and atoms coordinates (*.inpcrd*) files were generated, using the command *SaveAmberParm*.

2.2.3. Geometry optimization

Before performing a molecular dynamics simulation, a geometry optimization process must be carried out. This process promotes relaxation of the system, approximating it to the minimum local energy [73].

The energy of any molecular conformation can be calculated from its force field and the coordinates of the constituent atoms. Thus, the potential energy, called “total potential energy”, can be considered as the sum of mechanical contributions and non-binding interactions as represented on *Equation 2*.

$$E_{total\,potential} = \underbrace{E_{bond} + E_{\Theta} + E_{\omega}}_{E_{intramol}} + \underbrace{E_{vdw} + E_{el}}_{E_{intermol}} \qquad \text{(Equation 2)}$$

Where $E_{potential\ total}$ is the total potential energy of the system, $E_{Bonding}$ is the energy associated with the deformation of the bond length between two atoms, E_{Θ} is the energy associated with bond angles variation, E_{ω} is the energy associated with the torsion of the dihedral angles. The last three make up the intramolecular E. E_{vdw} and E_{el} correspond to the energy associated with non-binding interactions independent of the charge and dependent on the atomic charges, respectively. The last two make up the intermolecular E.

This process relaxes the initial atomic positions, avoiding undesirable steric repulsions and clashes that could result from excessively close or misdirected atoms in the modelling process. Jumping over the minimization step might affect the final configuration of the molecule and also its reactivity [73], leading to unrealistic systems.

The process consists in 4 steps through the creation of 4 inputs that differ on the atom’s constraints (*Fig.12*). The four were carried out sequentially to allow a maximum reduction in the energy of the system, aiming to reduce any unfavourable interactions caused by the automated placement of the water molecules by the *tLeap* program. So, it firstly started with the minimization of the geometry and position of the added water molecules, while the rest of the system was kept fixed. Secondly, the minimization of the hydrogen atoms, also added by the previous program, keeping the remaining atoms fixed. Then, the side chains were relaxed. Finally, the position of all atoms in the system was optimized. The obtained models were viewed using VMD – Visual Molecular Dynamics.

```
Min1_Water_Minimization
 &cntrl
   imin = 1,
   maxcyc = 500, ncyc = 250,
   ntb = 1, cut = 10, ntr = 1, ntxo = 1,
   restraint_wt = 50.0,
   restraintmask = ':*&!:WAT'
 /
```

```
Min2_Hydrogen_Minimization
 &cntrl
   imin = 1,
   maxcyc = 500, ncyc = 250,
   ntb = 1, cut = 10, ntr = 1,  ntxo = 1,
   restraint_wt = 50.0,
   restraintmask = ':*&!(:WAT|@H=)'
 /
```

```
Min3_Side_Chains_Minimization
 &cntrl
  imin = 1,
  maxcyc = 500, ncyc = 250,
  ntb = 1, cut = 10, ntr = 1, ntxo = 1,
  restraint_wt = 50.0,
  restraintmask = '@CA,C,N,O&!:WAT'
 /
```

```
Min4_Whole_System_Minimization
 &cntrl
  imin = 1,
  maxcyc = 500, ncyc = 250,
  ntb = 1, cut = 10, ntxo = 1,
 /
```

Figure 12 - Scripts used for the process of minimizing the geometry of both the protein and the peptide. &cntrl: initiates the variables block to run minimizations/MD; imin: flag to run minimization; maxcyc: maximum number of minimization cycles; ncyc: number of minimization cycles with steepest descent algorithm, since it is lower than the maxcyc, the remaining cycles were performed with the conjugated gradient algorithm; ntb: controls whether or not periodic boundaries are imposed in the system; cut: cut radius for nonbonded interactions; ntr: flag for restraining specified atoms in Cartesian space; ntxo: format of the final coordinates, velocities and box size written to the .rst file (=1, ASCII); restraint_wt: the weight for the positional restraints (50 kcal·mol^{-1}·Å^{-2}); restraintmask: specifies the restrained atom; / indicates the end of the variables block.

In all minimizations (imin = 1) the system was defined with a cut-off for intermolecular interactions (non-ligands) of 10.0 Å. A total of 500 cycles (maxcyc = 500), of which 250 use the “steepest descent” algorithm (ncyc = 250) was included and the remaining cycles used the Conjugate Gradient algorithm. After each minimization, an *.rst* file was generated with the atomic coordinates of the post-minimization structure. The last generated file (*.rst)* served as the basis for the subsequent molecular dynamics process.

SANDER (Simulated Annealing with NMR-Derived Energy Restraints) engine from the AMBER package used *.prmtop* and *.inpcrd* files generated by *tLeap* to run the scripts and perform energy minimizations and molecular dynamics.

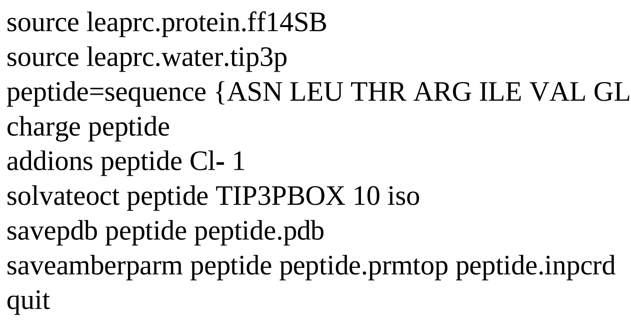
2.3. Protein-peptide Docking

Protein-peptide docking is a computational method that simulates covalent and non-covalent interactions between molecules (receptor and ligand) assembling a stable complex.

In order to further study the reaction mechanism of RVV-X and to predict the preferred orientation of the peptide to the target, a docking of the substrate cleavage region in the metalloproteinase was carried out.

As previously explained in the topic *2.1.2.1.*, the AP structure is still unknown, and it is believed to be disordered. Therefore, before carrying out the docking calculation, the AP was modeled. To this end, the last 4 residues of the AP (Asn-Leu-Thr-Arg) and the first 4 residues of the SP N-terminal domain (Ile-Val-Gly-Gly) were modeled. Thus, force field libraries were loaded and the *sequence* command within tLeap was used to build the 8-amino acid peptide (*Fig.13*).

The *sequence* command connects each residue to build the *peptide* unit. So, tLeap automatically recognizes ASN1 as the first residue (N-terminal with NH3$^+$) in the sequence, and GLY8 as the last residue (C-terminal end with a carboxylate).



```
source leaprc.protein.ff14SB
source leaprc.water.tip3p
peptide=sequence {ASN LEU THR ARG ILE VAL GLY GLY}
charge peptide
addions peptide Cl- 1
solvateoct peptide TIP3PBOX 10 iso
savepdb peptide peptide.pdb
saveamberparm peptide peptide.prmtop peptide.inpcrd
quit
```

Figure 13 - The script used for the generation of the peptide structure by tLeap.

In the end, topology parameters (*.prmtop*) and atoms coordinates (*.inpcrd*). files necessary for its minimizations were generated.

The peptide minimizations were carried out according to the same procedure used in topic *2.2.3*. With the substrate and target minimized, the next step consisted of protein-peptide docking.

Simulation of peptide docking was performed using HADDOCK (High Ambiguity Driven DOCKing proteins) [74], and HPepDock (Hierarchical Flexible Peptide Docking) [75, 76] online servers, to check the coherency of the prediction.

Both the RVV-X metalloproteinase (receptor) and peptide (ligand) PDB files were imported as input data for docking. Then, a restraint file was generated in which the receptor residues (His143, Zn417, Glu140) that interact with the ligand residues (Thr3, Arg4 and Ile5) were respectively specified, with a minimum and maximum distance of 2 and 6 Å, respectively.

Although the HADDOCK tool is one of the most used and cited docking tools per year, HPepDock webserver revealed a better interaction between the enzyme and the substate.

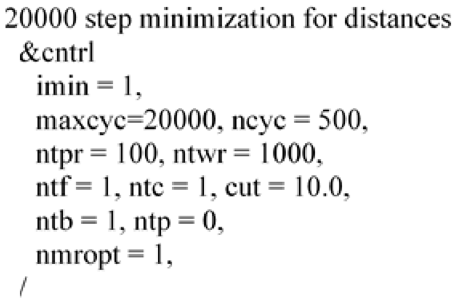
HPepDock is a webserver for protein-peptide docking which considers peptide flexibility. In a first step, it generates multiple peptide conformations with ModPep program. Then, the generated conformations are docked against the binding site of the target protein with a modified version of MDock program [76].

The most favourable interactions output from the top cluster were selected and later compared with the model proposed by Takeda et al., 2007 [38].

2.4. Molecular Dynamics (MD) Simulations

In order to evaluate which of the models was the most favourable and respective dynamic behaviours over time, Molecular Dynamics (aMD) simulations were performed using the AMBER18 software with the pmemd.MPI module.

First, the protein-peptide complexes were refined to remove undesirable contact points, as explained above. To carry out this process, a script represented in *Fig. 14* was used.



```
20000 step minimization for distances
 &cntrl
   imin = 1,
   maxcyc=20000, ncyc = 500,
   ntpr = 100, ntwr = 1000,
   ntf = 1, ntc = 1, cut = 10.0,
   ntb = 1, ntp = 0,
   nmropt = 1,
 /
```

Figure 14 - Script used for the process of minimizing the geometry of the protein-peptide complex.

So, the system was minimized during 20000 cycles, (maxcyc = 20000), of which 500 used the "steepest descent" algorithm (ncyc = 500) and the remaining cycles used the Conjugate Gradient algorithm. A cut-off for intermolecular interactions (non-ligands) of 10.0

Å was defined. After each minimization, a *.rst* file was generated with the atomic coordinates and was used for the following molecular dynamics process.

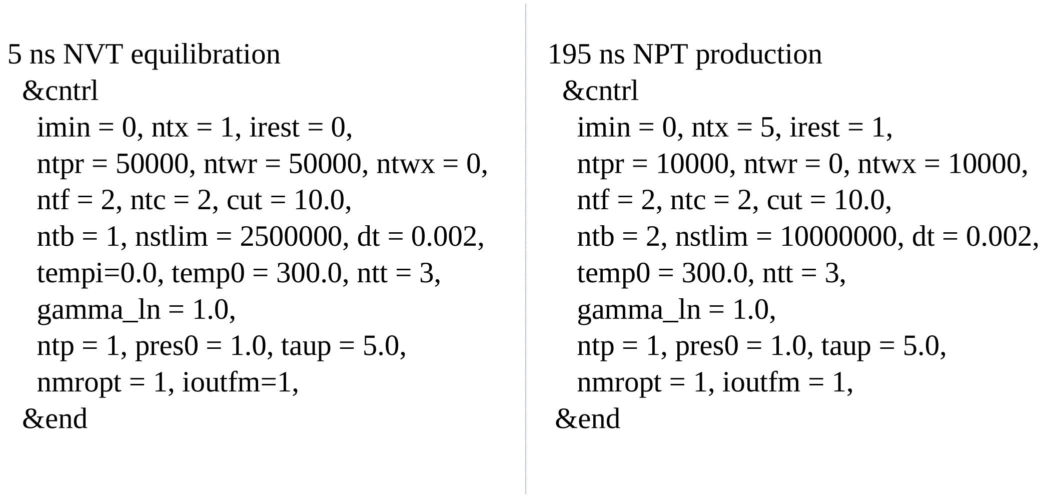
Then, the molecular dynamics simulations started with all the atoms being subjected to a slow heating procedure (equilibration) in which the system's temperature gradually increased from 0 K (tempi=0.0) to 300 K (temp0=300.0), using the Langevin thermostat (ntt = 3). The heating was performed over 5 ns at constant volume (ntb=1), followed by 155 ns of production at constant pressure (ntb=2), 1 bar, recording the trajectories every 10000 steps (ntwx=10000).

The frequency of collisions was established at 1 ps^{-1} (gamma_ln=1.0) and 5000000 simulation steps (ntslim = 5000000) with an integration time of 1 fs (dt = 0.001) were considered. Finally, a cut-off of 10.0 Å for non-ligand interactions was established, indicating that only interactions that fall within that radius were considered. If cut-off values were not used, the computational time would increase according to the number of atoms/interactions in the system.

In the third step, called production, occurred at constant temperature and pressure (ntb = 2), considering the isothermal-isobaric ensemble (NPT), allowing a more representative system density and therefore a more coherent simulation of the biological phenomena under study. In this case, only coordinates but not velocities were read (ntx=1) and 1000000 simulation steps were used (nstlim = 1000000), with 1 fs (dt = 0.001).

5 replicas (REP1-5) were run following the same steps for both solutions in order to assess the reliability of the system's dynamic behaviour.

Then, several files were created, corresponding to the energy information (*.mdout* and *.mdinfo*) every 50000 steps (ntpr=50000) and to the coordinates (*.nc* and *.mdcrd*) also every 50000 steps (ntwx=50000). The generated *.mdcrd* files contemplated a set of coordinates of each atom's nucleus from the system and its temporal evolution, that is, the trajectories of each atom. To carry out this process, two scripts were created (*Fig. 15)*.



```
5 ns NVT equilibration
 &cntrl
  imin = 0, ntx = 1, irest = 0,
  ntpr = 50000, ntwr = 50000, ntwx = 0,
  ntf = 2, ntc = 2, cut = 10.0,
  ntb = 1, nstlim = 2500000, dt = 0.002,
  tempi=0.0, temp0 = 300.0, ntt = 3,
  gamma_ln = 1.0,
  ntp = 1, pres0 = 1.0, taup = 5.0,
  nmropt = 1, ioutfm=1,
 &end
```

```
195 ns NPT production
 &cntrl
  imin = 0, ntx = 5, irest = 1,
  ntpr = 10000, ntwr = 0, ntwx = 10000,
  ntf = 2, ntc = 2, cut = 10.0,
  ntb = 2, nstlim = 10000000, dt = 0.002,
  temp0 = 300.0, ntt = 3,
  gamma_ln = 1.0,
  ntp = 1, pres0 = 1.0, taup = 5.0,
  nmropt = 1, ioutfm = 1,
&end
```

Figure 15 - Scripts used for the molecular dynamic simulations of the protein-peptide complex. On the left, the script of the heating phase is represented, and on the right the production process.

2.4.1. MD Analysis

To analyse the accuracy of the molecular dynamics produced in the previous step for each complex and respective replica, multiple *pdb* files were extracted from the *.nc* files at every 100 frames (from a total of 10000 frames) of the simulation time, leaving out the solvent. Results were visualized using the Pymol software, by loading the extracted pdbs.

Moreover, to depict the distribution of water molecules around the catalytic zinc, the trajectories were aligned, and occupancy water maps were computed using the VolMap plugin of VMD. This plugin creates occupancy maps, 3D grids, in which each grid point is set to be 1 or 0, if it finds one or more atoms or if it does not, respectively. When averaged over the entire trajectory, this map provides the fractional occupancy of that grid point.

After inspecting the trajectory pdb files, a series of analyses of different parameters, including calculation of distances and RMSDs, were carried out using CPPTRAJ program from the AMBER package. These calculations were carried out considering the initial structure.

Then, the average structure was calculated using the script represented in *figure 16* for the last 20 ns, which was necessary for the calculation of atomic fluctuations (RMSF).

```
trajin md_10.nc
autoimage
center origin :*byres
rms :1-426 out all_rms.data
average avg.rst7 rst7 :1-426 out
average avg.pdb pdb :1-426 out
run
```

Figure 16 - Script used to calculate the average structure from the last .nc file corresponding to the last 20 ns.

Once created, this average structure was minimized (*avg_min.rst*) using the *.rst7* file and its dry *.prmtop*, that is, without the solvent and counterions following the procedures above mentioned.

2.4.1.1. Root-mean-square deviation (RMSd) calculation

One of the first steps that must be taken to analyse the results produced by MD is the evaluation of the system's stability to ensure that equilibrium conditions are met.

RMSd (Root-mean-square deviation) is a standard measure used to compare the distance between two sets of atomic coordinates, i.e., how much a structure has diverged in the three-dimensional space. Therefore, these calculations were useful to estimate the average deviation of the protein atoms throughout the simulation from the reference structure, which, usually, corresponds to the first frame of the simulation or the input structure. As it allowed the inspection of possible displacements during the simulation, it was also used to verify issues related to the stability and rigidity/flexibility of the system under study [77].

In general, the higher the RMSd value, the more pronounced are the differences between structural conformations. Thus, for a structure to be considered stabilized the RMSd value in relation to a reference structure must be lower than 2.0 Å. However, in large structures it is known that the RMSD values may be significantly large [77, 78].

Considering two sets of *n* atoms coordinates for each structure (*v* and *w*), the RMSD between the two structures would be defined as:

$$\mathrm{RMSd}_{(\mathrm{v,w})} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}\left(v_x^i - w_x^i\right)^2 + \left(v_y^i - w_y^i\right)^2 + \left(v_z^i - w_z^i\right)^2} \qquad \text{(Equation 3)}$$

The RMSd analysis were carried out to depict structural deviations on the protein-peptide complex after equilibration and production simulations.

A script (*Fig. 17*) was created to calculate the RMSd for all the atoms in the protein as well as for the backbone atoms (N, Cα, C and O) in each frame of both metalloproteinase (residues 1-416) and peptide (residues 417-424). The trajectory files were given as input to be read by CPPTRAJ (e.g., *trajin md_1.nc*).

```
trajin md_1.nc
trajin md_2.nc
#(...) remaining trajectory files
autoimage
center origin :*byres
rms :1-416@CA,C,N,O out rms_metalloproteinase_bb.data
rms :1-416 out rms_metalloproteinase.data
rms :417-424 out rms_peptide.data
rms :417-424@CA,C,N,O out rms_peptide_bb.data
```

Figure 17 – The script created for RMSd calculations on both the metalloproteinase and the peptide.

2.4.1.2. Determination of distances

The distance variation between certain active site atoms during the simulation of molecular dynamics was analysed in order to check if the active centre was stable. In this step, the distances between the zinc atom (ZN) and the glutamate oxygens (Oε1 and Oε2) were analysed. Also, calculation of the distances between the zinc ion and the oxygen from the arginine of the substrate (Zn-Opep), between the catalytic water and the carbon of the arginine (OWat-Cpep) and between the catalytic glutamate and the nitrogen of the scissile peptide bond (Glu-Npep) were conducted. The script used for the calculation of the distances for example in the case of the coordinating His is represented in *figure 18*.

```
trajin md_1.nc
trajin md_2.nc
#(...) remaining trajectory files till the tenth .nc file
autoimage
center origin :*byres
distance d1 :425@ZN :140@OE1 out distance_zinc_glu140_1.data
distance d2 :425@ZN :140@OE2 out distance_zinc_glu140_2.data
```

Figure 18 - The script created for distance calculations.

2.4.1.3. Root-mean-square fluctuation (RMSF) calculation

After calculating the RMSD, it was possible to verify whether the structure of the complex under study fluctuated around a stable average conformation. From there, considering the average structure coordinates, the structural fluctuations could be calculated, that is, the RMSF.

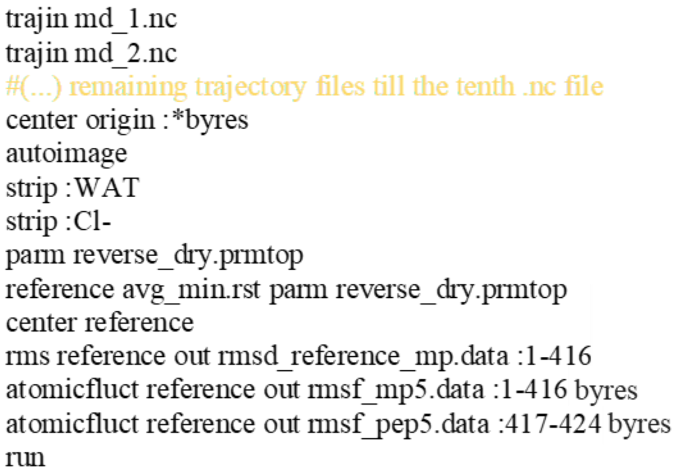
The root-mean-square fluctuation (RMSF) measures the fluctuations of a particular atomic position or group of atoms over time from a reference position, which was the average position.

This allows a description of the protein's flexibility, the changes suffered by the solvent exposed regions, the fluctuations of the residues that interact with the substrate and the variability of the substrate itself throughout the simulation. The RMSF value is given by the *Equation 4:*

$$\mathrm{RMSF}_{(\mathrm{i})} = \sqrt{\frac{1}{T}\sum_{t=1}^{T}\langle\left(r_i'(t) - r_i(t_{ref})^2\right)\rangle} \qquad \text{(Equation 4)}$$

Where r′ is the position of the atoms in the residue *i* after superimposition on the average structure, r_i is the position of the residue, T is the trajectory time, t_{ref} is the reference time , and the angle brackets designate the squared distance of atoms in the residue average.

A script (*Fig. 19*) was created to calculate the RMSF for both the protein (rmsf_mp.data - :1-416) and the peptide (rmsf_pep.data - :417-424) atoms using the atomicfluct command. The trajectory files were given as input to be read by CPPTRAJ (e.g., *trajin md_1.nc*).



```
trajin md_1.nc
trajin md_2.nc
#(...) remaining trajectory files till the tenth .nc file
center origin :*byres
autoimage
strip :WAT
strip :Cl-
parm reverse_dry.prmtop
reference avg_min.rst parm reverse_dry.prmtop
center reference
rms reference out rmsd_reference_mp.data :1-416
atomicfluct reference out rmsf_mp5.data :1-416 byres
atomicfluct reference out rmsf_pep5.data :417-424 byres
run
```

Figure 19 - The script created for RMSF calculations on both the metalloproteinase and the peptide.

2.4.2. Steered Molecular Dynamics (sMD)

Since the carbonyl of the zinc-binding substrate was farther away than what was favourable for the reaction to occur in all replicas, a sMD was performed to bring the oxygen coordinates of the arginine closer to the zinc. Here the second and fifth replicate (REP2 and REP5) whose molecular dynamics analysis proved to be one of the most stable were chosen.

SMD is a method that applies an external force to a reaction coordinate (one or more atoms), leading to its motion within a certain time at a constant velocity [79].

The SMD run was setup according to the Amber 2020 manual [73], by defining the jar variable to 1 (jar = 1). The coordinate's motion was carried along 14 ns from a starting value of X to an end value of Y between carbonyl oxygen atom of the peptide and Zn atom, decreasing 0.1 Å each step with a constant force of 5000. These restraints were specified in an additional "dist.RST" file.

2.5. A reaction mechanism study

The knowledge of the RVV-X reaction mechanism against FX is essential for the development of effective inhibitors.

For the study of large biomolecular systems, classical molecular dynamics simulations are typically used. However, molecular mechanics (MM) is not able to describe some processes that occur in the system, as is the case of the breaking or formation of chemical bonds and electronic rearrangements. To achieve this, it is necessary to resort to quantum mechanics (QM), which allows the study of the nature at the subatomic level by solving the Schrödinger equation [80, 81].

However, the Schrodinger equation has no analytic solution for atomic/molecular systems with more than one electron. It is possible to solve the Schrodinger equation resorting to numeric approximations [81], which nevertheless are very time-consuming if a reasonable accuracy (errors of 3 kcal.mol^{-1} or less) is necessary, even if a few hundred atoms are considered. Alternatively, the density functional theory (DFT) [82], another QM method to obtain the energy of a system, can reach such accuracy for a few hundred atoms with a more reduced computational power. Despite that, enzymes possess a small region, the reactive centre where electronic rearrangements occur, that is, chemical reactions.

Thus, in 1976, Warshel and Levitt [83] described the hybrid method QM/MM (Quantum Mechanics/Molecular mechanics), which combines both QM and MM methods, allowing for a more accurate study of systems with up to thousands of atoms.

Hybrid QM/MM methods consist of partitioning the system into layers (two or more) based on their contribution to the overall reaction process. Thus, each layer is treated with different theoretical levels. Accordingly, the reactive region which may contain only a few hundred atoms is treated using QM methods allowing for an electronic description of the system. On the other hand, the surrounding region (remaining part of the protein and solvent) which can contain thousands of atoms is described with MM, that is, with classical force fields such as AMBER [84].

The ONIOM (Our Own N-layer Integrated molecular Orbital molecular Mechanics) framework for QM/MM, developed by Morokuma and co-workers in 1996 [85], is commonly used to treat this type of systems. This method, among other things, allows to make geometry optimizations, calculate energies, and calculate frequencies.

2.5.1. Preparation of the starting structures

After running and analysing the structures resulted from the SMD, snapshots with the most suitable distances for the occurrence of the enzymatic reaction were taken as the initial structures. Also, the minimized structure of the model 2 (structure A) was used to study the mechanistic.

In order to reduce the computational cost of this process, the structures were truncated so that only the M catalytic domain and the substrate were present.

However, once again it is necessary to consider that biomolecules exist in an aqueous environment and as such, its effect were also included to realistically simulate the enzymatic reaction. Thus, the preparation of the input structures was carried out according to the procedures used in topic *2.2*, in which water molecules and chloride ions were added. In the end, parameters, and coordinates files necessary for the building of the QM/MM models were generated. However, to minimize the computational effort, the size of the solvent shell was reduced by defining a cap of water molecules within a distance of 6 Å from the protein surface, keeping the side chains of the outer residues solvated. In total, the systems consisted of 6954 (structure A) and 7081 (post-dynamics structures) atoms.

2.5.2. ONIOM QM/MM setup protocol

The calculations were prepared using molUP VMD's extension, which provides a full-featured graphical user interface (GUI) to the software Gaussian [86]. It was created to allow users to build the Gaussian input (*.com*) files and analyse the output (*.log*) of the generated calculations. The input files contain a panoply of data regarding the structure of the molecule and instructions used in the calculation (level of theory, basis set, type of calculation, atom coordinates, connectivity, and atom parameters).

The first step consisted in setting the ONIOM layers using molUP, the high-layer (HL) which is the layer that was treated by the QM approach, and the low-layer (LL), that was treated by the MM approach.

2.5.3. Definition of the QM/MM boundaries

The HL must include all the atoms that undergo bond formation and/or breaking processes, as well as those whose hybridization changes during the reaction. Here, it included part of the active site (Zn ion, the catalytic water molecule, the side chains of His139, His143, His149 and the Glu149 from the protein). In addition, it comprehends the substrate (carbonyl group of Arg and part of Ile). This layer consisted of 83 atoms, with a net charge of +1. The energy and forces in the QM part were described with a higher level of theory, the DFT level with a hybrid functional, the B3LYP in combination with the 6- 31G(d) basis set.

The LL included the rest of the protein and the solvent water molecules, which were treated with a lower level of theory, the AMBER *ff*14SB.

ONIOM automatically draws the boundaries between the two layers by adding "link" atoms (typically between the side chains of the residues around the active site and the α-carbon structure). Thus, the atoms from the QM region that were covalently linked to the those from the MM region were capped with hydrogen atoms to saturate their valences.

2.5.4. Optimization of the starting structures

All geometry optimizations were performed using Gaussian 09. In all cases, two optimization processes were carried out, in which the QM/MM interactions were treated differently.

The first optimization, where all atoms were free to optimize, was treated with the mechanical embedding formalism. On the other side, the second optimization, was treated with the electrostatic embedding formalism, to account the polarization of the QM subsystem by the MM atoms. Here, the MM region was divided into two sub-layers, an outer layer, in which the external cap of water molecules was kept fixed and an inner layer, in which the atoms were free to move and optimize. This approach is often used to reduce the computation effort during QM/MM optimization.

In the end, a structure similar enough to the reactant model was used as input for further calculations.

2.5.5. Determination of the stationary points

To study the energy profile of the reaction's steps and further search for the transition states, linear transit scans were performed. For example, the distance between the water oxygen atom and the scissile carbonyl carbon atom of the substrate was chosen as the reaction coordinate for the first step of the reaction considering the proposed reaction mechanism.

The putative transition structures were defined as the ones corresponding to the energy maxima in the Potential Energy surface (PES) profile of the reaction coordinates. These were subsequently fully optimized (apart from the frozen atoms), their vibrational frequencies were calculated to confirm that the structure obtained was indeed a transition state, and the vector was inspected to confirm that the transition state was indeed the one of interest (later definitively confirmed by IRC calculations).

Once a transition state structure was found, intrinsic reaction coordinate (IRC) calculations were performed to confirm the correct reactants and products of the reaction. The optimized products resulted from the IRC would correspond to the intermediates of the reaction.

These steps were repeated for the second step, the transference of a proton to the peptide's nitrogen and for the third step corresponding to the cleavage of the scissile peptide bond.

It is noteworthy to mention that the optimizations of geometry as well as the stationary points were obtained with the convergence criteria implemented in Gaussian.

2.6. Virtual Screening (VS)

2.6.1. Compound preparation

Virtual Screening (VS) is an in silico technique for the development of new drugs, which works by identifying the most promising molecules capable of binding to the enzyme receptor, to be subsequently tested in vitro [87].

This technique has already contributed with many bioactive substances on the market. Therefore, it has become fundamental to assist in the fast discovery and optimization of cheap drugs [50, 87, 88].

2.6.1.1. Dataset collection

An initial evaluation of the RVV-X binding pocket was performed and available PDB structures with co-crystallized ligands and similar active sites were also inspected. Of the analysed crystal structures, 6 complied with the conditions mentioned above (PDB IDs: 2DW0 [89], 2AIG [90], 1DTH [91], 1ATL [92], 3HY7 [93] and 2W12 [94]) and so, the SMILES codes from each active compound were collected from public databases, such as ChEMBl [95] , PubChem [96] and DrugBank [97], and then used as input for the following processes.

A validation set was created by an online server, the Directory of Useful Decoys–Enhanced (DUD–E) [98]. DUD-E contains a large collection of decoys, i.e., compounds whose bioactivity is unknown, but are assumed to be inactive. Decoys are often used to assess the docking program's competence to distinguish active molecules amid a large number of inactive ones [98, 99]. These molecules possessed active-like physicochemical properties, but with different topologies. These properties included the number of proton donors and acceptors (HBDs and HBAs, respectively), the number of hydrophobic atoms, the molecular weight, the number of rotating bonds, and the logP. However, the development of a discriminative model between active and inactive can be difficult when the physicochemical properties are too close to the actives [99].

In general, when active and inactive compounds fit in the model, these hits are called true positives and false positives, respectively. On the other hand, when the model is unable to find active or inactive compounds, these hits are called false negatives and true negatives, respectively [99].

The set of SMILES codes retrieved in the previous step was used as input to generate decoys, and so, in this work, since 6 active molecules were selected, 300 decoys were generated, making up a 1:50 ratio of active/decoys, the recommended quantity [99].

The decoy set was collected as 1D representations, however, for the docking process the 3D structures of the compounds are needed to provide the geometry of the molecule with connection lengths and angles [99]. Thus, to convert one-dimensional (1D) structures into their three-dimensional (3D) form, MarvinSketch application, a chemical editor from ChemAxon tools (**https://www.chemaxon.com**) was used and hydrogens were added under a physiological pH of 7.4. Finally, the 3D information was saved in the SDF format, to be readable by the docking programs.

2.6.2. Validation of the VS protocol

In general, molecular docking software use search algorithms to find the most realistic conformation of the ligand at the binding site; and scoring functions to estimate the strength of interactions (affinity) between the target and the ligand. These are fundamental components for running VS [88].

However, despite being widely used and promising, the results of virtual screening and docking tend to vary according to the type of target and ligand molecule, which makes it difficult to accurately compare the performance of different docking software [88].

In this project, the performance of three of the most widely used docking tools, GOLD (version 5.2.2) [100], AutoDock Vina using AutoDock tools [101] and rDOCK [102], was evaluated.

2.6.2.1. AutoDock Vina

Both the target and the ligands input files were prepared using AutoDockTools 1.5.6. The search space was defined considering the zinc ion as its centre. Here, files were converted to PDBQT format from the PDB one, and Gasteiger charges were applied. Recently the AD4 scoring function was implemented in Vina, so the affinity maps with all the forcefield information were generated for the docking of zinc pseudo atoms, using the AutoDock4Zn forcefield. The search space (grid box) size was centered on the zinc atom of the active site where the ligand is docked and set at 40×30×50 with a default grid point spacing of 0.375 Å.

Input files generated in the previous step, affinity maps and the ligand, were used. Each run was made using an exhaustiveness of 32, with the scoring function ad4.

2.6.2.2. Gold

Regarding the GOLD software, the parameters were adjusted so that the ligand binding site was defined by a cavity detection, whose research sphere was centred on the zinc atom and expanded up to a radius of 15 Å. It allowed for a search space big enough to accommodate the ligands. Each ligand was docked 20 times with the default parameters of the genetic algorithm (GA), using different scoring functions, Chemscore, ASP, CHEMPLP and Goldscore, and allowing ligand's total flexibility, including flipping of amide bonds.

2.6.2.3. rDock

A cavity around the active site of the target protein was defined within 15 Å of the Zinc atom. Search algorithm and scoring function parameters were kept as their default values. Each ligand was docked 20 times.

The 6 active molecules were firstly re-docked and compared to its original PDBs to assess the performance of the programs. Then, the actives and respective decoys were re-docked. In all programs, the standard docking protocol was adjusted in order to reproduce the experimentally known interaction poses.

Ligands were classified according to their affinity for the target's active site, resulting in a list in which the most promising compounds appeared at the top.

2.6.2.4. Evaluation metrics

Gold, Autodock and rDOCK returned a list of the actives and their corresponding decoys, based on the obtained scores. Through these lists, it was possible to compare the performance of these docking programs in the discrimination between the actives and decoys for the target protein, by determining the values of the enrichment factor (EF). EF is the percentage of ligands found on x% of the total uploaded compounds. The average EF 1%, EF 2%, EF 10%, and EF 20% was determined for the 306 compounds.

2.6.3. Ligand-Based Virtual Screening (LBVS)

2.6.3.1. Generation of Pharmacophore Models

In the ligand-based approach, a pharmacophore model was generated considering the common chemical functionalities of the active molecules [99].

As mentioned on the topic *2.2.*, 6 X-ray crystallographic structures of the active ligands were obtained from PDB database and the missing hydrogens for the organic compounds were added using ChemAxon tools. These ligands were further used to identify the most representative features from metalloproteinase inhibitors, using the pharmacophore model generator, PharmGist [103].

PharmaGist is a webserver used for ligand-based pharmacophore modelling, based on submitted active ligands. The generation of the pharmacophore model was carried out according to four steps. First, the uploaded actives were processed individually with respect to physicochemical features such as rotating bonds, HBAs or HBDs, H groups, aromatic rings and positively or negatively groups. Second, the input ligands were aligned and one of them was defined as a pivot molecule, that is, as rigid, and all the others were flexibly aligned according to the former one. The alignment was done in such a way that each part of the ligands was superimposed on the pivot molecule, and thus common features between them were defined. Third, the alignments were divided into subsets in which the pivot molecule matched the maximum number of ligands. Finally, the alignments from the previous steps were clustered, sorted by score and the generated pharmacophore models covering the chemical space of the active molecules were given as output [99].

In an attempt to align 3, 4, 5 and 6 active ligands, the program generated 40 pharmacophore models, 10 for each. For each one was attributed a score. The best scoring models were selected and analysed on PyMol, taking into account putative interactions with the protein target and, at the same time, ensuring a not very restrict chemical space.

Finally, the chosen pharmacophore model was the one that brought together the properties of 4 of the ligands (PDB IDs: BAT, 097, WR2 and GM6) with the best scoring (33.833) and with 15 spatial features (7 hydrophobic areas, 4 donors and 4 proton acceptors).

The important features of the selected pharmacophore model were indicated as spheres (location constraint of each feature), that were representatives of essential interaction points.

2.6.3.2. Database Screening

The selected pharmacophore model was submitted as a 3D query to ZINCpharmer, an online server that allows a quick search for purchasable drug-like and FDA approved drugs by filtering the ZINC database [104]. Although the pharmacophore model possessed 15 spatial features, not all were used for the screening process, due to the difficulty of matching them all. Thus, some query features were omitted to increase the pharmacophore fit, remaining only 3 proton acceptors, 3 donors and 1 hydrophobic region. Furthermore, a tolerance radius from the centre of each feature of 2 Å was considered.

Taking those pharmacophore features into account, molecules that matched the requirements of the model were collected from a list of 338836 hits (a list of fitting compounds).

2.6.3.3. Characterization of hits for experimental validation

In the previous step, a list of 338836 hits was obtained, however, it was impossible to evaluate all compounds experimentally.

To characterize the hit list, the RDKit software (https://www.rdkit.org/) was used with additional filters based on molecular chemical descriptors, which allowed to easily characterize compounds with atypical chemical properties (number of aromatic rings, number of rotating bonds, molecular weight, among others). One of the molecular descriptors used was Lipinski's “Rule of Five” [105] , which describes properties shared by approved orally administered drugs, namely a molecular weight of ≤500, a number of HBDs ≤5, HBAs ≤10, and a cLogP ≤5. Also, the “Rule of Three” introduced by Congreve et al., 2003 [106] that define lead-like compounds was used, which selected compounds with a molecular weight of ≤300, a number of HBDs ≤3, a number of HBAs ≤3 and a cLogP ≤3.

2.6.4. Structure-based Virtual Screening (SBVS)

Structure-based virtual screening (SBVS) attempts to predict the most favourable interaction between the target protein and each compound. As such, the 3D structure of the target protein must be available [88].

2.6.5. Chemical Libraries Docking and Clustering

The most reliable docking program was used to explore both the LB hit compounds and the chemical library of around 3.5 million compounds which were docked against the target's active site to find hit compounds as potential leads. The top 1000 ranked compounds were selected.

In addition, instead of evaluating structurally very similar candidates, which are not relevant at this stage, structurally diverse hits were selected in order to explore a wider chemical space necessary for the candidate's biological activity on the target. Thus, the chemical descriptors of the compounds from the lists were determined using 2D descriptors like extended-connectivity fingerprints (ECFPs) and pharmacophoric fingerprints (PF). ECFPs are circular topological fingerprints, designed for similarity searching, and molecular characterization. These were obtained with the GenerateMD tool of the Chemaxon's JChem software. The ECFP were compared, and the degree of similarity was evaluated by computing the Tanimoto coefficient using a cut-off of 0.5. The clustering and selection of representatives was carried out using the non-hierarchical clustering method Jarvis-Patrick (JP, implemented in ChemAxon software), which is based on similarity between neighbours. This method classifies similar compounds to obtain a diverse subset.

Of the LB obtained compounds, 20 of the 50 representatives (given on the output) were selected for experimental evaluation. In the second method, SB, following the same procedure, another 30 compounds were selected for further in-depth analysis and experimental validation.

Chapter III - Results & Discussion

3.1. Homology Modelling

3.1.1. RVV-X structure

3.1.1.1. Template Selection and alignments

Considering that the structure of the protein of interest (RVV-X) was not experimentally determined, its prediction was carried out using *in silico* methods, as is the case of homology modelling.

The primary sequence of the protein of interest, RVV-X from *Daboia russelii russelii*, was searched on the UniProtKB database. The sequences of the RVV-X HC, LC1 and LC2 were retrieved (ACs: K9JAW0, K9JAX2, and K9JCB2, respectively).

According to the Uniprot database, RVV-X HC is synthesized as a precursor protein of 619 amino acids in the venom gland, including a highly conserved signal peptide and pro-domain. This precursor is subsequently processed to produce a mature and active protein of 431 residues. The LC2 which corresponds to the “Factor X activator light chain 2” (K9JCB2) on the UniProtKB database, has 135 amino acids in its mature form. The LC1, which corresponds to the “Factor X activator light chain 1” (K9JAX2), contains 123 amino acids in its mature form. Unlike LC2, LC1 is not linked to HC.

As sequence similarity is one of the main important parameters to derive structural and functional information for the protein of interest, the retrieved sequences were used as query input for the homology modelling method on the SWISS-MODEL webserver.

Modeling results on the SWISS-MODEL webserver showed that the crystal structure with 2.91 Å resolution of the RVV-X from *D. Siamensis* venom was very similar to the target with a 97,64% identity, 87% coverage and a normalized QMEAN Z-score of -0.68. It was crystallized with Ilomastat inhibitor (PDB ID: GM6), Zn^{2+} (required for the target’s catalytic activity) and Ca^{2+} co-factors and oligosaccharides such as N-Acetylglucosamine (NAG) and N-acetylmuramic (NAM). The catalytic histidines coordinate the zinc ion, which is an extremely important trait as the catalytic activity of the protein depends on it. *Figure 20* represents the alignment between the target and the template heavy chain sequences.

From *figure 20*, it was possible to depict that the primary sequence is highly conserved between the two species, having undergone only a few changes between few amino acids. So, the key structural elements involved in the bioactivity of the protein were intact.

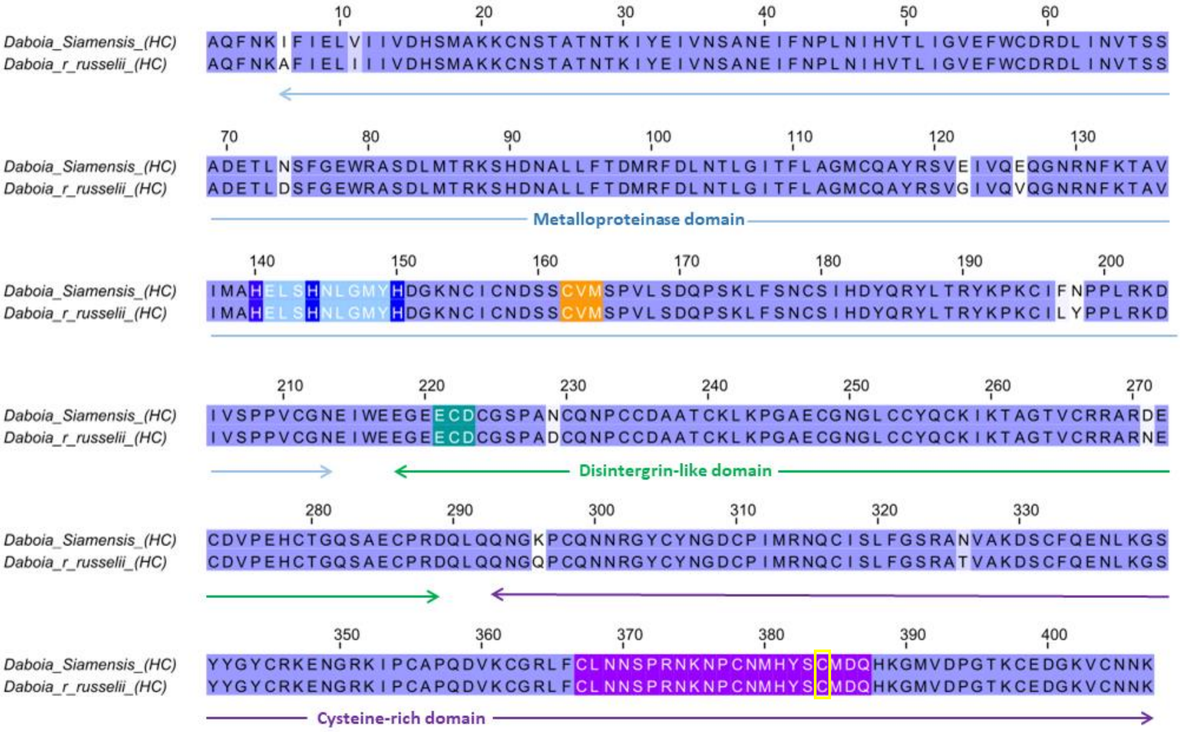


Figure 20 - Sequence alignment between RVV-X heavy chain from *Daboia siamensis*, used as template, and the target (Daboia_r_russelii). Important features are represented: zinc-binding site (light blue box) with its three histidines (dark blue), the conserved met-turn (orange box), the disintergrin domain (dark green box) and the hypervariable regions from the Cysteine-rich domain (purple box). The residue Cys384 from the HC (yellow box) located within the hypervariable region (HVR) forms an interchain disulfide bond with the Cys133 at the extended C-terminus of the LC2. The light purple boxes represent the residues that possess an identity percentage above 60%. Sequence alignment made using Jalview.

Thus, the global structure quality of the predicted model was good, meaning that this model is reliable for further analysis. For that reason, the X-ray crystal structure of RVV-X from *D. Siamensis* published by Takeda et al., 2007 [38] (PDB ID: 2E3X) was selected as a template, for the modelling of *D. r. russelii* RVV-X structure.

3.1.1.2. Model building and validation

The superimposition of the backbone of both the template and the target revealed remarkable similarities with a RMSD difference of 0.105 Å, except for an unmodeled loop (residues Gly60-Ile63), from the template's LC2. The surrounding region of this loop was also different in both structures. These results confirmed once again the reliability of the model. As such, the full structure of RVV-X was built by superimposition with the template chains (Fig. 21). The superimposition of the catalytic sites of both the target and template is shown in *figure 21*.

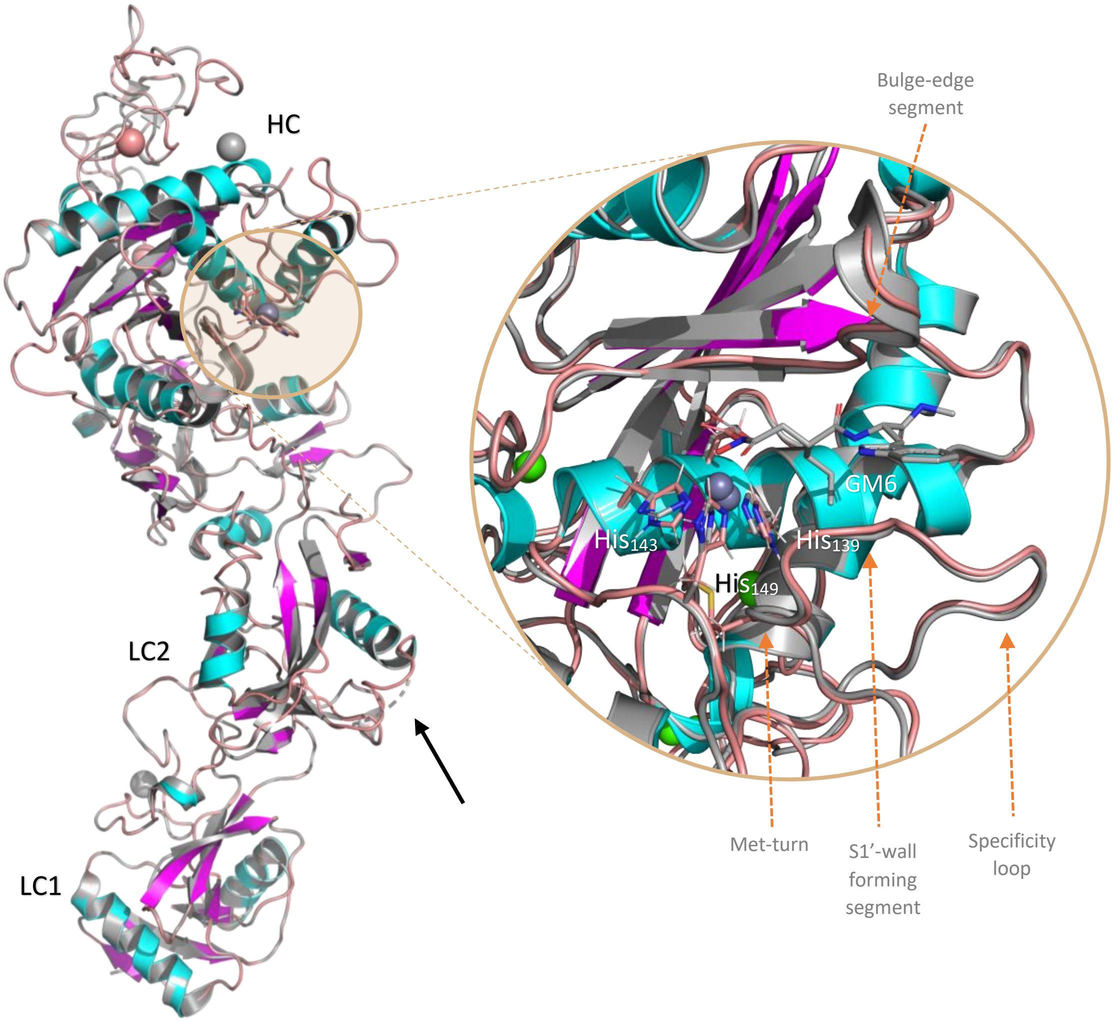


Figure 21 – Structural superimposition of the target (blue helixes, magenta b-sheets, and pink loops) and the template (gray). The arrow represents the unmodeled loop from the template's LC2. New cartoon representations were made using Pymol software. A closeup view of the catalytic centre: zinc coordinated by three histidines (sphere and sticks representations, respectively) in complex with the inhibitor, GM6 (stick representation). Orange arrows represent the structural elements important for the recognition/binding of the substrate/inhibitor.

The 3D model's HC was organized in four long α-helix, one of which represents the active site, and four-stranded β-sheet, but lacked the common antiparallel β-sheet (β4). The catalytic zinc ion tetrahedrally co-ordinated by His139, His143 and His149 and a putative water molecule (later modelled) was at the bottom of the flat active-site cleft, where the inhibitor was located.

Within the active centre there are several important elements for the recognition and binding of the substrate/inhibitor (*Fig. 21*), namely the bulge-edge segment (residues Gly107-Ala112) and a protruding loop segment between the α-helix (αB) and the β-sheet (β3) (Gln127-Lys133). These two structural elements make up the "upper rim" of the cleft of the active site, bordering, in turn, the primed and non-primed subsites. Finally, still in the active centre, there was also a segment (Pro165-Leu168) in the specificity loop that is commonly termed the “S1’-wall forming segment”.

The binding of the inhibitor (GM6) occurs in the groove that extends along the active site cleft, “S1’-wall forming segment” and the “bulge-edge segment”. Here, strong electrostatic interactions occur with the Zn ion and vast Van der Waals contacts occur within the specificity pocket (S1’).

The C-type lectin-like subunits (snaclecs), which constitute the LC1 and LC2 formed an intertwined dimer.

The catalytic water was manually modelled between the metallic ion and the catalytic glutamate residue which in turn polarizes the water for protein’s activity. This modelling was carried out considering the position of the inhibitor and substituting it (*Fig. 22*)

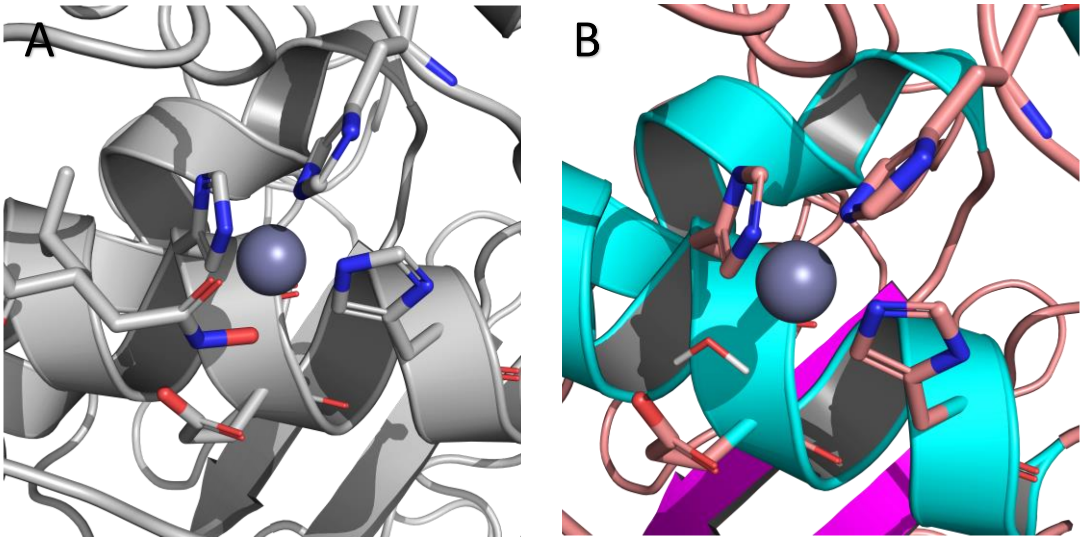


Figure 22 - Catalytic centre. (A) the catalytic centre of the template co-crystallized with the inhibitor, GM6. (B) the catalytic centre of the model with the catalytic water manually modelled.

The predicted model was further validated by several tools available. Its psi/phi torsion angles were checked using the Ramachandran plot form PROCHECK webserver (*Fig. 23*).

According to the results shown on *figure 23*, 84% of the residues were in favoured regions and 14% in allowed regions. However, 4 residues (0.7%) had Phi/Psi distribution in the disallowed region of the Ramachandran plot of both predicted model and the template. Given the low percentage of residues found in the disallowed region, the Ramachandran plot fulfilled the validation criteria.

The model showed good quality, however it has been submitted for further quality validation.

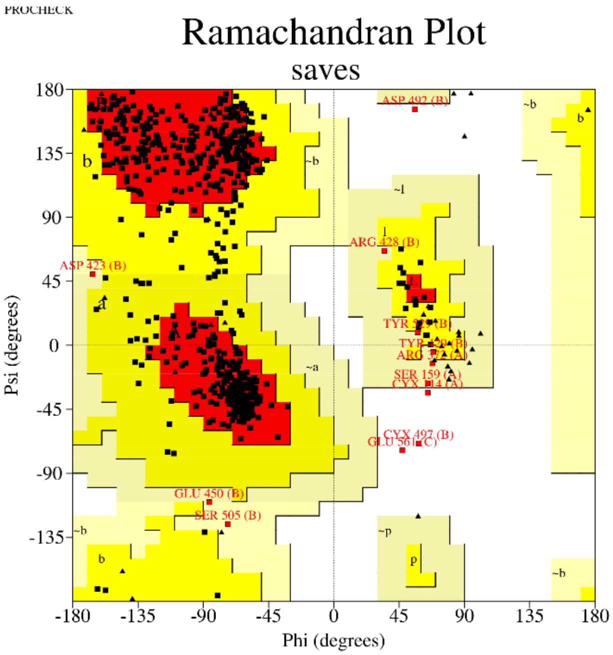


Figure 23 - Ramachandran plot of the RVV-X model. Colour red represents the most favoured regions; yellow, light yellow and white colours represent additional allowed, generously allowed, and disallowed regions, respectively.

ProSA webserver, another validation tool, compares the input structure with other experimentally determined structures of similar size, and checks if it is within a range of scores attributed to the latter. ProSA results showed almost identical Z-scores between the model's and template's heavy chain (-9.17 and -9.18, respectively) and first light chain (-5.25 and -5.20, respectively). Moreover, for the second light chain there was a small difference of 1.23 between the two (*Fig. 24*). ProSA also calculates knowledge-based energy profiles, and results showed that most of the residues that suffered bigger deviations on the model structure had an almost identical pattern with the template, further confirming its structural similarity with the crystal structure.

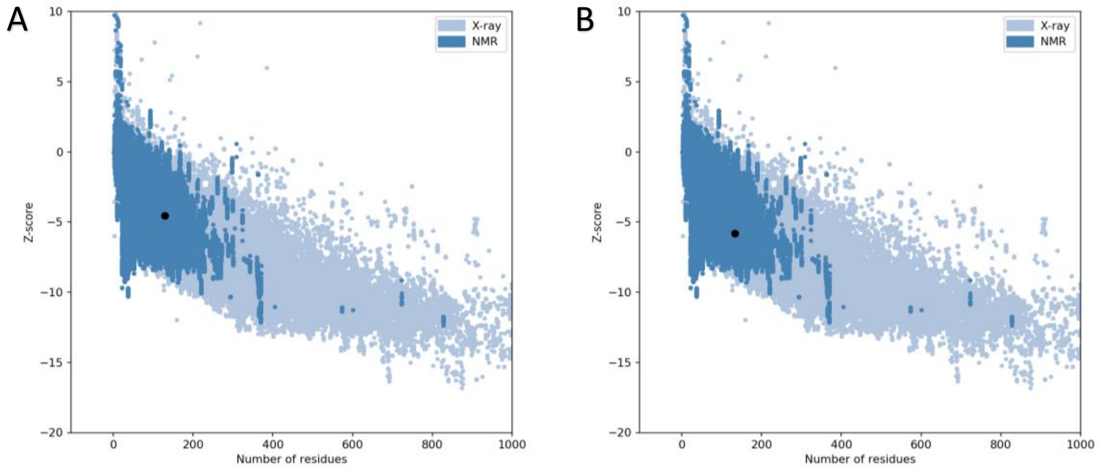


Figure 24 - ProSA Z-score (black spot) is −4.56 and -5.79, for the template (A) and for the mode (B), respectively. These values are within the range of the experimentally determined proteins by X-ray (light blue) and NMR (dark blue).

These validation tools allowed the confirmation of the low quantity of structural variations, and therefore the quality of the homology model, allowing to continue the study.

3.1.2. FX structure

3.1.2.1. Template Selection and alignments

The sequence of the human coagulation FX was obtained from UniprotKB database (AC: P00742), which contained 488 amino acids of which 40 corresponded to the signal peptide and to the propeptide domain. *Figure 25* represents the sequence of the FX mature protein (length: 448 residues).

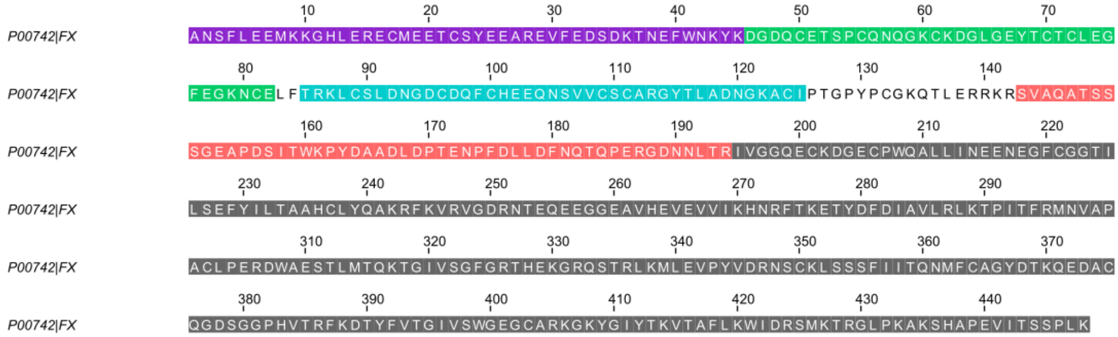

```
                            10          20          30          40          50          60          70
P00742|FX      ANSFLEEMKKGHLERECMEETCSYEEAREVFEDSDKTNEFWNKYKDGDQCETSPCQNQGKCKDGLGEYTCTCLEG
                       80          90         100         110         120         130         140
P00742|FX      FEGKNCELFTRKLCSLDNGDCDQFCHEEQNSVVCSCARGYTLADNGKACIPTGPYPCGKQTLERRKRSVAQATSS
                           160         170         180         190         200         210         220
P00742|FX      SGEAPDSITWKPYDAADLDPTENPFDLLDFNQTQPERGDNNLTRIVGGQECKDGECPWQALLINEENEGFCGGTI
                       230         240         250         260         270         280         290
P00742|FX      LSEFYILTAAHCLYQAKRFKVRVGDRNTEQEEGGEAVHEVEVVIKHNRFTKETYDFDIAVLRLKTPITFRMNVAP
                           310         320         330         340         350         360         370
P00742|FX      ACLPERDWAESTLMTQKTGIVSGFGRTHEKGRQSTRLKMLEVPYVDRNSCKLSSSFIITQNMFCAGYDTKQEDAC
                       380         390         400         410         420         430         440
P00742|FX      QGDSGGPHVTRFKDTYFVTGIVSWGEGCARKGKYGIYTKVTAFLKWIDRSMKTRGLPKAKSHAPEVITSSPLK
```

Figure 25 - Sequence of coagulation factor X (UniProtKB AC: P00742). FX domains are highlighted with different colors: Gla domain (purple), EGF1 (green), EGF2 (blue), AP (light red) and its heavy chain (dark gray).

As there was no experimental structure available for this protein either, the primary sequence retrieved previously was used for the search of a FX template. 100 hits were obtained with high identities, of which 23 had sequence identities above 90%, but none had all the domains neither all the residues that constitute FX.

Therefore, the search was made domain by domain using the same resources. In the end, 4 crystal structures (PDB ID: 2XC0, 1XKA, 2JKH and 1IOD) were selected to be used as templates for each domain of the FX. All the structures except for the one that represented the gla domain (PDB ID: 1IOD), which comprised a complex between the coagulation factor X binding protein from snake venom and the gla domain of bovine FX, were from activated FX (FXa). Hence, no structure of the FX zymogen with its activation peptide was found.

Despite sharing different primary sequences, serine proteases that are involved in the blood coagulation cascade have a very similar global folding and share a common catalytic triad. Thus, the SP domain of the activated FX (PDB ID: 2XC0) was compared with zymogen forms of SPs available in the PDB.

An active serine protease must have a substrate binding pocket (S1), a specific positioning of catalytic triad residues and a well formed oxyanionic hole, which accommodates the carbonyl oxygen of the substrate. Whereas, the zymogen usually has a hidden binding pocket or a malformed oxyanionic hole, and 3 loops (281-291, 330-339, 356-369) acquire a different conformation interfering with the catalytic region [107]. Through this comparison and based on the literature, it was possible to understand which of the searched structures fulfilled the requirements to be considered zymogens for further modelling.

Therefore, after analysing some zymogen structures available on PDB, the crystallographic structure of the Chymotrypsinogen A (PDB ID: 2CGA), with a resolution of 1.80 Å, published by Wang in 1987 [63], was used as template for the modelling of the zymogen form of FX SP domain. The alignment of the backbone atoms of the SP domain from Chymotripsinogen A and from FXa (PDB ID: 2XC0) revealed significant conformational differences in several solvent exposed loop regions, namely the activation loops, though the core region of both structures was similar with an overall RMS difference of 0.79 Å for the backbone atoms.

As there were still missing residues, including the activation peptide, Modeller 9v10 software was used to model each domain considering the alignment between their primary sequence and each template. The first output with the best scoring was chosen for each model.

3.1.3. Model Building and Validation

The chosen models were superimposed with its corresponding templates using PyMol and the assembly of the full FX structure was carried out manually. The modelled protein is depicted in *Figure 26*.

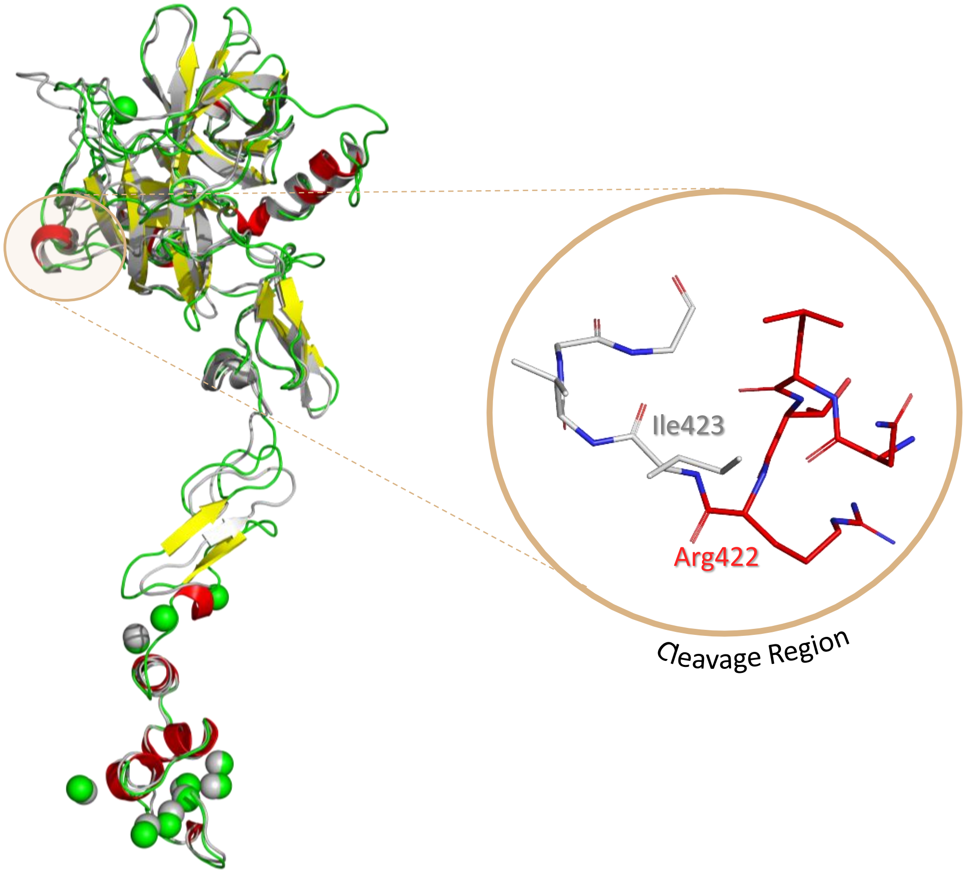


Figure 26 - Structural superimposition of the target (red helixes, yellow b-sheets, and green loops) and the templates (gray) – PDB ID: 2CGA, 2JKH, 1XKA and 1IOD. New cartoon representations were made using Pymol software.

The coagulation factor X HC is composed of two β-barrel subdomains with six antiparallel strands each, in the middle of which is the active site. Moreover, it contains four helices, of which two were missing in the homology model. This chain was followed by the LC, which contains two EGF-like domains and the N-terminal Gla domain. The EGF1 domain possessed two antiparallel β-hairpin motifs, and the EFG2 had only one. Between the two EGF domains, there was an inter-domain linker region, which might offer flexibility required for a proper orientation and folding, thus facilitating the docking process to the RVV-X. The gla domain contained seven Ca^{2+}-binding sites, with an additional one at the C-terminal helix region.

The PROCHECK Ramachandran plot represented on *figure 27* showed that 76.1% of the residues were in the core region, 19.9% in additional allowed, 3.4% in generously allowed and 0.6% in disallowed regions.

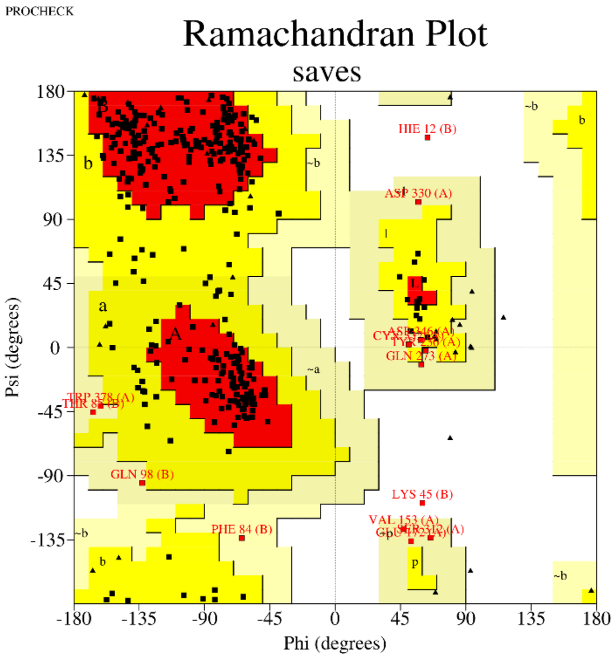


Figure 27 - Ramachandran plot of the FX model. Colour red represents the most favoured regions; yellow, light yellow and white colours represent additional allowed, generously allowed, and disallowed regions, respectively.

ProSA plot showed a Z-score of -5.59 and -3.84, for the heavy and light chains, respectively.

In summary, although there may exist some variations due to the absence of a complete crystallographic structure, the gathered results confirmed the quality of the homology model.

3.2. Geometry Optimization

The systems were prepared to create its topological parameters using tLeap. The returned outputs from both the RVV-X and the FX preparation are represented in *Annexes 1.A and B*.

After obtaining the topology and coordinates files for each model, the energy minimization process itself began.

3.2.1. RVV-X Refinement

Through the observation of *figure 28*, it was possible to verify that the overall folding of the model, did not suffer major conformational changes during the minimization process. What stood out the most was the positioning of certain counterions (Cl^-), as well as some loops that underwent few changes.

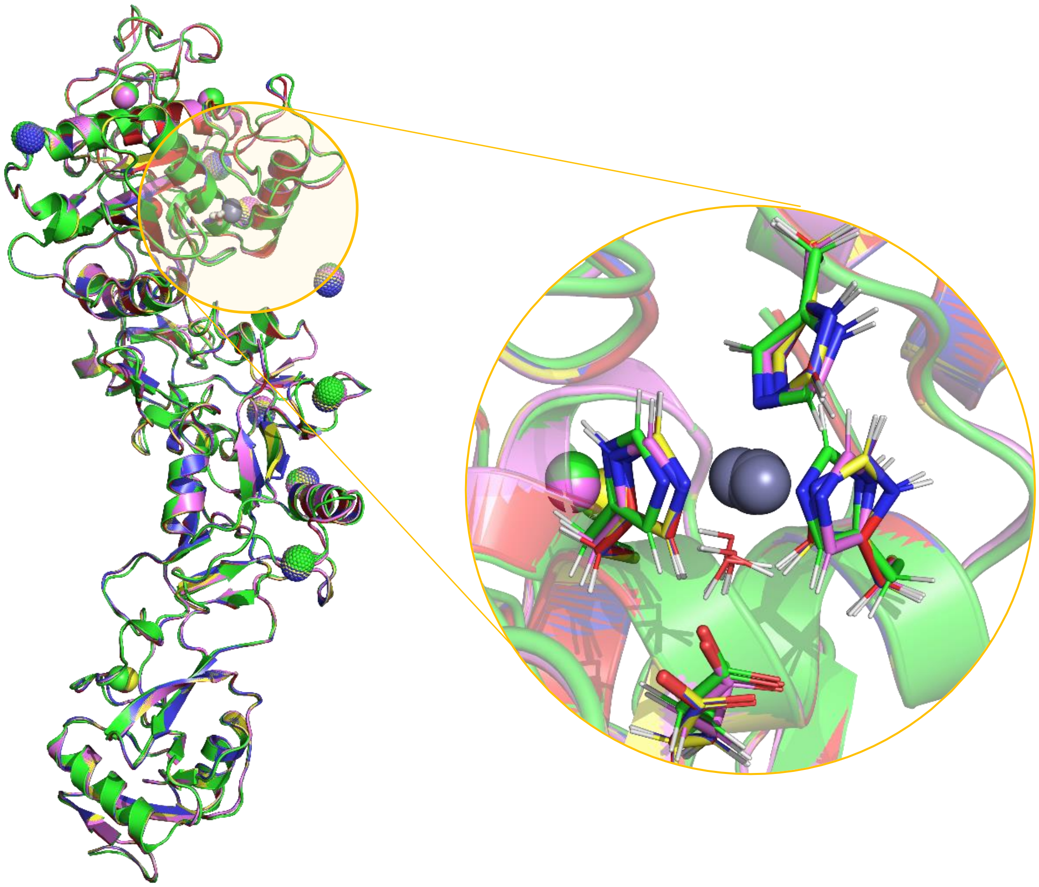


Figure 28 - Superimposed structures of the model (red) and respective minimizations, min1 (blue), min2 (yellow), min3 (violet), min4 (green) with a New Cartoon representation, using PyMOL. Chloride ions are represented by dots while calcium atoms are represented by solid spheres. Closeup view of the catalytic centre with its zinc (dark grey sphere) coordinated by the three histidines (sticks) and the water molecule (lines).

Considering that the chemical reaction takes place in the MP domain, the studies that followed focused only on this domain. Thus, the model MP domain was then subjected to molecular dynamic studies over a simulation time of 50 ns. This step allowed a full refinement of the system and to gain insight into its stability.

The RMSD profile, which evaluates the equilibrium and stability of the system under study is shown in *figure 29.* At the beginning of the simulation the RMSD backbone of the protein increased, converging after 30 ns.

The modelled protein exhibited a RMSD average value of 1.97 Å for the backbone, which indicated a stable dynamic behaviour of the structure as well as the equilibration state of the protein.

Furthermore, as shown in *figure 29*, the RMSF values of the residues, another stability indicator, ranged between 0.4 and 2.9 Å. Higher atomic fluctuations were mainly observed in the loops from the Dis-like and Cys-rich domains of the metalloproteinase. A sharp peak was observed around residues 369-376, which belong to the hypervariable region of the Cys-rich domain, known for its role on the recognition of the substrate.

At the N- and C- terminal loop regions, the RMSF values were higher, as these areas have propensity to undergo rearrangements.

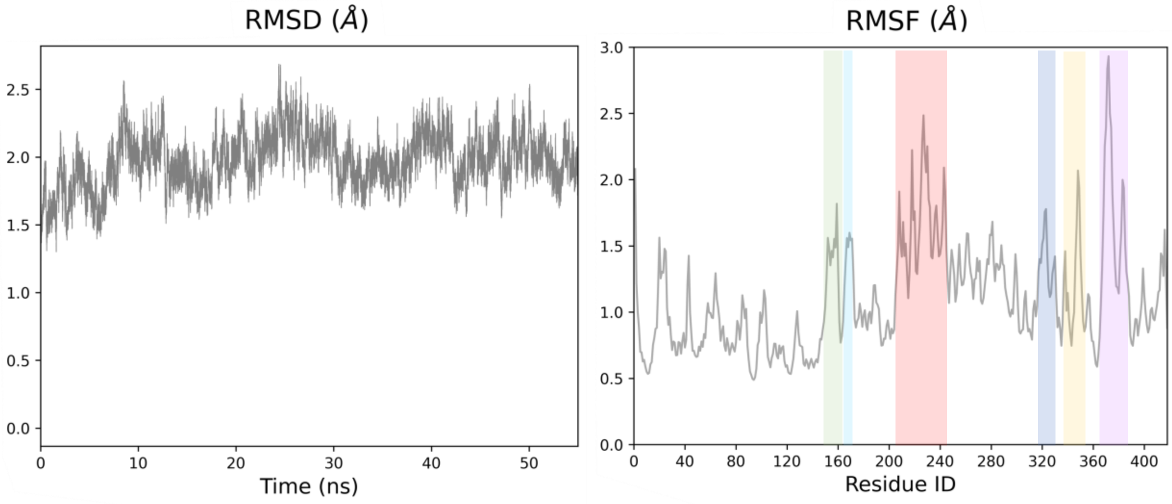


Figure 29 – Trajectory analysis: (left) RMSD and (right) RMSF of the protein's atoms. High and low atomic fluctuations with respect to the average position are represented by higher and lower RMSF values, respectively.

Thus, the created model provided a reliable base for the following analyses.

3.2.2. FX Refinement

A visual inspection of the superimposition of the minimized structures with the model, confirmed that the system did not suffer significant changes in the protein backbone. As expected, the position of the chloride counter-ions suffered little changes after the final minimization.

3.3. Protein-peptide Docking

The PDB file of the substrate with the last 4 residues of the activation peptide (Asn-Leu-Thr-Arg) and the first 4 N-terminal residues of the SP domain (Ile-Val-Gly-Gly) is represented in *figure 30*.

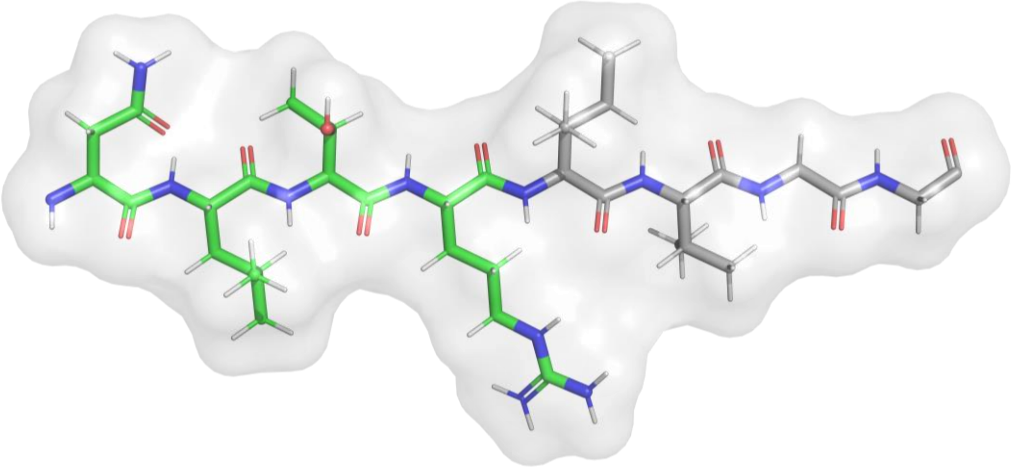


Figure 30 – Peptide created with the last 4 residues of the activation peptide (Asn-Leu-Thr-Arg) and the first 4 N-terminal residues of the SP domain (Ile-Val-Gly-Gly) to be docked on the receptor. Sticks and surface representation, using PyMOL.

The protein-peptide docking process was calculated using the HPepDock online software. This software generated 10 peptide docking models (poses) ranked accordingly to the lowest binding energy. These were visually inspected using PyMOL software and the first 4 best ranked peptide poses were selected. After analysing the protein-peptide interaction mechanism, it was possible to verify that 2 of the poses, which corresponded to opposite orientations of the peptide, had clearly the most favourable binding position in the binding site (*Fig. 30*). Furthermore, the results showed that the best ranked model between the selected ones, was the conformation 2, with a docking energy score of -795.03, while conformation 1 had -680.20. The two selected poses are presented on *figure 31*.

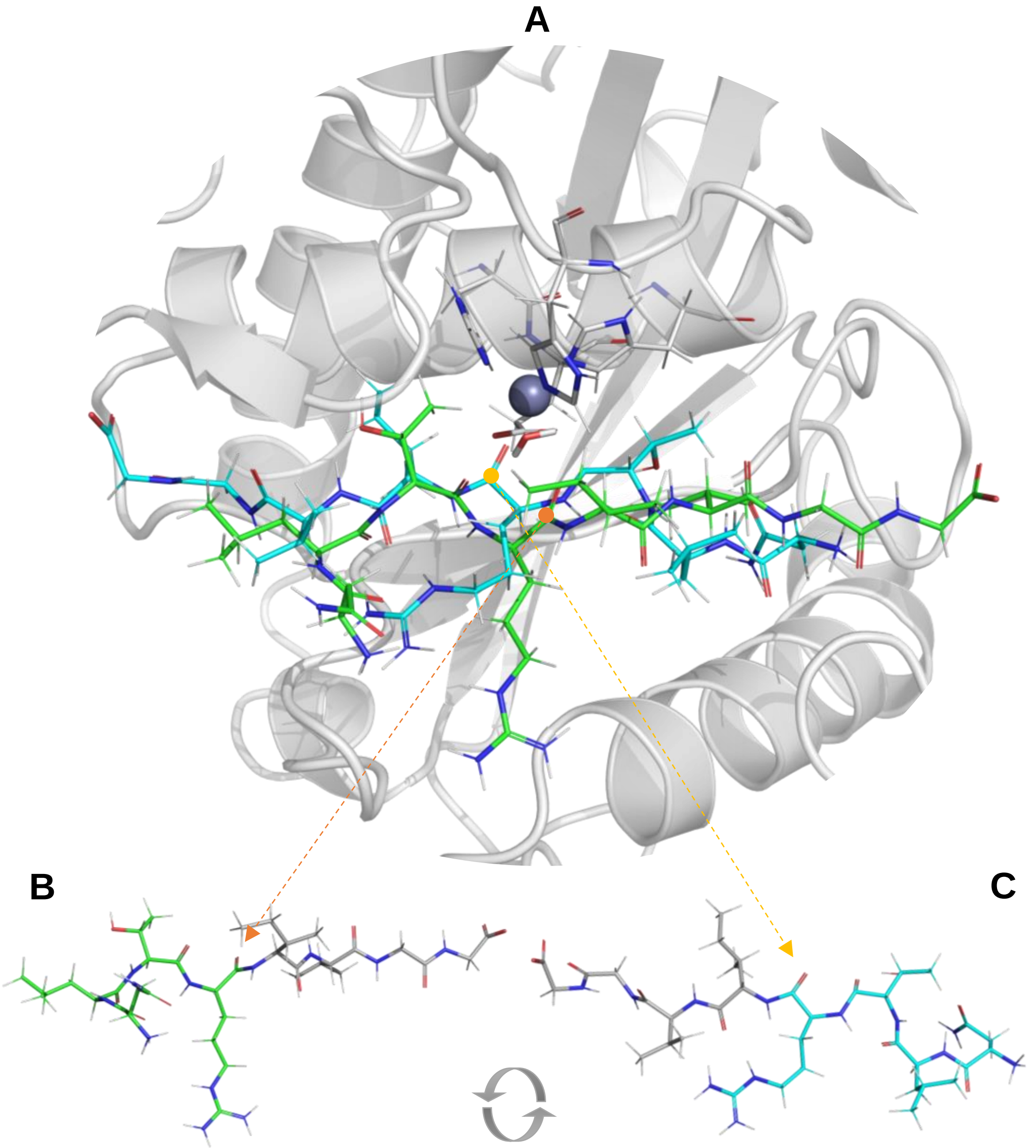


Figure 31 – (A) Superimposition of the predicted poses for conformation 1 (green sticks) and conformation 2 (cyan stick) in the binding pocket of RVV-X protein (grey cartoon) represented by PyMOL software. Amino acids from the peptides that interact with the catalytic zinc are labelled and represented by arrows (orange and yellow arrow, for conformation 1 and 2, respectively). (B) Conformation 1 with the reversed orientation. Green sticks correspond to the activation peptide residues, while grey sticks correspond to the amino residues of the FXa domain; (C) Conformation 2 with orientation in accordance with literature.

The preliminary results of the docking process allowed to observe slight readjustments on loops of conformation 1, namely before and after the highly conserved segment — Met-turn — as well as in the hypervariable region (HVR), both related to the recognition of the substrate (*Fig. 32*). Moreover, there is a difference of 0.43 Å between conformation 1 (Model 1) and the structure that was used as input for HPepDock. In contrast, conformation 2 (Model 2) seemed to exhibit greater stability, since there were no changes in the protein backbone, with a difference of only 0.09 Å regarding the input structure. In addition, the scissile carbonyl carbon to be cleaved from the conformation 1 was at a greater distance from the zinc (3.9 Å) than in conformation 2 (1.8 Å).

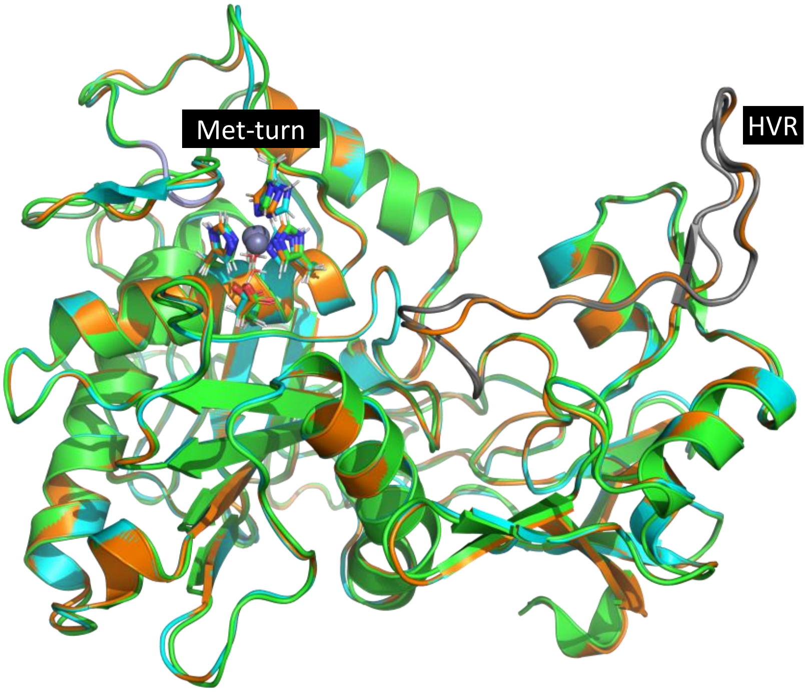


Figure 32 - Changes in the structure of proteins resulting from the peptide-protein docking can be observed through the superimposition of the models represented by PyMOL software. Conformation 1 model (green), Conformation 2 model (cyan), model (gray). Gray loop on the right represents the Met-turn, while the one on the left represents the HVR.

As represented on *figure 33*, the zinc ion from conformation 2 is well coordinated and the backbone of the cleavage residue — Arg422 and Ile423 — were correctly oriented to the cleavage process. This was not the case for the conformation 1, where the orientation of the cleavage residues does not allow the reaction to occur, being too far away from the catalytic glutamic. During the reaction, Glu140 acts as a base, deprotonating the catalytic water (Wat418) and transfers the proton to the key nitrogen (Ile423) of the scissile peptide bond.

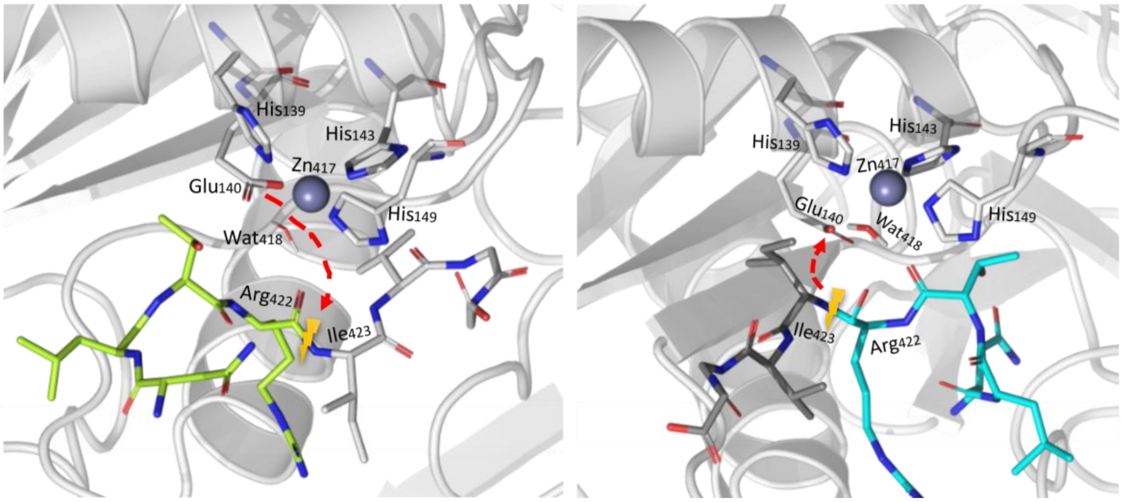


Figure 33 - The docking-predicted binding mode of model1 (left) and model 2 (right) peptides.

Takeda et. al. in 2007 determined the crystallographic structure of the *D. Siamensis* RVV-X, which served as a template for the enzyme of the present study and made a rigid docking model of the active FX, FXa, in the RVV-X. Curiously, the orientation of one of the resulting peptides (conformation 1) was in a reversed orientation when compared to the model proposed by Takeda in 2007 (*Fig. 34*). According to Takeda's model, the peptide would have same orientation as the peptide-like inhibitor GM6 [38], which was seen in conformation 2.

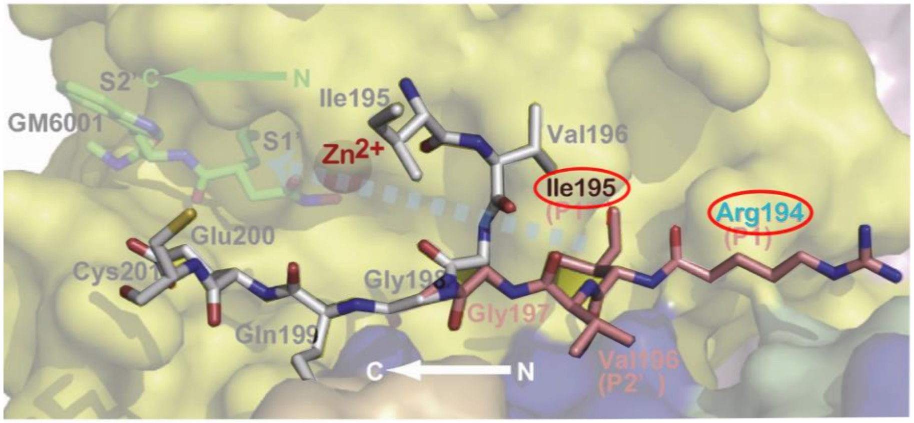


Figure 34 - Close up view of the docking model of RVV-X with both the FX and the FXa. The N-terminal residues of factor Xa are shown in white colour and those of the factor X (zymogen) are shown in light pink. Adapted from [38] and edited using Inkscape software.

This provides a first insight into the suitability of the binding poses. It seems that the model 2 which agrees with orientation proposed by Takeda in 2007 (*Fig. 34*), is more suitable. Despite having the same orientation, the Takeda model did not consider the pre-reactive conformation. In the Takeda's model, the FXa Ile of the scissile peptide bond is 16 Å apart from the catalytic Zn ion. Instead, the models obtained in this work display distances that could facilitate the catalysis to occur as previously mentioned.

Although these complexes provided a starting point for the development of this study, the conformational rearrangements that occur over time were not revealed. Thus, as these were static systems, it was necessary to resort to MD simulations for a deeper understanding of the structural adjusts caused by the docking of the peptide.

3.4. MD Trajectory Analysis

To understand the overall dynamics of the models, and to figure out which was the most suitable for reaction to occur, its structural changes were followed up over 170 ns (model 1) and 200 ns (model 2) of simulation.

To this end, several parameters were computed, such as the root-mean-square deviation (RMSD) to analyse the conformational stability, the root-mean-square-fluctuation (RMSF) to analyse the atomic fluctuations, and key distances to analyse the interactions of the active site residues with the peptide. These MD trajectory analyses were performed using the CPPTRAJ module of the AmberTools18 package and data was plotted using Matplotlib.

Moreover, both PyMOL and VMD programs were also used to visualize the resulting trajectory verifying whether the secondary structures of the protein-ligand complexes underwent major conformational changes during the simulations.

3.4.1. Structural analysis

Conformational changes can be elucidated by inspecting the secondary structure. After MD simulation, the five replicates of each model were superimposed. The superimposed structures are shown in *figure 35*.

During structural and conformational inspection, it was possible to infer that the system's backbone did not undergo significant changes.

It was also possible to verify that the active centre region remained rather stable throughout the simulations.

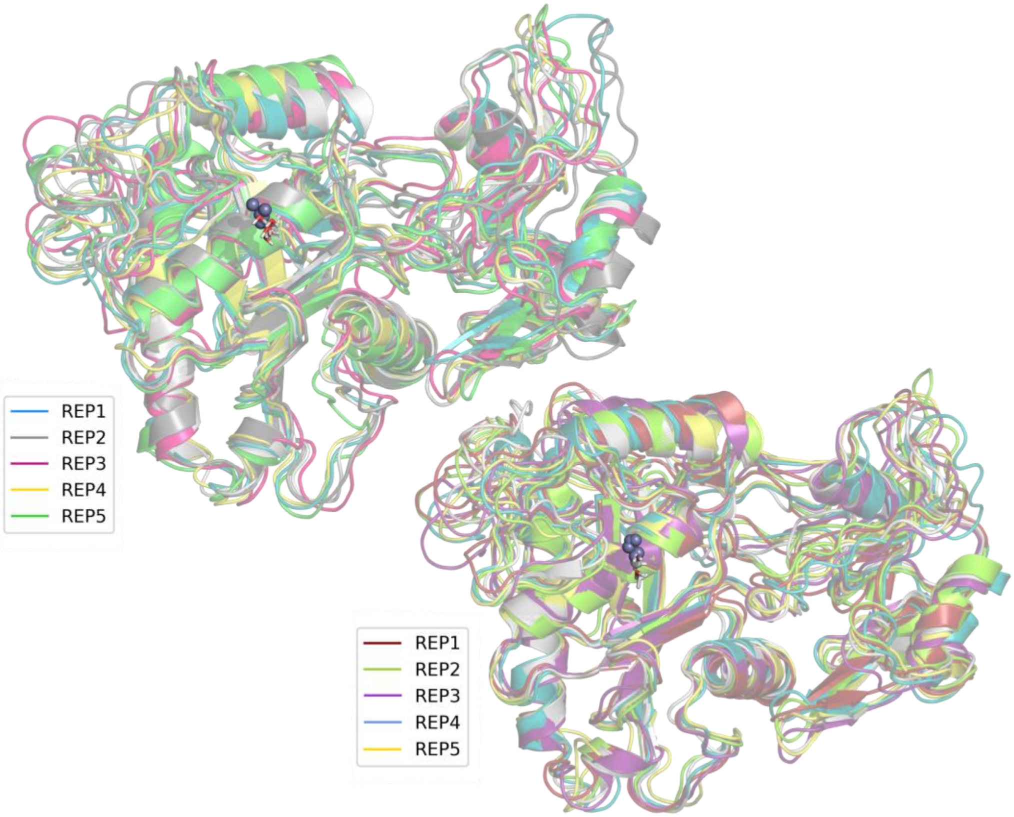


Figure 35 – Superimposition of the MD trajectory replica's last frame from model 1 (left) and model 2 (right). Protein structure was rendered as a 80% transparent new cartoon representation, the zinc ions as spheres and the water molecules as thick lines. Each protein color corresponds to a replicate.

The amino acid side chains as well as poorly defined regions such as loops and terminal chains exhibited larger variations due to higher conformational degrees of freedom. These deviations are mainly due to the presence of the solvent and of the peptide (not represented) since these interactions lead to slight changes on the structures. Conversely, the α-helices and β-sheets remain quite rigid during the simulation time due to strong hydrogen bonding, which limit structural fluctuations.

Thus, the secondary structures remained relatively identical throughout MD trajectory, while the loops, as would be expected, experience some flexibility.

Regarding the peptide, some rearrangements were observed at the side chains which underwent some rotations, in an attempt to relax its interactions with the enzyme (*Fig. 36*). Moreover, it was possible to verify that the peptide has moved from the catalytic centre in all the replicates. The case that stands out the most was the peptide in REP5 of model 1, which completely dissociated from the protein.

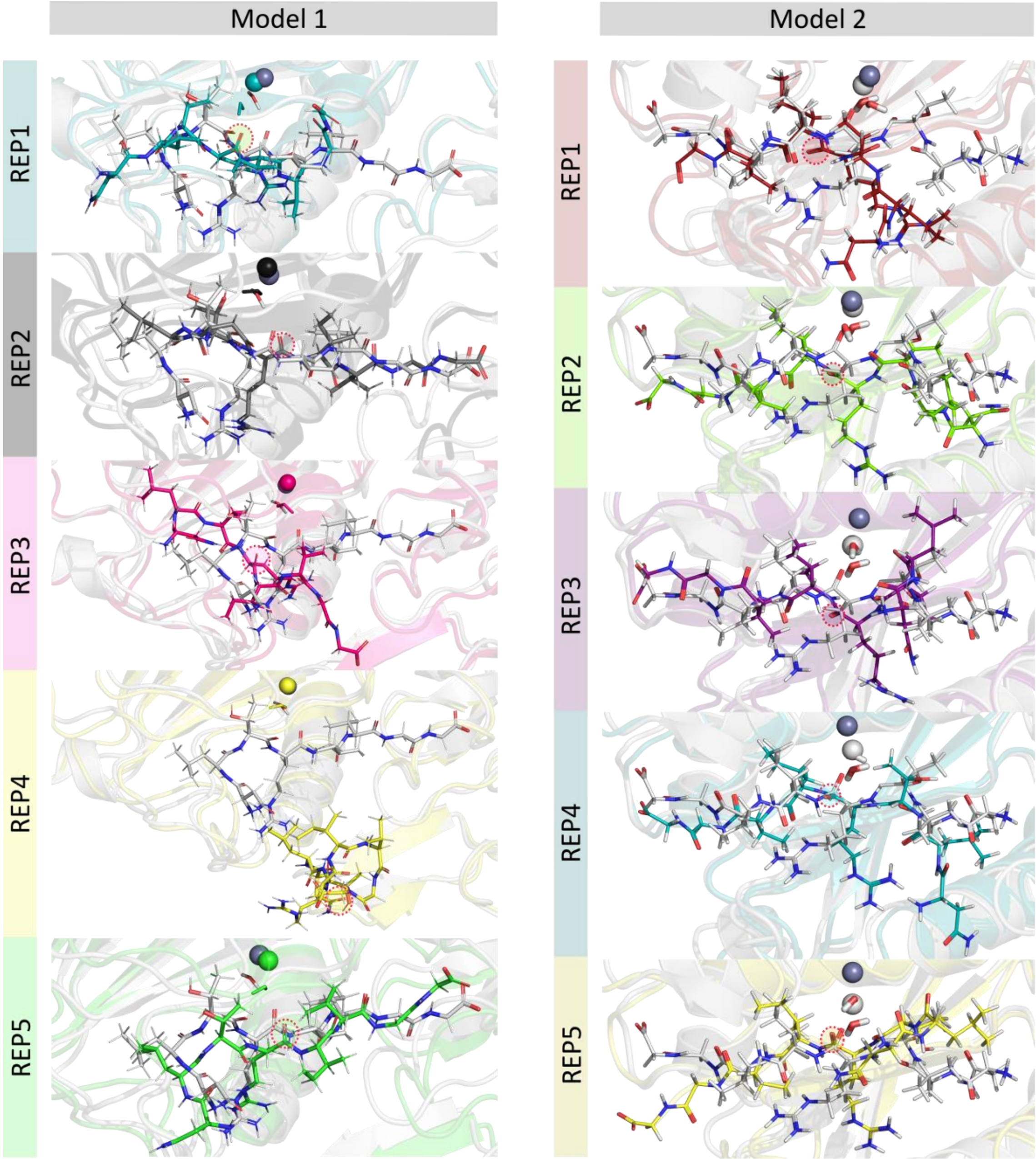


Figure 36 – Stick representation of the peptide superimposition of each replica's last frame with the original complex. The original complex of both systems is represented by transparent light grey new cartoon.

It should also be noted that in all the replicates the oxygen atom of the Arg422 carbonyl was pulled away from the zinc atom, and at the same time, suffered a rotation that impeded both atoms to point to each other (represented on *figure 36* by red dotted spheres). This can be a result of small readjustments in the residues that surround the Zn metal.

In 2002, Pelmenschikov and Siegbahn [108] demonstrated the importance of an assisting water molecule for the chemical reaction besides the catalytic one. This auxiliary water would act as an electrophilic agent to the carbonyl oxygen atom of the scissile peptide bond, stabilizing it.

At this stage, it was not yet known whether there was any auxiliary water near the catalytic site. Therefore, to see if any water entered the catalytic core along the trajectory, occupancy water maps were computed using the VolMap tool. Here, a threshold of 50% was used for each replicate, meaning that the water needs to remain at the same position at least during half of the time of the trajectory simulated. In only two replicates of model 2, REP2 and REP5, waters were found in the vicinity of the catalytic core, and none was found in any of the replicas of model 1.

In REP2 and 5 there was a water close to the catalytic Glu in 60% of the trajectory, which may have a stabilizing function towards Glu140 and Thr108, however no auxiliary water was found near the oxygen atom of the carbonyl group during a significant period of time.

3.4.2. RMSd

Despite not observing large conformational changes during the visual inspection over the MD trajectory, a quantitative analysis was carried out to assess their structural stability by analysing the backbone RMSd. To see how the complexes evolved, the starting structure of both systems was used as a reference for the calculations. The RMSDs of the simulated complexes are displayed in *figure 37* as a function of time.

The plots represented on the left (*Fig. 37*), confirmed the stability of all complex sets where the backbone RMSd increased at the beginning of the simulation (approximately 1.8 Å), after which it converged at approximately 20 ns, reaching a plateau. The data agreed with the backbone RMSd obtained for the initial RVV-X model, in absence of the ligand, which suggest that the presence of the substrate does not cause major changes in the conformational state of the protein. Therefore, it can be inferred from the plots that both orientations were stable at the protein level, not suffering high deviations, with an average RMSD trajectory value of 1.82 Å and 1.80 Å for model 1 and model 2, respectively.

The relative mobility of the peptide's backbone within the active site was also evaluated through RMSd calculations (*Fig. 37* right). The average RMSd trajectory values were 2.3 Å and 2.2 Å for model 1 and model2, respectively. These RMSd values can be attributed to more flexible regions as is the case of the amino and carboxyl terminal portions of the peptide which experience higher rotations.

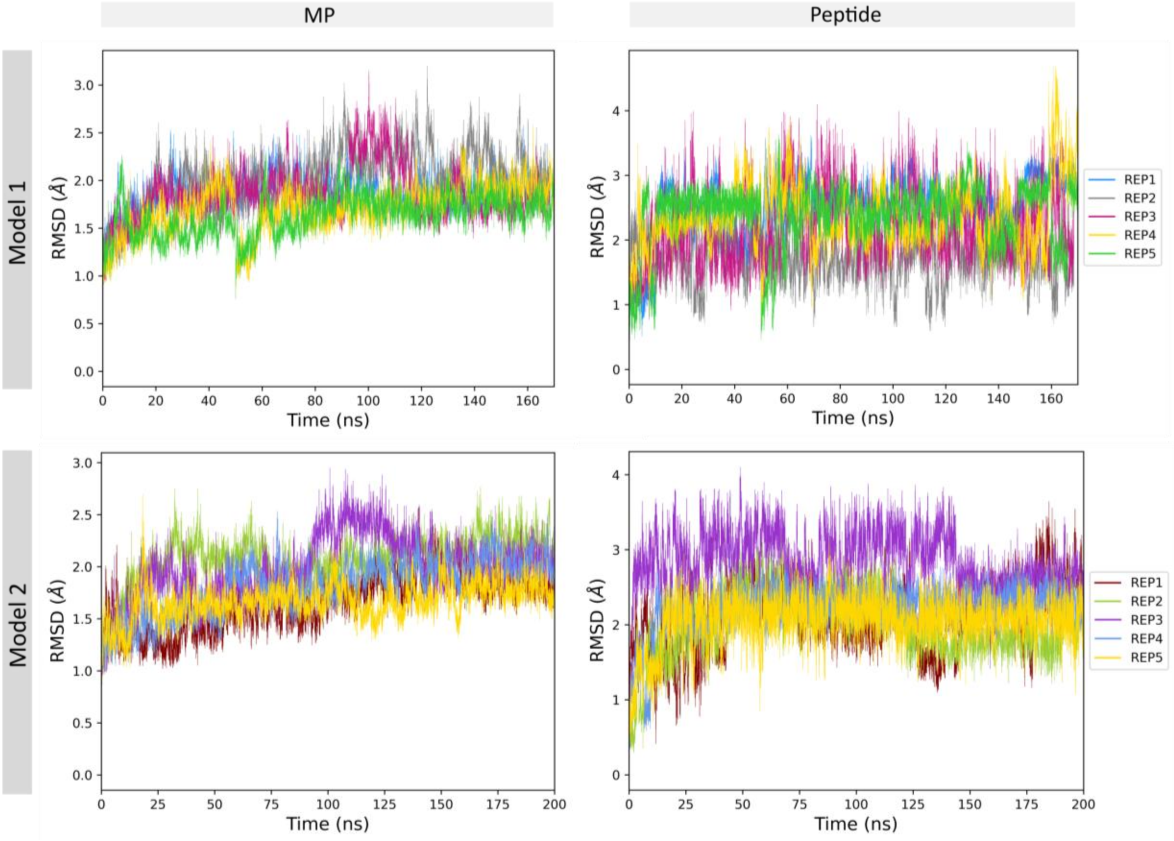


Figure 37 - Model1 and Model2: Backbone RMSD profiles as a function of time for the metalloproteinase (left) and for the peptides (right) of each model's replica. Higher RMSD value implies low stability of the protein structure.

The low RMSd allied with the high simulation time, show that the structures possess a stable folding conformation.

3.4.3. RMSF

To determine the effect of the substrate on the protein, the RMSF with respect to the average structure was calculated for all replicates of both models.

In line with the previous analyses, the RMSF profile of both systems (*Fig. 38*) do not significantly change during the simulation when compared to the initial structure, with a fluctuation difference of up to 2 Å. Moreover, the RMSF profile of both models was very similar to the RMSF of the initial structure.

However, REP3 of model 1 and REP1 of model 2 were the exceptions, which showed slightly higher loop fluctuations.

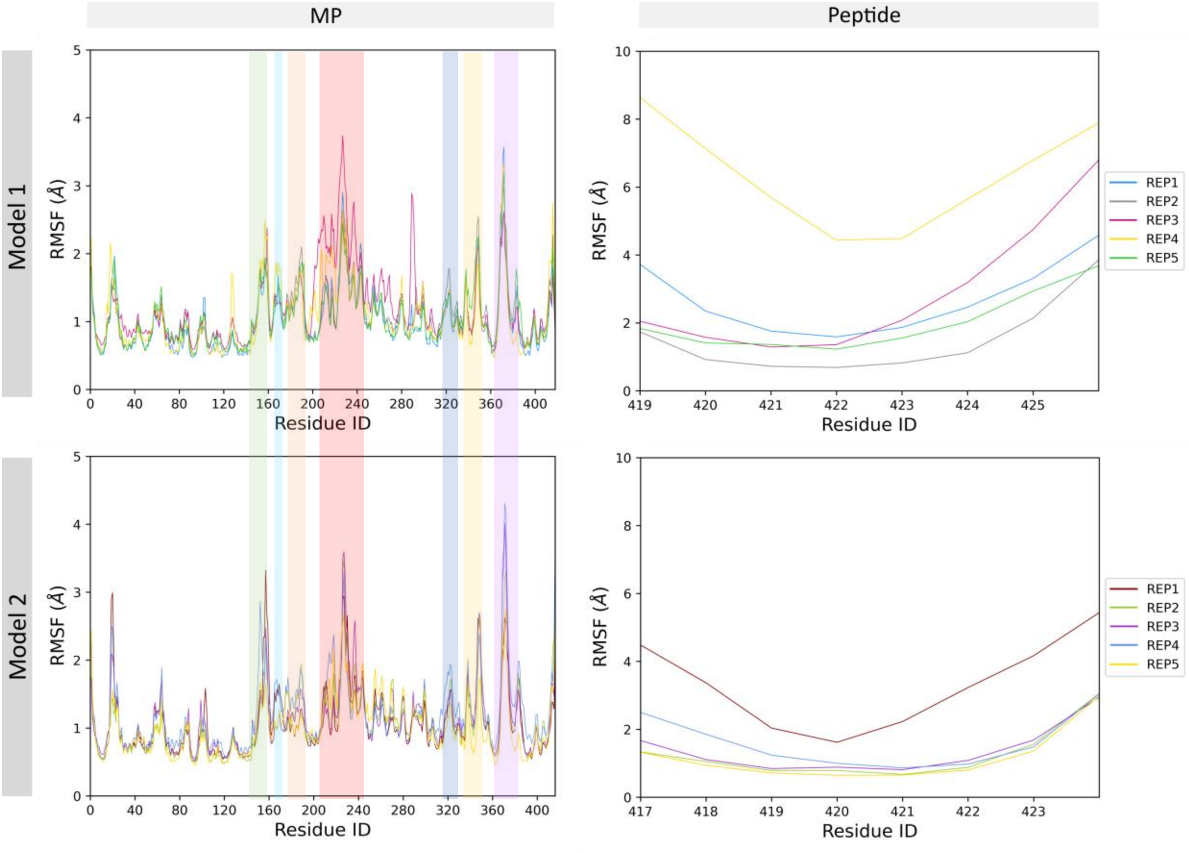


Figure 38 - Model1 and Model2: RMSF profiles of the metalloproteinase (left) and the peptide (right). Peaks are representatives of areas of high residual flexibility.

With respect to the metalloproteinase, it was possible to observe that both models exhibited a very similar behaviour regarding the flexibility between residues. These fluctuations mostly fell in loop regions, specifically in the domains that are likely to be sites that contribute to substrate's recognition, Dis-like and Cys-like domains as previously mentioned.

High fluctuations occurred in the loops placed before and after the conserved sequence "Met-turn", as stated before, at residues positions 151-160 (green colour), 170-173 (cyan colour) and 186-192 (orange colour). Also, as expected, the loop of the Dis-like domain, residues 224-245 (red color), and the loop regions of the Cys-like domain which include the HVR, residues 319-324 (navy blue), 338-350 (yellow colour), and 368-387 (violet colour) exhibited relatively large fluctuations during the simulations. The highest RMSF fluctuations were found around residues' number 228 and 373, both with a fluctuation value of approximately 3 Å. However, these regions were significantly more flexible/mobile on some replicates of conformation 2, with the highest peak at 4 Å at the HVR. These loop regions are represented on *figure 39*.

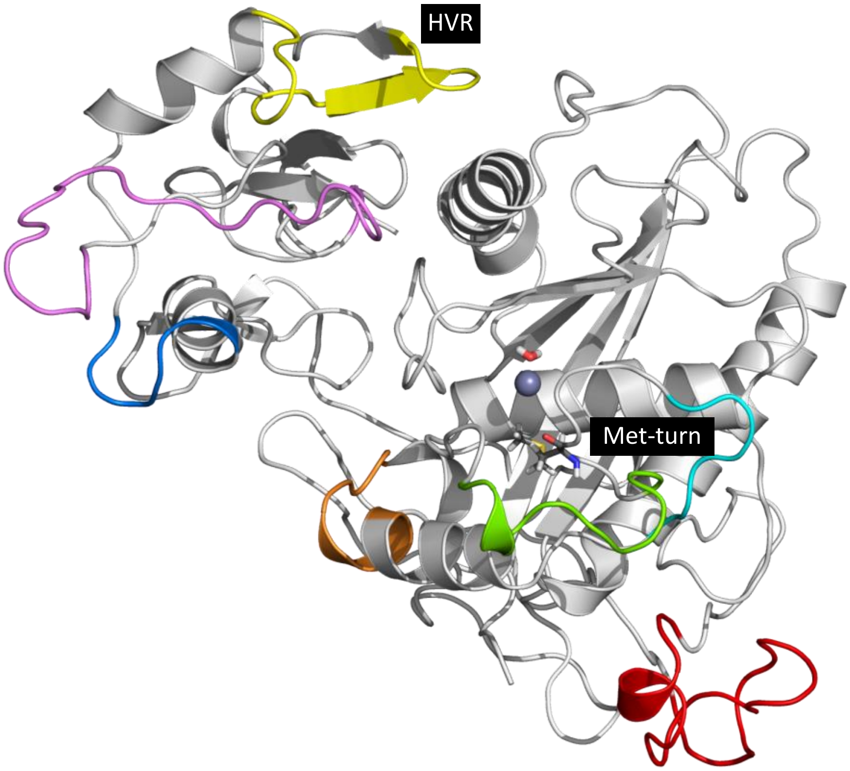


Figure 39 - Representation of the portions that undergo through greatest variations throughout the simulation based on the RMSF plots: yellow (338-350), orange (186-192), violet (368-387), navy blue (319-324), cyan (170-173), green (151-160) and red (224-245). The HVR is represented by the yellow portion and the met-turn is surrounded by the cyan and green loops.

Overall, the RMSF profile of both models displayed similar patterns suggesting that the binding mode of the peptide was alike.

With respect to the peptide, it was possible to verify that, in general, the residues of the conformation 2 underwent less fluctuations with lower RMSF values when compared to conformation 1.

Overall, the range of the atomic positional fluctuations in both models was within acceptance but the second seems to be more stable.

3.4.4. Distances

The distance distribution between the Zn ion with the carbonyl oxygen (Zn-Opep) and the catalytic water's oxygen with the carbonyl carbon (Owat-Cpep) of the peptide substrate of both models was also calculated. As can be seen from *figure 40*, they differed significantly even between replicates.

Regarding the Zn-Opep distance, with the exception of REP 1 and 5, conformation 1 distanced more than conformation 2, with REP 4 being the one that deviated the most extending in a range of 5–17 Å, showing that this peptide dissociates from the catalytic centre. A similar profile was seen for the Owat-Cpep distance. Moreover, when comparing the most stable replicates of conformation 1 (REP 1, 2, 3 and 4) with the ones from conformation 2 (REP 1, 2, 4 and 5), it was possible to notice that, in general, those of conformation 2 possess more suitable distances for the occurrence of the catalysis.

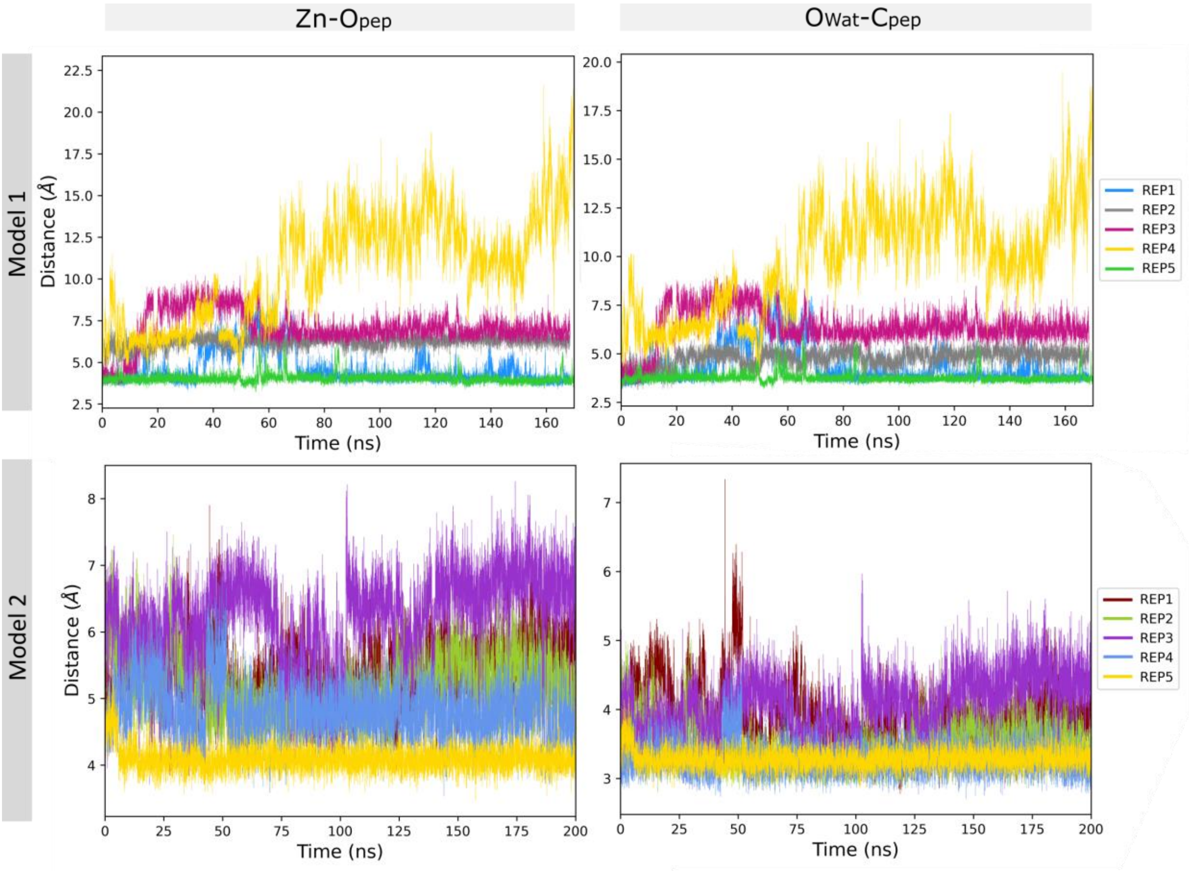


Figure 40 - Distance distribution between (left) Zn ion and the carbonyl oxygen and (right) catalytic water's oxygen and the carbonyl carbon of the scissile peptide bond on both models. 5 replicas are represented for each model.

On the other hand, from the Owat-Cpep distances, it was possible to confirm what was previously mentioned, as the distances in model 2 were smaller with an average value of 3.6 Å instead of 4.82 Å from model 1 (REP 4 not counted) and, consequently, more favourable.

Therefore, the plots above suggest that model 2 is more reliable and show favourable distances to the reaction mechanism.

Furthermore, the distances between the carboxylate side chain of Glu (Oε1 and Oε2) and the nitrogen of the peptide's Ile (Npep) were also measured (*Fig. 41*).

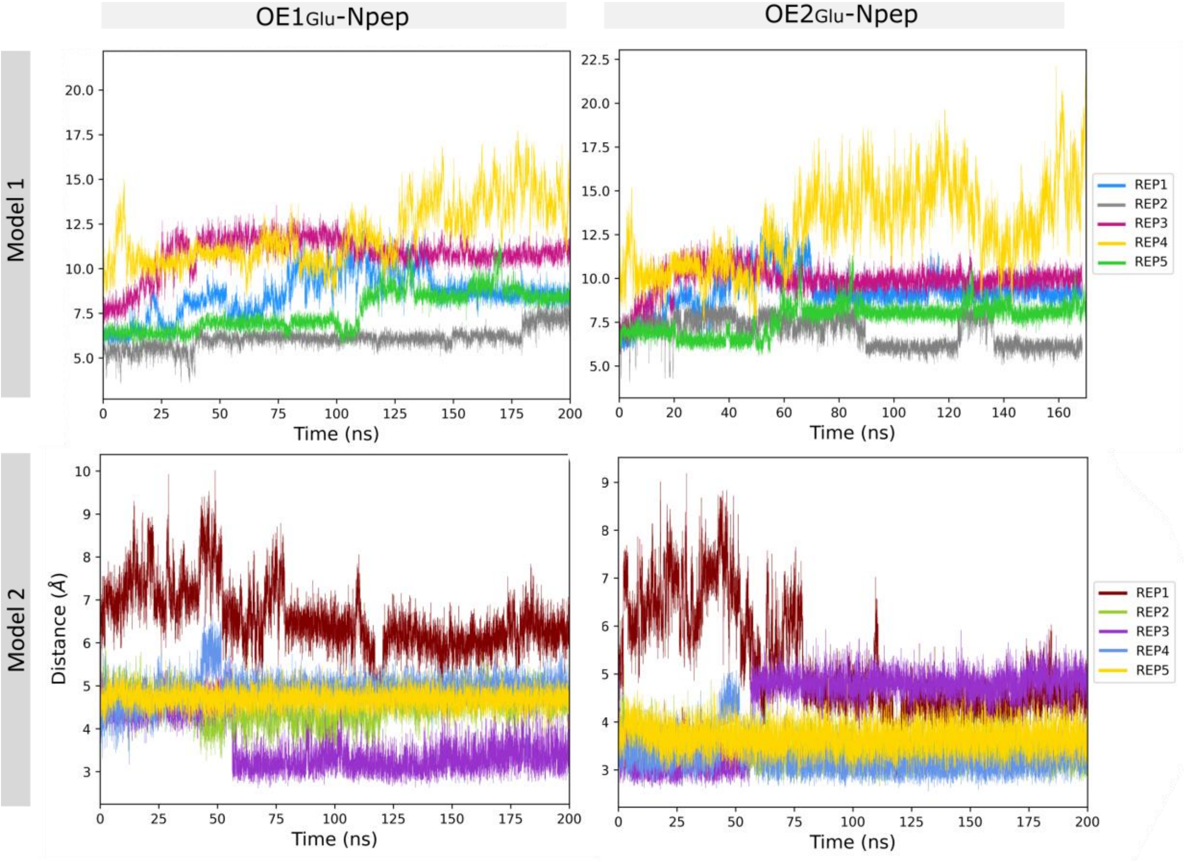


Figure 41 - Distance distribution between (left) GluOε1, (right) GluOε2 and the nitrogen of the scissile peptide bond and (right), respectively on both models. 5 replicas are represented for each model.

In the case of model 1, the distances cover a higher range of values, varying above 5 Å. Thus, the catalytic Glu140 and the Ile423 residue of the substrate are at distances that are not catalytically possible.

On the other hand, replicates from model 2 seem to be more suitable as O_{Glu}-N_{pep} distances vary between 3 and 5 Å. Even REP 1, which was at a very high distance at the beginning of the simulation, reached the plateau at around 5 Å after 60 ns.

Considering the obtained results, it is possible to conclude that model 2 is more suitable to the study of the reaction mechanism, as it is more stable and presents distances that are favourable for the catalysis to occur. Therefore, this model is the chosen one to study the reaction.

3.5. QM/MM simulations

3.5.1. Optimization of the starting structures

To test different options for the study of the reaction mechanism, several structures were used as starting points for the QM/MM calculations. One of these was the minimized structure of model 2 (structure A), prior to molecular dynamics. This structure did not possess any other water molecule near the catalytic centre than the catalytic one.

Catalytic distances of the starting structure are presented on *Table 2*.

Table 2 - Catalytic distances of the structure used for the QM/MM calculations.

	Min
Key distances (Å)	Structure A
Zn-O_{pep}	2.5
Zn-O_{wat}	1.97
O_{wat}- C_{pep}	2.85
OE2$_{Glu}$-N_{pep}	4.35
C_{pep}-N_{pep}	1.32

During the optimization process of structure A (*Fig. 42*), it was possible to verify that the metal ion maintains its tetrahedral geometry with the three His and the catalytic water, however there were some differences in the distances between Zn-O_{pep}, O_{wat}-C_{pep} and Oε2 $_{Glu}$-H1$_{wat}$. The carbonyl oxygen of the substrate got closer to Zn to a distance of 2.39 Å. As the Zn-O_{pep} distance decreased, the water molecule moved slightly away from the zinc ion (Zn-Owat). One of the water's proton was transferred to the Oε2 atom of Glu140 to a distance of 1.05 Å, making an hydrogen bond with the transferred proton of 1.48 Å. This water molecule was extremely polarized due to the interactions with the negatively charged carboxyl group of Glu140 base and the acidic Zn^{2+} and thus was converted into an Zn^{2+} - bound hydroxide ion.

Thus, a new reactant complex (ES) formed in which it is the hydroxide (OH^-) that functions as the nucleophilic species attacking the scissile peptide bond. This hydroxide nucleophile possesses a good position to attack the scissile peptide bond, however at a larger distance from the C_{pep} (3.03 Å) when compared to other studies (2.43-2.74 Å) [108, 109].

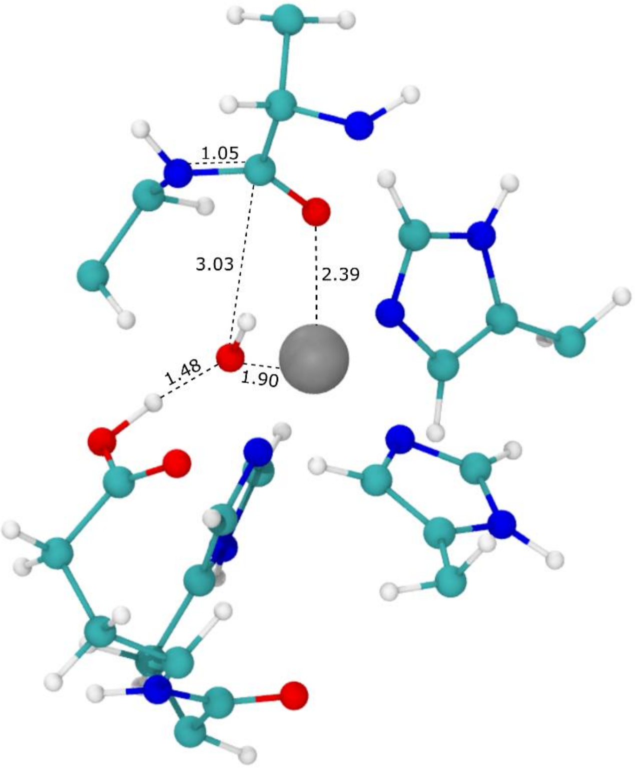


Figure 42 - Optimized structure of the reactant state (ES) for structure A and respective catalytic distances. Only the high-layer is shown.

In addition to structure A, several frames were extracted from the sMD simulation trajectories of REP5 and REP2 to test different catalytic distances and possible conditions for the catalysis to occur.

Considering that the active centre of RVV-X is exposed to the solvent, it can be inferred that the electrostatic interactions between solvent and protein can affect the stability of the reactive species. For that reason, several hypotheses were inspected. In addition to testing different initial structures with different catalytic distances, the impact of the presence of possible auxiliary water molecules in the active centre beyond the catalytic one was also studied. During MD analysis it was verified that in certain trajectories there were water molecules in close interaction with the catalytic glutamate making a hydrogen bond between it and the Thr108 residue. Another water was found near the carbonyl group of the scissile peptide bond. Their influence as to whether it stabilizes the atoms in the QM layer were studied.

During the optimization of the structures extracted from REP2 and REP5 , a decrease in the distance between the carbonyl oxygen of the peptide and the Zn ion (Zn-Opep) was verified, with a concomitant increase in the distance between the Zn and the water molecule (Zn-Owat). However, conversely to structure A, the water molecule is not spontaneously deprotonated, which means that these structures might have some feature that may disable this transference. It should be noted that the reactivity of the enzyme is influenced by all the interactions that occur in the system.

To proceed with the study of the reaction mechanism, the two-step reaction mechanism of peptide hydrolysis proposed for metalloproteinases shown in *figure 8* from Chapter I, was followed [109, 110].

3.5.2. First step – Nucleophilic attack

As the first step of the reaction is the nucleophilic attack on the carbonyl carbon of the substrate by the catalytic water, the distance between Owat-Cpep was chosen as the reaction coordinate to describe the potential energy surface (PES). This step resulted in the formation of the first transition state (TS1) as well as the first intermediate (INT1).

Along the scan, the distance between Owat-Cpep was shorten gradually to 1.74 Å and as this happened, it was possible to observe that the distance between the carbonyl oxygen of the substrate and the Zn ion also decreased to 1.97 Å, which is expectable as the carbonyl oxygen is receiving the negative charge of the hydroxide and becoming anionic.

This TS1 had a distorted pentacoordinate geometry with the three His, the catalytic water and the carbonyl oxygen of the substrate.

This first step yielded the higher energy barrier of 14.4 kcal.mol^{-1} at the B3LYP/6-31G(d) level of theory, which is in close agreement with other studies [108]. The corresponding optimized TS1 structure is represented in *figure 43*. The vibrational frequencies involving the reaction coordinate for the TS1 structure resulted in a single imaginary value of 192i cm^{-}

^{1}. This supports that the nucleophilic attack by the water molecule is most likely the rate-limiting step.

During the scan, the distance between the proton and the nitrogen atom of the scissile peptide bond (Npep) decreased from 3.86 to 3.04 Å. In parallel, there was a change in the substrate charge density as the carbonyl group became negatively charged. The intermediate INT1 was located at a reaction coordinate value of 1.49 Å and is also shown in *figure 43*.

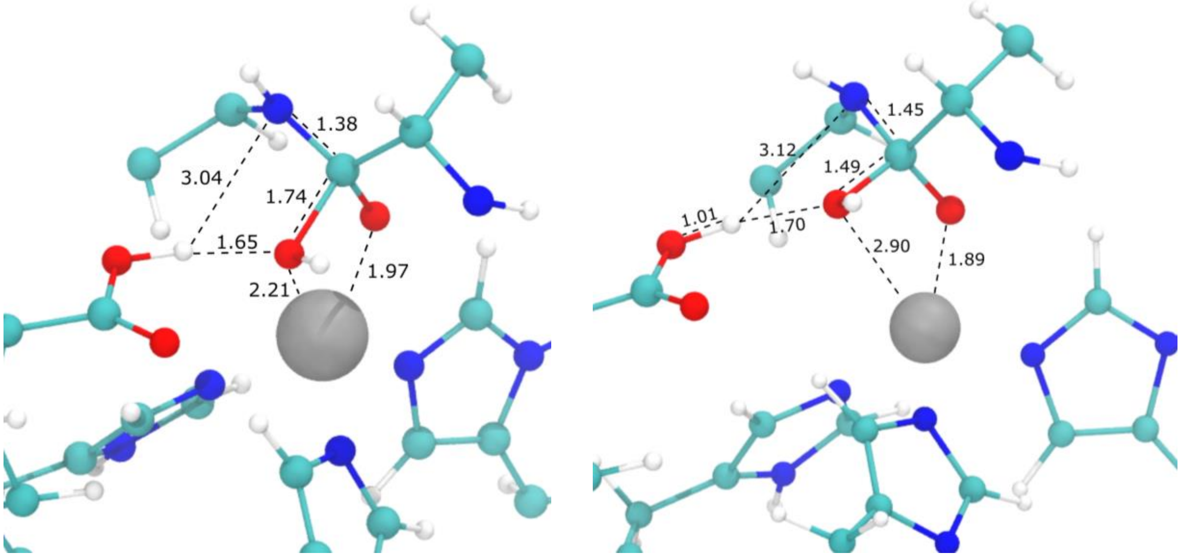


Figure 43 - Optimized structure of the transition state, TS1, which corresponds to the nucleophilic attack by the water molecule (distance of 1.74 Å) and the intermediate state, INT1, for structure A and respective catalytic distances. Only the high-layer is shown.

Although it was observed that the post-dynamic structures did not seem appropriate, a linear scan was performed following the same setup protocol used for structure A to assess the influence of the different conditions on the PES. The energy profile of the reactive structure A (black curve) represented on *figure 44* shows a transition state (maximum) and an intermediate (minimum).

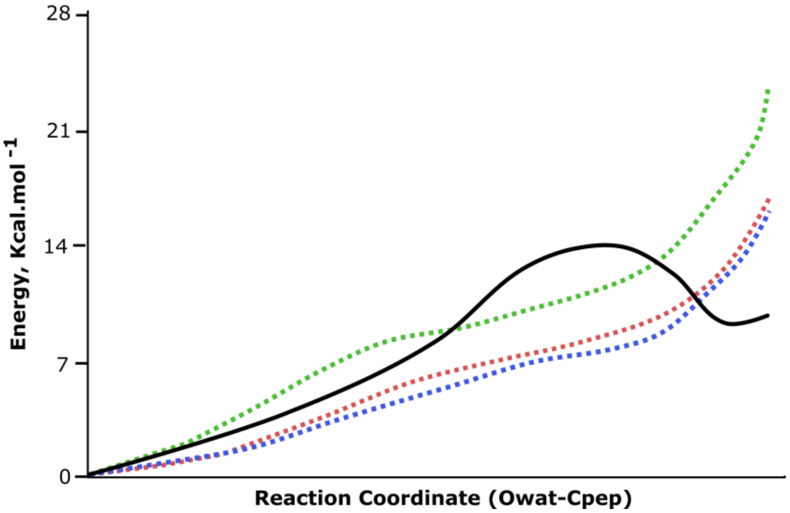


Figure 44 - Potential energy profiles for the linear scans (first step of the reaction) in the active site of the RVV-X. Black curve represents the scan from structure A; Green, red, and blue dotted curves represent the different frames tested in which the profiles do not reach a minimum.

On the contrary, the profiles of the post-molecular dynamics structures not only went up not reaching a minimum for INT1, but the overall barrier was higher. This might be because these structures are not pre-arranged for the reaction to occur. Although the reaction proceeds as expected, with no major differences at the QM layer, the MM layer underwent thorough some rearrangements rising the energy, which could be the reason for the energy profiles not reaching a minimum.

In the future, the geometry of the products in which the MM layer is already rearranged should be optimized , and a linear transit scan should be performed in the reverse direction.

Therefore, only the study of the mechanism for structure A will be reported throughout this work since it is necessary to find the underlying exact factors that prevent the reaction in the other structural conformations.

3.5.3. Second step – Proton transfer to the nitrogen atom of the substrate

In the second step, the side chain of Glu140, which acts as a proton shuttle, donates the proton taken from the catalytic water to the scissile-bond nitrogen. Thus, the distance between the Glu140 hydrogen and the scissile-bond nitrogen (OHGlu-Npep) was chosen as the reaction coordinate to perform a linear scan. This step resulted in the formation of the second transition state (TS2) and the second intermediate (INT2).

During this step, the proton taken from the water by the Glu140 was being transferred to the peptide's nitrogen. Furthermore, as expected, the N-C peptide distance gradually increased from 1.45 to 1.57 Å in the TS2 optimized geometry. In addition to this, it was possible to observe the formation of a new H-bond between the hydrogen of the Oε1 of Glu140 and the carboxyl group of the peptide at a distance of 1.75 Å.

Moreover, the bond between the substrate's carbonyl oxygen and the Zn^{2+} increased to 1.95 Å as it moves from oxyanion ($C\text{-}O^-$) to neutral oxygen (C=O).

The transition state corresponding to this step was located at the reaction coordinate value of 1.35 Å (*Fig. 45*). The calculated energy barrier is low, 8.0 $kcal.mol^{-1}$. The vibrational frequencies involving the reaction coordinate for the TS2 structure resulted in a single imaginary value of 1160i cm^{-1}.

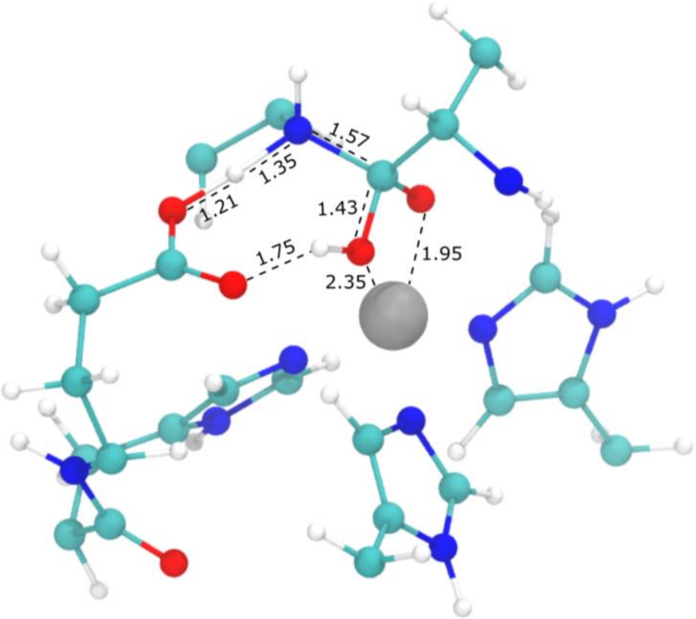


Figure 45 - Optimized structure of the second transition state (TS2) for structure A corresponding to the proton transfer and the cleavage of the bond C-N and respective catalytic distances. Only the high-layer is shown.

These events resulted in the C_{pep}-N_{pep} bond elongation to 1.68 Å leading to the intermediate INT2 (*Fig. 46*). In this step of the reaction, the proton was completely transferred to the peptide's nitrogen and the H-bonds are rearranged in order to protonate the catalytic Glu, as the substrate's C-terminal carboxylic acid is very prone to protonation the Glu140 and end up in the anionic form, due to the stabilization provided by the Zn^{2+} cofactor.

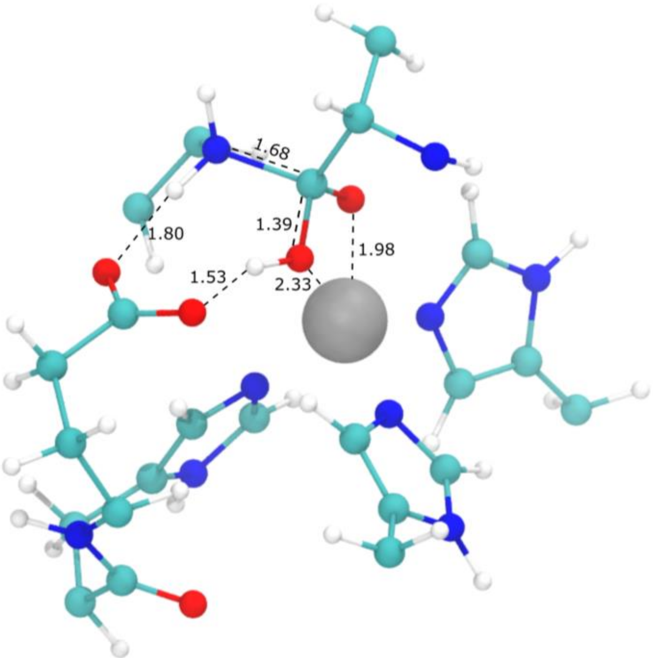


Figure 46 - Optimized structure of the second intermediate (INT2) for structure A and catalytic distances. Only the high-layer is shown.

In the last step, the C-N distance along the scissile bond was used as the reaction coordinate.

The barrier of this process was very low (less than 1 kcal.mol^{-1}). For that reason, the transition state corresponding to this reaction step was not further optimized, because the slope of the PES in the respective region was very flat and the associated barrier is insignificant at “biological temperature” (25-37 ºC), i.e. the body temperature range of the several Russell’s viper prey. Therefore, this step results in the formation of the third transition state (TS3) and the third intermediate state (INT3) (*Fig. 47*).

When passing from TS3 to the product, the Oε1 side chain of the acid Glu was protonated. The metal ion was penta-coordinated by the oxygen atoms of the C-product carboxyl group with distances of 2.33 and 1.99 Å and the three His139, 143 and 149.

In the INT3 it was possible to verify a drop in the energy profile (*Fig. 48*) which resulted from the cleavage process that led the C-N distance to 3.18 Å. The proton is completely transferred to the Glu.

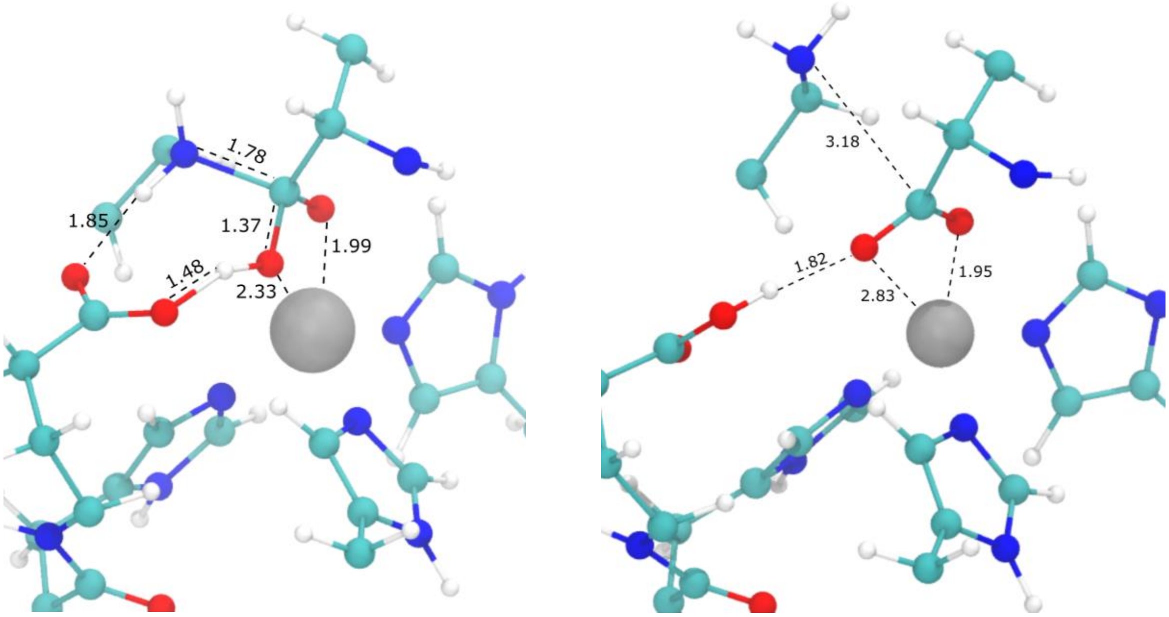


Figure 47 - Structures of the third intermediate state (TS3) (left) and the third intermediate state (INT3) (right) which corresponds to the release of the products for structure A and respective catalytic distances. Only the high-layer is shown.

At this stage the product was formed but the enzyme was not completely reconstituted, because it had a proton in the catalytic Glu. Thus, the deprotonation of the Glu and transference of the proton to the amine group are requirements for the protein to return to its active state to hydrolyse the next substrate after the release of the products. However, the steps related to the release of the product are very difficult to simulate and as they are physical processes these were not included in this study.

The energy profile for the overall reaction is represented on Fig. 48.

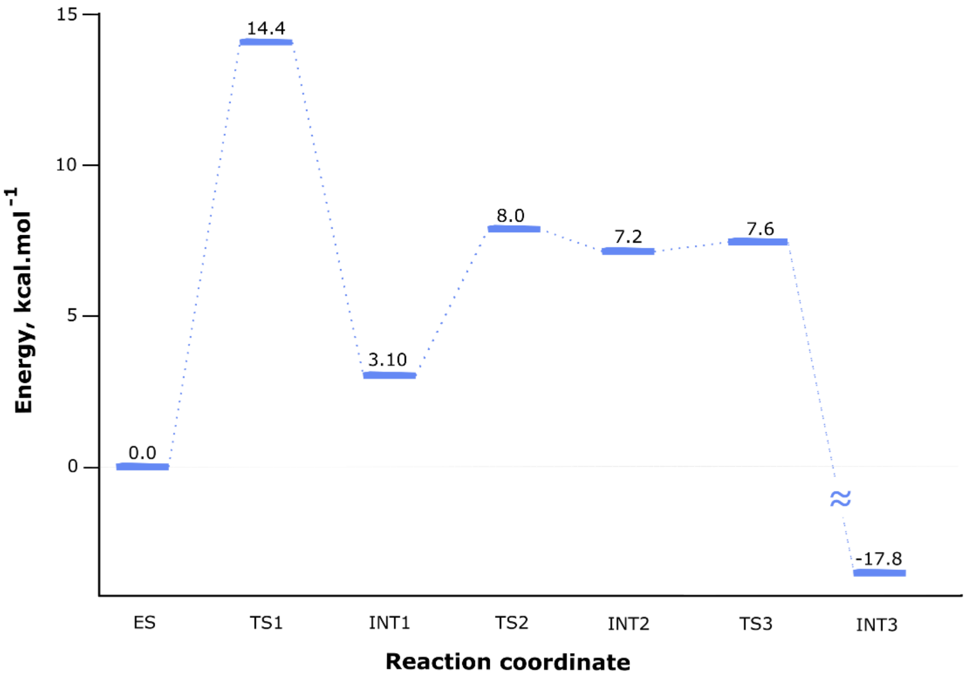


Figure 48 – ONIOM energy profile for the hydrolysis of the RVV-X substrate obtained at the QM(B3LYP/6-31G(d))/AMBER level of theory.

The variation on the key distances over the course of the reaction are represented on *Annex 2.*

Overall, the two-layered ONIOM model (B3LYP/6-31G(d):AMBER) seems to correctly reproduce the steps of the already proposed reaction mechanism. The potential energy profile for the complex as well as the optimized geometries are quite similar to those observed in other studies.

For structure A it would be interesting to model a second water molecule coordinating the peptides's carbonyl oxygen atom promoting an electrostatic stabilization to this group and include it in the QM layer. This would allow to verify if there is a reduction in the activation energy barrier.

For the seemingly non-reactive conformations, it would be necessary to perform a scan using the distance between the water's hydrogen and the calalytic Glu (Oε2Glu-Hwat). Alternatively, performing a double scan using the former distance and the distance between the water's oxygen and the carbonyl carbon of the substrate as reaction coordinates to describe the PES.

Moreover, according to Vasilevskaya et al. 2016 [12], a realistic zinc coordination sphere and their energies might depend on a proper solvent shell, that is, on its thickness. Thus, it would also be interesting to test different thicknesses of the water layer to inspect its influence on the PES.

3.6. VS

In this work, a set of commercially available drugs was used in two virtual screening approaches, LB and SB, to identify putative inhibitors compounds of the RVV-X.

3.6.1. Compound database preparation

Considering the similarity between binding sites of different snake venom metalloproteinases, 6 X-ray crystallographic structures containing co-crystallized small inhibitors present in the PDB database were selected (*Fig. 49*).

Batimastat, Marimastat, WR2 and Ilomastat belong to the group of hydroxamate-based peptidomimetics, which means they mimic the peptide structure of natural substrates, inhibiting the receptor by bidentate chelation of Zn^{2+}. 0GR and 0QI inhibitors, on the other hand, are carboxylate-based inhibitors. 0GR is naturally found in venoms at millimolar concentrations as an endogenous tripeptide (PyroGlu-Asn-Trp) [90]. The physicochemical properties of each compound are listed in *Table 3*.

Table 3 - Physicochemical properties of each ligand and its respective values. Lipinsky's Rule of 5 (MW < 500 $g.\,mol{-}1$, LogP < 5, HBD < 5, HBA < 10).

	BAT	**097**	**WR2**	**GM6**	**0GR**	**0QI**
Molecular Weight (g.mol^{-1})	477.6	331.41	455.6	388.5	411.5	339.5
Hydrogen Bond Donor	4	5	5	5	4	3
Hydrogen Bond Acceptor	6	5	7	4	5	5
Heavy Atom	32	23	31	28	30	23
Rotatable Bonds	12	8	11	9	9	9
Calculated logP	3.2	0.4	0.9	1.23	2.9	1.8
Lipinski's Rules of 5	✓	✓	✓	✓	✓	✓

Through observation of the values listed on *Table 3*, it is possible to infer that all the compounds respect the Lipinski's rule-of-five, suggesting that they are favourable regarding lead-likeness criteria.

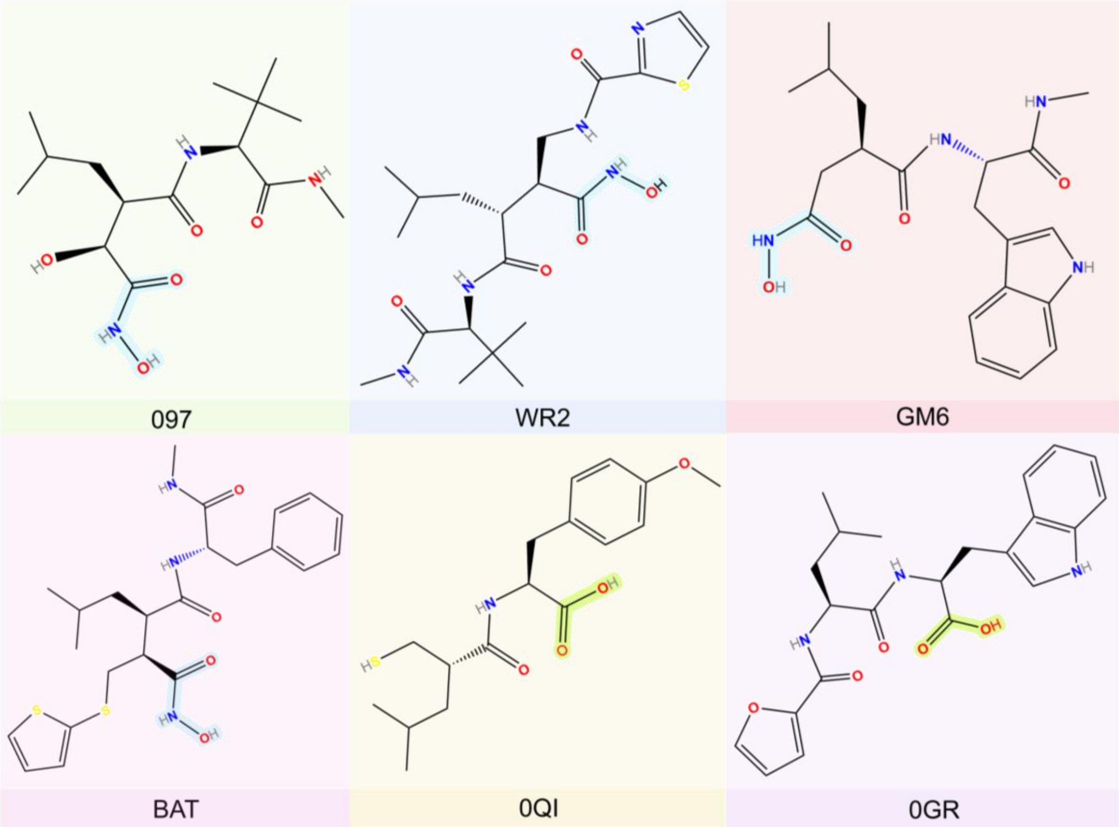


Figure 49 - Representative structures of metalloproteinases known inhibitors (Batimastat, Marimastat, Ilomastat, WR2, 0QI and 0GR) and respective PDB IDs, used throughout this work. The hydroxamate portion is highlighted in blue while the carboxylate portion is highlighted at green colour.

Both 0GR and 0QI inhibitors are placed between the bulged-edge segment and the hydrophobic S1'-specificity pocket. In both cases, the side chain of the Trp from 0GR and the 0-methyl phenol group from 0QI Tyrosine are found in the deep S1'-pocket of the target protein not filling it completely. Conversely, the BAT inhibitor fills this pocket completely. Moreover, the carboxylate group (COO^-) of 0GR and 0QI coordinates the metal ion. One of the oxygen atoms of the carboxylate group is close to the side chain of the catalytic Glu, which might play the role of a base in the reaction mechanism. Some polar side chains point towards the solvent in both cases.

As for hydroxamate-type inhibitors, Batimastat, Ilomastat, WR2 and Marimastat, it was possible to see that the hydroxamate group ($CONHO^-$) coordinates the metal ion in a bidentate chelating manner. In all cases, the inhibitor occupies the active-site cleft and is dominated by hydrogen interactions with the surrounding residues of the “S1’-wall forming segment” (residues Gly107-Ala112) and the “bulge-edge segment” (residues Pro166-Leu168). Isobutyl functional groups are often sitting outside the pocket.

3.6.2. Validation of the VS protocol

The co-crystallized inhibitory compounds (actives) were first redocked into the active site of the corresponding target protein to assess the performance of the docking programs when predicting the native pose. Here, besides rDOCK and AutoDock4, all the scoring functions implemented in the GOLD software, ChemPLP, ASP, GoldScore and ChemScore, were used.

The performance of the docking programs varied a lot. rDOCK and AutoDock4 have been proven unsuccessful regarding pose prediction because they were not able to re-dock the ligands in their experimental poses and it did not come close. On the contrary, GOLD's ASP and ChemPLP scoring functions were the ones that presented best accuracy to reproduce the native poses of the ligands, with ChemPLP offering a slightly higher ability. Therefore, the study continued with this scoring function.

It was found that the top-scoring pose of the redocked conformation of Marimastat perfectly superimposed on its co-crystallized form with a RMSD value of 0.9 Å. Inspection of the pose allowed to infer that the hydroxamate portion of the ligand agrees with the one in the crystallographic pose.

Marimastat was re-docked and completely superimposed on the crystallographic structure, whereas Batimastat, WR2 and GM6 had functional groups that were placed differently from the crystallographic structure. The deviations suffered by Batimastat and WR2 might be correlated with its size and flexibility, as the number of rotatable bonds of each ligand is up to twelve. As known, the greater the number of rotatable bonds, the lower the re-docking accuracy, independent of the docking program used [111]. In general, all scoring functions had great difficulty in re-docking the 0QI and 0GR ligands, leading to large deviations in the apolar and aromatic groups, which might account for the higher RMSD values.

Further, the performance of the docking programs was also evaluated to ensure the ability to discriminate the actives out of the decoys set.

The results were examined by calculating the enrichment factor (EF) value in different percent of top ranked molecules. The active compounds that were successfully identified in the top 1% (top 3 compounds), 2% (top 6 compounds), 10% (top 31 compounds) and 20% (top 61 compounds) were 0, 0, 1, and 4. For this reason, the best percent of ranked molecules was 20% as it could find almost all the active compounds. Based on these comparisons, GOLD ChemPLP's performance was superior offering higher performances regarding pose prediction as well as in the discrimination between active and inactive molecules.

Thus, Gold was employed as the docking program for the chemical libraries to be explored.

3.6.3. LBVS

3.6.3.1. Pharmacophore Model Generation

For the LBVS, a pharmacophore model was generated considering the common chemical functionalities of the known namely metalloprotease inhibitors.

Several pharmacophores were generated based on these active ligands, however, only the pharmacophore that combined the pharmacophoric properties of 4 of the ligands, that is, of the hydroxamate-based peptidomimetics (Marimastat, Ilomastat, Batimastat and WR2), was chosen. This choice was made considering that this pharmacophore possessed a higher score and the features of those 4 ligands completely superimposed between them. Furthermore, the inclusion of a 5th compound and/or a 6th compound would restrict very narrowly the search space.

The selected pharmacophore model (*Fig. 50*) comprises fifteen pharmacophoric characteristics, four HBA, four HBD and seven hydrophobic groups, which are found around the active site of RVV-X .

This hypothesis comprises the hydrogen bonding acceptor 5 (HA5) at 2.3 Å from the amine group of Leu167, while HA7 is at 1.8 Å from the amine group of Leu105. On the other hand, the hydrogen bonding donor (HD1) is 3.7 Å from the carbonyl of Pro165 and HD2 is 2.9 Å from the carbonyl of ASN 103. The hydrogen bonding donor HD3 is close to the O atom of Gly106. Thus, molecules aligned with these pharmacophoric properties can form hydrogen bonds with the adjacent amino acids from the enzyme.

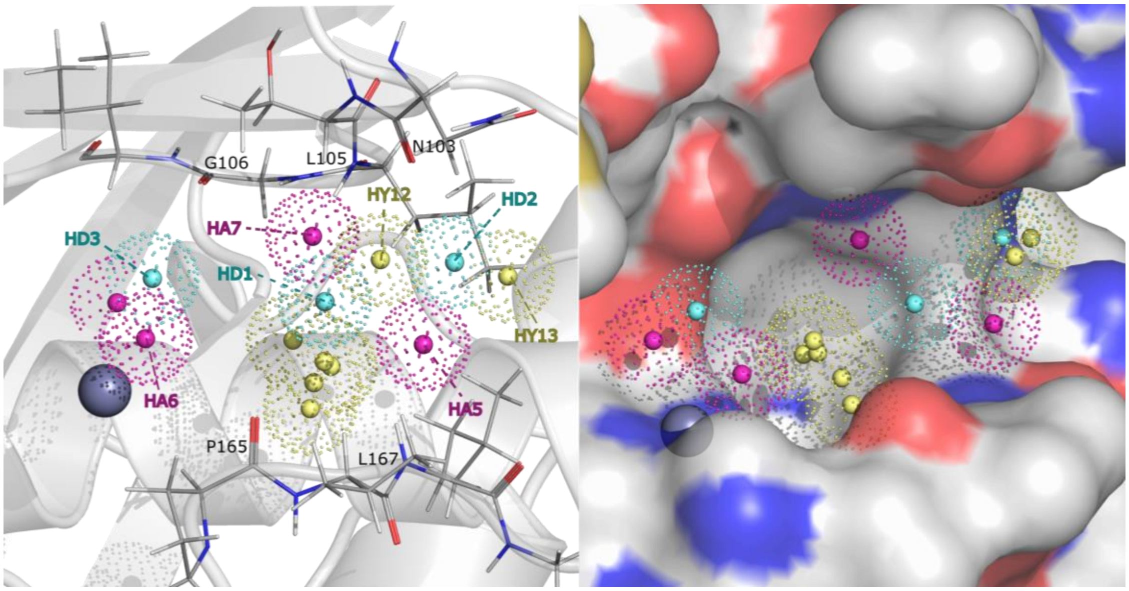


Figure 50 - Pharmacophoric features of the selected pharmacophore.

The hydrophobic groups (HY9, HY10, HY11, HY14 and HY15) are in a hydrophobic tunnel, specificity pocket, composed of Thr107, Leu108 and Ile136, known to be responsible for the substrate specificity. The hydrophobic group (HY14) is also close to the Leu167 side chain of the "S1'-wall forming segment", being able to interact with it and HY13 that points to the solvent. Thus, molecules aligned with these properties can promote hydrophobic interactions with the enzyme.

In summary, the compounds that have these characteristics may have the potential to inhibit RVV-X.

However, when the pharmacophore was submitted to ZincPharmer webserver, no output was returned and for that reason, after trying different pharmacophoric characteristics to obtain a significative number of hit compounds, only 7 of the 15 pharmacophoric characteristics were selected, namely HD1, HD2, HD3, HA5, HA6, HA7 and HYD14, and a radius of 2 Å was adjusted for further screening.

3.6.3.2. ZINC Database screening and characterization

This pharmacophore model was used as a query to select the compounds of the ligand-based screening over the ZINC database. Circa 339,000 compounds fit all the chosen features present in the pharmacophore. These hits allowed the construction of a new chemical library, based on their ligandability, specific for the metalloproteinase.

The compounds were further described and characterized by their physicochemical parameters, as shown in the next figures. These properties have significant influence on the pharmacokinetic properties of drug candidates.

The descriptors were computed for all the hit ligands. For each descriptor, compounds were assessed as to whether they were within the recommended range. The variability of the molecules retrieved from the Zinc database based on the pharmacophore model is represented in the following pictures.

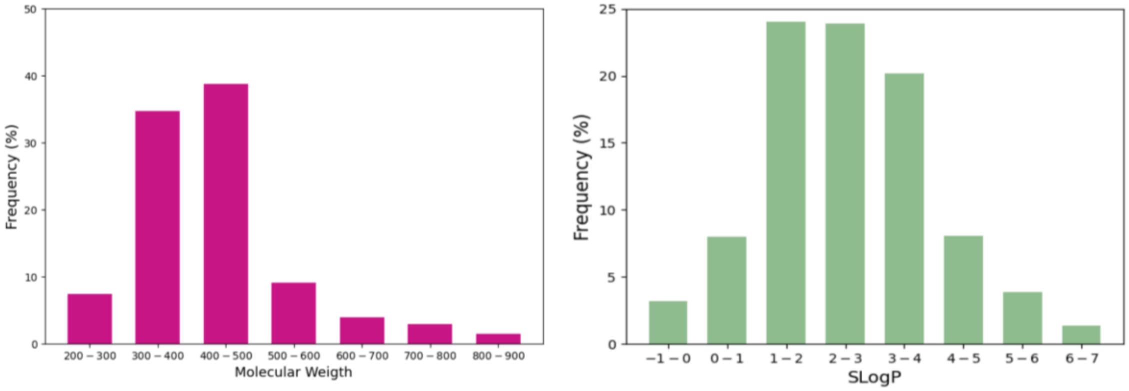


Figure 51 - Molecular weight (right) and LogP (left) distribution of all molecules retrieved from the Zinc database.

The results (*Fig. 51*) show that from the total number of retrieved compounds, 35% have a MW between 300-400 Da, while almost 40% are in the range of 400–500 Da. This indicates that most of the hit list of compounds obey the Rule of five, recommended for the oral bioavailability of a small molecule drug. As for the LogP, the results show that most retrieved compounds vary between 1-4 LogP, with a percentage of approximately 68 %, indicating that the compounds are mostly lipophilic.

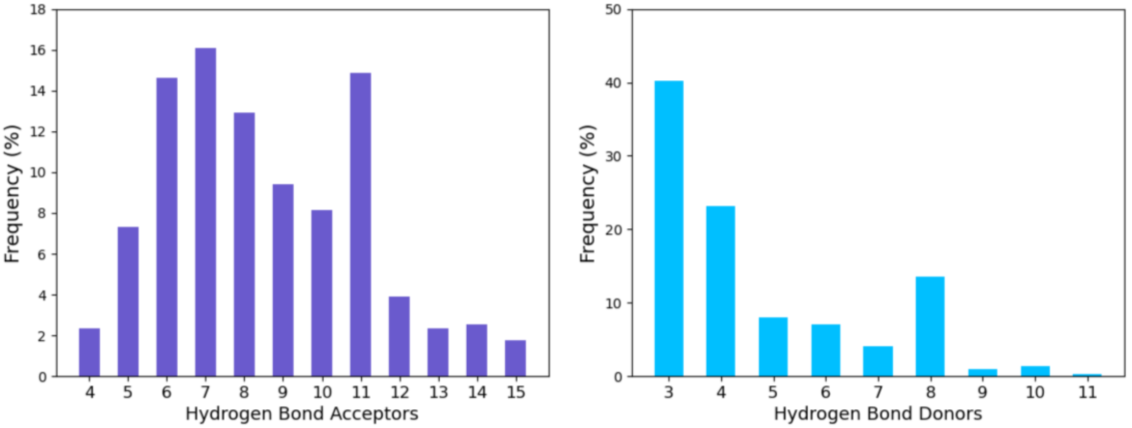


Figure 52 - HBA (right) and HBD (left) distribution of all molecules retrieved from the Zinc database.

Figure 52 presents the relative distribution of HBA and HBD number from the hit list of compounds. The results indicate that, regarding HBA, there is a wide range acceptor numbers, with a higher prevalence in compounds with 6 (14.5%), 7 (16%), 8 (13%) and 11 (15%) acceptors, which in total represents almost 59% of the set. While for HBD there is lower variance, with the majority having between 3 (40%) and 4 (24%) donors. Thus, the data is in agreement with the acceptable range proposed by the Lipinski's rules of 5.

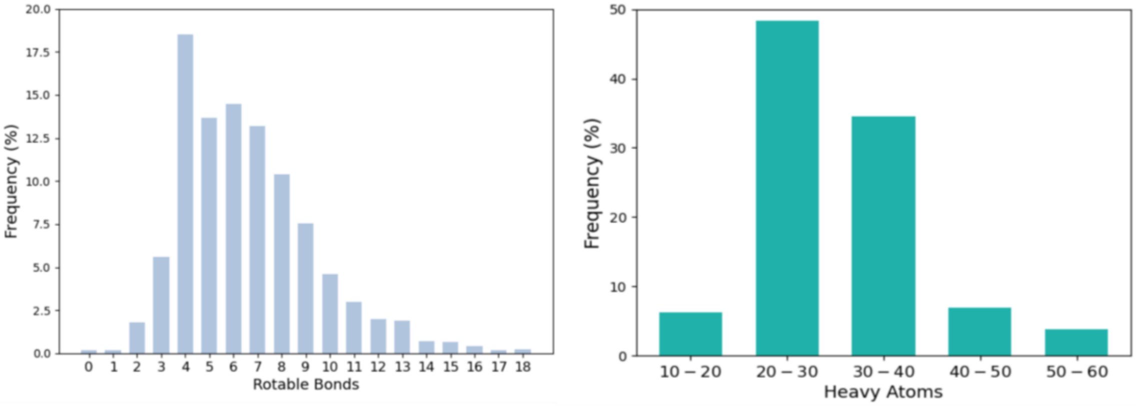


Figure 53 – Rotatable bonds (right) and Heavy atoms (left) distribution of all molecules retrieved from the Zinc database.

Figure 53 presents the relative distribution of rotational bonds number from the hit list of compounds. Results show that there is a higher prevalence of ligands with 4 (18%), 5 (13%), 6 (14%) and 7 (12.5%) rotational bonds, which in total represents almost 60% of the set. The remaining ligands possess 2 to 3 and 8 to 13 rotational bonds. As previously referred, the higher the number of rotatable bonds, the lower the accuracy in the docking process, considering that these compounds can adopt very different conformations.

Interestingly, most compounds have physicochemical properties in the acceptable range of the lead-likeliness of the Lipinski's Rules of 5.

3.6.3.3. Molecular Docking, characterization, and clustering

As there were to many hits, these were docked in the target's active centre. Molecules were then ranked according to their binding energy score.

To verify if the hits exhibit the same proprieties as the SBVS, the top 1000 were selected and characterized (*Fig. 54*). Here the molecular descriptors for the pharmacokinetic evaluation were also considered to understand which type of compounds the active centre prefers.

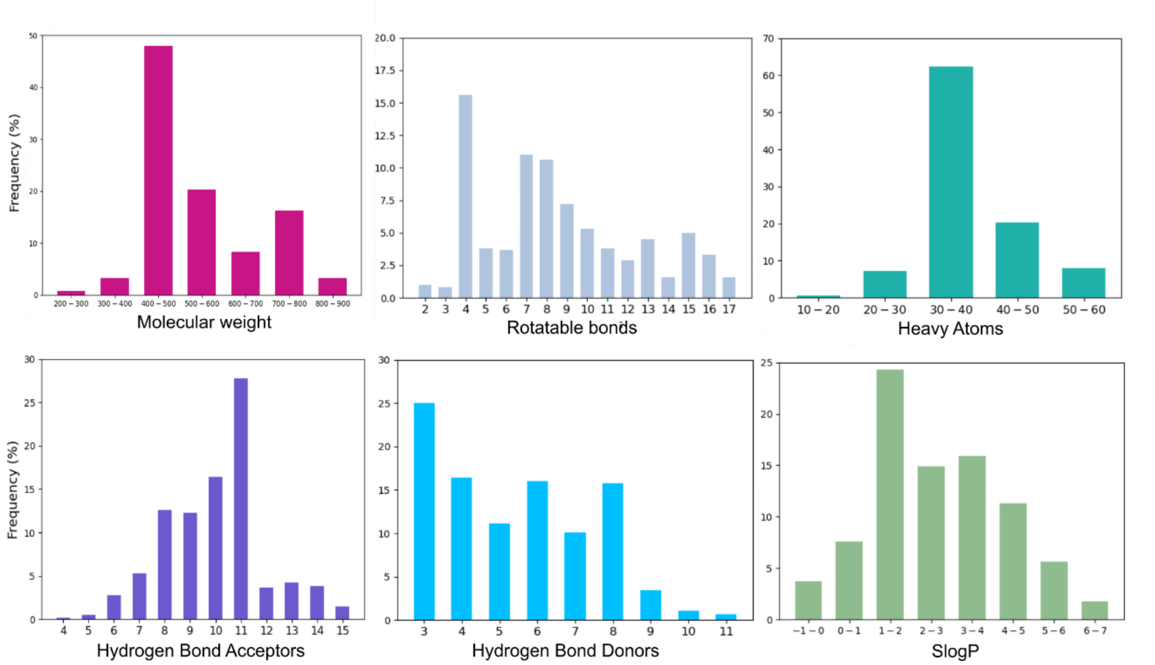


Figure 54 - Physicochemical properties of the top 1000 ranked compounds after docking.

According to the results shown on *figure 54*, it was possible to infer that the majority of the compounds that obtained the highest score did not fully respect Lipinski's rules, with more than 70% violating at least one of the rules. Furthermore, no compound met any of the requirements of the rules of 3.

Almost half of the dataset, 47.9% of the top 1000 compounds, had a MW between 400-500 Da, while 20.3% were in the range of 600–700 Da, which indicates that this enzyme can accommodate large molecules. The number of rotatable bonds varies a lot within the dataset, with the majority having 4 or between 7 and 8. Approximately 41% of the compounds have between 8-9 HBA and between 3-4 HBD agreeing with the Lipisnki’s rules. As for the LogP, the results show that 55% of the compounds vary between 1-4 LogP.

These compounds were then clustered by chemical similarity (*Fig. 55*) in order to check the diversity of the dataset. Here, the non-hierarchical hierarchical clustering method Jarvis-Patrick was used.

The output returned the population of each cluster, its centroids, i.e, the molecule that represents each cluster and molecules that showed different proprieties called singletons. At 60% similarity threshold (*Fig. 55*), the 1000 compounds were clustered into 53 centroids and 44 singletons, where 24 of the clusters contain only one to four molecules. Furthermore, the average and maximum dissimilarities within the whole dataset were 84.3% and 99%, respectively. The most populated clusters were the 2 (centroid 258), 3 (centroid 45), 18 (centroid 61) and 50 (centroid 209).

At 70% similarity threshold, these were clustered into 34 centroids and 16 singletons, with the average and maximum dissimilarities being also 84.3% and 99%, respectively with the most populated clusters being the 2 (centroid 362), 3 (centroid 121), 11 (centroid 45) and 31 (centroid 211).

The repartition chemical space of the dataset was assessed. Figure 55 shows a histogram of the chemical space resulted from the *jarp* tool.

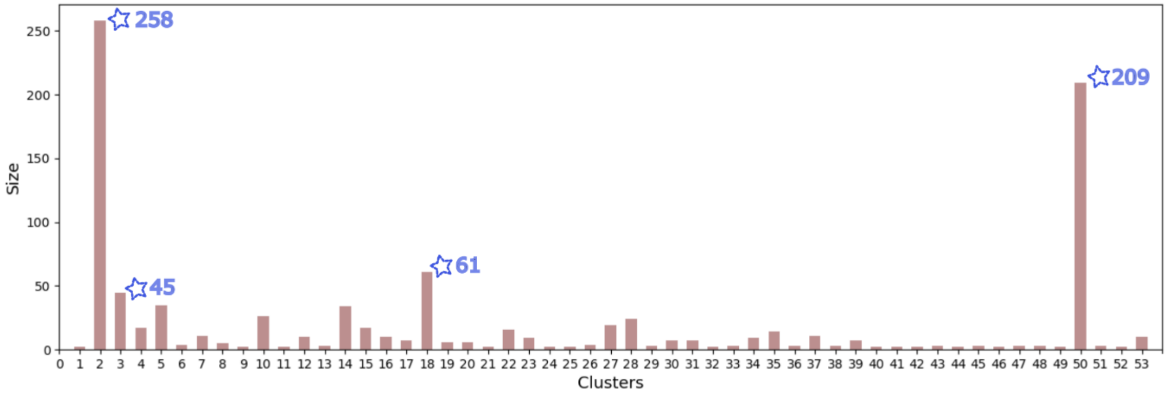


Figure 55 – Histogram of clusters sizes and the centroid of the most populated clusters (2, 3, 18 and 50) at 60% similarity threshold.

These results indicated that the metalloproteinase tends to fit better 2 scaffolds in its active site. This type of clustering tended to produce many singletons and few large clusters. To verify if indeed it was true, a hierarchical method was also used, Ward clustering also implemented in the Chemaxon package. This approach takes as input the chemical fingerprints calculated by the GenerateMD tool and measures the distance between the clusters based on the Euclidean distance.

Figure 56 shows a histogram of the chemical space resulted from the ward tool.

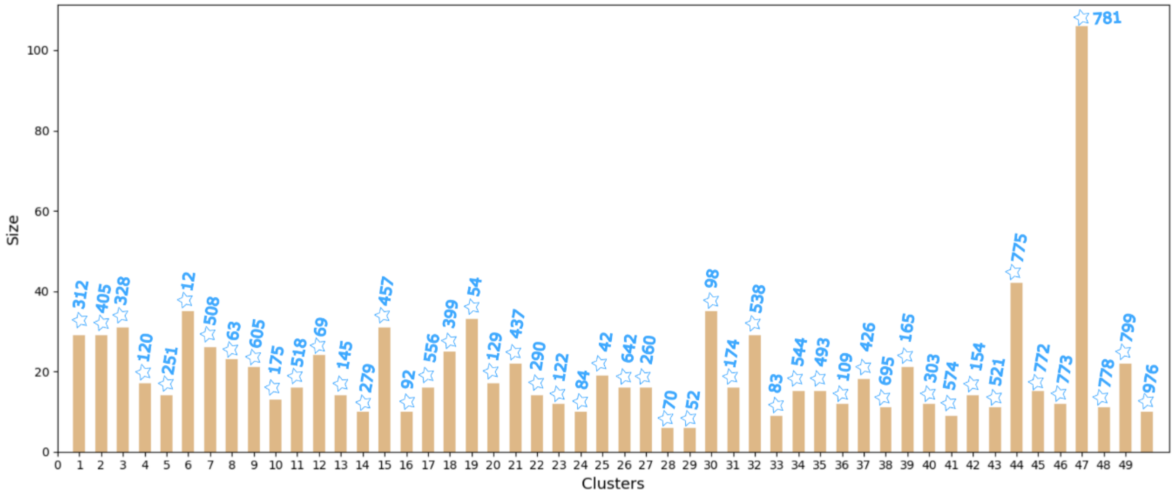


Figure 56 - Histogram of clusters sizes and the centroid of the most populated clusters represented by blue stars.

Results indicate that the studied dataset covers a large chemical space considering the well-distributed data with an average size of 20 molecules per cluster.

As a result of the comparison and after the clusterization of the molecules according to the Euclidean distances, an optimal number of 50 chemical representatives was obtained.

In order to select 20 molecules from the obtained dataset, meticulous visual inspection and analysis of the binding poses, molecular interactions with the active site and binding energy were carried out.

3.6.4. SBVS

In addition to the ligand library extracted from ZINC based on the metalloproteinase pharmacophore, the screening process was also repeated on a library of 3.5 million diverse compounds.

3.6.4.1. Molecular Docking, characterization and Clustering

The analyses were carried out using the same workflow as in the LB approach. *Figure 57* shows the repartition of the physicochemical properties of the top 1000 compounds.

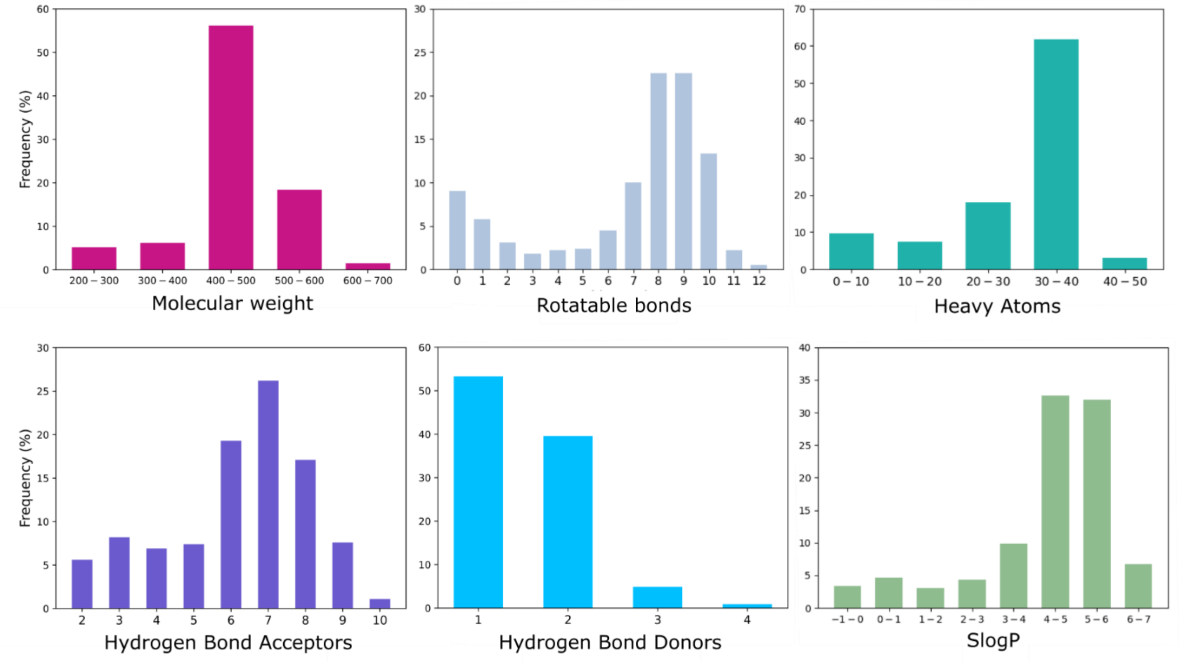


Figure 57 - Physicochemical properties of the top 1000 ranked compounds after docking.

According to the obtained data, 52.4% and 9.7% of the compounds totally fulfilled the requirements of the Lipinski's rules and the rules of 3, respectively.

56.2% of the top 1000 compounds had a MW between 400-500 Da, while 18.4% were in the range of 500–600 Da. The number of rotatable bonds varies a lot within the dataset, with the majority having between 8 and 10. Approximately 63% of the compounds have between 6-8 HBA and 93% have between 1-2 HBD both agreeing with the Lipisnki's rules and the last with the rules of 3. As for the LogP, the results show that 64.6% of the compounds vary between 4-6 LogP.

By comparing the physicochemical properties of the top 1000 compounds obtained in the two different approaches, it was possible to verify that, with the exception of the MW they showed different characteristics. Thus, it is reasonable to conclude that the two approaches allowed to explore a wider chemical space.

Following up the protocol used for the LBVS, these compounds were then clustered by chemical similarity (*Fig. 58*) using the Jarvis-Patrick method.

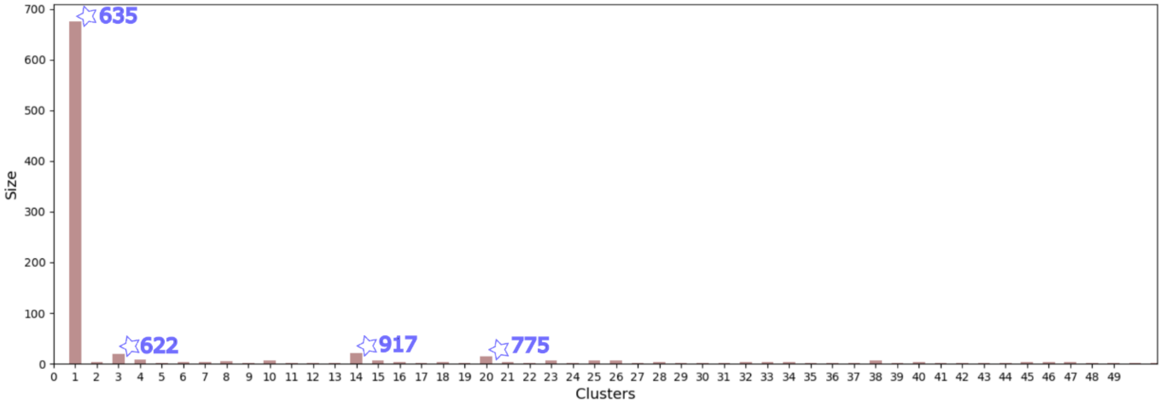


Figure 58 - Histogram of clusters sizes and the centroid of the most populated clusters (1, 3, 14 and 20) represented as slate blue at 70% similarity threshold.

At 70% similarity threshold, the 1000 compounds were clustered into 55 centroids and 114 singletons. The average and maximum dissimilarities within the global dataset were 82.8% and 100%, respectively.

At 60% similarity threshold, these were clustered into 63 centroids and 180 singletons. At 50% similarity threshold, these were clustered into 70 centroids and 257 singletons. In all the cases, singletons represent a large portion of the dataset. Furthermore, the average size of the obtained clusters is very small, with most of them containing between one to three molecules as occurred in the previous approach. For that reason, the *ward* tool was also used here.

Figure 59 shows a histogram of the clusters sizes resulted from the *ward* tool.

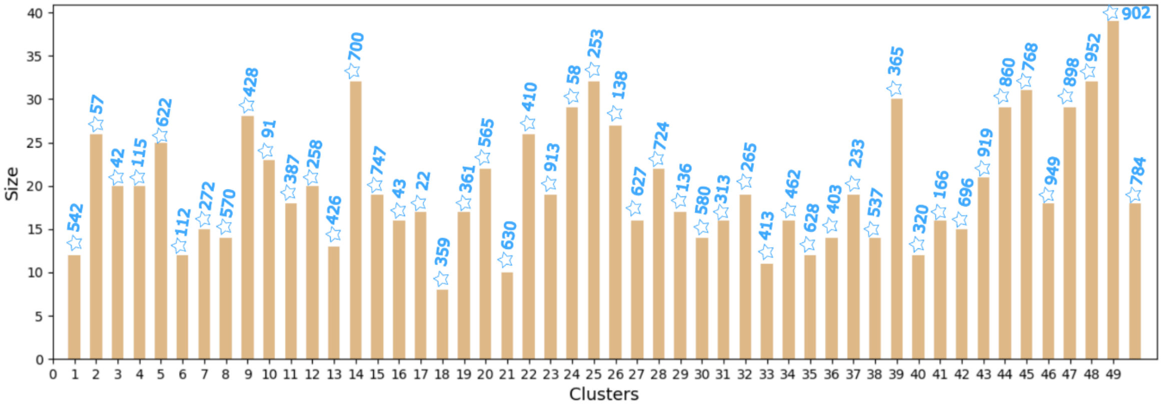


***Figure 59** - Histogram of clusters sizes based on their Euclidean distances and the centroid of the most populated clusters represented by blue stars.*

As shown in *Fig. 59*, the compounds are well-distributed indicating that this dataset covers a large chemical space. As in the previous approach, this clustering method revealed to be more specific of our study.

As a result, 50 chemical representatives were obtained. In order to select 30 molecules from the obtained dataset, meticulous visual inspection and analysis of the binding poses, molecular interactions with the active site and binding energy were carried out.

4. Conclusions

This work aimed to acquire deeper knowledge about the dynamics and reaction mechanism of RVV-X from *D. russelii russelii*, and to identify lead compounds against its enzymatic activity.

Here, *in silico* methods were used to predict the 3D structures of the RVV-X and FX. A 3D model comparable to the crystallographic structure of *D. siamensis* was assembled for the RVV-X. The stability of the protein complex was assessed with MD simulations. Overall, the zinc coordination sphere specific for this enzyme was parametrized as it has not been done before. Also, global folding of the complex remained stable during the t refinement steps, confirming the fitness of the model. Overall, a full structural model of the zymogenic form of the FX was developed and validated.

A protein-peptide docking method was carried out to understand the interactions between the target protein and its substrate. The most stable and suitable structure to the study of the reaction mechanism was the one that possessed a substrate orientation in agreement with the proposed by Takeda in 2007. Despite having the same orientation, the Takeda model did not consider the pre-reactive conformation, i.e, it did not take into account the atoms orientation and approximations that are needed to the enzymatic reaction occur. Instead, the chosen model does and displays distances that could facilitate the catalysis to occur. This model provided a good basis for the study of the mechanistic.

Furthermore, a computer-aided investigation of the reaction mechanism was performed. Overall, this study emphasizes the importance of zinc, the catalytic water molecule and glutamate for the catalysis to occur. It was also found that the nucleophilic species that promotes the attack to the scissile peptide bond is the hydroxide (HO^-). The first step of the mechanism had an activation barrier of 14.4 kcal.mol^{-1} and the second step had an energy barrier of 8.0 kcal.mol$^{-1.}$ These energy barriers agreed with the already proposed mechanism of zinc metalloproteinases.

Finally, two drug screening protocols were used, LB and SB. In both approaches, the compounds were docked on the target protein and the top 2000 compounds (1000 per screening) were then clustered according to their physicochemical properties and subsequently visually inspected.

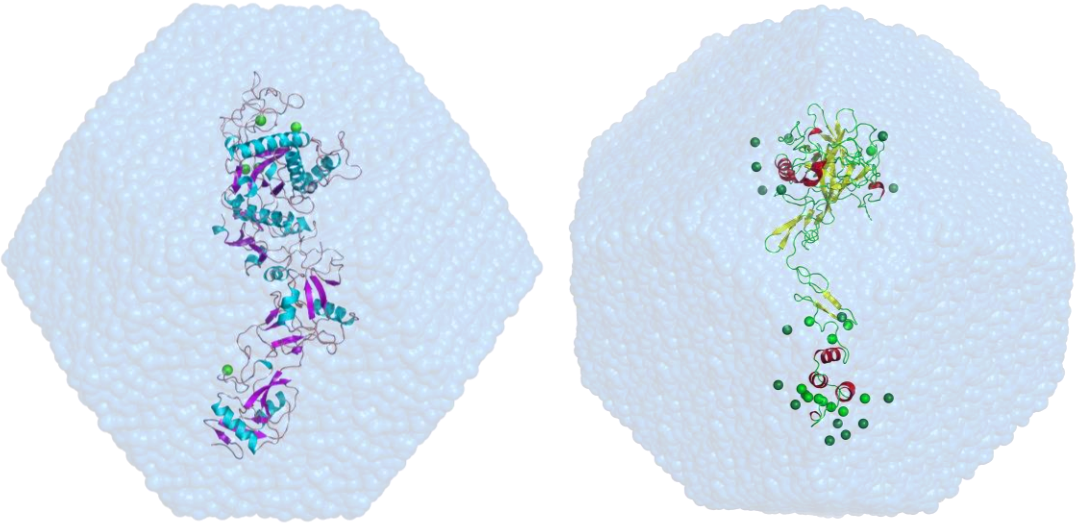
These analyses allowed to obtain more detailed information about the compound chemical proprieties that the active site prefers in terms of structure-activity relationship.

In the end, 50 putative metalloproteinase inhibitors with chemical diversity were selected, 20 from the LB method and 30 from the SB method. Soon, all the compounds will be tested in vitro at Prof. Ashis Mukherjee laboratory at the University of Tezpur, India.

In summary, the results of this study shed lights on the dynamics of the protein under study, disclosed its reaction mechanism and a list of potential hit inhibitors was created.

As future work, it would be necessary to carry out more steps of the QM/MM calculations, such as zero-point, thermal and entropic (rigid rotor/harmonic oscillator) corrections to obtain free energies for all species. Also, recalculate the final energies with higher theoretical levels, or with larger basis sets. As the reaction occurs on the surface of the protein, it is expected that the solvent dynamics has a significant role. Thus, it would also be necessary to test different thickness of the water layer to inspect its influence on the potential energy profiles.

5. Annexes



(1) Protein systems after simulation of physiological conditions. (A) Representation of the RVV-X (New Cartoon), chloride counterions ions (green spheres), and the solvent box (blue surface); (B) Representation of the FX (New Cartoon), chloride counterions ions (dark green spheres), calcium ions (light green spheres) and the solvent box (blue surface).

(2) Table with the key distances of the stationary points (ES, TS1, INT1, TS2, INT2, TS3, INT3) along the reaction pathway from the QM(B3LYP/6-31G(d))/AMBER calculations.

		Min	**REP5**		
Stationary point	Key distances (Å)	A	B1	B2	B3
ES	Zn-O$_{pep}$	2.39	2.14	2.15	2.22
	Zn-O$_{wat}$	1.97	2.10	2.11	2.07
	O$_{wat}$- C$_{pep}$	3.03	2.57	2.54	2.55
	OE2$_{Glu}$-N$_{pep}$	4.61	4.20	4.17	3.92
	C$_{pep}$-N$_{pep}$	1.05	1.87	1.85	1.47
TS1	Zn-O$_{pep}$	1.97			
	Zn-O$_{wat}$	2.21			
	O$_{wat}$- C$_{pep}$	1.74			
	OE2$_{Glu}$-N$_{pep}$	3.77			
	C$_{pep}$-N$_{pep}$	1.38			
INT1	Zn-O$_{pep}$	1.89			
	Zn-O$_{wat}$	2.9			
	O$_{wat}$- C$_{pep}$	1.49			
	OE2$_{Glu}$-N$_{pep}$	3.90			
	C$_{pep}$-N$_{pep}$	1.45			
TS2	Zn-O$_{pep}$	1.95			
	Zn-O$_{wat}$	2.35			
	O$_{wat}$- C$_{pep}$	1.43			
	OE2$_{Glu}$-N$_{pep}$	2.56			
	C$_{pep}$-N$_{pep}$	1.57			
INT2	Zn-O$_{pep}$	1.98			
	Zn-O$_{wat}$	2.33			
	O$_{wat}$- C$_{pep}$	1.39			
	OE2$_{Glu}$-N$_{pep}$	1.76			
	C$_{pep}$-N$_{pep}$	1.68			
TS3	Zn-O$_{pep}$	1.99			
	Zn-O$_{wat}$	2.33			
	O$_{wat}$- C$_{pep}$	1.37			
	OE2$_{Glu}$-N$_{pep}$	2.82			
	C$_{pep}$-N$_{pep}$	1.78			
INT3	Zn-O$_{pep}$	1.95			
	Zn-O$_{wat}$	2.83			
	O$_{wat}$- C$_{pep}$	1.20			
	OE2$_{Glu}$-N$_{pep}$	5.1			
	C$_{pep}$-N$_{pep}$	3.18			

6. Bibliography

1. Sunagar, K., et al., Deadly innovations: unraveling the molecular evolution of animal venoms. *Venom Genomics and Proteomics*; Springer: Dordrecht, The Netherlands, 2014: p. 1-23.
2. Gutiérrez, J.M., et al., Hemorrhage caused by snake venom metalloproteinases: a journey of discovery and understanding. *Toxins*, 2016. **8**(4): p. 93.
3. Vidal, N., et al., Evolution and diversification of the Toxicofera reptile venom system. *J Proteomics*, 2009. **72**: p. 127-36.
4. Casewell, N.R., Evolution: Gene Co-option Underpins Venom Protein Evolution. *Curr Biol*, 2017. **27**(13): p. R647-R649.
5. Cardoso, F.C., et al., Multifunctional toxins in snake venoms and therapeutic implications: from pain to hemorrhage and necrosis. *Front Ecol Evol*, 2019. **7**: p. 218.
6. Takeda, S., ADAM and ADAMTS family proteins and snake venom metalloproteinases: A structural overview. *Toxins*, 2016. **8**(5): p. 155.
7. Casewell, N.R., et al., Complex cocktails: the evolutionary novelty of venoms. *Trends Ecol Evol*, 2013. **28**(4): p. 219-229.
8. Kini, R.M. and C.Y. Koh, Metalloproteases affecting blood coagulation, fibrinolysis and platelet aggregation from snake venoms: Definition and nomenclature of interaction sites. *Toxins*, 2016. **8**(10): p. 284.
9. Mohamed Abd El-Aziz, T., A.G. Soares, and J.D. Stockand, Snake venoms in drug discovery: valuable therapeutic tools for life saving. *Toxins*, 2019. **11**(10): p. 564.
10. Fox, J.W. and S.M. Serrano, Insights into and speculations about snake venom metalloproteinase (SVMP) synthesis, folding and disulfide bond formation and their contribution to venom complexity. *The FEBS journal*, 2008. **275**(12): p. 3016-3030.
11. Moura-da-Silva, A.M., et al., Processing of snake venom metalloproteinases: generation of toxin diversity and enzyme inactivation. *Toxins*, 2016. **8**(6): p. 183.
12. Kalita, B., S.P. Mackessy, and A.K. Mukherjee, Proteomic analysis reveals geographic variation in venom composition of Russell's Viper in the Indian subcontinent: implications for clinical manifestations post-envenomation and antivenom treatment. *Expert review of proteomics*, 2018. **15**(10): p. 837-849.
13. Casewell, N.R., et al., Causes and consequences of snake venom variation. *Trends Pharmacol Sci*, 2020.
14. Laxme, R.S., et al., Beyond the 'big four': Venom profiling of the medically important yet neglected Indian snakes reveals disturbing antivenom deficiencies. *PLoS Negl Trop Dis*, 2019. **13**(12): p. e0007899.
15. Faisal, T., et al., Proteomics, functional characterization and antivenom neutralization of the venom of Pakistani Russell's viper (Daboia russelii) from the wild. *J Proteomics*, 2018. **183**: p. 1-13.
16. Albulescu, L.-O., et al., A therapeutic combination of two small molecule toxin inhibitors provides broad preclinical efficacy against viper snakebite. *Nat Commun*, 2020. **11**(1): p. 1-14.
17. Xie, C., et al., Neutralizing effects of small molecule inhibitors and metal chelators on coagulopathic Viperinae snake venom toxins. *bioRxiv*, 2020.
18. da Costa Neves-Ferreira, A.G., et al., 12 Enzyme Inhibitors in Reptile Venoms and Innate Immunity to Snake Venoms. 2009.

19. Valente, R.H., et al., BJ46a, a snake venom metalloproteinase inhibitor: Isolation, characterization, cloning and insights into its mechanism of action. *Eur J Biochem*, 2001. **268**(10): p. 3042-3052.
20. Aoki-Shioi, N., C.Y. Koh, and R.M. Kini, Natural inhibitors of snake venom metalloproteinases. Aust J Chem, 2020. **73**(4): p. 277-286.
21. Kasturiratne, A., et al., The Global Burden of Snakebite: A Literature Analysis and Modelling Based on Regional Estimates of Envenoming and Deaths. PLoS Med, 2008. **5**(11): p. e218.
22. Shorto, S., The Social Construction of Political Priority for Global Health Issues: A Case Study of Snakebite. 2016.
23. Williams, D.J., et al., Strategy for a globally coordinated response to a priority neglected tropical disease: Snakebite envenoming. *PLoS Negl Trop Dis*, 2019. **13**(2): p. e0007059.
24. Mukherjee, A.K., B. Kalita, and S.P. Mackessy, A proteomic analysis of Pakistan Daboia russelii russelii venom and assessment of potency of Indian polyvalent and monovalent antivenom. *J Proteomics*, 2016. **144**: p. 73-86.
25. Sharma, M., et al., Unveiling the complexities of Daboia russelii venom, a medically important snake of India, by tandem mass spectrometry. *Toxicon*, 2015. **107**: p. 266-281.
26. Senji Laxme, R.R., et al., Biogeographic venom variation in Russell's viper (Daboia russelii) and the preclinical inefficacy of antivenom therapy in snakebite hotspots. *PLoS Negl Trop Dis*, 2021. **15**(3): p. e0009247.
27. Thakur, R., A. Mukherjee, and M. Biotechnology, A brief appraisal on Russell's Viper venom (Daboia russelii russelii) proteinases. *Snake Venoms*, 2015: p. 1-18.
28. Pla, D., et al., Phylovenomics of Daboia russelii across the Indian subcontinent. Bioactivities and comparative in vivo neutralization and in vitro third-generation antivenomics of antivenoms against venoms from India, Bangladesh and Sri Lanka. *J Proteomics*, 2019. **207**: p. 103443.
29. Suntravat, M., et al., Effect of purified Russell's viper venom-factor X activator (RVV-X) on renal hemodynamics, renal functions, and coagulopathy in rats. *Toxicon*, 2011. **58**(3): p. 230-238.
30. Olaoba, O.T., et al., Snake Venom Metalloproteinases (SVMPs): A structure-function update. *Toxicon*: X, 2020. **7**: p. 100052.
31. Markland Jr, F.S. and S. Swenson, Snake venom metalloproteinases. *Toxicon*, 2013. **62**: p. 3-18.
32. Camacho, E., et al., Novel catalytically-inactive PII metalloproteinases from a viperid snake venom with substitutions in the canonical zinc-binding motif. *Toxins*, 2016. **8**(10): p. 292.
33. Yee, K.T., et al., Snake venom metalloproteinases and their peptide inhibitors from Myanmar Russell's viper venom. *Toxins*, 2017. **9**(1): p. 15.
34. Takeda, S., H. Takeya, and S. Iwanaga, Snake venom metalloproteinases: structure, function and relevance to the mammalian ADAM/ADAMTS family proteins. *Biochim Biophys Acta Proteins Proteom*, 2012. **1824**(1): p. 164-176.
35. Watanabe, L., et al., Amino acid sequence and crystal structure of BaP1, a metalloproteinase from Bothrops asper snake venom that exerts multiple tissue-damaging activities. *Protein Sci*, 2003. **12**(10): p. 2273-2281.
36. Laronha, H. and J. Caldeira, Structure and Function of Human Matrix Metalloproteinases. *Cells*, 2020. **9**: p. 1076.

37. Takeda, S., Structural aspects of the factor X activator RVV-X from Russell's viper venom, *Toxins and Hemostasis*. 2010, Springer. p. 465-484.
38. Takeda, S., T. Igarashi, and H. Mori, Crystal structure of RVV-X: An example of evolutionary gain of specificity by ADAM proteinases. *FEBS Lett*, 2007. **581**(30): p. 5859-5864.
39. Venkateswarlu, D., et al., Structure and dynamics of zymogen human blood coagulation factor X. *Biophys J,* 2002. **82**(3): p. 1190-1206.
40. INOUE, K. and T. MORITA, Identification of O-linked oligosaccharide chains in the activation peptides of blood coagulation factor X: The role of the carbohydrate moieties in the activation of factor X. *Eur J Biochem,* 1993. **218**(1): p. 153-163.
41. Rudolph, A.E., et al., The role of the factor X activation peptide: a deletion mutagenesis approach. *J Thromb Haemost*, 2002. **88**(11): p. 756-762.
42. Chattopadhyay, A. and D. Fair, Molecular recognition in the activation of human blood coagulation factor X. *J Biol Chem*, 1989. **264**(19): p. 11035-11043.
43. Morita, T., Structures and functions of snake venom CLPs (C-type lectin-like proteins) with anticoagulant-, procoagulant-, and platelet-modulating activities. *Toxicon*, 2005. **45**(8): p. 1099-1114.
44. Wang, S.X., et al., The Extended Interactions and Gla Domain of Blood Coagulation Factor Xa. *Biochemistry*, 2003. **42**(26): p. 7959-7966.
45. Sunnerhagen, M., et al., The Relative Orientation of Gla and EGF Domains in Coagulation Factor X Is Altered by Ca2+ Binding to the First EGF Domain. A Combined NMR− Small Angle X-ray Scattering Study. *Biochemistry*, 1996. **35**(36): p. 11547-11559.
46. Stojanovski, B.M., L.A. Pelc, and E. Di Cera, Role of the activation peptide in the mechanism of protein C activation. *Sci Rep*, 2020. **10**(1): p. 11079.
47. Parasuraman, S., *Protein data bank. J Pharmacol Pharmacother*, 2012. **3**(4): p. 351-352.
48. Sander, C. and R. Schneider, Database of homology-derived protein structures and the structural meaning of sequence alignment. *Proteins: Structure, Function, and Bioinformatics*, 1991. **9**(1): p. 56-68.
49. Bastos, V.A., et al., Natural inhibitors of snake venom metalloendopeptidases: history and current challenges. *Toxins*, 2016. **8**(9): p. 250.
50. N Cavasotto, C., Homology models in docking and high-throughput docking. *Curr Top Med Chem*, 2011. **11**(12): p. 1528-1534.
51. Consortium, T.U., UniProt: the universal protein knowledgebase in 2021. *Nucleic Acids Res*, 2020. **49**(D1): p. D480-D489.
52. UniProt, C., The universal protein resource (UniProt). *Nucleic Acids Res*, 2008. **36**(Database issue): p. D190-D195.
53. Waterhouse, A., et al., SWISS-MODEL: homology modelling of protein structures and complexes. *Nucleic Acids Res*, 2018. **46**(W1): p. W296-W303.
54. Biasini, M., et al., SWISS-MODEL: Modelling protein tertiary and quaternary structure using evolutionary information. *Nucleic Acids Res*, 2014. **42**.
55. Madden, T., The BLAST sequence analysis tool, in *The NCBI Handbook [Internet]*. 2nd edition. 2013, National Center for Biotechnology Information (US).
56. Remmert, M., et al., HHblits: Lightning-fast iterative protein sequence searching by HMM-HMM alignment. *Nature methods*, 2011. **9**: p. 173-5.
57. Benkert, P., M. Biasini, and T. Schwede, Toward the estimation of absolute quality of individual protein structure models. *Bioinformatics* (Oxford, England), 2011. **27**: p. 343-50.

58. DeLano, W.L., The PyMOL Molecular Graphics System. De-Lano Scientific, San Carlos, CA, USA. http://www. pymol. org, 2002.
59. Humphrey, W., A. Dalke, and K. Schulten, VMD: visual molecular dynamics. *J Mol Graph*, 1996. **14**(1): p. 33-38.
60. Laskowski, R.A., et al., PROCHECK: a program to check the stereochemical quality of protein structures. *J Appl Crystallog*, 1993. **26**(2): p. 283-291.
61. Wiederstein, M. and M.J. Sippl, ProSA-web: interactive web service for the recognition of errors in three-dimensional structures of proteins. *Nucleic acids Res*, 2007. **35**(suppl_2): p. W407-W410.
62. Sippl, M.J., Recognition of errors in three-dimensional structures of proteins. *Proteins: Structure, Function, and Bioinformatics*, 1993. **17**(4): p. 355-362.
63. Wang, D., W. Bode, and R. Huber, Bovine chymotrypsinogen A: X-ray crystal structure analysis and refinement of a new crystal form at 1.8 Å resolution. *J Mol Bio*, 1985. **185**(3): p. 595-624.
64. Gomis-Rüth, F., et al., The three-dimensional structure of the native ternary complex of bovine pancreatic procarboxypeptidase A with proproteinase E and chymotrysinogen C. *The EMBO journal*, 1995. **14**: p. 4387-94.
65. Eswar, N., et al., Protein structure modeling with MODELLER. *Methods Mol Biol*, 2008. **426**: p. 145-59.
66. Salomon-Ferrer, R., D.A. Case, and R.C. Walker, An overview of the Amber biomolecular simulation package. *Wiley Interdiscip Rev Comput Mol Sci*, 2013. **3**(2): p. 198-210.
67. Li, P. and K.M. Merz Jr, MCPB. py: A python based metal center parameter builder. 2016, *ACS Publications*.
68. Anandakrishnan, R., B. Aguilar, and A.V. Onufriev, H++ 3.0: automating p K prediction and the preparation of biomolecular structures for atomistic molecular modeling and simulations. *Nucleic acids Res*, 2012. **40**(W1): p. W537-W541.
69. Gordon, J.C., et al., H++: a server for estimating p K as and adding missing hydrogens to macromolecules. *Nucleic acids Res*, 2005. **33**(suppl_2): p. W368-W371.
70. Beck, A.D., Density-functional thermochemistry. III. The role of exact exchange. *J Chem Phys*, 1993. **98**(7): p. 5648-6.
71. Dill, J.D. and J.A. Pople, Self-consistent molecular orbital methods. XV. Extended Gaussian-type basis sets for lithium, beryllium, and boron. *J Chem Phys*, 1975. **62**(7): p. 2921-2923.
72. Zheng, G., et al., Gaussian 09. 2009, Gaussian Inc., Wallingford CT.
73. Case, D.A., et al., Amber 2020. 2020.
74. van Zundert, G.C.P., et al., The HADDOCK2.2 Web Server: User-Friendly Integrative Modeling of Biomolecular Complexes. *J Mol Biol*, 2016. **428**(4): p. 720-725.
75. Zhou, P., et al., HPEPDOCK: a web server for blind peptide-protein docking based on a hierarchical algorithm. *Nucleic Acids Res*, 2018. **46**(W1): p. W443-w450.
76. Zhou, P., et al., Hierarchical Flexible Peptide Docking by Conformer Generation and Ensemble Docking of Peptides. *J Chem Inf Model*, 2018. **58**(6): p. 1292-1302.
77. Sargsyan, K., C.d. Grauffel, and C. Lim, How molecular size impacts RMSD applications in molecular dynamics simulations. *J Chem Theory Comput*, 2017. **13**(4): p. 1518-1524.

78. Méndez, R., et al., Assessment of blind predictions of protein–protein interactions: current status of docking methods. *Proteins: Structure, Function, and Bioinformatics*, 2003. **52**(1): p. 51-67.
79. Do, P.-C., E.H. Lee, and L. Le, Steered Molecular Dynamics Simulation in Rational Drug Design. *J Chem Inf Model*, 2018. **58**(8): p. 1473-1482.
80. Senn, H.M. and W. Thiel, QM/MM methods for biomolecular systems. *Angewandte Chemie International Edition*, 2009. **48**(7): p. 1198-1229.
81. Groenhof, G., Introduction to QM/MM simulations. Biomolecular Simul, 2013: p. 43-66.
82. Hohenberg, P. and W. Kohn, Density functional theory (DFT). Phys. Rev, 1964. **136**: p. B864.
83. Warshel, A. and M. Levitt, Theoretical studies of enzymic reactions: Dielectric, electrostatic and steric stabilization of the carbonium ion in the reaction of lysozyme. *J Mol Bio*, 1976. **103**(2): p. 227-249.
84. Magalhães, R.P., H.S. Fernandes, and S.F. Sousa, Modelling enzymatic mechanisms with QM/MM approaches: current status and future challenges. *Israel J Chem*, 2020. **60**(7): p. 655-666.
85. Svensson, M., et al., ONIOM: a multilayered integrated MO+ MM method for geometry optimizations and single point energy predictions. A test for Diels− Alder reactions and Pt (P (t-Bu) 3) 2+ H2 oxidative addition. *J Phys Chem*, 1996. **100**(50): p. 19357-19363.
86. S. Fernandes, H., M.J. Ramos, and N. MFSA Cerqueira, molUP: A VMD plugin to handle QM and ONIOM calculations using the gaussian software. 2018, *Wiley Online Library.*
87. Maia, E.H.B., et al., Octopus: a platform for the virtual high-throughput screening of a pool of compounds against a set of molecular targets. *J Mol Model*, 2017. **23**(1): p. 26.
88. Maia, E.H.B., et al., Structure-Based Virtual Screening: From Classical to Artificial Intelligence. *Front Chem*, 2020. **8**(343).
89. Igarashi, T., et al., Crystal structures of catrocollastatin/VAP2B reveal a dynamic, modular architecture of ADAM/adamalysin/reprolysin family proteins. *FEBS Lett*, 2007. **581**(13): p. 2416-22.
90. Gomis-Rüth, F.X., et al., Structures of adamalysin II with peptidic inhibitors. Implications for the design of tumor necrosis factor alpha convertase inhibitors. *Protein Sci*, 1998. **7**(2): p. 283-292.
91. Botos, I., et al., Batimastat, a potent matrix mealloproteinase inhibitor, exhibits an unexpected mode of binding. *Proc Natl Acad Sci U.S.A*, 1996. **93**(7): p. 2749-2754.
92. Zhang, D., et al., Structural interaction of natural and synthetic inhibitors with the venom metalloproteinase, atrolysin C (form d). *Proc Natl Acad Sci U.S.A*, 1994. **91**(18): p. 8447-8451.
93. Tortorella, M.D., et al., Structural and inhibition analysis reveals the mechanism of selectivity of a series of aggrecanase inhibitors. *J Bio Chem*, 2009. **284**(36): p. 24185-24191.
94. Lingott, T., et al., High-resolution crystal structure of the snake venom metalloproteinase BaP1 complexed with a peptidomimetic: insight into inhibitor binding. *Biochemistry*, 2009. **48**(26): p. 6166-74.
95. Bento, A.P., et al., The ChEMBL bioactivity database: an update. *Nucleic acids Res*, 2014. **42**(D1): p. D1083-D1090.

96. Kim, S., et al., PubChem substance and compound databases. *Nucleic acids Res*, 2016. **44**(D1): p. D1202-D1213.
97. Wishart, D.S., et al., DrugBank 5.0: a major update to the DrugBank database for 2018. *Nucleic acids Res*, 2018. **46**(D1): p. D1074-D1082.
98. Mysinger, M.M., et al., Directory of useful decoys, enhanced (DUD-E): better ligands and decoys for better benchmarking. *J Med Chem*, 2012. **55**(14): p. 6582-6594.
99. Vuorinen, A. and D. Schuster, Methods for generating and applying pharmacophore models as virtual screening filters and for bioactivity profiling. *Methods*, 2015. **71**: p. 113-134.
100. Verdonk, M.L., et al., Improved protein–ligand docking using GOLD. *Proteins: Structure, Function, and Bioinformatics,* 2003. **52**(4): p. 609-623.
101. Huey, R., G.M. Morris, and S. Forli, Using AutoDock 4 and AutoDock vina with AutoDockTools: a tutorial. *The Scripps Research Institute, Molecular Graphics Lab*, 2012. **10550**: p. 92037.
102. Ruiz-Carmona, S., et al., rDock: A Fast, Versatile and Open Source Program for Docking Ligands to Proteins and Nucleic Acids. *PLoS Comput Biol*, 2014. **10**(4): p. e1003571.
103. Schneidman-Duhovny, D., et al., PharmaGist: a webserver for ligand-based pharmacophore detection. *Nucleic Acids Res*, 2008. **36**(suppl_2): p. W223-W228.
104. Sterling, T. and J.J. Irwin, ZINC 15–ligand discovery for everyone. *J Chem Inf Model*, 2015. **55**(11): p. 2324-2337.
105. Lipinski, C.A., et al., Experimental and computational approaches to estimate solubility and permeability in drug discovery and development settings*. Adv Drug Deliv Rev*, 2001. **46**(1): p. 3-26.
106. Congreve, M., et al., A ‘Rule of Three’ for fragment-based lead discovery?. *Drug Discovery Today*, 2003. **8**(19): p. 876-877.
107. Huntington, J.A., Slow thrombin is zymogen-like. *J Thromb Haemos*t, 2009. **7**: p. 159-164.
108. Pelmenschikov, V. and P.E.M. Siegbahn, Catalytic Mechanism of Matrix Metalloproteinases: Two-Layered ONIOM Study. *Inorg Chem*, 2002. **41**(22): p. 5659-5666.
109. Vasilevskaya, T., et al., Mechanism of proteolysis in matrix metalloproteinase-2 revealed by QM/MM modeling. Journal of computational chemistry, 2015. **36**.
110. Vasilevskaya, T., et al., Methodological aspects of QM/MM calculations: A case study on matrix metalloproteinase-2*. J Comput Chem*, 2016. **37**(19): p. 1801-9.
111. Chen, H., et al., On Evaluating Molecular-Docking Methods for Pose Prediction and Enrichment Factors. *J Chem Inf Model*, 2008. **46**: p. 401-15.