

UNIVERSITÀ DEGLI STUDI DI PAVIA

**DOTTORATO IN SCIENZE CHIMICHE E FARMACEUTICHE E
INNOVAZIONE INDUSTRIALE
(XXXVI Ciclo)**

Coordinatore: Chiar.mo Prof. Giorgio Colombo

**Insights into SARS-CoV-2:
Exploring Protein Dynamics and RNA Functionality
for future Therapeutic Strategies**

Tesi di Dottorato di
Alice Triveri

AA 2022/2023

Tutor

Prof. Filippo Doria

Co-tutor

Chiar.mo Giorgio Colombo

Table of Contents

| | |
|--|-----------|
| Abstract | 1 |
| <i>Preface: A Computational Evolutionary Journey together with SARS-CoV-2</i> | 1 |
| 1.1 Aim | 3 |
| <i>Introduction</i> | 6 |
| 1.2 Early stages of the COVID-19 pandemic: emergence and spread | 6 |
| 1.3 SARS-CoV-2 Virus | 8 |
| 1.3.1 Spike (S) protein | 11 |
| 1.3.2 Membrane (M) protein | 19 |
| 1.3.3 Nucleocapsid (N) protein | 19 |
| 1.3.4 Envelope (E) protein | 20 |
| 1.4 The emergence of SARS-CoV-2 Variants | 22 |
| 1.5 Therapies: small molecules, vaccines and monoclonal antibodies | 26 |
| 1.5.1 Small molecules | 26 |
| 1.5.2 Vaccines and monoclonal antibodies (mAbs) | 26 |
| 1.6 References | 34 |
| <i>Materials and methods</i> | 38 |
| 1.7 Molecular dynamics (MD) | 38 |
| 1.7.1 Force field | 41 |
| 1.7.2 Force field for RNA and DNA simulations | 43 |
| 1.7.3 Temperature and Pressure Coupling Schemes | 44 |
| 1.8 Enhanced sampling techniques for Accelerated Molecular Dynamics | 47 |
| 1.8.1 Rare Events, Separation of Timescales and Markovianity | 48 |
| 1.8.2 Potential and Free Energy Surfaces | 50 |
| 1.8.3 Metadynamics | 51 |
| 1.9 Epitope Prediction Method | 54 |
| 1.9.1 Energy Decomposition (ED) | 55 |
| 1.9.2 Matrix of Local Coupling Energies method (MLCE) | 56 |
| 1.10 Docking | 59 |
| 1.10.1 Theory of docking | 60 |
| 1.10.2 Search Algorithms | 60 |
| 1.10.3 Scoring functions | 61 |
| 1.10.4 Glide: grid-based ligand docking with energetics ⁴¹ | 63 |
| 1.11 References | 65 |
| <i>Proteins</i> | 67 |
| 1.12 Immunoreactivity of the WT SARS-CoV-2 spike protein | 69 |
| 1.12.1 Abstract | 69 |
| 1.12.2 Introduction | 70 |
| 1.12.3 Results and Discussion | 72 |
| 1.12.4 Methods | 81 |
| 1.12.5 References | 82 |
| 1.13 Immunoreactivity of the VOCs SARS-CoV-2 spike protein | 85 |

| | | |
|--|--|------------|
| 1.13.1 | Abstract | 85 |
| 1.13.2 | Introduction | 86 |
| 1.13.3 | Results | 90 |
| 1.13.4 | Discussion | 98 |
| 1.13.5 | Materials and Methods | 102 |
| 1.13.6 | References | 105 |
| Immunoreactivity of the Spike Protein: Conclusion | | 110 |
| 1.14 Structural dynamic differences of the VOCs SARS-CoV-2 spike protein | | 112 |
| 1.14.1 | Abstract | 112 |
| 1.14.2 | Introduction | 113 |
| 1.14.3 | Results | 116 |
| 1.14.4 | Discussion | 127 |
| 1.14.5 | Materials and Methods | 129 |
| 1.14.6 | References | 136 |
| Drug design on the Spike protein | | 140 |
| 1.15 Binding pocket conserved in the VOCs SARS-CoV-2 spike protein | | 142 |
| 1.15.1 | Abstract | 142 |
| 1.15.2 | Introduction | 142 |
| 1.15.3 | Results | 148 |
| 1.15.4 | Discussion | 158 |
| 1.15.5 | Materials and Methods | 159 |
| 1.15.6 | References | 162 |
| <i>DNA and RNA</i> | | 163 |
| 1.16 G4 Motifs as possible new drug targets | | 167 |
| 1.17 Unraveling the G-Quadruplex Mystery: Exploring the (un)folding mechanism | | 172 |
| 1.17.1 | G4 Folding and Unfolding Mechanisms. | 176 |
| 1.17.2 | Molecular Dynamics simulations of G4 folding and unfolding | 177 |
| 1.17.3 | Methodology | 180 |
| 1.17.4 | Observed unfolding pathways in different topologies | 182 |
| 1.17.5 | Discussion | 191 |
| 1.17.6 | Conclusion | 196 |
| 1.17.7 | Materials and Methods | 197 |
| 1.17.8 | References | 200 |
| <i>Conclusions and perspectives</i> | | 203 |

Abstract

The main goal of this thesis is to thoroughly understand the structural and dynamic properties of key biomolecular components of SARS-CoV-2 with the aim of developing pan-coronavirus methods. For this reason, the thesis divided into two parts: a protein-oriented exploration and a genetic-material-focused analysis.

The protein-oriented approach delves deep into the immunoreactive, dynamic, and structural characteristics of SARS-CoV-2 Spike protein and its variants. By investigating these intricate details, we aim to unravel pivotal insights that transcend individual variants, ultimately leading us to identify fundamental patterns applicable to a broader range of viral agents. This understanding serves as a basis to devising adaptable strategies capable of addressing both present and future pandemics.

On the other hand, the genetic approach deepens the possibility of targeting secondary structures of the virus RNA other than the classical ones, the G-quadruplexes (G4).

Through the synergy of these two distinct yet interlinked approaches, this thesis aims to construct a comprehensive framework for addressing the challenges posed by viral mechanisms. By combining knowledge from the world of proteins and genetics, we aspire to build a general computational strategy for the study of the fine mechanisms regulating viral-host interactions and viral life regulation.

Preface: A Computational Evolutionary Journey together with SARS-CoV-2

The outbreak of the novel coronavirus disease (COVID-19) caused by the severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) has ushered in an unprecedented global health crisis. Since its emergence, the virus has undergone continuous genetic changes, leading to the emergence of numerous variants with distinct biological characteristics, the so-called Variants of Concern (VOCs).

This thesis adopts a multi-faceted approach to investigate SARS-CoV-2: it leverages both protein-based and genetic methodologies to unravel the complex behavior of the virus to design new and innovative therapeutic strategies.

For this reason, the organization of the thesis will be reminiscent of this division, there will be a first part focused on the virus proteins (the **Protein approach**) and a second part concerning the RNA (the **Genetic approach**).

Among the proteins of the virus, the Spike protein (S protein) plays a pivotal role in viral entry and host immune recognition. Understanding the dynamics of the Spike protein and the effects of its mutations is essential for comprehending the virus's interaction with the host immune system and its impact on disease severity and transmission.

This thesis focuses on an in-depth analysis of the SARS-CoV-2 Spike protein and its mutations, aiming to shed light on their multifaceted roles in viral pathogenesis. Specifically, the research explores the effect of Spike protein mutations on the immune response, particularly the efficacy of monoclonal antibodies against different variants. Additionally, it investigates the influence of these mutations on the stability of the viral variants and their implications for virulence. Furthermore, this work delves into the presence of the fatty acid binding pocket within the Spike protein and examines its conservation across various SARS-CoV-2 variants. By addressing these critical aspects, this thesis aims to contribute to our understanding of SARS-CoV-2 evolution and host-virus interactions, with potential implications for the development of therapeutic interventions and public health strategies. Additionally, this protein approach examines conserved protein-protein interactions between different coronavirus, shedding light on essential molecular mechanisms governing virus-host interactions.

The urgent need to combat SARS-CoV-2 has boosted the search for new targets and innovative compounds, strengthening the arsenal of antiviral drugs. From a genetic perspective, a notable challenge is to target noncoding RNAs with small molecules, which requires a shift from conventional drug discovery methods.

Noncoding RNAs have gained considerable importance in drug discovery because of their central role in biological processes. Dysregulation of the functions of noncoding RNAs is directly associated with several diseases. Recognizing their potential as therapeutic targets represents an exciting frontier that could greatly expand drug development, particularly in the context of RNA viruses such as SARS-CoV-2. Targeting the RNA genome of SARS-CoV-2 with small molecule drugs extends its promise beyond the current pandemic. Despite the challenges posed by the complexity of these hybrid compounds, their clinical application and the identification of intricate RNA-based targets represent the future of medicinal chemistry.

In addition, G-quadruplexes (G4), non-canonical secondary structures formed by guanine-rich sequences, have emerged as highly conserved targets in viral genomes. Targeting these structures has potential for antiviral therapies. The goal was to delve deeper into the (un)folding mechanisms of G4s to develop precise tools that interact with them, thereby contributing to a comprehensive understanding and potential advancement of antiviral strategies.

1.1 Aim

The primary objective of the ongoing protein-focused research is to conduct an in-depth investigation into the SARS-CoV-2 Spike protein. This multifaceted approach involves distinct yet interconnected key objectives, each aiming to provide critical insights into the nature and behavior of this essential viral protein.

The first objective involves predicting immune recognition regions within the Spike protein. A straightforward yet robust structure-dynamics-energy based strategy has been developed for this purpose. This strategy is designed to comprehensively predict regions of the Spike protein that are likely involved in immune recognition. This insight holds immense potential in guiding the development of novel molecules for both vaccine and diagnostic purposes. Remarkably, this approach has successfully identified potentially reactive regions within the S protein stalk. These identified regions are currently undergoing experimental synthesis and testing, highlighting the translational impact of this research.

The second objective revolves around assessing immune response variability, specifically in response to mutations in the Spike protein. Understanding how these mutations affect the immune response is vital, especially in evaluating the efficacy of monoclonal antibodies against various SARS-CoV-2 variants. This research seeks to uncover the extent to which these mutations influence the immune system's ability to neutralize the virus, a critical aspect in the ongoing fight against the pandemic.

Additionally, investigating the stability of SARS-CoV-2 variants with Spike protein mutations constitutes another vital objective. This research seeks to delve into how these mutations impact the stability of the virus, its ability to persist, transmit, and potentially cause severe disease. By providing insights into the dynamic nature of viral evolution, this research strives to contribute valuable knowledge to the field of virology and aid in the development of informed public health strategies.

Lastly, this research involves exploring the presence and conservation of the fatty acid binding pocket within the Spike protein across various SARS-CoV-2 variants. This analysis aims to identify potential druggable targets for therapeutic interventions and assess their relevance in the context of viral evolution. Understanding the presence and variability of such pockets is crucial for identifying potential targets for drug development, an essential step towards effective therapeutic strategies.

In summary, this comprehensive protein-focused research endeavors to unravel critical aspects of the SARS-CoV-2 Spike protein, offering insights that can shape the development of diagnostics, therapeutics, and public health measures in the ongoing battle against the COVID-19 pandemic.

On the other end, the primary objective of **genetic approach** research is to develop a computational tool that can comprehensively characterize the intricate G-quadruplex (G4) folding landscape. These structures are not only present in the human genome but have been found also in the genomes of both DNA and RNA viruses, including human immunodeficiency virus-1 (HIV-1), Zika virus (ZIKV), hepatitis C virus (HCV), rhinovirus, Ebola virus (EBOV), etc.

Understanding their biological roles is crucial, as these structures play pivotal roles in gene regulation, DNA replication, transcriptional regulation, alternative splicing, and translational regulation, depending on the sequence and on the topology.

However, experimental study of the folding and unfolding mechanisms of G4 structures presents a significant challenge due to the complexity of the process. Unlike a simple funnel mechanism, the mechanism of G4 folding/unfolding follows a kinetic partition (KP) model, it contains deep competing free-energy minima (alternative folds, competing conformational basins or ensembles) separated by large free-energy barriers. Considering this, experimental identification of multiple competing folds populated during the folding process but vanishing at the thermodynamic equilibrium is difficult, since they may mutually overlap in the measurable signals during the folding, making them unresolvable.

By developing an advanced computational tool, this research aims to navigate through this intricate KP mechanism and shed light on the dynamic G4 folding landscape. Through computational simulations and analysis, it seeks to decipher the nuances of G4 folding, offering insights into how these structures modulate gene expression and regulation. This understanding will deepen our knowledge of biological processes in the human body but will also provide critical insights into the behavior of G4 structures within viral genomes.

Ultimately, the research endeavors to pave the way for potential therapeutic interventions and novel drug discovery strategies that target G4s, both in human and viral contexts, and further contributing to advancements in the field of genetic research.

At the end, by addressing these research objectives, this thesis aims to contribute to our understanding of SARS-CoV-2 biology and its interactions with the host immune system, ultimately offering valuable insights into the development of effective strategies to combat the ongoing COVID-19 pandemic.

Introduction

1.2 Early stages of the COVID-19 pandemic: emergence and spread

In the late 2019 (around mid-December) a cluster of patients in China’s Hubei Province, in the city of Wuhan began displaying symptoms of an unusual pneumonia-like illness that did not respond well to standard treatments. Among the first 27 documented hospitalized patients, most cases shared direct exposure to the Huanan Seafood Wholesale Market¹, a wet market located in downtown Wuhan, which sells not only seafood but also live animals, including poultry and wildlife.

Further investigations revealed that the first patients could be traced back to early December 2019.²

On December 31, the Wuhan Municipal Health Commission informed the World Health Organization (WHO) about an outbreak of pneumonia with an unknown cause (Summary of the events in Table 1).

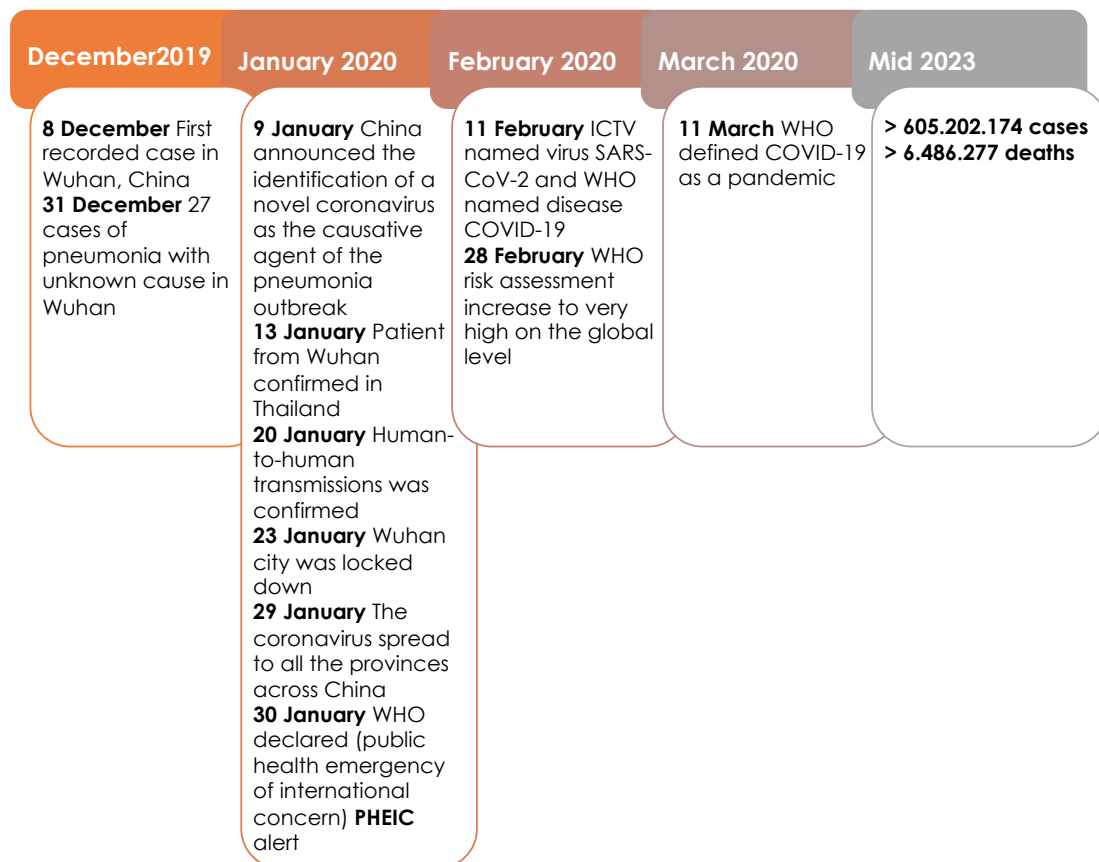


Table 1. Summary of the events of the COVID-19 Pandemic.

Subsequently, independent teams of Chinese scientists conducted RNA sequencing and virus isolation from samples taken from affected patients. Their research led them to identify the causative agent of this emerging disease as a previously unseen beta-coronavirus.³⁻⁵ The results of this etiological identification were publicly announced on January 9, 2020. On January 10, the first genome sequence of the new coronavirus was published on the Virological website, followed by the release of other nearly complete genome sequences determined by various research institutions through the GISAID database on January 12.⁶

This new coronavirus pneumonia quickly spread to other cities in Hubei Province, eventually reached different parts of China and within a month, it spread massively to all 34 provinces of the country. The number of confirmed cases suddenly increased, with thousands of new cases diagnosed every day by the end of January. On January 30, WHO declared the outbreak of the new coronavirus a public health emergency of international concern.⁷

Subsequently, on February 11, the International Committee on Virus Taxonomy named the new coronavirus "SARS-CoV-2" while the disease caused by it was named "COVID-19" by the WHO.⁸

The international spread of COVID-19 has accelerated since late February. The highly efficient transmission of SARS-CoV-2, coupled with extensive international travel, facilitated the rapid global spread of COVID-19. On March 11, 2020, WHO officially designated the global epidemic of COVID-19 as a pandemic (https://www.who.int/docs/default-source/coronaviruse/situation-reports/20200311-sitrep-51-covid-19.pdf?sfvrsn=1ba62e57_10).

In Italy, the first two cases of the pandemic were confirmed on January 30, 2020, when two tourists from China tested positive for SARS-CoV-2 in Rome. An outbreak of COVID-19 infections was subsequently detected on February 21, 2020, from 16 confirmed cases in Codogno (LO), Lombardy, increased to 60 the following day.

Although the genetic evidence suggests that SARS-CoV-2 is a naturally occurring virus that probably originated in animals, there is still no conclusion as to when and where the virus first entered humans. In fact, some of the early cases reported in Wuhan had no epidemiological link to the seafood market and it has been suggested that the market may not be the initial source of human infection with SARS-CoV-2.⁹ However, this remains a highly debated issue that falls beyond the scope of our discussion.

1.3 SARS-CoV-2 Virus

SARS-CoV-2 is a member of the Coronaviruses, a heterogeneous group of viruses that infect various animals and can lead to respiratory infections in humans, ranging from mild to severe. These viruses, according to differences in the genome sequence and serological reactions, are classified into four genera: *Alphacoronavirus*, *Betacoronavirus*, *Gammacoronavirus* and *Deltacoronavirus*. SARS-CoV-2 is classified as a *Betacoronavirus* (β -CoV) as other two other β -CoVs which have caused outbreaks: the severe acute respiratory syndrome coronavirus (SARS-CoV) in 2002 and the Middle East respiratory syndrome coronavirus (MERS-CoV) in 2012. These viruses originated in humans and caused severe respiratory illnesses, prompting concerns about the emergence of new coronaviruses as a significant public health concern in the 21st century. This family of viruses has a unique structural feature which resembles a solar crown due to the presence of Spike proteins on the virion surface.

One of the main characteristics of coronaviruses is the high rate of genetic recombination and mutation, which means that there are many different types of these viruses and that they can adapt to different hosts. Many of these can attack humans but usually (as in the case of these seven viruses: Human coronaviruses 229E, OC43, NL63 and HKU1) these are responsible for 10–30% of upper respiratory tract infections annually, characterized by mild respiratory illnesses, such as the common cold. Instead, SARS and MERS-CoV were able to cause severe human respiratory diseases, potentially resulting in high mortality. (In 2002–2003, SARS-CoV resulted in 8,096 reported cases and 774 deaths (case-fatality rate of ~10%). By the end of January 2020, 2,500 cases of Middle East respiratory syndrome and more than 800 associated deaths (case-fatality rate ~34%) were reported worldwide (https://www.who.int/health-topics/middle-east-respiratory-syndrome-coronavirus-mers#tab=tab_1)

.10

SARS-CoV-2 exhibits a genome sequence similarity of 79% with SARS-CoV and 50% with MERS-CoV. The genetic arrangement is consistent among all β -CoVs: the genome is approximately 30 kb in size and consists of 14 open reading frames (ORFs), which encode 29 viral proteins. The 5' end of the SARS-CoV-2 genome comprises about two-thirds of its length and encodes two overlapping polyproteins: pp1a and pp1ab. These polyproteins are cleaved by viral proteases, resulting in the production of 16 non-structural proteins (NSPs). The NSPs play a critical role in viral replication and transcription.

At the 3' terminus of the viral genome, four ORFs encode a standard set of structural proteins: Nucleocapsid (N) protein, Spike (S) protein, Membrane (M) protein, and Envelope (E) protein. These proteins are responsible for the assembly of viral particles and play a role in evading the host immune response. Between these structural genes, there are additional accessory genes that encode accessory proteins, including ORF3a, ORF3b, ORF6, ORF7a, ORF7b, ORF8b, ORF9b, and ORF14. These accessory proteins participate in the regulation of viral infection, although they may not be incorporated into the virion, except for ORF3a and ORF7a, which are considered structural proteins. Nevertheless, the molecular functions of many accessory proteins remain largely unknown owing to the lack of homologies to accessory proteins of other coronaviruses or to other known proteins.¹¹

The four structural proteins share 90% similarity with the corresponding SARS-CoV's proteins except the Spike which diverges (with only around 80% sequence identity due to longer protein length in the 2019-nCoV compared with both SARS and MERS-CoV). The non-structural proteins instead have greater than 85% amino acid sequence identity with SARS-CoV.¹²

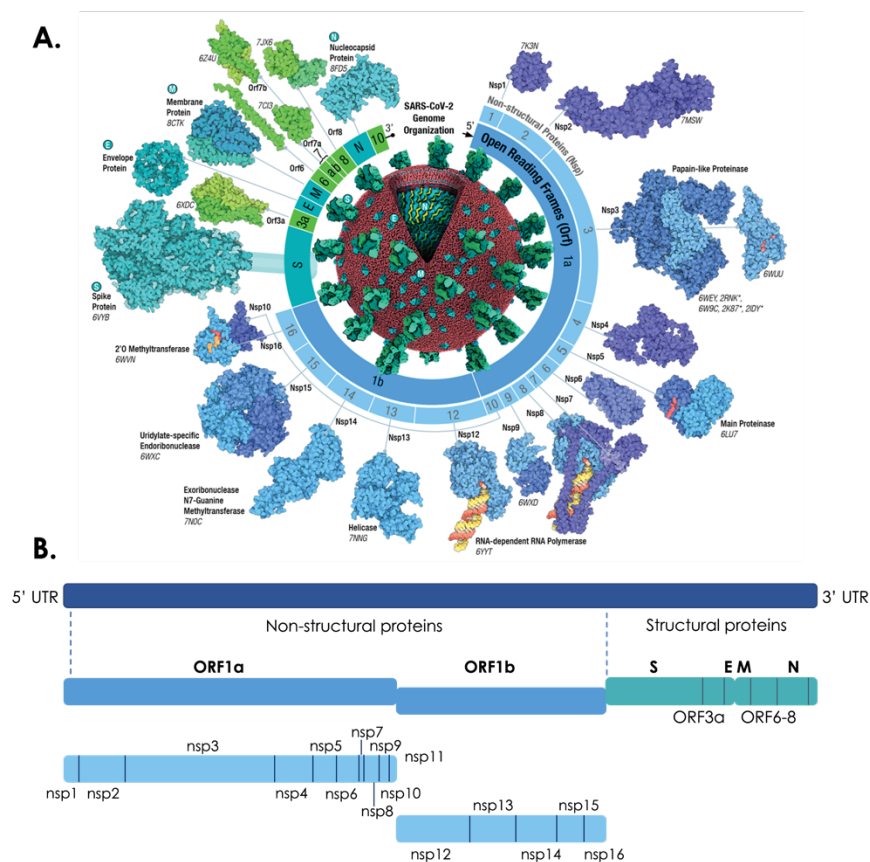


Figure 1. A. Architecture of the SARS-CoV-2 genome and proteome, including nsps derived from polyproteins or pp1a and pp1ab (shades of blue), virion structural proteins (turquoise), and open reading frame proteins (Orfs, shades of green). Polyprotein cleavage sites are indicated by inverted

triangles for Papain-like Proteinase (PLPro, black) and the Main Protease (nsp5, blue). The double-stranded RNA substrate-product complex of the RNA-dependent RNA polymerase (shown as the nsp7-nsp82-nsp12 heterotetramer and separately with only nsp12) is color coded (yellow: product strand, red: template strand). Transmembrane portions of the Spike S-protein are shown in cartoon form (light blue). The source of the structural models used for analyses for all study proteins are indicated. **B.** Sequence of the SARS-CoV-2 genome. Although these two previous betacoronavirus epidemics raised awareness of the need for clinically available therapeutic or preventive interventions, no treatments were ready to be used at the beginning of the pandemic. The development of effective intervention strategies relies on the knowledge of molecular and cellular mechanisms of coronavirus infections, which highlights the significance of studying virus–host interactions at the molecular level to identify targets for antiviral intervention and to elucidate critical viral and host determinants that are decisive for the development of severe disease.

Briefly, the SARS-CoV-2 life cycle¹³ may be described as follows (see also the **Figure 2**): in the first step the S protein on the outer surface of the virion binds the host receptor for attachment to the cell membrane, which is followed by viral and host cellular membrane fusion and the release of viral genomic RNA into the cells. Subsequently, host ribosomes are hijacked to produce the two viral replicase polyproteins, which can further be processed into 16 mature NSPs through two virus-encoding proteases: main protease (M^{Pro}) and papain-like protease (PL^{Pro}). These NSPs can assemble into the replication and transcription complex (RTC) to initiate viral RNA replication and transcription. The genomic RNA and structural proteins then assemble into mature progeny virions, which are subsequently released through exocytosis to initiate another round of infection. Viral proteins can individually perform important physiological roles, constitute the viral protein machinery for specific essential events in the viral life cycle or extensively interplay with the cellular factors

in the host immune response and pathogenesis. In the following, I will briefly describe the structures and biological roles of the key viral proteins.

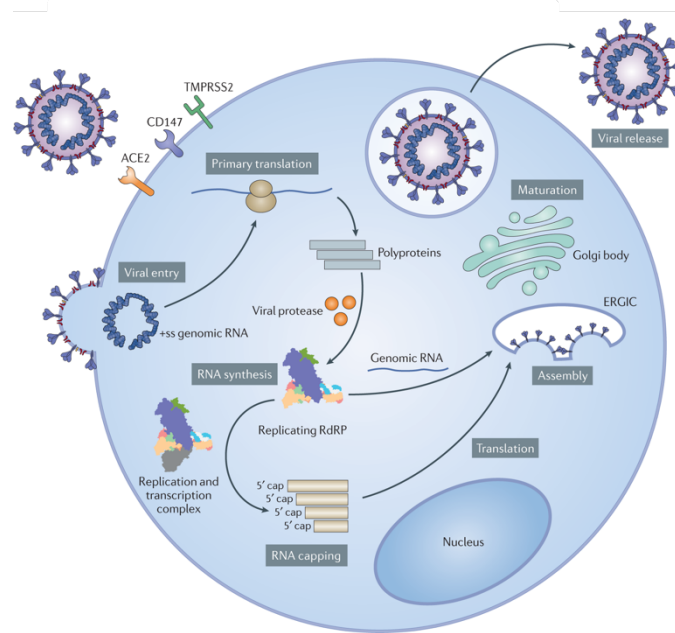


Figure 2. A. The life cycle of SARS-CoV-2, including viral entry, replication and transcription, assembly and release. SARS-CoV-2 enters host cells through an endocytosis pathway mediated by S protein–angiotensin-converting enzyme 2 (ACE2) interactions. Viral RNA enters the cytoplasm after the entry step, and then ORF1a or ORF1ab is translated by the host ribosome. The viral polyproteins are cleaved into NSPs and assemble themselves into the replication and transcription complexes. Subgenomic viral mRNAs (after capping) act as templates for viral protein translation. Progeny virions are assembled in the endoplasmic reticulum and Golgi body. Afterwards, the virions are exocytosed to complete the life cycle. ERGIC, endoplasmic reticulum–Golgi intermediate compartment; ExoN, exonuclease; HEL, helicase; Mac1, macrodomain 1; NendoU, uridine-specific endoribonuclease; NiRAN, nidovirus RNA-dependent RNA polymerase-associated nucleotidyltransferase; NMT, guanine-N7-methyltransferase; OMT, 2'-O-methyltransferase; PL2, papain-like protease 2; RBD, receptor-binding domain; RdRp, RNA-dependent RNA polymerase; SUD, SARS-unique domain; +ss, positive-sense single-stranded; TM, transmembrane; Ubl1, ubiquitin-like domain 1; UTR, untranslated region.

1.3.1 Spike (S) protein

The Spike (S) protein¹⁴ plays a crucial role in the life cycle of the virus, making it a significant target for various therapies, diagnostics, therapeutics, and vaccines. The S protein is a type I membrane protein, formed as a trimer composed of three identical monomers anchored to the viral membrane by its transmembrane segment. To initiate

infection, the S protein binds to the angiotensin-converting enzyme 2 (ACE2) receptor, undergoing structural rearrangements that promote fusion between the virus and the host cell. This protein is fully glycosylated with 22 N-linked glycosylation sites per protomer.

The full-length S protein in the original is synthesized as a single 1273 amino acid polypeptide chain and can be cleaved by a furin-like protease into two functional subunits, S1 and S2, which are responsible for mediating attachment to host cells and membrane fusion, respectively.

S1 contains the N-terminal domain (NTD), receptor-binding domain (RBD), and C-terminal domains (CTD1 and CTD2), while S2 includes the fusion peptide (FP), fusion-peptide proximal region (FPPR), heptad repeat 1 (HR1), central helix (CH), connector domain (CD), heptad repeat 2 (HR2), transmembrane segment (TM), and the cytoplasmic tail (CT), depicted in **Figure 3**.

In its native state, the S1 fragment forms a 'V' shaped architecture with the NTD at one arm and the RBD, CTD1 and CTD2 at the other, wrap around the central helical bundle formed by the prefusion S2 fragment, projecting the N-terminal end of HR1 toward the viral membrane.

In this configuration, the RBD can sample two distinct conformations: the open 'up' representing a receptor-accessible state and closed 'down' a receptor-inaccessible state (**Figure 4**). In the 'down' state, RBD angles are close to the central cavity of the trimer to shield the receptor-binding regions, while in the 'up' state, the RBD undergoes hinge-like conformational movement, exposing its determinant regions to recognize the human angiotensin-converting enzyme 2 (hACE2) receptor on the host cellular membrane, the state of which is considered to be less stable than in the 'down' state.

The three NTDs are located at the periphery of the trimer, each making contacts with the RBD from the adjacent protomer. The CTD1 and CTD2 pack underneath the RBD against S2 and between the two neighboring NTDs, indicating they could modulate these domains and play important roles in the structural rearrangements required for membrane fusion.

In the postfusion conformation (**Figure 3, E.**), S1 dissociates as a monomer and S2 adopts a rigid, baseball-bat-like shape (~220 Å long), with HR1 flips over forming a continuous long helix together with the CH. The connector domain (CD) and other segments contribute to the formation of helix bundles, ultimately facilitating membrane fusion.

The process of membrane fusion is triggered when the S1 subunit binds to hACE2 and leading to structural changes and shedding of S1: the interaction involves specific residues in both the RBD and hACE2, resulting in the activation of the HR1 and HR2 helices, which

form a stable six-helix bundle. This conformational change brings the viral and host cell membranes into proximity, facilitating membrane fusion and subsequent infection.

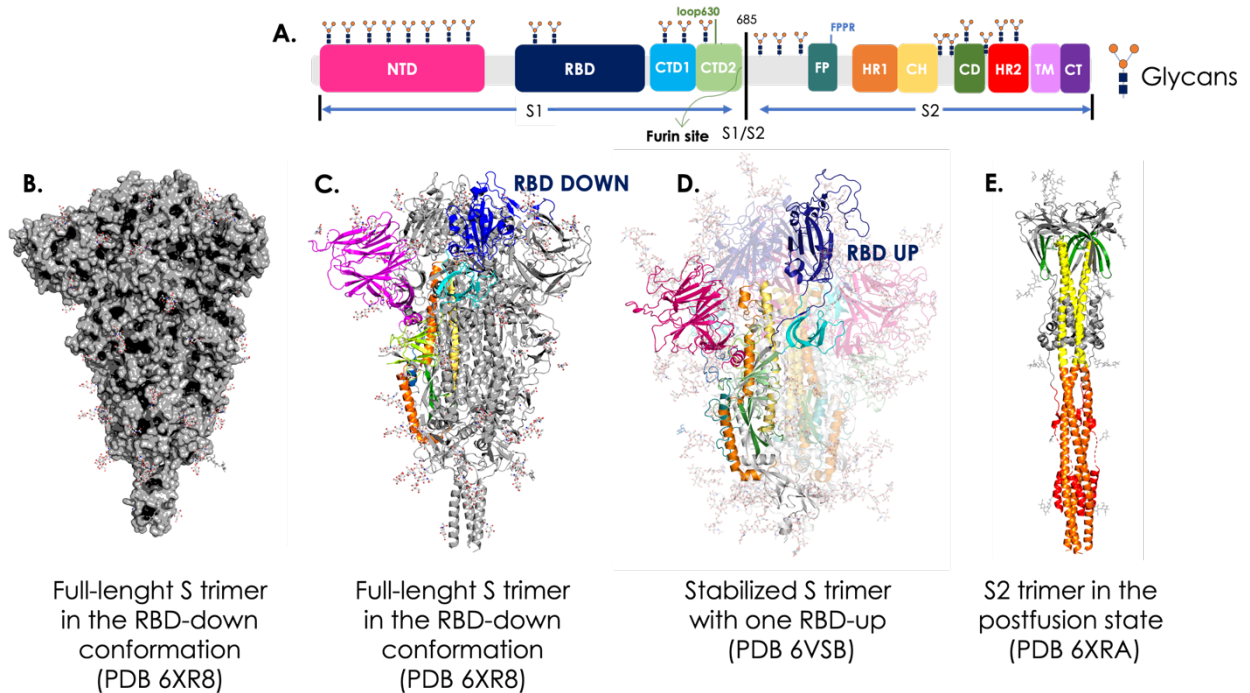


Figure 3. Distinct conformational states of the SARS-CoV-2 spike protein. **A.** Schematic representation of the SARS-CoV-2 spike protein organization. Segments of S1 and S2 include: NTD, N-terminal domain; RBD, receptor-binding domain; CTD1, C-terminal domain 1; CTD2, C-terminal domain 2; S1/S2, S1/S2 cleavage site; S20, S20 cleavage site; FP, fusion peptide; FPPR, fusion peptide proximal region; HR1, heptad repeat 1; CH, central helix region; CD, connector domain; HR2, heptad repeat 2; TM, transmembrane anchor; CT, cytoplasmic tail; and tree-like symbols for glycans. **B.** Viral SARS-CoV-2 S trimer in the prefusion conformation (PDB 6XR8). **C.** Cryo-EM structure of the full-length S trimer in the RBD-down conformation (PDB 6XR8). **D.** Stabilized S trimer with one RBD-up (PDB 6VSB). **E.** Cryo-EM structure of the full-length S2 trimer in the postfusion conformation (PDB 6XRA).

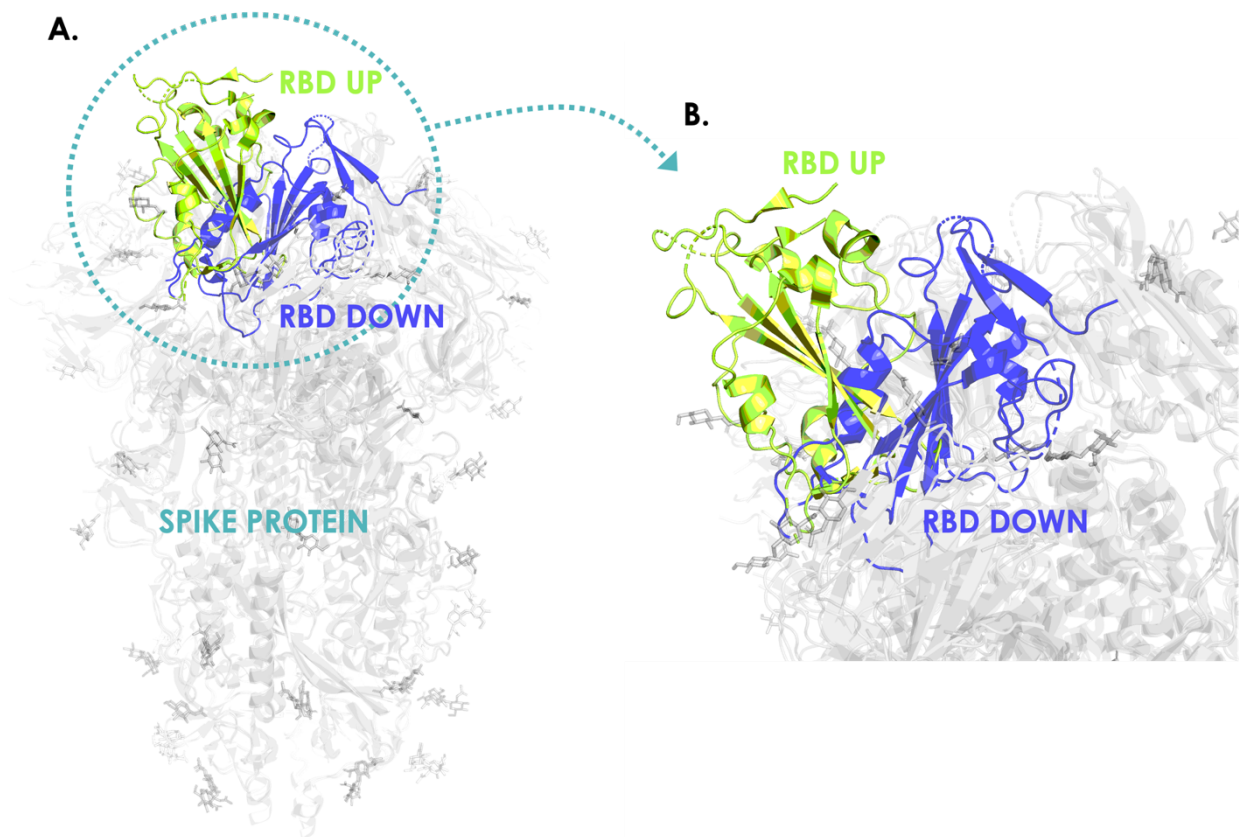


Figure 4. A. Superimposition of the full-length Spike protein in its conformations with one RBD in the DOWN configuration (in blue) and with the UP configuration (in green). **B.** Zoom on the RBD in DOWN (blue) and UP (green) configurations.

1.3.1.1 N-terminal domain (NTD)

The NTD (14-685 residues) is situated at the outer edge of the spike protein and extends away from the threefold axis. It can be subdivided into three regions: the top, core, and bottom regions. The core structure exhibits a galectin-like antiparallel β -sandwich fold, composed of one six-stranded β -sheet and another with seven strands. The top region consists of two antiparallel β strands connected by a short loop, while the bottom region is primarily composed of two short β sheets and a helix. The NTD is adorned with eight N-linked glycans, similar to those found in MERS.¹⁵ Though the exact function of the NTD in SARS-CoV-2 S is not fully understood, other coronaviruses have shown that NTD may be involved in recognizing sugars during initial attachment or specific protein receptors, or it might play a role in the prefusion-to-postfusion transition. The presence of NTD-targeted neutralizing antibodies (nAbs) isolated from SARS-CoV-2 infected patients,¹⁶ with potent effects at the nM level, suggests the domain's crucial functional role.

1.3.1.2 Receptor binding domain (RBD)

The RBD (319-541 residues) contains two subdomains: a five-stranded antiparallel β -sheet connected by short helices and loops, and an extended loop, named receptor binding motif (RBM). The overall structure of the SARS-CoV-2 RBD closely resembles that of the SARS-CoV RBD (with a RMSD of 1.2 Å for 174 aligned $C\alpha$ atoms). The RBM, which has more sequence variation, retains structural similarity (RMSD of 1.3 Å) with only a minor conformational change at its distal end. The binding mode of the SARS-CoV-2 RBD to ACE2 is almost identical to that observed in the previously determined structure of the SARS-CoV RBD-ACE2 complex.¹⁵

The extended RBM's gently concave outer surface interacts with the N-terminal helix of the peptidase domain (PD) of ACE2. The interaction involves hydrogen bonds and salt bridges between specific polar residues of the RBD and ACE2, contributing to receptor engagement. Mutations of key residues, such as N501Y, K417N, and E484K, found in fast-spreading variants, lead to enhanced affinity for ACE2 and immune evasion.

Obviously, the RBD is a dominant target of nAbs elicited by either natural infection or vaccination, confirming its pivotal role during infection, see Chapter 3.4.2.

1.3.1.3 C-terminal domains (CTDs)

The CTDs mainly consist of β -structures from segments of S1 and the N-terminal segment of S2 adjacent to the furin cleavage site (see **Figure 3**). CTD1 contains two antiparallel β -sheets with two strands and four strands, respectively. CTD2 also has two β -sheets: a four-stranded one and another four-stranded one that includes a strand from the S2 subunit.¹⁵

In the RBD-down conformation of the S trimer, a structural element in the CTD2, named the '630 loop', becomes well-ordered in the G614 variant while disordered in the Wuhan-Hu-1 strain. The structured 630 loop inserts into a gap between the NTD and CTD1 of the same protomer, stabilizing the CTD2 structure. This loop is also located near the S1/S2 boundary and the fusion peptide proximal region (FPPR) of a neighboring protomer. The FPPR and the 630 loop help retain the RBDs in the down conformation but move out of their positions when the adjacent RBD flips up. Thus, the CTDs, along with the FPPR and the 630 loop, play critical roles in modulating the fusogenic structural rearrangements of the S protein.¹⁷⁻¹⁹

1.3.1.4 S2 domain

In the prefusion conformation, three S2 subunits form a tight packing around a central three-stranded coiled-coil of approximately 140 Å long, formed by CH. Part of the HR1 and a segment of S2 (residues 758-784) adopt an α -helical conformation and assemble into a nine helix-bundle with the central coiled-coil, creating the most rigid part of the entire S trimer.

The CD region links CH and the C-terminal HR2 through a linker region. The FP forms a short helix and fits into a pocket between two neighboring S protomers. The structured FPPR clamps the prefusion S trimer in the closed, RBD-down conformation. The remaining HR2, TM and CT segments are disordered in the most S trimer structures but show low-resolution density in the cryo-ET reconstructions tilting away from the trimer's threefold axis at various angles (17° to 60°).

In the postfusion conformation, the HR1 and CH form a continuous α -helix with three copies of them assembling into a long central three-stranded coiled-coil. Part of the HR2 folds into α -helix and packs against the groove between two HR1-CH helices to form a six-helix bundle structure, resembling the postfusion organization of other viral fusion proteins. The CD remains unchanged from the prefusion conformation, as a three-stranded β -sheet covering the C-terminal end of HR1-CH helices.

Comparison of the prefusion and postfusion conformations of S suggests that HR1 undergoes significant rearrangements to form a coiled-coil, translocating its N-terminal closer to the target cell membrane to project the FP. Additionally, HR2 and the TM at its C-terminal end fold back to pack along the groove of the HR1-CH coiled-coil, forming the postfusion six-helical bundle. These refolding events bring the viral and target cell membranes into proximity, ultimately leading to membrane fusion. Notably, the postfusion S2 surface is decorated with five N-linked glycans arranged in a regular spacing, possibly serving to protect S2 from host immune responses.¹⁵

1.3.1.5 Roles of Glycans

Another key structural feature of the Spike protein is its extensive glycosylation as shown in **Figure 5**.

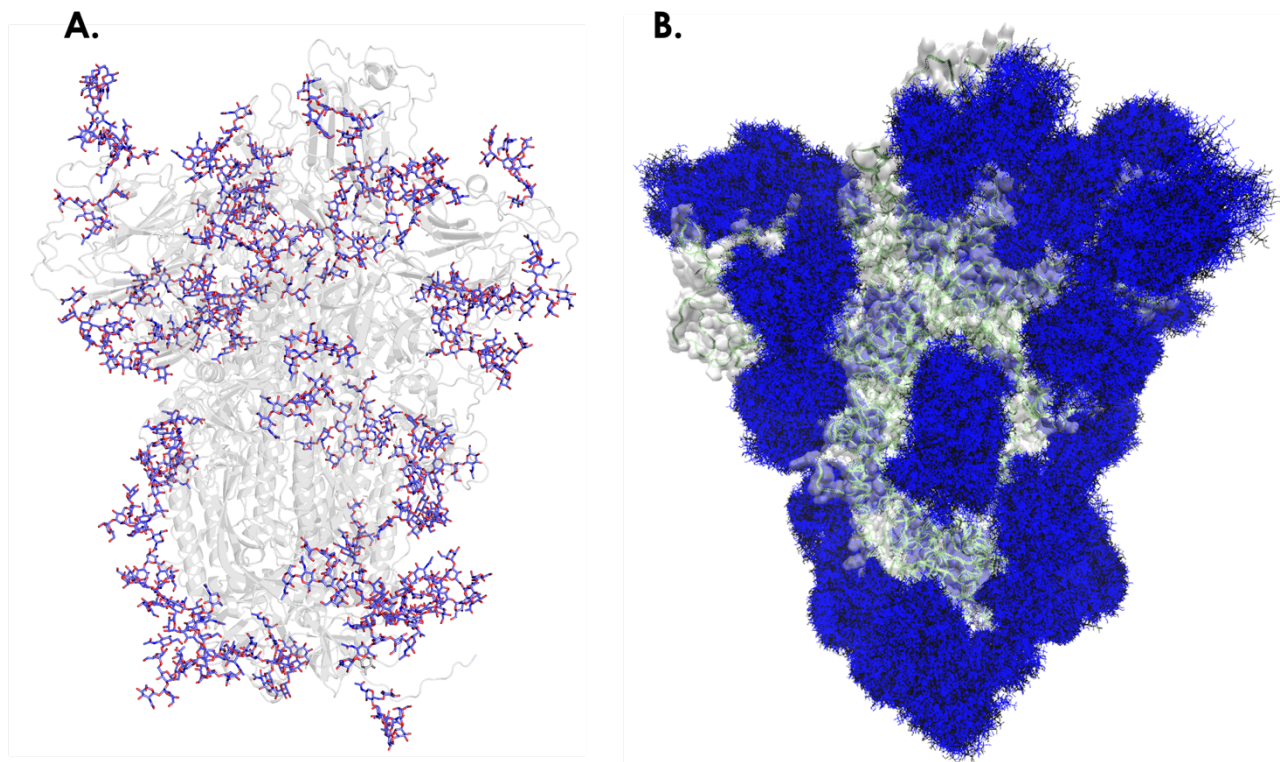


Figure 5. Glycan shield of the SARS-CoV-2 S protein. **A.** Glycosylated full-length model of the S protein in the open state. Protein is depicted with gray cartoons, where the *N*-/*O*-glycans are shown in sticks representation. **B.** Molecular representation. Glycans at several frames (namely, 500 frames, one every 10 ns from one replica) are represented with blue lines, whereas the protein is shown with cartoons and highlighted with a cyan transparent surface.

Glycan is a biomolecule present in numerous proteins and lipids, serving as a functional component. Usually, protein glycosylation plays a crucial role in viral pathogenesis, for example in the HIV-1 envelope spike (Env) protein where the surface is almost entirely covered in N-glycans (they account for more than half of the protein's molecular weight).²⁰⁻

26

The biological roles of the N-glycans present on the surface are very diverse and related to the glycoprotein's function.²⁰ They are crucial for facilitating viral entry through membrane fusion. During viral entry, envelope glycoproteins initiate the process through molecular recognition events with cell surface receptors. These interactions are often mediated by

specific N-glycan epitopes, further emphasizing the significant role of N-glycans in viral infection and entry mechanisms.^{20, 27-29}

In addition to their diverse biological roles, the N-glycans on viral envelope glycoproteins serve as an effective shield against the host immune system.^{20, 23, 24, 30} These complex carbohydrates which are non-immunogenic or weakly form a highly dense coating, making it difficult for the host immune system to recognize and neutralize the virus.

However, there are differences between various viruses in the effectiveness of this glycan shield. For instance, in HIV-1, the glycan shield has been proven to be highly effective, enabling the virus to evade the immune system effectively.^{20, 21} On the other hand, in the case of SARS and MERS, the glycan shield is not as efficient in evading the immune response, making these viruses more susceptible to immune recognition and attack.²¹ Furthermore, the glycosylation pattern in SARS and SARS-CoV-2 is different from HIV-1. These coronaviruses have a large presence of complex N-glycans, specifically the oligomannose type and specifically 22 predicted N-glycosylation sites per protomer, of which at least 17 have been found to be occupied plus at least two predicted O-glycosylation sites.³¹ These differences in glycosylation may contribute to the varying levels of immune evasion and recognition observed among different.

Beyond their shielding role, the N-glycans attached to N165 and N234 play a crucial structural role in regulating the conformational transitions of the receptor-binding domain (RBD) on the SARS-CoV-2 spike (S) protein. Through simulations, was observed that the deletion of these glycans destabilizes the RBD's 'up' conformation. These findings were validated by biolayer interferometry experiments, which showed a reduction in ACE2-binding and an increase in the RBD's "down" conformation. Furthermore, simulations revealed that the glycans act as camouflage for the SARS-CoV-2 S protein, enabling it to evade the host immune response effectively. Analyzing the glycan shield in detail, has been found that the stalk region is largely invulnerable, particularly to larger molecules, whereas the head region presents a more viable target for immune recognition. In addition, the receptor-binding motif (RBM) accessibility showed a significant difference in glycan shielding between the 'up' and 'down' RBD conformations.³²

Overall, the glycans on the viral envelope glycoproteins serve a dual role, being of paramount importance both as immune evasion mechanisms and as essential structural elements for virus infectivity.

Understanding the various roles of glycans in SARS-CoV-2 is crucial for developing effective strategies for therapeutics, vaccines, and treatments aimed at targeting the virus and mitigating the impact of COVID-19.

1.3.2 Membrane (M) protein

The membrane (M) protein is 221 amino acids long and has little similarity with the M proteins of other coronaviruses.³³ Its crucial role lies in the assembly of virions within the cell, taking place between the endoplasmic reticulum (ER) and the Golgi body. The M protein consists of three structural components: the N-terminal portion of the virion that protrudes from the membrane, which is sensitive to protease, binds to the surface of the virus, the transmembrane domains and there are two domains in the C-terminal. Adjacent to the transmembrane region in the third domain, an amphipathic domain is succeeded by a small hydrophilic region. This segment links to the host's viral or cytoplasmic membrane, facilitating virus assembly and maturation.³⁴ Over time, the cellular membrane protein transforms into a site where fresh viral particles are generated within the host cell. Furthermore, the M protein plays a vital role in fostering aggregation by interacting with the viral ribonucleoprotein and S glycoprotein during the budding process. Notably, this protein in SARS-CoV has also a protective glycosylated region that may be critical for host-virus interaction and this can be used to inhibit some important inflammatory proteins.³¹

A noteworthy attribute of the M protein is its ability to bind with all structural proteins³⁵: for instance, the interaction of the M protein with the nucleocapsid (N) protein contributes to the stability of the N protein.³⁶ On the other hand, when the S protein and the M protein bind to each other, changes occur that may affect how the virus interacts with the host cell and enters the cell.³¹

1.3.3 Nucleocapsid (N) protein

The N protein of SARS-CoV-2 consists of an N-terminal RNA-binding domain (NTD) and a C-terminal dimerization domain (CTD) and shares ~90% sequence identity with N protein of SARS-CoV. The regions located between the N-terminus and NTD, between NTD and CTD, and between CTD and the C-terminus of the N protein of SARS-CoV-2 (thereafter referred to as N protein) are predicted to be intrinsically disordered. At neutral pH, the N protein is positively charged (+24 e), consistent with its strong binding affinity with negatively charged RNA. The gel filtration and dynamic light scattering results further suggested the

oligomerization of N protein. Altogether, the sequence and structure features of N protein are similar to those of other proteins that have been reported to undergo liquid–liquid phase separation (LLPS) with nucleic acids. Thus, it is probable that the N protein may also undergo LLPS with viral genome RNA and potentially facilitate viral assembly.

The N protein serves as a multifunctional component crucial for transcription and replication processes.^{37, 38} This protein is required for the creation of ribonucleoproteins that regulate the replication and synthesis of the viral RNA genome.³⁸ The main function of the N protein is to bind to the RNA genome of the viral infection and package it into a long nucleocapsid, which is also known as ribonucleoprotein.³⁹ Most studies have shown that this protein affects host-pathogen interactions, including actin reactivation and host-cell cycle progression.⁴⁰ This protein is highly immunogenic and is present in large quantities during infection.⁴¹ Inside the virus, the N protein protects and stabilizes the viral RNA.⁴² Throughout the virus assembly process, it collaborates with viral membrane proteins and interfaces with the M protein.⁴³ Furthermore, it exerts influence on RNA folding, translation, and the progression of the cell cycle.⁴⁰ The N protein establishes associations with transcription and replication complexes within infected cells.⁴⁴

Evidence implies that this molecule might contribute to the pathophysiology of central nervous system (CNS) infections. One hypothesis proposes that the N protein could trigger the activation of toll-like receptors (TLR)3, TLR7, or TLR8, consequently initiating signaling pathways that enhance the activation of NF- κ B and NLRP3. This sequence of events could lead to a cytokine storm and subsequent inflammatory responses.⁴⁵ Consequently, these mechanisms could play a role in the development of various conditions such as cancer, coagulation disorders, neurodegenerative ailments, and cardiovascular diseases.⁴⁶ Notably, the N protein's potential involvement in the assembly of the SARS-CoV-2 virus offers valuable insights for the formulation of intervention strategies aimed at curbing the COVID-19 pandemic.⁴⁷

1.3.4 Envelope (E) protein

The envelope (E) protein consists of a chain of 10 to 74 amino acids, organized into three domains: a short hydrophilic N-terminal domain (NTD), a hydrophobic transmembrane domain (TMD), and a long hydrophilic C-terminal domain (CTD). It exists each in monomeric and pentameric forms. Alignment of the E protein in MERS-CoV, SARS-CoV, and SARS-CoV-2 revealed a tendency to accumulate a net positive charge balance in the CTD. This indicates a heightened stability within the topology from MERS-CoV to SARS-CoV-2,

probably contributing to the extended pathogenicity and heightened resistance to control determined in SARS-CoV-2.⁴⁸

Within viral cloth, this protein is typically found in about 20 copies.⁴⁹ Prior investigations have discovered that mutagenesis has an outstanding effect on the progression and dissemination of viral infections.⁵⁰ Specifically, viruses poor on this protein cannot infect the host cell and have a very low viral titer in the host cell.⁵¹ Positioned along the secretory pathways connecting the endoplasmic reticulum (ER) and the Golgi equipment of the host cellular,⁵² the E protein's C-terminal location is structurally enclosed in the virus's envelope. Consequently, it is closed to the envelope itself, in the end becoming encapsulated within it.⁵³ This phenomenon can restrict the host cell's capacity to replicate and disseminate the virus across the body. Despite its unique characteristic closing enigmatic, this small protein has the capacity to result in the formation of lipid droplets inside the virus. It accompanied by the M and N proteins is very important for the development and propagation of virus particles in SARS-CoV-2. Moreover, E protein interacts with host cell proteins and acts as an ion channel⁵¹ and because of the latter can act as a determinant for coronaviruses virulence and play an important role in its pathogenic process.⁵⁴ So, SARS-CoV-2 E protein plays an important role in virial pathogenesis, making it an excellent target for drug therapy.

1.4 The emergence of SARS-CoV-2 Variants

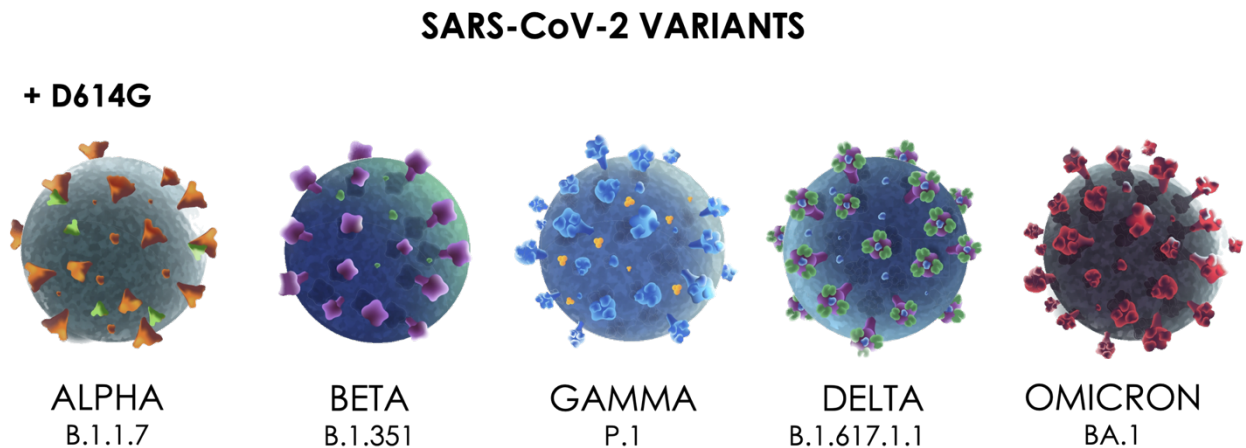


Figure 6. SARS-CoV-2 Variants of Concern (VOCs)

A crucial aspect to take into consideration when devising strategies against SARS-CoV-2 variants is their inherent nature as β -CoVs, which belong to the category of RNA viruses. These viruses store their genetic instructions in the form of single-stranded RNA, a knowingly delicate molecule prone to errors during the replication process. Consequently, these replication errors can trigger mutations within the viral genome. Indeed, an intrinsic feature of RNA viruses, including SARS-CoV-2, is their remarkable replication rate. This rapid replication leads to the swift generation of a multitude of viral particles within a short timeframe, heightening the likelihood of genetic errors during the copying of the genome. Together, unlike DNA viruses, many RNA viruses lack a robust proofreading mechanism during replication. Unlike the specialized enzymes present in DNA viruses that rectify replication mistakes, RNA viruses are deficient in this error-correction process, rendering them susceptible to accumulating mutations.

Furthermore, the host's immune system endeavors to combat the virus by detecting and targeting specific components. Here, mutations within the viral genome can sometimes confer a shield against the immune response, creating selective pressure that favors mutations facilitating evasion from the host's immune system. Along with this aspect, the environmental factors, such as changes in the host population, transmission dynamics, environmental conditions and vaccines should also be considered to the selection and spread of specific viral variants.

These mutations can also cause alterations in the virus's properties, affecting its ability to bind to host cells, replicate more efficiently, or achieve greater spread. Some mutations may

grant the virus a competitive advantage in survival or replication, thus potentially increasing their prevalence in the viral population over time.

It is crucial to acknowledge that not all mutations yield significant or substantial alterations in the virus's behavior. Many mutations have little to no impact on the virus's properties or its ability to cause disease. However, some mutations can lead to variations that affect factors like transmissibility, severity of illness, or vaccine effectiveness.

Combining all these factors (High Replication Rate, lack of Proofreading Mechanism, Immune Pressure, Evolutionary Advantage, Genetic Diversity and Environmental Factors) has led to the SARS-CoV-2 Variants.⁵⁵⁻⁵⁷ These changes, which can be mutations, insertions, or deletions of the amino acid sequence, affect the entire genome of the virus. But of particular concern is the fact that these mutations could affect the antigenicity of the S protein, which, given its function, is the main target of neutralizing antibodies against infection.

For almost a year the only noteworthy mutation in the S protein was D614G (Asp⁶¹⁴ → Gly) which immediately became the dominant strain throughout the world.⁵⁸ However, over time, this strain has further evolved, giving rise to several variants of concern (VOCs). This single-residue substitution correlates with elevated viral loads in infected patients and heightened infectivity of pseudotyped viruses. However, it does not exhibit a corresponding association with disease severity. The G614 virus demonstrates comparable susceptibility to neutralization by both convalescent human sera and sera from vaccinated hamsters.¹⁷ This suggests that vaccines containing the original D614 variant remain effective against the G614 virus.

Structural analyses reveal that the G614 virus possesses a more stable S trimer configuration compared to the original strain. This stability is attributed to the insertion of a loop (630 loop) into a wider gap in the G614 trimer, which is less accommodated in the D614 trimer due to its narrower gap. This loop stabilizes key domains and prevents premature S1 dissociation, contributing to the G614 variant's enhanced stability. The increased stability of the G614 variant influences its infectivity. The transition from a closed state to an RBD-up(s) conformation in a G614 trimer involves an order-disorder shift in the 630 loops, leading to slower transitions compared to the D614 trimer. This deceleration in transition rates, coupled with the stabilization of RBDs, explains the higher prevalence of the RBD-up conformation in the G614 variant. These insights elucidate the mechanisms underlying the enhanced infectivity of the G614 virus and its impact on viral behavior and vaccine efficacy.

Following the emergence of D614G, an amino acid substitution within the receptor-binding motif (RBM), N439K, was noted as increasing in frequency in Scotland in March 2020. This point mutation was followed by others, like Y453F, N501Y (substitution associated with greater transmissibility and a greater affinity for ACE2), E484K (increase the spike affinity to ACE2 and bring resistance to antibody neutralization targeting the original epitope) etc. and some deletion such as the deletion in the NTD, the $\Delta 69-70$ which became dominant and part of other changes in the genome.

In fact, after these first point mutations, in the fall of 2020 multiple SARS-CoV-2 variants began to circulate globally. The most notable are in the United Kingdom (UK), a new variant (the 20I/501Y.V1, or B.1.1.7), labelled as the **Alpha** Variant by the WHO, emerged this time with a high number of mutations. These mutations seemed to be associated with an increased risk of death compared with other variants.

In South Africa, the variant B.1.351 (known as 20H/501Y.V2, **Beta**), emerged independently of B.1.1.7 but shared some mutations with it.

In Brazil, the P.1 (20J/501Y.V3, **Gamma**) was first identified in four travelers from the Brazil (tested in Japan) and had 17 unique mutations including three in the receptor binding domain of the Spike protein, two shared with B.1.351, E484K and N501Y, the latter also shared with the strain of B.1.1.7, and a different mutation K417T which was K417N in the B1.351 strain.

All of them were labelled as Variants of Concern (VOC) because were associated with increased transmissibility or detrimental change in COVID-19 epidemiology, increased virulence, or different clinical disease presentation and decrease sensitivity to available vaccines or therapeutics.

After, the B.1.617.2 variant (AY, **Delta**) was first detected in India in late 2020 and in June 2021 became the dominant variant globally.

Of course, the assignment of these VOCs was disseminated with the presence of thousands of Variants of Interests (VOIs), such as the Epsilon (B.1.427 and B.1.429), Eta (B.1.525), Iota (B.1.526) Kappa (B.1.617.1) and Mu (B.1.621, B.1.621.1), etc.

Finally, the SARS-CoV-2 **Omicron** (B.1.1.529, BA.1) variant was first identified on November 24th, 2021, in South Africa and immediately declared VOC replacing the Delta variant. The omicron variant has a very large number of mutations, around 30-point mutations only in the Spike protein, combined with deletions and insertions of amino acids.

In the Table 1 below there is the summary of all the VOCs characteristics in term of transmissibility, severity and lethality and escape to immune response compared to the WT.

Table 1

| SARS-CoV-2 strain | Declaration of VOC | Extension | Transmissibility | Severity and lethality | Escape to immune response | Mutations found in the S protein gene |
|-------------------|---|-----------------------------------|---|---|---|---|
| WT | Wuhan (China) 07/01/2020 | Worldwide | $R_0 = 2.5$. Incubation period: 2–14 days, median 5.1 days. SAR: 0.7–75%. | 81% mild. 14% severe. 5% critical. 2.3% death. | PVE 95% for symptomatic infection. | |
| Alpha | United Kingdom 29/12/2020 | Europe, Oceania and North America | ↑Transmissibility (50% higher). SAR: 25.1%, 1.43–1.82 times higher. | ↑Severity and lethality. 1.55–1.73 times more lethality. | ↑Immune escape. PVE 89% for symptomatic infection, and 95% for hospitalization or death. | Del69-70, del144, N501Y, A570D, D614G, P681H, T716I, S982A, D1118H. |
| Beta | South Africa 29/12/2020 | Africa | ↑Transmissibility (50% higher, 2.5 times higher). SAR higher. | ↑Severity and lethality. | ↑Immune escape. PVE 84% for symptomatic infection, and 95% for hospitalization or death. | L18F, D80A, D215G, R246I, K417N, E484K, N501Y, D614G, A701V. |
| Gamma | Brazil 29/12/2020 | Latin America | ↑Transmissibility (1.7–2.4 times higher). $R_0 = 3.4$. SAR higher. | ↑Severity and lethality. | ↑Immune escape. PVE 84% for symptomatic infection, and 95% for hospitalization or death. | L18F, T20N, P26S, D138Y, R190S, K417T, E484K, N501Y, D614G, H655Y, T1027I, V1176F. |
| Delta | India 11/05/2021 | Worldwide | ↑↑Transmissibility (1.97 times higher). $R_0=7$. Intradomiciliary delta SAR (10.3–21%) 1.70 times higher than intradomiciliary alpha SAR. Shorter incubation period (median 4.5 days). Higher viral load, 2.5 times more in nasopharyngeal exudate and 15 times more in saliva. | ↑↑Severity and lethality. 2.20 times more hospitalization. 3.87 times more ICU admission. 2.37 times more lethality. | ↑↑Immune escape. PVE 87% for symptomatic infection, and 93% for hospitalization or death. | T19R, G142D , del156-157, R158G, K417N (delta plus), L452R, T478K, D614G, P681R. |
| Omicron | South Africa and Botswana 26/11/2021 | Worldwide | ↑↑↑Transmissibility (36.5% higher than delta). $R_0=10$. Intradomiciliary omicron SAR 15.8%-31% versus delta SAR 10.3–21%. Extradomiciliary omicron SAR 8.7% versus delta SAR 3.0%. 70-fold higher respiratory viral load at 24 hours in omicron than in original and delta strains. Shorter incubation period (median 3 days). | ↓Severity and lethality. 0.71 times less (29% less) hospitalization. 10-fold lower viral load in lung tissue at 24 hours in omicron than in original strain. | ↑↑↑Immune escape. PVE 10% for symptomatic infection, 49% if third dose. PVE 70% for hospitalization. 2.4–5.4 times higher risk of reinfection. | A67V, del69-70, T95I, G142D, del143-145, Y145D, del211, L212I, ins214EPE, G339D, S371L, S373P, S375F, K417N, N440K, G446S, S477N, T478K, E484A, Q493R, G496S, Q498R, N501Y, Y505H, T547K, D614G, H655Y, N679K, P681H, N764K, D796Y, N856K, Q954H, N969K, L981F. |

Table 1. Characteristics of the emerged Variants of Concern (VOCs). R_0 : basic reproductive number: average number of new cases generated (by contagion) from a single case. SAR: secondary attack rate: number of new cases of a disease among the total number of exposed susceptible people within a specific group (i.e., household or close contacts), that is, the proportion of contacts of a primary case who become ill. PVE: preventive vaccine effectiveness, in all cases after two doses with Comirnaty vaccine (based on messenger RNA technology).

1.5 Therapies: small molecules, vaccines and monoclonal antibodies

1.5.1 Small molecules

Since the outburst of the pandemic, researchers have explored various drugs as potential treatments for SARS-CoV-2. The first rapid approach was the repurposing of drugs already approved or advanced in clinical trials. Among these are remdesivir, a viral RNA polymerase inhibitor initially designed for Hepatitis C, chloroquine and hydroxychloroquine, recognized for malaria treatment, tocilizumab, a monoclonal antibody deployed in rheumatoid arthritis therapy, favipiravir, an anti-influenza medication, Kaletra, utilized in HIV treatment and more recently, masitinib, a kinase inhibitor applied in addressing mast cell tumors in animals.⁵⁹

Although these targeted repurposing strategies provide potentially rapid trajectories toward an approved treatment, in the meantime additional therapies for SARS-CoV-2 infection are needed to improve clinical efficacy, expand global drug supplies, and address the potential emergence of viral resistance.

Moreover, in more recent developments, the European Union has granted emergency use authorization for two additional treatments: Lagevrio (also referred to as molnupiravir, developed by Merck) and Paxlovid (developed by Pfizer). These medications are approved for treating adults with COVID-19 who are at an elevated risk of developing severe illness and do not require supplemental oxygen.

1.5.2 Vaccines and monoclonal antibodies (mAbs)

However, during this period, research has persistently advanced, leading to the creation of alternative treatments, notably including vaccines and the utilization of monoclonal antibodies. For the majority of widely used vaccines, the initial premise centered on targeting the vulnerability of SARS-CoV-2 viral transmission, specifically the interaction between the RBD of the spike protein and human angiotensin-converting enzyme 2 (ACE2). The strategy aimed to elicit robust levels of neutralizing antibodies (nAbs) against epitopes within this

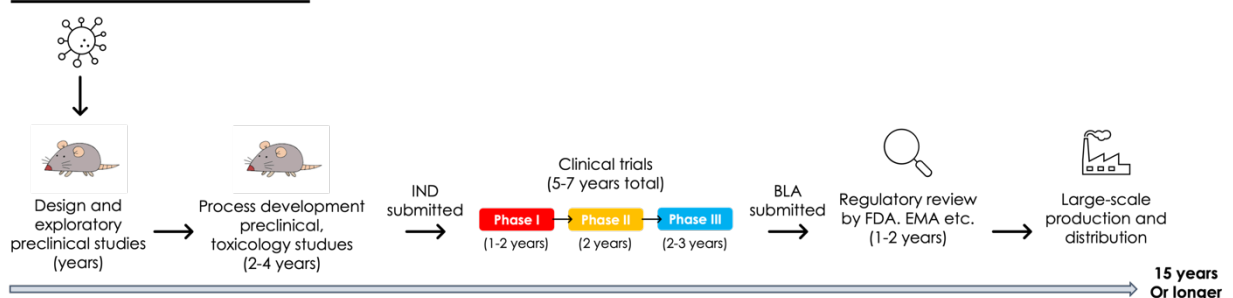
interface for effective immune response. Consequently, a diverse array of methodologies was employed to express the spike protein, either in stabilized or non-stabilized forms, to facilitate immune recognition.

Notably, the mRNA-based platforms (Pfizer, Moderna) and adenovirus-based platforms (AstraZeneca, Gamaleya, Johnson & Johnson), which have taken a prominent role in clinical development, have undergone rigorous experimental trials spanning various geographic and disease contexts over the course of decades.

Obviously, given the worldwide emergency, the vaccine against this virus has been fast-tracked. Traditional vaccine development is a lengthy process and has a development time of around 15 years (the process is summarized in the **Figure 7**). The SARS-CoV-2 pandemic has required rapid action and the development of vaccines in an unprecedented timeframe (**Figure 7, B.**). Knowledge of vaccine development from previous work on vaccine candidates for SARS- and MERS-CoV proved invaluable, allowing the initial exploratory vaccine design phase for SARS-CoV-2 to be significantly streamlined, resulting in substantial time savings. Consequently, the first clinical trial for a SARS-CoV-2 vaccine candidate commenced in March 2020 (NCT04283461).

The trial strategies were carefully orchestrated, with overlapping clinical phases and staggered trial initiation. Notably, several manufacturers have already embarked on commercial vaccine production, taking on potential risks, even before phase III trial results are available.

A. Traditional development



B. SARS-CoV-2 vaccine development

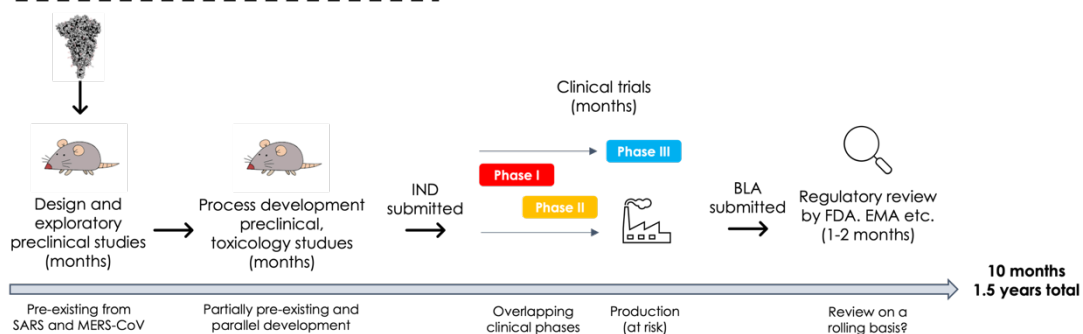


Figure 7. Vaccine development: A. Traditional vs B. for SARS-CoV-2. A. Traditional vaccine development typically take over 15 years, starting with an extensive discovery phase dedicated to vaccine design and preliminary preclinical investigations. Subsequently, a phase characterized by more formal preclinical experiments, toxicology assessments, and the development of production methods ensues. This intricate process involves the submission of an investigational new drug (IND) application, followed by the vaccine candidate progressing through phase I, II, and III trials. Upon successful completion of phase III trials and the fulfillment of predetermined endpoints, a biologics license application (BLA) is submitted for regulatory agency review, culminating in the vaccine's licensure. Large-scale production initiates thereafter.

B. The development of **vaccines for SARS-CoV-2** has undergone a rapid acceleration. Drawing from the knowledge acquired during the development of vaccines for SARS- and MERS-CoV, the initial discovery phase was bypassed. Instead, existing processes were adopted, and phase I/II trials were promptly launched. Phase III trials were initiated following an interim analysis of phase I/II results, with multiple clinical trial stages conducted concurrently. Simultaneously, vaccine manufacturers embarked on large-scale production of several vaccine candidates, despite the associated risks.

However, several platforms (the type of technology used to develop the vaccine) have been exploited for vaccine development: 'traditional' platform (inactivated or live-virus vaccines), or more 'innovative' as based on recombinant protein vaccines and vectored vaccines or RNA/DNA vaccines.⁶⁰ Currently, eight vaccines are approved for use in Europe: Bimervax (previously COVID-19 Vaccine HIPRA), Comirnaty, COVID-19 Vaccine (inactivated, adjuvanted) Valneva, Jcovden (previously COVID-19 Vaccine Janssen), Nuvaxovid, Spikevax (previously COVID-19 Vaccine Moderna), Vaxzevria (previously COVID-19 Vaccine AstraZeneca) and VidPrevtyn Beta (<https://www.ema.europa.eu/en/human-regulatory/overview/public-health-threats/coronavirus-disease-covid-19/covid-19-medicines#authorised-covid-19-vaccines-section>).

Among these, Comirnaty and Spikevax represent mRNA vaccines. These vaccines incorporate messenger RNA (mRNA) encoding the S protein, which is encapsulated within lipid nanoparticles. Once within the cell, a portion of this mRNA is broken down by RNAase enzymes, a defense mechanism against foreign RNA, while the remainder is translated to facilitate protein expression. Notably, the recognition of mRNA by sensors within the innate immune system leads to the generation of inflammatory mediators. It's essential to highlight that both mRNA translation and degradation occur within the cell's cytoplasm, preventing integration into the host genome, an occurrence that might arise with DNA adenovirus vector vaccines.

On the other hand, the adenoviral vector platform, utilized by vaccines such as Jcovden and Vaxzevria, involves a non-replicating viral vector bearing double-stranded DNA encoding the S protein. Upon cellular entry, the vector makes its way into the nucleus, where mRNA synthesis takes place. Once the mRNA exits the nucleus, protein synthesis proceeds as previously explained. The adenovirus functions both as a transporter and a protective barrier for the genetic material. Additionally, it targets cells of the innate immune system, thereby triggering both inflammatory and antigenic responses.

Lastly, Nuvaxovid, VidPrevtyn Beta, and Bimervax are recombinant protein vaccines. These vaccines contain the full-length prefusion recombinant S protein combined with an adjuvant called Matrix-M, which induces robust B-cell and T-cell responses.

While clinical studies have demonstrated the efficacy of vaccines in stimulating the production of antibodies, the duration of protection against SARS-CoV-2 is constrained, necessitating vaccine administration approximately every 4 to 6 months. Moreover, the effectiveness of vaccines is notably influenced by the specific variant of the virus that is currently in circulation. In fact, there are a lot of studies monitoring the efficacy of vaccines against different variants and the mRNA vaccines there are available adapted vaccines based of other strains.^{57, 61, 62}

In addition to vaccines but still based on the achilles heel of the virus, the spike protein, research has been based on monoclonal antibodies (mAbs).⁶³ Four mAb products targeting SARS-CoV-2 (bamlanivimab plus etesevimab, casirivimab plus imdevimab, sotrovimab, and bebtelovimab) have been granted Emergency Use Authorizations (EUA) by the Food and Drug Administration (FDA) for treating mild to moderate COVID-19 in outpatients.

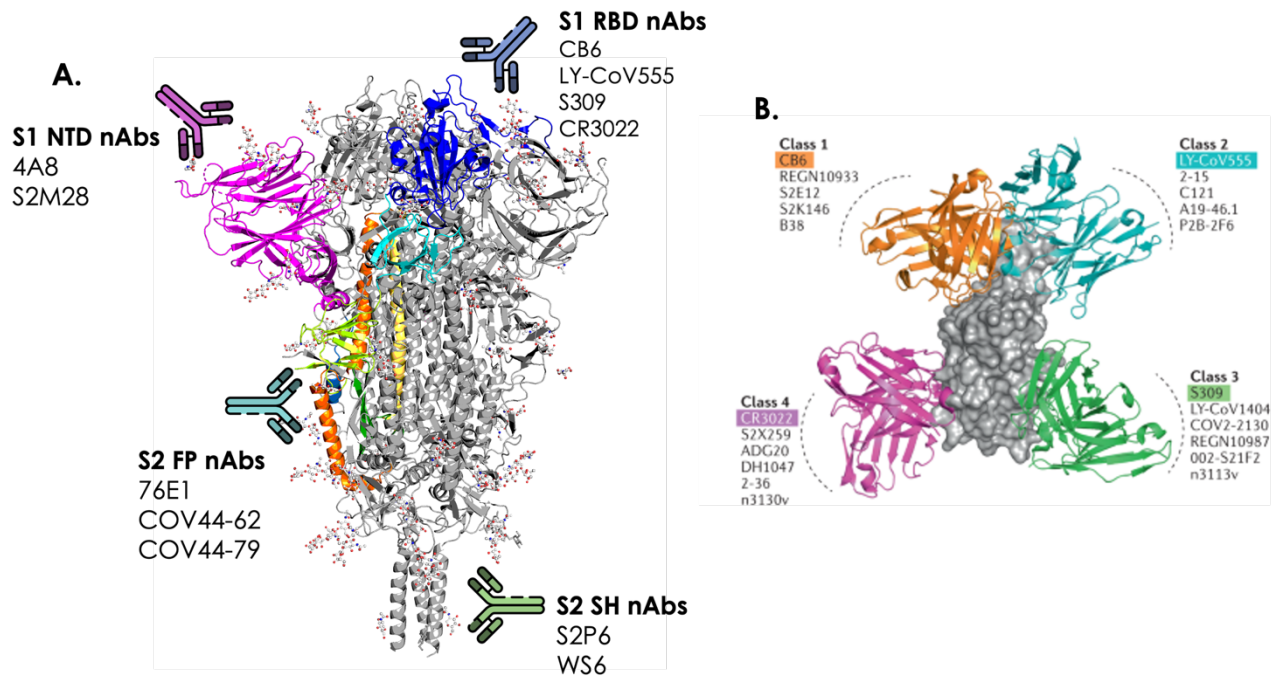


Figure 8. A. Different groups of neutralizing antibodies (nAbs) that target the S protein. Representative nAbs targeting the S1 N-terminal domain (NTD), S1 receptor-binding domain (RBD), and S2 stem helix (SH) and S2 fusion peptide (FP) regions. **B.** RBD-directed nAbs can be divided into four main classes depending on the epitopes they target in the RBD of the S protein. For each class, one representative nAb bound to the RBD monomer is shown: class 1, CB6; class 2, LY-CoV555; class 3, S309; class 4, CR3022. CD, connector domain; CH, central helix; CT, cytoplasmic tail; HR, heptad repeat; SD, subdomain; TM, transmembrane domain.

The mAbs differ from the polyclonal antibodies in that they selectively bind selectively a specific epitope of the antigen. Early research focused on the mAbs known to bind SARS- and MERS-CoV. Among the mAbs tested, only two gave promising results: CR3022 and S309. Many other mAbs failed to target the S protein, due to sequence and structural differences among the three coronaviruses. And interestingly, none of the two active mAbs bind to the ACE2 binding site. After, have been developed mAbs that can selectively bind to different areas of the spike, including the RBD, the NTD, in the S1 the SH region and the FPs.

The antibody 4A8 was among the earliest neutralizing antibodies (nAbs) discovered to target the N-terminal domain (NTD). The NTD's structural elements include five loops (N1–N5), where interactions with 4A8 primarily involve loops N3 and N5. Likewise, additional NTD-targeting mAbs like COV2-2676 and COV2-2489 can recognize the epitope formed loops N1, N3, and N5. Notably, most NTD-targeting antibodies do not hinder the action of antibodies targeting different sections of the S protein, such as the receptor-binding

domain (RBD). Therefore, an effective strategy to counter COVID-19 could involve combining NTD-targeting antibodies with those binding non-NTD regions of the S protein.

The majority of antibodies discovered against SARS-CoV-2 are directed at the RBD and can be categorized based on their specific targeted epitopes. Various classification systems have been proposed, with the most widely recognized being introduced by Barnes et al.⁶⁴ This system divides RBD-targeting antibodies into four distinct classes according to their binding interactions with the S protein.⁶⁵

Class 1 antibodies, block the ACE2 receptor and attach to the 'up' conformation of RBDs. The epitope on the RBD coincides with the receptor-binding motif (RBM). These antibodies, predominantly encoded by VH3-53 and VH3-66 germ lines, exclusively identify the RBD in its 'up' conformation. The primary mechanism of action for class 1 antibodies involves hindering the interaction between ACE2 and the S protein, thus resulting in significant neutralization effects. Other antibodies are part of this class such as S2E12 (one of the few class 1 mAbs that retains broad-spectrum neutralizing activity for all current VOCs),⁶⁶ CB6, REGN10933, B38 etc.⁶⁷

The mAbs in **class 2** are similar to those in class 1 on the basis of their binding to the RBM domain but they can bind both 'up' and 'down' conformations of the S protein. For example, LY-CoV555, which was isolated from a patient recovering from COVID-19, binds and neutralizes SARS-CoV-2 and displays protective efficacy against SARS-CoV-2 in clinical trials.⁶⁸

These two categories of monoclonal antibodies do not exhibit an exceptional capacity for wide-ranging inhibition against both SARS-CoV and other SARS-like coronaviruses. This is primarily due to the limited amino acid identity between SARS-CoV and SARS-CoV-2, with only a 59% overlap in the RBM region.

On the other hand, antibodies classified as **class 3** exhibit a distinct binding pattern, targeting regions outside the ACE2-binding site. Notably, these antibodies are capable of binding to the RBDs regardless of whether they are in the 'up' or 'down' conformation. Class 3 antibodies, exemplified by REGN10987, COV2-2130, 2-7, 1-57, A19-61.1, P2G3, S309, and LY-CoV1404, showcase robust neutralizing capabilities against various SARS-CoV-2 variants.^{67, 69-76} Among these antibodies, S309 recognizes epitopes comprised of residues that are remarkably conserved in both SARS-CoV and SARS-CoV-2 RBDs. This unique feature confers S309 with a broad cross-reactivity, enabling it to effectively target a wide range of variants across both viruses.⁷³

Class 4 antibodies, on the other hand, recognize an epitope within the RBD that displays a high degree of conservation. These antibodies are capable of binding to the RBD; however, they do not directly hinder the binding of ACE2 to the RBD. The epitope of these mAbs is located into a cryptic region, a resemblance that aligns with the cryptic epitope acknowledged by the CR3022 antibody, derived from a patient who had recuperated from SARS-CoV infection. The epitope targeted by class 4 antibodies exhibits a remarkable degree of conservation, sharing up to 86% similarity between SARS-CoV and SARS-CoV-2. This conservation allows CR3022 to effectively bind to both coronaviruses. Notably, due to the presence of a glycosylation site on N370 within the epitope on SARS-CoV, CR3022 binds more avidly to SARS-CoV than to SARS-CoV-2. It's worth mentioning that CR3022's binding to SARS-CoV-2 is contingent upon the 'up' conformation of at least two RBDs.

To summarize, antibodies targeting the RBD and belonging to class 1 and class 2 are likely to lose their effectiveness in neutralizing major variants of SARS-CoV-2 that carry new mutations in the RBM. On the contrary, antibodies falling within class 3 and class 4, which bind to highly conserved epitopes, show promise as potential candidates for neutralizing a variety of SARS-CoV-2 variants and other SARS-like coronaviruses. This observation implies that selecting such conserved epitopes for the design of vaccines has the potential to stimulate the production of robust, broad-spectrum antibodies. These antibodies could play a crucial role in tackling the ongoing COVID-19 pandemic and any potential future outbreaks.

An additional approach to achieving broad-spectrum protection against both SARS-CoV-2 and other human coronaviruses (HCoVs) involves targeting epitopes found in different regions of the S protein, such as the S2 domain where the epitopes seem to be more conserved than in the S1.⁷⁷ An illustrative example of this strategy is the antibody S2P6, which was derived from a COVID-19 patient in recovery. S2P6 demonstrates the ability to broadly neutralize all β -CoVs by specifically targeting the S2 subunit.⁷⁸ Upon further investigation, it was revealed that the epitope recognized by the S2P6 antibody is located within the S2 subunit's SH (stalk helix) region, spanning 14 amino acid residues from 1146 to 1159. Importantly, this epitope exhibits conservation across various β -CoVs. This innovative approach of targeting conserved regions within the S2 domain presents a potential avenue for developing neutralizing antibodies that can offer wide-ranging protection against SARS-CoV-2 and other related coronaviruses.

In addition to the previously mentioned SH region, the S2 fusion peptide (FP) region also exhibits significant conservation across all genera of coronaviruses. This suggests that the

FP epitope could be a promising target for the development of broad-spectrum antibodies. Recent discoveries have identified antibodies that exhibit potent broadly neutralizing activity against Alpha-CoVs, Beta but also Gamma and Delta.^{79, 80} For instance, antibodies COV44-62 and COV44-79, both isolated from individuals recovering from COVID-19, can bind to the S2 FP region.⁷⁹ Interestingly, these antibodies do not compete with S2P6, the previously mentioned antibody targeting the S2 SH region, for binding to the SARS-CoV-2 S protein. This intriguing observation suggests the potential for developing a bispecific antibody that combines the recognition of both the S2 SH and S2 FP regions. Such a bispecific antibody could offer enhanced neutralizing capabilities, targeting multiple conserved epitopes and thereby providing a more comprehensive defense against a broader range of coronaviruses.

Such understanding can be exploited to design and engineer improved antigens based on S, for instance by identifying antigenic domains that can be expressed in isolation or short sequences (epitopes) that can be mimicked by synthetic peptides: this would be a crucial first step in the selection and optimization of candidate vaccines and therapeutic antibodies (on top of those already in development), as well as in the development of additional serologic diagnostic tools. And this is what we have done in the first paper (see Chapter 5.1).

1.6 References

- (1) Huang, C.; Wang, Y.; Li, X.; Ren, L.; Zhao, J.; Hu, Y.; Zhang, L.; Fan, G.; Xu, J.; Gu, X.; et al. Clinical features of patients infected with 2019 novel coronavirus in Wuhan, China. *Lancet* **2020**, 395 (10223), 497-506. DOI: 10.1016/S0140-6736(20)30183-5.
- (2) Wu, Z.; McGoogan, J. M. Characteristics of and Important Lessons From the Coronavirus Disease 2019 (COVID-19) Outbreak in China: Summary of a Report of 72 314 Cases From the Chinese Center for Disease Control and Prevention. *JAMA* **2020**, 323 (13), 1239-1242. DOI: 10.1001/jama.2020.2648.
- (3) Zhu, N.; Zhang, D.; Wang, W.; Li, X.; Yang, B.; Song, J.; Zhao, X.; Huang, B.; Shi, W.; Lu, R.; et al. A Novel Coronavirus from Patients with Pneumonia in China, 2019. *N Engl J Med* **2020**, 382 (8), 727-733. DOI: 10.1056/NEJMoa2001017.
- (4) Wu, F.; Zhao, S.; Yu, B.; Chen, Y. M.; Wang, W.; Song, Z. G.; Hu, Y.; Tao, Z. W.; Tian, J. H.; Pei, Y. Y.; et al. Author Correction: A new coronavirus associated with human respiratory disease in China. *Nature* **2020**, 580 (7803), E7. DOI: 10.1038/s41586-020-2202-3.
- (5) Zhou, P.; Yang, X. L.; Wang, X. G.; Hu, B.; Zhang, L.; Zhang, W.; Si, H. R.; Zhu, Y.; Li, B.; Huang, C. L.; et al. A pneumonia outbreak associated with a new coronavirus of probable bat origin. *Nature* **2020**, 579 (7798), 270-273. DOI: 10.1038/s41586-020-2012-7.
- (6) Gralinski, L. E.; Menachery, V. D. Return of the Coronavirus: 2019-nCoV. *Viruses* **2020**, 12 (2). DOI: 10.3390/v12020135.
- (7) team, E. e. Note from the editors: World Health Organization declares novel coronavirus (2019-nCoV) sixth public health emergency of international concern. *Euro Surveill* **2020**, 25 (5). DOI: 10.2807/1560-7917.ES.2020.25.5.200131e.
- (8) Viruses, C. S. G. o. t. I. C. o. T. o. The species Severe acute respiratory syndrome-related coronavirus: classifying 2019-nCoV and naming it SARS-CoV-2. *Nat Microbiol* **2020**, 5 (4), 536-544. DOI: 10.1038/s41564-020-0695-z.
- (9) Li, Q.; Guan, X.; Wu, P.; Wang, X.; Zhou, L.; Tong, Y.; Ren, R.; Leung, K. S. M.; Lau, E. H. Y.; Wong, J. Y.; et al. Early Transmission Dynamics in Wuhan, China, of Novel Coronavirus-Infected Pneumonia. *N Engl J Med* **2020**, 382 (13), 1199-1207. DOI: 10.1056/NEJMoa2001316.
- (10) de Wit, E.; van Doremalen, N.; Falzarano, D.; Munster, V. J. SARS and MERS: recent insights into emerging coronaviruses. *Nat Rev Microbiol* **2016**, 14 (8), 523-534. DOI: 10.1038/nrmicro.2016.81.
- (11) Liu, D. X.; Fung, T. S.; Chong, K. K.; Shukla, A.; Hilgenfeld, R. Accessory proteins of SARS-CoV and other coronaviruses. *Antiviral Res* **2014**, 109, 97-109. DOI: 10.1016/j.antiviral.2014.06.013.
- (12) Lu, R.; Zhao, X.; Li, J.; Niu, P.; Yang, B.; Wu, H.; Wang, W.; Song, H.; Huang, B.; Zhu, N.; et al. Genomic characterisation and epidemiology of 2019 novel coronavirus: implications for virus origins and receptor binding. *Lancet* **2020**, 395 (10224), 565-574. DOI: 10.1016/S0140-6736(20)30251-8.
- (13) V'kovski, P.; Kratzel, A.; Steiner, S.; Stalder, H.; Thiel, V. Coronavirus biology and replication: implications for SARS-CoV-2. *Nat Rev Microbiol* **2021**, 19 (3), 155-170. DOI: 10.1038/s41579-020-00468-6.
- (14) Huang, Y.; Yang, C.; Xu, X. F.; Xu, W.; Liu, S. W. Structural and functional properties of SARS-CoV-2 spike protein: potential antiviral drug development for COVID-19. *Acta Pharmacol Sin* **2020**, 41 (9), 1141-1149. DOI: 10.1038/s41401-020-0485-4.
- (15) Zhang, J.; Xiao, T.; Cai, Y.; Chen, B. Structure of SARS-CoV-2 spike protein. *Curr Opin Virol* **2021**, 50, 173-182. DOI: 10.1016/j.coviro.2021.08.010.
- (16) Chi, X.; Yan, R.; Zhang, J.; Zhang, G.; Zhang, Y.; Hao, M.; Zhang, Z.; Fan, P.; Dong, Y.; Yang, Y.; et al. A neutralizing human antibody binds to the N-terminal domain of the Spike protein of SARS-CoV-2. *Science* **2020**, 369 (6504), 650-655. DOI: 10.1126/science.abc6952.
- (17) Zhang, J.; Cai, Y.; Xiao, T.; Lu, J.; Peng, H.; Sterling, S. M.; Walsh, R. M.; Rits-Volloch, S.; Zhu, H.; Woosley, A. N.; et al. Structural impact on SARS-CoV-2 spike protein by D614G substitution. *Science* **2021**, 372 (6541), 525-530. DOI: 10.1126/science.abf2303.
- (18) Zhang, J.; Xiao, T.; Cai, Y.; Lavine, C. L.; Peng, H.; Zhu, H.; Anand, K.; Tong, P.; Gautam, A.; Mayer, M. L.; et al. Membrane fusion and immune evasion by the spike protein of SARS-CoV-2 Delta variant. *Science* **2021**, 374 (6573), 1353-1360. DOI: 10.1126/science.abl9463.
- (19) Zhang, J.; Cai, Y.; Lavine, C. L.; Peng, H.; Zhu, H.; Anand, K.; Tong, P.; Gautam, A.; Mayer, M. L.; Rits-Volloch, S.; et al. Structural and functional impact by SARS-CoV-2 Omicron spike mutations. *Cell Rep* **2022**, 39 (4), 110729. DOI: 10.1016/j.celrep.2022.110729.
- (20) Watanabe, Y.; Bowden, T. A.; Wilson, I. A.; Crispin, M. Exploitation of glycosylation in enveloped virus pathobiology. *Biochim Biophys Acta Gen Subj* **2019**, 1863 (10), 1480-1497. DOI: 10.1016/j.bbagen.2019.05.012.

- (21) Watanabe, Y.; Berndsen, Z. T.; Raghvani, J.; Seabright, G. E.; Allen, J. D.; Pybus, O. G.; McLellan, J. S.; Wilson, I. A.; Bowden, T. A.; Ward, A. B.; et al. Vulnerabilities in coronavirus glycan shields despite extensive glycosylation. *Nat Commun* **2020**, *11* (1), 2688. DOI: 10.1038/s41467-020-16567-0.
- (22) Raman, R.; Tharakaraman, K.; Sasisekharan, V.; Sasisekharan, R. Glycan-protein interactions in viral pathogenesis. *Curr Opin Struct Biol* **2016**, *40*, 153-162. DOI: 10.1016/j.sbi.2016.10.003.
- (23) Doores, K. J.; Bonomelli, C.; Harvey, D. J.; Vasiljevic, S.; Dwek, R. A.; Burton, D. R.; Crispin, M.; Scanlan, C. N. Envelope glycans of immunodeficiency virions are almost entirely oligomannose antigens. *Proc Natl Acad Sci U S A* **2010**, *107* (31), 13800-13805. DOI: 10.1073/pnas.1006498107.
- (24) Crispin, M.; Ward, A. B.; Wilson, I. A. Structure and Immune Recognition of the HIV Glycan Shield. *Annu Rev Biophys* **2018**, *47*, 499-523. DOI: 10.1146/annurev-biophys-060414-034156.
- (25) Stewart-Jones, G. B.; Soto, C.; Lemmin, T.; Chuang, G. Y.; Druz, A.; Kong, R.; Thomas, P. V.; Wagh, K.; Zhou, T.; Behrens, A. J.; et al. Trimeric HIV-1-Env Structures Define Glycan Shields from Clades A, B, and G. *Cell* **2016**, *165* (4), 813-826. DOI: 10.1016/j.cell.2016.04.010.
- (26) Yang, M.; Huang, J.; Simon, R.; Wang, L. X.; MacKerell, A. D. Conformational Heterogeneity of the HIV Envelope Glycan Shield. *Sci Rep* **2017**, *7* (1), 4435. DOI: 10.1038/s41598-017-04532-9.
- (27) Cunningham, A. L.; Harman, A. N.; Donaghy, H. DC-SIGN 'AIDS' HIV immune evasion and infection. *Nat Immunol* **2007**, *8* (6), 556-558. DOI: 10.1038/ni0607-556.
- (28) Goncalves, A. R.; Moraz, M. L.; Pasquato, A.; Helenius, A.; Lozach, P. Y.; Kunz, S. Role of DC-SIGN in Lassa virus entry into human dendritic cells. *J Virol* **2013**, *87* (21), 11504-11515. DOI: 10.1128/JVI.01893-13.
- (29) Seitz, C.; Casalino, L.; Konecny, R.; Huber, G.; Amaro, R. E.; McCammon, J. A. Multiscale Simulations Examining Glycan Shield Effects on Drug Binding to Influenza Neuraminidase. *Biophys J* **2020**, *119* (11), 2275-2289. DOI: 10.1016/j.bpj.2020.10.024.
- (30) Altman, M. O.; Angel, M.; Košík, I.; Trovão, N. S.; Zost, S. J.; Gibbs, J. S.; Casalino, L.; Amaro, R. E.; Hensley, S. E.; Nelson, M. I.; et al. Human Influenza A Virus Hemagglutinin Glycan Evolution Follows a Temporal Pattern to a Glycan Limit. *mBio* **2019**, *10* (2). DOI: 10.1128/mBio.00204-19.
- (31) Shajahan, A.; Pepi, L. E.; Rouhani, D. S.; Heiss, C.; Azadi, P. Glycosylation of SARS-CoV-2: structural and functional insights. *Anal Bioanal Chem* **2021**, *413* (29), 7179-7193. DOI: 10.1007/s00216-021-03499-x.
- (32) Casalino, L.; Gaieb, Z.; Goldsmith, J. A.; Hjorth, C. K.; Dommer, A. C.; Harbison, A. M.; Fogarty, C. A.; Barros, E. P.; Taylor, B. C.; McLellan, J. S.; et al. Beyond Shielding: The Roles of Glycans in the SARS-CoV-2 Spike Protein. *ACS Cent Sci* **2020**, *6* (10), 1722-1734. DOI: 10.1021/acscentsci.0c01056.
- (33) Alharbi, S. N.; Alrefaei, A. F. Comparison of the SARS-CoV-2 (2019-nCoV) M protein with its counterparts of SARS-CoV and MERS-CoV species. *J King Saud Univ Sci* **2021**, *33* (2), 101335. DOI: 10.1016/j.jksus.2020.101335.
- (34) Gong, Y.; Qin, S.; Dai, L.; Tian, Z. The glycosylation in SARS-CoV-2 and its receptor ACE2. *Signal Transduct Target Ther* **2021**, *6* (1), 396. DOI: 10.1038/s41392-021-00809-8.
- (35) Wong, N. A.; Saier, M. H. The SARS-Coronavirus Infection Cycle: A Survey of Viral Membrane Proteins, Their Functional Interactions and Pathogenesis. *Int J Mol Sci* **2021**, *22* (3). DOI: 10.3390/ijms22031308.
- (36) Lu, S.; Ye, Q.; Singh, D.; Cao, Y.; Diedrich, J. K.; Yates, J. R.; Villa, E.; Cleveland, D. W.; Corbett, K. D. The SARS-CoV-2 nucleocapsid phosphoprotein forms mutually exclusive condensates with RNA and the membrane-associated M protein. *Nat Commun* **2021**, *12* (1), 502. DOI: 10.1038/s41467-020-20768-y.
- (37) Kang, S.; Yang, M.; Hong, Z.; Zhang, L.; Huang, Z.; Chen, X.; He, S.; Zhou, Z.; Chen, Q.; Yan, Y.; et al. Crystal structure of SARS-CoV-2 nucleocapsid protein RNA binding domain reveals potential unique drug targeting sites. *Acta Pharm Sin B* **2020**, *10* (7), 1228-1238. DOI: 10.1016/j.apsb.2020.04.009.
- (38) Cascarina, S. M.; Ross, E. D. A proposed role for the SARS-CoV-2 nucleocapsid protein in the formation and regulation of biomolecular condensates. *FASEB J* **2020**, *34* (8), 9832-9842. DOI: 10.1096/fj.202001351.
- (39) Perdikari, T. M.; Murthy, A. C.; Ryan, V. H.; Watters, S.; Naik, M. T.; Fawzi, N. L. SARS-CoV-2 nucleocapsid protein undergoes liquid-liquid phase separation stimulated by RNA and partitions into phases of human ribonucleoproteins. *bioRxiv* **2020**. DOI: 10.1101/2020.06.09.141101.
- (40) Gao, T.; Gao, Y.; Liu, X.; Nie, Z.; Sun, H.; Lin, K.; Peng, H.; Wang, S. Identification and functional analysis of the SARS-COV-2 nucleocapsid protein. *BMC Microbiol* **2021**, *21* (1), 58. DOI: 10.1186/s12866-021-02107-3.
- (41) Smits, V. A. J.; Hernández-Carralero, E.; Paz-Cabrera, M. C.; Cabrera, E.; Hernández-Reyes, Y.; Hernández-Fernaud, J. R.; Gillespie, D. A.; Salido, E.; Hernández-Porto, M.; Freire, R. The Nucleocapsid protein triggers the main humoral immune response in COVID-19 patients. *Biochem Biophys Res Commun* **2021**, *543*, 45-49. DOI: 10.1016/j.bbrc.2021.01.073.
- (42) Dai, L.; Gao, G. F. Viral targets for vaccines against COVID-19. *Nat Rev Immunol* **2021**, *21* (2), 73-82. DOI: 10.1038/s41577-020-00480-0.

- (43) Kumar, P.; Kumar, A.; Garg, N.; Giri, R. An insight into SARS-CoV-2 membrane protein interaction with spike, envelope, and nucleocapsid proteins. *J Biomol Struct Dyn* **2023**, *41* (3), 1062-1071. DOI: 10.1080/07391102.2021.2016490.
- (44) Wang, W.; Chen, J.; Yu, X.; Lan, H. Y. Signaling mechanisms of SARS-CoV-2 Nucleocapsid protein in viral infection, cell death and inflammation. *Int J Biol Sci* **2022**, *18* (12), 4704-4713. DOI: 10.7150/ijbs.72663.
- (45) Khanmohammadi, S.; Rezaei, N. Role of Toll-like receptors in the pathogenesis of COVID-19. *J Med Virol* **2021**, *93* (5), 2735-2739. DOI: 10.1002/jmv.26826.
- (46) Zhou, Y.; Little, P. J.; Downey, L.; Afroz, R.; Wu, Y.; Ta, H. T.; Xu, S.; Kamato, D. The Role of Toll-like Receptors in Atherothrombotic Cardiovascular Disease. *ACS Pharmacol Transl Sci* **2020**, *3* (3), 457-471. DOI: 10.1021/acspsci.9b00100.
- (47) Chen, H.; Cui, Y.; Han, X.; Hu, W.; Sun, M.; Zhang, Y.; Wang, P. H.; Song, G.; Chen, W.; Lou, J. Liquid-liquid phase separation by SARS-CoV-2 nucleocapsid protein and RNA. *Cell Res* **2020**, *30* (12), 1143-1145. DOI: 10.1038/s41422-020-00408-2.
- (48) Duart, G.; García-Murria, M. J.; Mingarro, I. The SARS-CoV-2 envelope (E) protein has evolved towards membrane topology robustness. *Biochim Biophys Acta Biomembr* **2021**, *1863* (7), 183608. DOI: 10.1016/j.bbmem.2021.183608.
- (49) Tilocca, B.; Soggiu, A.; Sanguinetti, M.; Babini, G.; De Maio, F.; Britti, D.; Zecconi, A.; Bonizzi, L.; Urbani, A.; Roncada, P. Immunoinformatic analysis of the SARS-CoV-2 envelope protein as a strategy to assess cross-protection against COVID-19. *Microbes Infect* **2020**, *22* (4-5), 182-187. DOI: 10.1016/j.micinf.2020.05.013.
- (50) Stodola, J. K.; Dubois, G.; Le Coupanec, A.; Desforges, M.; Talbot, P. J. The OC43 human coronavirus envelope protein is critical for infectious virus production and propagation in neuronal cells and is a determinant of neurovirulence and CNS pathology. *Virology* **2018**, *515*, 134-149. DOI: 10.1016/j.virol.2017.12.023.
- (51) Schoeman, D.; Fielding, B. C. Coronavirus envelope protein: current knowledge. *Virology* **2019**, *16* (1), 69. DOI: 10.1186/s12985-019-1182-0.
- (52) Torres, J.; Wang, J.; Parthasarathy, K.; Liu, D. X. The transmembrane oligomers of coronavirus protein E. *Biophys J* **2005**, *88* (2), 1283-1290. DOI: 10.1529/biophysj.104.051730.
- (53) Duart, G.; García-Murria, M. J.; Grau, B.; Acosta-Cáceres, J. M.; Martínez-Gil, L.; Mingarro, I. SARS-CoV-2 envelope protein topology in eukaryotic membranes. *Open Biol* **2020**, *10* (9), 200209. DOI: 10.1098/rsob.200209.
- (54) Zhou, S.; Lv, P.; Li, M.; Chen, Z.; Xin, H.; Reilly, S.; Zhang, X. SARS-CoV-2 E protein: Pathogenesis and potential therapeutic development. *Biomed Pharmacother* **2023**, *159*, 114242. DOI: 10.1016/j.biopha.2023.114242.
- (55) Chen, Y.; Liu, Q.; Zhou, L.; Zhou, Y.; Yan, H.; Lan, K. Emerging SARS-CoV-2 variants: Why, how, and what's next? *Cell Insight* **2022**, *1* (3), 100029. DOI: 10.1016/j.cellin.2022.100029.
- (56) Scovino, A. M.; Dahab, E. C.; Vieira, G. F.; Freire-de-Lima, L.; Freire-de-Lima, C. G.; Morrot, A. SARS-CoV-2's Variants of Concern: A Brief Characterization. *Front Immunol* **2022**, *13*, 834098. DOI: 10.3389/fimmu.2022.834098.
- (57) Harvey, W. T.; Carabelli, A. M.; Jackson, B.; Gupta, R. K.; Thomson, E. C.; Harrison, E. M.; Ludden, C.; Reeve, R.; Rambaut, A.; Peacock, S. J.; et al. SARS-CoV-2 variants, spike mutations and immune escape. *Nat Rev Microbiol* **2021**, *19* (7), 409-424. DOI: 10.1038/s41579-021-00573-0.
- (58) Korber, B.; Fischer, W. M.; Gnanakaran, S.; Yoon, H.; Theiler, J.; Abfalterer, W.; Hengartner, N.; Giorgi, E. E.; Bhattacharya, T.; Foley, B.; et al. Tracking Changes in SARS-CoV-2 Spike: Evidence that D614G Increases Infectivity of the COVID-19 Virus. *Cell* **2020**, *182* (4), 812-827.e819. DOI: 10.1016/j.cell.2020.06.043.
- (59) Warren, T. K.; Jordan, R.; Lo, M. K.; Ray, A. S.; Mackman, R. L.; Soloveva, V.; Siegel, D.; Perron, M.; Bannister, R.; Hui, H. C.; et al. Therapeutic efficacy of the small molecule GS-5734 against Ebola virus in rhesus monkeys. *Nature* **2016**, *531* (7594), 381-385. DOI: 10.1038/nature17180.
- (60) Dolgin, E. The tangled history of mRNA vaccines. *Nature* **2021**, *597* (7876), 318-324. DOI: 10.1038/d41586-021-02483-w.
- (61) Triveri, A.; Serapian, S. A.; Marchetti, F.; Doria, F.; Pavoni, S.; Cinquini, F.; Moroni, E.; Rasola, A.; Frigerio, F.; Colombo, G. SARS-CoV-2 Spike Protein Mutations and Escape from Antibodies: A Computational Model of Epitope Loss in Variants of Concern. *J Chem Inf Model* **2021**, *61* (9), 4687-4700. DOI: 10.1021/acs.jcim.1c00857.
- (62) VanBlargan, L. A.; Errico, J. M.; Halfmann, P. J.; Zost, S. J.; Crowe, J. E.; Purcell, L. A.; Kawaoka, Y.; Corti, D.; Fremont, D. H.; Diamond, M. S. An infectious SARS-CoV-2 B.1.1.529 Omicron virus escapes neutralization by therapeutic monoclonal antibodies. *Nat Med* **2022**, *28* (3), 490-495. DOI: 10.1038/s41591-021-01678-y.

- (63) Focosi, D.; McConnell, S.; Casadevall, A.; Cappello, E.; Valdiserra, G.; Tuccori, M. Monoclonal antibody therapies against SARS-CoV-2. *Lancet Infect Dis* **2022**, *22* (11), e311-e326. DOI: 10.1016/S1473-3099(22)00311-5.
- (64) Barnes, C. O.; Jette, C. A.; Abernathy, M. E.; Dam, K. A.; Esswein, S. R.; Gristick, H. B.; Malyutin, A. G.; Sharaf, N. G.; Huey-Tubman, K. E.; Lee, Y. E.; et al. SARS-CoV-2 neutralizing antibody structures inform therapeutic strategies. *Nature* **2020**, *588* (7839), 682-687. DOI: 10.1038/s41586-020-2852-1.
- (65) Chen, Y.; Zhao, X.; Zhou, H.; Zhu, H.; Jiang, S.; Wang, P. Broadly neutralizing antibodies to SARS-CoV-2 and other human coronaviruses. *Nat Rev Immunol* **2023**, *23* (3), 189-199. DOI: 10.1038/s41577-022-00784-3.
- (66) Starr, T. N.; Czudnochowski, N.; Liu, Z.; Zatta, F.; Park, Y. J.; Addetia, A.; Pinto, D.; Beltramello, M.; Hernandez, P.; Greaney, A. J.; et al. SARS-CoV-2 RBD antibodies that maximize breadth and resistance to escape. *Nature* **2021**, *597* (7874), 97-102. DOI: 10.1038/s41586-021-03807-6.
- (67) Zhou, T.; Wang, L.; Misasi, J.; Pegu, A.; Zhang, Y.; Harris, D. R.; Olia, A. S.; Talana, C. A.; Yang, E. S.; Chen, M.; et al. Structural basis for potent antibody neutralization of SARS-CoV-2 variants including B.1.1.529. *Science* **2022**, *376* (6591), eabn8897. DOI: 10.1126/science.abn8897.
- (68) Chen, P.; Nirula, A.; Heller, B.; Gottlieb, R. L.; Boscia, J.; Morris, J.; Huhn, G.; Cardona, J.; Mocherla, B.; Stosor, V.; et al. SARS-CoV-2 Neutralizing Antibody LY-CoV555 in Outpatients with Covid-19. *N Engl J Med* **2021**, *384* (3), 229-237. DOI: 10.1056/NEJMoa2029849.
- (69) Liu, L.; Wang, P.; Nair, M. S.; Yu, J.; Rapp, M.; Wang, Q.; Luo, Y.; Chan, J. F.; Sahi, V.; Figueroa, A.; et al. Potent neutralizing antibodies against multiple epitopes on SARS-CoV-2 spike. *Nature* **2020**, *584* (7821), 450-456. DOI: 10.1038/s41586-020-2571-7.
- (70) Hansen, J.; Baum, A.; Pascal, K. E.; Russo, V.; Giordano, S.; Wloga, E.; Fulton, B. O.; Yan, Y.; Koon, K.; Patel, K.; et al. Studies in humanized mice and convalescent humans yield a SARS-CoV-2 antibody cocktail. *Science* **2020**, *369* (6506), 1010-1014. DOI: 10.1126/science.abd0827.
- (71) Zost, S. J.; Gilchuk, P.; Case, J. B.; Binshtein, E.; Chen, R. E.; Nkolola, J. P.; Schäfer, A.; Reidy, J. X.; Trivette, A.; Nargi, R. S.; et al. Potently neutralizing and protective human antibodies against SARS-CoV-2. *Nature* **2020**, *584* (7821), 443-449. DOI: 10.1038/s41586-020-2548-6.
- (72) Cerutti, G.; Rapp, M.; Guo, Y.; Bahna, F.; Bimela, J.; Reddem, E. R.; Yu, J.; Wang, P.; Liu, L.; Huang, Y.; et al. Structural basis for accommodation of emerging B.1.351 and B.1.1.7 variants by two potent SARS-CoV-2 neutralizing antibodies. *Structure* **2021**, *29* (7), 655-663.e654. DOI: 10.1016/j.str.2021.05.014.
- (73) Pinto, D.; Park, Y. J.; Beltramello, M.; Walls, A. C.; Tortorici, M. A.; Bianchi, S.; Jaconi, S.; Culap, K.; Zatta, F.; De Marco, A.; et al. Cross-neutralization of SARS-CoV-2 by a human monoclonal SARS-CoV antibody. *Nature* **2020**, *583* (7815), 290-295. DOI: 10.1038/s41586-020-2349-y.
- (74) Westendorf, K.; Žentelis, S.; Wang, L.; Foster, D.; Vaillancourt, P.; Wiggin, M.; Lovett, E.; van der Lee, R.; Hendle, J.; Pustilnik, A.; et al. LY-CoV1404 (bebtelovimab) potently neutralizes SARS-CoV-2 variants. *Cell Rep* **2022**, *39* (7), 110812. DOI: 10.1016/j.celrep.2022.110812.
- (75) Wang, L.; Zhou, T.; Zhang, Y.; Yang, E. S.; Schramm, C. A.; Shi, W.; Pegu, A.; Oloniniyi, O. K.; Henry, A. R.; Darko, S.; et al. Ultrapotent antibodies against diverse and highly transmissible SARS-CoV-2 variants. *Science* **2021**, *373* (6556). DOI: 10.1126/science.abh1766.
- (76) Fenwick, C.; Turelli, P.; Ni, D.; Perez, L.; Lau, K.; Herate, C.; Marlin, R.; Lana, E.; Pellaton, C.; Raclot, C.; et al. Patient-derived monoclonal antibody neutralizes SARS-CoV-2 Omicron variants and confers full protection in monkeys. *Nat Microbiol* **2022**, *7* (9), 1376-1389. DOI: 10.1038/s41564-022-01198-6.
- (77) Shrestha, L. B.; Tedla, N.; Bull, R. A. Broadly-Neutralizing Antibodies Against Emerging SARS-CoV-2 Variants. *Front Immunol* **2021**, *12*, 752003. DOI: 10.3389/fimmu.2021.752003.
- (78) Pinto, D.; Sauer, M. M.; Czudnochowski, N.; Low, J. S.; Tortorici, M. A.; Housley, M. P.; Noack, J.; Walls, A. C.; Bowen, J. E.; Guarino, B.; et al. Broad betacoronavirus neutralization by a stem helix-specific human antibody. *Science* **2021**, *373* (6559), 1109-1116. DOI: 10.1126/science.abj3321.
- (79) Dacon, C.; Tucker, C.; Peng, L.; Lee, C. D.; Lin, T. H.; Yuan, M.; Cong, Y.; Wang, L.; Purser, L.; Williams, J. K.; et al. Broadly neutralizing antibodies target the coronavirus fusion peptide. *Science* **2022**, *377* (6607), 728-735. DOI: 10.1126/science.abq3773.
- (80) Low, J. S.; Jerak, J.; Tortorici, M. A.; McCallum, M.; Pinto, D.; Cassotta, A.; Foglierini, M.; Mele, F.; Abdelnabi, R.; Weynand, B.; et al. ACE2-binding exposes the SARS-CoV-2 fusion peptide to broadly neutralizing coronavirus antibodies. *Science* **2022**, *377* (6607), 735-742. DOI: 10.1126/science.abq2679.

Materials and methods

1.7 Molecular dynamics (MD)

The fundamental idea underlying a molecular dynamics (MD) simulation is simple: atoms and molecules interact and exhibit real-life movements, which MD simulations replicate. This notion entails that by inputting the spatial coordinates of all atoms within a biomolecular system (like a protein enveloped by water and potentially a lipid bilayer), it becomes feasible to calculate the impact or force exerted on each atom, resulting from interactions with all the surrounding atoms.

Computational simulations are meant to be a link connecting the minuscule dimensions and time frames of the microscopic realm to the broader, observable domain of the laboratory. Statistical mechanics, a fundamental branch of the physical sciences, equips us with the mathematical and theoretical apparatus necessary for establishing a seamless connection between the intricacies of the microscale and the substantial measurements of the macroscopic world.

In the context of a system comprising N particles, its characterization involves a set of atomic positions denoted as $(\vec{R} = \{\vec{R}_1, \dots, \vec{R}_N\})$ and corresponding relative momenta designated as $(\vec{P} = \{\vec{P}_1, \dots, \vec{P}_N\})$. Together, these parameters define the microscopic state of the system. This state can be visualized and represented as a singular point within a multidimensional space of $6N$ dimensions, referred to as the phase space (Γ). Consequently, a solitary point in this phase space corresponds to a specific microscopic configuration of the system, while an aggregation of points in Γ constitutes an ensemble.

MD simulations serve as a practical technique that generates a sequence of points within the phase space over time. In essence, MD simulations provide a sequence of diverse positions and momenta for the system, all belonging to the same ensemble. For any given microscopic state of the system within the phase space, it becomes feasible to estimate the value of an observable property A as a function of Γ , denoted as $A(\Gamma)$. This estimation is performed through either the ensemble average or the thermodynamic average calculation:

$$A_{obs} = \langle A \rangle_{ens} = \int A(\Gamma) \rho(\Gamma) d\Gamma \quad (0.1)$$

where $\rho(\Gamma)$ is the probability distribution function of collection of points Γ , and $d\Gamma = d\vec{R}_1 \dots d\vec{R}_N d\vec{P}_1 \dots d\vec{P}_N$. The probability distribution function depends on macroscopic

parameters that delineate the thermodynamic state of a system, such as the particle count (N), volume (V), temperature (T), and pressure (P). For instance, within the canonical ensemble (NVT), wherein N, V, and T remain constant, the probability distribution function adopts the structure of the Boltzmann distribution function:

$$\rho_{NVT} = \frac{e^{\frac{-H(\Gamma)}{K_B T}}}{Z} \quad (0.2)$$

where $H(\Gamma)$ is the classical Hamiltonian of the system defined as:

$$H(\Gamma) = H(\{\vec{R}_1\}, \{\vec{P}_1\}) = \sum_{I=1}^N \frac{\vec{P}_1^2}{2M_I} + U(\{\vec{R}_1\}) \quad (0.3)$$

Where \vec{R}_1 , \vec{P}_1 and M_I are the position, momentum and mass of the particle I, U is the potential energy, K_B is the Boltzmann constant and Z is the canonical partition function.

MD simulations offer a method to approximate ensemble averages by directly integrating Newton's equations of motion. This entails evolving the system over time, commencing from its microstate at time 0 and progressing to its microstate at time τ . Consequently, a sequence of microstates for the system, forming a trajectory of points in the phase space $\Gamma(t)$, is generated. From this trajectory, the time-averaged value of an observable ($\langle A \rangle_\tau$) can be computed, and it is connected to the ensemble average (denoted as $\langle A \rangle_{obs}$) according to the "ergodic hypothesis". In essence, this hypothesis posits that if the system evolves over an infinitely extended duration, it should be capable of visit all feasible states, thereby causing its behavior averaged across both time and the phase space to converge:

$$\lim_{\tau \rightarrow \infty} \langle A(\Gamma) \rangle_\tau = \lim_{\tau \rightarrow \infty} \frac{1}{\tau} \int_0^\tau A[\Gamma(t)] dt = \langle A(\Gamma) \rangle_\Gamma \quad (0.4)$$

The more extended the simulation time, the more accurately this equality is upheld. Consequently, employing MD simulations in the context of molecular biological systems provides a direct and practical approach to anticipate their average conduct and appraise macroscopic observables.

The progression of the system over time is determined by numerically integrating the second set of Newton's equations of motion, wherein atoms are modeled as point particles:

$$\vec{F}_i = M_i \vec{a}_i \quad (0.5)$$

where \vec{F}_i is the force acting on particle i , M_i is the mass of the particle i and \vec{a}_i the second derivative of the particle's position with respect to time t , i.e. its acceleration.

The atoms will move under the influence of the internal forces acting on them, which are derived from the potential energy of the system ($U(\vec{R})$):

$$\vec{F}_i = -\frac{\partial U(\vec{R})}{\partial \vec{R}_i} \quad (0.6)$$

Newton's equations serve as a comprehensive framework, definitively outlining the complete array of positions and velocities as functions of time. This intricate interplay precisely delineates the classical state of the system at any given time, t . An analytical solution to the equations of motion, as expressed in equation is acquired through a predetermined set of initial conditions dictating the positions and velocities of the particles. The formers are often derived from PDB crystal structures, NMR data, cryo-electron microscopy (cryo-EM), or homology modeling coordinates. On the other hand, the latter are typically generated in a random manner, adhering to the Maxwell-Boltzmann probability distribution at a designated temperature, T .

To facilitate this computational process, discretized numerical algorithms are enlisted to iteratively update the particles' positions and velocities at each discrete time step, denoted as Δt . This value is established at the beginning of the simulation and typically falls within the range of 1 to 2 femtoseconds (fs). This choice ensures the stable and precise integration of even the swiftest motions present within the system. The three open MD algorithms that are most widely used in molecular dynamics studies are the Verlet, leap-frog and Beeman algorithms.

The velocity-**Verlet** algorithm^{1, 2} exploits a truncated Taylor expansion, limited beyond the quadratic term, for the coordinates:

$$\vec{R}(t + \Delta t) \approx \vec{R}(t) + \vec{v}(t)\Delta t + \frac{\vec{F}(t)}{2m}\Delta t^2 \quad (0.7)$$

And the velocities \vec{v} considering the relation $\vec{P} = m\vec{v}$:

$$\vec{v}(t)\Delta t \approx \vec{v} + \frac{\vec{F}(t) + \vec{F}(t + \Delta t)}{2m}\Delta t \quad (0.8)$$

The **Leap-Frog** algorithm³ uses velocities at half-integer time steps to determine new particles' positions:

$$\vec{v}\left(t + \frac{\Delta t}{2}\right) = \vec{v}\left(t - \frac{\Delta t}{2}\right) + \frac{\vec{F}(t)}{m}\Delta t + O(\Delta t^3) \quad (0.9)$$

$$\vec{R}(t + \Delta t) = \vec{R}(t) + \vec{v}\left(t + \frac{\Delta t}{2}\right)\Delta t + O(\Delta t^3) \quad (0.20)$$

This algorithm calculates positions and forces at interleaved time points. Therefore, kinetic and potential energy are also not defined at the same time.

The **Beeman** algorithm looks very different, more complicated and requires more storage than the other two equivalent algorithms there is no reason to use it.

1.7.1 Force field

In force-field based MD, which is the classical MD, the force field constitutes a mathematical formulation that delineates how a system's energy hinges upon the coordinates of its constituent particles. It encompasses an analytical representation of the interatomic potential energy, denoted as $U(r_1, \dots, r_N)$, along with an ensemble of parameters incorporated into this functional form. Typically, these parameters are ascertained through ab initio or semi-empirical quantum mechanical calculations, or alternatively by fitting to experimental data encompassing techniques such as neutron scattering, X-ray and electron diffraction, NMR, infrared, Raman, and neutron spectroscopy, among others.

Molecules are essentially characterized as assemblies of atoms held together by simple elastic (harmonic) forces. A force field takes the place of the authentic potential energy with a streamlined model that holds validity within the specific simulated region. Ideally, a force field should possess the dual attributes of computational efficiency, allowing rapid evaluation, and sufficient intricacy to accurately reproduce the pertinent properties of the system under scrutiny.

Numerous force fields are documented in the scientific literature, differing in their levels of complexity and tailored to address various types of systems, for biomolecular applications such as AMBER,⁴ GROMOS,⁵ CHARMM⁶ etc.

However, a prototypical expression for a force field might be structured as follows:

$$\begin{aligned}
U = & \sum_{bonds} \frac{1}{2} k_b (r - r_0)^2 \\
& + \sum_{angles} \frac{1}{2} k_a (\theta - \theta_0)^2 + \sum_{torsions} \frac{V_n}{2} [1 + \cos(n\phi - \delta)] + \sum_{improper} V_{i\ mp} \\
& + \sum_{improper} 4\epsilon_{ij} \left(\frac{\sigma_{ij}^{12}}{r_{ij}^{12}} - \frac{\sigma_{ij}^6}{r_{ij}^6} \right) + \sum_{elec} \frac{q_i q_j}{r_{ij}}
\end{aligned} \tag{0.31}$$

where the first four terms refer to intramolecular or local contributions to the total energy (bond stretching, angle bending, and dihedral and improper torsions), and the last two terms serve to describe the non-bonded interactions corresponding to the repulsive and Van der Waals (vdW) interactions and the Coulomb potential for the electrostatic interactions.

In practical molecular dynamics (MD) simulations, finite systems are employed, necessitating a specialized approach to address potential issues arising from particles located at the system boundaries, which could coincide with the vacuum. This configuration can give rise to artifacts stemming from finite-size effects. To circumvent this challenge, periodic boundary conditions (PBC) are frequently adopted. Within PBC, the central system, the sole focus of explicit treatment, is surrounded by an infinite array of identical replicas of itself. This arrangement aims to replicate an infinite solution encompassing the system. When the system is infinitely replicated using PBC, long-range interactions, such as electrostatic interactions, pose a computational challenge due to their spatial extent potentially extending beyond the boundaries of the central image. Thus, an approach is needed that maintains both speed and accuracy. Fortunately, numerous algorithms have been devised and integrated to tackle this dilemma. Notably, the Ewald summation⁸⁷ and the particle mesh Ewald method (PME)⁸⁸ have gained widespread usage. These techniques effectively address long-range interactions of the form $1/r^n$, where $n \leq 3$, encompassing Coulombic interactions.

The fundamental principle underlying these methods involves partitioning the relevant potential into a short-range segment, typically addressed through a cutoff, and a long-range component. The latter entails Fourier transformation to handle the remaining interactions. In essence, these techniques successfully manage both the computational complexity and accuracy associated with long-range interactions, enabling the simulation of systems with periodic boundary conditions.

1.7.2 Force field for RNA and DNA simulations

A precise molecular mechanics force field is fundamental for conducting Molecular Dynamics simulations to achieve an accurate understanding of the structure and dynamics of biomolecules like DNA. While DNA force fields have seen periodic updates to enhance their alignment with available experimental data, they still retain parameters introduced over two decades ago. The recent surge in simulation durations has posed challenges for existing force fields in accurately describing biomolecules, including DNA. Consequently, extensive efforts have been initiated to enhance these DNA force fields, particularly focusing on modifying and refining dihedral angle parameters. However, other components of the force field, such as bond, angle, and nonbonded parameters, have not undergone significant adaptations. For example, in the case of current Amber DNA force fields (bsc0, bsc1, and OL15), nonbonded interactions still rely on Lennard-Jones parameters and partial charges introduced approximately 25 years ago by Cornell et al. This reliance raises concerns, such as the overestimation of binding affinity in protein-DNA complexes due to the limited accuracy of DNA's electrostatics.

In 2021 the group of Professor Zacharias introduced a novel DNA force field named Tumuc1, developed through parameterization based on cutting-edge quantum mechanical (QM) calculations and geometry optimization at the RI-MP2/def2-TZVP level. The force field's parameters have been determined by fitting the electrostatic potential derived from QM calculations on model-systems smaller than a single nucleotide. Additionally, QM frequency calculations were performed, and the modified Seminario method was applied consistently to ensure robust parameterization for bond- and angle terms. The dihedral angles within the Tumuc1 DNA force field have been parameterized by thoroughly scanning the QM potential energy landscapes. Subsequently, the dihedral angle parameters were fitted in a manner consistent with the derived bond, angle, and charge parameters. It's noteworthy that Tumuc1 utilizes the standard Lennard-Jones parameters for its treatment of nonbonded interactions. As demonstrated, the Tumuc1 DNA force field accurately replicates both the structural and dynamic behaviors of double-stranded B-DNA, showcasing remarkable alignment with experimental data. It exhibits significant enhancements in capturing the intricate structural details of DNA when compared to current force fields. Additionally, Tumuc1 effectively models the hybridization of single strands, accurately predicts hairpin folding, and offers a reliable description of protein–DNA complexes that closely matches experimental structures.

For a better understanding we refer to the original paper: “Tumuc1: A New Accurate DNA Force Field Consistent with High-Level Quantum Chemistry”.⁹

For the RNA, instead, we used the recommended χ OL3 RNA force field.¹⁰

1.7.3 Temperature and Pressure Coupling Schemes

One ensemble that can be effectively explored through MD simulations is the microcanonical ensemble, characterized by the preservation of the number of particles (N), volume (V), and total energy (E). Unfortunately, the microcanonical ensemble that comes out of a standard MD simulation does not correspond to the conditions under which most experiments are carried out.

However, for MD simulations to yield more meaningful insights that can be related to experimental observations, they are often combined with thermostats or barostats. These coupling methods play a crucial role in controlling temperature and pressure conditions within the simulated system.

In the case of a thermostat, two primary ensembles are commonly employed: Canonical Ensemble (NVT) and Isobaric-Isothermal Ensemble (NPT). In the NVT scheme, the volume (V) and temperature (T) are maintained as constants. This ensures that the system evolves while keeping its volume fixed and experiencing temperature fluctuations, simulating the behavior observed in many experimental settings. In NPT ensemble, the volume is allowed to change, while simultaneously maintaining a constant pressure (P) and temperature (T). This approach captures the behavior of systems that may experience volume fluctuations, such as in liquids and gases, while also keeping temperature conditions consistent.

To achieve these controlled temperature and pressure conditions, various thermostat and barostat algorithms have been developed and are commonly utilized in MD simulations. For a precise analysis on this topic refer to this review,¹¹ I would just like to present here the most important and well-known ones.

Langevin Thermostat:^{12, 13} this approach introduces a friction term, γ_i , in the equations of motion, simulating the interaction between the system and a heat bath, thereby controlling temperature fluctuations along with a stochastic random force acting on all particles. This is described by the following set of differential equations:

$$\frac{d\vec{R}_i}{dt} = \frac{\vec{P}_i}{M_i} \tag{0.42}$$

$$\frac{d\vec{P}_i}{dt} = \vec{F}_i - \gamma_i \vec{P}_i + \sigma \frac{\vec{R}_i}{\sqrt{dt}} \quad (0.53)$$

Where \vec{F}_i represents the force stemming from the interaction potential, and the final term, denoted as \vec{R}_i , corresponds to the contribution of the random force and \vec{P}_i is the particle momentum. The parameter σ represents the standard deviation of the random force and is linked to the frictional coefficient γ_i , through equation:

$$\sigma = \sqrt{2\gamma_i M_i K_B T} \quad (0.64)$$

Here, K_B is the Boltzmann's constant, M_i represents the mass of particle i , and T denotes the temperature. The random force is stochastically generated from a Gaussian distribution, infusing kinetic energy into the particles, and thus counterbalancing the dampening effect of negative frictional contributions. The Langevin thermostat is categorized as a stochastic ergodic thermostat, a classification denoting its capacity to regulate the system's temperature while ensuring its stochastic behavior encompasses the full phase space.

Nose-Hoover Thermostat:^{14, 15} using a chain of virtual particles, this method maintains temperature while better preserving the canonical distribution of particle velocities. The Langevin thermostat operates as a deterministic algorithm, wherein its ultimate formulation alters the equations of motion by incorporating a frictional force proportionate to the thermodynamic friction parameter $\xi \vec{P}_i$, and ξ represents the thermodynamic friction parameter and \vec{P}_i the momentum of each particle. The parameter ξ is a dynamic entity with its own momentum \vec{P}_ξ and is governed by its own equation of motion. The complete set of equations of motion, encompassing the equation for the heat bath parameter ξ , can be expressed as follows:

$$\vec{R}_i = \frac{\vec{P}_i}{M_i} \quad (0.75)$$

$$\vec{P}_i = \vec{F}_i - \xi \vec{P}_i \quad (0.86)$$

$$\xi = \frac{1}{Q} \left(\sum_i \frac{\vec{P}_i^2}{M_i} - g k_B T \right) = \frac{1}{Q} (T(t) - T_0) \quad (0.97)$$

where $(T(t) - T_0)$ is the difference between the actual temperature of the system and the reference one. Q is the thermal inertia parameter, also referred as the “mass” of the

oscillator ξ , which determines the rate of the heat transfer, i.e. the strength of the bath coupling:

$$Q = \frac{\tau^2 T_0}{4\pi^2} \quad (0.108)$$

where τ is the period of the oscillations of kinetic energy between the system and the reservoir.

For pressure control in NPT, barostats such as the Berendsen barostat and the Parrinello-Rahman barostat are frequently employed. These methods facilitate accurate control over pressure conditions within the simulation, allowing for the exploration of systems under diverse thermodynamic conditions.

Berendsen barostat:¹⁶ Berendsen and colleagues have introduced first-order coupling barostats, which work in tandem with temperature control methods. This coupling ensures that the pressure of the simulated system $P(t)$ is gradually adjusted towards the reference pressure P_0 using a time constant denoted as τ_p :

$$\frac{dP(t)}{dt} = \frac{1}{\tau_p} [P_0 - P(t)] \quad (0.19)$$

The coordinates x_{new} and the volume of the box, V_{new} , are rescaled by a scaling factor μ at every step, such that:

$$x_{new} = \mu x_{old} \quad (0.20)$$

$$V_{new} = \mu V_{old} \quad (0.211)$$

$$\mu = \sqrt[3]{1 - \frac{\beta \delta t}{\tau_p} (P_0 - P(t))} \quad (0.23)$$

Where β is the compressibility of the system. The Berendsen barostat is categorized as a weak coupling scheme, which renders it better suited for pressure equilibration purposes rather than for the actual MD production run. This is due to the potential for the length scaling inherent in the method to induce pronounced oscillations in the pressure, potentially disrupting the stability of the simulation.

Parrinello-Rahman barostat:^{17, 18} The pressure control method pioneered by Parrinello and Rahman formulated in the 1980s, facilitates adaptive adjustments to the simulation box's shape. This is achieved by introducing nine new variables into the system, which

correspond to the components of the unit cell vectors. These vectors are represented collectively by the matrix \mathbf{h} , where its columns consist of the three vectors \mathbf{a} , \mathbf{b} , and \mathbf{c} , which collectively define the box's shape. Consequently, the volume of the cell can be expressed as follows:

$$V = \det \mathbf{h} = \vec{\mathbf{a}} \cdot (\vec{\mathbf{b}} \times \vec{\mathbf{c}}) \quad (0.24)$$

The position \vec{R}_i of the particle i can be written in terms of \mathbf{h} and a column vector representing the scaled coordinates, $\vec{S}_i = [\xi_i \eta_i \zeta_i]$, with $0 \leq [\xi_i \eta_i \zeta_i] \leq 1$

$$\vec{R}_i = \mathbf{h} \vec{S}_i = \xi_i \vec{\mathbf{a}} + \eta_i \vec{\mathbf{b}} + \zeta_i \vec{\mathbf{c}} \quad (0.25)$$

The squared distance between particles i and j can be therefore rewritten as:

$$R_{ij}^2 = S_i^T \mathbf{G} S_j \quad (0.26)$$

where \mathbf{G} is the symmetric matrix, defined as the metric tensor:

$$\mathbf{G} = \mathbf{h}^T \mathbf{h} \quad (0.27)$$

With the introduction of the scaled coordinates \vec{S}_i for each atom i , the original Lagrangian of $3N$ variables becomes now an extended Lagrangian of $(3N + 9)$ variables, written as:

$$L_{PS} = \frac{1}{2} \sum_i M_i \dot{S}_i^T \mathbf{G} \dot{S}_i - \sum_{i < j} U(R_{ij}) + \frac{1}{2} W \text{Tr}(\dot{\mathbf{h}}^T \dot{\mathbf{h}}) - p_0 V \quad (0.28)$$

where $U(R_{ij})$ is the pair potential, p_0 is the reference external applied pressure, V is the unit cell volume, W is constant of proportionality (with mass dimensionality) of the kinetic term associated with the time variation of \mathbf{h} . The corresponding equations of motion are then derived for \vec{S}_i and \mathbf{h} .

1.8 Enhanced sampling techniques for Accelerated Molecular Dynamics

Molecular dynamics (MD) simulations have been extensively utilized over recent decades, finding applications in various domains such as materials science, chemistry, biology, geology, and more. These simulations offer a direct understanding of the temporal progression of molecular systems with a complete atomistic resolution. Particularly, with the widespread availability of parallel computing resources, researchers can now handle significantly large system sizes (simulating entire systems, viruses for example). However,

despite the availability of massively parallel resources, MD encounters a persistent timescale challenge that hinders its progress.

MD is confined to integration timesteps of a few femtoseconds, a limitation that can be somewhat alleviated through the implementation of multiple timestep algorithms. Yet, reaching the millisecond regime and beyond for any system with more than a few thousand atoms remains a considerable challenge. Unlike spatial dimensions, time operates sequentially, making the timescale problem less amenable to straightforward parallelization. The root cause and potential solutions to the timescale problem stem from the observation that many intriguing systems possess an energy landscape characterized by numerous metastable states separated by substantial kinetic barriers. Transgressing these barriers to explore new states becomes a rare event compared to the obligatory few femtosecond integration timestep of MD. Consequently, a multitude of enhanced sampling approaches have been proposed over the past two decades to accelerate system dynamics, enabling access to significantly extended timescales.

1.8.1 Rare Events, Separation of Timescales and Markovianity

Considering a system consisting of N atoms residing in a $2dN$ -dimensional phase space, where the dimensionality ranges from $d = 1$ to 3 , and N can vary from 1 to a few million. The MD simulation initiates from a given configuration, involving the numerical integration of Newton's laws of motion within a classical force-field or interatomic potential, all under a temperature T (or equivalently inverse temperature $\beta = \frac{1}{k_B T}$). A thermostat is employed to enforce the desired temperature. The simulation conditions can include constant temperature (within fluctuations), constant volume (V), or constant pressure (P), referred to as the NVT and NPT ensembles, respectively.

The processes of interest in our study typically occur on timescales significantly longer than the vibration period of individual atoms, which is usually in the order of a few femtoseconds. However, the rates at which these processes occur often extend to much slower timescales, ranging from microseconds to milliseconds. Although each degree of freedom within the system undergoes constant fluctuations with an average thermal energy of $\frac{1}{2}k_B T$, these fluctuations rarely align with the specific modes required for the desired event to occur. Hence, an event that might be readily observed and studied in a laboratory setting becomes a rare event in practice for the molecular dynamics.

Closely related to the concept of rare events are the notions of timescale separation and Markovianity. Timescale separation refers to the presence of a spectrum of timescales that can be clearly distinguished into distinct non-overlapping regimes. Essentially, the system traverses a landscape characterized by deep basins, and the decorrelation of dynamical variables within each basin occurs much faster compared to the typical basin escape times.

On the other hand, Markovianity signifies that before the system departs a stable basin, it has forgotten the specifics of how it entered the basin initially. This is reasonable because systems comprised of many interacting particles are highly sensitive to initial conditions. Consequently, even minute alterations in initial conditions before and shortly after entering the basin, often due to numerical noise in thermostating, rapidly lead to diverging trajectories within the basin. This renders the system's precise state effectively random within the basin after the molecular relaxation time has passed. Therefore, when mapping the system's trajectory into a list of states it visits, it becomes sensible to discuss a unique state-to-state timescale, characterized solely by two inputs: the identities of the state being exited, and the state being entered.

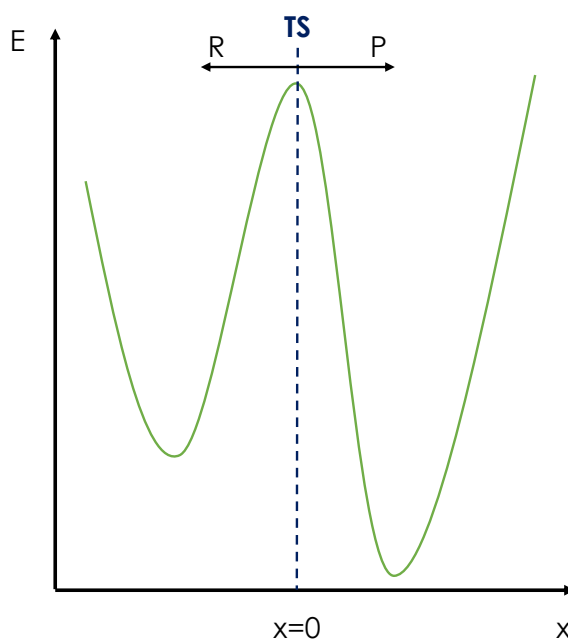


Figure 1. A schematic 1-d energy landscape where the x-axis denotes some reaction coordinate and y-axis is the energy. R, TS, and P stand for reactant, transition state, and product, respectively. Here the depth of either basin is much larger than kBT or the typical fluctuation in the energy associated with the coordinate x . As such, moving from basin R to P becomes a rare event, and only very occasionally the system visits the TS region.

1.8.2 Potential and Free Energy Surfaces

In the preceding subsection, we introduced the concept of a landscape or surface characterized by deep basins. To formalize this notion, one approach involves utilizing coordinates R from the entire $3N$ -dimensional configuration space (i.e., all atomic coordinates) and examining the total potential energy $U(R)$ of the system. A constant energy surface in this space is termed a potential energy surface (PES). PES has been extensively and effectively utilized for relatively small systems, especially at lower temperatures.¹⁹ At sufficiently low temperatures and with small system sizes, accelerating molecular dynamics becomes closely associated with identifying relevant saddle points on the PES, which the system must traverse while moving from one basin to another.

However, working with small system sizes is not always feasible, and characterizing or working with the PES computationally becomes prohibitively expensive. In high-dimensional systems, the PES contains an excessive number of saddle points, many of which are irrelevant to the dynamics of interest.²⁰ Additionally, it might be more appropriate and effective to describe the large number of visited states through an entropic description, especially at higher simulation temperatures, rather than exhaustively enumerating states as a PES description would require. Although one can still work with the PES, it necessitates separate approximations to calculate the appropriate entropic corrections.

As an alternative to focusing on the PES, one strategy (though not the only one) is to examine a low-dimensional free energy surface (FES), defined as a function of a small number of collective variables (CVs) denoted as $s = \{s_1, \dots, s_k\}$, where $k \ll N$. These collective variables represent "interesting" degrees of freedom or "reaction coordinates." They can be more than simple linear projections and may constitute complex nonlinear functions of all atomic coordinates. In terms of the potential energy $U(R)$, the free energy $F(s)$ is defined as follows:

$$F(s) = -\beta^{-1} \ln \int dR \delta(s - s(R)) e^{-\beta U(R)} \quad (0.29)$$

This definition differs from the conventional Helmholtz free energy solely by the term $\delta(s - s(R))$ which selects the region of phase space associated with a specific value of s , representing the collective variable. Excluding this term in the integration would yield the total Helmholtz free energy of the system.

In analogy with the PES description, one now examines basins in the FES, which is low-dimensional by design and hence easier to handle. Additionally, since the free energy

according to the provided definition incorporates temperature, one can explicitly address entropic effects. Of course, the methods based on the calculations of this FES are not the solution of all the problems because the reduction in dimensionality when transitioning from PES to FES is closely tied to a well-considered selection of a small number of effective collective variables. It often requires a priori knowledge of all possible existing and relevant deep stable basins in the system. It's important to note that knowledge of stable basins is generally a weaker requirement than knowledge of escape pathways. There are a lot of different Free Energy Surface Based Methods (Umbrella Sampling, Replica Exchange, etc.) but see in detail the theory and how all these methods work comes out of the scope of this thesis, we focus on the Metadynamics that the method chosen to study the unfolding of our systems.

1.8.3 Metadynamics

Metadynamics (MetaD) is a widely recognized method utilized to explore intricate Free Energy Surfaces (FESs) by constructing a time-dependent bias potential. Initially, a small set of relevant Collective Variables (CVs) is identified. To intensify sampling in regions of CV space that are seldom visited, a memory-dependent bias potential is progressively developed during the simulation as a function of these CVs. This bias typically takes the form of repulsive Gaussians added wherever the system visits in the CV space. Consequently, the system gradually avoids revisiting these areas, leading to an increase in fluctuations in the CVs. This discourages the system from becoming trapped in low free energy basins. At the conclusion of a MetaD run, the probability distribution of any observable—whether biased directly or not—can be computed using a reweighting procedure.²¹ This convenient reweighting capability is one of the many attributes of MetaD that has propelled its widespread adoption for calculating FESs.

So, the idea of MetaD is based on the systematically 'filling' the free energy minima of the metastable states in a controlled manner, allowing the system to explore all states. This entails initially selecting a low-dimensional Collective Variable based on chemical or physical intuition—a function of coordinates that assumes distinct values in all relevant metastable states. By using this CV, the probability distribution $P(x)$ can be transformed into a function of the CV:

$$P(s) = \int dx P(x) \delta(s - S(x)) \tag{0.30}$$

In this context, we assume $P(x)$ follows the canonical distribution associated with a potential energy function $V(x)$: $P(x) \propto e^{\left(\frac{-V(x)}{T}\right)}$ where T is the temperature (using units where the Boltzmann constant is one). A well-chosen CV manifests the metastable states as distinct, well-defined peaks in $P(s)$. The free energy as a function of s is:

$$F(s) = -T \log(P(s)) \tag{0.312}$$

showcases at least two clearly defined minima for a system with metastable states.

In an ideal scenario, for the molecular system under study, both a suitable CV and an approximation $B(s)$ of the negative free energy are known. Under such circumstances, the metastability problem can be regarded as resolved.

However, in the real world there are some issues:

- The structure of the free energy is unknown before the simulation, making the choice of a good $B(s)$ challenging.
- Identifying a good CV can be nontrivial, even if an intuitive CV distinguishing metastable states is constructed, it might not be suitable for describing transitions.
- In some cases, the relevant metastable states are themselves unknown, posing a significant challenge, particularly in studying conformational transitions in complex biomolecules.

MetaD is an algorithm that effectively addresses the first problem by iteratively constructing $B(s)$ during the simulation. While it doesn't provide a CV directly, it verifies the quality of a CV and can improve it for subsequent simulations. Additionally, MetaD allows for the simultaneous use of multiple CVs, providing more flexibility in their selection and even enabling the exploration of unknown metastable states in specific cases.

The fundamental concept of MetaD involves the systematically 'filling' the free energy minima with a Gaussian function with a width σ and height w is employed. Initially, the Gaussians are concentrated in the first free-energy minimum. These Gaussians induce significant fluctuations in the CV. Over time, the Gaussians progressively fill the first free-energy minimum until the system transitions to the second minimum. The CV starts diffusing freely between these minima once this transition occurs. The sum of these Gaussians compensates nearly precisely for the free energy, enabling an estimation of $F(s)$. The parameters w and σ can be adjusted to control the rate at which the free-energy landscape is filled and flattened.

Choosing larger Gaussians causes the bias to increase rapidly, but the system deviates significantly from equilibrium. Conversely, utilizing smaller Gaussians results in MetaD resembling a quasi-equilibrium process.

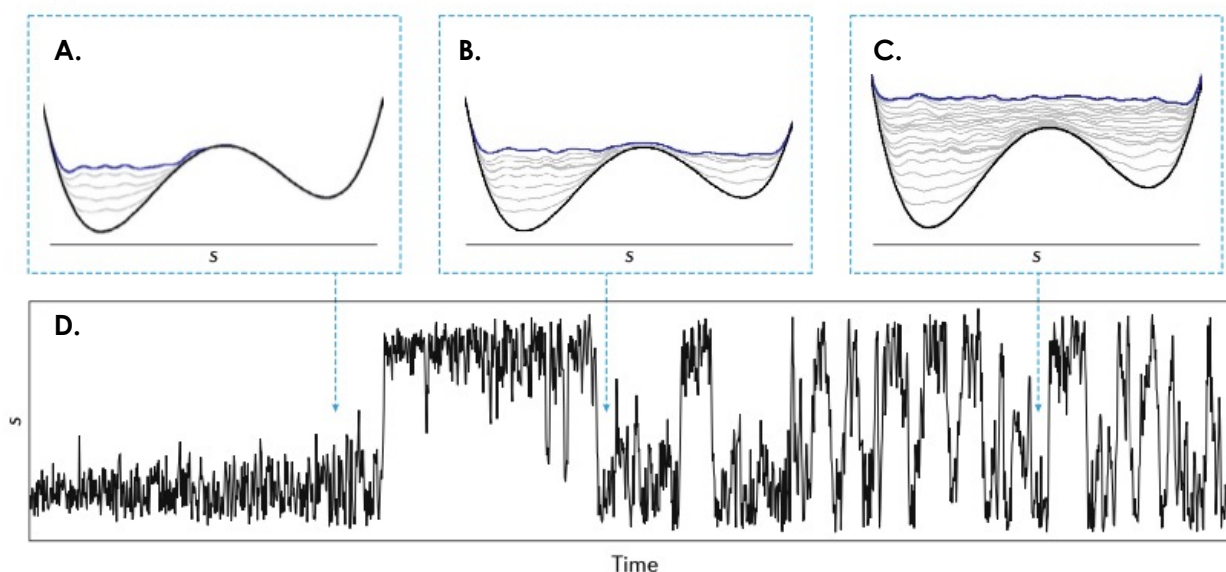


Figure 2. The working principles of metadynamics (MetaD). **A-C.** The sum of the free energy and of the MetaD bias potential (blue lines) at three different times marked by arrows in panel **D.**, along with the free energy (black lines). **D.** The CV s as a function of time in a MetaD simulation.

There are numerous additional aspects related to MetaD that could be discussed, including the various collective variables and the methods for their selection. Additionally, understanding how to calculate free energy and associated errors, along with an exploration of diverse methods linked to MetaD, could be covered. However, delving into these details is beyond the scope of this thesis. There are comprehensive reviews available that provide a detailed explanation of these parameters for those interested in a deeper understanding.²²

1.9 Epitope Prediction Method

Epitopes are sections of a protein that can be recognized by binding partners. Their sequences are often adaptable to mutations, indicating that they aren't crucial for stabilizing the protein's antigenic structure. Instead, they've evolved to continuously avoid detection by the host's immune system, while still maintaining the native structure necessary for the pathogen's function. Epitopes can exhibit flexibility and conformational changes. In essence, they aren't strongly involved in major stabilizing interactions within the protein. From a structural perspective, epitopes are exposed regions on the protein's surface, easily accessible for antibody binding. In the case of non-continuous epitopes, high-resolution X-ray structures of antigen-antibody complexes reveal that they consist of residues with spatial relationships defining a sizable region on the antigen's surface.

This approach combines an analysis of protein energetics from molecular dynamics (MD) simulations with topological data from contact matrices obtained from representative trajectory structures. The goal is to locate contiguous regions in the antigen's 3D conformation that have minimal interactions with the rest of the protein. These regions are likely candidates for dynamic modulation, important for recognition events.

The energy decomposition method underpins the energetics analysis, enabling the identification of significant residue-residue interactions for fold stability. The approach simplifies the noisy energy matrix through eigenvalue decomposition. We concentrate on the eigenvector's lowest eigenvalue components, revealing strong interaction centers. Applying this to the most populated structural cluster yields similar results to trajectory averaging. We validate this approach against experimental data, connecting protein stability with its energetic and topological traits.²³⁻²⁷

The pair energy-coupling map, filtered using topological insights, aids in recognizing local couplings with minimal energy interactions. Since low-energy couplings between distant residues are a product of distance-dependent energy functions, local low-energy couplings highlight sites where interaction networks aren't energetically optimized. These regions tend to interact with binding partners or tolerate mutations that preserve antigen structure. These areas often cluster on the protein's accessible surface. This concept is reminiscent of local frustration, seen near interaction sites on protein surfaces where high frustration is common.

This method has been optimized and extended to cover glycoproteins.

1.9.1 Energy Decomposition (ED)

The energy decomposition method is based on the calculation of the interaction matrix M_{ij} , which is determined by evaluating average, interresidue, nonbonded (van der Waals and electrostatics) interaction energies between residue pairs, calculated over the structures visited during an MD trajectory (The symmetric interaction matrix M_{ij} obtained from separate MM/GBSA calculations). For a protein of N residues, this calculation yields an $N \times N$ matrix. As stated above, the same results can be obtained by calculating the interaction matrix M_{ij} from the representative conformation of the most populated cluster, in the absence of major conformational changes.

The aim of the method is to obtain a simplified picture of the most relevant residue-residue interactions in a certain fold. The matrix M_{ij} is thus diagonalized and re-expressed in terms of eigenvalues and eigenvectors, in the form:

$$M_{ij} = \sum_{\alpha=1}^N \lambda_{\alpha} v_i^{\alpha} v_j^{\alpha} \quad (0.32)$$

where λ_{α} is the α -th eigenvalue and v_i^{α} is the i^{th} component of the corresponding eigenvector.

It was previously shown in a number of cases that eigenvector (v_i^{α}), also called *first eigenvector*, associated with the lowest eigenvalue λ_1 allows to identify most of the crucial aminoacids necessary for the stabilization of a protein fold, and consequently those aminoacids that are minimally coupled to such core. The latter were shown to correspond to potential interaction regions.

In the case of multidomain proteins such as S, the first eigenvector is not sufficient, and more eigenvectors are needed to capture the essential interactions for folding/stability and binding. The interaction matrix M_{ij} is thus decomposed instead via the alternative approach developed by Genoni *et al.*²³ In this scenario, the aim is to select the smallest set of N_e eigenvectors that cover the largest part of residues (i.e., components) with the minimum redundancy under the assumption that: (a) for each domain there should exist only one associated eigenvector recapitulating its most significant interactions; (b) each “domain eigenvector” has a block structure whereby its significant components correspond to the residues belonging to the identified domain; (c) combination of all significant blocks covers all residues in the protein. Matrix M_{ij} can thus be reformulated as a simplified matrix \tilde{M}_{ij} :

$$\tilde{M}_{ij} = \sum_{\alpha=1}^{N_e} \lambda_{\alpha} v_i^{\alpha} v_j^{\alpha} \quad (0.33)$$

(where this time the sum occurs over N_e essential eigenvectors instead of N residues). As detailed by Genoni *et al.*,²³ the essential folding matrix \tilde{M}_{ij} is subsequently further filtered through a symbolization process to emphasize the significant non-bonded interaction, yielding $\tilde{M}_{ij}^S, \tilde{M}_{ij}^S$ and finally subjected to a proper clustering procedure leading to domain identification.

The final simplified matrix \tilde{M}_{ij}^S resulting from domain decomposition thus only reports those residue pairs in the protomer that exhibit the strongest and weakest energetic interactions.

1.9.2 Matrix of Local Coupling Energies method (MLCE)

Final epitope predictions are made using the Matrix of Local Coupling Energies method (MLCE), in which analysis of a given protein's energetic properties is combined with that of its structural determinants. This approach allows to identify nonoptimized, contiguous regions on the protein surface that are deemed to have minimal coupling energies with the rest of the structure, and that have a greater propensity for recognition by Abs or other binding partners.

The MLCE procedure entails cross-comparison of the simplified pairwise residue-residue energy interaction matrix \tilde{M}_{ij} resulting from domain decomposition (*vide supra*) with a pairwise residue-residue contact matrix C_{ij} . The latter matrix namely considers a pair of residues to be spatially contiguous (i.e., 'in contact') if they are closer than an arbitrary 6.0 Å-threshold; contact distances are measured between C β atoms in the case of non-glycine aminoacid residues, H atoms in the case of glycine residues, and between C1 atoms in the case of glycan residues.

The Hadamard product of the two matrices yields the matrix of the local pairwise coupling energies $MLCE_{ij}$:

$$MLCE_{ij} = \tilde{M}_{ij} \cdot C_{ij} \quad (0.34)$$

Deriving the MLCE matrix allows to rank spatially contiguous residue pairs with respect to the strength of their energetic interactions (weakest to strongest). Selection of proximal pairs showing the weakest coupling with the rest of the protein ultimately defines putative epitopes; two distinct selections are carried out on the basis of two possible weakness

(softness) cutoffs (5% or 15%), corresponding to the top 5% or 15% spatially contiguous residue pairs with the lowest energetic interactions.

STRUCTURE SELECTION FROM MOLECULAR DYNAMICS (Clustering of MD Simulations)

Coordinates of the fully glycosylated SARS-CoV-2 S protein's are found using Clustering calculations conducted using the hierarchical agglomerative algorithm,²⁸ considering every 20th metatrajectory frame (*i.e.*, every 50 ps), based on the root-mean-square deviation of backbone heavy atoms of aminoacid residues composing the NTD and the RBD in all three protomers. Values of ϵ are chosen so that they provide the best compromise between maximizing cluster homogeneity, based on silhouette score, and ensuring at least 60-80% of the metatrajectory is covered by the three most populated clusters.

MINIMIZATION

A 200-step minimization of each structure is carried out using the default procedure (*i.e.*, steepest descent for 10 steps; then conjugate gradient) implemented in the MD engine *sander* in the *AMBER* software package (version 18).^{29, 30} Protomers are minimized using the generalized Born (GB) implicit solvent model as parametrized by Onufriev et al.³¹, with a universal 12.0 Å cutoff applied in the calculation of Lennard-Jones and Coulomb interactions (neither of which are calculated beyond this limit). For this stage, concentration of (implicit) mobile counterions in the GB model is set to 0.1 M, and the solvent-accessible surface area (SASA) is computed according to the LCPO method (linear combinations of pairwise overlaps).³²

MM/GBSA CALCULATIONS

MM/GBSA calculations³³ are performed on each of the three minimized 'RBD up' protomers using the dedicated *mm_pbsa.pl* utility in *AmberTools* (version 17). The purpose of these calculations is to obtain a breakdown of nonbonded energy interactions (*i.e.*, electrostatic, van der Waals, implicit solvation contributions and, in this case, 1-4 interactions) between every possible pair of residues in the protomer (aminoacids and monosaccharides alike): for a protomer composed of N residues, this leads to a symmetric $N \times N$ interaction matrix M_{ij} .³⁴

The implicit GB solvation model used in these calculations is identical to the one used in the preceding minimization step (*vide supra*), except that the implicit ion concentration is set to

0.0 M, and SASA is computed with the ICOSA method (based on icosahedra around each atom that are progressively refined to spheres).

1.10 Docking

In the pursuit of discovering novel therapeutic targets for drug development, the evolution of high-throughput methodologies, such as protein purification, crystallography, and nuclear magnetic resonance spectroscopy, has significantly enriched our understanding of protein structures and their interactions with ligands. These advancements have effectively paved the way for computational strategies to infiltrate every facet of contemporary drug discovery. These strategies encompass a wide array of techniques, among them being virtual screening (VS), which serves as a potent tool for identifying promising lead compounds, and methods tailored for the optimization of these leads. Notably, different from conventional experimental high-throughput screening (HTS), the VS approach offers a more streamlined and rational pathway to drug discovery. Its advantages, including cost-effectiveness and heightened efficiency, underscore its significance in the field.

VS strategies can be categorized into two main types: ligand-based and structure-based methods. Ligand-based methods, such as pharmacophore modeling and quantitative structure-activity relationship (QSAR) analysis, come into play when a set of active ligands is available and little structural information is known about the targets. On the other hand, for structure-based drug design, molecular docking stands out as the predominant approach and has been extensively utilized since the early 1980s.³⁵ Molecular docking studies have become an increasingly significant tool in pharmaceutical research, with various programs utilizing different algorithms developed to carry out these studies.

Molecular docking enables the modeling of atomic-level interactions between small molecules and proteins. This allows for a detailed understanding of small molecule behavior within the binding site of target proteins, as well as the elucidation of fundamental biochemical processes. The docking process entails two fundamental steps: the prediction of the ligand's conformation, position, and orientation within the binding site (referred to as the "pose"), and the evaluation of the binding affinity. These steps are intricately connected to sampling methods and scoring schemes.

Efficient docking processes benefit significantly from prior knowledge of the binding site's location. Often, this information is available before ligand docking, either from previous studies or through comparisons with functionally similar proteins or those co-crystallized with ligands. Alternatively, when binding site knowledge is lacking, cavity detection programs or online servers can be employed to predict potential active sites, a practice referred to as blind docking.

Early elucidation of ligand-receptor binding mechanisms centered around Fischer's lock-and-key theory,³⁶ where the ligand fits into the receptor like a lock and key. Initial docking methods were built upon this concept, treating both ligand and receptor as rigid entities. The subsequent "induced-fit" theory by Koshland³⁷ expanded on this, suggesting that protein active sites continually adapt as ligands interact with the protein, necessitating flexible treatment during docking for a more accurate depiction of binding events.

1.10.1 Theory of docking

Essentially, molecular docking tries to predict the structure of a ligand-receptor complex. This process involves two interconnected steps: first, sampling various conformations of the ligand within the protein's active site, and second, evaluating these conformations using a scoring function. The ideal outcome is that sampling algorithms accurately replicate the experimentally observed binding configuration, while the scoring function appropriately ranks it as the most favorable among all generated conformations. To provide a foundational understanding, let's briefly outline the fundamental principles of docking theory from these two perspectives.

1.10.2 Search Algorithms

Due to the vast number of potential binding modes resulting from six translational, rotational, and conformational degrees of freedom for both ligand and protein, exhaustive sampling is infeasible. Various search algorithms have been developed and are integral to molecular docking software. Generally, these algorithms could be divided in three main classes: Systematic search methods, Random or Stochastic methods and Simulation methods.

The **Systematic** search algorithms try to explore all the degrees of freedom in a molecule which is dictated by the rotations of the bonds and angles and size of increments. For example, Exhaustive search algorithms systematically explore ligand conformations by iteratively rotating all possible rotatable bonds within defined intervals. However, the extensive conformational space often makes exhaustive searches impractical. To address this challenge, advanced algorithms like GLIDE employ heuristics to prioritize regions of conformational space likely to yield favorable ligand poses. GLIDE utilizes precomputed grid representations of the target's shape and properties, alongside an initial set of low-energy ligand conformations based on torsion angles. The process begins with approximate

positioning and scoring to identify promising initial ligand poses, narrowing down the conformational space for a high-resolution docking search. This refined search involves ligand minimization through molecular mechanics energy calculations, followed by a Monte Carlo procedure to explore nearby torsional minima.

The **Stochastic** Methods randomly modify ligand conformations. Monte Carlo (MC) methods use energy-based criteria to accept or reject conformations based on transformations. Genetic algorithms (GA) encode ligand degrees of freedom as binary strings, subjecting them to genetic operators like mutation and crossover.

Molecular Dynamics (MD) MD **simulations** offer robust flexibility simulation, representing both ligand and protein movements effectively. However, MD's small progression steps may hinder sampling over high-energy barriers. MD is often used in combination with random search strategies to optimize local conformations.

In summary, molecular docking employs various sampling algorithms to explore ligand-receptor interactions, enabling the prediction of complex structures and aiding drug discovery efforts.

1.10.3 Scoring functions

Scoring functions play a crucial role in molecular docking by distinguishing favorable ligand-receptor binding poses from unfavorable ones within a reasonable computational time. These functions estimate the binding affinity between a protein and ligand, introducing various assumptions and simplifications. They can be categorized into force-field-based, empirical, and knowledge-based scoring functions.

Force-Field-Based Scoring Functions. Classical force-field-based scoring functions evaluate binding energy by summing non-bonded interactions, including electrostatic and van der Waals forces, as we see before.

$$\begin{aligned}
 V = & W_{vdw} \sum_{ij} \left(\frac{A_{ij}}{r_{ij}^{12}} - \frac{B_{ij}}{r_{ij}^6} \right) \\
 & + W_{hbond} \sum_{ij} E(t) \left(\frac{C_{ij}}{r_{ij}^{12}} - \frac{D_{ij}}{r_{ij}^{10}} \right) W_{elec} \sum_{ij} \frac{q_i q_j}{\epsilon(r_{ij}) r_{ij}} \\
 & + W_{solv} \sum_{ij} (S_i V_j + S_j V_i) e^{(-r_{ij}^2/2\sigma^2)}
 \end{aligned}
 \tag{0.35}$$

This is Extended force-field-based scoring function from AutoDock: For two atoms i, j , the pair-wise atomic energy is evaluated by the sum of van der Waals, hydrogen bond, coulomb energy and desolvation. $W_{vdw}, W_{elec}, W_{solv}$ are weight factor to calibrate the empirical free energy.³⁸

Empirical Scoring Functions. Empirical scoring functions decompose binding energy into components like hydrogen bonds, ionic interactions, hydrophobic effects, and binding entropy. Each component is assigned a coefficient, and the sum of these components yields a final score. Coefficients are determined through regression analysis using a test set of ligand-protein complexes with known affinities. While relatively simple, empirical scoring functions may lack generalization beyond their training set.

$$\begin{aligned} \Delta G = & \Delta G_0 + \Delta G_{rot} \times N_{rot} \\ & + \Delta G_{hb} \sum_{neutral\ H-bond} f(\Delta R, \Delta \alpha) + \Delta G_{io} \sum_{ion\ init.} f(\Delta R, \Delta \alpha) \\ & + \Delta G_{aro} \sum_{aro\ init.} f(\Delta R, \Delta \alpha) + \Delta G_{lipo} \sum_{lipo\ cont.} f^*(\Delta R) \end{aligned} \quad (0.36)$$

Empirical scoring function from FlexX. G is the estimated free energy of binding; G_0 is the regression constant; $G_{rot}, G_{hb}, G_{io}, G_{aro}$ and G_{lipo} are regression coefficients for each corresponding free energy term; $f(\Delta R, \Delta \alpha)$ is scaling function penalizing deviations from the ideal geometry; N_{rot} is the number of free rotate bonds that are immobilized in the complex.³⁹

Knowledge-Based Scoring Functions. These functions use statistical analysis of crystal structures to derive interatomic contact frequencies and distances between ligand and protein atoms. The assumption is that more favorable interactions occur more frequently. These distributions are converted into pairwise atom-type potentials, favoring preferred contacts and penalizing repulsive interactions. Knowledge-based functions are computationally efficient and can model uncommon interactions but may suffer from underrepresented interactions in training sets.

$$PMF_{score} = \sum_{\substack{kl \\ r < r_{cut\ off}^{ij}}} A_{ij}(r) \quad A_{ij}(r) = -k_B T \ln \left[f_{voll\ corr}^j(r) \frac{\rho_{seg}^{ij}(r)}{\rho_{bulk}^{ij}} \right] \quad (0.37)$$

Knowledge-based scoring functions PMF. k_B is the Boltzmann constant; T is the absolute temperature; r is the atom pair distance. $f_{voll\ corr}^j(r)$ is the ligand volume correction factor;

$\frac{\rho_{seg}^{ij}(r)}{\rho_{bulk}^{ij}}$ designates the radial distribution function of a protein atom of type i and a ligand atom of type j .⁴⁰

In summary, scoring functions are pivotal in molecular docking, guiding the selection of promising ligand-receptor interactions. Different types of scoring functions each have their strengths and limitations, reflecting the complex interplay between computational efficiency and accuracy in predicting binding affinities.

1.10.4 Glide: grid-based ligand docking with energetics⁴¹

Glide is a docking software designed to perform an accurate search of the positional, orientational and conformational space available to the ligand using hierarchical filters. It produces a set of initial ligand conformations, which correspond to minima in the torsion-angle space of the ligand, and it screens them over the entire phase space available to the ligand. Once the most promising poses are located inside the receptor, the ligand is minimized using a MM energy function which incorporates a distance-dependent dielectric model. Then, the lowest-energy poses undergo Monte Carlo procedure and then poses are evaluated with a scoring function.

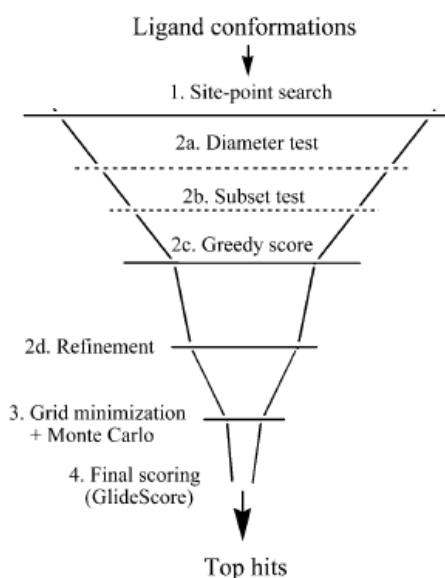


Figure 3. Schematic representation of the steps involved in Glide.

Glide can employ two different scoring functions: GlideScore 2.5 Standard-Precision (SP) is a more forgiving scoring function, which aims to identifying ligands that have a reasonable possibility to bind and is used mainly in screening libraries. On the other hand, GlideScore 2.5 Extra-Precision (XP) applies severe penalties on poses that violate physical chemistry

principles; its main application consists in lead optimization. The scoring functions used by Glide consider many types of interaction: lipophilic, hydrogen-bonding (both for two charged groups and for a charged group with a neutral one), metal-ligand, electrostatic and van der Waals interactions: furthermore, it introduces a solvation model, using explicit waters into the binding site (rather than using a continuum solvation model).

$$\begin{aligned}
 \Delta G_{bind} = & C_{lipo-lipo} \sum f(r_{lr}) + C_{hbond-neut-neut} \sum g(\Delta r)h(\Delta\alpha) \\
 & + C_{hbond-neut-charged} \sum g(\Delta r)h(\Delta\alpha) + C_{hbond-charged-charged} \sum g(\Delta r)h(\Delta\alpha) \\
 & + C_{max-metal-ion} \sum f(r_{lm}) + C_{rotb}H_{rotb} + C_{polar-phob}V_{polar-phob} \\
 & + C_{coul}E_{rotb} + C_{vdw}E_{vdw} + \text{solvation terms}
 \end{aligned}
 \tag{0.38}$$

1.11 References

- (1) Swope, W. C., Andersen, H.C., Berens, P.H. and Wilson, K.R. A computer simulation method for the calculation of equilibrium constants for the formation of physical clusters of molecules: Application to small water clusters. *J. Chem. Phys.* **1982**, *76*, 637-649.
- (2) Andersen, H. C. Rattle: A "velocity" version of the shake algorithm for molecular dynamics calculations. *J. Comput. Phys.* **1983**, *52*, 24-34.
- (3) Van Gunsteren, W. F. a. B., H.J.C. A Leap-frog Algorithm for Stochastic Dynamics. *Mol. Simul.* **1988**, *1*, 173-185.
- (4) Cornell, W. D., Cieplak, P., Bayly, C.I., Gould, I.R., Merz, K.M., Ferguson, D.M., Spellmeyer, D.C., Fox, T., Caldwell, J.W. and Kollman, P.A. A Second Generation Force Field for the Simulation of Proteins, Nucleic Acids, and Organic Molecules. *J. Am. Chem. Soc.* **1995**, *117*, 5179-5197.
- (5) Van Gunsteren, W. F. *Biomolecular simulation: the GROMOS96 manual and user guide.*; 1996.
- (6) MacKerell, A. D., Bashford, D., Bellott, M., Dunbrack, R.L., Evanseck, J.D., Field, M.J., Fischer, S., Gao, J., Guo, H., Ha, S. All-atom empirical potential for molecular modeling and dynamics studies of proteins. *J. Phys. Chem. B* **1998**, *102*, 3586-3616.
- (7) Ewald, P. P. Die Berechnung optischer und elektrostatischer Gitterpotentiale. *Annalen der Physik* **1921**, *369* 253-287.
- (8) Tom Darden; Darrin York; Pedersen, L. Particle mesh Ewald: An $N \cdot \log(N)$ method for Ewald sums in large systems. *J. Chem. Phys.* **1993**, *98*, 10089 DOI: <https://doi.org/10.1063/1.464397>.
- (9) Liebl, K.; Zacharias, M. Tumuc1: A New Accurate DNA Force Field Consistent with High-Level Quantum Chemistry. *J Chem Theory Comput* **2021**, *17* (11), 7096-7105. DOI: 10.1021/acs.jctc.1c00682.
- (10) Zgarbová, M.; Otyepka, M.; Sponer, J.; Mládek, A.; Banáš, P.; Cheatham, T. E.; Jurečka, P. Refinement of the Cornell et al. Nucleic Acids Force Field Based on Reference Quantum Chemical Calculations of Glycosidic Torsion Profiles. *J Chem Theory Comput* **2011**, *7* (9), 2886-2902. DOI: 10.1021/ct200162x.
- (11) Holm, C.; Kremer, K. <h1 data-test-id="paper-detail-title" style="border: 1px solid black; padding: 5px; font-family: "Roboto Slab", Georgia, serif; font-size: 30px; line-height: 32px; margin: 0px 0px 5px 0px; color: rgb(46, 55, 67); background-color: rgb(235, 236, 237);"> Thermostat Algorithms for Molecular Dynamics Simulations. *Advances in Polymer Science* **2005**, *173*, 105-147.
- (12) Adelman, S. A. a. D., J.D. Generalized Langevin Equation Approach for Atom-Solid-Surface Scattering - General Formulation for Classical Scattering Off Harmonic Solids. *J. Chem. Phys.* **1976**, *64*, 2375-2388.
- (13) Turq, P., Lantelme, F. and Friedman, H.L. Brownian dynamics: its application to ionic solutions. *J. Chem. Phys.* **1977**, *66*, 3039-3044.
- (14) Nosé, S. A unified formulation of the constant temperature molecular dynamics methods. *J. Chem. Phys.* **1984**, *81*, 511-519.
- (15) Hoover, W. G. Canonical dynamics: Equilibrium phase-space distributions. *Phys. Rev. A* **1985**, *31*, 1695-1697.
- (16) Berendsen, H. J. C.; Postma, J. P. M.; van Gunsteren, W. F.; Di Nola, A.; Haak, J. R. Molecular dynamics with coupling to an external bath. *Journal of Chemical Physics* **1984**, *81* (8), 3684-3690. DOI: <https://doi.org/10.1063/1.448118>.
- (17) Parrinello, M. a. R., A. Crystal Structure and Pair Potentials: a Molecular-Dynamics Study. *Phys. Rev. Lett.* **1980**, *45*, 1196-1199.
- (18) Parrinello, M. a. R., A. Polymorphic Transitions in Single-Crystals - a New Molecular-Dynamics Method. *J. Appl. Phys.* **1981**, *52*, 7182-7190.
- (19) de Souza, V. K.; Stevenson, J. D.; Niblett, S. P.; Farrell, J. D.; Wales, D. J. Defining and quantifying frustration in the energy landscape: Applications to atomic and molecular clusters, biomolecules, jammed and glassy systems. *J Chem Phys* **2017**, *146* (12), 124103. DOI: 10.1063/1.4977794.
- (20) Bolhuis, P. G.; Chandler, D.; Dellago, C.; Geissler, P. L. Transition path sampling: throwing ropes over rough mountain passes, in the dark. *Annu Rev Phys Chem* **2002**, *53*, 291-318. DOI: 10.1146/annurev.physchem.53.082301.113146.
- (21) Tiwary, P.; Parrinello, M. A time-independent free energy estimator for metadynamics. *J Phys Chem B* **2015**, *119* (3), 736-742. DOI: 10.1021/jp504920s.
- (22) Bussi, G., Laio, A. Using metadynamics to explore complex free-energy landscapes. *Nat Rev Phys* **2020**, *2*, 200-212 DOI: <https://doi.org/10.1038/s42254-020-0153-0>.
- (23) Scarabelli, G.; Morra, G.; Colombo, G. Predicting interaction sites from the energetics of isolated proteins: a new approach to epitope mapping. *Biophys J* **2010**, *98* (9), 1966-1975. DOI: 10.1016/j.bpj.2010.01.014.

- (24) Gourlay, L. J.; Peri, C.; Ferrer-Navarro, M.; Conchillo-Solé, O.; Gori, A.; Rinchai, D.; Thomas, R. J.; Champion, O. L.; Michell, S. L.; Kewcharoenwong, C.; et al. Exploiting the *Burkholderia pseudomallei* acute phase antigen BPSL2765 for structure-based epitope discovery/design in structural vaccinology. *Chem Biol* **2013**, *20* (9), 1147-1156. DOI: 10.1016/j.chembiol.2013.07.010.
- (25) Marchetti, F.; Capelli, R.; Rizzato, F.; Laio, A.; Colombo, G. The Subtle Trade-Off between Evolutionary and Energetic Constraints in Protein-Protein Interactions. *J Phys Chem Lett* **2019**, *10* (7), 1489-1497. DOI: 10.1021/acs.jpcclett.9b00191.
- (26) Paladino, A.; Woodford, M. R.; Backe, S. J.; Sager, R. A.; Kancherla, P.; Daneshvar, M. A.; Chen, V. Z.; Bourboulia, D.; Ahanin, E. F.; Prodromou, C.; et al. Chemical Perturbation of Oncogenic Protein Folding: from the Prediction of Locally Unstable Structures to the Design of Disruptors of Hsp90-Client Interactions. *Chemistry* **2020**, *26* (43), 9459-9465. DOI: 10.1002/chem.202000615.
- (27) Serapian, S. A.; Colombo, G. Designing Molecular Spanners to Throw in the Protein Networks. *Chemistry* **2020**, *26* (21), 4656-4670. DOI: 10.1002/chem.201904523.
- (28) Defays, D. An efficient algorithm for a complete link method. *Comput. J.* **1977**, *20* (4), 364-366.
- (29) Case, D. A.; Cheatham, T. E.; Darden, T.; Gohlke, H.; Luo, R.; Merz, K. M.; Onufriev, A.; Simmerling, C.; Wang, B.; Woods, R. J. The Amber biomolecular simulation programs. *J Comput Chem* **2005**, *26* (16), 1668-1688. DOI: 10.1002/jcc.20290.
- (30) *AMBER 2018*; 2018. (accessed).
- (31) Onufriev, A.; Bashford, D.; Case, D. A. Modification of the Generalized Born Model Suitable for Macromolecules. *J. Phys. Chem. B* **2000**, *104*, 3712-3720.
- (32) Weiser, J.; Shenkin, P. S.; Still, W. C. Approximate Atomic Surfaces from Linear Combinations of Pairwise Overlaps (LCPO). *J. Comput. Chem.* **1999**, *20*, 217-230.
- (33) Genheden, S.; Ryde, U. The MM/PBSA and MM/GBSA methods to estimate ligand-binding affinities. *Expert Opin Drug Discov* **2015**, *10* (5), 449-461. DOI: 10.1517/17460441.2015.1032936.
- (34) Morra, G.; Colombo, G. Relationship between energy distribution and fold stability: Insights from molecular dynamics simulations of native and mutant proteins. *Proteins* **2008**, *72* (2), 660-672. DOI: 10.1002/prot.21963.
- (35) Kuntz, I. D.; Blaney, J. M.; Oatley, S. J.; Langridge, R.; Ferrin, T. E. A geometric approach to macromolecule-ligand interactions. *J Mol Biol* **1982**, *161* (2), 269-288. DOI: 10.1016/0022-2836(82)90153-x.
- (36) Fischer, E. Einfluss der configuration auf die wirkung derenzyme. *Ber. Dt. Chem. Ges.* **1894**, *27*, 2985-2993.
- (37) Hammes, G. G. Multiple conformational changes in enzyme catalysis. *Biochemistry* **2002**, *41* (26), 8221-8228. DOI: 10.1021/bi0260839.
- (38) Morris, G. M.; Goodsell, D. S.; Halliday, R. S.; Huey, R.; Hart, W. E.; Belew, R. K.; Olson, A. J. Automated docking using a Lamarckian genetic algorithm and an empirical binding free energy function. *J. Comput. Chem.* **1998**, *19*, 1639-1662.
- (39) Rarey, M.; Kramer, B.; Lengauer, T.; Klebe, G. A fast flexible docking method using an incremental construction algorithm. *J Mol Biol* **1996**, *261* (3), 470-489. DOI: 10.1006/jmbi.1996.0477.
- (40) Muegge, I.; Martin, Y. C. A general and fast scoring function for protein-ligand interactions: a simplified potential approach. *J Med Chem* **1999**, *42* (5), 791-804. DOI: 10.1021/jm980536j.
- (41) Friesner, R. A.; Banks, J. L.; Murphy, R. B.; Halgren, T. A.; Klicic, J. J.; Mainz, D. T.; Repasky, M. P.; Knoll, E. H.; Shelley, M.; Perry, J. K.; et al. Glide: a new approach for rapid, accurate docking and scoring. 1. Method and assessment of docking accuracy. *J Med Chem* **2004**, *47* (7), 1739-1749. DOI: 10.1021/jm0306430.



Proteins

The primary purpose of the **protein approach** research is to comprehensively investigate the SARS-CoV-2 Spike protein, with a multifaceted approach encompassing the following key objectives:

Predicting immune recognition regions. Develop a simple and straightforward structure-dynamics-energy based strategy to comprehensively investigate the SARS-CoV-2 Spike protein to predict regions involved in immune recognition. This insight has the potential to guide the development of novel molecules for vaccine and diagnostic purposes. Notably, this approach has identified potentially reactive regions in the S protein stalk, currently undergoing experimental synthesis and testing.

This early analysis was covered in the paper entitled: "The Answer Lies in the Energy: How Simple Atomistic Molecular Dynamics Simulations May Hold the Key to Epitope Prediction on the Fully Glycosylated SARS-CoV-2 Spike Protein" ([Serapian S. A. et al.](#) J Phys Chem Lett. **2020** Oct 1;11(19):8084-8093).

Assessing Immune Response Variability. Analyze how mutations in the Spike protein impact the immune response, specifically by evaluating the efficacy of monoclonal antibodies against different SARS-CoV-2 variants. This research seeks to understand the extent to which these mutations affect the ability of the immune system to neutralize the virus.

This research was conducted in the paper with the title: "SARS-CoV-2 Spike Protein Mutations and Escape from Antibodies: a Computational Model of Epitope Loss in Variants of Concern" ([Triveri A. et al.](#) J Chem Inf Model. **2021** Sep 27;61(9):4687-4700).

Studying the Stability of Viral Variants. Investigate the stability of SARS-CoV-2 variants with Spike protein mutations and explore how these mutations influence the virus's ability to

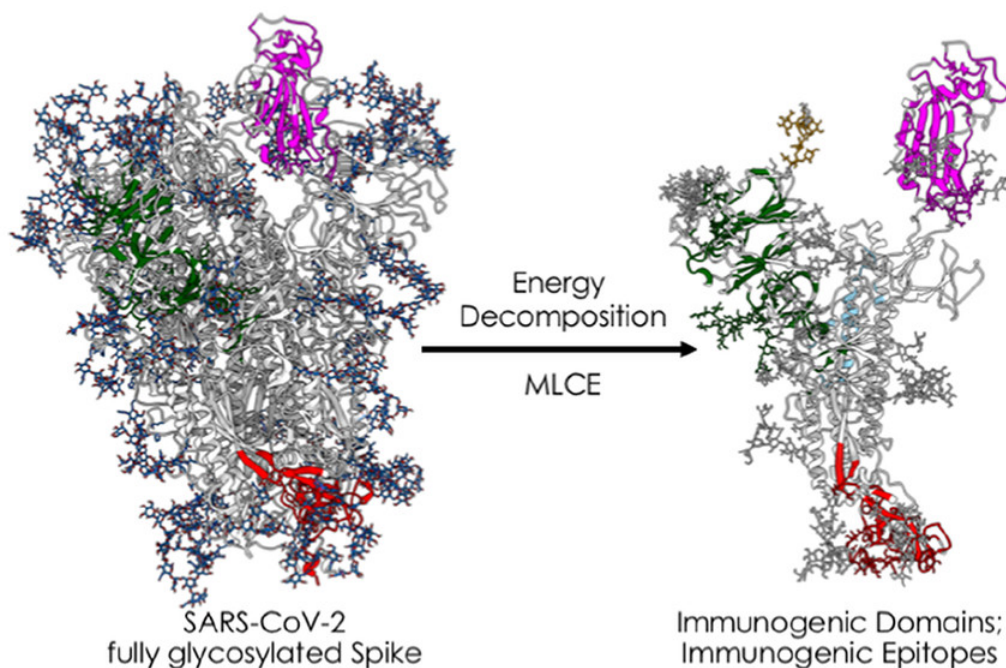
persist, transmit, and potentially cause severe disease. This research aims to provide insights into the dynamic nature of viral evolution.

This work has been published in the paper: “The Conformational Behaviour of SARS-Cov-2 Spike Protein Variants: Evolutionary Jumps In Sequence Reverberate In Structural Dynamic Differences” (Triveri A. et al. J Chem Theory Comput. **2023** Apr 11;19(7):2120-2134).

Exploring the Fatty Acid Binding Pocket. Examine the presence and conservation of the fatty acid binding pocket within the Spike protein across various SARS-CoV-2 variants. This analysis aims to identify potential druggable targets for therapeutic interventions and assess their relevance in the context of viral evolution. The manuscript of this latter research is in preparation.

1.12 Immunoreactivity of the WT SARS-CoV-2 spike protein

“The answer lies in the energy: how simple atomistic molecular dynamics simulations may hold the key to epitope prediction on the fully glycosylated SARS-CoV-2 spike protein”



1.12.1 Abstract

Herein, we use the original energy-decomposition approach outlined in the Method section to identify antigenic domains and antibody binding sites on the fully glycosylated S protein. Notably, our method relies solely on unbiased atomistic molecular dynamics simulations, eliminating the need for any prior knowledge of binding properties or arbitrary combinations of parameters extracted from simulations.

Our approach involves analyzing energy interactions among all intra-protomer amino acid and monosaccharide residue pairs, cross-comparing these interactions with structural data (i.e., residue-residue proximity). Through this analysis, we identify groups of spatially contiguous residues with weak energetic coupling to the rest of the protein, indicating potential immunogenic regions.

Validation of our results was achieved through comparison with experimentally confirmed structures of the S protein complexed with anti- or nanobodies. This validation process

enabled us to identify several subdomains with poor coupling, likely to accommodate multiple epitopes and possibly play a role in significant functional conformational changes.

Furthermore, we observed two distinct behaviors of the glycan shield. Glycans with stronger energetic coupling were found to be structurally relevant, providing protection to underlying peptidic epitopes. Conversely, glycans with weaker coupling might themselves be susceptible to antibody recognition. These predictions of immunoreactive regions offer a pathway to develop optimized antigens, such as recombinant subdomains and synthetic (glyco)peptidomimetics, with potential therapeutic applications. Additionally, employing similar predictive approaches could significantly enhance preparedness for future pandemic outbreaks.

1.12.2 Introduction

The knowledge acquired about recognition mechanisms and the determination of the detailed dynamic and structural characteristics of SARS-CoV-2 could help to be better prepared to tackle similar pandemics in the future by contrasting them more efficiently through the application of the same efficient and well-tested methods to new protein variants. More specifically, upon emergence of a new pathogen, generally portable computational methods could be advantageously exploited to rapidly identify and synthesize recombinant antigen or peptide-based vaccines.¹⁻¹²

For instance, the detailed dynamic and structural knowledge set the stage for understanding the molecular bases of S protein recognition by the host's immune system, providing information on which physico-chemical determinants are required to elicit functional antibodies.¹²⁻¹⁴ Such understanding could then be exploited to design and engineer improved antigens based on S, for instance by identifying antigenic domains that can be expressed in isolation or short sequences (epitopes) that can be mimicked by synthetic peptides¹⁵⁻²⁰: this would be a crucial first step in the selection and optimization of candidate vaccines and therapeutic antibodies (on top of those already in development), as well as in the development of additional serologic diagnostic tools.

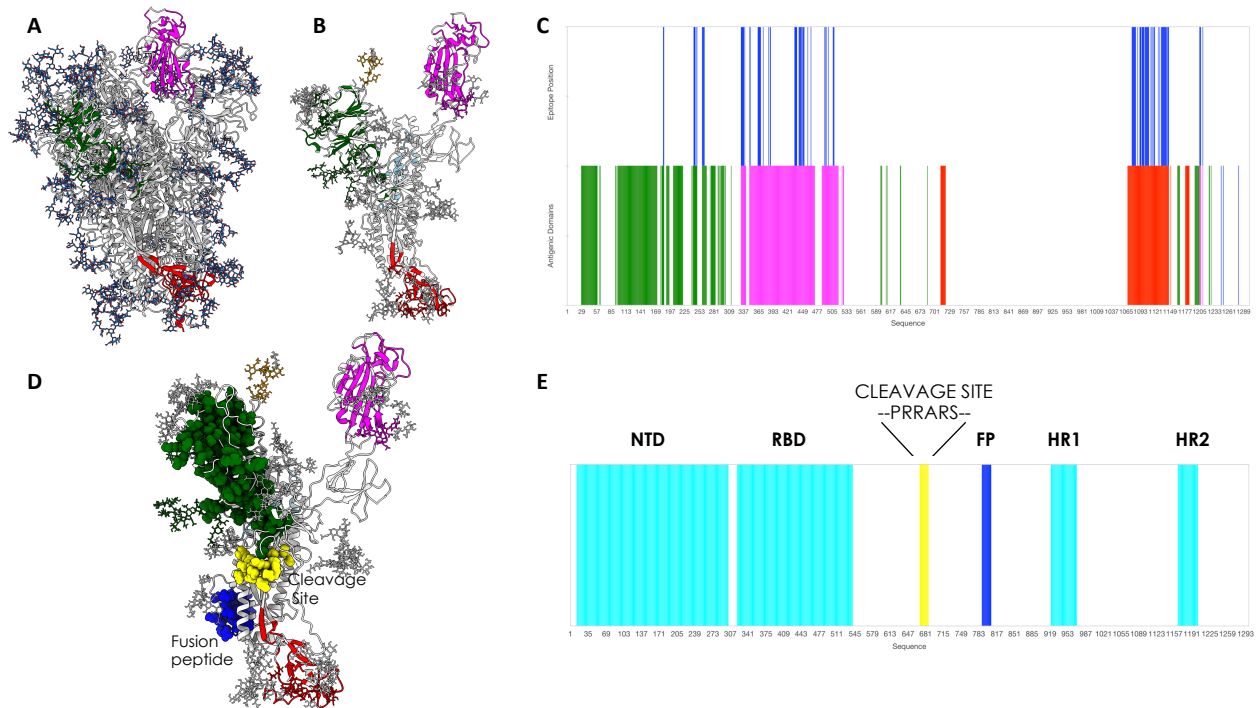


Figure 1. 3D structure, glycosylation and location of antigenic domains and epitopes on SARS-CoV-2 fully glycosylated Spike protein. **A.** The starting, fully glycosylated Spike protein trimer. The coating oligosaccharides are colored in dark blue. The predicted antigenic domains are colored on the structure of one protomer. **B.** Isolated protomer with the most antigenic domains, detected via MLCE with the 15% cutoff, highlighted in colors: the antigenic part in the N-Domain is dark green; the part in the RBD is magenta; the part in the C-terminal domain is dark red. Oligosaccharides that define or are part of antigenic domains are also colored. Oligosaccharides that have a structural role and show strong energetic coupling to the protein are in white. **C.** The predicted antigenic sequences projected on the sequence of the protein. The bottom line reports the sequences defined as antigenic domains, with the same color code as in **B**. The top bar reports the location of peptidic epitopes identified with the most restrictive definition. **D.** Physical interaction between the boundaries of the predicted antigenic domain in the N-terminal region and the cleavage site of S. This subfigure also shows the physical proximity of the predicted C-terminal uncoupled region with the fusion peptide. **E.** Domain organization of the spike protein projected on the sequence. Numbering and domain definitions obtained from UNIPROT (<https://www.uniprot.org/uniprot/P0DTC2>).

In this first work, we analyze representative 3D conformations of the full-length trimeric S protein in its fully glycosylated form (**Figure 1**), extracted from atomistic molecular dynamics (MD) simulations provided by the Woods group,^{13, 21} to predict immunogenic regions.

To this end, a simple *ab initio* epitope prediction method that we previously described for unmodified proteins is optimized and extended to cover glycoproteins.²²⁻²⁴ The method is based on the idea that antibody-recognition sites (epitopes) may correspond to localized regions only exhibiting low-intensity energetic coupling with the rest of the structure. Otherwise, putative interacting patches are hypothesized to be characterized by non-optimized intramolecular interactions with the remainder of the protein. Actual binding to an external partner such as an Ab is expected to occur if favorable intermolecular interactions determine a lower free energy for the bound than the unbound state.^{22, 24-25} Furthermore, minimal energetic coupling with the rest of the protein provides these subregions with greater conformational freedom to adapt to and be recognized by a binding partner, as well as improved tolerance to mutations at minimal energetic expense without affecting the protein's native organization and stability in a way that could be detrimental for the pathogen: all these properties are indeed hallmarks of Ab-binding epitopes.

This approach is indeed able to identify regions, also comprising carbohydrates, that recent structural immunology studies have shown to be effectively targeted by antibodies. On the same basis, our method predicts several additional potential immunogenic regions (currently still unexplored) that can then be used for generating optimized antigens, either in the form of recombinant isolated domains or as synthetic peptide epitopes. Finally, our results help shed light on the mechanistic bases of the large-conformational changes underpinning biologically relevant functions of the protein.

This method is one of the first that permits to discover epitopes in the presence of glycosylation (an aspect that is often overlooked), starting only from the analysis of the physico-chemical properties of the isolated antigen in solution. Importantly, the method does not require any prior knowledge of antibody binding sites of related antigenic homologs and does not need to be trained/tuned with data sets or ad hoc combinations of information on sequences, structures, SASA or geometric descriptors. The procedure is thus immediately and fully portable to other antigens.

1.12.3 Results and Discussion

To reveal the regions of the S protein that could be involved in antibody (Ab) binding, we employ a combination of the Energy Decomposition (ED) and MLCE (Matrix of Low Coupling Energies) methods, which we previously introduced and validated^{22-24, 26-34} and discuss in full in the Methods section.

Starting from 6 combined 400 ns replicas of atomistic molecular dynamics simulations of the fully glycosylated S protein in solution¹³ (built from PDB ID: 6VSB⁹), we isolate a representative frame from each of the three most populated clusters. ED and MLCE analyses of protein energetics assess the interactions that each amino acid and glycan residue in S protomers establishes with every other single residue in the same protomer. We compute the nonbonded part of the potential energy (van der Waals, electrostatic interactions, solvent effects) implicitly, via an MM/GBSA calculation (molecular mechanics/generalized Born and surface area continuum solvation³⁵), obtaining, for a protomer composed of N residues (including monosaccharide residues on glycans), a symmetric $N \times N$ interaction matrix M_{ij} . Eigenvalue decomposition of M_{ij} highlights the regions of strongest and weakest coupling. The map of pairwise energy couplings can then be filtered with topological information (namely, the residue-residue contact map) to identify localized networks of low-intensity coupling (i.e., clusters of spatially close residue pairs whose energetic coupling to the rest of the structure is weak and not energetically optimized through evolution).

In this model, when these fragments are located or near the surface, contiguous in space and weakly coupled to the protein's 'stability core', they represent potential interaction regions (i.e., epitopes). Once interacting vicinal residue pairs (i, j) are identified by cross-comparison with the residue-residue contact map (vide supra and Methods Section), identification of poorly coupled regions representing potential epitopes proceeds as follows. Residue pairs are firstly ranked in order of increasing interaction intensity (from weakest to strongest). Two distinct sets of energetically decoupled regions are then mapped by applying two distinct cutoffs ('softness thresholds') to the residue pair list: either from the first 15% or from the first 5% of the ranked pairs (i.e., the 5% or 15% of the residue pairs with the weakest energetic coupling).

The less restrictive 15% cutoff subdivides the full-length, fully folded S protein into potentially immunoreactive domains (see **Figure 1B.,C.** and Methods).^{23, 25, 27} The goal is to uncover regions that may normally be hidden from recognition by Abs in the native protein structure, but that can be experimentally expressed as isolated domains. Highly reactive neutralizing epitopes may in fact be present only in specific but transient conformations that are not immediately evident in the static X-ray and EM models of the protein or are not accessible even to large scale MD simulations. Presenting these (cryptic) regions for Ab binding through their isolated parent domains may prove more advantageous in developing new immunogens.^{23, 27}

The more stringent epitope definition (5% cutoff) narrows the focus on those (smaller) intra-domain regions that could be directly involved in forming the interface with Abs, and that can then be used to guide the engineering of optimized antigens in the form of synthetic epitope peptidomimetics. In this context, to be defined as epitopes, the energetically uncoupled regions must be at least 6-residue long.

Upon using the larger cutoff value, a large cluster of energetically unoptimized residue pairs localize at the Receptor Binding Domain, correctly identifying it as the most antigenic unit in the S protein's 'RBD up' protomer (**Figure 1B., C.** magenta colored domain). Interestingly, when the lowest energy-coupled residue pairs are mapped onto the 'up' RBD of all three 3D structures isolated from MD, there is a large overlap with regions recognized by Abs and nanobodies (revealed by recent X-ray and cryo-EM structures). Importantly, for example, our calculation correctly identifies the binding region of mAb CR302236 (PDB ID 6W41), known to target a cryptic epitope that is exposed only upon significant structural rearrangement of the protein¹² (**Figure 2** and **Figure 4**).

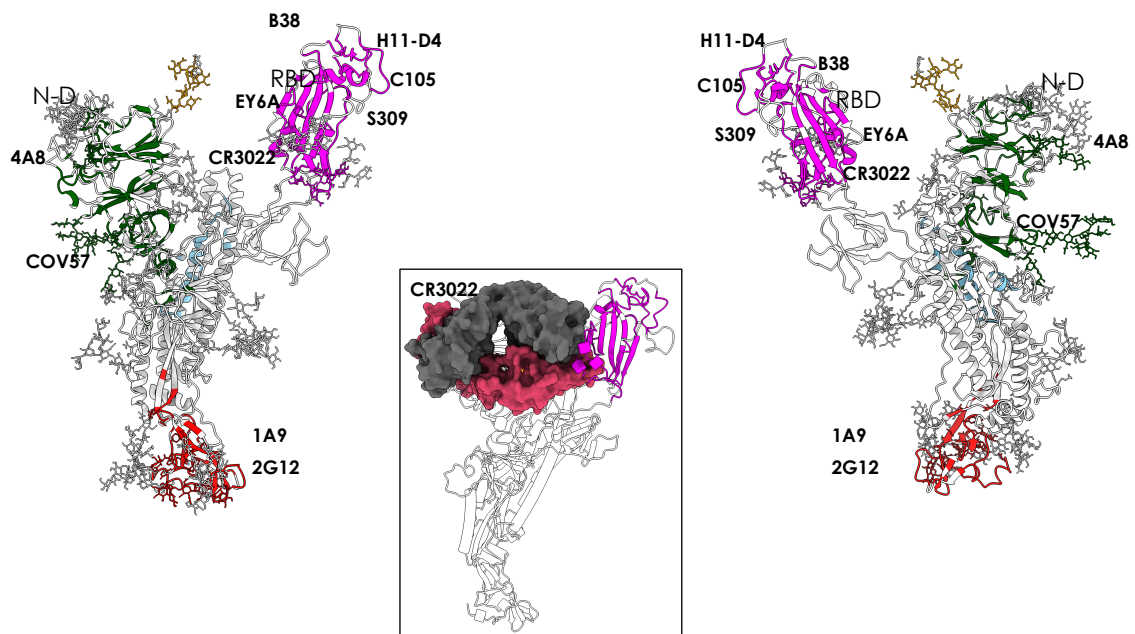


Figure 2. Antigenic domains and location of binding antibodies (in two different orientations of the same conformation of the protein). The clusters of residues defining antigenic domains (dark green in the N-domain, magenta in the RBD, red in the C-terminal region) and the positions of the various antibodies whose structures and interactions in complexes with the full-length protein have been described. The inset indicates the identification of the cryptic immunoreactive region binding CR3022.

A second domain that is found to host a large network of non-optimized interactions corresponds to the N-terminal domain (**Figure 1B., C., D.**). The latter has been shown to bind the antibody 4A8 (PDB ID 7C2L).

A third region predicted to be highly antigenic coincides with the central/C-terminal part of the S1A domain. In a recent cryo-EM study of polyclonal antibodies binding to the S protein, this substructure was shown to be in the vicinity of the density for COV57 Fab(s), a novel Ab whose neutralizing activity showed no correlation with that of RBD-targeting Abs³⁸ (**Figure 1B., D.**). We note here that MERS Ab 7D10 also binds in this region.³⁹

Furthermore, MLCE identifies a potentially highly reactive region in the S2 domain of the protein, in the CD region. This domain contains the epitope recently found to engage with 1A9,⁴⁰ an antibody recently shown to cross-react with S proteins of human, civet, and bat coronaviruses. This analysis also recognizes a potential antigenic region in a carbohydrate cluster located in the S2 domain of the protein: intriguingly, has been found that an oligosaccharide-containing epitope centered around this predicted region is targeted by the glycan-dependent antibody HIV-1 bnAb 2G12⁴¹ (**Figure 1, 2**).

Identification of energetically uncoupled domains also has mechanistic implications. Regions that are not involved in major intramolecular stabilization can be displaced from the biomolecule at minimal energetic costs, sustaining large-scale conformational changes that typically underpin its biological function. The boundary of the (uncoupled) N-terminal region (**Figure 1**, dark green domain) lies in physical proximity to the furin-targeted motif RRAR, which is essential for pre-activation of SARS-CoV-2 Spike through proteolysis. Thus, the large uncoupled region of the N-domain can synergize with (and favor, through domain displacement) cleavage of this motif, ultimately favoring detachment of S1-domain and release of the S2 fusion machinery.^{9-11, 42} Furthermore, the beta-sheet at the initial boundary of the C-terminal domain in S2 (Red domain in **Figure 1**) is in close physical proximity to the fusion peptide (**Figure 1D., E.**). Here, it would be reasonable to expect that exposure or conformational rearrangement of the C-terminal domain are favored by its non-optimized interactions with the core of the S protein stalk and would in turn optimally expose the fusion peptide favoring its integration with the host membrane.⁴²

Overall, these findings support the validity of our approach in identifying protein domains that can be aptly used as highly reactive immunogens, as they are most likely to be targeted by a humoral immune response. Our analysis predicts that regions other than the S protein RBD may represent alternative targets for neutralization or functional perturbation of SARS-CoV-2. On the one hand, this may be important considering the fact that RBD can also be

the target on non-neutralizing antibodies, e.g. 3022³⁶. Indeed, using cocktails of antibodies to target different regions of S has been proposed as a viable therapeutic option.³⁷

Turning to our more stringent definition of epitope, based exclusively on the top 5% of the most weakly coupled residue pairs (5% cutoff), we next focus on those regions of the S antigen that can be involved in forming contacts with antibodies.

Importantly, one predicted conformational epitope with sequence (348)A-(352)A-(375)S-(434)IAWNS(438)-(442)DSKVGG(447)-(449)YNYL(452)-(459)S-(465)E-(491)PLQS(494)-(496)Q-(507)PYR(509) encompasses regions of the S protein in contact with antibodies **C105** (6xcn.pdb)³⁸, **S309**(6wpt.pdb; 6wps.pdb),⁴³ **AB23** (7byr.pdb)⁴⁴; with nanobody **H11-D4** (6z43.pdb); and with a reported synthetic nanobody (7c8v.pdb) (**Figure 3**).

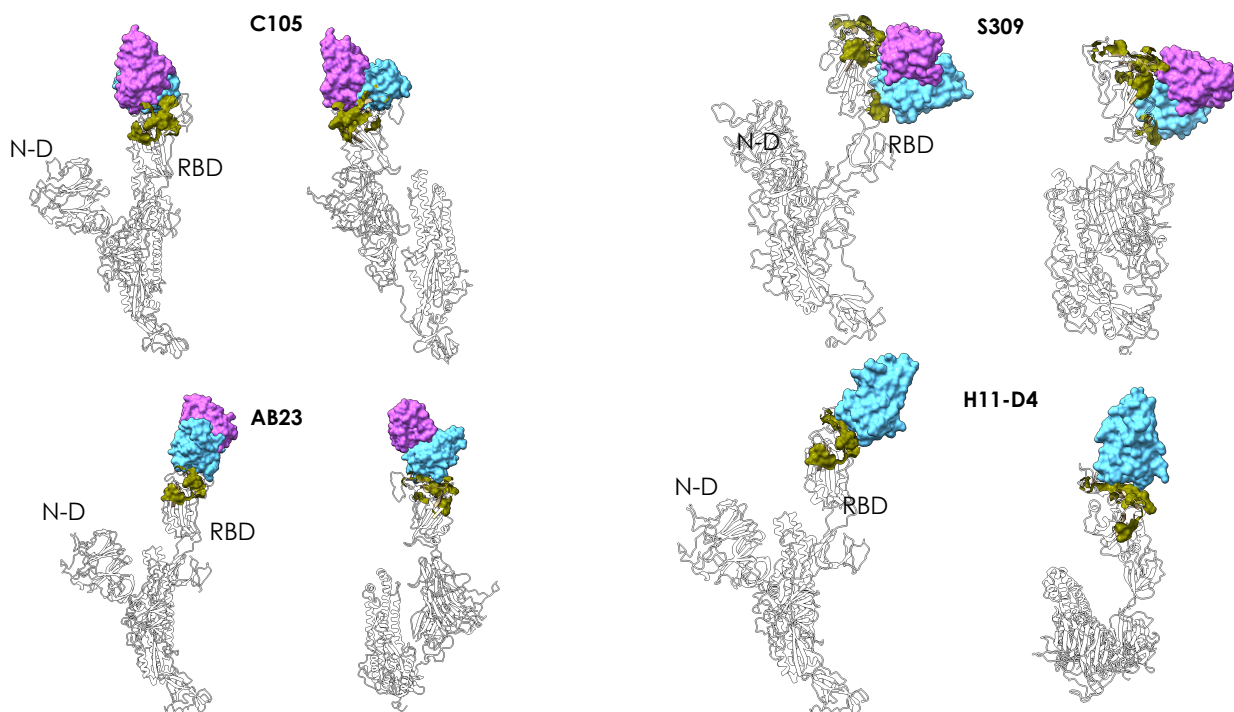


Figure 3. Peptidic epitopes predicted on the surface of the RBD using the restrictive definition of antigenic region and comparison with known Ab-complexes. The X-ray structures of the complexes between the various antibodies reported in the figure (C105, S309, AB23, and nanobody H11-D4) and the full-length Spike protein are superimposed to the structure of the protomer used here for prediction. The green surfaces indicate the location of MLCE epitope predictions. The Fabs of the antibodies or of the nanobody are depicted as accessible surfaces in shades of blue.

Interestingly, an additional predicted patch comprising a set of decorating carbohydrates is correctly predicted to be part of the interface with antibody S309 (6wpt.pdb; 6wps.pdb)⁴³, with aminoacidic sequence (332)ITNLC(336)-(361)C and with the (N334-linked) fucosylated

N-glycan chitobiose core (Man β 1-4GlcNAc β 1-4[Fuca α 1-6]GlcNAc β -Asn)⁴⁵. This predicted region sits notably close to the RBD interaction surface with ACE2.

Antibody EY6A (6zdh.pdb) binds the RBD in the region of the cryptic epitope described by Wilson and collaborators³⁶ (**Figure 4**).

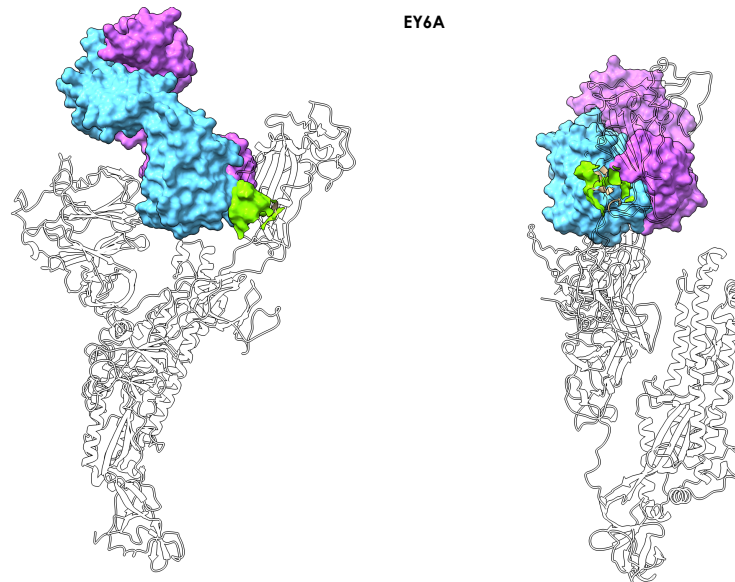


Figure 4. Antibody EY6A – Spike complex. The figure shows how Antibody **EY6A** (7byr.pdb) binds the RBD in the region of a cryptic epitope. The MLCE-predicted epitope region is shown in light green (lime) in two different orientations, indicating substantial contact formation with the antibody.

Importantly, our predicted patch (365)YSVLYN(370)-(384)PTKLN(388) covers a significant part of the epitope. Once again, it is worth remarking that identification of this potentially immunoreactive patch is simply and exclusively obtained from structural and energetic interaction data generated for a protomer of the glycosylated, isolated S protein, after unbiased MD simulation (see Methods section).

With the more restrictive epitope prediction cutoff we clearly identify a reactive area in the N-terminal domain of the Spike protein. The predicted patch (184)GN(185)-(242)LAL(244)-(246)R-(248)Y-(258)WTAGA(262) contains residues R246 and W258 which were described as central determinants for contact between the N-terminal domain and antibody **4A8** (7c2l.pdb)³⁷ (**Figure 5**).

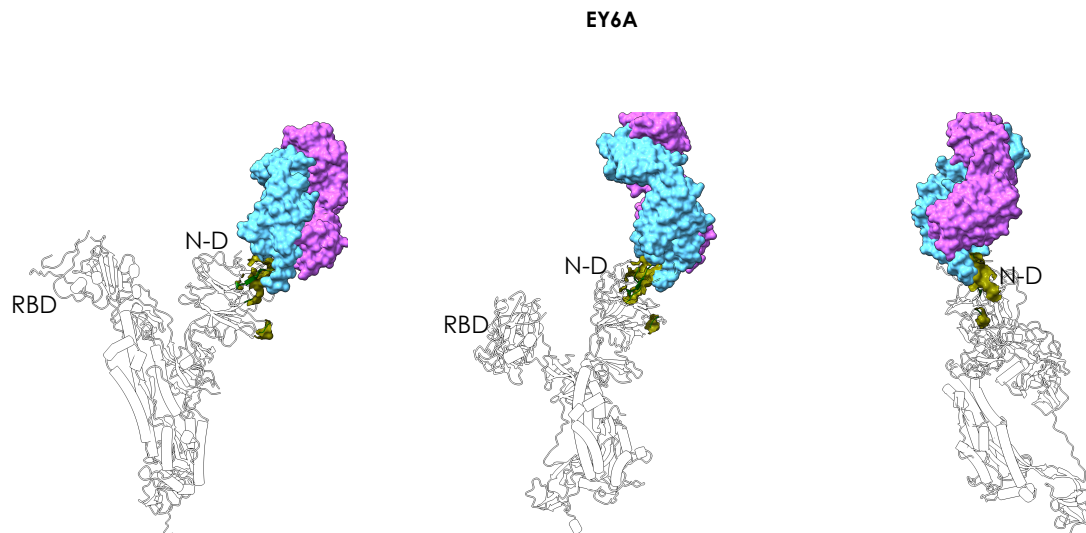


Figure 5. Antibody 4A8 – Spike complex. The figure shows how Antibody **4A8** binds the N-domain of Spike, supporting correct prediction of the epitope. The MLCE-predicted epitope region is shown in green in three different orientations, indicating substantial contact formation with the antibody. The Fab of the antibody is depicted as accessible surface in shades of blue. Finally, the restrictive prediction identifies the sequence spanning residues 1076-1146, which includes amino acids 1111-1130, experimentally identified as the epitope for the monoclonal mAb **1A9**⁴⁰. Specifically our identified reactive sequence is the following: (1076)TTAPAICH(1083)-(1087)A-(1092)REG(1094)-(1096)FVSNGHWFVTQRN(1108)-(1112)P-(1114)I-(1116)T-(1118)DN(1119)-(1126)C-(1129)V-(1132)IVNNTVYDPLQELD(1146).

In general, our approach is able to identify potential immunoreactive domains and epitopes of the Spike protein based only on structural and energetic information. Sequences predicted to be reactive using the restrictive epitope definition (5% cutoff) can be used for generating optimized antigens in the form of synthetic peptide epitopes. Engineering such epitopes would entail the synthesis of conformationally preorganized peptidomimetics of the ‘natural’ reactive regions, with intra- and extracellular stability enhanced through, e.g, a combination of natural and non-natural aminoacids, which could reproduce the main structural and energetic conditions required to elicit a humoral immune response, as well as constituting candidates for vaccine development. Furthermore, reactive peptides thus identified may be suitable for use as baits in serologic diagnostic applications (e.g., in ELISA assays and in microarrays), to capture and detect not only circulating antibodies that are expressed in response to SARS-CoV-2 infections but also those that are endowed with neutralization activity and thus potentially predicting the infection outcome. As a further application, these peptide-based baits can represent a useful tool for isolating new mAbs and the screening of small molecules for drug development.

One of the most significant aspects of this approach is that the S protein's entire glycan shield is explicitly accounted for in the prediction of the immunoreactive regions. Indeed, the various oligosaccharide chains appear to behave differently (see differential coloring of oligosaccharide chains in **Figure 1**). In light of their stronger energetic coupling to other areas of the protein, some of the glycans are not recognized as epitopes, and thus form an integral part of the stabilizing intramolecular interaction network of S (white chains in **Figure 1B.**); on the other hand, MLCE also identifies a second subset of poorly coupled oligosaccharides as potentially reactive epitopes (or part thereof) (colored oligosaccharide chains in **Figure 1B.**; carbohydrate cluster in S2 targeted by the glycan-dependent antibody HIV-1 bnAb 2G12, see **Figure 1B., 2**), highlighting potential vulnerable spots in the glycan shield that could be exploited to design novel immunoreagents and vaccine candidates.

The portion of the glycan shield falling within the former category thus mainly serves to *protect* the protein from recognition by antibodies and consequently enhances viral infectiousness, as well as providing extra structural support. Two such glycans are further exemplified in **Figure 6**.

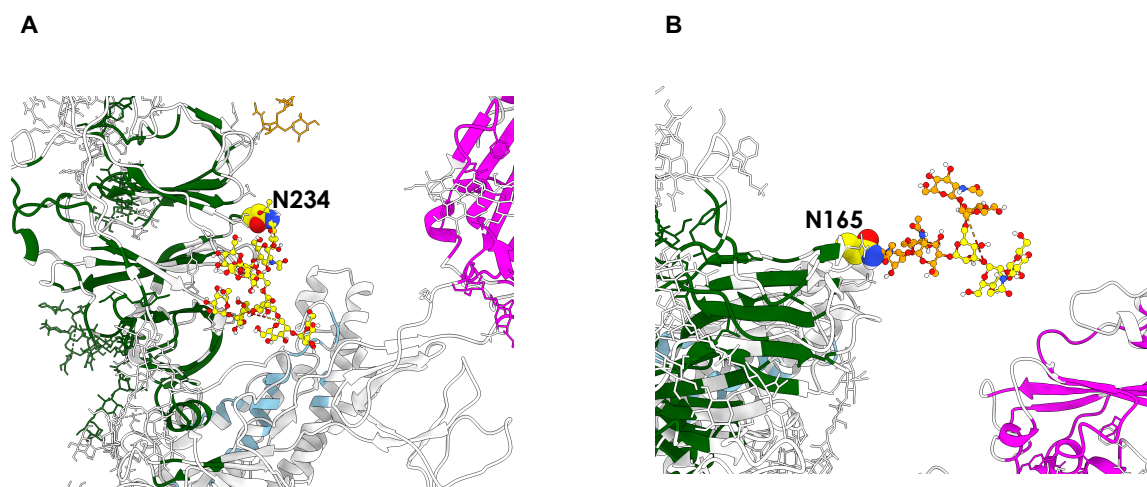


Figure 6. Glycans with different roles on Spike. A. The glycan chain attached to N234, which is predicted to be part of the networks of stabilizing interactions within the protein. **B.** The glycan chain attached to N165 is predicted to play a double role, a stabilizing one (yellow units) and an immunoreactive one (orange units).

The first is the entire oligosaccharide fragment bound to N234 (**Figure 6A.**), which is recognized by Amaro and coworkers as being crucial in ‘propping up’ the RBD.¹² Experimental deletion of *N*-glycans at this position by way of a mutation to Ala significantly modifies the conformational landscape of the protein’s RBD.⁴⁶ The second is the portion of

the N165-linked glycan whose subunits are rendered in yellow (**Figure 6B.**): consistent with experimental studies indicating N165-linked oligosaccharides as structural modulators,⁴⁶ we also find that the portion in question is *not* identified as a potential epitope, being consequently involved in diverting antibodies from targeting the region around N165 and thus preserving control of the S protein's structural dynamics.

Reflecting the multifaceted roles of the glycan shield, the remaining part of the N165-linked glycan (**Figure 6B; orange**) appears instead to belong to the category of glycans that *are* potentially able to act as epitopes, since, unlike the part in yellow, we do detect it to be decoupled from the rest of the protomer.

It is particularly significant to underline that MLCE, whose physical basis is to identify non-optimized interaction networks, detects peptidic epitopes even when they are in proximity of (optimized, non-immunogenic) shielding carbohydrates. In light of this, it is reasonable to suggest that the protective effect of these particular carbohydrates may be circumvented and neutralized by exposing the underlying peptidic substructures. Furthermore, information on oligosaccharides identified as epitope constituents can be exploited to design glycomimics or glycosylated peptides as synthetic epitopes.

The latter aspect is indeed particularly relevant: small synthetic molecules that mimic antigenic determinants (and effectively act as their minimal surrogates) offer enticing opportunities to develop immunoreagents with superior characteristics in terms of ease of handling, reproducibility of batch-to-batch production, ease of purification, sustainable cost, and better stability under a variety of conditions. Furthermore, production of these molecules greatly reduces the risk of cross-reactivity with any copurified antigens, which is instead rife when dealing with recombinant proteins. In contrast to smaller peptides or sugar-decorated peptidomimetics, a full-length recombinant antigenic protein (or any protein-based detection device) would typically require more stringent conditions (e.g., in terms of temperature and humidity) for storage, transport, and management in order to preserve the protein in its properly folded active form. The same would be true for other vaccinal solutions such as deactivated pathogens.

Overall, this work confirms how simple and transparent structural and physico-chemical understanding of the molecule that is the key player in SARS-CoV-2 viral infection can be harnessed to guide the prediction of (in some cases experimentally confirmed) regions, that are involved in immune recognition and to understand its molecular bases. Agreement with experiment confirms that knowledge generated in the process has the potential of being translated into new molecules for vaccine and diagnostic development. In this context, we

have also identified potentially reactive regions in the S protein stalk that are currently under experimental synthesis and testing.

Furthermore, potential functional implications offered by the approach are illustrated by the fact that domains/regions relevant for the protein's biological activation are naturally identified. This renders the approach well-suited to identify subtle functional variations in mutants of the S proteins. Finally, the possibility of accurately partitioning such a complex system in functional subunits could aptly be exploited in the parameterization of coarse-grained models to simulate the system at longer timescales.

This kind of structure-based computational approach can clearly expand the scope of simple structural analysis and molecular simulations. In applicative terms, generation of synthetic libraries based on predicted/identified epitopes (with possible addition of sugars) would boost selection and screening of antigens for vaccine development.

1.12.4 Methods

Structure Selection from Molecular Dynamics (RBD Clustering)

Coordinates of the fully glycosylated SARS-CoV-2 S protein's 'RBD up' protomer featured in this work originate from molecular dynamics (MD) simulations by Woods and coworkers,¹³ based on PDB ID: 6VSB. Throughout this work, we retain exactly the same forcefield parameters used by Woods *et al.* in their MD simulations: all residues except glycosylated asparagines are treated using the *ff14SB* forcefield,⁴⁷ whereas glycans and glycosylated asparagines are modeled using the *GLYCAM_06j* forcefield⁴⁵.

Clustering is based on root-mean-squared deviation of C_{α} atoms of the RBD domain in the 'RBD up' protomer and performed with the *cpptraj* utility in *AmberTools* (version 17)⁴⁸ after concatenating all six independent MD replicas and aligning them with the 'autoimage' command. The chosen method is the Hierarchical Agglomerative Algorithm⁴⁹, with an epsilon value of 0.5. From each of the three most populated clusters, we isolate one representative frame, from which we retain the 'RBD up' protomer and its glycans, whilst again using *cpptraj* to discard all solvent molecules, ions, and the two 'RBD down' protomers. All subsequent calculations on these three 'RBD up' protomer models (**Follows the step of MINIMIZATION, MM/GBSA CALCULATIONS, ENERGY DECOMPOSITION and MLCE calculations**) are listed chronologically in the subsections in the Methods section).

1.12.5 References

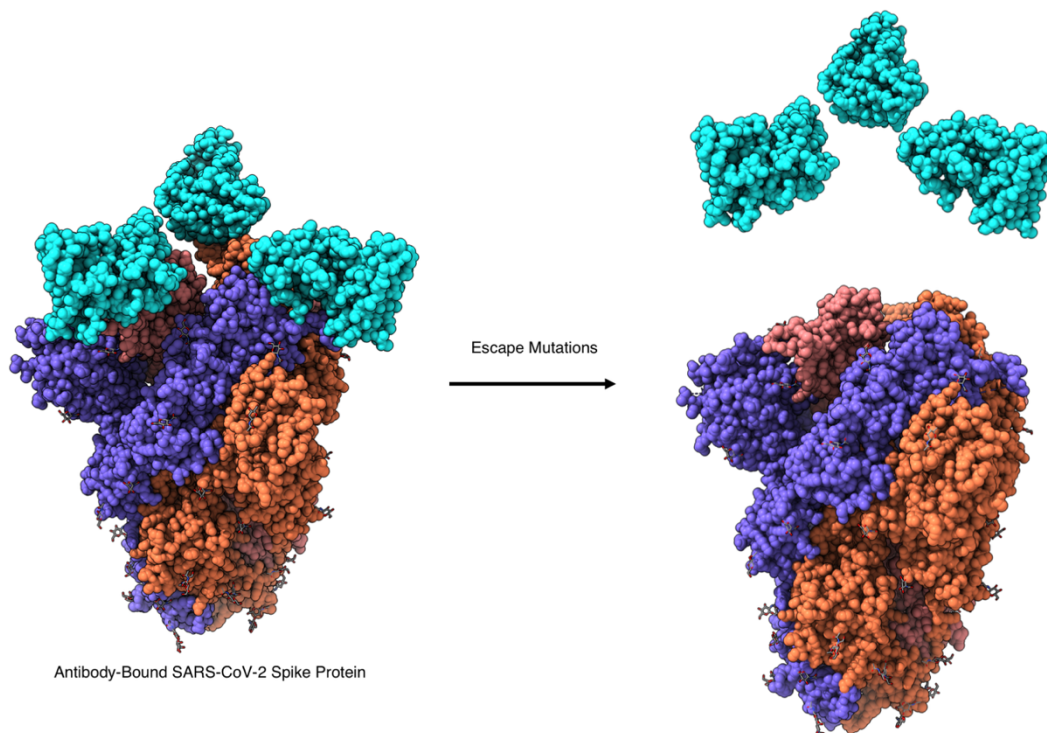
- (1) Kupferschmidt, K.; Cohen, J., Race to find COVID-19 treatments accelerates. *Science* 2020, 367 (6485), 1412.
- (2) Li, G.; De Clercq, E., Therapeutic options for the 2019 novel coronavirus (2019-nCoV). *Nat. Rev. Drug Disc.* 2020, 19, 149-150.
- (3) Liu, C.; Zhou, Q.; Li, Y.; Garner, L. V.; Watkins, S. P.; Carter, L. J.; Smoot, J.; Gregg, A. C.; Daniels, A. D.; Jervey, S.; Albaiu, D., Research and Development on Therapeutic Agents and Vaccines for COVID-19 and Related Human Coronavirus Diseases. *ACS Central Science* 2020, 6 (3), 315-331.
- (4) Romagnoli, S.; Peris, A.; De Gaudio, A. R.; Geppetti, P., SARS-CoV-2 and COVID-19: From the Bench to the Bedside. *Physiological Reviews* 2020, 100 (4), 1455-1466.
- (5) Andersen, K. G.; Rambaut, A.; Lipkin, W. I.; Holmes, E. C.; Garry, R. F., The proximal origin of SARS-CoV-2. *Nature Medicine* 2020, 26 (4), 450-452.
- (6) Weiss, S. R.; Leibowitz, J. L., Coronavirus pathogenesis. *Adv. Virus Res* 2011, 81, 85-164.
- (7) Yan, R.; Zhang, Y.; Li, Y.; Xia, L.; Guo, Y.; Zhou, Q., Structural basis for the recognition of the SARS-CoV-2 by full-length human ACE2. *Science* 2020, eabb2762.
- (8) Tortorici, M. A.; Veesler, D., Structural insights into coronavirus entry. *Adv Virus Res* 2019, 105, 93-116.
- (9) Wrapp, D.; Wang, N.; Corbett, K. S.; Goldsmith, J. A.; Hsieh, C.-L.; Abiona, O.; Graham, B. S.; McLellan, J. S., Cryo-EM structure of the 2019-nCoV spike in the prefusion conformation. *Science* 2020, eabb2507.
- (10) Walls, A. C.; Tortorici, M. A.; Bosch, B.-J.; Frenz, B.; Rottier, P. J. M.; DiMaio, F.; Rey, F. A.; Veesler, D., Cryo-electron microscopy structure of a coronavirus spike glycoprotein trimer. *Nature* 2016, 531 (7592), 114-117.
- (11) Walls, A. C.; Park, Y.-J.; Tortorici, M. A.; Wall, A.; McGuire, A. T.; Veesler, D., Structure, Function, and Antigenicity of the SARS-CoV-2 Spike Glycoprotein. *Cell* 2020, 181 (2), 281-292.e6.
- (12) Casalino, L.; Gaieb, Z.; Dommer, A. C.; Harbison, A. M.; Fogarty, C. A.; Barros, E. P.; Taylor, B. C.; Fadda, E.; Amaro, R. E., Shielding and Beyond: The Roles of Glycans in SARS-CoV-2 Spike Protein. *bioRxiv* 2020, 2020.06.11.146522.
- (13) Grant, O. C.; Montgomery, D.; Ito, K.; Woods, R. J., 3D Models of glycosylated SARS-CoV-2 spike protein suggest challenges and opportunities for vaccine development. *bioRxiv* 2020, 2020.04.07.030445.
- (14) Sikora, M.; von Bülow, S.; Blanc, F. E. C.; Gecht, M.; Covino, R.; Hummer, G., Map of SARS-CoV-2 spike epitopes not shielded by glycans. *bioRxiv* 2020, 2020.07.03.186825.
- (15) Peri, C.; Gagni, P.; Combi, F.; Gori, A.; Chiari, M.; Longhi, R.; Cretich, M.; Colombo, G., Rational epitope design for protein targeting. *ACS Chemical Biology* 2013, 8, 397-404.
- (16) Gourlay, L.; Peri, C.; Bolognesi, M.; Colombo, G., Structure and Computation in Immunoreagent Design: From Diagnostics to Vaccines. *Trends in Biotechnology* 2017, 35 (12), 1208-1220.
- (17) Smith, C. C.; Entwistle, S.; Willis, C.; Vensko, S.; Beck, W.; Garness, J.; Sambade, M.; Routh, E.; Olsen, K.; Kodysh, J.; O'Donnell, T.; Haber, C.; Heiss, K.; Stadler, V.; Garrison, E.; Grant, O. C.; Woods, R. J.; Heise, M.; Vincent, B. G.; Rubinsteyn, A., Landscape and Selection of Vaccine Epitopes in SARS-CoV-2. *bioRxiv : the preprint server for biology* 2020, 2020.06.04.135004.
- (18) De Gregorio, E.; Rappuoli, R., From empiricism to rational design: a personal perspective of the evolution of vaccine development. *Nat. Rev. Immunol.* 2014, 14 (7), 505-514.
- (19) Rappuoli, R.; Bottomley, M. J.; D'Oro, U.; Finco, O.; De Gregorio, E., Reverse vaccinology 2.0: Human immunology instructs vaccine antigen design. *The Journal of Experimental Medicine* 2016, 213, 469-481.
- (20) Thomas, S.; Dilbarova, R.; Rappuoli, R., Future Challenges for Vaccinologists. *Methods Mol Biol* 2016, 1403, 41-55.
- (21) Grant, O.; Woods, R. J., *Glycosylated Swiss-model molecular dynamics trajectory of SARS-CoV-2 spike glycoprotein*. 2020.
- (22) Scarabelli, G.; Morra, G.; Colombo, G., Predicting interaction sites from the energetics of isolated proteins: a new approach to epitope mapping. *Biophys. J.* 2010, 98 (9), 1966-1975.
- (23) Gourlay, L. J.; Peri, C.; Ferrer-Navarro, M.; Conchillo-Solé, O.; Gori, A.; Rinchai, D.; Thomas, R. J.; Champion, O. L.; Michell, S. L.; Kewcharoenwong, C.; Nithichanon, A.; Lassaux, P.; Perletti, L.; Longhi, R.; Lertmemongkolchai, G.; Titball, R. W.; Daura, X.; Colombo, G.; Bolognesi, M., Exploiting the *Burkholderia pseudomallei* Acute Phase Antigen BPSL2765 for Structure-Based Epitope Discovery/Design in Structural Vaccinology. *Chem. Biol.* 2013, 20, 1147-1156.
- (24) Marchetti, F.; Capelli, R.; Rizzato, F.; Laio, A.; Colombo, G., The Subtle Trade-Off between Evolutionary and Energetic Constraints in Protein-Protein Interactions. *J. Phys. Chem. Lett.* 2019, 10 (7), 1489-1497.

- (25) Genoni, A.; Morra, G.; Colombo, G., Identification of Domains in Protein Structures from the Analysis of Intramolecular Interactions. *J. Phys. Chem. B* **2012**, *116* (10), 3331-3343.
- (26) Soriani, M.; Petit, P.; Grifantini, R.; Petracca, R.; Gancitano, G.; Frigimelica, E.; Nardelli, F.; Garcia, C.; Spinelli, S.; Scarabelli, G.; Fiorucci, S.; Affentranger, R.; Ferrer-Navarro, M.; Zacharias, M.; Colombo, G.; Vuillard, L.; Daura, X.; Grandi, G., Exploiting antigenic diversity for vaccine design: the chlamydia ArtJ paradigm. *J. Biol. Chem.* **2010**, *285* (39), 30126-30138.
- (27) Lassaux, P.; Peri, C.; Ferrer-Navarro, M.; Gourlay, L.; Gori, A.; Conchillo-Solé, O.; Rinchai, D.; Lertmemongkolchai, G.; Longhi, R.; Daura, X.; Colombo, G.; Bolognesi, M., A structure-based strategy for epitope discovery in Burkholderia pseudomallei OppA antigen. *Structure* **2013**, *21*, 1-9.
- (28) Gourlay, L. J.; Lassaux, P.; Thomas, R. J.; Peri, C.; Conchillo-Sole, O.; Nithichanon, A.; Ferrer-Navarro, M.; Vila, J.; Daura, X.; Lertmemongkolchai, G.; Titball, R.; Colombo, G.; Bolognesi, M., Flagellar subunits as targets for structure-based epitope discovery approaches and melioidosis vaccine development. *Febs Journal* **2015**, *282*, 338-338.
- (29) Gourlay, L. J.; Thomas, R. J.; Peri, C.; Conchillo-Sole, O.; Ferrer-Navarro, M.; Nithichanon, A.; Vila, J.; Daura, X.; Lertmemongkolchai, G.; Titball, R.; Colombo, G.; Bolognesi, M., From crystal structure to in silico epitope discovery in the Burkholderia pseudomallei flagellar hook-associated protein FlgK. *Febs Journal* **2015**, *282* (7), 1319-1333.
- (30) Nithichanon, A.; Rinchai, D.; Gori, A.; Lassaux, P.; Peri, C.; Conchillo-Sole, O.; Ferrer-Navarro, M.; Gourlay, L. J.; Nardini, M.; Vila, J.; Daura, X.; Colombo, G.; Bolognesi, M.; Lertmemongkolchai, G., Sequence- and Structure-Based Immunoreactive Epitope Discovery for Burkholderia pseudomallei Flagellin. *Plos Neglected Tropical Diseases* **2015**, *9* (7).
- (31) Gori, A.; Peri, C.; Quilici, G.; Nithichanon, A.; Gaudesi, D.; Longhi, R.; Gourlay, L.; Bolognesi, M.; Lertmemongkolchai, G.; Musco, G.; Colombo, G., Flexible vs Rigid Epitope Conformations for Diagnostic- and Vaccine-Oriented Applications: Novel Insights from the Burkholderia pseudomallei BPSL2765 Pa13 Epitope. *Acs Infectious Diseases* **2016**, *2* (3), 221-230.
- (32) Gori, A.; Sola, L.; Gagni, P.; Bruni, G.; Liprino, M.; Peri, C.; Colombo, G.; Cretich, M.; Chiari, M., Screening Complex Biological Samples with Peptide Microarrays: The Favorable Impact of Probe Orientation via Chemoselective Immobilization Strategies on Clickable Polymeric Coatings. *Bioconjugate Chemistry* **2016**, *27* (11), 2669-2677.
- (33) Sola, L.; Gagni, P.; D'Annessa, I.; Capelli, R.; Bertino, C.; Romanato, A.; Damin, F.; Bergamaschi, G.; Marchisio, E.; Cuzzocrea, A.; Bombaci, M.; Grifantini, R.; Chiari, M.; Colombo, G.; Gori, A.; Cretich, M., Enhancing Antibody Serodiagnosis Using a Controlled Peptide Coimmobilization Strategy. *ACS Infectious Diseases* **2018**, *4* (6), 998-1006.
- (34) Bergamaschi, G.; Fassi, E. M. A.; Romanato, A.; D'Annessa, I.; Odinolfi, M. T.; Brambilla, D.; Damin, F.; Chiari, M.; Gori, A.; Colombo, G.; Cretich, M., Computational Analysis of Dengue Virus Envelope Protein (E) Reveals an Epitope with Flavivirus Immunodiagnostic Potential in Peptide Microarrays. *International journal of molecular sciences* **2019**, *20* (8), 1921.
- (35) Genheden, S.; Ryde, U., The MM/PBSA and MM/GBSA methods to estimate ligand-binding affinities. *Expert Opin Drug Discov.* **2015**, *10* (5), 449-461.
- (36) Yuan, M.; Liu, H.; Wu, N. C.; Lee, C.-C. D.; Zhu, X.; Zhao, F.; Huang, D.; Yu, W.; Hua, Y.; Tien, H.; Rogers, T. F.; Landais, E.; Sok, D.; Jardine, J. G.; Burton, D. R.; Wilson, I. A., Structural basis of a shared antibody response to SARS-CoV-2. *Science* **2020**, eabd2321.
- (37) Chi, X.; Yan, R.; Zhang, J.; Zhang, G.; Zhang, Y.; Hao, M.; Zhang, Z.; Fan, P.; Dong, Y.; Yang, Y.; Chen, Z.; Guo, Y.; Zhang, J.; Li, Y.; Song, X.; Chen, Y.; Xia, L.; Fu, L.; Hou, L.; Xu, J.; Yu, C.; Li, J.; Zhou, Q.; Chen, W., A neutralizing human antibody binds to the N-terminal domain of the Spike protein of SARS-CoV-2. *Science* **2020**, eabc6952.
- (38) Barnes, C. O.; West, A. P., Jr.; Huey-Tubman, K. E.; Hoffmann, M. A. G.; Sharaf, N. G.; Hoffman, P. R.; Koranda, N.; Gristick, H. B.; Gaebler, C.; Muecksch, F.; Lorenzi, J. C. C.; Finkin, S.; Hägglöf, T.; Hurley, A.; Millard, K. G.; Weisblum, Y.; Schmidt, F.; Hatzioannou, T.; Bieniasz, P. D.; Caskey, M.; Robbiani, D. F.; Nussenzweig, M. C.; Bjorkman, P. J., Structures of Human Antibodies Bound to SARS-CoV-2 Spike Reveal Common Epitopes and Recurrent Features of Antibodies. *Cell* **2020**, S0092-8674(20)30757-1.
- (39) Zhou, H.; Chen, Y.; Zhang, S.; Niu, P.; Qin, K.; Jia, W.; Huang, B.; Zhang, S.; Lan, J.; Zhang, L.; Tan, W.; Wang, X., Structural definition of a neutralization epitope on the N-terminal domain of MERS-CoV spike glycoprotein. *Nature Communications* **2019**, *10* (1), 3068.
- (40) Zheng, Z.; Monteil, V. M.; Maurer-Stroh, S.; Yew, C. W.; Leong, C.; Mohd-Ismail, N. K.; Arularasu, S. C.; Chow, V. T. K.; Pin, R. L. T.; Mirazimi, A.; Hong, W.; Tan, Y.-J., Monoclonal antibodies for the S2 subunit of spike of SARS-CoV cross-react with the newly-emerged SARS-CoV-2. *bioRxiv* **2020**, 2020.03.06.980037.
- (41) Acharya, P.; Williams, W.; Henderson, R.; Janowska, K.; Manne, K.; Parks, R.; Deyton, M.; Spreng, J.; Stalls, V.; Kopp, M.; Mansouri, K.; Edwards, R. J.; Meyerhoff, R. R.; Oguin, T.; Sempowski, G.; Saunders, K.

- Haynes, B. F., A glycan cluster on the SARS-CoV-2 spike ectodomain is recognized by Fab-dimerized glycan-reactive antibodies. *bioRxiv : the preprint server for biology* **2020**, 2020.06.30.178897.
- (42) Tang, T.; Bidon, M.; Jaimes, J. A.; Whittaker, G. R.; Daniel, S., Coronavirus membrane fusion mechanism offers a potential target for antiviral development. *Antiviral Research* **2020**, *178*, 104792.
- (43) Pinto, D.; Park, Y.-J.; Beltramello, M.; Walls, A. C.; Tortorici, M. A.; Bianchi, S.; Jaconi, S.; Culap, K.; Zatta, F.; De Marco, A.; Peter, A.; Guarino, B.; Spreafico, R.; Cameroni, E.; Case, J. B.; Chen, R. E.; Havenar-Daughton, C.; Snell, G.; Telenti, A.; Virgin, H. W.; Lanzavecchia, A.; Diamond, M. S.; Fink, K.; Veessler, D.; Corti, D., Cross-neutralization of SARS-CoV-2 by a human monoclonal SARS-CoV antibody. *Nature* **2020**, *583* (7815), 290-295.
- (44) Cao, Y.; Su, B.; Guo, X.; Sun, W.; Deng, Y.; Bao, L.; Zhu, Q.; Zhang, X.; Zheng, Y.; Geng, C.; Chai, X.; He, R.; Li, X.; Lv, Q.; Zhu, H.; Deng, W.; Xu, Y.; Wang, Y.; Qiao, L.; Tan, Y.; Song, L.; Wang, G.; Du, X.; Gao, N.; Liu, J.; Xiao, J.; Su, X.-d.; Du, Z.; Feng, Y.; Qin, C.; Qin, C.; Jin, R.; Xie, X. S., Potent Neutralizing Antibodies against SARS-CoV-2 Identified by High-Throughput Single-Cell Sequencing of Convalescent Patients' B Cells. *Cell* **2020**, *182* (1), 73-84.e16.
- (45) Kirschner, K. N.; Yongye, A. B.; Tschampel, S. M.; González-Outeiriño, J.; Daniels, C. R.; Foley, B. L.; Woods, R. J., GLYCAM06: A generalizable biomolecular force field. *Carbohydrates. Journal of Computational Chemistry* **2008**, *29* (4), 622-655.
- (46) Henderson, R.; Edwards, R. J.; Mansouri, K.; Janowska, K.; Stalls, V.; Kopp, M.; Haynes, B. F.; Acharya, P., Glycans on the SARS-CoV-2 Spike Control the Receptor Binding Domain Conformation. *bioRxiv* **2020**, 2020.06.26.173765.
- (47) Maier, J. A.; Martinez, C.; Kasavajhala, K.; Wickstrom, L.; Hauser, K. E.; Simmerling, C., ff14SB: Improving the Accuracy of Protein Side Chain and Backbone Parameters from ff99SB. *Journal of Chemical Theory and Computation* **2015**, *11* (8), 3696-3713.
- (48) Case, D. A.; Cheatham Iii, T. E.; Darden, T.; Gohlke, H.; Luo, R.; Merz Jr, K. M.; Onufriev, A.; Simmerling, C.; Wang, B.; Woods, R. J., The Amber biomolecular simulation programs. *Journal of Computational Chemistry* **2005**, *26* (16), 1668-1688.
- (49) Defays, D., An efficient algorithm for a complete link method. *The Computer Journal* **1977**, *20* (4), 364-366.
- (50) Onufriev, A.; Bashford, D.; Case, D. A., Modification of the Generalized Born Model Suitable for Macromolecules. *J. Phys. Chem. B* **2000**, *104*, 3712-3720.
- (51) Weiser, J.; Shenkin, P. S.; Still, W. C., Approximate Atomic Surfaces from Linear Combinations of Pairwise Overlaps (LCPO). *J. Comput. Chem.* **1999**, *20*, 217-230.
- (52) Tiana, G.; Simona, F.; De Mori, G. M. S.; Broglia, R. A.; Colombo, G., Understanding the determinants of stability and folding of small globular proteins from their energetics. *Protein Science* **2004**, *13* (1), 113-124.
- (53) Morra, G.; Colombo, G., Relationship between energy distribution and fold stability: Insights from molecular dynamics simulations of native and mutant proteins. *Proteins: Struct. Funct. and Bioinf.* **2008**, *72* (2), 660-672.

1.13 Immunoreactivity of the VOCs SARS-CoV-2 spike protein

“SARS-CoV-2 Spike Protein Mutations and Escape from Antibodies: a Computational Model of Epitope Loss in Variants of Concern”



1.13.1 Abstract

The SARS-CoV-2 spike (S) protein is prominently displayed on the viral surface, serving as the initial point of interaction between the virus and the host. Consequently, it is the primary target for COVID-19 vaccines. Over recent years, various versions of this protein have emerged as variants. These variants, with their ability to diminish or evade recognition by antibodies targeting the S protein, present a notable challenge to immunological treatments, raising serious concerns about their potential impact on vaccine effectiveness.

To develop a model able to predict the potential impact of S-protein mutations on antibody binding sites, we performed unbiased multi-microsecond molecular dynamics of several glycosylated S-protein variants and applied a straightforward structure-dynamics-energy based strategy to predict potential alterations in immunogenic regions for each variant. Remarkably, we successfully identified known epitopes on the reference D614G sequence.

By comparing our results, derived from isolated S-proteins in solution, with existing data on antibody binding and reactivity in the latest S variants, we demonstrated a consistent pattern: modifications in the S-protein correlated with the loss of potentially immunoreactive

regions. This finding directly aligns with the experimentally observed decreased ability of certain antibodies elicited against the dominant S-sequence to recognize these variants.

While our study centered on analyzing SARS-CoV-2 Spike variants, the computational epitope-prediction strategy we employed is transferable. It can be extended to study immunoreactivity in mutants of other proteins with characterized structures. This approach can significantly contribute to the development/selection of vaccines and antibodies adept at addressing emerging variants, ensuring a proactive response to evolving viral threats.

1.13.2 Introduction

In the general introduction has been mentioned that protein sequences evolve as a result of selective pressure to optimize function, create improved phenotypes, and introduce new advantageous traits. In pathogens like bacteria and viruses, sequences evolve via modifications such as point mutations, recombination and deletions/insertions to induce higher infectivity, more efficient replication, and ultimately escape from the host immune systems¹⁻⁷ and SARS-CoV-2 is no exception to these general rules. The spread of the virus to more than 200 million people worldwide, combined with the pressure determined by the reactions of immunocompetent populations, led to the emergence of “variants of concern”. In this context, attention has been focused on the SARS-CoV-2 spike protein (S), the large, heavily glycosylated class I trimeric fusion protein which mediates host cell recognition, binding and entry. Because it represents the first point of contact with the host, and given its crucial role in viral pathogenesis, the S protein has been the basis for the design of currently used vaccines effective at reducing viral spread, hospitalization and mortality rates.¹¹⁻¹⁶

While for almost one year the only notable mutation in S has been the D614G (Asp⁶¹⁴→Gly), which increases affinity for the cell receptor ACE2 and has immediately become dominant, novel S protein variants reported of late posed new potential challenges for efficacy of vaccination, antibody-based therapies and viral diffusion control. Three notable examples of such evolved S proteins are B.1.1.7 (the so-called UK or Alpha variant), 501Y.V2/B.1.351 (the South African or Beta variant), and B.1.1.28 (P.1, the Brazilian or Gamma variant). All such sequences contain various mutations due to nonsynonymous nucleotide changes in the RBD domain, including E484K, N501Y, and/or K417N.¹⁰ In B.1.1.7 and B.1.351, deletions are also present in the N-terminal Domains (**Figure 1**).

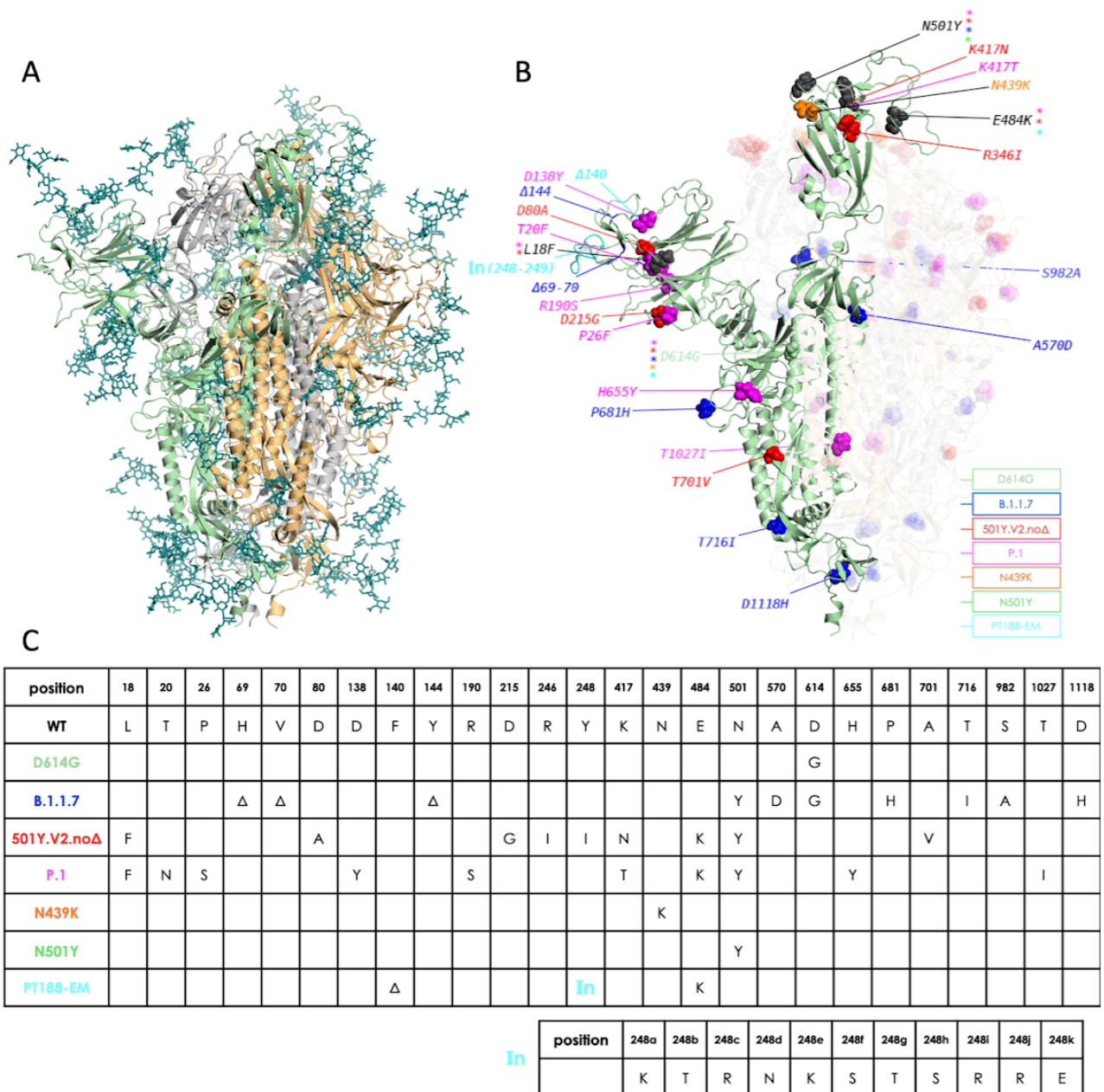


Figure 1. Overview of simulated variants (definitions in main text). **A.** The full-length, fully glycosylated trimeric structure corresponding to pdb code 6VSB. Protomer A (RBD “up”): secondary structure in green; protomers B and C (RBD “down”): grey and sand, respectively. Glycans’ C, N, and O atoms rendered as teal sticks. **B.** Positions and nature of mutations highlighted on protomer A of different variants. Mutant residues’ heavy atoms are rendered as spheres; a different color is assigned to each variant, as indicated in the legend. Mutations common to more than one variant are rendered and/or labeled in black, with colored asterisks denoting variants carrying the mutation. The insertion in the PT188-EM variant (cyan) is denoted by “In(248-249)”. Protomers B and C are also shown with their respective mutations but rendered with increased transparency for clarity; glycans are omitted; **C.** Synopsis of mutations on the different variants simulated in this work, including the 11-residue insertion in the PT188-EM variant.

Several studies showed how some of these circulating variants have reduced sensitivity to neutralizing antibodies targeting the RBD or to the NTD.^{10, 17-19} In this context, polyclonal antibodies contained in Convalescent Plasma (CP) from individuals infected with the D614G-containing SARS-CoV-2, showed reduced potency in neutralizing 501Y.V2/B.1.351 virus isolates.^{20, 21} Furthermore, antibodies elicited after vaccine treatment showed reduced neutralization of pseudoviruses bearing the mutations of the P.1 and 501Y.V2/B.1.351 variants.²² The same was observed for pseudoviruses with variations in S mimicking those of the B.1.1.7 lineage.^{22, 23} Yet, fortunately, it was shown that vaccine-generated antibody titers were sufficient to neutralize B.1.1.7 in sera from 40 BNT162b2-vaccinated individuals.²⁴

A crucial question for understanding the impact of S-protein evolution on the development of monoclonal antibody (mAb)-based and vaccine-based therapies, is whether we can develop a simple model to rationalize, and eventually predict, the effect of variations on the structural properties of S that ultimately underpin antibody recognition. Fundamentally, comparison across S-proteins mutants can help us understand the molecular basis of the protein's evolvability, furthering our grasp of the relationships between sequence, structure and (immuno)recognition. From the practical point of view, this knowledge could in principle be harnessed to design and engineer improved S-based antigens or multicomponent domain/peptide combinations, focusing for instance on those antibody binding regions, known as epitopes, that are predicted to be conserved in multiple variants.

Here, we apply the straightforward structure-dynamics-energy strategy to predict potentially immunogenic regions in representative 3D conformations of several variants of the full-length glycosylated trimeric S protein (**Figure 1**).

The selected S proteins represent some of the major variants of concern circulating at the time of setting up simulation. In this respect, the African variant we simulate, which is named 501Y.V2.no Δ , corresponds to the S lineage originally discovered in South Africa in late November 2020 by Tegally *et al.*²⁹. This S variant features the additional mutations L18F (in common with P.1) and R246I but does not feature the Δ 241-243 deletion, whose existence was still debated when the authors released their study in January 2021.²⁹ This variant has subsequently been referred to in several papers as B.1.351. The list of studied proteins is further enriched by a laboratory-evolved escape S-variant, obtained by Rappuoli and coworkers by co-incubating the SARS-CoV-2 virus with a highly neutralizing plasma from a COVID-19 convalescent patient. Interestingly, after several passages this strategy

generated a variant completely resistant to plasma neutralization. This “artificial” variant is labeled here as the PT188-EM variant.^{21, 22}

Conformations are extracted from independent atomistic molecular dynamics (MD) simulations totaling 4 μ s for each mutant. Our approach to the detection of epitopes on S, i.e., its antibody-binding protein regions, is based on the same concept as before: such sites should continuously evolve to escape immune recognition by the host without impairing the native protein structure required for viral function and survival. We previously showed (and experimentally confirmed) that these regions coincide with substructures that are not involved in major stabilizing *intramolecular* interactions with core protein residues that are important for its folding into a functional 3D structure.³⁰ In other words, Ab-interacting regions show minimal energetic coupling with the rest of the protein, which in turn should favor accumulation of escape mutations while preserving the antigen’s 3D structure. Furthermore, minimal intramolecular coupling provides epitopes with greater conformational freedom to adapt to and be recognized by a binding partner. Actual binding to an external partner such as an Ab is expected to occur if favorable intermolecular interactions determine a lower free energy for the bound state than for the unbound state.³⁰⁻³³

These concepts are analyzable by the MLCE approach (see Methods and the paper before). Starting from the characterization of the energy of pairwise interactions between all aminoacids and monosaccharides and filtering the resulting interaction map with structural information extracted from the same protein’s inter-residue contact map, MLCE identifies groups of spatially contiguous residues with poor energetic coupling to the rest of the protein as potential immunogenic regions. At the same time, groups of residues with high energetic coupling are identified as stabilization centers.

Upon comparing our results to recently reported characterization of Ab binding and reactivity, the analysis we report consistently shows that mutations, deletions, and/or insertions in S variants determine a reorganization of internal interactions leading to the loss of potentially immunoreactive regions on the surface. Encouragingly, these findings can be qualitatively reconnected to the decreased ability of some of the Abs elicited against the dominant S-sequence to recognize variants.

1.13.3 Results

To characterize the effects of mutations, deletions, and insertions on the definition of potential Ab-binding substructures in S variants, we apply, as before, a combination of the Energy Decomposition (ED) and MLCE (Matrix of Low Coupling Energies) methods to representative structures extracted from long timescale MD simulations of the S protein variants reported in **Figure 1**.

Briefly, we first run 4 independent 1 μ s long all-atom MD simulations of each variant of the full-length fully glycosylated S protein in solution (**Figure 1**) (each built from PDB ID: 6VSB¹¹). Next, for each variant, we concatenate individual trajectories into one a single 4 μ s metatrayjectory. Cluster analysis on each variant's metatrayjectory is then conducted to identify the 3 most representative conformations. These are then used to compute nonbonded pairwise potential energy terms (van der Waals, electrostatic interactions, solvent effects) obtaining, for a given variant with N aminoacid and monosaccharide residues, a symmetric $N \times N$ inter-residue interaction matrix. The three matrices extracted from a variant's trajectory are then weighted and averaged to yield an average nonbonded interaction matrix, M_{ij} . Upon eigenvalue decomposition of M_{ij} , eigenvectors associated with the most negative eigenvalues can help build a simplified version of M_{ij} that only highlights series of residues with high- and low-intensity couplings. The former represent residues acting as folding hotspots and responsible stabilizing the protein's 3D structure; the latter represent residue pairs with weak energetic coupling to the rest of the protein, whose mutation is expected not to impact S' structure and thus function. In this framework, once information contained in the simplified energy map is combined with information contained in the protein's residue-residue contact map, it permits to 'filter out' clusters of residues whose energetic coupling to the rest of the structure is weak *and* that are spatially contiguous. Such localized networks of low-intensity couplings, located in proximity of the protein surface represent potential interaction Ab-interaction regions, or epitopes. The reference S structure we use here is the dominant D614G variant. We analyze the results of epitope predictions we obtain on *isolated* S variants by comparing them against selected Spike-antibody complexes. To this end, we collected publicly available X-ray or Cryo-EM structural data of complexes between S and various Abs, reported in **Table 1 and Table S1**. Epitopes in experimental structures are defined as the sets of S protein residues within 5Å of any Ab residue. The experimental epitopes thus derived are used as the reference against which to compare epitopes predicted *in silico*.

| Antibody | PDB ID | D614G | B.1.1.7 | 501Y.V2.noΔ | P.1 | N439K | N501Y | PT188-EM |
|-----------|-----------|---|---|---|---|---|---|----------|
| REGN10987 | 6XDG | 10.1126/science.abd0827 | https://doi.org/10.1101/2021.02.18.431897 | https://doi.org/10.1101/2021.02.18.431897 501Y.V2.Δ | https://doi.org/10.1101/2021.03.01.433466 | https://doi.org/10.1101/2021.01.037 | https://doi.org/10.1002/1873-3468.14076 | |
| REGN10933 | 6XDG | 10.1126/science.abd0827 | https://doi.org/10.1101/2021.02.18.431897 | https://doi.org/10.1101/2021.02.18.431897 501Y.V2.Δ | https://doi.org/10.1101/2021.03.01.433466 | https://doi.org/10.1101/2021.01.037 | https://doi.org/10.1002/1873-3468.14076 | |
| LY-CoV555 | 7L3N | 10.1101/2020.09.30.318972 | https://doi.org/10.1038/s41586-021-03398-2 | https://doi.org/10.1038/s41586-021-03398-2 501Y.V2.Δ | | https://doi.org/10.1101/2021.01.037 | https://doi.org/10.1002/1873-3468.14076 | |
| S309 | 7JX3/6WPT | 10.1038/s41586-020-2349-v | https://doi.org/10.1038/s41586-021-03398-2 | https://doi.org/10.1038/s41586-021-03398-2 501Y.V2.Δ | https://doi.org/10.1101/2021.03.01.433466 | https://doi.org/10.1101/2021.01.037 | https://doi.org/10.1002/1873-3468.14076 | |
| C135 | 7K8Z | 10.1038/s41586-020-2852-1 | https://doi.org/10.1038/s41586-021-03398-2 | https://doi.org/10.1038/s41586-021-03398-2 501Y.V2.Δ | | https://doi.org/10.1101/2020.07.21.214759 | | |
| C144 | 7K90 | 10.1038/s41586-020-2852-1 | | https://doi.org/10.1101/2021.01.27.428478 501Y.V2.Δ | https://doi.org/10.1101/2020.07.21.214759 | https://doi.org/10.1101/2020.07.21.214759 | | |
| C121 | 7K90 | 10.1038/s41586-020-2852-1 | | https://doi.org/10.1101/2021.01.27.428478 501Y.V2.Δ | https://doi.org/10.1101/2021.03.01.433466 | https://doi.org/10.1101/2020.07.21.214759 | | |
| 4A8 | 7C2L | 10.1126/science.abc6952 | https://doi.org/10.1038/s41586-021-03398-2 | https://doi.org/10.1038/s41586-021-03398-2 501Y.V2.Δ | https://doi.org/10.1101/2021.06.003 | | https://doi.org/10.1101/2020.10.07.328302 | |
| DH1041 | 7LAA | 10.1101/2020.12.31.424729 | 10.1101/2020.12.31.424729 | https://doi.org/10.1101/2021.03.11.435037 | https://doi.org/10.1101/2021.03.002 | | | |
| DH1043 | 7LJR | 10.1101/2020.12.31.424729 | 10.1101/2020.12.31.424729 | https://doi.org/10.1101/2021.03.11.435037 | https://doi.org/10.1101/2021.03.002 | | | |
| DH1047 | 7LD1 | 10.1101/2020.12.31.424729 | 10.1101/2020.12.31.424729 | https://doi.org/10.1101/2021.03.11.435037 | https://doi.org/10.1101/2021.03.002 | | | |
| DH1050.1 | 7LCN | 10.1101/2020.12.31.424729 | 10.1101/2020.12.31.424729 | https://doi.org/10.1101/2021.03.11.435037 | https://doi.org/10.1101/2021.03.002 | | | |
| S2M11 | 7K43 | 10.1126/science.abc3354 | | https://www.citid.ca/m.ac.uk/wp-content/uploads/2021/02/POST-SUBMISSION_vaccine-DCv2-2.pdf | https://www.citid.ca/m.ac.uk/wp-content/uploads/2021/02/POST-SUBMISSION_vaccine-DCv2-2.pdf | https://www.citid.ca/m.ac.uk/wp-content/uploads/2021/02/POST-SUBMISSION_vaccine-DCv2-2.pdf | | |
| COVA1-16 | 7JMX | 10.1101/2020.08.02.233536 | https://doi.org/10.1038/s41586-021-03398-2 | https://doi.org/10.1038/s41586-021-03398-2 501Y.V2.Δ | https://doi.org/10.1101/2021.05.26.21257441 | | https://doi.org/10.1038/s41586-021-03398-2 | |
| B38 | 7BZ5 | 10.1126/science.abc2241 | 10.1016/j.chom.2021.03.002 | | | | https://doi.org/10.1002/1873-3468.14076 | |
| C002 | 7K8T | 10.1038/s41586-020-2852-1 | | | | | | |

| | | | | | | | | |
|-----|------|---|---|--|---|--|---|--|
| CB6 | 7C01 | 10.1038/s41586-020-2381-y | https://doi.org/10.1038/s41586-021-03398-2 | https://doi.org/10.1016/j.xcrn.2021.100255 501Y.V2.Δ | https://doi.org/10.1101/2021.03.01.433466 | | https://doi.org/10.1002/1873-3468.14076 | |
|-----|------|---|---|--|---|--|---|--|

Table 1. PDB IDs of the S-Ab complexes used to compare epitope predictions. For each Ab considered in this work (leftmost column), we report: PDB IDs of S-Ab Cryo-EM complexes used as experimental reference for our MLCE epitope predictions; and, where available, experimental studies reporting either that Ab's gain (yellow) or loss/absence of activity (blue) towards a particular variant. White cells indicate that experimental data is unavailable. * denotes experimental studies carried out on the 501Y.V2.noΔ S variant but with the Δ241-243 deletion.

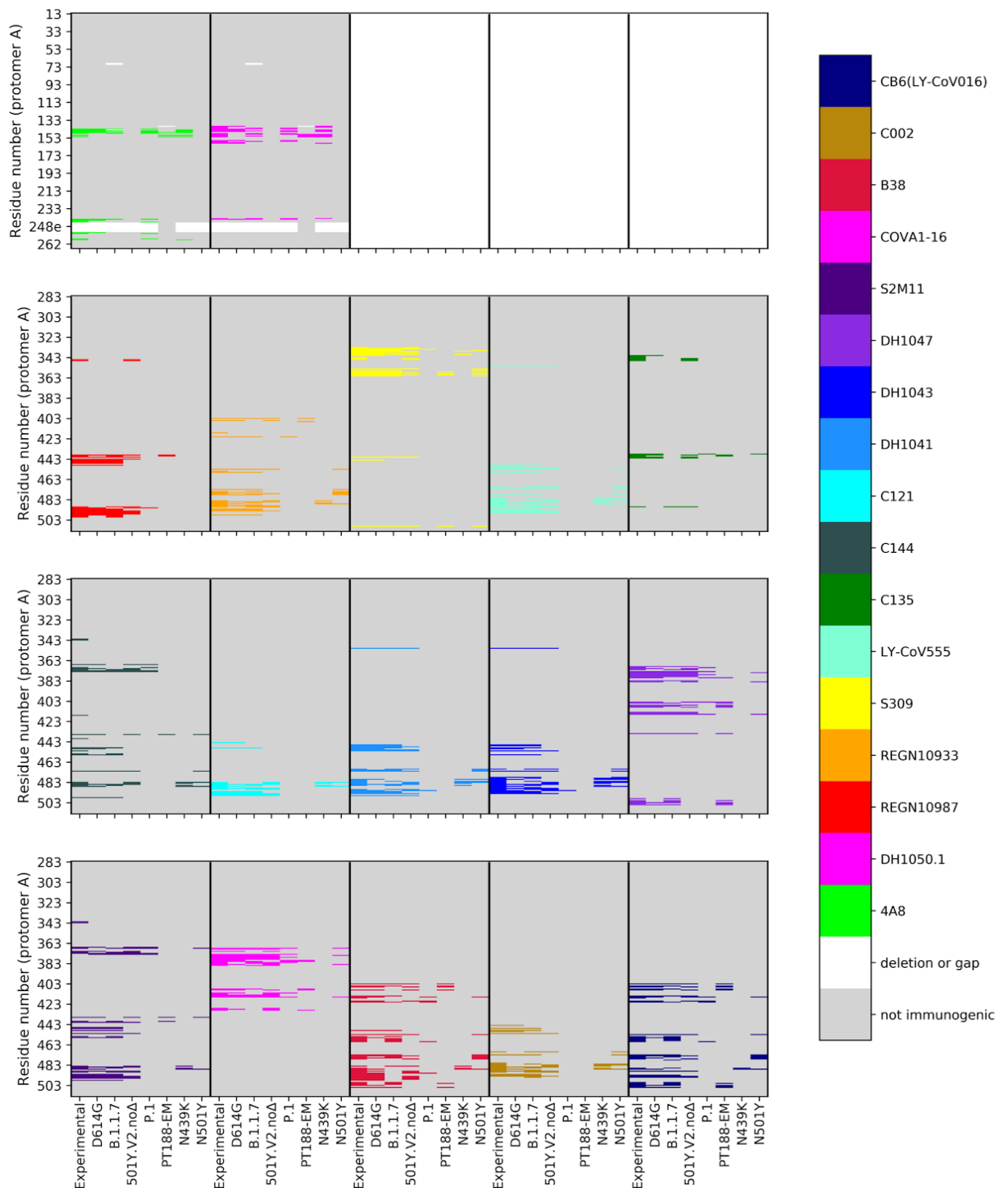


Figure 2. Mapping epitopes on each variant. Epitope mapping on S protomer A for the 2 NTD-targeting antibodies (two top left panels; *cf.* numbering on y-axis) and the 15 RBD-targeting antibodies (bottom three rows) considered in this study. In each panel, using a distinct color for each antibody (right palette), the experimentally (Cryo-EM or X-ray) detected residues that belong to an epitope (labeled “Experimental” on each panel’s x-axis) are compared to epitopes predicted *in silico* on each of the seven variants considered. Predicted immunogenic residues are colored according

to the Ab they would be targeted by. Non-immunogenic residues are shown in gray; gaps/insertions in white. The figure shows how the extent of the epitopes for the different antibodies varies in the distinct mutants.

In the remaining variants of concern, a diverse landscape of epitopes emerges. Several residues/regions that are predicted immunogenic in the reference S-protein disappear in the variants. Overall, this is observed for all the Abs considered. In this framework, after running an epitope prediction on each variant we monitor epitope conservation across variants through a *conservation ratio*: the number of residues in each predicted epitope for a given variant is divided by the number of residues in the corresponding experimental epitope in the reference S structure, which is defined based on the 5 Å threshold from its respective Ab, as discussed above. We define *epitope loss* when the conservation ratio is lower than 0.5; otherwise, the epitope is considered to be conserved. In **Table 2** and **Figure 3** we report such conservation ratios for each D614G S epitope on each simulated variant and confront them with available experimental data (at the time of writing) on the variant's reactivity towards the Ab that would be expected to bind to that particular epitope. Each cell in the table is color-coded according to the experimentally measured activity of the corresponding Ab on one of the given variants. If the Ab remains active, the cell is yellow. If the Ab has lost activity against that variant, the cell is blue. If experimental data is unavailable for a particular Ab on a particular variant, the cell is white.

| Antibodies | D614G | B.1.1.7 | 501Y.V 2.noΔ | P.1 | N439K | N501Y | PT188- EM |
|---------------------|-------|---------|-----------------|------|-------|-------|--------------|
| REGN10987 | 0.76 | 0.86 | 0.52 | 0.85 | 0.00 | 0.00 | 0.1 |
| REGN10933 | 0.65 | 0.78 | 0.39 | 0.04 | 0.13 | 0.30 | 0.09 |
| LY-CoV555 | 0.67 | 0.83 | 0.84 | 0.00 | 0.21 | 0.29 | 0.00 |
| S309 | 0.77 | 0.77 | 0.59 | 0.05 | 0.09 | 0.32 | 0.18 |
| C135 | 0.42 | 0.50 | 0.58 | 0.08 | 0.00 | 0.08 | 0.17 |
| C144 | 0.72 | 0.31 | 0.34 | 0.14 | 0.14 | 0.14 | 0.03 |
| C121 | 0.77 | 0.85 | 0.62 | 0.00 | 0.31 | 0.15 | 0.00 |
| 4A8 | 0.50 | 0.31 | 0.00 | 0.56 | 0.44 | 0.00 | 0.19 |
| DH1041 | 0.68 | 0.59 | 0.59 | 0.05 | 0.23 | 0.27 | 0.00 |
| DH1043 | 0.62 | 0.81 | 0.42 | 0.04 | 0.27 | 0.31 | 0.00 |
| DH1047 | 0.79 | 0.71 | 0.64 | 0.21 | 0.00 | 0.11 | 0.39 |
| DH1050.1 | 0.59 | 0.53 | 0.00 | 0.59 | 0.65 | 0.00 | 0.24 |
| S2M11 | 0.69 | 0.76 | 0.55 | 0.14 | 0.10 | 0.10 | 0.07 |
| COVA1-16 | 0.76 | 0.60 | 0.88 | 0.22 | 0.00 | 0.16 | 0.20 |
| B38 | 0.63 | 0.73 | 0.41 | 0.12 | 0.07 | 0.20 | 0.17 |
| C002 | 0.60 | 0.70 | 0.40 | 0.00 | 0.25 | 0.25 | 0.00 |
| CB6 (LY- CoV016) | 0.62 | 0.74 | 0.35 | 0.12 | 0.06 | 0.24 | 0.26 |

Table 2. Epitope predictions on each variant and epitope conservation ratio. Each cell reports an *epitope conservation ratio* for each S variant-Ab combination, relating *in silico* predictions to experimental epitopes from experimental Cryo-EM and/or crystal structures. Conservation ratios lower than 0.5 indicate epitope loss; otherwise, an epitope is considered to be conserved. Each cell in the table is color-coded according to the experimentally measured activity of the corresponding Ab on the respective variant. If the Ab remains active, the cell is yellow. If the Ab has lost activity against that variant, the cell is blue. If experimental data is unavailable for a particular Ab on a particular variant, the cell is white. Disagreement between predictions and experiment (i.e., blue and conservation ratio >0.5 or yellow and conservation ratio <0.5) is indicated by thick borders and dotted-line diagonal.

Analysis of **Table 2** clearly shows that the vast majority of blue cells, indicative of a loss of Ab reactivity, contain ratios lower than 0.5. This is an important validation of our prediction: whenever a variant's predicted epitope residues- *i.e.*, according to MLCE, contiguous residues uncoupled from the S protein core- shrink in number compared to D614G S, it is very likely that experimental data will also confirm that that variant evades Abs binding to the shrunk or lost epitopes. On the other hand, most cases for which Abs retain activity against a variant (yellow cells) are also confirmed by our prediction to retain their respective epitopes (conservation ratio > 0.5) with respect to D614G S. Disagreement between our

predictions and experiment only occurs in a minority of cases: corresponding cells are marked by thicker borders.

Analysis of Alpha and 501Y.V2.no Δ (South Africa; late November 2020) immediately shows that a large portion of predicted epitopes in the RBD are conserved compared to the reference D614G. Interestingly, however, we also observe a dramatic drop in the number of NTD residues predicted as epitopes for the 501Y.V2.no Δ . Epitope loss in the NTD, which was deemed to host a super-antigenic hotspot can help explain the ability for immune evasiveness observed for these two variants. In Alpha, the NTD epitope is largely conserved consistent with the conservation of activity of Abs targeting this region against the variant (**Table 2, Figure 2 and 3**).

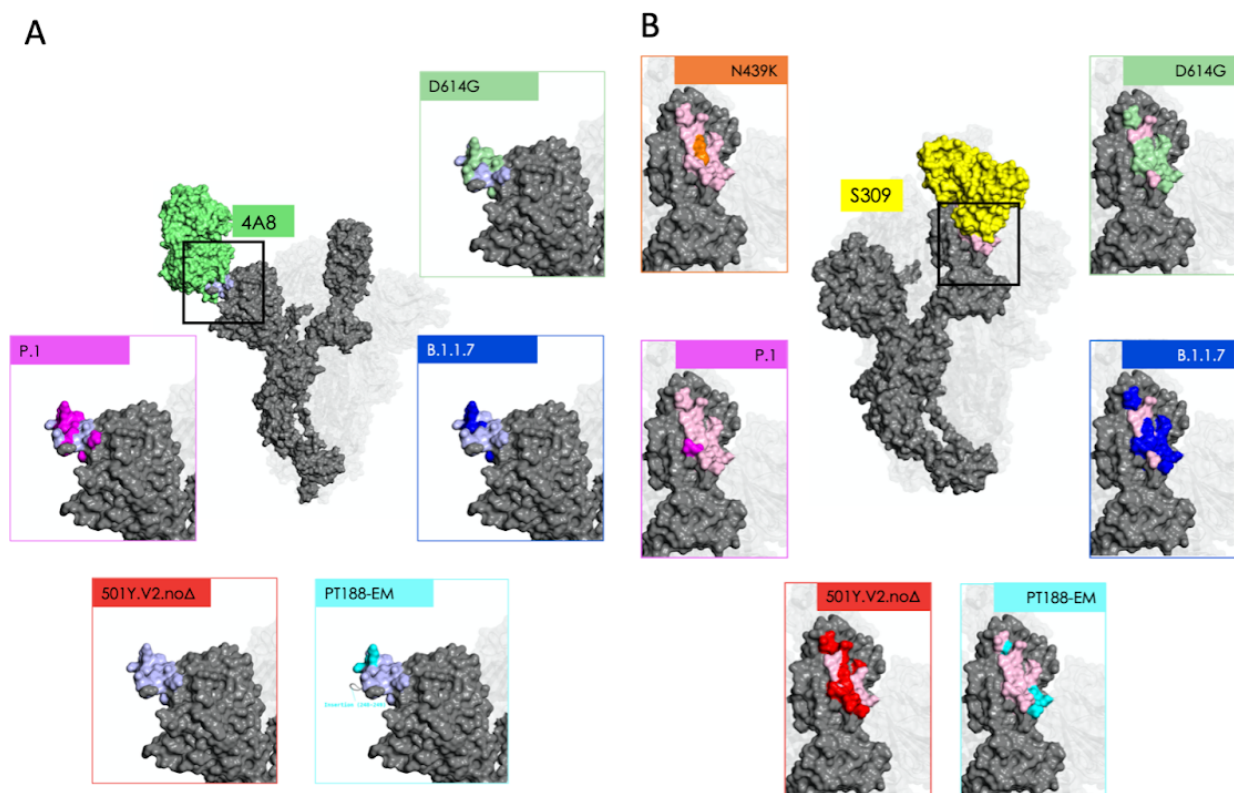


Figure 3. Mutations modify epitope identity. Central images in panels A and B depict the Cryo-EM structure of the antigen-binding fragments of two representative Abs bound to protomer A: 4A8 (panel A; Ab in green; experimental epitope in light blue); and S309 (panel B; Ab in yellow; experimental epitopes in light pink). Insets in each panel contrast the extent of the experimental epitope with epitopic residues predicted by the MLCE method (see main text) for five (panel A) or six (panel B) of the variants considered in this work: these residues are rendered using the same color code used for variants in **Figure 1**; residues in the experimental epitope not predicted by MLCE are rendered as in the central image (panel A: light blue; panel B: light pink). Other residues on the

S protein (not comprised in the experimental epitope) are rendered in gray. Glycans are omitted for clarity; positions of protomers B and C are shown for reference.

Importantly, conservation of a dominant part of the epitopes in the RBD still endows the two variants with reactivity against Abs directed to this domain, which may help explain the observed effectiveness of some convalescent plasma treatments and vaccines.^{13, 45}

Calculations on the Brazilian variant correctly indicate loss of immunoreactivity of several Abs as well as conserved reactivity of Abs 4A8 and S2M11. This variant is the only one for which our predictions of epitopes binding Abs of the DH family generally disagree with experimental data.

Finally, it is important to note that the “artificial” PT188-EM variant, evolved in the lab under the pressure of convalescent serum to evade Ab-effects, appears to have lost a very large number of protein epitopes (see **Table 2, Figure 2 and 3**). In particular, the insertion at residues 248 modifies the conformational properties of the region otherwise recognized by Ab 4A8. Therefore, the epitope to this antibody disappears from the predictions on the PT188-EM variant.²¹ Interestingly, in this case, the carbohydrate motifs coating the protein appear to host most of the uncoupled regions (117 carbohydrate moieties in the PT188-EM variant vs. 90 in the reference S-protein), pointing to a role of the glycan shield in protecting the protein from immune recognition, besides playing a key part in modulating interactions for ACE2 recognition and cell-entry.⁴⁶⁻⁵² (see **Figure 4**).

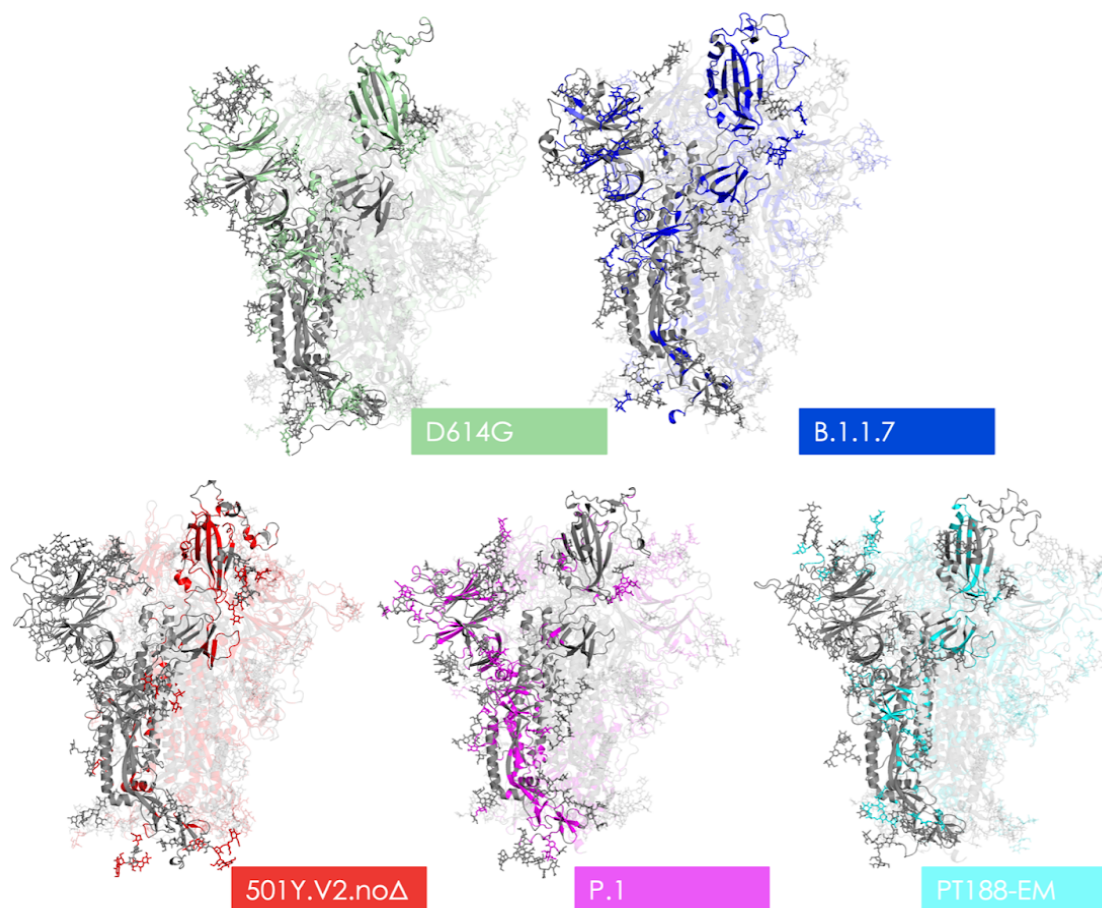


Figure 4. Structural representations of epitopes on different variants. The various structures depicted show the 3D structure of protomer A in gray. Residues rendered in the color assigned to their respective variant in **Figure 2** mark the locations of all predicted epitopes; areas in gray represent non-immunogenic regions. Glycan heavy atoms are rendered as sticks.

Importantly, mutants N439K, is correctly predicted as an escape variant from all Abs for which experimental data proved lower efficacy.

1.13.4 Discussion

In this work, we analyzed full-length models of 7 trimeric glycosylated SARS-CoV-2 S protein variants, derived from the prefusion conformation of the cryo-EM structure 6VSB, in which the Receptor Binding Domain of chain A (RBD-A) is in an “up” conformation, exposed to interaction with host cell receptors and potential targeting by Abs. The data from our energetic analyses can be aptly integrated in the characterization of the properties of S and other SARS-CoV-2 proteins from long scale simulations, such as those recently presented by Zimmerman et al.⁵³, Casalino et al.⁵⁰, Spinello et al.^{54, 55}, Oliveira et al.⁵⁶, Shoemark et al.⁵⁷, Wang et al.⁵⁸, and Fallon⁵⁹.

Our MLCE analysis of the full-length trimers correctly identifies a number of epitopes in the RBD that have been previously experimentally characterized. RBD is in fact targeted by the largest fraction of neutralizing antibodies. MLCE also identifies regions in the N-terminal domain (NTD), which are known to be targeted by different Abs, some of which potentially neutralize SARS-CoV-2¹⁹ and highlights putative immunoreactive substructures at the end of the S2 domain, where sugar-engaging Abs have recently been characterized (see **Table S2, Figure 4**).

As a caveat, it is worth pointing out the fact that antibodies can bind to conformations of the Spike protein that are different from the ones sampled here (as summarized in the General Introduction in the Chapter 3.4.2). Indeed, work by Casalino *et al.*⁵⁰, Zimmerman *et al.*⁵³, and by Fallon *et al.*⁵⁹ show that the protein can undergo dramatic structural changes. In our simulations, despite running 4 microseconds of all-atom MD simulation for each system, we could not observe such changes, if not in their initial stages. To be consistent in our comparative among the different species, we in fact decided to use the same protocol on every system and benchmarked the data obtained against available experiments. This may indeed partly limit the exploration of the conformational space available to this flexible protein, in turn somewhat limiting the prediction of immunogenic regions. We hypothesize that this is the reason behind the limited success we have with the Brazilian variant. The mutations, insertions, deletions in this sequence can expectedly favor the exploration of structures that are different from the ones we are considering here. We notice, however, that a significant number of experimental immunoreactivity data are correctly captured by our approach even on the Brazilian variant, supporting the validity of MLCE in this context.

Energy-based epitope prediction through the MLCE approach reveals a common theme across variants: the number and surface exposure of potentially immunoreactive regions decrease in S protein mutants compared to the reference D641G. In particular, the number of residues defining the epitope located in the long RBD loop (residues 417-503, recognized by many protective Abs) is much lower in mutants 501Y.V2.noΔ, B.1.1.28, and N439K (see **Figures 2, 3, Table S2**). Interestingly, in the case of B.1.1.7, which shows limited evasion, the loop is largely active in terms of immunoreactivity. In contrast, in the evading variant PT188-EM the entire loop disappears from the list of potential Ab-targets.

Potentially important contributions to the perturbation of epitopes' physico-chemical properties may be related to charge variations. Two striking examples are the loss or reduction of epitopes determined by the N439K and E484K mutations. Both cases involve

residues that are part of epitopes of a large number of antibodies and after these mutations the antibodies completely or partially lose their efficacies. In the case of the mutation N439K, it has been reported⁶⁰ that this variant maintains fitness while evading antibodies immunity. In fact, N439K RBD forms a new interaction with the human ACE2 receptor (hACE2) and has enhanced affinity for hACE2. The salt bridge at the RBD-hACE2 interface (RBD N439K:hACE2 E329) plausibly adds a strong interaction at the binding interface during viral cell entry. On the other hand, the N to K mutation determines stronger intra-Spike protein interactions which dramatically decrease the decoupling of this region from the core, making it substantially less prone to interaction with Abs.

The E484K mutation is of particular concern due to its location within nAb epitopes, and it has been shown to reduce or eliminate binding to many potent RBD-directed nAbs⁶¹.

Experimental characterization of Abs targeting the NTD revealed a site recognized by most Abs, located between the N3 and N5 loops of the domain. This epitope was correctly predicted in our previous work⁴³. Specifically, Lys147 and Arg246, known to be important in stabilizing interactions with the complementarity-determining regions of different Abs are correctly predicted as epitope elements.

On the other hand, sequence mutations in SARS-CoV-2 variants lead to the N3 and N5 NTD loops disappearing from the ensemble of Ab-binding substructures. This is observed computationally and is corroborated by recent experimental data by Veessler and coworkers⁶². Interestingly, these epitopes largely coincide with the regions where Alanine substitutions reduced affinity for antibodies 4A8, CM17, and CM25 (see⁶²) The impact of epitope loss in these regions is also confirmed by the observation that an engineered N3-N5 double mutant and native beta variant²⁹ both evade neutralization by mAbs CM25 and 4A8.

Interestingly, our approach correctly captures the epitopes for Abs, such as C121 and C144, that are known to engage different RBDs⁶³. Antibody C121, for instance, can bind to an RBD in the down conformation and to an adjacent RBD in the up conformation⁶³. In the structural paper, the epitope is reported to entail only residues in protomer A with the RBD in the up conformation. Contacts with the nearby RBD in the down conformation are made by Ab residues that are outside the complementary determining region. In this respect, our approach can correctly predict potential immunoreactive sequences even for Abs that would end up binding across different domains. MLCE in fact only aims to predict substructures on the antigen that can potentially be complexed by one or more Fabs. Focusing only on the

antigen, MLCE would not be able to predict whether different epitopes are targeted by the same or distinct Abs at the same time.

Finally, our strategy correctly predicts the loss of most epitopes in the lab-evolved escape variant described by Andreano *et al.*²¹ (see **Figure 4, Table S2**).

We propose a model for the study of Ab-reactivity of SARS-CoV-2 S protein variants that integrates sequence and structural information and incorporates dynamics and energetics into the analysis of the variation/loss of epitopes. Mutations in S variants determine the loss of epitopes and as a consequence can confer escape from antibodies. Upon sequence variation, the protein shifts to states characterized by different intramolecular interactions compared to the initial D614G structure; this transition decreases the number of energetically uncoupled substructures available for engaging interactors such as Abs. Unique to this model is the observation that mutations, insertions, and deletions exhibiting different immunoreactivity experimentally are consistently captured by the energy based decomposition of structures extracted from unbiased classical MD simulations of the glycosylated S protein *isolated* in solution, without any input of prior information on Ab-binding propensities. Although qualitative in nature and focused on the study of S variants of concern, our approach is general and immediately portable to other targets to provide physico-chemical information on the determinants of Abs recognition.

Since one of the fundamental goals of structural vaccinology is the identification and design of structures with optimized properties for immunoreactivity, development and validation of computational methods that help identify conserved vs. non-conserved epitope regions in different variants independently of whether structures of related protein-antibody complexes are available may hold great potential. In the case we have presented here, one may consider designing chimeras or multicomponent systems (peptide- or domain-based) presenting all (or most of) the conserved sequences that are predicted to be potentially Ab-reactive.

Furthermore, our results suggest that approaches like the one we presented here may be used prospectively as an aid in the analysis and characterization of emerging variants.

Though targeted experiments and design of mutants with tailored reactivities based on MLCE analysis are required to further validate these ideas and precisely define their progression to real-world applicability, our findings provide a new basis to understand how mutations could directly result in escape from immunorecognition.

1.13.5 Materials and Methods

Preparation of Spike Protein Variants

Fully glycosylated S protein variants simulated in this work were variously derived from simulations described by Grant *et al.*⁴⁷ based on the Cryo-EM structure of the WT S protein at PDB entry 6VSB¹¹, wherein one RBD is in the “up” conformation and the other two are “down”. All mutations, including the “reference” D614G, are introduced using the “mutations wizard” in the *PyMOL* molecular modeling package (Schrodinger LLC): rotamers of non-glycine side chains are chosen from the first suggested option for S protomer A, and then, where possible, we have sought to adopt the same rotamers for protomers B and C. Histidine tautomers and disulfide bridges are retained as in our reference simulations. In B.1.1.7 variant S protomers, mutant histidines 681 and 1118 are introduced with protonation at N ϵ 2, and mutant aspartate 570 side chains are left unprotonated. Mutant lysine 484 sidechains (B.1.1.28 variant; E484K variant) are left protonated.

Consistent with our reference simulations,^{43, 47} all three protomers are modeled without gaps, from Ala27 in the NTD to Asp1146 just downstream of heptapeptide repeat 1 (HR1); –NH₃⁺ and –COO[–] caps are added, respectively, at *N*- and *C*- termini of each protomer.

In the case of the B.1.1.7 variant, gaps left by deletions in all three protomers are replaced with artificially long C–N bonds; systems are then allowed to relax with a 400-step preminimization cycle *in vacuo* (200 steepest-descent + 200 conjugate gradient), using the *AMBER* platform’s *sander* utility (version 18)⁶⁴, in which harmonic positional restraints ($k = 5.0 \text{ kcal mol}^{-1} \text{ \AA}^{-2}$) are applied to all atoms except those in the five residues on either side of the gap. Distortions and clashes introduced with the glycosylated Ser13–Pro26 fragment are resolved using a similar approach.

The artificial PT188-EM was modeled following the methods described in²¹.

MD Simulation Details

After preparation, glycosylated S protein structures are solvated in a cuboidal box of TIP3P water molecules using *AMBER*’s *tleap* tool; where necessary, Na⁺ or Cl[–] ions are added accordingly to neutralize the charge. *N*-glycosylated asparagines and oligosaccharides are treated using the *GLYCAM-06j* forcefield⁶⁵, whereas ions are modeled with parameters by Joung and Cheatham⁶⁶. To all other (protein) atoms, we apply the *ff14SB* forcefield⁶⁷. Starting structures and topologies for all simulated variants are electronically provided.

On each glycosylated S protein variant, we conduct 4 independently replicated atomistic molecular dynamics simulations (MD), using the *AMBER* package (version 18): each replica

consists of two 300-step rounds of minimization, 2.069 ns preproduction, and 1 μ s production. The *sander* MD engine⁶⁴ is used into the earlier stages of preproduction; thereafter, we switch to the GPU-accelerated *pmemd.cuda*⁶⁴.

Details on MD production

The 1 μ s production stage is carried out in the NpT ensemble ($T = 300$ K; $p = 1$ atm) using a 2 fs time step; a cutoff of 8.0 Å is applied for the calculation of Lennard-Jones and Coulomb interactions alike. Coulomb interactions beyond this limit are computed using the Particle Mesh Ewald method⁶⁸. All bonds containing hydrogen are restrained using the *SHAKE* algorithm⁶⁹. Constant pressure is enforced *via* Berendsen's barostat⁷⁰ with a 1 ps relaxation time, whereas temperature is stabilized by Langevin's thermostat⁷¹ with a 5 ps⁻¹ collision frequency.

Details on MD preproduction

Prior to the production stage, every independent MD replica for every S variant goes through a series of preproduction steps, namely: minimization, solvent equilibration, system heating, and equilibration. The first two are conducted using the *sander* utility, after which the GPU-accelerated *pmemd.cuda* is invoked instead.

Minimization takes place in two 300-step rounds, the first 10 of which use the steepest-descent algorithm and the last 290 conjugate gradient. In the first round, we only minimize backbone H α and H1 hydrogens on aminoacids and monosaccharides, respectively, restraining all other atoms harmonically ($k = 5.0$ kcal mol⁻¹ Å⁻²). Thereafter, all atoms are released, including solvent and ions.

Solvent equilibration occurs over 9 ps with a time step of 1 fs; the ensemble is NVT , with temperatures in this case enforced by the Berendsen thermostat⁷⁰. Positions of non-solvent atoms are harmonically restrained ($k = 10$ kcal mol⁻¹ Å⁻²). Solvent molecules are assigned initial random velocities to match a temperature of 25 K. Fast heating to 400 K (coupling: 0.2 ps) is performed over the first 3 ps; the solvent is then retained at 400 K for another 3 ps; and cooled back down to 25 K over the last 3 ps, more slowly (coupling: 2.0). The cutoff for determining Lennard-Jones and Coulomb interactions remains at 8.0 Å for this and all subsequent stages, as does the Particle Mesh Ewald method⁶⁸ to determine Coulomb interactions beyond this cutoff. *SHAKE* constraints⁶⁹ are not applied at this stage, but are always present thereafter.

For system heating, the time step is increased to 2 fs and, whilst continuing in the NVT ensemble, temperatures are now enforced by the Langevin thermostat⁷¹ (which remains in

place for all subsequent stages). With an initial collision frequency of 0.75 ps^{-1} , the system is heated from 25 to 300 K over 20 ps: all atoms are free to move except aminoacids' C α atoms, which are positionally restrained with $k = 5 \text{ kcal mol}^{-1} \text{ \AA}^{-2}$.

For equilibration, the ensemble is switched to NpT ($p = 1 \text{ atm}$; Berendsen barostatcoupling: 1 ps), and the system is simulated for a further 2040 ps. The thermostat's collision frequency is kept lower than in the production stage (1 ps^{-1}). Restraints on C α atoms are lifted gradually: $k = 3.75 \text{ kcal mol}^{-1} \text{ \AA}^{-2}$ for the first 20 ps; $1.75 \text{ kcal mol}^{-1} \text{ \AA}^{-2}$ for the following 20 ps; none thereafter.

Clustering of MD Simulations

Following MD, each variant's 4 replicas are concatenated into a single $4 \mu\text{s}$ 'metatrajectory', desolvated, stripped of any ions, and aligned on backbone heavy atoms of all aminoacid residues, in all three protomers, that belong to neither the NTD nor the RBD according to domain definitions by Huang *et al.*⁷² Clustering calculations are then conducted using the hierarchical agglomerative algorithm⁷³, considering every 20th metatrajectory frame (*i.e.*, every 50 ps), based on the root-mean-square deviation of backbone heavy atoms of aminoacid residues composing the NTD and the RBD in all three protomers. Values of ϵ are chosen so that they provide the best compromise between maximizing cluster homogeneity, based on silhouette score, and ensuring at least 60-80% of the metatrajectory is covered by the three most populated clusters: this usually means $\epsilon=9-12$.

All the steps discussed in the previous paragraph are conducted using *AMBER's* postprocessing utility *cpptraj*.

MLCE method

Potential epitopes on each S variant are predicted using the Matrix of Low Coupling Energies (MLCE) method (of which we also provide a more detailed account in our previous work)⁴³. The procedure is automatically carried out by our own in-house code (<https://github.com/colombolab/MLCE>) which we have now rewritten to rely on the computationally more efficient MMPBSA.py utility⁷⁴ instead of *mm_pbsa.pl*.

The method is explained in detail in the Method section.

1.13.6 References

- (1) Dellus-Gur, E.; Toth-Petroczy, A.; Elias, M.; Tawfik, D. S. What Makes a Protein Fold Amenable to Functional Innovation? Fold Polarity and Stability Trade-Offs. *Journal of Molecular Biology* **2013**, *425*, 2609-2621.
- (2) Wellner, A.; Gurevich, M. R.; Tawfik, D. S. Mechanisms of Protein Sequence Divergence and Incompatibility. *Plos Genetics* **2013**, *9*, e1003665.
- (3) Toth-Petroczy, A.; Tawfik, D. S. The Robustness and Innovability of Protein Folds. *Current Opinion in Structural Biology* **2014**, *26*, 131-138.
- (4) Yanagida, H.; Gispán, A.; Kadouri, N.; Rozen, S.; Sharon, M.; Barkai, N.; Tawfik, D. S. The Evolutionary Potential of Phenotypic Mutations. *Plos Genetics* **2015**, *11*, e1005445.
- (5) Andreano, E.; Rappuoli, R. Sars-Cov-2 Escaped Natural Immunity, Raising Questions About Vaccines and Therapies. *Nature Medicine* **2021**, *27*, 759-765.
- (6) Pecetta, S.; Pizza, M.; Sala, C.; Andreano, E.; Pileri, P.; Troisi, M.; Pantano, E.; Manganaro, N.; Rappuoli, R. Antibodies, Epicenter of Sars-Cov-2 Immunology. *Cell Death & Differentiation* **2021**, *28*, 821-824.
- (7) Bloom, J. D.; Labthavikul, S. T.; Otey, C. R.; Arnold, F. H. Protein Stability Promotes Evolvability. *Proc. Natl. Acad. Sci. USA* **2006**, *103*, 5869-74.
- (8) Yuan, M.; Huang, D.; Lee, C.-C. D.; Wu, N. C.; Jackson, A. M.; Zhu, X.; Liu, H.; Peng, L.; van Gils, M. J.; Sanders, R. W.; Burton, D. R.; Reincke, S. M.; Prüss, H.; Kreye, J.; Nemazee, D.; Ward, A. B.; Wilson, I. A. Structural and Functional Ramifications of Antigenic Drift in Recent Sars-Cov-2 Variants. *Science* **2021**, eabh1139.
- (9) Altmann, D. M.; Boyton, R. J.; Beale, R. Immunity to Sars-Cov-2 Variants of Concern. *Science* **2021**, *371*, 1103-1104.
- (10) Harvey, W. T.; Carabelli, A. M.; Jackson, B.; Gupta, R. K.; Thomson, E. C.; Harrison, E. M.; Ludden, C.; Reeve, R.; Rambaut, A.; Peacock, S. J.; Robertson, D. L.; Consortium, C.-G. U. Sars-Cov-2 Variants, Spike Mutations and Immune Escape. *Nature Reviews Microbiology* **2021**, *19*, 409-424.
- (11) Wrapp, D.; Wang, N.; Corbett, K. S.; Goldsmith, J. A.; Hsieh, C.-L.; Abiona, O.; Graham, B. S.; McLellan, J. S. Cryo-Em Structure of the 2019-Ncov Spike in the Prefusion Conformation. *Science* **2020**, 1260-1263.
- (12) McDonald, I.; Murray, S. M.; Reynolds, C. J.; Altmann, D. M.; Boyton, R. J. Comparative Systematic Review and Meta-Analysis of Reactogenicity, Immunogenicity and Efficacy of Vaccines against Sars-Cov-2. *npj Vaccines* **2021**, *6*, 74.
- (13) Wall, E. C.; Wu, M.; Harvey, R.; Kelly, G.; Warchal, S.; Sawyer, C.; Daniels, R.; Hobson, P.; Hatipoglu, E.; Ngai, Y.; Hussain, S.; Nicod, J.; Goldstone, R.; Ambrose, K.; Hindmarsh, S.; Beale, R.; Riddell, A.; Gamblin, S.; Howell, M.; Kassiotis, G.; Libri, V.; Williams, B.; Swanton, C.; Gandhi, S.; Bauer, D. L. V. Neutralising Antibody Activity against Sars-Cov-2 Vocs B.1.617.2 and B.1.351 by Bnt162b2 Vaccination. *The Lancet* **2021**, *397*, 2331-2333.
- (14) Haas, E. J.; Angulo, F. J.; McLaughlin, J. M.; Anis, E.; Singer, S. R.; Khan, F.; Brooks, N.; Smaja, M.; Mircus, G.; Pan, K.; Southern, J.; Swerdlow, D. L.; Jodar, L.; Levy, Y.; Alroy-Preis, S. Impact and Effectiveness of Mrna Bnt162b2 Vaccine against Sars-Cov-2 Infections and Covid-19 Cases, Hospitalisations, and Deaths Following a Nationwide Vaccination Campaign in Israel: An Observational Study Using National Surveillance Data. *The Lancet* **2021**, *397*, 1819-1829.
- (15) Hsieh, C.-L.; Goldsmith, J. A.; Schaub, J. M.; DiVenere, A. M.; Kuo, H.-C.; Javanmardi, K.; Le, K. C.; Wrapp, D.; Lee, A. G.; Liu, Y.; Chou, C.-W.; Byrne, P. O.; Hjorth, C. K.; Johnson, N. V.; Ludes-Meyers, J.; Nguyen, A. W.; Park, J.; Wang, N.; Amengor, D.; Lavinder, J. J.; Ippolito, G. C.; Maynard, J. A.; Finkelstein, I. J.; McLellan, J. S. Structure-Based Design of Prefusion-Stabilized Sars-Cov-2 Spikes. *Science* **2020**, *369*, 1501.
- (16) Wang, X.; Du, Z.; Johnson, K. E.; Pasco, R. F.; Fox, S. J.; Lachmann, M.; McLellan, J. S.; Meyers, L. A. Effects of Covid-19 Vaccination Timing and Risk Prioritization on Mortality Rates, United States. *Emerging Infectious Diseases* **2021**, *27*, 1976-1979.
- (17) Greaney, A. J.; Starr, T. N.; Gilchuk, P.; Zost, S. J.; Binshtein, E.; Loes, A. N.; Hilton, S. K.; Huddleston, J.; Eguia, R.; Crawford, K. H. D.; Dingens, A. S.; Nargi, R. S.; Sutton, R. E.; Suryadevara, N.; Rothlauf, P. W.; Liu, Z.; Whelan, S. P. J.; Carnahan, R. H.; Crowe, J. E.; Bloom, J. D. Complete Mapping of Mutations to the Sars-Cov-2 Spike Receptor-Binding Domain That Escape Antibody Recognition. *Cell Host & Microbe* **2021**, *29*, 44-57.e9.
- (18) Starr, T. N.; Greaney, A. J.; Addetia, A.; Hannon, W. W.; Choudhary, M. C.; Dingens, A. S.; Li, J. Z.; Bloom, J. D. Prospective Mapping of Viral Mutations That Escape Antibodies Used to Treat Covid-19. *Science* **2021**, *371*, 850.

- (19) McCallum, M.; Bassi, J.; Marco, A. D.; Chen, A.; Walls, A. C.; Iulio, J. D.; Tortorici, M. A.; Navarro, M.-J.; Silacci-Fregni, C.; Saliba, C.; Agostini, M.; Pinto, D.; Culap, K.; Bianchi, S.; Jaconi, S.; Cameroni, E.; Bowen, J. E.; Tilles, S. W.; Pizzuto, M. S.; Guastalla, S. B.; Bona, G.; Pellanda, A. F.; Garzoni, C.; Van Voorhis, W. C.; Rosen, L. E.; Snell, G.; Telenti, A.; Virgin, H. W.; Piccoli, L.; Corti, D.; Veesler, D. Sars-Cov-2 Immune Evasion by Variant B.1.427/B.1.429. *bioRxiv* **2021**, 2021.03.31.437925.
- (20) Planas, D.; Bruel, T.; Grzelak, L.; Guivel-Benhassine, F.; Staropoli, I.; Porrot, F.; Planchais, C.; Buchrieser, J.; Rajah, M. M.; Bishop, E.; Albert, M.; Donati, F.; Prot, M.; Behillil, S.; Enouf, V.; Maquart, M.; Smati-Lafarge, M.; Varon, E.; Schortgen, F.; Yahyaoui, L.; Gonzalez, M.; De Sèze, J.; Péré, H.; Veyer, D.; Sève, A.; Simon-Lorière, E.; Fafi-Kremer, S.; Stefic, K.; Mouquet, H.; Hocqueloux, L.; van der Werf, S.; Prazuck, T.; Schwartz, O. Sensitivity of Infectious Sars-Cov-2 B.1.1.7 and B.1.351 Variants to Neutralizing Antibodies. *Nature Medicine* **2021**, 27, 917-924.
- (21) Andreano, E.; Piccini, G.; Licastro, D.; Casalino, L.; Johnson, N. V.; Paciello, I.; Monego, S. D.; Pantano, E.; Manganaro, N.; Manenti, A.; Manna, R.; Casa, E.; Hyseni, I.; Benincasa, L.; Montomoli, E.; Amaro, R. E.; McLellan, J. S.; Rappuoli, R. Sars-Cov-2 Escape in Vitro from a Highly Neutralizing Covid-19 Convalescent Plasma. *bioRxiv* **2020**, 2020.12.28.424451.
- (22) McCormick, K. D.; Jacobs, J. L.; Mellors, J. W. The Emerging Plasticity of Sars-Cov-2. *Science* **2021**, 371, 1306.
- (23) Wang, Z.; Schmidt, F.; Weisblum, Y.; Muecksch, F.; Barnes, C. O.; Finkin, S.; Schaefer-Babajew, D.; Cipolla, M.; Gaebler, C.; Lieberman, J. A.; Oliveira, T. Y.; Yang, Z.; Abernathy, M. E.; Huey-Tubman, K. E.; Hurley, A.; Turroja, M.; West, K. A.; Gordon, K.; Millard, K. G.; Ramos, V.; Da Silva, J.; Xu, J.; Colbert, R. A.; Patel, R.; Dizon, J.; Unson-O'Brien, C.; Shimeliovich, I.; Gazumyan, A.; Caskey, M.; Bjorkman, P. J.; Casellas, R.; Hatziioannou, T.; Bieniasz, P. D.; Nussenzweig, M. C. Mrna Vaccine-Elicited Antibodies to Sars-Cov-2 and Circulating Variants. *Nature* **2021**, 592, 616-622.
- (24) Muik, A.; Wallisch, A.-K.; Sängler, B.; Swanson, K. A.; Mühl, J.; Chen, W.; Cai, H.; Maurus, D.; Sarkar, R.; Türeci, Ö.; Dormitzer, P. R.; Şahin, U. Neutralization of Sars-Cov-2 Lineage B.1.1.7 Pseudovirus by Bnt162b2 Vaccine–Elicited Human Sera. *Science* **2021**, 371, 1152.
- (25) Abu-Raddad, L. J.; Chemaitelly, H.; Butt, A. A. Effectiveness of the Bnt162b2 Covid-19 Vaccine against the B.1.1.7 and B.1.351 Variants. *New England Journal of Medicine* **2021**.
- (26) Tarke, A.; Sidney, J.; Methot, N.; Yu, E. D.; Zhang, Y.; Dan, J. M.; Goodwin, B.; Rubiro, P.; Sutherland, A.; Wang, E.; Frazier, A.; Ramirez, S. I.; Rawlings, S. A.; Smith, D. M.; da Silva Antunes, R.; Peters, B.; Scheuermann, R. H.; Weiskopf, D.; Crotty, S.; Grifoni, A.; Sette, A. Impact of Sars-Cov-2 Variants on the Total Cd4⁺ and Cd8⁺ T Cell Reactivity in Infected or Vaccinated Individuals. *Cell Reports Medicine*.
- (27) Control., E. C. f. D. P. a. Assessing Sars-Cov-2 Circulation, Variants of Concern, Non-Pharmaceutical Interventions and Vaccine Rollout in the Eu/Eea. *ECDC: Stockholm* **2021**, 15th update.
- (28) Madhi, S. A.; Baillie, V.; Cutland, C. L.; Voysey, M.; Koen, A. L.; Fairlie, L.; Padayachee, S. D.; Dheda, K.; Barnabas, S. L.; Bhorat, Q. E.; Briner, C.; Kwatra, G.; Ahmed, K.; Aley, P.; Bhikha, S.; Bhiman, J. N.; Bhorat, A. a. E.; du Plessis, J.; Esmail, A.; Groenewald, M.; Horne, E.; Hwa, S.-H.; Jose, A.; Lambe, T.; Laubscher, M.; Malahleha, M.; Masenya, M.; Masilela, M.; McKenzie, S.; Molapo, K.; Moultrie, A.; Oelofse, S.; Patel, F.; Pillay, S.; Rhead, S.; Rodel, H.; Rossouw, L.; Taoushanis, C.; Tegally, H.; Thombrayil, A.; van Eck, S.; Wibmer, C. K.; Durham, N. M.; Kelly, E. J.; Villafana, T. L.; Gilbert, S.; Pollard, A. J.; de Oliveira, T.; Moore, P. L.; Sigal, A.; Izu, A. Efficacy of the Chadox1 Ncov-19 Covid-19 Vaccine against the B.1.351 Variant. *New England Journal of Medicine* **2021**, 384, 1885-1898.
- (29) Tegally, H.; Wilkinson, E.; Giovanetti, M.; Iranzadeh, A.; Fonseca, V.; Giandhari, J.; Doolabh, D.; Pillay, S.; San, E. J.; Msomi, N.; Mlisana, K.; von Gottberg, A.; Walaza, S.; Allam, M.; Ismail, A.; Mohale, T.; Glass, A. J.; Engelbrecht, S.; Van Zyl, G.; Preiser, W.; Petruccione, F.; Sigal, A.; Hardie, D.; Marais, G.; Hsiao, N.-y.; Korsman, S.; Davies, M.-A.; Tyers, L.; Mudau, I.; York, D.; Maslo, C.; Goedhals, D.; Abrahams, S.; Laguda-Akingba, O.; Alisoltani-Dehkordi, A.; Godzik, A.; Wibmer, C. K.; Sewell, B. T.; Lourenço, J.; Alcantara, L. C. J.; Kosakovsky Pond, S. L.; Weaver, S.; Martin, D.; Lessells, R. J.; Bhiman, J. N.; Williamson, C.; de Oliveira, T. Detection of a Sars-Cov-2 Variant of Concern in South Africa. *Nature* **2021**, 592, 438-443.
- (30) Scarabelli, G.; Morra, G.; Colombo, G. Predicting Interaction Sited from the Energetics of Isolated Proteins: A New Approach to Epitope Mapping. *Biophys. J.* **2010**, 98, 1966-1975.
- (31) Gourlay, L. J.; Peri, C.; Ferrer-Navarro, M.; Conchillo-Solé, O.; Gori, A.; Rinchai, D.; Thomas, R. J.; Champion, O. L.; Michell, S. L.; Kewcharoenwong, C.; Nithichanon, A.; Lassaux, P.; Perletti, L.; Longhi, R.; Lertmemongkolchai, G.; Titball, R. W.; Daura, X.; Colombo, G.; Bolognesi, M. Exploiting the Burkholderia Pseudomallei Acute Phase Antigen Bpsl2765 for Structure-Based Epitope Discovery/Design in Structural Vaccinology. *Chem. Biol.* **2013**, 20, 1147-1156.

- (32) Lassaux, P.; Peri, C.; Ferrer-Navarro, M.; Gourlay, L.; Gori, A.; Conchillo-Solé, O.; Rinchai, D.; Lertmemongkolchai, G.; Longhi, R.; Daura, X.; Colombo, G.; Bolognesi, M. A Structure-Based Strategy for Epitope Discovery in Burkholderia Pseudomallei Oppa Antigen. *Structure* **2013**, *21*, 1-9.
- (33) Gourlay, L.; Peri, C.; Bolognesi, M.; Colombo, G. Structure and Computation in Immunoreagent Design: From Diagnostics to Vaccines. *Trends in Biotechnology* **2017**, *35*, 1208-1220.
- (34) Capelli, R.; Serapian, S. A.; Colombo, G. Computational Epitope Prediction and Design for Antibody Development and Detection. *Meth. Mol. Biol.* **2021**, *In press*.
- (35) Marchetti, F.; Capelli, R.; Rizzato, F.; Laio, A.; Colombo, G. The Subtle Trade-Off between Evolutionary and Energetic Constraints in Protein-Protein Interactions. *J. Phys. Chem. Lett.* **2019**, *10*, 1489-1497.
- (36) Soriani, M.; Petit, P.; Grifantini, R.; Petracca, R.; Gancitano, G.; Frigimelica, E.; Nardelli, F.; Garcia, C.; Spinelli, S.; Scarabelli, G.; Fiorucci, S.; Affentranger, R.; Ferrer-Navarro, M.; Zacharias, M.; Colombo, G.; Vuillard, L.; Daura, X.; Grandi, G. Exploiting Antigenic Diversity for Vaccine Design: The Chlamydia Artj Paradigm. *J. Biol. Chem.* **2010**, *285*, 30126-30138.
- (37) Gourlay, L. J.; Lassaux, P.; Thomas, R. J.; Peri, C.; Conchillo-Sole, O.; Nithichanon, A.; Ferrer-Navarro, M.; Vila, J.; Daura, X.; Lertmemongkolchai, G.; Titball, R.; Colombo, G.; Bolognesi, M. Flagellar Subunits as Targets for Structure-Based Epitope Discovery Approaches and Melioidosis Vaccine Development. *Febs Journal* **2015**, *282*, 338-338.
- (38) Gourlay, L. J.; Thomas, R. J.; Peri, C.; Conchillo-Sole, O.; Ferrer-Navarro, M.; Nithichanon, A.; Vila, J.; Daura, X.; Lertmemongkolchai, G.; Titball, R.; Colombo, G.; Bolognesi, M. From Crystal Structure to in Silico Epitope Discovery in the Burkholderia Pseudomallei Flagellar Hook-Associated Protein Flgk. *Febs Journal* **2015**, *282*, 1319-1333.
- (39) Nithichanon, A.; Rinchai, D.; Gori, A.; Lassaux, P.; Peri, C.; Conchillo-Sole, O.; Ferrer-Navarro, M.; Gourlay, L. J.; Nardini, M.; Vila, J.; Daura, X.; Colombo, G.; Bolognesi, M.; Lertmemonkolchai, G. Sequence- and Structure-Based Immunoreactive Epitope Discovery for Burkholderia Pseudomallei Flagellin. *Plos Neglected Tropical Diseases* **2015**, *9*.
- (40) Gori, A.; Peri, C.; Quilici, G.; Nithichanon, A.; Gaudesi, D.; Longhi, R.; Gourlay, L.; Bolognesi, M.; Lertmemongkolchai, G.; Musco, G.; Colombo, G. Flexible Vs Rigid Epitope Conformations for Diagnostic- and Vaccine-Oriented Applications: Novel Insights from the Burkholderia Pseudomallei BpsI2765 Pa13 Epitope. *Acs Infectious Diseases* **2016**, *2*, 221-230.
- (41) Gori, A.; Sola, L.; Gagni, P.; Bruni, G.; Liprino, M.; Peri, C.; Colombo, G.; Cretich, M.; Chiari, M. Screening Complex Biological Samples with Peptide Microarrays: The Favorable Impact of Probe Orientation Via Chemoselective Immobilization Strategies on Clickable Polymeric Coatings. *Bioconjugate Chemistry* **2016**, *27*, 2669-2677.
- (42) Sola, L.; Gagni, P.; D'Annessa, I.; Capelli, R.; Bertino, C.; Romanato, A.; Damin, F.; Bergamaschi, G.; Marchisio, E.; Cuzzocrea, A.; Bombaci, M.; Grifantini, R.; Chiari, M.; Colombo, G.; Gori, A.; Cretich, M. Enhancing Antibody Serodiagnosis Using a Controlled Peptide Coimmobilization Strategy. *ACS Infectious Diseases* **2018**, *4*, 998-1006.
- (43) Serapian, S. A.; Marchetti, F.; Triveri, A.; Morra, G.; Meli, M.; Moroni, E.; Sautto, G. A.; Rasola, A.; Colombo, G. The Answer Lies in the Energy: How Simple Atomistic Molecular Dynamics Simulations May Hold the Key to Epitope Prediction on the Fully Glycosylated Sars-Cov-2 Spike Protein. *The Journal of Physical Chemistry Letters* **2020**, 8084-8093.
- (44) McCallum, M.; De Marco, A.; Lempp, F. A.; Tortorici, M. A.; Pinto, D.; Walls, A. C.; Beltramello, M.; Chen, A.; Liu, Z.; Zatta, F.; Zepeda, S.; di Iulio, J.; Bowen, J. E.; Montiel-Ruiz, M.; Zhou, J.; Rosen, L. E.; Bianchi, S.; Guarino, B.; Fregni, C. S.; Abdelnabi, R.; Foo, S.-Y. C.; Rothlauf, P. W.; Bloyet, L.-M.; Benigni, F.; Cameroni, E.; Neyts, J.; Riva, A.; Snell, G.; Telenti, A.; Whelan, S. P. J.; Virgin, H. W.; Corti, D.; Pizzuto, M. S.; Veesler, D. N-Terminal Domain Antigenic Mapping Reveals a Site of Vulnerability for Sars-Cov-2. *Cell* **2021**, *184*, 2332-2347.e16.
- (45) Yadav, P. D.; Sapkal, G. N.; Ella, R.; Sahay, R. R.; Nyayanit, D. A.; Patil, D. Y.; Deshpande, G.; Shete, A. M.; Gupta, N.; Mohan, V. K.; Abraham, P.; Panda, S.; Bhargava, B. Neutralization against B.1.351 and B.1.617.2 with Sera of Covid-19 Recovered Cases and Vaccinees of Bbv152. *bioRxiv* **2021**, 2021.06.05.447177.
- (46) Grant, O. C.; Montgomery, D.; Ito, K.; Woods, R. J. 3d Models of Glycosylated Sars-Cov-2 Spike Protein Suggest Challenges and Opportunities for Vaccine Development. *bioRxiv* **2020**, 2020.04.07.030445.
- (47) Grant, O. C.; Montgomery, D.; Ito, K.; Woods, R. J. Analysis of the Sars-Cov-2 Spike Protein Glycan Shield Reveals Implications for Immune Recognition. *Sci. Rep.* **2020**, *10*, 14991
- (48) Smith, C. C.; Entwistle, S.; Willis, C.; Vensko, S.; Beck, W.; Garness, J.; Sambade, M.; Routh, E.; Olsen, K.; Kodysh, J.; O'Donnell, T.; Haber, C.; Heiss, K.; Stadler, V.; Garrison, E.; Grant, O. C.; Woods, R. J.; Heise, M.; Vincent, B. G.; Rubinsteyn, A. Landscape and Selection of Vaccine Epitopes in Sars-Cov-2. *bioRxiv : the preprint server for biology* **2020**, 2020.06.04.135004.

- (49) Zhao, P.; Praissman, J. L.; Grant, O. C.; Cai, Y.; Xiao, T.; Rosenbalm, K. E.; Aoki, K.; Kellman, B. P.; Bridger, R.; Barouch, D. H.; Brindley, M. A.; Lewis, N. E.; Tiemeyer, M.; Chen, B.; Woods, R. J.; Wells, L. Virus-Receptor Interactions of Glycosylated Sars-Cov-2 Spike and Human Ace2 Receptor. *Cell host & microbe* **2020**, *28*, 586-601.e6.
- (50) Casalino, L.; Gaieb, Z.; Goldsmith, J. A.; Hjorth, C. K.; Dommer, A. C.; Harbison, A. M.; Fogarty, C. A.; Barros, E. P.; Taylor, B. C.; McLellan, J. S.; Fadda, E.; Amaro, R. E. Beyond Shielding: The Roles of Glycans in the Sars-Cov-2 Spike Protein. *ACS central science* **2020**, *6*, 1722-1734.
- (51) Barros, E. P.; Casalino, L.; Gaieb, Z.; Dommer, A. C.; Wang, Y.; Fallon, L.; Raguette, L.; Belfon, K.; Simmerling, C.; Amaro, R. E. The Flexibility of Ace2 in the Context of Sars-Cov-2 Infection. *Biophysical Journal* **2021**, *120*, 1072-1084.
- (52) Casalino, L.; Dommer, A.; Gaieb, Z.; Barros, E. P.; Sztain, T.; Ahn, S.-H.; Trifan, A.; Brace, A.; Bogetti, A.; Ma, H.; Lee, H.; Turilli, M.; Khalid, S.; Chong, L.; Simmerling, C.; Hardy, D. J.; Maia, J. D. C.; Phillips, J. C.; Kurth, T.; Stern, A.; Huang, L.; McCalpin, J.; Tatineni, M.; Gibbs, T.; Stone, J. E.; Jha, S.; Ramanathan, A.; Amaro, R. E. Ai-Driven Multiscale Simulations Illuminate Mechanisms of Sars-Cov-2 Spike Dynamics. *bioRxiv : the preprint server for biology* **2020**, 2020.11.19.390187.
- (53) Zimmerman, M. I.; Porter, J. R.; Ward, M. D.; Singh, S.; Vithani, N.; Meller, A.; Mallimadugula, U. L.; Kuhn, C. E.; Borowsky, J. H.; Wiewiora, R. P.; Hurley, M. F. D.; Harbison, A. M.; Fogarty, C. A.; Coffland, J. E.; Fadda, E.; Voelz, V. A.; Chodera, J. D.; Bowman, G. R. Sars-Cov-2 Simulations Go Exascale to Predict Dramatic Spike Opening and Cryptic Pockets across the Proteome. *Nature chemistry* **2021**, *13*, 651-659.
- (54) Spinello, A.; Saltalamacchia, A.; Magistrato, A. Is the Rigidity of Sars-Cov-2 Spike Receptor-Binding Motif the Hallmark for Its Enhanced Infectivity? Insights from All-Atom Simulations. *The Journal of Physical Chemistry Letters* **2020**, *11*, 4785-4790.
- (55) Spinello, A.; Saltalamacchia, A.; Borišek, J.; Magistrato, A. Allosteric Cross-Talk among Spike's Receptor-Binding Domain Mutations of the Sars-Cov-2 South African Variant Triggers an Effective Hijacking of Human Cell Receptor. *The Journal of Physical Chemistry Letters* **2021**, *12*, 5987-5993.
- (56) Oliveira, A. S. F.; Ibarra, A. A.; Bermudez, I.; Casalino, L.; Gaieb, Z.; Shoemark, D. K.; Gallagher, T.; Sessions, R. B.; Amaro, R. E.; Mulholland, A. J. A Potential Interaction between the Sars-Cov-2 Spike Protein and Nicotinic Acetylcholine Receptors. *Biophysical journal* **2021**, *120*, 983-993.
- (57) Shoemark, D. K.; Colenso, C. K.; Toelzer, C.; Gupta, K.; Sessions, R. B.; Davidson, A. D.; Berger, I.; Schaffitzel, C.; Spencer, J.; Mulholland, A. J. Molecular Simulations Suggest Vitamins, Retinoids and Steroids as Ligands of the Free Fatty Acid Pocket of the Sars-Cov-2 Spike Protein**. *Angewandte Chemie International Edition* **2021**, *60*, 7098-7110.
- (58) Wang, Y.; Fallon, L.; Raguette, L.; Budhan, S.; Belfon, K.; Stepanenko, D.; Helbock, S.; Varghese, S.; Simmerling, C. Receptor Binding May Directly Activate the Fusion Machinery in Coronavirus Spike Glycoproteins. *bioRxiv* **2021**, 2021.05.10.443496.
- (59) Fallon, L.; Belfon, K. A. A.; Raguette, L.; Wang, Y.; Stepanenko, D.; Cuomo, A.; Guerra, J.; Budhan, S.; Varghese, S.; Corbo, C. P.; Rizzo, R. C.; Simmerling, C. Free Energy Landscapes from Sars-Cov-2 Spike Glycoprotein Simulations Suggest That Rbd Opening Can Be Modulated Via Interactions in an Allosteric Pocket. *Journal of the American Chemical Society* **2021**, *143*, 11349-11360.
- (60) Thomson, E. C.; Rosen, L. E.; Shepherd, J. G.; Spreafico, R.; da Silva Filipe, A.; Wojcechowskyj, J. A.; Davis, C.; Piccoli, L.; Pascall, D. J.; Dillen, J.; Lytras, S.; Czudnochowski, N.; Shah, R.; Meury, M.; Jesudason, N.; De Marco, A.; Li, K.; Bassi, J.; O'Toole, A.; Pinto, D.; Colquhoun, R. M.; Culap, K.; Jackson, B.; Zatta, F.; Rambaut, A.; Jaconi, S.; Sreenu, V. B.; Nix, J.; Zhang, I.; Jarrett, R. F.; Glass, W. G.; Beltramelio, M.; Nomikou, K.; Pizzuto, M.; Tong, L.; Cameroni, E.; Croll, T. I.; Johnson, N.; Di Iulio, J.; Wickenhagen, A.; Ceschi, A.; Harbison, A. M.; Mair, D.; Ferrari, P.; Smollett, K.; Sallusto, F.; Carmichael, S.; Garzoni, C.; Nichols, J.; Galli, M.; Hughes, J.; Riva, A.; Ho, A.; Schiuma, M.; Semple, M. G.; Openshaw, P. J. M.; Fadda, E.; Baillie, J. K.; Chodera, J. D.; Rihn, S. J.; Lycett, S. J.; Virgin, H. W.; Telenti, A.; Corti, D.; Robertson, D. L.; Snell, G. Circulating Sars-Cov-2 Spike N439k Variants Maintain Fitness While Evading Antibody-Mediated Immunity. *Cell* **2021**, *184*, 1171-1187.e20.
- (61) Gobeil, S. M. C.; Janowska, K.; McDowell, S.; Mansouri, K.; Parks, R.; Stalls, V.; Kopp, M. F.; Manne, K.; Li, D.; Wiehe, K.; Saunders, K. O.; Edwards, R. J.; Korber, B.; Haynes, B. F.; Henderson, R.; Acharya, P. Effect of Natural Mutations of Sars-Cov-2 on Spike Structure, Conformation, and Antigenicity. *Science* **2021**, eabi6226.
- (62) Voss, W. N.; Hou, Y. J.; Johnson, N. V.; Delidakis, G.; Kim, J. E.; Javanmardi, K.; Horton, A. P.; Bartzoka, F.; Paresi, C. J.; Tanno, Y.; Chou, C.-W.; Abbasi, S. A.; Pickens, W.; George, K.; Boutz, D. R.; Towers, D. M.; McDaniel, J. R.; Billick, D.; Goike, J.; Rowe, L.; Batra, D.; Pohl, J.; Lee, J.; Gangappa, S.; Sambhara, S.; Gadush, M.; Wang, N.; Person, M. D.; Iverson, B. L.; Gollihar, J. D.; Dye, J. M.; Herbert, A. S.; Finkelstein, I. J.; Baric, R. S.; McLellan, J. S.; Georgiou, G.; Lavinder, J. J.; Ippolito, G. C. Prevalent, Protective, and Convergent Igg Recognition of Sars-Cov-2 Non-Rbd Spike Epitopes. *Science* **2021**, *372*, 1108.

- (63) Barnes, C. O.; West, A. P., Jr.; Huey-Tubman, K. E.; Hoffmann, M. A. G.; Sharaf, N. G.; Hoffman, P. R.; Koranda, N.; Gristick, H. B.; Gaebler, C.; Muecksch, F.; Lorenzi, J. C. C.; Finkin, S.; Hägglöf, T.; Hurley, A.; Millard, K. G.; Weisblum, Y.; Schmidt, F.; Hatzioannou, T.; Bieniasz, P. D.; Caskey, M.; Robbiani, D. F.; Nussenzweig, M. C.; Bjorkman, P. J. Structures of Human Antibodies Bound to Sars-Cov-2 Spike Reveal Common Epitopes and Recurrent Features of Antibodies. *Cell* **2020**, S0092-8674(20)30757-1.
- (64) Case, D. A.; Cerutti, D. S.; Cheatham, T. E. I.; Darden, T. A.; Duke, R. E.; Giese, T. J.; Gohlke, H.; Goetz, A. W.; Greene, D.; Homeyer, N.; Izadi, S.; Kovalenko, A.; Lee, T. S.; LeGrand, S.; Li, P. L., C.; Liu, J.; Luchko, T.; Luo, R.; Mermelstein, D.; Merz, K. M.; Monard, G.; Nguyen, H.; Omelyan, I.; Onufriev, A.; Pan, F.; Qi, R.; Roe, D. R.; Roitberg, A.; Sagui, C.; Simmerling, C. L.; Botello-Smith, W. M.; Swails, J.; Walker, R. C.; Wang, J.; Wolf, R. M.; Wu, X.; Xiao, L.; York, D. M.; Kollman, P. A. Amber 2018. *University of California, San Francisco* **2018**.
- (65) Kirschner, K. N.; Yongye, A. B.; Tschampel, S. M.; González-Outeiriño, J.; Daniels, C. R.; Foley, B. L.; Woods, R. J. Glycam06: A Generalizable Biomolecular Force Field. *Carbohydrates. Journal of Computational Chemistry* **2008**, *29*, 622-655.
- (66) Joung, I. S.; Cheatham, T. E. Determination of Alkali and Halide Monovalent Ion Parameters for Use in Explicitly Solvated Biomolecular Simulations. *The Journal of Physical Chemistry B* **2008**, *112*, 9020-9041.
- (67) Maier, J. A.; Martinez, C.; Kasavajhala, K.; Wickstrom, L.; Hauser, K. E.; Simmerling, C. Ff14sb: Improving the Accuracy of Protein Side Chain and Backbone Parameters from Ff99sb. *Journal of Chemical Theory and Computation* **2015**, *11*, 3696-3713.
- (68) Darden, T.; York, D.; Pedersen, L. Particle Mesh Ewald: An N-Log(N) Method for Ewald Sums in Large Systems. *J. Chem. Phys.* **1993**, *98*.
- (69) Miyamoto, S.; Kollman, P. A. Settle: An Analytical Version of the Shake and Rattle Algorithms for Rigid Water Models. *J. Comp. Chem.* **1992**, *13*, 952-962.
- (70) Berendsen, H. J. C.; Postma, J. P. M.; van Gunsteren, W. F.; Di Nola, A.; Haak, J. R. Molecular Dynamics with Coupling to an External Bath. *J. Chem. Phys.* **1984**, *81*, 3684-3690.
- (71) Loncharich, R. J.; Brooks, B. R.; Pastor, R. W. Langevin Dynamics of Peptides: The Frictional Dependence of Isomerization Rates of N-Acetylalanyl-N'-Methylamide. *Biopolymers* **1992**, *32*, 523-535.
- (72) Huang, Y.; Yang, C.; Xu, X.-f.; Xu, W.; Liu, S.-w. Structural and Functional Properties of Sars-Cov-2 Spike Protein: Potential Antivirus Drug Development for Covid-19. *Acta Pharmacologica Sinica* **2020**, *41*, 1141-1149.
- (73) Defays, D. An Efficient Algorithm for a Complete Link Method. *The Computer Journal* **1977**, *20*, 364-366.
- (74) Miller, B. R.; McGee, T. D.; Swails, J. M.; Homeyer, N.; Gohlke, H.; Roitberg, A. E. Mmpbsa.Py: An Efficient Program for End-State Free Energy Calculations. *Journal of Chemical Theory and Computation* **2012**, *8*, 3314-3321.
- (75) Onufriev, A.; Bashford, D.; Case, D. A. Modification of the Generalized Born Model Suitable for Macromolecules. *J. Phys. Chem. B.* **2000**, *104*, 3712-3720.
- (76) Genheden, S.; Ryde, U. The Mm/Pbsa and Mm/Gbsa Methods to Estimate Ligand-Binding Affinities. *Expert Opin Drug Discov.* **2015**, *10*, 449-461.
- (77) Genoni, A.; Morra, G.; Colombo, G. Identification of Domains in Protein Structures from the Analysis of Intramolecular Interactions. *J. Phys. Chem. B.* **2012**, *116*, 3331-3343.

Immunoreactivity of the Spike Protein: Conclusion

In conclusion, in this first section (**Immunoreactivity of the Spike Protein**), we utilized a novel energy-decomposition approach detailed in the Method section to comprehensively identify antigenic domains and antibody binding sites within the fully glycosylated SARS-CoV-2 Spike protein. A key advantage of our method is its reliance solely on unbiased atomistic molecular dynamics simulations, eliminating the need for prior knowledge of binding properties or ad hoc combinations of simulation-derived parameters. Our approach involved a meticulous analysis of energy interactions among all intra-protomer amino acid and monosaccharide residue pairs, cross-referenced with structural data, particularly residue-residue proximity. This method enabled the identification of spatially contiguous residues displaying weak energetic coupling within the protein, pinpointing potential immunogenic regions. Validation of our findings was conducted through a comparison with experimentally confirmed structures of the S protein complexed with anti- or nanobodies. This validation procedure facilitated the identification of subdomains with poor energetic coupling, likely accommodating multiple epitopes and potentially contributing to significant functional conformational changes.

Furthermore, our investigation unveiled distinct behaviors of the glycan shield associated with the Spike protein. Glycans with stronger energetic coupling were found to be structurally significant, providing protection to underlying peptidic epitopes. Conversely, glycans with weaker coupling could be susceptible to antibody recognition.

These predictions of immunoreactive regions pave the way for the development of optimized antigens, including recombinant subdomains and synthetic (glyco)peptidomimetics, holding promise for therapeutic applications. Additionally, similar predictive approaches can bolster preparedness for future pandemic outbreaks.

As the SARS-CoV-2 spike protein stands as a primary target for COVID-19 vaccines, the emergence of variants capable of evading antibody recognition raises critical concerns about the effectiveness of immunological treatments. So, this computational model can be used to predict the impact of S protein mutations on antibody binding sites. Thereby, it has successfully identified known epitopes from the reference structure and correlated mutations with loss of potential immunoreactive regions. Moreover, the versatility of our computational epitope prediction strategy extends its applicability to the study of immunoreactivity in mutants of other characterized proteins, promisingly contributing to the development and

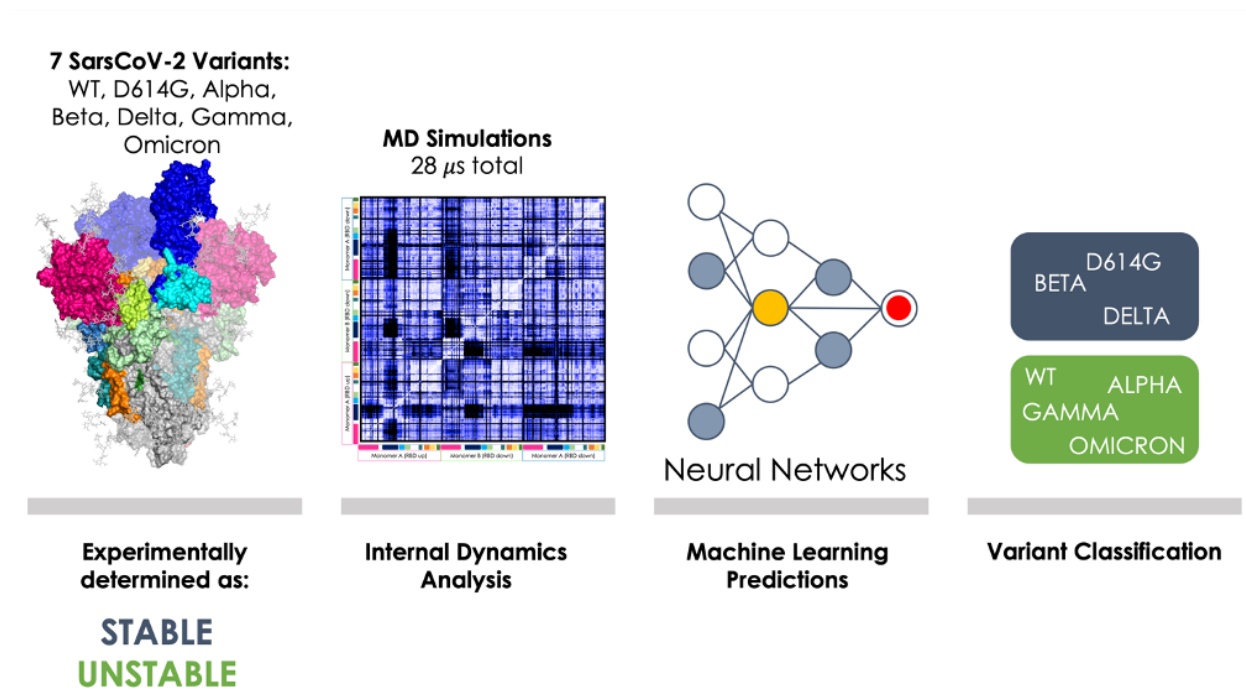
proactive selection of vaccines and antibodies that can address emerging variants and ensure effective responses to evolving viral threats.

In the pursuit of understanding virus immunoreactivity, a remarkable aspect that has come to light is the astonishing lack of rigor in the proliferation of variants. Over the initial three years of the pandemic, SARS-CoV-2 underwent rapid evolution. What stood out was the initial evolution of the virus, which appeared to advance through substantial sequence changes rather than the gradual accumulation of point mutations on existing variants.

Having assessed the impact of these mutations on antibody recognition, we now delve into whether this non-linear mutational trajectory is reflected in variations in the conformational dynamics of the SARS-CoV-2 Spike protein. Our objective is to comprehensively grasp the intricate interplay between the evolving mutational landscape and the functional dynamics of the Spike protein, crucial for advancing our knowledge and strategies in combating the evolving virus.

1.14 Structural dynamic differences of the VOCs SARS-CoV-2 spike protein

“The Conformational Behaviour of SARS-Cov-2 Spike Protein Variants: Evolutionary Jumps In Sequence Reverberate In Structural Dynamic Differences”



1.14.1 Abstract

To understand how this non-linear mutational process reverberates in variations of the conformational dynamics of the SARS-CoV-2 Spike protein, we run extensive microsecond-scale MD simulations of seven distinct variants of the protein in their fully glycosylated state and set out to elucidate possible links between the mutational spectrum of the S-protein and the structural dynamics of the respective variant, at the global and local levels. The results reveal that mutation-dependent structural and dynamic modulations mostly consist of increased coordinated motions in variants that acquire stability and in an increased internal flexibility in variants that are less stable. Importantly, a limited number of functionally important substructures (the Receptor Binding Domain, RBD, in particular) share the same time of movements in all variants, indicating efficient preorganization for functional regions dedicated to host-interactions.

Our results support a model in which the internal dynamics of the S-proteins from different strains varies in a way that reflects the observed random and non-stepwise jumps in

sequence evolution, while conserving the functionally oriented traits of conformational dynamics necessary to support productive interactions with host receptors.

1.14.2 Introduction

Viruses are known to evolve, and we said that SARS-CoV-2 is no exception. While it was initially expected that new mutants would descend from existing ones through a stepwise process in which new mutations are implanted on successful sequences, sequencing data showed that the newer and more efficient variants (e.g. Omicron and alike) harbor a notably large number of mutations.^{24, 26}

This represents a key peculiarity of SARS-CoV-2: as noted by Bloom and colleagues (The New York Times “We Study Virus Evolution. Here’s Where We Think the Coronavirus Is Going.” March 28, 2022. <https://www.nytimes.com/interactive/2022/03/28/opinion/coronavirus-mutation-future.html>), the virus seemed to defy common knowledge with its variants emerging through big evolutionary jumps, at least in the initial steps of diffusion. In this context, it is important to note that there is a great sequence difference between one of the earlier most infective variants, namely Delta, and the later ones, i.e. Omicron.

The salient features of the evolution of viral variants of concern can effectively be traced to the evolution of the sequence of the Spike protein. In this context, the history of VOC development has already been elucidated. In early March 2020 the first point mutation appeared, a single amino acid change caused by an A-to-G nucleotide mutation at position 23,403 in the Wuhan reference strain. This mutation gave rise to the emergence of the dominant D614G Spike variant, which rapidly spread from Europe to North America, Oceania, and Asia.³²⁻³⁵ After this first one, an increased level of surveillance and sequencing contributed to reveal novel variants.

Among the ones that have been brought to attention in the last couple of years, the one labelled 20I/501Y.V1 or B.1.1.7, commonly named Alpha variant, was initially found in the UK and was associated with an increased risk of infection and death.³⁶ In South Africa, variant B.1.351 (known as 20H/501Y.V2, Beta) emerged independently from B.1.1.7 but shared some mutations with it.³⁷ Next, The P.1 variant (20J/501Y.V3, Gamma) was first identified in travelers from Brazil and featured 17 unique mutations including three in the receptor binding domain of the Spike protein, two shared with B.1.351, E484K and N501Y, the latter also shared with the strain of B.1.1.7, and a different mutation K417T which was K417N in the B1.351, Beta strain.³⁸ The B.1.617.2 variant (AY, Delta) was first detected in

India in late 2020, where it was responsible for a huge surge in the number of cases, and in June 2021 it became the dominant variant globally.³⁹ The SARS-CoV-2 Omicron (B.1.1.529, BA.1) variant was first identified on November 24th, 2021, in South Africa and immediately declared VOC replacing the Delta variant. The Omicron variant has a very large number of mutations, around 30-point mutations in the Spike protein alone, combined with deletions and insertions of amino acids.¹⁵

The Spike protein perturbations associated to the different variants described are summarized **Figure 1** (list of the ones studied in this paper). Further variants have emerged and continue to emerge due to the pressure exerted by the virus to adapt and to survive in an increasingly immunized population, such as the Epsilon (B.1.427 and B.1.429), Eta (B.1.525), Iota (B.1.526) Kappa (B.1.617.1) and Mu (B.1.621, B.1.621.1), etc (<https://www.who.int/activities/tracking-SARS-CoV-2-variants>).

The differences between the above-mentioned variants have been studied diffusely. Veessler et al. linked the conformational properties to plasma neutralizing activities.⁴⁰

A paper by the Amaro and Freeman groups showed that Omicron specifically modified its positive surface charge to improve interactions with heparan sulfates and ACE2.⁴¹ This effect was related to enhanced binding rates to charged glycolyx molecules. Other studies have shown that specific mutations in the RBD can also be correlated to increased affinity for ACE2.⁴²⁻⁴⁵

The importance of long-range modulation of S-dynamics was demonstrated to be fundamental in response to the binding of endogenous molecules, such as fatty acids, that were proven to preorganize the RBD for attachment to the receptor.⁴⁶⁻⁵⁰

Here, we ask whether the significant sequence differences observed for the various strains reverberate in changes in the traits of long-range structural dynamics of the Spike protein by comparing seven different mutant sequences. Specifically, we analyze how the dynamics is modulated by mutations (compared to the initial Wuhan variant, the Wild Type (WT) in our model) both at the level of global and local motions, specifically focusing on substructures that are important for Spike functions (i.e. recognition of the human receptor angiotensin-converting enzyme 2 (ACE2) and conformational reorganization of the architecture to favor host-virus membrane fusion).

To progress along this avenue, we address various aspects of this problem by analyzing and comparing atomistic simulations of S-protein mutants reported in **Figure 1**, in their fully glycosylated form. Starting from the atomistic resolution investigation of internal fluctuations and analysis of the coordination in the motion of different domains, we demonstrate that

different mutants show distinctive dynamic traits, which can be qualitatively correlated to their relative stability properties.

Our results also indicate that the structural dynamics of S-proteins from different strains varies in a way that appears to reflect the jumps in sequence evolution observed at the initial stages of diffusion of the virus.

The Structural organization of the SARS-CoV-2 Spike protein and localization of the mutations are represent in **Figure 1**.

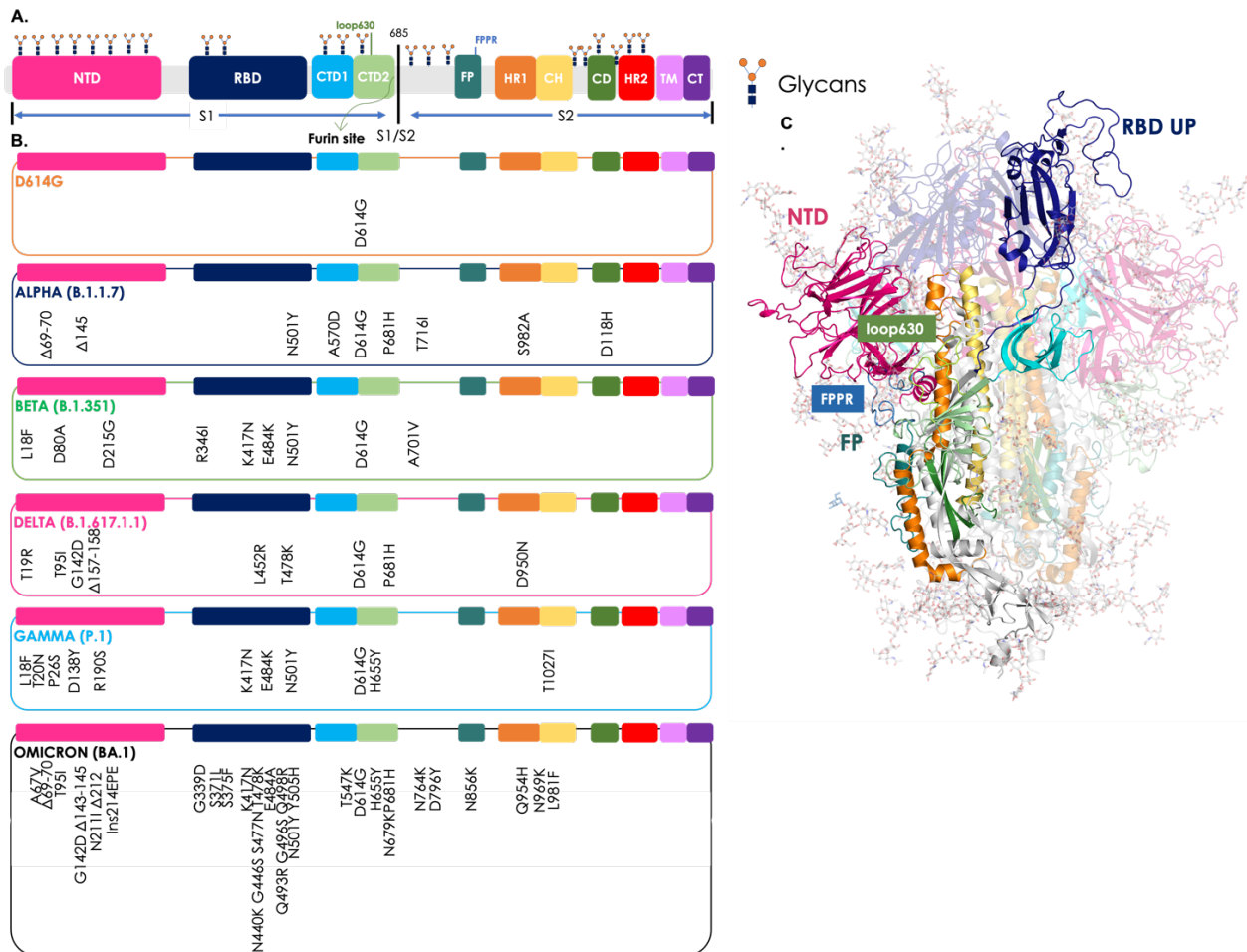


Figure 1. Sequence mutations and structural organization of the Spike protein variants studied. **A.** Colored-block representation of the sequence of the full-length SARS-CoV-2 Spike protein (from PDB ID 6VSB), and its subdivision in the various domains of the S1 and S2 regions: N-terminal domain (NTD, 14-306), receptor binding domain (RBD, 319-528), C-terminal domain 1 (CTD1, 529-591), C-terminal domain 2 (CTD2, 592-686), loop 630 (loop630, 620-640) furin cleavage site (S1/S2), fusion peptide (FP, 788-834), fusion peptide proximal region (FPPR, 828-853), heptad repeat 1 (HR1, 910-984), central helix (CH, 985–1034), connecting domain (CD, 1035-1068). Representative icons (in orange and blue) for glycans in their positions. **B.** Positions of all mutations, deletions (Δ) and insertion (ins) from the amino-acid sequence of Wuhan in the relative domain of

the virus' variants studied. The color code of the VOC is: D614G in orange, Alpha (B.1.1.7) in blue, Beta (B.1.351) in green, Delta (B.1.617.1.1) in pink, Gamma (P.1) in light blue, Omicron (BA.1) in black. **C.** The full-length, fully glycosylated trimeric structure corresponding to pdb code 6VSB. Protomer A (RBD "up"): secondary structures are colored by domain as reported above in point (a); protomers B and C (RBD "down") are in transparency. Glycans' C, N, and O atoms rendered as teal sticks. In the figure the loop630 is shown in light green and FPPR in blue.

1.14.3 Results

Mutations Modulate the Global Internal Dynamics of the Spike Protein Variants.

First, we notice that mutations have been shown to impact on S-protein stability. Comparative experimental characterization of SARS-CoV-2 VOCs identifies two sets of variants: stable proteins (shown to elute as a single peak in SDS–polyacrylamide gel electrophoresis) which comprise the D614G, Beta and Delta variants;^{17, 18, 122} and unstable proteins (shown to elute as two or more peaks, some of which corresponding to aggregated species due to unfolding/misfolding), which entail the WT, Alpha, Gamma, and Omicron.¹²² Interestingly, Omicron represents one of the most unstable species.¹⁹

To explore whether dynamic signatures exist that can be related to the observed trends in stability, we set out to characterize residue-pair Distance Fluctuations (DFs) among all aminoacid-pairs in the various proteins.¹²³⁻¹²⁷ This calculation, which reports the mean-square fluctuation of the inter-residue distance between any two residues in the protein, informs on the effect of sequence variations on the internal dynamics of the protein. In particular, an increase of global internal flexibility (overall decreased pair coordination) can be related to an enhanced tendency to support transitions to states alternative to the native one. In this framework, sequence alterations reverberate in a differential capacity of the protein to populate the native basin.

Given the complexity of the system under exam and the expectedly wide structural variations involved, our aim is not to sample large conformational changes (or even unfolding pathways and mechanisms), but to provide a simple dynamic-based approximation of global stability. Furthermore, DF analysis can potentially highlight substructures and (ensembles of) residues that respond differently to sequence variations.

We first comparatively analyzed the WT vs. the D614G variant.

The overall DF matrices shows the block character typically observed for multidomain proteins, reflecting the alternation of regions of small and large inter-residue distance fluctuations. It is immediately evident that the D614G mutant displays patterns of residue-

pair coordination that are significantly more diffuse than in the case of the WT protein (the Wuhan sequence). Indeed, in D614G higher coordination appears to extend to the whole 3D structure of the protein (**Figure 2**, **Figure 3**). The pervasive enhancement of pair-coordination can contribute to stabilize the protein in the 3D structure of the native state. Breaking-up the extensive networks of low fluctuating residue-pairs in D614G can expectedly require a higher energy contribution than in the case of the native sequence. Extension of the analysis to the Delta variant confirms the trend for more stable proteins to be characterized by more diffuse networks of highly-coordinated residue-pairs. The same trends hold for the final stabilized variant studied herein, namely the Beta variant, B.1.351 (**Figure 2**).

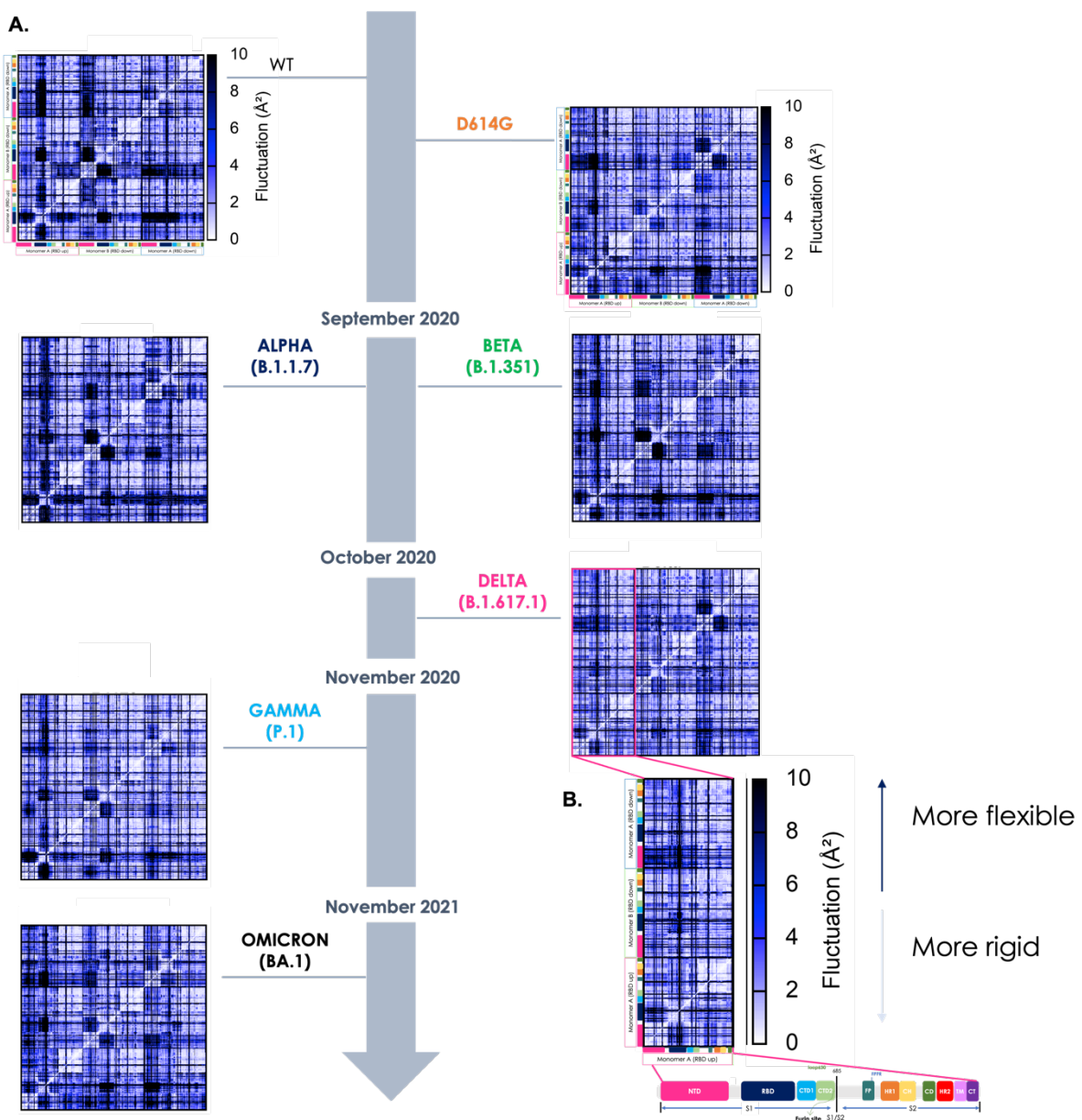


Figure 2. Characterization of the internal dynamics and flexibility of the various Spike mutants in terms of residue-pair fluctuations. **A.** Matrix of residue-pair distance fluctuations (DFs) among all aminoacid-pairs in all the variants. The VOCs are represented on the timeline according to the appearance and colored in their color-code described in **Figure 1**. The x- and y-axes show the sequences colored per domain as in **Figure 1**. Residue pairs with fluctuations between 0 and 2 Å are white, between 2 and 8 Å blue, and larger than 8 Å black. **B.** represent a zoom in on the submatrix of monomer A (with the RBD up): this shows pair fluctuations within monomer A and with the other two monomers (RBD down), specifically for of the Delta variant. The colors (from white to black) represent the intensity of the fluctuation: the clearer the matrix pixel, the more intense the coordination between the aminoacid-pairs and ultimately more rigid the (region of the) protein; the darker the color, the higher the distance fluctuation indicative of lower coordination. The inset also reports a zoom in on the domain partitioning of single Spike protomers.

Analysis of the finer details of the matrices can aptly highlight detailed sequence-dependent modulations of the S-proteins. In this context, of particular interest is the finding that in D614G, Beta and Delta, an increasing coordination with the rest of the protein is observed for the RBD in the “up” conformation, the one required for interaction with human cell.¹²⁸ In this model, mutations induce an overall change in the S-dynamic states that significantly preorganized the protein for recognition of its receptors.

Strikingly, the analysis of DF distributions in the Alpha, Gamma and Omicron, together with WT, shows a trend pointing to increased internal flexibility: larger pair-fluctuations are indeed generally observed. Importantly, the Omicron variant, which experimentally was shown to be one of the least stable and most infective mutants, turned out to be the protein with the larger internal flexibility. Here, the residues of monomer A, in particular, become completely uncoordinated with the rest of the protein.

Interestingly, in all these cases, the RBD in the “up” conformation is seen to maintain similar coordination patterns with the rest of the protein as those observed above for the stabilized mutants.

These data suggest a pattern whereby increased flexibility can be viewed as a double-edged sword. While determining a degree of structural instability, flexibility in general supports the exploration of dynamic states that facilitate conformational conversions. The ability to sample different states eventually increases the probability for displaying the RBDs in the proper orientation for interaction with host-receptor, while at the same time supporting the large conformational rearrangements in the stem region that are required for subsequent membrane fusion.

To provide a more direct structural picture of how global internal dynamics is modulated upon mutation, we set out to calculate the variation in flexibility in the various mutants, using the Wuhan WT protein as a reference. In this context, we carried out a point-by-point subtraction of the DF matrix of each mutant from the matrix of the WT. The resulting difference matrix is further manipulated by calculating the sum of all values in each column: as each column corresponds to one residue, the calculation returns a compact description of the increase or decrease of flexibility for each residue in the mutant with respect to the WT. The data are then projected on the structure as reported in **Figure 3**: a pervasive increase of coordination is clearly observed for Delta, while in contrast a marked increase of flexibility is noticed for Omicron.

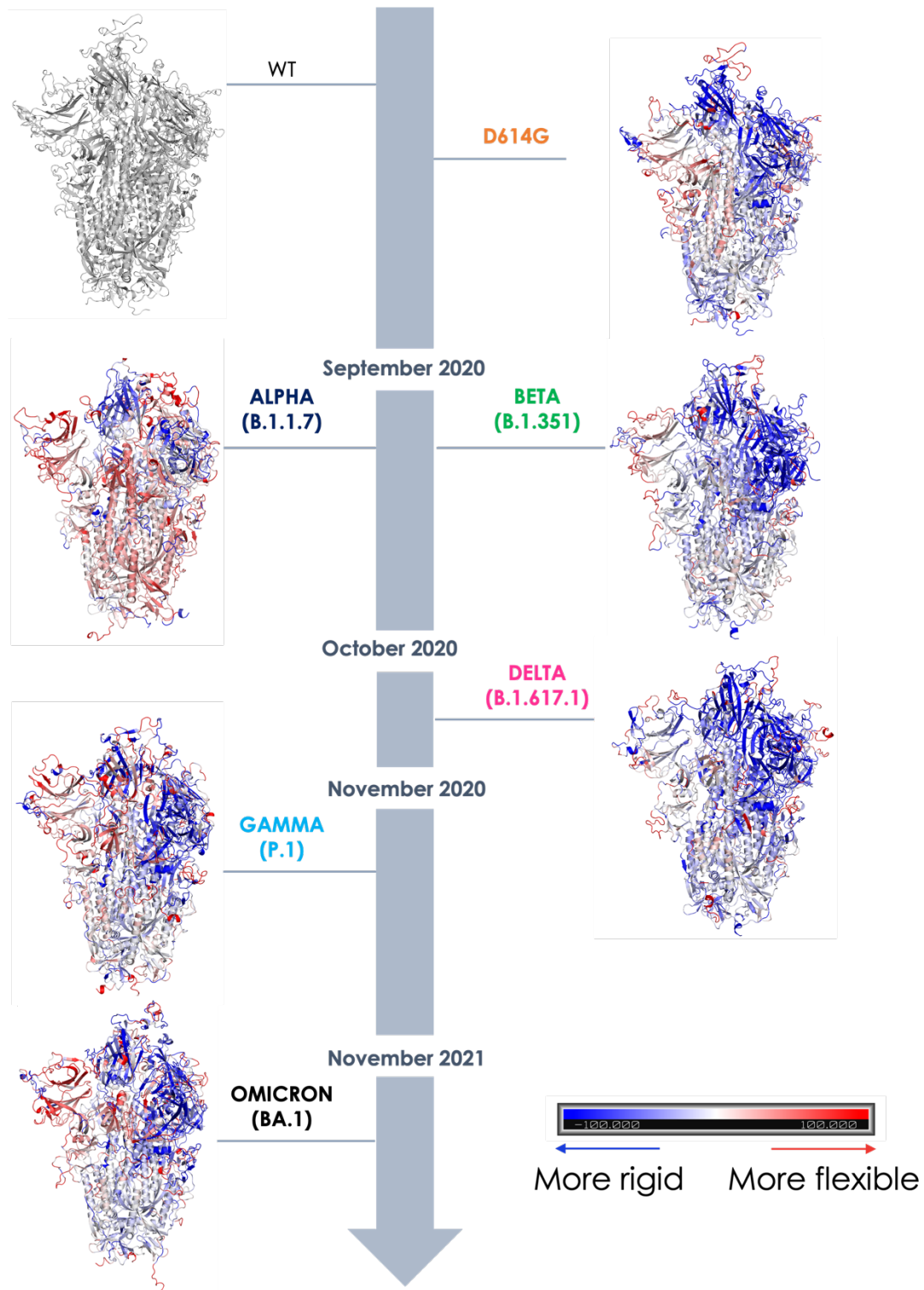


Figure 3. Structural projection of flexibility differences with respect to the WT Wuhan mutant. Point-by-point subtraction of the DF matrix of the WT from the matrix of each variant. The larger the fluctuation value for each residue is for the variant with respect to the WT, the more the domain is colored red. Conversely, the lower the value is in the variant, the more the protein is colored blue: thus blue-colored domains represent areas that are stiffer overall, i.e. where coordination is greater, and this is clearly seen in the more stable variants (D614G, Beta, Delta). Conversely, the appearance

of red areas report on the lower co-ordination and therefore the greater the flexibility of the variants (see Alpha, Gamma and Omicron).

Summarizing, the sequence-dependent modulation of Spike's internal dynamics, characterized in terms of the degree of coordination in residue pairs, can be related to the tendency for the structure to sample alternative dynamic states, while maintaining the RBD preorganized to interact with its host receptors. On the one hand, stabilization of the native structure would be expected to maximize the display of RBD for interaction (a case exemplified by the Delta variant); on the other hand, increased flexibility of the native state, which would aptly lead to destabilization of the structure, could favor the exploration of states that organize the RBD for ACE2-recognition and subsequent structural transitions in the stem region (a case exemplified by the Omicron variant). Finally, our analysis points into the direction of a dynamic behavior of the different variants that appears to follow a (random) stepwise pattern of differentiation similar to that observed for the evolution and selection of mutations.

Machine Learning Classification of Variant Dynamics-Stability Relationships.

The results described above identify distinct internal dynamic profiles of the S-protein as a function of sequence and define a possible link between the degree of coordination and emerging (in)stability in VOCs. However, these results are still qualitative and rely on an attentive critical investigation of the features of the DF matrices. To put the analysis of dynamics, and the possibility to relate them to specific features, on a more quantitative ground, we set out to develop a Machine Learning (ML) approach capable to classify the variants as "STABLE" or "UNSTABLE" simply based on the input of information on internal dynamics. To this end, we resorted to image recognition methods: in this context, the above reported DF matrices are considered as images to classify. The advantage of using the whole DF matrix as an input image is that it compactly reports on the internal dynamics-state of the protein as a whole. It is important to notice that in this framework, small modifications in the sequence that may reverberate in large scale coordination modifications can potentially be efficiently identified.¹²³

We use a Convolutional Neural Networks (CNN) approach. Specifically, we start from the VGG19 model, extensively tested in classification problems and easy to import into in-house Python scripts from the Tensorflow (TF) library.¹²⁹ Moreover, VGG19 shows one of the best compromises between computational cost and accuracy, especially with GPU compiled TF.¹²⁹ We introduced modifications to the VGG19 model to increase the dimensions of the

layers consistent with the pixel number of the input images. The layout of the model is reported in **Figure 4A**.

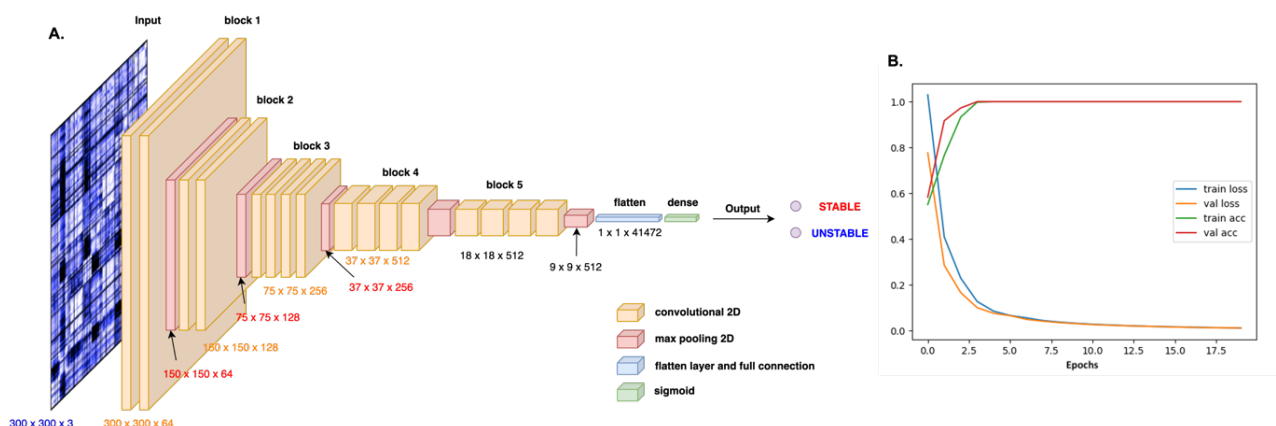


Figure 4. Machine learning approach. (A) The architecture of modified VGG19 model. The pixel image of DF-matrix is transformed in mathematical values which are submitted through convolutional and max pooling layers to reduce the computational cost and overfitting during the model training. Flatten fully-connected layer is used to converge all the data towards the dense output layer where the sigmoid function acts in order to provide the binary output. (B) Performance evaluation during training and validation of the model. Accuracy achieves values close to 1.0 during both training and validation steps, while in both the two cases the loss function decreases towards 0. The combination of these two information assures that the model is well trained and suitable for next evaluation on test dataset.

In our approach, the images depicting the DFs were prepared with a typical resolution of 300x300 pixels and used as input for the multiple layers of the model where they are processed by alternating convolutions and max-pooling operations, until achieving the last flattened and fully-connected layers which provide the final output.

To train the model, DF images from the last equilibrated 200ns of each of the four replicas were used as data sets. Variants were firstly divided in two sets according with the known stability: D614G, Beta, Delta were initially considered as STABLE; WT, Alpha, Gamma and Omicron were labeled UNSTABLE. The final goal of the model is in fact to classify a certain protein as STABLE or UNSTABLE, based *only* on the image of one (or more) respective DF matrices.

To prepare the datasets we extracted a DF image each 10ns and considering the number of replicas for each variant, we end up with a total of 672 images. Starting from this dataset

we operated a manual random separation between test (20%), train (64%) and validation (16%) sets (see Materials and Methods for more details).

Our dataset was thus composed as follows:

- Training set: 430 DF matrix images (215 STABLE and 215 UNSTABLE)
- Validation set: 108 DF matrix images (54 STABLE and 54 UNSTABLE)
- Test set: 134 DF matrix images (67 STABLE and 67 UNSTABLE)

The performance evaluation of our method on training and validation sets are shown in the following **Figure 4B**. In both the cases we can highlight a mutual fit and convergence between validation and training accuracies as well as losses. By evaluating our trained model with the test set, we got 100% of accuracy, taking only 3s to scan all the 134 DF test images. This result is corroborated by the confusion matrix which showed that amongst 134 total images, 67 test entries were correctly classified as STABLE while the other 67 as UNSTABLE. We also calculated the Cohen's kappa coefficient obtaining a κ value equal to 1.

Next, we moved on to feed the model and predictor with new (and completely unseen and unrelated) sets of data. The new set included DF matrices that were calculated on parts of the trajectories that were not used for either the training or testing reported above. In particular, we selected matrices calculated even on the less equilibrated parts of the trajectories, specifically the ones at the beginning of the production. The new set was thus composed by 20 DF matrix images of which 10 came from STABLE variants, while another 10 from UNSTABLE variants. Interestingly, the model was able to predict all the cases with 100% of accuracy in both STABLE and UNSTABLE entries.

The model thus proves able to provide a direct labeling of DF matrices establishing a link between internal dynamics and the property used for classification (stability in this case): from the physical point of view, the model associates a more diffuse and pervasive pattern of internal coordination to the increased stability of the relative protein, speeding up MD analysis and removing human bias in the classification of distinct variants of human proteins.

The Dynamics of RBD and Functional Substructures in Different Mutants.

The above reported analyses indicate that the common trait in the dynamics of all the different variants entails the preorganized presentation of the RBD. Indeed, such motions underlie binding to the human ACE2 and are thus key for viral entry. Here, we focus on the characterization of the dynamics of the RBDs in the different variants.

To this end, we monitored the distributions of two variables that recapitulate the main motions of the RBDs with respect to the rest of the protein: the first is the distance between the centers of mass (COMs) of the spike core and of the RBD (see Materials and Methods); the second is the 3D angle between these two parts of the spike protein (see **Figure 5** and its different subpanels).

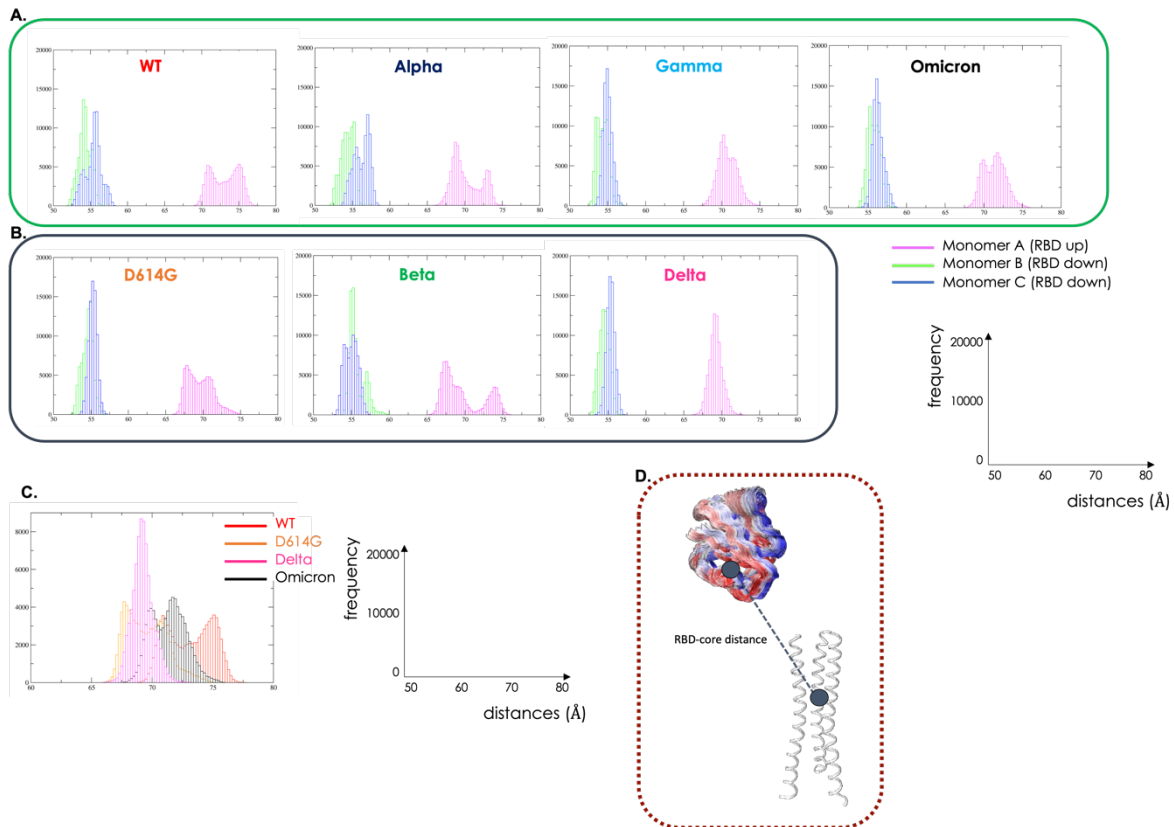


Figure 5. The Dynamics of RBD. In green hues the UNSTABLE variants and in grey hues the STABLE variants. **A.** Distance between the centers of mass (COMs) of the spike core and of the RBD of the UNSTABLE variants (WT, Alpha, Gamma and Omicron) represented in histograms. In magenta the fluctuation of monomer A with the RBD in the UP position, in green the monomer B (RBD down) and in blue the monomer C (RBD down). **B.** Distance between the centers of mass (COMs) of the spike core and of the RBD of the STABLE variants (D614G, Beta and Delta). The graphs report on the x-axis the distances (Å) and on the y-axis the frequencies. **C.** In this panel, we reported the comparisons among the distances in the most representative variants (WT, D614G, Delta and Omicron). **D.** Structural representation reporting a simplified cartoon representation of the variables mentioned.

The COMs distance analysis (**Figure 5A., B.**) shows that there is a tendency for the variant that determine a jump in infectivity, specifically Delta (and to some extent Gamma and Omicron) VOC which then became dominant on the background of existing variants, to have

the RBD in the “up” conformation populating a more restricted part of the conformational landscape.

This tendency can clearly be seen on moving from WT to D614 to the Delta variant (**Figure 5B.**). Strikingly, Delta shows the RBD “up” populating a restricted portion of available configurations, sticking out of the protein in the direction of possible interaction partners. Albeit to a more limited extent, this is observed also for the Omicron variant. In general, and consistent with the results presented above, the dynamics of the very infectious Omicron variant seem to combine the increase in global conformational flexibility that favors functional conformational transitions with an almost optimal ability to present the RBD for targeting human cell receptors. Both aspects are clearly advantageous for the virus.

Similar trends are observed also for the angle-distributions described above (**Figure 5C., D.**).

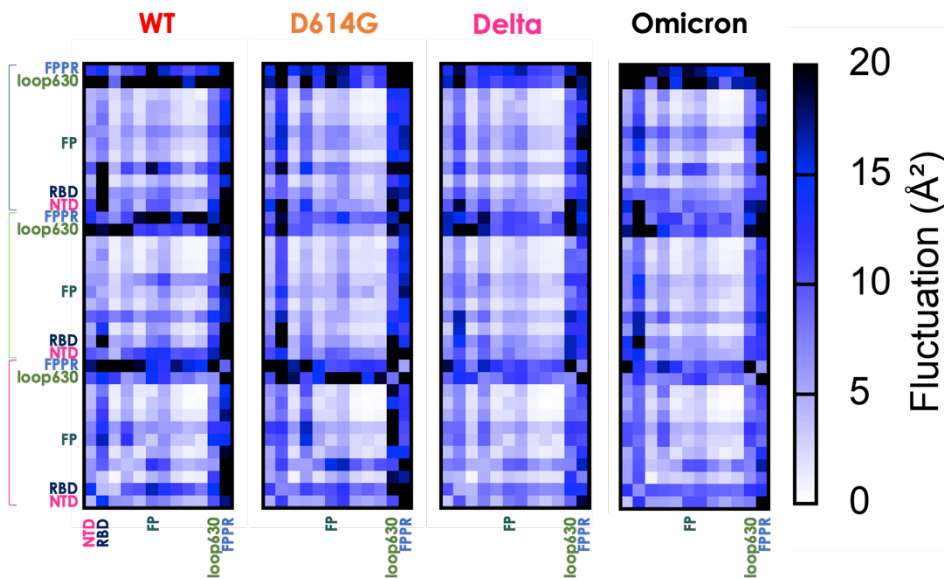
We next moved on to analyze the dynamic behavior of the Fusion Peptide (FP) and of the region proximal to it, namely Fusion Peptide Proximal Region (FPPR). This site is important for the step following attachment to the cell receptor and to prompt the large conformational changes that eventually lead to the Spike-driven membrane fusion.^{14, 130} We used a simplified representation of internal dynamics and coordination patterns, in which the average value of the coordination of all residues within a certain domain with all other substructures is considered. The coordination matrix is thus represented as a simplified block matrix, in which single blocks report on the overall coordination between structurally-defined subdomains. Interestingly, both FP and FPPR turn out to increase their dynamic coordination with the rest of the protein upon moving from the WT to all different variants (**Figure 6A, B.**). Importantly, coordination of these regions is particularly diffuse in the Delta and Omicron variants, indicating that the substructure may be particularly efficient in sensing variations (such as binding to the receptor) at other regions of the protein.

Similar considerations can be applied to loop630, a substructure important for the stabilization of the S-protein in the RBD “up” conformation. This substructure was identified in cryo-EM to fold to an ordered structure on passing from the WT to D614G.¹³¹

In the variants associated with higher stability, we notice a diffuse coordination for loop 630 in monomers B and C with respect to the variants with lower stability. It is interesting to observe here that the starting structure for this loop is disordered for all variants. Interestingly, Omicron shows a peculiar behavior, in agreement with experimental data¹⁹: while the protein is overall more flexible (see above), loop630 is seen to coordinate with the

NTD and CTD1 domains within the same protomer. This also reverberates in the larger amount of ordered secondary structure for loop630 observed for Omicron (**Figure 6B**).

A. Monomer A (RBD up)



B. Omicron

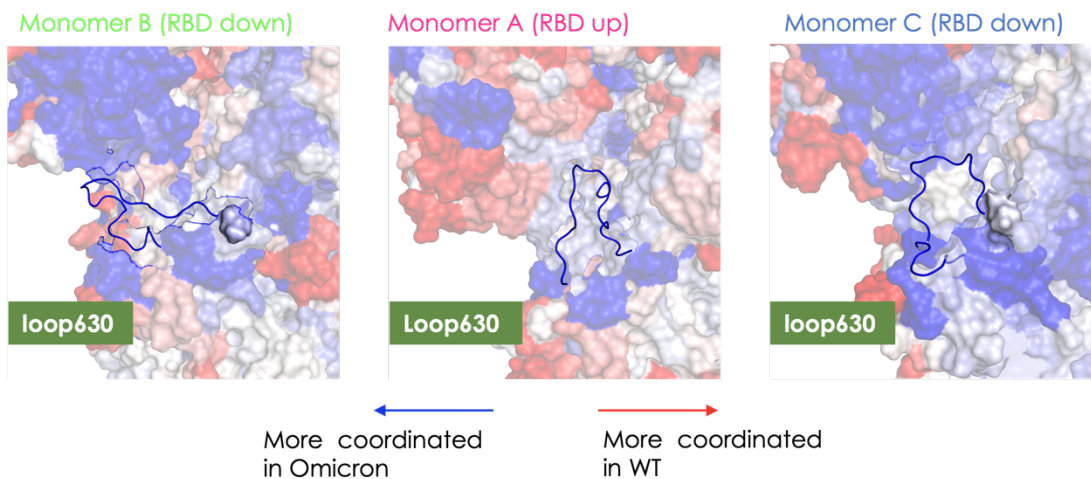


Figure 6. Coordinated motions in terms distance fluctuations of structural sub-blocks. A. the simplified block matrix of the coordination matrix in which single blocks report on the overall coordination between structurally-defined subdomains. We report here only the DF in blocks for monomer A (RBD up) of WT, D614G, Delta and Omicron. The matrix is divided considering all the domains of the Spike protein (reported in **Figure 1**). Specifically, NTD, RBD, FP, loop630 and FPPR are highlighted using the color-code of **Figure 1**.

B. Loop630 on the Omicron Spike protein: loop630 is a segment which seems to be important for the stabilization of the S-protein in the RBD UP conformation. This substructure was identified in cryo-EM to fold to an ordered structure on passing from the WT to D614G. The figure reports the point-by-point subtraction of the DF matrix of Omicron variant from the matrix of the WT, as defined in **Figure 3**. Considering almost all of it blue, it can be noted that the loop is certainly much more

coordinated in this variant than in original strand, supporting the importance of this loop for the stability of the configuration with one RBD in the UP position (configuration that can be related to infectivity).

1.14.4 Discussion

In this paper, we carried out an extensive analysis of different mutants of the SARS-CoV-2 Spike protein. Our aim here is not to thoroughly sample large-scale conformational changes through MD simulations (which given the number and complexity of the systems under exam is out of reach) but to shed light on the traits of microscopic dynamics, determined by sequence changes, that can be related to the modulation of motions and differences the native state dynamics of the mutants. Importantly, the aforementioned modulations can be revealed even in the absence of major conformational changes. In this context, we note that nanosecond/microsecond timescale residue fluctuations and modulation of protein flexibility have been linked in other cases to the regulation of protein stabilities and activities.¹³²⁻¹³⁶

Based on our results, we developed a model relating S-proteins' internal dynamic traits to the sequence modification paths followed by distinct VOCs during their evolution in the first couple of years of diffusion of the virus.

Interestingly, we notice that at the global level (whole protein), the dynamics of the different variants appear to change following the path of somewhat random evolution that characterizes the underlying sequences. In other words, following the time-line of emergence of the various VOCs, one would expect a stepwise modulation of the structural dynamics of their respective S-proteins. However, as noted by Bloom and colleagues (The New York Times “We Study Virus Evolution. Here’s Where We Think the Coronavirus Is Going.” March 28, 2022. <https://www.nytimes.com/interactive/2022/03/28/opinion/coronavirus-mutation-future.html>), at least initially (2.5 years on the evolutionary scale of a virus can conceivably be considered an early stage) viral evolution selected advantageous sequences through big jumps. Consistent with sequences, the structural dynamics of the S-protein appear to follow this trend.

In terms of advantage to spread and survive the challenges of an increasingly trained human immune system, which aims to get rid of the SARS-CoV-2 virus, these random changes could help the virus better escape acquired immunity, while maintaining (or increasing) its ability to interact with host cell receptors. Such mechanisms can also provide

the virus with an efficient way to scan for sequences with convenient functionally-oriented motions. Indeed, a linear, steady-state evolution would be less efficient at exploring the sequence landscape, potentially limiting the capacity to overcome extensive vaccination.

Our dynamics-based results show that the internal coordination of the S-protein can be reconnected to its degree of stability or instability: analyzing pair-distance fluctuations we found specific patterns of extensive residue-pair coordination, particularly pervasive of the whole protein in the variants (D614G, Beta and Delta) that are experimentally shown to be more stable (these variants elute as single peaks in SDS-page gel electrophoresis).^{18, 19, 122, 131}

In the case of Omicron, Beta and Gamma, as well as in the WT, a more globally uncoordinated and generally flexible dynamics is observed, which may be considered as a factor favoring structural instability. In terms of viral evolution and diffusion, both of these two aspects can be advantageous: increased coordination/stability guarantees persistence of the protein in the active structure in the environment; flexibility, on the other hand, would support a more efficient scan of conformations among which the ones able to recognize and bind ACE2 (and/or other human receptors) can be selected. Extensive flexibility and increased instability could also facilitate the large structural rearrangements of the S-protein required to drive the fusion of the membranes of the host and virus.

In this framework, it is also important to underline that specific functional substructures share the same dynamic traits throughout all variants: these include the motions of the Receptor Binding Domain (RBD), the Fusion Peptide (FP) and the region preceding it (FPPR), as well as loop630, whose motions stabilize the display of the RBD in the active conformation.^{14, 130, 131}

On the basis of our internal dynamics analyses, we also developed a Machine Learning classification method that allows us to label variants based on a visual representation of their dynamics, automatically reconnecting sequences with biophysical properties.

Overall, we propose a model whereby the jumps in sequence evolution that have characterized the first years of SARS-CoV-2 diffusion are reflected in the variations of the microscopic native dynamics of the encoded S-proteins. In this model, the events of Spike dynamics modification are not sequential and deterministic. A critical feature of our model is that, while we observe a direct coupling between the motions of the RBD, the FP and FPPR (hinting to a conserved conformational preorganization of these functionally fundamental substructures), the global dynamics of the rest of the protein appears to rearrange to provide increased stability or increased flexibility.

These two factors can be considered alternative mechanisms to favor S-protein/ACE2 interactions, being at the same time convenient and advantageous.

Our approach may represent a means to characterizing the dynamic properties of different forms of S-protein from distinct VOCs: dynamics are modulated by sequence modifications but retain the traits necessary for the selection of conformational states that favor receptor recognition and binding. The ML model can conveniently intervene in the classification of potentially emerging new variants as STABLE or UNSTABLE, linking this property to the ability of the protein to guide viral-host recognition and infection. Together with other approaches based on sequence analysis, evolutionary investigations, and the application of different studies of structure-function relationships, this could enrich our knowledge of the physico-chemical determinants of evolution of certain protein forms, relating them to their functions in the context of viral diffusion. While based on the case of SARS-CoV-2 Spike protein, our models and considerations are fully general, and readily transferable to other targets and contexts.

1.14.5 Materials and Methods

Preparation of Spike Protein Variants

Fully glycosylated S protein variants simulated in this work were variously derived from simulations described by Grant *et al.*¹³⁷ based on the Cryo-EM structure of the WT S protein at PDB entry 6VSB¹²⁸, wherein one RBD is in the “up” conformation and the other two are “down”. All the variants’ mutations are introduced as discussed before using the “mutations wizard” in the *PyMOL* molecular modeling package (Schrodinger LLC).

The same method for deletions was used to model deletions of Delta (del157-158) and Omicron (del69-70, del143-145, del212).

In the case of Omicron there is also an insertion of three new amino acids (214EPE). This was modelled again using Pymol by inserting the three amino acids into the sequence and then relaxing the system with a 400-step preminimization cycle *in vacuo* (200 steepest-descent + 200 conjugate gradient), using the *AMBER* platform’s *sander* utility (version 18)¹¹⁰, in which harmonic positional restraints ($k = 5.0 \text{ kcal mol}^{-1} \text{ \AA}^{-2}$) are applied to all atoms except those in the five residues on either side of the insertion.

MD Simulation Details

After preparation, glycosylated S protein structures are solvated in a cuboidal box of TIP3P water molecules using *AMBER*’s *tleap* tool; where necessary, Na⁺ or Cl⁻ ions are added

accordingly to neutralize the charge. *N*-glycosylated asparagines and oligosaccharides are treated using the *GLYCAM-06j* forcefield¹³⁸, whereas ions are modeled with parameters by Joung and Cheatham.¹³⁹ To all other (protein) atoms, we apply the *ff14SB* forcefield¹⁴⁰. Starting structures and topologies for all simulated variants are electronically provided.

On each glycosylated S protein variant, we conduct 4 independently replicated atomistic molecular dynamics simulations (MD), using the *AMBER* package (version 18): each replica consists of two 300-step rounds of minimization, 2.069 ns preproduction, and 1 μ s production. The *sander* MD engine¹¹⁰ is used into the earlier stages of preproduction; thereafter, we switch to the GPU-accelerated *pmemd.cuda*.¹¹⁰

Details on MD preproduction

Prior to the production stage, every independent MD replica for every S variant goes through a series of preproduction steps, namely: minimization, solvent equilibration, system heating, and equilibration. The first two are conducted using the *sander* utility, after which the GPU-accelerated *pmemd.cuda* is invoked instead.

Minimization takes place in two 300-step rounds, the first 10 of which use the steepest-descent algorithm and the last 290 conjugate gradient. In the first round, we only minimize backbone H α and H1 hydrogens on aminoacids and monosaccharides, respectively, restraining all other atoms harmonically ($k = 5.0 \text{ kcal mol}^{-1} \text{ \AA}^{-2}$). Thereafter, all atoms are released, including solvent and ions.

Solvent equilibration occurs over 9 ps with a time step of 1 fs; the ensemble is *NVT*, with temperatures in this case enforced by the Berendsen thermostat¹⁴¹. Positions of non-solvent atoms are harmonically restrained ($k = 10 \text{ kcal mol}^{-1} \text{ \AA}^{-2}$). Solvent molecules are assigned initial random velocities to match a temperature of 25 K. Fast heating to 400 K (coupling: 0.2 ps) is performed over the first 3 ps; the solvent is then retained at 400 K for another 3 ps; and cooled back down to 25 K over the last 3 ps, more slowly (coupling: 2.0). The cutoff for determining Lennard-Jones and Coulomb interactions remains at 8.0 \AA for this and all subsequent stages, as does the Particle Mesh Ewald method⁸⁸ to determine Coulomb interactions beyond this cutoff. *SHAKE* constraints⁷⁰ are not applied at this stage but are always present thereafter.

For system heating, the time step is increased to 2 fs and, whilst continuing in the *NVT* ensemble, temperatures are now enforced by the Langevin thermostat⁷¹ (which remains in place for all subsequent stages). With an initial collision frequency of 0.75 ps^{-1} , the system

is heated from 25 to 300 K over 20 ps: all atoms are free to move except aminoacids' C α atoms, which are positionally restrained with $k = 5 \text{ kcal mol}^{-1} \text{ \AA}^{-2}$.

For equilibration, the ensemble is switched to NpT ($p = 1 \text{ atm}$; Berendsen barostatcoupling: 1 ps), and the system is simulated for a further 2040 ps. The thermostat's collision frequency is kept lower than in the production stage (1 ps^{-1}). Restraints on C α atoms are lifted gradually: $k = 3.75 \text{ kcal mol}^{-1} \text{ \AA}^{-2}$ for the first 20 ps; $1.75 \text{ kcal mol}^{-1} \text{ \AA}^{-2}$ for the following 20 ps; none thereafter.

Details on MD production

The 1 μs production stage is carried out in the NpT ensemble ($T = 300 \text{ K}$; $p = 1 \text{ atm}$) using a 2 fs time step; a cutoff of 8.0 \AA is applied for the calculation of Lennard-Jones and Coulomb interactions alike. Coulomb interactions beyond this limit are computed using the Particle Mesh Ewald method⁶⁸. All bonds containing hydrogen are restrained using the *SHAKE* algorithm¹⁴². Constant pressure is enforced *via* Berendsen's barostat⁹⁶ with a 1 ps relaxation time, whereas temperature is stabilized by Langevin's thermostat¹⁴¹ with a 5 ps^{-1} collision frequency.

Residue-pair distance fluctuations (DFs)

To understand the impact mutations on the internal dynamics of SARS-CoV-2 we conducted the distance fluctuation analysis.

To compute the matrix of distance fluctuations, we used the 4 μs metatrajectory available for each studied system, obtained by concatenating the MD replicas of each specific protein: in this framework, each element of the matrix corresponds to the DF parameters:

$$DF_{ij} = \langle (d_{ij} - \langle d_{ij} \rangle)^2 \rangle$$

Where d_{ij} is the time-depended distance of the C α atoms of amino acids i and j and the brackets indicate the time-average over the trajectory. The advance of this parameter is its invariant nature under translations and rotations of the molecules and, different from the covariance matrix, does not depend on the choice of a particular protein reference structure.

DF was calculated for every pair of residues during the trajectory. This parameter characterizes residues that move in a coordinated fashion, and it is actually able to reflect the presence of specific coordination patterns and quasi-rigid domains motion in the protein of interest. In particular, pairs of amino acids belonging to the same quasi-rigid domain or highly coordinated at a distance are associated with small distance fluctuations and vice versa.

Distance fluctuations (DFs) in BLOCKS

To further analyze the coordination patterns among distinct subdomains of the protein, we first subdivided the structure into domains, also called blocks in our definition, according to the annotation reported in **Figure 6A**. Here we evaluate the degree of interdomain coordination among different blocks and the contribution of each single block to the overall internal dynamics of the protein. DF for each domain (block) is calculated from the full DF matrices reported above. The latter is in fact simplified by combining the contributions of residues assigned to a certain domain (block) based on the sequence definition from **Figure 6A**. The cumulative DF value associated with each block is then obtained by averaging all the terms for each residue grouped in the block.

Difference between Distance fluctuations matrices

To further compare fluctuation matrices, we calculated the difference matrix, obtainable by subtracting the matrix for one particular protein from the DF matrix of Wuhan (WT) molecule, used as a reference for all such calculations. To account for sequence differences due to deletions and insertions, we simply considered the DF matrices of all the common structures among the proteins, to obtain matrices of the same dimensions. The values of the various difference matrices, reporting on how the internal dynamics of a variant changes with respect to the WT, are then summed by column: the obtained parameter reports on the increased or decreased global coordination of the residue corresponding to that column, with respect to the WT. The parameter is then projected with using the color code reported in **Figure 3** on the 3D structure.

RBD fluctuations

To follow the fluctuations of the RBD during the simulation, we focused sampling along a two-dimensional progress coordinate: 1) the difference in the center of mass of the spike core to the RBD (distances parameter) and 2) the angle defined by these two regions of the Spike protein (angles).

Distances

We used the CPPTRAJ and the command distances (<https://amberhub.chpc.utah.edu/distance/>) to calculate the distances between the center of mass of atoms in “mask1” to atoms in “mask2”. The atoms in “mask1” are the atoms of the RBD and the “mask2” includes residues of the core of the Spike (849-881, 945-1045 of each protomer of the protein).

Angles

To construct the angle between these two masks we used first the vector command of CPPTRAJ (<https://amberhub.chpc.utah.edu/vector/>) to keep track of a vector value (and its origin) of each mask over the trajectory and after we perform the vector product (<https://amberhub.chpc.utah.edu/vectormath/>) to get the angles between the two previously calculated vectors (using the option “dotangle” to calculate angle from dot-product between the two vectors; vectors will be normalized).

CNN-ML

Preparing DF-images: The trajectories from the MD-simulations were directly submitted to the DF-matrix calculation using the above reported procedure. Specifically, we extracted the DF each 10ns starting from the very first not equilibrated ones, till the last of the dynamics. We ended up with a total number of 2816 DF-matrices. We then used an in-house developed *Gnuplot* script to prepare the images with a dimension of 300X300 pixels using a white-blue-black color palette. Colors tending towards white indicates DF of $\sim 0 \text{ \AA}^2$, while black ones indicate DF of $\sim 10 \text{ \AA}^2$. The halfway point (i.e., DF of $\sim 5 \text{ \AA}^2$) is represented with blue nuances.

Preparing the CNN-model: Image recognition through Convolutional Neural Networks (CNN) was elaborated using a modified version of the readily available VGG19 model, since it demonstrated to be one of the best compromises between computational cost and accuracy and can be directly imported in Python using Tensorflow (TF).¹²⁹

The architecture of VGG19 model was maintained unaltered, while we modified the dimensions of layers in order to accommodate the 300x300 pixels of the input DF-image. Furthermore, the imported images were again rescaled to the dimension of the VGG19 layers and normalized according to the standard pixel values which can range from 0 to 255. This step aims to exclude possible scaling errors introduced during the *Gnuplot* image preparation from the numerical matrix.

We set the classification mode to ‘binary’ (class_mode) and the number of samples propagated through the network was set to 32 (batch_size). To provide a measure for goodness of the method we used *ImageNet* (For ImageNet see: <https://image-net.org/about.php>) weights as widely recognized to be a standard for images classification problems.

The last layer of our VGG19 modified model provides the prediction output using a single layer on which the ‘sigmoid’ activation function $\sigma(z)$ acts:

$$\sigma(z) = \frac{1}{1 + e^z}$$

We selected this function since it is the standard for binary classifications: given its existence only between 0 and 1, it constitutes the natural choice for binary problems. We compiled the model by using a ‘binary cross-entropy’ loss function and using the ‘Adam’ algorithm for the stochastic optimization.¹⁴³ Lastly, in order to avoid model overfitting, we introduced an early stopping monitor which stops training the model if the validation loss starts increasing during five consecutive epochs. However, we never experienced strong increases in propagation of loss function to justify the intervention of the monitor. Moreover, the specific placement of the five max pooling layers reduces the computation time and memory usage, by limiting also the probability to get into overfitting issues.

Training of the model and test with internal data: To train the model, we selected the DF-images coming from the last equilibrated 200ns of each replica, for a total number of 672 images, which were manually divided between test (20%), train (64%) and validation (16%) sets. Within these sets we operated a manual classification in order to define the two main classes of interest in our model: STABLE variants (Beta, D614G, Delta, and Delta⁺) and UNSTABLE variants (WT, Omicron, Alpha, Gamma). We trained our model for 20 epochs using the datasets and we obtained complete training and validation in 53 seconds, with an average of 153 ms/step. This result is extremely promising and is mainly due to the TF parallelization of using GPU.

We next tested the just trained model with data arriving from the same equilibrated portion of dynamics, but not used during the train and validation steps. We got 100% of accuracy, taking only 3s to scan all the 134 DF test images. This result was also checked through classification report and confusion matrix analysis, in order to validate the goodness of the predictions.

Test of model with external data: We submitted to the trained model a new dataset prepared by taking 20 unseen DF matrix images from molecular dynamics replicas of the variants involved in this study and never used for the previous training of the model. We selected 10 images coming from the STABLE variants and 10 from the UNSTABLE ones. The manual choice we operated was specifically directed in order to choose within the first non-equilibrated parts of each trajectory. The aim was to prove that – once the model is trained – our method can be extended to other new variants without the need to use long MD simulations. Those images were firstly submitted to the same scaling and normalization steps as performed for the other sets of data (see above). The only difference we introduced

was on the number of samples propagated through the network, which was set to 1 (batch_size) since we need to predict each submitted image. Moreover, according with the just printed classification report, we were able to assign the UNSTABLE class if the prediction assumes values above 0.5, while if below the STABLE class was inferred.

Again, the test on external data was extremely fast and only took 3s to complete all the 20 classifications, with a final accuracy of 100% for each of them.

1.14.6 References

- (1) Wrapp, D.; Wang, N.; Corbett, K. S.; Goldsmith, J. A.; Hsieh, C. L.; Abiona, O.; Graham, B. S.; McLellan, J. S. Cryo-EM structure of the 2019-nCoV spike in the prefusion conformation. *Science* **2020**, *367* (6483), 1260-1263. DOI: 10.1126/science.abb2507.
- (2) Walls, A. C.; Park, Y. J.; Tortorici, M. A.; Wall, A.; McGuire, A. T.; Veerles, D. Structure, Function, and Antigenicity of the SARS-CoV-2 Spike Glycoprotein. *Cell* **2020**, *183* (6), 1735. DOI: 10.1016/j.cell.2020.11.032.
- (3) Iacob, S.; Iacob, D. G. SARS-CoV-2 Treatment Approaches: Numerous Options, No Certainty for a Versatile Virus. *Front Pharmacol* **2020**, *11*, 1224. DOI: 10.3389/fphar.2020.01224.
- (4) Zhang, J.; Zeng, H.; Gu, J.; Li, H.; Zheng, L.; Zou, Q. Progress and Prospects on Vaccine Development against SARS-CoV-2. *Vaccines (Basel)* **2020**, *8* (2). DOI: 10.3390/vaccines8020153.
- (5) Zhang, Z.; Shen, Q.; Chang, H. Vaccines for COVID-19: A Systematic Review of Immunogenicity, Current Development, and Future Prospects. *Front Immunol* **2022**, *13*, 843928. DOI: 10.3389/fimmu.2022.843928.
- (6) Scarabel, L.; Guardascione, M.; Dal Bo, M.; Toffoli, G. Pharmacological strategies to prevent SARS-CoV-2 infection and treat the early phases of COVID-19. *Int J Infect Dis* **2021**, *104*, 441-451. DOI: 10.1016/j.ijid.2021.01.035.
- (7) Owen, D. R.; Allerton, C. M. N.; Anderson, A. S.; Aschenbrenner, L.; Avery, M.; Berritt, S.; Boras, B.; Cardin, R. D.; Carlo, A.; Coffman, K. J.; et al. An oral SARS-CoV-2 M. *Science* **2021**, *374* (6575), 1586-1593. DOI: 10.1126/science.abl4784.
- (8) Dellus-Gur, E.; Toth-Petroczy, A.; Elias, M.; Tawfik, D. S. What makes a protein fold amenable to functional innovation? Fold polarity and stability trade-offs. *J Mol Biol* **2013**, *425* (14), 2609-2621. DOI: 10.1016/j.jmb.2013.03.033.
- (9) Wellner, A.; Raites Gurevich, M.; Tawfik, D. S. Mechanisms of protein sequence divergence and incompatibility. *PLoS Genet* **2013**, *9* (7), e1003665. DOI: 10.1371/journal.pgen.1003665.
- (10) Tóth-Petróczy, A.; Tawfik, D. S. The robustness and innovability of protein folds. *Curr Opin Struct Biol* **2014**, *26*, 131-138. DOI: 10.1016/j.sbi.2014.06.007.
- (11) Yanagida, H.; Gispan, A.; Kadouri, N.; Rozen, S.; Sharon, M.; Barkai, N.; Tawfik, D. S. The Evolutionary Potential of Phenotypic Mutations. *PLoS Genet* **2015**, *11* (8), e1005445. DOI: 10.1371/journal.pgen.1005445.
- (12) Pecetta, S.; Pizza, M.; Sala, C.; Andreano, E.; Pileri, P.; Troisi, M.; Pantano, E.; Manganaro, N.; Rappuoli, R. Antibodies, epicenter of SARS-CoV-2 immunology. *Cell Death Differ* **2021**, *28* (2), 821-824. DOI: 10.1038/s41418-020-00711-w.
- (13) Andreano, E.; Rappuoli, R. SARS-CoV-2 escaped natural immunity, raising questions about vaccines and therapies. *Nat Med* **2021**, *27* (5), 759-761. DOI: 10.1038/s41591-021-01347-0.
- (14) Bloom, J. D.; Labthavikul, S. T.; Otey, C. R.; Arnold, F. H. Protein stability promotes evolvability. *Proc Natl Acad Sci U S A* **2006**, *103* (15), 5869-5874. DOI: 10.1073/pnas.0510098103.
- (15) Harvey, W. T.; Carabelli, A. M.; Jackson, B.; Gupta, R. K.; Thomson, E. C.; Harrison, E. M.; Ludden, C.; Reeve, R.; Rambaut, A.; Peacock, S. J.; et al. SARS-CoV-2 variants, spike mutations and immune escape. *Nat Rev Microbiol* **2021**, *19* (7), 409-424. DOI: 10.1038/s41579-021-00573-0.
- (16) Mittal, A.; Khattry, A.; Verma, V. Structural and antigenic variations in the spike protein of emerging SARS-CoV-2 variants. *PLoS Pathog* **2022**, *18* (2), e1010260. DOI: 10.1371/journal.ppat.1010260.
- (17) Serapian, S. A.; Marchetti, F.; Triveri, A.; Morra, G.; Meli, M.; Moroni, E.; Sautto, G. A.; Rasola, A.; Colombo, G. The Answer Lies in the Energy: How Simple Atomistic Molecular Dynamics Simulations May Hold the Key to Epitope Prediction on the Fully Glycosylated SARS-CoV-2 Spike Protein. *J Phys Chem Lett* **2020**, *11* (19), 8084-8093. DOI: 10.1021/acs.jpcclett.0c02341.
- (18) Triveri, A.; Serapian, S. A.; Marchetti, F.; Doria, F.; Pavoni, S.; Cinquini, F.; Moroni, E.; Rasola, A.; Frigerio, F.; Colombo, G. SARS-CoV-2 Spike Protein Mutations and Escape from Antibodies: A Computational Model of Epitope Loss in Variants of Concern. *J Chem Inf Model* **2021**, *61* (9), 4687-4700. DOI: 10.1021/acs.jcim.1c00857.
- (19) Scarabelli, G.; Morra, G.; Colombo, G. Predicting interaction sites from the energetics of isolated proteins: a new approach to epitope mapping. *Biophys J* **2010**, *98* (9), 1966-1975. DOI: 10.1016/j.bpj.2010.01.014.
- (20) Fan, Y.; Li, X.; Zhang, L.; Wan, S.; Zhou, F. SARS-CoV-2 Omicron variant: recent progress and future perspectives. *Signal Transduct Target Ther* **2022**, *7* (1), 141. DOI: 10.1038/s41392-022-00997-x.
- (21) Gobeil, S. M.; Henderson, R.; Stalls, V.; Janowska, K.; Huang, X.; May, A.; Speakman, M.; Beaudoin, E.; Manne, K.; Li, D.; et al. Structural diversity of the SARS-CoV-2 Omicron spike. *Mol Cell* **2022**, *82* (11), 2050-2068. DOI: 10.1016/j.molcel.2022.03.028.
- (22) Stalls, V.; Lindenberger, J.; Gobeil, S. M.; Henderson, R.; Parks, R.; Barr, M.; Deyton, M.; Martin, M.; Janowska, K.; Huang, X.; et al. Cryo-EM structures of SARS-CoV-2 Omicron BA.2 spike. *Cell Rep* **2022**, *39* (13), 111009. DOI: 10.1016/j.celrep.2022.111009.
- (23) Mannar, D.; Saville, J. W.; Zhu, X.; Srivastava, S. S.; Berezuk, A. M.; Tuttle, K. S.; Marquez, A. C.; Sekirov, I.; Subramaniam, S. SARS-CoV-2 Omicron variant: Antibody evasion and cryo-EM structure of spike protein-ACE2 complex. *Science* **2022**, *375* (6582), 760-764. DOI: 10.1126/science.abn7760.

- (24) Chen, J.; Wang, R.; Gilby, N. B.; Wei, G. W. Omicron Variant (B.1.1.529): Infectivity, Vaccine Breakthrough, and Antibody Resistance. *J Chem Inf Model* **2022**, *62* (2), 412-422. DOI: 10.1021/acs.jcim.1c01451.
- (25) Luo, R.; Delaunay-Moisan, A.; Timmis, K.; Danchin, A. SARS-CoV-2 biology and variants: anticipation of viral evolution and what needs to be done. *Environ Microbiol* **2021**, *23* (5), 2339-2363. DOI: 10.1111/1462-2920.15487.
- (26) McCallum, M.; Czudnochowski, N.; Rosen, L. E.; Zepeda, S. K.; Bowen, J. E.; Walls, A. C.; Hauser, K.; Joshi, A.; Stewart, C.; Dillen, J. R.; et al. Structural basis of SARS-CoV-2 Omicron immune evasion and receptor engagement. *Science* **2022**, *375* (6583), 864-868. DOI: 10.1126/science.abn8652.
- (27) Huang, Y.; Yang, C.; Xu, X. F.; Xu, W.; Liu, S. W. Structural and functional properties of SARS-CoV-2 spike protein: potential antiviral drug development for COVID-19. *Acta Pharmacol Sin* **2020**, *41* (9), 1141-1149. DOI: 10.1038/s41401-020-0485-4.
- (28) Wall, E. C.; Wu, M.; Harvey, R.; Kelly, G.; Warchal, S.; Sawyer, C.; Daniels, R.; Hobson, P.; Hatipoglu, E.; Ngai, Y.; et al. Neutralising antibody activity against SARS-CoV-2 VOCs B.1.617.2 and B.1.351 by BNT162b2 vaccination. *Lancet* **2021**, *397* (10292), 2331-2333. DOI: 10.1016/S0140-6736(21)01290-3.
- (29) Haas, E. J.; Angulo, F. J.; McLaughlin, J. M.; Anis, E.; Singer, S. R.; Khan, F.; Brooks, N.; Smaja, M.; Mircus, G.; Pan, K.; et al. Impact and effectiveness of mRNA BNT162b2 vaccine against SARS-CoV-2 infections and COVID-19 cases, hospitalisations, and deaths following a nationwide vaccination campaign in Israel: an observational study using national surveillance data. *Lancet* **2021**, *397* (10287), 1819-1829. DOI: 10.1016/S0140-6736(21)00947-8.
- (30) Hsieh, C. L.; Goldsmith, J. A.; Schaub, J. M.; DiVenere, A. M.; Kuo, H. C.; Javanmardi, K.; Le, K. C.; Wrapp, D.; Lee, A. G.; Liu, Y.; et al. Structure-based design of prefusion-stabilized SARS-CoV-2 spikes. *Science* **2020**, *369* (6510), 1501-1505. DOI: 10.1126/science.abd0826.
- (31) Wang, X.; Du, Z.; Johnson, K. E.; Pasco, R. F.; Fox, S. J.; Lachmann, M.; McLellan, J. S.; Meyers, L. A. Effects of COVID-19 Vaccination Timing and Risk Prioritization on Mortality Rates, United States. *Emerg Infect Dis* **2021**, *27* (7), 1976-1979. DOI: 10.3201/eid2707.210118.
- (32) Zhang, L.; Jackson, C. B.; Mou, H.; Ojha, A.; Peng, H.; Quinlan, B. D.; Rangarajan, E. S.; Pan, A.; Vanderheiden, A.; Suthar, M. S.; et al. SARS-CoV-2 spike-protein D614G mutation increases virion spike density and infectivity. *Nat Commun* **2020**, *11* (1), 6013. DOI: 10.1038/s41467-020-19808-4.
- (33) Zhou, B.; Thao, T. T. N.; Hoffmann, D.; Taddeo, A.; Ebert, N.; Labroussaa, F.; Pohlmann, A.; King, J.; Steiner, S.; Kelly, J. N.; et al. SARS-CoV-2 spike D614G change enhances replication and transmission. *Nature* **2021**, *592* (7852), 122-127. DOI: 10.1038/s41586-021-03361-1.
- (34) Zhang, J.; Cai, Y.; Xiao, T.; Lu, J.; Peng, H.; Sterling, S. M.; Walsh, R. M.; Rits-Volloch, S.; Zhu, H.; Woosley, A. N.; et al. Structural impact on SARS-CoV-2 spike protein by D614G substitution. *Science* **2021**, *372* (6541), 525-530. DOI: 10.1126/science.abf2303.
- (35) Bhattacharya, M.; Chatterjee, S.; Sharma, A. R.; Agoramorthy, G.; Chakraborty, C. D614G mutation and SARS-CoV-2: impact on S-protein structure, function, infectivity, and immunity. *Appl Microbiol Biotechnol* **2021**, *105* (24), 9035-9045. DOI: 10.1007/s00253-021-11676-2.
- (36) Grabowski, F.; Preibisch, G.; Giziński, S.; Kocharczyk, M.; Lipniacki, T. SARS-CoV-2 Variant of Concern 202012/01 Has about Twofold Replicative Advantage and Acquires Concerning Mutations. *Viruses* **2021**, *13* (3). DOI: 10.3390/v13030392.
- (37) Tegally, H.; Wilkinson, E.; Giovanetti, M.; Iranzadeh, A.; Fonseca, V.; Giandhari, J.; Doolabh, D.; Pillay, S.; San, E. J.; Msomi, N.; et al. Detection of a SARS-CoV-2 variant of concern in South Africa. *Nature* **2021**, *592* (7854), 438-443. DOI: 10.1038/s41586-021-03402-9.
- (38) Voloch, C. M.; da Silva Francisco, R.; de Almeida, L. G. P.; Cardoso, C. C.; Brustolini, O. J.; Gerber, A. L.; Guimarães, A. P. C.; Mariani, D.; da Costa, R. M.; Ferreira, O. C.; et al. Genomic characterization of a novel SARS-CoV-2 lineage from Rio de Janeiro, Brazil. *J Virol* **2021**, *95* (10). DOI: 10.1128/JVI.00119-21.
- (39) Mlcochova, P.; Kemp, S. A.; Dhar, M. S.; Papa, G.; Meng, B.; Ferreira, I. A. T. M.; Datir, R.; Collier, D. A.; Albecka, A.; Singh, S.; et al. SARS-CoV-2 B.1.617.2 Delta variant replication and immune evasion. *Nature* **2021**, *599* (7883), 114-119. DOI: 10.1038/s41586-021-03944-y.
- (40) Bowen, J. E.; Park, Y. J.; Stewart, C.; Brown, J. T.; Sharkey, W. K.; Walls, A. C.; Joshi, A.; Sprouse, K. R.; McCallum, M.; Tortorici, M. A.; et al. SARS-CoV-2 spike conformation determines plasma neutralizing activity elicited by a wide panel of human vaccines. *Sci Immunol* **2022**, eadf1421. DOI: 10.1126/sciimmunol.adf1421.
- (41) Sang Hoon Kim University of North Carolina at Chapel Hill , F. L. K. U. o. C., San Diego , Mia A. Rosenfeld University of California, San Diego , Lane Votapka University of California, San Diego , Lorenzo Casalino University of California, San Diego , Micah Papanikolas University of North Carolina at Chapel Hill , Rommie E. Amaro University of California, San Diego , Ronit Freeman University of North Carolina at Chapel Hill. Positively bound: Remapping of Increased Positive Charge Drives SARS-CoV-2 Spike Evolution to Optimize its Binding to Cell Surface Receptors. *ChemRxiv* **Oct 19, 2022** DOI: 10.26434/chemrxiv-2022-dmqq3.

- (42) Spinello, A.; Saltalamacchia, A.; Borišek, J.; Magistrato, A. Allosteric Cross-Talk among Spike's Receptor-Binding Domain Mutations of the SARS-CoV-2 South African Variant Triggers an Effective Hijacking of Human Cell Receptor. *J Phys Chem Lett* **2021**, *12* (25), 5987-5993. DOI: 10.1021/acs.jpcllett.1c01415.
- (43) Yang, Y.; Zhang, Y.; Qu, Y.; Zhang, C.; Liu, X. W.; Zhao, M.; Mu, Y.; Li, W. Key residues of the receptor binding domain in the spike protein of SARS-CoV-2 mediating the interactions with ACE2: a molecular dynamics study. *Nanoscale* **2021**, *13* (20), 9364-9370. DOI: 10.1039/d1nr01672e.
- (44) Ma, S.; Li, H.; Yang, J.; Yu, K. Molecular simulation studies of the interactions between the human/pangolin/cat/bat ACE2 and the receptor binding domain of the SARS-CoV-2 spike protein. *Biochimie* **2021**, *187*, 1-13. DOI: 10.1016/j.biochi.2021.05.001.
- (45) Nguyen, H. L.; Thai, N. Q.; Nguyen, P. H.; Li, M. S. SARS-CoV-2 Omicron Variant Binds to Human Cells More Strongly than the Wild Type: Evidence from Molecular Dynamics Simulation. *J Phys Chem B* **2022**, *126* (25), 4669-4678. DOI: 10.1021/acs.jpcc.2c01048.
- (46) Sofia F Oliveira, A.; Shoemark, D. K.; Avila Ibarra, A.; Davidson, A. D.; Berger, I.; Schaffitzel, C.; Mulholland, A. J. The fatty acid site is coupled to functional motifs in the SARS-CoV-2 spike protein and modulates spike allosteric behaviour. *Comput Struct Biotechnol J* **2022**, *20*, 139-147. DOI: 10.1016/j.csbj.2021.12.011.
- (47) Shoemark, D. K.; Colenso, C. K.; Toelzer, C.; Gupta, K.; Sessions, R. B.; Davidson, A. D.; Berger, I.; Schaffitzel, C.; Spencer, J.; Mulholland, A. J. Molecular Simulations suggest Vitamins, Retinoids and Steroids as Ligands of the Free Fatty Acid Pocket of the SARS-CoV-2 Spike Protein*. *Angew Chem Int Ed Engl* **2021**, *60* (13), 7098-7110. DOI: 10.1002/anie.202015639.
- (48) Gupta, K.; Toelzer, C.; Williamson, M. K.; Shoemark, D. K.; Oliveira, A. S. F.; Matthews, D. A.; Almuqrin, A.; Staufer, O.; Yadav, S. K. N.; Borucu, U.; et al. Structural insights in cell-type specific evolution of intra-host diversity by SARS-CoV-2. *Nat Commun* **2022**, *13* (1), 222. DOI: 10.1038/s41467-021-27881-6.
- (49) Verkhivker, G. M.; Di Paola, L. Integrated Biophysical Modeling of the SARS-CoV-2 Spike Protein Binding and Allosteric Interactions with Antibodies. *J Phys Chem B* **2021**, *125* (18), 4596-4619. DOI: 10.1021/acs.jpcc.1c00395.
- (50) Verkhivker, G. M.; Agajanian, S.; Oztas, D. Y.; Gupta, G. Landscape-Based Mutational Sensitivity Cartography and Network Community Analysis of the SARS-CoV-2 Spike Protein Structures: Quantifying Functional Effects of the Circulating D614G Variant. *ACS Omega* **2021**, *6* (24), 16216-16233. DOI: 10.1021/acsomega.1c02336.
- (51) Casalino, L.; Gaieb, Z.; Goldsmith, J. A.; Hjorth, C. K.; Dommer, A. C.; Harbison, A. M.; Fogarty, C. A.; Barros, E. P.; Taylor, B. C.; McLellan, J. S.; et al. Beyond Shielding: The Roles of Glycans in the SARS-CoV-2 Spike Protein. *ACS Cent Sci* **2020**, *6* (10), 1722-1734. DOI: 10.1021/acscentsci.0c01056.
- (52) Sztain, T.; Ahn, S. H.; Bogetti, A. T.; Casalino, L.; Goldsmith, J. A.; Seitz, E.; McCool, R. S.; Kearns, F. L.; Acosta-Reyes, F.; Maji, S.; et al. A glycan gate controls opening of the SARS-CoV-2 spike protein. *Nat Chem* **2021**, *13* (10), 963-968. DOI: 10.1038/s41557-021-00758-3.
- (53) Zhang, J.; Xiao, T.; Cai, Y.; Lavine, C. L.; Peng, H.; Zhu, H.; Anand, K.; Tong, P.; Gautam, A.; Mayer, M. L.; et al. Membrane fusion and immune evasion by the spike protein of SARS-CoV-2 Delta variant. *Science* **2021**, *374* (6573), 1353-1360. DOI: 10.1126/science.abl9463.
- (54) Cai, Y.; Zhang, J.; Xiao, T.; Lavine, C. L.; Rawson, S.; Peng, H.; Zhu, H.; Anand, K.; Tong, P.; Gautam, A.; et al. Structural basis for enhanced infectivity and immune evasion of SARS-CoV-2 variants. *Science* **2021**, *373* (6555), 642-648. DOI: 10.1126/science.abi9745.
- (55) Zhang, J.; Cai, Y.; Lavine, C. L.; Peng, H.; Zhu, H.; Anand, K.; Tong, P.; Gautam, A.; Mayer, M. L.; Rits-Volloch, S.; et al. Structural and functional impact by SARS-CoV-2 Omicron spike mutations. *Cell Rep* **2022**, *39* (4), 110729. DOI: 10.1016/j.celrep.2022.110729.
- (56) Morra, G.; Potestio, R.; Micheletti, C.; Colombo, G. Corresponding functional dynamics across the Hsp90 Chaperone family: insights from a multiscale analysis of MD simulations. *PLoS Comput Biol* **2012**, *8* (3), e1002433. DOI: 10.1371/journal.pcbi.1002433.
- (57) Corrada, D.; Morra, G.; Colombo, G. Investigating allostery in molecular recognition: insights from a computational study of multiple antibody-antigen complexes. *J Phys Chem B* **2013**, *117* (2), 535-552. DOI: 10.1021/jp310753z.
- (58) Paladino, A.; Morra, G.; Colombo, G. Structural Stability and Flexibility Direct the Selection of Activating Mutations in Epidermal Growth Factor Receptor Kinase. *J Chem Inf Model* **2015**, *55* (7), 1377-1387. DOI: 10.1021/acs.jcim.5b00270.
- (59) Rehn, A.; Moroni, E.; Zierer, B. K.; Toppel, F.; Morra, G.; John, C.; Richter, K.; Colombo, G.; Buchner, J. Allosteric Regulation Points Control the Conformational Dynamics of the Molecular Chaperone Hsp90. *J Mol Biol* **2016**, *428* (22), 4559-4571. DOI: 10.1016/j.jmb.2016.09.014.
- (60) Triveri, A.; Sanchez-Martin, C.; Torielli, L.; Serapian, S. A.; Marchetti, F.; D'Acerno, G.; Pirota, V.; Castelli, M.; Moroni, E.; Ferraro, M.; et al. Protein Allostery and Ligand Design: Computational Design Meets Experiments to Discover Novel Chemical Probes. *J Mol Biol* **2022**, *434* (17), 167468. DOI: 10.1016/j.jmb.2022.167468.

- (61) Karen Simonyan , A. Z. Very Deep Convolutional Networks for Large-Scale Image Recognition. *arxiv* **2015**. DOI: <https://doi.org/10.48550/arXiv.1409.1556>.
- (62) Pal, D. Spike protein fusion loop controls SARS-CoV-2 fusogenicity and infectivity. *J Struct Biol* **2021**, *213* (2), 107713. DOI: 10.1016/j.jsb.2021.107713.
- (63) Grant, O. C.; Montgomery, D.; Ito, K.; Woods, R. J. Analysis of the SARS-CoV-2 spike protein glycan shield reveals implications for immune recognition. *Sci Rep* **2020**, *10* (1), 14991. DOI: 10.1038/s41598-020-71748-7.
- (64) *AMBER 2018*; 2018. (accessed).
- (65) Kirschner, K. N.; Yongye, A. B.; Tschampel, S. M.; González-Outeiriño, J.; Daniels, C. R.; Foley, B. L.; Woods, R. J. GLYCAM06: a generalizable biomolecular force field. *Carbohydrates. J Comput Chem* **2008**, *29* (4), 622-655. DOI: 10.1002/jcc.20820.
- (66) Joung, I. S.; Cheatham, T. E. Determination of alkali and halide monovalent ion parameters for use in explicitly solvated biomolecular simulations. *J Phys Chem B* **2008**, *112* (30), 9020-9041. DOI: 10.1021/jp8001614.
- (67) Maier, J. A.; Martinez, C.; Kasavajhala, K.; Wickstrom, L.; Hauser, K. E.; Simmerling, C. ff14SB: Improving the Accuracy of Protein Side Chain and Backbone Parameters from ff99SB. *J Chem Theory Comput* **2015**, *11* (8), 3696-3713. DOI: 10.1021/acs.jctc.5b00255.
- (68) Loncharich, R. J.; Brooks, B. R.; Pastor, R. W. Langevin dynamics of peptides: the frictional dependence of isomerization rates of N-acetylalanyl-N'-methylamide. *Biopolymers* **1992**, *32* (5), 523-535. DOI: 10.1002/bip.360320508.
- (69) Tom Darden; Darrin York; Pedersen, L. Particle mesh Ewald: An $N \cdot \log(N)$ method for Ewald sums in large systems. *J. Chem. Phys.* **1993**, *98*, 10089 DOI: <https://doi.org/10.1063/1.464397>.
- (70) Shuichi Miyamoto ; Kollman, P. A. Settle: An analytical version of the SHAKE and RATTLE algorithm for rigid water models. *Journal of Computational Chemistry* **1992**, *13* (8), 952-962
- (71) Berendsen, H. J. C.; Postma, J. P. M.; van Gunsteren, W. F.; Di Nola, A.; Haak, J. R. Molecular dynamics with coupling to an external bath. *Journal of Chemical Physics* **1984**, *81* (8), 3684-3690. DOI: <https://doi.org/10.1063/1.448118>.
- (72) Diederik P. Kingma, J. B. Adam: A Method for Stochastic Optimization. In 3rd International Conference for Learning Representations, San Diego; 2015.

Drug design on the Spike protein

Indeed, the fascination with the Spike protein extends well beyond its role in understanding and developing vaccines or monoclonal therapies. It has emerged as a pivotal focal point in drug design, signifying a key objective in the pursuit of drug development therapies.

The S prominence as the initial point of contact between the virus and host cells makes it a prime target for therapeutic interventions. Efforts to design molecules that can interact with the spike protein aim to disrupt its function, prevent viral entry into host cells, inhibit its structural changes, or neutralize its ability to evade the immune system. The ultimate goal is to develop potent antiviral drugs that can be used in the treatment of COVID-19 and potentially future coronavirus outbreaks.

A captivating illustration of this *pan*-coronavirus approach lies in the discovery of a free fatty acid binding pocket (FFBP) embedded within the SARS-CoV-2 Spike protein. Within each protomer, this pocket resides at the interface between two adjacent Receptor Binding Domains (RBDs), presenting an intriguing avenue for potential therapeutic intervention because this pocket could regulate the conformational changes of the Spike protein forcing the DOWN configuration of the RBD which is the inactive one. The revelation of this pocket emerged during the early stages of the pandemic through cryo-electron microscopy (cryo-EM) analysis of the SARS-CoV-2 Spike protein's structure. Subsequent experimental validation confirmed the presence of a fatty acid akin in weight to linoleic acid (LA) within this cryptic pocket.

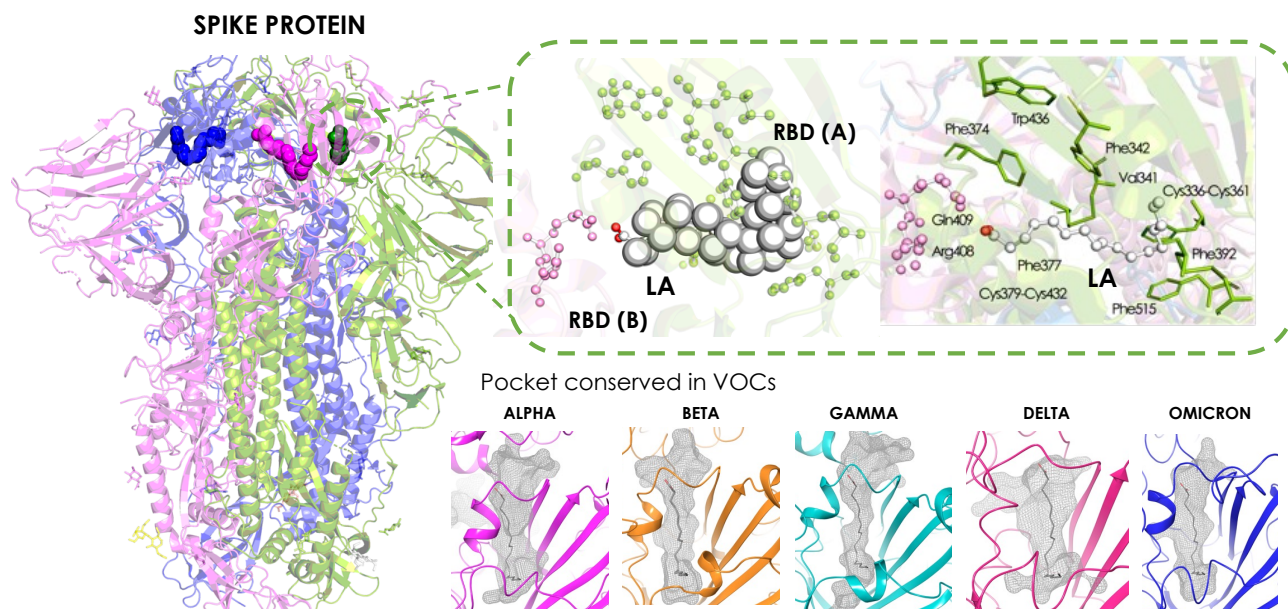
It's worth noting that the cryptic nature of this pocket implies its visibility only within specific conformational ensembles. These ensembles, transient in nature, might not be readily evident from a static 3D protein structure. The FFBP thus emerges as a potential target for drug development, emphasizing the importance of exploring diverse avenues in the ongoing quest for effective therapeutics against evolving viral threats.

Additionally, it's worth highlighting this binding pocket is conserved not only in SARS-CoV-2 but also in its predecessors, SARS-CoV and MERS-CoV. This points to a potentially fundamental and conserved structural feature across various coronaviruses, underlining its significance and potential as a therapeutic target. But, from our computational simulations we have also suggested the persistence of this pocket across all the Variants of Concern. The pocket's consistent presence in these variants, which have garnered significant attention due to their impact on transmissibility and immune evasion, emphasizes its potential as a stable and druggable site.

This remarkable conservation across different coronaviruses and variants reinforces the promising prospect of targeting the FFBP for therapeutic purposes, providing a potential avenue for the development of broad-spectrum antiviral drugs effective against multiple coronaviruses. It underscores the significance of advancing research in this direction to explore and harness the therapeutic potential offered by this intriguing structural feature.

1.15 Binding pocket conserved in the VOCs SARS-CoV-2 spike protein

“Drug Design for a conserved cryptic pocket in SARS-CoV-2 Spike protein variants”



1.15.1 Abstract

Our study assessed the druggability of the free fatty acid binding pocket as a potential target for antiviral drugs against SARS-CoV-2. Using molecular dynamics, we devised a minimization protocol enabling simulation of the opening of the free fatty acid binding pocket in the context of the D614G variant's molecular dynamics, creating space for ligand re-docking. Screening across three compound libraries identified promising compounds. Fragments strategically placed within the pocket are undergoing evaluation for drug design. Notably, some compounds maintained stability even with certain mutations. Preliminary results demonstrated successful binding of fragments and enhanced activity when combined.

Also, molecular dynamics revealed specific frames in variants (Alpha, Beta, Gamma) satisfying prerequisites for ligand binding, suggesting their potential as drug targets. Despite challenges, our findings support further exploration of this promising binding site.

1.15.2 Introduction

In the past few years, a great deal of effort has been put into developing therapies to counteract the virus. In the general introductory section (Chapter 3.4.1), we highlighted a range of therapies that have been developed and approved for treating SARS-CoV-2

infection. These therapies encompass various approaches, among these there is the use of small molecules including Antiviral Drugs such as Paxlovid (nirmatrelvir and ritonavir), Lagevrio (molnupiravir), and Veklury (containing remdesivir, an inhibitor of viral RNA polymerase). These drugs target specific components like the RNA-dependent RNA polymerase, pivotal for generating new viral RNA copies, and viral cysteine proteases, namely 3CLpro or Mpro, and papain-like cysteine protease (PLpro). These proteases play a crucial role in breaking down polyproteins translated from the viral genome into essential non-structural proteins needed for packaging the nascent virion and supporting viral replication. By inhibiting these proteases, viral replication is hindered. Notably, these proteases are less prone to mutation, rendering the development of antiviral treatments an appealing strategy against COVID-19.

Another category of molecules consists of Immune Modulators, drugs designed to activate, enhance, or suppress immune functions. Particularly in COVID-19 cases, immune hyperactivity can exacerbate the disease, and these modulators serve to mitigate this hyperinflammatory response. FDA-approved Immune Modulators include Kineret (anakinra, an Interleukin-1 receptor antagonist), Olumiant (baricitinib), Actemra (tocilizumab), and Gohibic (vilobelimab). (<https://www.fda.gov/drugs/emergency-preparedness-drugs/coronavirus-covid-19-drugs>).

It's important to note that the focus of drug development isn't limited to these targets, and extensive efforts have been dedicated to creating compounds that interact with diverse targets.

A fascinating example is the identification of a free fatty acid binding pocket (FFBP) found within the Spike protein.¹ Each protomer features one of these pockets at the interface between two adjacent RBDs. The existence of this pocket was initially observed during the early stages of the pandemic through the cryo-EM structure of the SARS-CoV-2 Spike protein. Subsequent experiments involving liquid chromatography coupled with mass spectrometry (LC-MS) confirmed the presence of a fatty acid similar in weight to linoleic acid (LA) within this pocket.¹

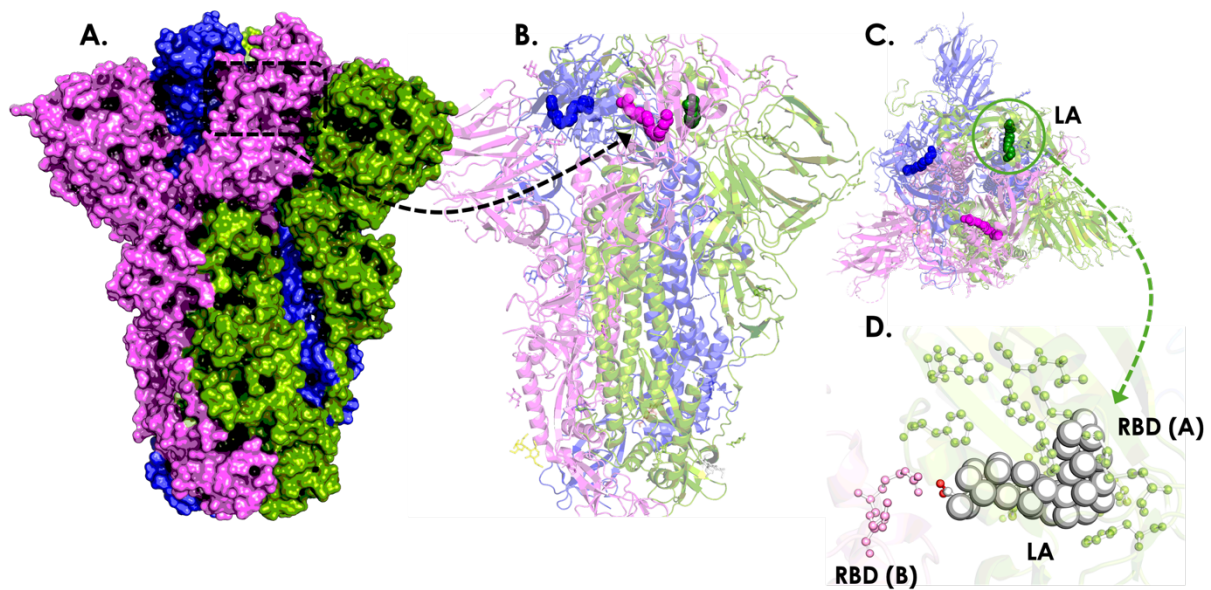


Figure 1. **A.** and **B.** side view **C.** top view of the experimental cryo-EM Spike protein bound to three linoleic acid molecules (represented with spheres). **D.** Free fatty acid binding pocket (FFBP): the residues from the RBDs involved in the interactions with the linoleic acid (LA) (shown in spheres) are highlighted and represented with balls and sticks representation, in light green the residue from the RBD in which the LA is located and in pink from the adjacent monomer.

This pocket has been described as "cryptic" meaning that it's discernible only within specific conformational ensembles. Such ensembles, though transiently formed, might not be immediately apparent from the static 3D protein structure. Specifically, to accommodate linoleic acid (LA), four essential features must coexist. The pocket primarily comprises hydrophobic residues (like Phe, Val, and Leu) enabling LA accommodation but to stabilize LA's negatively charged carboxylic acid head, charged residues are crucial. Through the cryo-EM structure (PDB: 6ZB5), it's evident that adjacent protomer residues Arg408 and Gln409 interact with LA's charged component, serving as "anchor residues".¹

Notably, these anchor residues are initially positioned 10 Å away from the binding pocket in the apo structure. Consequently, for Arg408 and Gln409 to effectively secure LA within the pocket, the receptor-binding domain must draw nearer to the binding pocket.

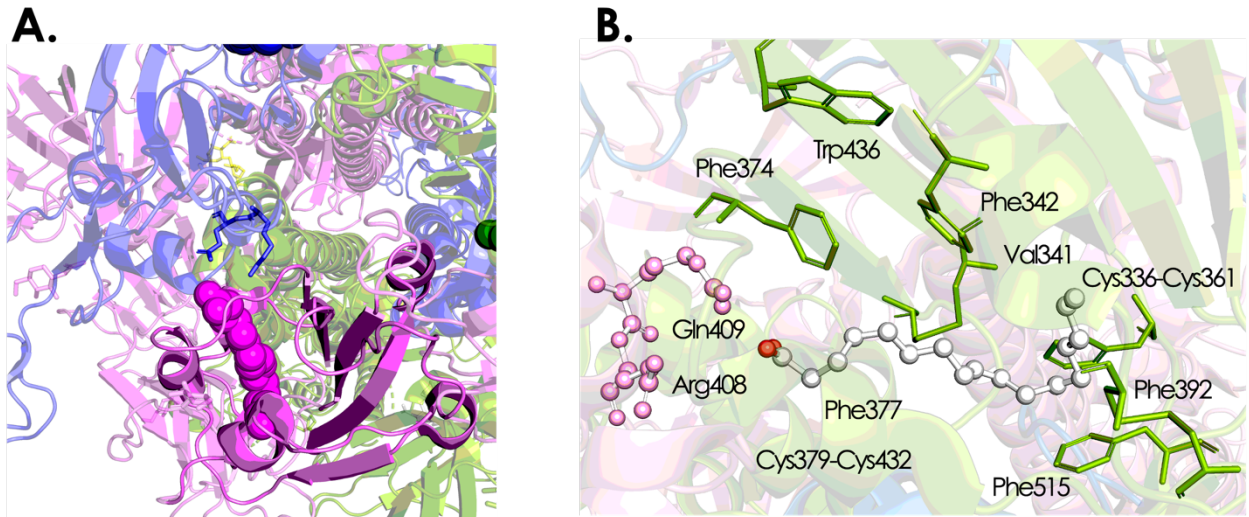


Figure 2. A. Free fatty acid binding pocket (FFBP): the residues from the RBDs involved in the interactions with the linoleic acid (LA) (shown in spheres) are highlighted and represented with sticks representation (in blue the “anchor residues”). **B.** LA interactions with amino acids in the binding pocket. The acidic LA headgroup is in the vicinity of an arginine (Arg408) and a glutamine (Gln409). Hydrophobic LA binding pocket in a sticks representation in green.

Additionally, comparing the LA-bound structure with previous SARS-CoV-2 apo S structures in the closed conformation unveils a gating helix situated at the pocket entrance. This helix, spanning Tyr365 to Tyr369, shifts by approximately 6 Å upon LA binding, causing the pocket to open. This structural shift leads to the compaction of the trimer architecture within the region formed by the three RBDs, resulting in a locked S structure.

A pivotal observation is that, in the presence of linoleic acid, about 70% of the cryo-EM revealed Spike proteins are in the closed conformation. This contrasts with the prior state where approximately 60 to 75% of S trimers were open. This shift could be attributed to linoleic acid stabilizing the closed conformation, possibly leading to reduced binding to the ACE2 receptor. Interest in this pocket stems from the stabilizing influence of LA on the closed state.¹

Docking experiments explored whether natural compounds, such as vitamins and retinoids, could interact beneficially within the pocket.² Docking scores indicate that some of these molecules might potentially bind to the fatty acid pocket, behaving similarly to LA in stabilizing the closed state. A research group conducted molecular dynamics simulations involving S protein complexed with LA, cholesterol (known to bind the spike protein), and dexamethasone (a corticosteroid anti-inflammatory). In the closed state simulations, all three molecules remained bound as expected. In the open conformation, differing outcomes

were observed at distinct binding sites. While cholesterol and dexamethasone displayed less stability than linoleic acid, the overall results suggest the potential druggability of the free fatty acid pocket and theoretically, molecules binding to this pocket could stabilize the closed state, preventing RBD-ACE2 receptor interactions.²

Importantly, this pocket is conserved in other human coronaviruses (HCoVs) like SARS-CoV and MERS-CoV. In MERS-CoV, while the conserved charged residues R408/Q409 are absent, N501 and K502 could potentially serve as alternative anchor points.^{3, 4}

Also, in a recent study, this pocket has been studied employed dynamical-nonequilibrium molecular dynamics (D-NEMD) simulations⁵⁻⁸. This investigation demonstrated the allosteric connection of the FA site with key functional elements involved in membrane fusion or antigenic epitopes.^{4, 9} These simulations elucidated that the absence of linoleic acid (LA) at the fatty acid (FA) sites triggers substantial structural changes at a distance, affecting the receptor-binding motif (RBM), N-terminal domain (NTD), furin cleavage site, and the regions surrounding the fusion peptide (FP). Furthermore, the D-NEMD simulations shed light on distinct allosteric and dynamic behaviors observed in the WT compared to the D614G. Later, they employ the same D-NEMD simulations to investigate and analyze the response of four spike variants (Alpha (B.1.1.7), Delta (B.1.617.2), Delta Plus (B1.617.2-AY1) and Omicron BA.1 (B.1.1.529) variants respect to the original spike to the removal of LA. This research reveals substantial variations in the allosteric response to fatty acid binding among SARS-CoV-2 variants. These distinctions hold notable functional implications concerning the regulation of viral infectivity via LA. Additionally, these findings may impact endeavors aimed at targeting the FA site using natural compounds, repurposed drugs, or ligands specifically designed for this purpose.

So, here we ask if this FFBP, even if it is transient, can be observed in long-time scale simulations of the fully glycosylated D614G Spike protein without the ligand and in all the other Variants of Concern (VOCs). Additionally, we aim to determine whether this observed pocket meets the specific requirements for binding with fatty acids.

To accomplish this, first we identify suitable pockets within the Spike protein than resemble the essential features and then we formulate an efficient selection protocol. The primary goal here is to replicate the experimental positioning of linoleic acid as seen in the reference structure (PDB 6ZB5) and this involves understanding the precise binding pose and interactions between LA and the pocket.

Then, we use the stereochemical information derived from the pocket to guide high throughput virtual screening (HTVS) (<https://www.schrodinger.com/htvs-hit-finding-and->

[evaluation-course-page](#)) of small molecule and fragment libraries. This step aims to identify potential compounds or fragments that can interact favorably with the identified pocket, possibly mimicking the stabilizing effects of linoleic acid on the Spike protein's closed conformation.

Our primary objective in this study is to identify molecules that can establish favorable interactions resembling those formed by linoleic acid within the Spike protein. To achieve this, we developed a protocol that was applied to both the D614G variant and a variant with the additional K417 mutation, which is particularly noteworthy within the FFBP among variants of concern. Furthermore, we extended our investigation to different SARS-CoV-2 variants, including the Wild type, Alpha (B.1.1.7), Beta (B.1.351), Gamma (P.1), Delta (B.1.617.2), Delta+ (B.1.617.2.1), Mu (B.1.621), and Omicron (B.1.1.529). We explored the existence of the free fatty acid binding pocket using a combination of molecular dynamics simulations and structural representatives from the most populated conformational clusters for each variant mentioned.

Our protocol facilitated the docking of a substantial number of molecules into the linoleic acid pocket in the D614G variant. From these docking studies, we identified certain small molecules or fragments that ranked highly in terms of their potential to interact with the binding pocket. Moreover, we are currently conducting experimental testing to determine if these selected fragments are indeed capable of binding to the Spike protein inside the binding pocket.

Additionally, we successfully located the Free Fatty Acid Binding Pocket in some of the SARS-CoV-2 variants. Given the expected conservation of this pocket across different variants, further investigation into this potential target holds promise for future research.

In summary, this section aims to develop a computational approach to identify cryptic pocket from MD simulations and use this understanding to design molecules or fragments that can mimic linoleic acid's interactions within the Spike protein, offering potential avenues for drug development or therapeutic intervention against various SARS-CoV-2 variants.

1.15.3 Results

Pocket investigation in the D614G variant

In the pursuit to evaluate the feasibility of targeting the Spike protein and to investigate the potential value of the linoleic acid (LA) pocket defined in the previously described crystal structure (PDB 6ZB5),¹ we initiated an analysis of extensive molecular dynamics (MD) simulations. This MD simulations were performed without the ligand in the pocket to prove the effective presence of the pocket, even if only transient.

Initially, our approach involved identifying snapshots within the MD simulation of the D614G SARS-CoV-2 fully glycosylated Spike protein that closely resembled the reference structure (PDB 6ZB5). To achieve this, we employed a straightforward metric, the root-mean-square distance (RMSD), which quantifies the average spatial disparity between atoms or residues when comparing two superimposed protein structures. First, the trajectory has been aligned trajectories following the procedure described in the clustering methodology in the Materials and Methods section and subsequently, our focus centered on the side chains of Arg408 and Gln409, for which we computed the RMSD values.

From the MD simulation, approximately 120 frames exhibited an RMSD value less than 2 Å. Among these, we pinpointed the frame displaying the lowest RMSD, measuring at 1.4553 Å. This frame was identified as the closest match to the crystallized Spike protein structure containing LA. This meticulous selection process enables us to establish a basis for further investigations and analyses, laying the groundwork for assessing the potential druggability of the Spike protein and the viability of targeting the LA pocket for therapeutic interventions. Then, to create a complex between LA and the selected MD simulation frame, we employed a stepwise approach. First, we aligned the chosen MD simulation frame with the reference Spike protein-LA complex using PyMol¹⁰. Then, we added the LA molecule into the frame manually superimposing the structures. Notably, within our frame, the LA molecule was positioned in close proximity to the alpha-helix of the pocket, causing some steric clashes and several hydrophobic residues were oriented toward the LA molecule, which led to a reduction in the accessible volume inside the pocket (See Image 3).

So, to optimize the binding of the ligand within the pocket and to mimic the displacement of the gating helix to accommodate the LA molecule, we devised a minimization protocol. The protocol is detailed in Material and Method section in the docking subsection. After the minimization, the steric clashes between the LA molecule and the side chains of the amino acids within the pocket were visibly reduced. The LA molecule was positioned to facilitate

both hydrogen bonding and electrostatic interactions with key residues, including Arg408, Gln409, and Lys417.

Following this initial minimization, a second minimization step was performed. However, in this step, a constraint was applied to the ligand, keeping it fixed while allowing the alpha helix to move. Two additional minimization steps were performed under these conditions to obtain the final LA-S protein complex.

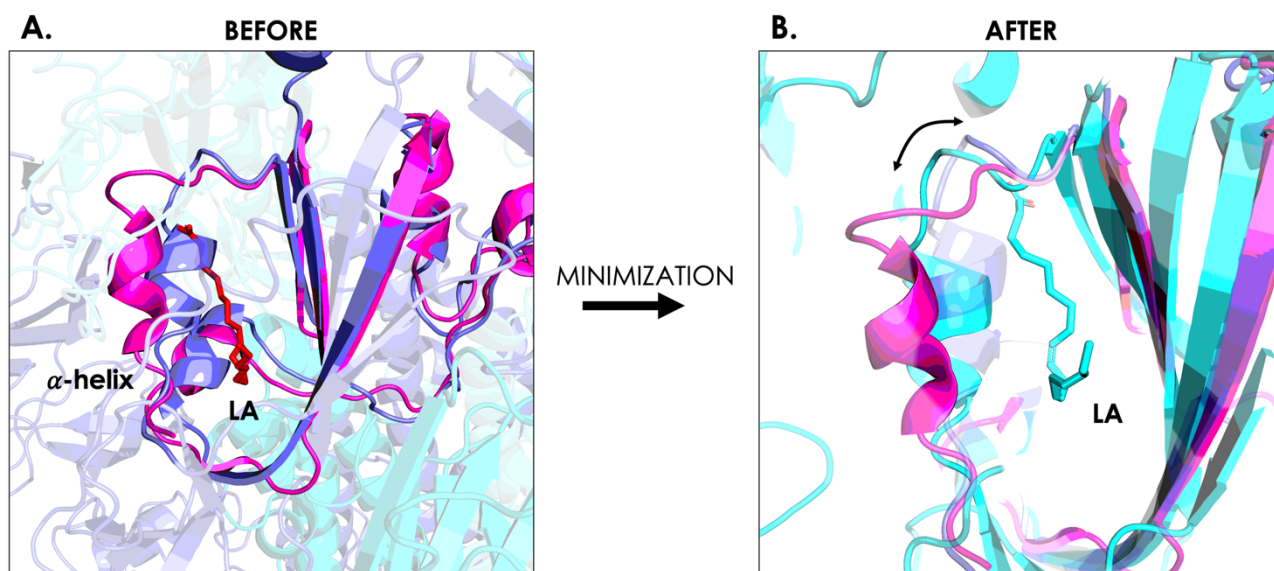


Figure 3. Minimization protocol. **A.** Superimposition of the fatty acid binding pocket from the crystal structure (PDB 6ZB5) in pink and in violet our frame extracted from the simulation (from the D614G variant). In red the linoleic acid from the crystal structure. **B.** Superimposition after the minimization: in violet the original structure of the frame from the MD, in pink from the crystal structure and in light blue the MD-frame after the minimization (see the movement of the α -helix).

To assess the accuracy of our computational approach in replicating the experimental cryo-EM structure, we conducted a docking experiment using Glide (from Schrodinger Suite¹⁴) with the minimized LA-S protein complex. From this, we obtained a considerable number of poses where the linoleic acid molecule interacted either electrostatically or through hydrogen bonds with the charged residues Arg408, Gln409, and Lys417 with a RMSD, compared to the experimental structure, comprised between 2.5 and 3 Å.

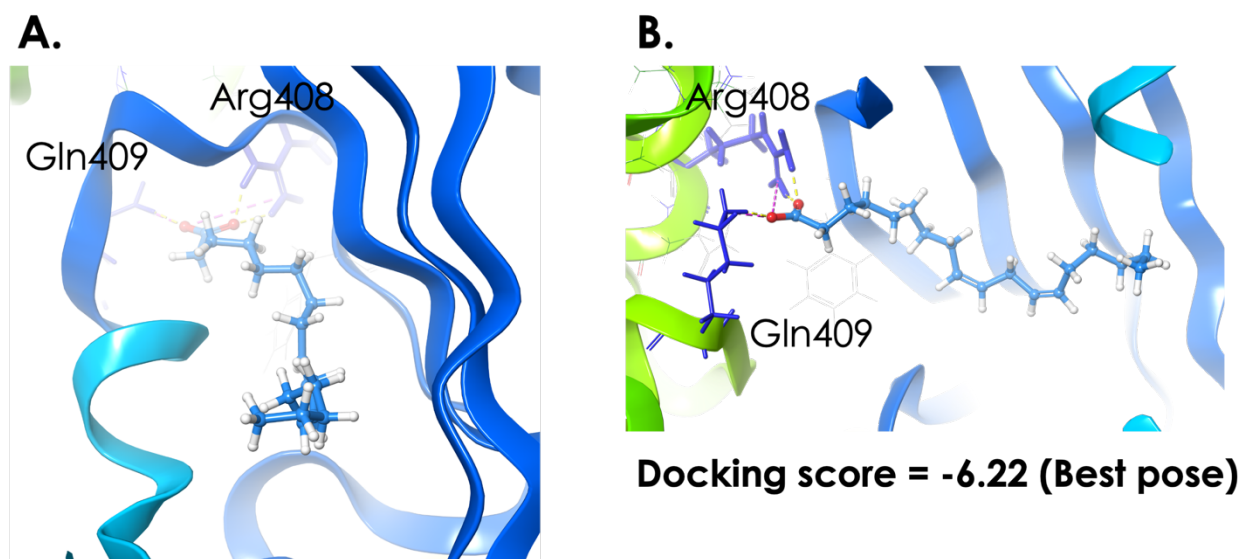


Figure 4. Docking results of the linoleic acid molecule. A. and B. Two views of the same docking pose of the linoleic acid inside the minimized binding pocket. Highlighted in purple and yellow the electrostatic and hydrogen bonding interactions, respectively, with the «anchor residues» Arg408 and Gln409 from the adjacent monomer.

These results provide an assessment of the computational model's ability to generate a ligand binding pose comparable to the experimental cryo-EM structure, shedding light on the potential binding interactions within the LA-Spike protein complex.

Docking of small and drug-like molecules

Once the starting structure of the protein with the “open” pocket is obtained, we moved on identifying candidates that could mimic the LA interactions inside the pocket and possibly its behavior. This docking process involved small molecules with drug-like properties screening two different libraries of approximately 1,000 compounds each (we used two libraries composed of small, drug-like molecules, for a total of 2000 compounds, for details see Materials and Methods). HTVS was used to generate one pose per ligand. Following this initial screening, we conducted further refinement of the results by filtering ligands based on their ability to establish hydrogen bonds with the charged residues Arg1678, Gln1679, and Lys417 within the binding pocket.

For each library, we identified 22 ligands that achieved high docking scores and exhibited favorable ligand efficiency. Below, you will find descriptions of some poses from highly ranked ligands, as well as lists of the selected ligands for each library (**Figures 5**).

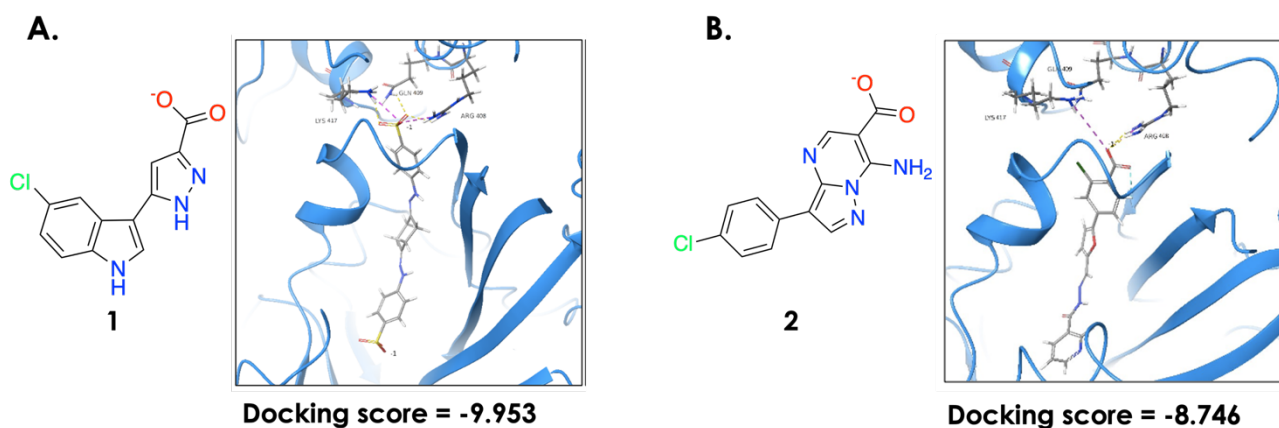


Figure 5. Docking results with small and drug-like molecules. A. structure of the best ligand from library 360 and pose in the fatty acid binding pocket. **B.** Structure of the best ligand from the library 400 and pose.

These results represent a crucial step in identifying potential drug candidates or small molecules that could interact effectively with the Spike protein's binding pocket, offering promising avenues for further exploration and development.

Docking of fragments

In the final phase of our study, we focused on docking small fragments from a library containing approximately 4,300 compounds. Given the nature of these small fragments, it was unlikely that any single molecule could establish both polar/electrostatic interactions with the anchor residues and hydrophobic interactions with the rest of the pocket simultaneously.

To manage this complexity, we employed a two-tiered filtering approach after performing HTVS. The selection of fragments was based on two types of interactions:

1. Type P (Polar Interactions): Fragments of this type were chosen primarily for their ability to form polar interactions, such as hydrogen bonds, with specific residues within the pocket, including Arg408, Gln409, and Lys417.

2. Type H (Hydrophobic Interactions): able to do interactions with residues like Phe342 or Tyr365, two residues are located at the opposite end of the binding pocket and belong to the C protomer (as shown in **Figure 6**).

This dual approach allowed us to identify fragments that could contribute to the overall binding affinity through polar or hydrophobic interactions within the binding pocket, considering the spatial distribution of key residues. From the two distinct sets of fragments,

we carefully selected a total of 30 fragments from the former category and 11 from the latter (**Figure 6**).

These fragments represent valuable candidates for further investigation and potential development as ligands for the Spike protein's binding pocket.

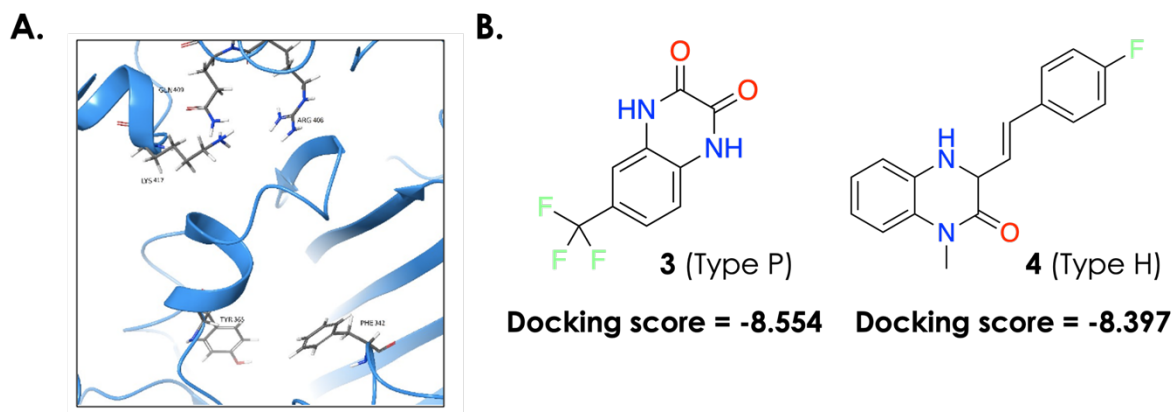


Figure 6. Docking results with fragments. **A.** Residues selected for the selection of the two types of fragments. **B. 3** Structure of the best ligand from Type P and **4** Type H. (Polar Interactions: fragments able to form polar interactions, such as hydrogen bonds, with Arg408, Gln409, and Lys417.) Type H (Hydrophobic Interactions): able to do interactions with residues like Phe342 or Tyr365.

We also explored the potential synergistic interactions between fragments of type P (polar interactions with anchor residues) and type H (hydrophobic interactions with the hydrophobic region of the pocket). The goal was to identify combinations of these fragments that could work together to mimic the behavior of linoleic acid more effectively within the binding pocket (**Figure 7**). During this evaluation, an interesting observation emerged. In some pairs of fragments, we noticed that the aromatic ring of one fragment overlapped with the aromatic ring of the other fragment (**Figure 7, B.**). This observation suggested that it might be possible to synthesize a single molecule by combining these two fragments. Such a synthesized molecule could potentially retain both types of interactions-polar and hydrophobic-within the binding pocket.

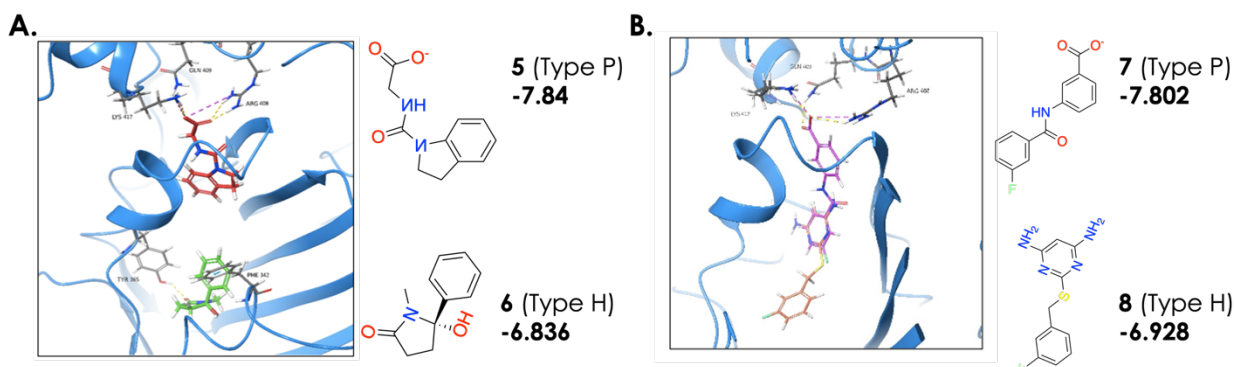


Figure 7. Combination of fragments. **A.** Example of two fragments combined: the fragments are **5** (red) and **6** (green). The picture also shows some residues inside the pocket involved in important interactions: R408, Q409 and K417 at the top; F342 and Y365 at the bottom. **B.** Example of two fragments that have aromatic rings superimposed: fragments are **7** (purple) and **8** (orange).

This finding opens intriguing possibilities for designing novel compounds that could more closely mimic the behavior of linoleic acid and enhance their binding affinity to the Spike protein's pocket. It highlights the potential for innovative drug design strategies to target the Spike protein effectively.

SARS-CoV-2 Variants: effects on the LA binding pocket

Docking K417 mutant structures.

Dealing with mutations in the Spike protein presents a significant challenge in the design of drugs and vaccines against SARS-CoV-2.¹¹ Fortunately, among the variants we examined, only one mutation occurs within the free fatty acid binding pocket. Specifically, the Gamma variant carries the K417T mutation, while both the Alpha and Omicron variants have a mutation resulting in the substitution of the residue to N. This mutation involves the substitution of Lysine (a positively charged amino acid) with either Threonine or Asparagine, both of which are polar but neutral residues. The aim here was to assess how the docking results would change with the substitution of a positively charged residue with two polar neutral residues. As expected, the top-ranked ligands in these mutated structures predominantly formed hydrogen bonds with the anchor residues rather than electrostatic interactions, although some maintained limited electrostatic interactions with Arg408 and Gln409. Interestingly, the ligands selected for the K417 mutant structures still exhibited acceptable values of docking score and ligand efficiency, suggesting that some ligands might be effective against multiple variants.

These findings suggest that in theory, if a candidate drug can successfully bind to the free fatty acid binding pocket in the D614G variant, it might also exhibit binding affinity to other variants. This potential cross-reactivity is a promising avenue to explore.

Pocket search in different variants.

As previously mentioned, one of the primary challenges in developing drugs or vaccines for SARS-CoV-2 is their limited efficacy against various virus variants. Ideally, an effective drug should target a site that remains consistent across multiple variants. To address this, we investigated the free fatty acid binding pocket in eight different variants: Wild type, Alpha (B.1.1.7), Beta (B.1.351), Gamma (P.1), Delta (B.1.617.2), Delta+ (B.1.617.2.1), Mu (B.1.621), and Omicron (B.1.1.529).

Our approach began by considering the most populated cluster for each variant, as it should provide the most representative data from the simulations. Initially, we explored pockets using the SiteMap tool¹³ within the entire protein structure and identified multiple pockets for each variant. Importantly, at least one of these pockets encompassed the region where the ligand LA binds.

To develop a model suitable for all the variants, we started considering the WT. However, when we restricted the pocket search to the vicinity of the LA molecule in the WT, SiteMap did not identify that region as a pocket (**Figure 8**). Instead, it identified a hydrophobic region behind the ligand, likely due to the orientation of hydrophobic and bulky residues within the pocket (such as some bulky residues such as Phe377 which clash sterically with the LA molecule). In these clusters, these residues pointed toward the ligand, causing steric clashes and preventing SiteMap from recognizing the region as a suitable pocket. Conversely, the region identified by SiteMap was formed by smaller and hydrophobic residues, potentially offering better accommodation for a small molecule.

Nevertheless, after applying our minimization protocol (see Materials and Methods), steric clashes were reduced, and SiteMap identified the pocket around the ligand (**Figure 8**). This occurred because, after the minimization, the hydrophobic residues that initially pointed toward LA tended to reorient themselves away from the ligand, allowing the linoleic acid molecule to fit inside the pocket with fewer steric clashes. To assess if the minimization produced a volume comparable to experimental data, we used POVME for quantitative results.

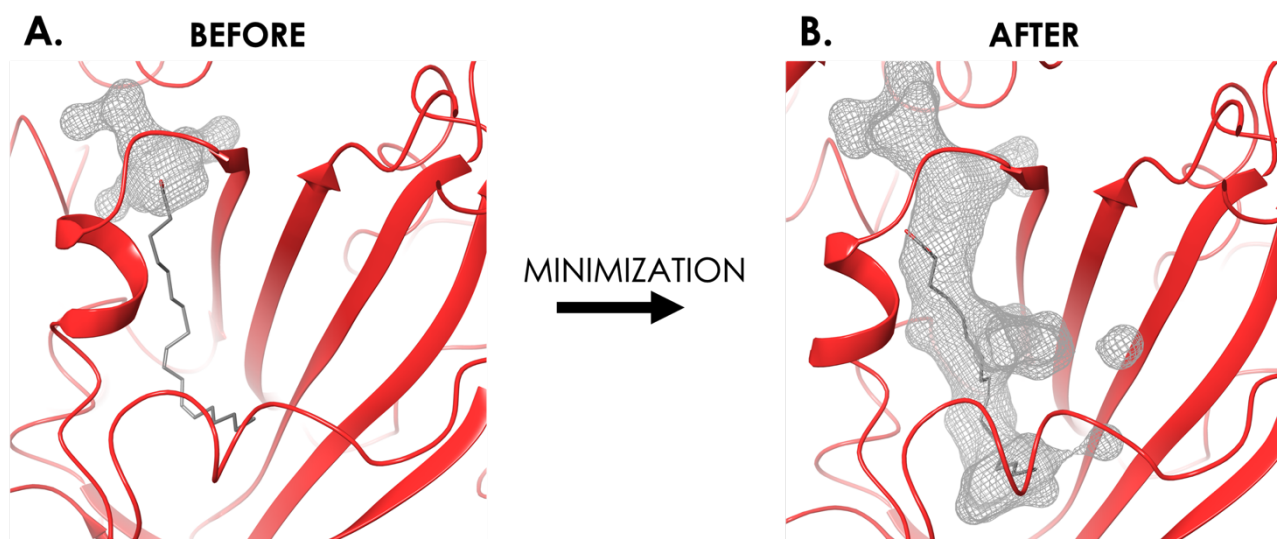


Figure 8. SiteMap results for the Wild Type variant **A.** before the minimization and **B.** after the minimization (the grey mesh represents the volume of the pocket). The protein structure is the most populated cluster from the MD simulation for the WT variant.

It's important to note that our SiteMap analysis involved only a single step of our minimization protocol. Given that this single step cannot fully replicate a 6 Å displacement of the helix, it was expected that both volumes and surface areas for the clusters would be smaller compared to the reference structure. Indeed, the data confirmed this hypothesis.

Then we apply the same protocol to all the variants and we found out that some variants (Alpha, Beta, and Omicron) exhibited relatively high volumes ($> 220 \text{ \AA}^3$) and surface areas ($> 270 \text{ \AA}^2$) (**Figure 9, Table 1**). This could be due to the FFBP in the most populated cluster already having a wider volume compared to other variants that yielded less favorable results, such as the Wild Type. Alternatively, during the minimization step, the residues in these three variants might have moved farther away from the ligand compared to other variants.

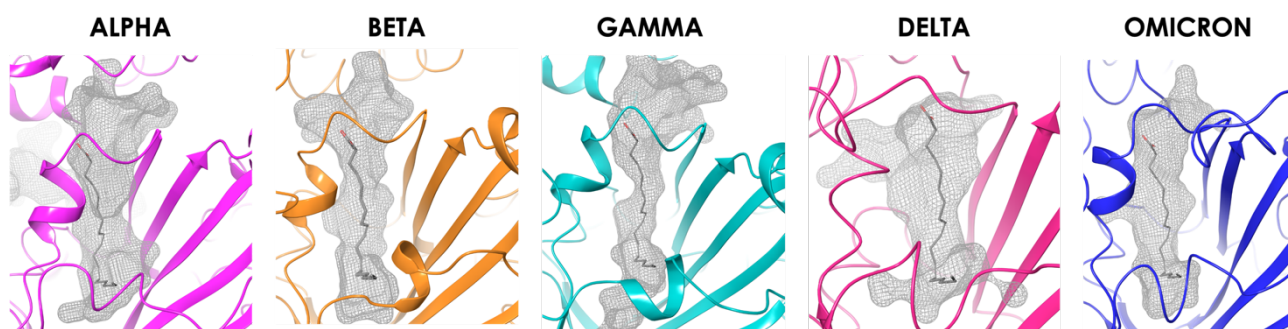


Table 1

| Variant | Volume (Å ³) | Surface (Å ²) |
|------------------|--------------------------|---------------------------|
| Reference (6ZB5) | 355 | 318 |
| Wild Type | 165 | 254 |
| D614G | 190 | 286 |
| Alpha | 225 | 275 |
| Beta | 243 | 301 |
| Gamma | 174 | 231 |
| Delta | 201 | 257 |
| Omicron | 241 | 298 |

Figure 9. SiteMap results for the variants after the minimization.

Table 1. Quantitative results of the pocket volume calculated using POVME software for the reference structure (PDB 6ZB5) compared with all the other variants).

Given that the target is a transient pocket, it is reasonable to assume that in the most populated cluster of an MD simulation of the Spike protein without the LA molecule, the residues inside the pocket are not oriented correctly to accommodate the ligand. To address this, we investigated MD simulations of fully glycosylated Spike proteins to gain a better understanding of the pocket's opening. We examined the trajectories of all the variants mentioned before. The analysis began with aligning the trajectories on the backbone atoms that are not part of the binding site. We then conducted an RMSD analysis to identify frames within the 80,000-frame MD simulation that met all the criteria for binding linoleic acid. These criteria encompassed both the orientation of charged residues (Arg408 and Gln409) and the displacement of the helix. We calculated the RMSD between the frames and the reference structure, considering six residues: two from the α -helix (Val367 and Leu368 from the adjacent protomer), two from the β -sheet (Phe377 and Ile434 from the adjacent protomer), and the two charged residues.

Interestingly, we discovered frames in three variants (Alpha, Beta and Gamma) that resembled the experimental structure. In these frames, the charged residues pointed toward the charged head of LA, establishing positive interactions with it (**Figure 10**).

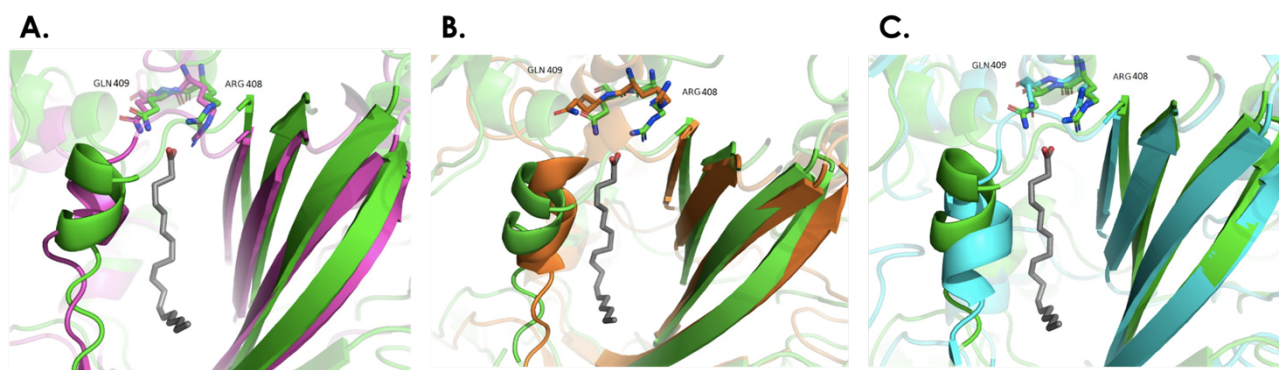


Figure 10. **A.** Comparison between the reference structure (PDB 6ZB5) (green) and the frame 64833 (RMSD: 2.3929 Å) from the MD simulation of the Alpha variant (purple). **B.** and the frame 57798 (RMSD = 2.0577 Å) from the MD simulation of the Beta variant (orange). **C.** and the frame 57798 (RMSD = 1.7447 Å) from the MD simulation of the Gamma variant (turquoise).

Moreover, the bulky residue Phenylalanine (residue 377 in the WT, Beta, and Gamma variants and residue 374 in the Alpha variant), which typically resides in the center of the binding pocket, was not as closely positioned to the LA molecule as in the majority of frames. While the side chain of this residue still did not precisely match the reference structure, it was plausible that the movement of the benzene of the Phe was influenced by the presence of the LA molecule inside the binding site.

Another notable feature of these frames was that the volume of the binding pocket appeared comparable to the experimental structure. Although the FFBP in these frames was not as wide as in the reference structure, it still provided sufficient space to accommodate a small molecule. This observation suggests that the pocket might indeed be a druggable site.

It's crucial to recognize that these frames were only sparsely populated, which is expected for a cryptic pocket. Nonetheless, our results suggest that the presence of LA (or a similar ligand) might be necessary for the opening of the pocket in most variants. But in a few cases, such as Alpha, Beta and Gamma, we were able to identify the transient pocket, albeit infrequently. This underscores the importance of extensive and unbiased sampling of Spike protein conformations and the value of integrating multiple computational approaches to investigate complex biological problems.

1.15.4 Discussion

Our research primarily focused on evaluating the potential druggability of the free fatty acid binding pocket, which presents an intriguing target for antiviral drug development. To achieve this, we devised a minimization protocol capable of simulating the opening of the free fatty acid binding pocket within the context of the D614G variant's molecular dynamics simulation. Although we were unable to precisely replicate the 6 Å displacement observed in experimental studies involving the unglycosylated Spike protein, we did observe significant helix movement, providing ample space for the ligand (LA) to re-dock within the structure. This enabled us to generate a complex similar to the one observed in the cryo-electron microscopy (cryo-EM) structure.

Our screening efforts, involving three different compound libraries, identified certain compounds that maintained critical interactions with charged residues, which are essential for proficiently locking the ligand inside the binding site. Furthermore, we selected combinations of fragments positioned in various regions of the pocket, which could serve as a foundation for fragment-based drug design. Currently, both individual fragments and combinations thereof are undergoing experimental evaluation to assess their ability to bind within the targeted pocket. But we can anticipate that it turns out that some fragments are able to bind the protein, and using the combined fragments increases its effectiveness.

Notably, when we considered mutations of K417 (found in the Beta, Gamma and Omicron variants), the docking results for the three libraries remained relatively stable. This suggests that certain highly ranked ligands could potentially serve as effective drugs against multiple variants, regardless of the presence or absence of this mutation. By analyzing molecular dynamics simulations of different variants (Wild Type, Alpha, Beta, Gamma, Mu and Omicron), we identified specific frames in some VOCs where the stereochemical prerequisites for binding, as revealed by the cryo-EM structure, were satisfied. These prerequisites included the orientation of charged residues and the helix displacement. Although the number of frames meeting these criteria was limited, as expected for a transient pocket in the absence of a ligand, this finding is crucial. It indicates that in the Alpha, Beta and Gamma variants, the pocket can transiently open even without a ligand, suggesting that these variants, in particular, could potentially accommodate a ligand within the free fatty acid binding pocket, making them promising drug targets.

In summary, our findings suggest that the free fatty acid binding pocket holds promise as a target for neutralizing various SARS-CoV-2 variants. However, it's important to

acknowledge that this binding site has limitations, mainly related to the specific conditions required for it to open. Nonetheless, our ability to identify structures in the molecular dynamic simulations of different variants with these key features reinforces the importance of further investigating this binding site.

1.15.5 Materials and Methods

Clustering

Clustering was performed using the Cpptraj¹² program in AMBER 20. The metatrajectories of the SARS-CoV-2 variants (Wild Type, D614G, Alpha, Beta, Gamma, Delta, Mu and Omicron) were built starting from 4 independent replicas, each one of 1 μ s, for a total of 80000 frames. The newly obtained trajectories were aligned on the backbone atoms of residues that do not belong to the RBD and the N-terminal. Then, clustering was performed on the remaining regions (RBD and N-terminal), using a hierarchical agglomerative algorithm. An ϵ value was chosen to obtain a number of clusters that covers majority of the trajectory, approximately 8-10 clusters. These types of clusters were used for the SiteMap¹³ analysis.

For the RMSD analysis of the variants we performed clustering in the same conditions but aligning the metatrajectories on the backbone atoms that do not belong to free fatty acid binding pocket.

Analysis of MD simulations of multiple Variants of Concern

MD simulations were analysed using AMBER 20: the cpptraj¹² program was used to perform a root-mean-square deviation (RMSD) analysis of the simulation the fully glycosylated S protein of different SARS-CoV-2 VOCs (18) (<https://amberhub.chpc.utah.edu/cpptraj/>). The variants we investigated are: Wild type, B.1.1.7 (alpha), B.1.351 (beta), P.1 (gamma), B.1617.2 (delta), and B.1.1.529 (omicron). We calculated RMSD considering the backbone atoms of residues B:Arg408, B:Gln409, C:Val367, C:Leu368, C:Phe377 and C:Ile434 using the experimental structure (PDB: 6zb5) as a reference.

Docking procedure

A single frame was extracted from the MD simulation of the fully glycosylated Spike protein (D614G variant) and this structure was structurally aligned with the reference WT Spike containing the LA¹⁰ molecules (PDB: 6ZB5) using PyMol. After, the linoleic acid molecule in the reference PDB was transferred to the D614G Spike, glycans were removed and NLN

residues (which are Asparagine residues covalently bound to glycans) were converted in ASN: this step was necessary, as Maestro¹⁴ is not able to recognize the nomenclature used to define glycans according to the force field used in MD simulations. The complex was uploaded with Maestro¹⁴ and prepared using its Protein Preparation Wizard¹⁵: beside the default options, we selected “fill in missing side chains using Prime”, to add the missing hydrogens on the former NLN residues converted to ASN. After that, we performed four minimization steps using MacroModel¹⁶ with the default force field (OPLS3e). In the first minimization step we constructed a substructure of freely moving atoms, selecting the LA molecule as the center and expanding the selection to residues within 5 Å. Then, we constructed two more shells: the first one has a radius of 6 Å (starting from LA) and the residues are constrained with a force constant of 100 kJ/(mol Å²), whereas in the second shell (with a radius of 8 Å) residues are constrained with a force constant of 200 kJ/(mol Å²). The following minimization steps were performed with the addition of a constraint with a force constant of 100 kJ/(mol Å²) on the ligand; after initial evaluations, the substructure of freely moving residues was increased to 6, in order to include the helix inside of it. Therefore, the shell constrained with a force constant of 100 kJ/(mol Å²) was increased as well to 7 Å.

Libraries used for high throughput virtual screening

To perform high throughput virtual screening (HTVS) inside the LA pocket we used three different libraries. Two of them are composed of small, drug-like molecules, for a total of 2000 compounds, the two set we chose are named “drug_like_decoys_avg_360mw_1_epik” and “drug_like_decoys_avg_400mw_1_epik” and they were downloaded from the Schrödinger website (<https://www.schrodinger.com/products/glide>). They were prepared using the LigPrep tool¹⁷: the OLPS4 force field was used considering a pH interval between 6 and 8, using Epik¹⁵, pose for every ligand was generated. The third library is a custom-made proprietary library of small molecules and fragments covering a large chemical space. It was prepared using the LigPrep¹⁷ program in Maestro: using the FF OLPS3e, all the possible state were generated for the pH interval 6-8 and all the possible combinations were generated (at most 32 per ligand).

POVME 3.0

POVME (POcket Volume MEasurer)¹⁸⁻²⁰ is an algorithm developed to determine both the volume and surface area of a binding pocket using a MD simulation trajectory. To achieve that, an inclusion region must be defined: the region might be a sphere, a cylinder, or a box

of an appropriate size and at the right coordinates. Then, one can define an exclusion region as well, to refine the region and avoid overestimation in the results; furthermore, the algorithm automatically removes any points that are too close to the receptor. Alternatively, one could use the structure or trajectory of the ligand- receptor complex: once the ligand is specified, the program is able to calculate the volume and surface of the region where the ligand is placed.

Characterization of the free fatty acid binding pocket on the minimized clusters.

The PDB files of the most populated, minimized clusters for different variants underwent a POVME analysis to determine the volume of the binding pocket around the ligand. The default parameters in the input were used (https://github.com/POVME/POVME/blob/master/POVME/examples/ligand_example/sample_POVME_input.ini) the ligand was defined as EIC and both volume and surface area were calculated. As a reference structure we used PDB 6ZB5 and performed the same analysis.

1.15.6 References

- (1) Toelzer, C.; Gupta, K.; Yadav, S. K. N.; Borucu, U.; Davidson, A. D.; Kavanagh Williamson, M.; Shoemark, D. K.; Garzoni, F.; Staufer, O.; Milligan, R.; et al. Free fatty acid binding pocket in the locked structure of SARS-CoV-2 spike protein. *Science* **2020**, *370* (6517), 725-730. DOI: 10.1126/science.abd3255.
- (2) Shoemark, D. K.; Colenso, C. K.; Toelzer, C.; Gupta, K.; Sessions, R. B.; Davidson, A. D.; Berger, I.; Schaffitzel, C.; Spencer, J.; Mulholland, A. J. Molecular Simulations suggest Vitamins, Retinoids and Steroids as Ligands of the Free Fatty Acid Pocket of the SARS-CoV-2 Spike Protein*. *Angew Chem Int Ed Engl* **2021**, *60* (13), 7098-7110. DOI: 10.1002/anie.202015639.
- (3) Toelzer, C.; Gupta, K.; Yadav, S. K. N.; Hodgson, L.; Williamson, M. K.; Buzas, D.; Borucu, U.; Powers, K.; Stenner, R.; Vasileiou, K.; et al. The free fatty acid-binding pocket is a conserved hallmark in pathogenic β -coronavirus spike proteins from SARS-CoV to Omicron. *Sci Adv* **2022**, *8* (47), eadc9179. DOI: 10.1126/sciadv.adc9179.
- (4) Sofia F Oliveira, A.; Shoemark, D. K.; Avila Ibarra, A.; Davidson, A. D.; Berger, I.; Schaffitzel, C.; Mulholland, A. J. The fatty acid site is coupled to functional motifs in the SARS-CoV-2 spike protein and modulates spike allosteric behaviour. *Comput Struct Biotechnol J* **2022**, *20*, 139-147. DOI: 10.1016/j.csbj.2021.12.011.
- (5) G. Ciccotti, G. J., I. R. McDonald. Thought-experiments by molecular dynamics . . . *J Stat Phys* **1979**, *21*, 1– 21.
- (6) Ciccotti, G. in *Computer Simulation in Material Science. P. V. Meyer M, Ed. (Kluwer Academic Publishers)* **1991**, *21*, 119–137
- (7) G. Ciccotti, M. F. Non-equilibrium by molecular dynamics: a dynamical approach . . . *Mol Simul* **2016**, *42*, 1385 – 1400.
- (8) Oliveira, A. S. F.; Ciccotti, G.; Haider, S.; Mulholland, A. J. Dynamical nonequilibrium molecular dynamics reveals the structural basis for allostery and signal propagation in biomolecular systems. *Eur Phys J B* **2021**, *94* (7), 144. DOI: 10.1140/epjb/s10051-021-00157-0.
- (9) Gupta, K.; Toelzer, C.; Williamson, M. K.; Shoemark, D. K.; Oliveira, A. S. F.; Matthews, D. A.; Almuqrin, A.; Staufer, O.; Yadav, S. K. N.; Borucu, U.; et al. Structural insights in cell-type specific evolution of intra-host diversity by SARS-CoV-2. *Nat Commun* **2022**, *13* (1), 222. DOI: 10.1038/s41467-021-27881-6.
- (10) *PyMOL Molecular Graphics System, Version 2.5.0 Schrödinger, LLC.* ; (accessed).
- (11) Carabelli, A. M.; Peacock, T. P.; Thorne, L. G.; Harvey, W. T.; Hughes, J.; Peacock, S. J.; Barclay, W. S.; de Silva, T. I.; Towers, G. J.; Robertson, D. L.; et al. SARS-CoV-2 variant biology: immune escape, transmission and fitness. *Nat Rev Microbiol* **2023**, *21* (3), 162-177. DOI: 10.1038/s41579-022-00841-7.
- (12) Roe, D. R.; Cheatham, T. E. PTRAJ and CPPTRAJ: Software for Processing and Analysis of Molecular Dynamics Trajectory Data. *J Chem Theory Comput* **2013**, *9* (7), 3084-3095. DOI: 10.1021/ct400341p.
- (13) *Schrödinger Release 2020-4: SiteMap, Schrödinger, LLC, New York, NY, 2020.* ; (accessed).
- (14) *Schrödinger Release 2020-4: Maestro, Schrödinger, LLC, New York, NY, 2020.* ; 2020. (accessed).
- (15) *Schrödinger Release 2020-4: Protein Preparation Wizard; Epik, Schrödinger, LLC, New York, NY, 2020; Impact, Schrödinger, LLC, New York, NY; Prime, Schrödinger, LLC, New York, NY, 2020.* ; (accessed).
- (16) *Schrödinger Release 2020-4: MacroModel, Schrödinger, LLC, New York, NY, 2020.* ; (accessed).
- (17) *Schrödinger Release 2020-4: LigPrep, Schrödinger, LLC, New York, NY, 2020.* ; (accessed).
- (18) Durrant, J. D.; de Oliveira, C. A.; McCammon, J. A. POVME: an algorithm for measuring binding-pocket volumes. *J Mol Graph Model* **2011**, *29* (5), 773-776. DOI: 10.1016/j.jmgm.2010.10.007.
- (19) Durrant, J. D.; Votapka, L.; Sørensen, J.; Amaro, R. E. POVME 2.0: An Enhanced Tool for Determining Pocket Shape and Volume Characteristics. *J Chem Theory Comput* **2014**, *10* (11), 5047-5056. DOI: 10.1021/ct500381c.
- (20) Wagner, J. R.; Sørensen, J.; Hensley, N.; Wong, C.; Zhu, C.; Perison, T.; Amaro, R. E. POVME 3.0: Software for Mapping Binding Pocket Flexibility. *J Chem Theory Comput* **2017**, *13* (9), 4584-4592. DOI: 10.1021/acs.jctc.7b00500.



Genetic Approach

DNA and RNA

In contemporary medicinal chemistry, one of the emerging major challenges is the attempt to target RNA (in particular noncoding sequences) with small-molecules.

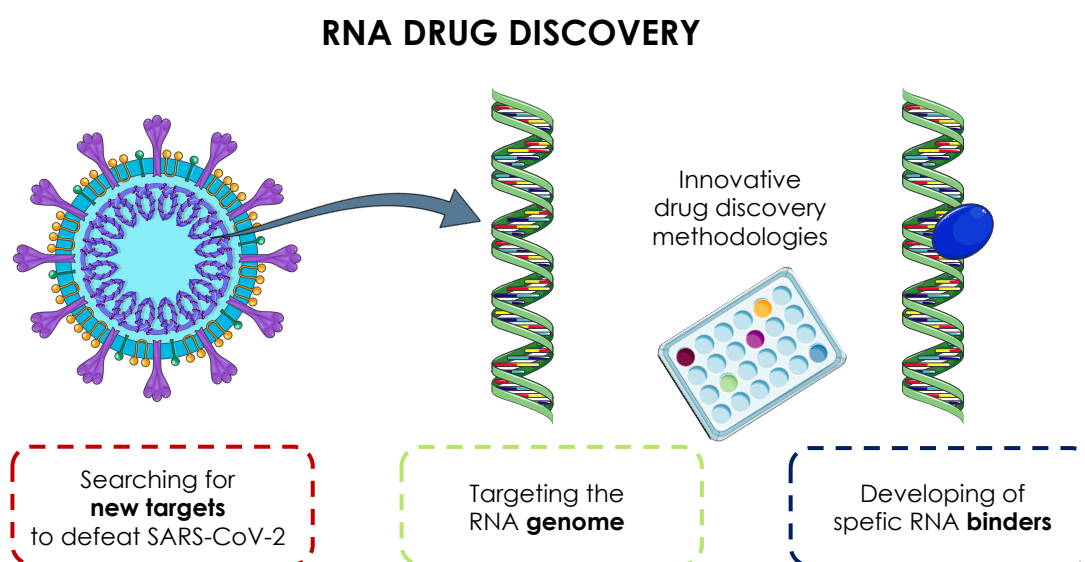


Figure 1. Schematic representation of the drug discovery process aimed at devising effective strategies against SARS-CoV-2. The initial step involves identifying novel targets, distinct from conventional ones. For instance, focusing on the RNA genome as a target allows for the subsequent development of specialized RNA binders. Targeting RNA using small molecules shows significant potential for substantial therapeutic advancements, particularly with emerging chemical approaches like ribonuclease targeting chimeras (RIBOTACs).

This pursuit necessitates a fundamental shift in perspective compared to traditional drug discovery approaches. In this context, the recent work by Disney and colleagues,¹ featured in ACS Central Science, takes center stage. Their groundbreaking research encompasses the identification of specific ligands tailored for targeting SARS-CoV-2 RNA, the

development of ribonuclease targeting chimeras (RIBOTAC), and an exhaustive exploration of intracellular mechanisms of action through the utilization of chemical biology tools. This study underscores the transformative potential of novel chemical modalities that could shape the landscape of future therapies.

The inspiration for this research stems from the urgent need to fight SARS-CoV-2 and, while substantial efforts have been devoted to repurposing existing drugs, conducting large-scale screenings, and expediting vaccine development, there is an evident demand for fresh targets and innovative bioactive compounds to reinforce the arsenal of antiviral drugs.

Noncoding RNAs, once relegated as mere intermediaries in the gene expression process, have emerged as promising possible candidates for drug discovery. These RNA molecules play pivotal roles in essential biological processes, including transcription, translation, and gene expression regulation. Pertinently, dysregulations in noncoding RNA expression and functions have been directly linked to various pathologies, encompassing neurological disorders, cardiovascular diseases, and cancers. Remarkably, over 70% of the human genome is transcribed into noncoding RNAs, while only 1.5% codes for proteins. Recognizing that only a fraction of these proteins serves as the targets of currently available drugs, the inclusion of noncoding RNAs as potential therapeutic targets presents an exciting frontier that could substantially expand the field of drug development.

The ongoing COVID-19 pandemic underscores the promise of RNA-targeted therapies for addressing infections caused by RNA viruses. Prior investigations have explored RNA-targeted therapies within the context of HIV, HCV, and influenza viruses, setting the stage for the pursuit of RNA-targeted interventions as a potent strategy against viral infections.

Over the years, significant progress has been achieved in the discovery of RNA-targeting drugs, beginning with the advent of RNA ligands such as aminoglycosides and tetracyclines in the 1940s, and more recently, oxazolidinones. This progress has been marked by a diverse array of innovative drug discovery strategies, including the introduction of *Inforna*, a lead identification approach for identifying highly specific RNA ligands. Inforna combines two-dimensional combinatorial screening (2-DCS) with structure-activity relationships through sequencing (StARTS), enabling the anticipation of affinity and selectivity within RNA libraries, thus streamlining the identification of compounds targeting specific RNA secondary structures.

Furthermore, various screening technologies, such as microarrays and fluorescence-based assays, have been employed in tandem with structure-based design principles to yield selective RNA ligands. These endeavors have contributed to the construction of

valuable databases, exemplified by R-BIND, which are poised to accelerate the drug discovery process in the realm of RNA targeting.

To comprehensively understand the interactions between small molecules and RNA, researchers have leveraged tools from structural biology and chemical probing, including NMR and SHAPE technologies. Collectively, these efforts have yielded promising results, culminating in the market introduction of Risdiplam, a mRNA splicing modulator against spinal muscular atrophy (SMA).

The adaptability of drug discovery tools enables their customization to tackle the intricate challenges posed by RNA targets. For a considerable period, there was a prevailing belief that RNA, due to its highly polar and solvent-exposed nature, was resistant to small molecule binding.² The skepticism stemmed from the notion that compounds might lack sufficient binding energy to surmount the water barrier and effectively reach the surface of RNA. This adaptability can be harnessed in conjunction with the development of original drug modalities, such as precise genome editing, modified peptides, oligonucleotides, macrocycles, and various conjugates. These innovations open new avenues for target modulation. In the study by Disney and colleagues,¹ these technologies were deployed for the identification of precise tools targeting specific RNAs and facilitating RNA degradation, in addition to elucidating the intracellular molecular mechanisms of action.³

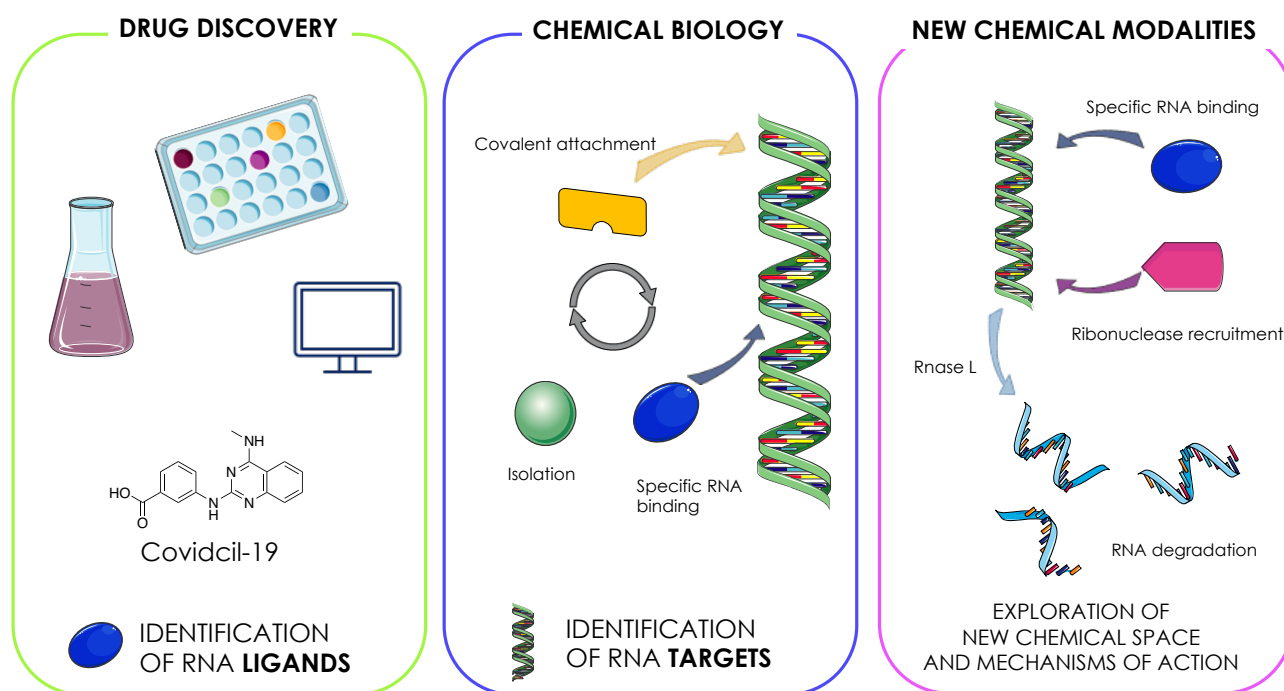


Figure 2. On the left: drug discovery tools encompassing screening, structure-based design, and molecular modeling. These methods aid in identifying potent and precise RNA ligands like Covidicil-19, which targets SARS-CoV-2 FSE RNA. Following this, chemical biology approaches (center) are

employed to delve into the intracellular mechanism of action. Modified compounds are engineered to covalently bind the target, allowing for straightforward isolation and identification. Subsequently, on the right, novel chemical modalities are devised to extend functionality, such as target-specific degradation.

The research conducted by Disney and collaborators started with a screening of an RNA-focused small molecule collection housed in the Inforna database. This screening was directed towards the SARS-CoV-2 frameshifting element RNA, which contains a 1 × 1 nucleotide UU internal loop critical for frameshifting element function. Subsequent investigations led to the discovery of Covidcil-19, an RNA binder with nanomolar affinity for the target RNA. Covidcil-19 exhibited the capacity to stabilize the folded state of the RNA hairpin and impede frameshifting within cells.

To further probe the intracellular target, chemical cross-linking and isolation by pull-down (Chem-CLIP) were employed. Chem-CLIP, a proximity-based reaction, covalently links an RNA-binding small molecule to its target, enabling precise mapping of binding interactions. Using a conjugate known as Covidcil-CLIP, which covalently linked to the RNA target and facilitated immunoprecipitation, a 3-fold increase in targeted RNA in the pull-down fraction was observed, affirming the affinity and selectivity for the frameshifting element RNA.

Finally, Covidcil-19 was modified to yield a RIBOTAC compound, a novel chemical modality reminiscent of PROTACs developed for protein targeting and degradation. RIBOTACs involve the covalent attachment of an RNA binder to a chemical compound capable of recruiting a cellular ribonuclease, thereby inducing cleavage and degradation of the targeted RNA. This innovative approach resulted in the creation of the Covidcil RIBOTAC, which demonstrated the ability to induce targeted cleavage and degradation of the entire SARS-CoV construct. Further optimization efforts enhanced the bioactivity of the RIBOTAC compound by at least tenfold, as corroborated by intracellular luciferase reporter assays.

While the battle against SARS-CoV-2 entails the exploration of multiple strategies and therapeutic avenues, the study's focus on targeting the viral RNA genome with small-molecule drugs is promising, with potential applications beyond the current pandemic. This research shows the feasibility of drugging the SARS-CoV-2 RNA genome and underscores the potential for RNA-targeted therapies in drug discovery. Recent successes with new RNA binders making their way to clinical trials underscore the growing potential of RNA targeting in the near future. However, the intricate nature of these hybrid compounds, featuring

diverse moieties designed to harness various intracellular mechanisms, presents challenges related to scale-up, formulation, stability, and toxicity. Although the clinical application of such complex chemical tools and the targeting of challenging RNA-based targets may appear daunting, the recent progress seen in clinical trials of PROTACs suggests that these novel drug modalities may indeed represent the future of medicinal chemistry.

1.16 G4 Motifs as possible new drug targets

Although the genomes of viruses undergo rapid mutations, the secondary structural elements are relatively conserved and therefore can be used as potential antiviral targets.

In this regard, another type of novel and highly conserved target is represented by G-quadruplexes (G4s), which are non-canonical secondary structures formed by guanine-rich sequences within nucleic acids. G4s have undergone extensive research within the human genome. Recently several studies have identified G4 structures within the genomes of both DNA and RNA viruses, including human immunodeficiency virus-1 (HIV-1), Zika virus (ZIKV), hepatitis C virus (HCV), rhinovirus, Ebola virus (EBOV), influenza virus, human papillomavirus (HPV), herpes simplex virus 1 (HSV-1), Epstein-Barr virus (EBV), and human cytomegalovirus (HCMV).⁴⁻⁶

Viral G4s are known to play crucial roles in regulating genome replication, maintaining genome integrity, and controlling processes like transcription and translation. This makes them potential targets for antiviral therapies.

In these years of pandemic, several studies have predicted many putative G4-forming sequences (PQSs) in the genome of SARS-CoV-2 and proposed as potential therapeutic targets.⁷⁻¹⁰ However, these results exhibit significant variations due to the diverse prediction software and algorithms employed.

Cui et al.¹¹ employed the quadruplex-forming G-rich sequences mapper (QGRS) to identify 14 PQSs in the positive RNA strand. These PQSs were characterized by G-rich sequences with patterns of $G_2N_xG_2N_yG_2N_zG_2$ and loop sizes ranging from 0 to 12. Their findings suggested that SARS-CoV-2 harbors fewer PQSs compared to SARS-CoV, which could contribute to the faster replication rate of SARS-CoV-2. In another paper, Panera et al.⁸ utilized QGRS and identified 25 PQSs, with loop sizes ranging from 0 to 36. Ji et al.¹² corroborated these 25 PQSs and noted that PQSs at specific genome positions were well-preserved across the coronavirus family.

Subsequent studies employed various bioinformatics prediction tools such as G4CatchAll, pqsfinder, G4Hunter Web, and G4screener to identify additional PQSs, including some in the negative strand.^{7, 13} G4-iM Grinder analysis of the SARS-CoV-2 genome unveiled a total of 323 PQSs: while most had low scores, seven PQSs exhibited scores exceeding 30, signifying a high probability of G4 formation and four of these PQSs were conserved in SARS-coronaviruses and Bat-CoV¹⁴ (Table 2).

In a recent study, Josu'e Carvalho et al. analyzed over 200,000 SARS-CoV-2 genome sequences from across five continents. They identified highly conserved PQSs at specific positions, emphasizing their potential significance.¹⁵ All these sequences are showed in **Table 1** and **2**.

Table 1

| | G4 position | Gene | Sequence | QGRS mapper score | G4Hunter score | G4-iM grinder Score |
|----|-------------|---------|--|-------------------|----------------|---------------------|
| 1 | +236 | 5'- UTR | GGUUUCGUCCGGGUGUGACCGAAAGGUAAGAUGG | | | |
| 2 | +353 | Nsp1 | GGCUUUGGAGACUCCGUGGAGGAGG | 16 | 0.64 | |
| 3 | +359 | Nsp1 | GGAGACUCCGUGGAGGAGG | | | 30 |
| 4 | +370 | Nsp1 | GGAGGAGGUCUUAUCAGAGG | | | 30 |
| 5 | +545 | Nsp1 | GGCAUUCAGUACGGUCGUAGUGGUGAGACACUUGG | | | |
| 6 | +644 | Nsp1 | GGUAAUAAAGGAGCUGGUGG | 15 | 0.8 | 30 |
| 7 | +1463 | Nsp2 | GGUGGUCGCACUAUUGCCUUUGGAGG | 6 | 0.423 | |
| 8 | +1574 | Nsp2 | GGUGUUGUUGGAGAAGGUUCCGAAGG | 18 | 0.615 | |
| 9 | +2714 | Nsp2 | GGCGGUGCACCAACAAAGGUUACUUUUGG | 10 | 0.31 | |
| 10 | +3467 | Nsp3 | GGAGGAGGUGUUGCAGG | 15 | 1 | 34 |
| 11 | +4162 | Nsp3 | GGUUAUACCUACUAAAAAGGCUGGUGG | 6 | 0.37 | |
| 12 | +4255 | Nsp3 | GGGUCAGGGUUUAAAUGGUUACACUGUAGAGGAGG | | | 31 |
| 13 | +4261 | Nsp3 | GGGUUUAAAUGGUUACACUGUAGAGGAGG | 10 | 0.933 | |
| 14 | +4262 | Nsp3 | GGUUUAAAUGGUUACACUGUAGAGGAGG | 10 | | |
| 15 | +5036 | Nsp3 | GGACAACAGUUUGGUCCAACUUAUUUGGAUGG | | | |
| 16 | +8687 | Nsp4 | GGAUACAAGGCUAUUGAUGGUGG | 14 | 0.652 | |
| 17 | +10,058 | Nsp5 | GGUUUUAGAAAAUGGCAUUCUCAUCUGGUAAGUUGAGG | | | |
| 18 | +10,255 | Nsp5 | GGUACAGGCUGGUAAUGUUAACUCAGG | | | |
| 19 | +10,261 | Nsp5 | GGCUGGUAAUGUUAACUCAGGGUUUAUUGG | 9 | 0.6 | |
| 20 | +13,385 | Nsp10 | GGUAUGUGGAAAGGUUAUUGG | 19 | 1.048 | 31 |
| 21 | +14,947 | Nsp12 | GGUUUUCCAUUUAAUAAAUGGGGUAAAGG | 4 | 0.714 | |
| 22 | +15,208 | Nsp12 | GGAACAAGCAAUUCUAUGGUGGUUGG | 6 | 0.678 | |
| 23 | +15,448 | Nsp12 | GGCGGUUCACUAUAUGUUAACCAGGUGG | 3 | 0.345 | |
| 24 | +18,296 | Nsp14 | GGAUUGGCUUCGAUGUCGAGGGG | 9 | 1.043 | |
| 25 | +20,869 | Nsp16 | GGUGCUGGUUCUGAUAAAAGGAGUUGCACCAGG | | | |
| 26 | +22,316 | S | GGUGAUUCUUCUUCAGGUUGGACAGCUGG | 10 | 0.448 | |
| 27 | +24,215 | S | GGUUGGACCUUUGGUGCAGG | 17 | 0.6 | |
| 28 | +24,268 | S | GGCUUAUAGGUUUAAUGGUUAUUGG | 19 | 0.625 | |
| 29 | +25,197 | S | GGCCAUGGUACAUUUGGCUAGG | 17 | 0.455 | |
| 30 | +25,951 | ORF3a | GGUGGUUAUACUGAAAAUUGGAAUCUGG | 8 | 0.69 | |
| 31 | +26,746 | M | GGAUCACCGGUGGAAUUGCUAUCGCAAUGG | 7 | 0.33 | |
| 32 | +28,613 | N | GGAACUGGGCCAGAAGCUGGACUCCCUAUGG | | | |
| 33 | +28,781 | N | GGCUUCUACGCAGAAGGGAGCAGAGGCGG | 9 | 0.655 | |
| 34 | +28,903 | N | GGCUGGCAAUUGGCGG | 18 | 0.867 | 34 |
| 35 | +29,123 | N | GGAAUUUUUGGGGACCAGG | 14 | 1.053 | |
| 36 | +29,234 | N | GGCAUGGAAGUCACACCUUCGGGAACGUGG | 11 | 0.467 | |
| 37 | +29,254 | N | GGGAACGUGGUUGACCUACACAGGUGCCAUCAAAUUGG | | | |
| 38 | -165 | | GGCCUCGGUGAAAAUGUGGUGG | 13 | 0.591 | |
| 39 | -1591 | | GGGGUGCAUUUCGUGAUUUUGGGG | | 1.280 | |
| 40 | -2987 | | GGUCUGGUCAGAAUAGUGCCAUGGAGUGG | 9 | 0.483 | |
| 41 | -6822 | | GGUUGGUAAACCAACACCAUUAGUGGGUUGG | 6 | 0.433 | |

| | | | | |
|----|---------|-----------------------------|----|-------|
| 42 | -11,440 | GGCGGUGGUUUAGCACUAACUCUGG | 7 | 0.48 |
| 43 | -13,136 | GGUUAAGUGGUGGUCUAGG | 16 | 0.842 |
| 44 | -13,963 | GGAUCUGGGUAAGGAAGG | 19 | 1.111 |
| 45 | -16,623 | GGAUUUGGAUGAUCUAUGUGGCAACGG | 14 | 0.556 |
| 46 | -19,856 | GGUGAUAGAGGUUUUGUGGUGG | | |
| 47 | -19,865 | GGUGAUAGAGGUUUUGUGGUGGUUUGG | 19 | 0.92 |
| 48 | -19,874 | GGUUUGUGGUGGUUUGG | | |
| 49 | -23,877 | GGAUUUGGUUUGGUUUGG | 19 | 0.941 |
| 50 | -25,003 | GGUGGAAUGUGGUAGG | 17 | 1.063 |
| 51 | -27,432 | GGGGCUUUUAGAGGCAUGAGUAGG | 13 | 1.042 |
| 52 | -29,867 | GGUUGGUUUUGUUACCUGGGGAAGG | 13 | 0.783 |

Table 2

| | G4 position | Gene | Sequence | Found in other members of the Coronaviridae family |
|----|-------------|-------|--------------------------------|--|
| 1 | +353 | Nsp1 | GGCUUUGGAGACUCCGUGGAGGAGG | SARS-CoV Bat-CoV SARS-CoV Bat-CoV |
| 2 | +644 | Nsp1 | GGUAAUAAAGGAGCUGGUGG | BtRI-BetaCoV BtRs-BetaCoV BtRf-BetaCoV Rhinolophus-affinis-coronavirus |
| 3 | +644 | Nsp2 | GGCGGUGCACCAACAAAGGUUACUUUUGG | Bat-CoV |
| 4 | +644 | Nsp3 | GGAGGAGGUGUUGCAGG | BtRt-BetaCoV |
| 5 | +644 | Nsp4 | GGAUACAAGGCUAUUGAUGGUGG | Bat-CoV |
| 6 | +644 | Nsp5 | GGCUGGUAAUGUUAACUCAGGGUUUUGG | Bat-CoV SARS-CoV |
| 7 | +644 | Nsp10 | GGUAUGUGGAAAGGUUAUGG | Bat-CoV Rhinolophus-affinis-coronavirus SARS-CoV |
| 8 | +644 | Nsp12 | GGUUUCCAUUUAAUAAAUGGGGUAAGG | Bat-CoV BtRs-BetaCoV Rhinolophus-affinis-coronavirus SARS-CoV |
| 9 | +644 | Nsp12 | GGCGGUUCACUAUAUGUUAACCAGGUGG | Bat-CoV BtRI-BetaCoV BtRs-BetaCoV Rhinolophus-affinis-coronavirus SARS-CoV |
| 10 | +644 | S | GGUUGGACCUUUGGUGCAGG | SARS-CoV Bat-CoV Bat-CoV |
| 11 | +644 | S | GGCUUAUAGGUUUAAUGGUUAUUGG | BtRf-BetaCoV Rhinolophus-affinis-coronavirus SARS-CoV |
| 12 | +644 | S | GGCCAUGGUACAUUUGGCUAGG | SARS-CoV Bat-CoV |
| 13 | +644 | M | GGAUCACCGGUGGAAUUGCUAUCGCAAUGG | SARS-CoV Bat-CoV |
| 14 | +644 | N | GGCUUCUACGCAGAAGGGAGCAGAGGCGG | SARS-CoV Bat-CoV Rhinolophus-affinis-coronavirus |
| 15 | +644 | N | GGCUGGCAAUGGCGG | SARS-CoV Bat-CoV |

Table 1. PQSs that have been predicted or verified, all of which contain two G-quartets once folded.

Table 2. Conserved PQSs among coronaviruses.

Given discrepancies stemming from diverse prediction software and algorithms, these findings necessitate rigorous experimental validation.

The roles played by RNA G4s in the SARS-CoV-2 life cycle encompass various processes, as illustrated in **Figure 3**, including translation of Nsps and structural proteins, RNA transcription, RNA replication, and genome packaging.

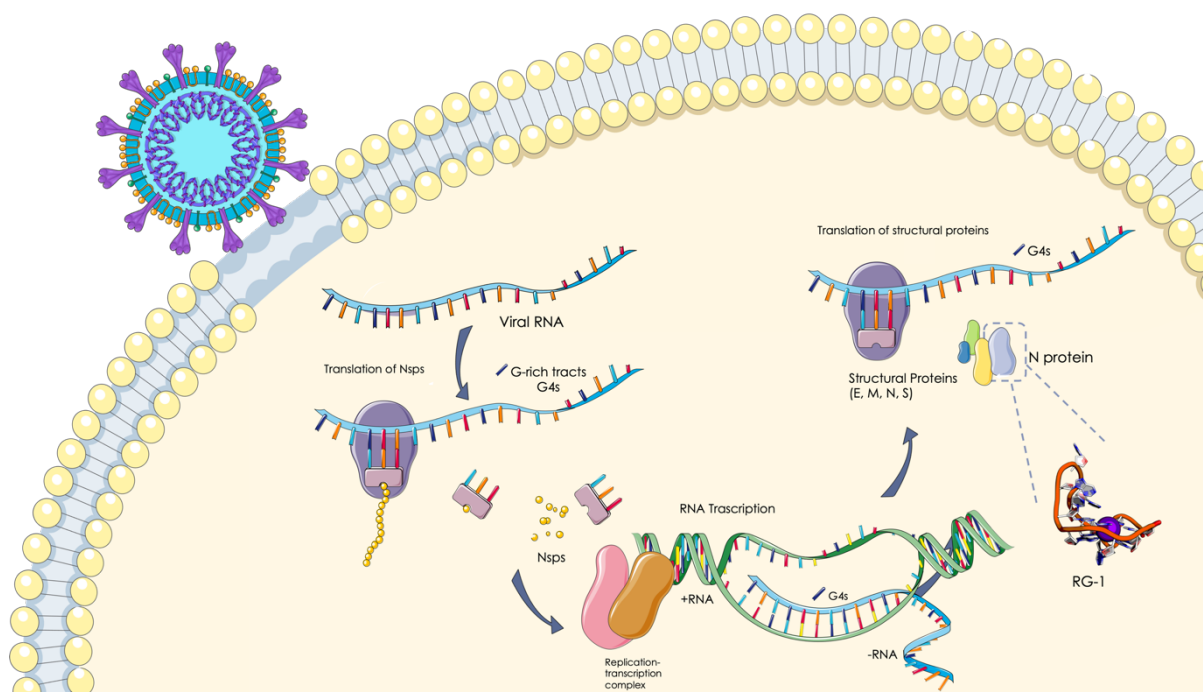


Figure 3. Potential roles of RNA G4s in the SARS-CoV-2 life cycle, detailing various stages of viral infection: release of the viral genome; Translation of Nsps; RNA transcription; RNA replication; Translation of structural proteins; Packaging of the genome; Formation of new virion; Release by exocytosis. RNA G4s within the virus may play a crucial role in modulating the efficiency of steps 2–6 and could serve as a target for antiviral therapy. For example, the G4-specific ligand PDS has the potential to stabilize RG-1 (at genome position 28,903) and diminish the translation of the nucleocapsid (N) protein.

Notably, the first computationally derived G4 structure of RG-1 at position 28,903 (a sequence in the coding region of the SARS-CoV-2 N protein) was recently reported, utilizing a multiscale approach combining quantum and classical molecular modeling¹⁶ and this is the structure that we used for our research. Recently, Zhao et al. conducted a comprehensive investigation using various experimental techniques, including fluorescence turn-on assays, circular dichroism (CD), nuclear magnetic resonance (NMR), and fluorescence resonance energy transfer (FRET), to demonstrate experimentally the formation of this G4 structure by RG-1 in vitro. Notably, this study marked the first confirmation of a PQS folding into a stable unimolecular G4 structure within live cells.

Subsequently, known molecules capable of stabilizing these structures were tested on these PQSs. For example, Pyridostatin (PDS), were shown to stabilize RG-1 and significantly reduce the expression levels of the N protein both in vitro and in vivo. Moreover, PDS was shown to promote the formation of the G4 structure and seems to inhibit N protein expression by targeting RG-1.

Conventional molecular dynamics simulations (MD) supported experimental findings by confirming the folding of RG-1 into a parallel G4 conformation, comprising two rigid tetrads and a flexible peripheral loop.¹⁷ Additionally, docking complexes have been identified of PDS and CX-3543 (quarfloxin, a G4 ligand derived from fluoroquinolones) with RG-1. These ligands bind to the G-quartet through $\pi - \pi$ stacking interactions and are primarily driven by dispersion and hydrophobic interactions. Both PDS and CX-3543 hold potential as valuable compounds for stabilizing G4 arrangements, particularly the parallel conformation of RG-1.

Berberine, a planar molecule with an extended π delocalized system, also exhibits the ability to interact with G-quartets of G4 structures through $\pi - \pi$ stacking interactions. Berberine, which has been used in traditional Chinese medicine for centuries and possesses antiviral, anti-allergic, and anti-inflammatory properties, was recently investigated for its binding properties to RG-1.¹⁸ Results indicated that berberine can interact with RG-1. Furthermore, it was observed that two berberine molecules bind to one RG-1 molecule. Additionally, RG-1 maintains a parallel conformation in both the ligand-free and ligand-bound states. These findings suggest that SARS-CoV-2 G4s may serve as promising therapeutic targets in the fight against this virus.

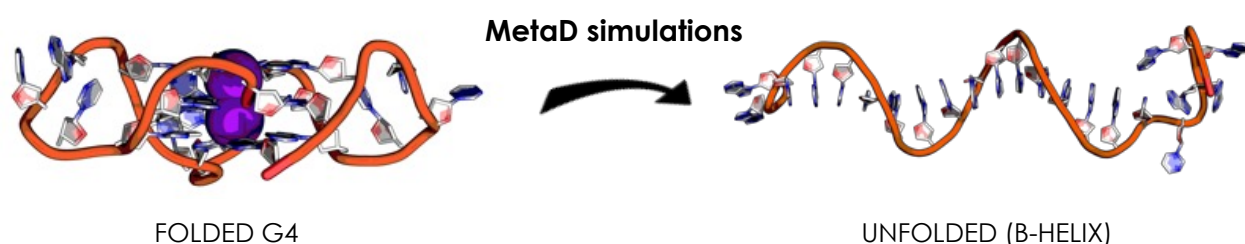
In addition to G4s in the SARS-CoV-2 genome, host G4s that regulate the expression of proteins crucial for virus entry can also be targeted for antiviral therapy. For example, there is a G-rich tract in the promoter of the human Tmprss2 gene which encode the transmembrane serine protease TMPRSS2, a crucial host factor for SARS-CoV-2 infection. This tract is capable of forming a G4 structure in the presence of K^+ , significantly affecting the transcription level of Tmprss2. PDS has been demonstrated to bind a G4 found in this region, PQS-675, and to attenuate the infection of SARS-CoV-2 pseudoviruses in human lung cells.¹⁹ These results indicated that Tmprss2 RNA G4 is a potential target for SARS-CoV-2 inhibition.

Apart from the G4 structure itself, interaction partners of G4 may also be targeted: G4-interacting proteins of SARS-CoV-2 from both the virus and the host cell have been discussed as potential targets of interest. For example, the SUD (Nsp3), Nsp13 and N proteins of SARS-CoV-2 are the main G4 regulators encoded by the virus. In addition, host

proteins that have been confirmed to be members of the RNA interactome of SARS-CoV-2 and have previously been shown to act as G4 regulators in other viruses, such as helicases, hnRNPs, nucleolin, and CNBP, are also presented because these proteins may serve as potential pharmacological targets to interfere with the normal functions of viral G4s. Since the structures of some G4-binding proteins are already known, it is possible to design inhibitors against these proteins to inhibit their interactions with G4s.²⁰

Finally, due to their high stability, convenient operation, and low cost, G4 structures could also be used to detect SARS-CoV-2.²¹

1.17 Unraveling the G-Quadruplex Mystery: Exploring the (un)folding mechanism



The above reported considerations suggest that G4s represent an intriguing target for the development of antiviral drugs. However, the discovery that DNA and RNA sequences can adopt secondary structures beyond the conventional alpha-helix dates to the early 1900s and the idea that G4s could serve as therapeutic targets was initially conceived in the 1960s and 1970s. Since then, considerable efforts have been made to unravel the complexity of these structures and exploit their therapeutic potential.

However, G4s hold importance beyond therapeutic applications, as these structures are associated with DNA damage and genome instability. Consequently, considerable efforts have been devoted to understanding the role of G4s in cancer biology. This exploration has led to the evaluation of small-molecule ligands that target G4s: in fact, more than 1,000 molecules have been developed with the capability to interact with G4 targets.

Despite these extensive efforts, only a limit number of ligands have progressed to the clinical phase, and none are used as drugs. The primary obstacle lies in their modest ability to differentiate between individual G4 structures, often resulting in undesirable side effects.

A fundamental challenge, therefore, is to improve the selectivity of these molecules. For this reason, identifying specific conformations of the G4, which could be targeted by small molecules, may help designing new G4-ligands.

To this end, we need to understand the determinants of folding and stability of the structures of G4s. DNA and RNA sequences rich in guanine bases possess the inherent capacity to self-assemble into coplanar, cyclic structures known as tetrads, visualized in **Figure 4**. These tetrads consist of four guanine bases held together by eight Hoogsteen hydrogen bonds and are referred to as G-tetrads or G-quartet. These G-tetrads can stack one atop the other forming G4s, as visualized in **Figure 4**. These stacked tetrads are stabilized by $\pi - \pi$ stacking interactions (nonbonded attractive forces).

The stability of the G4 structure is further enhanced by the presence of positively charged ions (e.g., K^+ and Na^+), which play a crucial role by coordinating two consecutive G-tetrads. This coordination helps shield the electrostatic repulsion between the carbonyl oxygens of guanine bases.

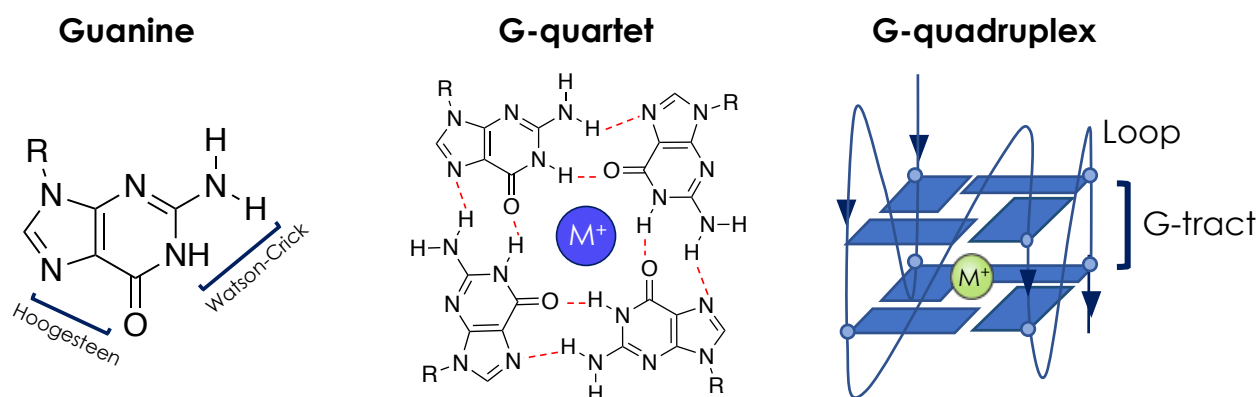


Figure 4. G-quadruplex (G4) structure: A. Structure of the nucleobase guanine. B. Chemical structures of G-quartet. Structural arrangement of the G-quartet, highlighting the hydrogen bonding network between the Hoogsteen and Watson–Crick faces of the coplanar guanine bases with a centrally placed metal ion (M^+). C. G4 formed by the stacking of two G-quartets/tetrads.

G4-DNA structures exhibit remarkable polymorphism. They can adopt a diverse range of topologies (**Figure 5**). These different topologies will be described in detail below. The structural variations of G4s are influenced by various factors: the number of DNA strands involved (ranging from one to four), the orientation of these strands (parallel, antiparallel, or hybrid), the glycosidic conformation of guanine bases (syn or anti), the length of intervening loops, and the stretches of guanine bases within the structure.

In contrast, G4-RNA structures are more constrained in their diversity, with a tendency toward predominantly parallel topologies. This limitation primarily arises from the anti-conformation of the glycosidic bonds in RNA and the presence of an additional 2'-OH group, which leads to enhanced hydrogen bond networks within the structure.

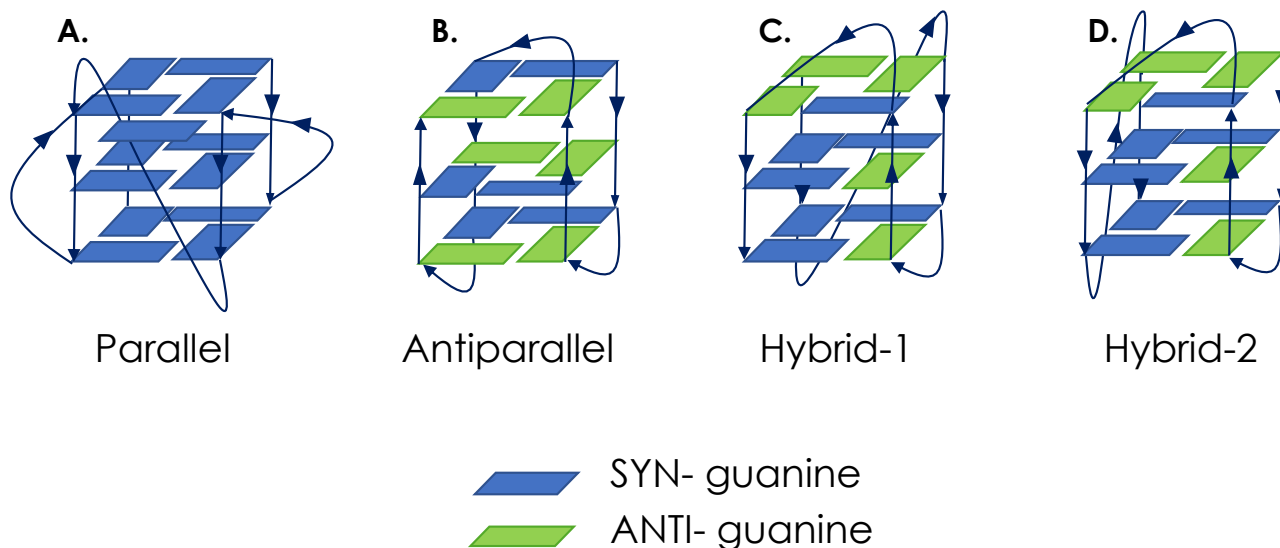


Figure 5. Schematic representation of the human telomeric G4-DNA folding topologies. **A.** Parallel or propeller type, as identified by X-ray in presence of K^+ . **B.** Antiparallel or basket-like, as detected in Na^+ solution. **C.** Hybrid type 1 and **D.** hybrid type 2, both found in K^+ solution. Syn and anti guanines glycosidic bond orientation are colored in blue and green, respectively.

Intramolecular G4s (GQs) consist of loops, which are segments of nucleotides that connect the G-strands (also referred to as G-stretches or columns). These loops serve as integral components in defining the GQ structure and its characteristics.

So-called propeller loops connect two neighboring G-strands that run in the same direction, which is known as the parallel orientation. They are aptly named "propeller loops" due to their role in maintaining this parallel alignment.

Lateral loops, on the other hand, connect two neighboring G-strands that run in an antiparallel orientation. These loops facilitate interactions between antiparallel strands.

Another topology is the diagonal and here the loops have the distinct feature of connecting two antiparallel G-strands that are positioned across from each other within the G4.

The directionality of the G-strands is closely linked to the syn/anti conformation of the guanine bases' glycosidic torsional angles χ . Specifically, if two guanines within a quartet are in mutually parallel strands, they will exhibit the same χ conformation. Conversely, if

these guanines are situated in antiparallel strands, they will display opposite χ conformations.

The human telomeric sequence d-(GGGTTA)_n is a prime example of such a highly polymorphic sequence, forming at least six known stable GQ conformations.²²⁻²⁹

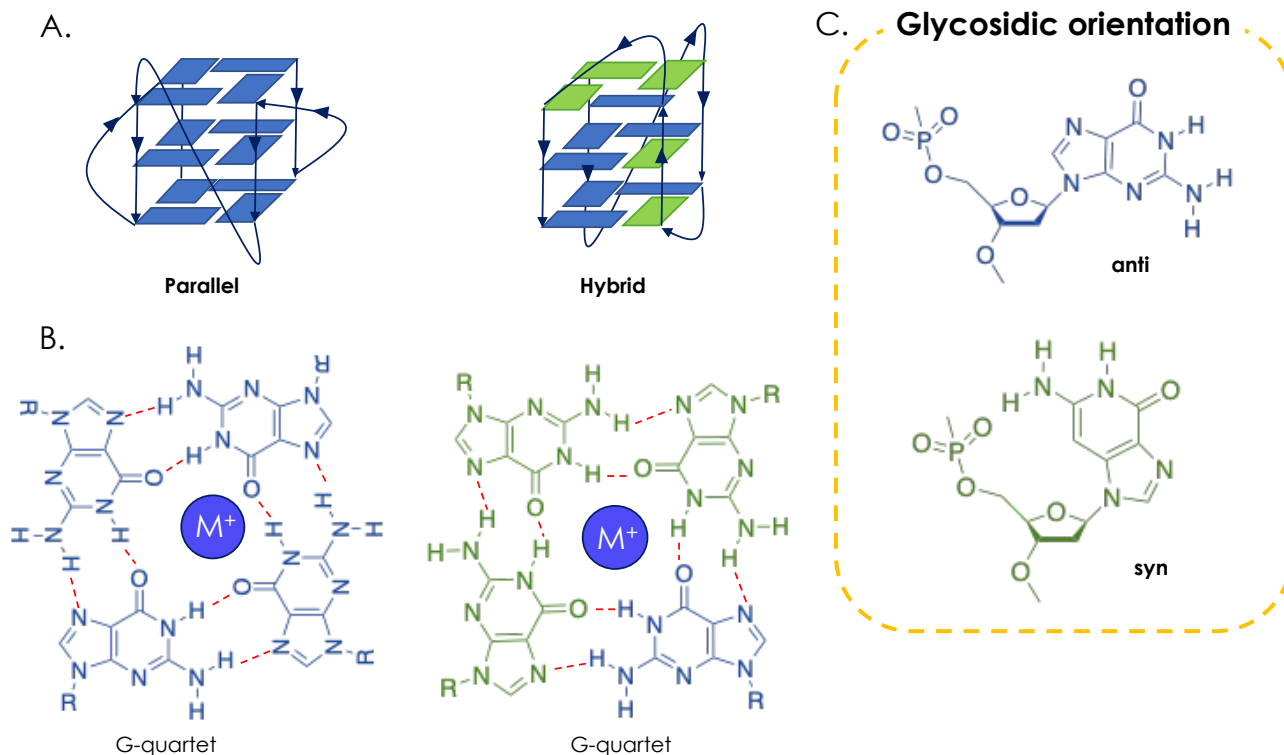


Figure 6. Insights into the various G4 topologies: **A.** Presents schematic representations of three-quartet parallel and hybrid G4s. Bases in anti-conformation are shaded in blue, syn in green, and blue spheres represent metal cations (M⁺). **B.** Structural formulas of the top quartets of G4s from **A.**, with hydrogen bonds indicated by dashed lines in red. **C.** Shows the anti and syn guanosine orientations.

Thus, the goal here is to develop a computational method to explore the intricate processes of G4s structural organization, exploring their unfolding mechanisms to ultimately discover the fundamental principles that govern the intricate stabilization process of G4 structures. The overall goal of this work is to gain a comprehensive understanding of these dynamic structures, and thereby enable future development of precise and selective tools that can interact with G4s. In fact, through meticulous examination and characterization of the unfolding mechanisms, we aim to shed light on the factors and conditions influencing the stability, dynamics, and structural polymorphism of G4s.

1.17.1 G4 Folding and Unfolding Mechanisms.

Recent studies have analyzed the complexity of G4 folding, suggesting that it can be most accurately described as a process characterized by state model, also called “kinetic partitioning landscape”³⁰.

This process cannot be described by the original "folding funnel" mechanism that has occasionally been used in the context of intramolecular DNA G4s because a simple funnel may not accurately capture the folding dynamics of G4s. In a conventional funneled free-energy landscape, molecules smoothly decrease in energy and configurational entropy as they progress towards the native state. This progression is marked by a continuous increase in the number of native contacts. In the idealized model, there are no significant kinetic traps, featuring local energy minima, substantially deeper than thermal fluctuations, along the way. Such a funneled landscape would lead to fast folding events and minimal frustration due to non-native interactions.

In contrast, the folding landscape for G4 DNA sequences, as evidenced by numerous experiments, do not align with this funnel-like behavior. G4s can exhibit folding timescales spanning up to several days, indicating a more complex folding landscape. Supposedly, this landscape includes deep, competing free-energy minima, often representing alternative folds, or competing conformational ensembles separated by substantial free-energy barriers. Only a fraction of the molecules folds directly into the native basin of attraction (NBA), which is the most populated at thermodynamic equilibrium. The majority of molecules within the ensemble initially fold into competing (non-native) basins of attraction (CBA), where they become trapped. This interplay between native and non-native basins slows down the folding process significantly.

Investigating such folding processes is inherently more challenging compared to fast, funnel-like folding. The same argument applies to protein folding, which in any case can neither be described as a direct transition from unfolded to folded. Instead, it is studded with intermediate states, which are separated by a notable energy barrier.

The precise mechanisms governing transitions between different basins are system dependent and hardly ever fully understood at the atomistic level. Only the folding of very small so-called fast-folding proteins could be resolved in reasonable detail.³¹ It remains unclear to what extent misfolded molecules must unfold before transitioning to another basin. Additionally, it's essential to recognize that the classification of folded, misfolded, and

unfolded states or structures depends on the resolution and definitions of specific experimental methods.

The folding of G4s has been divided into two distinct stages. The first is marked by the rapid folding of the initial G4 ensemble: during this phase, various G4 structures are formed, but they tend to be predominantly misfolded. These misfolded G4s may exhibit non-native syn/anti combinations of guanine bases or a reduced number of tetrads when compared to fully folded G4s.

The second stage of the folding process involves the gradual refinement of the initial GQ population, leading to the eventual formation of the native GQ structures. This refinement process occurs slowly and is responsible for transitioning the initial, mostly misfolded GQs into the final native GQs. These structural transitions occur through the various other structural ensembles, as incomplete or perturbed GQs, G-triplexes, G-hairpins, and cross-like species.

It is important to note that the nature of these intermediate ensembles can vary depending on external factors. Factors such as temperature, ionic strength, the presence of cosolvents, the binding of ligands, and the specific types of cations in the environment all play a role in influencing the structural transitions and kinetics of GQ folding.

1.17.2 Molecular Dynamics simulations of G4 folding and unfolding

In this study, we use all-atom explicit-solvent enhanced-sampling molecular dynamics (MD) simulations to explore the unfolding dynamics of different topologies of G4s: parallel (both from DNA and RNA sequences), antiparallel and hybrid.

Our research reveals several key findings. Our primary goal was to establish an enhanced sampling scheme to investigate the unfolding of G4. Given the long lifetime of the folded G4, it is basically impossible to sample the unfolding of G4 in conventional simulations. While there have been other attempts to use enhanced sampling to this end,^{32, 33} we wanted to establish a new method, which does not bias the unfolding towards specific pathways.

Generally, MD simulations serve as valuable tools for the in-depth study of transient ensembles and dynamic processes, offering detailed spatial and temporal resolution. In the context of investigating various aspects of G4 folding, MD simulations have been employed multiple times.³⁴⁻⁴⁷

Standard MD simulations are, however, constrained by limitations in affordable sampling. To overcome this limitation, researchers often turn to enhanced sampling methods. These

methods can broadly be categorized into two groups: those that modify the total energy of the system and those that operate on specific low-dimensional projections, known as collective variables (CV), on the free energy surface (FES). Examples of the former category include replica-exchange methods while metadynamics falls into the latter category.

In our present research, we used MD simulations to investigate the unfolding of G4 structures: to accelerate the sampling process, we employ metadynamics.⁴⁸ Our simulations specifically target the unfolding of a complete G4 structure with different topologies (parallel, antiparallel and hybrid) to discover the respective unfolding pathways.

The challenge lies in unraveling the intricate unfolding mechanism of G4s these secondary structures because this process follows diverse and complex pathways, sometimes referred to as "kinetic partitioning". This term aptly allows the existence of multiple routes from the unfolded state to the folded state. Consequently, obtaining experimental proof of the precise folding mechanism remains elusive: isolating intermediate states proves exceedingly difficult due to their inherent instability and sensitivity to the specific experimental conditions employed. Nevertheless, various proposals for both folding and unfolding mechanisms have been made, using both experimental and computational methods.

The folding process potentially includes a series of steps, commencing with the collapse of unfolded species into hairpin-like structures with G:G base pairs.⁴⁹⁻⁵² These structures subsequently transform into antiparallel G4s,^{53, 54} transitioning progressively into triplex intermediates^{35, 55, 56} before adopting the proper G4 conformation⁵³. This model follows a sequential pathway and does not account for side-reactions or 'branched' pathways, avoiding confusion with 'parallel' reactions in chemical kinetics when discussing mechanisms.

Alternative models propose branched mechanisms where the unfolded species, intermediates (triplex or other), and the G4 are in equilibrium, either with or without intermediates.⁵⁶⁻⁵⁸ Based on MD simulations, Stadlbauer et al. suggest a complex branched mechanism for the folding of human telomeric sequences.^{40, 55} Previous studies have proposed various G4 folding pathways, including intermediates such as hairpins and triplexes.

In 2016, Gabelica et al.⁵⁹ proposed a generalized folding landscape based on experimentally obtained kinetic data, which unveil a complex mechanism with several branched reactions, resulting in multiple 'branches' within the folding landscape.

However, Gabelica and coworkers were unable to definitively confirm or rule out the presence of a triplex structure. Additionally, molecular modeling indicated that both hairpins

and G-triplexes remain stable only when their constituent strands are in an antiparallel orientation. This suggests that these intermediates likely assume an antiparallel configuration during the folding process.

The folding landscape of G4 differs from the traditional sequential folding pathway, demonstrating the influence of misfolded species on the overall folding kinetics. Nucleic acids, particularly G4 folding, display a distinctive folding energy landscape due to numerous hydrogen bonds and ionic interactions, creating significant barriers between ensembles. Consequently, misfolded structures persist on a second to minute timescale, impacting the overall folding dynamics, setting nucleic acids apart from proteins in terms of folding behavior.

In a very recent paper, Stadlbauer and Šponer³³ use a combination of the REST2 scheme coupled with well-tempered metadynamics simulations to provide evidence in the folding process of a parallel topology. They suggested the initial formation of a compacted, coil-like ensemble of G-strands as the starting point for the folding of a fully parallel three-quartet G4, specifically (GGGA)₃GGG. From this ensemble, the G4 structure gradually emerges via multiple pathways, devoid of a distinct intermediate structure. The folding progresses through a sequence of incremental conformational alterations, involving cross-like structures, hairpins, slip-stranded arrangements, and two-quartet G4 ensembles.

This coil-like ensemble is believed to coordinate at least one monovalent cation. The simulations also suggest that an essential early phase of the folding process involves the stacking of guanine units in G-tracts (G-strands). These stacked G-tracts then aggregate into larger, semi-rigid blocks, streamlining the exploration of conformational space by the DNA chain, akin to the diffusion collision model observed in protein folding. Notably, isolated G-hairpins prove unstable but can find support through additional interactions within the compacted coil-like ensemble.

A noteworthy observation is that each folding event follows a distinct trajectory: there is just the common initial step which typically involves the formation of a coil ensemble, coordinated with at least one ion. This coil ensemble is a broad set of compacted structures, stabilized by hydrogen-bond interactions between guanine bases, and was previously proposed as a potential intermediate in parallel G4 folding.

Subsequently, they predominantly observed the formation of hairpin ensembles, slip-stranded ensembles, and two-tetrads G4s along these trajectories. Ideal cross-hairpin ensembles were formed in only two folding trajectories. However, the coil ensemble frequently exhibited cross-like states, featuring interaction between four guanines rather

than six. Another variant observed was a tilted ensemble with fully stacked G-tracts, resembling a tilted hairpin ensemble but not entirely parallel. While the ideal tilted ensemble was rarely observed, structures with a tilted-like configuration and four interacting guanines were more common.

Triplex structures emerged in two folding events, occurring just before the complete formation of a G4 in one of them. The other four folding events bypassed the triplexes through different routes. In the final stages of folding, once two quartets were fully formed, the third layer typically comprised two or three guanines. The remaining guanines were either exposed to the solvent or stacked with the loop adenines. After all guanines were folded and the last tetrad was formed, the second cation coordinated between the tetrads, serving as the concluding step. Standard unbiased simulations initiated from nearly-folded snapshots confirmed this chronological order, with the final quartet forming first, followed by coordination of the second cation.

Clearly, folding can occur with very different pathways. Accordingly, the same holds true for the unfolding of these structures.

1.17.3 Methodology

Our study involved enhanced-sampling unfolding simulations of different topologies of G4 (parallel both from DNA and RNA sequences, antiparallel and hybrid). The primary objective was to establish a method to investigate their unfolding mechanisms and to assess their stabilities. To achieve this, we employed metadynamics (MetaD) using the coordination number as a collective variable (CV), which effectively captured the G4 unfolding event in every simulation of every topology.

The CV that has been used is the coordination number, CONUM, which calculates the number of contacts between two groups of atoms. For a better explanation of this property see the Materials and Methods section. For the evaluation of the eventual existence of a contact, we used a well-established approach: we evaluated a logistic function based on the respective pair distance d_{ij} .

In this MetaD protocol we used as group A the heavy atoms of DNA (or RNA) sequence and as Group B the oxygen of the water molecules surrounding the G4, as show in **Figure 7**.

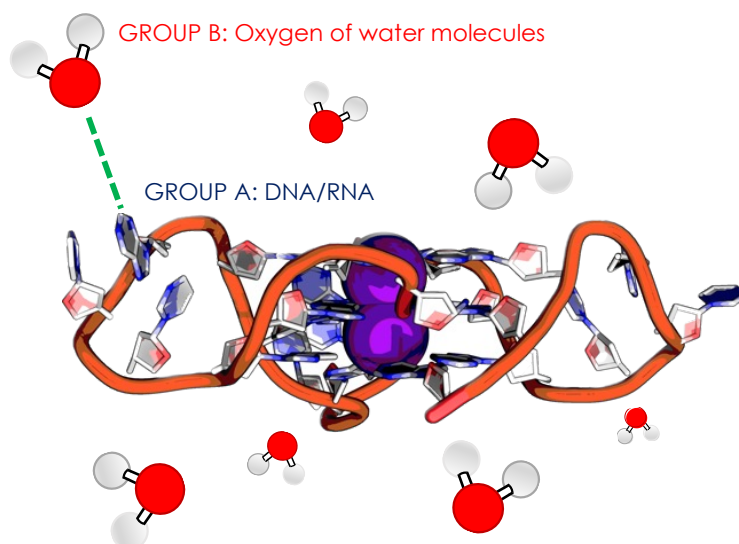


Figure 7. Choice of Coordinate Variables for MetaD. The figure illustrates the selected coordinate variables for the simulations, the CONUM, highlighting two key groups: oxygen atoms of water molecules (depicted in red), atoms comprising the DNA or RNA sequence (depicted in grey) and the magnesium ions (depicted in violet, present in the crystal structure (PDB 1KF1) and fundamental for the stability of the structure).

For our study, we utilized the telomeric 22-mer sequence d[AGGG(TTAGGG)3], capable of forming an anti-parallel topology in the presence of Na^+ (143D), while in the presence of K^+ (1KF1), it adopts a parallel-type topology. Intriguingly, the telomere G4 in a K^+ solution assumes a hybrid-type topology, deviating from its crystalline state. Phan and Petal investigated the topology in telomeric G4 using the d[TAGGG(TTAGGG)3A] sequence, where adenine was added to the 3'-end of the native sequence d[TAGGG(TTAGGG)3] (2GKU) and we used these sequences to simulate the hybrid topology (**Figure 8**).

For each topology, we carried out 5 replicas, each lasting 200 ns. However, it was observed that to capture the unfolding of the structure, a simulation duration of 100 ns would have been sufficient. The unfolding process and relevant dynamics were effectively captured within this timeframe, indicating that a shorter simulation duration could have provided equally meaningful insights into the unfolding mechanisms for the respective topologies.

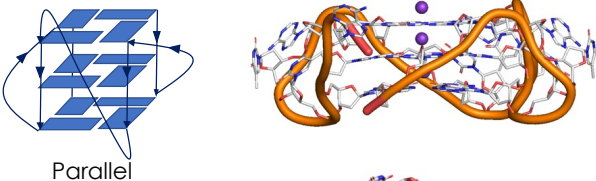
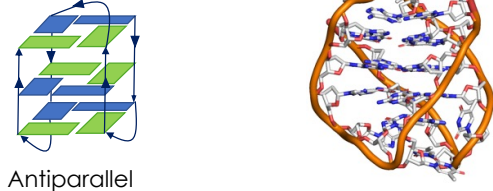
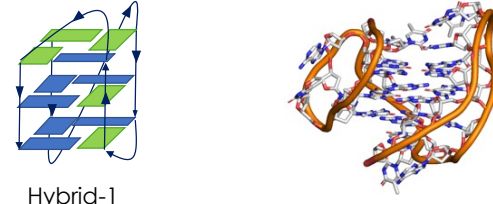
| SEQUENCE | PDB ID | STRUCTURE |
|---|--------|--|
| <u>AGGGTAGGGTAGGGTAGGG</u> <i>Native</i> | 1KF1 |  Parallel |
| <u>AGGGTAGGGTAGGGTAGGG</u> <i>Native</i> | 143D |  Antiparallel |
| <u>TTGGGTAGGGTAGGGTAGGG</u> <u>A</u> <i>Modified</i> | 2GKU |  Hybrid-1 |

Figure 8. G4s Topologies. For the parallel topology we used the telomeric 22-mer sequence d[AGGG(TTAGGG)3], known to adopt a parallel conformation in the presence of K^+ (PDB ID: 1KF1) and the anti-parallel conformation in the presence of Na^+ (PDB ID: 143D). For the hybrid topology, we chose the d[TAGGG(TTAGGG)3A] sequence, where adenine was added to the 3'-end of the native sequence d[TAGGG(TTAGGG)3] (PDB ID: 2GKU). We simulated every topology coordinated with K^+ in the central cavity. For each topology, we conducted 5 replicas, each lasting 200 ns.

1.17.4 Observed unfolding pathways in different topologies

Generally, we found that our method is always able to unfold G4. By changing the strength of the bias, we may decide to modulate the unfolding of the G4, faster or less fast. We choose the bias height to be able to perform multiple simulation runs for the different replicas within a reasonable computation time, i.e., around 100 ns. For two of three topologies we saw very diverse unfolding pathways, whereas for the third topology, the antiparallel, the unfolding trajectories were very similar.

To better understand the differences but also the similarities between the various topologies of structures we divide the results according to the respective topologies, i.e., parallel, antiparallel and hybrid (as described in earlier sections).

Parallel topology

We observed diverse unfolding pathways in different simulation runs with this topology. However, certain consistent species were identified across all replicas. Specifically, prior to

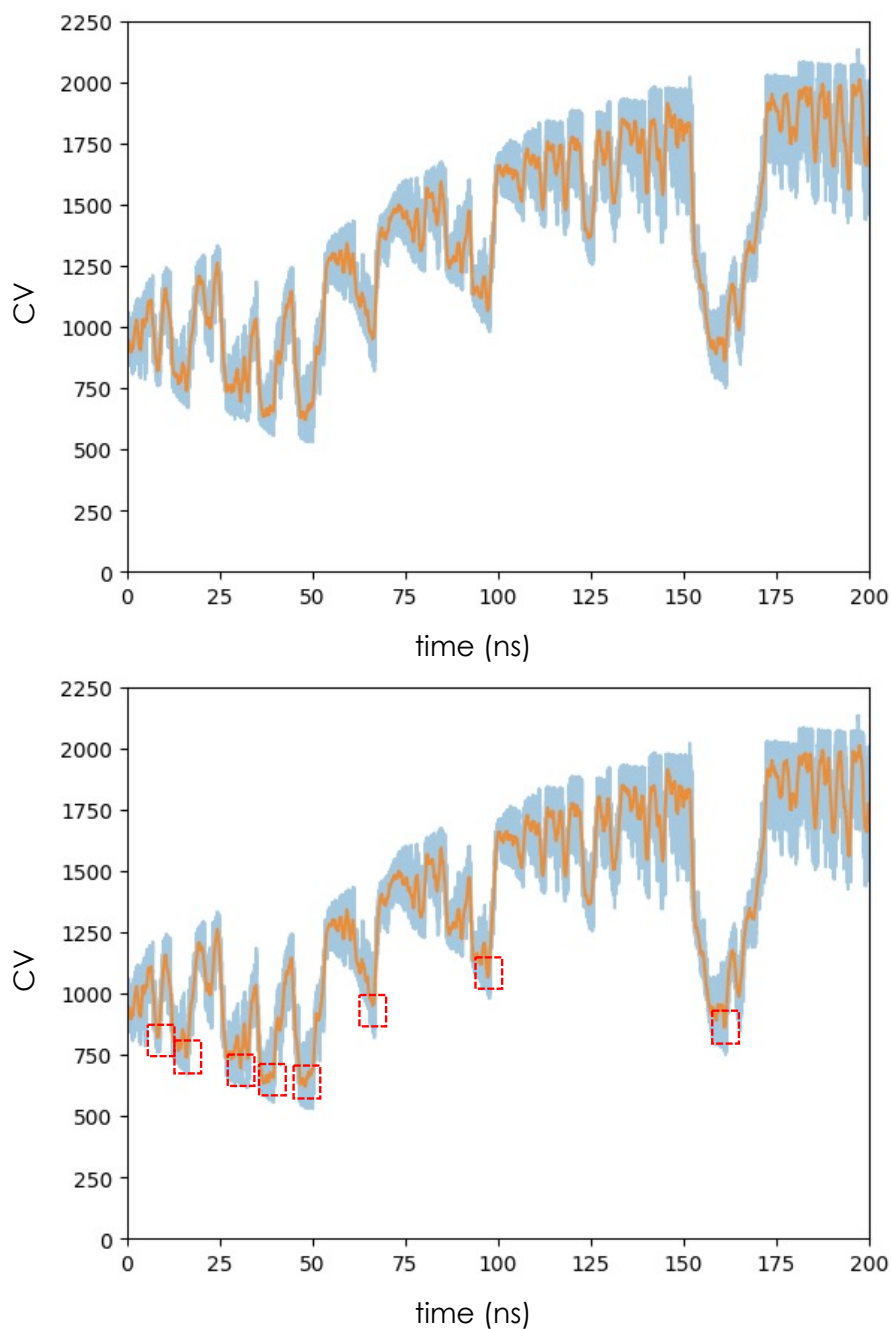
the G4 destabilization, a consistent event was noted involving the terminal 5' adenine (DA5') stacking onto the terminal tetrad.

As this adenine shifts away and ceases its interaction with the tetrads, it marks the initiation of G4 breakdown. This crucial event is present in all replicas. The differing factor lies in whether it is the terminal G directly bound to the terminal adenine or a G from the opposite tetrad. This initial movement leads to the formation of a stable structure that can either revert to reform the G4 or proceed further by shedding an additional guanine from the same tetrad. This progression results in a "two-layered G4" structure characterized by two tetrads and the other one with only two guanines (Figure 9).

Subsequent steps in the unfolding process may follow different paths but they all lead to the loss of a second tetrad. This transition ultimately results in a structure with only one remaining intact tetrad, which was originally the central tetrad of the G4. At this stage, the structure becomes unstable and rapidly unfolds, leading to the disintegration of the last tetrad.

However, an interesting observation at this point is that stacking interactions between two guanines persist for a significant duration during this process, which could mean that they act as semi-rigid building blocks during the folding process.

The development of the value of the coordination number (CV) as a function of time can be seen in Graph 1. Here, for simplicity's sake, we report the value of CV as a function of time for the first replica. From this, we have extrapolated the structures corresponding to the various minima of the simulation and we report these structures in Figure 9.



Graph 1. CV vs time (ns). MetaD simulation of replica 1, coordination number (CV) in function of the time (ns). The dashed rectangles correspond to the minima of the CV from which the structures representing the intermediate states in the unfolding process were extracted. The last rectangle (with a CV value around 800) corresponds to the B-helix.

FOLDED G4

Interaction with DA5'

1G away

2G away

2 TETRADS

1 TETRADS, 3G

1 TETRADS, 2G

Various disordered states

UNFOLDED

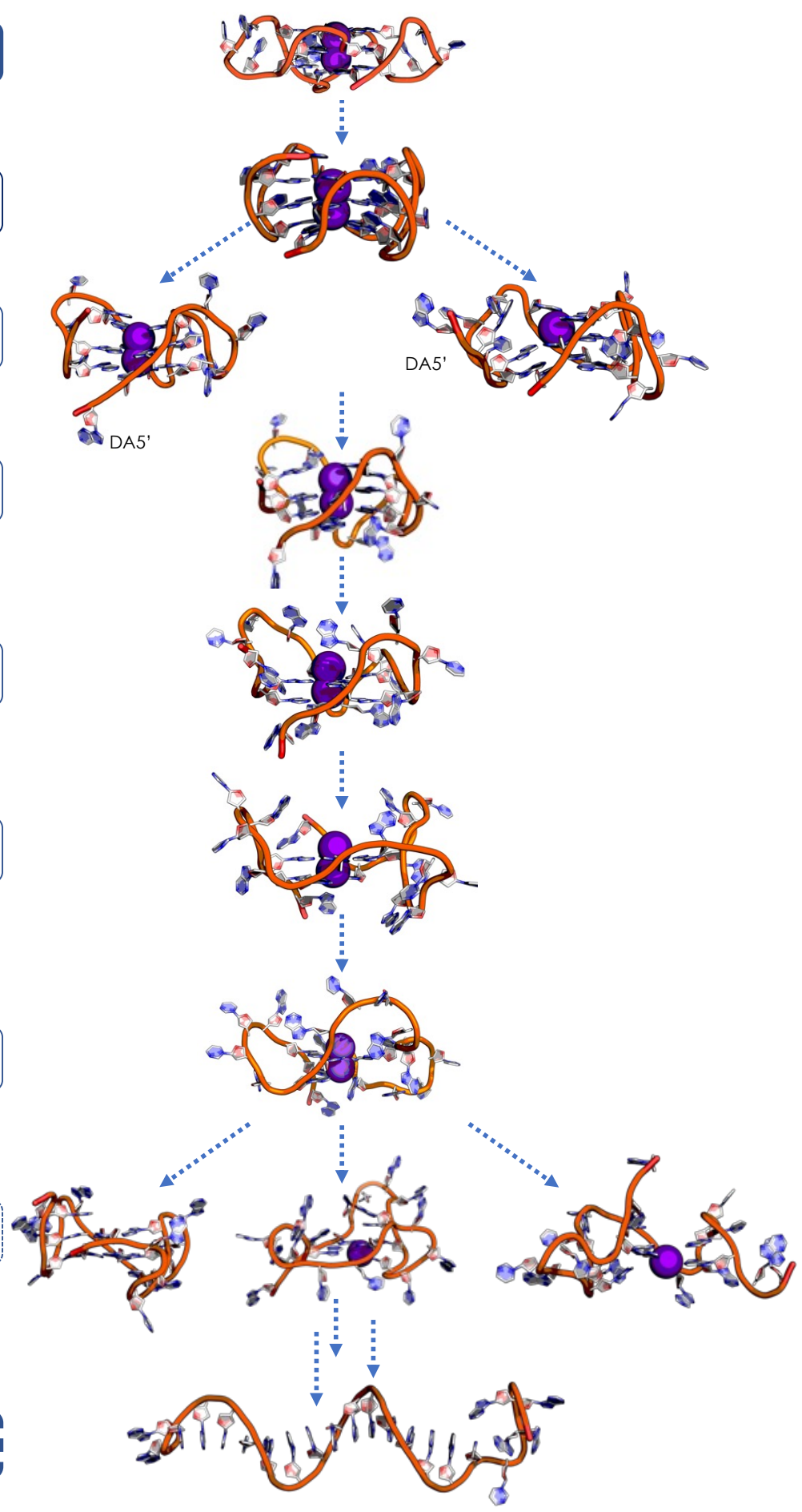


Figure 9. Proposed Unfolding Path for the Parallel Topology. The depicted steps represent the common states observed across various replicas, although the specific pathways to these intermediate states may vary from one replica to another. The initial cluster consistently represents the perfectly folded G4, with a notable configuration where one adenine (specifically, the adenine at the 5' end, DA5) is positioned atop the tetrad; this intermediate state is termed "interaction with DA5". Subsequently, a consistent event is the loss of one guanine, which can manifest in either of the terminal tetrads, "1G away". This leads to the loss of the second on the same tetrad ("2G away") and then the whole tetrad ("2 tetrads"). The same thing happens to the other terminal tetrad where, first, one G moves away, and this structure with one tetrad formed and another with only 3 guanines remains for a while, ("1 tetrad, 3G") and could be stabilized by the same adenine as before (DA5'). But after this adenine moves away and another guanine from the plane with only 3 Gs also moves away, ("1 tetrad, 2G") the structure begins to break down completely. It's important to note that numerous unstable intermediate states may exist before the complete rupture of the ex-G4. Notably, a state that consistently emerges is the B-helix, characterized by a remarkably low coordination number, even lower than that of the folded G4. This state is a significant observation in the unfolding pathway.

It is worth noting that the unfolding process can skip the formation of a G-triplex structure, aligning with the findings of a recent study. The unfolding mechanism emerging from these simulations seems precisely to require two steps: first, a very slow destruction of the first terminal tetrad that leads to the formation of rather stable structures that can reform the G4. But once the other outer tetrad is destroyed this process leads to the immediate breakdown of the structure and the formation of disordered coils. It is interesting to note that we obtained in every simulation the complete strand of B-DNA helix and the solvation number is almost the same as the folded G4 explaining the stability of these structures.

The same unfolding mechanism was observed in simulations of the human telomeric RNA G4 (called TERRA G4, PDB ID 6HHJ).

In contrast, the G4 structure found in the virus necessitates further analysis because, despite its stability during classical MD simulations, it exhibits significant instability when subjected to MetaD simulations. Hence, careful parameter modulation for the MetaD simulation is essential: the structure is less stable considering that it is formed by two tetrads instead than three, as in the telomeric sequences. Nonetheless, the unfolding mechanism appears to resemble that of the other RNA structure.⁶¹ It's important to note that these two RNA sequences (the human telomeric and the virus sequences) were simulated using different force fields, highlighting the need for a nuanced understanding of the specific force field impact on simulation outcomes.

Antiparallel topology

Antiparallel topology presents a more straightforward and consistent mechanism across all replicas. Specifically, the G4 initiation of breakdown begins when the 5' end initiates movement, disrupting interactions within the terminal tetrad. In this scenario, as DA5' moves away, the guanine linked to it also shifts, resulting in the disruption of that tetrad. Due to the antiparallel topology, all guanines are interconnected through the same loop. Consequently, what is observed is a simultaneous disengagement of all overlapping Gs constituting the G4 structure. This leads to the formation of a transient triplex momentarily, followed by the detachment of another guanine strand, ultimately resulting in a structure resembling the one depicted in the **Figure 10**.

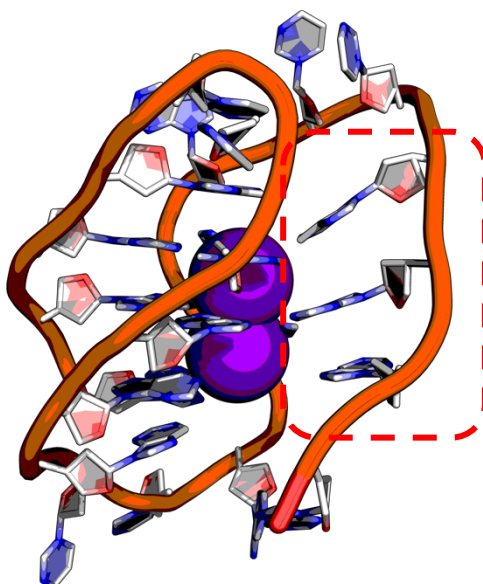


Figure 10. Triplex-like Structure. The rectangle highlights three guanines that undergo a bending motion, forming a triplex-like structure. These guanines are subsequently displaced away from the plane, contributing to the unfolding process of the G4 structure.

Before breaking down, G4 forms a high-order structure, adding to the interactions of the guanines of the tetrads also those of some adenines (**Figure 11**). These very stable structures have a lower coordination number than G4 itself because they are obviously much more closed and can interact with fewer water molecules. These very high-order structures are present in all the replica.

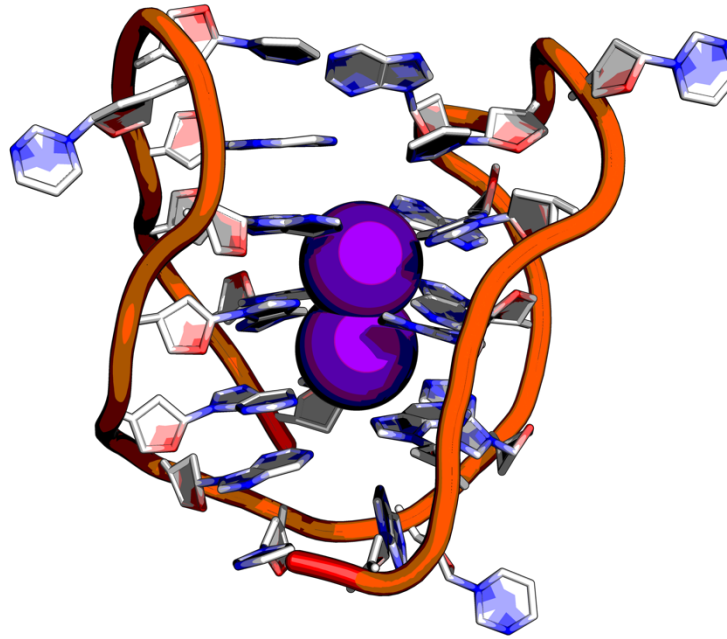


Figure 11. Pre-unfolding: Formation of High-Order G4 Structure. Prior to undergoing unraveling, the G4 structure attains a high-order conformation, incorporating interactions not only amongst the guanines within the tetrads but also involving certain adenines. These exceptionally stable structures exhibit a lower coordination number compared to the G4 structure, owing to their notably compact nature, allowing interactions with fewer water molecules. Remarkably, these high-order structures are consistently observed in all replicas, underscoring their significant presence in the unfolding pathway.

Despite this, a pair of interactions persists, maintained between adjacent guanines and the two underlying pairs of Gs. This continuity arises because these guanines are bound together by the same loop (Figure 12).

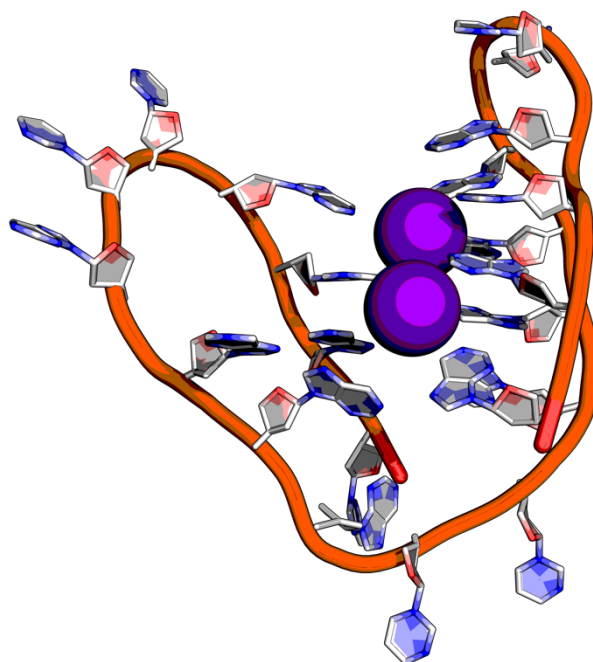


Figure 12. Intermediate State Preceding Quadruplex Dissociation: This intermediate state occurs just before the complete dissociation of the quadruplex structure. Consecutive guanine pairs within the same loop remain in close interaction, representing a transitional arrangement in the unfolding process.

In one replica, on the other hand, the interaction between pairs of facing guanines is maintained and not those linked by the same loop.

In the case of this topology, the mechanism appears less stepwise, and the sequential breakdown of tetrads is not readily discernible. Instead, the rupture occurs mainly because a column of guanine folds over and moves away.

Hybrid topology

The hybrid topology, on the other hand, is much more reminiscent of the mechanism of the parallel topology. Again, the mechanism begins with the loss of a guanine from one of the tetrads. This intermediate structure, however, is stabilized by an adenine that is placed on top of that tetrad. In this case, however, the mechanism seems as if after the loss of this guanine, the two overlying Gs are taken away from the loop to which they are bound, like a strand slippage.

But in the simulations of this topology, we can clearly see the formation of the triplex (in replica 3 and 4) **Figure 13**.

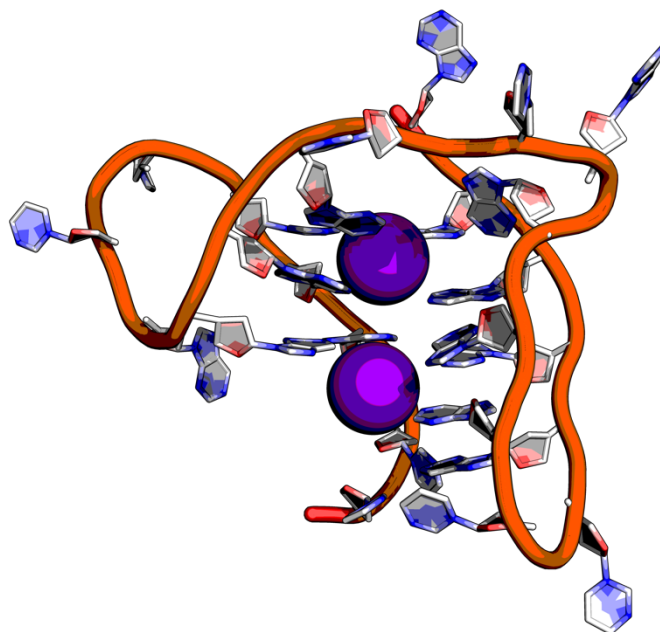


Figure 13. Triplex-like intermediate state found in all the replica of the hybrid topology.

The initial change in the coordination number, which first begins to rise and then has an abrupt descent is since the G4 begins to open and only two planes of the tetrads remain intact (the other one has only three guanines left to interact) but soon after the loop closes again forming much more ordered and compact structures than the initial structure.

This unfolding pathway seems more stepwise than the antiparallel topology but less so than in the parallel topology, where the breakage occurred from tetrad to tetrad, in this case there is the initial loss of a guanine from one of the tetrads but then one of the guanine columns is seen to move away and at that point the triplex formed is not stable and the G4 immediately unfolds.

In the replica 5, on the other hand, there is an initial opening of G4, only two tetrads remain formed and one with 2 Gs. Subsequently, all three tetrads close and reform and a high-order structure is formed, resembling that of antiparallel topology. When this structure breaks down, and it does so starting from the bottom loop, the structure collapses and opens without the intermediate formation of the G-triplex.

1.17.5 Discussion

Our primary and crucial objective here was to develop an efficient method capable of unfolding G4 sequences within a reasonable simulation timeframe. To achieve this, we utilized enhanced-sampling techniques, specifically employing MetaD simulations. Through this innovative approach, we successfully achieved the unfolding of various G4 topologies within approximately 100 ns, and this duration could be easily accelerated or decelerated by changing the bias height.

We applied this developed method to investigate the diverse topologies of the human telomeric G4 (including parallel, antiparallel, and hybrid conformations). Our aim was to glean valuable insights into the process of G4 unfolding, comprehend the importance of specific intermediates, and, at its core, visualize these intermediate structures. Through this exploration, we sought to anticipate potential structures that might manifest during the unfolding process.

In this regard, the MetaD method demonstrated its effectiveness by consistently facilitating the unfolding of the G4 structure across all simulation replicas. Additionally, the choice of the CV for MetaD appeared optimal for achieving this purpose, contributing to the successful unfolding of G4 structures.

In all our simulations, we consistently employed the same collective variable, namely, the count of contacts between the DNA sequence and water molecules. The selection of a collective variable is a critical aspect of MetaD simulations, as it carries the risk of influencing the simulation if chosen incorrectly. By focusing on the number of contacts between the DNA sequence and water, we aimed to minimize such influences and ensure that the dynamics progressed naturally. Furthermore, our findings suggest that the CV we selected was not biased toward a specific unfolding mechanism: this is evident from the observed diversity in unfolding mechanisms for two out of the three G4 topologies investigated. The variability in unfolding mechanisms underscores the robustness and versatility of the chosen CV in capturing the diverse unfolding pathways for different G4 conformations.

Indeed, we plan to analyze and explore different collective variables in the near future, aiming to gain a more comprehensive understanding of the unfolding mechanisms.

For example, the radius of gyration (R_g) is a commonly utilized metric in G4 simulations and sometimes even in experimental studies. It is often presumed that high R_g corresponds to unfolded states, while the lowest R_g corresponds to the compact, fully-folded G4 structure. However, recently folding simulations revealed a distribution of R_g where the

coiled ensemble exhibited a lower R_g compared to the native G4. These results are in-line with studies on other complex conformation transitions, such as the Coil-Globule transition. Quoika and coworkers⁶⁰ found that R_g may potentially not capture the transition barriers not well and is therefore not a good collective variable, eventually. This challenges the notion that R_g alone is a sufficiently discriminative metric for distinguishing between folded and unfolded states.

Therefore, taking into account the solvation of molecules might be an alternative metric to this end, as it is unbiased with respect to certain intramolecular arrangements. By doing so, we aimed to anticipate the potential structures that could arise during the unfolding process. In this context, the MetaD method emerged as highly effective, consistently facilitating the unfolding of the G4 structure across all simulation replicas. It is evident, however, that achieving a comprehensive understanding of the entire unfolding of G4 through unbiased MD simulations remains currently beyond the computational capabilities of contemporary computers and force fields.

An important insight from our simulations is the presence of numerous intermediate states for each topology examined, highlighting the presence of a multitude of competing states and the absence of a single stable structure for these sequences, which agrees with the kinetic partition mechanism proposed.

The comprehension of the unfolding mechanism, akin to the intricacies of the folding process, is far from straightforward, and experimental data supporting any proposed mechanism remain elusive. Experimental investigation of this pathway is challenging due to its inherent complexity, characterized by the existence of numerous mutually exclusive intermediate states. Consequently, it is apparent that the formation mechanism involves more than a simple transition from a folded to an unfolded state. As a result, the reverse journey is equally complex, contrasting significantly with the relative simplicity of G4-folded β -helix transitions.

Notably, our simulations revealed distinct unfolding mechanisms across the various G4 topologies. Let's delve into each topology individually.

Parallel Topology: A Multifaceted Unfolding Mechanism

The parallel topology of G4s is undeniably fascinating due to its experimental stability, making it a crucial focus for understanding G4 unfolding mechanisms and advancing therapeutic strategies, particularly in the context of RNA-G4s, because it is the only topology found in RNA sequences.

Our unfolding simulations of this structure revealed a hierarchical mechanism, characterized by various potential paths. However, amidst this diversity, commonalities emerged across all simulations. A pivotal event in all simulations was the loss of a guanine from one of the terminal tetrads, leading to the formation of a highly stable intermediate state.

This state featured two tetrads, one with three guanines, a state consistently observed in every replication. The significance of this intermediate state lies in its ability to either reform the G4 structure or lead to the loss of an additional guanine from that plane. This intermediate state has also been simulated with MD classical simulations where it showed to be perfectly stable.

Remarkably, this "tilted" G state aligns precisely with the state predicted for interaction with G-clamps, a class of molecules known to destabilize G4 structures. The displacement of a guanine from the tetrad initiates the formation of four hydrogen bonds between the molecule and the guanine, exemplified by the name G-clamp. Recognizing this, we are actively working on developing computational methods to study the unfolding mechanism with these G-clamp molecules.

Another shared intermediate state observed in the simulations is the loss of a second guanine from the same tetrad after the first guanine is lost. This structure, featuring two tetrads and only two guanines, was demonstrated to be unstable in classical MD simulations, ultimately leading to the complete breakage of the G4. The rupture, however, can follow diverse mechanisms, forming a multitude of distinct intermediate states.

A recurring observation across nearly all replications was the rupture of the central tetrad after the two terminal tetrads were disrupted. This suggests a sequential pathway for the G4s to fold into this parallel topology, beginning with the formation of one tetrad, followed by the formation of another, and culminating in the formation of the last tetrad. These final states represent merely the end points, and the pathways to reach them are incredibly varied, encompassing an extensive sampling of markedly distinct intermediate states.

Additionally, we noted that during the simulation of the unfolding of this topology, G-triplexes, which were hypothesized as potential intermediate structures, were not observed. This observation aligns with a recent study by Sponer and colleagues, where the folding of parallel structures similarly did not yield the isolation of G-triplex intermediates.

Implications for Unfolding Pathways: DNA vs RNA G4s

The intricate unfolding mechanisms unveiled in our simulations for parallel-stranded DNA G4s hold potential implications for the unfolding processes of their RNA counterparts. The hierarchical mechanism observed, accompanied by stable intermediate states marked by guanine loss from terminal tetrads, could likely represent pivotal stages in the unfolding pathway shared by both DNA and RNA G4s.

The striking similarity in the unfolding pathways between DNA and RNA G4s suggests a fundamental commonality in their structural dynamics. The loss of guanines from the terminal tetrads, observed consistently in our simulations, might be a characteristic feature of G4 unfolding irrespective of the nucleic acid type.

Understanding these shared unfolding processes is key for gaining comprehensive insights into G4s dynamics, with implications extending to diverse biological contexts. This understanding may enable the development of targeted strategies for disrupting G4 structures, a promising avenue for potential therapeutic interventions in various diseases. Further research exploring the parallelism in folding and unfolding mechanisms between DNA and RNA G4s could shed additional light on their structural and functional roles in biological systems.

Hybrid and Antiparallel Topologies: Unfolding Mechanisms

The unfolding mechanism for the hybrid topology closely mirrors that of the parallel topology. In both cases, the process unfolds consecutively, usually initiated by the loss of a guanine from one of the tetrads. An intriguing observation is the stabilization of intermediate states, often involving interactions with adenines in the loop. In some simulations, a unique intermediate structure known as the G-triplex forms, resulting from a strand slippage-like event where a column of guanines is pulled away.

In the case of the antiparallel topology, the mechanism diverges significantly from the other topologies. The unfolding process involves the formation of distinct intermediate structures characterized by stacked columns of bases. These stacked structures exhibit a notably lower coordination number than the initial G4 structure. Subsequently, the G4 unravels through a straightforward opening of the sequence, often leaving behind hairpin-like structures in certain replicas, a consequence of the G4's inherent structural features with consecutively bound guanines.

Understanding these distinct unfolding pathways sheds light on the dynamic behavior of hybrid and antiparallel G4s. The identification of these intermediate states and their stability

provides valuable insights into the complex and varied nature of G4 dynamics. Further exploration of these pathways and their implications can deepen our knowing of the biological roles and potential therapeutic targets associated with G4 structures.

Exploring Unfolding Events and FES Complexity

Efficiently exploring the unfolding events and the free energy surface (FES) for such intricate systems remains a challenge. In this study, the MetaD approach, guides simulations towards the unfolding of these structures.

However, the complexity of the FES in the studied systems presents hurdles in achieving fully converged conformational ensembles. Additionally, interpreting these free energies proves challenging for understanding the unfolding mechanism comprehensively. The choice of CVs, dictating the biasing direction for conformational sampling, profoundly influences sampled states and unfolding pathways. Despite robustly capturing the formation of highly unstable species, calculating the ΔG of the unfolding of these structures and discriminating between different states or identifying commonalities across various topologies remain elusive at this stage. Exploration of the conformational space of the unfolded DNA is generally very challenging. Independent of the chosen CV, the population of the unfolded state may only be approximated.

Acknowledging the intricate G4 folding landscapes, this study underscores that a single simple CV could be useful to encapsulating these complex processes. However, the choice of the CV offers valuable insights into the potential existence of alternative pathways. Future endeavors will focus on optimizing folding CVs to enrich our understanding of these intricate pathways.

Challenges in Force Field Representation of G4 Unfolding Landscape

While the current simulations offer valuable insights into the G4 unfolding landscape, they are inherently constrained by limitations in force-field accuracy, chosen CVs, and achievable sampling.

In the context of DNA simulations, we used a newly developed force field. Given that this force field has been newly parametrized from scratch, it could not have been fully established, yet. However, the first validations look very promising. Notably, an acknowledged issue with force fields is the tendency to overestimate ion-ion repulsion within the G4 channel due to the absence of polarization/charge transfer effects with fixed point charges force fields. This overestimation might be addressed in the future through the adoption of polarizable force fields.

Moreover, the force field might potentially inadequately stabilize propeller loops, a concern that has been tentatively raised in the past. Previous studies have indicated notably fast unfolding for structures with propeller loops, and longer lifetimes for structures with specific loop configurations.

Previous unbiased and temperature-accelerated MD studies reported notably fast unfolding (short lifetimes) for all-anti hairpins, triplexes, and G4s (the latter simulated in the absence of cations) with propeller loops. Longer lifetimes were observed for structures with propeller loops with at least one guanosine in syn, and structures with lateral and diagonal loops. However, these studies used a different force field.

The precise aspects of the force field contributing to this imbalanced description of propeller loops remain unclear.

However, unraveling the exact components of the force field responsible for imbalanced descriptions, particularly concerning propeller loops, remains a challenge. Efforts to improve these representations, perhaps through dihedral parametrizations or advancements in force-field designs, will be crucial for a more accurate portrayal of G4 folding dynamics.

1.17.6 Conclusion

We have demonstrated that using the MetaD simulations along with the coordination number as a collective variable in our method holds significant promise for characterizing the unfolding pathways of G4s. This approach not only provides insights into the unfolding pathways but also enables the identification of transient ensembles that link major free energy states. This effort is pivotal in advancing our understanding of G4 properties and their functional roles.

Experimental approaches often face challenges in capturing short-lived intermediates and transitory ensembles in the dynamics of G4s. Consequently, complementary modeling and simulation studies become essential tools to unravel these intricate dynamics, offering a more comprehensive view of G4 behavior and function. Integrating computational methodologies with experimental data empowers us to explore and interpret the dynamic behavior of G4 structures, providing valuable insights that would be challenging to obtain through experimental techniques alone.

In this study, we presented a comprehensive set of all-atom enhanced-sampling unfolding simulations targeting various topologies of both DNA and RNA G4s. Our simulations shed light on the unfolding mechanism of parallel G4s, revealing a multi-pathway process

involving the tetrads. These finding challenges conventional literature models, which often depict simple intermediates like the Hoogsteen triplex in the folding process.

Moreover, our investigations underscore the considerable differences in unfolding mechanisms across different G4 topologies. This insight into topology-specific unfolding mechanisms holds immense promise for the design of molecules that can selectively target distinct G4 topologies. By leveraging these novel insights, we can pave the way for the development of specific molecules tailored to modulate G4 structures, offering exciting prospects for therapeutic interventions and advancing our fundamental understanding of G4 biology.

1.17.7 Materials and Methods

Starting structures

For the DNA we used three different topologies:

- Parallel: 1KF1
- Antiparallel: 143D
- Hybrid: 2GKU

For the RNA we used the parallel TERRA (PDB 6HHJ) and the structure find in SARS-CoV-2 genome (RG-1).

| PDB ID | Number of replica | Time (ns) |
|--------|-------------------|-----------|
| 1KF1 | 5 | 200 |
| 143D | 5 | 200 |
| 2GKU | 5 | 200 |
| 6HHJ | 5 | 200 |
| RG-1 | 5 | 200 |

Metadynamics protocol

The employed CVs for the MetaD biasing were based on the coordination number which count the number of contacts between two groups.

This keyword can be used to calculate the number of contacts between two groups of atoms and is defined as:

$$C = \sum_{i \in A} \sum_{j \in B} s_{ij}$$

where s_{ij} is 1 if the contact between atoms i and j is formed, zero otherwise. In actuality, s_{ij} is replaced with a (sigmoidal) switching function to ensure that the calculated CV has continuous derivatives. The default switching function is:

$$s_{ij} = \frac{1 - \left(\frac{r_{ij} - d_0}{r_0}\right)^n}{1 - \left(\frac{r_{ij} - d_0}{r_0}\right)^m}$$

Gromacs package (versions 2022.2) patched with PLUMED (versions 2.8.2) was used to run the MetaD simulations. The parameter, coordinate, and simulation settings files, reference structures, and PLUMED input files are available in the Supplementary data.

The entire protocol to perform the MetaD simulation is the following:

1. Initial minimization of the system solvated and with ions (50000 steps);
2. 100-ps of nvt equilibration;
3. 100-pns of npt equilibration;
4. Production MetaD using PLUMED with these parameters:

dna: GROUP NDX_FILE=G4.ndx NDX_GROUP=DNA

sol: GROUP NDX_FILE=G4.ndx NDX_GROUP=OW

conum: COORDINATION GROUPA=dna GROUPB=sol R_0=0.3 NLIST
NL_CUTOFF=0.5 NL_STRIDE=100

boost: METAD ARG=conum SIGMA=20 HEIGHT=.25 PACE=1000

PRINT ARG=conum,boost.* STRIDE=1000 FILE=colvar

RMSD calculation

The Root Mean Square Deviation (RMSD) measures how much a certain molecular structure deviates from a reference geometry. For a general molecular structure with N atoms, we can compute the RMSD between a certain conformation (r) and a reference structure (r_{ref}) via the following equation:

$$RMSD = \sqrt{\frac{1}{N} \sum_{i=1}^N (r_i - r_i^{ref})^2}$$

where r_i^{ref} represents the set of coordinates of the i -th atom in the reference structure, while r_i is the set of coordinates of the same atom but belonging to the structure that is going to

be compared to the reference. In summary, the RMSD is the square root of the average of the squared distances between atoms.

To compute this analysis, we used the `gmx rms` command which compares two structures by computing the RMSD.

Cluster analysis

To For clustering, to find the most representative structures for every replica, we used the Gromos Clustering Algorithm (Algorithm as described in Daura et al. (Angew. Chem. Int. Ed. 1999, 38, pp 236-240)) and it works in brief:

- Count number of neighbors using cut-off based on RMSD
- Take structure with largest number of neighbors with all its neighbors as cluster and eliminate it from the pool of clusters
- Repeat the same steps for the remaining structures.

We have done the clustering in several ways but the one that seems to best describe and separate the intermediate states is to do the analysis with respect to the orientation of the guanines that constitute the G4 during the simulation.

1.17.8 References

- (1) Haniff, H. S.; Tong, Y.; Liu, X.; Chen, J. L.; Suresh, B. M.; Andrews, R. J.; Peterson, J. M.; O'Leary, C. A.; Benhamou, R. I.; Moss, W. N.; et al. Targeting the SARS-CoV-2 RNA Genome with Small Molecule Binders and Ribonuclease Targeting Chimera (RIBOTAC) Degraders. *ACS Cent Sci* **2020**, *6* (10), 1713-1721. DOI: 10.1021/acscentsci.0c00984.
- (2) Garber, K. Drugging RNA. *Nat Biotechnol* **2023**, *41* (6), 745-749. DOI: 10.1038/s41587-023-01790-z.
- (3) Di Giorgio, A.; Duca, M. New Chemical Modalities Enabling Specific RNA Targeting and Degradation: Application to SARS-CoV-2 RNA. *ACS Cent Sci* **2020**, *6* (10), 1647-1650. DOI: 10.1021/acscentsci.0c01187.
- (4) Lavezzo, E.; Berselli, M.; Frasson, I.; Perrone, R.; Palù, G.; Brazzale, A. R.; Richter, S. N.; Toppo, S. G-quadruplex forming sequences in the genome of all known human viruses: A comprehensive guide. *PLoS Comput Biol* **2018**, *14* (12), e1006675. DOI: 10.1371/journal.pcbi.1006675.
- (5) Ruggiero, E.; Richter, S. N. Viral G-quadruplexes: New frontiers in virus pathogenesis and antiviral therapy. *Annu Rep Med Chem* **2020**, *54*, 101-131. DOI: 10.1016/bs.armc.2020.04.001.
- (6) Abiri, A.; Lavigne, M.; Rezaei, M.; Nikzad, S.; Zare, P.; Mergny, J. L.; Rahimi, H. R. Unlocking G-Quadruplexes as Antiviral Targets. *Pharmacol Rev* **2021**, *73* (3), 897-923. DOI: 10.1124/pharmrev.120.000230.
- (7) Zhang, R.; Xiao, K.; Gu, Y.; Liu, H.; Sun, X. Whole Genome Identification of Potential G-Quadruplexes and Analysis of the G-Quadruplex Binding Domain for SARS-CoV-2. *Front Genet* **2020**, *11*, 587829. DOI: 10.3389/fgene.2020.587829.
- (8) Panera, N.; Tozzi, A. E.; Alisi, A. The G-Quadruplex/Helicase World as a Potential Antiviral Approach Against COVID-19. *Drugs* **2020**, *80* (10), 941-946. DOI: 10.1007/s40265-020-01321-z.
- (9) Shao, X.; Zhang, W.; Umar, M. I.; Wong, H. Y.; Seng, Z.; Xie, Y.; Zhang, Y.; Yang, L.; Kwok, C. K.; Deng, X. RNA G-Quadruplex Structures Mediate Gene Regulation in Bacteria. *mBio* **2020**, *11* (1). DOI: 10.1128/mBio.02926-19.
- (10) Zhao, C.; Qin, G.; Niu, J.; Wang, Z.; Wang, C.; Ren, J.; Qu, X. Targeting RNA G-Quadruplex in SARS-CoV-2: A Promising Therapeutic Target for COVID-19? *Angew Chem Int Ed Engl* **2021**, *60* (1), 432-438. DOI: 10.1002/anie.202011419.
- (11) Cui, H.; Zhang, L. G-Quadruplexes Are Present in Human Coronaviruses Including SARS-CoV-2. *Front Microbiol* **2020**, *11*, 567317. DOI: 10.3389/fmicb.2020.567317.
- (12) Ji, D.; Juhas, M.; Tsang, C. M.; Kwok, C. K.; Li, Y.; Zhang, Y. Discovery of G-quadruplex-forming sequences in SARS-CoV-2. *Brief Bioinform* **2021**, *22* (2), 1150-1160. DOI: 10.1093/bib/bbaa114.
- (13) Bartas, M.; Brázda, V.; Bohálová, N.; Cantara, A.; Volná, A.; Stachurová, T.; Malachová, K.; Jagelská, E. B.; Porubiaková, O.; Červeň, J.; et al. In-Depth Bioinformatic Analyses of. *Front Microbiol* **2020**, *11*, 1583. DOI: 10.3389/fmicb.2020.01583.
- (14) Belmonte-Reche, E.; Serrano-Chacón, I.; Gonzalez, C.; Gallo, J.; Bañobre-López, M. Potential G-quadruplexes and i-Motifs in the SARS-CoV-2. *PLoS One* **2021**, *16* (6), e0250654. DOI: 10.1371/journal.pone.0250654.
- (15) Carvalho, J.; Lopes-Nunes, J.; Figueiredo, J.; Santos, T.; Miranda, A.; Riscado, M.; Sousa, F.; Duarte, A. P.; Socorro, S.; Tomaz, C. T.; et al. Molecular Beacon Assay Development for Severe Acute Respiratory Syndrome Coronavirus 2 Detection. *Sensors (Basel)* **2021**, *21* (21). DOI: 10.3390/s21217015.
- (16) Miclot, T.; Hognon, C.; Bignon, E.; Terenzi, A.; Marazzi, M.; Barone, G.; Monari, A. Structure and Dynamics of RNA Guanine Quadruplexes in SARS-CoV-2 Genome. Original Strategies against Emerging Viruses. *J Phys Chem Lett* **2021**, *12* (42), 10277-10283. DOI: 10.1021/acs.jpcllett.1c03071.
- (17) Luisa D'Anna, T. M., Emmanuelle Bignon, Ugo Perricone, Giampaolo Barone, Antonio Monari and Alessio Terenzi. Resolving a guanine-quadruplex structure in the SARS-CoV-2 genome through circular dichroism and multiscale molecular modeling. *Chem. Sci.* , **2023**.
- (18) Oliva, R.; Mukherjee, S.; Manisegaran, M.; Campanile, M.; Del Vecchio, P.; Petraccone, L.; Winter, R. Binding Properties of RNA Quadruplex of SARS-CoV-2 to Berberine Compared to Telomeric DNA Quadruplex. *Int J Mol Sci* **2022**, *23* (10). DOI: 10.3390/ijms23105690.
- (19) Shen, L. W.; Qian, M. Q.; Yu, K.; Narva, S.; Yu, F.; Wu, Y. L.; Zhang, W. Inhibition of Influenza A virus propagation by benzoselenoxanthenes stabilizing TMPRSS2 Gene G-quadruplex and hence down-regulating TMPRSS2 expression. *Sci Rep* **2020**, *10* (1), 7635. DOI: 10.1038/s41598-020-64368-8.
- (20) Zhai, L. Y.; Su, A. M.; Liu, J. F.; Zhao, J. J.; Xi, X. G.; Hou, X. M. Recent advances in applying G-quadruplex for SARS-CoV-2 targeting and diagnosis: A review. *Int J Biol Macromol* **2022**, *221*, 1476-1490. DOI: 10.1016/j.ijbiomac.2022.09.152.
- (21) Xi, H.; Juhas, M.; Zhang, Y. G-quadruplex based biosensor: A potential tool for SARS-CoV-2 detection. *Biosens Bioelectron* **2020**, *167*, 112494. DOI: 10.1016/j.bios.2020.112494.
- (22) Dai, J.; Carver, M.; Yang, D. Polymorphism of human telomeric quadruplex structures. *Biochimie* **2008**, *90* (8), 1172-1183. DOI: 10.1016/j.biochi.2008.02.026.

- (23) Wang, Y.; Patel, D. J. Solution structure of the human telomeric repeat d[AG3(T2AG3)3] G-tetraplex. *Structure* **1993**, *1* (4), 263-282. DOI: 10.1016/0969-2126(93)90015-9.
- (24) Lim, K. W.; Amrane, S.; Bouaziz, S.; Xu, W.; Mu, Y.; Patel, D. J.; Luu, K. N.; Phan, A. T. Structure of the human telomere in K⁺ solution: a stable basket-type G-quadruplex with only two G-tetrad layers. *J Am Chem Soc* **2009**, *131* (12), 4301-4309. DOI: 10.1021/ja807503g.
- (25) Phan, A. T.; Kuryavii, V.; Luu, K. N.; Patel, D. J. Structure of two intramolecular G-quadruplexes formed by natural human telomere sequences in K⁺ solution. *Nucleic Acids Res* **2007**, *35* (19), 6517-6525. DOI: 10.1093/nar/gkm706.
- (26) Parkinson, G. N.; Lee, M. P.; Neidle, S. Crystal structure of parallel quadruplexes from human telomeric DNA. *Nature* **2002**, *417* (6891), 876-880. DOI: 10.1038/nature755.
- (27) Lim, K. W.; Ng, V. C.; Martín-Pintado, N.; Heddi, B.; Phan, A. T. Structure of the human telomere in Na⁺ solution: an antiparallel (2+2) G-quadruplex scaffold reveals additional diversity. *Nucleic Acids Res* **2013**, *41* (22), 10556-10562. DOI: 10.1093/nar/gkt771.
- (28) Dai, J.; Punchihewa, C.; Ambrus, A.; Chen, D.; Jones, R. A.; Yang, D. Structure of the intramolecular human telomeric G-quadruplex in potassium solution: a novel adenine triple formation. *Nucleic Acids Res* **2007**, *35* (7), 2440-2450. DOI: 10.1093/nar/gkm009.
- (29) Palacký, J.; Vorlíčková, M.; Kejnovská, I.; Mojžeš, P. Polymorphism of human telomeric quadruplex structure controlled by DNA concentration: a Raman study. *Nucleic Acids Res* **2013**, *41* (2), 1005-1016. DOI: 10.1093/nar/gks1135.
- (30) Šponer, J.; Bussi, G.; Stadlbauer, P.; Kührová, P.; Banáš, P.; Islam, B.; Haider, S.; Neidle, S.; Otyepka, M. Folding of guanine quadruplex molecules-funnel-like mechanism or kinetic partitioning? An overview from MD simulation studies. *Biochim Biophys Acta Gen Subj* **2017**, *1861* (5 Pt B), 1246-1263. DOI: 10.1016/j.bbagen.2016.12.008.
- (31) Lindorff-Larsen, K.; Piana, S.; Dror, R. O.; Shaw, D. E. How fast-folding proteins fold. *Science* **2011**, *334* (6055), 517-520. DOI: 10.1126/science.1208351.
- (32) Stadlbauer, P.; Mlýnský, V.; Krepl, M.; Šponer, J. Complexity of Guanine Quadruplex Unfolding Pathways Revealed by Atomistic Pulling Simulations. *J Chem Inf Model* **2023**, *63* (15), 4716-4731. DOI: 10.1021/acs.jcim.3c00171.
- (33) Pavlína Pokorná, V. M., Giovanni Bussi, Jiří Šponer, Petr Stadlbauer. Parallel G-quadruplex folds via multiple paths involving G-tract stacking and structuring from coil ensemble. 2023.
- (34) Stefl, R.; Cheatham, T. E.; Spacková, N.; Fadrná, E.; Berger, I.; Koca, J.; Sponer, J. Formation pathways of a guanine-quadruplex DNA revealed by molecular dynamics and thermodynamic analysis of the substates. *Biophys J* **2003**, *85* (3), 1787-1804. DOI: 10.1016/S0006-3495(03)74608-6.
- (35) Limongelli, V.; De Tito, S.; Cerofolini, L.; Fragai, M.; Pagano, B.; Trotta, R.; Cosconati, S.; Marinelli, L.; Novellino, E.; Bertini, I.; et al. The G-triplex DNA. *Angew Chem Int Ed Engl* **2013**, *52* (8), 2269-2273. DOI: 10.1002/anie.201206522.
- (36) Bian, Y.; Tan, C.; Wang, J.; Sheng, Y.; Zhang, J.; Wang, W. Atomistic picture for the folding pathway of a hybrid-1 type human telomeric DNA G-quadruplex. *PLoS Comput Biol* **2014**, *10* (4), e1003562. DOI: 10.1371/journal.pcbi.1003562.
- (37) Yang, C.; Kulkarni, M.; Lim, M.; Pak, Y. Insilico direct folding of thrombin-binding aptamer G-quadruplex at all-atom level. *Nucleic Acids Res* **2017**, *45* (22), 12648-12656. DOI: 10.1093/nar/gkx1079.
- (38) Stadlbauer, P.; Kührová, P.; Vicherek, L.; Banáš, P.; Otyepka, M.; Trantírek, L.; Šponer, J. Parallel G-triplexes and G-hairpins as potential transitory ensembles in the folding of parallel-stranded DNA G-Quadruplexes. *Nucleic Acids Res* **2019**, *47* (14), 7276-7293. DOI: 10.1093/nar/gkz610.
- (39) Rocca, R.; Palazzesi, F.; Amato, J.; Costa, G.; Ortuso, F.; Pagano, B.; Randazzo, A.; Novellino, E.; Alcaro, S.; Moraca, F.; et al. Folding intermediate states of the parallel human telomeric G-quadruplex DNA explored using Well-Tempered Metadynamics. *Sci Rep* **2020**, *10* (1), 3176. DOI: 10.1038/s41598-020-59774-x.
- (40) Stadlbauer, P.; Kührová, P.; Banáš, P.; Koča, J.; Bussi, G.; Trantírek, L.; Otyepka, M.; Šponer, J. Hairpins participating in folding of human telomeric sequence quadruplexes studied by standard and T-REMD simulations. *Nucleic Acids Res* **2015**, *43* (20), 9626-9644. DOI: 10.1093/nar/gkv994.
- (41) Bergues-Pupo, A. E.; Arias-Gonzalez, J. R.; Morón, M. C.; Fiasconaro, A.; Falo, F. Role of the central cations in the mechanical unfolding of DNA and RNA G-quadruplexes. *Nucleic Acids Res* **2015**, *43* (15), 7638-7647. DOI: 10.1093/nar/gkv690.
- (42) Kogut, M.; Kleist, C.; Czub, J. Molecular dynamics simulations reveal the balance of forces governing the formation of a guanine tetrad-a common structural unit of G-quadruplex DNA. *Nucleic Acids Res* **2016**, *44* (7), 3020-3030. DOI: 10.1093/nar/gkw160.
- (43) Gajarský, M.; Živković, M. L.; Stadlbauer, P.; Pagano, B.; Fiala, R.; Amato, J.; Tomáška, L.; Šponer, J.; Plavec, J.; Trantírek, L. Structure of a Stable G-Hairpin. *J Am Chem Soc* **2017**, *139* (10), 3591-3594. DOI: 10.1021/jacs.6b10786.
- (44) Bian, Y.; Ren, W.; Song, F.; Yu, J.; Wang, J. Exploration of the folding dynamics of human telomeric G-quadruplex with a hybrid atomistic structure-based model. *J Chem Phys* **2018**, *148* (20), 204107. DOI: 10.1063/1.5028498.

- (45) Bian, Y.; Song, F.; Zhang, J.; Yu, J.; Wang, J.; Wang, W. Insights into the Kinetic Partitioning Folding Dynamics of the Human Telomeric G-Quadruplex from Molecular Simulations and Machine Learning. *J Chem Theory Comput* **2020**, *16* (9), 5936-5947. DOI: 10.1021/acs.jctc.0c00340.
- (46) Kejnovská, I.; Stadlbauer, P.; Trantírek, L.; Renčíuk, D.; Gajarský, M.; Krafčík, D.; Palacký, J.; Bednářová, K.; Šponer, J.; Mergny, J. L.; et al. G-Quadruplex Formation by DNA Sequences Deficient in Guanines: Two Tetrad Parallel Quadruplexes Do Not Fold Intramolecularly. *Chemistry* **2021**, *27* (47), 12115-12125. DOI: 10.1002/chem.202100895.
- (47) Stadlbauer, P.; Islam, B.; Otyepka, M.; Chen, J.; Monchaud, D.; Zhou, J.; Mergny, J. L.; Šponer, J. Insights into G-Quadruplex-Hemin Dynamics Using Atomistic Simulations: Implications for Reactivity and Folding. *J Chem Theory Comput* **2021**, *17* (3), 1883-1899. DOI: 10.1021/acs.jctc.0c01176.
- (48) Bussi, G.; Laio, A. Using metadynamics to explore complex free-energy landscapes. *Nat Rev Phys* **2020**, *2*, 200-212 DOI: <https://doi.org/10.1038/s42254-020-0153-0>.
- (49) Čeru, S.; Šket, P.; Prislán, I.; Lah, J.; Plavec, J. A new pathway of DNA G-quadruplex formation. *Angew Chem Int Ed Engl* **2014**, *53* (19), 4881-4884. DOI: 10.1002/anie.201400531.
- (50) Mashimo, T.; Sugiyama, H. Folding pathways of human telomeric hybrid G-quadruplex structure. *Nucleic Acids Symp Ser (Oxf)* **2007**, (51), 239-240. DOI: 10.1093/nass/nrm120.
- (51) Mashimo, T.; Sannohe, Y.; Yagi, H.; Sugiyama, H. Folding pathways of hybrid-1 and hybrid-2 G-quadruplex structures. *Nucleic Acids Symp Ser (Oxf)* **2008**, (52), 409-410. DOI: 10.1093/nass/nrn208.
- (52) Rajendran, A.; Endo, M.; Hidaka, K.; Sugiyama, H. Direct and single-molecule visualization of the solution-state structures of G-hairpin and G-triplex intermediates. *Angew Chem Int Ed Engl* **2014**, *53* (16), 4107-4112. DOI: 10.1002/anie.201308903.
- (53) Gray, R. D.; Trent, J. O.; Chaires, J. B. Folding and unfolding pathways of the human telomeric G-quadruplex. *J Mol Biol* **2014**, *426* (8), 1629-1650. DOI: 10.1016/j.jmb.2014.01.009.
- (54) Su, D. G.; Fang, H.; Gross, M. L.; Taylor, J. S. Photocrosslinking of human telomeric G-quadruplex loops by anti cyclobutane thymine dimer formation. *Proc Natl Acad Sci U S A* **2009**, *106* (31), 12861-12866. DOI: 10.1073/pnas.0902386106.
- (55) Stadlbauer, P.; Trantírek, L.; Cheatham, T. E.; Koča, J.; Sponer, J. Triplex intermediates in folding of human telomeric quadruplexes probed by microsecond-scale molecular dynamics simulations. *Biochimie* **2014**, *105*, 22-35. DOI: 10.1016/j.biochi.2014.07.009.
- (56) Koirala, D.; Mashimo, T.; Sannohe, Y.; Yu, Z.; Mao, H.; Sugiyama, H. Intramolecular folding in three tandem guanine repeats of human telomeric DNA. *Chem Commun (Camb)* **2012**, *48* (14), 2006-2008. DOI: 10.1039/c2cc16752b.
- (57) Lee, J. Y.; Okumus, B.; Kim, D. S.; Ha, T. Extreme conformational diversity in human telomeric DNA. *Proc Natl Acad Sci U S A* **2005**, *102* (52), 18938-18943. DOI: 10.1073/pnas.0506144102.
- (58) Bončina, M.; Lah, J.; Prislán, I.; Vesnaver, G. Energetic basis of human telomeric DNA folding into G-quadruplex structures. *J Am Chem Soc* **2012**, *134* (23), 9657-9663. DOI: 10.1021/ja300605n.
- (59) Marchand, A.; Gabelica, V. Folding and misfolding pathways of G-quadruplex DNA. *Nucleic Acids Res* **2016**, *44* (22), 10999-11012. DOI: 10.1093/nar/gkw970.
- (60) Quoika, P. K.; Fernández-Quintero, M. L.; Podewitz, M.; Hofer, F.; Liedl, K. R. Implementation of the Freely Jointed Chain Model to Assess Kinetics and Thermodynamics of Thermosensitive Coil-Globule Transition by Markov States. *J Phys Chem B* **2021**, *125* (18), 4898-4909. DOI: 10.1021/acs.jpcc.1c01946.
- (61) Zhang A. Y. Q., Balasubramanian S. The Kinetics and Folding Pathways of Intramolecular G-Quadruplex Nucleic Acids. *JACS* **2012**, *134*, 19297-19308. DOI: 10.1021/ja309851t.

Conclusions and perspectives

The COVID-19 pandemic, caused by the novel coronavirus SARS-CoV-2, has presented an unprecedented global health crisis, demanding urgent research and innovative strategies to combat the virus and its emerging variants. This thesis embraced a comprehensive approach, integrating both protein-based and genetic methodologies to unravel the complexities of SARS-CoV-2 and design innovative therapeutic strategies.

Protein Approach.

Understanding the SARS-CoV-2 Spike protein and its mutations is pivotal, given its critical role in viral entry and immune recognition. This research successfully predicted immune recognition regions within the Spike protein, guiding the development of potential vaccine candidates and diagnostic targets. Additionally, it investigated immune response variability to Spike protein mutations, providing insights into the efficacy of monoclonal antibodies against different variants. The study also delved into the stability of viral variants with Spike protein mutations, offering critical knowledge for effective public health strategies. The exploration of the fatty acid binding pocket within the Spike protein across variants identified potential druggable targets, furthering the potential for therapeutic interventions.

Genetic Approach

The genetic approach focused on characterizing the G-quadruplex (G4) folding landscape, a crucial secondary structure present in both human and viral genomes. By developing a computational tool to navigate the complex G4 folding mechanisms, this research unveiled the dynamic G4 folding landscape. Understanding the biological roles of G4 structures, particularly within viral genomes, is a significant step toward potential therapeutic interventions and drug discovery strategies.

The dual approach taken in this thesis, encompassing both protein and genetic perspectives, contributes substantially to our understanding of SARS-CoV-2 biology and its interactions with the host immune system. By addressing critical research objectives, this work paves the way for the development of effective strategies to combat COVID-19 and potentially other emerging viral threats. Ultimately, these findings contribute to the ongoing global efforts to mitigate the impact of the pandemic and enhance preparedness for future viral outbreaks.