In silico analysis of pathogen-host interactions at molecular level

by

Klevia Dishnica

A thesis submitted for the degree of

Doctor of Philosophy

University of Verona

2024

"In science, knowledge and understanding no longer appear quickly.

Time, patience, trial and error are all essential ingredients in any screening process."

Nobel Laureate: Satoshi Ōmura | Japan, Medicine 2015

UNIVERSITY OF VERONA

DEPARTMENT OF

BIOTECHNOLOGIES

PHD SCHOOL

Natural and Engineering Sciences

With funding by

University of Verona

CYCLE / YEAR of initial enrolment XXXVI**/2020**

Coordinator: **Prof. Dr. Flavia Guzzo**
Advisor: **Prof. Dr. Alejandro Giorgetti**

Ph.D. Candidate**: Klevia Dishnica**

A mio padre, mia mamma e mia sorella.

A mio marito Florenc.

A mio figlio Boiken, anima mia.

Acknowledgement

This thesis is a testament to the collective effort, guidance, and support of many individuals who have contributed to this path, making it not only possible but deeply enriching.

First and foremost, I would like to express my deepest gratitude to my advisor Prof. Alejandro Giorgetti for his guidance and support through the course of my PhD journey. Your mentorship has been motivating and has deeply influenced both my professional and personal growth. I sincerely appreciate your belief in me and the important role you have played in this transformative experience. Thank you for being not just a mentor; you are an example of how to combine intellectual rigor with humanity. I will carry these lessons with me throughout my life and career.

To my parents and sister, whose unconditional love and support have been the basis of my journey. Your sacrifices and faith in me have been my greatest source of strength and motivation. I appreciate your support during difficult times and your joy in celebrating my accomplishments. This achievement would not have been possible without your patience, wisdom, and belief in my dreams.

To my friends, you have made this experience unforgettable. Thank you for filling my life with joy and positivity. Particular thanks to Rui and Anisa for their professional and emotional support.

To my husband, and my in-laws, I am deeply thankful for your love and encouragement. To my husband, your patience and understanding have given me strength and comfort me through every hard time. Thank you for believing in me, even when I doubted myself. I am extremely thankful for your constant presence by my side. To my in-laws, your kindness and support have made this journey smoother and more fulfilling.

To my son, Boiken, who came into this world during one of the challenging yet rewarding times of my life. Thank you for being my little ray of sunshine and for inspiring me to push forward every single day. This accomplishment is as much yours as it is mine.

Finally, to Italy, for the most beautiful chapter of my life. The past 10 years were like puzzle challenging but ultimately rewarding. Thank you for all the opportunities that helped me grow and fulfill my dreams.

Abstract (Italian)

Questo studio esplora aspetti cruciali di SARS-CoV-2 e Strongyloides stercoralis, concentrandosi su variazioni strutturali, scoperta di farmaci, sfide diagnostiche e strategie terapeutiche. Inizialmente, esamina l'emergere di un'inserzione ricorrente nel dominio N-terminale della glicoproteina spike di SARS-CoV-2, evidenziando le sue implicazioni per l'evoluzione virale e la progettazione dei vaccini. Un esame basato sull'interazione per la scoperta di farmaci chiarisce i meccanismi degli inibitori noti di SARS-CoV-2 e identifica nuovi potenziali scaffolds di composti, fornendo una base per futuri sviluppi terapeutici. Inoltre, lo studio affronta la ridotta sensibilità dei test antigenici per il rilevamento delle infezioni da variante Omicron di SARS-CoV-2 attraverso un'analisi approfondita dei dati reali, sottolineando la necessità di migliorare i metodi diagnostici. Basandosi su questi risultati, la ricerca integra approcci multidisciplinari per affrontare le sfide in corso nella gestione delle malattie infettive. L'inserzione ricorrente nel dominio N-terminale della proteina spike di SARS-CoV-2 è analizzata per il suo ruolo nell'elusione del sistema immunitario e il suo potenziale come bersaglio per nuovi antivirali. Il metodo di screening basato sull'interazione colma le lacune tra gli inibitori noti di SARS-CoV-2 e i potenziali candidati farmacologici, dimostrando il potere dei modelli computazionali nell'accelerare la scoperta di farmaci. L'analisi della sensibilità dei test antigenici nel contesto delle infezioni da variante Omicron evidenzia le carenze critiche nelle tecnologie diagnostiche attuali e sottolinea la necessità di un continuo adattamento dei protocolli di test. Nel campo della parassitologia, il profilo proteomico dettagliato delle larve infettive di terzo stadio di Strongyloides stercoralis fornisce una preziosa risorsa per identificare nuovi bersagli molecolari e comprendere i meccanismi adattativi del parassita. L'esplorazione del targeting del recettore GluCl illustra il potenziale del riposizionamento dei farmaci, offrendo soluzioni pratiche per accelerare la disponibilità di trattamenti efficaci per l'infezione da Strongyloides stercoralis. Questo approccio integrato non solo avanza la conoscenza scientifica, ma propone anche strategie concrete per migliorare il controllo delle malattie e i risultati per i pazienti.

Abstract (English)

This study explores critical aspects of SARS-CoV-2 and Strongyloides stercoralis, focusing on structural variations, drug discovery, diagnostic challenges, and therapeutic strategies. Initially, it examines the emergence of a recurrent insertion in the N-terminal domain of the SARS-CoV-2 spike glycoprotein, highlighting its implications for viral evolution and vaccine design. An interaction-based drug discovery screen elucidates the mechanisms of known SARS-CoV-2 inhibitors and identifies potential novel compound scaffolds, providing a foundation for future therapeutic developments. Additionally, the study addresses the reduced sensitivity of antigen tests for detecting Omicron SARS-CoV-2 infections through extensive real-life data analysis, underscoring the need for improved diagnostic methods. Building on these findings, the research integrates multi-disciplinary approaches to tackle ongoing challenges in infectious disease management. The recurrent insertion in the SARS-CoV-2 spike protein N-terminal domain is analyzed for its role in immune evasion and its potential as a target for novel antivirals. The interaction-based screening method bridges gaps between known SARS-CoV-2 inhibitors and prospective drug candidates, demonstrating the power of computational models in expediting drug discovery. The analysis of antigen test sensitivity in the context of Omicron variant infections highlights critical shortcomings in current diagnostic technologies and emphasizes the necessity for continuous adaptation of testing protocols. In parasitology, the detailed proteomic profiling of Strongyloides stercoralis infective larvae provides a valuable resource for identifying new molecular targets and understanding the parasite's adaptive mechanisms. The exploration of GluCl receptor targeting illustrates the potential of drug repurposing pipeline, offering practical solutions to accelerate the availability of effective treatments for Strongyloides stercoralis infection. This integrated approach not only advances scientific knowledge but also proposes tangible strategies for enhancing disease control and improving patient outcomes.

Contents

General introduction

Emerging infectious diseases pose a significant and growing challenge to global public health [1]. These conditions arise from newly recognized pathogens in a population or involve familiar agents affecting new or larger populations or geographic regions. They are caused by various organisms, including bacteria, viruses, fungi, or parasites, and can spread rapidly and unpredictably [2]. Understanding transmission dynamics in the early stages of an outbreak is crucial for an effective public health response [3]. Estimating changes in transmission over time reveals the speed of spread and identifies infection hotspots, which is essential for evaluating control measures such as social distancing, lockdowns, and vaccination campaigns [4]. By analyzing how interventions influence transmission, alternative strategies can be designed to manage outbreaks more effectively. This comprehensive approach ensures that public health responses are data-driven, timely, and adaptable to evolving circumstances.

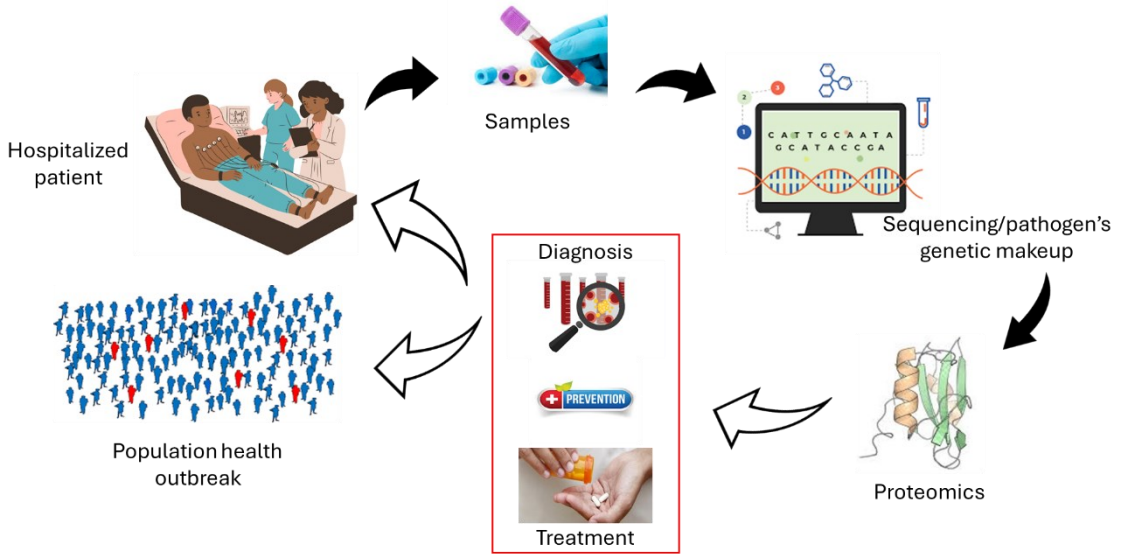


Figure 1.1: *Workflow demonstrating the use of patient samples for sequencing and proteomics analysis to inform diagnosis and treatment strategies. Insights from hospitalized patients and population outbreaks are integrated to guide interventions, creating a feedback loop for improving public and individual health outcomes.*

As the global impact of the COVID-19 pandemic, driven by SARS-CoV-2, has highlighted an urgent need for rapid advancements in viral diagnostics, therapeutic strategies, and a deeper understanding of viral evolution. This thesis contributes to this expanding field of research, exploring multiple dimensions of SARS-CoV-2 and related pathogens through computational, structural, and proteomic approaches.

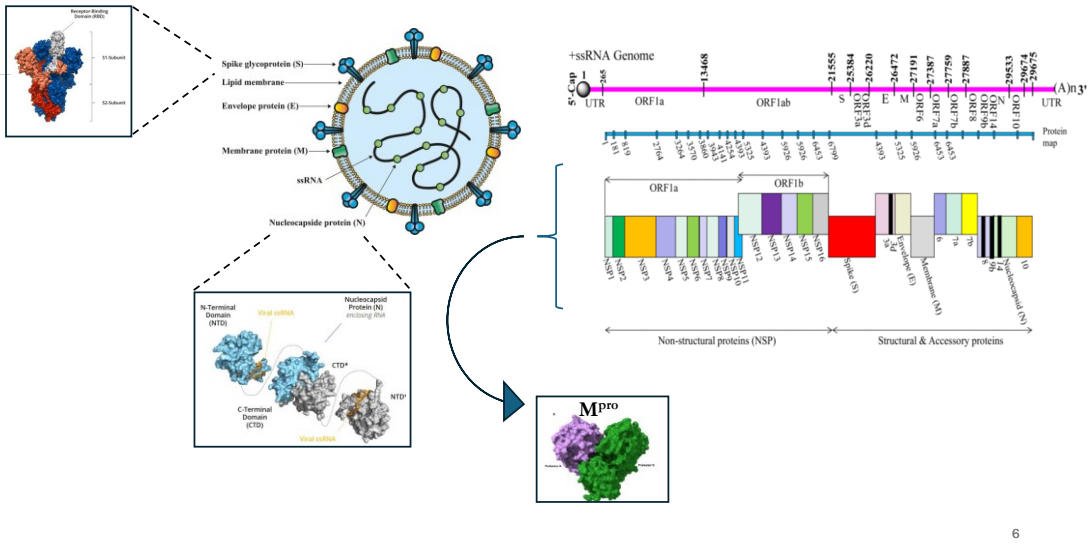
In late December 2019, an unexpected outbreak of respiratory illness emerged in Wuhan, Hubei Province, China. The symptoms, including fever, sore throat, and respiratory difficulties, initially

resembled viral pneumonia [5]. However, genomic analysis of respiratory specimens identified the causative agent as a novel coronavirus, initially named 2019-nCoV. Subsequently, the International Committee on Taxonomy of Viruses (ICTV) reclassified it as SARS-CoV-2 due to its close genetic relationship to the SARS-CoV virus [6]. The disease caused by this virus was officially named Coronavirus Disease 2019 (COVID-19). On March 11, 2020, the World Health Organization (WHO) declared COVID-19 a pandemic in light of its rapid global spread. By March 30, 2021, the virus had resulted in approximately 128 million confirmed cases and over 2.8 million deaths worldwide. Coronaviruses belong to a broad family potentially fatal diseases such as acute respiratory distress syndrome (ARDS) and organ failure [7].

SARS-CoV-2 is classified under the order Nidovirales, subfamily Orthocoronavirinae, and has four genera: Alphacoronavirus, Betacoronavirus, Gammacoronavirus, and Deltacoronavirus [8]. Viruses within the Nidovirales order share several structural features, including a conserved genomic organization, replicase gene located downstream of the 5'-UTR, and a unique mechanism of gene expression involving ribosomal frameshifting [8]. Seven coronaviruses are known to infect humans, including 229E, NL63, OC43, HKU1, MERS-CoV, SARS-CoV, and SARS-CoV-2. Of these, 229E, NL63, HKU1, and OC43 typically cause mild to moderate upper respiratory illnesses, while MERS-CoV, SARS-CoV, and SARS-CoV-2 are more severe, potentially causing fatal lower respiratory tract infections [9]. SARS-CoV-2 shares 88% of its genetic sequence with bat-SL-CoVZC45 and bat-SL-CoVZXC21, 79% with SARS-CoV, and around 50% with MERS-CoV, suggesting it likely originated in bats [6], [10]. Understanding these genetic similarities and differences can help develop effective treatment strategies for the pandemic. MERS-CoV was first identified in Jordan, Saudi Arabia, in 2012 and has since led to over 24,000 cases across 27 countries, predominantly in Saudi Arabia [11]. SARS-CoV was identified in China in 2003 and is also believed to have originated in bats, with an animal intermediary. The SARS outbreak was contained in mid-2003 through isolation and quarantine [12]. However, a few cases since then have occurred due to laboratory accidents and related exposures. SARS-CoV-2 is a novel strain that has not previously infected humans.

The SARS-CoV-2 genome is approximately 26-32 kb in length and includes various open reading frames (ORFs), sharing structural similarities with other human coronaviruses (HCoVs) [13]. These viruses are enveloped and display surface projections in the form of spike proteins. Their genome is an unsegmented, single-stranded positive-sense RNA with a 5' cap and a 3' poly(A) tail, which serves as functional mRNA for translation of the replicase polyproteins. Two-thirds of the genome, near the 5' end, encodes the replicase gene known as Open Reading Frame 1a and 1ab (ORF1ab), which produces nonstructural proteins (nsps) called pp1a and pp1ab polyproteins. The two polyproteins are cleaved by viral proteases, a papain-like protease (PLpro) and a 3C-like protease (3CLpro), to

generate 16 NSPs and to form the replication and transcription machinery. The pp1a non-structural protein corresponds to NSP1 to NSP11 and pp1ab non-structural protein comprises of NSP12 to NSP16. These nsps, such as the RNA-dependent RNA polymerase (nsp12) and helicase (nsp13), are essential components of the viral replication-transcription complex (RTC) responsible for synthesizing viral RNA and subgenomic RNAs required for structural protein production. The cleavage specificity of M^{pro} (Leu-Gln↓(Ser, Ala, Gly)) is unique to coronaviruses, ensuring minimal overlap with human proteases and making it a highly specific and conserved drug target. Without M^{pro}, the polyproteins remain unprocessed, preventing the formation of a functional RTC and halting viral replication. The pivotal role that M^{pro} plays in regulating viral replication and transcription makes it an attractive drug target, as its inhibition can effectively disrupt the SARS-CoV-2 life cycle and prevent the spread of infection. The remaining 10 kb, located near the 3' end, encodes structural proteins (S, E, M, and N) and nine accessory proteins encoded by ORF3a, ORF3d, ORF6, ORF7a, ORF7b, ORF8, ORF9b, ORF14, and ORF10 genes [14].



6

Figure 1.2: *SARS-CoV-2 structure and genomic organization. The illustration depicts the structural proteins (S, E, M, N) of SARS-CoV-2 surrounding its +ssRNA genome. The genome map outlines ORF1a/1b encoding non-structural proteins, along with structural and accessory proteins. Key features include the Spike protein's receptor-binding domain, Nucleocapsid domains (NTD, CTD), and the Main Protease (*M^{pro}*), a potential drug target.*

Given the critical role of M^{pro} in viral replication and transcription, advancing drug discovery for SARS-CoV-2, particularly targeting the main protease (M^{pro}), has emerged as a key focus.

In this context, my research, titled "An interaction-based drug discovery study to identify potential inhibitors of M^{pro}," contributes significantly to this field by employing an interaction-based drug discovery approach that identified 692 potential inhibitors targeting M^{pro}. By screening protein-ligand complexes from the Protein Data Bank (PDB), the investigation successfully predicted both known inhibitors, such as Dasatinib and Amodiaquine, and novel compound scaffolds, expanding the chemical diversity of potential therapeutic candidates. Notably, 17% of the top 100 predictions were validated using existing data, including four FDA-approved drugs, underscoring the method's utility in drug repurposing. Additionally, this work revealed a triplet hydrogen bond motif in the M^{pro} active site involving Gly143, Ser144, and Cys145, which is critical for ligand binding and protease inhibition. This discovery provides a deeper understanding of M^{pro}'s interaction dynamics and can guide the design of more effective inhibitors. While computational in nature, the findings highlight the need for rigorous experimental validation to confirm the efficacy and safety of these predictions, offering a robust framework for accelerating the development of targeted therapies against SARS-CoV-2.In addition to epidemiological data, structural and computational biology play an essential role in managing infectious diseases.

Mapping mutations onto 3D structures of proteins helps us understand the structure-function relationship in detail [5], which is critical for managing infectious diseases. At the atomic level, this adds information regarding protein stability and dynamics, offering better predictions about structural impacts. Recent progress in structural genomic consortiums has brought molecular docking and molecular dynamics (MD) simulation to the forefront, reducing reliance on labor-intensive experimental techniques. Computational MD analysis, using force fields for atoms in a macromolecule, helps us understand molecular motion and elucidates small differences caused by variations. Therefore, describing a molecule at the atomic level becomes an indispensable method, potentially bypassing experimental difficulties. MD can explain significant changes in the binding affinity of a macromolecule upon mutation and drug response. Integrating molecular insights with epidemiological data enhances the development of targeted treatments and intervention strategies, ultimately leading to more effective control and mitigation of infectious disease outbreaks.

In this perspective, the combined insights from my research "Emergence of a recurrent insertion in the N-terminal domain of the SARS-CoV-2 spike glycoprotein" and "Wide Real-Life Data Support Reduced Sensitivity of Antigen Tests for Omicron SARS-CoV-2 Infections" I performed on SARS-CoV-2 highlight the role of the SARS-CoV-2 N (nucleocapsid) protein and its significant impact on diagnostic efficacy and viral behavior.

The study "Emergence of a Recurrent Insertion in the N-terminal Domain of the SARS-CoV-2 Spike Glycoprotein" investigates a unique recurrent insertion region (RIR1) in the spike protein's N-terminal domain (NTD), located between codons 213–216, observed in at least 49 independent instances across multiple SARS-CoV-2 lineages. These insertions, reflecting convergent evolution, do not directly overlap with major antibody epitopes but may indirectly influence the structure of the spike protein and its interaction with the ACE2 receptor, impacting infectivity and immune escape. Found in Variants of Concern (VOCs) such as Alpha, Delta, and Omicron, RIR1 pairs with other mutations to enhance transmissibility and immune evasion, exemplified by Omicron's insertion (S:ins214EPE), which is accompanied by numerous spike mutations conferring significant immune escape and altered tropism. Molecular dynamics simulations played a critical role in assessing the structural and functional consequences of these insertions, using coarse-grained models to explore their impact on the spike protein's stability, flexibility, and interactions. The findings suggest that RIR1 insertions, likely arising from recombination events or replication errors, are maintained by structural constraints and evolutionary advantages. This work underscores the importance of genomic surveillance and structural analysis in understanding SARS-CoV-2 adaptation and informing vaccine and therapeutic strategies.

On the other hand, the work "Wide Real-Life Data Support Reduced Sensitivity of Antigen Tests for Omicron SARS-CoV-2 Infections" explores the diagnostic challenges posed by SARS-CoV-2 variants, focusing on antigen diagnostic tests (ADTs) during the transition from the Delta to the Omicron variant. The N protein, a core structural component of SARS-CoV-2, serves as a critical target for many ADTs. However, it is prone to mutations, particularly in variants like Delta and Omicron. These mutations may alter the protein's structure, which can influence the accuracy of rapid antigen tests and contribute to immune escape mechanisms. The study on the N protein specifically shows that Omicron has developed unique mutations in this protein, affecting ADT performance due to potential changes in antigen recognition. For instance, mutations like P80R and D343G in Omicron's N protein were identified as causing structural and dynamic shifts, potentially leading to a decrease in ADT sensitivity due to less efficient antibody binding. The decrease in ADT sensitivity from Delta to Omicron (63% to 33%) underscores the influence of these mutations on diagnostic accuracy, as newer variants present increased viral loads yet evade detection through altered N protein structures.

By integrating computational and structural biology methods, the research investigates how mutations in both the spike and N proteins can affect the virus's immune evasion and diagnostic detectability, with computational models elucidating the implications of each mutation. The presence of intrinsically disordered regions (IDRs) in the N protein further complicates this, as mutations in

these areas could impact the protein's stability, RNA binding affinity, and overall infectivity. This underscores the need for diagnostic tools that can account for these mutations to maintain accurate detection across emerging SARS-CoV-2 variants.

Beyond SARS-CoV-2, this thesis also expands to the neglected tropical disease strongyloidiasis, detailing novel proteomic insights into *Strongyloides stercoralis (S. stercoralis)* infective larvae.

There are about fifty species of parasitic nematodes in the genus Strongyloides, and they infect everything from humans to frogs. The majority of these species only have one or a small number of host species. Human infection by S. stercoralis, known as strongyloidiasis, is often referred to as a "disease of disadvantage" due to its prevalence in low-resource areas with inadequate sanitation [15]. Current estimates suggest that over 600 million people worldwide are infected, primarily in tropical and subtropical regions [16]. However, the actual prevalence is likely underestimated due to diagnostic difficulties, including low larval counts in stool samples and generally low parasite loads. Its life-cycle is complex, alternating between cycles of free-living and parasitic stages. Humans acquire the infection through the penetration of the intact skin by infective filariform larvae (iL3) present in contaminated soil which, once in the host, migrate through different organs. During migration, the larvae moult until they become adult worms, which ultimately settle in the small intestine. Once there, the parthenogenetic females deposit eggs that hatch in rhabditiform larvae (L1), which are then excreted in stools and initiate the free-living cycle. However, some L1 undergo an auto-infective cycle, i.e. mature into invasive filariform larvae, in the large intestine and penetrate the intestinal mucosa or the perianal skin to continue the parasitic life-cycle. This peculiar life-cycle allows S. stercoralis to perpetuate the infection, in the absence of treatment, potentially indefinitely [17].

Diagnosing and treating S. stercoralis infection remains challenging due to limitations in existing diagnostic methods and therapeutic strategies. Stool-based techniques, including microscopy and culture methods, lack sensitivity in chronic infections with low larval output, while serologic tests like ELISA and NIE-LIPS are more sensitive but prone to cross-reactivity [18]. Molecular diagnostics, such as PCR, offer improved detection but still face technical constraints. To enhance accuracy, a composite diagnostic approach combining fecal, serologic, and molecular methods is recommended. Ivermectin is the first-line treatment due to its high efficacy and tolerability, though repeated or alternative administration routes may be necessary for severe cases or immunocompromised patients [19]. Albendazole is a secondary option, and moxidectin shows potential as a new therapy with advantages like resistance mitigation and dose-independent efficacy [19]. Continued research such as advanced proteomic characterization to identify immunogenic proteins as potential diagnostic

markers.is critical to refine diagnostics, optimize treatments, and address the unique needs of vulnerable populations, ensuring effective management of this neglected disease.

The understanding of protein functions is fundamental in drug discovery, with computational tools and methodologies playing an increasingly crucial role in this field [20]. Proteins are essential components of biological systems, involved in virtually all cellular processes. They are the primary targets for most therapeutic drugs, underscoring the need to understand their activities and functions in drug discovery. Analytical methods that accurately determine protein activities and functions are, therefore, critical for developing new treatments. While mass spectrometry-based proteomics provides detailed data on protein dynamics and interactions, the integration of computational analysis is what truly amplifies the potential of this technology [20]. This synergy between proteomics and computational analysis is exemplified in the study of the "Somatic proteome of *S. stercoralis* infective third-stage larvae", which combines these methodologies to uncover crucial insights into protein functions, addressing challenges in diagnosis and treatment.

The study leverages mass spectrometry-based proteomics and computational methodologies to advance the understanding of *S. stercoralis* protein functions, crucial for addressing challenges in diagnosis and treatment.

The integration of molecular insights with computational approaches drives advancements in understanding *S. stercoralis* biology and facilitates the development of diagnostic and therapeutic strategies. Using high-throughput tandem mass spectrometry (LC-MS/MS), the study generated a comprehensive dataset of the proteome of *S. stercoralis* infective larvae (iL3), identifying 430 proteins, 187 of which were previously uncharacterized. To interpret this wealth of molecular data, computational tools were employed alongside manual annotation, which played a crucial role in refining and validating the automated analyses. Functional annotation through Gene Ontology (GO) and InterPro databases enabled the classification of proteins based on their molecular functions, biological processes, and cellular components, shedding light on the biological roles of key proteins involved in parasite survival, host interaction, and pathogenesis.

Immunoinformatics tools such as BepiPred-2.0 were used to predict linear B-cell epitopes, identifying immunogenic proteins capable of eliciting an immune response. Manual annotation was instrumental in verifying these predictions, ensuring the selection of relevant targets. Homology analysis further prioritized *S. stercoralis*-specific proteins by comparing them to human and pathogen proteomes, reducing the likelihood of cross-reactivity and enhancing the specificity of diagnostic markers. Additionally, by structural modeling was able to map the predicted epitopes onto protein structures,

validating their accessibility for antibody binding and confirming their potential as vaccine candidates or diagnostic markers.

By integrating experimental proteomics with computational methodologies and manual annotation, the study provided a dynamic understanding of the *S. stercoralis* proteome, emphasizing critical protein functions such as oxidoreductase and peptidase activities, which are essential for the parasite's survival and infectivity. This comprehensive approach not only identified novel diagnostic and therapeutic targets but also demonstrated the value of combining automated computational tools with manual annotation for accurate and meaningful data interpretation. It underscores the importance of such integrative methods in advancing parasitic disease research, paving the way for more precise diagnostics and effective treatments. This is particularly important in combating emerging infectious diseases and pandemics, where rapid and accurate diagnosis, along with effective treatment options, is essential. Additionally, computational tools help identify new proteins at different stages of infectious diseases, providing a dynamic understanding of how infections progress and how the body's response changes over time [21]. This insight is crucial for developing stage-specific diagnostics and treatments, ensuring timely and targeted interventions.

This need for rapid drug development was starkly highlighted during the COVID-19 pandemic. The emergence of COVID-19 in late 2019 and its rapid global spread highlighted the urgent need for effective treatments and vaccines [22]. The traditional drug discovery process, spanning several years and extensive laboratory testing, was insufficient for the immediate demand. This scenario underscored the critical role of virtual screening in accelerating drug discovery and repurposing efforts. Similarly, strongyloidiasis, caused by the parasitic worm *S. stercoralis*, presents a pressing need for new therapies due to the growing risk of resistance to the most used drugs, like ivermectin. As the primary treatment for strongyloidiasis, ivermectin's potential resistance necessitates the identification of alternative compounds.

The drug development process extensively uses virtual screening for scaffold hopping, lead optimization, and lead identification, offering a quick and low-cost alternative to high-throughput screening for novel pharmaceuticals.

Virtual screening approaches fall into two main categories: ligand-based (e.g., ligand similarity) and structure-based (e.g., ligand docking). Protein-ligand docking uses the three-dimensional structure of the target protein to predict binding modes and affinities of ligands, while ligand similarity methods capitalize on the likelihood that similar ligands will exhibit similar activity. The exponential growth in computational biology and experimental protein structure determination has significantly increased novel drug identification [21]. This process saves time and money by determining a drug's stability,

safety, and efficacy through computational study alongside experimental work. Consequently, the computational approach has become a key component for integrating and analyzing all available knowledge.

In my study "Targeting GluCl Receptor: Drug Repurposing Strategies for Strongyloides stercoralis Infection", virtual screening was employed to identify potential therapeutic compounds targeting the *S. stercoralis* GLUCL protein, a crucial receptor in the parasite's neurobiology. By leveraging structure-based approaches such as protein-ligand docking, we explored the three-dimensional structure of the GLUCL protein to predict binding modes and affinities of various ligands. This method allowed us to identify several promising compounds that demonstrated strong binding potential, highlighting their viability as candidates for alternative therapies to ivermectin. However, as with any virtual screening effort, these findings require rigorous experimental validation to confirm their efficacy, safety, and stability. This step is essential to ensure the reliability of these compounds as potential treatments for strongyloidiasis and to address the growing concern of drug resistance. The integration of computational methods and experimental follow-up provides a robust framework for advancing therapeutic solutions for this neglected tropical disease.

Together, these studies provide a comprehensive view of viral evolution, therapeutic discovery, and diagnostic precision, with applications spanning pandemic preparedness and tropical disease management. My thesis aims to bridge gaps in SARS-CoV-2 research and contribute to broader virology and bioinformatics fields, supporting future response strategies for infectious diseases

About this thesis

This thesis is divided into two main parts. In Part I - ***Comprehensive Insights into SARS-CoV-2 Mutations, Therapeutic Targets, and Diagnostic Challenges***, I present three distinct studies that together build a coherent picture of the ongoing challenges and advancements in the fight against COVID-19.

The ongoing evolution of SARS-CoV-2 poses significant challenges and offers critical opportunities for understanding and combating the virus. Recent research has delved into various facets of the virus's behavior, including genetic mutations, drug discovery, and diagnostic accuracy. These studies provide valuable insights into the emergence of specific spike glycoprotein mutations, innovative methods for identifying effective inhibitors, and the efficacy of diagnostic tests against new variants like Omicron. Collectively, they highlight the necessity of continuous research and adaptive strategies in the global effort to manage and mitigate the impact of COVID-19.

Firstly, in Chapter **'Emergence of a Recurrent Insertion in the N-terminal Domain of the SARS-CoV-2 Spike Glycoprotein'** the study examines the emergence of a recurrent insertion in the N-terminal domain (NTD) of the SARS-CoV-2 spike glycoprotein. The insertion, identified in various SARS-CoV-2 variants, suggests potential impacts on viral behavior, including changes in transmissibility, immune evasion, and vaccine effectiveness. By analyzing the genetic sequences of these variants, the study provides insights into how such mutations could influence the pandemic's trajectory and underscores the need for continuous genomic surveillance.

Then, in Chapter '**An Interaction-Based Drug Discovery Screen Explains Known SARS-CoV-2 Inhibitors and Predicts New Compound Scaffolds'** I present a building on the understanding of viral mutations, this research focuses on a drug discovery approach that utilizes interaction-based screening to identify compounds that inhibit the function of M^{pro} protein in the genome of SARS-CoV-2. The study validates known inhibitors and discovers new potential drug scaffolds by mapping interactions between viral proteins and small molecules. This method enhances the understanding of how these compounds interfere with the virus's replication process, offering a strategic framework for developing effective antiviral therapies against COVID-19.

Finally, in Chapter **'Wide Real-Life Data Support Reduced Sensitivity of Antigen Tests for Omicron SARS-CoV-2'** in light of the evolving virus and efforts to control its spread, this study evaluates the performance of antigen tests in detecting the Omicron variant of SARS-CoV-2 using extensive real-life data. The findings indicate a reduced sensitivity of these tests when identifying Omicron infections compared to previous variants. The study highlights the implications for public

health strategies, emphasizing the necessity for updated testing protocols and technologies to ensure accurate and timely diagnosis in the context of evolving viral mutations.

These interconnected studies underscore the dynamic nature of SARS-CoV-2 and the multifaceted approach required to address its challenges. From understanding mutations to discovering new therapeutic agents and refining diagnostic tools, this research collectively advances our capability to manage and mitigate the impact of COVID-19.

In Part II - ***Proteomic Analysis and Drug Repurposing Strategies for Strongyloides stercoralis: Novel Insights and Therapeutic Approaches***, I delve into the proteomic analysis of *Strongyloides stercoralis (S. stercoralis)* to uncover new information about the proteins expressed by infective third-stage larvae. This part of the thesis explores the somatic proteome of these parasitic nematodes, aiming to enhance our understanding of their biology, host interaction mechanisms, and potential vulnerabilities. Such insights are crucial for developing new strategies to combat strongyloidiasis, a parasitic disease with significant global health implications.

By characterizing the proteomic landscape of *S. stercoralis* infective third-stage larvae, this research identifies key proteins that may play essential roles in the parasite's survival, infectivity, and adaptation to host environments. This proteomic profiling not only broadens our knowledge of *S. stercoralis* biology but also opens up new avenues for targeted therapeutic interventions and diagnostic tool development to improve the management and treatment of strongyloidiasis.

The chapter **Targeting GluCl Receptor: Drug Repurposing Strategies for *Strongyloides stercoralis* Infection** focuses on identifying new therapeutic strategies by repurposing existing drugs that target the glutamate-gated chloride channel (GluCl) receptor, a critical component in the neurobiology of *S. stercoralis*. This research explores how known drugs can be repurposed to inhibit the GluCl receptor, offering a cost-effective and expedited pathway to new treatments. By leveraging existing pharmacological knowledge and drugs, this approach aims to accelerate the development of effective therapies for strongyloidiasis, thereby improving patient outcomes and addressing a significant public health concern.

Together, these studies in Part II provide novel insights and practical strategies for understanding and combating *S. stercoralis* infections, contributing to the broader effort to control parasitic diseases globally.

PART I

Comprehensive Insights into SARS-CoV-2 Mutations, Therapeutic Targets, and Diagnostic Challenges

Emergence of a recurrent insertion in the N-terminal domain of the SARS-CoV-2 spike glycoprotein

This chapter describes my contribution to: Gerdol M, ***Dishnica K****, Giorgetti A. Emergence of a recurrent insertion in the N-terminal domain of the SARS-CoV-2 spike glycoprotein. Virus Res. 2022 Mar;310:198674. doi: 10.1016/j.virusres.2022.198674. Epub 2022 Jan 10. PMID: 35021068; PMCID: PMC8743576.* [23]

Introduction

Coronaviruses generally accumulate mutations at a much lower rate than other RNA viruses, thanks to the efficient proofreading exonuclease activity exerted by nsp14, in complex the activator protein nsp10 [24] [25]. As a result, the rate of molecular evolution of SARS-CoV-2 is currently estimated (as of January 5th, 2022, based on GISAID data [26]), to be close to 25 substitutions/year per genome, i.e. 8.36×10^{-4} substitutions/site/year, which is slightly higher than previous estimates for human endemic coronaviruses [27]. Consistently with comparative genomics data obtained from other members of the *Sarbecovirus* subgenus, such mutations are not evenly distributed across the genome, but they are disproportionally located in the S gene, which encodes the spike glycoprotein. It is also worth noting that the S gene undergoes frequent recombination events, likely as a result of naturally occurring co-infections in the animal viral reservoirs [28], and that these events are also theoretically possible among different SARS-CoV-2 lineages [29]. The encoded transmembrane protein forms a homotrimer and plays a fundamental role in the interaction between the virus and host cells, promoting viral entry though the interaction with different membrane receptors [30]. In the case of

SARS-CoV-2 and of the closely related SARS-CoV responsible of the 2002–2004 outbreak, such receptor is represented by the angiotensin converting enzyme 2 (ACE2) [31], [32].
While most of these mutations have little or no phenotypic impact at all, some may significantly influence viral transmissibility and the ability of the virus to escape host immune response. The causes underpinning such phenotypic effects may either lie in an increased viral shedding, in the alteration of the binding affinity between the spike receptor binding domain (RBD) and the host ACE2 receptor, or in the modification of key antibody epitopes. The most striking example of a non-synonymous mutation which had a dramatic impact on the dynamics of the pandemics is most certainly represented by S:D614G. This mutation, which was not present in the ancestral lineage that caused the Wuhan outbreak, emerged in the very early phases of the pandemics, quickly becoming dominant worldwide [33], most likely due to an increased packing of functional spike protein into the virion [34].

Even though the mutation rate of the SARS-CoV-2 genome remained relatively stable throughout 2020, growing evidence soon started to point out the presence of shared mutations across multiple independent lineages, suggesting ongoing convergent evolution and possible signatures of host adaptation [35]. While early investigations failed to identify evidence of increased transmissibility associated with such recurrent mutations [36], the nearly contemporary independent emergence of three variants sharing the non-synonymous substitution N501Y in the spike protein started to raise serious concerns about the possible involvement of this mutation in increasing viral infectivity. While the functional role of N501Y still remains to be fully elucidated, structural modeling points towards a possible function in the stabilization of the spike protein in the open conformation, which may increase ACE2 binding, especially in combination with other mutations targeting the RBD [37], [38], [39].

B.1.1.7 (the alpha variant, according to WHO labeling), one of the emerging lineages carrying S:N501Y, spread in southeastern England in early 2020 and quickly became dominant in Europe. Despite being significantly more transmissible than wild-type genotypes [40], alpha was not associated with significant immune escape from the neutralizing activity of convalescent or vaccinated sera [41], [42], [43], [44]. On the other hand, some point mutations present in the spike NTD, i.e. the deletion of a codon in position 144, led to full escape from the activity of a few NTD-directed monoclonal antibodies [45].
Two other major lineages carrying N501Y, designated as variants of concerns (VOCs) in early 2021, i.e. B.1.351 (beta) and P.1 (gamma), were linked with major outbreaks in geographical regions with

very high estimated seroprevalence, i.e. in the Eastern Cape region (South Africa) [46] and in Manaus (Amazonas, Brazil) [47], respectively. Both variants were characterized by a constellation of non-synonymous mutations and accelerated rates of evolution, which suggested that their selection might have occurred in immunocompromised patients with persistent viral infection [48]. Among the many features shared by beta and gamma, the most remarkable one was the presence of two additional RBD mutations, i.e. E484K and K417N/K417T. The former one has been identified as a key player in antibody escape, due to its presence in a major epitope recognized by class II RBD-directed antibodies [49], [50], [51]. On the other hand, mutations of K417, located in an epitope recognized by class I antibodies, are thought to provide a minor contribution to polyclonal antibody response escape [49] and to possibly stabilize, together with E484K and N501Y, the interaction between the RBD and the ACE2 receptor [39]. Due to the possible negative impacts of these emerging variants on ongoing vaccination campaigns [52], [53], the focus placed on molecular surveillance significantly increased throughout 2021.

In the spring of 2021, the lineage B.1.617.2 (delta) was internationally recognized as the fourth VOC. Like the three previously mentioned variants, delta carried several non-synonymous mutations in the S gene, including L452R, which is located in a major class III RBD-directed antibody epitope [53] and allows to completely escape the neutralizing activity of several monoclonal antibodies (mAbs) [51]. Following its initial association with the surge of infections that occurred in India in early 2021 [54], this variant rapidly spread worldwide and replaced alpha, which strongly suggested a higher intrinsic transmissibility [55], possibly due to a more efficient cleavage site between the S1 and S2 subunits [56]. At the same time, delta was also found to be endowed with significant immune escape properties, which resulted in reduced sensitivity towards the sera of convalescent and vaccinated individuals [57] and in reduced vaccine effectiveness, in particular after the first dose [58]. Although delta became dominant worldwide in the second half of 2021, a novel variant, designed as B.1.1.529, started to quickly spread in the Gauteng province (South Africa) in November 2021, outcompeting delta. This fitness advantage has been tentatively linked with a substantial ability to evade immunity from previous infection [59], which might be consistent with the high number of non-synonymous mutations and indels observed in the S gene compared with the reference SARS-CoV-2 genome. These include a number of previously described RBD mutations associated with the aforementioned VOC, such as T478K and N501Y, plus E484A, which suggested significant immune evasion properties, later confirmed by a number of in vitro studies [60], [61], [62], [63]. Based on early epidemiological data and on the growing number of imported cases reported abroad, on December 1st, 2021 WHO included B.1.1.529 in the list of VOCs under the "omicron" designation.

Several VOCs and VOIs (including alpha, beta, delta and omicron) carry spike deletions in the NTD. Such deletions were previously shown to often occur in distinct NTD sites, named Recurrent Deletion Regions (RDR), arising in different geographical backgrounds, in independent viral lineages. Some RDR sites display a significant overlap with known immune epitopes, suggesting that they may drive antibody escape [64]. Comparatively, prior to the emergence of omicron, which carries a three amino acids-long insertion (S:ins214EPE) in the NTD of the spike protein, very little attention had been directed towards insertions. Nevertheless, such events are known to have played a fundamental role in the past evolution of SARS-CoV-2 spike protein by allowing, among the other things, the acquisition of a furin-like cleavage site, which is an uncommon feature in bat coronaviruses. This short motif, which is thought to be a key pathogenicity determinant [65], is indeed completely absent in the closely related *Sarbecovirus* RaTG13 [66] and only partly present in the recently described RmYN02 [67], [68] and RacCS203 [69].

The present work reports the independent occurrence of at least 49 distinct insertion events at the very same NTD site, located between Val213 and Leu216, which will be hereafter referred to as Recurrent Insertion Region 1 (RIR1). The transient international spread of the RIR1 insertion-carrying lineages A.2.5 and B.1.214.2, the presence of S:ins214EPE in omicron and the identification of several insertions at this site in the alpha and delta lineages point out that more attention should be put towards the functional characterization of these codon acquisitions in the near future.

Materials and methods

Sequence data analysis

The global frequency of insertion and deletion mutations mapped on the SARS-CoV-2 S gene was retrieved, based on GISAID data [26], from https://mendel.bii.a-star.edu.sg/ (last accessed on January 5th, 2022; credit to Raphael Tze Chuen Lee). Disruptive insertion and deletion mutations (i.e. those that interrupted the open reading frame of the S gene) and insertions carrying undetermined amino acids were discarded. Genomes carrying insertions at any position between codons 213 and 216 were grouped based on the inserted nucleotide sequence. Each group was assigned a code based on progressive Roman numerals, following their chronological order of identification; variants of the same insertion including SNPs, which were detected for insertion III, IV and XLI, were disregarded. The nucleotide sequences of representative entries for each of the identified insertions were aligned with the Wuhan-Hu-1 isolate SARS-CoV-2 reference sequence (GenBank ID: NC_045512.2) using MUSCLE [70] in the MEGA X environment [71], initially preserving codon boundaries. The multiple sequence alignment was then manually refined to reflect the most probable location of the insertion

within each codon. Each event was consequently classified as a phase 0, phase I or phase II insertion, annotating insertions with ambiguous placement.

All SARS-CoV-2 genome data used for phylogenetic inference in this study were retrieved from GISAID [26]. In detail, all available sequenced genomes belonging to the lineage A.2.5, to the related sublineages A.2.5.1, A.2.5.2 and A.2.5.3, and to the sister lineage A.2.4 were downloaded, along with associated metadata. While all available GISAID entries were considered for reporting observation frequencies, only high quality genomes (i.e. those listed as "complete" and "high coverage") associated with a sampling date were taken into account for further analysis. Genomes containing long stretches of Ns (i.e. comprising more than 25 consecutive undetermined nucleotides) were discarded. The reference isolate Wuhan-Hu-1 was also included for tree rooting purposes. Note that several genome sequences from Panama with sampling date anterior to November 2021 were disregarded due to the unreliability of associated metadata (i.e. the sampling dates appeared to be inconsistent with the very small genetic distances with recent isolates belonging to the same lineage). Overall, the A.2.5-focused datasets included 1283 sequences.

SARS-CoV-2 genomes were analyzed with the nextstrain *augur* pipeline (https://github.com/nextstrain/augur). Briefly, nucleotide sequences were aligned with MAFFT [72] and the resulting multiple sequence alignment was used as an input for a maximum likelihood phylogenetic inference analysis, carried out with FastTree [73] under a generalized time reversible model of molecular evolution. The resulting tree was further refined in the *augur* environment with treetime v.0.8.1 [74] using sampling date metadata, generating a time-calibrated tree. The phylogenetic tree was rooted based on the oldest available genotype, which in this case was Wuhan-Hu-1, and graphically rendered using FigTree v.1.1.4.

A root-to-tip genetic distance analysis was performed by plotting the sampling dates against the total number of nucleotide substitutions (excluding insertions and deletions) observed in genomes belonging to the A.2.5 lineage and related sublineages. These were calculated with MEGA X [71], compared with the reference genotype Wuhan-Hu-1. The global average genome-wide mutation rate of SARS-CoV-2, roughly equivalent to 25 substitutions per year, was retrieved from GISAID (as of January 5th, 2022).

System setup of coarse-grained models

The simulations on the wild-type spike protein were carried out considering the crystallographic structure deposited in PDB (accession ID: 6XR8) [75]. A few missing portions were modeled with Swiss Model [76] in order not to compromise the molecular dynamic properties of the protein. Homology modeling was performed to obtain the 3D structure of the spike protein of A.2.5 with

Swiss Model [76], using the EPI_ISL_1,502,836 GISAID entry as a reference. The protein structure was converted to a coarse-grained Martini representation using the martinize.py script [77]. The coarse-grained protein coordinates were then positioned in the center of a simulation box of size $23\times23\times23$ nm^3.

The Martini coarse-grained force field with an Elastic Network (CG-ElNeDyn) [77] was used for running the molecular dynamics simulations through the Gromacs 2019.3 package [78]. The analyses were run using isothermal-isobaric NPT ensemble equilibrium simulations. The temperature for each group (protein, water and ions) was kept constant at 315 K using V-rescale thermostat [79] with a coupling constant of 1.0 ps. The pressure was isotropically controlled by a Parrinello-Rahman barostat [80] at a reference of 1 bar with a coupling constant of 12.0 ps and compressibility of 3×10^{-4}. Non-bonded interactions were used in their shifted form with electrostatic interactions shifted to zero in the range of 0–1.1 nm. A time step of 20 fs was used with neighbor lists updated every 20 steps. Periodic boundary conditions were used in the x, y and z axes. ~4μs were collected for the simulations of the wild type and mutant (i.e. A.2.5) spike proteins, respectively. The root mean square deviation (RMSD) of backbone beads, the root mean square fluctuations (RMSF) and the radius of gyration (RGYR) were calculated using the gmx rms, rmsf and gyrate modules from the Gromacs package [78]. Principal component analysis (PCA), computed with MDAnalysis, was restricted to backbone beads, as it is less perturbed by statistical noise and provides significant characterization of the essential space motions [81].

To visualize the direction and extent of the principal motions of the simulated systems, a porcupine plot analysis was performed using the modevectors.py script in Pymol [82].

Results and discussion

Presence of a recurrent insertion region (RIR1) in the N-terminal domain of SARS-CoV-2 spike protein

The analysis of the genomic data deposited in GISAID revealed that, before the emergence of omicron in November 2021, S gene insertions (excluding those that disrupted the open reading frame) were present in just a minor fraction of all sequenced SARS-CoV-2 genomes, i.e. roughly 0.3% of the total. Overall, the frequency of observation of spike deletions was more than 500 folds higher than spike insertions, even though this ratio is now rapidly changing due to the spread of omicron. As previously reported by other authors, most deletions occur in specific sites of the N-terminal domain, including the four previously identified Recurrent Deletion Regions (RDR) 1, 2, 3 and 4, associated with several widespread VOCs and VOIs (**Figure 3.1**) [64], and the deletion which

characterizes the delta variant, occurring at positions 157/158. This is consistent with the higher rate of mutation observed for the S1 region (which includes the NTD and RBD) in human coronaviruses compared with the more slowly evolving S2 subunit.

Despite their lower frequency of observation, insertions do not occur randomly in the S gene. In fact, the overwhelming majority of the insertion mutations mapped so far in SARS-CoV-2 S gene target the NTD, being in most cases identified at a specific site, located between codons 213 and 216 (**Figure 3.1**). However, this figure might be an underestimate due to the frequent use of reference-based insertion-unaware algorithms for SARS-CoV-2 genome assembly, especially during the early phases

of the pandemics. Due to the convergent finding of such insertions in independent viral lineages (see below), this region will be hereafter named Recurrent Insertion Region 1 (RIR1).

Even though insertions were observed at several other spike sites, RIR1 was the only one where multiple insertions have independently occurred in different lineages. The only other spike insertions site with more than 1000 occurrences among the sequenced SARS-CoV-2 genomes (as of January 5th, 2022) is ins145T, found in the VOI mu [83] (**Figure 3.1**).

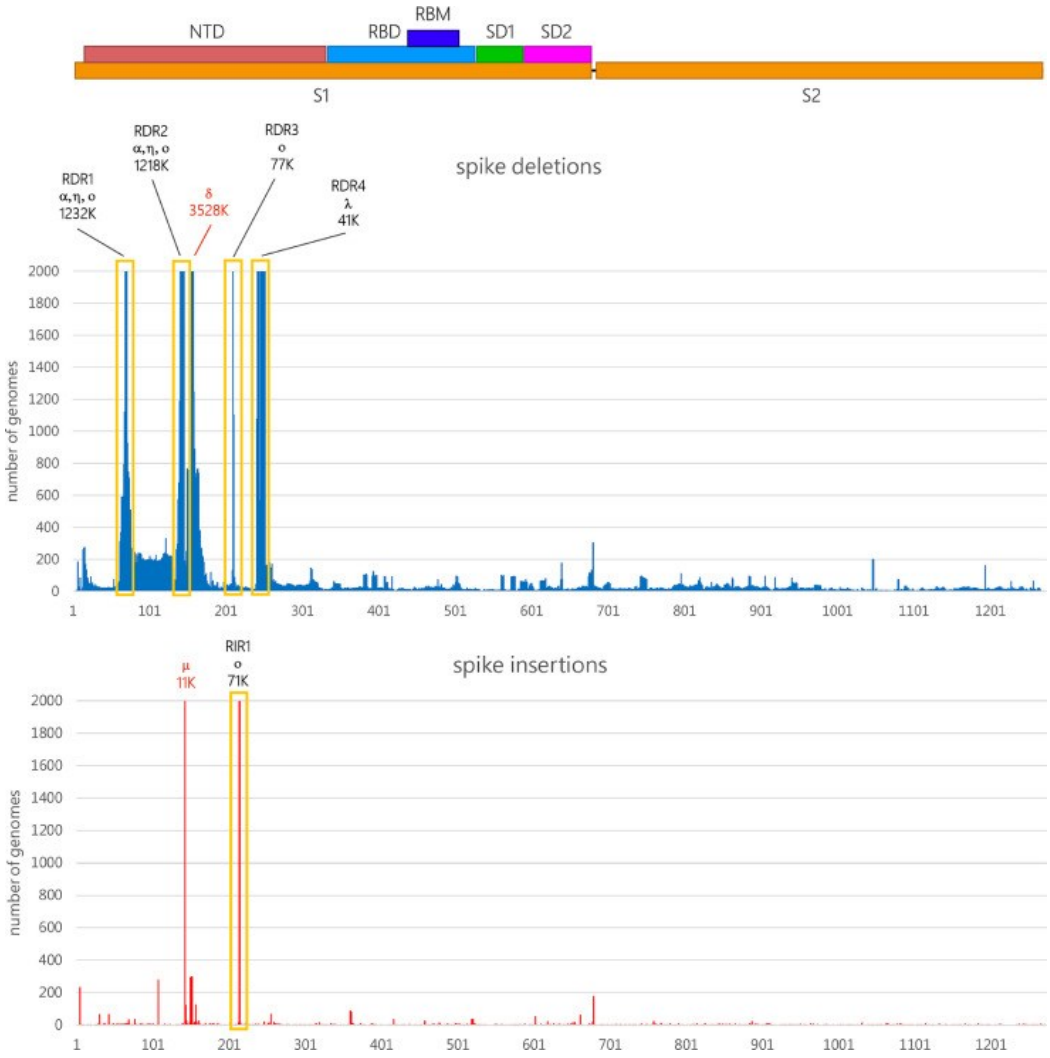


Figure 3.1: *Schematic representation of the SARS-CoV-2 protein, with indication of the two functional S1 and S2 subunits, which are separated by a furin-like proteolytic cleavage site, the N-terminal domain (NTD), the receptor binding domain (RBD) and receptor binding motif (RBM), the SD1 and SD2 subdomains. The absolute number of observed deletion mutations along the S-gene are reported (https://mendel.bii.a-star.edu.sg/ was last accessed on January 5th, 2022). Bars were truncated at 2000 observed genomes; in such cases, the approximate absolute number of observations is reported above the truncated bars, together with the main VOCs and VOIs associated with each indel, indicated with a Greek alphabet letter. The position of RDR1-RDR4 from a previous study* [64]*, as well as the deletion 157/158 characterizing the delta variant and the ins145T insertion characterizing the mu variant, are reported.*

RIR1 insertions independently emerged in multiple viral lineages

As of January 5th, 2022 RIR1 insertions could be documented as the result of at least 49 independent events that occurred in different branches of the SARS-CoV-2 phylogenetic tree, which strongly suggests convergent evolution. Even though the length of the insertion spanned from one to eight

codons (**Figure 3.2**), the overwhelming majority of the genomes with RIR1 insertions only included three codons (**Table 3.1**).

The most prominent viral variant carrying an insertion at RIR1 (the XLI 215:EPE insertion, see **Table 3.1**) is undoubtedly the emerging VOC omicron (which includes the lineages B.1.1.529, BA.1, BA.2 and BA.3), which, as of January 5th, 2022, is rapidly outcompeting delta worldwide. Although omicron was first detected on November 8th, 2021, time-calibrated Bayesian phylogenetic analyses suggest that it might have been spreading undetected in Southern Africa since early October [84]. The XLI insertion is paired with three small deletions in the spike NTD: (i) Δ69/70 at RDR1, which has been previously suggested to act as a "permissive" mutation to compensate otherwise slightly deleterious immune escape mutations [85]); (ii) Δ143/144/145 at RDR2, known to fall within a relevant antibody epitope [64];

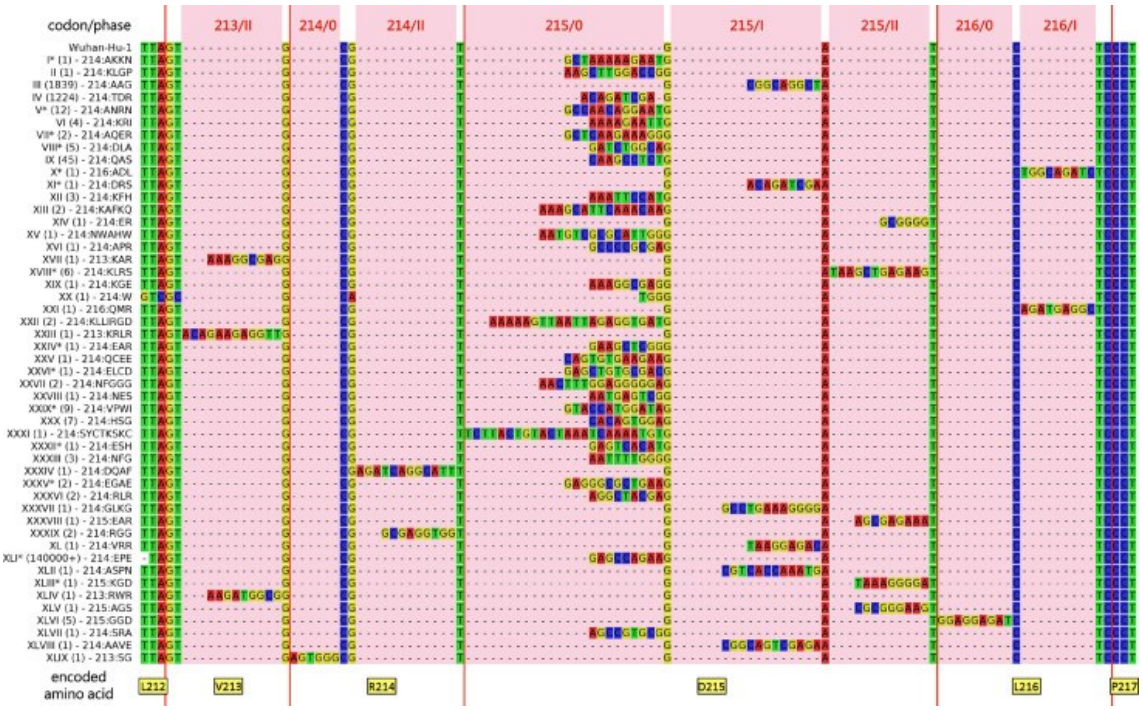


Figure 3.2: *Multiple sequence alignment of the nucleotide sequences of the SARS-CoV-2 S gene of the viral lineages characterized by an insertion at RIR1, compared with the reference sequence Wuhan Hu-1. The multiple sequence alignment only displays a small portion of the S gene and of the encoded spike protein, zoomed-in and centered on RIR1 (i.e. codons 212–217). Red vertical bars indicate codon boundaries, with the encoded amino acids (in the Wuhan Hu-1 reference sequence) indicated below. The number of observed GISAID entries for each insertion as well as the encoded amino acid sequences are shown near the insertion name. Please note that the exact position of all insertion could not be unambiguously detected in all cases; those with ambiguous placement are marked with an asterisk (see* ***Table 3.1*** *for details).*

(iii) Δ212, with unknown functional significance, located close to the RIR1 site. In addition, omicron carries an unprecedented number of non-synonymous mutations in the S1 subunit, some of which

had been previously described in other VOCs and VOIs and linked either with antibody escape, improved ACE2 binding or proteolytic cleavage. For example, these mutations include E484A, N501Y and P681H: the first one involves a residue also mutated in beta, gamma and mu (E484K), which plays a major role in polyclonal sera escape [86]. The second one is shared by alpha, beta, gamma and mu, and may increase ACE2 binding [37], [38], [39]. The third one, shared with alpha and mu, involves a residue mutated also in delta, where the substitution of proline with arginine dramatically enhances spike cleavage and viral fusogenicity [56].

This unusual pattern of mutations results in significant immune escape in vitro and in an enhanced reinfection potential in vivo [60], [61], [62], [63]. Moreover, omicron displays an altered cellular tropism and cell entry mechanism, which depend on the acquisition of an enhanced ability to rely on the TMPRSS2-independent endosomal route [63], [87]. It is presently unclear whether and to which extent the XLI insertion provides a contribution to the unique biological properties of this variant.

Unlike omicron, most RIR1 insertions were associated with very small local clusters that did not lead to further spread. However, the lineages A.2.5 (insertion III) and B.1.214.2 (insertion IV) were associated with a significant community spread (**Table 3.1**), reaching high prevalence in some geographical regions during 2021. While A.2.5 will be discussed in detail as a case study in the following section, it is worth briefly reporting here the transient spread of B.1.214.2. Following an initial importation from central Africa to Europe in late 2020, this lineage accounted for a non-negligible fraction of the covid-19 cases recorded in Belgium and Switzerland between March and April 2021. The spread of B.1.214.2, which led to over 1000 documented infections worldwide, was followed by a significant drop in its frequency of observation, which occurred in parallel with the rise of alpha, and further declined when delta became dominant. This lineage has not been detected since early July 2021 and can be thus provisionally considered as extinct. Insertion IV, which results in the addition of the TDR tripeptide between R214 and D215, was associated with the presence of two other non-synonymous spike mutations located on the RBD (i.e. Q414K and N450K). These have been previously linked with a moderate increase in RBD stability [88] and with immune escape both towards a few mAbs and towards convalescent sera [89], respectively. Due to the lack of functional data, it is presently unknown whether insertion IV and the other aforementioned mutations endowed this lineage with improved transmissibility or with increased potential for reinfection.

Several other insertion events at RIR1 occurred in lineages identified as VOCs or VOIs by WHO, CDC, ECDC or PHE, including some that have been recently de-escalated to the status of variants under monitoring. In detail, insertion V was found in twelve viral genomes belonging to the gamma lineage, sequenced in different Brazilian states and Guyana between December 2020 and April 2021,

indicating the presence of community transmission in the region. As reported in a previous work [90], these genomes belong to a monophyletic P.1-like clade that appears to be basal to P.1. The highly transmissible alpha lineage, which became dominant in Europe and quickly spread worldwide in early 2021 [91], before the rise of delta, was associated with at least six independent insertions at RIR1 (insertion XII, XIII, XIV, XVI, XXI and XLII) between February and November 2021 (**Table 3.1**, **Figure 3.2**). A single RIR1 insertion (XXXI) was recorded in August 2021 in the lineage B.1.525 (eta) in the United Kingdom [92] and two genomes characterized by the presence of insertion VII belonging to B.1.429 (epsilon) [93] were sequenced in California in January 2021. Several recently identified insertions at RIR1 are associated with delta (i.e. insertion XXIV, XXVI, XXVII, XXVIII, XXIX, XXX, XXXIII, XXXIV, XXXV, XXXVI, XXVII, XXXVIII, XXXIX, XL, XLIII, XLIV, XLV, XLVI, XLVII, XLVIII and XLIX). None of these have led to significant community spread to date, even though some were linked to small clusters of infections in England (**Table 3.1**, **Figure 3.2**). Albeit not directly linked with variants designated as VOCs or VOIs, other RIR1 insertions were associated with the presence of immunologically relevant spike mutations. This is the case of insertion IX (lineage B.1.639), which is characterized by the contemporary presence of E484K, T478K and by the deletions Δ69/70 (found in RDR1) and Δ144 (found in RDR2), which are shared by several VOCs and VOIs. Curiously, like the omicron insertion XLI, insertions XV, XXI, XXXI, XXXII and XLII also targeted viral genomes carrying both non-synonymous spike mutations at E484 and deletions at RDR1, suggesting a possible role of RIR1 insertions in compensating otherwise slightly deleterious mutations, like previously hypothesized for RDR1 itself [85].

Taking into account the limited efforts carried out by several countries in genomic surveillance throughout 2020 and 2021, the insertions reported in **Table 3.1** and **Figure 3.2** may just represent a fraction of those that emerged at RIR1 during the course of the pandemics. Although it was possible to unambiguously ascertain the exact placement of just 34 RIR1 insertions (see **Table 3.1**), most of them were in-frame, occurring at phase 0 between codons 214 and 215 (17 out of 34 cases, i.e. 50%), between codons 213 and 214 or between codons 215 and 216 (one case each) with no effect on neighboring codons. However, others were out-of-frame, occurring either at phase I (i.e. between the first and the second nucleotide of a codon) or at phase II (i.e. between the second and the third nucleotide of codon) (**Figure 3.2**). In detail, three insertions were observed at phase II within codon 213, two at phase II within codon 214, five and four at phase I and II, respectively, within codon 215, and a single one at phase I within codon 216. In such cases, the placement of the insertion often determined a non-synonymous mutation of the residues flanking RIR1 either at the N- or at the C-terminal side (**Table 3.1**). It is also worth noting that the omicron insertion XLI was associated with a three nucleotides-long, out-of-frame proximal deletion, which affected codons 211 and 212,

resulting in the deletion of a single amino acid. Similar deletions are uncommon in other lineages carrying RIR1 insertions, as they have been previously observed in a single other case, i.e. insertion XXII, which displays a Δ210 deletion.

Although the origins of the 49 RIR1 insertions was not investigated in the present study, other authors have previously suggested that they may result from the incorporation either of other regions of the SARS-CoV-2 genome itself, of host mRNAs [87], or of portions of genomic RNA of other endemic coronaviruses co-infecting the host [94].

***Table 3.1**: Summary of the 49 independent RIR1 insertions found in the SARS-CoV-2 genome, ordered by the earliest date of detection, as of January 5th, 2022.*

Designation	Insertion	Insertion type	Lineage	GISAID entries	Other spike mutations	Earliest detection	Latest detection
I	214:AKKN	ambiguous (codon 215 phase0/I)	B	1	none	Mar 5th, 2020	Mar 5th, 2020
II	214:KLGP	in-frame (codon 215 phase 0)	B.1.177	1	E154K, A222V, D614G	Nov 13th, 2020	Nov 13th, 2020
III	214:AAG	out-of-frame (codon 215 phase I)	A.2.5	1844	L141del, G142del, V143del, D215Y, L452R, D614G	Nov 20th, 2020	Nov 5th, 2021
IV	214:TDR	in-frame (codon 215 phase 0)	B.1.214.2	1228	Q414K, N450K, D614G, T716I	Nov 22nd, 2020	Jun 28th, 2021
V	214:ANRN	ambiguous (codon 215 phase0/I)	γ(P.1)	12	L18F, P26S, D138Y, K417T, E484K, N501Y, D614G, D1139H, V1176F	Dec 23rd, 2020	Apr 5th, 2021
VI	214:KRI	in-frame (codon 215 phase 0)	B	4	V367F, E990A	Dec 28th, 2020	Mar 15th, 2021
VII	214:AQER	ambiguous (codon 215 phase0/I)	ε(B.1.429)	2	S13I, P26S, S98F, W152C, L452R, D614G, T1027I	Jan 15th, 2021	Jan 18th, 2021
VIII	214:DLA	ambiguous (codon 215 phase0/I/II, codon 216 phase 0/I/II)	B.1.2	5	D614G	Jan 17th, 2021	Feb 2nd, 2021
IX	214:QAS	in-frame (codon 215 phase 0)	B.1.639	45	H69del, V70del, Y144del, M153T, T478K, E484K, D614G, T859N, D936Y	Jan 19th, 2021	Nov 4th, 2021
X	216:ADL	ambiguous (codon 216 phase I/II)	B.1.2	1	D614G	Jan 25th, 2021	Jan 25th, 2021
XI	214:DRS	out-of-frame (codon 215 phaseI/II)	B.1	1	D215N, V382L, D614G, M1237I	Feb 1st, 2021	Feb 1st, 2021
XII	214:KFH	in-frame (codon 215 phase 0)	α(B.1.1.7)	3	H69del, V70del, Y144del, N501Y, A570D, D614G, P681H, T716I, S982A, D1118H	Feb 12th, 2021	Feb 22nd, 2021
XIII	214:KAFKQ	in-frame (codon 215 phase 0)	α(B.1.1.7)	2	H69del, V70del, Y144del, A262S, N501Y, A570D, D614G, P681H, T716I, S982A, D1118H	Feb 25th, 2021	Feb 25th, 2021
XIV	214:ER	out-of-frame (codon 215 phase II)	α(B.1.1.7)	1	H69del, V70del, Y144del, D215G, N501Y, A570D, D614G, P681H, I712V, T716I, S982A,	Mar 11th, 2021	Mar 11th, 2021
XV	214:NWAHW	in-frame (codon 215 phase 0)	B.1.547	1	T19I, T95I, H69del, V70del, D614G, E484A, A879T, T1027I	Mar 22nd, 2021	Mar 22nd, 2021
XVI	214:APR	ambiguous (codon 215 phase0/I)	α(B.1.1.7)	1	H69del, V70del, Y144del, A262S, N501Y, A570D, D614G, P681H, T716I, S982A, D1118H	Mar 31st, 2021	Mar 31st, 2021
XVII	213:KAR	out-of-frame (codon 213 phase II)	B.1.177	1	A222V, A262S, P272L, D614G, P681H, M1229I	Apr 2nd, 2021	Apr 2nd, 2021
XVIII	214:KLRS	ambiguous (codon 215 phase II, codon 216 phase 0)	A.28	6	T76I, D215S, N501T, F592S, H655Y	Apr 23rd, 2021	Jun 22nd, 2021
XIX	214:KGE	in-frame (codon 215 phase 0)	B.1.1.519	1	T732A, T478K, D614G, P681H	Apr 24th, 2021	Apr 24th, 2021
XX	214:W	in-frame (codon 215 phase 0)	B.1	1	T19I, F140del, P139del, L141del, G142del, V143del, Y144del, Y145del, I210del, L242del, A243del, L244del, T470N, S494P, D614G, H655Y, T859N	May 14th, 2021	May 14th, 2021
XXI	216:QMR	out-of-frame (codon 216 phase I)	α(B.1.1.7)	1	L5F, S13I, H69del, V70del, Y144del, D215R, E484K, N501Y, A570D, D614G, P681H, T716I, S982A, D1118H	May 25th, 2021	May 25th, 2021
XXII	214:KLLIRGD	in-frame (codon 215 phase 0)	B.1	2	Q14del, C15del, V16del, N17del, L18del, W64R, T95I, C136Y, N137del, L141del, G142del, V143del, W152R, I210del, G252V, T415A, N440T, E484K, D614G, H655Y, P681H, T859N, Q1011H, G1219C	Jun 4th, 2021	Jun 4th, 2021
XXIII	213:KRLR	out-of-frame (codon 213 phase II)	B.1.1	1	R214Q, D614G, E484D	Jun 6th, 2021	Jun 6th, 2021
XXIV	214:EAR	ambiguous (codon 215 phase I/II/III)	δ(B.1.617.2)	1	T19R, N137K, G142D, E156G, F157del, R158del, L452R, T478K, E484Q, D614G, P681R,	Jun 8th, 2021	Jun 8th, 2021
XXV	214:QCEE	in-frame (codon 215 phase 0)	B.1.247	1	P209S, A222V, T572I	Jul 15th, 2021	Jul 15th, 2021
XXVI	214:ELCD	ambiguous (codon 215 phase I/II/III)	δ(AY.12)	1	T19R, T95I, E156G, F157del, R158del, L452R, T478K, D614G, P681R, D950N	Jul 17th, 2021	Jul 17th, 2022
XXVII	214:NFGGG	in-frame (codon 215 phase 0)	δ(AY.4)	2	T19R, E156G, F157del, R158del, L452R, T478K, D614G, P681R, D950N	Jul 22nd, 2021	Jul 22nd, 2021
XXVIII	214:NES	in-frame (codon 215 phase 0)	δ(AY.16)	1	T19R, G142D, E156G, F157del, R158del, L452R, T478K, D614G, P681R, D950N	Jul 30th, 2021	Jul 30th, 2021
XXIX	214:VPWI	ambiguous (codon 215 phase 0/I)	δ(AY.4)	9	T19R, T95I, G142D, E156G, F157del, R158del, L452R, T478K, D614G, P681R,V622F,	Aug 4th, 2020	Aug 21st, 2021
XXX	214:HSG	in-frame (codon 215 phase 0)	δ(AY.4)	7	T19R, T95I, D138H, G142D, E156G, F157del, R158del, L452R, T478K, D614G, P681R,	Aug 4th, 2021	Aug 25st, 2021
XXXI	214:SYCTKSKC	in-frame (codon 215 phase 0)	η(B.1.525)	1	A67V, H69del, V70del, Y144del, E484K, D614G, A653V, N679del, Q677H, F888L	Aug 5th, 2021	Aug 5th, 2021
XXXII	214:ESH	ambiguous (codon 215 phase 0/I/II)	B.1.240	1	C15F, L141del, G142del, V143del, Y144del, L242del, A243del, L244del, G446V, E484A, D614G, A688V, V1176F	Aug 6th, 2021	Aug 6th, 2021
XXXIII	214:NFG	in-frame (codon 215 phase 0)	δ(AY.25)	3	T19R, S112L, G142D, E156G, F157del, R158del, L452R, T478K, D614G, P681R, D950N	Aug 20th, 2021	Sep 14th, 2021
XXXIV	214:DQAF	out-of-frame (codon 214 phase II)	δ(B.1.617.2)	1	T19R, G142D, E156G, F157del, R158del, A222V, V289I, L452R, T478K, D614G, P681R,	Aug 21st, 2021	Aug 21st, 2021
XXXV	214:EGAE	ambiguous (codon 215 phase 0/I/II)	δ(AY.4)	2	T19R, T95I, G142D, E156G, F157del, R158del, L452R, T478K, D614G, P681R, D950N	Sep 2nd, 2021	Sep 2nd, 2021
XXXVI	214:RLR	in-frame (codon 215 phase 0)	δ(AY.3)	2	T19R, G142D, E156G, F157del, R158del, L452R, T478K, D614G, P681R, D950N	Sep 20th, 2021	Sep 20th, 2021
XXXVII	214:GLKG	out-of-frame (codon 215 phase I)	δ(AY.101)	1	T19R, T95I, G142D, E156G, F157del, R158del, L452R, T478K,D614G, P681R, D950N	Oct 1st, 2021	Oct 1st, 2021
XXXVIII	215:EAR	out-of-frame (codon 215 phase II)	δ(AY.4)	1	T19R, T95I, E156G, F157del, R158del, D215N, L452R, T478K, D614G, P681R, D950N	Oct 3rd, 2021	Oct 3rd, 2021
XXXIX	214:RGG	out-of-frame (codon 214 phase II)	δ(AY.100)	2	T19R, T95I, G142D, E156G, F157del, R158del, N394S, D614G, P681R, D950N	Oct 19th, 2021	Nov 8th, 2021
XL	214:VRR	out-of-frame (codon 215 phase I)	δ(AY.113)	1	T19R, G142D, E156G, F157del, R158del, D215H, L452R, D614G, P681R, D950N	Oct 26th, 2021	Oct 26th, 2021
XLI	214:EPE	ambiguous (codon 215 phase 0/I/II)	ο (BA.1/BA.2/BA.3)	>140000	A67V, H69del, V70del, T95I, G142D, V143del, Y144del, Y145del, N211del, L212I, G339D, S371L, S373P, S375F, G446S, S477N, T478K, E484A, Q493R, G496S, Q498R, N501Y, Y505H,T547K, D614G, H655Y, P681H, N679K, N764K, N679K, D796Y, N856K, Q954H, N969K, L981F	Nov 8th, 2021	Jan 2nd, 2022
XLII	214:ASPN	out-of-frame (codon 215 phase I)	α(B.1.1.7)	1	H69del, V70del, Y144del, L242del, A243del, L244del, G446V, E484K, Q498K, N501Y, A570D, D614G, P681H, T716I, W886L, S982A, D1118H	Nov 14th, 2021	Nov 14th, 2021
XLIII	215:KGD	ambiguos (codon 215 phase II, codon 216 phase 0)	δ(AY.4.2)	1	T19R,T95I,Y145H,E156G,F157del,R158del,G142D,A222V,L452R,T478K,D614G,P681R,D9	Nov 18th, 2021	Jan 3rd, 2021
XLIV	213:RWR	out-of-frame (codon 213 phase II)	δ(AY.4)	2	T19R, T95I, G142D, E156G, F157del, R158del, L452R, T478K, D614G, P681R, D950N	Nov 19th, 2021	Nov 22nd, 2021
XLV	215:AGS	out-of-frame (codon 215 phase II)	δ(AY.122)	1	T19R, G142D, E156G, F157del, R158del, L452R, T478K, D614G, P681R, D950N	Nov 29th, 2021	Nov 29th, 2021
XLVI	215:GGD	in-frame (codon 216 phase 0)	δ(AY.25)	5	T19R, G142D, E156G, F157del, R158del, L452R, T478K, D614G, P681R, D950N	Dec 5th, 2021	Dec 7th, 2021
XLVII	214:SRA	in-frame (codon 215 phase 0)	δ(AY.4.2.1)	1	T19R, V36F, T95I, G142D, Y145H, E156G, F157del, R158del, A222V, L452R, T478K, P681R, D614G, D950N	Dec 6th, 2021	Dec 6th, 2021
XLVIII	214:AAVE	out-of-frame (codon 215 phase I)	δ(AY.4)	1	T19R, T95I, G142D, E156G, F157del, R158del, D215N, N282S, L452R, T478K, D614G, P681R, D950N, A1020S	Dec 9th, 2021	Dec 9th, 2021
XLIX	213:SG	in-frame (codon 214 phase 0)	δ(AY.4)	1	T19R, T95I, G142D, E156G, F157del, R158del, L452R, T478K, D614G, P681R, D950N	Dec 16th, 2021	Dec 16th, 2021

*as of January 5th, 2022

These events would be most likely explained by poorly understood copy-choice recombination processes occurring during viral genome replication [95], [96]. Nevertheless, we caution that the short length of RIR1 insertions (usually nine nucleotides) is in most cases not sufficient to unequivocally establish the origins of the inserted nucleotide sequence, since several randomly occurring identical

sequence matches are expected to be found in a broad range of living organisms. On the other hand, when RIR1 insertions are relatively long, such as in the case of the S:ins214GLTSKRN insertion, seemingly acquired in vitro though repeated passages in Vero cell cultures, the robustness of such inferences might be significantly higher [87], [97].

Mutational pattern of A.2.5 lineage

As mentioned above, the only two lineages carrying insertions at RIR1 with solid evidence of widespread community transmission before the global emergence of omicron were A.2.5 and B.1.214.2. The inserted amino acid sequence found in A.2.5 is AAG, as the result of the phase I out-of-frame insertion of the nucleotide sequence CGTCAGGCTA within codon 215, which determines the non-synonymous substitution of Asp215 to Tyr (as a result of a GAT->TAT codon replacement) (**Table 3.1**, **Figure 3.2**).

Besides the insertion at RIR1, A.2.5 also displays the deletion of three codons (Δ141–143) in RDR2 (**Table 3.1**), sometimes extending to codon 144. This region has been previously implicated in antibody escape [64] and shows deletions in some relevant VOCs and VOIs, including alpha, omicron and eta. In particular, Δ144 appears to largely explain the resistance towards several NTD-directed mAbs displayed by alpha in vitro [45]. Moreover, the insertion at RIR1 is also combined with L452R, a key mutation that confers resistance towards class III RDB-directed antibodies [49], including LY-CoV555, the basis for the formulation of the commercial mAb bamlanivimab developed by Eli Lilly [51]. Among the lineages currently or previously designated as VOCs and VOIs, L452R is also found in delta, kappa and epsilon. Like the overwhelming majority of the variants circulating in 2021, A.2.5 is characterized by the presence of the prevalent mutation D614G. Although no other spike mutations are widespread in A.2.5, the A.2.5.3 sublineage acquired S477N, shared with omicron and known to strengthen the binding with the ACE2 receptor [98]. Overall, this mutation is associated with ~2% of all A.2.5 genomes (**Figure 3.3A**). Other relevant spike non-synonymous mutations, known to significantly alter either ACE2 binding or antibody recognition, were only seldom detected: K417T, N501Y and E484K (which are the hallmark spike mutations of gamma) were simultaneously found in a single genome (EPI_ISL_2,305,075, see **Figure 3.3A**) sequenced in Texas in May 2021. E484K was found in three additional cases (two in the United States, one in Canada) in April 2021, and N501Y was detected in nine additional cases (eight in the United States, one in Canada) between March and May 2021. Interestingly, in one such cases N501Y was paired with E484Q, which is found in kappa and determines reduced antibody sensitivity, even though not synergistically with L452R [99]. The acquisition of the mutation P681H, known to increase the efficiency of the furin-like

cleavage site was documented in 8 cases, 6 of which also displayed N501Y. Such insertions occurred independently in different branches of the A.2.5 evolutionary tree, indicating convergent evolution (see **Figure 3.3a**). Other lineage-defining non-synonymous mutations of A.2.5 are placed in other genomic locations. These included K1657E, F3071Y, T3255I (shared with delta and mu) and H3580Q in ORF1a; P1000L in ORF1b; S74F and G196V in ORF3a; S197L and M234I (shared with iota) in N (see **Figure 3.3b**). The functional consequences of these point mutations are presently unknown.

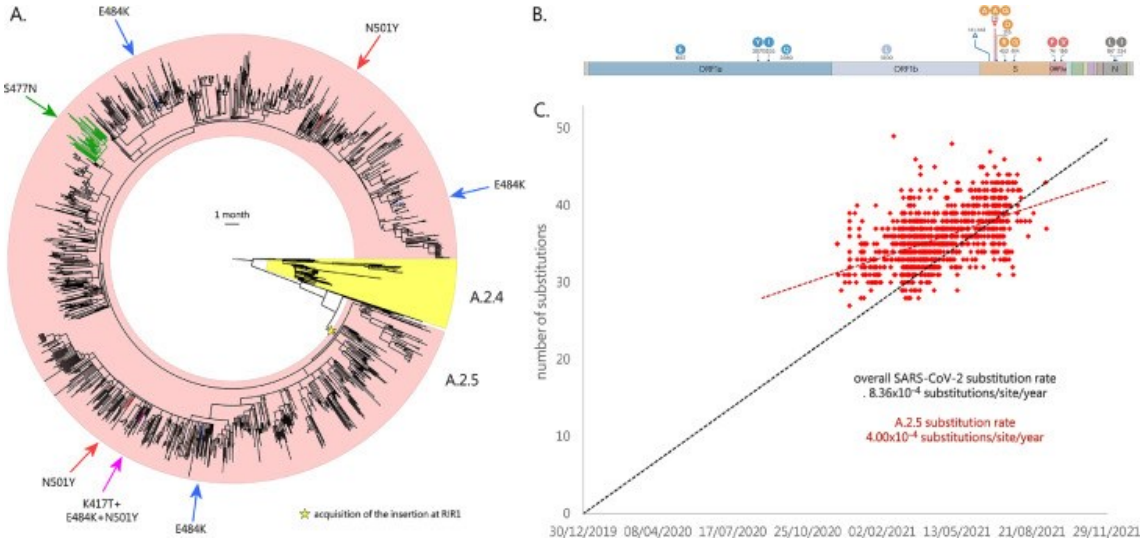


Figure 3.3: Panel A*: circular time tree exemplifying the phylogeny of the A.2.5 lineage related sublineages. Only high quality, complete genomes have been included. The Wuhan-Hu-1 strain was used to root the tree; the sister lineage A.2.4 is also indicated. The acquisition of relevant spike mutations placed in the receptor binding domain (i.e. S477N, K417T, E484K and N501Y) is marked with arrows. Please note that the monophyletic clade linked with the acquisition of S477N corresponds to the A.2.5.3 sublineage.* ***Panel B****: key mutations associated with the A.2. lineages. Genes associated with mutations (compared with the reference strain Wuhan-Hu-1) are indicated; only mutations detected in > 50% of the genomes belonging to this lineage and associated sublineages are shown. Modified from https://outbreak.info/.* ***Panel C****: root-to-tip genetic distance (number of nucleotide substitutions) of the genomes belonging to the A.2.5 lineage and related sublineages, compared with the reference genome Wuhan-Hu-1. The black dashed line represents the average rate of mutation of all SARS-CoV-2 sequenced genomes, according to GISAID (i.e. 25 substitutions per genome per year, as of January 5th, 2022). The red dashed line represent the rate of mutation computed for A.2.5. Note that insertions and deletions were excluded from this calculation.*

Root-to-tip genetic distance analysis revealed that the overall nucleotide substitution rate observed in the A.2.5 lineage (and related sublineages) was significantly lower than the average substitution rate computed for SARS-CoV-2 (based on GISAID data), as evidenced by the markedly different slope of the regression line (see **Figure 3.3c**). This was consistent with a substitution rate equal to 4.00×10^{-4} substitutions/site/year, i.e. roughly 12 substitutions/genome/year. Nevertheless, the A.2.5 SARS-CoV-2 genomes detected in the earliest phases of the spread of this lineage (i.e.

December 2020) were linked with a number of substitutions significantly higher than the average number of substitutions found in the same period in other SARS-CoV-2 lineages (i.e. ~33 vs ~25).

Emergence and international spread of A.2.5

A.2.5 belongs to one of the very few surviving children lineages of the ancestral lineage A, which, after several months of limited global spread, has led to a few major clusters of infections in 2021, such as the one which involved A.23.1 in Uganda [100]. A.2.5 stems from A.2.4, the dominant lineage in the Panama pandemics during the first half of 2020 [101]. The first documented cases can be traced back to late November 2020, all within a 100 km^2 area around the capital city Panamá. However, the precise timing of the emergence of A.2.5, along with the acquisition of insertion III at RIR1 and of the other associated mutations described in the previous section, is presently unclear due to the insufficient molecular surveillance carried out in Central America. To date, less than 1300 out of nearly 500 K covid-19 cases reported in Panama have been selected for viral characterization by sequencing, i.e. less than 0.3% of the total, far below of the threshold that would be sufficient to track emerging variants [102]. The presence of a number of genomes sampled in El Salvador and Guatemala, two countries where genomic surveillance has been virtually non-existing in 2020, in the earliest-branching clade belonging to A.2.5 (**Figure 3.3a**), leaves the precise geographical origins of this lineage unclear.

Nevertheless, A.2.5 undoubtedly underwent expansion in Panama between December 2020 and February 2021, as revealed by the increase in estimated prevalence from ~60 to ~95%. Interestingly, A.2.5 has been linked with clinically documented reinfections in individuals previously infected by the A.2.4 lineage, which is consistent with the presence of the constellation of non-synonymous spike mutations reported in the previous section, some of which may have immune escape properties [103]. The A.2.5 lineage likely spread very early also in the neighboring countries: while investigations carried out in August 2020 failed to identify A.2.5 in Costa Rica [104], the prevalence of this lineage in the country reached 30% between March and June 2021, with the establishment of large clusters of community transmission (**Figure 3.4**). A.2.5 may have undergone a similar spread in other countries in central America, including Belize, Honduras, El Salvador, Guatemala and Mexico, where multiple cases have been detected, starting from the spring of 2021 (**Figure 3.4**).

The remarkable spread of SARS-CoV-2 in Central America was connected with a significant number of exported cases, which have sometimes led to clusters of infection abroad. The first evidence of the detection of A.2.5 in southern America dates to December 1st 2020, in Ecuador. In this country, the acquisition of the spike mutation S477N, mentioned in the previous section, later led to the

establishment of the A.2.5.3 sublineage (**Figure 3.3** and **Figure 3.4**). Reports in other Latin American countries remain sporadic, but it is worth noting that A.2.5 genomes have been so far sequenced in Argentina, Suriname, Guyana, Grenada, Dominican Republic, Sint Marteen, Cayman Islands, Chile, Colombia, Venezuela, Brazil and Paraguay (**Figure 3.4**). The earliest cases exported in other continents were reported with similar timing in UAE (December 27th, 2020), Philippines (December 30th, 2020) and Australia (January 11th, 2021), which is consistent with the period with the highest incidence of covid-19 infections documented in Panama.

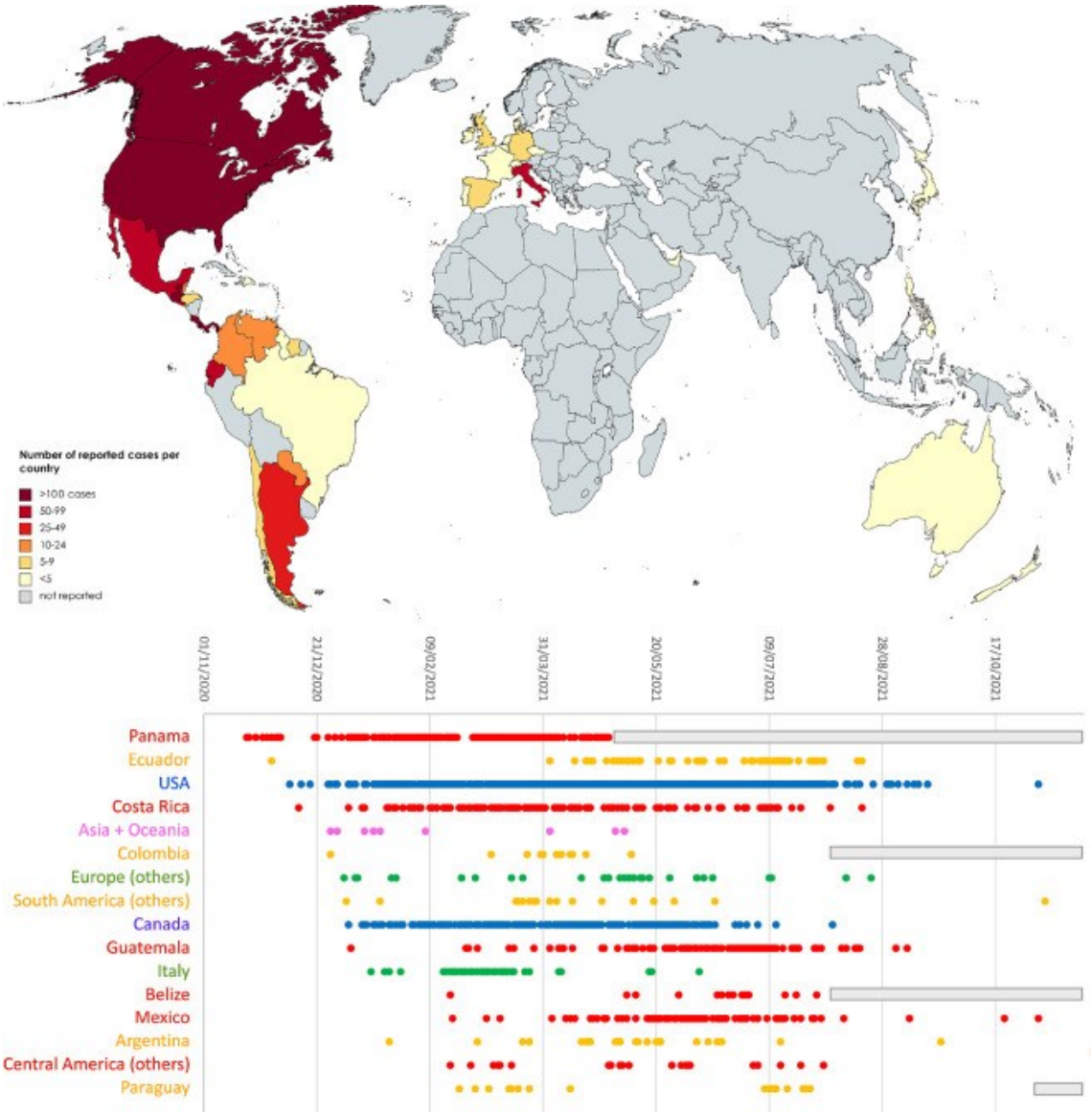


***Figure 3.4: Upper panel**: global spread of A.2.5 and related sublineages. **Lower panel**: detailed timing of the detection of sequenced genomes belonging to A.2.5 and related sublineages in different countries. Only countries with >= 10 unique days of detection are reported, whereas the others were collapsed in geographic macroareas (i.e. Asia + Oceania, Europe, South America and Central America). The reported dates refer to the dates of sampling reported in GISAID. gray boxes indicate periods of time with no sequencing data available for a given country.*

Cases linked with A.2.5 in Europe were identified in Luxembourg, Portugal, Germany, Italy United Kingdom, Czech Republic, France, Belgium, Ireland, Switzerland, Denmark, Netherlands and Spain. In most cases these did not lead to significant community transmission, with the exception of the cluster of cases linked with the A.2.5.2 sublineage recorded in Campania (central Italy) in February-March 2021 (**Figure 3.4**). Similarly, imported cases have most certainly led to local cluster of infections in different areas of the United States and Canada, starting from late 2020 (**Figure 3.3**). Nevertheless, the prevalence of A.2.5 in Northern America never exceeded 0.5%). No SARS-CoV-2 infections linked with A.2.5 have been identified to date in the African continent.

The global frequency of observation of A.2.5 and related sublineages underwent a rapid decline in the second half of 2021, in parallel with the global spread of delta. Just four genomes belonging to this lineage have been sequenced after October 1st, 2021, with the most recent GISAID entry (EPI_ISL_6960593) sampled in Brazil on November 8th, 2021 (**Figure 3.4**). The lack of recent sequencing data from Panama and other countries from Central and Southern America presently does not allow ascertaining whether A.2.5 disappeared in a similar fashion to what occurred for B.1.214.2 during the summer of 2021.

Impact of RIR1 insertions on the structure of the spike glycoprotein

RIR1 is located in a loop which connects the spike NTD β strands 15 and 16, a region which, unlike most RDRs, does not show any overlap with any known major NTD antigenic sites [105], [106]. Hence, the involvement of the insertions reported in this manuscript in antibody escape is unlikely, even though the possibility that this modification may lead to paired structural alterations at distantly related sites, leading to a reduced surface accessibility of canonical antibody epitopes cannot be ruled out. Moreover, the possibility that RIR1 insertions might significantly affect T-cell epitopes remains to be investigated, considering that the majority of T-cell response appears to be directed towards the spike NTD and the S2 subunit [107]. Comparative genomics investigations carried out on other viruses belonging to the *Sarbecovirus* subgenus revealed that the RIR1 has been previously prone to structural alterations during the radiation of bat coronaviruses [108]. In fact, in comparison with the spike proteins of other bat coronaviruses, RmYN02, RacCS203, BANAL-116 and BANAL-247 [67], [68], [109], which are among the closest known relatives to SARS-CoV-2 when genomic recombination is taken into account [110], comprise an insertion of four codons in a position close to RIR1.

Most certainly, the spread of the A.2.5 and B.1.214.2 lineages in different geographical contexts between late 2020 and early 2021, as well as the recent rapid global spread of omicron, indicate that RIR1 insertions are unlikely to have a detrimental impact on the three-dimensional structure of the spike protein or to significantly reduce the infectivity of these variants. At the same time, the well-defined length of the insertions (in the overwhelming majority of cases 3 or 4 codons) suggests that some critical structural constraints, that may prevent the selection of shorter/longer insertions or limit their associated evolutionary benefits, might exist. Several spike mutations located in the NTD can affect the structural organization of the spike protein, altering the stability of the interaction between the RBD and the ACE2 receptor, or its accessibility to antibody recognition. For instance, the NTD Δ69/70 deletion, which, like RIR1, is found in multiple independent lineages, does not determine a significant antibody escape in vitro [45]. However, it is thought to have an important impact on the structure of the spike protein, by compensating otherwise deleterious escape mutations [85]. In light of these observations, some NTD indels apparently not related with immune escape may act as permissive mutations, by compensating small infectivity deficits associated with other RBD mutations (i.e. L452R in A.2.5, Q414K and N450K in B.1.214.2, N440K, G446S, S477N, T478K, E484A, Q493R, Q498R, N501Y and Y505H in omicron).

Interestingly, the insertion of a seven amino-acid long peptide at RIR1 in SARS-CoV-2 through passages in Vero cell cultures has been recently implicated in enhanced in vitro infectivity, which may be linked with an increase in the positive charge of NTD surface [97]. According to the authors, this insertion (which is not reported in **Table 3.1** due to its laboratory origin) might have increased the affinity of the spike NTD to heparin, bringing viral particles in close proximity with host cells, thereby favoring the interaction with ACE2. While RIR1 insertions rarely share significant pairwise similarity both at the nucleotide and at the amino acid level (**Figure 3.2**, **Table 3.1**), we tested whether the amino acids found in the 49 RIR1 insertions were over-represented compared to expectations (assuming no codon usage bias). As shown in **Supplementary Figure S3.1**, the basic amino acids arginine and lysine were the most abundant ones (accounting for over 20% of total observations), followed by alanine, glycine and glutamic acid. Overall, lysine was the amino acid characterized by the highest observed/expected ratio (i.e. close to 3) and also arginine showed a moderate increase in frequency compared to expectations, supporting the conclusions by Shiliaev and colleagues about the benefits of acquiring basic residues at RIR1 for viral infectivity. Nevertheless, the two negatively charged residues (i.e. glutamic acid, found in the omicron 214:EPE XLI insertion, and aspartic acid) also had a positive observed/expected ratio, raising the question as to whether such benefits may apply to all charged residues. On the other hand, several amino acids with hydrophobic side chains (e.g. I, M, T, V and Y in particular) were strongly under-represented.

To preliminarily investigate the impact of RIR1 insertions on the structure of the spike protein, we applied molecular dynamics simulations, a well-known technique able to capture and study the dynamical properties of proteins and to assess the effects of mutations, deletions and insertions [111]. In this case, we have used a coarse-grained force-field to compare the structural and/or dynamical differences between the spike proteins from the wild-type virus and from the A.2.5 lineage. After 2μs of simulations, the RMSD of the backbone atoms relative to the equivalent initial structures (which represents a global measure of protein fluctuations) was calculated as a function of time to evaluate the stability of MD simulations equilibrium in the two systems. No significant global displacement was detected for any of the two protein models compared with the initial structure, as most of the RMSD values only displayed fluctuations in a range between 0.35 Å and 0.45 Å. Similarly, the presence of a few spike mutations in A.2.5 only led to minor changes in the compactness of the protein, as suggested by the differences of about 1 Å found in the average RYGR values among the two models (**Supplementary Figure S3.2**). On the other hand, some fluctuations were visible in the RMSF of the A.2.5 spike protein model, in particular in the regions which harbored non-synonymous mutations compared with the wild-type protein.

To understand changes in the direction of motions of the two systems under analysis, PCA was performed on the last 2μs of the simulations, the time after which the systems reached the equilibration state. The analysis was then restricted to the backbone beads, as they are less perturbed by statistical noise, providing at the same time a significant characterization of the essential space motions [81]. The diagonalization of the covariance matrix of fluctuations of the residues belonging to the backbone resulted in a set of eigenvalues, which were plotted in decreasing order against the corresponding eigenvector indices. The first few eigenvectors corresponded to concerted motions that quickly decreased in amplitude to reach some constrained and more localized fluctuations. Here we present the principal modes along the first eigenvector (**Figure 3.5a**), which covers about 25% of the motions of the protein. Consistently with the placement of non-synonymous mutations (**Figure 3.3b**), this analysis revealed that A.2.5 exhibited some changes in the fluctuations in regions belonging to the NTD and RBD, which are shown in red in **Figure 3.5b**. This indicates that the presence of the mutations and of the insertion at RIR1 may induce local structural and dynamical changes on the spike protein, highlighting the usefulness of performing studies on the dynamical properties of insertions upon their emergence in variants with widespread circulation.

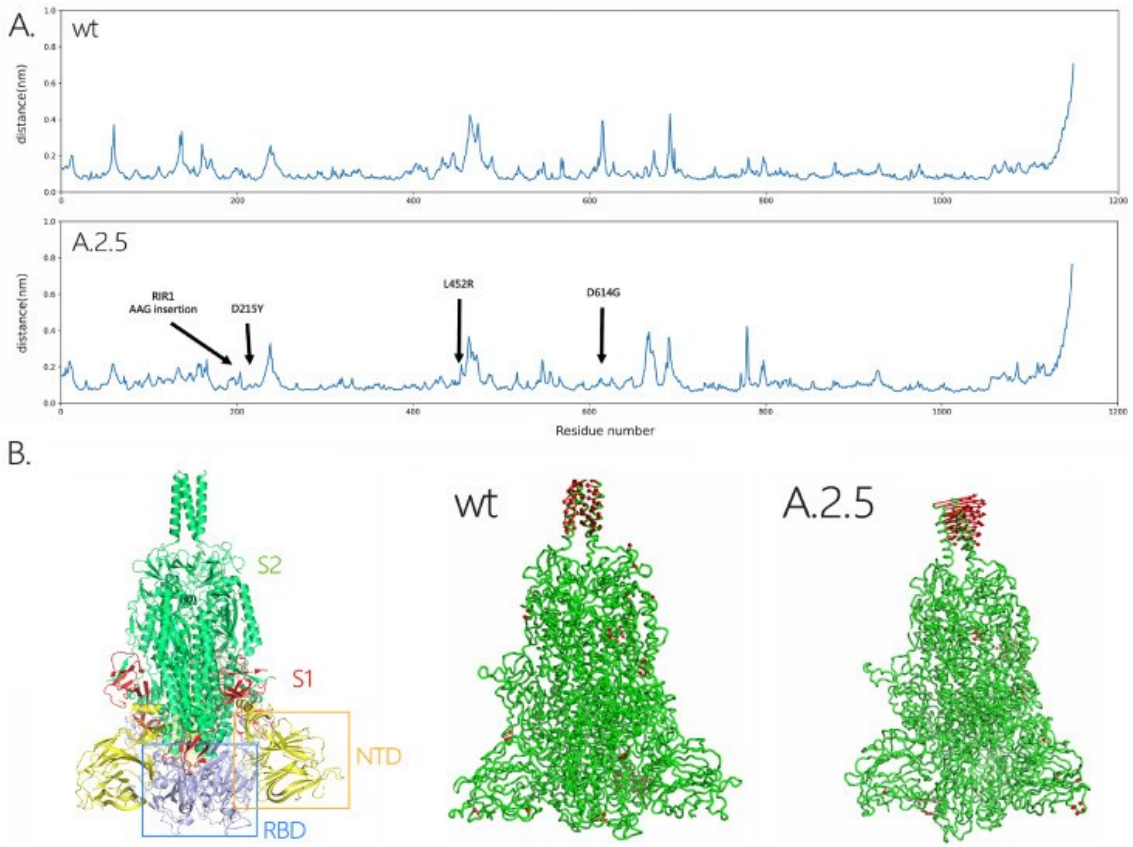


***Figure 3.5: Panel A**: RMSF plot for the models of the wild-type and A.2.5 SARS-CoV-2 spike proteins, with indication of the point and insertion mutations present in the two viral lineages target of his study, compared with the wild type virus. **Panel B**: Three-dimensional structural models obtained for the wild type and A.2.5 spike proteins. The location of the NTD and RBD (within the S1 subunit) and of the S2 subunit in the spike trimer are shown at the left-hand side. The regions where the most significant fluctuations are marked in red.*

Conclusions

The SARS-CoV-2 genome continues to accumulate mutations at a relatively constant rate, occasionally originating new VOCs and VOIs as a result of continued high viral circulation and natural selection. Prior to November 2021, the insertions at RIR1 documented in this work had only led to the emergence of two viral lineages with widespread transient distribution, i.e. B.1.214.2 and A.2.5, which now appear to be extinct. However, the presence of a RIR1 insertion in the emerging VOC omicron, together with the recurrent independent occurrence of this phenomenon by convergent evolution in multiple viral lineages (including alpha delta), suggests that RIR1 insertions may be linked with an evolutionary advantage, whose magnitude is presently unclear.

In absence of functional data, the role of RIR1 insertions can be only speculated. Based on the lack of overlap with known immune epitopes their involvement in immune escape phenomena appears unlikely, even though their impact on T-cell response remains to be investigated. Similarly, the previously hypothesized role of NTD insertions in enhancing viral infectivity by promoting the

interaction with host cell membranes is only partly supported by the over-representation of lysine and arginine residues in RIR1 inserts. On the other hand, we observe a correlation between the presence of RIR1 insertions, RDR deletions and several non-synonymous mutations found in the RBD with known impact on immune evasion, enhanced ACE2 binding and transmissibility. This, together with the predicted impact of RIR1 on the structure of the spike protein, may suggest a possible role as a permissive mutation in compensating otherwise slightly disadvantageous non-synonymous spike RBD mutations. Undoubtedly, our observations strongly suggest that the functional and structural impact of these insertions, with particular focus on omicron, should be the subject of in-depth studies.

An interaction-based drug discovery screen explains known SARS-CoV-2 inhibitors and predicts new compound scaffolds

This chapter describes my contribution to: Schake, P., ***Dishnica, K****., Kaiser, F., Leberecht, C., Haupt, V. J., & Schroeder, M. (2023). An interaction-based drug discovery screen explains known SARS-CoV-2 inhibitors and predicts new compound scaffolds. Scientific Reports, 13(1), 9204.* https://doi.org/10.1038/s41598-023-35671-x [112]

Introduction

The COVID-19 pandemic, which started in Wuhan (China) and then spread worldwide, has caused almost 609 million infections and more than 6 million deaths as of September 2022 (World Health Organization). Its causative agent the Severe Acute Respiratory Syndrome Coronavirus 2 (SARS-CoV-2) belongs to the Coronaviridae family of single-stranded positive-sense RNA viruse [113], [114]. Other viruses of the same family, namely the Severe Acute Respiratory Syndrome Coronavirus (SARS-CoV) and the Middle East Respiratory Syndrome coronavirus (MERS-CoV) [115] already led to epidemics in 2002/3 and 2012 respectively [116]. Due to the severity of the current outbreak, the scientific community has undergone huge efforts to experimentally determine SARS-CoV-2 genome sequences and three-dimensional structures as fast as possible. The unseen amount of publicly available data on a single virus is the groundwork for developing virus-specific drugs that could end the current pandemic. The SARS-CoV-2 genome encodes for structural proteins and non-structural proteins such as $3CL^{pro}$, PL^{pro}, helicase, and RNA-dependent RNA polymerase [117]. The four non-structural proteins mentioned above are key enzymes in the viral cycle [118].

The Main protease (M^{pro}) is being studied a lot in terms of structural and functional properties because of its high similarity, with significant conservation in the cleavage site, shared with SARS-CoV [119]. It is an enzyme involved in the processing of polyprotein which is translated from viral RNA [120]. Therefore, the inhibition of M^{pro} would ultimately suppress viral replication. Furthermore, there are no human proteases with a similar cleavage specificity as M^{pro}, making it very unlikely for M^{pro} inhibitors to be toxic [121]. Considering this evidence, we will put the main effort into the SARS-CoV-2 target M^{pro}.

In general, there are two main groups of methods that aim to identify new drugs for a given target, such as M^{pro}, which are computational and experimental approaches [122].

The wide range of in vitro experimental approaches performed to manage the pandemic includes studies aiming to determine appropriate drug targets [123], newly developed experimental methods to validate predicted drugs [124], [125], [126], experiments to uncover drug mechanisms [127], [128], [129], and high throughput drug repurposing experiments [130]. One of the most important outcomes of experimental approaches is the development of the by-now-approved drug Paxlovid, a combination of nirmatrelvir [131] and ritonavir, for treating COVID-19 patients with a very high risk of severe illness [132]. Furthermore, Boceprevir and GC-376 are identified as potent SARS-CoV-2 main protease inhibitors [133]. Nevertheless, experimental approaches in drug discovery require a high level of training, are expensive, and are generally less suited to perform large throughput studies to evaluate extensive compound libraries [134]. The above-mentioned drug Paxlovid for example is a derivative of a drug that was already developed as a potential SARS-CoV-1 inhibitor[131].

Besides in vitro approaches aiming to identify potential new drugs, others are aiming to detect three-dimensional active site structures and compound binding modes. Structures obtained and published in the protein database (PDB) early on showed compound fragments in complex with M^{pro}. They revealed the importance of the residues His41 and Cys145 that comprise the catalytic dyad similar to M^{pro} of SARS-CoV-1 [135], [136]. Further work disclosed that in M^{pro} an oxyanion hole is composed of Gly143, partly Ser144, and Cys145 [121], [137] implying that a promising drug candidate should be able to interact covalently or noncovalently with at least one of these residues. However, these structures should be used with caution. It was shown that especially the M^{pro} structures generated with high-throughput methods are often lacking the representation of a possible important water molecule that could serve as a third catalytic residue and that the models are not on par with other structures in the PDB [138]. In addition, most structures are generated at temperatures of 100 K and thus are representing an active site configuration that is non-physiological, leading to errors such as the previously mentioned missing water molecule [139]. Nonetheless, structural approaches are

extremely important to get insights into protein function and have already uncovered the mechanism of the FDA-approved SARS-CoV-2 inhibitor Remdesivir [140].

To cope with the problems of experimental approaches and to make use of the available data, computer-aided approaches in drug discovery are becoming more and more popular and important [122]. Interestingly, the most prominent examples of in silico drug screenings against COVID-19 seem to be based on molecular docking or molecular dynamic algorithms. Benefitting from the increased computational power, molecular docking algorithms are now suitable to screen giga-sized compound libraries against a single protein target. Such studies are testing tens of billions of compounds and are predicting a wide range of chemically diverse compounds [141], [142]. Most screened libraries are focused on known drugs and their relatives, but other recent approaches are screening against libraries of natural compounds to increase the search space [134], [143], [144].

Still, the major drawback of most in silico screenings is the lack of proper prediction validation resulting in only modest outcomes of huge screenings and no fast and global solution for the current pandemic [145].

By using a large amount of available data on the main protease of SARS-CoV-2, we want to address the above-mentioned problems. First, available M^{pro} compound complexes are extracted from the PDB and their binding patterns get analyzed by the Protein–Ligand Interaction Profiler (PLIP) [146]. Second, all protein–ligand complexes in the PDB are screened to detect similar binding patterns and predict potential inhibitors. Since we noted a drastic increase in publicly available data after the screen was done we decided to use this information for a further validation step. The data available in the PDB, before and after the screen, is depicted as a timeline in **Figure 4.1**. Using this data and M^{pro} binding affinity values from ChEMBL we were able to semi-automatically validate the predictions. Following these steps, the predictions are not dependent on pure chemical properties and therefore expected to be very diverse, leading to potential interesting and never considered findings. The automated part of the validation does not require any wet lab work and only depends on publicly available data. The pipeline is summarized in **Figure 4.2**.

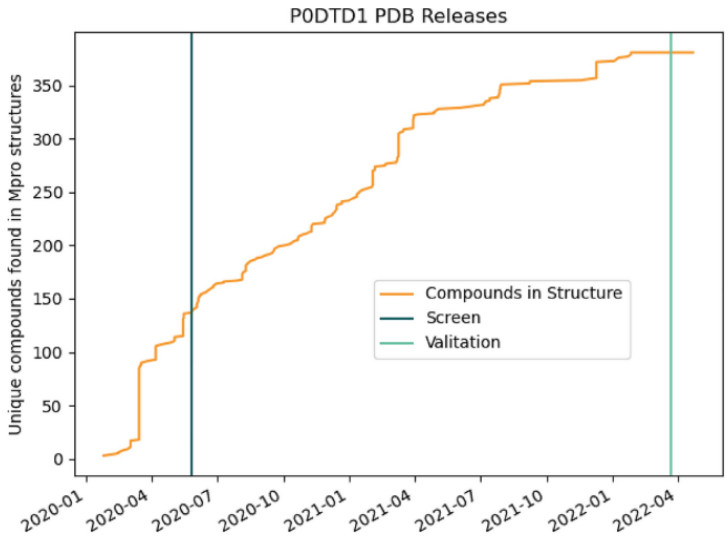


Figure 4.1: *Unique compounds released in complex with Mpro in the PDB. Structures are searched by the UniProt ID P0DTD1 and filtered for interactions with Mpro. Horizontal lines mark the days of Mpro inhibitor prediction and validation by data available in the PDB and ChEMBL.*

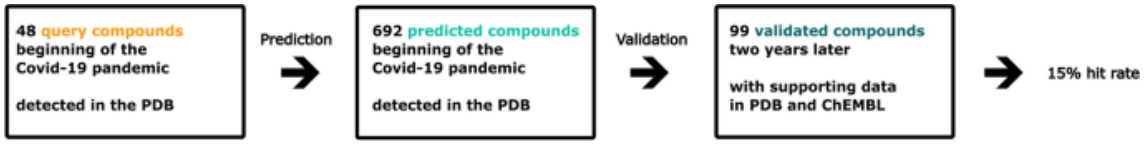


Figure 4.2: *Graphical abstract. The pipeline consists of three major steps. First (left panel) 48 query complexes of Mpro with co-crystallized ligands are extracted from the PDB. Second (middle panel) the interaction patterns are transformed into one-dimensional fingerprints and screened against the full PDB database resulting in 692 predicted compounds. Third (right panel) these predictions are validated with publicly available data leading to 99 validated compounds that are associated with SARS-CoV-2. The validation implicates a hit rate of at least 15%.*

This way, we were able to predict 692 unique potential Mpro inhibitors and validated 17% of the top 100 predictions retrospectively by publicly available data. The predictions cover a large chemical space and have great potential as lead compounds targeting Mpro. Within the top 100 predictions, we identified 4 already FDA-approved drugs that are currently under investigation for the treatment of the COVID-19 disease. The analysis of specific binding patterns within all available Mpro compound complexes in the PDB confirmed the importance of potential drugs interacting with the catalytic dyad of Mpro's active site. We furthermore detected an interesting pattern of three almost perpendicular hydrogen bonds interacting with hydrogen donors of an oxyanion hole within the active side. Our work contributes to the scientific community's efforts to detect potential lead compounds for a given protein target in a fast and reliable way.

Methods

Data extraction and prefiltering

A search of the PDB for M^{pro} on 21 March 2020 returned a set of 140 compounds found in complex with the protein. Those were filtered in two major steps. First generic and promiscuous compounds were filtered out using an in-house blacklist. Second, only those that bound the catalytic binding site of M^{pro} were considered, leaving only 48 compound- M^{pro} complexes. These 48 complexes served as input for an interaction-based screening using the PharmAI DiscoveryEngine (Version 2021.03, date 21 March 2021, https://www.pharm.ai). The small molecules in the PDB were set as target library for the predictions of the DiscoveryEngine.

Interaction based screening

In these screening approaches the way a given ligand is interacting with a protein is extracted using software, such as the Protein–Ligand Interaction Profiler (PLIP) [147] from three-dimensional complex data as provided by the protein database (PDB) [148] as well as geometric matches of ligand and binding site. The interactions are afterward converted into one-dimensional vectors (interaction fingerprints). Such interaction fingerprints can be compared with others using comparison schemes, such as the Tanimoto similarity index or comparable techniques, to screen large databases. The screen returned 740 unique compounds. Similar screening strategies have been used in [149], [150], [151].

Prediction evaluation and visualization

48 predicted compounds, which were already in complex with M^{pro}, were removed, resulting in 692 compounds. For these compounds, chemical fingerprints were computed using the Morgan fingerprint radius 2 and 512 bits [152]. The similarity of compounds was computed with the Tanimoto score, i.e. $|A \cap B| / |A \cup B|$ where A and B are two vectors. A random set of 400 compounds was created to determine a cut-off for dissimilar compounds. 200 were selected from the total of all 35.153 compounds in PDB and 200 from the total of 2.157.379 compounds in ChEMBL (March 2022). There was no overlap between the two groups. Pairwise Tanimoto scores were computed, and their distribution indicated that 99% of pairs have a Tanimoto score of less than 0.25. Thus, 0.25 was used as a cut-off for dissimilar compounds. Compounds were clustered using hierarchical clustering with single linkage from scipy [153]. They were visualized as a heatmap (**Figure 4.3**) with the cut-off of 0.25 to indicate dissimilar compounds. The multiple correspondence analysis and empirical cumulative density functions (**Figure 4.4** and **Figure 4.5**) were computed using scipy [153]. Interactions of compounds to M^{pro} were extracted from PDB files using PLIP 2.2.0 [147] and

visualized in Pymol [82]. The hydrogen bond triple motif was flagged if PLIP identified a hydrogen bond in M^{pro} residue 143, 144, and 145.

To validate the results, we searched PDB and ChEMBL for compounds known to interact with M^{pro} to compare those with our predictions. PDB and ChEMBL were searched for the M^{pro} Uniprot ID P0DTD1 on 9 March 2022 and 22 March 2022, respectively. PDB returned 471 unique compounds and ChEMBL 7.221. All considered PDB structures are generated by X-Ray Diffraction with a resolution of at least 2.4 Å (**see Suppl Appendix Table 4.1**). All interactions in ChEMBL are from the same screen (CHEMBL4495582) and results are reported as M^{pro} inhibition percentage at 20 µM by FRET kind of response from peptide substrate [154]. Inhibitory activity was normalized to the one of Zn-Pyrithione as the positive control (100%) and DMSO as the negative control (0%). For the confirmation of valid hits, we assumed that reported compounds with values above 0% inhibition are at least weakly active.

Results

Structure-based drug screening for M^{pro} reveals 692 potential inhibitors

To identify repositioning candidates for the inhibition of M^{pro}, predictions were provided by PharmAI (Dresden, Germany) as a result of an interaction-based screening. The screening revealed 692 potential M^{pro} inhibitors within the PDB. The predictions are further evaluated in three steps. First, their chemical properties are analyzed in terms of similarity to each other and known M^{pro} inhibitors. Here, we aim to find a heterogeneous set of predictions that cover chemical scaffolds beyond the already known ones with the potential of inhibiting M^{pro}. Such novel predictions may function as the basis for further evaluation and drug design. Our analysis revealed that the predictions are indeed very heterogeneous and do cover a large chemical space. Second, the predictions are searched for already known binders that are found in the PDB or ChEMBL to get a first idea of the predictive performance of the screen and to include publicly available data. Furthermore, predictions of high importance as already FDA-approved drugs are checked for an association as a M^{pro} inhibitor or COVID-19 drug in general. By that, we can confirm that 17% of our top 100 predictions have evidence of binding M^{pro}. Furthermore, 12 compounds are known to interact with other viral proteins of the replicase polyprotein 1ab, and we identify multiple FDA-approved drugs that are potential COVID-19 drug candidates. Third, we analyzed compound-M^{pro} binding patterns to detect potentially important binding modes and recognized a potentially important tripled hydrogen bond pattern.

Predicted compounds are heterogeneous

The chemical properties of 692 predicted compounds were evaluated. To get a first impression of the chemical relations in the large prediction set, we created a heatmap of their pairwise chemical similarities. All similarities are calculated as the Tanimoto similarity score of Morgan chemical fingerprints which is a 2D descriptor (see "Methods"). Such an analysis gives insights into how chemically diverse a set of compounds is. For example, similar compounds would form one or a few big clusters in such a heatmap while dissimilar ones would form none or multiple very small clusters. Ideally, the predicted compounds consist of new scaffolds covering a large chemical space. An outcome like this can give new insights into chemical species that should be considered as the groundwork for further drug design approaches.

Comparing chemical species is a challenging task and is usually done by transferring string representations of the compound into vector representations that can be compared by metrics such as the Tanimoto similarity index. Since all of such approaches come with their own benefits and drawbacks, we benchmarked the used combination of the Morgan fingerprint with radius 2 and 512-bit representation combined with the Tanimoto similarity index. Evaluating the similarity of 400 randomly selected compounds (**Figure 4.4**) revealed that 99% have a similarity of less than 25% suggesting that this is a meaningful cut-off to consider compounds related/unrelated.

The heatmap analysis (**Figure 4.3**) revealed that in all but one case only small clusters are formed. Similarities below 25% are whited out since those compounds can be treated as unrelated. The big cluster (118 out of 692 compounds) consists primarily of deoxyadenosine monophosphate derivatives. This result is not surprising since the already FDA-tested drug Remdesivir and its active metabolite GS-441524 are adenosine derivatives as well. These types of inhibitors are already shown to successfully inhibit viral replication. Other derivatives e.g. Cordycepin yield M^{pro} binding affinity [155], [156], [157]. This gives further support for the predicted compounds. Nonetheless, the majority of compounds are unrelated, suggesting that the predictions are indeed chemically diverse.

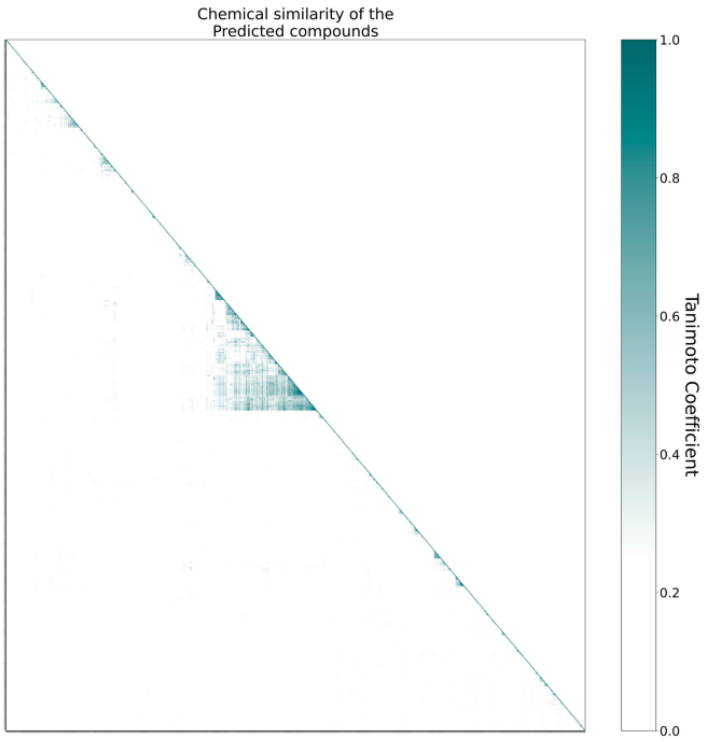


Figure 4.3: *Chemical similarity heatmap of the 692 predicted compounds. Since the underlying matrix is symmetric, the upper triangle is not shown explicitly. The analysis reveals little redundancy and a broad spectrum of scaffolds. The big cluster (middle) consists of compounds similar to deoxyadenosine monophosphate which is a group known to bind* M^{pro}.

How do the predictions relate to known inhibitors?

In general, predictions that cover a large chemical space are more likely to reveal interesting and novel scaffolds that can even be more important than a high hit rate [158]. **Figure 4.5** shows the multiple correspondence analysis (MCA) applied to the chemical Morgan fingerprints of our predictions and all compounds with structures available in the PDB where they are in complex with M^{pro}. Given in blue is the kernel density estimate (KDE), i.e. the probability distribution, of the PDB M^{pro} binders, orange dots mark the predictions, green dots mark query compounds, and magenta dots mark validated predictions. The analysis implies that the predictions fill a larger chemical space compared to the known binders and query compounds. Most of them are found in high-density regions of the known binders, which supports the overall approach since they do not form a whole new chemical space. The same holds true for validated predictions. However, we indeed identified compounds that are beyond the chemical space of known binders.

To access the heterogeneity of the predicted compounds even further we computed the pairwise similarity of 400 randomly selected compounds (200 ChEMBL, 200 PDB). The result is shown in the top panel of **Figure 4.4**. Only the set of query compounds seems to show some degree of

homogeneity with a mean chemical similarity of 0.23, which is still below our prior defined threshold. The randomly selected compounds, predictions, and known M^{pro} PDB binders have mean similarities around 0.125.

In summary, the predicted compounds seem to be as heterogeneous as known and tested M^{pro} binders while containing new scaffolds that may contribute to future efforts in developing a M^{pro}-specific anti-COVID-19 drug.

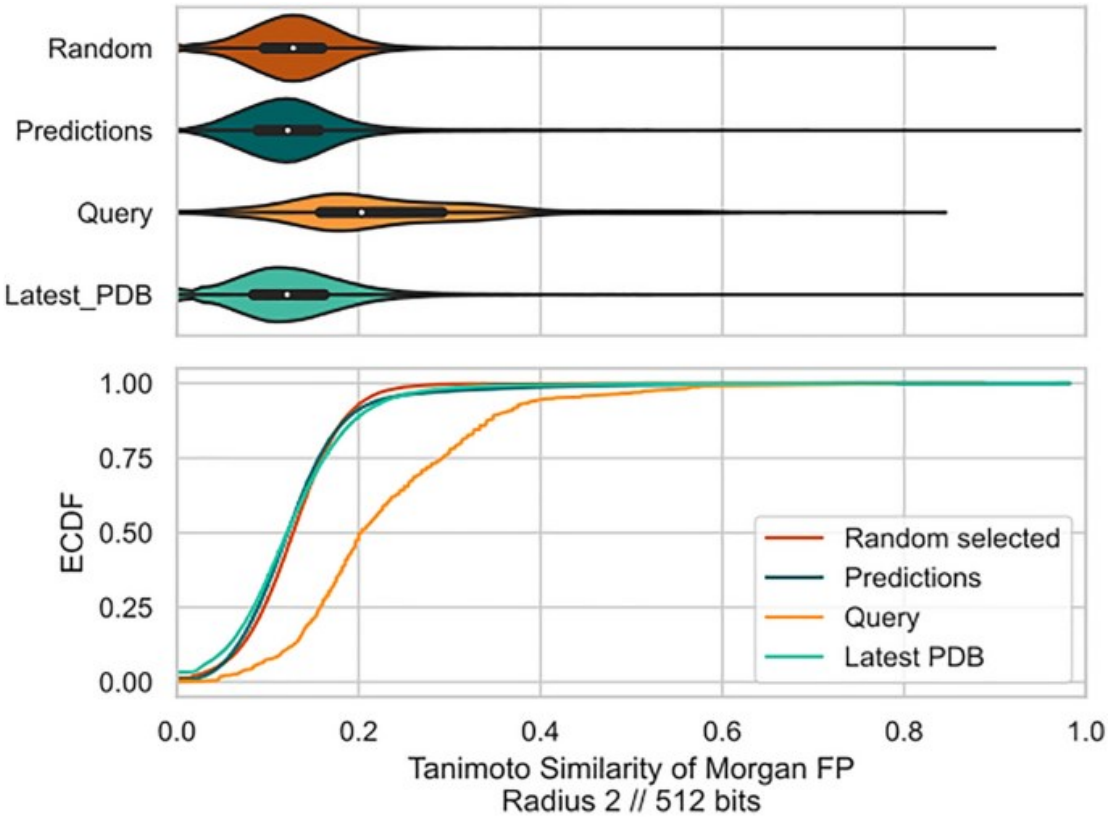


Figure 4.4: *Pairwise chemical similarity of predicted, random, latest PDB, and query compounds.* ***Top:*** *violine plot.* ***Bottom****: empirical cumulative density function (ECDF) of similarities. Query compounds are more similar to each other than predictions, which are as similar to each other as a random set of compounds. This indicates that predictions substantially expand from the queries and cover a vast chemical space. 99% of random compounds have a similarity of less than 0.25 suggesting that 0.25 is a meaningful cut-off to consider compounds unrelated.*

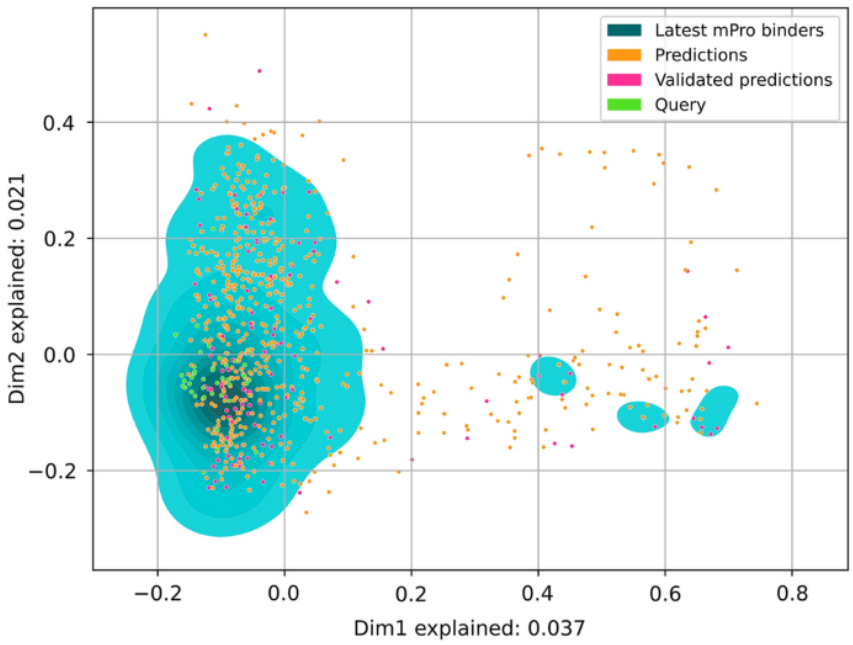


Figure 4.5: *Multiple correspondence analysis (MCA) of predicted- (orange dots), validated- (magenta dots), query- (green dots), and known- (blue surface) Mᵖʳᵒ binders. The axes of the MCA plot represent the dimensions of the data with the highest amount of explained variance. The analysis reveals that the predictions do cover a bigger chemical space than the known Mᵖʳᵒ binders with structures available in the PDB.*

The validation with publicly available data revealed a hit rate of 17%

After evaluating the predictions based on their chemical features, we aimed to validate them. Doing this for more than 600 compounds in vitro is a huge effort and we, therefore, make use of the astonishing amount of publicly available data on SARS-CoV-2. Here we have three principal approaches: first we extracted all compounds that are found co-crystallized with SARS-CoV-2 viral proteins in the PDB. **Figure 4.1** gives an overview of structures published with the UniProt ID P0DTD1 that are co-crystallized with Mᵖʳᵒ. Second, we searched ChEMBL for released affinity values of experiments with the target Mᵖʳᵒ (CHEMBL4523582). For this section of the analysis, ChEMBL was selected due to its accessibility and the thorough curation of the provided data. Lastly, we evaluated FDA-approved predicted drugs by literature search.

Compounds are considered to be validated in PDB if a structure is available with a predicted compound in complex with the protein target Mᵖʳᵒ. In addition to these four compounds, we identified another 12 which are found in complex with other proteins of the replicase polyprotein 1ab (**see Suppl Appendix Table 1**). After the screening was performed in 2020, 420 new structures of Mᵖʳᵒ were released, which serve as a basis for this part of the validation.

Since PDB is very limited due to its small number of available compounds (34,204) we investigated our results against ChEMBL as well. ChEMBL was searched for activity evidence on the reported

predictions and M^{pro}. Interestingly, to date, there is only data of a single high throughput screening on M^{pro} available in ChEMBL. For a total of 100 compounds, there is activity evidence, however only inhibition percentage values at 20 μM compound concentration are provided. Out of those 100 compounds, 76 show relative inhibition of > 10%, 30 more than 20%, and 11 more than 30%. It is therefore hard to judge if those are strong (nanomolar binders) or compounds that are only weakly interacting with M^{pro}. Detailed information on the predictions and validation data can be found in **Suppl Appendix Table 1**.

Nonetheless, the compounds are active which gives evidence beyond estimated interaction patterns, and even non-nanomolar binders are potential foundations for further drug optimization. Strangely, there is hardly any overlap between compounds found in ChEMBL and PDB even though M^{pro} is currently one of the most studied proteins. Among all 99 validated compounds, only 7 are found to have activity values reported in ChEMBL and a structure in complex with an viral protein available in the PDB. The lack of more activity data in ChEMBL can be attributed to the fact that ChEMBL has a very strict and standardized review procedure.

In summary, the performed in silico screening has an in vitro hit rate of 15% within all 692 predicted compounds and a hit rate of 17% within the top 100 predictions, ranked by p-values (**Table 4.1**). Thus, there is substantial evidence that the predictions are indeed valid drug candidates against SARS-CoV-2.

Table 4.1: *The top predictions are highly enriched in independently validated M^{pro} binders. Validation is done by evaluating with identical compounds that show inhibitory activity in ChEMBL or found in complex with M^{pro} in the PDB. Given values for PDB and ChEMBL validation do not consider any overlap.*

	PDB	ChEMBL	Both
Top 100	*2 (2%)*	*15 (15%)*	*17 (17%)*
All 692	*4 (0.5%)*	*100 (14%)*	*99 (15%)*

Further evaluation supports prior findings on four FDA-approved drugs

Next, we want to get a deeper understanding of these predictions. We assess them by the interaction motifs present in the query structures and predictions, by highlighting the two most strongly validated

predictions with evidence in both ChEMBL and PDB, and third by evaluating predictions of FDA-approved drugs with literature or clinical trial evidence as anti-COVID drugs.

Among the top 100 predictions, four are approved for use in humans by the U.S. food and drug administration (FDA), which are Flavin mononucleotide, Amodiaquine, Dasatinib, and Adenosine (**Figure 4.6**). Flavin mononucleotide (FMN) is an orange-red food color additive and is predicted in complex with UbiX from the psychrophilic bacterium colwellia psychrerythraea (PDB:4REH) [159]. In [160] authors gave evidence about the usage of riboflavin supplementation to decrease inflammation in COVID-19 patients. The malaria drug Amodiaquine is predicted in complex with human histamine N-methyltransferase (HNMT), which is a histamine-inactivating enzyme (PDB:2AOU) [161]. Amodiaquine was found to block SARS-CoV-2 infection with an EC50 value of 0.13 μM and was already proposed as a potential candidate against the early phases of the infection [162]. It was furthermore predicted to be a fruitful inhibitor of M^{pro} in a molecular docking study performed in [163]. Dasatinib is a known tyrosine kinase inhibiting drug approved for use in patients with chronic myelogenous leukemia and is predicted in complex with the human SH2-kinase domain (PDB:4XEY) [164]. In a clinical case, Dasatinib (100 mg/day) reduced fever, and a duplicate swab test came out negative two weeks later [165].

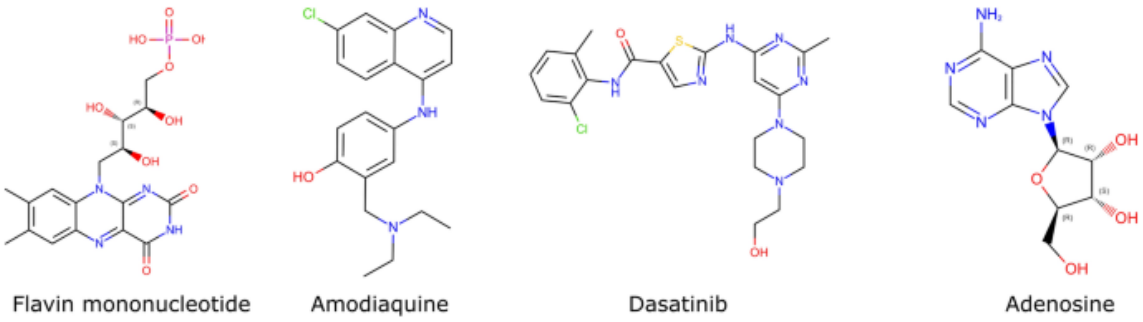


Figure 4.6: *Structures of four FDA-approved predictions with evidence on COVID-19. All are part of the top 100 predictions.*

However, it was unclear with which protein target the drug was interacting [166]. Furthermore, Dasatinib in combination with Quercetin reduces lung inflammation in SARS-CoV-2 infected hamsters and mice [137] and is now in phase two of clinical trials as an anti-inflammatory drug in patients with moderate and severe COVID-19 (https://clinicaltrials.gov/ct2/show/NCT04830735). Adenosine is an organic body-own compound and showed promising anti-inflammatory effects in COVID-19 patients when inhaled [167], [168]. In addition, the adenosine analog cordycepin was found to potently inhibit viral replication of resistant SARS-CoV-2 strains with an in vitro EC50 value of only 2 μM. Despite the existing evidence of viral inhibition, the specific mechanisms of

action for all four molecules remain unclear, necessitating the need for an in vitro demonstration of M^{pro} inhibition.

The evaluation of recently released PDB M^{pro} structures reveals a common interaction pattern

In addition to using recently published data on M^{pro} to validate inhibitor predictions, the data was used to get supplemental insights on the binding mode. Starting from the most high-level perspective on the interactions we calculated the frequency of each main interaction type. It was previously shown that the most frequent interaction type in the PDB are hydrophobic interactions [169]. As depicted in **Figure 4.7**, the most frequent interaction types among M^{pro} binders are hydrogen bonds followed by hydrophobic interactions and water bridges. There is some specificity in the compound M^{pro} interactions compared to what is generally present in the PDB.

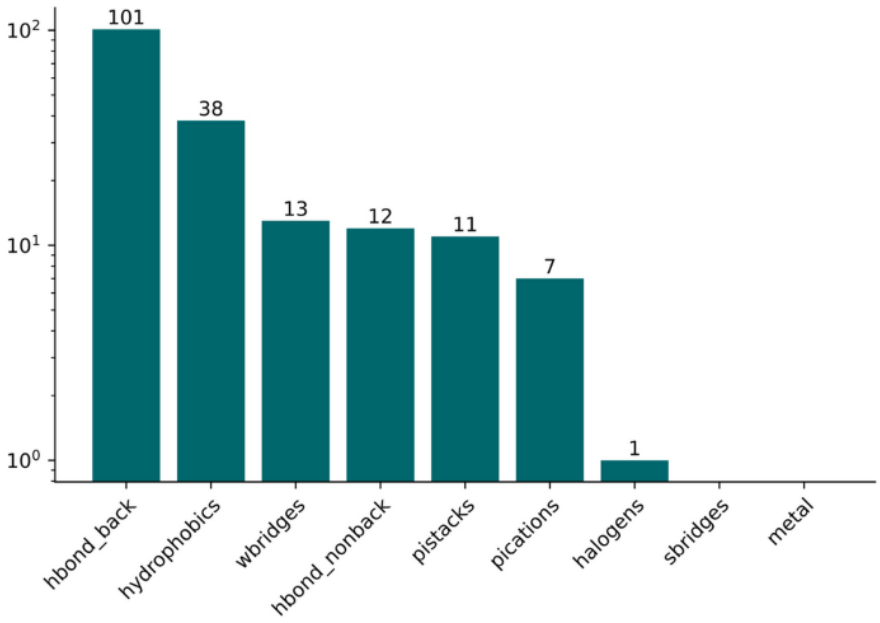


Figure 4.7: *Interaction types present in 48 query compounds.*

Not surprisingly, a total of 121 out of 471 unique compounds are interacting with one or both amino acids composing the catalytic dyad. Notably, the His41 residue exhibited a diverse range of interactions, with 39 pi-stacking interactions, and 23 hydrophobic interactions dominating the scene. Additionally, hydrogen bonds (8), pi-cation interactions (7), water bridges (4), salt bridges (2), and even halogen bonds (1) were also detected, providing a complex and intriguing picture of the binding interactions at play. Interestingly, Cys145 displayed a clear preference for hydrogen bonding interactions, with a remarkable 73 compounds interacting via this mode. Other interaction types, such

as water bridges (2) and hydrophobic interactions (1), were also observed, hinting at the complexity and diversity of the catalytic dyad's interactions with ligands.

Further investigation on M^{pro} binding modes results in the identification of a potentially interesting triplet hydrogen bond pattern present in 35 out of 471 structures.

In **Figure 4.8**, we showcased six examples that were used as input for the compound predictions. The compounds form three hydrogen bonds with the residues Gly143, Ser144, and Cys145. This finding is in agreement with what is reported in [136]. Here they found, that co-crystallized electrophilic ligands tend to form either two or three hydrogen bonds with Gly 143, Ser 144, or Cys 145. A similar pattern was previously reported by in [121] and is an addition to the importance of interactions with the catalytic dyad composed of His41 and Cys145. This triplet interaction is of major importance for the protease function since Gly143, Ser144, or Cys145 do function as hydrogen bonding donors of the oxyanion hole present in M^{pro}'s active side [170]. Therefore, we expect compounds that are able to dive deeply into the pocket and form interactions with those residues will efficiently inhibit the protease.

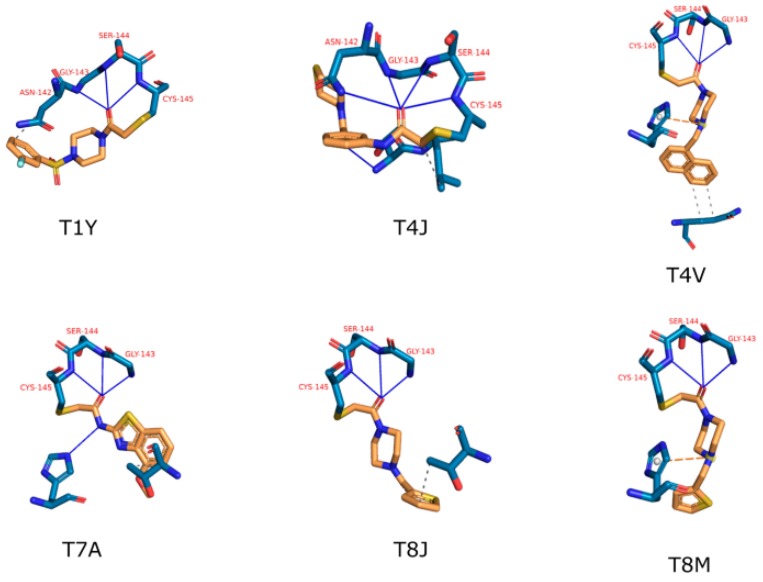


Figure 4.8: *Protein (blue) compound (orange) interactions of selected compounds. Blue lines mark hydrogen bonds, orange dashed lines mark pi-cation interactions, and dashed grey lines mark hydrophobic interactions. The three-letter codes refer to PDB chemical ids. Residues are indicated in red. A specific motif of three nearly perpendicular hydrogen bonds is present in six of the 48 query compounds.*

Turning the attention to our drug candidates, we identified a very similar pattern in three predicted structures (**Figure 4.9**), all of which are complexes with FDA-approved drugs. These cherry-picked examples show the opportunity of detecting similar patterns in different proteins by interaction-based prediction methods.

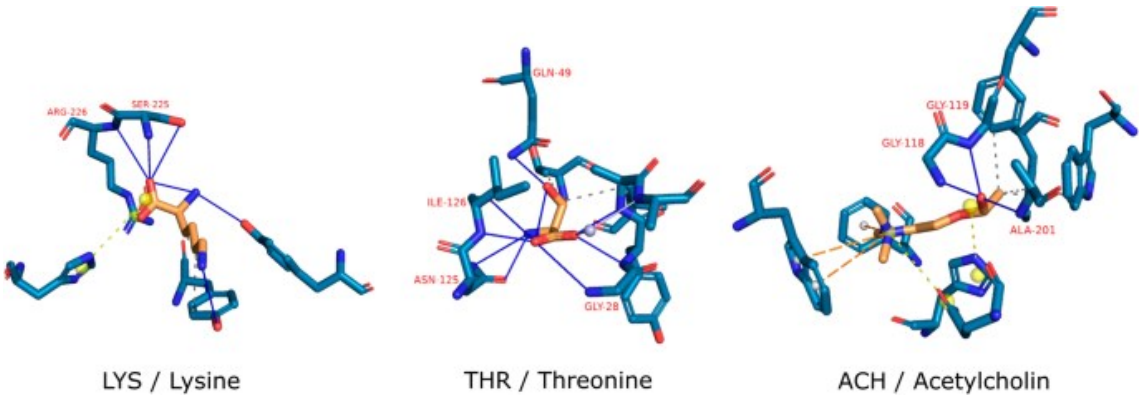


Figure 4.9: *Protein (blue) compound (orange) interactions of selected compounds. Blue lines mark hydrogen bonds, dashed orange lines mark pi-cation interactions, dashed yellow lines mark salt bridges, and dashed grey lines mark hydrophobic interactions. Residues are indicated in red. The three-letter codes refer to PDB chemical ids. Interacting proteins from left to right are: SET domain lysine methyltransferases (UniProt: Q43088), aspartokinase (UniProt: P9WPX3), and acetylcholinesterase (UniProt: P04052). The triple hydrogen motif is present in multiple predictions as well as in 35 out of 471 M^{pro} complexes in PDB.*

Discussion

The current COVID-19 pandemic exemplifies that fast-spreading diseases are a serious threat to modern society. By structure-based drug repurposing, we can predict a chemically diverse set of potential lead compounds against the main protease of SARS-CoV-2 with a success rate of 17%. Within the set of validated compounds, we identified several FDA-approved drugs, of which some are currently tested in clinical trials against SARS-CoV-2. Furthermore, we exploited the binding mode of known M^{pro} inhibitors and revealed the potential importance of a triplet hydrogen bond pattern for the protein–compound interaction.

Performing in silico drug screenings is a challenging task and comes with its own benefits and drawbacks. In contrast to wet lab studies, they are rather inexpensive, safe, and cheap. However, the result is only a prediction that requires experimental validation. Several researchers took the challenge of the COVID-19 pandemic and applied their very own algorithms aiming to predict fruitful drug candidates for multiple viral targets. Nonetheless, several of these studies do lack any kind of validation leaving the reader of such articles to judge themselves on how trustworthy the results in general are. Others created a full pipeline starting from in silico predictions which are then meticulously experimentally tested on important parameters, such as binding, cytotoxicity, metabolic stability, or oral receptivity.

Drug repurposing already led to some successes in the context of the COVID-19 pandemic. In [131], authors proved that by chemically modifying and improving a predicted lead compound an efficient drug against a given disease can be developed. Their drug Nirmatrelvir is now conditionally approved

in the EU and US. Even though this is a great success, their lead compound was already predicted as a potential drug against the SARS-CoV-1 outbreak in 2002. Still, it shows that experts in the field can rapidly develop potent drugs in a relatively short period of time when starting from an appropriate lead molecule. Following this assumption, we aimed to predict a chemically diverse set of potential M^{pro} inhibitors with our interaction-based approach. In doing so, the chances to detect so far unknown but potentially very important compound scaffolds are increased, giving more value to the predictions. We are able to show that the predictions are not only little redundant but furthermore cover a large chemical space including so far untested scaffolds. This is especially important considering that the query compounds used as the input for the prediction are far more homogeneous compared to the predictions and validated predictions. The same holds true for validated predictions, suggesting, that the scientific community is already heavily increasing the diversity of tested small molecules against COVID-19. Moreover, it is a proof of concept, that chemically diverse small molecules can still be effective as inhibitors for the same protein target.

This opens the gates for further developments based on our predictions. The most limiting factor is the availability of compounds in the PDB that are the only ones considered in the screen due to the requirement of protein–compound complexes as input for the algorithm.

Furthermore, the herein presented method aims to predict small molecules targeting a specific active site and does not allow for reliable predictions on molecules targeting e.g. allosteric binding sites. However, these can be included in a screen if interaction data is available in the PDB. By using publicly available data, we have created an intermediate approach that yields more trustworthy results than comparable in silico approaches but is not as powerful as those who considered experimental validation. With a hit rate of at least 17% within the top 100 predictions and 15% overall, the algorithm performance is substantial compared to similar approaches [142].

The evaluation of FDA-approved drugs within the predictions revealed the potential of the method to generate new hypotheses on drug mechanisms. All compounds are predicted to inhibit the main proteases of the Sars-CoV-2 virus and should therefore prevent viral replication. Through literature research, we identified articles on four FDA-approved drugs, showing beneficial effects in COVID-19 patients, that are within our top 100 predictions, and none of those reported any drug mechanism. The drugs Riboflavin, Amodiaquine, Dasatinib, and Adenosine have shown anti-inflammatory effects in COVID-19 patients or in-cell antiviral activity [160], [162], [166], [167], [168]. This raises the question of whether reduced viral replication mediated by the inhibition of M^{pro} as predicted by us is responsible for the reduced inflammation.

Ascorbic acid on the other hand is one of our validated and FDA-approved predictions but there is evidence that it is not applicable as a COVID-19 drug due to its inefficiency in infected patients [171], [172]. This exemplifies the limitations of the approach. Even if a drug does bind and eventually inhibits a target protein, there is no guarantee that it could function as a drug. Factors such as cell permeability, half-time, or other mechanisms can counteract the inhibitory properties of a compound. That can not be tested in a pure in silico fashion and does require wet lab work.

Anyway, the elephant in the room here is the other 82% of the predictions without validation. So far, there is no evidence of these compounds interacting with M^{pro} found in the PDB or ChEMBL. Therefore, this set of compounds may contain fruitful new lead scaffolds and their identification does require further experimental validation and evaluation.

Supplementary analysis on interaction patterns of recently released M^{pro}-compound complexes reveals a triplet hydrogen bond that could explain stable interactions and efficient inhibition. Compounds with such a binding mode do interact with all neighboring residues of the oxyanion hole (Gly143, Ser144, Cys145) and are therefore blocking its catalytic function. Since only 13% of the M^{pro} complexes in the PDB do show such a pattern, further investigations are required to test if those do have lower binding energy as we expect. Still, similar patterns are reported by different research groups highlighting the importance of further investigations regarding its importance on M^{pro} inhibition.

Conclusions

With our work on SARS-CoV-2, we can show that our interaction-based prediction method has great potential to predict a diverse set of potential lead compounds for a given protein target. Starting from a relatively homogeneous and small set of compound fragments bound to the main proteases of SARS-CoV-2, we predicted a chemically diverse set of potential inhibitors. Overall, we produced lead compound predictions at a very high hit rate by our interaction-based approach and were able to perform a first validation without the requirement of additional wet-lab work.

In this work, we benefited from the data-rich situation, but the method is applicable as long as there are complexes of the target protein bound to a compound available in the PDB. That way, we can provide a foundation for further lead optimization for lots of disease-associated proteins enhancing the drug development process.

Data availability

The interaction data used as input for the predictions can be found in **Supplementary Table S4.1** column "Query PDB ID:Chemical ID". The corresponding PDB files are publicly available from the

PDB (https://www.rcsb.org). All resulting predictions can be found in **Suppl Appendix Table 1** column "Hit PDB ID:Chemical ID".

Wide Real-Life Data Support Reduced Sensitivity of Antigen Tests for Omicron SARS-CoV-2 Infections

This chapter describes my contribution to: Piubelli, C., Treggiari, D., Lavezzari, D., Deiana, M., ***Dishnica, K.,*** *Tosato, E. M. S., ... & Castilletti, C. (2024). Wide Real-Life Data Support Reduced Sensitivity of Antigen Tests for Omicron SARS-CoV-2 Infections. Viruses, 16(5), 657.* [173]

Introduction

As of 31 February 2024, over 775 million confirmed cases of the novel coronavirus disease (COVID-19) caused by severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) and over 7 million deaths have been reported globally [174]. Although on 5 May 2023, the World Health Organization (WHO) declared the end of the COVID-19 global public health emergency, WHO still recommends governments to maintain a monitoring system, in case new variants emerge and cause another surge. Moreover, the accurate identification of people infected with SARS-CoV-2 is an essential prerequisite for facilitating the early initiation of therapy to reduce disease progression and for limiting the community spread of the infection [175].

The appearance and evolution of new variants with novel mutations require the monitoring of the available diagnostic methods for the detection of SARS-CoV-2 infection, based on both molecular and antigen testing. With the emergence of Omicron in particular, the effectiveness of antigen diagnostic tests (ADTs) was questioned. Diagnostic test sensitivity is a major criterion for detecting individuals infected with SARS-CoV-2 as fast as possible [176]. Most commercially available ADTs are based on the detection of the Nucleocapsid (N) protein, one of the four major structural proteins of SARS-CoV-2 [177], which has proven to be a good diagnostic target due to its high conservation rate [178], [179], [180], [181]. However, mutations also affect this gene. In fact, ADTs were developed

for the original SARS-CoV-2 N protein, and since the initial phases of the pandemic, new viral variants have been identified with specific patterns of mutations that could affect their detection due to epitope modification.

In October 2021, the Omicron variant (B.1.1.529) emerged in South Africa and started to be the dominant SARS-CoV-2 variant worldwide [84], [182], [183]. Omicron and its descendent sub-variants drew particular attention due to the high number of mutations. Their higher transmissibility and immune escape ability were assessed compared to the Delta (B.1.617.2) variant [59]. But how alteration in the N protein could influence antigen recognition by diagnostic tests has never been clarified. Omicron sub-variants have extensive mutations in its spike (S) and N proteins [184]. Mutations in the Nucleocapsid gene may lead to protein conformational changes that affect the target binding site of the ADT. This could, theoretically, alter the performance of the ADT in detecting this variant [185], [186], [187], [188]. The rapid global emergence and dominance of the Omicron variant highlighted the importance of understanding the performance of ADTs in real-world settings. Some in vitro studies suggested that the performance of rapid ADTs did not differ between the Delta and Omicron variants [189], [190], while studies using clinical specimens suggested a possible decrease in antigen tests' sensitivity for the Omicron variant [191], [192], [193], [194].

In addition to the variability of the N protein, in the early months of 2022, some studies hypothesized that Omicron variant infection could present a higher level of detectable viral RNA in the mouth than in the nose, with a positive predictive value of 100% in the saliva compared with 86% in mid-turbinate swabs [195]. These findings were supported by data from other labs, describing altered tissue tropism for the Omicron variant [196]. Another study did not support a preferred sample type for Omicron detection, but suggested a heterogeneous distribution of viral RNA in the nose and mouth [197], indicating that the choice of the sampling site still remains a controversial issue.

Based on the above considerations, further studies are needed to monitor the performance of these diagnostic tests in order to maintain accurate diagnoses throughout the evolution of the Omicron variant. Therefore, the aim of our study was to directly compare the results of ADTs with those of corresponding molecular tests in the same subjects, in a cohort of about 5000 patients attending our hospital during two epidemic waves, dominated by the Delta and Omicron variants, respectively. Moreover, we compared the viral loads present in the nasal nostrils and in the anterior oral cavity of Omicron-infected patients in order to assess whether nasal swab collection, which is generally the preferred practice for ADT testing, could still be suitable for Omicron descending variants, or whether it should be switched to a mouth swab, a sample type that could also reduce patient discomfort. Finally, an *in silico* study was performed to evaluate the effect of mutations on the

conformation of the N protein in the Omicron and Delta variants and their possible impact on molecular recognition by ADTs.

Materials and methods

This paper refers to the STARD 2015 guidelines [198] for the evaluation and reporting of diagnostic test accuracy.

Study Population

The test performance assessment included samples from 5175 subjects, either symptomatic or asymptomatic, who were referred to the IRCCS Sacro Cuore Don Calabria Hospital (Italy) between 1 October 2021 and 15 July 2022 for SARS-CoV-2 testing, most of whom were tested prior to hospital procedures or were contacts of infected persons. The enrolled subjects were assigned to either the Delta (1 October 2021 to 15 January 2022; n = 2726) or Omicron (from 16 January 2022 to 15 July 2022; n = 2449) wave, according to the viral variant dominating in the Veneto Region in the corresponding period. No information on the presence of symptoms was available. The inclusion criterion was the availability of results from both a SARS-CoV-2 ADT and RT-PCR on two parallel samples collected on the same day. According to the hospital's procedures, for each person, two different nasal/nasopharyngeal swab samples were concomitantly collected by trained healthcare personnel, one for ADT, according to the manufacturer's instructions, and the other for routine SARS-CoV-2 RT-PCR using eSwab® (COPAN Diagnostics Inc., Murrieta, CA, USA). Both samples were processed in the laboratory of the IRCCS Sacro Cuore Don Calabria Hospital within two hours of sample collection. Data were retrieved from the database of the internal Laboratory Information Management System (LIMS), including the date of collection, study patient code, age, sex, type of test assay, and result for SARS-CoV-2 testing.

For the comparison of the nose vs. mouth swabs, 61 subjects verified as positive for SARS-CoV-2, according to either a molecular or antigen test, were recruited during the Omicron period. For each subject, one mouth (buccal, internal cheeks, MS) and one nasal (anterior nares, NS) swab were collected in parallel by healthcare staff with eSwab® (Copan, Brescia, Italy). Both samples were analysed by RT-PCR for SARS-CoV-2 detection.

Ethics

The study was conducted in accordance with the ethical principles of the Declaration of Helsinki. Subjects or their legal representatives provided written informed consent. The study was approved by the local Ethics Committee (Comitato Etico per la Sperimentazione Clinica delle Province di Verona e Rovigo), protocol n° 17058/2022.

SARS-CoV-2 Antigen Diagnostic Tests

During the study, different ADTs were used for diagnostic purposes. Their main characteristics are summarized in **Table 5.1**. Each ADT test was applied by a nasal or nasopharyngeal swab according to the manufacturer's instructions, as indicated in **Table 5.1**. For the comparison among the different types of assays, the 6 tests used were grouped as follows:

ADT Group	3	3	2	1	1	1
Commercial Name	LIAISON® SARS-CoV-2 Ag	MAGLUMI® SARS-CoV-2 Ag	FREND™ COVID-19 Ag	Green Spring SARS-CoV-2 Antigen Rapid Test Kit	Panbio™ COVID-19 Ag Rapid Test Device	STANDARD Q COVID-19 Ag Test 2.0
Manufacture	DiaSorin, Inc	Shenzhen New Industries Biomedical Engineering Co., Ltd. China	NanoEntek, South Korea	Shenzhen Lvshiyuan Biotechnology Co. Ltd., China	Abbott Rapid Diagnostics Jena GmbH, Germany	SD Biosensor, Inc. Korea
Assay Type	Laboratory chemiluminescence-based ADT	Laboratory-chemiluminescence-based ADT	Microfluidic-based rapid ADT	Lateral Flow, Immunochromatography Rapid ADT	Lateral Flow, Immunochromatography Rapid ADT	Lateral Flow, Immunochromatography Rapid ADT
Target Protein	N	N	N	N	N	N
Sampling Site	Nasal swab	Nasophar. Swab	Nasophar. Swab	Nasal swab	Nasal swab	Nasal swab
Sensitivity	99.0%	97.7%	94.12%	96.77%	98.1%	94.94%
Specificity	98.0%	99.6%	94.12%	100%	99.8%	100%

Table 5.1: *Characteristics of SARS-CoV-2 ADTs used in this study. Sensitivity and specificity as reported by test manufacturers [ECDC, COVID-19 In Vitro Diagnostic Devices and Test Methods Database, available at https://covid-19-diagnostics.jrc.ec.europa.eu/devices/, (accessed on 30 December 2023)].*

- Group 1 ADT: Lateral Flow Immunochromatography rapid assay
- Group 2 ADT: Microfluidic-based rapid assay
- Group 3 ADT: Chemiluminescence-based assay

All the different ADTs mentioned were used without preference during the study period, according to the working needs of the laboratory.

SARS-CoV-2 RT-PCR Analysis

The swab specimens were analysed by routine SARS-CoV-2 RT-PCR. Briefly, RNA was extracted from 200µL of eSwabs medium using the automated Microlab Nimbus workstation (Hamilton, Reno, NV, USA) coupled to a Kingfisher Presto system (Thermo Fisher Scientific, Waltham, MA, USA) or using the EZ1 Advanced XL instrument with EZ1 DSP Virus Kit (Qiagen, Hilden, Germany) according to the manufacturer's instructions.

RT-PCR was performed using the Bosphore SARS-CoV-2/Flu/RSV IVD panel (Anatolia geneworks, Sultanbeyli/İstanbul, Turkey), targeting the Orf 1a/b and N genes, using a CFX96 Touch Real-Time PCR Detection System (Bio-Rad Laboratories S.r.l., Segrate/Mi, Italy). The amplification cycle threshold (Ct) was determined using CFX Maestro (Bio-Rad). Alternatively, the Real-Time PCR SARS-CoV-2 Panel Kit using NeuMoDx istrument (Qiagen Italia, Milan, Italy) was employed, targeting the N and Nsp2 genes. The Ct value for the N target was used as a proxy of the viral load in the corresponding sample. Cellular RnaseP mRNA was used as am endogenous control for the RT-PCR.

SARS-CoV-2 Genome Sequencing

Genomic sequencing for SARS-CoV-2 variant or lineage identification was applied to RT-PCR-positive samples from 168 patients from the Delta and Omicron waves. Reverse-transcription was performed with the SuperScript™ VILO™ Master Mix (Thermo Fisher Scientific, Waltham, MA, USA) in 20 µL of reaction volume, as per the user manual. The SARS-CoV-2 genome was amplified, according to the manufacturer's instructions, with the Ion AmpliSeq™ SARS-CoV-2 Insight Research Panel (Thermo Fisher Scientific, Waltham, MA, USA), with two primer pools protocol covering the whole SARS-CoV-2 genome. Amplified fragments were used to prepare barcoded libraries for massive parallel sequencing using the Ion AmpliSeq™ Library Kit Plus (Thermo Fisher Scientific, MA, USA), as reported in the user guide. The barcoded libraries were purified with Agencourt™ AMPure™ XP Reagent (Beckman Coulter), eluted in 50 µL of TE buffer, analysed on the 4150 TapeStation System (Agilent, Santa Clara, CA, USA) (average size 250–400 bp), and quantified by a Qubit™ Fluorometer (Thermo Fisher Scientific, Waltham, MA, USA). Each of the prepared libraries was diluted to 100 pM and pooled together; 30 pM of the library's pool was loaded on the Ion Chef™ Instrument (Thermo Fisher Scientific, Waltham, MA, USA) for clonal amplification and chip loading. The clonally amplified libraries were, shortly afterwards, subjected to next-generation sequencing on the Ion GeneStudio™ S5 System (Thermo Fisher Scientific, Waltham, MA, USA), on the Ion 520 or 530 chips.

Bioinformatic Analysis of Genome Sequences

The sequencing results were analysed in the Torrent Suite™ Software (v 5.14.1) using the SARS-CoV-2 plugins [i.e., generateConsensus, SARS-CoV_2_annotateSnpEff, SARS-CoV_2_variantCaller, SARS-CoV_2_coverageAnalysis (Thermo Fisher Scientific, Waltham, MA, USA)] with standard configuration. BAM files were visualized in the Integrative Genomic Viewer (IGV). FASTA consensus files were used for a lineage analysis with the Pangolin COVID-19 Lineage Assigner **https://pangolin.cog-uk.io** (accessed on 30 December 2023); sequences that passed the QC during Nextclade v2.8.1 analysis **https://clades.nextstrain.org/** (accessed on 30 December 2023) were further submitted to GISAID **https://gisaid.org/** (accessed on 30 December 2023). Specific sample mutations of the N gene were obtained from the CoV-GLUE database **http://cov-glue.cvr.gla.ac.uk/#/home** (accessed on 30 December 2023), and their frequency of occurrence was determined.

Nucleocapsid (N) Protein Mutation Analysis

The N-terminal domain (NTD) and C-terminal domain (CTD) X-ray structures were retrieved from the Protein Data Bank (PDB), with the accession IDs 6VYO and 6WZO, respectively [177]. The mutations in the proteins were mapped using the PyMOL software (v2.4.1) [82]. For the intermediate linker region (LKR), we used AlphaFold2 [199] for modelling the full-length protein (**Supplementary Figures S5.1 and S5.2**). After the mutation mapping, we ran the InterfaceResidues.py script to identify if they belonged to the dimerization interface. Using the Mutagenesis Wizard function in PyMOL, we changed residue S310 from Serine to Cysteine (S310C) (**Supplementary Figures S5.3 and S5.4**). The rotamer chosen for the substitution also did not cause any conflicts. This process was repeated for both chains in the dimer. We estimated the variation in protein folding free energy (ΔΔG) brought on by mutations, carrying out a qualitative evaluation of the N protein stability by using the webservers DynaMut [200] and DynaMut2 [201]. Additionally, using the DynaMut2 tool to compute the changes in vibrational entropy (ΔΔSVibENCoM), we investigated the potential impact on the monomer's flexibility. Due to the uncertainties in the modelling of the intrinsically disordered regions present in the LKR, this analysis was performed on the solved domains of the protein, i.e., the RNA-binding domain (i.e., NTD) and the N dimerization domain (i.e., CTD) of the N protein structures (PDB IDs 6wzo and 6vyo, respectively).

Statistical Analysis

Continuous variables were summarized with means, standard deviations (SD), and ranges (confidence interval, CI), while count variables were summarized with absolute and percentage frequencies. The

normality distribution of the data was assessed using the Shapiro–Wilk test. A comparison of the N gene Ct values between groups was performed using the Wilcoxon test. A comparison of the sensitivity and specificity of ADTs between the Delta and Omicron periods was performed using the two-sample Z-test for proportions. A comparison of the Ct values across time and sampling sites was performed, stratifying the analyses accordingly. R v. 4.2.3 [202], Graphpad Prism v. 10.1.0(316) (GraphPad Software, Boston, MA, USA) and SAS (SAS 9.4 Software, USA) were used to perform statistical analyses.

Results

Evaluation of ADT Performance in Delta versus Omicron VOCs Period

In order to analyse the performance of the ADTs during the Delta and Omicron waves, we retrospectively evaluated data collected from 5175 patients subjected to both ADTs and RT-PCR tests for SARS-CoV-2 infection in the period from 1 October 2021 to 15 July 2022. The demographic characteristics of the patients are summarized in **Table 5.2**. According to the data on the prevalence of viral variants in our region (Veneto, Italy) [203], we divided our study into two periods: the first, from 1 October 2021 to 15 January 2022, when Delta was predominant (Delta wave), and the second, from 16 January 2022 to 15 July 2022, when Omicron was predominant (Omicron wave). Taking into account the samples with an RT-PCR positive result, we evaluated the Ct values of the N gene and compared the results from ADT positive (+) and negative (−) specimens in the two periods. As expected, significant differences in the Ct values were observed between the ADT+ and ADT− samples, with a significantly lower median Ct value in the first group for both periods ($p < 0.0001$, for both Delta and Omicron periods, **Figure 5.1a**, **Figure 5.1b**). When comparing the two waves, significant differences were observed in the median Ct values detected for RT-PCR+/RADT+ ($p = 0.0009$) between the Delta and Omicron periods, indicating a generally higher viral load during the Omicron wave, detected at the RNA level for samples with a positive ADT, whereas no differences were found for RT-PCR+/RADT− (**Figure 5.1c**).

Table 5.2: *Descriptive statistics of the ADT study population. A total of 5175 subjects were considered. Subjects were divided according to Delta and Omicron waves.*

Demographics	***Delta Wave***		***Omicron Wave***	
	Count (n)	**Value (%)**	**Count (n)**	**Value (%)**
Population	2726		2449	
Female	1319	48.38	1216	49.65
Male	1407	51.61	1233	50.34
Age (years)	***Female***	***Male***	***Female***	***Male***

Lower 95% CI	46.71	46.81	39.01	41.20
Upper 95% CI	49.68	49.71	42.02	44.33
Median	48.00	52.00	37.00	41.00

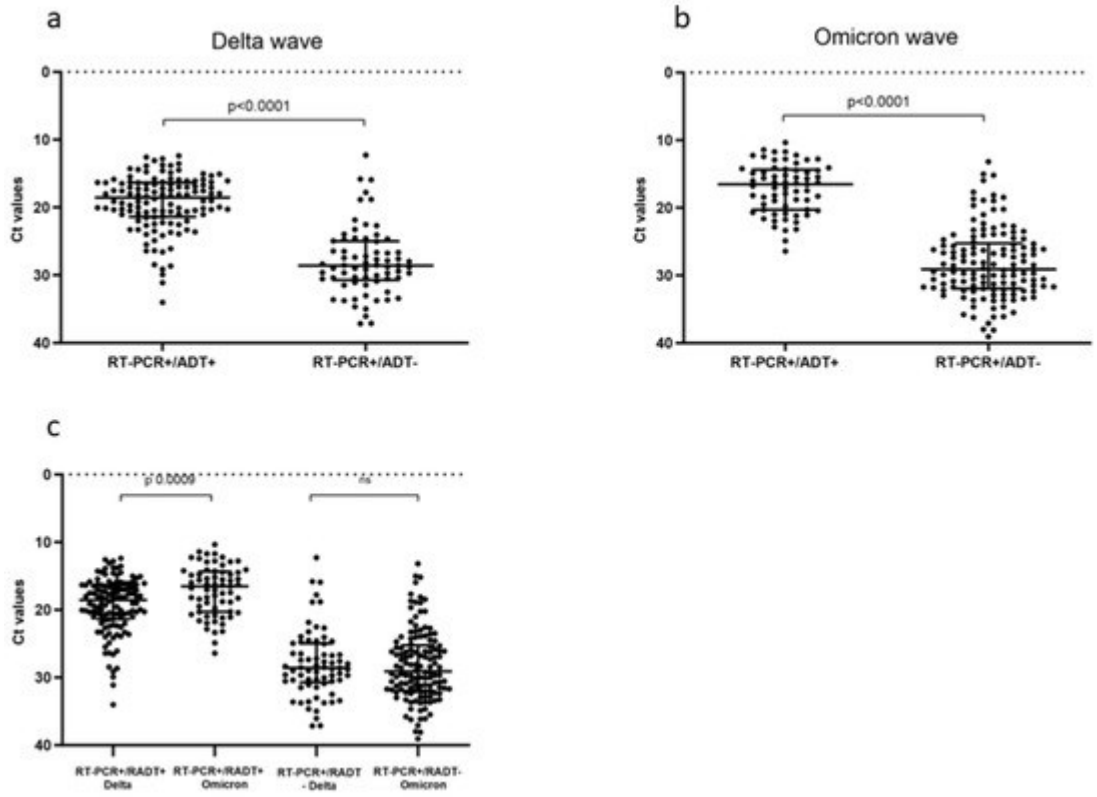


Figure 5.1: *Comparison of Ct values of SARS-CoV-2 RT-PCR according to ADT results. (**a**) Panel shows results during Delta and (**b**) panel during Omicron waves, respectively. (**c**) Comparison of Ct values of ADT-positive and -negative results for Delta and Omicron waves. Each dot plot represents an individual Ct value, error bars represent median with interquartile range (IQR). Wilcoxon test was applied to compare the difference of Ct values between the two groups. $p < 0.05$ was accepted as significant difference.*

We then compared the diagnostic performances of the ADTs between the two SARS-CoV-2 variant periods, using RT-PCR as a reference, and the results are reported in **Table 5.3**. During the Delta wave, 122 out of 2726 swabs (4.4%) tested positive by ADT and RT-PCR, and 2512 (92.1%) tested negative by both assays (overall concordance: 96.6%). We found 92 discordant samples (3.4%): 70 (2.5%) that tested negative by ADT and positive by RT-PCR, whereas 22 samples (0.8%) tested positive by ADT and negative by RT-PCR (**Table 5.3**). The sensitivity and specificity of ADTs during the Delta wave were 64% (95% CI, 56 to 70) and 99% (95% CI, 99 to 99), respectively (**Table 5.3**).

Throughout the Omicron wave, 65 out of 2449 swabs (2.6%) were positive by ADT and RT-PCR, and 2253 (91.9%) tested negative by both assays (overall concordance: 94.6%). We found 131

discordant samples (5.3%), of which 130 (5.3%) tested negative by ADT and positive by RT-PCR, and 1 positive by ADT and negative by RT-PCR. ADTs carried out during the Omicron wave achieved an overall sensitivity and specificity of 33.3% (95% CI, 26.8 to 44) and 100% (95% CI, 99.8 to 100), respectively (**Table 5.3**).

			Delta		**Omicron**		**Delta**		**Omicron**	
	Overall	**Group 1**	**TP/(TP + FN)**	**SE (95% CI)**	**TP/(TP + FN)**	**SE (95% CI)**	**TN/(TN + FP)**	**SP (95% CI)**	**TN/(TN + FP)**	**SP (95% CI)**
Group 2	122/192	27/54	73/112	65.2 (55.6, 73.9)	22/62	35.5 (23.7, 48.7)	1822/1838	99.1 (98.6, 99.5)	614/614	100 (99.4, 100)
Group 3	63.5 (56.3, 74.0)	50 (36.1, 63.9)	22/26	84.6 (65.1, 95.6)	11/37	29.7 (15.9, 47.0)	141/146	96.6 (92.2, 98.9)	143/143	100 (97.5, 100)
p	65/195	32/96								
	33.3 (26.8, 44.0)	33.3 (24.0, 43.7)								
	<0.001	0.045	<0.001		<0.001		549/550	99.8 (99.0, 100)	1496/1497	99.9 (99.6, 100)
	2512/2534									
	99.1 (98.7, 99.5)									
	2253/2254									
	99.9 (99.8, 100)									
	<0.001	0.460	0.020		<0.001					

Table 5.3: *SARS-CoV-2 ADT results during Delta and Omicron waves. RT-PCR results have been used as reference for calculation of sensitivity and specificity of ADT. Results were also stratified based on the assay type of the test as follow: Group 1 (Lateral Flow Immuno-chromatography rapid assay), Group 2 (Microfluidic-based rapid assay), and Group 3 (Chemiluminescence-based assay). SE: sensitivity, SP: specificity, TN: true negative, TP: true positive, FN: false negative, FP: false positive, and CI: confidence interval.*

For both sensitivity and specificity, the differences between the overall performances of the two tests during the Omicron and Delta periods were found to be statistically significant (both p-value < 0.001).The Positive and Negative Predictive values (PPVs and NPVs) of the ADTs for the two variants were calculated, considering the prevalence of SARS-CoV-2 infections during the two waves according to the GIMBE foundation, Italy, "**https://www.gimbe.org/** (accessed on 30 December 2023)", i.e., 2.6% during the Delta and 2.2% during the Omicron wave. The PPVs and NPVs resulted in being 65% (95% CI, 57 to 73) and 99% (95% CI, 98 to 99) for the Delta wave and 94% (95% CI, 85 to 98) and 98% (95% CI, 97 to 98) for the Omicron wave, respectively.

We compared the sensitivity and specificity of the Delta and Omicron periods in more detail by stratifying the different rapid antigen tests according to three types of assay. Specifically, Group 1 included all ADTs based on lateral flow immunochromatography rapid assays, Group 2 referred to microfluidic-based ADTs, and Group 3 included all chemiluminescence-based ADTs. As shown in **Table 5.3**, the results showed reduced sensitivities for each group of ADT during the Omicron period. Due to the small number of samples in split groups, statistical significance was only achieved for the overall analysis.

Evaluation of Nucleocapsid Protein Mutations in Delta and Omicron Variants

In order to assess whether specific mutations in the Delta and/or Omicron variants may affect the structure and function of the N protein, we analysed the amino acid sequence variations translated from the SARS-CoV-2 whole genome data, available from positive swabs in our study. Data were collected from 168 patients at the IRCCS Sacro Cuore Don Calabria Hospital in both the Delta and Omicron waves and were submitted to GISAID database. The protein is structured into three principal regions crucial for its activity: an N-terminal domain (NTD) responsible for RNA binding, a C-terminal domain (CTD) involved in dimerization, and an intermediate linker region (LKR) with a serine- and arginine-rich (SR-rich) motif [204], which, when phosphorylated, can regulate discontinuous transcription during the early stages of replication [205].

In the analysed sequences, a total of 33 different mutations were detected in the N protein sequence of the Delta and Omicron samples with respect to the original Wuhan sequence, and the LKR turned out to be the most affected region (**Figure 5.2a**). The frequency of each mutation is shown in **Figure 5.2b**. We observed that both Delta and Omicron VOCs showed exclusive mutations, e.g., D343G, P80R, and others among the most frequent ones are exclusive to the Omicron variant. Although the structure of the NTD and CTD domains of the N protein were solved, the full-length structure remains difficult to obtain due to protein stability issues and the presence of intrinsically disordered regions (IDRs) [206].

We performed *in silico* modelling of the full-length protein (**Supplementary Figures S5.1**) to evaluate the structural locations of mutations in the Delta and Omicron variants. Due to the uncertainties in the modelling of the IDRs, we focused on the experimentally solved 3D structures of the NTD and CTD (**Figure 5.3a**. and **Figure 5.3b**., respectively). We predicted the differences in folding free energy (ΔΔG) and vibrational entropy (ΔΔSVibENCoM) between the wild type and mutants in order to better understand how mutations may affect the protein stability (**Supplementary Table S5.1**). A positive ΔΔG indicates an increased stability, whereas a negative ΔΔG indicates a decreased stability.

A negative ΔΔSVibENCoM indicates an increase in protein rigidification, while a positive ΔΔSVibENCoM implies an increase in protein structure flexibility. An increase in terms of protein folding energy was predicted for P80R and H300Y and a decrease for D343G and S310C. The D343G mutation was shown to have the most negative ΔΔG, while P80R had the highest positive ΔΔG. These two mutations, both exclusive to the Omicron VOC, were connected with the largest increase (D343G) and decrease (P80R) in vibrational entropy, indicating an effect of these mutations on the Omicron VOC's structural/dynamic properties, which could lead to different recognitions by antibody-based detection systems.

Mouth versus Nose Viral Load in Omicron-Infected Patients

In order to assess the SARS-CoV-2 viral load at different sites (nose and mouth) during the Omicron wave, a total of 61 symptomatic patients (30 female, 31 male; mean ages of 43 and 44 years, respectively) were tested by RT-PCR in both the mouth and nose. Fifty-one subjects reported mild symptoms. In 49 out of the 61 patients (80% of the total population), the samples were collected less than 4 days after the infection diagnosis, and in 27 (44%) of these, the samples were collected at the onset of symptoms or the following day. Fifty-seven out of the sixty-one patients were positive for at least one of the two swabs (nasal or oral). In particular, 43 of them were RT-PCR positive on both sites, 12 patients were positive only in the nose, and 2 only in the mouth.

Four patients resulted in being negative in both sites, but all of them were sampled more than 6 days after symptoms onset. So, the number of RT-PCR-positive NS was higher than that of positive BS. In line with these results, when analyzing the samples' Ct, we found that the nose site presented lower Ct values, corresponding to a higher viral load compared to the mouth (Wilcoxon test, $p < 0.001$) (**Figure 5.4a**.). Possible changes in the viral load during the year were also investigated. We found that the nose was the site where the virus was more likely to be detected [significant results for the months of March ($p = 0.004$) and April ($p < 0.001$), **Figure 5.4b**.]. When performing a breakdown of the data according to days after symptoms onset, a higher viral load was always detected in the nose with respect to the mouth [significance at 1 day ($p = 0.006$), 2 days ($p = 0.016$), 4 days or more ($p = 0.002$), **Figure 5.4c**.]. After day 5 from symptoms onset, few samples showed a positive signal. Moreover, we analysed the Ct trend after symptoms onset at the two sampling sites based on the different identified Omicron subvariants (AY.4, BA.1, BA.1.1, BA.2, BA.2.9, BA.2.18, BA.5.1, and BA.5.2, with BA.2 being the most frequent). **Supplementary Figure S5.5** shows that none of the subvariants showed a higher presence in one of the two collections sites.

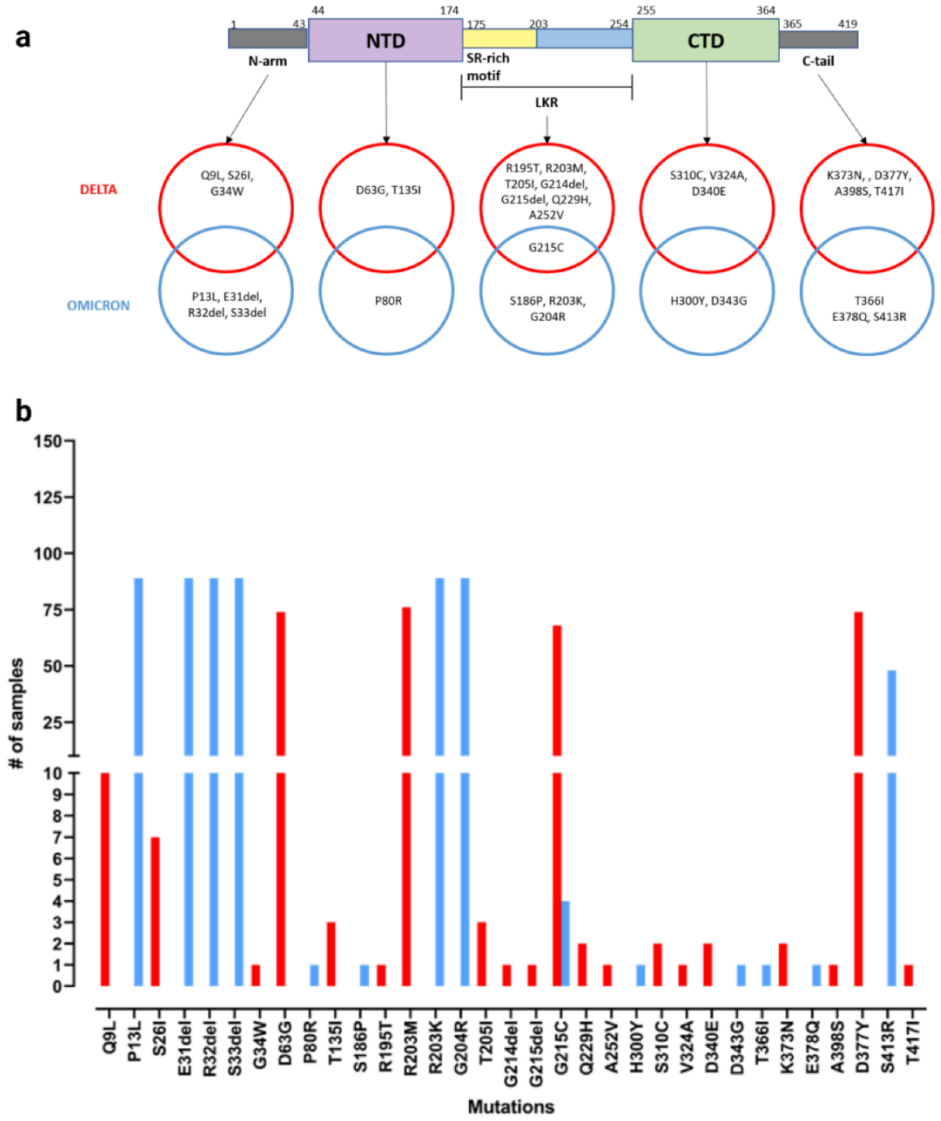


Figure 5.2: *Mutation in N protein of Delta and Omicron sequences identified in infected subjects included in the study. (**a**) Panel represents the modular structure of the SARS-CoV-2 N protein with mutations identified for Delta (in red) and Omicron (in blue) variants. (**b**) Panel shows mutation frequencies in Delta (in red) and Omicron (in blue) variants.*

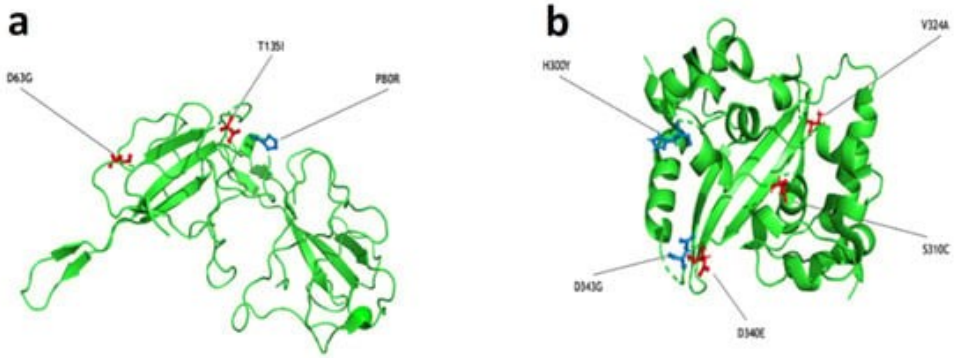


Figure 5.3: *Spatial representation of Delta and Omicron mutations within the N protein. The 3D structure of the NTD (**a**) and CTD (**b**) are reported. NTD and CTD structures are associated with the PDB IDs 6vyo and 6wzo, respectively. Mutations in red are characteristic of the Delta variant, whereas the blue ones belong to the Omicron variant.*

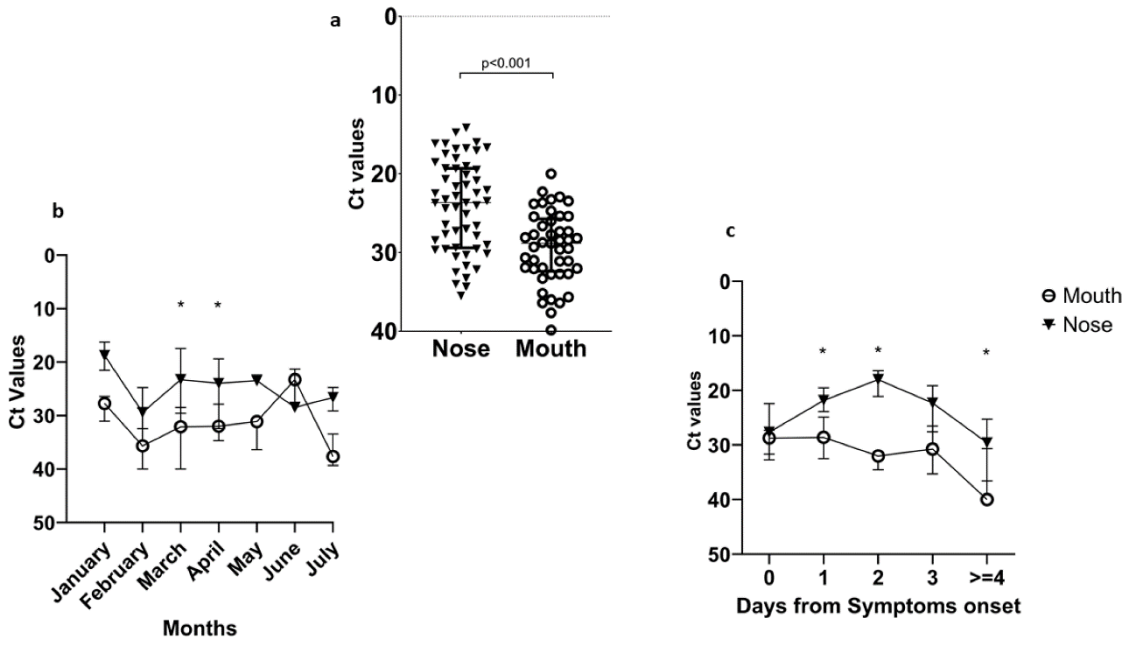


Figure 5.4: *Viral load during the Omicron wave in different upper respiratory tract sampling sites. (**a**) Shows the Ct values detected in the nose (left) and in the mouth (right). Viral load dynamics in mouth and nose. (**b**) Shows the median Ct values and the IQR detected in the mouth (triangle) and in the nose (circle) across the year. (**c**) Shows the median Ct values and the IQR detected in the mouth (circle) and in the nose (triangle) based on days after symptoms.*

Discussion

The genome of Omicron subvariants contains more than 50 mutations [207], many of which have been associated with an increased transmissibility, variable disease severity, and the potential to evade immune responses acquired after SARS-CoV-2 vaccination or infection with a previous variant. Few studies have attempted to investigate the impact of mutations in the N protein on the diagnostic performance of ADTs, with conflicting results [181], [189], [190], [191], [192], [193], [194]. Due to a possible change in the tropism, it has been suggested that the detection of the Omicron variant could be favoured in oral swabs compared to nasal swabs [195]. In the present study, we monitored the performance of the ADTs in a real-world scenario, studying a cohort of more than 5000 subjects across the Delta and Omicron waves. We also assessed the viral load of Omicron at different sites (specifically, the nose and mouth). Moreover, an *in silico* study at the amino acid level was performed to investigate the possible effect of mutations on conformational changes in the Omicron and Delta variants' N protein, which may affect its recognition by antigen tests.

Our results indicate that, as expected, for both the Delta and Omicron waves, the ADTs could only detect samples with a relatively higher viral load compared to the molecular test based on RT-PCR (**Figure 5.1a** and **Figure 5.1b**). All the used ADTs targeted the N protein, and the Ct values analysis was performed focusing on the N gene. Significant differences were observed in the Ct values detected for antigenic-positive samples (RT-PCR+/RADT+) between the Delta and Omicron periods, indicating even a higher viral loads at the RNA level in ADT-positive samples for the Omicron period compared to the Delta period (**Figure 5.1c**). No differences were found in the viral load of the antigenic-negative samples (RT-PCR+/RADT−).

Importantly, when evaluating the diagnostic performance of the ADTs between the two SARS-CoV-2 variant periods (**Table 5.3**), the ADTs showed a decrease of about 30% in sensitivity during the Omicron compared to the Delta period, accompanied by a slight but significant increase in specificity (**Table 5.3**). As the Ct values of ADT-negative samples were similar in the Delta and Omicron waves and ADT-positive samples presented even lower Ct in Omicron compared to Delta, we can conclude that the decrease in ADT sensitivity for Omicron was not due to a lower viral load, but was more likely due to a change in the N protein. A possible theory is that this reduced sensitivity was due to a reduced recognition of N antigen by the ADT; this hypothesis is supported by the analysis of the N protein structure *in silico*. In fact, to investigate the possible effect of mutations present in the N protein on the ADTs' performances [208], we evaluated the SARS-CoV-2 whole genome data obtained from positive swabs of 168 patients from both the Delta and Omicron waves. Focusing the mutation analysis on the N protein amino acid sequence, i.e., the target of ADTs, we observed mutations localized in the NTD, the LKR, and the CTD (**Figure 5.2**). Our results confirmed the literature data, showing that mutations in the SARS-CoV-2 N protein mainly accumulate within intrinsically disordered regions, probably due to the functional importance of the NTD and CTD [204], [206], [209]. In the Omicron LKR region, the co-occurring mutations R203K and G204R are the most common mutations, with a frequency of >60% across all sequences [194], [206], [207]. Within the NTD and CTD domains, the folding free energy and vibrational entropy analysis indicated that P80R and D343G, both exclusively present in the Omicron variant, were shown to putatively alter the dynamic properties of the protein (**Supplementary Table S5.1**), strongly suggesting that these mutations may affect the N-protein stability and dynamicity and reduce the performance of antigenic assays. Moreover, from the literature data, the P13L and E378Q mutations also present in the N and C arms of Omicron variants, respectively, were predicted to destabilize the N protein [208].

An additional hypothesis that could explain the observed variations in the ADT sensitivity is the different amounts of nucleocapsid protein that could be shed during infections with different virus

variants. Rao et al. showed that Omicron samples had lower ratios of antigen to RNA compared to Delta, which leads to a possible explanation for this result when using Ct values as a reference [191].

To establish whether a reduced sensitivity of ADTs could be due to a shift in viral tropism with a preferential location in the mouth compared to the nose, we evaluated the viral load in nasal nostrils and buccal swabs in Omicron-infected patients. The choice of NS and MS was performed in order to reduce the patients' discomfort. For this purpose, 61 patients from the Omicron wave underwent RT-PCR testing on swabs collected from both sites. We found that the nasal site had significantly lower Ct values and, therefore, a higher viral load than the oral site. Furthermore, when the viral load was examined according to the different periods in which SARS-CoV-2 Omicron subvariants were prevalent, the nose was always confirmed as the sample type in which the virus was more detectable, especially during the first 3 days after symptoms onset (**Figure 5.3**). Molecular characterization showed that the preferred virus localization in the nose vs mouth was independent from the specific Omicron subvariant.

Overall, our results indicate that the nose is the best sampling site to maximize virus detection for a diagnosis of Omicron infection. Nasal mid-turbinate swabs can be used for both ADTs and molecular assays to provide safe and reliable results as an alternative to nasopharyngeal swabs, in an effort to reduce patient discomfort [210], [211].

This study has several limitations. First, the lack of clinical characteristics of the patients included in the comparison of ADTs' sensitivity in Omicron vs. Delta infections. In fact, previous studies have shown a very low sensitivity of rapid antigen tests in asymptomatic patients, and only a moderate decrease in sensitivity for symptomatic Omicron infections [191], [192], [193], [194], [212]. Second, 6.8% of the analysed samples derived from multiple hospital accesses were from individuals who participated more than once in the study, so this could be a possible confounder. Third, the use of different ADTs throughout the study may have introduced some bias into the analysis, due to possible heterogeneous results from the different ADTs. However, the stratification analysis according to the type of assay and the high number of participants provided confidence in the reliability of the results and their interpretation, indicating a substantially reduced sensitivity of ADTs for Omicron infections.

Conclusions

In conclusion, real-life data from a large number of subjects strongly support the evidence of a substantially reduced detection rate of Omicron infections by ADTs, confirming and extending circumstantial evidence from previous studies.

This drop in sensitivity should be taken into consideration in establishing testing strategies and monitoring infection prevalence. The emergence of new variants, as well as new mutations affecting the N protein structure, might further affect ADTs' diagnostic performance, which could require assay revalidation to maintain efficient and reliable screening and diagnostic strategy programs. Our study suggests that ADTs should be adapted to better detect Omicron-descending variants.

PART II

Proteomic Analysis and Drug Repurposing Strategies for Strongyloides stercoralis: Novel Insights and Therapeutic Approaches

Novel insights into the somatic proteome of Strongyloides stercoralis infective third-stage larvae

This chapter describes my contribution to: Dishnica, K., Piubelli, C., Manfredi, M., Kondaveeti, R. T., Longoni, S. S., Degani, M., ... & Tiberti, N. (2023). Novel insights into the somatic proteome of Strongyloides stercoralis infective third-stage larvae. Parasites & Vectors, 16(1), 45. [213]

Introduction

Human strongyloidiasis caused by *Strongyloides stercoralis (S. stercoralis)* is a soil-transmitted helminthiasis that has recently been listed by the WHO among the tropical neglected diseases requiring control actions in endemic areas [214]. Strongyloidiasis is estimated to affect about 600 million people worldwide [16], mostly in tropical and subtropical regions. However, foci of autochthonous strongyloidiasis have also been reported in temperate areas, including Italy, Spain, Japan, Australia and USA [215]. *S. stercoralis* belongs to the phylum Nematoda, clade IV [216]. Its life-cycle is complex, alternating between cycles of free-living and parasitic stages. Humans acquire the infection through the penetration of the intact skin by infective filariform larvae (iL3) present in contaminated soil which, once in the host, migrate through different organs. During migration, the larvae moult until they become adult worms, which ultimately settle in the small intestine. Once there, the parthenogenetic females deposit eggs that hatch in rhabditiform larvae (L1), which are then excreted in stools and initiate the free-living cycle. However, some L1 undergo an auto-infective cycle, i.e.

mature into invasive filariform larvae, in the large intestine and penetrate the intestinal mucosa or the perianal skin to continue the parasitic life-cycle. This peculiar life-cycle allows *S. stercoralis* to perpetuate the infection, in the absence of treatment, potentially indefinitely [217].

In immunocompetent subjects, the infection mostly leads to a chronic indolent condition; however, changes in the host immune status can cause a dramatic increase in parasite burden, known as hyper-infection or dissemination, which can be life-threatening [217].

The diagnosis of strongyloidiasis is challenging, with most available methods presenting variable sensitivity [217], [218]. The most sensitive diagnostic tools are serological immunoassays [219], [220]]. Most commercial assays are based on crude larval antigens, which reduce their specificity and result in a high batch-to-batch variability. The development of assays based on recombinant antigens represents a very promising strategy to avoid the need of constant supply of parasites, to overcome the variability of the antigenic source and to reduce cross reactions with other helminths, factors that affect the performance of current serological tests [217]. Indeed, a novel commercial enzyme-linked immunosorbent assay (ELISA) based on the detection of two recombinant antigens, Ss-NIE and Ss-IR, has recently been developed and evaluated on cryopreserved samples. The test has shown variable accuracy in two different studies, probably due to the lack of a diagnostic gold standard, and has yet to be tested prospectively. Nonetheless, at present it is among the most sensitive and specific serological tests for strongyloidiasis, further highlighting the potential of recombinant antigens for serodiagnosis [221], [222].

To date, only a few proteomics studies have been conducted to elucidate the molecular mechanisms associated with *Strongyloides* parasitism or to highlight novel immunological markers [17], [223], [224], [225], [226], [227], [228]. Consequently, our knowledge of *S. stercoralis* proteome is still limited. In order to highlight novel targets to improve current serodiagnosis and treatment, it is fundamental to expand the molecular understanding derived from 'omics studies beyond the current state of the art. Indeed, an in depth characterization of the *S. stercoralis* infective larvae proteome might reveal on one hand novel players in the mechanisms of host–pathogen interaction and on the other hand potentially immunogenic proteins to be used for the development of novel diagnostic serological tests, as well as target proteins for new therapeutics.

The aim of this study was to expand the characterization of *S. stercoralis* iL3 proteome as established by high-throughput proteomics, combining automatic search strategies and manual annotation.

Graphical Abstract

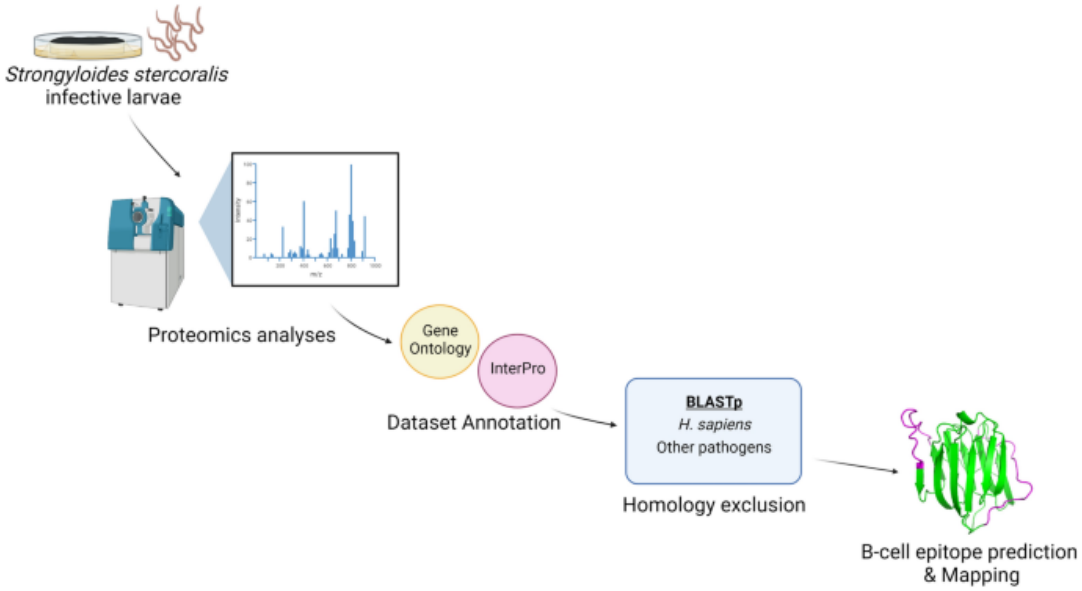


Figure 6.1: *This workflow illustrates the analysis of S. stercoralis infective larvae through proteomics, followed by dataset annotation using Gene Ontology and InterPro, homology exclusion via BLASTp against H. sapiens and other pathogens, and culminating in B-cell epitope prediction and mapping for identifying potential antigenic targets.*

Methods

Larvae isolation, protein extraction and digestion

S. stercoralis larvae were obtained from a human subject. Fresh stools mixed with charcoal and saline were cultivated using the agar plate culture method . iL3 larvae were harvested after 3 days of culture [229], concentrated by centrifugation and incubated with phosphate buffered saline (PBS) supplemented with 100 U/ml penicillin, 100 µg/ml streptomycin and 0.625 µg/ml amphotericin B (all from Gibco, Thermo Fisher Scientific, Waltham, MA, USA) for 2 h at 4 °C. Larvae were then washed twice with cold PBS, counted under the microscope and stored at − 80 °C for future use.

A pellet of 10,000 iL3 was re-suspended in 0.1% RapiGest SF (Waters Corporation, Milford, MA, USA) in 0.1 M triethylammonium bicarbonate buffer (TEAB) pH 8.0, sonicated with breaks on ice and incubated for 10 min at 80 °C, following a protocol reported in [230]. The sample was then

centrifuged at 14,000 *g* for 10 min at 4 °C and the supernatant recovered. Protein concentration was determined by the Qubit protein assay (Life Technologies, Thermo Fisher Scientific). A 75-µg sample of proteins was reduced with 50 mM Tris-(2-carboxyethyl)phosphine hydrochloride (TCEP), alkylated with 15 mM iodoacetamide and digested with 0.25 µg/µl sequencing grade-modified trypsin (Roche, Basel, Switzerland; 1:25 protease to protein ratio). The sample was incubated with 1% trifluoroacetic acid for 45 min at 37 °C to cleave the RapiGest SF surfactant, cleaned with C18 spin columns (Pierce™, Thermo Fisher Scientific) and dried under vacuum prior to liquid chromatography-tandem mass spectrometry (LC–MS/MS) analyses.

Protein identification by LC–MS/MS

Trypsin-digested protein samples were analysed with a micro-LC system (Eksigent Technologies, Dublin, CA, USA) coupled with the TripleTOF 5600+ system (Sciex, Concord, ON, Canada) equipped with a DuoSpray ion source (Sciex). The stationary phase was a Halo C18 column (0.5 × 100 mm, 2.7 µm; Eksigent Technologies). The mobile phase was a mixture of 0.1% (v/v) formic acid in water (phase A) and 0.1% (v/v) formic acid in acetonitrile (phase B), eluting at a flow rate of 15.0 µl/min at an increasing concentration of solvent B from 2% to 40% in 30 min. Samples were also analysed with nano liquid chromatography using an Acclaim PepMap C18 column 2 µm, 75 µm × 150 mm (Thermo Fisher Scientific) and injection volume of 2 µl. The flow rate was 300 nl/min, phase A was 0.1% formic acid/water and phase B was 80% acetonitrile/0.1% formic acid/20% water. A 2-h gradient was used (3–45%). Identification was performed using a data-dependent acquisition (DDA) method: the MS analysis was carried out using a mass range of 100–1500 Da (time-of-flight scan with an accumulation time of 0.25 s), followed by a MS/MS product ion scan from 200 to 1250 Da (accumulation time of 5.0 ms) with the abundance threshold set at 30 cps (35 candidate ions can be monitored during each cycle) [231]. The MS data were acquired with Analyst TF 1.7 (Sciex). The DDA files were searched using Protein Pilot software v. 4.2 (Sciex) and Mascot v. 2.4 (Matrix Science Inc., Boston, MA, USA) using trypsin as the enzyme, with two missed cleavages, a search tolerance of 50 ppm for the peptide mass tolerance and 0.1 Da for the MS/MS tolerance [232]. Searches were performed using the UniProt Swiss-Prot database for *S. stercoralis* (version 01/02/2020, taxon: 6248, proteome ID: UP000035681, protein count: 12,978), with a false discovery rate (FDR) fixed at 1%. The MS proteomics data have been deposited in the ProteomeXchange Consortium via the PRIDE [233] partner repository with the dataset identifier PXD037243.

Bioinformatic Analyses

The bioinformatics analyses were carried out by automatizing a standard protocol for the annotation. The protocol included: (i) a BLASTp [234] search using default parameters against Uniprot 2022 [235]; (ii) each protein identifier was then connected with its associated GO terms [236] via UniProt and GO terms were then organized using QuickGO tool [237]; and (iii) proteins were then classified into families or domains, and important sites were predicted, retrieving this information by InterPro [238]. In order to identify possible candidates for immunogenic epitope prediction, the annotated dataset of proteins was then investigated using BLASTp for homology with the human protein database and with a list encompassing 29 clinically relevant pathogens (24 helminths and 5 *Plasmodium* spp.) that might co-infect individuals with strongyloidiasis (**Figure 6.1**). The threshold for considering an *S. stercoralis* protein as having low homology with proteins of human or with those of other pathogens' origin was empirically established based on the BLASTp e-value obtained for the L3NieAg.01 (AC: Q9UA16), which is known to have a good specificity when used in serodiagnosis [221]. Thus, a BLASTp e-value threshold of 4E-25 and 2E-30 was applied for the comparison with *H. sapiens* or with other pathogens, respectively.

Linear B-cell epitopes were predicted from protein sequences using the different web-based tools available via the Immune Epitope Database Analysis Resource (IEDB; available at http://tools.iedb.org/main/). The following physicochemical properties of individual residues were explored and scored: beta-turn, surface accessibility, antigenicity and hydrophilicity, as already reported in the literature [239]. All residues having an individual score equal or higher than the average protein score were highlighted. In parallel, prediction was also performed using BepiPred-2.0, which combines a hidden Markov model (HMM) with an amino acid propensity scale [240]. Proteins of potential interest for bearing B-cell epitopes were then manually analysed and selected based on the following parameters: (i) sequences of at least 8 amino acids; (ii) a BepiPred-2.0 score > 0.5 (range 0–1); and (iii) at least three physicochemical properties above their thresholds (calculated as the mean of the scores of all individual residues). The specific sequences of interest were highlighted and visualized in the proteins three-dimensional model using Pymol v2.4.1 [82]. Due to the lack of structural characterization of *S. stercoralis* proteins on Protein Data Bank (PDB) [241], selected proteins of interest containing predicted epitopes were structurally predicted using AlphaFold [199].

Results and discussion

In the present study, we analysed the proteome of *S. stercoralis* infective larvae by LC-MS/MS and performed a semi-automated annotation of the dataset to achieve a more in depth characterization of larval proteome and to predict potential immunogenic proteins of interest for the development of new sero-diagnostic tools. The study flowchart is reported in **Figure 6.1**. Our high-throughput MS analysis identified 430 proteins (2 unique peptides, 1% FDR), which to the best of our knowledge is the largest experimental proteome of *S. stercoralis* iL3 reported to date (**Supplementary Table S6.1**). Indeed, only one study had previously employed untargeted proteomics to investigate the *S. stercoralis* iL3 proteome; however, due to the lack of a reference genome at that time, only 26 proteins were identified [223].

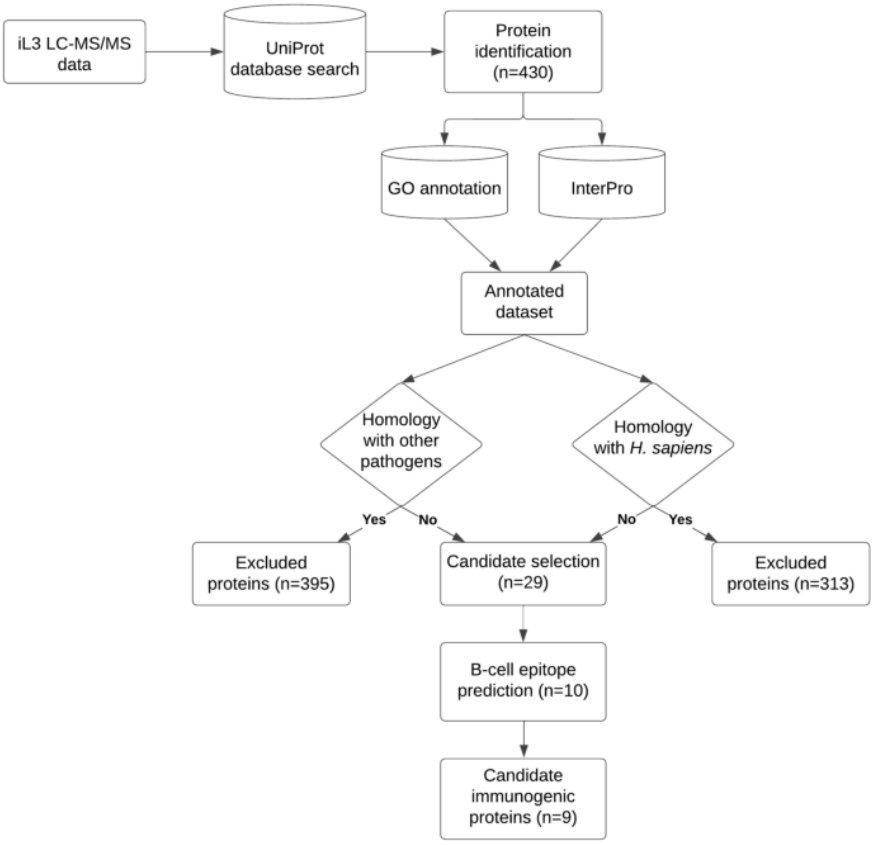


Figure 6.1: *Study flowchart. Pipeline followed in the present study. GO, Gene ontology; iL3, infective filariform larvae; LC-MS/MS, Liquid chromatography-tandem mass spectrometry. Other pathogens include: Ancylostoma duodenale; Ancylostoma ceylanicum; Necator americanus; Ascaris lumbricoides; Trichuris trichiura; Toxocara canis; Loa loa; Mansonella perstans; Mansonella ozzardi; Wuchereria bancrofti; Onchocerca volvulus; Brugia malayi; Brugia timori; Dirofilaria immitis; Dirofilaria repens; Trichinella spiralis; Taenia saginata; Taenia solium; Echinococcus granulosus; Hymenolepis nana; Schistosoma mansoni; Schistosoma haematobium; Schistosoma japonicum; Fasciola hepatica; Plasmodium falciparum; Plasmodium vivax; Plasmodium ovale; Plasmodium malariae; Plasmodium knowlesi.*

A high-quality draft genome of *S. stercoralis* was assembled in 2016 (42.6 Mb) and predicted to contain 13,098 protein-coding genes [17], facilitating the annotation of 'omics data, although a reference genome has yet to be assembled. Hunt et al. performed an in depth investigation of the genomic bases of parasitism in the *Strongyloides* clade by comparing distinct life stages of different *Strongyloides* species, the closely related *Parastrogyloides trichosuri* and the free-living *Rhabditophanes* at the genome, transcriptome and proteome level [17]. This comparison has allowed researchers to propose protein categories with a putative role in parasitism that are expanded in the *S. stercoralis* genome or abundantly transcribed in iL3. These include proteinases (astacins—metallopeptidases, aspartic proteases, prolyl oligopeptidase), protease inhibitors, SCP/TAPS proteins, transthyretin-like proteins and acetylcholinesterases [17], [242]. Interestingly, the same protein families were also identified in *Strongyloides venezuelensis* iL3 [226]. Similarly, next generation RNA sequencing was employed to evaluate the association between larval development in an *S. stercoralis* laboratory strain (i.e. PV001) and the expression of specific genes homologous of *Caenorhabditis elegans*, in which they were reported to be involved in dauer arrest or activation [243].

In our dataset, 43% of the identified protein sequences (i.e. 187 protein matches) corresponded to uncharacterized proteins according to UniProt database 2022 for *S. stercoralis*. In order to achieve a better characterization of the dataset we performed a semi-automated annotation through GO and InterPro functional analyses (**Supplementary Table S6.1**). The cellular component (CC) GO analysis highlighted a prevalence of membrane and mitochondrial proteins, which together accounted for > 40% of the annotated terms (**Figure 6.2a**). Almost half (47%) of the molecular function (MF) GO terms had binding activities, with nucleic acid and nucleotide binding being the most represented sub-categories, while 40% were associated with enzymatic activities (**Figure 6.2b**). Intriguingly, within this latter group, the most represented term corresponded to oxidoreductase activity, accounting for 33% of all GO terms associated with catalytic activities. It could be speculated that infective larvae might need to counteract the oxidative stress either derived from their particularly active cellular metabolism or as a defence mechanism against the host immune response [244], [245]. Interestingly, we identified three of the four major antioxidant enzyme families involved in the response against reactive oxygen species (ROS), namely glutathione peroxidase, superoxide dismutase and peroxiredoxin/thioredoxin. Antioxidant enzymes are known to be important in nematodes, and an evolutionary analysis has been recently published [244], [245]. Antioxidative enzymes were also identified in the excretory-secretory products (ESPs) from the different life stages of *Strongyloides ratti* [225] and *S. venezuelensis* iL3 [228]; it would be interesting to evaluate whether the expression of these proteins is modulated during parasite development. The biological process (BP) GO analysis also highlighted that *S. stercoralis* iL3 larvae express a high number of proteins involved in metabolic

(42%) and cellular processes (44%) (**Figure 6.2c**). Notably, the most represented cellular processes were translation and cellular respiration, which is in agreement with the high number of nucleic acid/nucleotide binding proteins and mitochondrial proteins and further supports the observation of a highly active metabolic state of infective larvae. The InterPro analysis allowed a further classification of the identified proteins, including those uncharacterized, either into families or on the basis of the presence of specific domains within their amino acid sequence. The most frequent InterPro domain and family entries are reported in **Figure 6.3**, while the entire annotation is available as **Supplementary Table S6.1**. The CAP domain, SCP domain and thioredoxin domain were the most commonly represented protein domains, while several proteins were annotated as belonging to cysteine-rich secretory, transthyretin-like or peptidase protein families, making them the most represented protein families in the iL3 proteome. Overall, our functional analysis provides experimental evidence that confirms previous data on the most represented proteins associated with *S. stercoralis* parasitism, as inferred from genomic and transcriptomic data [17], [242], [246], [247], as well as with proteomics analyses of *S. ratti* and *S. venezuelensis* iL3 ESP [225], [228]. In particular, in our dataset we identified a high proportion of proteins with peptidase activity; such proteins have already been highlighted as potentially involved in parasitism as they are upregulated in the adult parasitic female stage of *S. ratti* and *S. stercoralis* [17]. Indeed, these proteins, including metalloproteases and metallopeptidases (such as astacin-like proteins), are involved in tissue degradation. This is a fundamental process in the initial phases of the infection for the penetration of host tissues and in parasite migration through the host body—even though peptidases could also contribute to immune evasion [248]. Other protein categories known to be associated with *S. stercoralis* parasitism and identified in our study include: (i) galectins, involved in pathogen adhesion to the host cells and activation of host innate and adaptive immunity [249]; (ii) transthyretin-like proteins; and (iii) SCP/TAPS-/CAP-domain containing proteins, with putative immunomodulatory properties in parasitic nematodes [250]. The expansion of SCP/TAPS coding genes in *Strongyloides* and *Parasitrongyloides* compared to *Rabditophanes* suggests that their gene products might be associated with human parasitism [17]. In our study, we identified 11 proteins either as SCP domain-containing proteins ($n = 7$) or as uncharacterized proteins containing the SCP-domain according to the InterPro analysis (IPR034113). Most of the protein categories that had already been proposed as associated with iL3 parasitism were thus experimentally confirmed in our proteomics study of *S. stercoralis* iL3 with 43 protein matches (**Supplementary Table S6.2**). It is worth noting that almost 50% of these proteins were uncharacterized and were assigned to those categories only following the GO and InterPro semi-automated annotation. The importance of focussing 'omics studies not only on known and characterized genes and proteins, but especially on those "novel" or

uncharacterized ones was already highlighted more than 10 years ago. Such an approach can achieve a more in depth knowledge of the molecular mechanisms associated with pathology but also with pathogen development [251], especially for organisms whose genome and proteome are not fully annotated, such as *S. stercoralis*.

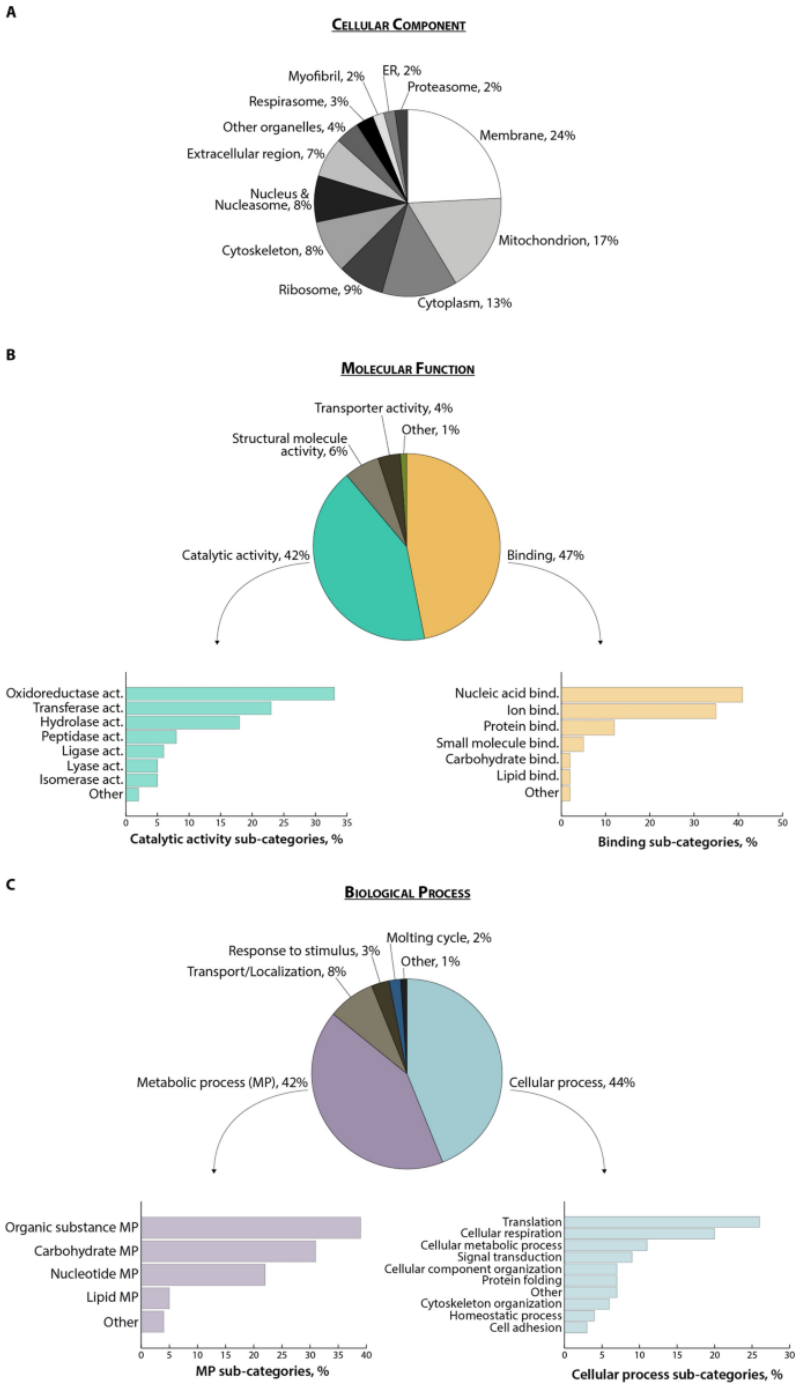


Figure 6.2: *Gene ontology results. Frequency of the GO terms for the three categories cellular component (**A**), molecular function (**B**) and biological process (**C**) across identified proteins. ER, Endoplasmic reticulum*

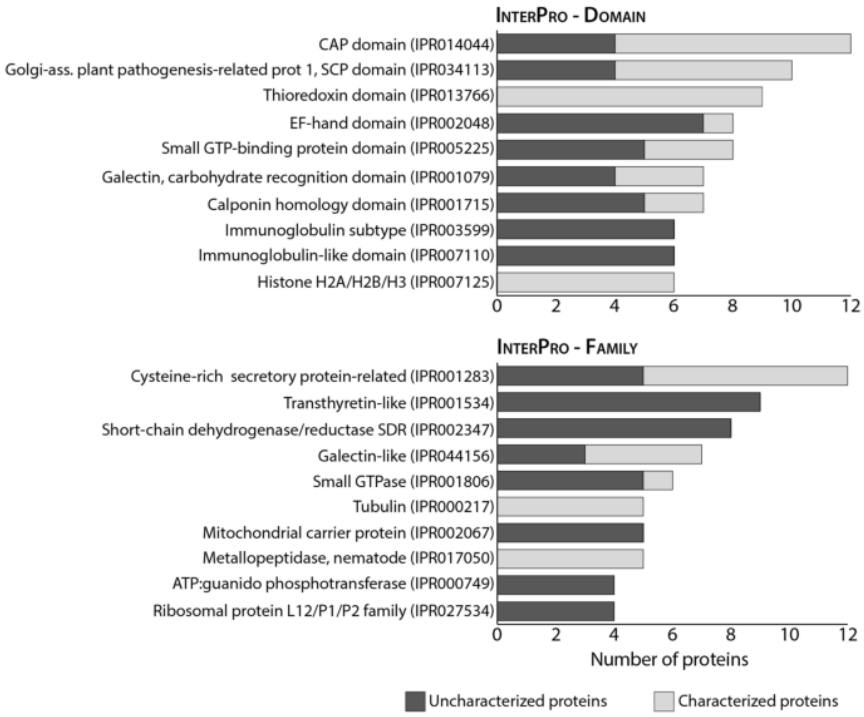


Figure 6.3: *InterPro annotation results. The top 10 most frequent InterPro terms for the categories domain and family represented among all identified proteins (n = 430), as established by InterPro annotation. For each term, the number of uncharacterized and characterized proteins is represented in different color shades. The complete annotation is reported in* **Supplementary Table S6.1**

The objective of our study was not only to improve our knowledge of the iL3 proteome with novel experimental evidence, but also to predict—among those identified from a clinical isolate—potential immunogenic proteins that could be useful for the development of novel serological tests for the accurate diagnosis of human strongyloidiasis or as vaccine candidates, as potentially recognized by antibodies present in patients' serum.

A few studies dating back to the 1990s reported the investigation of the humoral immune response associated with the intensity of infection and the detection of immunoreactive iL3 polypeptides [252], [253], [254]. However, only a couple of studies applied immuno-proteomics MS/MS-based approaches to also identify the immunogenic proteins recognized by antibodies from infected subjects [224], [255]. Indeed, Rodpai and colleagues confirmed by immunoblotting the high frequency of some protein bands that had previously been reported as immunoreactive [252], [253], [254], [256], but also identified them by tandem MS based on protein homology with *S. ratti* [255]. In particular, they identified a 26-kDa band corresponding to 14–3-3 protein and a 29-kDa band

corresponding to ADP/ATP translocase 4. Additional antigenic proteins were further identified by the same group after they had improved sample separation through two-dimensional gel electrophoresis prior to immunoblotting [224]; the majority of these proteins were also identified in the present study.

In silico approaches can be used as an alternative, or a complement, to the experimental identification of immuno-reactive proteins. The advent of immunoinformatics has actually led to the development of a number of tools that can assist researchers in B-cell epitope prediction. Moreover, it has been shown that using multiple prediction methods results in a more accurate epitope prediction than using individual tools [239], [257]. In agreement with this, in the present study we combined the use of a machine learning-based algorithm (i.e. BepiPred-2.0 [240]) and the evaluation of several physicochemical residue properties to predict linear B-cell epitopes. In order to avoid the selection of proteins highly conserved across helminths or similar to human ones, we first excluded all those having high homology, as established by our BLASTp analysis (**Supplementary Table S6.3**). Among the 29 proteins showing limited homology with *Homo sapiens* or other pathogens of clinical relevance, we selected 10 for use in the prediction of the presence of B-cell epitopes (**Table 6.1**). This selection was based on the following criteria: (i) proteins already highlighted as potentially associated with *S. stercoralis* parasitism [17], [242] or as immunogenic [224]; (ii) extracellular or plasma membrane proteins as per CC GO terms; (iii) proteins with peptidase activity according to the MF GO terms; and (iv) proteins associated with relevant InterPro domain, family or homologous superfamily (namely transthyretin-like domain, CAP domain, galectin, cysteine-rich, peptidase, protease inhibitors). According to UniProt, 60% of the selected proteins are already characterized, while the remaining 40% are still uncharacterized (**Table 6.1**). The six characterized proteins included three SCP domain-containing proteins (ACs: A0A0K0E6J0, A0A0K0EG68, A0A0K0DTP5), galectin (AC: A0A0K0ECK4), NTR domain-containing protein (AC: A0A0K0EMX1) and L3NieAg.01 (AC: Q9UA16 also known as Ss-NIE). It is worth noting that galectins have already been reported to be involved in host–pathogen interaction and to display immuno-regulatory properties in *S. ratti* [258]. Also, several commercial and in-house assays already use the recombinant Ss-NIE for *S. stercoralis* serodiagnosis [18], [221], [259], [260], [261]. Ss-NIE is also included in a commercial research use only (RUO) serological test, together with Ss-IR, which was not identified in our dataset [221]. The presence of Ss-NIE within our selection further supports the validity of our approach. In our study we also identified most of the proteins already highlighted as potentially immunogenic by Rodpai and colleagues [224], [255]. However, since these proteins displayed high homology with

either human or other related pathogen proteins, their analysis for epitope prediction was not pursued (**Supplementary Table S6.2** and **Supplementary Table S6.3**).

Table 6.1: *List of potentially immunogenic proteins and the predicted B-cell epitopes*

Protein AC	Protein name	e-value[a] vs Homo sapiens	e-value[a] vs other pathogens[b]	Protein properties relevant for selection	Epitope sequence
A0A0K0E6J0	SCP domain-containing	7.00E−24	Minimum 2E−25	Extracellular region [GO:0005576]	[105]VTQPPRPTARPFSRNPE
			Maximum 5.5	Integral component of membrane [GO:0016021]	[128]KPAPRPTIPPKTAKPG
				IPR014044: CAP domain	[147]APPNNRIDPMYIPNPDE
A0A0K0DY51	Uncharacterized protein	3.00E−06	Minimum 3E−11	IPR035940: CAP superfamily [17]	[16]ESKNEEVHPT
			Maximum 7.7	IPR018244: Allergen V5/Tpx-1-related, conserved site [17]	[40]AVEPPAETPAE
					[72]PVETTTETP
					[140]PVETPAETSVDAPTENPTEVSADVPSTE
					[186]SVPEQSVEKIEEPSVTEVQCP
A0A0K0ECK4	Galectin	1.00E−13	Minimum 1E−22	IPR001079: Galectin, carbohydrate recognition domain	[10]TYDHIPKESYSIQPR
			Maximum 7.5		[91]KWQHEERTPKGNPFK
A0A0K0EG68	SCP domain-containing protein	7.00E−10	Minimum 2E−11	IPR001283: Cysteine-rich secretory protein-related	[66]RPTNRPINKKPIKKPNNKPK
			Maximum 1.8	IPR014044: CAP domain [17]	[105]PKPPGPRPKPPG
					[123]GPRPKPPG
					[137]GPKPKPTTTKPKPKPTTTKPKPKPTTTKP KPTQPPT

A0A0K0EM	NTR domain-	3.00E−05	Minimum 7E−15	Extracellular region [GO:0005576]	[125]MSPEKSPRYIYPPE
			Maximum 8.0	IPR001820: Protease inhibitor I35 (TIMP) [17]	[145]EVKNNLRTN
Q9UA16	L3NieAg.01	4.00E−25	Minimum 2E−30	Extracellular region [GO:0005576]	[75]YNYDNDKA
			Maximum 4.0	IPR014044: CAP domain [17]	[131]LEHDPKNRIE
A0A0K0E2F4	Uncharacterized	1.00E + 00	Minimum 1.1E−02	Integral component of membrane [GO:0016021]	[40]IDNQPAYV
			Maximum 4.1	IPR007863: Peptidase M16, C-terminal [17]	[75]HKIPHEPKASAREGVDGDEEDGASDTF
A0A0K0DTP	SCP domain-	1.00E−13	Minimum 4E−16	IPR014044: CAP domain [17]	[108]KQHNYDRDT
			Maximum 5.9		
A0A0K0ELA9	Uncharacterized	2.90E−01	Minimum 4E−13	Integral component of membrane [GO:0016021]	[44]FGKKDFSTKDLEPKNLKD
			Maximum 8.6	IPR001534: Transthyretin-like [17]	
A0A0K0E132	Uncharacterized	8.40E−01	Minimum 2E−19	Integral component of membrane [GO:0016021]	No B-cell epitope found
			Maximum 7.3	IPR001534: Transthyretin-like [17]	

The B-cell epitope prediction highlighted that nine out of the 10 selected proteins contained epitopes with high consensus across the different tools employed (**Table 6.1; Figure 6.4; Supplementary Figures S6.1–S6.8**); these were therefore considered as potentially immunogenic. The remaining protein (AC A0A0K0E132, uncharacterized protein) did not display potentially immunogenic epitopes as per our analysis. The structural models, together with a confidence estimation as per AlphaFold, are reported in **Figure 6.4** and **Supplementary Figures S6.1–S6.8**. In agreement with the results obtained from different web-based prediction tools, all epitopes were exposed to the external environment, thus potentially accessible for antibody binding. However, some epitopes fell within regions of the structure which was modelled with low confidence. This could be explained by

the fact that immunogenic epitopes often fall within highly variable regions, and there is a lower confidence in the structure as predicted by AlphaFold.

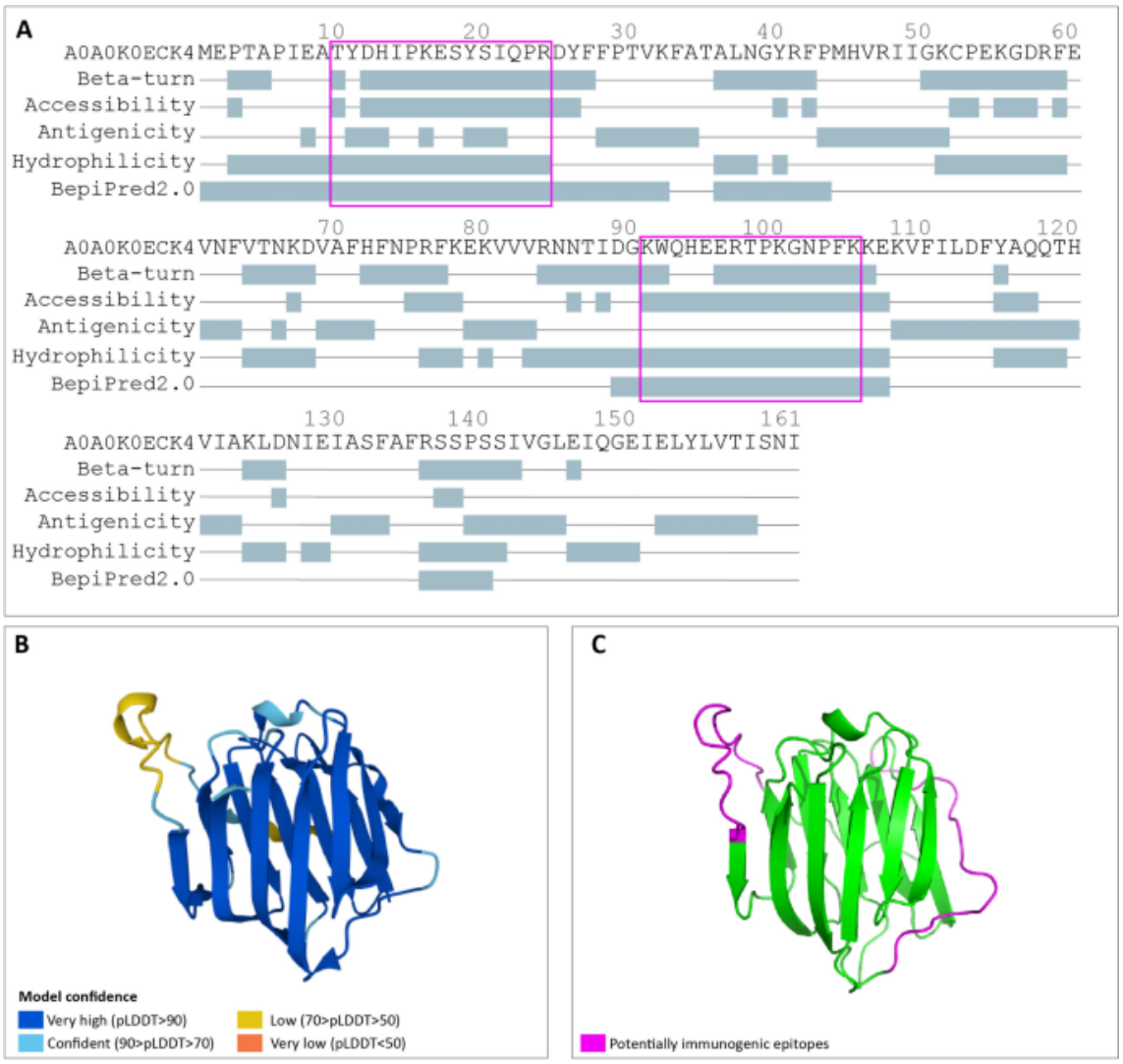


Figure 6.4: *B-cell epitope prediction results. The results for the protein A0A0K0ECK4—galectin are reported as an example.* ***a)*** *FASTA sequence showing the results obtained with each tool (Chou & Fasman Beta-Turn Prediction; Emini Surface Accessibility Prediction; Kolaskar & Tongaonkar Antigenicity; Parker Hydrophilicity Prediction, BepiPred2.0; all available via http://tools.iedb.org/bcell/). All residues having a score above their threshold are highlighted in grey. The purple squares indicate the sequences highlighted as being potentially immunogenic as reported in the Methods section.* ***b)*** *Protein structures as predicted by AlphaFold showing the model confidence.* ***c)*** *Mapping of the potentially immunogenic epitopes on the protein structure. The same images for all other selected proteins are reported in* **Supplementary Figures S6.1–S6.8**

A recent work employed a reverse in silico approach to predict immunogenic proteins from the *S. stercoralis* proteome available in UniProt [262]. However, none of the proteins proposed as potentially

immunogenic was identified in our dataset, probably because the analysis was performed on the entire *S. stercoralis* proteome, without taking into account the parasite developmental stage.

In the present study we did not perform a comparison of protein expression between larval developmental stages, as has been done at the transcriptomic level or for other *Strongyloides* species [17], [225], [242], [243], [247], thus we cannot speculate on the role of iL3 proteins in larval development. However, a comparison using quantitative proteomics of different larval stages might contribute to corroborate these transcriptomics data and might identify novel proteins potentially involved in parasitism and/or in parasite development that could be of interest for the development of novel disease control strategies. Similarly, investigations should be extended to the study of ESPs released from *S. stercoralis* iL3, as has already done for *S. ratti* [225] and *S. venezuelensis* [228], as these could highlight additional candidates for serodiagnosis. Proteomics data on *S. stercoralis* are still limited, and the reference database and proteome are in continuous evolution. Therefore, some proteins ID here reported might change in the future.

Conclusions

In conclusion, we provide the largest experimental dataset of the *S. stercoralis* iL3 proteome. By presenting, for the first time, an extensive proteomics dataset from the analysis of iL3 isolated from a clinical sample, our study brings knowledge on the *S. stercoralis* proteome to a level comparable to our knowledge on its close relatives *S. ratti* and *S. venezuelensis* [263]. These data may be useful for future studies as they represent a step towards filling the current gap in experimental proteomics data. Indeed, a broader expertise about protein expression in *S. stercoralis* larvae, as well as their modulation during different developmental stages, will be essential for identifying novel therapeutic and vaccine targets.

Our semi-automated annotation allowed us to confirm the presence—at the proteome level—of protein categories potentially involved in parasitism that to date were only inferred from genomics and transcriptomics data. Moreover, additional protein groups deserving further investigation, such as oxidoreductases, were also highlighted. Finally, we also propose a number of immunogenic protein candidates that, if experimentally confirmed, might be considered in the future for the development of novel serological diagnostic tests that could make the diagnosis of this neglected tropical disease more reliable and accurate.

Availability of data and materials

The mass spectrometry proteomics data have been deposited to the ProteomeXchange Consortium via the PRIDE partner repository with the dataset identifier PXD037243.

Targeting GluCl Receptor: Drug Repurposing Strategies for Strongyloides stercoralis Infection

Introduction

Strongyloides stercoralis (S. stercoralis), also known as threadworm, is a soil-transmitted human parasite that belongs to the nematode group called roundworms[]. It is found nearly worldwide, excluding only the extreme north and south and poses significant health risks, particularly in immunocompromised individuals. However, the true global burden of this infection is often underestimated due to a lack of precise data from endemic regions. Consequently, *S. stercoralis* remains one of the most overlooked parasitic infections among the "neglected tropical diseases" (NTDs)[]. Soil-transmitted helminth (STH) infections are among the most common infections worldwide and affect the poorest and most deprived communities. They are transmitted by eggs present in human faeces which in turn contaminate soil in areas where sanitation is poor[]. The infection with *S. stercoralis* is challenging to diagnose due to its often- asymptomatic nature and can persist for years, leading to severe complications if untreated. To treat STHs, including Strongyloides, a variety of anthelmintic medications are available. Ivermectin is the current treatment recommended by the US Centers for Disease Control and Prevention for Strongyloides infection, with albendazole serving as a backup [264]. Strongyloidiasis infections frequently result in secondary infections with other STHs, hence in these situations a broad-spectrum anthelmintic is preferable; this also holds true in veterinary

contexts. Given the relative rarity of disseminated infections, it is difficult to see how controlled trials of various therapies can be established. Human Strongyloides infections can spread, and in these cases, therapy other than anthelmintic treatment is required. However, what is ideal and optimum is not well established [265].The question of when and where drug resistance will develop is the same as it is with any anti-parasitic medication treatment. Ivermectin targets the GluCl receptor in parasitic nematodes. It binds to GluCls, causing an influx of chloride ions into the cells, leading to hyperpolarization and subsequent paralysis of the parasite. This paralysis prevents the parasite from moving and feeding, ultimately leading to its death [266], [267], [268], [269].

Developing novel treatments for *S. stercoralis* is a demanding, time-consuming, and costly process with low success rates. To address these challenges, computational approaches such as drug repositioning are increasingly utilized. Drug repositioning involves identifying new therapeutic uses for both FDA-approved drugs and those under preclinical investigation, thereby reducing the cost and time required for drug development due to their known safety profiles and therapeutic potentials in other diseases. In this study, we propose a computational approach to screen both FDA-approved drugs and drugs under preclinical investigation for other disease but maybe potentially effective also against *S. stercoralis*. Our target protein, the glutamate-gated chloride channel (GluCl), was modeled by homology modeling. Drugs with promising binding affinities underwent further studies to validate their potential as effective treatments. This methodology not only aims to identify new treatment options for *S. stercoralis* but also exemplifies the broader utility of drug repositioning in addressing parasitic diseases.

Methodology

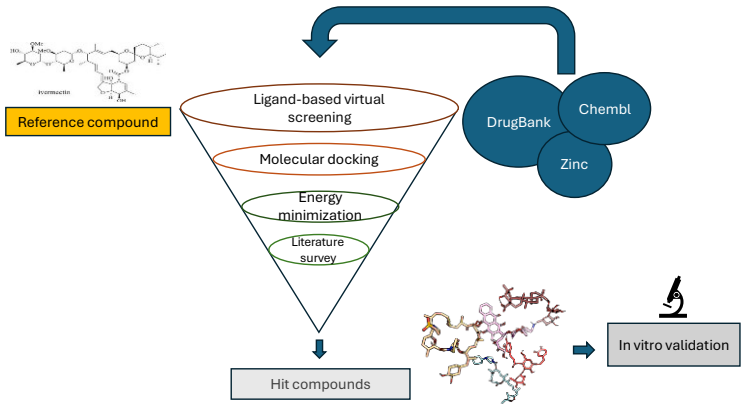


Figure 7.1: *The research flow chart illustrates the structured hierarchy and sequence of steps in our newly conceptualized project. It outlines the pipeline from initial stages to final outcomes.*

Structural Modeling of the Glutamate Chloride Channel in *Strongyloides*

Modeling the glutamate chloride channel is a crucial step in our drug repurposing strategy, as it allows us to explore how existing drugs might interact with this receptor in *Strongyloides*. By understanding the structural details of this channel, we can better evaluate potential therapeutic agents and optimize drug repurposing efforts to address parasitic infections.

We employed the Swiss-Model server [76], [270], a tool for generating high-quality structural models based on sequence similarity. For our modeling efforts, we used the crystal structure of the glutamate chloride channel from Caenorhabditis elegans, represented by PDB id 3RIA, as the template. This template was selected due to its high resolution and relevance, as the glutamate chloride channel in C. elegans shares approximately 51% sequence identity with the corresponding protein in Strongyloides.

Ligand Screening

Ivermectin is commonly known as an inhibitor of glutamate-gated chloride channels and is used as a standard drug in the treatment of strongyloidiasis [264]. The SMILES format and chemical structure of ivermectin were retrieved from the FDA (https://www.fda.gov/). SwissSimilarity, an online platform that allows the identification of chemical hits from FDA and other libraries based on a reference structure [271], [272], was utilized in this study. Ivermectin served as the standard template to screen not only FDA-approved drugs but also compounds from other chemical libraries, like Chembl active compounds and Zinc, for potential treatment of strongyloidiasis. All screened drugs were ranked according to their predicted score values.

Molecular docking

Before conducting our docking experiments, all the screened drugs were sketched using the RDKit library (https://www.rdkit.org/). A crucial preliminary step involved selecting the most promising drugs to dock against the GluCl receptor, given the extensive number of molecules generated by the screening process. Our selection was not solely based on the similarity scores calculated by Swiss Similarity; we also computed Tanimoto similarity [273] based on Morgan [152] and RDKit fingerprints to ensure a thorough evaluation. Our assumption is that the Similarity Score calculated by Swiss Similarity provides a low score to molecules that however could have favorable interactions with the target (protein of interest), at least from a structural point of view. Additionally, Smina [274] was employed for energy minimization of each ligand using default settings. The docking experiments were then performed on all selected compounds against the GluCl receptor. The binding pocket of the target protein (GluCl) was identified utilizing Plip [147]. For the docking experiments, the grid box dimensions were set to X = 10.31, Y = 91.36, and Z = 20.81, with the default exhaustiveness

value of 8. The grid box size was adequately adjusted around the binding pocket residues to allow the ligands sufficient freedom of movement within the search space. Each screened drug was docked separately against the target protein. The resulting complexes were analyzed to observe their binding conformational poses against the target protein, aiming to identify the best docking results. The generated docked complexes were evaluated based on the lowest binding energy (kcal/mol) values and the binding interaction patterns between the ligands and the receptor. The graphical depictions of all the docked complexes were created using PyMOL [82].

Results and discussion

Structural characterization of GluCl-IVM complex

We have characterized the interactions between ivermectin (IVM) and the GluCl receptor, providing detailed insights into the binding mechanisms. The left panel of **Figure 7.2** presents the overall structure of the GluCl receptor, with IVM prominently bound within its active site. An inset offers a closer view of the binding pocket, highlighting specific residues involved in the interaction. The right panel features a detailed table of hydrophobic interactions and hydrogen bonds between IVM and the GluCl receptor. Key residues such as LEU, ILE, and PHE are noted for their participation in hydrophobic interactions, while hydrogen bonds involve residues like ASN and ASP. This thorough visualization underscores the critical interactions that contribute to the ligand's binding affinity and stability within the receptor's active site, providing valuable insights for future drug design and development efforts.

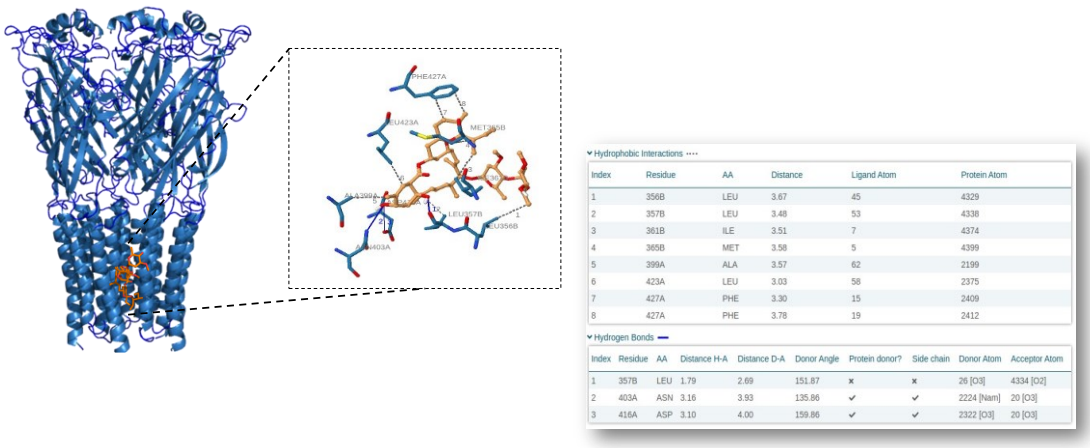


Figure 7.2: *Binding Interactions of Ligands within the GluCl Receptor.*

Superimposition of screened drugs within active region of GluCl

With the aim of verifying the drug binding configurations within GluCl's active binding site, all docked structures were overlaid. According to the binding pocket analysis, all of the molecules that were screened were narrowed down in the binding pocket of GluCl and bound with residues that were comparable but had different conformational poses. The docking reliability and expected outcomes were further validated by the binding of all molecules at the same position (**Figure 7.3**).

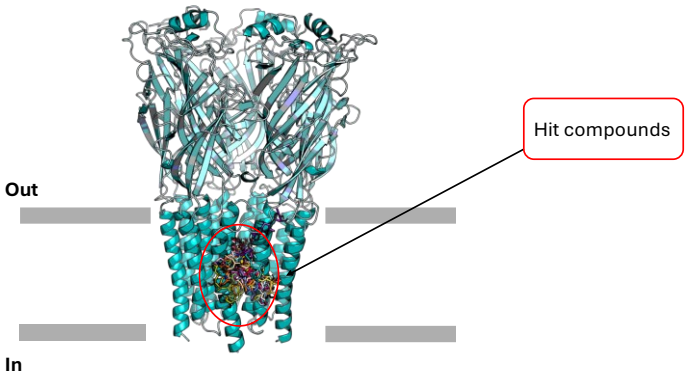


Figure 7.3: *Superimposition of all screened drugs.*

Selection Criteria Based on Affinity Scores

After docking the molecules, the criterion we utilized to select a subset of molecules for further in vitro investigation was the affinity score of the ivermectin (IVM)-GluCl complex, which is the drug currently in use. We selected only those molecules that exhibited an affinity score higher than that of IVM. The bar chart (**Figure 7.4**) presents the minimized affinity values for the screened compounds against the target protein. The x-axis represents the minimized affinity in kcal/mol, with more negative values indicating stronger binding affinities. The y-axis lists the identifiers of the compounds, including both CHEMBL IDs and other identifiers. From the chart, it is evident that all compounds exhibit varying degrees of binding affinities, with several compounds showing stronger interactions than IVM (e.g., CHEMBL4279943 and CHEMBL3289640). These higher affinity scores suggest a potential for greater efficacy in binding to the target protein, thereby guiding the selection of promising candidates for further experimental validation. The consistent measurement of minimized affinities provides a comparative basis for evaluating the potential of each compound in the docking experiments.

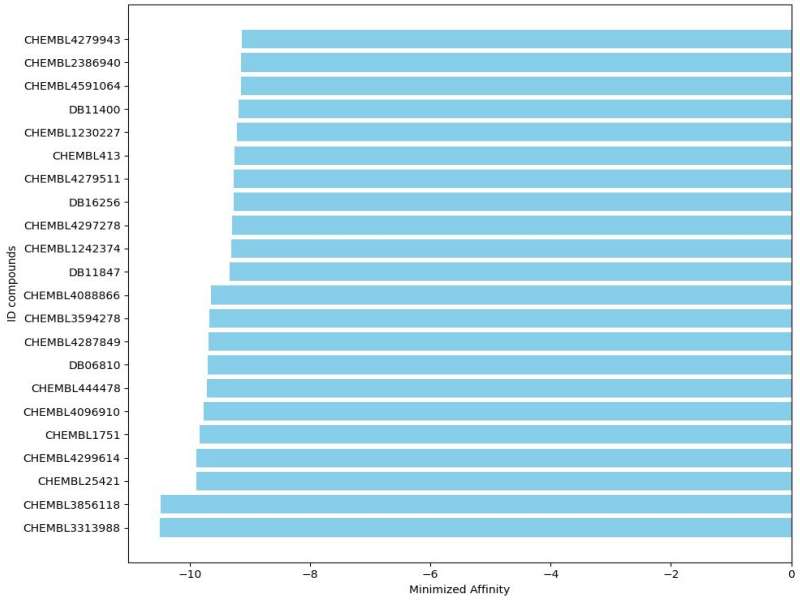


Figure 7.4: *Affinity score of all screened drugs against GluCl receptor.*

Conclusions

Through a combination of molecular docking, virtual screening, and interaction analysis, we have identified several promising candidates that exhibit strong binding affinities and favorable interaction profiles with the target protein. The predicted hit compounds warrant further investigation through in vitro and in vivo studies to confirm their efficacy and safety profiles. By identifying these alternative compounds, we aim to expand the therapeutic options available for diseases where ivermectin is currently used, potentially overcoming limitations associated with resistance or side effects.

General Conclusions

The results of this research provide significant insights into various aspects of infectious diseases and their treatment strategies. The studies encompassed in this thesis have advanced our understanding of SARS-CoV-2, particularly in the context of drug discovery and the impact of viral mutations on diagnostic efficacy, as well as provided novel perspectives on parasitic infections such as *S. stercoralis*. The investigation using interaction-based drug discovery screens has highlighted the utility of this approach in identifying and explaining known inhibitors of SARS-CoV-2, while also predicting new compound scaffolds, thereby showcasing its potential in the rapid identification of therapeutic candidates. This method has proven essential in mapping the interactions between viral proteins and potential inhibitors, laying a foundation for targeted drug development that could lead to more effective treatments for COVID-19. Furthermore, the emergence of a recurrent insertion in the N-terminal domain of the SARS-CoV-2 spike glycoprotein has significant implications for the virus's transmissibility and immune evasion capabilities. This discovery sheds light on the adaptive mechanisms of SARS-CoV-2, illustrating how genetic variations can influence viral behavior and impact the efficacy of vaccines and therapeutic antibodies. Understanding these mutations is crucial for anticipating potential changes in the virus that could affect public health measures and treatment strategies. Additionally, novel insights into the somatic proteome of *S. stercoralis* infective third-stage larvae have provided a deeper understanding of the molecular underpinnings of parasitic infection, revealing critical information about the proteins involved in the infective process and the parasite's survival mechanisms within the host. This knowledge is pivotal for developing new therapeutic interventions and diagnostic tools for strongyloidiasis. Moreover, real-life data supporting the reduced sensitivity of antigen tests for detecting Omicron SARS-CoV-2 infections highlight the challenges posed by the continuous evolution of the virus, emphasizing the need for ongoing evaluation and adaptation of diagnostic tools to ensure their effectiveness against new variants. This finding underscores the importance of robust and flexible testing strategies in managing the pandemic. Lastly, the exploration of drug repurposing strategies targeting the GluCl receptor for *S. stercoralis* infection offers promising avenues for treatment by leveraging the known safety profiles and mechanisms of action of existing drugs. This approach demonstrates the potential of repurposing in addressing neglected tropical diseases, where the development of new drugs may be economically and logistically challenging. Collectively, the findings of these studies underscore the importance of a multifaceted approach to infectious disease research, combining computational, molecular, and real-world data analyses. The insights gained from these investigations enhance our understanding of pathogens and inform the development of more effective diagnostic, therapeutic, and preventive measures. Future research should continue to focus on the dynamic interactions between pathogens

and hosts, the impact of genetic variations on disease progression, and the potential of innovative strategies like drug repurposing to address emerging and neglected infections. This thesis highlights the critical need for continued vigilance and adaptability in the face of evolving infectious diseases and underscores the importance of interdisciplinary collaboration in tackling these complex challenges.

Bibliography

[1] K. E. Jones *et al.*, 'Global trends in emerging infectious diseases', *Nature*, vol. 451, no. 7181, pp. 990–993, 2008.

[2] J. M. Van Seventer and N. S. Hochberg, 'Principles of infectious diseases: transmission, diagnosis, prevention, and control', *International encyclopedia of public health*, p. 22, 2017.

[3] A. J. Kucharski *et al.*, 'Early dynamics of transmission and control of COVID-19: a mathematical modelling study', *The lancet infectious diseases*, vol. 20, no. 5, pp. 553–558, 2020.

[4] R. M. Anderson *et al.*, 'The SARS-CoV-2 pandemic: Remaining uncertainties in our understanding of the epidemiology and transmission dynamics of the virus, and challenges to be overcome', *Interface Focus*, vol. 11, no. 6, p. 20210008, 2021.

[5] P. Zhou *et al.*, 'A pneumonia outbreak associated with a new coronavirus of probable bat origin', *nature*, vol. 579, no. 7798, pp. 270–273, 2020.

[6] J. F.-W. Chan *et al.*, 'Genomic characterization of the 2019 novel human-pathogenic coronavirus isolated from a patient with atypical pneumonia after visiting Wuhan', *Emerging microbes & infections*, vol. 9, no. 1, pp. 221–236, 2020.

[7] A. J. Rodriguez-Morales *et al.*, 'History is repeating itself: Probable zoonotic spillover as the cause of the 2019 novel Coronavirus Epidemic', *Infez Med*, vol. 28, no. 1, pp. 3–5, 2020.

[8] Y. Wang, M. Grunewald, and S. Perlman, 'Coronaviruses: an updated overview of their replication and pathogenesis', *Coronaviruses: Methods and Protocols*, pp. 1–29, 2020.

[9] D. X. Liu, J. Q. Liang, and T. S. Fung, 'Human coronavirus-229E,-OC43,-NL63, and-HKU1 (Coronaviridae)', *Encyclopedia of virology*, p. 428, 2021.

[10] A. K. Cordes, W. M. Rehrauer, M. A. Accola, B. Wölk, B. Hilfrich, and A. Heim, 'Fully automated detection and differentiation of pandemic and endemic coronaviruses (NL63, 229E, HKU1, OC43 and SARS-CoV-2) on the hologic panther fusion', *Journal of Medical Virology*, vol. 93, no. 7, pp. 4438–4445, 2021.

[11] B. Hijawi *et al.*, 'Novel coronavirus infections in Jordan, April 2012: epidemiological findings from a retrospective investigation', *EMHJ-Eastern Mediterranean Health Journal, 19 (supp. 1), S12-S18, 2013*, 2013.

[12] Z. A. Memish *et al.*, 'Human infection with MERS coronavirus after exposure to infected camels, Saudi Arabia, 2013', *Emerging infectious diseases*, vol. 20, no. 6, p. 1012, 2014.

[13] Z. Song *et al.*, 'From SARS to MERS, thrusting coronaviruses into the spotlight', *viruses*, vol. 11, no. 1, p. 59, 2019.

[14] C. W. Nelson *et al.*, 'Dynamically evolving novel overlapping gene as a factor in the SARS-CoV-2 pandemic', *Elife*, vol. 9, p. e59633, 2020.

[15] M. Beknazarova, H. Whiley, and K. Ross, 'Strongyloidiasis: a disease of socioeconomic disadvantage', *International journal of environmental research and public health*, vol. 13, no. 5, p. 517, 2016.

[16] D. Buonfrate *et al.*, 'The global prevalence of Strongyloides stercoralis infection', *Pathogens*, vol. 9, no. 6, p. 468, Jun. 2020, doi: 10.3390/pathogens9060468.

[17] V. L. Hunt *et al.*, 'The genomic basis of parasitism in the Strongyloides clade of nematodes', *Nat Genet*, vol. 48, no. 3, pp. 299–307, Mar. 2016, doi: 10.1038/ng.3495.

[18] Z. Bisoffi *et al.*, 'Diagnostic accuracy of five serologic tests for Strongyloides stercoralis infection', *PLoS Negl Trop Dis*, vol. 8, no. 1, p. 38, 2014, doi: 10.1371/journal.pntd.0002640.

[19] D. Buonfrate, P. Rodari, B. Barda, W. Page, L. Einsiedel, and M. R. Watts, 'Current pharmacotherapeutic strategies for Strongyloidiasis and the complications in its treatment', *Expert Opinion on Pharmacotherapy*, vol. 23, no. 14, pp. 1617–1628, 2022.
[20] P. Śledź and A. Caflisch, 'Protein structure-based drug design: from docking to molecular dynamics', *Current opinion in structural biology*, vol. 48, pp. 93–102, 2018.
[21] G. Sliwoski, S. Kothiwale, J. Meiler, and E. W. Lowe, 'Computational methods in drug discovery', *Pharmacological reviews*, vol. 66, no. 1, pp. 334–395, 2014.
[22] S. Muralidar, S. V. Ambi, S. Sekaran, and U. M. Krishnan, 'The emergence of COVID-19 as a global pandemic: Understanding the epidemiology, immune response and potential therapeutic targets of SARS-CoV-2', *Biochimie*, vol. 179, pp. 85–100, 2020.
[23] M. Gerdol, K. Dishnica, and A. Giorgetti, 'Emergence of a recurrent insertion in the N-terminal domain of the SARS-CoV-2 spike glycoprotein', *Virus Research*, vol. 310, p. 198674, Mar. 2022, doi: 10.1016/J.VIRUSRES.2022.198674.
[24] Y. Ma *et al.*, 'Structural basis and functional analysis of the SARS coronavirus nsp14-nsp10 complex', *Proceedings of the National Academy of Sciences of the United States of America*, vol. 112, no. 30, pp. 9436–9441, Jul. 2015, doi: 10.1073/PNAS.1508686112.
[25] M. R. Denison, R. L. Graham, E. F. Donaldson, L. D. Eckerle, and R. S. Baric, 'Coronaviruses: an RNA proofreading machine regulates replication fidelity and diversity', *RNA biology*, vol. 8, no. 2, pp. 270–279, 2011.
[26] Y. Shu and J. McCauley, 'GISAID: Global initiative on sharing all influenza data – from vision to reality', *Eurosurveillance*, vol. 22, no. 13, Mar. 2017, doi: 10.2807/1560-7917.ES.2017.22.13.30494.
[27] L. Ren *et al.*, 'Genetic drift of human coronavirus OC43 spike gene during adaptive evolution', *Scientific Reports*, vol. 5, Jun. 2015, doi: 10.1038/SREP11451.
[28] M. F. Boni *et al.*, 'Evolutionary origins of the SARS-CoV-2 sarbecovirus lineage responsible for the COVID-19 pandemic', *Nature Microbiology*, vol. 5, no. 11, pp. 1408–1417, Nov. 2020, doi: 10.1038/S41564-020-0771-4.
[29] D. VanInsberghe, A. S. Neish, A. C. Lowen, and K. Koelle, 'Recombinant SARS-CoV-2 genomes circulated at low levels over the first year of the pandemic', *Virus Evolution*, vol. 7, no. 2, 2021, doi: 10.1093/VE/VEAB059.
[30] J. K. Millet, J. A. Jaimes, and G. R. Whittaker, 'Molecular diversity of coronavirus host cell entry receptors', *FEMS Microbiology Reviews*, vol. 45, no. 3, May 2021, doi: 10.1093/FEMSRE/FUAA057.
[31] W. Ren *et al.*, 'Difference in Receptor Usage between Severe Acute Respiratory Syndrome (SARS) Coronavirus and SARS-Like Coronavirus of Bat Origin', *Journal of Virology*, vol. 82, no. 4, pp. 1899–1907, Feb. 2008, doi: 10.1128/JVI.01085-07.
[32] A. C. Walls, Y. J. Park, M. A. Tortorici, A. Wall, A. T. McGuire, and D. Veesler, 'Structure, Function, and Antigenicity of the SARS-CoV-2 Spike Glycoprotein', *Cell*, vol. 181, no. 2, pp. 281-292.e6, Apr. 2020, doi: 10.1016/J.CELL.2020.02.058.
[33] B. Korber *et al.*, 'Tracking Changes in SARS-CoV-2 Spike: Evidence that D614G Increases Infectivity of the COVID-19 Virus', *Cell*, vol. 182, no. 4, pp. 812-827.e19, Aug. 2020, doi: 10.1016/J.CELL.2020.06.043.
[34] L. Zhang *et al.*, 'SARS-CoV-2 spike-protein D614G mutation increases virion spike density and infectivity', *Nature Communications*, vol. 11, no. 1, Dec. 2020, doi: 10.1038/S41467-020-19808-4.
[35] L. van Dorp *et al.*, 'Emergence of genomic diversity and recurrent mutations in SARS-CoV-2', *Infection, Genetics and Evolution*, vol. 83, Sep. 2020, doi: 10.1016/J.MEEGID.2020.104351.

[36] L. van Dorp, D. Richard, C. C. S. Tan, L. P. Shaw, M. Acman, and F. Balloux, 'No evidence for increased transmissibility from recurrent mutations in SARS-CoV-2', *Nature Communications*, vol. 11, no. 1, Dec. 2020, doi: 10.1038/S41467-020-19818-2.
[37] N. Teruel, O. Mailhot, and R. J. Najmanovich, 'Modelling conformational state dynamics and its role on infection for SARS-CoV-2 Spike protein variants', *PLoS Computational Biology*, vol. 17, no. 8, Aug. 2021, doi: 10.1371/JOURNAL.PCBI.1009286.
[38] X. Zhu *et al.*, 'Cryo-electron microscopy structures of the N501Y SARS-CoV-2 spike protein in complex with ACE2 and 2 potent neutralizing antibodies', *PLoS Biology*, vol. 19, no. 4, Apr. 2021, doi: 10.1371/JOURNAL.PBIO.3001237.
[39] G. Nelson, O. Buzko, P. Spilman, K. Niazi, S. Rabizadeh, and P. Soon-Shiong, 'Molecular dynamic simulation reveals E484K mutation enhances spike RBD-ACE2 affinity and the combination of E484K, K417N and N501Y mutations (501Y. V2 variant) induces conformational change greater than N501Y mutant alone, potentially resulting in an escape mutant', *BioRxiv*, pp. 2021–01, 2021.
[40] N. G. Davies *et al.*, 'Estimated transmissibility and impact of SARS-CoV-2 lineage B.1.1.7 in England', *Science*, vol. 372, no. 6538, Apr. 2021, doi: 10.1126/SCIENCE.ABG3055.
[41] Y. Lustig *et al.*, 'Neutralizing Response against Variants after SARS-CoV-2 Infection and One Dose of BNT162b2', *New England Journal of Medicine*, vol. 384, no. 25, pp. 2453–2454, Jun. 2021, doi: 10.1056/NEJMC2104036.
[42] G.-L. Wang *et al.*, 'Susceptibility of Circulating SARS-CoV-2 Variants to Neutralization', *New England Journal of Medicine*, vol. 384, no. 24, pp. 2354–2356, Jun. 2021, doi: 10.1056/NEJMC2103022.
[43] Z. Wang *et al.*, 'mRNA vaccine-elicited antibodies to SARS-CoV-2 and circulating variants', *Nature*, vol. 592, no. 7855, pp. 616–622, Apr. 2021, doi: 10.1038/S41586-021-03324-6.
[44] X. Xie *et al.*, 'Neutralization of SARS-CoV-2 spike 69/70 deletion, E484K and N501Y variants by BNT162b2 vaccine-elicited sera', *Nature Medicine*, vol. 27, no. 4, pp. 620–621, Apr. 2021, doi: 10.1038/S41591-021-01270-4.
[45] P. Wang *et al.*, 'Antibody resistance of SARS-CoV-2 variants B.1.351 and B.1.1.7', *Nature*, vol. 593, no. 7857, pp. 130–135, May 2021, doi: 10.1038/S41586-021-03398-2.
[46] W. Sykes *et al.*, 'Prevalence of anti-SARS-CoV-2 antibodies among blood donors in Northern Cape, KwaZulu-Natal, Eastern Cape, and Free State provinces of South Africa in January 2021.', *Research square*, Feb. 2021, doi: 10.21203/RS.3.RS-233375/V1.
[47] E. C. Sabino *et al.*, 'Resurgence of COVID-19 in Manaus, Brazil, despite high seroprevalence', *The Lancet*, vol. 397, no. 10273, pp. 452–455, Feb. 2021, doi: 10.1016/S0140-6736(21)00183-5.
[48] B. Choi *et al.*, 'Persistence and Evolution of SARS-CoV-2 in an Immunocompromised Host', *New England Journal of Medicine*, vol. 383, no. 23, pp. 2291–2293, Dec. 2020, doi: 10.1056/NEJMC2031364.
[49] A. J. Greaney *et al.*, 'Mapping mutations to the SARS-CoV-2 RBD that escape binding by different classes of antibodies', *Nature Communications*, vol. 12, no. 1, Dec. 2021, doi: 10.1038/S41467-021-24435-8.
[50] T. N. Starr *et al.*, 'Deep Mutational Scanning of SARS-CoV-2 Receptor Binding Domain Reveals Constraints on Folding and ACE2 Binding', *Cell*, vol. 182, no. 5, pp. 1295-1310.e20, Sep. 2020, doi: 10.1016/J.CELL.2020.08.012.
[51] T. N. Starr, A. J. Greaney, A. S. Dingens, and J. D. Bloom, 'Complete map of SARS-CoV-2 RBD mutations that escape the monoclonal antibody LY-CoV555 and its cocktail with LY-CoV016', *Cell Reports Medicine*, vol. 2, no. 4, Apr. 2021, doi: 10.1016/J.XCRM.2021.100255.

[52] S. A. Madhi *et al.*, 'Efficacy of the ChAdOx1 nCoV-19 Covid-19 Vaccine against the B.1.351 Variant', *New England Journal of Medicine*, vol. 384, no. 20, pp. 1885–1898, May 2021, doi: 10.1056/NEJMOA2102214.
[53] X. Shen *et al.*, 'Neutralization of SARS-CoV-2 Variants B.1.429 and B.1.351', *New England Journal of Medicine*, vol. 384, no. 24, pp. 2352–2354, Jun. 2021, doi: 10.1056/NEJMC2103740.
[54] M. S. Dhar *et al.*, 'Genomic characterization and epidemiology of an emerging SARS-CoV-2 variant in Delhi, India', *Science*, vol. 374, no. 6570, pp. 995–999, Nov. 2021, doi: 10.1126/SCIENCE.ABJ9932.
[55] J. Dagpunar, 'Interim estimates of increased transmissibility, growth rate, and reproduction number of the Covid-19 B. 1.617. 2 variant of concern in the United Kingdom', *MedRxiv*, pp. 2021–06, 2021.
[56] A. Saito *et al.*, 'Enhanced fusogenicity and pathogenicity of SARS-CoV-2 Delta P681R mutation', *Nature*, vol. 602, no. 7896, pp. 300–306, Feb. 2022, doi: 10.1038/S41586-021-04266-9.
[57] D. Planas *et al.*, 'Reduced sensitivity of SARS-CoV-2 variant Delta to antibody neutralization', *Nature*, vol. 596, no. 7871, pp. 276–280, Aug. 2021, doi: 10.1038/S41586-021-03777-9.
[58] J. L. Bernal *et al.*, 'Effectiveness of Covid-19 Vaccines against the B.1.617.2 (Delta) Variant', *New England Journal of Medicine*, vol. 385, no. 7, pp. 585–594, Aug. 2021, doi: 10.1056/NEJMOA2108891.
[59] J. R. C. Pulliam *et al.*, 'Increased risk of SARS-CoV-2 reinfection associated with emergence of Omicron in South Africa', *Science*, vol. 376, no. 6593, May 2022, doi: 10.1126/SCIENCE.ABN4947.
[60] S. Cele *et al.*, 'SARS-CoV-2 Omicron has extensive but incomplete escape of Pfizer BNT162b2 elicited neutralization and requires ACE2 for infection.', *medRxiv : the preprint server for health sciences*, Dec. 2021, doi: 10.1101/2021.12.08.21267417.
[61] W. F. Garcia-Beltran *et al.*, 'mRNA-based COVID-19 vaccine boosters induce neutralizing immunity against SARS-CoV-2 Omicron variant', *Cell*, vol. 185, no. 3, pp. 457-466.e4, Feb. 2022, doi: 10.1016/J.CELL.2021.12.033.
[62] F. Schmidt *et al.*, 'Plasma Neutralization of the SARS-CoV-2 Omicron Variant', *New England Journal of Medicine*, vol. 386, no. 6, pp. 599–601, Feb. 2022, doi: 10.1056/NEJMC2119641.
[63] B. J. Willett *et al.*, 'The hyper-transmissible SARS-CoV-2 Omicron variant exhibits significant antigenic change, vaccine escape and a switch in cell entry mechanism', *medRxiv*, 2022, doi: 10.1101/2022.01.03.21268111.
[64] K. R. McCarthy *et al.*, 'Recurrent deletions in the SARS-CoV-2 spike glycoprotein drive antibody escape', *Science*, vol. 371, no. 6534, pp. 1139–1142, Mar. 2021, doi: 10.1126/SCIENCE.ABF6950.
[65] B. A. Johnson *et al.*, 'Loss of furin cleavage site attenuates SARS-CoV-2 pathogenesis', *Nature*, vol. 591, no. 7849, pp. 293–299, Mar. 2021, doi: 10.1038/S41586-021-03237-4.
[66] X. Y. Ge *et al.*, 'Coexistence of multiple coronaviruses in several bat colonies in an abandoned mineshaft', *Virologica Sinica*, vol. 31, no. 1, pp. 31–40, Feb. 2016, doi: 10.1007/S12250-016-3713-9.
[67] H. Zhou *et al.*, 'A Novel Bat Coronavirus Closely Related to SARS-CoV-2 Contains Natural Insertions at the S1/S2 Cleavage Site of the Spike Protein', *Current Biology*, vol. 30, no. 11, pp. 2196-2203.e3, Jun. 2020, doi: 10.1016/J.CUB.2020.05.023.

[68] H. Zhou *et al.*, 'Identification of novel bat coronaviruses sheds light on the evolutionary origins of SARS-CoV-2 and related viruses', *Cell*, vol. 184, no. 17, pp. 4380-4391.e14, Aug. 2021, doi: 10.1016/J.CELL.2021.06.008.
[69] S. Wacharapluesadee *et al.*, 'Evidence for SARS-CoV-2 related coronaviruses circulating in bats and pangolins in Southeast Asia', *Nature Communications*, vol. 12, no. 1, Dec. 2021, doi: 10.1038/S41467-021-21240-1.
[70] R. C. Edgar, 'MUSCLE: Multiple sequence alignment with high accuracy and high throughput', *Nucleic Acids Research*, vol. 32, no. 5, pp. 1792–1797, 2004, doi: 10.1093/NAR/GKH340.
[71] S. Kumar, G. Stecher, M. Li, C. Knyaz, and K. Tamura, 'MEGA X: Molecular evolutionary genetics analysis across computing platforms', *Molecular Biology and Evolution*, vol. 35, no. 6, pp. 1547–1549, Jun. 2018, doi: 10.1093/MOLBEV/MSY096.
[72] K. Katoh, K. Misawa, K. I. Kuma, and T. Miyata, 'MAFFT: A novel method for rapid multiple sequence alignment based on fast Fourier transform', *Nucleic Acids Research*, vol. 30, no. 14, pp. 3059–3066, Jul. 2002, doi: 10.1093/NAR/GKF436.
[73] M. N. Price, P. S. Dehal, and A. P. Arkin, 'FastTree 2 - Approximately maximum-likelihood trees for large alignments', *PLoS ONE*, vol. 5, no. 3, Mar. 2010, doi: 10.1371/JOURNAL.PONE.0009490.
[74] P. Sagulenko, V. Puller, and R. A. Neher, 'TreeTime: Maximum-likelihood phylodynamic analysis', *Virus Evolution*, vol. 4, no. 1, Jan. 2018, doi: 10.1093/VE/VEX042.
[75] Y. Cai *et al.*, 'Distinct conformational states of SARS-CoV-2 spike protein', *Science*, vol. 369, no. 6511, Sep. 2020, doi: 10.1126/SCIENCE.ABD4251.
[76] A. Waterhouse *et al.*, 'SWISS-MODEL: Homology modelling of protein structures and complexes', *Nucleic Acids Research*, vol. 46, no. W1, pp. W296–W303, Jul. 2018, doi: 10.1093/NAR/GKY427.
[77] R. Aguayo-Ortiz, C. Chávez-García, J. E. Straub, and L. Dominguez, 'Characterizing the structural ensemble of γ-secretase using a multiscale molecular dynamics approach', *Chemical Science*, vol. 8, no. 8, pp. 5576–5584, 2017, doi: 10.1039/C7SC00980A.
[78] D. V. D. Spoel, E. Lindahl, B. Hess, G. Groenhof, A. E. Mark, and H. J. C. Berendsen, 'GROMACS: Fast, flexible, and free', *Journal of Computational Chemistry*, vol. 26, no. 16, pp. 1701–1718, Dec. 2005, doi: 10.1002/JCC.20291.
[79] G. Bussi, D. Donadio, and M. Parrinello, 'Canonical sampling through velocity rescaling', *Journal of Chemical Physics*, vol. 126, no. 1, 2007, doi: 10.1063/1.2408420.
[80] R. Martoňák, A. Laio, and M. Parrinello, 'Predicting Crystal Structures: The Parrinello-Rahman Method Revisited', *Physical Review Letters*, vol. 90, no. 7, p. 4, 2003, doi: 10.1103/PHYSREVLETT.90.075503.
[81] N. Michaud-Agrawal, E. J. Denning, T. B. Woolf, and O. Beckstein, 'MDAnalysis: A toolkit for the analysis of molecular dynamics simulations', *Journal of Computational Chemistry*, vol. 32, no. 10, pp. 2319–2327, Jul. 2011, doi: 10.1002/JCC.21787.
[82] W. L. DeLano and others, 'Pymol: An open-source molecular graphics tool', *CCP4 Newsl. Protein Crystallogr*, vol. 40, no. 1, pp. 82–92, 2002.
[83] K. Laiton-Donato *et al.*, 'Characterization of the emerging B.1.621 variant of interest of SARS-CoV-2', *Infection, Genetics and Evolution*, vol. 95, Nov. 2021, doi: 10.1016/j.meegid.2021.105038.
[84] R. Viana *et al.*, 'Rapid epidemic expansion of the SARS-CoV-2 Omicron variant in southern Africa', *Nature*, vol. 603, no. 7902, pp. 679–686, Mar. 2022, doi: 10.1038/S41586-022-04411-Y.

[85] B. Meng *et al.*, 'Recurrent emergence of SARS-CoV-2 spike deletion H69/V70 and its role in the Alpha variant B.1.1.7', *Cell Reports*, vol. 35, no. 13, Jun. 2021, doi: 10.1016/j.celrep.2021.109292.
[86] W. T. Harvey *et al.*, 'SARS-CoV-2 variants, spike mutations and immune escape', *Nature Reviews Microbiology*, vol. 19, no. 7, pp. 409–424, Jul. 2021, doi: 10.1038/S41579-021-00573-0.
[87] T. P. Peacock *et al.*, 'The SARS-CoV-2 variant, Omicron, shows rapid replication in human primary nasal epithelial cultures and efficiently uses the endosomal route of entry', *BioRxiv*, vol. 10, no. 2021.12, pp. 31–474653, 2022.
[88] S. Teng, A. Sobitan, R. Rhoades, D. Liu, and Q. Tang, 'Systemic effects of missense mutations on SARS-CoV-2 spike glycoprotein stability and receptor-binding affinity', *Briefings in Bioinformatics*, vol. 22, no. 2, pp. 1239–1253, Mar. 2021, doi: 10.1093/BIB/BBAA233.
[89] Z. Liu *et al.*, 'Identification of SARS-CoV-2 spike mutations that attenuate monoclonal and serum antibody neutralization', *Cell Host and Microbe*, vol. 29, no. 3, pp. 477-488.e4, Mar. 2021, doi: 10.1016/J.CHOM.2021.01.014.
[90] P. C. Resende *et al.*, 'The ongoing evolution of variants of concern and interest of SARS-CoV-2 in Brazil revealed by convergent indels in the amino (N)-terminal domain of the spike protein', *Virus Evolution*, vol. 7, no. 2, 2021, doi: 10.1093/VE/VEAB069.
[91] E. Volz *et al.*, 'Assessing transmissibility of SARS-CoV-2 lineage B.1.1.7 in England', *Nature*, vol. 593, no. 7858, pp. 266–269, May 2021, doi: 10.1038/S41586-021-03470-X.
[92] E. A. Ozer *et al.*, 'Coincident rapid expansion of two SARS-CoV-2 lineages with enhanced infectivity in Nigeria.', *medRxiv : the preprint server for health sciences*, Jul. 2021, doi: 10.1101/2021.04.09.21255206.
[93] W. Zhang, B. D. Davis, S. S. Chen, J. M. S. Martinez, J. T. Plummer, and E. Vail, 'Emergence of a Novel SARS-CoV-2 Variant in Southern California', *JAMA - Journal of the American Medical Association*, vol. 325, no. 13, pp. 1324–1326, Apr. 2021, doi: 10.1001/JAMA.2021.1612.
[94] A. Venkatakrishnan *et al.*, 'Omicron variant of SARS-CoV-2 harbors a unique insertion mutation of putative viral or human genomic origin', 2021.
[95] B. S. Chrisman *et al.*, 'Indels in SARS-CoV-2 occur at template-switching hotspots', *BioData Mining*, vol. 14, no. 1, Dec. 2021, doi: 10.1186/S13040-021-00251-0.
[96] S. K. Garushyants, I. B. Rogozin, and E. V. Koonin, 'Template switching and duplications in SARS-CoV-2 genomes give rise to insertion variants that merit monitoring', *Communications Biology*, vol. 4, no. 1, Dec. 2021, doi: 10.1038/S42003-021-02858-9.
[97] N. Shiliaev *et al.*, 'Natural and Recombinant SARS-CoV-2 Isolates Rapidly Evolve In Vitro to Higher Infectivity through More Efficient Binding to Heparan Sulfate and Reduced S1/S2 Cleavage', *Journal of Virology*, vol. 95, no. 21, Oct. 2021, doi: 10.1128/JVI.01357-21.
[98] A. Singh, G. Steinkellner, K. Köchl, K. Gruber, and C. C. Gruber, 'Serine 477 plays a crucial role in the interaction of the SARS-CoV-2 spike protein with the human receptor ACE2', *Scientific Reports*, vol. 11, no. 1, Dec. 2021, doi: 10.1038/S41598-021-83761-5.
[99] I. A. T. M. Ferreira *et al.*, 'SARS-CoV-2 B.1.617 Mutations L452R and E484Q Are Not Synergistic for Antibody Evasion', *Journal of Infectious Diseases*, vol. 224, no. 6, pp. 989–994, Sep. 2021, doi: 10.1093/INFDIS/JIAB368.
[100] D. L. Bugembe *et al.*, 'Emergence and spread of a SARS-CoV-2 lineage A variant (A.23.1) with altered spike protein in Uganda', *Nature Microbiology*, vol. 6, no. 8, pp. 1094–1101, Aug. 2021, doi: 10.1038/S41564-021-00933-9.

[101] D. Franco *et al.*, 'Early transmission dynamics, spread, and genomic characterization of SARS-CoV-2 in Panama', *Emerging Infectious Diseases*, vol. 27, no. 2, pp. 612–615, Feb. 2021, doi: 10.3201/EID2702.203767.

[102] D. Vavrek, L. Speroni, K. J. Curnow, M. Oberholzer, V. Moeder, and P. G. Febbo, 'Genomic surveillance at scale is required to detect newly emerging strains at an early timepoint', *MedRxiv*, pp. 2021–01, 2021.

[103] Y. Díaz *et al.*, 'SARS-CoV-2 reinfection with a virus harboring mutation in the Spike and the Nucleocapsid proteins in Panama', *International Journal of Infectious Diseases*, vol. 108, pp. 588–591, Jul. 2021, doi: 10.1016/j.ijid.2021.06.004.

[104] J. A. Molina-Mora *et al.*, 'SARS-CoV-2 genomic surveillance in Costa Rica: Evidence of a divergent population and an increased detection of a spike T1117I mutation', *Infection, Genetics and Evolution*, vol. 92, Aug. 2021, doi: 10.1016/J.MEEGID.2021.104872.

[105] G. Cerutti *et al.*, 'Potent SARS-CoV-2 neutralizing antibodies directed against spike N-terminal domain target a single supersite', *Cell Host and Microbe*, vol. 29, no. 5, pp. 819-833.e7, May 2021, doi: 10.1016/J.CHOM.2021.03.005.

[106] M. McCallum *et al.*, 'N-terminal domain antigenic mapping reveals a site of vulnerability for SARS-CoV-2', *Cell*, vol. 184, no. 9, pp. 2332-2347.e16, Apr. 2021, doi: 10.1016/J.CELL.2021.03.028.

[107] A. Tarke *et al.*, 'Comprehensive analysis of T cell immunodominance and immunoprevalence of SARS-CoV-2 epitopes in COVID-19 cases', *Cell Reports Medicine*, vol. 2, no. 2, Feb. 2021, doi: 10.1016/J.XCRM.2021.100204.

[108] R. F. Garry *et al.*, 'Spike protein mutations in novel SARS-CoV-2 'variants of concern'commonly occur in or near indels', *image*, vol. 881, no. 1147, p. 85, 2021.

[109] S. Temmam *et al.*, 'Coronaviruses with a SARS-CoV-2-like receptor-binding domain allowing ACE2-mediated entry into human cells isolated from bats of Indochinese peninsula', 2021.

[110] S. Lytras, J. Hughes, W. Xia, X. Jiang, and D. L. Robertson, 'Exploring the natural origins of SARS-CoV-2', *bioRxiv*, p. 2021.01.22.427830, 2021, doi: 10.1101/2021.01.22.427830.

[111] T. Hansson, C. Oostenbrink, and W. F. V. Gunsteren, 'Molecular dynamics simulations', *Current Opinion in Structural Biology*, vol. 12, no. 2, pp. 190–196, Apr. 2002, doi: 10.1016/S0959-440X(02)00308-1.

[112] P. Schake, K. Dishnica, F. Kaiser, C. Leberecht, V. J. Haupt, and M. Schroeder, 'An interaction-based drug discovery screen explains known SARS-CoV-2 inhibitors and predicts new compound scaffolds', *Scientific Reports 2023 13:1*, vol. 13, no. 1, pp. 1–13, Jun. 2023, doi: 10.1038/s41598-023-35671-x.

[113] P. J. Walker *et al.*, 'Changes to virus taxonomy and the Statutes ratified by the International Committee on Taxonomy of Viruses (2020)', *Arch. Virol.*, vol. 165, no. 11, pp. 2737–2748, Nov. 2020, doi: 10.1007/s00705-020-04752-x.

[114] A. E. Gorbalenya *et al.*, 'Coronaviridae Study Group of the International Committee on Taxonomy of Viruses. The species severe acute respiratory syndrome-related coronavirus: classifying 2019-nCoV and naming it SARS-CoV-2', *Nat. Microbiol*, vol. 5, no. 4, pp. 536–544, 2020.

[115] A. M. Zaki, S. van Boheemen, T. M. Bestebroer, A. D. M. E. Osterhaus, and R. A. M. Fouchier, 'Isolation of a novel coronavirus from a man with pneumonia in Saudi Arabia', *N. Engl. J. Med.*, vol. 367, no. 19, pp. 1814–1820, Nov. 2012, doi: 10.1056/nejmoa1211721.

[116] S. Su *et al.*, 'Epidemiology, genetic recombination, and pathogenesis of coronaviruses', *Trends Microbiol.*, vol. 24, no. 6, pp. 490–502, Jun. 2016, doi: 10.1016/j.tim.2016.03.003.

[117] G. Li and E. D. Clercq, 'Therapeutic options for the 2019 novel coronavirus (2019-nCoV)', *Nat. Rev. Drug Discov.*, vol. 19, no. 3, pp. 149–150, Mar. 2020, doi: 10.1038/d41573-020-00016-0.

[118] A. Zumla, J. F. W. Chan, E. I. Azhar, D. S. C. Hui, and K. Y. Yuen, 'Coronaviruses — Drug discovery and therapeutic options', *Nat. Rev. Drug Discov.*, vol. 15, no. 5, pp. 327–347, May 2016, doi: 10.1038/nrd.2015.37.

[119] A. A. T. Naqvi *et al.*, 'Insights into SARS-CoV-2 genome, structure, evolution, pathogenesis and therapies: Structural genomics approach', *Biochim. Biophys. Acta BBA-Mol. Basis Dis.*, vol. 1866, no. 10, p. 165878, Oct. 2020, doi: 10.1016/j.bbadis.2020.165878.

[120] Z. Jin *et al.*, 'Structure of Mpro from SARS-CoV-2 and discovery of its inhibitors', *Nature*, vol. 582, no. 7811, pp. 289–293, Jun. 2020, doi: 10.1038/s41586-020-2223-y.

[121] L. Zhang *et al.*, 'Crystal structure of SARS-CoV-2 main protease provides a basis for design of improved α-ketoamide inhibitors', *Science*, vol. 368, no. 6489, pp. 409–412, Apr. 2020, doi: 10.1126/science.abb3405.

[122] Y. L. Ng, C. K. Salim, and J. J. H. Chu, 'Drug repurposing for COVID-19: Approaches, challenges and promising candidates', *Pharmacol. Ther.*, vol. 228, Dec. 2021, doi: 10.1016/j.pharmthera.2021.107930.

[123] J. Wei *et al.*, 'Genome-wide CRISPR screens reveal host factors critical for SARS-CoV-2 infection', *Cell*, vol. 184, no. 1, pp. 76-91.e13, Jan. 2021, doi: 10.1016/j.cell.2020.10.028.

[124] A. Wahl *et al.*, 'SARS-CoV-2 infection is effectively treated and prevented by EIDD-2801', *Nature*, vol. 591, no. 7850, pp. 451–457, Mar. 2021, doi: 10.1038/s41586-021-03312-w.

[125] N. Drayman *et al.*, 'Masitinib is a broad coronavirus 3CL inhibitor that blocks replication of SARS-CoV-2', *Science*, vol. 373, no. 6557, pp. 931–936, Aug. 2021, doi: 10.1126/science.abg5827.

[126] J. Qiao *et al.*, 'SARS-CoV-2 Mpro inhibitors with antiviral activity in a transgenic mouse model', *Science*, vol. 371, no. 6536, pp. 1374–1378, Mar. 2021, doi: 10.1126/science.abf1611.

[127] T. A. Tummino *et al.*, 'Drug-induced phospholipidosis confounds drug repurposing for SARS-CoV-2', *Science*, vol. 373, no. 6554, pp. 541–547, Jul. 2021, doi: 10.1126/science.abi4708.

[128] K. Amporndanai *et al.*, 'Inhibition mechanism of SARS-CoV-2 main protease by ebselen and its derivatives', *Nat. Commun.*, vol. 12, no. 1, p. 3061, Dec. 2021, doi: 10.1038/s41467-021-23313-7.

[129] M. Thoms *et al.*, 'Structural basis for translational shutdown and immune evasion by the Nsp1 protein of SARS-CoV-2', *Science*, vol. 369, no. 6508, pp. 1249–1255, Sep. 2020, doi: 10.1126/science.abc8665.

[130] L. Riva *et al.*, 'Discovery of SARS-CoV-2 antiviral drugs through large-scale compound repurposing', *Nature*, vol. 586, no. 7827, pp. 113–119, Oct. 2020, doi: 10.1038/s41586-020-2577-1.

[131] D. R. Owen *et al.*, 'An oral SARS-CoV-2 M pro inhibitor clinical candidate for the treatment of COVID-19', *Science*, vol. 374, no. 6575, pp. 1586–1593, Dec. 2021, doi: 10.1126/science.abl4784.

[132] E. Mahase, 'Covid-19: Pfizer's paxlovid is 89% effective in patients at risk of serious illness, company reports', *BMJ*, vol. 375, p. n2713, Nov. 2021, doi: 10.1136/bmj.n2713.

[133] C. Ma *et al.*, 'Boceprevir, GC-376, and calpain inhibitors II, XII inhibit SARS-CoV-2 viral replication by targeting the viral main protease', *Cell Res.*, vol. 30, no. 8, pp. 678–692, Aug. 2020, doi: 10.1038/s41422-020-0356-z.

[134] C. S. Adamson, K. Chibale, R. J. M. Goss, M. Jaspars, D. J. Newman, and R. A. Dorrington, 'Antiviral drug discovery: Preparing for the next pandemic', *Chem. Soc. Rev.*, vol. 50, no. 6, pp. 3647–3655, Mar. 2021, doi: 10.1039/d0cs01118e.

[135] K. Anand, J. Ziebuhr, P. Wadhwani, J. R. Mesters, and R. Hilgenfeld, 'Coronavirus main proteinase (3CL pro) structure: Basis for design of anti-SARS drugs', *Science*, vol. 300, no. 5626, pp. 1763–1767, Jun. 2003, doi: 10.1126/science.1085658.

[136] A. Douangamath *et al.*, 'Crystallographic and electrophilic fragment screening of the SARS-CoV-2 main protease', *Nat. Commun.*, vol. 11, no. 1, p. 5047, Dec. 2020, doi: 10.1038/s41467-020-18709-w.

[137] J. Lee *et al.*, 'Crystallographic structure of wild-type SARS-CoV-2 main protease acyl-enzyme intermediate with physiological C-terminal autoprocessing site', *Nat. Commun.*, vol. 11, no. 1, p. 5877, Dec. 2020, doi: 10.1038/s41467-020-19662-4.

[138] M. Jaskolski *et al.*, 'Crystallographic models of SARS-CoV-2 3CL pro : In-depth assessment of structure quality and validation', *IUCrJ*, vol. 8, pp. 238–256, 2021, doi: 10.1107/s2052252521001159.

[139] D. W. Kneller *et al.*, 'Structural plasticity of SARS-CoV-2 3CL Mpro active site cavity revealed by room temperature X-ray crystallography', *Nat. Commun.*, vol. 11, no. 1, p. 3202, Dec. 2020, doi: 10.1038/s41467-020-16954-7.

[140] W. Yin *et al.*, 'Structural basis for inhibition of the RNA-dependent RNA polymerase from SARS-CoV-2 by remdesivir', *Science*, vol. 368, no. 6498, pp. 1499–1504, Jun. 2020, doi: 10.1126/science.abc1560.

[141] A. T. Ton, F. Gentile, M. Hsing, F. Ban, and A. Cherkasov, 'Rapid identification of potential inhibitors of SARS-CoV-2 main protease by deep docking of 1.3 billion compounds', *Mol. Inform.*, vol. 39, no. 8, p. 2000028, Aug. 2020, doi: 10.1002/minf.202000028.

[142] F. Gentile *et al.*, 'Automated discovery of noncovalent inhibitors of SARS-CoV-2 main protease by consensus deep docking of 40 billion small molecules', *Chem. Sci.*, vol. 12, no. 48, pp. 15960–15974, Dec. 2021, doi: 10.1039/d1sc05579h.

[143] S. Bharadwaj, A. Dubey, U. Yadava, S. K. Mishra, S. G. Kang, and V. D. Dwivedi, 'Exploration of natural compounds with anti-SARS-CoV-2 activity via inhibition of SARS-CoV-2 Mpro', *Brief. Bioinform.*, vol. 22, no. 2, pp. 1361–1377, Mar. 2021, doi: 10.1093/bib/bbaa382.

[144] M. T. ul Qamar, S. M. Alqahtani, M. A. Alamri, and L. L. Chen, 'Structural basis of SARS-CoV-2 3CLpro and anti-COVID-19 drug discovery from medicinal plants', *J. Pharm. Anal.*, vol. 10, no. 4, pp. 313–319, Aug. 2020, doi: 10.1016/j.jpha.2020.03.009.

[145] E. N. Muratov *et al.*, 'A critical overview of computational approaches employed for COVID-19 drug discovery', *Chem. Soc. Rev.*, vol. 50, no. 16, pp. 9121–9151, Aug. 2021, doi: 10.1039/d0cs01065k.

[146] M. F. Adasme *et al.*, 'PLIP 2021: Expanding the scope of the protein–ligand interaction profiler to DNA and RNA', *Nucleic Acids Res.*, vol. 49, no. W1, pp. W530–W534, Jul. 2021, doi: 10.1093/nar/gkab294.

[147] S. Salentin, S. Schreiber, V. J. Haupt, M. F. Adasme, and M. Schroeder, 'PLIP: Fully automated protein–ligand interaction profiler', *Nucleic Acids Res.*, vol. 43, no. W1, pp. W443–W447, 2015, doi: 10.1093/nar/gkv315.

[148] H. Berman, K. Henrick, H. Nakamura, and J. L. Markley, 'The worldwide Protein Data Bank (wwPDB): Ensuring a single, uniform archive of PDB data', *Nucleic Acids Res.*, vol. 35, no. SUPPL. 1, pp. D301–D303, Jan. 2007, doi: 10.1093/nar/gkl971.

[149] S. Salentin *et al.*, 'From malaria to cancer: Computational drug repositioning of amodiaquine using PLIP interaction patterns', *Sci. Rep.*, vol. 7, no. 1, p. 11401, Dec. 2017, doi: 10.1038/s41598-017-11924-4.
[150] M. F. Adasme *et al.*, 'Structure-based drug repositioning explains ibrutinib as VEGFR2 inhibitor', *PLoS ONE*, vol. 15, no. 5, May 2020, doi: 10.1371/journal.pone.0233089.
[151] M. F. Adasme *et al.*, 'Repositioned drugs for chagas disease unveiled via structure-based drug repositioning', *Int. J. Mol. Sci.*, vol. 21, no. 22, p. 8809, Nov. 2020, doi: 10.3390/ijms21228809.
[152] D. Rogers and M. Hahn, 'Extended-connectivity fingerprints', *J. Chem. Inf. Model.*, vol. 50, no. 5, pp. 742–754, May 2010, doi: 10.1021/ci100050t.
[153] P. Virtanen *et al.*, 'SciPy 1.0: Fundamental algorithms for scientific computing in Python', *Nat. Methods*, vol. 17, no. 3, pp. 261–272, Mar. 2020, doi: 10.1038/s41592-019-0686-2.
[154] M. Kuzikov *et al.*, 'Identification of Inhibitors of SARS-CoV-2 3CL-pro enzymatic activity using a small molecule in vitro repurposing screen', *ACS Pharmacol. Transl. Sci.*, vol. 4, no. 3, pp. 1096–1110, Jun. 2021, doi: 10.1021/acsptsci.0c00216.
[155] A. Sonousi, H. A. Mahran, I. M. Ibrahim, M. N. Ibrahim, A. A. Elfiky, and W. M. Elshemey, 'Novel adenosine derivatives against SARS-CoV-2 RNA-dependent RNA polymerase: An in silico perspective', *Pharmacol. Rep.*, vol. 73, no. 6, pp. 1754–1764, Dec. 2021, doi: 10.1007/s43440-021-00300-9.
[156] A. M. Rabie, 'Potent inhibitory activities of the adenosine analogue cordycepin on SARS-CoV-2 replication', *ACS Omega*, vol. 7, no. 3, pp. 2960–2969, Jan. 2022, doi: 10.1021/acsomega.1c05998.
[157] D. Tian *et al.*, 'An update review of emerging small-molecule therapeutic options for COVID-19', *Biomed. Pharmacother.*, vol. 137, May 2021, doi: 10.1016/j.biopha.2021.111313.
[158] T. Zhu *et al.*, 'Hit identification and optimization in virtual screening: Practical recommendations based on a critical literature analysis', *J. Med. Chem.*, vol. 56, no. 17, pp. 6560–6572, Sep. 2013, doi: 10.1021/jm301916b.
[159] H. Do *et al.*, 'Crystal structure of UbiX, an aromatic acid decarboxylase from the psychrophilic bacterium Colwellia psychrerythraea that undergoes FMN-induced conformational changes', *Sci. Rep.*, vol. 5, p. 8196, Feb. 2015, doi: 10.1038/srep08196.
[160] R. A. Akasov *et al.*, 'Riboflavin for COVID-19 adjuvant treatment in patients with mental health disorders: Observational study', *Front. Pharmacol.*, vol. 13, Mar. 2022, doi: 10.3389/fphar.2022.755745.
[161] J. R. Horton, K. Sawada, M. Nishibori, and X. Cheng, 'Structural basis for inhibition of histamine N-methyltransferase by diverse drugs', *J. Mol. Biol.*, vol. 353, no. 2, pp. 334–344, Oct. 2005, doi: 10.1016/j.jmb.2005.08.040.
[162] G. Bocci *et al.*, 'Virtual and in vitro antiviral screening revive therapeutic drugs for COVID-19', *ACS Pharmacol. Transl. Sci.*, vol. 3, no. 6, pp. 1278–1292, Dec. 2020, doi: 10.1021/acsptsci.0c00131.
[163] M. Hagar, H. A. Ahmed, G. Aljohani, and O. A. Alhaddad, 'Investigation of some antiviral n-heterocycles as COVID 19 drug: Molecular docking and DFT calculations', *Int. J. Mol. Sci.*, vol. 21, no. 11, p. 3922, Jun. 2020, doi: 10.3390/ijms21113922.
[164] S. Lorenz, P. Deng, O. Hantschel, G. Superti-Furga, and J. Kuriyan, 'Crystal structure of an SH2-kinase construct of c-Abl and effect of the SH2 domain on kinase activity', *Biochem. J.*, vol. 468, no. 2, pp. 283–291, 2015, doi: 10.1042/bj20141492.
[165] E. Abruzzese, L. Luciano, F. D'Agostino, M. M. Trawinska, F. Pane, and P. de Fabritiis, 'SARS-CoV-2 (COVID-19) and chronic myeloid leukemia (CML): A case report and review of ABL

kinase involvement in viral infection', *Mediterr. J. Hematol. Infect. Dis.*, vol. 12, no. 1, 2020, doi: 10.4084/mjhid.2020.031.
[166] R. Xiang *et al.*, 'Recent advances in developing small-molecule inhibitors against SARS-CoV-2', *Acta Pharm. Sin. B*, vol. 12, no. 4, pp. 1591–1623, Apr. 2022, doi: 10.1016/j.apsb.2021.06.016.
[167] M. Caracciolo *et al.*, 'Efficacy and effect of inhaled adenosine treatment in hospitalized COVID-19 patients', *Front. Immunol.*, vol. 12, p. 613070, Mar. 2021, doi: 10.3389/fimmu.2021.613070.
[168] C. Falcone *et al.*, 'Can adenosine fight COVID-19 acute respiratory distress syndrome?', *J. Clin. Med.*, vol. 9, no. 9, p. 3045, Sep. 2020, doi: 10.3390/jcm9093045.
[169] R. F. D. Freitas and M. Schapira, 'A systematic analysis of atomic protein–ligand interactions in the PDB', *MedChemComm*, vol. 8, no. 10, pp. 1970–1981, 2017, doi: 10.1039/c7md00381a.
[170] J. Tan *et al.*, 'pH-dependent conformational flexibility of the SARS-CoV main proteinase (Mpro) dimer: Molecular dynamics simulations and multiple X-ray structure analyses', *J. Mol. Biol.*, vol. 354, no. 1, pp. 25–40, Nov. 2005, doi: 10.1016/j.jmb.2005.09.012.
[171] K. A. Sulaiman *et al.*, 'Ascorbic acid as an adjunctive therapy in critically ill patients with COVID-19: A propensity score matched study', *Sci. Rep.*, vol. 11, no. 1, p. 17648, Dec. 2021, doi: 10.1038/s41598-021-96703-y.
[172] S. Thomas *et al.*, 'Effect of high-dose zinc and ascorbic acid supplementation vs usual care on symptom length and reduction among ambulatory patients with SARS-CoV-2 infection: The COVID A to Z randomized clinical trial', *JAMA Netw. Open*, vol. 4, no. 2, Feb. 2021, doi: 10.1001/jamanetworkopen.2021.0369.
[173] C. Piubelli *et al.*, 'Wide Real-Life Data Support Reduced Sensitivity of Antigen Tests for Omicron SARS-CoV-2 Infections', *Viruses*, vol. 16, no. 5, p. 657, May 2024, doi: 10.3390/V16050657/S1.
[174] W. H. Organization, 'WHO Coronavirus (COVID-19) Dashboard'. 2023. [Online]. Available: https://data.who.int/dashboards/covid19/
[175] B. J. Tromberg *et al.*, 'Rapid Scaling Up of Covid-19 Diagnostic Testing in the United States — The NIH RADx Initiative', *New England Journal of Medicine*, vol. 383, no. 11, pp. 1071–1077, Sep. 2020, doi: 10.1056/NEJMSR2022263.
[176] S. Fourati *et al.*, 'Performance of six rapid diagnostic tests for SARS-CoV-2 antigen detection and implications for practical use', *Journal of Clinical Virology*, vol. 142, Sep. 2021, doi: 10.1016/J.JCV.2021.104930.
[177] Z. Bai, Y. Cao, W. Liu, and J. Li, 'The sars-cov-2 nucleocapsid protein and its role in viral structure, biological functions, and a potential target for drug or vaccine mitigation', *Viruses*, vol. 13, no. 6, Jun. 2021, doi: 10.3390/V13061115.
[178] B. Diao *et al.*, 'Accuracy of a nucleocapsid protein antigen rapid test in the diagnosis of SARS-CoV-2 infection', *Clinical Microbiology and Infection*, vol. 27, no. 2, p. 289.e1-289.e4, Feb. 2021, doi: 10.1016/J.CMI.2020.09.057.
[179] G. C. Mak *et al.*, 'Evaluation of rapid antigen test for detection of SARS-CoV-2 virus', *Journal of Clinical Virology*, vol. 129, Aug. 2020, doi: 10.1016/J.JCV.2020.104500.
[180] L. Porte *et al.*, 'Evaluation of a novel antigen-based rapid detection test for the diagnosis of SARS-CoV-2 in respiratory samples', *International Journal of Infectious Diseases*, vol. 99, pp. 328–333, Oct. 2020, doi: 10.1016/J.IJID.2020.05.098.

[181] W. Song *et al.*, 'The role of SARS-CoV-2 N protein in diagnosis and vaccination in the context of emerging variants: present status and prospects', *Frontiers in Microbiology*, vol. 14, 2023, doi: 10.3389/FMICB.2023.1217567.
[182] 'Assessment of the Further Spread and Potential Impact of the SARS-CoV-2 Omicron Variant of Concern in the EU/EEA, 19th Update'. [Online]. Available: https://www.ecdc.europa.eu/en/publications-data/covid-19-omicron-risk-assessment-further-emergence-and-potential-impact
[183] N. I. for C. Diseases, 'The Daily Covid-19 Effective Reproductive Number (R) in South Africa'. 2021. [Online]. Available: https://www.nicd.ac.za/wp-content/uploads/2021/12/COVID-19-Effective-Reproductive-Number-in-South-Africa-week-51.pdf
[184] L. Wang and G. Cheng, 'Sequence analysis of the emerging SARS-CoV-2 variant Omicron in South Africa', *Journal of Medical Virology*, vol. 94, no. 4, pp. 1728–1733, Apr. 2022, doi: 10.1002/JMV.27516.
[185] Q. Yang, A. A. S. Syed, A. Fahira, and Y. Shi, 'Structural Analysis of the SARS-CoV-2 Omicron Variant Proteins', *Research*, vol. 2021, Jan. 2021, doi: 10.34133/2021/9769586.
[186] E. Boehm, I. Kronig, R. A. Neher, I. Eckerle, P. Vetter, and L. Kaiser, 'Novel SARS-CoV-2 variants: the pandemics within the pandemic', *Clinical Microbiology and Infection*, vol. 27, no. 8, pp. 1109–1117, Aug. 2021, doi: 10.1016/J.CMI.2021.05.022.
[187] C. rong Wu, W. chao Yin, Y. Jiang, and H. E. Xu, 'Structure genomics of SARS-CoV-2 and its Omicron variant: drug design templates for COVID-19', *Acta Pharmacologica Sinica*, vol. 43, no. 12, pp. 3021–3033, Dec. 2022, doi: 10.1038/S41401-021-00851-W.
[188] V. M. Ferré, N. Peiffer-Smadja, B. Visseaux, D. Descamps, J. Ghosn, and C. Charpentier, 'Omicron SARS-CoV-2 variant: What we know and what we don't', *Anaesthesia Critical Care and Pain Medicine*, vol. 41, no. 1, Feb. 2022, doi: 10.1016/J.ACCPM.2021.100998.
[189] J. Deerain *et al.*, 'Assessment of the Analytical Sensitivity of 10 Lateral Flow Devices against the SARS-CoV-2 Omicron Variant', *Journal of Clinical Microbiology*, vol. 60, no. 2, Feb. 2022, doi: 10.1128/JCM.02479-21.
[190] S. Stanley *et al.*, 'Limit of Detection for Rapid Antigen Testing of the SARS-CoV-2 Omicron and Delta Variants of Concern Using Live-Virus Culture', *Journal of Clinical Microbiology*, vol. 60, no. 5, May 2022, doi: 10.1128/JCM.00140-22.
[191] A. Rao *et al.*, 'Sensitivity of rapid antigen tests against SARS-CoV-2 Omicron and Delta variants', *Journal of Clinical Microbiology*, vol. 61, no. 10, Oct. 2023, doi: 10.1128/JCM.00138-23.
[192] K. Widyasari and S. Kim, 'Efficacy of novel SARS-CoV-2 rapid antigen tests in the era of omicron outbreak', *PLoS ONE*, vol. 18, no. 8 August, Aug. 2023, doi: 10.1371/JOURNAL.PONE.0289990.
[193] A. Osterman *et al.*, 'Impaired detection of omicron by SARS-CoV-2 rapid antigen tests', *Medical Microbiology and Immunology*, vol. 211, no. 2–3, pp. 105–117, Jun. 2022, doi: 10.1007/S00430-022-00730-Z.
[194] E. Schuit *et al.*, 'Diagnostic accuracy of covid-19 rapid antigen tests with unsupervised self-sampling in people with symptoms in the omicron period: Cross sectional study', *The BMJ*, 2022, doi: 10.1136/BMJ-2022-071215.
[195] G. Marais *et al.*, 'Improved oral detection is a characteristic of Omicron infection and has implications for clinical sampling and tissue tropism', *Journal of Clinical Virology*, vol. 152, Jul. 2022, doi: 10.1016/J.JCV.2022.105170.

[196] K. P. Y. Hui *et al.*, 'Replication of SARS-CoV-2 Omicron BA.2 variant in ex vivo cultures of the human upper and lower respiratory tract', *eBioMedicine*, vol. 83, Sep. 2022, doi: 10.1016/J.EBIOM.2022.104232.

[197] J. Lin *et al.*, 'Where is Omicron? Comparison of SARS-CoV-2 RT-PCR and Antigen Test Sensitivity at Commonly Sampled Anatomic Sites Over the Course of Disease.', *medRxiv : the preprint server for health sciences*, Feb. 2022, doi: 10.1101/2022.02.08.22270685.

[198] P. M. Bossuyt *et al.*, 'STARD 2015: An updated list of essential items for reporting diagnostic accuracy studies', *The BMJ*, vol. 351, Oct. 2015, doi: 10.1136/BMJ.H5527.

[199] J. Jumper *et al.*, 'Highly accurate protein structure prediction with AlphaFold', *Nature*, vol. 596, no. 7873, pp. 583–589, Aug. 2021, doi: 10.1038/S41586-021-03819-2.

[200] C. H. M. Rodrigues, D. E. V. Pires, and D. B. Ascher, 'DynaMut: Predicting the impact of mutations on protein conformation, flexibility and stability', *Nucleic Acids Research*, vol. 46, no. W1, pp. W350–W355, Jul. 2018, doi: 10.1093/NAR/GKY300.

[201] Y. Chen, H. Lu, N. Zhang, Z. Zhu, S. Wang, and M. Li, 'PremPS: Predicting the impact of missense mutations on protein stability', *PLoS Computational Biology*, vol. 16, no. 12 December, Dec. 2020, doi: 10.1371/JOURNAL.PCBI.1008543.

[202] R Foundation for Statistical Computing, 'R: A Language and Environment for Statistical Computing'. Vienna, Austria, 2022. [Online]. Available: https://www.R-project.org/

[203] 'Sorveglianza Integrata COVID-19: I Principali Dati Nazionali'. [Online]. Available: https://www.epicentro.iss.it/coronavirus/sars-cov-2-sorveglianza-dati

[204] Q. Ye, S. Lu, and K. D. Corbett, 'Structural Basis for SARS-CoV-2 Nucleocapsid Protein Recognition by Single-Domain Antibodies', *Frontiers in Immunology*, vol. 12, Jul. 2021, doi: 10.3389/FIMMU.2021.719037.

[205] C. Wu *et al.*, 'Characterization of SARS-CoV-2 nucleocapsid protein reveals multiple functional consequences of the C-terminal domain', *iScience*, vol. 24, no. 6, Jun. 2021, doi: 10.1016/J.ISCI.2021.102681.

[206] J. Cubuk *et al.*, 'The SARS-CoV-2 nucleocapsid protein is dynamic, disordered, and phase separates with RNA', *Nature Communications*, vol. 12, no. 1, Dec. 2021, doi: 10.1038/S41467-021-21953-3.

[207] Y. Zhou, H. Zhi, and Y. Teng, 'The outbreak of SARS-CoV-2 Omicron lineages, immune escape, and vaccine effectivity', *Journal of Medical Virology*, vol. 95, no. 1, Jan. 2023, doi: 10.1002/JMV.28138.

[208] A. Sharma, S. Balda, M. Apreja, K. Kataria, N. Capalash, and P. Sharma, 'COVID-19 Diagnosis: Current and Future Techniques', *International Journal of Biological Macromolecules*, vol. 193, pp. 1835–1844, Dec. 2021, doi: 10.1016/J.IJBIOMAC.2021.11.016.

[209] A. Isaacs *et al.*, 'Nucleocapsid Specific Diagnostics for the Detection of Divergent SARS-CoV-2 Variants', *Frontiers in Immunology*, vol. 13, Jun. 2022, doi: 10.3389/FIMMU.2022.926262.

[210] C. Y. Chu *et al.*, 'Performance of saliva and mid-turbinate swabs for detection of the beta variant in South Africa', *The Lancet Infectious Diseases*, vol. 21, no. 10, p. 1354, Oct. 2021, doi: 10.1016/S1473-3099(21)00405-9.

[211] Y.-P. Tu *et al.*, 'Swabs Collected by Patients or Health Care Workers for SARS-CoV-2 Testing', *New England Journal of Medicine*, vol. 383, no. 5, pp. 494–496, Jul. 2020, doi: 10.1056/NEJMC2016321.

[212] R. P. Venekamp *et al.*, 'Diagnostic accuracy of SARS-CoV-2 rapid antigen self-tests in asymptomatic individuals in the omicron period: a cross-sectional study', *Clinical*

Microbiology and Infection, vol. 29, no. 3, p. 391.e1-391.e7, Mar. 2023, doi: 10.1016/J.CMI.2022.11.004.

[213] K. Dishnica *et al.*, 'Novel insights into the somatic proteome of Strongyloides stercoralis infective third-stage larvae', *Parasites and Vectors*, vol. 16, no. 1, pp. 1–12, Dec. 2023, doi: 10.1186/S13071-023-05675-7/FIGURES/4.

[214] World Health Organization, 'Ending the Neglect to Attain the Sustainable Development Goals: A Road Map for Neglected Tropical Diseases 2021–2030'. 2021. [Online]. Available: https://www.who.int/publications/i/item/9789240010352

[215] F. Schär *et al.*, 'Strongyloides stercoralis: global distribution and risk factors', *PLoS Negl Trop Dis*, vol. 7, no. 7, 2013, doi: 10.1371/journal.pntd.0002288.

[216] A. Coghlan *et al.*, 'Comparative genomics of the major parasitic worms', *Nat Genet*, vol. 51, no. 1, pp. 163–174, Jan. 2019, doi: 10.1038/s41588-018-0262-1.

[217] T. B. Nutman, 'Human infection with Strongyloides stercoralis and other related Strongyloides species', *Parasitology*, vol. 144, no. 3, pp. 263–273, Mar. 2017, doi: 10.1017/s0031182016000834.

[218] F. Tamarozzi, S. S. Longoni, C. Mazzi, E. Rizzi, R. Noordin, and D. Buonfrate, 'The accuracy of a recombinant antigen immunochromatographic test for the detection of Strongyloides stercoralis infection in migrants from sub-Saharan Africa', *Parasit Vectors*, vol. 15, no. 1, p. 142, Dec. 2022, doi: 10.1186/s13071-022-05249-z.

[219] D. Buonfrate *et al.*, 'Prevalence of strongyloidiasis in a cohort of migrants in Italy and accuracy of a novel ELISA assay for S. stercoralis infection, a cross-sectional study', *Microorganisms.*, vol. 9, no. 2, p. 401, Feb. 2021, doi: 10.3390/microorganisms9020401.

[220] N. W. Anderson *et al.*, 'Comparison of three immunoassays for detection of antibodies to Strongyloides stercoralis', *Clin Vaccine Immunol*, vol. 21, no. 5, pp. 732–736, 2014, doi: 10.1128/cvi.00041-14.

[221] W. J. Sears and T. B. Nutman, 'Strongy detect: Preliminary validation of a prototype recombinant Ss-NIE/Ss-IR based ELISA to detect Strongyloides stercoralis infection', *PLoS Negl Trop Dis*, vol. 16, no. 1, Jan. 2022, doi: 10.1371/journal.pntd.0010126.

[222] F. Tamarozzi *et al.*, 'Diagnostic accuracy of a novel enzyme-linked immunoassay for the detection of IgG and IgG4 against Strongyloides stercoralis based on the recombinant antigens NIE/SsIR', *Parasit Vectors*, vol. 14, no. 1, p. 412, Dec. 2021, doi: 10.1186/s13071-021-04916-x.

[223] A. Marcilla *et al.*, 'Proteomic analysis of Strongyloides stercoralis L3 larvae', *Parasitology*, vol. 137, no. 10, pp. 1577–1583, Sep. 2010, doi: 10.1017/s0031182010000314.

[224] R. Rodpai *et al.*, 'Identification of antigenic proteins in Strongyloides stercoralis by proteomic analysis', *Parasitol Res*, vol. 116, no. 6, pp. 1687–1693, Jun. 2017, doi: 10.1007/s00436-017-5443-9.

[225] H. Soblik *et al.*, 'Life cycle stage-resolved proteomic analysis of the excretome/secretome from Strongyloides ratti–identification of stage-specific proteases', *Mol Cell Proteomics*, vol. 10, no. 12, Dec. 2011, doi: 10.1074/mcp.m111.010157.

[226] P. D. M. Fonseca *et al.*, 'Shotgun proteomics of Strongyloides venezuelensis infective third stage larvae: Insights into host-parasite interaction and novel targets for diagnostics', *Mol Biochem Parasitol*, vol. 235, Jan. 2020, doi: 10.1016/j.molbiopara.2019.111249.

[227] M. A. Corral *et al.*, 'Potential immunological markers for diagnosis of human strongyloidiasis using heterologous antigens', *Parasitology*, vol. 144, no. 2, pp. 124–130, Feb. 2017, doi: 10.1017/s0031182016001645.

[228] Y. Maeda *et al.*, 'Secretome analysis of Strongyloides venezuelensis parasitic stages reveals that soluble and insoluble proteins are involved in its parasitism', *Parasit Vectors*, vol. 12, no. 1, p. 21, Jan. 2019, doi: 10.1186/s13071-018-3266-x.

[229] J. B. Lok, 'Strongyloides stercoralis: a model for translational research on parasitic nematode biology', *WormBook.*, pp. 1–18, 2007, doi: 10.1895/wormbook.1.134.1.

[230] V. Dozio and J. C. Sanchez, 'Profiling the proteomic inflammatory state of human astrocytes using DIA mass spectrometry', *J Neuroinflammation*, vol. 15, no. 1, p. 331, Nov. 2018, doi: 10.1186/s12974-018-1371-6.

[231] J. Brandi *et al.*, 'Exploring the wound healing, anti-inflammatory, anti-pathogenic and proteomic effects of lactic acid bacteria on keratinocytes', *Sci Rep*, vol. 10, no. 1, p. 11572, Dec. 2020, doi: 10.1038/s41598-020-68483-4.

[232] S. Martinotti *et al.*, 'HMGB1 osteo-modulatory action on osteosarcoma SaOS-2 cell line: an integrated study from biochemical and -omics approaches', *J Cell Biochem*, vol. 117, pp. 2559–2569, Nov. 2016, doi: 10.1002/jcb.25549.

[233] Y. Perez-Riverol *et al.*, 'The PRIDE database and related tools and resources in 2019: Improving support for quantification data', *Nucleic Acids Res*, vol. 47, no. D1, pp. D442–D450, Jan. 2019, doi: 10.1093/nar/gky1106.

[234] S. McGinnis and T. L. Madden, 'Blast: at the core of a powerful and diverse set of sequence analysis tools', *Nucleic Acids Res*, vol. 32, no. WEB SERVER ISS., pp. W20-25, Jul. 2004, doi: 10.1093/nar/gkh435.

[235] A. Bateman *et al.*, 'Uniprot: the universal protein knowledgebase in 2021', *Nucleic Acids Res*, vol. 49, no. D1, pp. D480–D489, Jan. 2021, doi: 10.1093/nar/gkaa1100.

[236] M. A. Harris *et al.*, 'The gene ontology (GO) database and informatics resource', *Nucleic Acids Res*, vol. 32, no. DATABASE ISS., pp. D258-261, Jan. 2004, doi: 10.1093/nar/gkh036.

[237] D. Binns, E. Dimmer, R. Huntley, D. Barrell, C. O'Donovan, and R. Apweiler, 'QuickGo: a web-based tool for gene ontology searching', *Bioinformatics*, vol. 25, no. 22, pp. 3045–3046, Nov. 2009, doi: 10.1093/bioinformatics/btp536.

[238] M. Blum *et al.*, 'The InterPro protein families and domains database: 20 years on', *Nucleic Acids Res*, vol. 49, no. D1, pp. D344–D354, Jan. 2021, doi: 10.1093/nar/gkaa977.

[239] N. Zobayer, A. A. Hossain, and M. A. Rahman, 'A combined view of B-cell epitope features in antigens', *Bioinformation*, vol. 15, no. 7, pp. 530–534, Jul. 2019, doi: 10.6026/97320630015530.

[240] M. C. Jespersen, B. Peters, M. Nielsen, and P. Marcatili, 'BepiPred-2.0: improving sequence-based B-cell epitope prediction using conformational epitopes', *Nucleic Acids Res*, vol. 45, no. W1, pp. W24–W29, Jul. 2017, doi: 10.1093/nar/gkx346.

[241] T. Cokelaer, D. Pultz, L. M. Harder, J. Serra-Musach, J. Saez-Rodriguez, and A. Valencia, 'BioServices: a common Python package to access biological web services programmatically', *Bioinformatics*, vol. 29, no. 24, pp. 3241–3242, Dec. 2013, doi: 10.1093/bioinformatics/btt547.

[242] V. L. Hunt, I. J. Tsai, M. E. Selkirk, and M. Viney, 'The genome of Strongyloides spp. gives insights into protein families with a putative role in nematode parasitism', *Parasitology*, vol. 144, no. 3, pp. 343–358, Mar. 2017, doi: 10.1017/s0031182016001554.

[243] J. D. Stoltzfus, S. Minot, M. Berriman, T. J. Nolan, and J. B. Lok, 'RNAseq analysis of the parasitic nematode Strongyloides stercoralis reveals divergent regulation of canonical dauer pathways', *PLoS Negl Trop Dis*, vol. 6, no. 10, Oct. 2012, doi: 10.1371/journal.pntd.0001854.

[244] L. Xu, J. Yang, M. Xu, D. Shan, Z. Wu, and D. Yuan, 'Speciation and adaptive evolution reshape antioxidant enzymatic system diversity across the phylum Nematoda', *BMC Biol*, vol. 18, no. 1, p. 181, Dec. 2020, doi: 10.1186/s12915-020-00896-z.

[245] K. Henkle-Dührsen and A. Kampkötter, 'Antioxidant enzyme families in parasitic nematodes', *Mol Biochem Parasitol*, vol. 114, no. 2, pp. 129–142, 2001, doi: 10.1016/s0166-6851(01)00252-3.

[246] A. Marcilla *et al.*, 'The transcriptome analysis of Strongyloides stercoralis L3i larvae reveals targets for intervention in a neglected disease', *PLoS Negl Trop Dis*, vol. 6, no. 2, Feb. 2012, doi: 10.1371/journal.pntd.0001513.

[247] E. Pomari *et al.*, 'Identification of miRNAs of Strongyloides stercoralis L1 and iL3 larvae isolated from human stool', *Sci Rep*, vol. 12, no. 1, p. 9957, Dec. 2022, doi: 10.1038/s41598-022-14185-y.

[248] R. Varatharajalu, V. Parandaman, M. Ndao, J. F. Andersen, and F. A. Neva, 'Strongyloides stercoralis excretory/secretory protein strongylastacin specifically recognized by IgE antibodies in infected human sera', *Microbiol Immunol*, vol. 55, no. 2, pp. 115–122, Feb. 2011, doi: 10.1111/j.1348-0421.2010.00289.x.

[249] K. Donskow-ŁYsoniewska, M. Maruszewska-Cheruiyot, and M. Stear, 'The interaction of host and nematode galectins influences the outcome of gastrointestinal nematode infections', *Parasitology*, vol. 148, no. 6, pp. 648–654, May 2021, doi: 10.1017/s003118202100007x.

[250] C. Cantacessi and R. B. Gasser, 'SCP/TAPS proteins in helminths–where to from now?', *Mol Cell Probes*, vol. 26, no. 1, pp. 54–59, Feb. 2012, doi: 10.1016/j.mcp.2011.10.001.

[251] K. Pawłowski, 'Uncharacterized/hypothetical proteins in biomedical "omics" experiments: is novelty being swept under the carpet?', *Brief Funct Genomic Proteomic*, vol. 7, no. 4, pp. 283–290, 2008, doi: 10.1093/bfgp/eln033.

[252] Y. Sato, F. Inoue, R. Matsuyama, and Y. Shiroma, 'Immunoblot analysis of antibodies in human strongyloidiasis', *Trans R Soc Trop Med Hyg*, vol. 84, no. 3, pp. 403–406, 1990, doi: 10.1016/0035-9203(90)90337-e.

[253] N. S. Atkins, D. J. Conway, J. F. Lindo, J. W. Bailey, and D. A. P. Bundy, 'L3 antigen-specific antibody isotype responses in human strongyloidiasis: correlations with larval output', *Parasite Immunol*, vol. 21, no. 10, pp. 517–526, 1999, doi: 10.1046/j.1365-3024.1999.00248.x.

[254] D. J. Conway *et al.*, 'Serum IgG reactivity with 41-, 31-, and 28-kDa larval proteins of Strongyloides stercoralis in individuals with strongyloidiasis', *J Infect Dis*, vol. 168, no. 3, pp. 784–787, 1993, doi: 10.1093/infdis/168.3.784.

[255] R. Rodpai *et al.*, 'Strongyloides stercoralis diagnostic polypeptides for human strongyloidiasis and their proteomic analysis', *Parasitol Res*, vol. 115, no. 10, pp. 4007–4012, Oct. 2016, doi: 10.1007/s00436-016-5170-7.

[256] A. P. Sudré, R. C. Siqueira, M. G. M. Barreto, R. H. S. Peralta, H. W. Macedo, and J. M. Peralta, 'Identification of a 26-kDa protein fraction as an important antigen for application in the immunodiagnosis of strongyloidiasis', *Parasitol Res*, vol. 101, no. 4, pp. 1117–1123, Sep. 2007, doi: 10.1007/s00436-007-0596-6.

[257] L. Potocnakova, M. Bhide, and L. B. Pulzova, 'An introduction to B-cell epitope mapping and in silico epitope prediction', *J Immunol Res*, vol. 2016, p. 6760830, 2016, doi: 10.1155/2016/6760830.

[258] D. Ditgen *et al.*, 'Comparative characterization of two galectins excreted-secreted from intestine-dwelling parasitic versus free-living females of the soil-transmitted nematode

Strongyloides', *Mol Biochem Parasitol*, vol. 225, pp. 73–83, Oct. 2018, doi: 10.1016/j.molbiopara.2018.08.008.
[259] M. H. Yunus, N. Arifin, D. Balachandra, N. S. Anuar, and R. Noordin, 'Lateral flow dipstick test for serodiagnosis of strongyloidiasis', *Am J Trop Med Hyg*, vol. 101, no. 2, pp. 432–435, 2019, doi: 10.4269/ajtmh.19-0053.
[260] D. Balachandra *et al.*, 'A new antigen detection ELISA for the diagnosis of strongyloides infection', *Acta Trop*, vol. 221, Sep. 2021, doi: 10.1016/j.actatropica.2021.105986.
[261] L. N. Rascoe, C. Price, S. H. Shin, I. McAuliffe, J. W. Priest, and S. Handali, 'Development of Ss-NIE-1 recombinant antigen based assays for immunodiagnosis of strongyloidiasis', *PLoS Negl Trop Dis*, vol. 9, no. 4, Apr. 2015, doi: 10.1371/journal.pntd.0003694.
[262] M. F. Culma, 'Strongyloides stercoralis proteome: A reverse approach to the identification of potential immunogenic candidates', *Microb Pathog*, vol. 152, Mar. 2021, doi: 10.1016/j.micpath.2020.104545.
[263] T. G. Jaleta and J. B. Lok, 'Advances in the molecular and cellular biology of Strongyloides spp', *Curr Trop Med Rep*, vol. 6, no. 4, pp. 161–178, Dec. 2019, doi: 10.1007/s40475-019-00186-x.
[264] G. Benz, R. Roncalli, and S. Gross, 'Use of ivermectin in cattle, sheep, goats, and swine', in *Ivermectin and abamectin*, Springer, 1989, pp. 215–229.
[265] T. J Nolan and J. B Lok, 'Macrocyclic lactones in the treatment and control of parasitism in small companion animals', *Current pharmaceutical biotechnology*, vol. 13, no. 6, pp. 1078–1094, 2012.
[266] E. A. Ottesen and W. Campbell, 'Ivermectin in human medicine', *Journal of antimicrobial chemotherapy*, vol. 34, no. 2, pp. 195–203, 1994.
[267] E. A. Ottesen, B. Duke, M. Karam, and K. Behbehani, 'Strategies and tools for the control/elimination of lymphatic filariasis.', *Bulletin of the world Health Organization*, vol. 75, no. 6, p. 491, 1997.
[268] M. Turner and J. Schaeffer, 'Mode of action of ivermectin', in *Ivermectin and abamectin*, Springer, 1989, pp. 73–88.
[269] T. G. Geary *et al.*, 'Haemonchus contortus: ivermectin-induced paralysis of the pharynx', *Experimental parasitology*, vol. 77, no. 1, pp. 88–96, 1993.
[270] S. Bienert *et al.*, 'The SWISS-MODEL Repository—new features and functionality', *Nucleic acids research*, vol. 45, no. D1, pp. D313–D319, 2017.
[271] M. E. Bragina, A. Daina, M. A. Perez, O. Michielin, and V. Zoete, 'The SwissSimilarity 2021 web tool: novel chemical libraries and additional methods for an enhanced ligand-based virtual screening experience', *International Journal of Molecular Sciences*, vol. 23, no. 2, p. 811, 2022.
[272] V. Zoete, A. Daina, C. Bovigny, and O. Michielin, 'SwissSimilarity: a web tool for low to ultra high throughput ligand-based virtual screening'. ACS Publications, 2016.
[273] D. Bajusz, A. Rácz, and K. Héberger, 'Why is Tanimoto index an appropriate choice for fingerprint-based similarity calculations?', *Journal of cheminformatics*, vol. 7, pp. 1–13, 2015.
[274] D. R. Koes, M. P. Baumgartner, and C. J. Camacho, 'Lessons learned in empirical scoring with smina from the CSAR 2011 benchmarking exercise', *Journal of chemical information and modeling*, vol. 53, no. 8, pp. 1893–1904, 2013.

Abbreviations

3CLpro: 3C-like Protease

ACE2: Angiotensin Converting Enzyme 2

ADT: Antigen Diagnostic Tests

ARDS: Acute Respiratory Distress Syndrome

ASN: Asparagine

ASP: Aspartic Acid

BP: Biological Process

CC: Cellular Component

CTD: C-Terminal Domain

ECDF: Empirical Cumulative Density Function

ESP: Excretory-Secretory Product

FDA: Food And Drug Administration

GluCl: Glutammate gated chloride channel

GO: Gene Ontology

HNMT: Human Histamine N-Methyltransferase

ICTV: International Committee On Taxonomy of Viruses

IDRs: Intrinsically Disordered Regions

iL3: Infective Filariform Larvae

ILE: Isoleucine

IQR: Interquartile Range

IVM: Ivermectin

KDE: Kernel Density Estimate

L1: Rhabditiform Larva

LC–MS/MS: Liquid Chromatography-tandem Mass Spectrometry

LEU: Leucine

mAbs: Monoclonal Antibodies

MCA: Multiple Correspondence Analysis

MD: Molecular Dynamics

MERS-CoV: Middle East Respiratory Syndrome CoronaVirus

MF: Molecular Function

M^{pro}: Main protease

NTD: N-terminal Domain

ORFs: Open Reading Frames

PBS: Phosphate Buffered Saline

PCA: Principal Component Analysis

PDB: Protein Data Bank

PHE: Phenylalanine

PLIP: Protein–Ligand Interaction Profiler

PLpro: Papain-Like Protease

RBD: Receptor Binding Domain

RBM: Receptor Binding Motif

RDR: Recurrent Deletion Region

RGYR: Radius of Gyration

RIR1: Recurrent Insertion Region 1

RMSD: Root Mean Square Deviation of Backbone Beads

RMSF: Root Mean Square Fluctuations

ROS: Reactive Oxygen Species

RTC: Replication-Transcription Complex

SARS-CoV: Severe Acute Respiratory Syndrome Coronavirus

SD: Standard Deviations

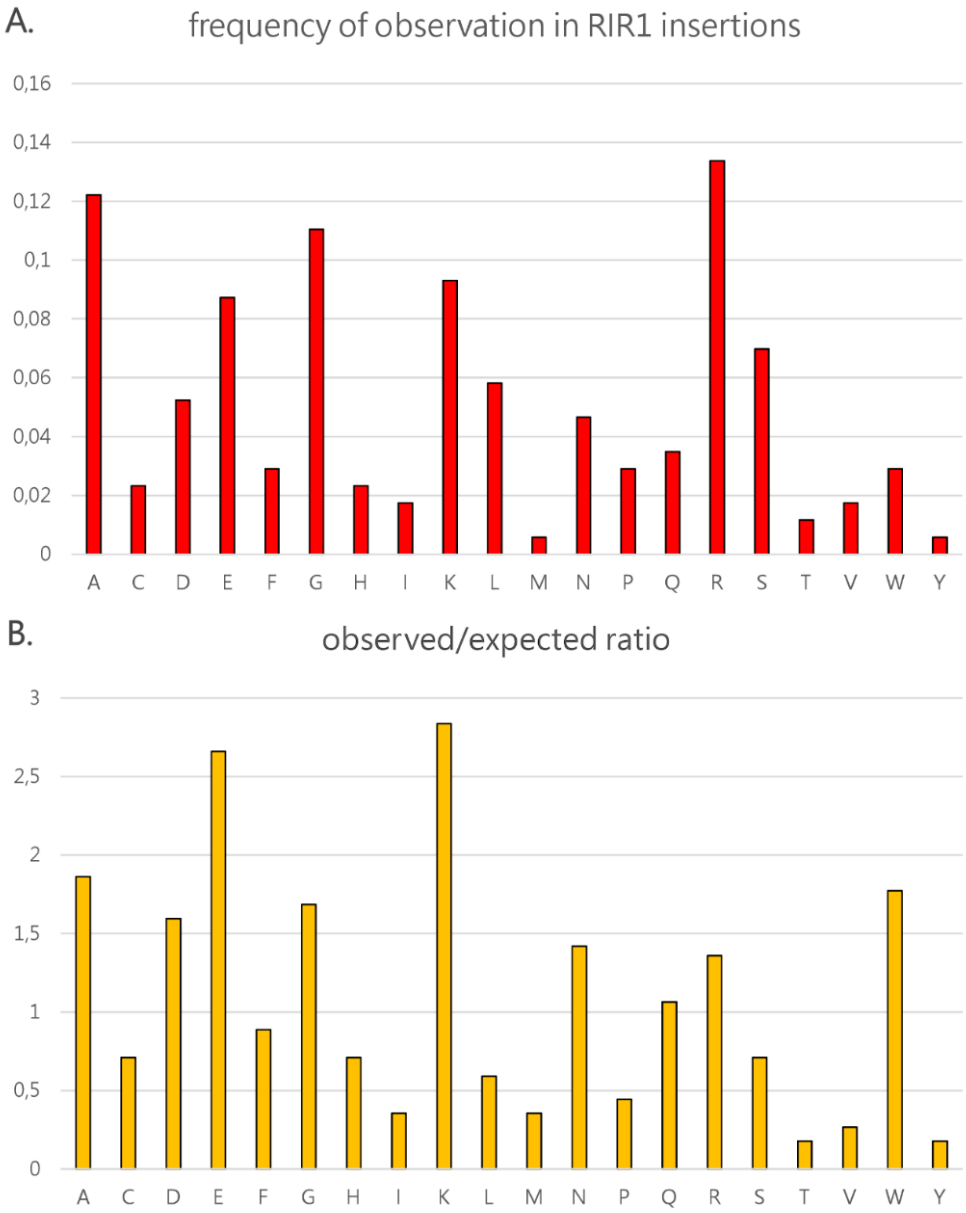
STH: Soil-Transmitted Helminth

VOC: Variant of Concern

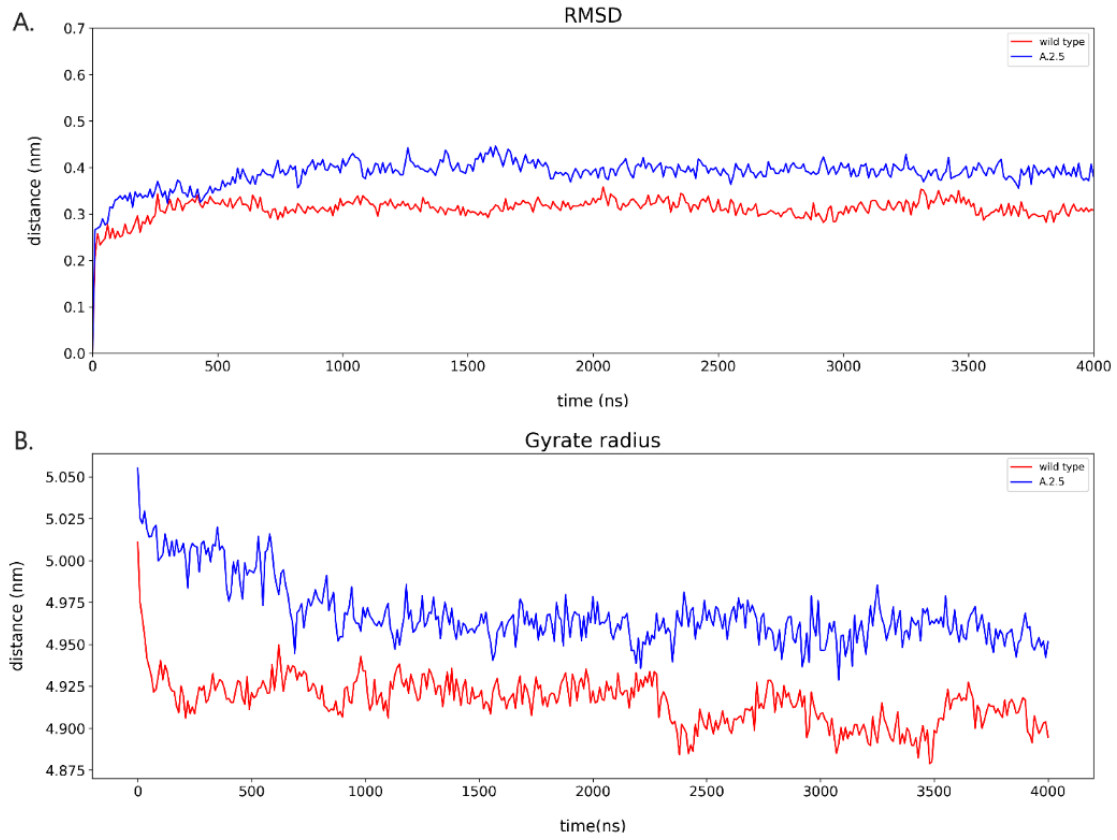
VOI: Variant of Interest

Supplementary materials

Supplementary material Chapter 3: Emergence of a recurrent insertion in the N-terminal domain of the SARS-CoV-2 spike glycoprotein



Supplementary Figure S3.1: Panel A*: observed frequency of each amino acid in the 49 RIR1 insertions.* ***Panel B****: observed/expected ratios for each amino acid, calculated based under the assumption that no codon usage bias was present. Amino acids showing a ratio > 1 were over-represented compared with expectations, whereas those showing a ratio < 1 were under-represented.*

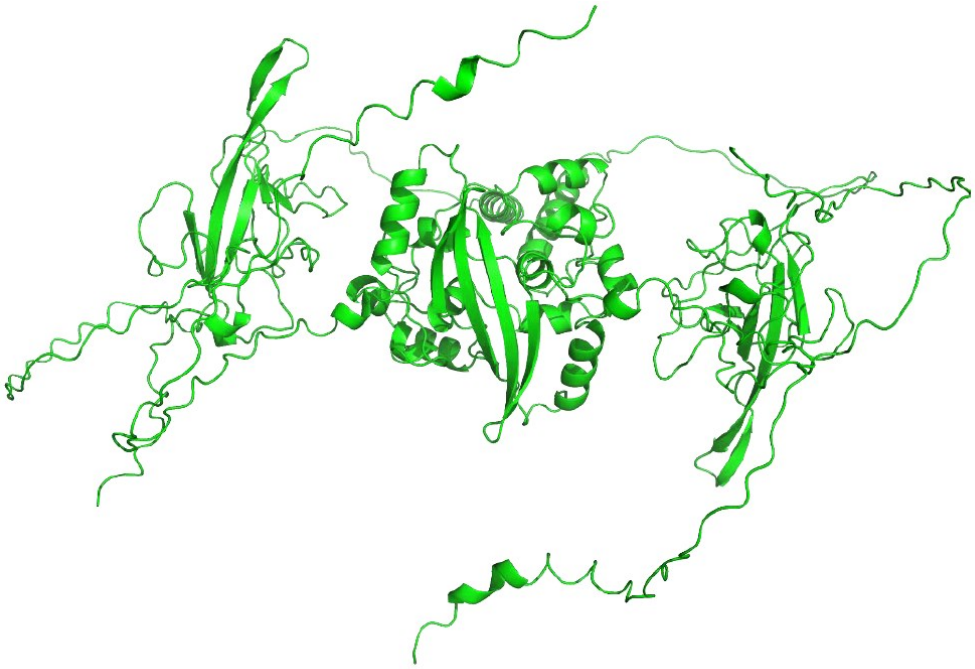


***Supplementary Figure S3.2**: Panel A: RMSD plot of the three models of the SARS-CoV-2 spike protein (wild-type and A.2.5), as a function of simulated time. The systems reach equilibrium after 1.5 µs. Panel B: variation of RGYR observed over time for the three models of the SARS-CoV-2 spike protein (wild-type and A.2.5) during the MD simulation.*

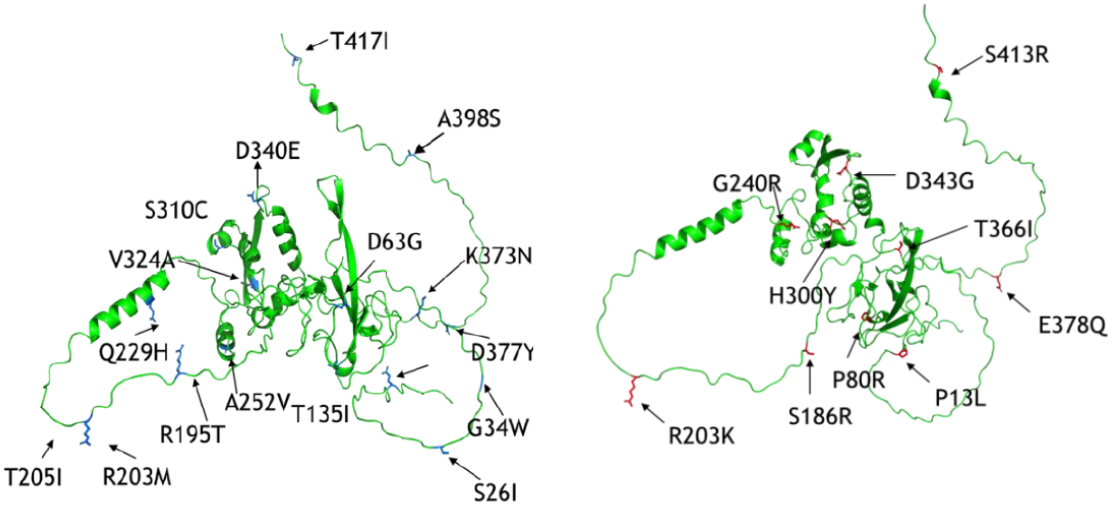
Supplementary material Chapter 4: An interaction-based drug discovery screen explains known SARS-CoV-2 inhibitors and predicts new compound scaffolds

***Supplementary Table S4.1:** Link to online material*

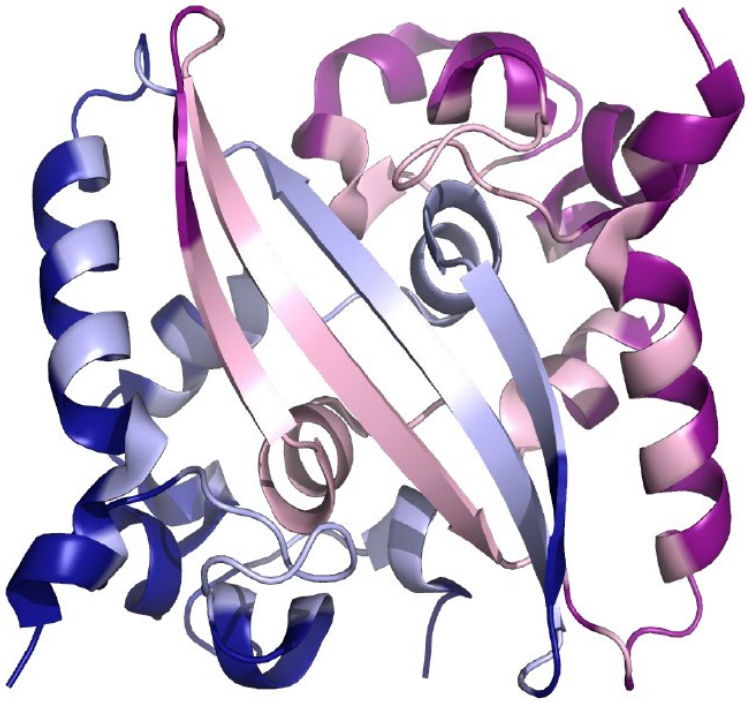
Supplementary material Chapter 5: Wide Real-Life Data Support Reduced Sensitivity of Antigen Tests for Omicron SARS-CoV-2 Infections



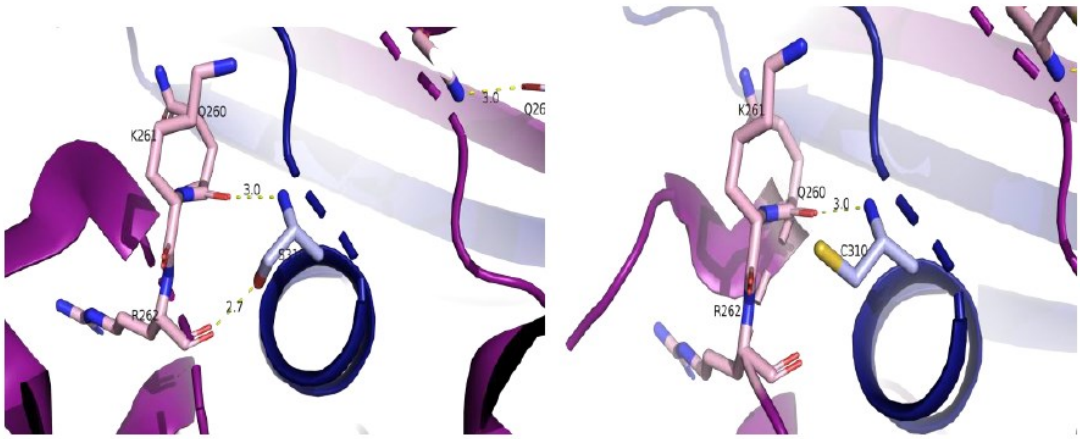
***Supplementary Figure S5.1**: Model of the full-length dimer N protein using AlphaFold2*



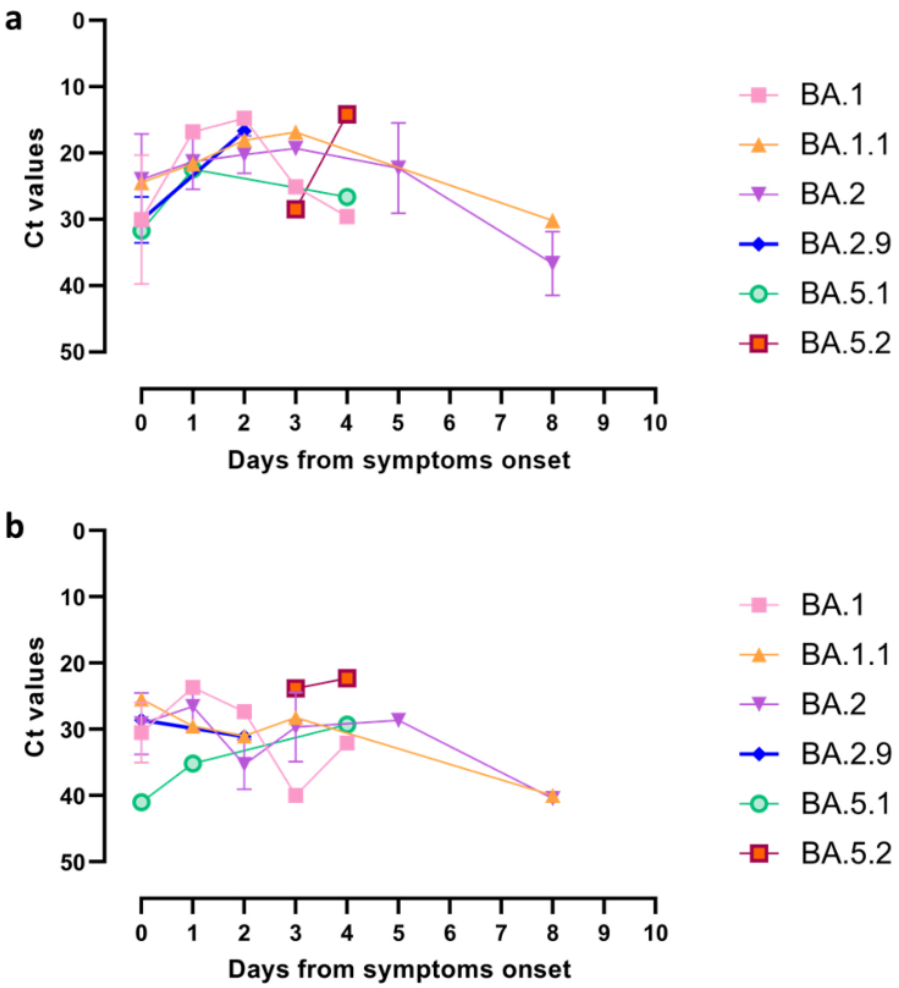
***Supplementary Figure S5.2**: Mutations mapped in the N monomer protein structure predicted by AlphaFold. The mutations are color-coded with blue representing those exclusive to the Delta variant, red representing those exclusive to the Omicron variant.*



Supplementary Figure S5.3: *CTD domain dimer of the N protein structure. Light pink: residues belonging to the dark blue chain (A). Light purple: residues belonging to the dark purple chain (B). CTD domain PDB ID: 6wzo.*



Supplementary Figure S5.4*: Zoom in showing the S310 mutated into Cysteine (S310C) and re-evaluation of the interactions after the substitution. On the left side, interactions occur involving the residue positioned at Ser310 with Gln 260, Lys 261, and Arg 262. On the right, there is a representation of amino acid substitution. We observed only minor difference of one less bond between Arginine 262 and Cysteine.*



***Supplementary Figure S5.5**: Mean Ct values and Standard Deviation for each SARS-CoV2 lineage detected in the nose (A) and in the mouth (B) based on days after symptoms onset.*

Mutant	PDB id	ΔΔG (kcal/mol) DynaMut2	ΔΔG (kcal/mol) DynaMut	ΔΔSvibENCoM (kcal.mol-1.K-1)
P80R	6vyo	0,41 kcal/mol	1,460 kcal/mol	-0.522 kcal.mol-1.K-1
H300Y	6wzo	1,16 kcal/mol	0.895 kcal/mol	-0,131 kcal.mol-1.K-1
S310C	6wzo	-0,43 kcal/mol	-0.217 kcal/mol	0.113 kcal.mol-1.K-1
D343G	6wzo	-0,28 kcal/mol	-0.958 kcal/mol	0.130 kcal.mol-1.K-1

***Supplementary Table S5.1**: M⊥⊥G and ⊥⊥SvibENCoM calculations using DynaMut and DynaMut2 webservers. The positive ⊥⊥G indicates increased stability, while negative ⊥⊥G indicates decreased stability. A negative ⊥⊥SVibENCoM implies an increase in protein rigidification, while a positive ⊥⊥SVibENCoM implies an increase in the flexibility of protein structure.*

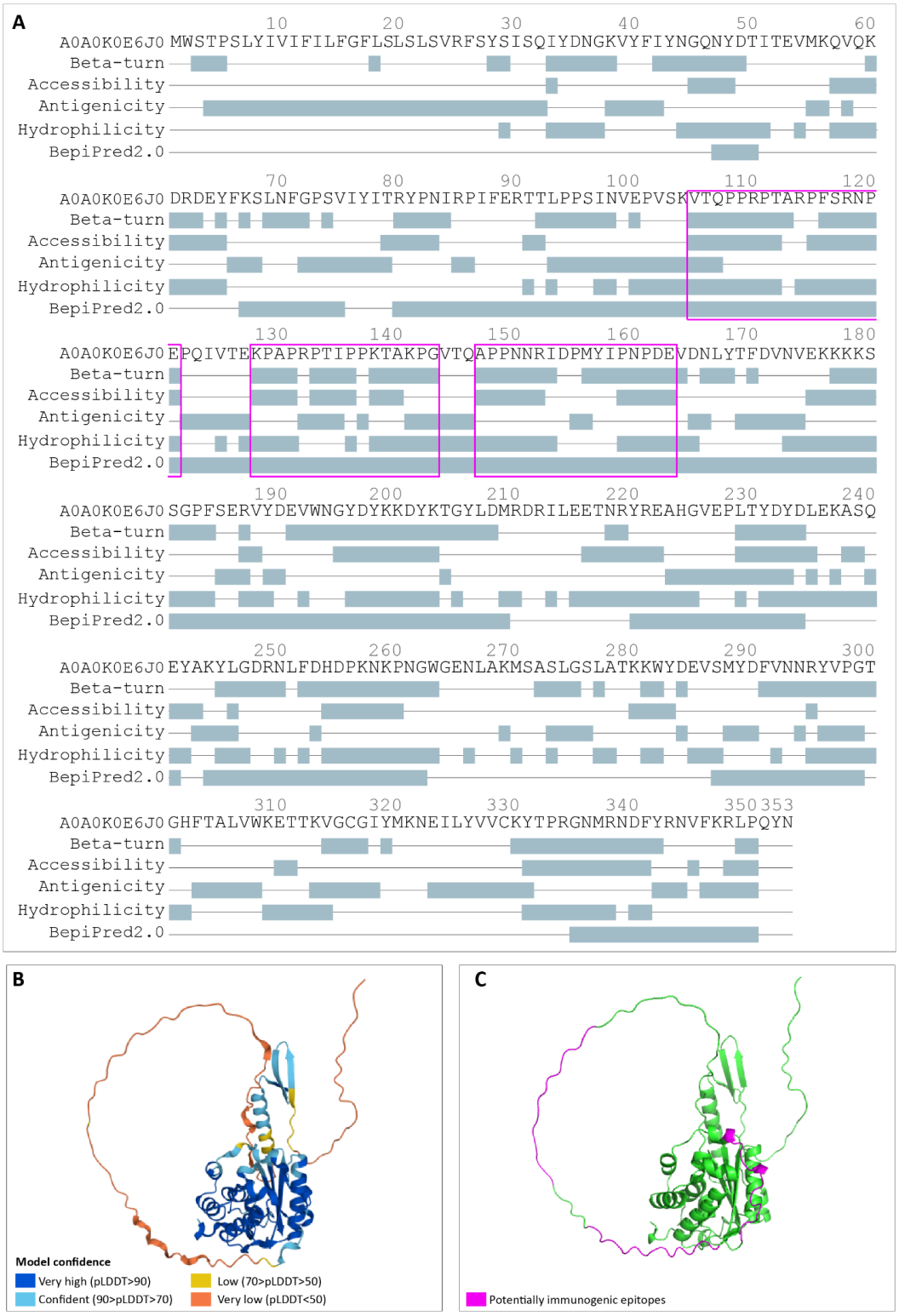
Supplementary material Chapter 6: Novel insights into the somatic proteome of Strongyloides stercoralis infective third-stage larvae

Supplementary Table S6.1: *Annotated dataset. The dataset includes peptide list, protein identification, gene ontology annotation and InterPro annotation.* *Link to online material*

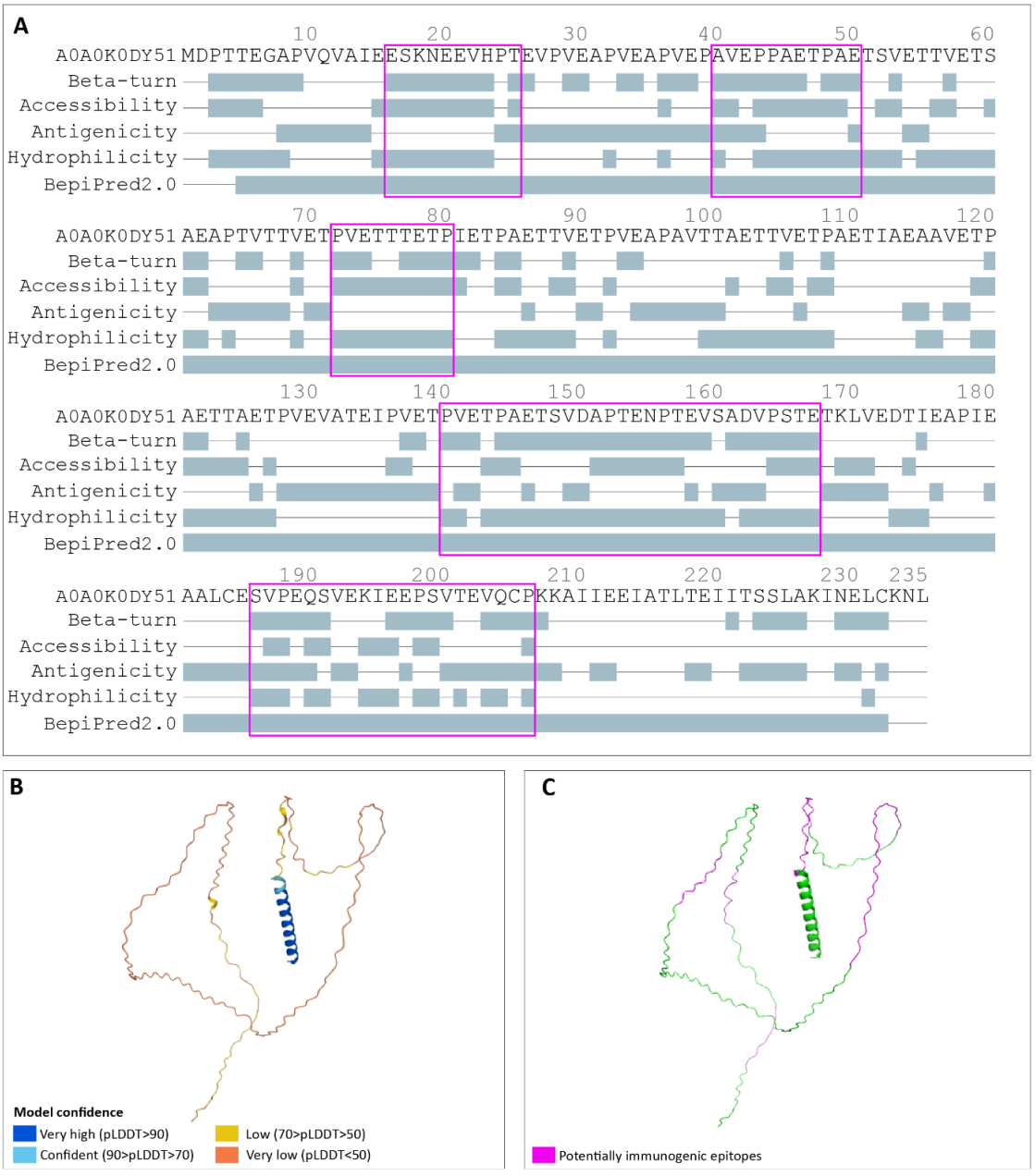
Supplementary Table S6.2: *Proteins identified in the present study and already reported in the literature as: (i) associated with Strongyloides parasitism; (ii) part of iL3 proteome; (iii) potentially immunogenic.* *Link to online material*

Supplementary Table S6.3: *Homology with Homo sapiens and other pathogens of clinical importance as potentially responsible for co-infections with S. stercoralis.* *Link to online material*

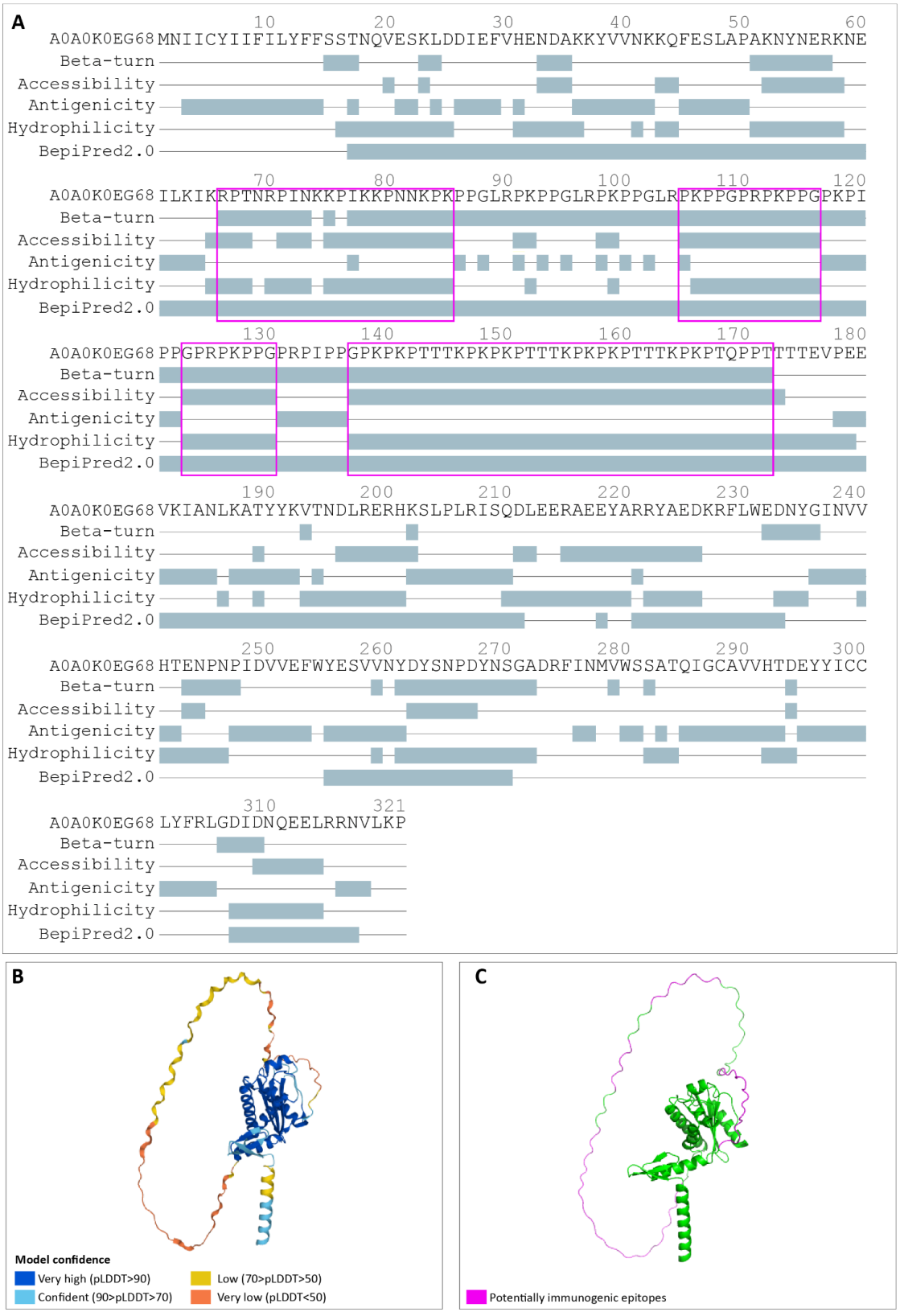
Supplementary Figures S6.1-S6.8: *B-cell epitope prediction results. For each figure: A) FASTA sequence showing the results obtained with each tool (Chou &Fasman Beta-Turn Prediction; Emini Surface Accessibility Prediction; Kolaskar & Tongaonkar Antigenicity; Parker Hydrophilicity Prediction, BepiPred2.0; all available via http://tools.iedb.org/bcell/); all residues having a score above their threshold are highlighted in grey. The purple squares indicate the sequences highlighted as potentially immunogenic as reported in the methods section. B) Protein structures as predicted by AlphaFold showing the model confidence. C) Mapping of the potentially immunogenic epitopes on the protein structure. Figure S1. B-cell epitope prediction resultsfor the protein A0A0K0E6J0 - SCP domain-containing protein.*



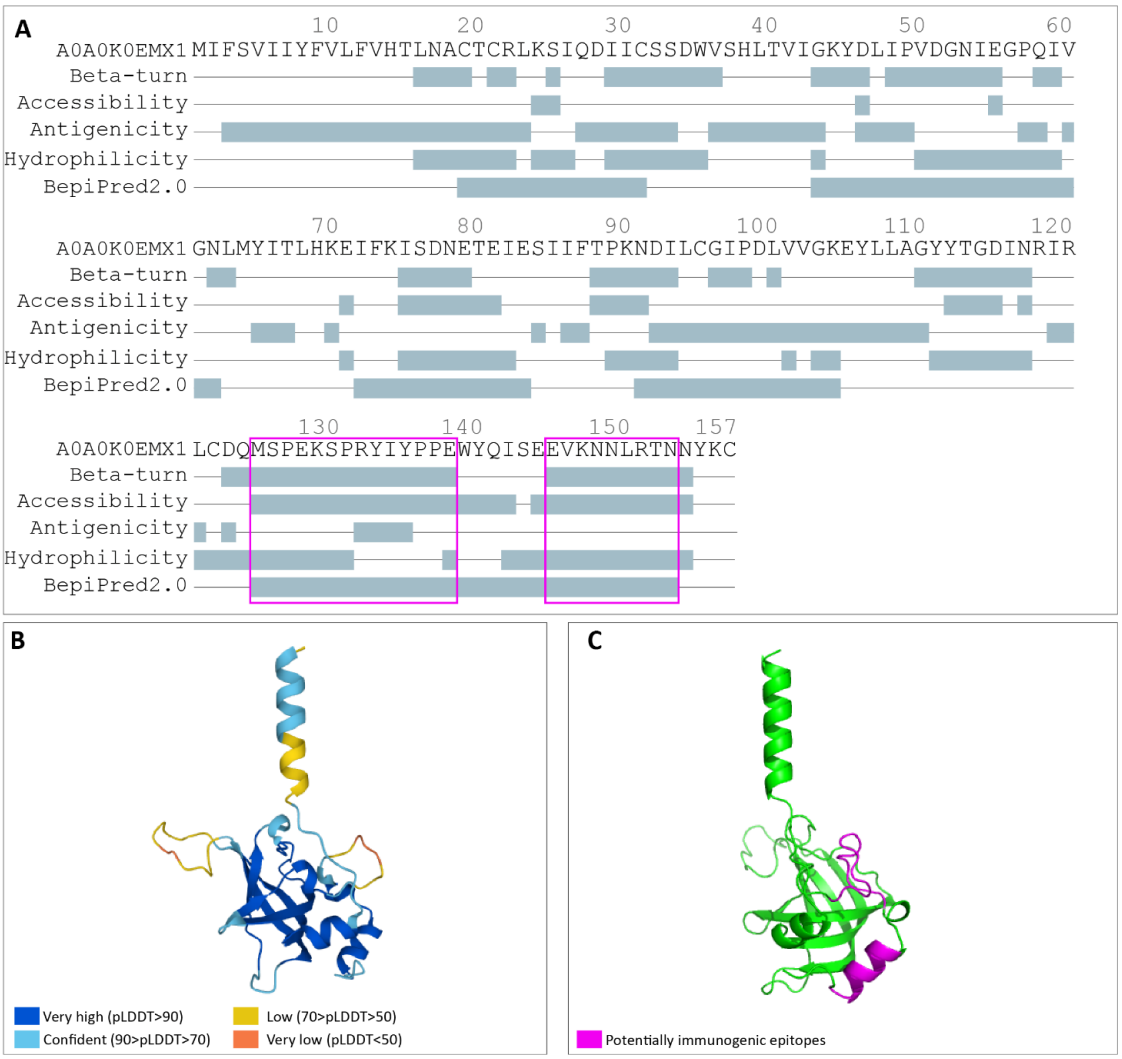
Supplementary Figures S6.1: *B-cell epitope prediction results for the protein A0A0K0E6J0 - SCP domain-containing protein.*



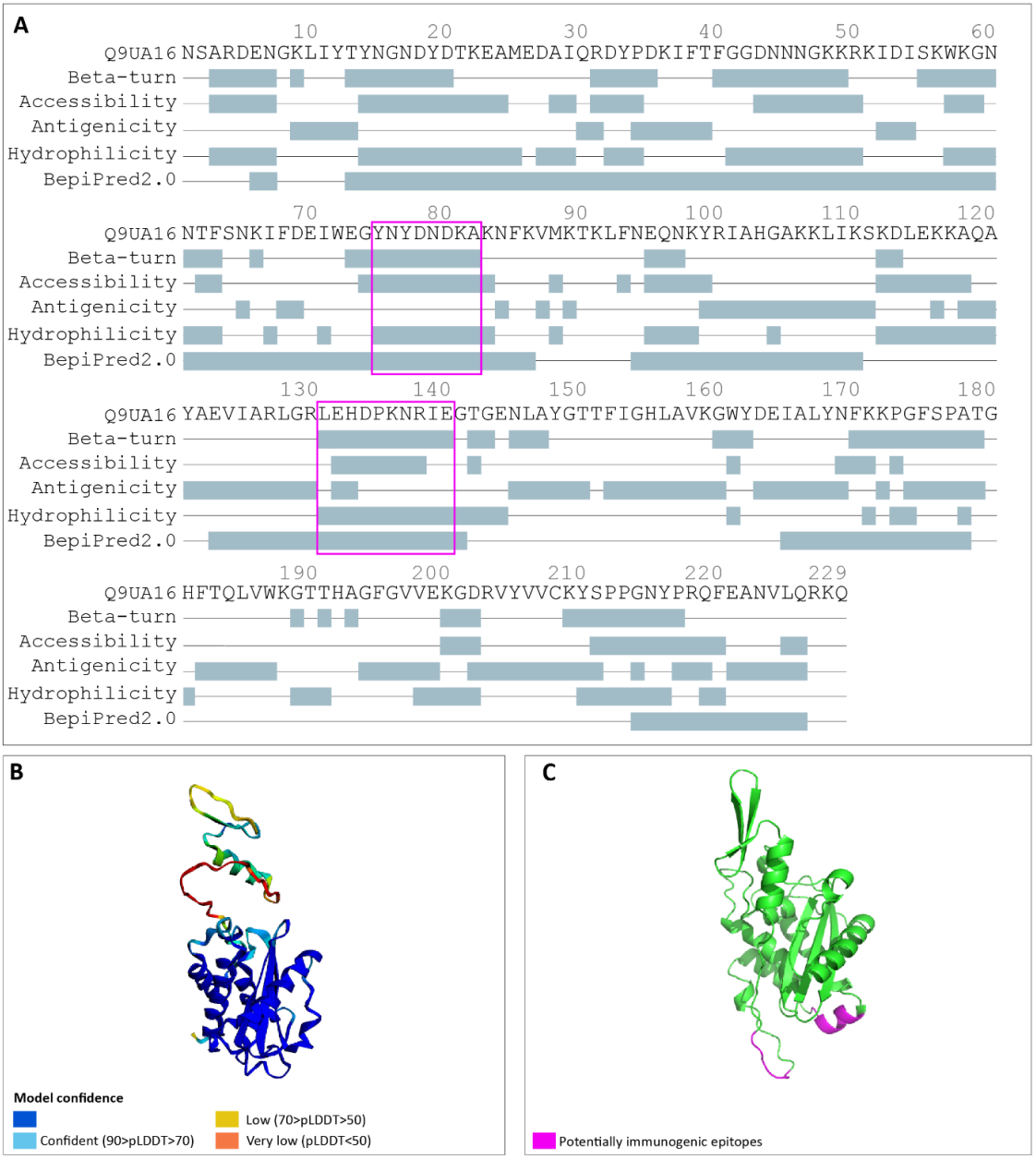
Supplementary Figures S6.2: *B-cell epitope prediction results for the protein A0A0K0DY51 - Uncharacterized protein.*



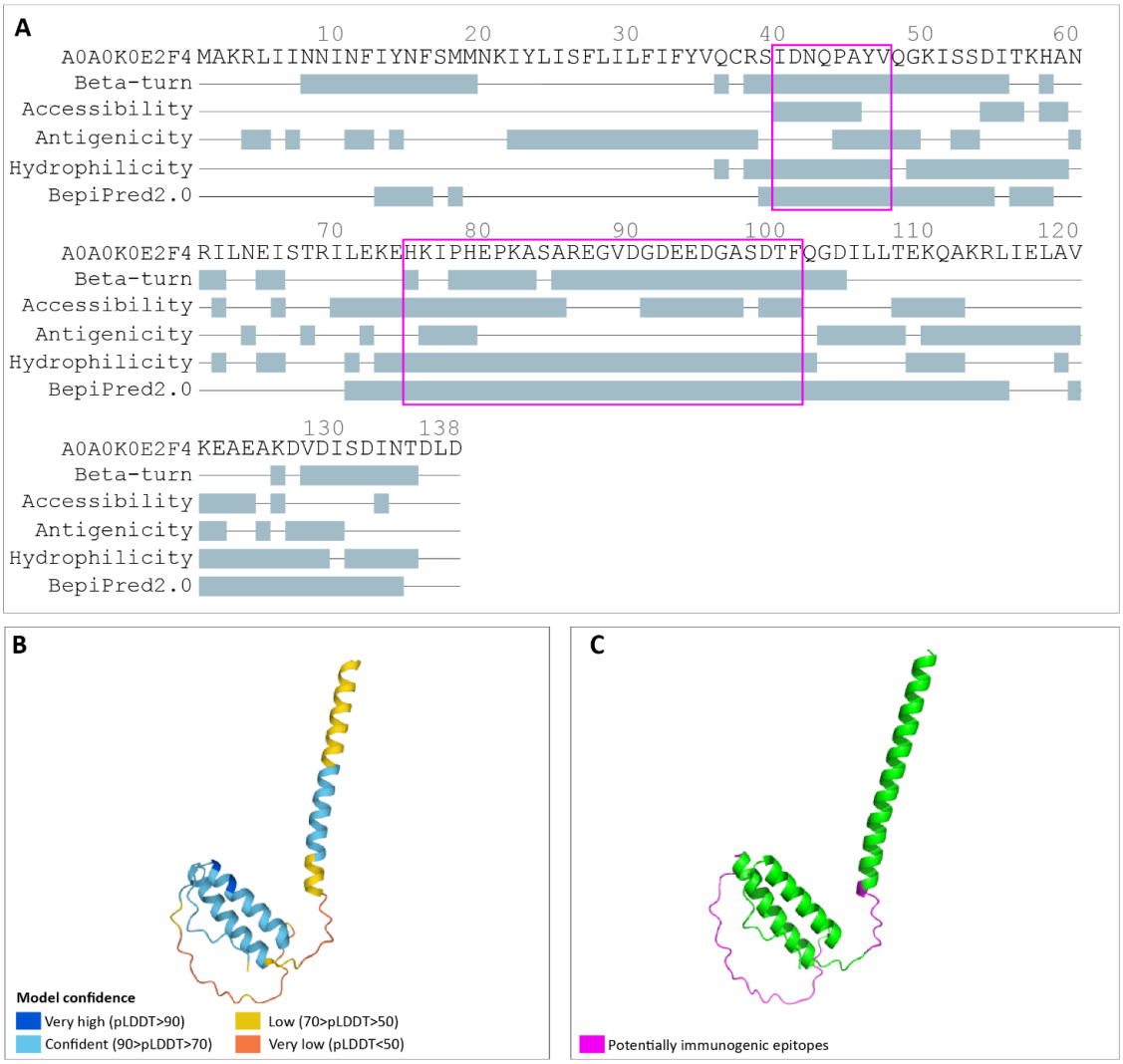
Supplementary Figures S6.3: *B-cell epitope prediction results for the protein A0A0K0EG68 - SCP domain-containing protein.*



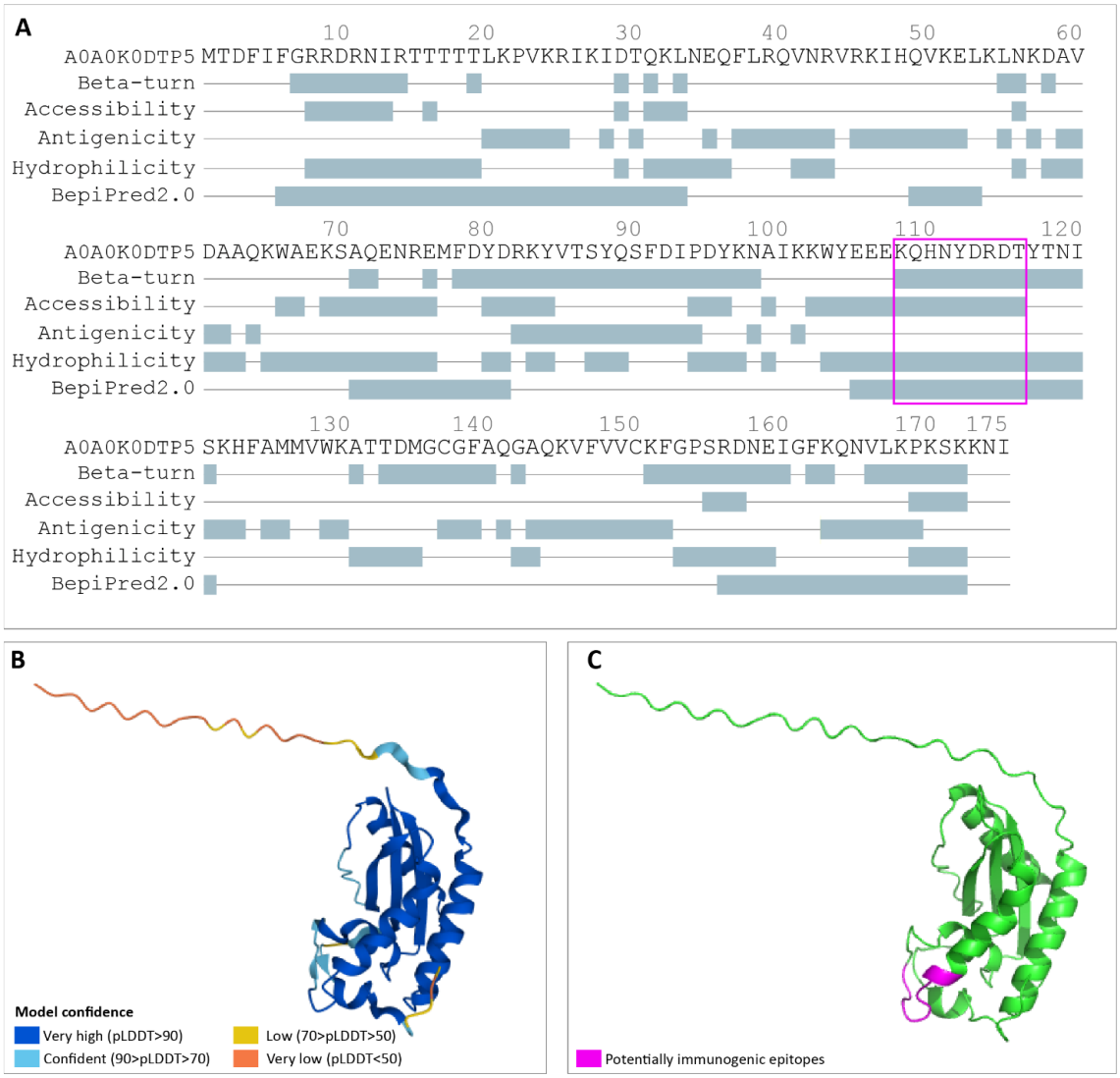
Supplementary Figures S6.4: *B-cell epitope prediction results for the protein A0A0K0EMX1 - NTR domain-containing protein.*



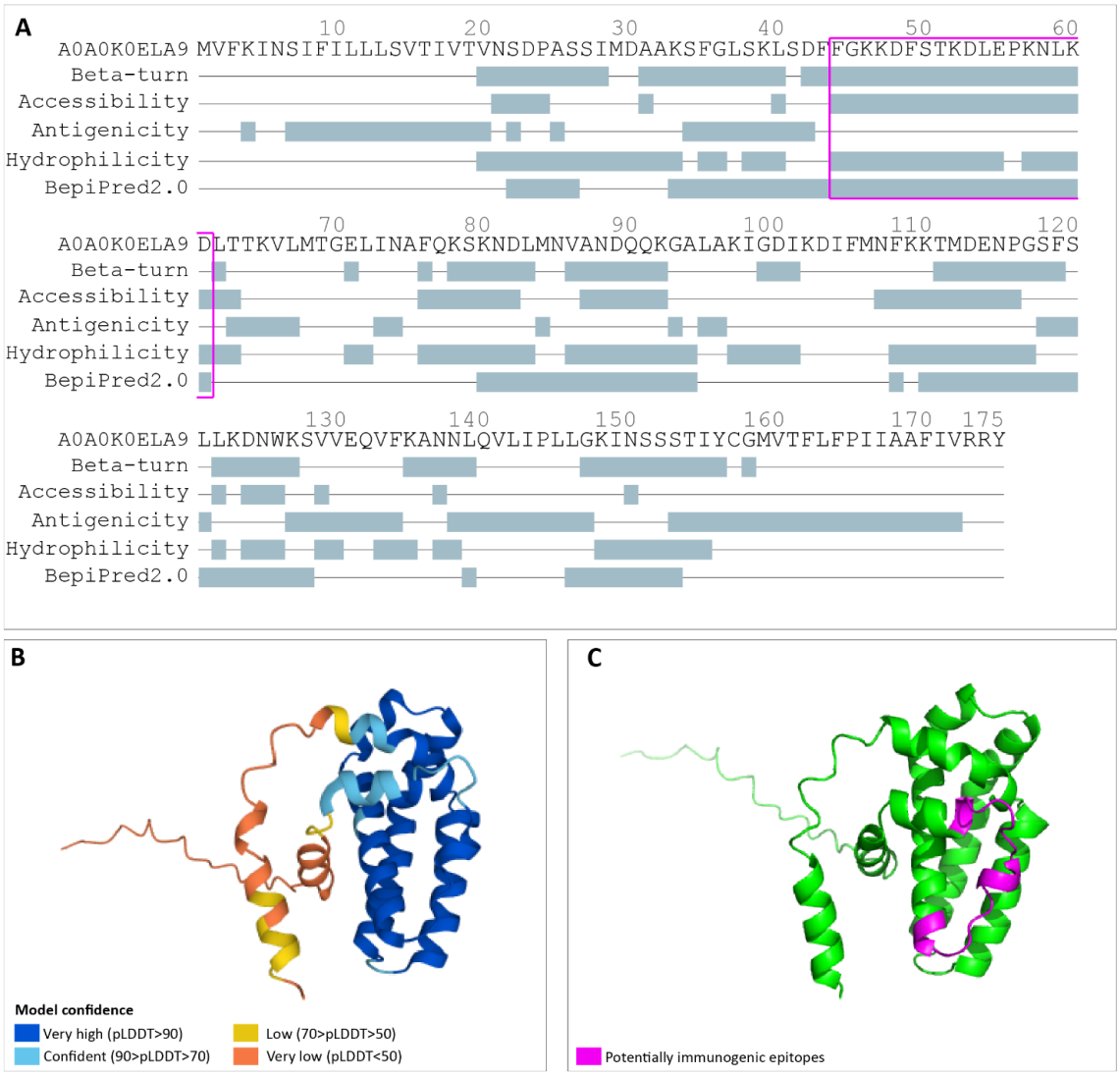
Supplementary Figures S6.5: *B-cell epitope prediction results for the protein Q9UA16 - L3NieAg.01.*



Supplementary Figures S6.6: *B-cell epitope prediction results for the protein A0A0K0E2F4 - Uncharacterized protein.*



Supplementary Figures S6.7: *B-cell epitope prediction results for the protein A0A0K0DTP5 - SCP domain-containing protein.*



Supplementary Figures S6.8: *B-cell epitope prediction results for the protein A0A0K0ELA9 - Uncharacterized protein.*