DEPARTMENT OF
LIFE SCIENCES

NOVA SCHOOL OF
SCIENCE & TECHNOLOGY

ANDRÉ RAFAEL FERREIRA SALGUEIRO

BSc in Health Sciences

ARTIFICIAL INTELLIGENCE-BASED DESIGN OF ANTIBODY-LIKE ENGINEERED PROTEIN SCAFFOLDS

MASTER IN COMPUTATIONAL BIOLOGY AND BIOINFORMATICS

NOVA University Lisbon
September, 2024

NOVA SCHOOL OF
SCIENCE & TECHNOLOGY

DEPARTMENT OF
LIFE SCIENCES

ARTIFICIAL INTELLIGENCE-BASED DESIGN OF ANTIBODY-LIKE ENGINEERED PROTEIN SCAFFOLDS

ANDRÉ RAFAEL FERREIRA SALGUEIRO

BSc in Health Sciences

Adviser: Diana Lousa
Assistant Researcher, ITQB NOVA

Co-adviser: Leonardo Vanneschi
Full Professor, NOVA IMS

Examination Committee

Chair: Paula Maria Theriaga Mendes Bernardo Gonçalves
Associate Professor, FCT NOVA

Rapporteur: Manuel Nuno de Sousa Pereira Simões de Melo
Assistant Researcher, ITQB NOVA

Adviser: Diana Andreia Pereira Lousa
Assistant Researcher, ITQB NOVA

MASTER IN COMPUTATIONAL BIOLOGY AND BIOINFORMATICS

NOVA University Lisbon
September, 2024

Artificial Intelligence-based Design of Antibody-like Engineered Protein Scaffolds

This document was created with the (pdf/Xe/Lua)LaTeX processor and the NOVAthesis template (v7.1.18) [1].

Acknowledgements

I would like to express my deepest gratitude to my supervisors, Doctor Diana Lousa and Professor Leonardo Vanneschi, for their invaluable guidance, support, and encouragement throughout this project. In particular, a thank you to Doctor Diana Lousa, that acompanied me during my stay at the Protein Modeling Lab. Their expertise, insights have been fundamental in shaping the direction of my research, and their constructive feedback has pushed me to think critically and strive for excellence. I am truly fortunate to have had this opportunity.

A special thank you goes to PhD student Pedro Moreira, whose assistance was indispensable during the course of my research. Your willingness to share your knowledge about Protein Design and your readiness to help at every stage made a significant impact on my progress. I greatly appreciate your time, patience, and the insightful discussions we had, which were crucial to overcoming many challenges.

I would also like to extend my heartfelt thanks to the entire Protein Modeling Lab team. The collaborative atmosphere and constant willingness to help made the lab a desirable working space. In particular, a thank you to Rita, whose tips on Molecular Dynamics made my life easier in the early days; to João, whose insights on this process of writing a thesis for our Master helped me very much; to Benedita, Diogo, Fábio and Madalena, that shared these months of adventure with me as fellow Master students; to Carolina and Catarina, for also sharing the Lab with me. I am grateful for the friendships, shared knowledge, and camaraderie that have made my time here so enriching and enjoyable.

To my family and friends, your unwavering support and belief in me have been a constant source of strength. I am especially thankful to my parents for their unconditional love and encouragement, which have carried me through the most challenging times of this journey.

Lastly, to my lovely girlfriend, Jéssica, words cannot express how grateful I am for your endless patience, understanding, and love throughout this entire process. Your support has meant the world to me, and I could not have reached this point without you by my side.

”

“Algorithms don’t replace scientists; they empower them to make discoveries that were previously impossible.”

— Ewan Birney, Somewhere in an interview or speach (prominent bioinformatician and director of the European Bioinformatics Institute (EMBL-EBI))

Abstract

Viral pandemics have profoundly impacted human societies, underscoring the urgent need for enhanced pandemic preparedness. With limited antiviral treatments, developing new strategies applicable to various viruses is crucial. Monobodies, small proteins with unique characteristics, have emerged as promising antiviral candidates. This thesis aims to create a computational framework streamlining monobody development, positing that language models can generate a large, diverse monobody library suitable for further design.

The research began with the 10^{th} human FN3 domain, using sequence search and the MSA Transformer language model to expand the monobody library. Structures were predicted and docked with the SARS-CoV-2 RBD as proof of concept, followed by optimization. Interface metrics were calculated to filter promising results. Top-performing complexes underwent MD simulations, alongside simulations of RBDs from SARS-CoV-2 variants BA.2.86 and JN.1.

Findings showed the MSA Transformer significantly enhanced the selection process for target-specific antiviral discovery. The most effective approach involved starting from the predicted docked complex structure, with improvements through optimizing the entire sequence.

MD simulations revealed highly dynamic RBDs of SARS-CoV-2 variants, particularly in loop regions. Principal Component Analysis indicated shifts in conformational states, potentially impacting viral infectivity and immune evasion, emphasizing the importance of monitoring mutations and their structural consequences.

This research demonstrates that language models can effectively create a large, diverse monobody library, integrable into a broader antiviral design framework. These advances could significantly enhance pandemic preparedness by streamlining target-specific antiviral development.

Keywords: Monobodies, Protein Language Models, Antiviral Design, Molecular Dynamics Simulations, Viral Variants

Resumo

As pandemias virais têm impactado profundamente as sociedades humanas, sublinhando a necessidade urgente de melhor preparação pandémica. Com tratamentos antivirais limitados, é crucial desenvolver novas estratégias aplicáveis a vários vírus. Os monobodies, pequenas proteínas com características únicas, surgiram como promissores candidatos antivirais. Esta tese visa criar uma estrutura computacional que simplifique o desenvolvimento de monobodies, propondo que modelos de linguagem podem gerar uma biblioteca de monobodies grande e diversificada, adequada para design adicional.

A investigação começou com o 10º domínio FN3 humano, utilizando pesquisa de sequências e o modelo de linguagem MSA Transformer para expandir a biblioteca de monobodies. As estruturas foram previstas e acopladas com o RBD do SARS-CoV-2 como prova de conceito, seguidas de otimização. Calcularam-se métricas de interface para filtrar resultados promissores. Os complexos com melhor desempenho foram submetidos a simulações de MD, juntamente com simulações de RBDs das variantes BA.2.86 e JN.1 do SARS-CoV-2.

Os resultados mostraram que o MSA Transformer melhorou significativamente o processo de seleção para a descoberta antiviral específica. A abordagem mais eficaz envolveu começar pela estrutura do complexo acoplado previsto, com melhorias através da otimização de toda a sequência. As simulações de MD revelaram RBDs altamente dinâmicos das variantes do SARS-CoV-2, particularmente nas regiões de loop. A Análise de Componentes Principais indicou mudanças nos estados conformacionais, potencialmente impactando a infecciosidade viral e evasão imune, enfatizando a importância de monitorizar mutações e suas consequências estruturais.

Esta investigação demonstra que modelos de linguagem podem criar eficazmente uma biblioteca de monobodies grande e diversificada, integrável num quadro mais amplo de design antiviral. Estes avanços poderão melhorar significativamente a preparação pandémica, agilizando o desenvolvimento de antivirais específicos.

Palavras-chave: Monobodies, Modelos de Linguagem para Proteínas, 'Design' Antiviral, Simulações de Dinâmica Molecular, Variantes Virais

Contents

List of Figures

List of Tables

Acronyms

$\Delta\Delta$G	delta-delta-G *(pp. 40, 56, 59, 61–64)*
ACE2	Angiotensin Converting Enzyme 2 *(pp. 2–5, 7–10, 29, 45)*
ALEPS	Antibody-like Engineered Protein Scaffolds *(p. 18)*
BUns	Buried Unsatisfied interface H-bonds *(pp. 40, 56, 59, 61, 63, 64)*
COVID-19	Coronavirus Disease 19 *(pp. 2–4, 7, 8, 10–12, 14)*
ECDC	European Centre for Disease prevention and Control *(p. 10)*
FN3	Fibronectin type III domain *(pp. 13, 37, 48, 61)*
IpTM	Interface predicted Template Modelling score *(pp. 40, 56, 57, 59, 61, 63, 64)*
MD	Molecular Dynamics *(pp. 9, 10, 21–24, 26–31, 43, 45, 61, 68)*
MSA	Multiple Sequence Alignment *(pp. 32, 33)*
pAE	predicted Aligned Error *(pp. 39, 40, 52, 57)*
PBC	Periodic Boundary Conditions *(pp. 22–24)*
PCA	Principal Component Analysis *(pp. 28, 31, 45, 68)*
plDDT	predicted local Distance Difference Test *(pp. 39, 40, 52, 56, 57)*
PPI	Protein-Protein Interactions *(pp. 13, 35, 36)*
pTM	predicted Template Modelling score *(pp. 39, 40, 52, 56, 57)*
RBD	Receptor Binding Domain *(pp. 7–10, 29–31, 39, 43–45, 55–57, 61, 62, 68)*
RBM	Receptor Binding Motif *(pp. 9, 29, 39, 44, 45, 53, 68)*
RMSD	Root Mean Square Deviation *(pp. 27, 28, 31, 44, 61, 62)*
RMSF	Root Mean Square Fluctuation *(pp. 28, 31, 44)*

RSV Respiratory Syncytial Virus *(pp. 12, 15)*

SARS-CoV-2 Severe Acute Respiratory Syndrome Coronavirus 2 *(pp. 2–4, 9, 12, 14, 15, 18, 29–31, 39, 43, 45, 67, 68)*

SC Shape Complementarity *(pp. 40, 56, 59, 61, 63, 64)*

WHO World Health Organization *(p. 4)*

WT Wild Type *(pp. 8, 29, 45)*

Chemical Symbols

Cl^- Negative chlorine ion *(p. 29)*

Na^+ Positive sodium ion *(p. 29)*

1

Introduction

1.1 Viral Pandemics Throughout History

Throughout history, viral pandemics have profoundly impacted human societies, causing widespread mortality and societal disruption. These pandemics, characterized by the global spread of infectious diseases, have repeatedly tested the resilience of humanity. From the ancient world to modern times, viral pandemics have shaped the course of history in significant ways.

One of the earliest recorded viral pandemics was the Antonine Plague, which struck the Roman Empire in 165 AD. Believed to have been caused by either smallpox or measles, it resulted in the deaths of up to five million people [2].

The 20^{th} century witnessed one of the most devastating viral pandemics in history: the Spanish Flu of 1918-1919. Caused by the H1N1 influenza A virus, it infected about a third of the global population and resulted in an estimated 50 to 100 million deaths [3]. Unlike outbreaks caused by typical influenza viruses, the Spanish Flu had a high mortality rate among young adults, contributing to its devastating impact.

Later in the 20^{th} century, the HIV/AIDS pandemic, first identified in the early 1980s but originating in central Africa in the late 1970s, emerged as one of the most significant global health crises in recent history. It has infected over 75 million people worldwide and caused approximately 32 million deaths, primarily through the exchange of bodily fluids and sharing of needles. Despite advances in antiretroviral therapy, HIV/AIDS continues to pose a substantial public health challenge, underscoring the ongoing need for effective prevention and treatment strategies [4].

In the 21^{st} century, several viral outbreaks have also raised global alarm, as exemplified by the H1N1 influenza pandemic of 2009, also known as the Swine Flu. Caused by a novel strain of the H1N1 virus, it spread rapidly, infecting millions worldwide, although its mortality rate was much lower than that of the Spanish Flu [4].

A few years later, between 2014 and 2016, the Ebola virus outbreaks in West Africa demonstrated the lethal potential of viral outbreaks. Characterized by severe hemorrhagic fever, the Ebola virus disease caused over 11 thousand deaths [5].

More recently, Coronavirus Disease 19 (COVID-19) erupted, rapidly becoming the largest pandemic in recent years. Originated in China in late 2019, the COVID-19 pandemic has had a major impact on the world's population, being responsible for millions of deaths.

1.2 The COVID-19 Pandemic

Severe Acute Respiratory Syndrome Coronavirus 2 (SARS-CoV-2) (Figure 1.1), the virus responsible for COVID-19, is a highly transmissible and pathogenic coronavirus that emerged in late 2019 in Wuhan, China. It has since caused a global pandemic, significantly impacting human health and public safety. As a member of the beta-coronavirus genus, SARS-CoV-2 is classified within group 2B, closely related to the Severe Acute Respiratory Syndrome Coronavirus and Middle East Respiratory Syndrome Coronavirus [6–9].

SARS-CoV-2 is highly contagious, primarily spreading through respiratory droplets when an infected person coughs, sneezes, or talks. Also, the virus is capable to spread by touching surfaces contaminated with the virus [9] and then touching the face, especially the mouth, nose, or eyes, although this way of transmission is rarer. The virus's rapid transmission is facilitated by its ability to infect individuals who are asymptomatic or pre-symptomatic, making containment challenging. The virus primarily targets the respiratory system, particularly the lungs. It enters the body through Angiotensin Converting Enzyme

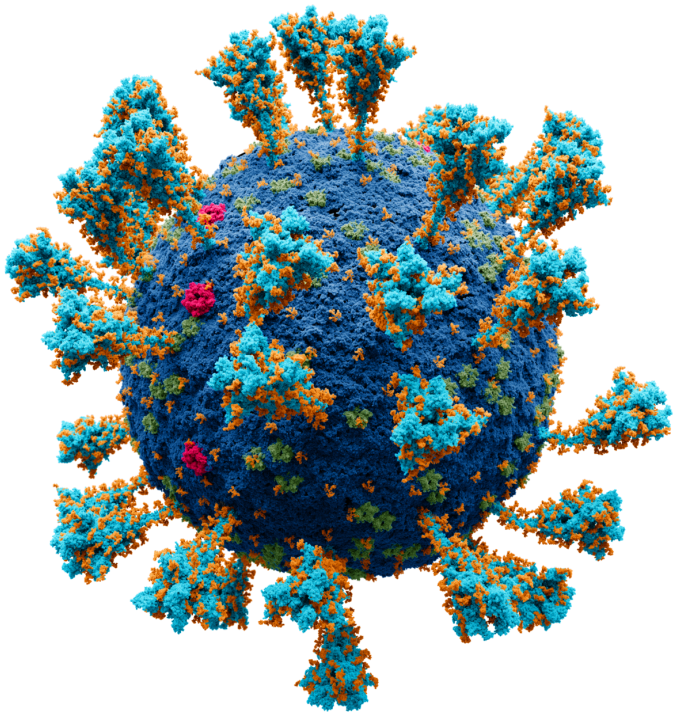
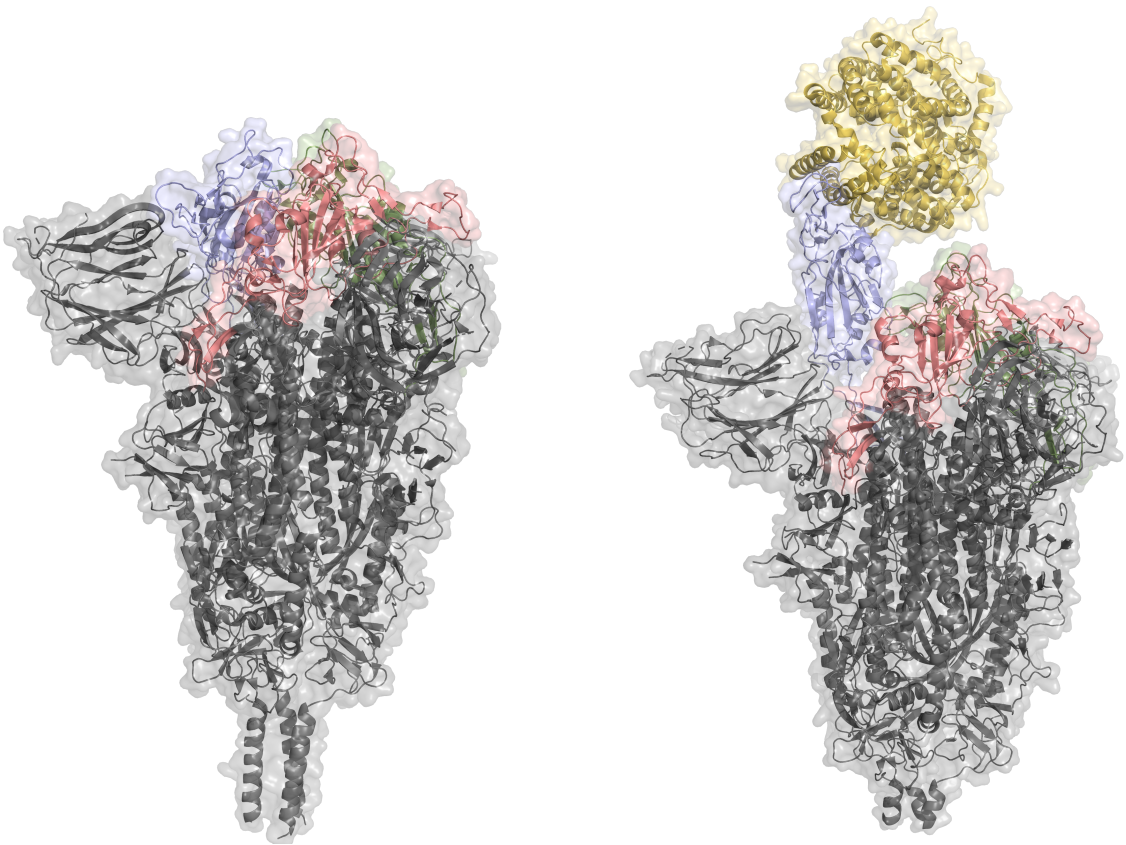


Figure 1.1: Atomic model of the external structure of SARS-CoV-2. The membrane is depicted in cobalt; the E protein is depicted in crimson; the M protein is depicted in green; glucose (glycan) is depicted in orange; the Spike glycoprotein is depicted in turquoise. This figure was adapted from `https://commons.wikimedia.org/wiki/File:Coronavirus._SARS-CoV-2.png`.

2 (ACE2) receptors, which are abundantly expressed on the surface of lung epithelial cells, as well as in other tissues such as the heart, kidneys, and gastrointestinal tract. Upon binding to ACE2, a process mediated by the Spike protein (Figure 1.2), the virus gains entry into the host cell, where it replicates and spreads, causing varying degrees of respiratory illness [8–10].

COVID-19, the disease caused by SARS-CoV-2, presents with a wide spectrum of symptoms ranging from mild to severe. Common symptoms include fever, a common early sign of infection; cough, often dry and persistent; shortness of breath, indicating more severe disease and potential lung involvement; and fatigue, which can be profound and debilitating. In addition to these typical symptoms, some patients experience atypical manifestations, including a sudden and severe loss of smell (anosmia) and complete or partial loss of taste (ageusia). Other symptoms may include headache, muscle pain, sore throat, congestion or runny nose, nausea or vomiting, and diarrhea. In severe cases, COVID-19 can lead to complications such as acute respiratory distress syndrome, multi-organ failure, septic shock, and death. Certain populations, including the elderly and



(a) **(b)**

Figure 1.2: Representation of the SARS-CoV-2 Spike protein. **(a)**: Trimeric form of the protein, with its Receptor Binding Domains in "closed" conformation (PDB ID: 6XR8). **(b)**: Trimeric form of the protein, with one of its Receptor Binding Domains in "open" conformation, bound to the ACE2 receptor, in yellow (PDB ID: 7DF4). Receptor Binding Domains of each monomer are colorized; other Spike protein structures are in gray. This figure is adapted from [11].

those with underlying health conditions like cardiovascular disease, diabetes, chronic respiratory disease, and hypertension, are at higher risk of severe outcomes [11, 12].

The outbreak was first noted in Wuhan, China in November 2019 and the World Health Organization (WHO) reported the first case on December 31^{st}, 2019. WHO declared the outbreak a Public Health Emergency of International Concern on January 30^{th}, 2020 and on March 11^{th}, 2020, the outbreak was declared a global pandemic [8, 9]. According to data from WHO, as of September 8^{th}, 2024, there were more than 776 million confirmed cases of COVID-19 [13], including over 7 million deaths worldwide [14].

Treatment is primarily supportive, however prognosis is dismal in those who need invasive ventilation. Trials are ongoing to discover effective vaccines and drugs to combat the disease. Preventive strategies aim at reducing transmission through contact tracing, hand washing, face masks and government-led lockdowns [8, 9].

1.2.1 Emergence of Variants

Since the onset of the COVID-19 pandemic, the SARS-CoV-2 virus has undergone a continuous process of evolution and diversification [15], giving rise to multiple variants (Figures 1.3 and 1.4). These variants represent a natural part of virus evolution, with each one carrying unique genetic changes (Table 1.1). Some of these variants have drawn significant attention from the global health community due to their potential impact on the course of the pandemic. These concerns revolve around several key factors, including changes in transmissibility, severity, and the effectiveness of vaccines and treatments against these variants. The emergence of new variants underscores the importance of ongoing vigilance and research to better understand their properties and implications.

1.2.1.1 Alpha (B.1.1.7)

The Alpha variant first appeared in late 2020 and was predominant in the United Kingdom in the early months of 2021. N501Y, A570D, del69/70, del144, P681H, T716I, S982A, and D1118H were among the Spike protein mutations found in it. By strengthening its affinity for ACE2, N501Y, a frequent mutation in other variants, boosted transmissibility

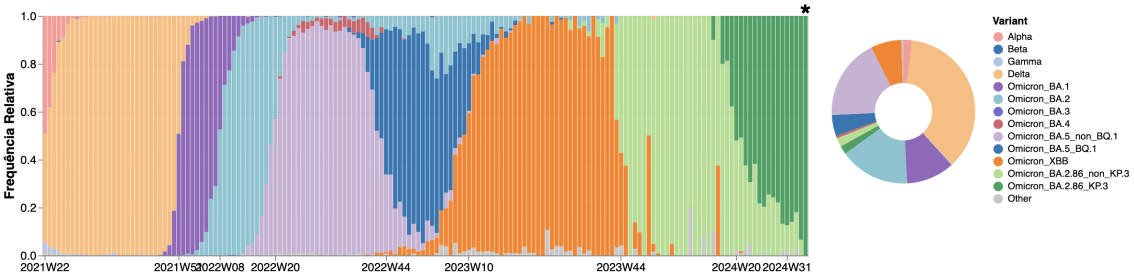
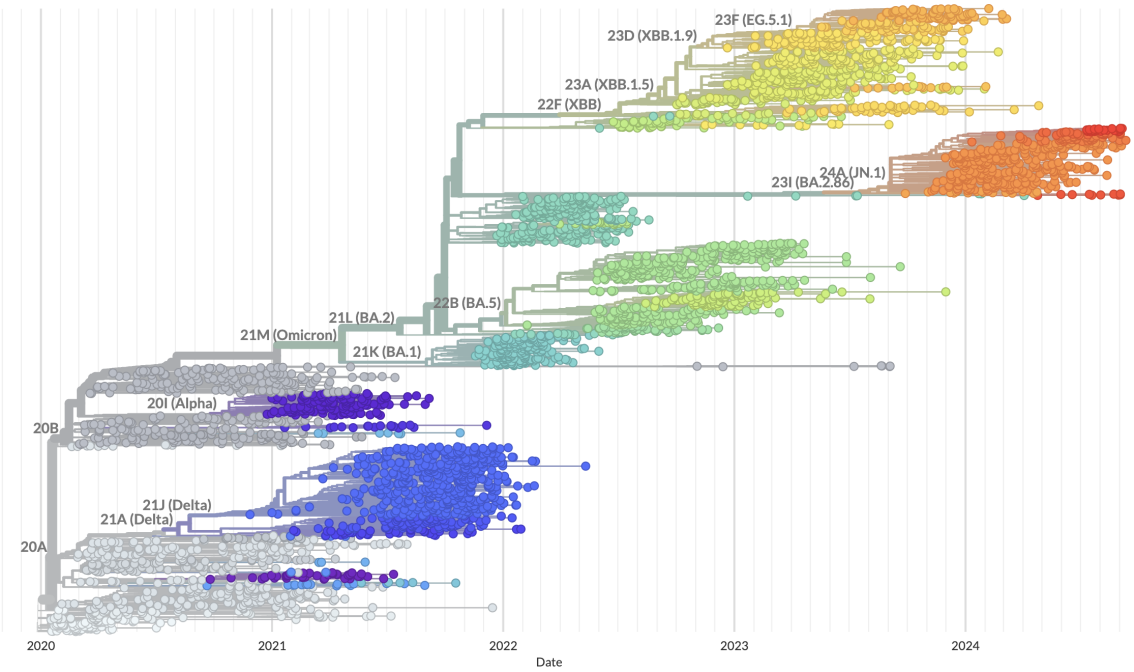
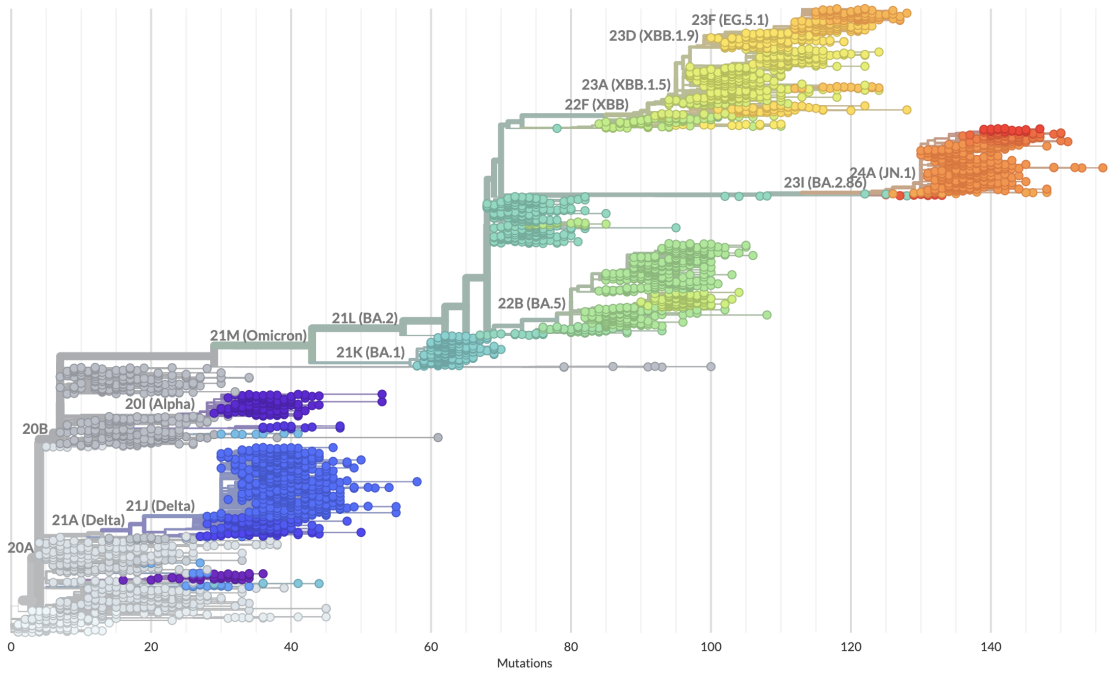


Figure 1.3: Relative frequencies of SARS-CoV-2 variants over time in Portugal. On the X axis, the time scale ranging from week 22 of 2021 to week 31 of 2024, indicated with an asterisk (*). On the Y axis, the relative frequency of the SARS-CoV-2 variants, ranging from 0.0 to 1.0. This figure was adapted from `https://insaflu.insa.pt/covid19/`, accessed on September 27^{th}, 2024.



(a)



(b)

Figure 1.4: Global SARS-CoV-2 phylogenetic tree. Variants from the same clade are represented in the same color. **(a)**: time evolution since January 1^{st}, 2020. **(b)**: divergence in number of mutations. This figure is adapted from `https://nextstrain.org/ncov/gisaid/global/all-time?dmin=2020-01-01`, enabled by data from GISAID updated September 24^{th}, 2024.

[16]. Increased ACE2 binding and resistance to neutralization by antibodies that target the initial epitope were also brought about by a subsequent E484K mutation [17]. Research conducted *in vitro* revealed that serum from individuals vaccinated with ChAdOx1 or BNT162b2 showed less antibody neutralization of B.1.1.7 [18, 19], but some other studies

Table 1.1: Key Spike protein mutations for SARS-CoV-2 variants compared to the wildtype. **X** indicates the presence of the mutation.

Mutation Residue / Position	Alpha (B.1.1.7)	Beta (B.1.351)	Delta (B.1.617.2)	Omicron (B.1.1.529)
T19R			X	
del69–70	X		X	
D80A		X		
G142D			X	X
del144	X			
del156–157			X	
R158G			X	
D215G		X		
del241–243		X		
G339D				X
S373P				X
S375F				X
K417N		X		X
N440K				X
L452R			X	
S477N				X
T478K			X	X
E484A				X
E484K	X	X		
Q498R				X
N501Y	X	X		X
Y505H				X
A570D	X			
D614G	X	X	X	X
H655Y				X
N679K				X
P681H	X			X
P681R			X	
A701V		X		
T716I	X			
N764K				X
D796Y				X
D950N			X	
Q954H				X
N969K				X
S982A	X			
D1118H	X			

showed no significant difference [20–22].

The efficacy of vaccines varies depending on the kind and region. BNT162b2 demonstrated a 97.4% efficacy against severe cases and an 89.5% efficacy against the illness in Qatar [23]. In Brazil, even with a single dosage, ChAdOx1 and CoronaVac decreased fatalities and severe cases in older populations [24]. In the United Kingdom, despite lower neutralization, ChAdOx1 showed about 70% efficacy against symptomatic cases, indicating the possibility of additional protective mechanisms like complement activation, antibody-dependent NK cell activation, or Spike protein-specific T cell induction [25].

1.2.1.2 Beta (B.1.351)

During the second wave of the epidemic, the Beta variant was first discovered in South Africa and was found to be prevalent in multiple provinces. Eight Spike protein mutations, including K417N, E484K, and N501Y, are present in this variant. These mutations affect important receptor-binding sites and may have an effect on antibody neutralization [26]. Whether these mutations raise ACE2 receptor affinity is unknown. Additionally, mutations in the Spike's N-terminal domain can prevent neutralizing antibodies from working [27].

A South African study that examined the neutralization efficiency of Beta variants in different waves using plasma from COVID-19 patients found that plasma form patients from the first wave of infections in South Africa had an 8.4-fold reduced neutralization against the Beta subvariant 501Y.V2, 501Y.V2 plasma showed a 4.1-fold decrease against the first-wave virus. Additionally, 501Y.V2 plasma had only a 2.3-fold reduction against the first-wave virus, but first-wave plasma was 15.1 times less effective against the 501Y.V2 Beta subvariant [28].

The efficacy of vaccines varied. In young South African populations, the Oxford-AstraZeneca vaccine only provided 10% protection against mild-to-moderate disease caused by the Beta variant [29]. The Johnson & Johnson vaccine, on the other hand, demonstrated a 64% protection against moderate-to-severe disease [30]. According to a small Israel study, the Pfizer/BioNTech vaccine seemed less effective against Beta [31], but in Qatar, it showed 75% effectiveness against the variant, with 97.4% effectiveness against severe and fatal cases [23].

1.2.1.3 Delta (B.1.617.2)

Originating in India, the B.1.617 variant was first detected in October 2020 and became a dominant variant in some regions. Its affinity for ACE2 receptors is partly attributed to the Spike protein mutation D614G. There are sub-lineages within B.1.617, the most transmissible of which is B.1.617.2 - also termed the Delta variant - which became the most dominant variant worldwide until late 2021. The Spike protein has several mutations, including those in the N-terminal domain (R158G, T19R, G142D, del156-157), S2 region (D950N), and furin cleavage site (Receptor Binding Domain (RBD) - L452R and T478K). These mutations may improve replication and lessen recognition by neutralizing antibodies [32, 33].

Research on the effectiveness of the Pfizer/BioNTech and Oxford-AstraZeneca vaccines against Delta revealed a decreased susceptibility to antibody neutralization [32, 34], but a study in Israel found similar effectiveness post-vaccination as in clinical trials [35]. The effectiveness of vaccinations against outbreaks caused by the Delta variant was also assessed in China, where internal transmission was rare because of a zero-tolerance policy. Adenovirus vector vaccine Cansino's Ad5 and inactivated virus vaccines CoronaVac/Sinovac and BBIBPCorV were 100% effective against severe cases of COVID-19 and 74.6% effective against symptomatic cases [36]. Following two doses, Pfizer/BioNTech

had approximately 83% effectiveness (97% after the third dose), Oxford-AstraZeneca had approximately 80% effectiveness, and CoronaVac/Sinovac had approximately 65% effectiveness (63% after the third dose), according to a meta-analysis of vaccine effectiveness data. The vaccines' respective efficaciousness against severe cases of the illness was roughly 98% for Pfizer/BioNTech, 91% for Oxford-AstraZeneca, and 75% for CoronaVac/Sinovac [37].

1.2.1.4 Omicron (B.1.1.529)

The Omicron variant was first detected in Botswana and South Africa in November 2021 and quickly spread to other nations. Because of several deletions, including 69-70del, this variant is more difficult to identify with RT-PCR assays. This variant has over 30 mutations in the Spike protein, roughly 15 of which are in the RBD. These mutations are linked to immune escape and decreased neutralization by vaccine-induced antibodies, and they increase transmissibility by improving the Spike protein's affinity for the ACE2 receptor [38–41].

Since the original Omicron variant, several Omicron subvariants have emerged, including BA.1 (B.1.1.529.1), BA.2 (B.1.1.529.2), BA.4 (B.1.1.529.4), and BA.5 (B.1.1.529.5). These subvariants have continued to show increased transmissibility and immune evasion compared to earlier Omicron strains [42]. More recently, two Omicron subvariants called BA.2.86 [43] and JN.1 [44] have become dominant in many regions. These subvariants are believed to be even more transmissible and better able to evade immunity than previous Omicron strains. The rapid emergence and spread of Omicron and its subvariants have been a major challenge in the COVID-19 pandemic. While Omicron may cause less severe disease on average, its high transmissibility and immune evasion have led to surges of cases globally.

Research indicates that B.1.1.529 is effectively neutralized by the third dose of the Pfizer/BioNTech and Moderna mRNA vaccines, whereas the first and second doses offer little to no neutralization [45]. Efficient neutralization is also achieved with heterologous boosters, such as CoronaVac/Sinovac with Pfizer/BioNTech [46] or Oxford-AstraZeneca with Pfizer/BioNTech [47].

In terms of its ability to infect host cells, B.1.1.529 outperforms the Wild Type (WT) and even the Delta variant *in vitro* [45] . In some areas, Omicron has a growth advantage over Delta, which has resulted in a notable rise in clinical cases [40].

Interestingly, compared to other variants, infections caused by the Omicron variant typically result in milder symptoms. Populations in nations with highly developed vaccination programs enjoy elevated levels of vaccine protection. Omicron's spread causes more cases but does not always result in higher death rates in areas with low vaccination rates, suggesting lower virulence. Omicron's diminished ability to replicate is ascribed to its ineffective cleavage by the host protease TMPRSS2, a consequence of Spike gene mutations. Because of this, host proteases are unable to recognize it [48].

1.2.1.5 Insights from Molecular Dynamics Simulations on the Impact of Mutations Present in Variants

Molecular Dynamics (MD) simulations have provided valuable insights into the conformational dynamics of the RBD across various SARS-CoV-2 variants. MD simulations of the RBD from the Alpha variant revealed a shift in the open/closed equilibrium towards more open conformations by approximately 20%. This shift is driven by interactions such as the triple π-stacking between residues Y489–F456–Y473 and the hydrogen bond between Y489 and Y473, which stabilize the open conformation. In contrast, the closed conformation is supported by hydrophobic interactions involving residues Y501, V483, and F486, which prevent the formation of the E484–R403 salt bridge [49]. Han *et al.* [50] further elucidated that the Y501 mutation significantly enhances the Alpha variant's binding affinity to ACE2, leading to increased transmissibility. Their MD simulations revealed that this mutation induces increased flexibility in the Spike protein's RBD, facilitating a more efficient interaction with the host receptor, which is crucial for the variant's enhanced infectivity.

The Beta variant demonstrates a similar 20% shift towards open conformations, with stabilization mechanisms akin to those observed in the Alpha variant, including the π-stacking interactions and hydrogen bonds that support the open conformation. However, the E484K mutation disrupts the E484–R403 salt bridge, essential for maintaining the closed conformation. Despite this disruption, the Beta variant retains the ability to adopt a closed conformation due to compensatory effects from the N501Y mutation, which enhances other stabilizing interactions [49]. Gobeil *et al.* [51] provided further insights into these dynamics, demonstrating that the E484K mutation alters the electrostatic properties of the RBD, impacting its interaction with antibodies and contributing to immune escape. Their MD simulations revealed that the Beta variant's RBD exhibits similar stabilization mechanisms in the open conformation as observed in Alpha, but the E484K mutation leads to more substantial changes in the electrostatic surface, affecting antibody binding and immune response.

For the Delta variant, MD simulations show the presence of two Receptor Binding Motif (RBM) conformations, with one corresponding to the open conformation. Unlike other variants, however, the Delta RBD does not exhibit a traditional closed conformation. Instead, it features an alternative open conformation termed "reversed," which highlights the RBM region's flexibility. This reversed state acts as a two-way hinge, shifting to the side of the RBD, potentially offering advantages over the wild-type open state by concealing the RBM ridge region from antibody recognition while maintaining an open ACE2 binding surface for infection [49]. Socher *et al.* [52] expanded on these findings, showing that the Delta variant, particularly with the D614G mutation, exhibits increased flexibility and altered interaction dynamics with ACE2. Their simulations demonstrated that the D614G mutation enhances the Spike protein's flexibility, leading to more robust binding to ACE2 and contributing to the Delta variant's increased transmissibility and partial resistance to

neutralizing antibodies.

In the case of the Omicron variant, MD simulations reveal a predominant open conformation, with a notable 50% shift in the open/closed equilibrium towards more open states. Stabilizing interactions such as triple π-stacking and hydrogen bonds are present in the Omicron variant, similar to those in Alpha, Beta, and Delta variants. However, the closed conformation observed in other variants is notably distinct in Omicron [49]. Socher *et al.* [52] provided additional insights, highlighting that mutations such as S371L, S373P, and S375F in Omicron significantly increase RBD flexibility. This increased flexibility impacts ACE2 binding, enhancing the variant's ability to evade neutralizing antibodies and contributing to its high transmissibility. Their MD simulations also revealed that Omicron exhibits greater conformational heterogeneity compared to earlier variants, underscoring its enhanced ability to adapt and evade immune detection.

1.2.2 Viral Surveillance

To address the challenge of keeping track of emerging variants, organizations like the European Centre for Disease prevention and Control (ECDC) categorize and monitor these variants based on their specific characteristics [53]. This categorization allows for a systematic approach to tracking the evolution of the virus and identifying those variants that may pose particular risks. The categories established by the ECDC typically encompass various aspects, such as the variant's potential to spread more easily, its resistance to existing immunity, and its impact on disease severity. By categorizing these variants, health authorities can prioritize resources and responses accordingly, whether it involves modifying vaccination strategies, implementing public health measures, or conducting further research to develop countermeasures.

This ongoing monitoring and categorization process is critical to adapt and refine public health strategies in response to the virus's ever-evolving nature. It also plays a pivotal role in ensuring that vaccines and treatments remain effective against emerging variants, thereby contributing to the global effort to control and ultimately end the COVID-19 pandemic. The collaborative efforts of international organizations, researchers, and healthcare professionals are essential in this endeavor to stay one step ahead of the virus's mutations and to safeguard public health.

1.3 Approaches to Combat Viruses

The ongoing battle against viral infections has driven researchers and scientists to explore various strategies to combat these pathogens effectively. From the development of vaccines to the exploitation of small molecules and biologics, the scientific community has made significant strides in understanding and combating viral infections. This section will delve into the different approaches used to combat viruses, highlighting their potential and limitations.

1.3.1 Vaccines

Vaccines are a key preventive approach to combat viral diseases. They work by exposing the immune system to a weakened or inactivated form of the virus or to a specific antigen, triggering the body to produce antibodies and memory cells that can quickly recognize and fight the real virus upon future exposure. Vaccines can provide long-lasting immunity and help control the spread of viral outbreaks. Developing effective vaccines, especially for novel viruses, is a critical priority in combating viral threats [54, 55].

There are several types of vaccine technologies, including traditional approaches like live-attenuated, inactivated, and subunit vaccines, as well as modern platforms such as viral vector, mRNA, and DNA vaccines. Live-attenuated vaccines use weakened, non-virulent strains of the virus to induce an immune response, while inactivated vaccines use killed or inactivated viral particles. Subunit vaccines use purified viral proteins or peptides to elicit an immune response. Viral vector vaccines use a harmless virus to deliver viral genes and trigger an immune response, while mRNA vaccines use synthetic messenger RNA to instruct cells to produce viral proteins and stimulate immunity [56].

Vaccines have been successful in eradicating diseases like smallpox and significantly reducing the incidence of diseases like polio, measles, and influenza. However, the development of effective vaccines can be challenging, especially for rapidly mutating viruses like influenza and HIV [54].

1.3.2 Small Molecules

Small molecule drugs target specific steps in the viral lifecycle to disrupt infection and replication. Key mechanisms of action include entry inhibitors that bind to viral surface proteins or interfere with host cell receptors to block attachment and entry into host cells, replication inhibitors that inhibit viral enzymes like polymerases, proteases, and helicases essential for replication or disrupt viral genome synthesis, and assembly and release inhibitors that prevent proper assembly of viral particles or interfere with the release of newly formed virions from infected cells [57].

One example of a successful small molecule antiviral is remdesivir, which was developed by Gilead Sciences and approved for the treatment of COVID-19. Remdesivir is a nucleotide analog that inhibits the viral RNA-dependent RNA polymerase, disrupting viral replication [57, 58].

Small molecules offer advantages such as ease of administration, potential for oral bioavailability, and the ability to target specific viral proteins. However, the development of small molecules can be time-consuming and costly, and they may face challenges such as drug resistance and off-target effects [57, 58].

1.3.3 Biologics

Biologics are a broader category of therapeutic products created through the manipulation of biological sources, rather than chemical synthesis. Biologics include vaccines, gene therapies, cell therapies, engineered tissues, hormones, antibodies and other protein-based pharmaceuticals (Table A.1). These molecules can be used to target specific aspects of viral infection or modulate the host's immune response [54].

Antibodies can be leveraged in various ways to combat viral infections. Therapeutic monoclonal antibodies are produced in the lab to target specific viral epitopes and can neutralize viruses, block entry, or recruit immune effector functions. They can be used to treat active infections or provide short-term prophylaxis. For example, the monoclonal antibody cocktail REGEN-COV, developed by Regeneron, has been authorized for emergency use in the treatment of mild-to-moderate COVID-19, palivizumab is used for Respiratory Syncytial Virus (RSV) and interferons for hepatitis C virus [54].

In addition to therapeutic monoclonal antibodies, passive immunization is another approach that can provide immediate, temporary protection against viral infections. This involves administering antibodies derived from the plasma of individuals who have recovered from a viral infection (convalescent sera) or from hyperimmunized donors. This passive transfer of antibodies can be useful for protecting high-risk populations or providing a stopgap measure until vaccines become available [59].

Researchers have also explored ways to engineer antibodies to enhance their therapeutic properties. Techniques like affinity maturation can improve the binding affinity of antibodies to their viral targets, increasing their neutralizing potency. Bispecific antibodies, which can bind to two different epitopes simultaneously, can broaden the range of viruses that can be neutralized. Additionally, antibody-drug conjugates combine antibodies with cytotoxic payloads, allowing for targeted delivery of potent antiviral agents [59].

Nanobodies, a unique subclass of antibodies derived from camelids, have also gained attention as promising antiviral agents. These single-domain antibodies are smaller and more stable than conventional antibodies, allowing them to bind to viral epitopes that may be inaccessible to larger molecules. Their ease of production and potential for high specificity make them a valuable tool in antiviral therapy. For instance, nanobodies targeting the spike protein of SARS-CoV-2 have demonstrated potent neutralizing activity and are being explored for therapeutic use against COVID-19 [60].

Despite the challenges, the biopharmaceutical industry has experienced rapid growth in recent years, driven by the profound impact of biologics in treating a wide range of diseases. As our understanding of biology and biotechnology continues to advance, the future of biologics, including antibody-based approaches, holds great promise for innovative and effective therapies [54].

1.3.4 Monobodies

Monobodies are a subgroup of biologics that represent a class of engineered binding proteins which have gained prominence in various biomedical applications owing to their unique properties and versatile functionalities. Derived from the tenth human Fibronectin type III domain (FN3) (Figure 1.5), monobodies exhibit robust stability, solubility, target-binding specificity, and the possibility of being expressed in bacteria, making them ideal candidates for diverse research and therapeutic endeavors. The compact and modular structure of monobodies, consisting of a 7 β-sandwich fold stabilized by disulfide bonds and 6 loops of which 2 are the main responsible for the binding affinity (loop between β-strands B and C (BC loop) and the loop between β-strands F and G (FG loop)), enables facile engineering for enhanced affinity and specificity toward a wide range of target molecules. Moreover, their small size (~10 kDa) and monomeric nature render monobodies advantageous over conventional antibodies, offering improved tissue penetration, reduced immunogenicity, and enhanced stability in harsh physiological conditions. As such, monobodies have emerged as invaluable tools for probing Protein-Protein Interactions (PPI), deciphering complex biological pathways, and accelerating drug discovery efforts [61, 62].

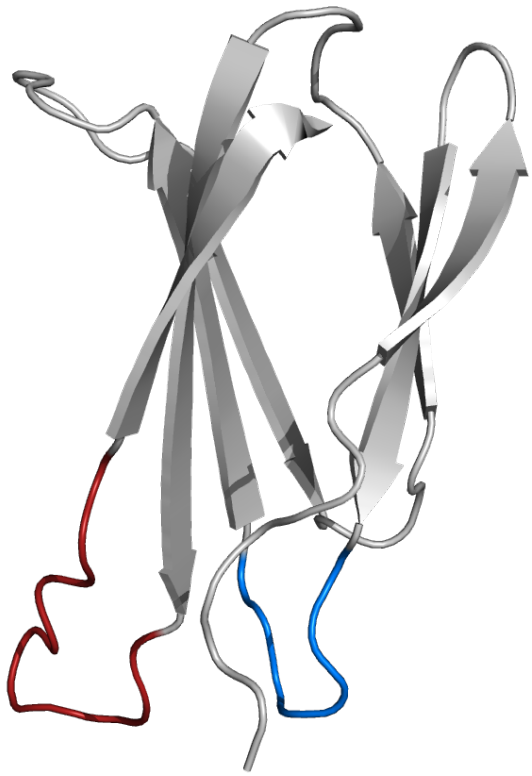


Figure 1.5: Tenth domain of human FN3, PDB ID: 1TTG. In blue is represented the BC loop; in red is represented the FG loop.

The development of monobodies typically involves rational design strategies or directed evolution approaches, aimed at optimizing their binding affinity, specificity, and biophysical properties. Rational design relies on computational modeling and structure-guided engineering to predict and enhance monobody-target interactions through site-directed mutagenesis or grafting of target-binding loops. Conversely, directed evolution harnesses the power of genetic diversity and iterative selection processes to evolve monobodies with desired binding properties from diverse combinatorial libraries. Recent advancements in both approaches have facilitated the generation of high-affinity monobodies targeting various antigens, including proteins, peptides, small molecules, and even post-translational modifications. Furthermore, the advent of next-generation sequencing and bioinformatics tools has expedited the characterization and optimization of monobody libraries, enabling rapid screening and identification of lead candidates for downstream applications [61, 62].

In addition to their utility as research reagents, monobodies hold great promise as therapeutic agents for the treatment of various diseases, including cancer, autoimmune disorders, and infectious diseases. Their ability to target specific disease-associated proteins with high affinity and selectivity makes them attractive candidates for precision medicine and targeted drug delivery. Moreover, the modular nature of monobodies facilitates the fusion with effector moieties, such as toxins, cytokines, or imaging probes, to impart additional functionalities and therapeutic modalities. With ongoing advancements in monobody engineering and characterization, coupled with the increasing understanding of their structural dynamics and biological mechanisms, the prospects for exploiting monobodies in biomedical research and clinical applications continue to expand, heralding a new era of precision therapeutics and molecular diagnostics [61, 62].

1.4 How Can We Improve Pandemic Preparedness?

If we look back, we realize that in a short space of time a single species of virus, SARS-CoV-2, was able to diversify to such an extent that, five years on, we are faced with strains that have completely distinct characteristics to those that were initially identified in 2019. These characteristics include the ability to evade the immune system, infectivity, fusogenicity, among others, which play a key role in the severity of the disease with which the virus is associated. However, just as COVID-19 appeared overnight, a different virus could give rise to a new pandemic with completely distinct characteristics to those we have been familiar with for the last few years.

Given that it is not possible to predict precisely which virus or viruses will cause the next pandemic, it is essential to develop new methodologies to combat a wide variety of viruses, allowing us to be prepared for any new pandemic that may arise. In other words, how can we develop a strategy that can be applied to different viruses or different variants of a virus? Because they can adapt to various targets, monoclonal antibodies have this potential. However, since their discovery, development and manufacturing

processes are complex, time-consuming, and expensive, their clinical application is limited. Alternatively, engineered protein scaffolds have recently emerged as a viable option. Since they are smaller than monoclonal antibodies and can be expressed in bacteria, their application to this problem becomes more advantageous. This adaptability, along with their potential for oral administration and reduced immunogenicity, makes engineered protein scaffolds an exciting avenue for developing innovative and effective treatments for viral diseases.

1.4.1 EvaMobs: Research for Antiviral Biopharmaceuticals

In the face of globalization and climate change, the world is increasingly vulnerable to new viral outbreaks. To prevent future pandemics, rapid and effective antiviral development is crucial. EvaMobs [63], an innovative project led by the Instituto de Tecnologia Química e Biológica António Xavier (ITQB NOVA) in Portugal, is spearheading a novel approach to antiviral therapy. By utilizing evolvable monobodies, or "Mobs," EvaMobs aims to create a flexible, efficient platform for the discovery and production of new antiviral agents.

As explained before, Mobs are small proteins designed to have a high affinity for various viruses. Unlike traditional antiviral methods that rely on monoclonal antibodies or designed miniproteins, Mobs offer several advantages which allow for the rapid generation of specific molecules capable of neutralizing particular viruses.

The EvaMobs project is structured into four main work packages (WPs), each focusing on a critical aspect of antiviral development [64].

1.4.1.1 WP1: Discovery of New Mobs

In this phase, EvaMobs employs artificial intelligence and physics-based computational biology methods to develop a discovery framework for creating Mobs. This framework will be tested against four target viruses: SARS-CoV-2, RSV, Influenza A virus, and Zika virus. By generating and mutating Mob sequences, predicting their structures and interactions with viral components, and screening the best candidates, EvaMobs aims to continuously refine its computational methods. This will enhance the framework's ability to design Mobs that effectively bind to and neutralize these viruses, making them suitable for drug development [64].

1.4.1.2 WP2: Production of Specific Mobs

EvaMobs implements a production pipeline to evaluate the binding activity, thermal stability, and aggregation propensity of Mobs. The best-performing Mobs will be optimized for large-scale production and formulated for preclinical and clinical trials. The final Mob candidates will be produced under Good Manufacturing Practices conditions, preparing them for Phase I clinical trials [64].

1.4.1.3 WP3: Preclinical Validation

In this stage, at least 25 Mobs will be tested for antiviral effectiveness against the four target viruses in cellular models. The most promising candidates will undergo further preclinical evaluation in animal models to assess their impact on virus replication, clinical outcomes, and immune responses. These studies will provide comprehensive data on safety and effective dosage levels, culminating in a thorough preclinical report [64].

1.4.1.4 WP4: Phase I Clinical Validation

The final phase involves a Phase I clinical trial to evaluate the safety of a lead antiviral Mob in humans. This trial will be conducted following International Council for Harmonisation guidelines, with preparation and submission to regulatory authorities. Upon approval, the trial will commence, and data will be collected and analyzed to produce a comprehensive Clinical Study Report, providing the final efficacy assessment of the Mob.

ITQB NOVA collaborates with several European research institutions and biotechnology companies, leveraging their expertise in computational biology, protein engineering, and antiviral drug development [64].

1.4.1.5 EvaMobs Scope

In essence, by creating broad-spectrum antivirals that can target a wide range of viruses, EvaMobs aims to shorten the development timeline for effective treatments. This project not only addresses current viral threats but also builds a flexible platform for pandemic preparedness, ensuring a rapid response to future outbreaks. Through its innovative approach and international collaboration, EvaMobs is poised to make a significant contribution to antiviral research and global health security.

This thesis work is part of WP1: Discovery of New Mobs. The scope of this work includes the study of the application of languages models to generate new Mobs, as described in the next sections.

1.4.2 Strategies for the Generation of Target-specific Biopharmaceuticals

Pandemic preparedness is a very complex challenge, especially because one cannot be certain about when and where the next pandemic will arise and which pathogen will be responsible for the disease in question. Whenever a new virus appears, a suitable viral target that is essential for viral replication or entry should be identified, studied and targeted, and there are two main strategies that can be used to computationally design target-specific antivirals.

1.4.2.1 *Ab initio* Binder Design

Ab initio design strategies involve the computational creation of compounds that bind to specific targets from the ground up. This method relies heavily on advanced

computational tools, such as RFdiffusion [65], Chroma [66], EvoDiff [67] and Genie 2 [68], which have demonstrated efficacy in generating target-specific antivirals by predicting how compounds will interact with viral targets. These tools utilize sophisticated algorithms to model molecular interactions and can generate novel compounds that are optimized for binding to specific viral proteins [69].

Moreover, the success of *ab initio* designs is contingent upon the accuracy of the computational models used. Recent advances in machine learning and artificial intelligence have significantly improved the predictive capabilities of these models, allowing for a more accurate prediction of molecular interactions. For instance, the integration of deep learning techniques has enabled the identification of potential binding sites on viral proteins with greater accuracy, thereby enhancing the feasibility of *ab initio* designs [70, 71]. As these technologies continue to evolve, they hold the promise of accelerating the development of novel antivirals, particularly in response to emerging pathogens [72].

Another important consideration in *ab initio* design is the need for iterative refinement. Once initial compounds are designed, they must undergo rigorous testing to assess their binding affinity and specificity. This often involves high-throughput screening methods and structural biology techniques, such as X-ray crystallography or cryo-electron microscopy, to visualize the interactions at an atomic level [73], which allows for the improvement of the designed compounds, ultimately leading to the identification of highly effective antiviral candidates. However, a critical limitation of the *ab initio* design approach is the prerequisite knowledge of the binding sites on the target protein. Understanding the epitope region is essential before any design can be effectively initiated. This necessity for prior knowledge can slow down the drug development process, as extensive preliminary studies are required to identify these critical regions. Furthermore, if the target undergoes mutations or if novel viral strains emerge, the initial designs may become obsolete, necessitating a new round of discovery and design. This limitation underscores the potential advantages of alternative strategies that can bypass the need for extensive prior knowledge of the target's binding sites.

1.4.2.2 Library "Fishing"

In contrast, the library "fishing" strategy circumvents the need for prior knowledge of epitope regions by utilizing a vast library of antiviral compounds that exhibit significant sequence, structure, and conformational diversity. This approach can be likened to 'fishing' for the optimal antiviral within a "lake" of potential candidates, using the viral target as "bait". By screening this extensive library, one can identify the most effective antiviral candidates that bind to the target, allowing for a more flexible and rapid response to emerging viral threats. Once promising candidates are identified, further studies can elucidate the nature of the interactions between the antiviral compounds and their targets, paving the way for iterative redesign and optimization of these therapeutics.

The advantages of library "fishing" extend beyond just the identification of effective

binders; it also fosters the discovery of novel mechanisms of action. By exploring diverse compounds, researchers can uncover unique binding interactions that may not have been anticipated through traditional design approaches. This can lead to the development of antivirals that target previously unexploited viral pathways, thus broadening the therapeutic arsenal available for combating viral infections. Furthermore, the use of high-throughput screening technologies has made it feasible to evaluate millions of compounds in a relatively short timeframe, significantly expediting the drug discovery process.

Additionally, the library "fishing" strategy can be enhanced through the incorporation of advanced computational techniques, such as virtual screening and molecular docking simulations [74]. These tools allow researchers to prioritize compounds based on predicted binding affinities before proceeding to experimental validation, thereby streamlining the drug discovery pipeline. The combination of high-throughput screening with computational methods not only increases the efficiency of the discovery process but also improves the likelihood of identifying high-quality antiviral candidates that can be rapidly developed into therapeutics.

1.4.2.3 Monobody Library Development for Pandemic Preparedness

In this sense, with the intent to increase our pandemic preparedness and streamline the development of target specific antivirals, the main aim of this project is to develop a monobody library that shows both sequence and structure diversity, allowing it to be applicable to finding the best monobody to target any given viral protein using a "fishing" strategy. However, if one compiles the monobodies already described in the literature, this library's size would be small and it would be rather poor, since that there are few monobodies structures described in the literature - and the ones that are described show low diversity. Therefore, there is the need to enrich that library.

In this thesis, we pose the hypothesis that language models can be used to create a large and diverse monobody library, that is adequate to serve as a starting point for design approaches.

1.5 Objective

Given the fact that there are too few monobodies described in the literature, the primary aim of this project is to create a computational framework that streamlines the development of these Antibody-like Engineered Protein Scaffolds (ALEPS) through the creation of a monobody library, by synergizing computational structural biology with protein design tools, software suites and machine learning techniques. The project will explore the application of docking studies to facilitate the efficient optimization of these molecules, ensuring their precise binding to specific targets. To validate the feasibility of this approach, the project will focus on the design of ALEPS tailored to bind and neutralize the SARS-CoV-2. Ultimately, the designed ALEPS will be produced and

rigorously validated by collaborative partners, marking a significant step forward in the development of innovative therapeutic agents.

2

Theory and Methods

2.1 Computational Structural Biology

Computational structural biology is a rapidly advancing field that leverages powerful computational techniques to study the structure, dynamics, and function of biological macromolecules, particularly proteins. These computational approaches, combined with experimental data, provide valuable insights into the complex behavior of biomolecules and their interactions at the atomic level. One of the most widely used computational methods in structural biology is molecular dynamics simulation, which offers a unique window into the dynamic behavior of proteins and other biological systems.

2.1.1 Molecular Dynamics Simulations

Molecular Dynamics (MD) simulation is a powerful computational technique indispensable in molecular modeling and simulation. It offers a unique window into the dynamic behavior of atoms and molecules, allowing the investigation of a wide range of physical, chemical, and biological phenomena at the atomic and molecular scale [75]. At its core, MD simulation involves solving classical equations of motion to predict the trajectories of individual particles within a system [76]. Over the past few decades, MD has evolved from a niche method in theoretical chemistry and condensed matter physics to a widely adopted and highly versatile tool with applications spanning various scientific disciplines, including chemistry, physics, materials science, biology, and engineering.

To accurately simulate molecular systems, certain approximations are necessary to treat atoms as classical particles. One fundamental approximation is the Born-Oppenheimer approximation, which separates the motion of nuclei and electrons in a molecule. The approximation is based on the assumption that the mass of nuclei is much larger than that of electrons, allowing the separation of the electronic and nuclear wavefunctions. The total wavefunction can be written as:

$$\Psi(R, r) = \psi_{\text{nuc}}(R) \cdot \psi_{\text{elec}}(r; R) \tag{2.1}$$

Here, $\Psi(R, r)$ is the total wavefunction of the system, R represents the nuclear coordinates, and r represents the electronic coordinates. The nuclear wavefunction $\psi_{\text{nuc}}(R)$ and the electronic wavefunction $\psi_{\text{elec}}(r; R)$ are treated separately, with the electronic wavefunction depending parametrically on the nuclear positions. This approximation simplifies the problem by allowing the electrons to be considered in their ground state for any given nuclear configuration, enabling the simulation to focus on classical trajectories of the nuclei.

The behavior of the particles is governed by Newton's equations of motion, which describe how the position and velocity of each particle in a system evolve over time under the influence of forces. These equations relate the force F_i acting on a particle to its acceleration a_i times its mass m_i, forming the basis for predicting the trajectories of particles:

$$F_i = m_i a_i \tag{2.2}$$

where the acceleration a_i is the second derivative of the particle's position r_i with respect to time:

$$a_i = \frac{d^2 r_i}{dt^2} \tag{2.3}$$

The force on a particle is typically derived from a potential energy function that characterizes the interactions between particles in the system. Specifically, the force on each particle is calculated as the negative gradient of the potential energy with respect to the particle's position:

$$F_i = -\nabla_{r_i} V(r_1, r_2, \ldots, r_N) \tag{2.4}$$

This potential energy function typically includes several components that account for both bonded and non-bonded interactions. The bonded interactions include terms for bond stretching V_{bond}, angle bending V_{angle}, and dihedral (torsional) interactions V_{dihedral}. The non-bonded interactions consist of van der Waals forces V_{vdW} and electrostatic interactions V_{coulomb}. The overall potential energy function can be expressed as:

$$V = \sum_{\text{bonds}} V_{\text{bond}} + \sum_{\text{angles}} V_{\text{angle}} + \sum_{\text{dihedrals}} V_{\text{dihedral}} + \sum_{i<j} (V_{\text{vdW}} + V_{\text{coulomb}}) \tag{2.5}$$

By solving the equations of motion numerically, MD simulations can track the dynamic behavior of particles over time, providing insights into the molecular mechanisms underlying a wide range of physical, chemical, and biological processes [76].

Next I will delve into four key ideas that are fundamental to better understand MD simulations: Periodic Boundary Conditions (PBC), Energy Minimization, Integration Algorithms and Analysis (for a more complete discussion see reference [76]).

2.1.1.1 Periodic Boundary Conditions

In MD simulations, a central challenge is effectively addressing system boundaries within a finite space. Achieving accurate results for complex biological systems necessitates striking a balance between using a sufficiently large simulation box to minimize collisions with box walls and the computational expense associated with this approach. An alternative is employing a smaller box, concentrating particles near the edge, and introducing different forces on these particles compared to those at the center.

A third and more widely used approach to mitigate boundary effects and efficiently simulate an infinite system in MD simulations is to adopt PBC. PBC involves replicating the simulation box throughout space, forming an infinite lattice, ensuring that surrounding atoms influence each atom in the central cell without boundary effects. Despite the absence of physical replicate boxes, mathematical operations simulate periodicity. This periodicity enables particles exiting one side of the simulation cell to seamlessly re-enter from the opposite side, providing a conducive environment for particle movement (Figure 2.1). The setup of the simulation box plays a crucial role in the simulation and it is important to have a setup that recreates a realistic bulk environment, frequently incorporating solvent molecules like water and other important components, such as salt, to mimic real experimental conditions.

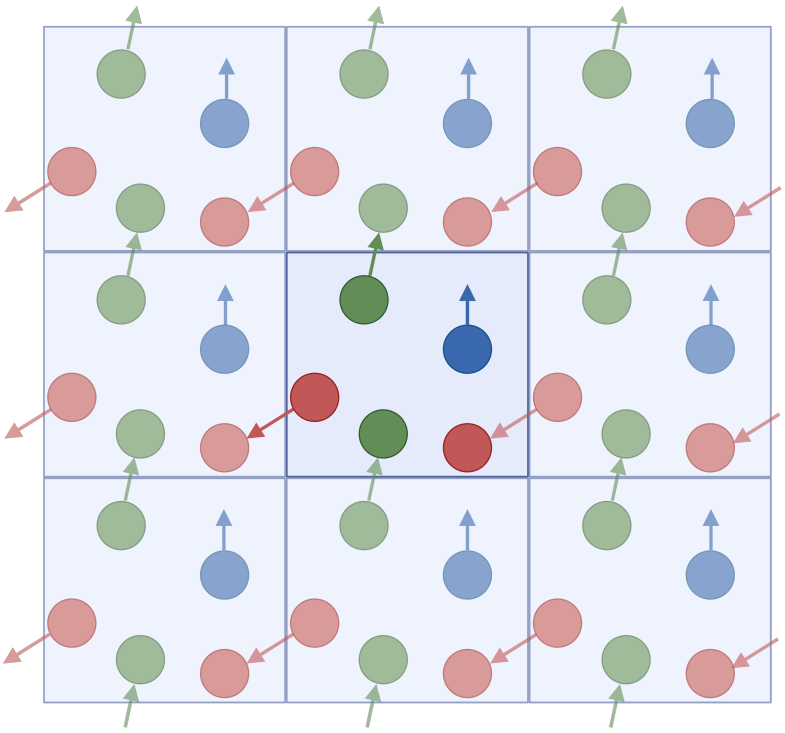


Figure 2.1: Representation of periodic boundary conditions in 2D. Whenever an entity exits from the top of its box, there is an image of that entity that enters the box from the bottom, keeping the number of entities in the box constant. For 3D, imagine the existence of one copy of this image on each side along an axis that is perpendicular to the represented 2D surface and the possibility of entities entering or exiting their box in the direction of that axis.

Various simulation box shapes compatible with PBC exist, with the choice dependent on the system's geometry. While the cubic box is a simple and common choice, it may not be ideal for approximately spherical molecules. Alternatives like the truncated octahedron and rhombic dodecahedron, approximating spherical shapes, prove more suitable for simulations involving spherical molecules, requiring fewer solvent molecules and accounting for protein rotation, which can lead to the appearance of periodic images, complicating the simulation. In this dissertation, all simulations use a rhombic dodecahedron-shaped box.

When implementing PBC, a "non-bonded cutoff" is applied. The cutoff ensures that interactions between distant particles are ignored, significantly reducing the number of calculations and therefore reducing computational costs. However, it is important to ensure that the cutoff distance is carefully chosen so that interactions between a particle and its periodic images remain negligible. The cutoff must also adhere to the "minimum image convention", meaning it cannot exceed half the length of the shortest box vector, thereby maintaining accurate and efficient calculations within the simulation.

2.1.1.2 Energy Minimization

Energy minimization is a crucial preparatory step in MD simulations, ensuring that the system starts from a stable and realistic configuration. Since the system is constructed as a box with PBC filled with water molecules and ions, inside of which our molecule is set, the primary objective of energy minimization is to optimize the positions of atoms within the system, thereby reducing steric clashes, unfavorable interactions, and overall potential energy. This process brings the system to a local energy minimum, which is essential for achieving a stable starting point for subsequent simulations (Figure 2.2).

The potential energy landscape of a molecular system is typically complex, with many local minima corresponding to different conformations. During energy minimization, the goal is to find a nearby local minimum rather than the global minimum, which would require exploring the entire energy landscape. By optimizing the atomic positions, energy minimization helps to alleviate unrealistic geometries that may arise from initial model building or from assembling the system. These unrealistic geometries could lead to significant forces that would cause instability if not corrected before the MD simulation.

Energy minimization is achieved through the application of numerical minimization algorithms, such as the steepest descent method or the conjugate gradient method. These algorithms iteratively adjust the atomic coordinates based on the forces acting on each atom, gradually reducing the system's potential energy.

The steepest descent method is one of the simplest and most commonly used minimization techniques. It works by moving the atoms in the direction of the negative gradient of the potential energy function, which is the direction in which the energy decreases most rapidly. Mathematically, the update for the position of each atom r_i at each iteration t is given by:

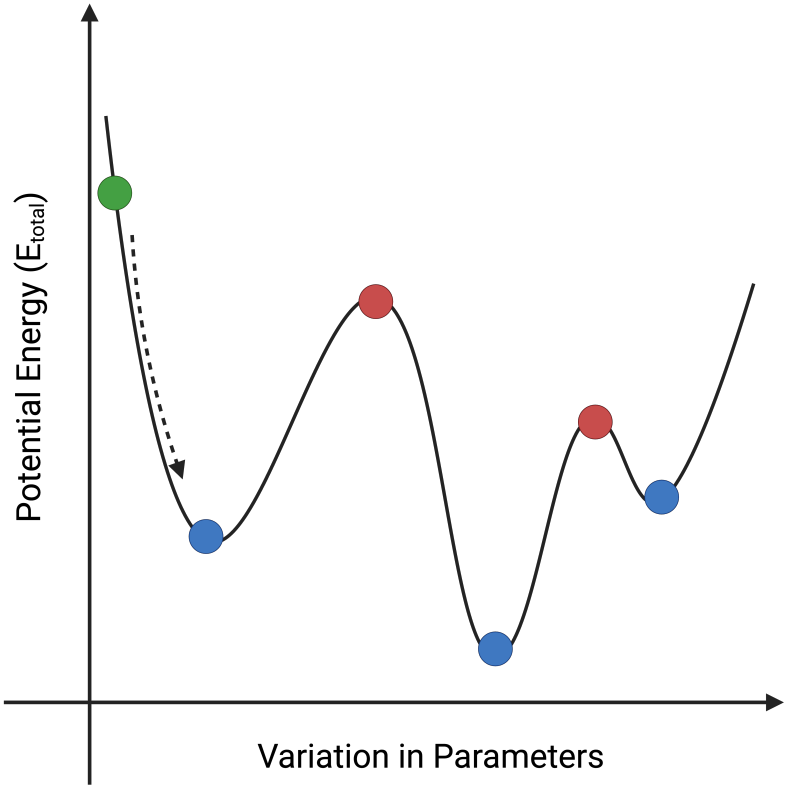


Figure 2.2: Representation of the different phases of an atomic structure during minimization of its energy. By optimizing the positions of atoms through variation of parameters, the system gets closer to an energy minimum (**blue**). In some energy landscapes there may exist energy barriers (**red**) between the local and global energy minima, making it difficult to always achieve the global minimum. Therefore, the landed minimum depends on the position of the starting conformation in the landscape (**green**).

$$r_i(t+1) = r_i(t) - \alpha \nabla E(r_i(t)) \tag{2.6}$$

where $\nabla E(r_i(t))$ is the gradient of the potential energy function with respect to the position of atom i, and α is a step size parameter that controls the magnitude of the displacement. The step size α must be chosen carefully; if it is too large, the method may overshoot the minimum, leading to instability, while if it is too small, convergence may be slow.

The steepest descent method is straightforward and robust, making it particularly useful for initial energy minimization when the system may be far from equilibrium. However, it can become inefficient as the system approaches the minimum because the energy landscape typically flattens out near the minimum, causing the gradient to decrease and the steps to become smaller.

An alternative to the steepest descent method is the conjugate gradient method, which is more sophisticated and generally converges more quickly, especially in the vicinity of the minimum. The conjugate gradient method not only considers the current gradient but also takes into account the history of previous steps to determine the search direction, allowing it to avoid the zigzagging behavior that can occur in the steepest descent method.

The number of minimization steps required to reach a satisfactory configuration depends on the initial state of the system and the desired level of convergence. Convergence is typically monitored by observing the changes in potential energy or the magnitude of the forces acting on the atoms. A well-minimized structure is essential as it serves as a reliable and stable starting point for the subsequent MD simulation where the system will advance through time, reducing the likelihood of instabilities or unphysical behavior during the production run.

2.1.1.3 Integration Algorithms

In MD simulations, numerical integration algorithms are crucial for advancing the system through time by calculating updated positions and velocities of particles at discrete time steps. Among the most widely used integration methods are the Verlet and Leapfrog algorithms, which approximate solutions to Newton's equations of motion. The Verlet algorithm, known for its simplicity and energy conservation over long simulations, updates particle positions with the formula:

$$r(t + \Delta t) = 2r(t) - r(t - \Delta t) + \Delta t^2 \frac{F(t)}{m} \tag{2.7}$$

Here, $r(t)$ is the position at time t, Δt is the time step, $F(t)$ is the force, and m is the particle's mass. This method relies on past and current positions and forces to predict future positions. The Leapfrog algorithm, a variant that updates positions and velocities in a staggered fashion, is given by:

$$v\left(t + \frac{\Delta t}{2}\right) = v\left(t - \frac{\Delta t}{2}\right) + \Delta t \frac{F(t)}{m} \tag{2.8}$$

$$r(t + \Delta t) = r(t) + \Delta t v\left(t + \frac{\Delta t}{2}\right) \tag{2.9}$$

In this scheme, velocities are updated at half-time steps, which enhances stability and accuracy, particularly useful when precise control of kinetic energy is needed.

Thermodynamic control within MD simulations is managed through various algorithms for temperature and pressure regulation. By regulating and keeping these variables constant, one is able to reproduce more realistic experimental data obtained in the same temperature and pressure conditions. An example of an algorithm used for temperature regulation is the Berendsen thermostat, which adjusts particle velocities to control temperature with the formula:

$$v_i(t + \Delta t) = v_i(t) \times \sqrt{1 + \frac{\Delta t}{\tau}\left(\frac{T_0}{T(t)} - 1\right)} \tag{2.10}$$

where $v_i(t)$ is the velocity of particle i, T_0 is the target temperature, $T(t)$ is the instantaneous temperature, and τ is the relaxation time. While effective for smooth temperature

adjustments, the Berendsen thermostat does not fully sample the canonical ensemble, which can limit its applicability for precise thermodynamic studies.

An alternative approach for temperature regulation is the V-rescale thermostat, which is a stochastic velocity rescaling algorithm. It corrects the canonical ensemble sampling limitations of the Berendsen thermostat by introducing stochastic noise into the temperature coupling, ensuring proper sampling of the canonical ensemble. The temperature control is applied by scaling the velocities using the following equation:

$$v_i(t + \Delta t) = v_i(t) \times \sqrt{1 + \frac{\Delta t}{\tau}\left(\frac{T_0}{T(t)} - 1\right) + \frac{2\Delta t}{\tau}\frac{k_B T_0}{m_i} W_i} \tag{2.11}$$

where k_B is the Boltzmann constant, m_i is the mass of particle i, and W_i is a Gaussian random number with a mean of 0 and a variance of 1. The added stochastic term allows the system to fluctuate around the target temperature while maintaining correct ensemble properties.

Pressure control can be achieved using a barostat like the Parrinello-Rahman barostat, which maintains a desired pressure by dynamically adjusting the simulation box dimensions. The equations governing this barostat are:

$$G = hh^T \tag{2.12}$$

$$\frac{d^2G}{dt^2} = \frac{1}{W}\left(P - P_{\text{ext}}\right) G \tag{2.13}$$

Here, G is the matrix of box vectors, h is the box matrix, W is a mass-like parameter, P is the internal pressure tensor, and P_{ext} is the external pressure tensor. By adjusting box dimensions to match the target pressure, the Parrinello-Rahman barostat enables accurate simulations of systems under varying pressure conditions, including those with anisotropic stress. Properly tuning these algorithms ensures stability and accuracy, making them essential for reliable molecular dynamics simulations.

2.1.1.4 Analysis

Once the MD simulation is complete, a comprehensive analysis of the trajectory is performed to extract valuable insights about the system's behavior. This analysis involves examining various properties, such as the system's energy, temperature, pressure, and structural changes over time. These parameters help in understanding how the system evolves, providing a window into the thermodynamic stability and the dynamics of molecular interactions within the simulated environment.

One of the key steps in this analysis is the calculation of the Root Mean Square Deviation (RMSD), which measures the average deviation of the atomic positions from a reference structure, typically the initial configuration. RMSD is a crucial indicator of the structural stability of the system, allowing for the identification of conformational changes, equilibration times, and overall system stability. The RMSD is defined as:

$$\text{RMSD}(t) = \sqrt{\frac{1}{N}\sum_{i=1}^{N}\left(r_i(t) - r_i^{\text{ref}}\right)^2} \tag{2.14}$$

where N is the number of atoms, $r_i(t)$ is the position of atom i at time t, and r_i^{ref} is the reference position of atom i. The RMSD provides a global measure of how much the structure deviates from the reference over time, making it a useful tool for assessing the structural drift or stability of the system.

In addition to RMSD, the Root Mean Square Fluctuation (RMSF) is calculated to evaluate the flexibility of individual residues or atoms over the course of the simulation. RMSF provides insights into which parts of the molecule are most dynamic, highlighting regions of flexibility or rigidity, which can be important for understanding the functional dynamics of the system. The RMSF is defined as:

$$\text{RMSF}(i) = \sqrt{\frac{1}{T}\sum_{t=1}^{T}\left(r_i(t) - \langle r_i \rangle\right)^2} \tag{2.15}$$

where T is the total number of time frames, $r_i(t)$ is the position of atom i at time t, and $\langle r_i \rangle$ is the time-averaged position of atom i. This equation quantifies the fluctuations of each atom or residue around its average position, providing a per-atom or per-residue measure of flexibility over the simulation time.

In addition to these standard analyses, Principal Component Analysis (PCA) is a powerful statistical tool used to reduce the dimensionality of the MD simulation data while preserving the most critical information about the system's motion. PCA identifies the principal components, or eigenvectors, which correspond to the directions of maximum variance in the atomic motions. By projecting the trajectory onto these principal components, it is possible to focus on the most significant collective motions within the system, such as large-scale conformational changes. PCA is particularly useful for identifying correlated motions and exploring the dynamic behavior of complex molecular systems beyond what can be observed from individual atomic fluctuations.

The process of performing PCA begins with the construction of a covariance matrix from the atomic coordinates, which captures the correlated motions of the atoms. The eigenvectors and eigenvalues of this matrix are then calculated, where the eigenvectors represent the principal components, and the eigenvalues correspond to the amount of variance explained by each component. By analyzing the first few principal components, which usually account for the majority of the variance, it is possible to visualize and interpret the most significant conformational changes in the system. PCA thus provides a deeper understanding of the underlying mechanisms driving the observed dynamics, making it an invaluable tool in the analysis of MD simulations.

Visualization tools and specialized software, such as PyMOL [77] and GROMACS [78], are often used to analyze and interpret the results of MD simulations. These tools provide functionalities for performing PCA, calculating RMSD and RMSF, and visualizing

structural changes, enabling one to gain comprehensive insights into the behavior and dynamics of molecular systems.

As an example of additional analyses that can be performed, the secondary structure elements of proteins, such as alpha-helices and beta-sheets, can be analyzed to track how these structures evolve over time. Although this was not conducted in this dissertation it is a common analysis in protein simulations to understand folding and stability, further extending the insights gained from MD simulations.

2.2 Molecular Dynamics Simulations of SARS-CoV-2 Variants

2.2.1 Simulation Setup

2.2.1.1 System Preparation

The initial structure of the WT RBD was sourced from PDB ID: 6M0J, representing an ACE2-bound conformation of RBD. Since we aimed to simulate the unbound form of the RBD, ACE2 was excluded from the coordinate file that was used as a starting point. For the study of the SARS-CoV-2 variants BA.2.86 and JN.1, the respective RBD structures were generated by mutating specific residues as described in the literature in the WT RBD using PyMOL; whereas for the study of the selected monobody:RBD complexes obtained in the computational design pipeline described below, the output structure resultant from the design pipeline was used as input for the MD simulations.

It is noteworthy to highlight the absence of glycosylations in our simulation systems. Although much of the Spike protein surface is densely glycosylated, the RBD itself contains only a single glycosylation site, located distantly from the RBM region, which had previously shown to be the most dynamic region in simulations performed by our lab. Notably, while literature suggests that neighboring glycosylation sites can shield the RBD, this glycan shield is contingent upon the RBD's conformational change in the complete Spike protein. Specifically, when the RBD is in the down state, the glycan shield conceals it, but when in the up state, the RBD emerges from the glycan shield, presenting a fully accessible RBM. Our study focuses on this up state, where the RBM is fully exposed, and glycans are no longer influential in the observed dynamics, potentially playing a role in modulating binding to ACE2.

Considering that the inclusion of glycans would introduce complexities in sampling, we adopted a reductionist approach by simulating the RBD without glycosylations. Simulations were performed in a water environment, with the RBD structure placed in a truncated dodecahedron box filled with water molecules (maintaining a minimum distance of 1.2 nm between the protein and box walls). The system's total charge was neutralized with Na^+ ions, and additional Na^+ and Cl^- ions were added to achieve an ionic strength of 0.1 M.

2.2.1.2 MD Runs

MD simulations were conducted using the GROMACS [78] 2021.7 package, employing the Amber14 forcefield and TIP3P water model for all atomistic simulations.

Prior to production runs, the system underwent energy minimization using the steepest descent method for a maximum of 50,000 steps, with position restraints on heteroatom positions based on crystallographic coordinates using a force constant of 1,000 kJ/mol in the X, Y and Z positions. A five-stage initialization process was then implemented over 500 ps for 5 different replicates for each variant.

In the first stage, the simulation was conducted over 100 ps and the system's temperature was coupled with the Berendsen algorithm with a reference temperature of 300 K and a temperature coupling constant of 0.01 ps. No pressure coupling was applied, and velocities were generated at 300 K.

For the second stage, which was extended for 100 ps, all simulation parameters, except the temperature coupling constant, remained unchanged. The temperature coupling utilized the Berendsen algorithm with a slightly higher temperature coupling constant of 0.1 ps . Like the first step, there was no pressure coupling, and velocity generation was omitted, since this block aims to start from the positions and velocities generated in the previous one.

In the third stage, the simulation continued for 100 ps. Temperature coupling was kept with the Berendsen algorithm at a constant of 0.1 ps. Pressure coupling was introduced using the Berendsen algorithm with a coupling constant of 5.0 ps. Like the previous steps, there was no velocity generation.

The fourth step also involved an extension for 100 ps. The temperature coupling method shifted to the v-rescale algorithm with a constant of 0.1 ps, and pressure coupling used the Parrinello-Rahman algorithm with a coupling constant of 5.0 ps. As in previous steps, no velocity generation occurred.

In the fifth and last stage, the simulation continued for 100 ps. Temperature coupling was kept with the v-rescale algorithm at a constant of 0.1 ps, and pressure coupling used the Parrinello-Rahman algorithm with a coupling constant of 5.0 ps. As before, there was no velocity generation.

For the SARS-CoV-2 variants, after the five-stage initialization process, simulations were conducted for 7 μs across the five replicates of each variant, with the first four microseconds considered as equilibration time. For the monobody:RBD complexes, after the five-stage initialization process, simulations were conducted for 1 μs across the 3 replicates of each variant. Analysis was performed on the frames after equilibration.

2.2.2 Analysis of the System Properties

The visualization and rendering of simulation snapshots were accomplished through the utilization of the PyMOL molecular graphic viewer [77]. Analyses of the simulations were conducted using GROMACS tools [78] and the MDAnalysis package [79]. The

creation of all plots was executed using the Matplotlib library [80] in in-house Python scripts; for a better plot visualization, a moving average window size of 100 ps and 5 ps was used for the smoothing of the RMSD plots of the SARS-CoV-2 RBD simulations and for the monobody:RBD complexes, respectively, whereas a moving average window size of 5 residues was used for the smoothing of the RMSF plots.

2.2.3 Analysis of the Conformational Dynamics

Principal Component Analysis (PCA) is a statistical technique used to reduce the dimensionality of data by identifying the principal components that capture the most significant collective motions within the system, such as large-scale conformational changes, in the context of MD simulations. Here, PCA was performed on the $(3N - 6)$-dimensional space of conformational coordinates derived from the SARS-CoV-2 RBD MD simulations, where N represents the number of RBD residues. Prior to the PCA, each conformation was adjusted for translation and rotation relative to the core Cα atoms of the RBD. Using MDAnalysis [79], principal components were extracted from the trajectories, focusing solely on the coordinates of the RBD Cα atoms. The first two principal components were chosen to project the RBD structures from each simulation frame into a two-dimensional space, facilitating a simplified representation of the RBD conformational space for both variants.

The two-dimensional density and energy landscape within the principal component space were computed. The probability density function for each trajectory projection was estimated using a Gaussian kernel density estimator from LandscapeTools' *get_density* program [81] (accessible at `https://www.itqb.unl.pt/labs/molecular-simulation/in-house-software`), employing a uniform grid with a mesh size of 0.5 Å.

To analyze the energy surface landscapes, energy minima and their corresponding basins were determined. A basin was defined as the collection of conformations for which the steepest descent path on the energy surface leads to a specific minimum [81]. Steepest descent paths for each grid cell were computed, with each conformation following the path of its corresponding grid cell. Landscape regions with energy values exceeding 6 k_BT were excluded, resulting in the final set of basins for each dataset.

2.3 Protein Design

Protein design is the process of engineering proteins with specific structures, functions, or properties by manipulating their amino acid sequences. The ultimate goal of this field is to create proteins with enhanced or entirely new functionalities, such as improved catalytic activity, stability, or binding affinity. Achieving this requires a deep understanding of the relationship between sequence and structure, as well as the ability to predict how changes in the sequence will influence the protein's behavior. To tackle the complexity of this task, one can utilize a range of computational methods, including physics-based

energy functions, statistical potentials, machine learning techniques or a combination of the different types of approaches. These tools enable the efficient exploration of sequence space to identify protein variants that exhibit the desired characteristics. Whether through rational design, directed evolution, or *de novo* design, protein design holds immense potential for advancements in medicine, biotechnology, and basic research.

2.3.1 Sequence Generation Using Language Models

Protein sequence generation using language models is an emerging field at the intersection of computational biology and artificial intelligence. Traditional methods of protein design rely heavily on experimental data and evolutionary insights, which are often time-consuming and costly. However, advancements in language models, particularly those inspired by natural language processing, have opened new avenues for protein sequence prediction and design. By training on vast datasets of known protein sequences, these models learn the underlying patterns and rules that govern protein structure and function, enabling them to generate novel sequences that could potentially exhibit desired properties.

The core idea behind using language models for protein sequence generation is their ability to capture long-range dependencies and complex relationships within sequences. Just as these models can generate coherent and contextually relevant text, they can produce biologically plausible protein sequences by understanding the grammar and syntax of amino acid arrangements. Techniques such as transformer architectures have proven particularly effective in this domain, allowing for the consideration of multiple layers of contextual information. These models not only predict the next amino acid in a sequence but can also evaluate the likelihood of a sequence to fold into a functional protein structure, which is crucial for practical applications.

One of the most promising applications of protein sequence generation using language models is in the design of therapeutic proteins and enzymes. By tailoring sequences to achieve specific binding affinities or catalytic activities, researchers can develop proteins with enhanced efficacy and stability for medical and industrial purposes. Furthermore, these models can expedite the discovery of new proteins by generating diverse sequence libraries that can be experimentally tested for desired characteristics. As the field progresses, integrating language models with experimental validation and feedback will likely lead to more accurate and efficient protein design, ultimately transforming our approach to biotechnology and synthetic biology.

One example of such models is MSA Transformer [82]. The MSA Transformer model, proposed by Rao *et al.* [82], is a deep learning model specifically designed to harness the evolutionary information contained within a Multiple Sequence Alignment (MSA) for protein sequence analysis. The core of this model is the MSA Transformer block (Figure 2.3), which extends the standard Transformer architecture by incorporating specialized attention mechanisms to capture the intricate relationships within MSAs. Each block

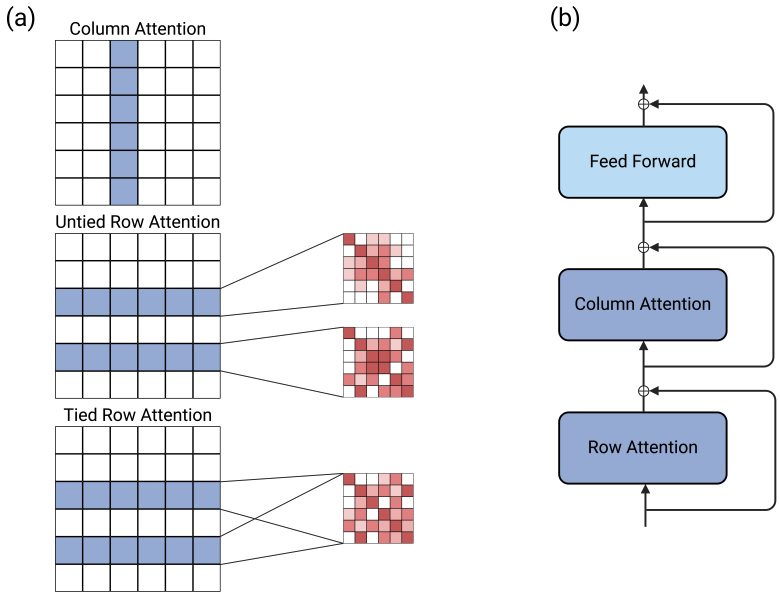


Figure 2.3: Representation of the MSA Transformer architecture. **(a)**: The sparsity structure of attention in the MSA Transformer, where attention is constrained to operate over rows and columns, reduces computational complexity from $O(M^2L^2)$ to $O(LM^2) + O(ML^2)$, with M as the number of rows and L as the number of columns in the MSA. Untied row attention uses distinct attention maps for each sequence in the MSA, while tied row attention employs a single attention map for all sequences, simplifying the contact structure. Ablation studies evaluate both approaches, with the final model utilizing tied attention. **(b)**: The architecture of a single MSA Transformer block from the final model. Some ablation studies explored variations in the ordering of row and column attention. This figure is adapted from [82].

consists of a series of layers that include multi-head self-attention and feedforward networks. However, what sets the MSA Transformer apart is its dual attention mechanism, which operates both across sequences (row attention) and along the positions within each sequence (column attention). Row attention allows the model to attend to different sequences within the MSA at the same alignment position, capturing how evolutionary variations correlate across sequences. Column attention, on the other hand, operates along the length of each individual sequence, enabling the model to understand dependencies between different positions within the same sequence.

A key innovation in the MSA Transformer is the distinction between tied and untied row attention. Tied row attention uses a single attention head shared across all sequences in the MSA, assuming uniform relationships between sequences at a given position. This approach reduces the number of parameters and enhances computational efficiency. In contrast, untied row attention allows each sequence to have its own set of attention weights, providing a more flexible and nuanced model that can capture diverse and complex relationships between sequences. Although untied row attention is more computationally demanding, it offers a more expressive model capable of capturing intricate evolutionary patterns.

When processing a MSA, the MSA Transformer first encodes each sequence with positional information to preserve the order of residues. The encoded sequences are

then passed through multiple layers of the MSA Transformer block, where both row and column attention mechanisms are applied. Row attention, whether tied or untied, focuses on capturing evolutionary correlations across sequences, while column attention captures the intra-sequence dependencies crucial for understanding protein structure and function. This architecture allows the MSA Transformer to generate representations that encapsulate both the evolutionary history of protein families and the structural dependencies within individual sequences.

2.3.2 Sequence Design Using Deep Learning Methods

Protein sequence design using deep learning methods represents a significant leap forward in the field of bioengineering and synthetic biology. Traditional protein engineering approaches often involve laborious and iterative processes of mutation and selection, heavily dependent on prior knowledge and experimental data. In contrast, deep learning methods harness the power of large-scale data and sophisticated algorithms to predict and generate protein sequences with desired properties. By training on extensive datasets of known protein sequences and structures, deep learning models can uncover intricate patterns and relationships that are not immediately apparent to human researchers, by processing and analyzing the complex, high-dimensional data associated with proteins, learning the critical features that influence their folding and function, thus enabling more efficient and innovative protein design.

The practical applications of protein sequence design using deep learning are vast and transformative. In the pharmaceutical industry, these methods are being used to develop novel therapeutic proteins and antibodies with improved efficacy and reduced side effects. Furthermore, deep learning approaches facilitate the rapid exploration of sequence space, generating diverse candidate proteins that can be experimentally tested and validated. As these technologies continue to advance, they promise to accelerate the pace of discovery and innovation in protein engineering, ultimately leading to breakthroughs in medicine, environmental sustainability, and beyond.

ProteinMPNN [83] is an example of a deep learning model designed for the *de novo* design and optimization of protein sequences that can fold into specific three-dimensional structures. The model is built around a message-passing neural network (hence the name) architecture, which is well-suited for handling graph-structured data like protein structures. In ProteinMPNN, the protein's 3D structure is represented as a graph, where each amino acid residue is a node, and the edges represent the spatial or sequential connections between these residues. This graph-based representation allows the model to consider both the local environment of each residue and its interactions with distant parts of the protein, which is crucial for accurate protein design.

The core of ProteinMPNN's architecture (Figure 2.4) is its message-passing mechanism, which facilitates the exchange of information (or "messages") between nodes along the edges of the graph. During each iteration of message passing, each node aggregates

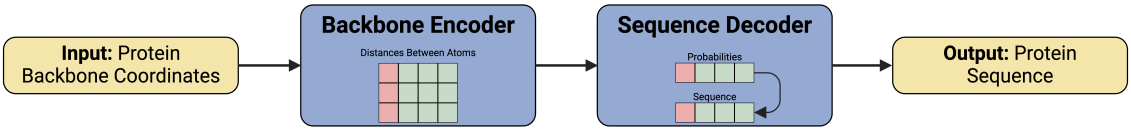


Figure 2.4: Representation of the ProteinMPNN architecture. ProteinMPNN operates through a systematic process to design and optimize protein sequences. It begins with a 3D protein structure represented as a graph, where each residue is a node and spatial or sequential connections between residues are edges. The model then applies iterative message-passing, where information is exchanged between nodes to update residue representations and capture both local and global structural contexts. Using these refined representations, ProteinMPNN predicts amino acid sequences that are likely to fold into the specified target structure. To ensure the quality of the predicted sequences, energy-based scoring functions evaluate their stability and compatibility with the desired fold. The final output is an optimized protein sequence designed to effectively adopt the input 3D structure. This figure is adapted from [83].

information from its neighboring nodes, updating its representation to capture both local and global structural contexts. This process enables the model to understand complex dependencies within the protein structure, allowing it to predict how changes in one part of the sequence might impact the overall structure and function of the protein.

The sequence design process in ProteinMPNN involves generating or optimizing a protein sequence that can fold into a specified target structure. The model takes a partially defined protein structure, often represented by a backbone scaffold of the desired 3D fold, as input. Through the message-passing mechanism, ProteinMPNN predicts the most likely amino acid sequence that will adopt this structure. The model iteratively refines its sequence predictions, ensuring that the generated sequence is both energetically favorable and structurally compatible with the target fold. To further validate the quality of the designed sequences, ProteinMPNN incorporates energy-based scoring functions that assess factors such as residue-residue interactions and steric clashes, which are critical for stable protein folding.

ProteinMPNN is trained on large datasets of protein structures, learning the relationship between sequence and structure across diverse protein families. This training allows the model to generalize well to new structures, making it capable of designing sequences for novel folds that have not been previously observed. The model's architecture enables it to efficiently explore the vast sequence space, narrowing down to sequences that are likely to fold correctly while meeting specific functional criteria. ProteinMPNN's ability to capture both local and global structural features, combined with its capacity to generalize across different protein families, makes it a powerful tool for applications in drug discovery, synthetic biology, and the engineering of proteins with new or enhanced functions.

2.3.3 Exploiting Protein-Protein Interaction Fingerprints in Protein Design

Exploiting PPI fingerprints in protein design is an innovative approach that leverages the specific interaction patterns between proteins to guide the creation of new proteins with desired functions. PPI are fundamental to virtually all biological processes, including signal transduction, immune responses, and cellular metabolism. By understanding and mimicking these interactions, one can design proteins that precisely modulate these

processes. PPI fingerprints, which are detailed maps of interaction sites and the nature of interactions between protein surfaces, provide crucial insights into how proteins recognize and bind to each other. These fingerprints can be used as blueprints for engineering proteins with tailored interaction profiles.

The use of PPI fingerprints in protein design involves several advanced computational techniques. Machine learning algorithms can analyze large datasets of known PPI to identify common features and critical residues responsible for binding specificity and affinity. These features can then inform the design of new proteins or the modification of existing ones to enhance or inhibit specific interactions. For instance, by altering amino acids at key interaction sites, scientists can create proteins that bind more tightly to their targets or disrupt pathological interactions in disease contexts. This approach not only enhances the precision of protein design but also expands the range of potential therapeutic and industrial applications.

One of the most promising applications of PPI fingerprint-based design is in the development of novel therapeutics. By designing proteins that can specifically interfere with disease-related PPI, it is possible to develop treatments for conditions such as cancer, autoimmune and infectious diseases. Additionally, this method can be used to create synthetic proteins that enhance beneficial interactions, such as those involved in immune responses or metabolic pathways. As computational tools and databases continue to improve, the integration of PPI fingerprints into protein design will likely lead to even more sophisticated and effective protein-based solutions for a wide array of challenges.

One tool that exploits PPI fingerprints is MaSIF [84]. MaSIF is a deep learning model designed for the analysis and prediction of protein structures, focusing particularly on the identification of protein-binding sites. The architecture of MaSIF (Figure 2.5) is tailored to capture spatial and geometric features of protein surfaces, leveraging a graph-based approach to represent and process these structures effectively.

At its core, MaSIF employs a graph convolutional network to encode the spatial arrangement of atoms on the protein surface. The model represents the protein structure as a 3D graph, where nodes correspond to atoms and edges represent spatial proximities or interactions between atoms. This graph-based representation allows MaSIF to model the local environment around each atom, capturing both the geometric and chemical context crucial for understanding protein function.

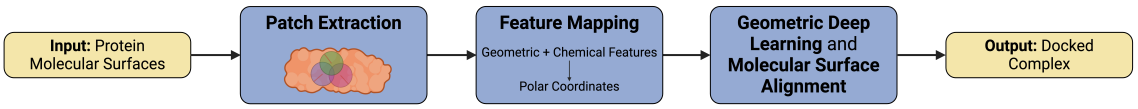


Figure 2.5: Representation of the MaSIF's conceptual design and implementation. Given a protein structure, its surface is segmented into overlapping radial patches of a fixed geodesic radius for analysis in MaSIF. Each patch includes geometric and chemical features mapped onto the protein surface using polar geodesic coordinates. MaSIF uses geometric deep learning with convolutional neural networks to generate fingerprint descriptors for surface patches, followed by specialized layers to analyze these descriptors; patches of interacting molecules are then aligned by applying translations and rotations to maximize geometric and chemical complementarity. This figure is adapted from [84].

MaSIF processes protein structures through several stages. First, it uses a 3D grid-based approach to discretize the protein surface into a spatial grid, which captures the local geometric context of the protein. The grid data is then integrated with the graph-based representation, enhancing the model's ability to learn from both spatial and topological features. The graph convolutional layers aggregate information from neighboring nodes, allowing the model to learn complex patterns and interactions within the protein structure.

A key feature of MaSIF is its ability to learn from surface patches, where the model extracts local regions of the protein surface and analyzes them for potential binding sites. MaSIF identifies these patches based on both their geometric and chemical properties, focusing on concave regions, flat surfaces, and charged areas. Using surface alignment, MaSIF compares these patches to potential interaction partners, aligning compatible surfaces by matching geometrically and chemically complementary regions. This approach enables MaSIF to generate docked complexes by positioning the surfaces together in a way that maximizes complementarity, facilitating the prediction of protein-ligand and protein-protein interactions.

The model is trained on large datasets of protein structures with annotated binding sites, enabling it to generalize across different proteins and predict binding sites in novel structures. By using surface alignment to generate docked complexes, MaSIF can propose potential interaction sites between proteins or ligands, making it a powerful tool for predicting protein-protein interactions and understanding protein function at a high level of detail.

MaSIF's architecture is designed to handle large-scale protein structures efficiently. The model's use of graph convolutions allows it to process complex 3D spatial data, capturing detailed structural features while maintaining computational efficiency. This makes MaSIF particularly effective in analyzing protein surfaces and predicting how proteins will interact with other molecules.

2.4 Generation, Selection and Refinement of RBD-targeting Monobodies

Figure 2.6 depicts a general overview of our protein design pipeline.

Starting from the 10^{th} FN3 (PDB ID: 1TTG), sequence search was performed using BLAST [85] and FoldSeek [86]. Hits over 50% identity were retrieved. In-house Python scripts were used to manipulate and analyze the downloaded sequences. Those sequences were then aligned using Clustal Omega [87].

2.4.1 Sequence Generation of Monobodies Using MSA Transformer

The aligned sequences obtained previously served as input for the MSA Transformer, which was applied using the iterative masking approach of Sgarbossa *et al.* [88]. The

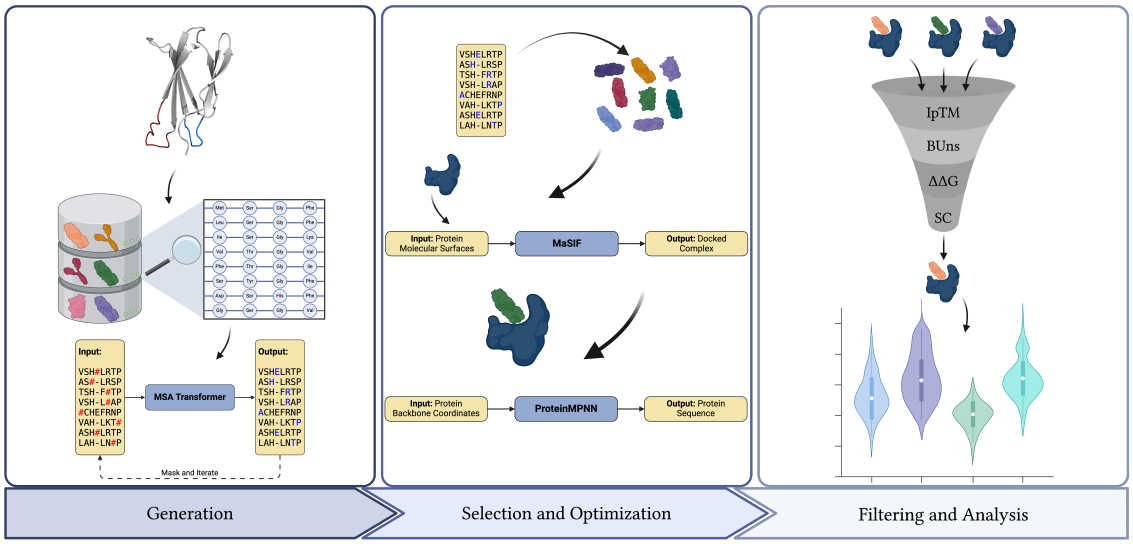


Figure 2.6: Overall representation of the protein design pipeline applied in this thesis. **Generation**: starting from PDB ID: 1TTG, sequence search was performed using BLAST and FoldSeek; obtained sequences were aligned using Clustal Omega and used as input for MSA Transformer (in alternative to the MSA Transformer-generated dataset, sequences from the sequence search step were tested). **Selection and Optimization**: generated sequences were clustered using MMseqs2 and representatives from each cluster proceded to structure prediction using ColabFold; MaSIF was used to dock these structures with SARS-CoV-2 RBD as proof of concept (in the alternative approach, docking was done with ColabFold); obtained complexes proceded to sequence optimization using ProteinMPNN. **Filtering and Analysis**: optimized complexes were filtered according to AlphaFold and Rosetta interface metrics; results were used to compare the performance of different approaches and datasets.

model was run in three batches of 200 iterations of sequence generation using different numbers of combinations and sequences per combinations, as described in Table 2.1.

2.4.2 Clustering of Generated Sequences and Structure Prediction

The monobody sequences generated by the MSA Transformer [88] were clustered using MMseqs2 [89]. To guarantee that the program would not generate neither several clusters of few sequences nor few clusters of several sequences, a 75% identity was used. This means that two sequences are to be considered as members of the same cluster if their identity is equal or greater than 75%. Using the same tool, a cluster representative sequence was chosen using a coverage of 95%. This means that the sequence representative is a real member of the cluster that shares a part of the sequence and represents at least 95% of the other members of the cluster.

For the process of structure prediction, we used ColabFold [90]. For every cluster

Table 2.1: Summary of the parametrization of the iterative masking approach for sequence generation.

Batch	Number of Combinations	Number of Sequences per Combination	Number of Iterations
1	100	200	200
2	100	500	200
3	250	250	200

representative sequence, three structures were predicted using only one seed, without template and without relaxation. The predicted structures were ranked by the program and only the best raked one was considered for the further steps. The results were filtered according to the ColabFold confidence metrics predicted local Distance Difference Test (plDDT), predicted Aligned Error (pAE), predicted Template Modelling score (pTM) and their predicted secondary structure as described in Table 2.2.

2.4.3 Selection of RBD-interacting Monobodies Using MaSIF

For docking the predicted structures with the target, we used MaSIF [84]. Using scripts available in the authors' online repository (`https://github.com/LPDI-EPFL/masif`) as a starting point and using in-house Python scripts to apply the tool to our use case, the docking pipeline is as follows:

1. Find core residues in the middle of the interface;
2. Interface residues' descriptor extraction;
3. Extract docking sites;
4. Docking.

Because we wanted the best targeting possible to the SARS-CoV-2 RBM, Glu493 RBD residue was chosen. Since this residue is central in the region we wanted to target, choosing it would allow MaSIF to find better docking positions.

After this step, the docking results were filtered based on the MaSIF score, consisting on the probability output from the deep learning model (the closer to the value 1 the better, meaning that it is predicted a higher binding probability), and on the alignment fitness, which measures how close the docked poses and their target are spatially. All results with MaSIF score under 0.99 and alignment fitness under 0.5 were filtered out.

2.4.4 Sequence Optimization Using ProteinMPNN

For refining the monobody sequences in the interface region, we used ProteinMPNN [83]. In this step, two strategies were tested:

1. Enabling the mutation of up to 6 random residues in the BC and FG loops, excluding Proline residues;

Table 2.2: Summary of the cutoff values of the metrics for filtering the ColabFold predicted structures.

Confidence Metrics	Cutoff Value
plDDT	> 90%
pAE	< 6Å
pTM	> 0.60
Number of β-strands	7

2. Enabling the mutation of all but Proline residues in the BC and FG loops.

For both cases, each monobody sequence was set to generate 500 new monobody sequences by applying mutations as described. All generated monobodies went under ProteinMPNN fast relax, to correct their conformations in complex with the target after applying the mutations.

2.4.5 Filtering and Selection of Designed Monobodies

The final selection of the best monobodies is done by a two-stage filtering. In the first stage, we use AlphaFold to predict and rank the structures of protein complexes [91]. We filtered the docked complexes using a Interface predicted Template Modelling score (IpTM) cutoff of 0.75. In the second stage, we use Rosetta software suite (available in `https://www.rosettacommons.org/software`) to calculate various metrics for the interface between the monobody and the target. We filtered the docked complexes using a Shape Complementarity (SC) cutoff of 0.65, a Buried Unsatisfied interface H-bonds (BUns) cutoff of 4.0 and a delta-delta-G ($\Delta\Delta G$) cutoff of -20.0. Table 2.3 summarizes the metrics and cutoff values of this filtering process.

2.4.6 Alternative Monobody Identification Strategy Using AlphaFold

In this alternative approach, we used ColabFold [90] for multimer structure prediction [92].

The target sequence was added to every cluster representative sequence as a new chain (hence multimer structure prediction). As before, three structures were predicted using only one seed, without template and without relaxation. The predicted structures were ranked by the program and only the best ranked one was considered for the further steps. The results were filtered according to the ColabFold confidence metrics plDDT, pAE and pTM as described in Table 2.4.

Table 2.3: Summary of the cutoff values of the AlphaFold and Rosetta metrics for filtering the structures of the docked complexes.

Tool	Confidence Metrics	Cutoff Value
AlphaFold	IpTM	> 0.75
Rosetta	SC	> 0.65
Rosetta	BUns	≤ 4.0
Rosetta	$\Delta\Delta G$	$\leq$ -20.0

Table 2.4: Summary of the cutoff values of the metrics for filtering the ColabFold predicted structures.

Confidence Metrics	Cutoff Value
plDDT	$> 90\%$
pAE	< 10Å
pTM	> 0.50

For the rest of the pipeline, all the steps are the same as before except for the sequence optimization using ProteinMPNN [83]. Here, each monobody sequence was set to generate 500 new monobody sequences by applying mutations without being restricted to the BC and FG loops, i.e., allowing the redesign of the entire sequences.

3

Results and Discussion

3.1 Molecular Dynamics Simulations of SARS-CoV-2 Variants

Understanding the dynamic behavior of viral targets at the molecular level is crucial to increase the chances of success in the design of antivirals. MD simulations provide detailed insights into the conformational flexibility and structural stability of these targets, which are often not apparent from static crystal structures alone. By simulating the atomic movements over time, MD simulations help in identifying transient binding pockets, conformational changes upon ligand binding, and potential allosteric sites that could be targeted by novel antivirals. This dynamic perspective is essential for designing inhibitors that can effectively bind to and inhibit viral proteins under physiological conditions, thus increasing the likelihood of therapeutic success.

Moreover, MD simulations allow for the assessment of the impact of mutations on the viral target's structure and function. Given the high mutation rates of many viruses, this capability is particularly valuable for anticipating resistance mechanisms and designing broad-spectrum antivirals. Through MD simulations, one can evaluate how different mutations might alter the binding affinity of potential antivirals, enabling the preemptive modification of antiviral candidates to maintain efficacy.

Therefore, incorporating MD simulations into the monobody design process not only enhances the understanding of viral target dynamics but also provides a robust framework for developing more effective and durable monobodies specific for that target in the future. In specific, BA.2.86 and JN.1 SARS-CoV-2 variants were selected for being the most relevant and predominant variants as of the writing of this document, and MD simulations of their respective RBD were conducted. In particular, BA.2.86 is the ancestor variant of JN.1, being the only difference between the two a single mutation in the RBD: L455S in the Spike protein (or L123S, for RBD residues). Therefore, the objective was to understand the viral evolution of SARS-CoV-2 and the extent of the impact of mutations in the dynamics of the RBD by comparing our results with the findings of Valério *et al.* [49].

3.1.1 Analysis of the System Properties

Looking at Figure 3.1, one of the first things that is possible to observe is that the RBD of both variants is very dynamic. For BA.2.86, RMSD values vary around the 5 Å range throughout all replicates with more abrupt peaks, which does not occur for JN.1, where some replicates show variation around the 3 Å and 4 Å range and with overall less abrupt peaks. In gross terms, this means that JN.1 converges to an equilibrium state that is less distant from the initial reference state when compared with BA.2.86; and that after the considered starting point of the equilibrium stage, 4 μs, JN.1 shows more stability with smoother curves than BA.2.86, which presents rough peaks still after 4 μs.

For the RMSF, however, the differences are not as significant. Although there is an increase in the RMSF value of one BA.2.86 replicate in the lower region of the loop present in the RBM compared to JN.1, overall this per residue analysis shows similar results for both variants, even for position 123, the position where occurs the only mutation that originated the JN.1 variant from its predecessor BA.2.86. With this analysis, it is possible to conclude that both variants are equally flexible on loop regions and stable on core residues, having the mutation L123S caused no noticeable impact on this property.

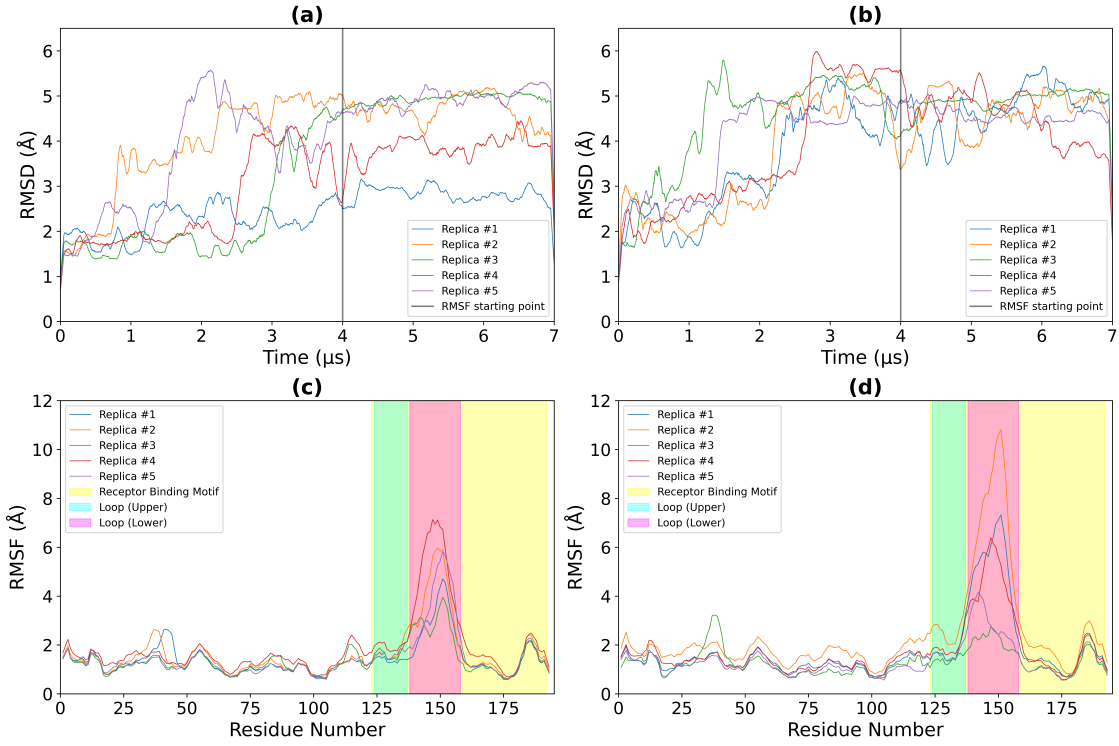


Figure 3.1: Root Mean Square Deviations and Root Mean Square Fluctuations for the JN.1 (**(a)** and **(c)**, respectively) and BA.2.86 (**(b)** and **(d)**, respectively) RBD variants MD simulations. For the RMSD plots: on the X axis is depicted the simulation time in μs; on the Y axis is depicted the RMSD in Å. For the RMSF plots: on the X axis are depicted the residue numbers; on the Y axis is depicted the RMSF in Å.

3.1.2 Analysis of the Conformational Dynamics

Regarding the two-dimensional PCA of SARS-CoV-2 RBD conformational dynamics in water, our simulations (Figure 3.2) share some similarities with those of Valério *et al.* [49]. Valério *et al.* conducted MD simulations of the Omicron RBD and found no clear prevalence of two distinct sets of RBM conformations. Their PCA identified a deep basin cluster corresponding to the WT open conformation, while only shallow basins were associated with the closed conformation. The open conformation accounted for approximately 95% of the configurations sampled during the simulations of the Omicron variant [49].

In our case, the simulations of both the BA.2.86 and JN.1 Omicron subvariants similarly do not reveal a clear prevalence of two distinct sets of RBM conformations. However, significant differences were observed in the free energy landscapes and the distribution of conformational states. For BA.2.86, the open conformation accounts for approximately 70% of the sampled conformations, whereas for JN.1, 63% of the sampled conformations correspond to the open state (Table A.2). Both variants exhibit lower free energy values for the most prevalent conformations compared to the values reported by Valério *et al.* [49].

This observed decrease in the percentage of open conformations (and the consequent increase in the percentage of closed conformations) suggests that these variants' RBDs may exhibit reduced availability for interaction with ACE2. The lower prevalence of the open conformation, which is the state necessary for effective binding to ACE2, implies a potential reduction in viral infectivity for these Omicron subvariants. However, this shift in conformational dynamics also has another critical implication: the reduced presence of the open conformation may lead to decreased exposure of key epitopes that are typically targeted by antibodies and antivirals.

The conformational dynamics observed in these viral variants are directly linked to the mutations that characterize new strains. Each mutation can subtly or significantly

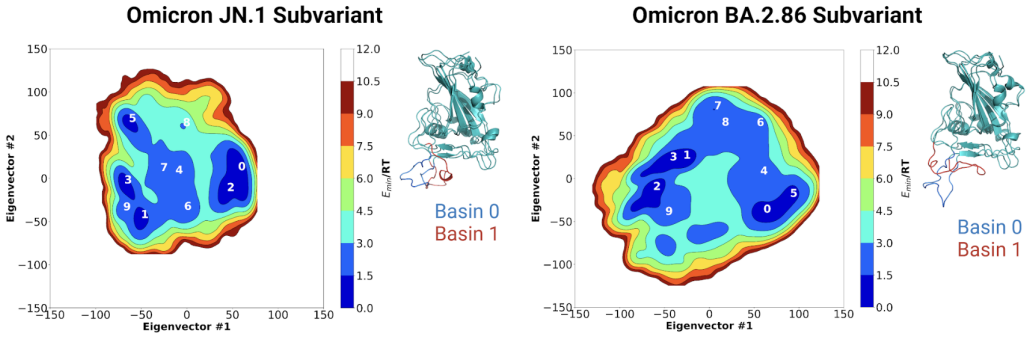


Figure 3.2: Two-dimension PCA analysis of SARS-CoV-2 RBD conformational dynamics in water. Plots of the first two principal components determined from the Cα backbone of the JN.1 (**left**) and BA.2.86 (**right**) RBD variants. Basins of the 10 most prevalent conformations are numbered. Snapshots of the most prevalent structures for selected open and closed basins are also shown. The ridge regions of the open and closed snapshots are colored in blue and red, respectively.

alter the structural landscape of viral proteins, influencing not only their interaction with host receptors but also their susceptibility to neutralization by the host immune system. In the context of viral evolution, these findings underscore the critical role that mutations play in enabling viruses to adapt to selective pressures, such as host immune responses and antiviral therapies.

The emergence of new viral variants, which exhibit altered conformational dynamics, highlights the potential for viruses to continuously evolve mechanisms to evade immunity. This evolutionary process is driven by mutations that may reduce the effectiveness of therapeutic interventions designed based on earlier viral strains. The percentage reduction of open conformations, as observed in our results, could signify a strategic adaptation by these variants to avoid detection and neutralization, posing a challenge to current public health measures.

These findings emphasize the urgent need for enhanced pandemic preparedness across a broad range of viral pathogens. As viruses continue to mutate, it is crucial to monitor not only genetic changes but also the structural and functional consequences of these mutations. Understanding how new variants impact conformational dynamics can provide valuable insights into a virus's potential to evade immunity and inform the development of next-generation therapeutics. Continuous surveillance and the ability to rapidly adapt in response to emerging variants are essential to mitigate the impact of future viral outbreaks and maintain control over pandemics.

3.2 Generation, Selection and Refinement of RBD-targeting Monobodies

In order to be able to generate new monobody sequences with confidence using a language model, we need to feed the model with as many monobody sequences as possible. The more sequences we feed the model, the better trained it will be to generate new sequences that can be considered as members of the same family as the ones we fed it with. Given that there are not that many monobody sequences described in the literature, we needed to perform a sequence search step prior to the sequence generation.

Using a 50% identity threshold allowed us to have a starting pool of monobody-like proteins that are not that distinct from our monobody of reference 1TTG, avoiding major structural and functional differences but still considering slight conformational and surface dissimilarities. If, on one hand, we used a higher threshold (75%, for example), we would obtain less sequences to input to the language model despite being similar to our monobody of reference. If, on the other hand, we used a lower threshold (35%, for example), we would obtain more sequences to input to the model but they would represent proteins more structural and functionally different from our reference than desired.

This sequence search step using a 50% identity threshold allowed us to obtain 3,050 monobody-like protein sequences (Figure 3.3).

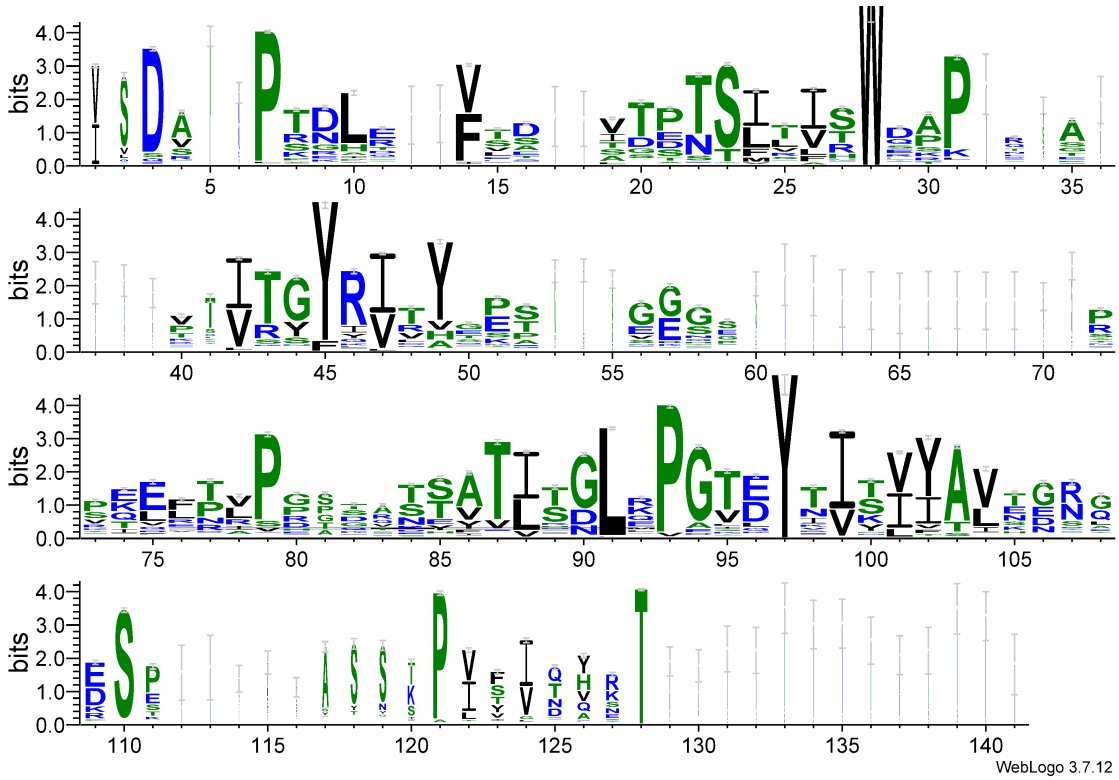


Figure 3.3: WebLogo [93] of the sequence search step sequences. The bigger the letter for a given residue, the more conserved it is in that position. Hydrophilic residues are depicted in blue; hydrophobic residues are depicted in black; neutral residues are depicted in green.

Figure 3.3 highlights key patterns of sequence conservation across the aligned monobody sequences, shedding light on the functional and structural importance of different regions. Notably, the highly conserved residues, correspond to the protein's β-strands. This suggests that these positions are likely crucial for maintaining the structural integrity of the monobodies, as β-strands play a key role in forming the stable β-sheet framework typical of monobody structures. The conservation in these regions underscores their importance in stabilizing the overall fold of the protein and ensuring proper functionality.

In contrast, more variable regions, correspond to loops, which are generally more tolerant to sequence diversity. These regions might allow for flexibility in interactions or adaptability to different binding targets without compromising the monobody's structural core. This variability could also facilitate evolution in response to different functional requirements or binding environments.

Overall, the combination of highly conserved β-strand regions and more flexible loop regions reflects the balance between maintaining a robust structural framework and allowing for functional adaptability. The WebLogo underscores the critical role of conserved residues in supporting the β-sandwich structure, while also highlighting areas of potential flexibility that could be exploited for designing monobodies with varied or enhanced binding properties.

3.2.1 Sequence Generation of Monobodies Using MSA Transformer

To generate new monobody-like protein sequences, we applied MSA Transformer [82] using an iterative masking approach [88]. In this approach, random positions of the input sequences are masked, and the model tries to use the context of the rest of the input multiple sequence alignment to predict which residues (or gaps) were in the masked positions originally (Figure 3.4). This process occurs iteratively, increasing the pool of generated sequences each time the model runs.

We decided to use this approach because it had been already tested for the FN3 family and also because this generation method based on MSA Transformer outperformed Potts models, proving to be a strong candidate for protein sequence generation and protein design [88].

From the 3,050 monobody-like sequences obtained in the previous sequence search step, the model outputted a total of 1,169,115 unique protein sequences, ranging from 32 to 141 residues in length (Figure 3.5).

Figure 3.5 illustrates the diversity of sequence lengths centered around the monobody average size of 90 residues. It is shown a broad distribution of sequences, with a noticeable concentration around the 85 to 95 residue range, indicating significant diversity in sequence lengths close to the monobody reference size. The highest frequencies occur within this range, suggesting that sequences slightly shorter or longer than 90 residues are the most common.

While sequences of exactly 90 residues do not dominate, their neighboring bins show substantial counts, signifying considerable variation. The diversity diminishes as sequence lengths deviate further from the 90-residue mark. For instance, shorter sequences (e.g., 32-61 residues) and longer sequences (over 100 residues) occur much less frequently, reflecting the lower prevalence of extreme sizes. This overall distribution underscores the capability of generating sequence size diversity of by the MSA Transformer.

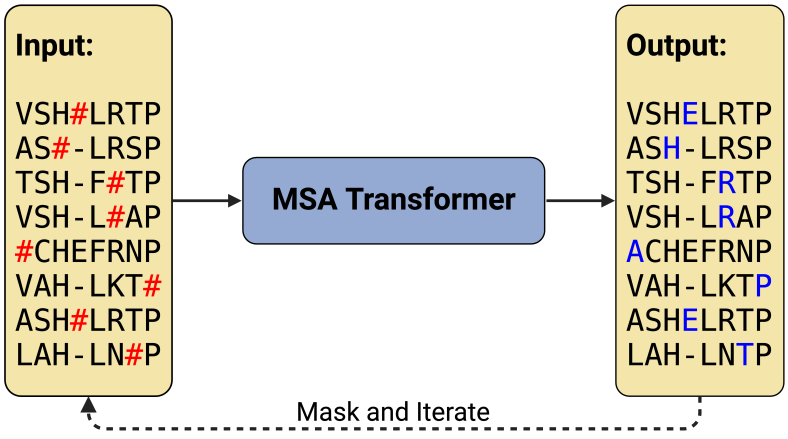


Figure 3.4: Iterative masking approach to generate sequences using MSA Transformer. The red hashtag (#) stands for a masked amino acid, while blue uppercase letters stand for predicted amino acids at the masked positions. This figure is adapted from [88].

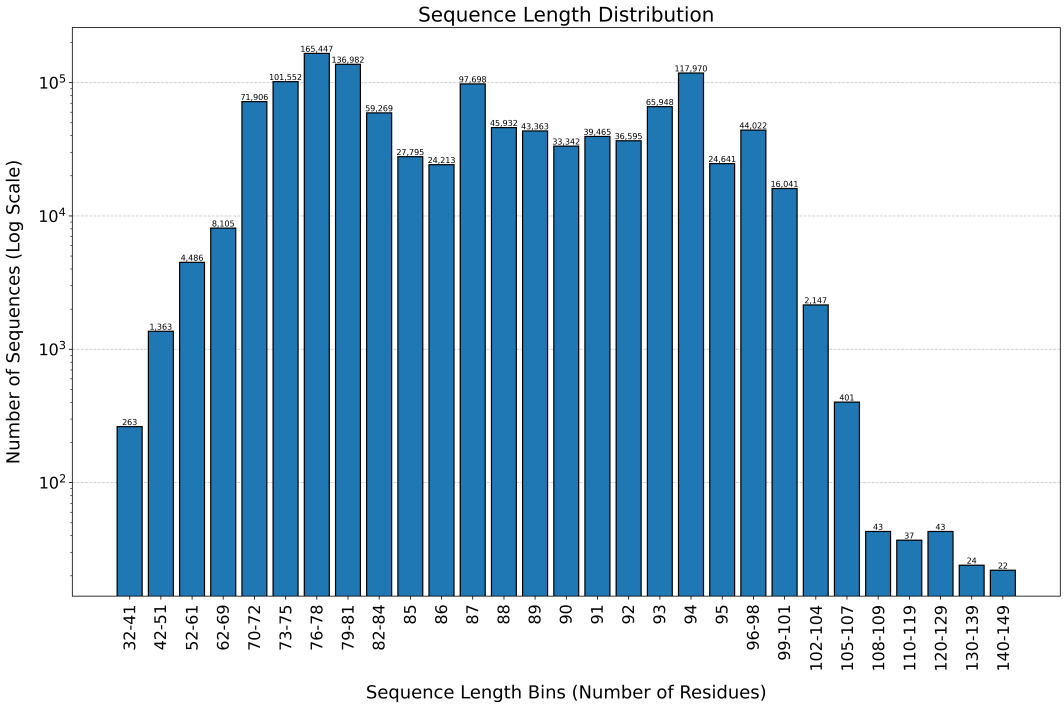


Figure 3.5: Length distribution of the monobody sequences generated by the MSA Transformer [88]. X axis values, representing the number of residues of each sequence, are divided in bins with different sizes; Y axis values, representing the number of sequences, are represented in logarithmic scale. The minimum and maximum achieved sizes are 32 and 141 residues, respectively, with a frequency of 2 and 1, respectively. The most frequent size is 94 with a frequency of 117,970.

3.2.2 Clustering of Generated Sequences and Structure Prediction

Protein structure prediction is a very computationally demanding process, which reflects in the time it takes to run. Given that the objective of this work was to streamline the development of monobodies, we needed to opt for a strategy to reduce the computational time without compromising the accuracy and reliability of the structure prediction step.

Because of this, we decided to perform sequence clustering to narrow down the number of sequences whose structures were to be predicted. Sequence clustering groups similar sequences based on identity, thereby allowing us to consolidate sequences into clusters, significantly reducing computational costs while preserving the overall structural diversity of the dataset.

Here, a compromise must be made between the number of clusters desired (and consequently the number of structures to be predicted) and the size of each cluster. The larger the cluster, the more sequence variability it contains, which reduces the confidence in a representative sequence accurately reflecting all members. Simply, if we aim for fewer clusters and thus fewer sequences to predict, the clusters will be larger and more heterogeneous; if we aim for smaller, more homogeneous clusters, the number of clusters—and thus the number of sequences to predict—will be larger. Therefore, aiming to reduce the number of structures to around 10,000, we selected a threshold of 75% sequence identity. This threshold clustered the 1,169,115 sequences generated by the MSA Transformer into

9,369 clusters of varying sizes, ranging from 1 to 72,816 members (Figure 3.6).

To select a representative sequence from each cluster, a sequence coverage threshold of 95% was applied. A higher coverage threshold would result in an attempt to find a "one-fits-all" representative, which would only work if a higher sequence identity threshold were used; a lower coverage threshold would result in choosing a representative that might not adequately represent a significant portion of the cluster. This balance ensured that each representative sequence closely reflected the majority of its cluster.

The clustering results depicted in Figure 3.6 indicate a significant skew towards smaller clusters. Most sequences fall into clusters containing between 1 and 50 members, with the largest number of clusters being singletons (clusters of size 1) and small clusters (size 2–10). This suggests that, while the majority of sequences are highly distinct or form only small similarity groups, there are still sequences that cluster into larger groups of hundreds or even thousands of members. Notably, the largest clusters, containing over 1,000 members, although few in number, make a substantial contribution to computational savings. By predicting the structure of just one representative from these large clusters, we avoid the need to predict the structure of thousands of similar sequences.

This distribution aligns with our goal of reducing the number of sequences for structure prediction, demonstrating that the clustering threshold of 75% identity successfully consolidated sequences while maintaining diversity (Figure 3.7). It reflects a balance between minimizing the number of clusters and ensuring each representative is a meaningful

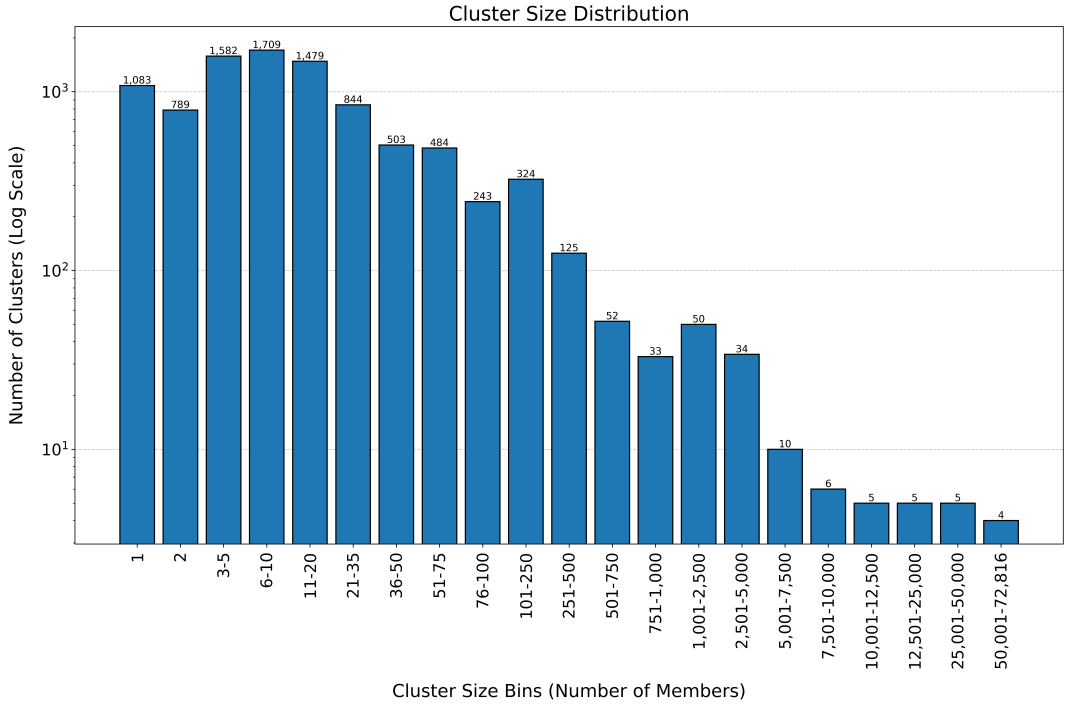


Figure 3.6: Size distribution of the clusters outputted by MMseqs2 [89]. X axis values, representing the number of members in each cluster, are divided in bins with different sizes; Y axis values, representing the number of clusters, are represented in logarithmic scale. The minimum and maximum cluster sizes are 1 and 72,816 members, respectively.

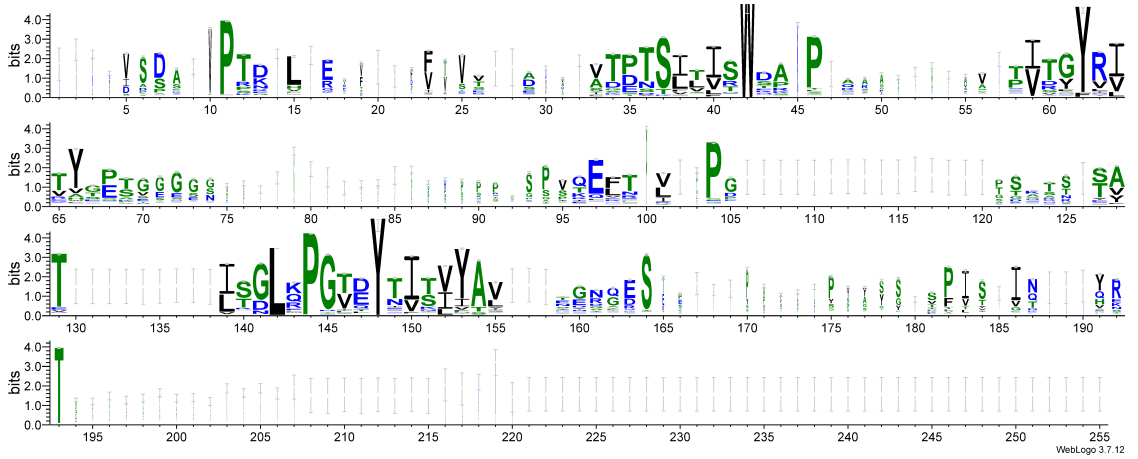


Figure 3.7: WebLogo [93] of the cluster's sequence representatives. The bigger the letter for a given residue, the more conserved it is in that position. Hydrophilic residues are depicted in blue; hydrophobic residues are depicted in black; Neutral residues are depicted in green.

proxy for the sequences in its cluster. Additionally, the presence of very large clusters (e.g., over 500 members) highlights areas of high sequence similarity, likely representing conserved structural motifs or functional domains. These clusters are critical in reducing computational costs without sacrificing the integrity of the analysis.

The WebLogo for this second set of monobody sequences (Figure 3.7) reveals important insights into the conservation patterns when compared to the previous set.

In Figure 3.7, certain regions remain highly conserved, as also noticed in Figure 3.3.These conserved regions likely still correspond to β-strand structures, reinforcing their essential role in maintaining the protein's structural framework. The conservation in these regions reflects that, despite reducing the number of sequences through clustering, the core structural features of the monobodies remain unchanged, supporting the reliability of this clustering strategy.

In terms of overall sequence diversity, the second WebLogo (Figure 3.7) shows slightly less diversity in some conserved regions and more variability in less conserved regions compared to the first WebLogo (Figure 3.3). This indicates that clustering with a 75% identity threshold on the new sequences generated by the MSA Transformer has succeeded in preserving the important structural features, while still incorporating some of the inherent variability of the monobody sequences. Additionally, it demonstrates the trade-off between reducing the number of sequences and maintaining sequence variability in non-conserved regions, especially in surface-exposed loops that could influence binding specificities.

For the protein structure prediction process, we decided to use ColabFold [90] because it offers accelerated prediction of protein structures and complexes by combining the fast homology search of MMseqs2 with AlphaFold2 or RoseTTAFold. Also, ColabFold's 40-60-fold faster search and optimized model utilization enables prediction of close to a thousand structures per day on a server with one graphics processing unit [90]. Altogether, using ColabFold allowed us to decrease the time needed even further without compromising

on the confidence of the results.

Although the generated sequences were obtained from a single monobody of reference, it still existed the possibility of some of those sequences not folding into the wanted monobody structure. Being it because of the size of the sequences, or because of the always present uncertainty and error (from the MSA Transformer or from ColabFold), some predicted structures were, in fact, undesirable for our purpose. Thus, a filtering process was needed. We applied this filtering according to the values as presented in Table 2.2.

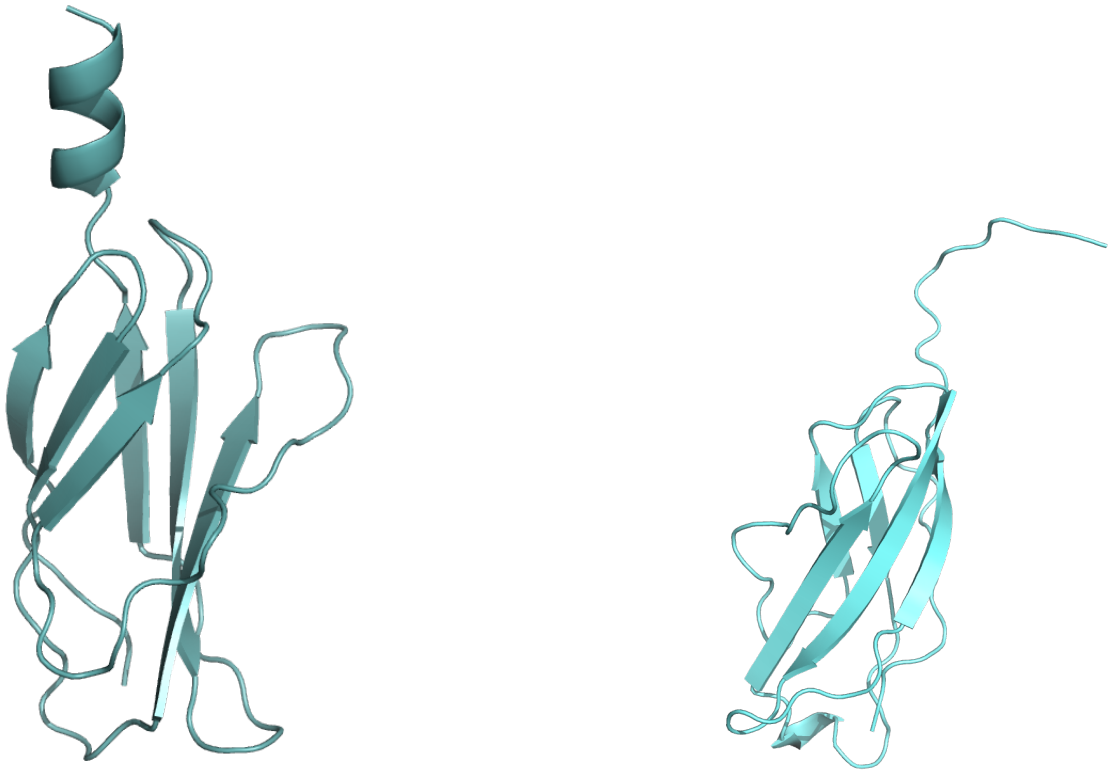
plDDT is a per-residue confidence score that corresponds to the model's predicted score on the lDDT-Cα metric [94]. This measure estimates whether the predicted residue has similar distances to neighboring Cα atoms (within 15 Å) in agreement with distances in the true structure. It ranges from 0% to 100% - the higher the value, the more confident the prediction is.

pAE estimates the expected positional error for each residue in a predicted protein structure if it was aligned to a corresponding residue in the true protein structure. This measurement helps assess the confidence in the relative positions and orientations of different parts of the predicted protein model. Values range from 0 to 35 Å - the lower the value, the more confident the prediction is.

pTM is generated from the template modeling (TM) score, which assesses the accuracy of a protein's overall structure. This score is an integrated assessment of how closely the projected structure matches the hypothetical true structure. It ranges from 0 to 1 - a pTM score above 0.5 indicates that the overall projected fold for the protein is likely comparable to the genuine structure, while a value below 0.5 shows that the predicted structure is probably erroneous.

The number of β-strands was also taken in consideration in this filtering process. As explained in the Introduction subsection 1.3.4, monobodies consist of a 7 β-strand and 6 loop structure with a sequence of around 90 in length. Therefore, those structures which were predicted to have less (or more) than 7 β-strands were filtered out. By removing the structures with less than 7 β-strands, we were also removing the smaller sequences (the ones that were not long enough to achieve the desired number of β-strands and loops), therefore removing the structures without the FG loop needed for the binding; by removing the structures with more than 7 β-strands, we were also removing the longer sequences (the ones that were able to produce more than the characteristic number of secondary structure elements of a typical monobody), therefore removing the structures that could potentially have undesired immunogenic characteristics. Figure 3.8 depicts two examples of proteins that were filtered out in this stage.

By the end of the structure prediction filtering stage, only 2,898 monobody predicted structures out of the 9,369 moved to the docking step.



(a) (b)

Figure 3.8: Examples of aberrant monobodies. These are examples of proteins that were filtered out due to their secondary structure.

3.2.3 Selection of RBD-interacting Monobodies Using MaSIF

As explained before, the typical monobody structure is composed of 7 β-strands with 6 loops between them. Given their nature, the loops are the regions of the protein that can be altered the most without compromising the stability of the protein. Having that in consideration and also the fact that the positions of the BC, DE and FG loops approximately correspond to those of CDR1, 2 and 3, respectively, of the immunoglobulin VH domain [61] (Figure 3.9), we decided to establish the docking interface with the BC and FG loops of the monobody, as the example depicted in Figure 3.10.

Using MaSIF [84], we calculated the fingerprints and established the interface between these monobody regions and the target RBM.

Anticipating the computational power that would be needed in the following sequence optimization step, we used a stricter MaSIF score cutoff value of 0.99 with an alignment fitness cutoff value of 0.5, to reduce the pool of monobodies by their target binding potential. Applying these values, only 25 out of the 2,898 monobody structures passed to the sequence refinement step.

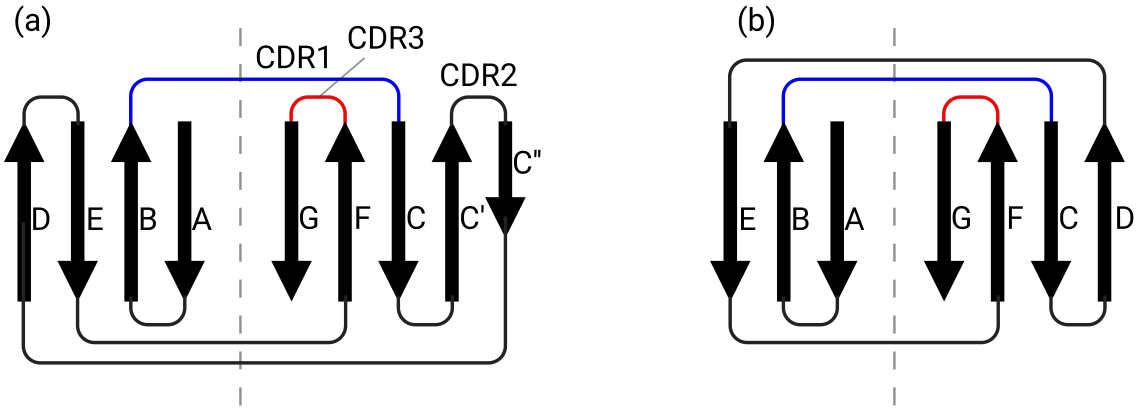
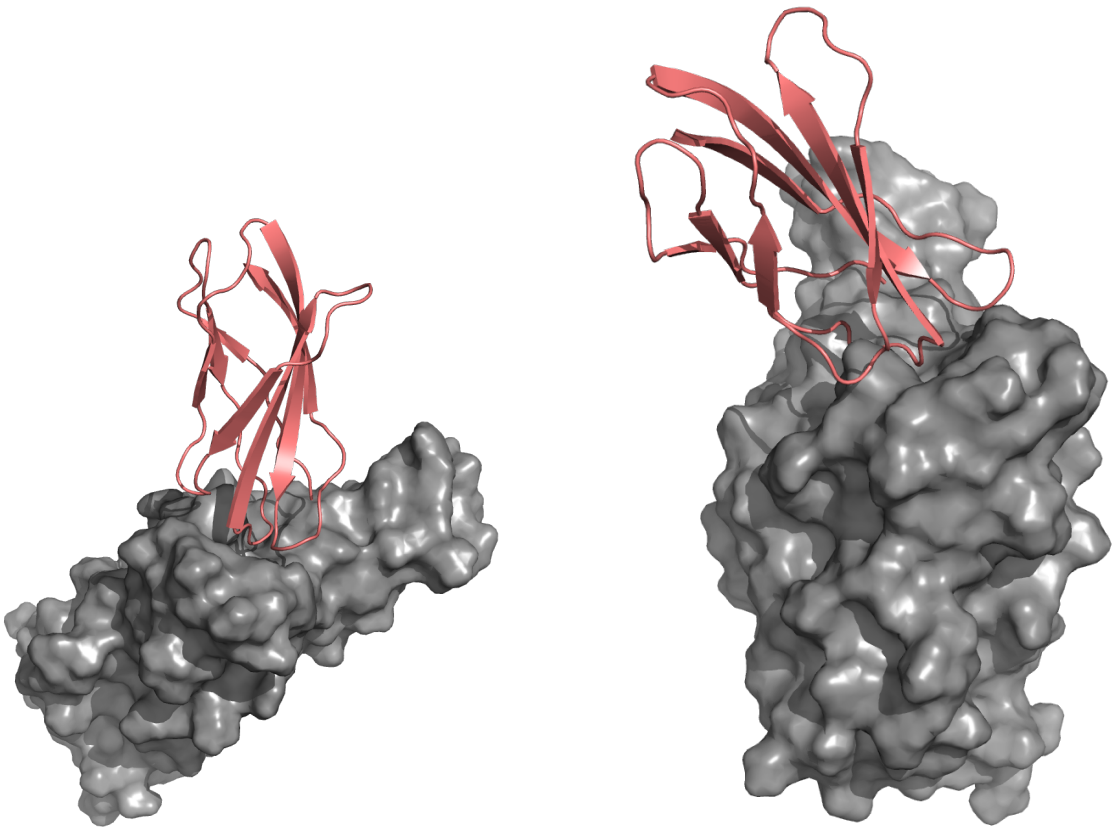


Figure 3.9: Schematic comparison of the secondary structures of an immunoglobulin VH domain **(a)** and the FN3 domain **(b)**. In blue is represented the BC loops; in red is represented the FG loops. Figure adapted from [61].



(a) **(b)**

Figure 3.10: Example of a docked complex with MaSIF that passed the filtering stage. In red is represented the monobody structure; in gray is represented the WT SARS-CoV-2 RBD. Subfigures **(a)** and **(b)** represent the same complex visualized from different angles.

3.2.4 Sequence Optimization Using ProteinMPNN

Sequence optimization is the step where we try to improve the interface interactions between the monobody and the RBD by implementing mutations in the monobodies' residue sequences. For our framework, we decided to test two different approaches: a more conservative approach where we allowed 6 random residues in the BC and FC loops to be mutated without mutating Proline residues, and a less conservative approach where we allowed all but Proline residues in the BC and FG loops to be mutated.

The reason why we chose to not mutate Proline residues in the monobody sequence during the optimization process is because of their unique properties and structural significance. Given its structure, Proline has the ability to act as a "kink" in the protein backbone, affecting the local conformation and potentially disrupting the binding interface if mutated. Its role in maintaining structural integrity make it crucial for the monobody's native structure and function. Additionally, preserving Proline residues reduces the risk of off-target effects, where changes in the monobody's structure or function inadvertently affect other biological processes, and improves sequence recovery, ensuring that the optimization process is more targeted and specific to the desired binding affinity. By not mutating Proline residues, we can maintain the monobody's biological activity and binding affinity, ultimately leading to more effective optimization of the interface region.

Given the computational power available, we were aiming to generate between ten thousand and fifteen thousand new sequences with ProteinMPNN [90] with relaxation. With that in consideration, we previously applied more strict MaSIF [84] cutoff values and here we set each one of the 25 monobody sequences to generate 500 new ones by applying the mutations as described before. By doing this, we were allocating the available computational power to generate sequence diversity only on the proteins that seemed more promising in the docking studies.

When performing the sequence optimization with ProteinMPNN, we decided to include one relaxation cycle for each generated sequence. Simply, this relaxation cycle is an energy minimization step. Since we were generating new sequences for a given input fold, the relaxation cycle is important to ensure the "re-accommodation" of the structure of the new monobody sequence to the target, minimizing the overall energy of the system. If a relaxation cycle was not applied, the input fold would be wrongly considered as the structure of the newly generated monobody sequences.

In this sequence optimization stage, 12,500 different monobodies were generated in total from mutating the 25 sequences, per each one of the ProteinMPNN application approaches.

3.2.5 Filtering and Selection of Designed Monobodies

Until this point, we have been able to generate monobodies with different sequences, different structural conformations and different binding interactions with the target. Therefore, it was necessary to select which might be the best candidates to proceed to further

wet lab validation.

For this final selection, we selected four different metrics from AlphaFold and Rosetta (Table 2.3). Since we were working with docked complexes, we focus on the AlphaFold IpTM metric, which consists on the pTM score applied to the interface residues for assessing the topological similarity of protein structures, rather than the plDDT metric, which would be more adequate to monomer studies. For the Rosetta metrics, our focus is on the SC, BUns and $\Delta\Delta G$ metrics.

SC is a metric that represent the structural fit between a binder and a target. By using it, we were able to select only the monobodies whose structure shape better complements the one of the target. Values range from 0 to 1 - the higher the value, the better the fitting is.

BUns, is the number of hydrogen atoms buried in the interface region that do not establish any inter-chain hydrogen bond. It reflects the wasted potential of forming more interactions to stabilize the interface. The closer the value is to 0, the less wasted interaction potential exists.

$\Delta\Delta G$ is the Rosetta's predicted binding free energy of the complex, expressed in Rosetta Energy Units. The lower the energy, the more favorable the binding is.

By the end of this filtering stage, only 97 (0.8%) out of the 12,500 monobody predicted structures cleared all cutoff values for the first and more conservative sequence refinement approach (6 mutations per loop) and only 9 (0.07%) out of the 12,500 monobody predicted structures cleared all cutoff values for the second and less conservative sequence refinement approach (full loop redesign). The immediate conclusion that may be drawn from these results is the fact that, in those cases when we are designing proteins to dock in a desired, specific way to the target, knowing the protein's interface region, the best way to address the optimization of the interface seems to be with a more conservative approach. The reason behind this may be the amount of residue substitutions that we are applying - by applying more mutations, some of the interactions established in the docking process ended up being worsen (or even destroyed) rather than improved.

However, when we do not know exactly how and where our protein should interact and bind to the target - for instance in cases when we are dealing with a newly discovered viral protein, different from other better known viral proteins - it is not feasible to "force" the docking to occur in a specific way and between specific residues. In those cases, the most promising approach is to start from the predicted structure of the docked complex.

3.2.6 Alternative Monobody Identification Strategy Using AlphaFold

Rather than using a docking tool to dock each monobody structure to the target, we also considered to "dock" the monobodies to the target taking advantage of the AlphaFold multimer structure prediction capabilities using ColabFold [90]. By doing that, we are not docking the monobody structure with the RBD *per se*, but rather predicting the structure of

the complex monobody:RBD directly. Therefore, we do not choose the interface residues of our interest and let it be completely predicted instead.

For the case of multimer structure prediction, the filtering was done according to the values as presented in Table 2.4.

Given that we were predicting complex structures rather than monomeric structures (a monobody alone), we applied less strict cutoff values of pAE and pTM without compromising on the plDDT confidence level. Since we intended to further redesign and optimize the interface with ProteinMPNN [83], IpTM values were not considered for filtering in this stage despite being calculated by ColabFold.

By the end of the multimer structure prediction filtering stage, only 6 predicted structures out of the 9,369 moved to the sequence refinement step. Two examples are depicted in Figure 3.11. As expected, the number of structures that cleared the cutoff values was lower than the 2,898 monobody predicted structures in the monomeric structure prediction step, even having established less strict cutoff values. That was because of the increasing number of degrees of uncertainty when predicting complex structures, such as the existence of an interface and/or the larger number of residues to consider.

One major detail that was noted was the fact that the 6 structures that passed the filtering stage were around 10 residues in length shorter than the average 90-residue-long monobody sequence. This occurrence resulted in the selection of monobody-like structures with a 6 β-sandwich fold instead of the typical secondary structure composed of 7 β-strands. Therefore, the missing G β-strand is the cause of the absence of the FG loop in these monobody-like proteins. Although these proteins presented this characteristic, we hypothesized that the fact that these monobody-like proteins being the ones that cleared the plDDT, pAE and pTM filtering stage could mean that they may exhibit some increased target-binding potential, and decided to move forward with them.

From this point onward, these monobody-like structures will be simply referred as monobodies, for the sake of this dissertation.

According to the ColabFold multimer predictions, the interaction between the monobodies and the RBD is not established through the monobodies' loops as noted in the example in Figure 3.11. Therefore, we decided not to restrain the mutations to only occur in these regions of the monobody, in opposition of our previous ProteinMPNN [90] applications in the section 3.2.4. However, to maintain consistency for further approach comparison, each one of the 6 monobody sequences was also set to generate 500 new ones, which also went under ProteinMPNN fast relax. This generated a total of 3 000 new monobodies to proceed to the final filtering step.

Although the starting point was not the same, we have also been able to generate monobodies with different sequences, different structural conformations and different binding interactions with the target until this point in this approach. Therefore, it was also necessary to select which might be the best candidates to proceed to further wet lab validation. For this selection, we decided to use the same metrics and cutoff values as before (Table 2.3), making the results comparable. This time, 1,132 (37.7%) out of the

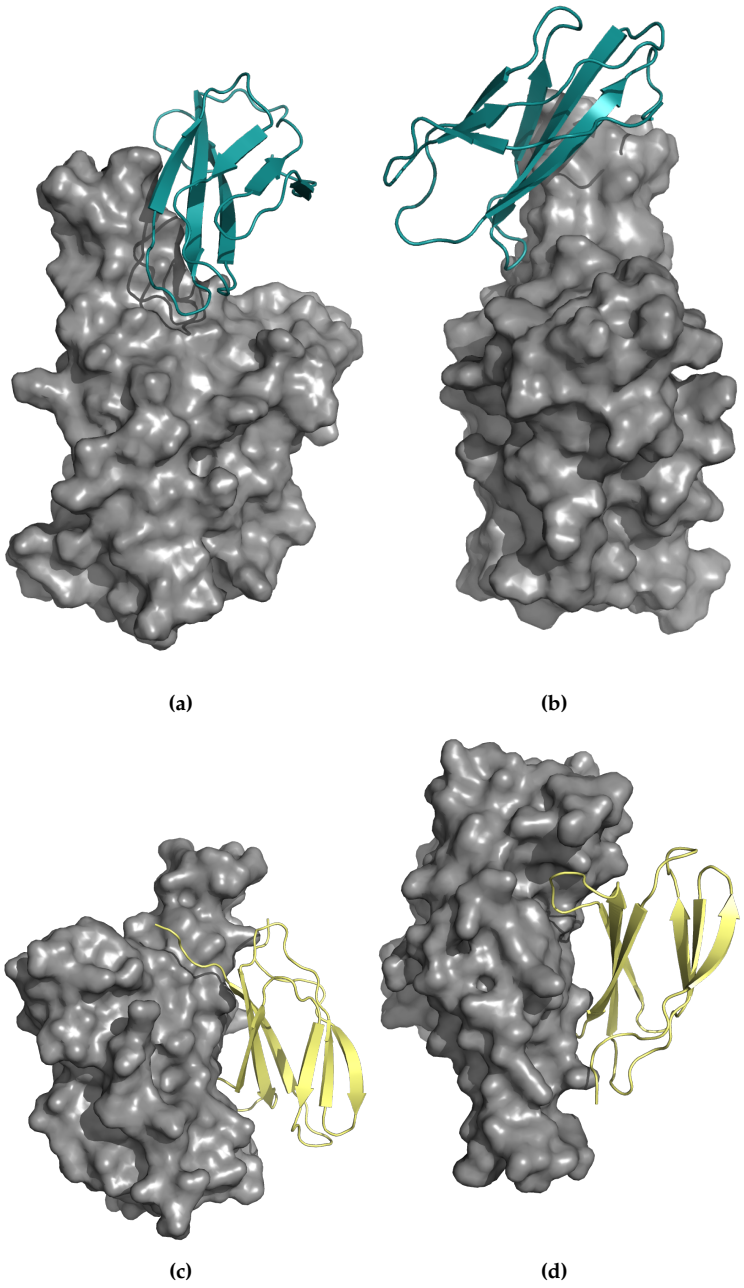


Figure 3.11: Examples of two docked complexes with ColabFold that passed the multimer structure prediction filtering stage. In blue and yellow are represented the monobody structures; in gray is represented the WT SARS-CoV-2 RBD. Subfigures **(a)** and **(b)** represent the same complex visualized from different angles; subfigures **(c)** and **(d)** represent the same complex visualized from different angles.

3,000 monobodies cleared all of the cutoff values. Immediately, one can conclude that the starting point makes a big difference. Despite the predicted binding region of the viral protein is the same as the one considered in the previous approach (except for a specific case, depicted in Figures 3.11c and 3.11d), the spatial orientation of the monobody is completely different, and so are the monobody residues that compose the interface region.

3.2.7 Select and Test the Best Approach

Altogether, unless one is very sure about exactly how and where the monobody should interact and bind to the viral target, the best approach to tackle the problem of target-specific monobody selection is to start from the predicted structure of the docked complex and allow the complete sequence design for a better sequence optimization (Tables 3.1, 3.2 and Figures 3.12, 3.13). Combining this approach with the initial selection of monobodies with only 6 β-strands resulted in the increased number of proteins that cleared all the filters.

Looking at Tables 3.1 and 3.2 and Figures 3.12 and 3.13, one may conclude that monobodies with 6 β-strands seem to be able to bind better to the target, and their binding interaction is improved more effectively by applying sequence optimization allowing mutations on the entire monobody sequence.

The overall performance of the ColabFold approach highlights the advantage of complex structure prediction coupled with full sequence re-design in enhancing the binding characteristics of the monobodies. ColabFold displays the highest IpTM and SC values, indicating better structural predictions and interface complementarity, which suggests improved binding. MaSIF approaches, on the other hand, show lower IpTM and average SC values, indicating less effective binding improvement compared to the ColabFold approach. In terms of BUns and $\Delta\Delta G$, ColabFold approach also outperforms MaSIF aproaches all around, indicating more stable protein-ligand interactions from an energy perspective. This demonstrates that while ColabFold excels in both structural complementarity and energetic stability, MaSIF falls behind in these crucial metrics for effective

Table 3.1: Comparison of AlphaFold and Rosetta metrics averages between approaches for all tested monobodies.

Approach	Number of monobodies	Average IpTM	Average SC	Average BUns	Average $\Delta\Delta G$
MaSIF + 6 loop mutations	12500	0.4685	0.6791	3.8217	-19.5984
MaSIF + full loop re-design	12500	0.444	0.6966	2.7714	-19.0977
ColabFold + full sequence re-design	3000	0.8052	0.6749	2.8749	-22.8753

Table 3.2: Comparison of AlphaFold and Rosetta metrics averages between approaches only for monobodies that passed the filtering stage.

Approach	Number of monobodies	Average IpTM	Average SC	Average BUns	Average $\Delta\Delta G$
MaSIF + 6 loop mutations	97	0.7721	0.6931	3.122	-23.4938
MaSIF + full loop re-design	9	0.7663	0.703	2.237	-24.4212
ColabFold + full sequence re-design	1132	0.8273	0.7166	1.629	-26.0076

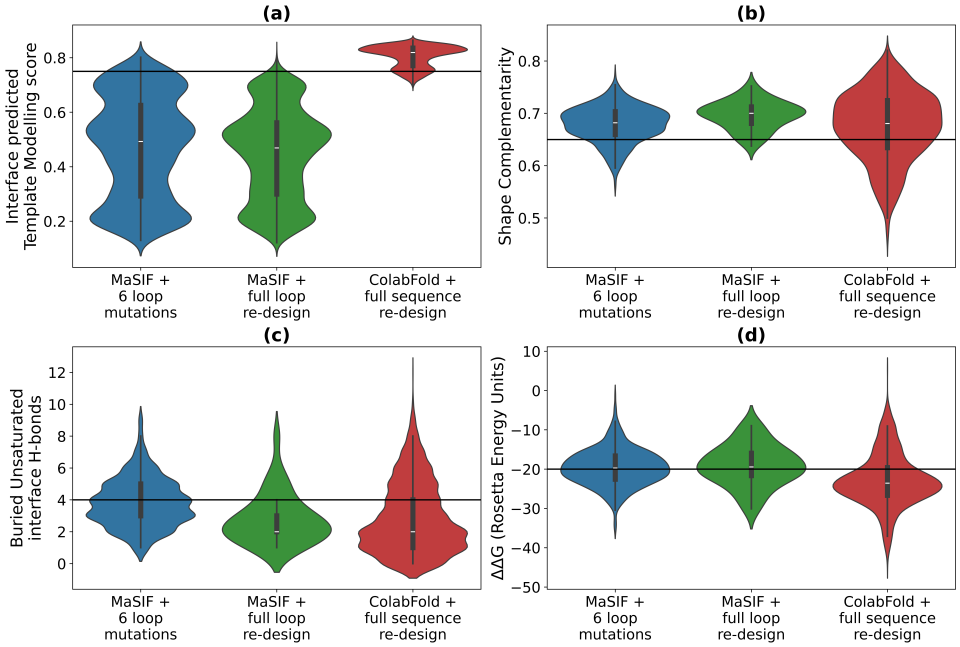


Figure 3.12: Comparison of AlphaFold and Rosetta metrics distributions between approaches for all tested monobodies. **(a)**: Violin distribution of the IpTM scores for each approach. **(b)**: Violin distribution of the SC scores for each approach. **(c)**: Violin distribution of the BUns scores for each approach. **(d)**: Violin distribution of the ΔΔG scores for each approach. The plots of each approach are colored equally in the different graphs for better visual perception. The horizontal lines mark the threshold values for each metric, detailed in Table 2.3.

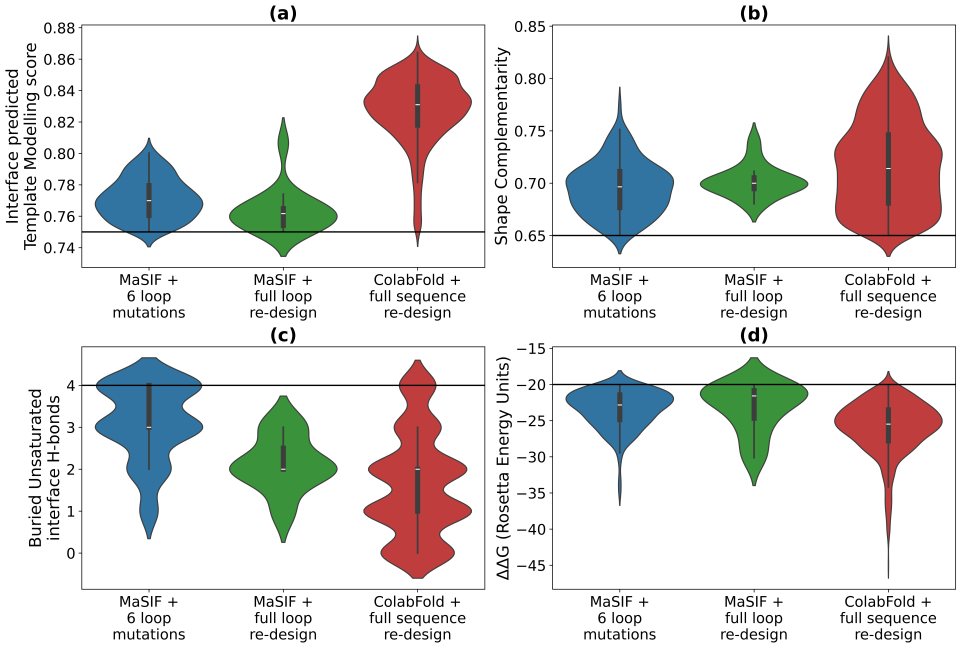


Figure 3.13: Comparison of AlphaFold and Rosetta metrics distributions between approaches only for monobodies that passed the filtering stage. **(a)**: Violin distribution of the IpTM scores for each approach. **(b)**: Violin distribution of the SC scores for each approach. **(c)**: Violin distribution of the BUns scores for each approach. **(d)**: Violin distribution of the ΔΔG scores for each approach. The plots of each approach are colored equally in the different graphs for better visual perception. The horizontal lines mark the threshold values for each metric, detailed in Table 2.3.

binding.

However, wet lab validation is lacking to confirm this hypothesis. One major problem that may arise from using monobodies with 6 β-strands as antivirals is the immunogenicity, as may happen with antibodies [95]. Since these monobodies differ in some extent from the human FN3, there exists the possibility of these proteins being recognized as non-human and therefore trigger the immune system when presented as an antiviral. A strategy to cope with this potential problem would be to "humanize" these monobodies, making them to be recognized as a human protein instead of as an antigen.

3.2.7.1 Molecular Dynamics Simulations of the Best Performing Complexes

To analyze the dynamics of the best-performing complex for each interface metric, MD simulations were conducted. By simulating the dynamics of these complexes, we gain a more comprehensive understanding of their behavior under physiologically relevant conditions, allowing us to observe the flexibility and adaptability of the monobody at the interface. This dynamic assessment is critical to capturing how well the monobody can maintain binding under solvent effects and other molecular interactions. Additionally, MD simulations help uncover any structural rearrangements or conformational changes that may either enhance or destabilize the interaction, which are not immediately apparent in the static models generated by the design pipeline.

The RMSD plots (Figure 3.14) reveal some key insights into the stability of each complex. For the best IpTM metric, the complex RMSD (Figure 3.14a) remains remarkably stable across all replicates, with minimal deviations observed (within 2 Å). However, the monobody RMSD (Figure 3.14b) shows greater variability, particularly in replicate 1, suggesting possible local flexibility within the monobody. Interestingly, the RBD RMSD (Figure 3.14c) remains low, indicating that the receptor's structure is stable and does not contribute significantly to the variability seen in the monobody.

For the best SC metric, while the complex RMSD (Figure 3.14d) remains consistently low, the monobody (Figure 3.14e) and RBD (Figure 3.14f) RMSD display slightly increased variability, especially towards the end of the simulation. This could be indicative of some adaptive movement or induced fit behavior, where both the monobody and RBD slightly adjust to optimize binding.

In the case of the best BUns metric, the overall complex RMSD (Figure 3.14g) suggests strong stability across replicates, but like the other metrics, monobody RMSD (Figure 3.14h) exhibits more variation. The RBD RMSD (Figure 3.14i) remains relatively stable, reinforcing that the interaction is largely driven by the monobody's adaptability without significant perturbations to the RBD.

Notably, the best $\Delta\Delta$G metric shows a unique pattern. The complex RMSD (Figure 3.14j) initially exhibits a spike in replicate 1 around 0.4 μs, followed by stabilization, suggesting a temporary rearrangement. The monobody RMSD (Figure 3.14k) here presents the highest degree of variation across all metrics, which could be linked to the fact that $\Delta\Delta$G

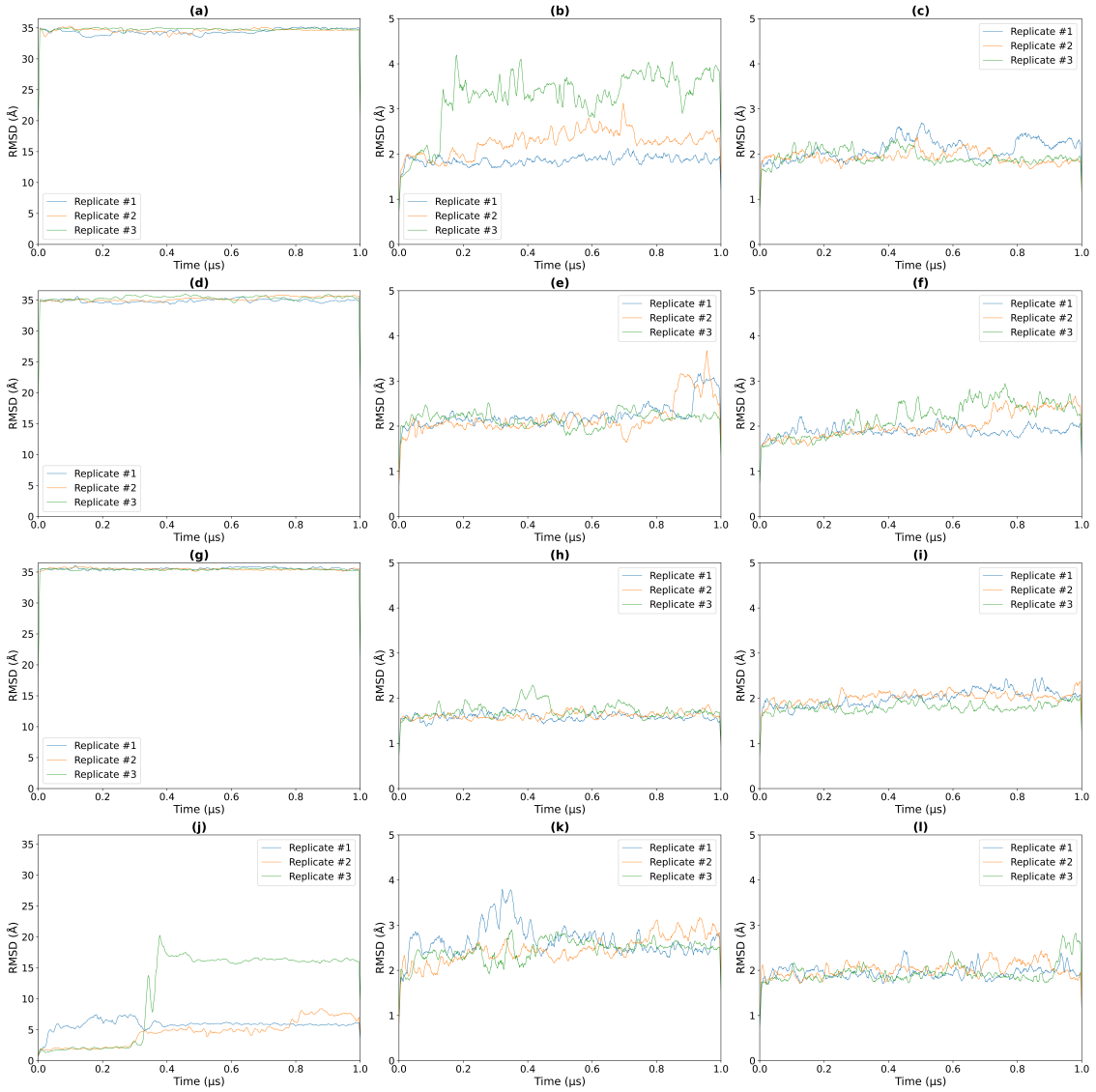


Figure 3.14: Root Mean Square Deviations for the MD simulations of the monobodies and RBD, individually and in complex, with best IpTM, SC, BUns and ΔΔG. The three columns of plots correspond to the RMSD of the complex (**(a)**, **(d)**, **(g)**, **(j)**), the monobody (**(b)**, **(e)**, **(h)**, **(k)**) and the RBD (**(c)**, **(f)**, **(i)**, **(l)**); the four rows of plots correspond to the RMSD of the top scoring complexes, monobodies and RBDs for IpTM (**(a)**, **(b)**, **(c)**, respectively), SC (**(d)**, **(e)**, **(f)**, respectively), BUns (**(g)**, **(h)**, **(i)**, respectively) and ΔΔG (**(j)**, **(k)**, **(l)**, respectively) metrics, respectively. On the X axis is depicted the simulation time in μs; on the Y axis is depicted the RMSD in Å.

accounts for energy differences and may introduce more flexibility in the monobody's binding mode. Despite these variations, the RBD RMSD (Figure 3.14l) remains steady, highlighting that these rearrangements are likely confined to the monobody.

Overall, these simulations suggest that while the monobody is flexible, the interactions with the RBD are consistently maintained, reflecting a robust binding interface. The variations observed in the monobody, particularly in metrics such as ΔΔG, may contribute to enhanced adaptability, potentially allowing the monobody to better accommodate the RBD's surface features over time. This stability and adaptability are promising for

therapeutic or diagnostic applications, demonstrating resilience in maintaining target engagement.

3.2.8 Compare Monobody Dataset Performance

Building on the findings from the section 3.2.7, where multiple protein design approaches were compared, it was established that the most effective strategy for target-specific monobody selection is to begin with the predicted structure of the docked complex and allow for complete sequence re-design to optimize the resulting monobody sequences. However, one critical question may arise regarding the starting pool of monobody sequences on which this established approach is applied: Does the application of the MSA Transformer to BLAST and FoldSeek-derived datasets further enhance the performance of monobody selection for target-specific antiviral discovery, compared to using these sequence search methods alone? This question seeks to determine whether integrating the MSA Transformer can provide additional benefits in refining monobody sequences.

While BLAST and FoldSeek are instrumental in identifying homologous sequences and structural analogs, they may fall short in capturing the intricate evolutionary patterns that are crucial for optimizing monobody sequences. The MSA Transformer, with its ability to model these evolutionary contexts through deep learning, could potentially uncover hidden functional motifs and sequence correlations that enhance the monobody's binding affinity and specificity.

To test that, the best approach was also applied to the complexes containing the 3,050 monobody sequences obtained from the sequence search step using the same settings and filtering cutoff values as before (Tables 2.3 and 2.4).

Looking at Tables 3.3 and 3.4, we can notice that the average monobody generated using the MSA Transformer dataset performs better than the average monobody generated using the sequence search dataset and that the number of monobodies that cleared all the interface metric filtering is more than 8 times greater for the MSA Transformer dataset. Moreover, the violin distributions in Figures 3.15 and 3.16 allow us to conclude that using MSA Transformer to increase the size of the initial dataset for the protein design pipeline further enhances the performance of monobody selection for target-specific antiviral discovery, by generating not only monobodies that have better interface metrics performance, when further re-designed, but also a greater number of monobodies that clear the interface filters.

Table 3.3: Comparison of AlphaFold and Rosetta metrics averages between datasets for all tested monobodies.

Dataset	Number of monobodies	Average IpTM	Average SC	Average BUns	Average $\Delta\Delta G$
MSA Transformer	3000	0.8052	0.6749	2.8749	-22.8753
Sequence Search	2500	0.7982	0.5801	2.8929	-18.7715

Table 3.4: Comparison of AlphaFold and Rosetta metrics averages between datasets only for monobodies that passed the filtering stage.

Dataset	Number of monobodies	Average IpTM	Average SC	Average BUns	Average ΔΔG
MSA Transformer	1132	0.8273	0.7166	1.629	-26.0076
Sequence Search	134	0.8259	0.6737	2.8731	-24.9678

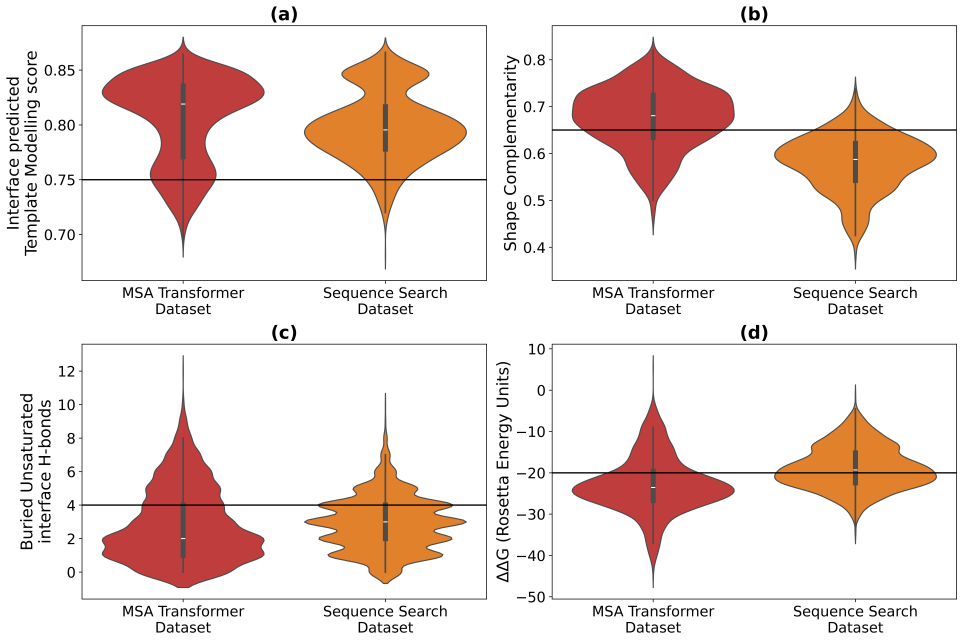


Figure 3.15: Comparison of AlphaFold and Rosetta metrics distributions between datasets for all tested monobodies. **(a)**: Violin distribution of the IpTM scores for each dataset. **(b)**: Violin distribution of the SC scores for each dataset. **(c)**: Violin distribution of the BUns scores for each dataset. **(d)**: Violin distribution of the ΔΔG scores for each dataset. The plots of each dataset are colored equally in the different graphs for better visual perception. The horizontal lines mark the threshold values for each metric, detailed in Table 2.3.

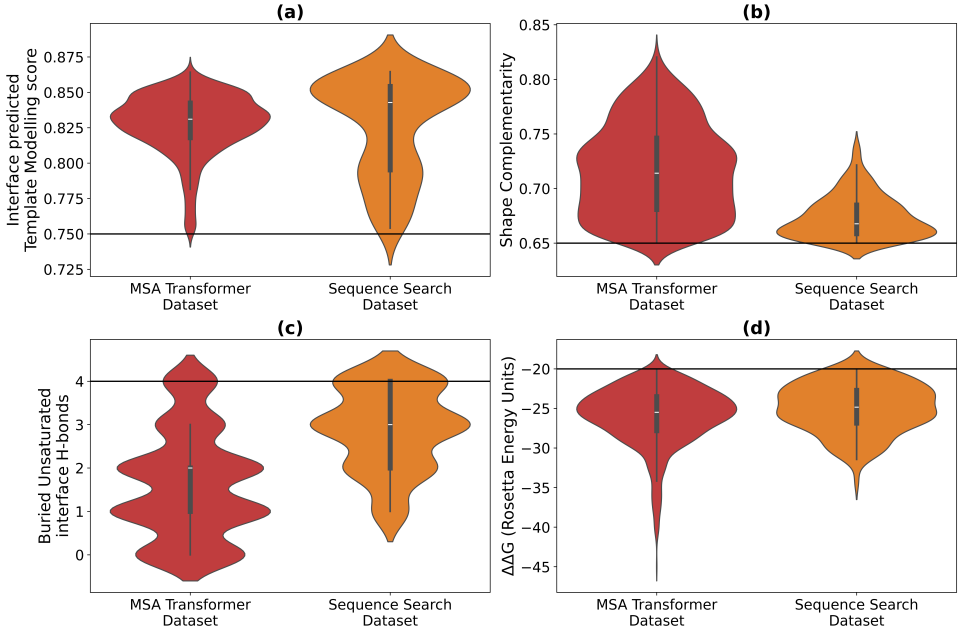


Figure 3.16: Comparison of AlphaFold and Rosetta metrics distributions between datasets only for monobodies that passed the filtering stage. **(a)**: Violin distribution of the IpTM scores for each dataset. **(b)**: Violin distribution of the SC scores for each dataset. **(c)**: Violin distribution of the BUns scores for each dataset. **(d)**: Violin distribution of the $\Delta\Delta$G scores for each dataset. The plots of each dataset are colored equally in the different graphs for better visual perception. The horizontal lines mark the threshold values for each metric, detailed in Table 2.3.

4

Conclusion

Since the early days of mankind, several viruses emerged and caused major pandemics, costing the lives of several people. The now well-known SARS-CoV-2, was not an exception. Literature shows that SARS-CoV-2 is one of the most adaptive viruses that exist, which is reflected on the continuous appearance of variants, each being more capable and fit than the previous. Having appeared almost overnight and having become responsible for millions of deaths - the majority of them in its early months and years – the COVID-19 outbreak made clear that humanity is not prepared for a future pandemic, and the next one may be right around the corner.

Current antiviral solutions have their advantages and can be very effective once adapted, or discovered, and used to combat a given virus. However, their research and production costs are very high, and, among other disadvantages, the time it takes to go from the appearance of the virus to a good antiviral solution is still far from ideal. Monobodies come as an alternative that can be used to overcome the limitations of the current antiviral solutions. Exhibiting robust stability, solubility and the possibility of being expressed in bacteria, monobodies present themselves as ideal candidates for diverse research and therapeutic endeavors. Also, monobodies exhibit the possibility of being designed and adapted for an increased target-binding specificity. Through the design of monobodies, virtually any viral protein of interest can be effectivelly targeted. By coupling the concept of monobody design with computational tools, monobodies can be selected, adapted, optimized and screened in a much larger scale than through experimental techniques.

The main downside is the fact that there are too few monobodies already described in the literature, which limits design approaches that are based on an existing scaffold library, such as the one that will be explored in the EvaMobs project.

Therefore, the primary aim of this thesis was to create a computational framework that streamlines the development of this class of antivirals through the creation of a monobody library, by synergizing computational structural biology with protein design tools, software suites and machine learning techniques. More specifically, we posed the hypothesis that language models can be used to create a large and diverse monobody library, that is adequate to serve as a starting point for protein design approaches.

Using MSA Transformer, we were able to generate a pool of over 1.1 million monobody-like sequences that proved to be a great starting point for the following protein design framework.

Different strategies for docking the generated monobodies with the SARS-CoV-2 RBD as proof of concept were tested. MaSIF was used to establish the docking between wanted specific residues, whereas ColabFold was used to completely predict the interacting residues. After further sequence optimization with ProteinMPNN and filtering with AlphaFold and Rosetta, results showed that the best approach to tackle the problem of target-specific monobody selection is to start from the predicted structure of the docked complex and that the monobody binding capability is improved more effectively by applying sequence optimization allowing mutations on the entire monobody sequence. Interestingly, according to our results, it seems that monobodies with 6 β-strands can potentially bind better to the target than monobodies with the more common 7 β-sandwich fold.

Answering the posed hypothesis, with our results we could conclude that language models can be used to create a large and diverse monobody library, that is adequate to serve as a starting point for protein design approaches. More specifically, using MSA Transformer to increase the size of the initial dataset for the protein design pipeline further enhanced the performance of monobody selection for target-specific antiviral discovery, by generating not only monobodies that have better interface metrics performance, when further re-designed, but also a greater number of monobodies that clear the interface filters.

In parallel, MD simulations of the SARS-CoV-2 BA.2.86 and JN.1 variants' RBDs were performed to further understand the dynamics of these proteins, revealing significant flexibility, particularly in the loops of the RBM. Subsequent PCA was conducted to compare our results with existing literature on other variants and to identify the most predominant conformations of these variants' RBDs, which is essential for applying the framework developed in this thesis to future work. The PCA results revealed unique shifts in the distribution of open and closed states in these variants, providing insights into how mutations affect viral protein flexibility and dynamics. The observed decrease in the percentage of open conformations, coupled with the lower free energy of these states, suggests that these variants might exhibit reduced receptor binding availability, potentially impacting viral infectivity and immune evasion. This observation highlights the critical influence of mutations on the structural behavior of the viral proteins, reinforcing the importance of monitoring these conformational shifts as a key element of pandemic preparedness.

Crucially, the insights from the PCA can inform monobody design strategies by providing a deeper understanding of the variant-specific structural dynamics that affect receptor accessibility. Combining these structural dynamics insights with the monobody library generated by language models offers an adaptive antiviral strategy capable of quickly responding to new viral variants. This integrated framework allows for the rapid

development of monobodies that can be specifically tailored to different viral protein conformations, enhancing their efficacy in targeting viral variants. As a result, the computational pipeline established in this work paves the way for an adaptable and scalable approach to antiviral discovery, where monobody design can evolve in tandem with the emergence of new viral threats, ensuring a quicker and more targeted response to future pandemics.

Bibliography

[1] J. M. Lourenço. *The NOVAthesis LaTeX Template User's Manual*. NOVA University Lisbon. 2021. URL: https://github.com/joaomlourenco/novathesis/raw/main/template.pdf (cit. on p. i).

[2] J. Horgan. "Antonine plague". In: *Ancient History Encyclopedia* (2019). URL: https://www.worldhistory.org/Antonine_Plague/ (visited on 2024-07-20) (cit. on p. 1).

[3] A. Aassve et al. "Epidemics and trust: The case of the Spanish Flu". In: *Health economics* 30.4 (2021), pp. 840–857. DOI: 10.1002/hec.4218 (cit. on p. 1).

[4] S. Sampath et al. "Pandemics throughout the history". In: *Cureus* 13.9 (2021). DOI: 10.7759/cureus.18136 (cit. on p. 1).

[5] P. Bhadoria, G. Gupta, and A. Agarwal. "Viral pandemics in the past two decades: an overview". In: *Journal of family medicine and primary care* 10.8 (2021), pp. 2745–2750. DOI: 10.4103/jfmpc.jfmpc_2071_20 (cit. on p. 1).

[6] X.-Y. Ge et al. "Isolation and characterization of a bat SARS-like coronavirus that uses the ACE2 receptor". In: *Nature* 503.7477 (2013), pp. 535–538. DOI: 10.1038/nature12711 (cit. on p. 2).

[7] N. Zhu et al. "A novel coronavirus from patients with pneumonia in China, 2019". In: *New England journal of medicine* 382.8 (2020), pp. 727–733. DOI: 10.1056/NEJMoa2001017 (cit. on p. 2).

[8] B. Hu et al. "Characteristics of SARS-CoV-2 and COVID-19". In: *Nature reviews microbiology* 19.3 (2021), pp. 141–154. DOI: 10.1038/s41579-020-00459-7 (cit. on pp. 2–4).

[9] M. T. Adil et al. "SARS-CoV-2 and the pandemic of COVID-19". In: *Postgraduate medical journal* 97.1144 (2021), pp. 110–116. DOI: 10.1136/postgradmedj-2020-138386 (cit. on pp. 2–4).

[10] M. Hoffmann et al. "SARS-CoV-2 cell entry depends on ACE2 and TMPRSS2 and is blocked by a clinically proven protease inhibitor". In: *cell* 181.2 (2020), pp. 271–280. DOI: 10.1016/j.cell.2020.02.052 (cit. on p. 3).

[11] C. Huang et al. "Clinical features of patients infected with 2019 novel coronavirus in Wuhan, China". In: *The lancet* 395.10223 (2020), pp. 497–506. DOI: 10.1016/S0140-6736(20)30183-5 (cit. on pp. 3, 4).

[12] W.-j. Guan et al. "Clinical characteristics of coronavirus disease 2019 in China". In: *New England journal of medicine* 382.18 (2020), pp. 1708–1720. DOI: 10.1056/NEJMoa2002032 (cit. on p. 4).

[13] W. H. Organization. *WHO COVID-19 dashboard - Cases*. URL: https://data.who.int/dashboards/covid19/cases?n=c (visited on 2024-09-27) (cit. on p. 4).

[14] W. H. Organization. *WHO COVID-19 dashboard - Deaths*. URL: https://data.who.int/dashboards/covid19/deaths?n=c (visited on 2024-09-27) (cit. on p. 4).

[15] H. Brüssow. "COVID-19: emergence and mutational diversification of SARS-CoV-2". In: *Microbial Biotechnology* 14.3 (2021), pp. 756–768. DOI: 10.1111/1751-7915.13800 (cit. on p. 4).

[16] E. Volz et al. "Assessing transmissibility of SARS-CoV-2 lineage B. 1.1. 7 in England". In: *Nature* 593.7858 (2021), pp. 266–269. DOI: 10.1038/s41586-021-03470-x (cit. on p. 5).

[17] F. Tian et al. "N501Y mutation of spike protein in SARS-CoV-2 strengthens its binding to receptor ACE2". In: *elife* 10 (2021), e69091. DOI: 10.7554/eLife.69091 (cit. on p. 5).

[18] D. A. Collier et al. "Sensitivity of SARS-CoV-2 B. 1.1. 7 to mRNA vaccine-elicited antibodies". In: *Nature* 593.7857 (2021), pp. 136–141. DOI: 10.1038/s41586-021-03412-7 (cit. on p. 5).

[19] P. Supasa et al. "Reduced neutralization of SARS-CoV-2 B. 1.1. 7 variant by convalescent and vaccine sera". In: *Cell* 184.8 (2021), pp. 2201–2211. DOI: 10.1016/j.cell.2021.02.033 (cit. on p. 5).

[20] X. Xie et al. "Neutralization of SARS-CoV-2 spike 69/70 deletion, E484K and N501Y variants by BNT162b2 vaccine-elicited sera". In: *Nature medicine* 27.4 (2021), pp. 620–621. DOI: 10.1038/s41591-021-01270-4 (cit. on p. 6).

[21] Y. Liu et al. "Neutralizing activity of BNT162b2-elicited serum". In: *New England Journal of Medicine* 384.15 (2021), pp. 1466–1468. DOI: 10.1056/NEJMc2102017 (cit. on p. 6).

[22] Y. Chen et al. "Serum neutralising activity against SARS-CoV-2 variants elicited by CoronaVac". In: *The Lancet. Infectious Diseases* 21.8 (2021), p. 1071. DOI: 10.1016/S1473-3099(21)00287-5 (cit. on p. 6).

[23] L. J. Abu-Raddad, H. Chemaitelly, and A. A. Butt. "Effectiveness of the BNT162b2 Covid-19 Vaccine against the B. 1.1. 7 and B. 1.351 Variants". In: *New England Journal of Medicine* 385.2 (2021), pp. 187–189. DOI: 10.1056/NEJMc2104974 (cit. on pp. 6, 7).

[24] W. M. de Souza et al. "Clusters of SARS-CoV-2 lineage B. 1.1. 7 infection after vaccination with adenovirus-vectored and inactivated vaccines". In: *Viruses* 13.11 (2021), p. 2127. DOI: 10.3390/v13112127 (cit. on p. 6).

[25] K. R. Emary et al. "Efficacy of ChAdOx1 nCoV-19 (AZD1222) vaccine against SARS-CoV-2 variant of concern 202012/01 (B. 1.1. 7): an exploratory analysis of a randomised controlled trial". In: *The Lancet* 397.10282 (2021), pp. 1351–1362. DOI: 10.1016/S0140-6736(21)00628-0 (cit. on p. 6).

[26] M. Alenquer et al. "Signatures in SARS-CoV-2 spike protein conferring escape to neutralizing antibodies". In: *PLoS pathogens* 17.8 (2021), e1009772. DOI: 10.1371/journal.ppat.1009772 (cit. on p. 7).

[27] H. Tegally et al. "Detection of a SARS-CoV-2 variant of concern in South Africa". In: *Nature* 592.7854 (2021), pp. 438–443. DOI: 10.1038/s41586-021-03402-9 (cit. on p. 7).

[28] S. Cele et al. "Escape of SARS-CoV-2 501Y. V2 from neutralization by convalescent plasma". In: *Nature* 593.7857 (2021), pp. 142–146. DOI: 10.1038/s41586-021-03471-w (cit. on p. 7).

[29] S. A. Madhi et al. "Efficacy of the ChAdOx1 nCoV-19 Covid-19 vaccine against the B. 1.351 variant". In: *New England Journal of Medicine* 384.20 (2021), pp. 1885–1898. DOI: 10.1056/NEJMoa2102214 (cit. on p. 7).

[30] J. Sadoff et al. "Safety and efficacy of single-dose Ad26. COV2. S vaccine against Covid-19". In: *New England Journal of Medicine* 384.23 (2021), pp. 2187–2201. DOI: 10.1056/NEJMoa2101544 (cit. on p. 7).

[31] T. Kustin et al. "Evidence for increased breakthrough rates of SARS-CoV-2 variants of concern in BNT162b2-mRNA-vaccinated individuals". In: *Nature medicine* 27.8 (2021), pp. 1379–1384. DOI: 10.1038/s41591-021-01413-7 (cit. on p. 7).

[32] D. Planas et al. "Reduced sensitivity of SARS-CoV-2 variant Delta to antibody neutralization". In: *Nature* 596.7871 (2021), pp. 276–280. DOI: 10.1038/s41586-021-03777-9 (cit. on p. 7).

[33] M. Li, F. Lou, and H. Fan. "SARS-CoV-2 Variants of Concern Delta: a great challenge to prevention and control of COVID-19". In: *Signal Transduction and Targeted Therapy* 6.1 (2021), p. 349. DOI: 10.1038/s41392-021-00767-1 (cit. on p. 7).

[34] P. Mlcochova et al. "SARS-CoV-2 B. 1.617. 2 Delta variant replication and immune evasion". In: *Nature* 599.7883 (2021), pp. 114–119. DOI: 10.1038/s41586-021-03944-y (cit. on p. 7).

[35] J. Lopez Bernal et al. "Effectiveness of Covid-19 vaccines against the B. 1.617. 2 (Delta) variant". In: *New England Journal of Medicine* 385.7 (2021), pp. 585–594. DOI: 10.1056/NEJMoa2108891 (cit. on p. 7).

[36] C. Ma et al. "Effectiveness of adenovirus type 5 vectored and inactivated COVID-19 vaccines against symptomatic COVID-19, COVID-19 pneumonia, and severe COVID-19 caused by the B. 1.617. 2 (Delta) variant: evidence from an outbreak in Yunnan, China, 2021". In: *Vaccine* 40.20 (2022), pp. 2869–2874. DOI: 10.1016/j.vaccine.2022.03.067 (cit. on p. 7).

[37] A. Pormohammad et al. "Effectiveness of COVID-19 vaccines against Delta (B. 1.617. 2) variant: a systematic review and meta-analysis of clinical studies". In: *Vaccines* 10.1 (2021), p. 23. DOI: 10.3390/vaccines10010023 (cit. on p. 8).

[38] A. J. Greaney et al. "Complete mapping of mutations to the SARS-CoV-2 spike receptor-binding domain that escape antibody recognition". In: *Cell host & microbe* 29.1 (2021), pp. 44–57. DOI: 10.1016/j.chom.2020.11.007 (cit. on p. 8).

[39] W. F. Garcia-Beltran et al. "Multiple SARS-CoV-2 variants escape neutralization by vaccine-induced humoral immunity". In: *Cell* 184.9 (2021), pp. 2372–2383. DOI: 10.1016/j.cell.2021.03.013 (cit. on p. 8).

[40] R. Viana et al. "Rapid epidemic expansion of the SARS-CoV-2 Omicron variant in southern Africa". In: *Nature* 603.7902 (2022), pp. 679–686. DOI: 10.1038/s41586-022-04411-y (cit. on p. 8).

[41] W. T. Harvey et al. "SARS-CoV-2 variants, spike mutations and immune escape". In: *Nature Reviews Microbiology* 19.7 (2021), pp. 409–424. DOI: 10.1038/s41579-021-00573-0 (cit. on p. 8).

[42] Y. Fan et al. "SARS-CoV-2 Omicron variant: recent progress and future perspectives". In: *Signal transduction and targeted therapy* 7.1 (2022), pp. 1–11. DOI: 10.1038/s41392-022-00997-x (cit. on p. 8).

[43] P. Qu et al. "Immune evasion, infectivity, and fusogenicity of SARS-CoV-2 BA. 2.86 and FLip variants". In: *Cell* 187.3 (2024), pp. 585–595. DOI: 10.1016/j.cell.2023.12.026 (cit. on p. 8).

[44] S. Yang et al. "Fast evolution of SARS-CoV-2 BA. 2.86 to JN. 1 under heavy immune pressure". In: *The Lancet Infectious Diseases* 24.2 (2024), e70–e72. DOI: 10.1016/S1473-3099(23)00744-2 (cit. on p. 8).

[45] W. F. Garcia-Beltran et al. "mRNA-based COVID-19 vaccine boosters induce neutralizing immunity against SARS-CoV-2 Omicron variant". In: *Cell* 185.3 (2022), pp. 457–466. DOI: 10.1016/j.cell.2021.12.033 (cit. on p. 8).

[46] G. R. Campos et al. "Booster dose of BNT162b2 after two doses of CoronaVac improves neutralization of SARS-CoV-2 Omicron variant". In: *Communications medicine* 2.1 (2022), p. 76. DOI: 10.1038/s43856-022-00141-4 (cit. on p. 8).

[47] N. Andrews et al. "Covid-19 vaccine effectiveness against the Omicron (B. 1.1. 529) variant". In: *New England Journal of Medicine* 386.16 (2022), pp. 1532–1546. DOI: 10.1056/NEJMoa2119451 (cit. on p. 8).

[48] H. Shuai et al. "Attenuated replication and pathogenicity of SARS-CoV-2 B. 1.1. 529 Omicron". In: *Nature* 603.7902 (2022), pp. 693–699. DOI: 10.1038/s41586-022-04442-5 (cit. on p. 8).

[49] M. Valério et al. "SARS-CoV-2 variants impact RBD conformational dynamics and ACE2 accessibility". In: *Frontiers in medical technology* 4 (2022), p. 1009451. DOI: 10.3389/fmedt.2022.1009451 (cit. on pp. 9, 10, 43, 45).

[50] P. Han et al. "Receptor binding and complex structures of human ACE2 to spike RBD from omicron and delta SARS-CoV-2". In: *Cell* 185.4 (2022), pp. 630–640. DOI: 10.1016/j.cell.2022.01.001 (cit. on p. 9).

[51] S. M.-C. Gobeil et al. "Effect of natural mutations of SARS-CoV-2 on spike structure, conformation, and antigenicity". In: *Science* 373.6555 (2021), eabi6226. DOI: 10.1126/science.abi6226 (cit. on p. 9).

[52] E. Socher et al. "Molecular dynamics simulations of the delta and omicron SARS-CoV-2 spike–ACE2 complexes reveal distinct changes between both variants". In: *Computational and structural biotechnology journal* 20 (2022), pp. 1168–1176. DOI: 10.1016/j.csbj.2022.02.015 (cit. on pp. 9, 10).

[53] E. C. for Disease Prevention and Control. *SARS-CoV-2 variants of concern as of 30 August 2024*. URL: https://www.ecdc.europa.eu/en/covid-19/variants-concern (visited on 2024-09-27) (cit. on p. 10).

[54] G. Pantaleo et al. "Antibodies to combat viral infections: development strategies and progress". In: *Nature Reviews Drug Discovery* 21.9 (2022), pp. 676–696. DOI: 10.1038/s41573-022-00495-3 (cit. on pp. 11, 12).

[55] K. L. Hopkins et al. "Community-based approaches to increase COVID-19 vaccine uptake and demand: lessons learned from four UNICEF-supported interventions". In: *Vaccines* 11.7 (2023), p. 1180. DOI: 10.3390/vaccines11071180 (cit. on p. 11).

[56] D. Ao et al. "Strategies for the development and approval of COVID-19 vaccines and therapeutics in the post-pandemic period". In: *Signal Transduction and Targeted Therapy* 8.1 (2023), p. 466. DOI: 10.1038/s41392-023-01724-w (cit. on p. 11).

[57] S.-C. Tsai et al. "Approaches towards fighting the COVID-19 pandemic". In: *International journal of molecular medicine* 47.1 (2021), pp. 3–22. DOI: 10.3892/ijmm.2020.4794 (cit. on p. 11).

[58] T. Kelesidis et al. "How to approach and treat viral infections in ICU patients". In: *BMC infectious diseases* 14 (2014), pp. 1–12. DOI: 10.1186/1471-2334-14-321 (cit. on p. 11).

[59] H. Marcotte and L. Hammarström. "Passive immunization: toward magic bullets". In: *Mucosal immunology*. Elsevier, 2015, pp. 1403–1434. DOI: 10.1016/B978-0-12-415847-4.00071-9 (cit. on p. 12).

[60] D. Wrapp et al. "Structural basis for potent neutralization of betacoronaviruses by single-domain camelid antibodies". In: *Cell* 181.5 (2020), pp. 1004–1015. DOI: 10.1016/j.cell.2020.04.031 (cit. on p. 12).

[61] A. Koide et al. "The fibronectin type III domain as a scaffold for novel binding proteins". In: *Journal of molecular biology* 284.4 (1998), pp. 1141–1151. DOI: 10.1006/jmbi.1998.2238 (cit. on pp. 13, 14, 53, 54).

[62] F. Sha et al. "Monobodies and other synthetic binding proteins for expanding protein science". In: *Protein Science* 26.5 (2017), pp. 910–924. DOI: 10.1002/pro.3148 (cit. on pp. 13, 14).

[63] *EvaMobs: Strengthening the EU's pandemic preparedness*. URL: https://evamobs.eu/ (visited on 2024-07-20) (cit. on p. 15).

[64] *EvaMobs: About our project*. URL: https://evamobs.eu/our-project/ (visited on 2024-07-20) (cit. on pp. 15, 16).

[65] J. L. Watson et al. "De novo design of protein structure and function with RFdiffusion". In: *Nature* 620.7976 (2023), pp. 1089–1100. DOI: 10.1038/s41586-023-06415-8 (cit. on p. 17).

[66] J. B. Ingraham et al. "Illuminating protein space with a programmable generative model". In: *Nature* 623.7989 (2023), pp. 1070–1078. DOI: 10.1038/s41586-023-06728-8 (cit. on p. 17).

[67] S. Alamdari et al. "Protein generation with evolutionary diffusion: sequence is all you need". In: *bioRxiv* (2023), pp. 2023–09. DOI: 10.1101/2023.09.11.556673 (cit. on p. 17).

[68] Y. Lin et al. "Out of Many, One: Designing and Scaffolding Proteins at the Scale of the Structural Universe with Genie 2". In: *arXiv preprint arXiv:2405.15489* (2024). DOI: 10.48550/arXiv.2405.15489 (cit. on p. 17).

[69] G. Schneider and D. E. Clark. "Automated de novo drug design: are we nearly there yet?" In: *Angewandte Chemie International Edition* 58.32 (2019), pp. 10792–10803. DOI: 10.1002/anie.201814681 (cit. on p. 17).

[70] J. Jumper et al. "Highly accurate protein structure prediction with AlphaFold". In: *Nature* 596.7873 (2021), pp. 583–589. DOI: 10.1038/s41586-021-03819-2 (cit. on p. 17).

[71] A. W. Senior et al. "Improved protein structure prediction using potentials from deep learning". In: *Nature* 577.7792 (2020), pp. 706–710. DOI: 10.1038/s41586-019-1923-7 (cit. on p. 17).

[72] B. Tang et al. "AI-aided design of novel targeted covalent inhibitors against SARS-CoV-2". In: *Biomolecules* 12.6 (2022), p. 746. DOI: 10.3390/biom12060746 (cit. on p. 17).

[73] T. Zhou et al. "Cryo-EM structures of SARS-CoV-2 spike without and with ACE2 reveal a pH-dependent switch to mediate endosomal positioning of receptor-binding domains". In: *Cell host & microbe* 28.6 (2020), pp. 867–879. DOI: 10.1016/j.chom.2020.11.004 (cit. on p. 17).

[74] S. Brogi. "Computational approaches for drug discovery". In: *Molecules* 24.17 (2019), p. 3061. DOI: doi.org/10.3390/molecules24173061 (cit. on p. 18).

[75] S. Patodia, A. Bagaria, and D. Chopra. "Molecular dynamics simulation of proteins: a brief overview". In: *Journal of Physical Chemistry & Biophysics* 4.6 (2014), p. 1. DOI: 10.4172/2161-0398.1000166 (cit. on p. 21).

[76] A. R. Leach. *Molecular modelling: principles and applications*. Ed. by P. education. Prentice Hall, 2001. ISBN: 9780582382107. URL: https://books.google.pt/books?id=kB7jsbV-uhkC (cit. on pp. 21, 22).

[77] L. Schrödinger and W. DeLano. *The PyMOL Molecular Graphics System*. Version 2.5.0. 2022-03-17. URL: http://www.pymol.org/pymol (cit. on pp. 28, 30).

[78] M. J. Abraham et al. "GROMACS: High performance molecular simulations through multi-level parallelism from laptops to supercomputers". In: *SoftwareX* 1 (2015), pp. 19–25. DOI: 10.1016/j.softx.2015.06.001 (cit. on pp. 28, 30).

[79] N. Michaud-Agrawal et al. "MDAnalysis: a toolkit for the analysis of molecular dynamics simulations". In: *Journal of computational chemistry* 32.10 (2011), pp. 2319–2327. DOI: 10.1002/jcc.21787 (cit. on pp. 30, 31).

[80] J. D. Hunter. "Matplotlib: A 2D graphics environment". In: *Computing in science & engineering* 9.03 (2007), pp. 90–95. DOI: 10.1109/MCSE.2007.55 (cit. on p. 31).

[81] S. R. Campos and A. M. Baptista. "Conformational analysis in a multidimensional energy landscape: study of an arginylglutamate repeat". In: *The Journal of Physical Chemistry B* 113.49 (2009), pp. 15989–16001. DOI: 10.1021/jp902991u (cit. on p. 31).

[82] R. Rao et al. "MSA Transformer". In: *bioRxiv* (2021). DOI: 10.1101/2021.02.12.430858 (cit. on pp. 32, 33, 48).

[83] J. Dauparas et al. "Robust deep learning–based protein sequence design using ProteinMPNN". In: *Science* 378.6615 (2022), pp. 49–56. DOI: 10.1126/science.add2187 (cit. on pp. 34, 35, 39, 41, 57).

[84] P Gainza et al. "Deciphering interaction fingerprints from protein molecular surfaces using geometric deep learning". In: *Nature Methods* 17.2 (2020), pp. 184–192. DOI: 10.1038/s41592-019-0666-6 (cit. on pp. 36, 39, 53, 55).

[85] S. F. Altschul et al. "Basic local alignment search tool". In: *Journal of molecular biology* 215.3 (1990), pp. 403–410. DOI: 10.1016/S0022-2836(05)80360-2 (cit. on p. 37).

[86] M. Van Kempen et al. "Fast and accurate protein structure search with Foldseek". In: *Nature Biotechnology* 42.2 (2024), pp. 243–246. DOI: 10.1038/s41587-023-01773-0 (cit. on p. 37).

[87] F. Sievers and D. G. Higgins. "Clustal Omega, accurate alignment of very large numbers of sequences". In: *Multiple sequence alignment methods* (2014), pp. 105–116. DOI: 10.1007/978-1-62703-646-7_6 (cit. on p. 37).

[88] D. Sgarbossa, U. Lupo, and A.-F. Bitbol. "Generative power of a protein language model trained on multiple sequence alignments". In: *Elife* 12 (2023), e79854. DOI: 10.7554/eLife.79854 (cit. on pp. 37, 38, 48, 49).

[89] M. Steinegger and J. Söding. "MMseqs2 enables sensitive protein sequence searching for the analysis of massive data sets". In: *Nature biotechnology* 35.11 (2017), pp. 1026–1028. DOI: 10.1038/nbt.3988 (cit. on pp. 38, 50).

[90] M. Mirdita et al. "ColabFold: making protein folding accessible to all". In: *Nature methods* 19.6 (2022), pp. 679–682. DOI: 10.1038/s41592-022-01488-1 (cit. on pp. 38, 40, 51, 55–57).

[91] J. P. Roney and S. Ovchinnikov. "State-of-the-art estimation of protein model accuracy using AlphaFold". In: *Physical Review Letters* 129.23 (2022), p. 238101. DOI: 10.1103/PhysRevLett.129.238101 (cit. on p. 40).

[92] R. Evans et al. "Protein complex prediction with AlphaFold-Multimer". In: *biorxiv* (2021), pp. 2021–10. DOI: 10.1101/2021.10.04.463034 (cit. on p. 40).

[93] G. E. Crooks et al. "WebLogo: a sequence logo generator". In: *Genome research* 14.6 (2004), pp. 1188–1190. DOI: 10.1101/gr.849004 (cit. on pp. 47, 51).

[94] V. Mariani et al. "lDDT: a local superposition-free score for comparing protein structures and models using distance difference tests". In: *Bioinformatics* 29.21 (2013), pp. 2722–2728. DOI: 10.1093/bioinformatics/btt473 (cit. on p. 52).

[95] W. Y. K. Hwang and J. Foote. "Immunogenicity of engineered antibodies". In: *Methods* 36.1 (2005), pp. 3–10. DOI: 10.1016/j.ymeth.2005.01.001 (cit. on p. 61).

A

Supplementary Tables

Table A.1: Summary of differences between three main types of biologics: antibodies, nanobodies and monobodies.

Feature	Antibodies	Nanobodies	Monobodies
Structure	- Y-shaped proteins composed of two heavy chains and two light chains. - Each chain has variable and constant regions.	- Consist only of the variable region of a single heavy chain. - Lack light chains.	- Synthetic proteins derived from fibronectin type III domains. - Do not have light and heavy chains.
Origin	- Naturally produced by the immune system of vertebrates (e.g., humans mice). - Can be engineered using hybridoma technology or phage display.	- Naturally produced by camelids (e.g., camels, llamas). - Can be isolated from these animals or produced synthetically.	- Entirely synthetic and engineered in the laboratory. - Derived from modifying the fibronectin type III domain found in humans and other organisms.
Binding Properties	- High specificity and affinity for a wide variety of antigens. - Antigen-binding site accommodates diverse shapes and sizes.	- High specificity and affinity. - Small size allows access to hidden or inaccessible epitopes. - Effective in targeting enzymes, ion channels, and intracellular proteins.	- Versatile binding surface engineered to recognize various protein targets. - Capable of binding to challenging protein surfaces.

Continued on next page

Table A.1 – *Continued from previous page*

Feature	Antibodies	Nanobodies	Monobodies
Stability and Production	- Relatively stable but sensitive to temperature and pH changes. - Require mammalian cell systems for production (costly and time-consuming).	- Highly stable even under extreme conditions (e.g., high temperatures, denaturing environments). - Can be produced efficiently in microbial systems like bacteria or yeast.	- Highly stable and easy to produce using bacterial expression systems. - Small size and simple structure facilitate manipulation and production.
Applications	- Widely used in research, diagnostics, and therapeutics. - Basis for diagnostic tests like ELISA and immunohistochemistry. - Employed as therapeutic agents in treating cancer, autoimmune disorders, and infectious diseases.	- Utilized in research and emerging therapeutics. - Suitable for diagnostic imaging and targeting inaccessible epitopes. - Explored as therapeutic agents due to favorable properties.	- Primarily used in research for studying protein function and interactions. - Useful in structural biology and protein engineering. - Potential applications as therapeutic agents being explored.
Size	- Approximately 150 kDa.	- Approximately 15 kDa.	- Approximately 10-15 kDa.

Table A.2: Energy surface landscape analysis from 2D PCA of SARS-CoV-2 RBD conformational dynamics in water. Energy minima, frame percentage, and conformation state for each of the basins are also given.

Basin	Free Energy	<E>/k_BT	S/R	Emin/k_BT	Frame Percentage	Conformation State
BA.2.86 Variant						
00	-6.25	1.53	7.78	0.82	16.28	Open
01	-6.09	1.01	7.10	-0.00	13.86	Closed
02	-6.06	0.90	6.95	0.01	13.42	Open
03	-5.72	1.18	6.90	0.29	9.65	Open
04	-5.40	2.64	8.04	1.94	6.95	Open
05	-5.33	1.73	7.06	1.08	6.50	Open
06	-5.33	2.37	7.69	1.60	6.48	Closed
07	-5.16	2.11	7.27	1.49	5.49	Closed
08	-5.06	2.23	7.29	1.75	4.94	Open
09	-4.59	2.12	6.71	1.52	3.09	Open
10	-4.56	1.74	6.31	1.02	3.05	Closed
11	-4.48	2.64	7.12	1.84	2.77	Open
12	-4.35	3.04	7.39	2.37	2.44	Open
13	-4.23	3.13	7.36	2.24	2.18	Open
14	-3.89	2.01	5.90	1.52	1.54	Closed
15	-3.70	3.03	6.73	2.37	1.28	Open
16	-1.05	4.64	5.69	4.22	0.09	Open
JN.1 Variant						
00	-6.60	0.94	7.54	-0.00	33.51	Open
01	-5.53	1.79	7.33	0.91	11.61	Closed
02	-5.38	1.61	6.99	0.87	9.95	Open
03	-5.35	1.52	6.87	0.76	9.69	Closed
04	-5.33	2.33	7.66	1.61	9.48	Open
05	-5.32	1.92	7.24	0.90	9.32	Open
06	-5.11	2.40	7.51	1.86	7.58	Closed
07	-4.69	2.77	7.47	2.07	4.99	Closed
08	-3.36	3.68	7.03	2.93	1.32	Closed
09	-3.23	2.47	5.70	2.05	1.16	Closed
10	-2.68	3.54	6.22	3.17	0.67	Closed
11	-2.30	4.00	6.30	3.45	0.48	Open
12	-1.25	3.15	4.40	3.10	0.16	Open
13	0.30	5.75	5.45	5.55	0.08	Closed