University of Kent School of Biosciences

MSc Degree by research in Computational Biology

Charlotte Thorpe

Oct 2021

(93 pages)

A comparative analysis of similar respiratory viruses to determine a cause for their differential phenotypes

Pandemic viruses have plagued humanity since records began. Recent years have seen viruses re-emerge from the past with high sequence identity to previous strains despite significant differences in virulence and pathogenicity. This research focuses on two different viruses. Firstly, investigating the determinants of pathogenicity in influenza A(H1N1) using a novel approach developed at the University of Kent to identify differentially conserved positions (DCPs). DCPs are specific positions within the virus protein that are one amino acid in group 1 and a different amino acid in group 2, meaning they have no structural purpose but may play a role in pathogenicity. This will be explored as it may explain how differences in pathogenicity can arise between related viruses. Secondly, this project considers the adaptation of SARS-CoV-2 to the serine protease drugs camostat and nafamostat by analysing sequence data provided by collaborators at Frankfurt University. N233S, T293I, and Q250P are three influenza A(H1N1) DCPs that were concluded to cause likely effects to the protein structure or function and therefore, may well be causing the differences in phenotypes and therefore the differences in disease severity. Furthermore, S50L, A222V, D614G, A653V, T732I and A879V are SARS-CoV-2 spike protein mutations that were found to cause likely effects to the protein structure or function, signifying they may contribute to resistance to camostat or nafamostat. This project has found several mutations of interest between closely related viruses that provides insight as to how mutations in viral genomes can cause differences in phenotypes.

Declaration

No part of this thesis has been submitted in support of an application for any degree or other qualification of the University of Kent, or any other University or Institution of learning.

Acknowledgements

I offer my gratitude to my research supervisors Professor Mark Wass and Professor Martin Michaelis of the University of Kent School of Biosciences. Despite a difficult year of online learning and zoom calls they made themselves available to me whenever I needed help or advice. I would like to thank them both for their constant support and encouragement which enabled me to develop my confidence in my scientific research and writing skills to produce a dissertation I am very proud of.

I would also like to thank Dr. Magdalena Antczak and Mr Jake McGreig for their support at the beginning of the year with coding in Python, and to Dr. Gary Thompson for running the training sessions. Your patience and reassurance helped me to master the basics and gave me the foundation I needed to get results from my research. A special thanks goes to my good friend and fellow researcher at the University of Kent, Miss Paige Policelli, for her endless encouragement throughout my time at Kent, without whom my accomplishments would not have been possible. Finally, I must thank my partner, Mr Tom Jefferis, and my family for providing me with their eternal love and admiration – I hope to make them proud.

Table of Contents

List of Figures

List of Tables

Abbreviations

SARS-CoV, severe acute respiratory syndrome associated coronavirus

SARS-CoV-2, severe acute respiratory syndrome associated coronavirus-2

MERS-CoV, Middle East respiratory syndrome associated coronavirus

WHO, World Health Organization

HPAI, highly pathogenic avian influenza

RNA, ribonucleic acid

NSP, non-structural protein

ORF, open reading frame

RBD, receptor binding domain

VAT, virus analysis tool

DCP, differentially conserved position

VLP, virus-like particle

RTC, replication and transcription complex

TMPRSS2, transmembrane protease serine 2

IL-6, interleukin 6

RdRp, RNA-dependent RNA polymerase

PBD, Protein Data Bank

Neu5Ac, N-acetylneuraminic acid

Neuc5Gc, N-glycolylneuraminic acid

CFR, case fatality rate

HA, haemagglutinin

NP, nucleoprotein

PB1, polymerase basic protein 1

PB2, polymerase basic protein 2

PA, polymerase acidic protein

NA, neuraminidase

MP, matrix protein

S-OIV, swine-origin influenza virus

S pro, spike protein

M pro, matrix protein

E pro, envelope protein

N pro, nucleocapsid protein

ACE2, angiotensin converting enzyme 2

ExoN, exonuclease

ER, endoplasmic reticulum

ARDS, acute respiratory distress syndrome

NCBI, National Center for Biotechnology Information

NBD, N-terminal binding domain

FCS, furin-like cleavage site

RNP, ribonucleoprotein

RBS, receptor binding site

SDP, specificity determining positions

Abstract

Pandemic viruses have plagued humanity since records began. Recent years have seen viruses re-emerge from the past with high sequence identity to previous strains despite significant differences in virulence and pathogenicity. This research focuses on two different viruses. Firstly, investigating the determinants of pathogenicity in influenza A(H1N1) using a novel approach developed at the University of Kent to identify differentially conserved positions (DCPs). DCPs are specific positions within the virus protein that are one amino acid in group 1 and a different amino acid in group 2, meaning they have no structural purpose but may play a role in pathogenicity. This will be explored as it may explain how differences in pathogenicity can arise between related viruses. Secondly, this project considers the adaptation of SARS-CoV-2 to the serine protease drugs camostat and nafamostat by analysing sequence data provided by collaborators at Frankfurt University. N233S, T293I, and Q250P are three influenza A(H1N1) DCPs that were concluded to cause likely effects to the protein structure or function and therefore, may well be causing the differences in phenotypes and therefore the differences in disease severity. Furthermore, S50L, A222V, D614G, A653V, T732I and A879V are SARS-CoV-2 spike protein mutations that were found to cause likely effects to the protein structure or function, signifying they may contribute to resistance to camostat or nafamostat. This project has found several mutations of interest between closely related viruses that provides insight as to how mutations in viral genomes can cause differences in phenotypes.

1 Introduction

1.1 Background of Respiratory Viruses

Respiratory viruses are one of the most common causative agents of human disease, with respiratory infection being a top five cause of mortality worldwide (Tyrrell, et al., 2017). Viral respiratory infections, such as those caused by influenza virus, occur in yearly endemics and are kept under constant surveillance alongside other seasonal respiratory viruses. The Influenza Surveillance Team at Public Health England's National Infection Service monitors influenza activity and publishes weekly activity updates throughout the winter months. It also has a role in monitoring emerging viruses such as MERS-CoV and several avian influenzas. Respiratory viruses tend to have a worse effect on those in the extremes of age as well as immunocompromised individuals due to decline in immune function and these cases are published by the EU-27's standardised death rate for respiratory diseases (Eurostat, 2020). In 2017, The World Health Organization (WHO) found that every year up to 650 000 deaths occur relating to respiratory infection from influenza virus (World Health Organization, 2017). These figures are particularly alarming given the marked increase in death compared with previous years. In addition to this, recent years have seen an increase in respiratory virus emergence which poses a threat to worldwide public health security.

In late 2019, a novel coronavirus emerged known as severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) which spread rapidly around the globe and gained pandemic status in March 2020. COVID-19 is the resulting disease from SARS-CoV-2 infection and, at the time of writing, has caused over 4.4 million deaths worldwide (Center for Systems Science

and Engineering (CSSE) , 2021). SARS-CoV-2 is also the third coronavirus to have emerged amongst humans since 2002 with the first being severe acute respiratory syndrome coronavirus (SARS-CoV), and the second being Middle Eastern respiratory syndrome-related coronavirus (MERS-CoV). Consequently, the devastation SARS-CoV-2 is causing worldwide has exposed cracks in the surface of society, leaving an overwhelming reminder of the fragile nature of humankind. Although, it is likely similar pandemic events will occur at increasing frequency in the future due to several factors including increased global travel, urbanisation, and climate change. Zoonotic diseases pose a global health threat due to the close connection between animals and people which would result in mass transmission. Zoonotic viruses have adapted to a specific host and are generally kept amongst that population; however, there are occasions where transmission to a human host occurs which poses a risk of human-to-human transmission. The highly pathogenic avian influenza (HPAI) A(H5N1) has been reportedly transmitted to humans in recent years with the WHO reporting 861 human cases since 2003 (World Health Organization, 2020). The main public health concern is the possibility that these viruses acquire an adaptive mutation that improves viral replication in the human respiratory tract, thus, providing it with a selective advantage.

Viruses must successfully undergo cellular and systemic virus-host interactions in order to replicate and cause disease. There are several host factors that determine the pathogenicity of a particular virus including: age at time of infection, route of infection, and cytokine induction (Rouse & Sehrawat, 2010). There are also virulence factors that interdepend on host susceptibility to spread and modify host defences. Furthermore, novel viruses often have newly acquired surface glycoproteins and are therefore, able to evade the host immune response. It is the combination of these factors that allows novel respiratory viruses to

become highly infectious and able to spread on a global scale in a matter of weeks. Many pandemics have occurred throughout human history including the "Spanish Flu" H1N1 pandemic of 1918-1919 which spread to one third of the world's population, resulting in an estimated 50 million deaths worldwide (Barry, 2005). The significance of this pandemic was in how deadly it was and the unusually high mortality rate for young adults. Despite scientists putting forward hundreds of possible explanations for the high mortality rate, there are still no answers to suggest a reason for this.

Given this, it was troubling to see the unexpected emergence of a pandemic H1N1 strain in 2009 known as "Swine Flu" which became the newest influenza pandemic in 41 years. Despite being a progeny of the 1918 H1N1 virus, the clinical features of the 2009 pandemic were milder and the mortality rate was much lower. This raises the question of why do two related viruses have such different phenotypes? The same question can be applied to several viruses throughout history including SARS-CoV with SARS-CoV-2 which forms part of my investigation as can be seen in later sections. With several examples of closely related viruses having very different phenotypes, the genetic reason for this must be investigated and comparisons made to find a cause for the differences in prognoses and virulence within the viral genome.

1.2 Influenza Virus A

1.2.1 Structure

Influenza A is one of four influenza viruses belonging to the family Orthomyxoviridae. This negative-sense RNA virus is separated into subtypes and categorised by an H number depending on the type of haemagglutinin and an N number depending on the type of

neuraminidase. Haemagglutinin and neuraminidase are two large glycoproteins found on the surface of the viral envelope each encoded for by segments 4 and 6 respectively. There are 18 known H subtypes and 11 known N subtypes but not all cause infection in humans. A total of eight RNA segments encode ten essential viral proteins in addition to several strain-dependent accessory proteins to make up each virion (Figure 1). Segment 1 encodes RNA polymerase subunit PB2, segment 2 encodes polymerase subunit PB1 and, segment 3 encodes RNA polymerase subunit PA and PA-X protein. PB2, PB1 and, PA together form the polymerase complex which is responsible for translation and transcription of the viral RNA. In some strains of influenza A, the PB1 gene encodes the small accessory protein PB1-F2 which is found in the +1 reading frame (Chen, et al., 2001). Segment 5 encodes a nucleoprotein which forms a major part of the ribonucleoprotein (RNP) complex of which associates with the polymerase complex to become transcriptionally active (Noda & Kawaoka, 2010). The M gene at segment 7 encodes two proteins: M1 which is a matrix protein and M2 which is a membrane protein. Finally, segment 8 encodes two non-structural proteins NS1 and NS2 (also known as nuclear export protein).

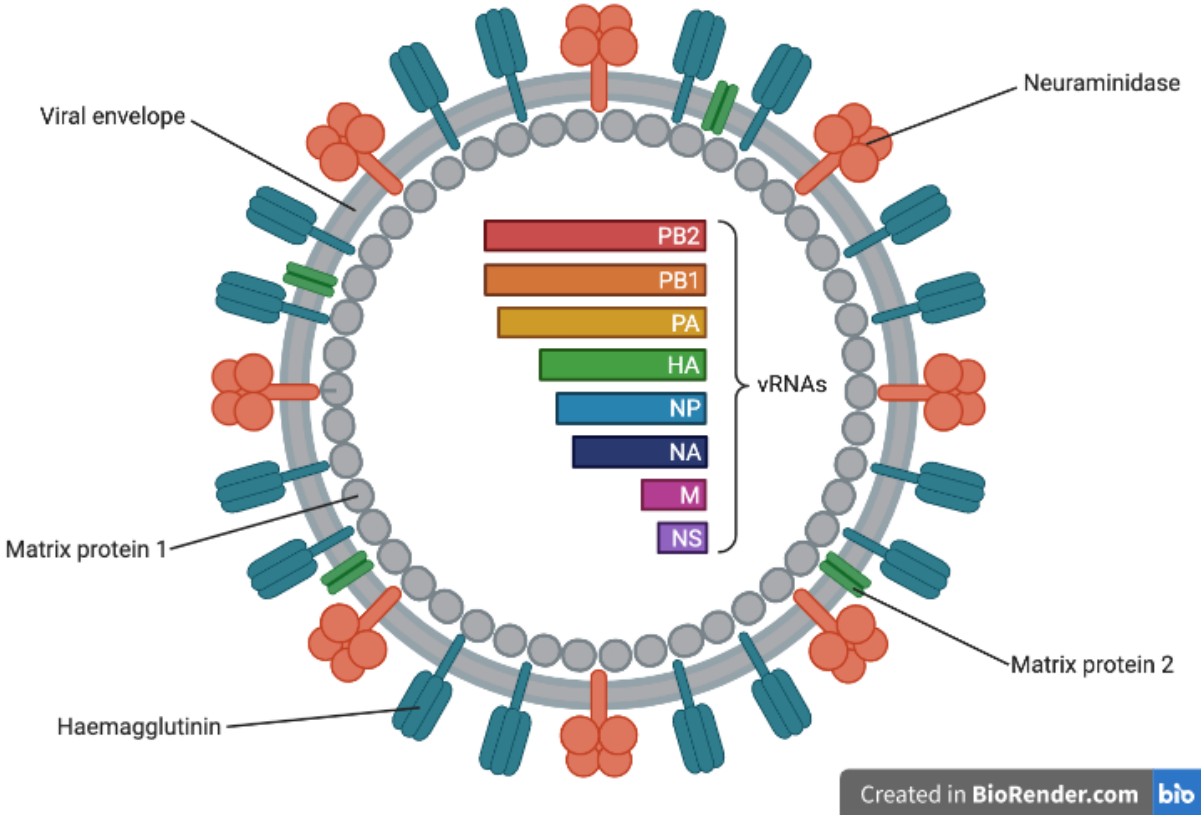


Figure 1: Structure of Influenza A Virus

Created with BioRender.com. In the centre of the virus are the eight segments of viral RNA (vRNA), each encoding a total of ten essential viral proteins and several strain-dependent accessory proteins. Matrix protein 1 surrounds the vRNAs and matrix protein 2 (green) is embedded into the viral envelope alongside neuraminidase (orange) and haemagglutinin (teal).

The surface glycoprotein haemagglutinin is coded by segment 4 of the viral RNA. It is a trimer of identical subunits with each monomer consisting of an HA0 polypeptide chain, along with membrane-distal HA1 and, membrane-proximal HA2 which are linked by a disulfide bridge (Boonstra, et al., 2018). HA1 engages sialic acid, a derivative of neuraminic acid which is widely expressed in higher vertebrates (Stencel-Baerenwald, et al., 2014). Neuraminic acid is often modified through acetylation to form *N*-acetylneuraminic acid (Neu5Ac) which can also be further hydroxylated to *N*-glycolneuraminic acid (Neu5Gc) (Stencel-Baerenwald, et al., 2014). HA1 binds most commonly to Neu5Ac which involves formation of α2,3-linked and α2,6-linked sialic acid attached to galactose, with avian influenza primarily binding α2,3-

linked sialic acid, and human influenza preferentially binding α2,6-linked sialic acid (Rogers, et al., 1983). When binding, Neu5Ac inserts deeply into the carbohydrate binding site of HA1 where two hydrogen bonds form between residues, and the glycerol and *N*-acetyl chain are contained in a hydrophobic pocket in the binding site (Stencel-Baerenwald, et al., 2014). Rotation around the glycosidic bond allows the galactose molecule to have a *cis* or *trans* conformation, with avian influenza virus haemagglutinin commonly bound in a *trans* conformation and human influenza virus commonly bound in a *cis* conformation - thus increasing its affinity for α-2,6-linked sialic acid (Stencel-Baerenwald, et al., 2014); (Xiong, et al., 2013).

1.2.2 Influenza replication

For viral pathogenesis to occur, the virus must successfully undergo cellular and systemic virus-host interactions to replicate itself and cause disease. HA1 from human influenza viruses preferentially binds α-2,6-linked sialic acid residues which are commonly found on human respiratory epithelial cells. This multivalent binding event is tight even though the individual affinity of each sialyl moiety to the HA binding site may be weak (Takemoto, et al., 1996). Protease cleaves haemagglutinin which causes the virus to be internalised into intracellular compartments either through the most common mechanism of clathrin-dependent endocytosis or through a clathrin- and caveolin- independent pathway (Stencel-Baerenwald, et al., 2014). These compartments, such as endosomes, have a low pH which triggers refolding of haemagglutinin and the resulting fusion of the viral envelope with the endosome membrane via a protein catalysed membrane fusion process (Lakadamyali, et al., 2004). HA refolding involves a conformational change through the release of the N-terminal

peptide of HA2 from the hydrophobic pocket enabling it to insert into the viral and target membrane (Gaudin, et al., 1995); (Leikina, et al., 2002).

Matrix protein 2 ion channels allow protons to move through the viral envelope and acidify the core of the virus, causing the core to disassemble and release viral RNA and core proteins. Viral RNA molecules, accessory proteins, and RNA-dependent RNA polymerase (RdRp) are released into the cytoplasm. Core proteins and viral RNA form a complex that gets transported into the cell nucleus where RdRp starts transcribing complementary positive-sense viral RNA. Viral RNA is either exported into the cytoplasm and translated, or it remains in the nucleus. Negative-sense viral RNAs, RdRp and other viral proteins are assembled into a virion. Newly synthesised viral proteins are either secreted through the Golgi apparatus onto the cell surface or transported back to the nucleus to bind viral RNA to form new genome particles (Dou, et al., 2018). The role of neuraminidase comes in the final stage of viral infection when it cleaves sialic acids from cellular receptors to prevent the virion aggregating consequently allowing the sudden rupture or gradual extrusion of virion progeny to spread to new target cells (Palese, et al., 1974).

1.2.3 Seasonal Influenza

Influenza spreads yearly with seasonal flu outbreaks often peaking between December and February, and can last as late as May. This cyclic occurrence is attributed to the antigenic variability of influenza A which occurs as antigenic drifts or antigenic shifts. Antigenic drifts are caused by mutations during replication occurring in haemagglutinin and neuraminidase genes. These mutations only cause minor variability and are usually responsible for the

seasonal flu endemics whereas, antigenic shift causes major variability which often leads to pandemics. The segmented genome of the Influenza A virus allows easy exchange of gene segments between viruses which can evade host immune response mechanisms as it produces novel antigens, known as gene reassortment (Chin, et al., 2016). It is now known that gene reassortment events occurred to produce the influenza A(H1N1)pdm09 virus in addition to A(H3N2) which dominated the 2017-2018 North American influenza season (Potter, et al., 2019).

Seasonal influenza is easily transmitted through droplets and direct contact resulting in an average of 25-50 million symptomatic cases in the United States alone each year with 20,000 of those resulting in death (Thompson, et al., 2004). Most cases are mild and managed symptomatically however, there is still a resounding number of deaths occurring despite there being a vaccine available. This can be put down to antigenic mismatching because of antigenic variability as well as uptake of the vaccine as this needs to occur annually. Although, emergence of new viruses can also lead to increased deaths as there is little to no pre-existing immunity in a population. Public Health England also monitors novel respiratory avian viruses including A(H7N9) and A(H5N6) both of which emerged in 2013, as well as A(H5N1) which emerged in 2003. As of August 2021, 1,568 cases of A(H7N9) have been reported in humans with a total of 616 deaths (CFR of 39.2%) and 863 cases of A(H5N1) have been reported with 456 deaths (CFR of 53%) (World Health Organization, 2021). Surveillance of these influenza viruses are vital as some subtypes are known to be highly pathogenic avian influenzas (HPAI) and if human-to-human transmission arises it is likely to result in a deadly pandemic, given the already high CFR's these subtypes are showing.

1.2.4 Comparison of Influenza Pandemics

Three influenza pandemics occurred in the 1900s with A(H1N1) in 1918, A(H2N2) in 1957 and A(H3N2) in 1968. Influenza pandemics tend to arise when an influenza virus spreading from human to human develops a new haemagglutinin molecule, however pandemic emergence remains poorly understood. Gene reassortment in influenza typically occurs when avian influenza virus swaps its genes with a human influenza virus, creating viruses with novel surface antigens that can spread in a human population (Glezen, 1996). Usually, avian influenzas are unable to infect humans since humans do not possess the α2,3-sialyllactose (NeuAc-2,3Gal) receptors that avian influenza viruses preferentially bind to (Rogers, et al., 1983). Mutations can alter the receptor binding specificity of avian viruses which may lead to increased human transmission therefore, allowing human-to-human transmission.

One study found that a mutation to the H1 gene in influenza at E190D left the virus capable of binding to avian and mammalian receptors, whereas if it possesses both E190D and D225G it can only bind to mammalian α2,6-linked sialic acid (Glaser, et al., 2005). There is evidence to suggest that gene reassortment events between avian influenza and human influenza occurred in both pandemic viruses A(H2N2) of 1957 and A(H3N2) of 1968 (Webster, et al., 1993). The 1957 H2N2 virus had both haemagglutinin and neuraminidase genes of avian origin, whereas the 1968 H3N2 virus had just avian haemagglutinin (Kawaoka, et al., 1989). It was because of these newly acquired surface glycoproteins that the viruses were able to escape herd immunity. Often when there is a gene reassortment event like this between avian and human influenzas, receptor binding specificity also changes from α2,3-linked to α2,6-linked sialic acid.

In 2009, the world saw the emergence of a novel H1N1 Swine-Origin Influenza Virus (S-OIV) which was first sequenced in April 2009. It is widely accepted that this pandemic virus emerged from a population in pigs from a small region in Mexico (Mena, et al., 2016). Phylogenetic analysis of this triple-reassortant virus has found it contains genes from avian (PB2 and PA), human H3N2 (PB1) and swine (HA, NP, and NS) lineages. It is thought to have emerged amongst the human population around January 2009 and that its polymerase genes alongside haemagglutinin (HA), nucleoprotein (NP) and the non-structural proteins (NS) genes emerged from an already circulating North American triple-reassortant virus found among swine population (Trifonov, et al., 2009). The neuraminidase (NA) and matrix protein (MP) genes are known to come from the Eurasian avian-like swine H1N1 influenza virus (Trifonov, et al., 2009); (Dawood, et al., 2009).

It was suggested that an intermediate host is required for genetic reassortment to take place and in the case of H1N1pdm09 the host was pigs (swine) since they can be infected by human and avian influenzas, having both SAα-2,3Gal and SAα-2,6Gal receptors (Ito, et al., 1998). Most cases of A(H1N1)pdm09 were mild and self-limiting with the total number of lab-confirmed deaths at 18,449 however this figure is likely to be much higher with estimates of around 200,000 (World Health Organization, 2011). This new H1N1 virus primarily infected children and young people, which is unlike many other influenza A infections because it spread rapidly and had a sudden onset which is dissimilar to other closely related viruses.

The 2009 pandemic highlighted a stark contrast to the 1918 H1N1 pandemic which is seen as the deadliest pandemic in recorded history. It is not clear how the 1918 “Spanish” flu began

and why it become so deadly. An estimated 500 million people may have been affected by the virus which at the time was one-third of the world's population (Frost, 1920). It is said to have originated in Haskell County, Kansas in January 1918 in an army camp from which it soon spread however this has been disputed (Barry, 2005). The pandemic came in three waves: the first wave occurring in March 1918 in North America, Asia, and Europe; the second wave being the deadliest and occurring in September to November 1918; the third wave emerging in early 1919. Each wave covered more locations than the previous until the entire globe was affected.

Typically, disease severity results from an interplay between host resistance and virulence of the virus. In the case of the 1918 flu, cytokine storm was a frequent occurrence where there is an uncontrolled release of proinflammatory cytokines which can cause multisystem organ failure and death due to the sudden release in large quantities (Peiris, et al., 2010). This phenomenon is thought to be a cause of the high death rate of the 1918 pandemic virus in younger individuals. There are many hypotheses as to why the 1918 pandemic was so deadly; some scientists say that a climate anomaly affecting migration of disease vectors increased the spread of the disease through bodies of water (More, et al., 2020). However, it could be that viral infection was no more aggressive than other influenza strains and that the poor hygiene, lack of antibiotics and overcrowded hospitals were to blame for the secondary bacterial superinfection which resulted in most of the deaths. Furthermore, advanced modern medical care and an effective public health strategy for dealing with pandemics is said to have prevented the 2009 pandemic from escalating to scales seen in 1918.

1.3 Coronaviruses

1.3.1 Structure

Coronaviruses are a group of related respiratory viruses from the sub-family Orthocoronavirinae that can cause illness such as Middle East respiratory syndrome (MERS), severe acute respiratory syndrome (SARS) and COVID-19. They are positive sense single stranded RNA viruses that can infect mammals and avian species depending on the genera of which there are four: alphacoronavirus, betacoronavirus, gammacoronavirus and deltacoronavirus. Alphacoronavirus and betacoronavirus exclusively affect mammals whereas, gammacoronavirus and deltacoronavirus affect a wider range of hosts including birds. Human coronaviruses have existed in the population for many decades and can cause seasonal respiratory infections alongside influenza virus however, these are mild when compared with the recent SARS-CoV, MERS-CoV and SARS-CoV-2 which are highly pathogenic. Human coronaviruses alone amount to 15-30% of all annual respiratory tract infections (Fehr & Perlman, 2015).

The coronavirus genome encodes five major open reading frames (ORFs): a 5'-leader-UTR-replicase (ORF1ab), the spike (S) protein, the membrane (M) protein, the envelope (E) protein, and the aforementioned nucleocapsid (N) protein (Figure 2). ORF1ab occupies two-thirds of the genome and encodes the replicase polyprotein 1ab (pp1ab). In turn, pp1ab is divided into 15-16 non-structural proteins (NSPs) which forms most of the viral replication and transcription complex (RTC) (V'kovski, et al., 2020). There are also several accessory proteins which are not essential for replication to occur but are known to have a role in pathogenicity (Zhao, et al., 2012). They have the largest genome among all RNA viruses ranging from 27 to

32 kb, which gets packaged by a helical capsid by nucleocapsid protein (N) and further encapsulated by an envelope (Brian & Baric, 2005). The M protein and E protein are involved in virus assembly and organisation, and the S protein mediates entry into host cells.

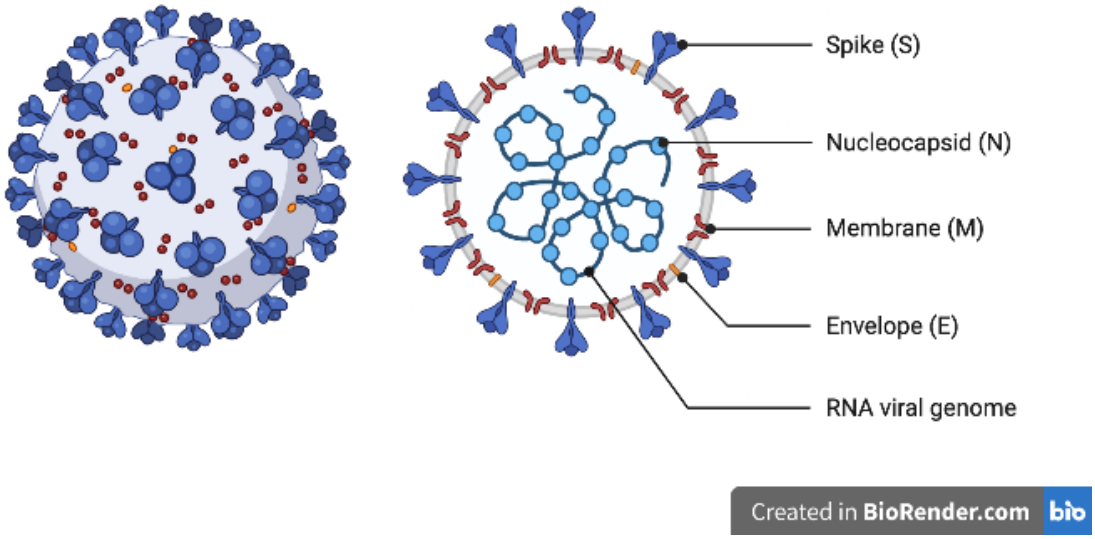


Figure 2: Human Coronavirus Structure

Created with BioRender.com. In the centre of the virus is the RNA viral genome and nucleocapsid protein which is surrounded by the viral envelope. The membrane protein (red) and the envelope protein (orange) are embedded in the viral envelope alongside the spike protein (purple) which protrudes out of the virus.

1.3.2 Coronavirus replication

The S protein is composed of two subunits: S1 and S2. The S1 subunit contains the receptor binding domain (RBD) and forms the head of the spike and the S2 subunit forms the stem which provides anchorage to the envelope (V'kovski, et al., 2020). To infect a human host, the spike protein must initially bind to the cellular entry receptors which in the case of SARS-CoV and SARS-CoV-2 is angiotensin-converting enzyme 2 (ACE2). Acid-dependent proteolytic cleavage of the S protein under the action of proteases, including cathepsin family and transmembrane protease serine 2 (TMPRSS2), allows the viral membrane to fuse with the

host cell membrane (Fehr & Perlman, 2015). TMPRSS2 is expressed in the human respiratory tract which is how SARS-CoV, and SARS-CoV-2 can spread easily from human-to-human. Cleavage of the S protein occurs in two stages with the first resulting in separation of the RBD and fusion domains, and the second exposing the fusion peptide which inserts into the membrane. A bundle is formed which allows the ultimate release of viral genome into the host cytoplasm.

Upon entry the two overlapping ORFs, ORF1a and ORF1b, can be directly translated by the host cell ribosomes to produce polyproteins, pp1a and pp1ab. Pp1ab is a result of a -1 ribosomal frameshift caused by utilisation of a slippery sequence and an RNA pseudoknot at the end of pp1a which allows for the continuous translation of both ORF1a and ORF1b (Khrustalev, et al., 2020). The polyproteins contain their own proteases located within NSP3 (PLpro) and NSP5 (3CLpro) which cleave and release the 16 individual NSPs. Some of the NSPs coalesce to form a replicase-transcriptase complex (RTC) that allows RNA synthesis to occur with the aid of specific NSP enzymes: RNA-dependent RNA polymerase (RdRp) encoded for by NSP12, RNA helicase encoded for by NSP13, the exoribonuclease (ExoN) encoded for by NSP14, and 2'-O-methyltransferase encoded for by NSP16 (Boopathi, et al., 2020); (V'kovski, et al., 2020). Following replication and RNA synthesis, the structural proteins S, M and E are translated and inserted into the endoplasmic reticulum (ER) where they move along the secretory pathway into the Golgi intermediate compartment. The M protein controls virus assembly but needs to be expressed alongside E protein to form virus-like particles (VLPs). S protein get incorporated into the virion and the M protein binding the nucleocapsid marks the end of virion assembly. Secretory vesicles release the progeny viruses via exocytosis into the cytoplasm where they can go on to infect other host cells.

1.3.3 Comparison of SARS, MERS, and COVID-19

The past two decades has seen three highly pathogenic coronaviruses emerge all resulting in significant loss of life with the former, SARS-CoV-2, bringing about economic crisis. In late 2002, SARS began spreading rapidly around the world having emerged in Guangdong Province, China. The World Health Organization estimates the number of SARS cases from 1 November 2002 to 31 July 2003 to be 8096 with 774 deaths which makes the case fatality rate (CFR) 9.6%; other estimates have put the CFR at 11% (World Health Organization, 2015); (Chan-Yeung & Xu, 2003). MERS emerged in the Middle East in 2012 and as of June 2021 a total of 2574 cases have been confirmed with 886 deaths, giving a CFR of 34.4% (World Health Organization, 2021). Finally, COVID-19 was identified to have emerged in Wuhan, China in December 2019 and spread rapidly worldwide causing a pandemic. At the time of writing, there have been over 209 million confirmed cases of COVID-19 (World Health Organization, 2021). The levels of human devastation caused by these human coronaviruses are immeasurable. When compared, SARS-CoV (R_0 of 3) has a lower transmissibility than SARS-CoV-2 (R_0 ranging from 1.8 to 3.6) (Liu, et al., 2020) but SARS-CoV-2 has a lower mortality rate (10.8% to 4.6% respectively) (Caldaria, et al., 2020) and more patients required hospitalisation with SARS-CoV. Despite these differences, the SARS-CoV and SARS-CoV-2 genome shares around 80% sequence identity (Abdelrahman, et al., 2020), with one strain of SARS-CoV having a homological similarity of 99.8% with SARS-CoV-2 (Song, et al., 2005).

As mentioned previously, there is a strain of SARS-CoV derived from palm civets which is very similar to SARS-CoV-2. This is suggestive that this strain, in particular, can switch between a

palm civet host or a human host thereby making palm civets a likely intermediate host for SARS-CoV (Guan, et al., 2003). It is also known that bats are a natural reservoir for SARS-CoV with certain strains isolated from Chinese horseshoe bats having 88-92% identity to human coronaviruses (Ren, et al., 2006). There are several theories of how SARS-CoV-2 emerged with one suggesting a zoonotic transfer between humans and bats as seen with SARS-CoV. A sample of bat coronavirus RaTG13 taken from a horseshoe bat had a 96.2% sequence identity with SARS-CoV-2 suggesting a close relationship (Zhou, et al., 2020). Another theory is that SARS-CoV-2 comes from pangolins which are illegally imported into Guangdong province often carrying coronaviruses. Although sequence identity is still high for pangolin-CoV at 91% it is still less than that of RaTG13 and some have suggested that pangolins may be a natural intermediate host for both SARS-CoV-2 and RaTG13 (Zhang, et al., 2020). Moreover, MERS-CoV is thought to also originate from bats with dromedary camels acting as the intermediate hosts (Killerby, et al., 2020). A summary table comparing COVID-19 and SARS can be seen below (Table 1).

Table 1: Summary of the main aspects of COVID-19, SARS, and MERS (Zhu, et al. 2020)

Table shows a comparison of the three main coronavirus illnesses: COVID-19 caused by SARS-CoV-2, SARS caused by SARS-CoV, and MERS caused by MERS-CoV. It compares the R_0 number which indicates how contagious the disease is, CFR which is the case fatality rate, number of cases, number of fatalities at time of writing, where the virus emerged from, what the natural reservoir organism is, and finally what organism is the host organism.

	COVID-19	**SARS**	**MERS**
R_0	3.28	3	0.69
CFR	2.1%	9.6%	34.3%
Cases	209,876,613	8096	2553
Fatalities	4,400,284	774	876

Origin	Wuhan, China	Guangdong, China	Saudi Arabia
Natural reservoir	Bat	Bat	Bat
Intermediate host	Pangolins	Palm civets	Dromedary camels

Most cases of COVID-19 have resulted in full recovery, but some may present with severe symptoms that have likely developed in the later stages of disease (>10 days). In these critical cases, patients develop acute respiratory death syndrome (ARDS) and organ-failure rather quickly which often leads to death (Chinese Preventative Medicine Association, 2020). One of the major causes of ARDS is cytokine storm which is a term applied to dysregulated cytokine release leading to disease aggravation (Chousterman, et al., 2017). At the early stage of infection by SARS-CoV, MERS-CoV, or SARS-CoV-2 delayed release of cytokines and chemokines occurs which results in high concentrations of pro-inflammatory cytokines. These high concentrations attract neutrophils, monocytes and other inflammatory cells leading to excessive infiltration into lung tissue resulting in injury. Increased levels of inflammatory cytokines and chemokines in the blood is highly correlated with disease mortality of COVID-19 (Ye, et al., 2020). The key issue with this is that it cannot be predicted and often occurs rapidly, with no effective treatment against it.

1.4 Differentially Conserved Positions (DCPs)

Researchers at the University of Kent developed an approach to identify the determinants of key phenotypic differences between related viruses building on the finding of differentially conserved positions (DCPs) (Pappalardo M., 2016) (Bojkova, et al., 2021). DCPs are specific positions within the virus protein that are one amino acid in group 1 and a different amino acid in group 2 (Figure 3). It is known that conserved positions within a sequence of DNA or

RNA are likely to have function relevance, but it was postulated by researchers that differential conservation may cause functional differences (Rausell, et al., 2010).

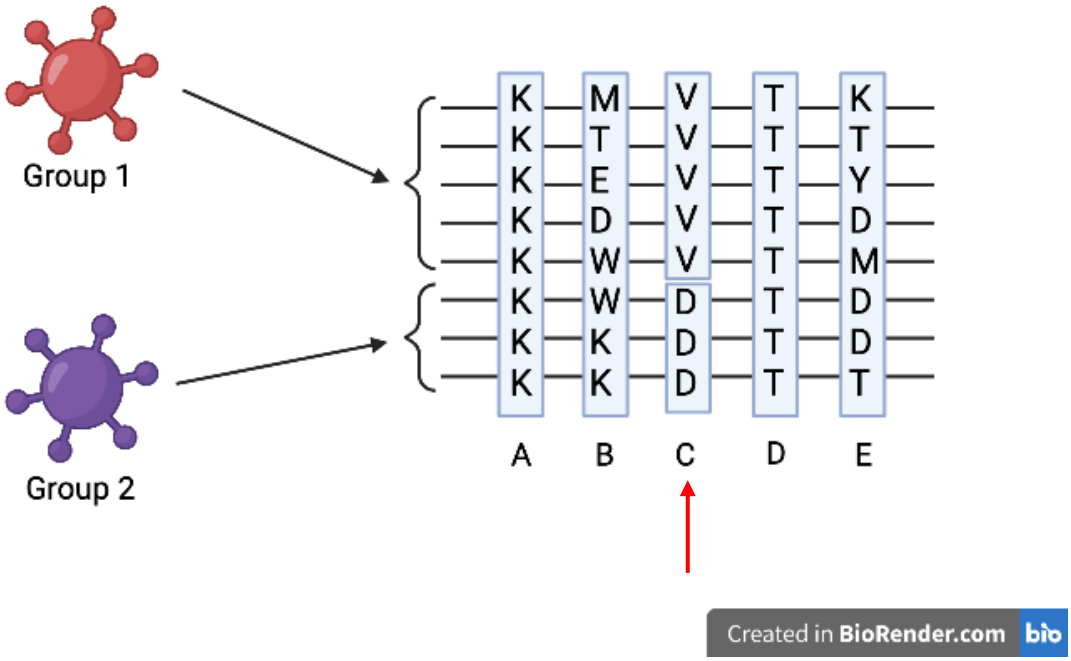


Figure 3: Summary of DCP identification by VAT

Created with BioRender.com. A schematic explanation of DCP identification whereby there are specific positions within the virus protein that are one amino acid in group 1 and a different amino acid in group 2 (column C). Group 1 and 2 refer to similar viruses which in this case was strains taken from two different time periods. Columns A and D represent full conservation whereby the amino acids are all the same, and columns B and E represent variation whereby there is no common amino acid.

This method was first established through research determining the differences in human pathogenicity between Reston viruses and other Ebolaviruses (Pappalardo M., 2016). Reston viruses are not pathogenic in humans whereas Ebolaviruses are, and this was highlighted by an outbreak of Ebola virus disease in Western Africa between 2013 and 2016 which resulted in 28,616 reported cases and 11,310 deaths (www.who.int). Pappalardo (2016) identified specificity determining positions (SDPs) which are positions that are conserved but differ between protein subfamilies. These SDPs contained amino acid changes between Reston viruses and Ebolaviruses that could explain the differences in pathogenicity. Bojkova (2021)

was able to apply this method to SARS-CoV and SARS-CoV-2 to identify DCPs between these related viruses to determine why phenotypic differences arise between them.

1.5 Research Aims

The key issues discussed previously refer to repeating emergence of similar viruses, each with severe socio-economic consequences in addition to an ever-increasing risk of HPAIs. Nevertheless, when a novel virus emerges such as SARS-CoV-2 or A(H1N1)09pdm it is generally phylogenetically similar to its predecessor therefore, giving them many similarities. However, differences in viral genomes arise which influences virulence and pathogenicity - often in very different ways than previously seen. My research focused on two different viruses. First, I investigated the determinants of pathogenicity in influenza through DCP analysis using a virus analysis tool (VAT) and *in silico* modelling. The aim of this part of the project was to build on previous research into DCP analysis and to apply these tested methods to influenza viruses. I chose two groups of similar influenza viruses that show differences in pathogenicity and apply these to the VAT method. If DCPs were present between the similar influenza viruses, *in silico* protein structure modelling was utilised to confirm whether this mutation is likely to have an effect on protein structure or function. If a significant effect on protein structure or function occured, it suggests that this position possibly has a role in causing the different phenotypes. As this mutation is still present, it ought to have had to withstand various selection pressures over time meaning it must have been conserved because it has an important functional role in the protein. It is hoped that the result of this analysis will be DCPs which would explain why phenotypic differences arise.

Second, I considered the adaptation of SARS-CoV-2 to two potential COVID-19 drugs: camostat and nafamostat. Sequence data from the FFM1 strain of SARS-CoV-2 was analysed to investigate the methods of resistance to the serine protease inhibitors Camostat mesylate and Nafamostat mesylate. SARS-CoV-2 was isolated from a patient and cultured in the human intestinal caco-2 cell line where separate samples were introduced individually to camostat and nafamostat and their resulting genomes were sequenced. These drugs are known to block TMPRSS2 activity which is an important activator of SARS-CoV-2 so thereby prevents virus entry (Hoffmann, et al., 2021). In particular, the S protein was investigated since it utilises the host ACE2 and TMPRSS2 cells to gain entry through cleavage and activation. The sequence data was compared with the SARS-CoV-2 reference genome taken from UCSC Genome Browser (Kent, et al. 2002). Mixed variant populations whereby there is less than 90% of the sequencing reads at that base in agreement were flagged for further research as well as mixed variant populations. These flagged positions will be analysed using *in silico* modelling and ligand prediction software to determine whether these mutations are likely to have a structural or functional effect on the protein. If any significant structural or functional effects are found these must be contributing to resistance to camostat or nafamostat.

2 Methods and Materials

2.1 Sequence collection

Influenza virus H1N1 sequences were obtained from the NCBI Influenza Virus Database (Bao, et al., 2008). All the sequences came from human-only hosts and were the full-length complete sequence. The SARS-Cov-2 strain FFM1 was passaged in Caco2 cells [Hoehl et al., 2020] in the absence or presence of increasing concentrations of camostat or nafamostat. SARS-CoV-2 sequencing was performed by Public Health England.

2.2 Identification of differentially conserved amino acid sequence positions

A novel approach developed by Wass-Michaelis Lab at the University of Kent was utilised to identify DCPs (Pappalardo M., 2016) (Bojkova, et al., 2021). Sequences collected in 2.1 were kept separately as group 1 and group 2 based off their different features (e.g. collection date) and stored by protein as a text file in FASTA format. Each protein text file from group 1 and group 2 is inputted and DCPs are returned if they are present for that protein. This can be repeated for every other protein or the whole genome.

2.3 FFM1 analysis

The sequenced FFM1 strain of SARS-CoV-2 was analysed using a python script (see Appendix 6) to identify and flag bases within the sequenced genome that did not match the reference SARS-CoV-2 genome. This script loaded the sequences into the python data frame which created a dictionary of all the nucleotides within the sequence and if the nucleotide for a specific position did not match the reference sequence nucleotide it would store that position

in a new .csv file. Positions whereby less than 90% of the sequencing reads at that base in agreement will also be flagged as these are inconsistencies.

2.4 Retrieving protein structures

Protein models were taken from several sources with the main source being the RCSB Protein Data Bank (Berman, et al., 2000). Some proteins were modelled using Phyre2 which generated a 3D structure prediction based off evolutionary relationships and sequence homology (Kelley, et al., 2015). Furthermore, the SARS-CoV-2 proteins that lacked suitable templates were taken from DeepMind which predicts protein structure using its AlphaFold AI system (Jumper, et al., 2020).

2.5 *In silico* modelling

In silico modelling was used to visualise and analyse proteins using the molecular visualisation software PyMol (Schrödinger & DeLano, 2020). This was used first and foremost to identify whether DCPs influenced protein structure or function. It was also used to analyse adaptations of FFM1 to the drugs camostat and nafamostat and see whether these also had impacts on protein structure and function. DynaMut2 was also used as a tool to predict the impact of a mutation on protein dynamics and stability, as well as to have a more comprehensive look at amino acid interactions (Rodrigues, et al., 2020).

2.6 Predicting ligand binding sites

3DLigandSite was used to predict the likely interaction residues between monomers and ligands, to predict a ligand binding site (Wass, et al., 2010). These ligand binding sites were confirmed using the program Firestar which uses structural templates and alignment reliability to predict functionally important residues (López, et al., 2007).

3 Results

3.1 Overview of DCP findings

The H1N1 sequences (Section 2.1) were sorted into two groups depending on their collection date, with sequences in group 1 (WT) ranging from January 1918 until December 1950 and sequences in group 2 (mutant) ranging from September 2008 until July 2009. The significance of these dates was to ensure that the right H1N1 variant was being analysed since the 1918 H1N1 virus was in circulation up until the 1950s when a new H1N1 variant began circulating. The aim was to compare the 1918 pandemic H1N1 virus with the 2009 pandemic H1N1 virus so the collection date had to be chosen carefully. The average number of sequences collected in group 1 was 51.25 compared with the average number in group 2 at 499.33, so even with a range of 32 years in group 1 the lack of availability of sequencing technology made this type of data scarce. This data was inputted into the virus analysis tool (VAT) provide by Wass-Michaelis lab.

The virus analysis tool (VAT) works as a result of inputting two sets of sequences from related viruses to find residue conservation at each point in the individual sequences using a BLOSUM62 matrix. Sequences submitted in FASTA format are further separated into individual proteins before being run through the program which then identifies unique sequences and generates the sequence alignment. The alignments are then split into groups and the Jenson Shannon Divergence calculated from this (Capra J.A., 2007). If the residue at that position has a conservation score >0.8 for both sequence groups, then it is a DCP and returned. A total of eight DCPs were found throughout the entire genome: three were found in the HA gene, two were found in the NA gene, two were found in the PB1 gene with one on

the +1 reading frame accessory protein PB1-F2, and one was found in the RNA polymerase subunit PA-X (Table 2). Further analysis of these DCPs shows that the eight DCPs formed just 2.63% of the total number of residues in the influenza genome (Table 3).

Table 2: Summary of DCP results for Influenza A(H1N1)

A total of 8 DCPs were found through using the VAT and can be seen in the table. DCP refers to the position within the sequence alignment that mutation takes place. WT refers to the position within the viral genome of sequences taken from January 1918 to December 1950. Mutant refers to the position within the viral genome of sequences taken from September 2008 and July 2009. The alignment position tells you where in the sequence alignment these DCPs can be found. The BLOSUM score tells you the probability of this mutation occurring in nature (a positive score is given to likely substitutions). The protein refers to which influenza A protein the DCP is found.

DCP	WT	Mutant	Alignment Pos	BLOSUM	Protein
A20T	A11	T11	20	0	HA
N233S	N223	S224	233	1	HA
T293I	T283	I284	293	-1	HA
S95R	S95	R95	95	-1	NA
Q250P	Q250	P250	250	-1	NA
L11Q	L11	Q11	11	-2	PB1-F2
E583D	E581	D581	583	2	PB1
R195K	R195	K195	195	2	PA-X

Table 3: Summary of H1N1 sequence information per protein

This table shows how many sequences were in the data set for each protein. These sequences were inputted into VAT and from that, the DCPs were found. It also shows the length of each protein and how many DCPs were identified as a percentage of the whole sequence. A total of 8 DCPs were identified which is 2.63% of the whole influenza A genome.

Protein	Seqeunces in Dataset	Protein Length	DCPs Identified	% of Residues DCPs
Hemagglutinin	1819	566	3	0.53
Neuraminidase	941	470	2	0.43
Nucleoprotein	374	498	0	0.00
Matrix Protein 1	236	252	0	0.00
Matrix Protein 2	218	97	0	0.00
Non-structural Protein 1	512	230	0	0.00
Non-structural Protein 2 (Nuclear Export Protein)	215	121	0	0.00
Polymerase Acidic Protein	732	716	0	0.00
PA-X	59	232	1	0.43
Polymerase Basic Protein 1	656	757	1	0.13
Polymerase Basic Protein 2	751	759	0	0.00
PB1-F2	94	90	1	1.11
Total			**8**	**2.63**

3.2 Structural Analysis of DCPs

This section is dedicated to the discussion of results pertaining to the structural analysis of DCPs found in section 2.2. to characterise the different phenotypes found between similar viruses.

3.2.1 Haemagglutinin DCPs

The haemagglutinin DCPs account for just 0.71% of the total residues in this protein out of a dataset containing 1819 sequences (Table 3).

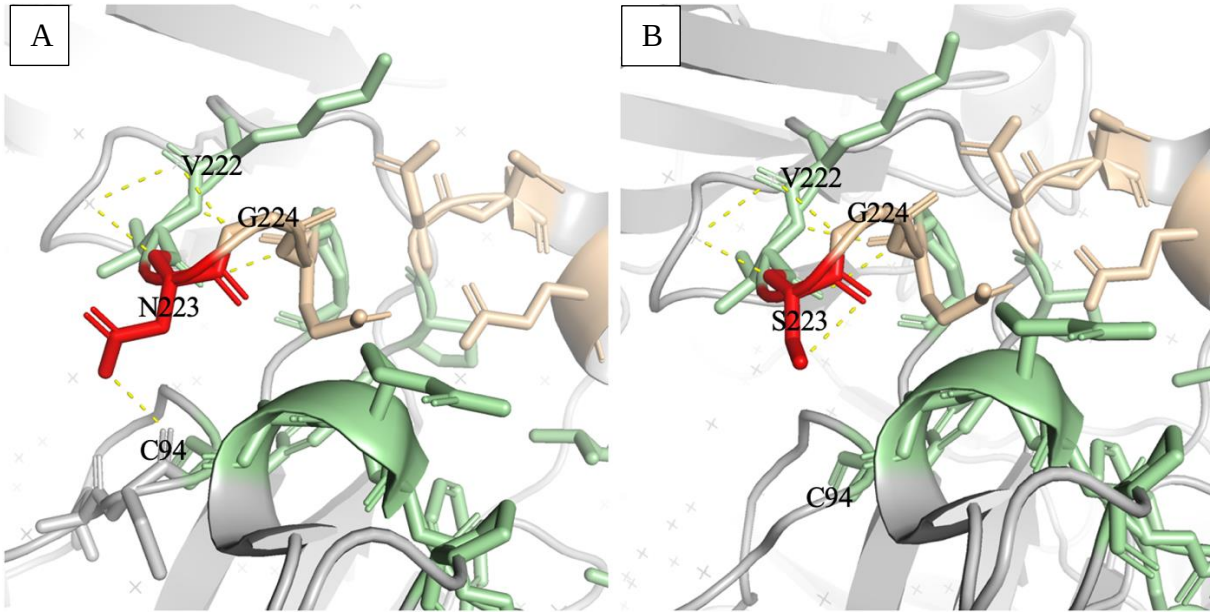


Figure 4: Structure of haemagglutinin DCP N233S

Panel A: 1918-1950 (WT) haemagglutinin protein. DCP N233S at position N223 (red) which forms one hydrogen bond (yellow) with a water molecule and residue C94. Panel B: 2008-2009 (mutant) haemagglutinin protein. DCP N233S now as residue S223 (red) which has lost the hydrogen bond (yellow) with C94. Pale green colour relates to cluster 1 and wheat colour represents cluster 2 as predicted by 3DLigandSite.

The WT form of the DCP N233S can be seen above on the left (Figure 4). This residue forms one hydrogen bond with a water molecule and another with the similarly polar residue C94 which is found within a neighbouring loop. It is found within a loop in the protein structure,

within the binding sites of cluster 1 and cluster 2. These clusters are predicted by 3DLigandSite to be binding sites for molecules and ions, or protein ligands so may play an important functional role. However, it has a machine learning generated probability score of 0.05 which suggests that it is non-binding as the benchmark average probability score for binding residues is 0.33. Despite this, it is still found very close to these binding sites which have a predicted 594 and 102 bound ligands respectively which makes them highly likely to be true clusters. It is unexpected that despite being within such proximity to two predicted clusters that it does not form part of these binding sites.

The mutant form of the DCP N233S can be seen on the right and is predicted to be a conservative change (Figure 4). This residue, serine, has similar properties to the WT asparagine as it too is polar therefore making this substitution a biological favoured one. Serine is smaller in size than asparagine and no clashing occurs between the DCP and surrounding residues within a 5 angstrom area. However, the hydrogen bond with residue C94 is lost which results in alteration to side chain interactions. Cysteine works to link fragments within a polypeptide chain thus increasing stability. Here the hydrogen bond may coordinate the structure in this region by holding the two loops in proximity; loss of this bond will result in decreased stability. In addition, the serine residue forms a hydrogen bond with itself. It is likely that there will be an effect on stability of the protein structure given the alteration of side chain interactions with its surrounding residues.

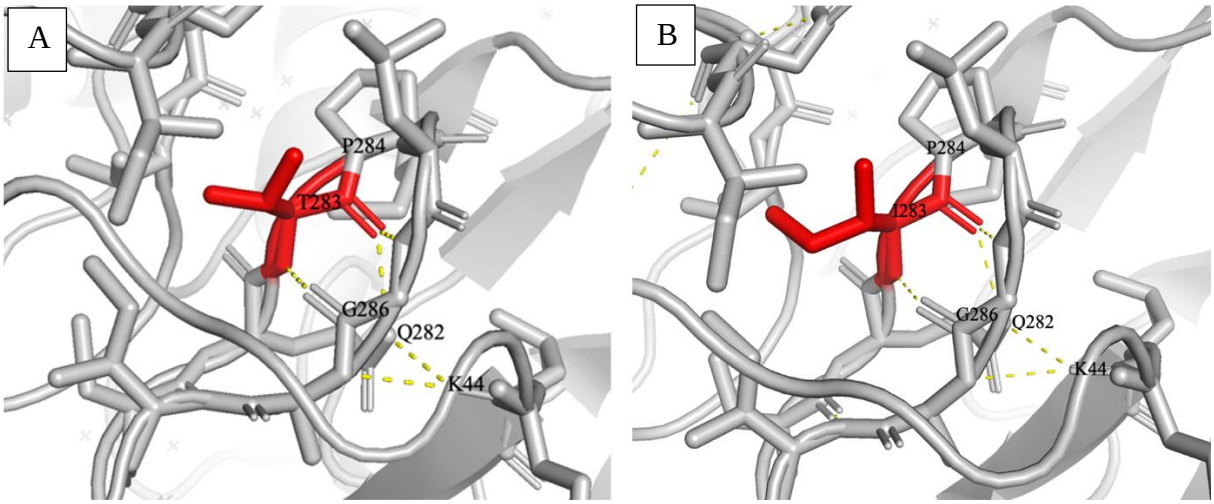


Figure 5: Structure of haemagglutinin DCP T293I

Panel A: 1918-1950 (WT) haemagglutinin protein. DCP T293I at position T283 (red) which forms a total of three hydrogen bonds (yellow): two with residue G286 and one with residue Q282. Panel B: 2008-2009 (mutant) haemagglutinin protein. DCP T293I now as residue I293 (red) which keeps all hydrogen bonds (yellow) with G286 and Q282. 3DLigandSite predicts no clusters in this area.

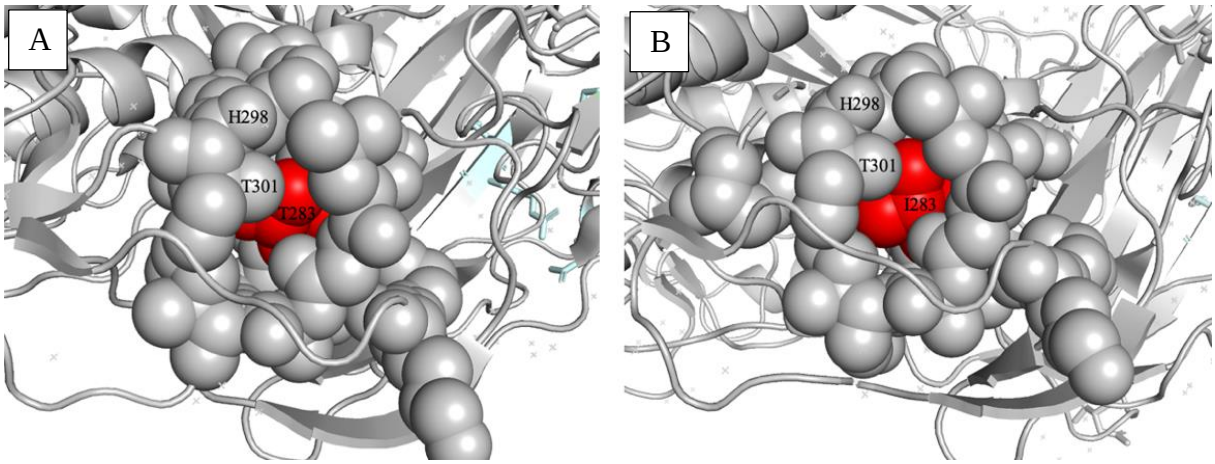


Figure 6: Structure of haemagglutinin DCP T293I with sphere representation

Panel A: 1918-1950 (WT) haemagglutinin protein with sphere representation to show interactions with surrounding residues. DCP T293I at position T283 (red) shown to be in a crowded region of the protein in particular being very close with residues H298 and T301. Panel B: 2008-2009 (mutant) haemagglutinin protein. DCP T293I now as residue I293 (red) which now overlaps with residues H298 and T301. 3DLigandSite predicts no clusters in this area.

The WT residue of T293I can be seen above on the top left (Figure 5). It forms a total of three hydrogen bonds: two with residue G286 and one with residue Q282. This DCP is found within a loop in the protein structure just after a beta sheet and is not found near any predicted

binding sites. However, the surrounding area is very crowded. This residue, threonine, is considered to be slightly polar so will likely substitute with other polar or small amino acids.

The mutant form of the DCP T293I can be seen above on the top right (Figure 5) and is a polar to nonpolar change which can often be a dangerous transition in terms of changes to the protein structure. Isoleucine is much larger than the WT caused by an increase in the size of the side chain resulting in clashes with surrounding residues H298 and T301 which can also be seen above (Figure 6). The outcome of this will be conformational changes that may result in denaturation due to the overlapping residues causing a loss of functional activity.

3.2.2 Neuraminidase DCPs

The neuraminidase DCPs account for 0.43% of the total residues in this protein out of a dataset containing 941 sequences (Table 3).

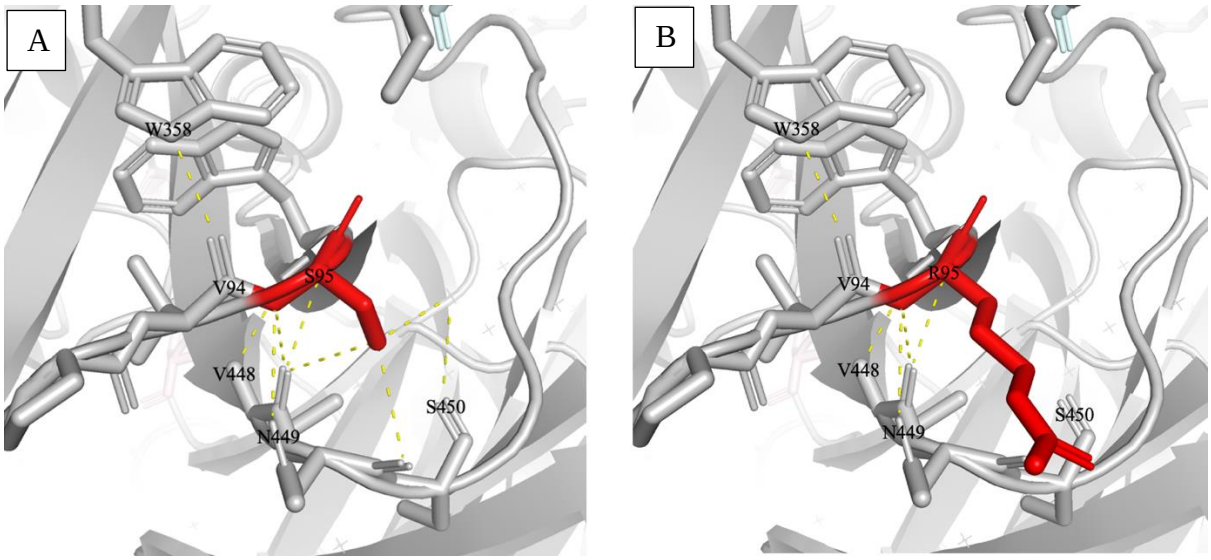


Figure 7: Structure of neuraminidase DCP S95R

Panel A: 1918-1950 (WT) neuraminidase protein. DCP S95R at position S95 (red) which forms a total of five hydrogen bonds (yellow) including one with a water molecule, two with residue V448, and two with residue N449. Panel B: 2008-2009 (mutant) neuraminidase protein. DCP S95R now as residue R95 (red) which loses three

hydrogen bonds (yellow) and just two remain at residues V448 and N449. There are three predicted ligand binding sites close to this region but only one can be seen (light blue).

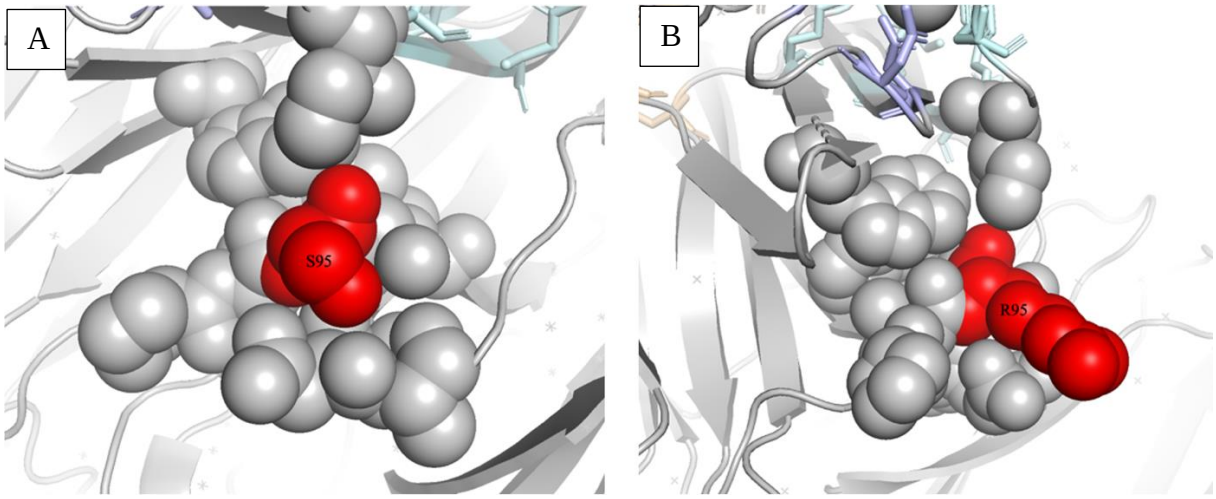


Figure 8: Structure of neuraminidase DCP S95R with sphere representation

Panel A: 1918-1950 (WT) neuraminidase protein with sphere representation to show interactions with surrounding residues. DCP S95R at position S95 (red) shown to be in a fairly crowded region of the protein. Panel B: 2008-2009 (mutant) neuraminidase protein. DCP S95R now as residue S95 (red) which still in the crowded pocket but due to the angle of the side chain it does not overlap any surrounding residues. There are three predicted ligand binding sites close to this region (light blue, wheat, light purple).

The WT residue of the DCP S95R can be seen above on the top left panel (Figure 7). It is found within a loop in the protein structure, just before the start of a beta sheet. The amino acid serine is slightly polar and uncharged. It forms a total of five hydrogen bonds with its surrounding residues, including one with a water molecule, two with residue V448, and two with residue N449. The area surrounding the residue (within 5Å) is particularly crowded. The mutated form of the DCP S95R at position R95 can be seen above on the top right (Figure 8) and it goes from being uncharged to positively charged which will result in stability changes and changes to the pH of the solution the protein is in. The program DynaMut2 (Rodrigues, et al., 2020) proposes the predicted stability change for this mutation is -0.17 kcal/mol which suggests this mutation is only slightly destabilising. It may have little to no effect and therefore, not be selected for.

Furthermore, it loses three of the original hydrogen bonds with residues in a neighbouring loop will likely decrease stability of the protein slightly since these polar contacts may be holding the two loops together although there are still two remaining ones (V448 and N449) which may counteract this change. Arginine has a much larger side chain than serine yet, this conformation shows no overlap to occur between surrounding residues due to it projecting outwards across the surface of the protein, so no clashing occurs between the DCP and surrounding residues within a 5 angstrom area (Figure 8). Therefore, the loss of hydrogen bonds will reduce stability as these are in place to pull the molecules together. The result of this will be structural changes and potential denaturation of the protein as the normal shape can become deformed.

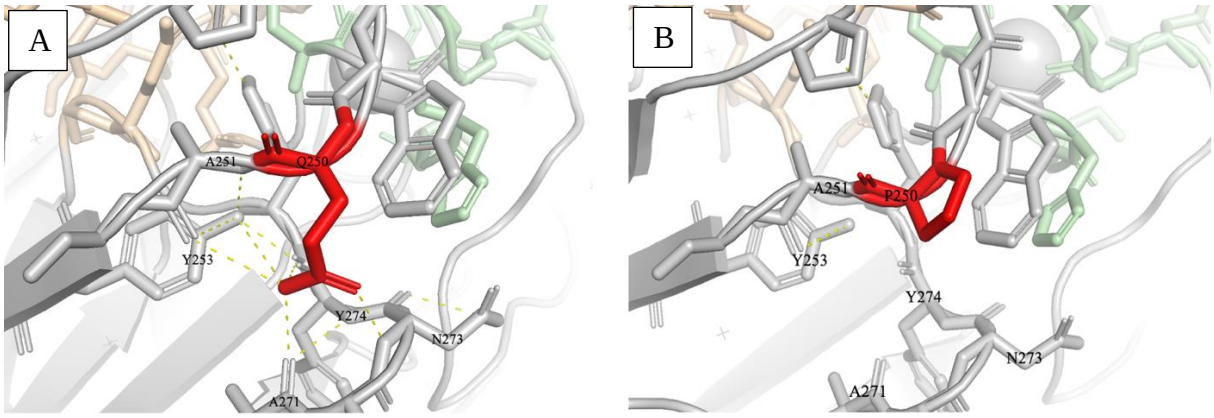


Figure 9: Structure of neuraminidase DCP Q250P

Panel A: 1918-1950 (WT) neuraminidase protein. DCP Q250P at position Q250 (red) which forms a total of five hydrogen bonds (yellow) with five different residues: Y253, A251, A271, N273, and Y274. Panel B: 2008-2009 (mutant) neuraminidase protein. DCP Q250P now as residue P250 (red) which results in the loss of all 5 hydrogen bonds (yellow) with surrounding residues. The pale green and wheat colours in the background represents clusters 1 and 2 respectively, as predicted by 3DLigandSite.

The WT for the DCP Q250P is shown above on the left (Figure 9). It is found in a loop in the protein structure just before a beta sheet. Q250 forms a total of five hydrogen bonds with five different residues: Y253, A251, A271, N273, and Y274. The mutated version of the DCP

Q250P can be seen above on the right (Figure 9) and is a polar to nonpolar change which often results in serious irreversible changes to the protein structure. DynaMut2 proposes the predicted stability change for this mutation is -0.12 kcal/mol which suggests this change in polarity is only slightly destabilising meaning it may have little to no effect and therefore not be selected for.

Significantly, all five hydrogen bonds are lost in the mutated form which would be expected to cause a substantial decrease in stability since hydrogen bonds contribute highly to protein stability but this is not the case. One of these polar contacts forms with residue Y253 which is situated within a beta sheet and the other four form with residues in a neighbouring loop. This is highly suggestive that these polar contacts play a role in stabilising the structure, particularly by holding the two loops in proximity and thereby coordinating the surrounding structures at this point. Finally, proline is a smaller amino acid than glutamine despite it containing a pyrrolidine ring structure which often implements conformational restrictions which, in combination with the loss of hydrogen bonds and polarity changes, results in changes to side chain interactions conferring different biochemical and structural properties. This could lead to loss of protein function.

3.2.3 RNA-directed RNA polymerase catalytic subunit (PB1) DCPs

The final DCP that was analysed is the RdRp residue E583D (Figure 10) which accounted for 0.13% of total residues in this protein, out of 656 sequences in the dataset (Table 3).

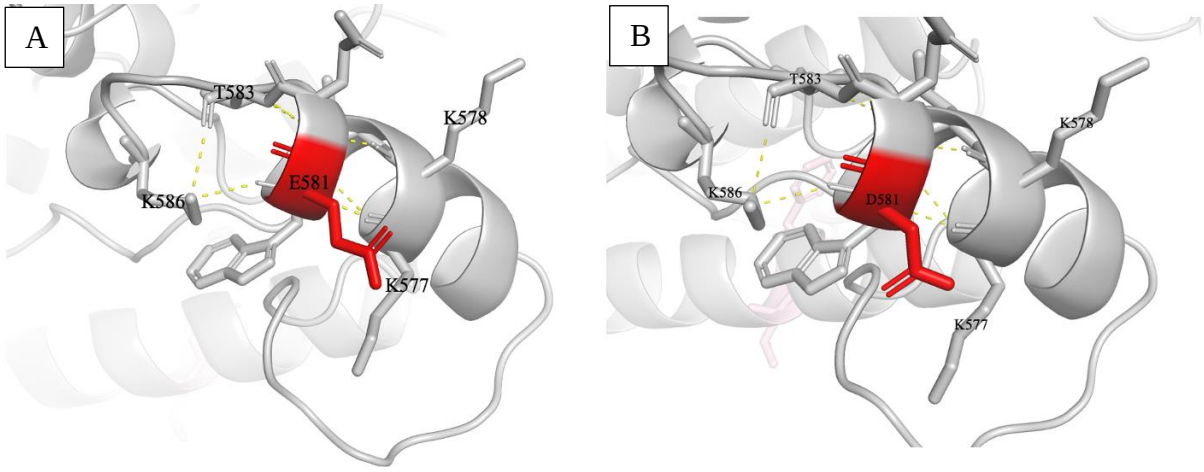


Figure 10: Structure of RdRp DCP E583D

Panel A: 1918-1950 (WT) RdRp protein. DCP E583D at position E581 (red) which does not form any hydrogen bonds (yellow) from the side chain and just having one hydrogen bond within the alpha helix at residue K577. Panel B: 2008-2009 (mutant) RdRp protein. DCP E583D now as residue D581 (red) which does not gain any hydrogen bonds from the side chain although the one hydrogen bond (yellow) with K577 remains which could be structural.

The WT for the DCP E583D can be seen above (Figure 10). It is found towards the end of an alpha helix structure which is stabilised by a regular formation of hydrogen bonds. There are no hydrogen bonds formed from the side chain which projects outwards across the surface of the protein, but there is one hydrogen bond within the alpha helix with residue K577. Since alpha helices are stabilised by hydrogen bonds within the main chain it can be assumed that this hydrogen bond with K577 is just there to stabilise this structure. This hydrogen bond is carried over in the mutation which further supports the suggestion that this hydrogen bond is purely for structural support.

The mutant form of the DCP E583D can also be seen above, on the right (Figure 10) and this amino acid substitution was predicted to be a conservative change since both amino acids are negatively charged and similarly sized. The BLOSUM score for this substitution is +2 which also suggests this type of substitution is biologically likely to occur. The mutant is slightly larger with one extra methylene group although it appears the side chains of glutamic acid

and aspartic acid do not interact with its surrounding residues since they do not form polar contacts from the sidechain, and projects outwards across the surface of the protein. The number of hydrogen bonds remains the same also. Overall, this is a conservative substitution, and it can be assumed that this mutation is unlikely to have any effect on the protein structure and/or function.

3.3 Overview of FFM1 results

The betacoronavirus hCoV-19/Germany/HE-FFM1/2020 (EPI_ISL_452218) was taken by Goethe University Hospital Frankfurt and passaged (by collaborators at Frankfurt University) in a caco-2 cell line where it was exposed to increasing concentrations of camostat or nafamostat. The virus adapted to drugs will be referred to as FFM1-camostat and FFM1-nafamostat to indicate the drug they have been adapted to. After sequencing it was found that the FFM1-camostat virus shared a 99.96% identity to FFM1 (EPI_ISL_452218) and the FFM1-nafamostat virus shared a 99.95% identity to FFM1 (EPI_ISL_452218). A total of 20 sequence positions were flagged for FFM1-camostat (Table 4) and of these 9 were mixed populations and 11 were base changes. A total of 24 sequence positions were flagged for FFM1-nafamostat (Table 5) and of these 7 were mixed populations, 11 were base changes and 6 had very low percentage sequencing reads. These changes either had no change (conservative) due to the degenerate nature of the genetic code or they resulted in amino acid substitutions.

Table 4: Summary of FFM1-camostat flagged residues

This table shows the positions that were returned and stored in a .csv file for further investigation. The program was a python script which returned positions in the sequence that had less than 90% of sequencing reading in agreement at that base as well as flagging mixed variant populations. 'Pos' refers to the position within the viral

genome and 'RefN' refers to the corresponding base in the reference genome. It then shows what the nucleotide change is when FFM1 is adapted to camostat and then it has the corresponding codon and amino acid changes. There is also a BLOSUM score which tells you the probability of this mutation occurring in nature (a positive score is given to likely substitutions). It includes which protein the mutation occurs in including what model was used in in silico modelling as well as the source. AA position refers to the position within the entire genome and then in brackets is the position within the specific protein.

Pos	RefN	Nucleotide change	Codon change	AA change	BLOSUM Score	Protein	Model	Source	AA position
2167	G	G > T	AAG > AAU	K > N	0	NSP2	nsp2	DeepMind	K634 (K453)
3518	G	G > T	GUU > UUU	V > F	-1	NSP3	6vxs	PDB	V1085 (V266)
4084	C	C > T	GAC > GAU	D > D	6	NSP3	6vxs	PDB	D1273 (D454)
12704	G	G > T	GUU > UUU	V > F	-1	NSP9	6wxd	PDB	V4147 (V6)
12706	T	T > G	GUU > GUG	V > V	4	NSP9	6wxd	PDB	V4147 (V6)
12797	G	G > A	GGU > AGU	G > S	0	NSP9	6wxd	PDB	G4178 (G37)
13381	C	C > T	GUC > GUU	V > V	4	NSP10	6w6l	PDB	V4372 (V118)
17423	A	A > G	UAU > UGU	Y > C	-2	NSP13 (Hel)	c5wwpB	Phyre2	Y5719 (Y394)
17678	C	C > T	ACG > AUG	T > M	-1	NSP13 (Hel)	c5wwpB	Phyre2	T5804 (T479)
20148	C	C > T	UUC > UUU	F > F	6	NSP15 (uridylate	6vww	PDB	F6627 (F174)
21711	C	C > T	UCA > UUA	S > L	-2	S	6vsb	PDB	S50
23403	A	A > G	GAU > GGU	D > G	-1	S	6vsb	PDB	D614
23520	C	C > T	GCU > GUU	A > V	0	S	6vsb	PDB	A653
23536	C	C > T	AAC > AAU	N > N	6	S	6vsb	PDB	N658
23895	C	C > T	ACC > AUC	T > I	-1	S	6vsb	PDB	T778
24198	C	C > T	GCG > GUG	A > V	0	S	6vsb	PDB	A879
24520	A	A > C	AAA > AAC	K > N	0	S	6vsb	PDB	K986
27272	T	T > C	GUU > GCU	V > A	0	ORF6	N/A	N/A	V24
28512	C	C > T	CAG > UAG	Q > *	N/A	ORF9c	N/A	N/A	Q77
28854	C	C > T	UCA > UUA	S > L	-2	ORF9c	N/A	N/A	Q41

Table 5: Summary of FFM1-nafamostat flagged residues

This table shows the positions that were returned and stored in a .csv file for further investigation. The program was a python script which returned positions in the sequence that had less than 90% of sequencing reading in agreement at that base as well as flagging mixed variant populations. 'Pos' refers to the position within the viral genome and 'RefN' refers to the corresponding base in the reference genome. It then shows what the nucleotide change is when FFM1 is adapted to nafamostat and then it has the corresponding codon and amino acid changes. There is also a BLOSUM score which tells you the probability of this mutation occurring in nature (a positive score is given to likely substitutions). It includes which protein the mutation occurs in including what model was used in in silico modelling as well as the source. AA position refers to the position within the entire genome and then in brackets is the position within the specific protein.

Pos	RefN	Nucleotide change	Codon chang	AA change	BLOSUM Score	Protein	Model	Source	AA Position
19	C	C > T	UUC > UUU	F > F	6	NSP1	N/A	N/A	
934	C	C > T	GAC > GAU	D > D	6	NSP2	nsp2	DeepMind	D223 (D42)
3518	G	G > T	GUU > UUU	V > F	-1	NSP3	6vxs	PDB	V1085 (V266)
6935	T	T > C	UCU > CCU	S > P	-1	NSP3	6vxs	PDB	S2224 (S1405)
10537	C	C > T	UAC > UAU	Y > Y	7	NSP5	6y2e	PDB	Y3424 (Y160)
12704	G	G > T	GUU > UUU	V > F	-1	NSP9	6wxd	PDB	V4147 (V6)
12797	G	G > A	GGU > AGU	G > S	0	NSP9	6wxd	PDB	G4178 (G37)
17423	A	A > G	UAU > UGU	Y > C	-2	NSP13 (Hel)	c5wwpB	Phyre2	Y5719 (Y394)
21765	T					S	6vsb	PDB	I68
21766	A					S	6vsb	PDB	I68
21767	C					S	6vsb	PDB	H69
21768	A					S	6vsb	PDB	H69
21769	T					S	6vsb	PDB	H69
21770	G					S	6vsb	PDB	V70
22227	C	C > T	GCU > GUU	A > V	0	S	6vsb	PDB	A222
22326	C	C > T	UCU > UUU	S > F	-2	S	6vsb	PDB	S255
22423	T	T > C	GAU > GAC	D > D	6	S	6vsb	PDB	D287
23280	C	C > T	ACU > AUU	T > I	-1	S	6vsb	PDB	T573
23757	C	C > T	ACC > AUC	T > I	-1	S	6vsb	PDB	T732
27272	T	T > C	GUU > GCU	V > A	0	ORF6	N/A	N/A	V24
28854	C	C > T	UCA > UUA	S > L	-2	ORF9c	N/A	N/A	Q41

3.3.1 Adaption to Camostat

Camostat mesylate is clinically proven to prevent SARS-CoV-2 entry into Caco-2 cells by preventing the S protein binding with TMPRSS-2. There was a total of 7 spike protein mutations found in the FFM1-camostat strain with 1 resulting in no amino acid change (Table 6). Positions 21711, 23403, 23520, 23895, and 24190 will all be further analysed in this section. Position 23536 resulted in no amino acid change due to the degenerative nature of the genetic code. Position 24520 is not found within the S protein structure so it cannot be analysed further.

Table 6: Amino acid changes to FFM1-camostat spike protein residues

This table is an edited version of table 4 that just shows the spike protein mutations. It includes 'Pos' which refers to the position within the viral genome and 'RefN' refers to the corresponding base in the reference genome. It then shows what the nucleotide change is when FFM1 is adapted to camostat and then it has the corresponding codon and amino acid changes. There is also a BLOSUM score which tells you the probability of this mutation occurring in nature (a positive score is given to likely substitutions). Finally, it tells you the specific amino acid this mutation corresponds to.

Pos	RefN	Nucleotide change	Codon change	AA change	BLOSUM	Protein	AA position
21711	C	C > T	UCA > UUA	S > L	-2	S	S50
23403	A	A > G	GAU > GGU	D > G	-1	S	D614
23520	C	C > T	GCU > GUU	A > V	0	S	A653
23536	C	C > T	AAC > AAU	N > N	6	S	N658
23895	C	C > T	ACC > AUC	T > I	-1	S	T778
24198	C	C > T	GCG > GUG	A > V	0	S	A879
24520	A	A > C	AAA > AAC	K > N	0	S	K986

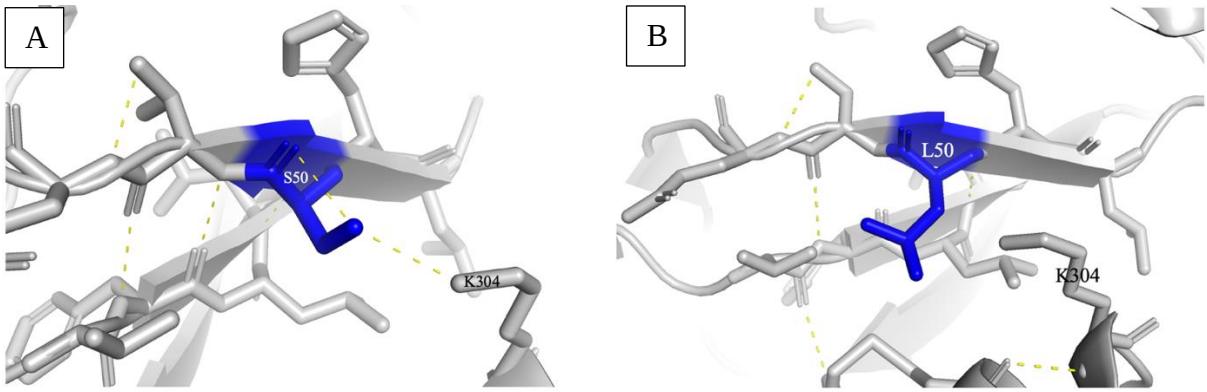


Figure 11: Structure of S protein residue S50L

Panel A: WT S protein taken from the reference sequence. Residue S50 (blue) can be seen to form one hydrogen bond (yellow) with itself and another with residue K304. Panel B: mutant S protein in response to camostat adaptation at position L50 (blue) which results in loss of hydrogen bond (yellow) with itself and K304.

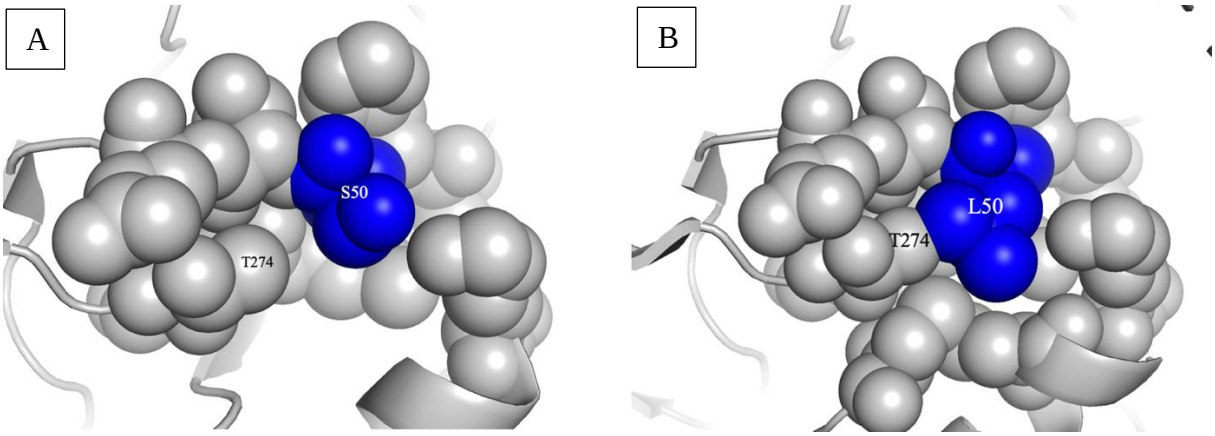


***Figure 12*: Structure of S protein residue S50L with sphere representation**

Panel A: WT S protein taken from the reference sequence with sphere representation to show interactions with surrounding residues. Residue S50 (blue) can be seen in a spacious pocket of the protein and does not overlap any proteins. Panel B: mutant S protein in response to camostat adaptation at position L50 (blue) which causes side chain to increase in size causing it to overlap with residue T274.

The WT residue for S50L can be seen above on the left (Figure 11). It is located towards the end of a beta sheet and its surrounding area is spacious. S50 forms one hydrogen bond with itself and another with residue K304 which is found in an adjacent secondary structure element. The mutated form of the residue can be seen above on the right (Figure 11) and this substitution causes a polarity change since the WT serine is polar and the mutant leucine is nonpolar. DynaMut2 suggests the predicted stability change for this mutation is 0.05 kcal/mol which suggests this mutation is slightly stabilising however it is likely this has little to no effect. This substitution does have a low BLOSUM score of -2 indicating this alignment was found to occur less often than by chance.

Also, both polar contacts that form with the WT are lost when mutated which will affect how the protein side chain interacts with its surrounding residues. In particular, the side chain will be able to take on more conformations when mutated because it loses a polar contact with itself and the neighbouring residue K304. Leucine has a larger side chain than the WT serine which causes there to be amino acid clashes with residue T274 (Figure 12). The overall effects of these changes will likely be conformational changes to the protein structure as a result of loss of hydrogen bonds which will cause decreased stability. This can lead to protein denaturation if the structure becomes too destabilised.

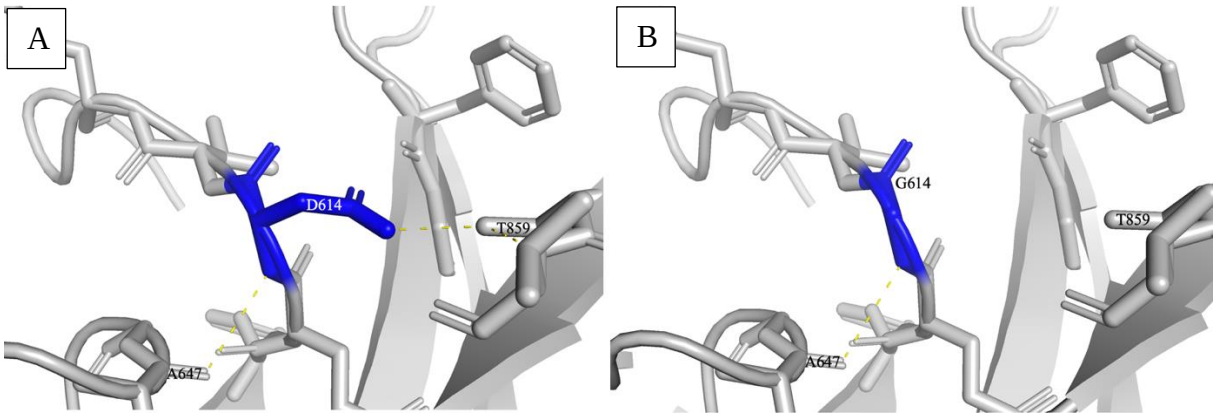


Figure 13: Structure of S protein residue D614G

Panel A: WT S protein taken from the reference sequence. Residue D614 (blue) can be seen to form two hydrogen bonds (yellow), one with A647 and another with T859. Panel B: mutant S protein in response to camostat adaptation at position G614 (blue) which results in loss of hydrogen bond (yellow) with T859.

The WT residue for D614G can be seen above on the left (Figure 13). It is located within a loop in the protein structure and the surrounding area is spacious. The WT residue, aspartic acid, is found in its negatively charged side chain state called aspartate. It forms two hydrogen bonds, one with A647 which is a hydrophobic residue and another with T859 which is polar uncharged. The mutant form of the residue D614G can also be seen above, on the right (Figure 13) and results in a change from a negatively charged aspartic acid residue to a nonpolar glycine residue. This amino acid substitution has a -1 score on the BLOSUM62 matrix indicating this alignment was found to occur less often than by chance.

The predicted stability change for this mutation is -0.3 kcal/mol as predicted by DynaMut2 which suggests this mutation is destabilising and usually removed by natural selection but since it has survived various selection pressures it is likely to have a key functional role. It also loses a hydrogen bond between D614 in S1 and T859 in S2 which has been speculated to promote S1 shedding. The result of this may cause increased flexibility in G614 as it is no longer bound in one conformation. D614G is a well-documented missense mutation that occurred during the COVID-19 pandemic, resulting in the formation of the G clade. Thus, it is known this mutation affects protein structure and stability however, whether it played a role in camostat resistance is unclear.

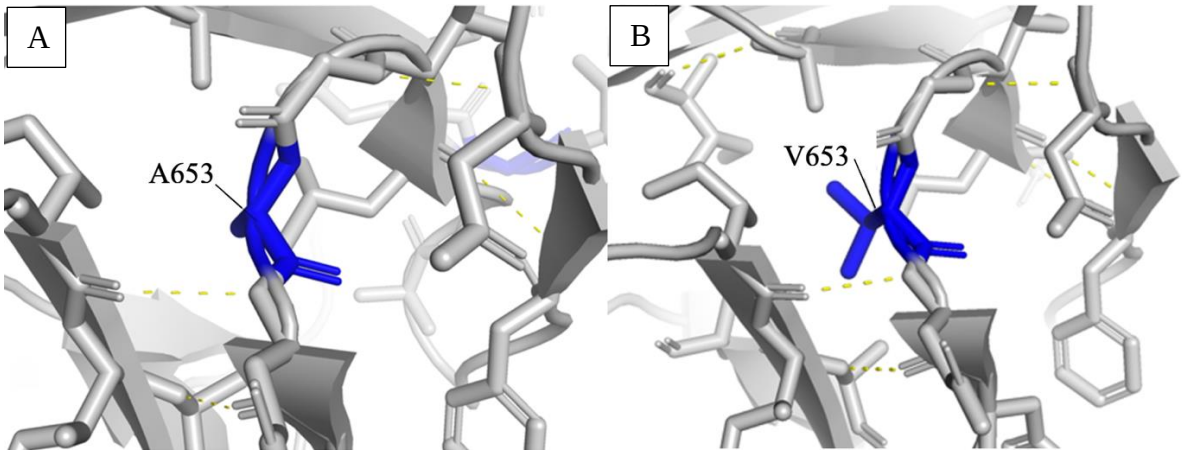


Figure 14: Structure of S-protein residue A653V

Panel A: WT S protein taken from the reference sequence. Residue A653 (blue) can be seen in the middle of a protein loop and does not form any hydrogen bonds with surrounding residues. Panel B: mutant S protein in response to camostat adaptation at position V653 (blue) which results in a larger side chain, but no hydrogen bonds gained.

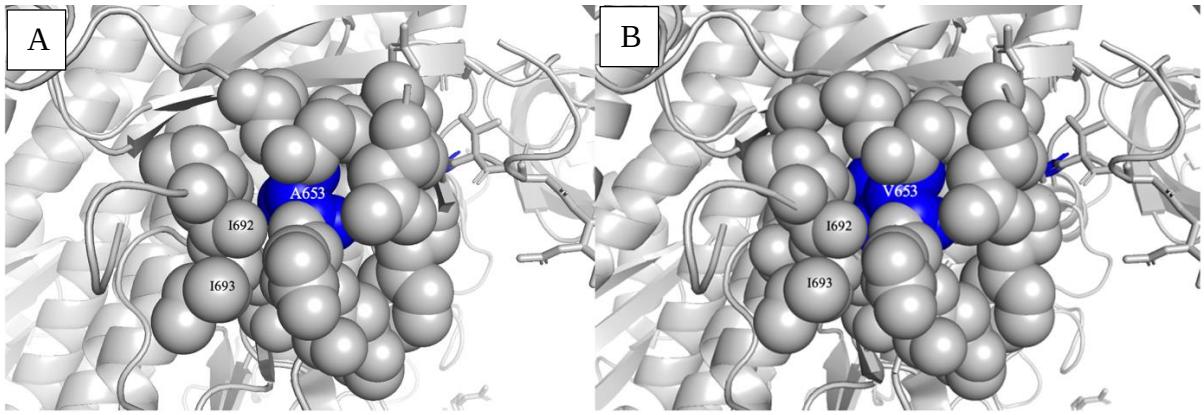


Figure 15: Structure of S-protein residue A653V with sphere representation

Panel A: WT S protein taken from the reference sequence with sphere representation to show interactions with surrounding residues. Residue A653 (blue) can be seen in a crowded pocket of the protein and is found very close to residues I692 and I693. Panel B: mutant S protein in response to camostat adaptation at position V653 (blue) which causes side chain to increase in size causing it to overlap with residues I692 and I693.

The WT form of the residue A653V can be seen above on the left (Figure 14). It is found within a loop in the protein structure in a crowded region. The mutant for the residue A653V can also be seen above, on the right (Figure 14), and this is predicted to be a conservative change since both amino acids are nonpolar. It has a BLOSUM matrix score of 0 signifying the frequency of this alignment occurs as expected by chance, which is also expected for a

conservative change. Although, there are differences to the size of the side chains between these two residues, with valine having a much larger side chain compared to alanine, this results in clashes between surrounding residues given the crowded area nearby. Specifically, V653 clashes with residues I692 and I693 (Figure 15) which will result in conformational changes to the protein structure.

A653V is also found to be near a predicted furin-like cleavage site (FCS) which is known to contribute to viral pathogenesis of SARS-CoV-2 (Xia, et al., 2020). The result of A653V is a possible change to protein structure or function. Given the conformational change that occurs as a result of the increased side chain length, and the structural changes that have survived selection pressures over time, this must be indicative of resistance to camostat. This change is expected given its proximity to a binding site of such importance.

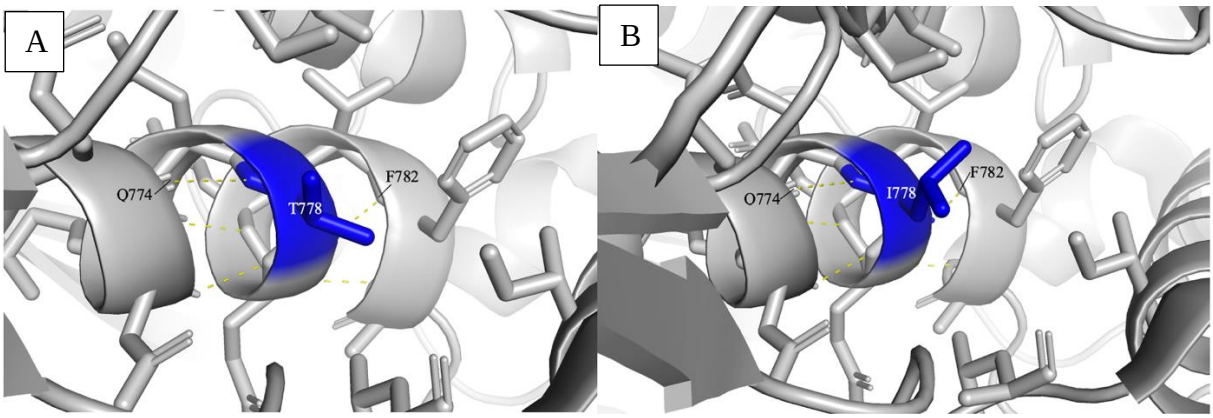


Figure 16: Structure of S-protein residue T778I

Panel A: WT S protein taken from the reference sequence. Residue T778 (blue) can be seen in the middle of an alpha helix and does not form any hydrogen bonds from its side chain with surrounding residues, but it forms two hydrogen bonds (yellow) from within the helix with positions Q774 and F782. Panel B: mutant S protein in response to camostat adaptation at position I778 (blue) which results in no hydrogen bonds (yellow) lost suggesting these are structural.

The WT for the residue T778I can be seen above on the left (Figure 16). It is located on the protein surface within an alpha helix and forms two hydrogen bonds. One of these bonds are with Q774 which is a polar uncharged residue and the other is with F782 which is hydrophobic. The mutant for the residue T778I can also be seen above, on the right (Figure 16). This is a polar to nonpolar change which causes the chemical environment to become more hydrophobic, often being a dangerous transition. The BLOSUM matrix score for this substitution is -1 meaning this alignment was found to occur less often than by chance. This position within the FFM1-camostat sequence was flagged as having a mixed population and this could contribute to resistance as it does not need to have 100% sequencing reads in agreement to have a role.

This mutation lies within an alpha helix, but threonine residues tend to prefer a beta sheet confirmation since its C-beta branching causes bulkiness and restricts the movement making it difficult for it to adopt an alpha helical conformation. There is also a difference in size between the WT and the mutant with isoleucine being slightly larger than threonine although, no clashing is observed between the DCP and surrounding residues within a 5 angstrom area since the side chain projects outwards across the surface of the protein. Overall, this mutation is unlikely to have any effect on protein structure and/or function which is unexpected given the polar to nonpolar change, as well as it being a mixed population variant.

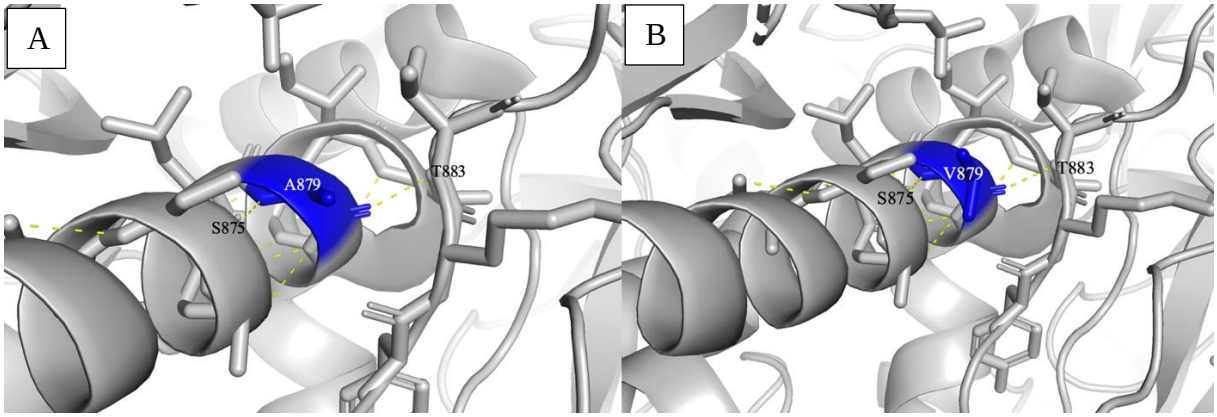


Figure 17: Structure of S-protein residue A879V

Panel A: WT S protein taken from the reference sequence. Residue A879 (blue) can be seen in the middle of an alpha helix and does not form any hydrogen bonds from its side chain with surrounding residues, but it forms two hydrogen bonds (yellow) from within the helix with positions S875 and T883. Panel B: mutant S protein in response to camostat adaptation at position V879 (blue) which results in no hydrogen bonds (yellow) lost suggesting they are structural.

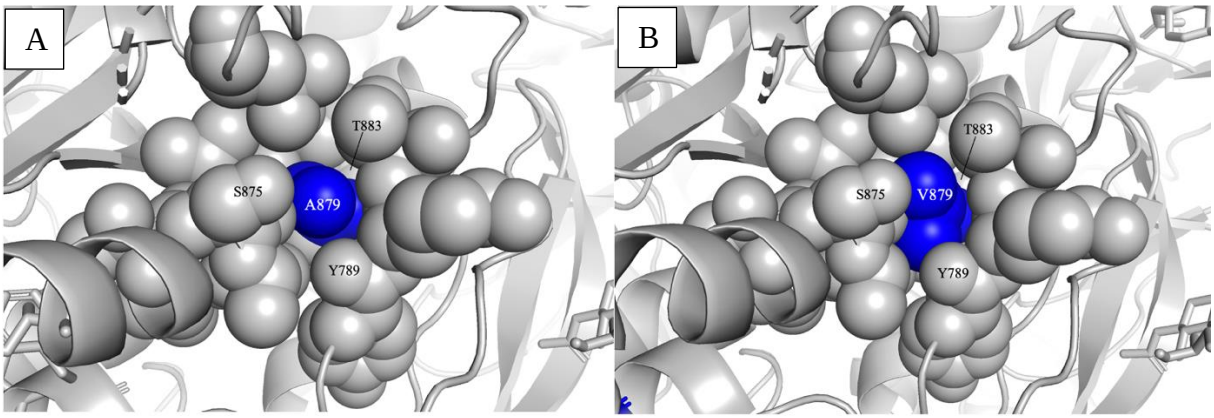


Figure 18: Structure of S-protein residue A879V with sphere representation

Panel A: WT S protein taken from the reference sequence with sphere representation to show interactions with surrounding residues. Residue A879 (blue) can be seen in a crowded pocket of the protein and is found very close to residues T883, S875, and Y789. Panel B: mutant S protein in response to camostat adaptation at position V879 (blue) which causes side chain to increase in size causing it to overlap with residues T883, S875, and Y789.

The WT for the residue A879V can be seen above on the left (Figure 17). It is found on the protein surface within an alpha helix and the surrounding area is very crowded because of this helix. The hydrogen bonds it forms are within the backbone so are not relevant to the mutation. The mutant for the residue A879V can also be seen above, on the right (Figure 17) and this is predicted to be a conservative change as both residues are nonpolar. The score

for this substitution is 0 on the BLOSUM matrix which means the frequency of this alignment occurs as expected by chance and this is a common score for conservative changes. This position within the camostat-adapted FFM1 sequence was also flagged as having a mixed population thus potentially contributing to resistance as it does not need to have 100% sequencing reads in agreement to have a role.

This mutation occurs within an alpha helix which are structures stabilised by hydrogen bonds. Since no polar contacts are gained or lost during the mutation, it can be said that these hydrogen bonds are present to stabilise the alpha helix structure. There is, however, a difference in size between the two amino acids even though they are both small but since the surrounding area is very crowded these can often cause clashing between the DCP and surrounding residues within a 5 angstrom area. Valine has one more carbon in its side chain which results in clashing between residues Y789, S875, and T883 (Figure 18) which is significant as this will produce a conformational change. It is likely this change affects protein structure as a result of the overlapping which can result in denaturation, and the fact it has survived years of selection pressures indicates its presence must aid camostat resistance.

3.3.2 Adaption to Nafamostat

Nafamostat mesylate is a very similar drug to camostat mesylate except it has a much more rigid structure owing to it possessing more aromatic rings and being overall slightly shorter than camostat (Zhu, et al., 2021). It can still prevent SARS-CoV-2 infection by blocking TMPRSS-2 activity. A total of 12 mutations occurred in FFM1-nafamostat in the S protein which includes 6 residues that had low sequencing reads (>36%) as well as 3 mixed

populations and 3 base changes (Table 7). Positions 21765-21770 were flagged because they have a very low percentage of sequencing reads at that base as well as having relatively low sequencing depth compared to other positions. Position 22227 is the well documented A222V mutation (Figure 19) (Hodcroft E., 2021). Positions 22326, 22423 and 24107 showed mixed populations when sequenced. Position 23280 is a T573I mutation (Figure 27) and position 23757 is a T732I mutation (Figure 20).

Table 7: Amino acid changes to FFM1-nafamostat spike protein residues

This table shows the positions that were returned and stored in a .csv file for further investigation. The program was a python script which returned positions in the sequence that had less than 90% of sequencing reading in agreement at that base as well as flagging mixed variant populations. 'Pos' refers to the position within the viral genome and 'RefN' refers to the corresponding base in the reference genome. Depth refers to the total number of bases sequenced and aligned at a given reference base position. A, C, T, G refer to the DNA bases and the number below it is what percentage of that base was found at that read – as you can see for some of these it was below 90%. It then has the protein and what amino acid position this mutation occurs at within the spike protein.

Pos	RefN	Depth	A	C	G	T	Protein	AA Position
21765	T	4433	0	0	0	35.935	S	I68
21766	A	3554	19.584	0	0	0	S	I68
21767	C	3552	0	20.073	0	0	S	H69
21768	A	3531	19.541	0	0	0	S	H69
21769	T	3597	0	0	0	21.3233	S	H69
21770	G	3626	0	0	18.45	0	S	V70
22227	C	6138	0	0	0	99.5764	S	A222
22326	C	7217	0	83.788	0	16.1147	S	S255
22423	T	7220	0	13.38	0	86.5512	S	D287
23280	C	15940	0	0	0	99.542	S	T573
23757	C	12448	0	0	0	99.6787	S	T732
24107	C	5849	0	15.849	0	84.1511	S	L849

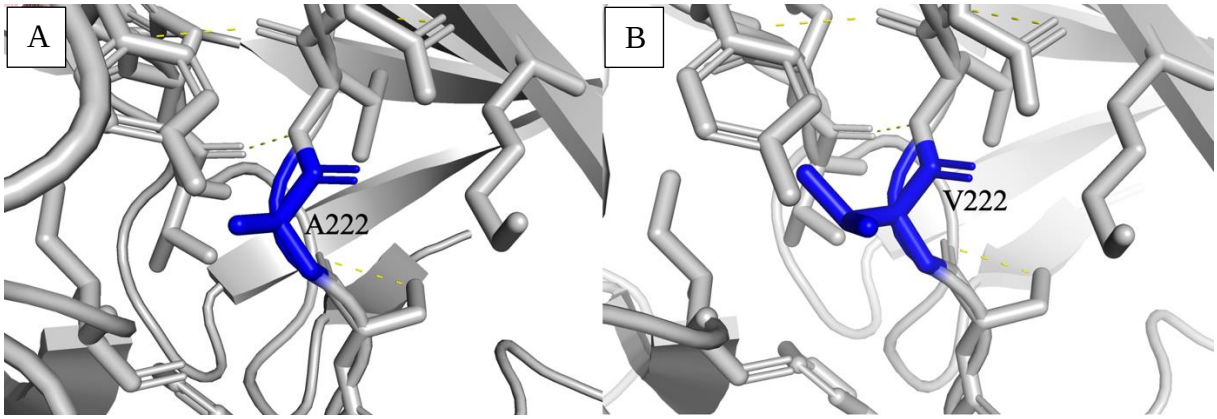


Figure 19: Structure of S-protein residue A222V

Panel A: WT S protein taken from the reference sequence. Residue A222 (blue) can be seen in the middle of a loop and does not form any hydrogen bonds with surrounding residues. Panel B: mutant S protein in response to nafamostat adaptation at position V222 (blue) which results in an increase in side chain length but this does not affect surrounding residues as the region is spacious.

The WT for the residue A222V can be seen above on the left (Figure 19). This residue is located in the S protein domain A which is not known to play any role in receptor binding; therefore, any mutations may not have such a major effect on protein structure or function. It is found within a loop in the protein structure and its surrounding area is quite crowded. The mutant for the residue A222V can also be seen above, on the right (Figure 19) and is expected to be a conservative change given both residues are nonpolar. The score for this substitution is 0 on the BLOSUM62 matrix which means the frequency of this alignment occurs as expected by chance, which is also expected for a conservative change.

Valine does have a slightly larger hydrocarbon chain, however; this does not result in overlapping due to the conformation as its projecting outwards across the surface of the protein. Despite this, the A222V mutation is well documented as the variant that dominated the European coronavirus cases in autumn 2020. This is unexpected given there are no noticeable changes that may affect the protein structure and/or function from my analysis.

However, DynaMut2 predicts a stability change of 0.3 kcal/mol which makes this a stabilising mutation. Therefore, since this mutation has managed to remain despite selection pressures, it must have a reason for being present and contributes to nafamostat resistance through this increased stability.

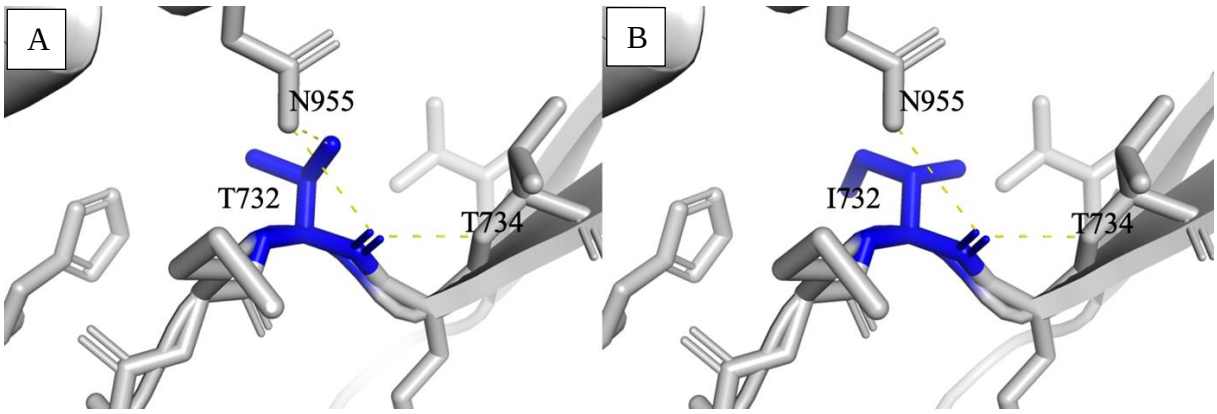


Figure 20: Structure of S protein residue T732I

Panel A: WT S protein taken from the reference sequence. Residue T732 (blue) can be seen in a loop within the protein structure, and it forms three hydrogen bonds (yellow) total with surrounding residues N955 which has two and one with T734. Panel B: mutant S protein in response to nafamostat adaptation at position I732 (blue) which results in one hydrogen bond (yellow) lost with N955 due to the difference in side chain properties.

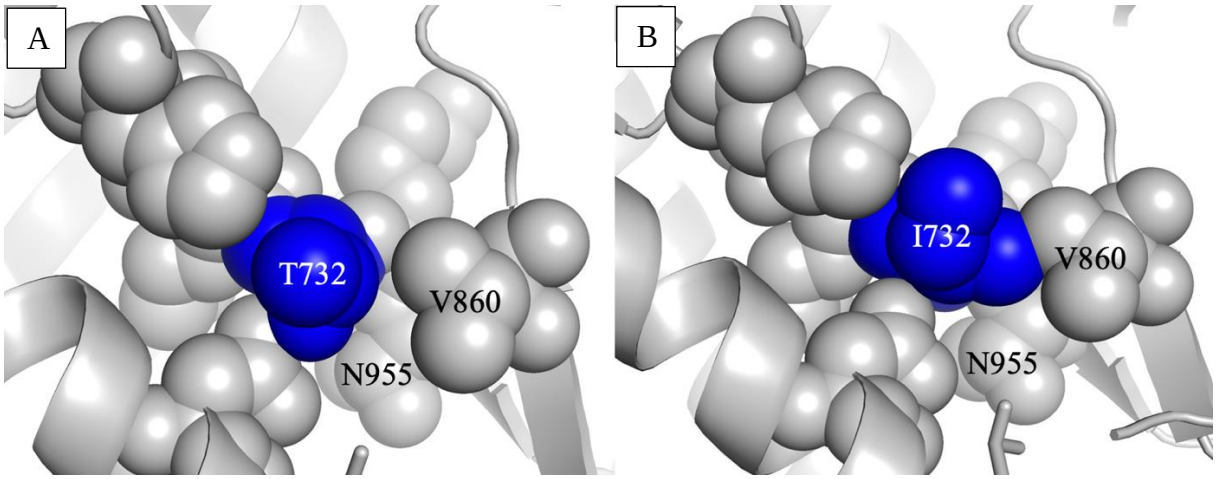


Figure 21: Structure of S protein residue T732I with sphere representation

Panel A: WT S protein taken from the reference sequence with sphere representation to show interactions with surrounding residues. Residue T732 (blue) can be seen in a crowded pocket of the protein and is found very close to residues V860 and N955. Panel B: mutant S protein in response to nafamostat adaptation at position I732 (blue) which causes side chain to increase in size causing it to overlap with both residues V860 and N955.

The WT for the residue T732I can be seen above on the left (Figure 20). It is found in a loop in the protein structure just before a beta sheet. It forms a total of three hydrogen bonds with its surrounding residues: two with the residue N955 which is a polar residue found in a neighbouring alpha helix and one with T734 which is also a polar residue found in a neighbouring beta sheet. The mutant for the residue T732I can also be seen above, on the right (Figure 20). This is a polar to nonpolar change which causes the chemical environment to become more hydrophobic. The BLOSUM matrix score for this substitution is -1 meaning this alignment was found to occur less often than by chance.

One hydrogen bond with N955 is lost owing to the difference in side-chain properties. It is likely these polar contacts are important in holding these structures in proximity however when the polar contact formed by the threonine side chain with N955 is lost, it does not affect this conformation since there is still another contact remaining with
N955. Furthermore, there is a difference in size between the two residues with isoleucine having one hydrocarbon extra in its side chain. This results in clashing between residues V860 and N955, the latter of which formed a hydrogen bond with T732 (Figure 21). As a consequence of this, it will likely produce a conformational change as it is now no longer able to fold as it should in this region which can lead to protein denaturation and therefore loss of function.

3.3.3 Overlapping mutations

Six mutations occurred in both FFM1-camostat and FFM1-nafamostat adaptation which are highlighted in Table 8. Since these drugs are both serine protease inhibitors, these positions

are of importance because they could cause resistance through similar mechanisms. It is also important to note that none of these overlapping mutations occur in the S protein which could mean several mutations acting together may possibly cause resistance to camostat and nafamostat. Position 3518 is unable to be structurally analysed as it is not present in the structure of NSP3. Positions 12797, 12704, and 17423 will be evaluated below. Positions 27272 and 28854 are also unable to be structurally analysed as there was no structures available for these proteins at the time of writing.

Table 8: Amino acid changes present in both FFM1-camostat and FFM1-nafamostat

This table shows the positions that were returned and stored in a .csv file for further investigation. The program was a python script which returned positions in the sequence that had less than 90% of sequencing reading in agreement at that base as well as flagging mixed variant populations. 'Pos' refers to the position within the viral genome and 'RefN' refers to the corresponding base in the reference genome. It then shows what the nucleotide change is when FFM1 is adapted to camostat and nafamostat and then it has the corresponding codon and amino acid changes. There is also a BLOSUM score which tells you the probability of this mutation occurring in nature (a positive score is given to likely substitutions). It includes which protein the mutation occurs and the amino acid position this occurs at.

Pos	RefN	Nucleotide change	Codon change	AA change	BLOSUM	Protein	AA position
3518	G	G > T	GUU > UUU	V > F	-1	NSP3	V266
12704	G	G > T	GUU > UUU	V > F	-1	NSP9	V7
12797	G	G > A	GGU > AGU	G > S	0	NSP9	G37
17423	A	A > G	UAU > UGU	Y > C	-2	NSP13 (Hel)	Y396
27272	T	T > C	GUU > GCU	V > A	0	ORF6	V24
28854	C	C > T	UCA > UUA	S > L	-2	ORF9c	Q41

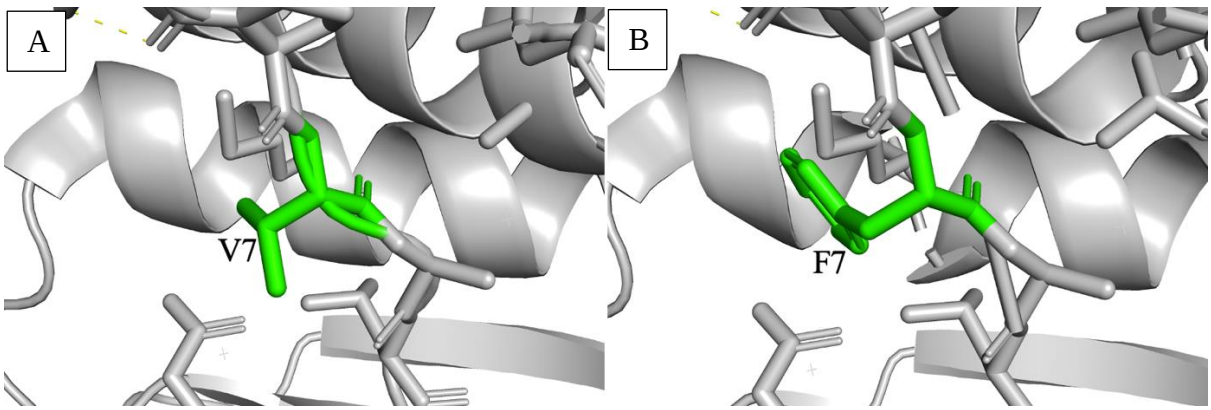


Figure 22: Structure of NSP9 residue V7F

Panel A: WT NSP9 protein taken from the reference sequence. Residue V7 (green) can be seen in a loop within the protein structure, and it does not form any hydrogen bonds with surrounding residues. Panel B: mutant NSP9 protein in response to camostat and nafamostat adaptation at position F7 (green) which results in a significant increase in side chain size as there is the addition of an aromatic ring.

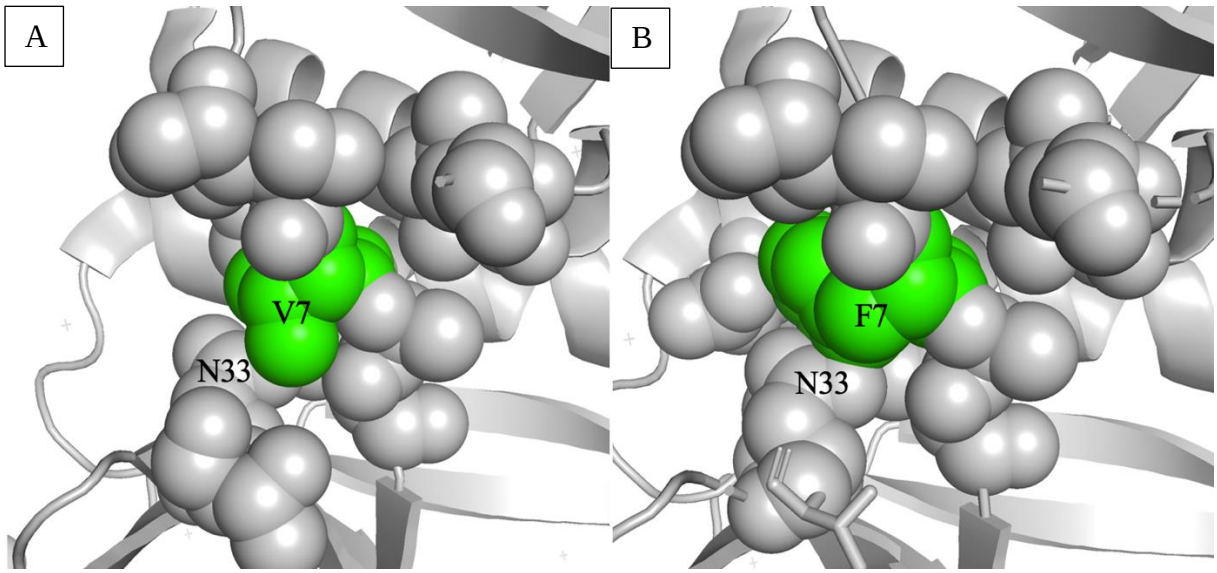


Figure 23: Structure of NSP9 residue V7F with sphere representation

Panel A: WT NSP9 protein taken from the reference sequence with sphere representation to show interactions with surrounding residues. Residue V7 (green) can be seen in a crowded pocket of the protein and is found very close to residue N33. Panel B: mutant NSP9 protein in response to camostat and nafamostat adaptation at position F7 (green) which causes side chain to increase in size causing it to overlap with N33.

The WT for the residue V7F can be seen above on the left (Figure 22). It is found within a loop in the protein structure and its surrounding area is quite crowded owing to several alpha helices. The mutant for the residue V7F can also be seen above, on the right (Figure 22). This mutation is conservative in the way that both the WT and the mutant are both hydrophobic residues. However, phenylalanine has an aromatic ring in its side chain and therefore there is a large increase in size from that of valine. This causes clashing between F7 and residues N33 and M101 which is present behind the structure (Figure 23). Nevertheless, this mutation is present in FFM1 at initial sequencing before adaptation to camostat and nafamostat, so it is not a mutation because of the adaptation process.

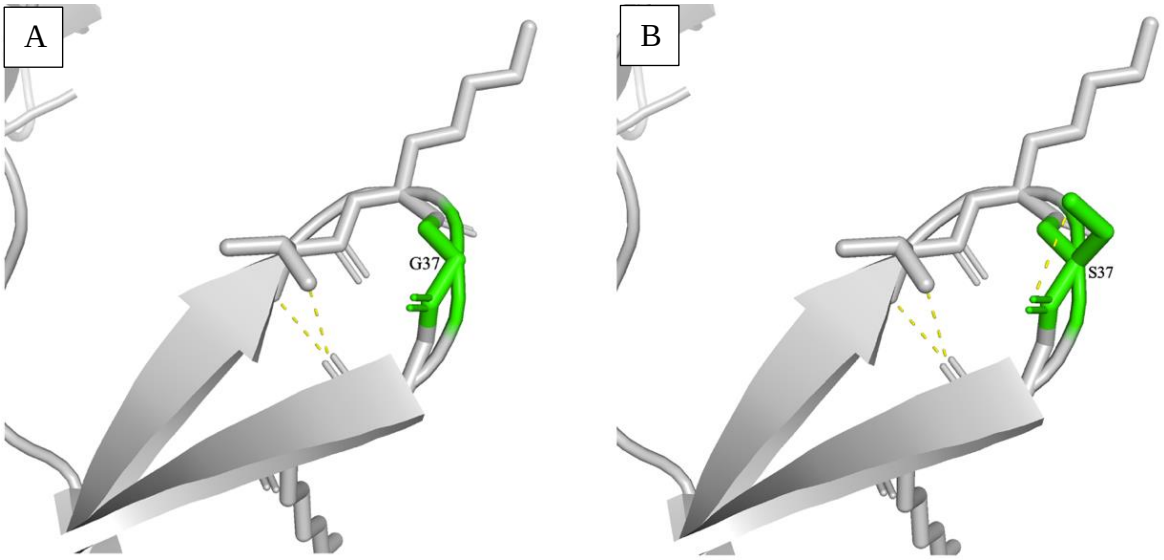


Figure 24: Structure of NSP9 residue G37S

Panel A: WT NSP9 protein taken from the reference sequence. Residue G37 (green) can be seen in a loop within the protein structure, and it does not form any hydrogen bonds with surrounding residues. Panel B: mutant NSP9 protein in response to camostat and nafamostat adaptation at position S37 (green) which results in formation of a hydrogen bond (yellow) with itself.

The WT residue of the mutation G37S can be seen above on the left panel (Figure 24). It is found within a loop in the protein structure, between the middle of two beta sheets. The mutated form of this can be seen above on the right at position S37 (Figure 24). This mutation is found to be a conservative change given glycine and serine are both similarly sized, and the BLOSUM matrix score for this substitution is 0 meaning the frequency of this alignment occurs as expected by chance, which is also expected for a conservative change. There is a change to the polar contacts with serine forming a hydrogen bond with its side chain and the results of this in terms of impact to its surrounding environment are minimal There is also a slight difference in size of side chains between glycine and serine with serine having a longer hydrocarbon chain. Therefore, it can be said that this mutation is unlikely to have any effect on protein structure or function which is as expected for a conservative change.

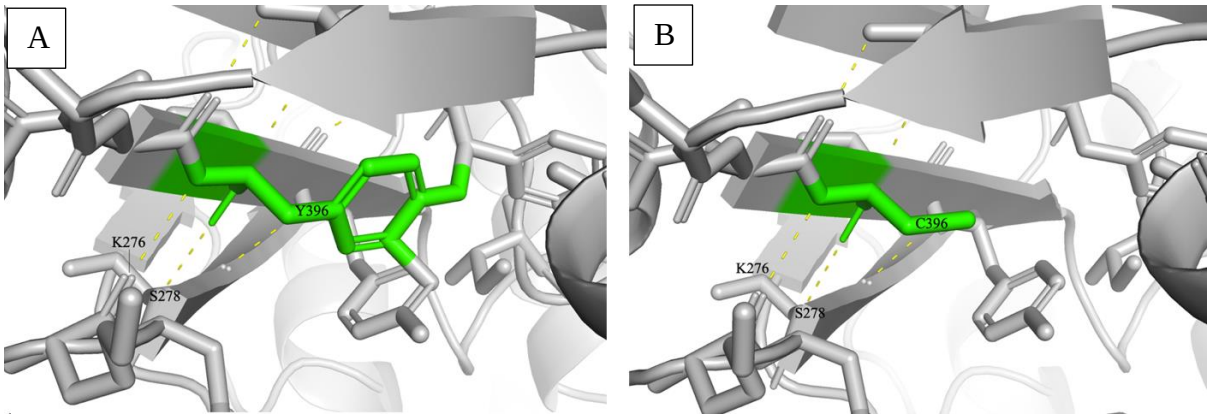


Figure 25: Structure of NSP13 residue Y396C

Panel A: WT NSP13 protein taken from the reference sequence. Residue Y396C (green) can be seen at the start of a beta sheet within the protein structure, and it forms two hydrogen bonds (yellow) with surrounding residues K276 and S278. Panel B: mutant NSP13 protein in response to camostat and nafamostat adaptation at position C396 (green) which results in a significant decrease in side chain size as there is the removal of an aromatic ring. No changes to hydrogen bonds (yellow).

The WT residue of the mutation Y396C can be seen above on the left panel (Figure 25). It is found within a beta sheet in the protein structure. It forms two hydrogen bonds, one with polar residue K276 and another with polar residue S278. The mutated form of this can be seen above on the right at position C396 (Figure 25). This change in amino acids is one from tyrosine which has an aromatic side chain to cysteine which has a hydroxyl containing side chain. The BLOSUM matrix score for this substitution is -2 which means that this alignment of residues was found to occur less often than by chance, meaning it is an unlikely substitution. Despite this, there are no changes to the number of hydrogen bonds and the result of the residue going from large to small means there is no observable clashing between residues in its surrounding environment. It is also worth noting that this particular mutation is present in FFM1 at initial sequencing before adaptation to camostat and nafamostat, so it is not a mutation as a result of the adaptation process.

Further to these mutations, 2 residues were found to overlap between FFM1-camostat and the virus control sequence (Table 9), one of which was the D614G mutation analysed previously in section 3.3.1. The virus control was FFM1 that was passaged in caco-2 cells in the absence of camostat and nafamostat. The mutations that arise in both the virus control and FFM1-camostat may just occur to the adaptation process of the virus being cultured in the cell line rather than as an adaptation to resistance. However, it is still important to investigate these further. Position 17678 is a substitution mutation T481M which will be examined below.

Table 9: Amino acid changes present in both FFM1-camostat and the virus control

This table shows the positions that were returned and stored in a .csv file for further investigation. The program was a python script which returned positions in the sequence that had less than 90% of sequencing reading in agreement at that base as well as flagging mixed variant populations. 'Pos' refers to the position within the viral genome and 'RefN' refers to the corresponding base in the reference genome. It then shows what the nucleotide change is when FFM1 is adapted to camostat as well as the virus control sequence and then it has the corresponding codon and amino acid changes. There is also a BLOSUM score which tells you the probability of this mutation occurring in nature (a positive score is given to likely substitutions). It includes which protein the mutation occurs and the AA position within the protein.

Pos	RefN	Nucleotide change	Codon change	AA change	BLOSUM	Protein	AA position
17678	C	C > T	ACG > AUG	T > M	-1	NSP13 (Hel)	T481
23403	A	A > G	GAU > GGU	D > G	-1	S	D614

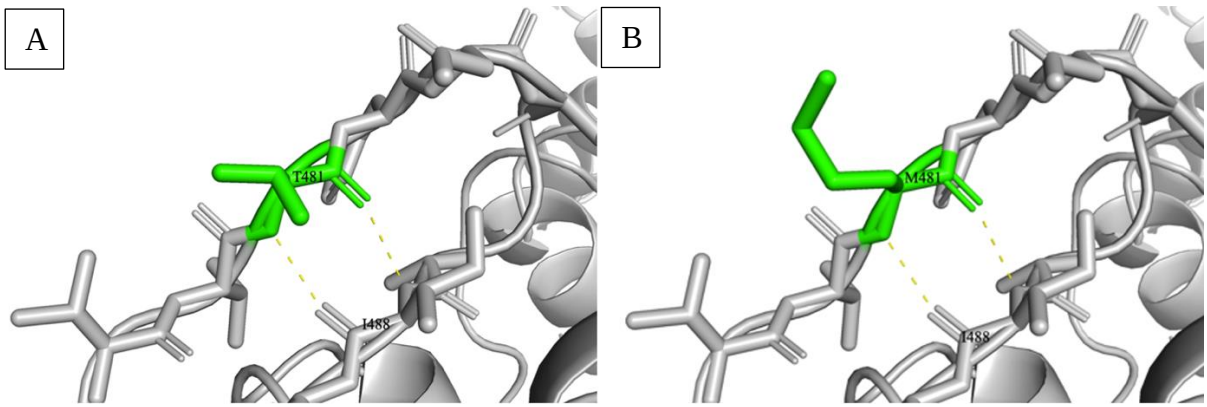


Figure 26: Structure of NSP13 residue T481M

Panel A: WT NSP13 protein taken from the reference sequence. Residue T481 (green) can be seen within a loop in the protein structure, and it forms two hydrogen bonds (yellow) with residue I488. Panel B: mutant NSP13

protein in response to camostat and nafamostat adaptation at position M481 (green) which results in an increase in side chain size but this projects out away from the protein structure. No changes to hydrogen bonds (yellow).

The WT residue of the mutation T481M can be seen above on the left panel (Figure 26). It is found within a loop in the protein structure. It forms two hydrogen bonds, both with the hydrophobic residue I488. The mutated form of this can be seen above on the right at position M481 (Figure 26). This amino acid change was from a polar threonine residue to a nonpolar methionine, which has a BLOSUM matrix score of -1 meaning this substitution was found to occur less often than by chance thus making this an unlikely change. The side chain projects outwards across the surface of the protein so there is no chance this will overlap with surrounding residues. It forms two polar contacts with a neighbouring loop, and these are not affected by this mutation which means these polar contacts are holding the two loops in proximity and therefore coordinating this section of the protein. It can be said that this mutation is unlikely to have any effect on protein structure or function and, therefore, the likelihood of this mutation causing camostat resistance is low. This result is as expected since a resistance mutation would not be present in the control sequence as well as the camostat-adapted FFM1.

In addition, one residue was found to overlap between FFM1-nafamostat and the virus control sequence (Table 10). As mentioned above, the control is FFM1 passaged with no drug present, so it is important to determine whether this mutation arises as a result of adaptations to being cultured in a cell line. Position 232380 is a substitution mutation in the S protein at T573I which will be examined below.

Table 10: Amino acid changes present in both FFM1-nafamostat and the virus control

This table shows the positions that were returned and stored in a .csv file for further investigation. The program was a python script which returned positions in the sequence that had less than 90% of sequencing reading in agreement at that base as well as flagging mixed variant populations. 'Pos' refers to the position within the viral genome and 'RefN' refers to the corresponding base in the reference genome. It then shows what the nucleotide change is when FFM1 is adapted to nafamostat as well as the virus control sequence and then it has the corresponding codon and amino acid changes. There is also a BLOSUM score which tells you the probability of this mutation occurring in nature (a positive score is given to likely substitutions). It includes which protein the mutation occurs and the AA position within the protein.

Pos	RefN	Nucleotide change	Codon change	AA change	BLOSUM	Protein	AA position
23280	C	C > T	ACU > AUU	T > I	-1	S	T573

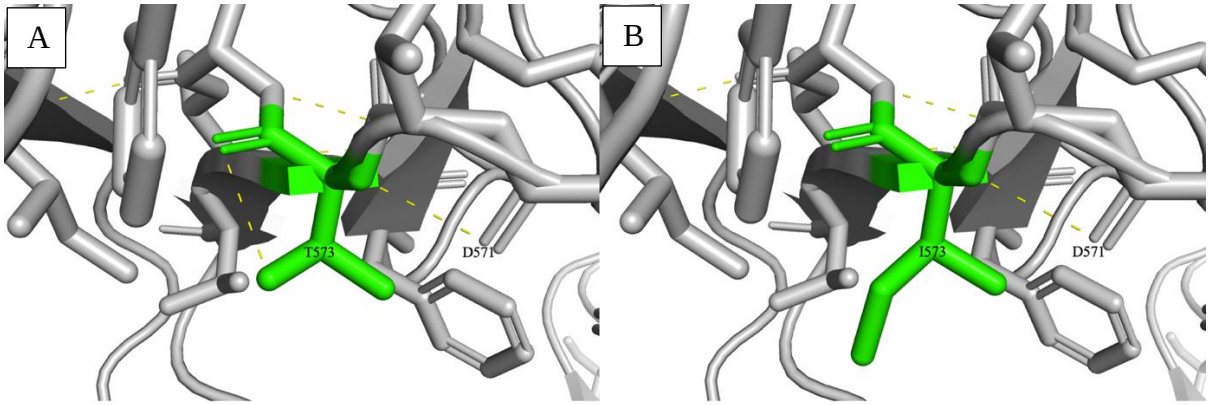


Figure 27: Structure of S protein residue T573I

Panel A: WT S protein taken from the reference sequence. Residue T573 (green) can be seen at the start of a beta sheet within the protein structure, and it forms one hydrogen bond (yellow) with its side chain. Panel B: mutant S protein in response to camostat and nafamostat adaptation at position I573 (green) which results in an increase in size of side chain, but this projects outwards away from the structure of the protein. The hydrogen bond (yellow) with itself is also lost.

The WT residue of the mutation T573I can be seen above on the left panel (Figure 27). It is found at the start of beta sheet in the protein structure and the side chain projects outwards across the surface of the protein. It forms one hydrogen bond with its side chain. The mutated form of this can be seen above on the right at position I573 (Figure 27) and is a polar to nonpolar change which causes the chemical environment to become more hydrophobic. The BLOSUM matrix score for this substitution is -1 indicating this alignment was found to occur less often than by chance. Threonine forms a polar contact with its side chain which is lost in

the isoleucine mutant which could result in increased flexibility since threonine is C-beta branched which causes bulkiness and restricts the movement.

There is also a difference in size between the WT and the mutant with isoleucine being slightly larger than threonine although, no clashing is observed between the DCP and surrounding residues within a 5 angstrom area since the side chain projects outwards across the surface of the protein. It can be said that this mutation is unlikely to have any effect on protein structure and/or function and therefore, the likelihood of this mutation causing nafamostat resistance is low. This result is as expected since a resistance mutation would not be present in the control sequence as well as the nafamostat-adapted FFM1.

4 Discussion

4.1 Key Findings

The first aim of this project was to investigate the determinants of pathogenicity in influenza, specifically between influenza A (H1N1) from the Spanish Flu pandemic of 1918 with influenza A (H1N1) from the Swine Flu pandemic of 2009. The second aim was to identify mutations in FFM1 that cause resistance to camostat and nafamostat: two potential COVID-19 drugs. It was important to establish whether these mutations had any effect on protein structure or function since if they do, these imply they are more likely to cause the differences in phenotypes between the similar viruses. They have been classified with regards to whether there will be a likely, possible, or unlikely effect.

Of the eight identified influenza A DCPs, A20T and L11Q were located close to the N-terminal domains of their proteins so there was no structure available at these positions, and R195K could not be found in any structures of PA-X. Of the 5 remaining DCPs, N233S, T293I, and Q250P were concluded to cause likely effects to the protein structure or function. The DCP S95R was concluded to possibly cause an effect to the protein structure or function and finally the DCP E583D was concluded to have no effect on protein structure or function.

In terms of analysis of FFM1-camostat, out of eleven S protein mutations just N658N was found to have no amino acid change, and K986N was not present in any S protein structural models. Four mutations were concluded to cause likely effects to the protein structure or

function: S50L, D614G, A653V and A879V. Finally, just T778I was found to have no effect on protein structure or function. In terms of analysis of FFM1-nafamostat, out of twelve S protein mutations just D287D resulted in a conservative change. Furthermore, six mutations were found to have low sequencing reads and were not present in the structure which is suggestive of a deletion mutation. Both A222V and T732I were concluded to cause likely effects to the protein structure or function, and the mutation T573I was thought to have no effect. Finally, the mutations S255D and T732I were unable to be analysed due their lack of presence in the S protein structure.

The results indicate that sequence specific mutations do occur between similar respiratory viruses and that these may well be causing the differences in phenotypes and therefore the differences in disease severity.

4.2 Implications of DCP Analysis of Influenza A

It is known that the influenza A DCP N233S is situated close to the receptor binding site (RBS) of haemagglutinin as well as being close to E190D and G225D which are known adaptive mutations in the RBS and changes the binding preference from avian to human receptor (Glaser, et al., 2005). Although, it is said that it is likely each HA subtype has a different set of mutations which causes this shift in binding preference. It was reported that an S223N mutation in H5N1 increased the binding affinity to human receptors which suggests a reason as to why 1918 H1N1 was so transmissible as this contained N223 (Du, et al., 2020). My research has shown much the opposite, that a N223S mutation decreases affinity for SAα-2,6Gal receptors which must have contributed to H1N1pdm09 being less widespread and

much less deadly than its 1918 counterpart. This agrees with my hypothesis that these DCPs could provide a reason for differences in disease severity between related viruses.

An analysis of murine monoclonal antibodies (MAbs) found that these broadly reactive antibodies were able to inhibit or bind N1 neuraminidase in a variety of subtypes. It was found that residues G249 and Q250 likely formed part of the binding site of MAb AG7C and that the substitutions G249K and Q250P affects this MAb (Rijal, et al., 2020). These site-specific substitution mutations were thought to stop binding to N1 neuraminidase and prevent inhibition of specific strains of N1 subtypes such as A/Brisbane/59/2007 which allowed them to infect host cells (Wan, et al., 2015). Therefore, this Q250P mutation must play a role in blocking the immune response and enabling infection into the host cell. Since the change from Q250 to P250 is destabilising and still present in several strains of H1N1, it must have provided a functional improvement that provided a benefit such as being able to block binding with MAbs as explained above. This would confer a difference in disease severity, however, this data does not fit in with the theory that the Spanish Flu H1N1 was more virulent than H1N1pdm09 as this is showing much the opposite case.

4.3 Limitations and Recommendations for DCP Analysis of Influenza A

The reliability of this data is impacted by a shortage of sequenced strains of H1N1 pre 1977 which marked the development of Sanger sequencing (Sanger, et al., 1977). As a result, there were many more strains available for VAT analysis for group B strains of H1N1pdm09 than for the group A strains from pre-1950. If more sequences were available, perhaps a wider image could be seen which may reveal more DCPs for even more influenza proteins. Otherwise, a

different time period could have been investigated such as 1950-2008 or 2010-present which may have acquired more sequences. Furthermore, three identified DCPs could not be analysed for two reasons: being located too close to the protein N-terminal domain such as the case for A20T from haemagglutinin and L11Q from PB1-F2, the other reason was that the protein was not fully modelled so had low sequence identity such as the case for the PA-X DCP R195K. There is also little information out there on these mutations. Therefore, it can be difficult to draw specific conclusions given the lack of research on these residues. Despite these limitations, my results are still valid because they identify differences in related viruses in the form of DCPs that may account for differences in disease severity. Furthermore, several of these DCPs are proven to cause conformational or stability changes which affect virus infectivity.

Additionally, it is interesting to see a large difference between the number of DCPs found in influenza A(H1N1) compared with researchers who utilised the same method on other viruses. Bojkova (2021) found there to be 1243 DCPs when comparing the 22 SARS-CoV-2 proteins with SARS-CoV which makes up around 89% of positions encoding different amino acids. This number represents 13% of all residues in the SARS-CoV-2 genome which is a significant increase on my data which shows my DCPs represent just 2.63% of all residues in the influenza A genome. This could relate to variation in the virus sequences and the fact that the compared H1N1 sequences from all years had very high conservation levels. Importantly, Influenza RNA from 1918 was collected from lung tissue samples of a woman that had been preserved in Alaskan permafrost and the complete sequence was published 87 years later. Moreover, the group 1 sequences from 1918-1950 represents more than 100 years of data and at the time were sequenced using older technologies before the widely accepted Sanger

method was established. It is the combination of these factors that could have prevented more DCPs from being identified.

If more time was available, it may be worth investigating the differences in how haemagglutinin binds to avian receptors compared with human receptors and what structural changes occur. From this, it might give more information on mutations in this area such as N223S and what the specific effects on binding would be as it can be concluded this DCP confers to differences in disease severity, but the specific mechanism is unknown. Another limitation comes from VAT which is only able to find DCPs between two groups but if it was able to identify DCPs between three or more groups there may have been more DCPs found. For example, I would have compared group A as 1918-1950, group B as 1951-2008, and group C as 2009-present which would have provided a complete timeline to show the effects of changes to H1N1 proteins over 100 years. Consequently, there were many obstacles when it came to comparing H1N1 sequences, especially when choosing the right time groups as some of these had >95% similarity to each other so was difficult to gather DCPs from the program.

4.4 Implications of FFM1 drug resistance analysis

The FFM1 strain of SARS-CoV-2 was cultured through Caco-2 cells in increasing concentrations of camostat and nafamostat which resulted in increasing resistance over every passage. Human SARS-CoV-2 entry is mediated by the spike (S) glycoprotein which in turn is highly dependent on ACE2 and TMPRSS-2 expression. Cleavage sites within the S protein are processed by TMPRSS-2 which initiates priming to fuse the S protein with the host membrane. If TMPRSS-2 is blocked by camostat or nafamostat, the virus cannot be activated and cannot

go on to infect the host cells. Therefore, there was a focus on S protein mutations. It is important to add that these mutations may be working together to exhibit resistance to camostat or nafamostat.

The mutation S50L in the SARS-CoV-2 spike protein was predicted to result in conformational changes to the protein structure. One paper analysing several SARS-CoV-2 S protein substitutions found mutants S50L and H49Y produced the largest reduction in total free energy in its open state at -7.34 and -5.29 kcal/mol respectively (Laha, et al., 2020). This infers a large stability change occurs and that this variant is stable in nature. Another paper also calculates the predicted stability change of S50L to be $\Delta\Delta G = -2.614$ kcal/mol which results in stabilising effects on the entire S protein (Teng, et al., 2021). The effects of this stability change are an increased resistance of SARS-CoV-2 which fits in with my hypothesis that this mutation could confer to camostat resistance.

Researchers have shown that D614G spike protein mutation alters the fitness of SARS-CoV-2 (Plante, et al., 2021). It also marked the start of the G clade as this mutation conferred higher infection and transmission rates. It has been suggested that this mutation increases the transmissibility of SARS-CoV-2 because of a higher viral load of G614 observed in COVID-19 patients (Plante, et al., 2021). In addition, increased viral infectivity in the upper respiratory tract was observed through the loss of a hydrogen bond with T859 which can be seen in Figure 13. The result of which shifts the conformation of the RBD to an open configuration which increases binding with ACE2 thus increasing the ability of SARS-CoV-2 to establish infection (Yurkovetskiy, et al., 2020). However, D614G does not occur in FFM1 before sequencing since it is only present in FFM1-camostat and not FFM1-nafamostat. Therefore, it may play

somewhat of a role in camostat resistance as perhaps the higher viral load shown in G614 may increase the likelihood of resistance occurring and being carried over future passages.

There is little published research on the S protein variant A653V. However, it is found to be near a furin-like cleavage site (FCS) that has been postulated to hold responsibility for the high infectivity and transmissibility of SARS-CoV-2 (Coutard, et al., 2020). The FCS PRRAR is inserted between residues 680-690 in the spike protein and plays a critical role in SARS-CoV-2 replication and pathogenesis (Peacock, et al., 2021). A653V is found deep in a hydrophobic pocket in the protein and is likely to interact with residues in the FCS which may result in side chain alterations when the FCS is not inserted or is inserted with a mutation. These results build on the predicted conformational changes expected by A653V which are likely to cause differences in resistance to camostat.

The S protein A222V mutation is another example of a very well researched mutation having emerged in the summer of 2020 which formed the GV clade. It is found in the NTD of the spike protein and does not have a role in receptor binding that is known although, it is understood to affect protein stability (Hodcroft, et al., 2021). One paper predicts A222V to be stabilising with a predicted stability change of 0.95 kcal/mol by the SDM server and 0.91 kcal/mol by the DUET server (Jacob, et al., 2021). This supports my stability change data predicted by DynaMut2 that suggests this mutation is stabilising at a measurement of 0.3 kcal/mol. The same paper suggests that the A222V mutation could be working in combination with other mutations such as D614G and S477N to affect spike stability and increase ACE2 binding (Jacob, et al., 2021). However, this does not support my data since the A222V mutation occurs in FFM1-nafamostat only and D614G only occurs in FFM1-camostat only.

Therefore, these mutations do not occur at the same time so this cannot be the reason for having resistance mutations.

The S protein mutation T573I was observed alongside H49Y and D614G in a Mexican lineage of SARS-CoV-2 (Sixto-Lopez, et al., 2021). T573I is a polar to nonpolar change which changes the surroundings to become more hydrophobic and so offsetting the chemical nature of the protein. This article suggests that I573 was a more compact version of the protein and so this mutation must increase structural stability as well as the region surrounding it, specifically amino acids I587-I720 (Sixto-Lopez, et al., 2021). These results build on the theory that mutations occurring away from the RBD are likely to have structural or conformational changes. It was also found in this article that T573I increased the magnitude and direction of the protein, thus increasing mobility. Even though this mutation is beneficial for the S protein, it cannot be said that it confers to nafamostat-resistance, otherwise it would not also be present in the virus control.

4.5 Limitations and Recommendations of FFM1 drug resistance analysis

Lots of information can be found about SARS-CoV-2 variants as there are hundreds of known variants. However, for some of the less frequently observed variants there is little to no information on these which makes it difficult to relate my observations to the existing literature. This also means I am unable to provide alternative explanations of my results. Another limitation came from the availability, or lack thereof, of SARS-CoV-2 structures. Many of the smaller proteins such as NSP2 and NSP6 were not present in PDB and produced a low confidence score in Phyre2, so AlphaFold was utilised, which is an AI system that predicted

several protein structures of SARS-CoV-2 including ones that were previously unknown such as ORF3a. However, even this provided its limitations when it came to ORF6 and ORF9c which had no structures available at the time of research. Consequently, I was unable to model the mutations that were found in these proteins.

Future work would include the potential to isolate these mutations to test whether they are in fact capable of causing resistance to camostat or nafamostat. Since resistance testing has been carried out to identify the mutations using sequencing methods, the next step would be using cell-based assays to determine the effect on viral drug susceptibility. This would involve using the mutations that have been predicted in this research to potentially cause resistance to camostat and/or nafamostat and to check that these do in fact cause resistance. These mutations would need to be induced individually at first to see if they are able to cause mutation on their own or several mutations would be expressed at the same time to see if these mutations act together to produce this resistance. The result of this would be able to aid future drug design and allow for similar methods to be carried out on other existing drugs in order to treat infectious diseases. The hope for COVID-19 is that it can be used to reduce mortality rates of infection especially in countries that have not fully equipped a vaccine programme.

4.6 Conclusion

To summarise, this project has successfully managed to identify sequence associated differences between related viruses and established a connection to phenotypic differences. *In silico* modelling revealed four DCPs were found in H1N1 proteins that resulted in conformational changes that had not been removed by natural selection. Their presence

indicates a beneficial purpose for this which I have predicted to be a reason as to why the 1918 Spanish Flu was far deadlier and more easily transmissible when compared to that of the 2009 H1N1 Swine Flu pandemic. The use of the VAT provided by Wass-Michaelis Lab at the University of Kent has proved it can be applied to various related viruses to identify these differentially conserved positions.

Additionally, *in silico* modelling revealed four FFM1 mutations have been identified that could cause camostat resistance amongst two mutations that are thought to cause resistance to nafamostat. These mutations all result in conformational changes or changes to the way the side chain interacts with surrounding ligands. As above, these are not deleterious mutations as they have not been removed by natural selection so therefore must cause resistance by an unknown mechanism.

My project has successfully identified variants of interest that are likely to be the cause of differential phenotypes between two closely related viruses, between that of H1N1 from 1918 with H1N1pdm09, or that of SARS-CoV and SARS-CoV-2. Therefore, my research question has been answered, however because of my findings, it has revealed many more questions that need to be answered through further investigation and future studies. There are also several ways this data can be taken and used for further study including for public health initiatives to help reduce spreading of viral respiratory diseases throughout communities.

Bibliography

Abdelrahman, Z., Li, M. & Wang, X., 2020. *Comparative Review of SARS-CoV-2, SARS-CoV, MERS-CoV, and Influenza A Respiratory Viruses.* [Online]
Available at: https://doi.org/10.3389/fimmu.2020.552909
[Accessed 19 August 2021].

Bao, Y., Bolotov, P., Dernovoy, D. & Kiryutin, B., 2008. The Influenza Virus Resource at the National Center for Biotechnology Information. *J. Virol.,* 82(2), pp. 596-601.

Barry, J. M., 2005. *The great influenza.* New York: Penguin Books.
Berman, H. M. et al., 2000. The Protein Data Bank. *Nucleic Acids Research,* 28(1), pp. 235-242.

Bojkova, D., McGreig, J. E., McLaughlin, K. & Masterson, S. G., 2021. Differentially conserved amino acid positions may reflect differences in SARS-CoV-2 and SARS-CoV behaviour. *Bioinformatics,* 37(16), pp. 2282-2288.

Boonstra, S. et al., 2018. Hemagglutinin-Mediated Membrane Fusion: A Biophysical Perspective. *Annual Review of Biophysics,* Volume 47, pp. 153-173.

Boopathi, S., Poma, A. B. & Kolandaivel, P., 2020. Novel 2019 coronavirus structure, mechanism of action, antiviral drug promises and rule out against its treatment. *J Biomol Struct Dyn.,* pp. 1-10.

Brian, D. A. & Baric, R. S., 2005. Coronavirus genome structure and replication. *Curr Top Microbiol Immunol.,* Volume 287, pp. 1-30.

Caldaria, A. et al., 2020. COVID-19 and SARS: Differences and similarities. *Dermatologic Therapy,* Volume e13395.

Capra J.A., S. M., 2007. Predicting functionally important residues from sequence conservation. *Bioinformatics,* Volume 23, pp. 1875-1882.

Center for Systems Science and Engineering (CSSE) , 2021. *COVID-19 Dashboard.* [Online] Available at: https://www.arcgis.com/apps/dashboards/bda7594740fd40299423467b48e9ecf6 [Accessed 2021 August 19].

Chan-Yeung, M. & Xu, R.-H., 2003. SARS: epidemiology. *Respirology,* 8(1), pp. S9-S14.

Chen, W., Calvo, P. A., Malide, D. & Gibbs, J., 2001. A novel influenza A virus mitochondrial protein that induces cell death. *Nat. Med.*, 7(12), pp. 1306-1312.

Chin, A. W. H. et al., 2016. Recombinant influenza virus with a pandemic H2N2 polymerase complex has a higher adaptive potential than one with seasonal H2N2 polymerase complex. *Journal of General Virology,* 97(3), pp. 611-619.

Chinese Preventative Medicine Association, 2020. An update on the epidemiological characteristics of novel coronavirus pneumonia (COVID-19). *Chin J Epidemiol.,* 41(2), pp. 139-144.

Chousterman, B. G., Swirski, F. K. & Weber, G. F., 2017. Cytokine storm and sepsis disease pathogenesis. *Seminars Immunopathol.*, 39(5), pp. 517-528.

Clarke, D. M., Loo, T. W. & MacLennan, D. H., 1990. Functional Consequences of Alterations to Polar Amino Acids Located in the Transmembrane Domain of the Ca2+-ATPase of Sarcoplasmic Reticulum. *The Journal of Biological Chemistry,* 265(11), pp. 6262-6267.

Coutard, B., Valle, C., de Lamballerie, X. & Canard, B., 2020. The spike glycoprotein of the new coronavirus 2019-nCoV contains a furin-like cleavage site absent in CoV of the same clade. *Antiviral Res.,* Volume 176, p. 104742.

Dawood, F. S. et al., 2009. Emergence of a novel swine-origin influenza A(H1N1) virus in humans. *N Engl J Med.,* 360(25), pp. 2605-2615.

Dou, D. et al., 2018. Influenza A Virus Cell Entry, Replication, Virion Assembly and Movement. *Frontiers in Immunology,* Volume 9, p. 1581.

Du, W., Wolfert, M. A., Peeters, B. & van Kuppeveld, F. J. M., 2020. Mutation of the second sialic acid-binding site of influenza A virus neuraminidase drives compensatory mutations in hemagglutinin. *PLoS Pathogens,* 16(8), p. e1008816.

Eurostat, 2020. *Health in the European Union - facts and figures,* s.l.: Europa.

Fehr, A. R. & Perlman, S., 2015. Coronavirus: An Overview of Their Replication and Pathogenesis. *Coronaviruses,* Volume 1282, pp. 1-23.

Frost, W. H., 1920. Statistics of Influenza Morbidity: With Special Reference to Certain Factors in Case Incidence and Case Fatality.. *JSTOR,* 35(11), pp. 584-597.

Gaudin, Y., Ruigrok, R. W. & Brunner, J., 1995. Low-pH induced conformational changes in viral fusion proteins: implications for the fusion mechanism. *Journal of General Virology,* 76(7), pp. 1541-1556.

Glaser, L., Stevens, J. & Zamarin, D., 2005. A single amino acid substitution in 1918 influenza virus hemagglutinin changes receptor binding specificity.. *Journal of Virology,* Volume 79, pp. 11533-11536.

Glezen, W. P., 1996. Emerging infections: pandemic influenza. *Epidemiol Rev,* 18(1), pp. 64-76.

Guan, Y., Zheng, B. J., He, Y. Q. & Liu, X. L., 2003. Isolation adn charachterization of viruses related to the SARS coronavirus from animals in southern China.. *Science,* Volume 302, pp. 276-278.

Hodcroft E., e. a., 2021. Emergence and spread of a SARS-CoV-2 variant through Europe in the summer of 2020. *Nature,* 595(7869), pp. 707-712.

Hodcroft, E. B., Zuber, M., Nadeau, S. & Vaughan, T. G., 2021. Spread of a SARS-CoV-2 variant through Europe in the summer of 2020. *Nature,* Volume 595, pp. 707-712.

Hoffmann, M. et al., 2021. Camostat mesylate inhibits SARS-CoV-2 activation by TMPRSS2-related proteases and its metabolite GBPA exerts antiviral activity. *EBioMedicine,* Volume 65, p. 103255.

Ito, T. et al., 1998. Molecular basis for the generation in pigs of influenza A viruses with pandemic potential.. *J Virol.*, Volume 72, pp. 7367-7373.

Jacob, J. J. et al., 2021. Evolutionary Tracking of SARS-CoV-2 Genetic Variants Highlights an Intricate Balance of Stabilizing and Destabilizing Mutations. *ASM Journals*, 12(4).

Jumper, J. et al., 2020. *Computational predictions of protein structure associated with COVID-19.* [Online]
Available at: https://deepmind.com/research/open-source/computational-predictions-of-protein-structures-associated-with-COVID-19
[Accessed 23 August 2021].

Kawaoka, Y., Krauss, S. & Webster, R. G., 1989. Avian-to-human transmission of the PB1 gene of influenza A viruses in the 1957 and 1968 pandemics. *Journal of Virology*, 63(11), pp. 4603-4608.

Kelley, L. A. et al., 2015. The Phyre2 web portal for protein modeling, prediction and analysis. *Nature Protocols*, Volume 10, pp. 845-858.

Kent, W. J. et al. 2002. The human genome browser at UCSC. *Genome Res*, Volume 12(6), pp. 996-1006.

Khrustalev, V. V. et al., 2020. Translation-Associated Mutational U-Pressure in the First ORF of SARS-CoV-2 and Other Coronaviruses. *Frontiers in Microbiology*, Volume 11, p. 559165.

Killerby, M. E. et al., 2020. Middle East respiratory syndrome coronavirus transmission. *Emerg Infect Dis.*, Volume 26, pp. 191-198.

Laha, S., Chakraborty, J., Das, S. & Manna, S. K., 2020. Characterizations of SARS-CoV-2 mutational profile, spike protein stability and viral transmission. *Infecr Genet Evol.*, Volume 85, p. 104445.

Lakadamyali, M., Rust, M. J. & Zhuang, X., 2004. Endocytosis of influenza viruses. *Microbes and Infection*, 6(10), pp. 929-936.

Leikina, E. et al., 2002. Reversible stages of the low-pH-triggered conformational change in influenza virus hemagglutinin. *The EMBO Journal*, 21(21), pp. 5701-5710.

Liu, B., Li, M. & Zhou, Z., 2020. Can we use interleukin-6 (IL-6) blockade for coronavirus disease 2019 (COVID-19)-induced cytokine release syndrome (CRS)?. *J Autoimmun.*, Volume 111, p. 102452.

Liu, Y., Gayle, A. A., Wilder-Smith, A. & Rocklov, J., 2020. The reproductive number of COVID-19 is higher compared to SARS coronavirus. *Journal of Travel Medicine,* 27(2), p. taaa021.

López, G., Valencia, A. & Tress, M. L., 2007. firestar—prediction of functionally important residues using structural templates and alignment reliability. *Nucelic Acids Research,* Volume 35, pp. 573-577.

Mena, I. et al., 2016. Origins of the 2009 H1N1 influenza pandemic in swine in Mexico. *eLife,* Volume 5, p. e16777.

More, A. F. et al., 2020. The Impact of a Six-Year Climate Anomaly on the "Spanish Flu" Pandemic and WWI. *Geohealth,* 4(9), p. e2020GH000277.

Noda, T. & Kawaoka, Y., 2010. Structure of Influenza Virus Ribonucleoprotein Complexes and Their Packaging into Virions. *Rev Med Virol.,* 20(6), pp. 380-391.

Ord, M., Faustova, I. & Loog, M., 2020. The sequence at Spike S1/S2 site enables cleavage by furin and phospho-regulation in SARS-CoV2 but not in SARS-CoV1 or MERS-CoV. *Scientific Reports,* Volume 10, p. 16944.

Palese, P., Tobita, K., Ueda, M. & Compans, R. W., 1974. Characterization of temperature sensitive influenza virus mutants defective in neuraminidase. *Virology,* 61(2), pp. 397-410.

Pappalardo M., e. a., 2016. Conserved differences in protein sequence determine the human pathogenicity of Ebolaviruses. *Sci. Rep.,* Volume 6, p. 23743.

Peacock, T. P. et al., 2021. The furin cleavage site in the SARS-CoV-2 spike protein is required for transmission in ferrets. *Nature Microbiology,* Volume 6, pp. 899-909.

Peiris, J. S. M., Hui, K. P. Y. & Yen, H.-L., 2010. Host response to influenza virus: protection versus immunopathology. *Curr Opin Immunol.,* 22(4), pp. 475-481.

Plante, J. A., Liu, Y., Liu, J. & Xia, H., 2021. Spike mutation D614G alters SARS-CoV-2 fitness. *Nature,* Volume 592, pp. 116-121.

Potter, B. I. et al., 2019. Evolution and rapid spread of a reassortant A(H3N2) virus that predominated the 2017–2018 influenza season. *Virus Evolution,* 5(2), p. vez046.

Rausell, A., Juan, D. & Pazos, F., 2010. Protein interactions and ligand binding: from protein subfamilies to functional specificity. *Proc Natl Acad Sci U S A,* 107(5), pp. 1995-2000.

Ren, W., Li, W., Yu, M. & Hao, P., 2006. Full-length genome sequences of two SARS-like coronaviruses in horseshoe bats and genetic variation analysis.. *Journal of General Virology,* Volume 87, pp. 3355-3359.

Rijal, P., Wang, B. B., Tan, T. K. & Schimanski, L., 2020. Broadly Inhibiting Antineuraminidase Monoclonal Antibodies Induced by Trivalent Influenza Vaccine and H7N9 Infection in Humans. *J Virol.*, 94(4), pp. e01182-19.

Rodrigues, C. H. M., Pires, D. E. V. & Ascher, D. B., 2020. DynaMut2: Assessing changes in stability and flexibility upon single and multiple point missense mutations. *Tools for Protein Science,* 30(1), pp. 60-69.

Rogers, G. N., Pritchett, T. J., Lane, J. L. & Paulson, J. C., 1983. Differential sensitivity of human, avian, and equine influenza A viruses to a glycoprotein inhibitor of infection: selection of receptor specific variants. *Virology,* 131(2), pp. 394-408.

Rouse, B. T. & Sehrawat, S., 2010. Immunity and immunopathology to viruses: what decides the outcome?. *Nature Reviews Immunology,* Volume 10, pp. 514-526.

Sanger, F., Nicklen, S. & Coulson, A. R., 1977. DNA sequencing with chain-terminating inhibitors. *Proc Natl Acad Sci U S A.*, 74(12), pp. 5463-5467.

Schrödinger, L. & DeLano, W., 2020. *PyMOL.* [Online]
Available at: http://www.pymol.org/pymol
[Accessed 23 August 2021].

Sixto-Lopez, Y., Correa-Basurto, J., Bello, M. & Landeros-Rivera, B., 2021. Structural insights into SARS-CoV-2 spike protein and its natural mutants found in Mexican population. *Scientific Reports,* Volume 11, p. 4659.

Song, H.-D.et al., 2005. Cross-host evolution of severe acute respiratory syndrome coronavirus in palm civet and human. *PNAS,* 102(7), pp. 2430-2435.

Stencel-Baerenwald, J. E. et al., 2014. The sweet spot: defining virus-sialic acid interactions. *Nature Reviews Microbiology,* 12(11), pp. 739-749.

Takemoto, D. K., Skehel, J. J. & Wiley, D. C., 1996. A surface plasmon resonance assay for the binding of influenza virus hemagglutinin to its sialic acid receptor. *Virology,* 217(2), pp. 452-458.

Teng, S. et al., 2021. Systemic effects of missense mutations on SARS-CoV-2 spike glycoprotein stability and receptor-binding affinity. *Briefings in Bioinformatics,* 22(2), pp. 1239-1253.

Thompson, W. W. et al., 2004. Influenza-associated hospitilizations in the United States. *JAMA,* 292(11), pp. 1333-1340.

Trifonov, V., Khiabanian, H., Greenbaum, B. & Rabadan, R., 2009. The origin of the recent swine influenza A(H1N1) virus infecting humans. *Euro Surveill.,* 14(17), p. pii=19193.

Tyrrell, C. S. B., Allen, J. L. Y. & Carson, G., 2017. Influenza and other emerging respiratory viruses. *Medicine,* pp. 781-787.

V'kovski, P. et al., 2020. Coronavirus biology and replication: implications for SARS-CoV-2. *Nature Reviews Microbiology,* Volume 19, pp. 155-170.

Wan, H., Yang, H., Shore, D. A. & Garten, R. J., 2015. Structural characterization of a protective epitope spanning A(H1N1)pdm09 influenza virus neuraminidase monomers. *Nat Commun.,* 10(6), p. 6114.

Wass, M. N., Kelley, L. A. & Sternberg, M. J. E., 2010. 3DLigandSite: predicting ligand-binding sites using similar structures. *Nucleic Acids Res.,* Volume 38, pp. 469-473.

Webster, R. G. et al., 1993. Influenza--a model of an emerging virus disease. *Intervirology*, 35(1-4), pp. 16-25.

World Health Organization, 2011. *Pandemic influeza A (H1N1).* [Online]
Available at:
https://www.who.int/csr/resources/publications/swineflu/h1n1_donor_032011.pdf
[Accessed 21 August 2021].

World Health Organization, 2015. *Summary of probable SARS cases with onset of illness from 1 November 2002 to 31 July 2003.* [Online]
Available at: https://www.who.int/publications/m/item/summary-of-probable-sars-cases-with-onset-of-illness-from-1-november-2002-to-31-july-2003
[Accessed 21 August 2021].

World Health Organization, 2017. *Up to 650 000 people die of respiratory diseases linked to seasonal flu each year.*. [Online]
Available at: https://www.who.int/news/item/13-12-2017-up-to-650-000-people-die-of-respiratory-diseases-linked-to-seasonal-flu-each-year
[Accessed 19 August 2021].

World Health Organization, 2020. *Cumulative number of confirmed human cases for avian influenza A(H5N1) reported to WHO, 2003-2020.* [Online]
Available at:
https://www.who.int/influenza/human_animal_interface/2020_MAY_tableH5N1.pdf
[Accessed 19 August 2021].

World Health Organization, 2021. *Avian Influenza Weekly Update Number 805.* [Online]
Available at: https://www.who.int/docs/default-source/wpro---documents/emergency/surveillance/avian-influenza/ai-20210813.pdf?sfvrsn=30d65594_157
[Accessed 20 August 2021].

World Health Organization, 2021. *MERS Situation Update.* [Online]
Available at: http://www.emro.who.int/health-topics/mers-cov/mers-outbreaks.html
[Accessed 21 August 2021].

World Health Organization, 2021. *WHO Coronavirus (COVID-19) Dashboard.* [Online] Available at: https://covid19.who.int/
[Accessed 21 August 2021].

Xia, S., Lan, Q., Su, S. & Wang, X., 2020. The role of furin cleavage site in SARS-CoV-2 spike protein-mediated membrane fusion in the presence or absence of trypsin. *Signal Transduction and Targeted Therapy,* 5(92).

Xiong, X., Coombs, P. J., Martin, S. R. & Liu, J., 2013. Receptor binding by a ferret-transmissible H5 avian influenza virus. *Nature,* 497(7449), pp. 392-396.

Ye, Q., Wang, B. & Mao, J., 2020. The pathogenesis and treatment of the `Cytokine Storm' in COVID-19. *Journal of Infection,* 80(6), pp. 607-613.

Yurkovetskiy, L., Wang, X., Pascal, K. E. & Tomkins-Tinch, C., 2020. Structural and functional analysis of the D614G SARS-CoV-2 spike protein variant. *Cell,* Volume 183, pp. 739-751.

Zhang, T., Wu, Q. & Zhang, Z., 2020. Probably Pangolin Origin of SARS-CoV-2 Associated with the COVID-19 Outbreak. *Curr Biol.,* Volume 30, pp. 1346-1351.

Zhao, L., Jha, B. & Wu, A., 2012. Antagonism of the interferon-induced OAS-RNase L pathway by murine coronavirus ns2 protein is required for virus replication and liver pathology. *Cell Host & Microbe,* Volume 11, pp. 607-616.

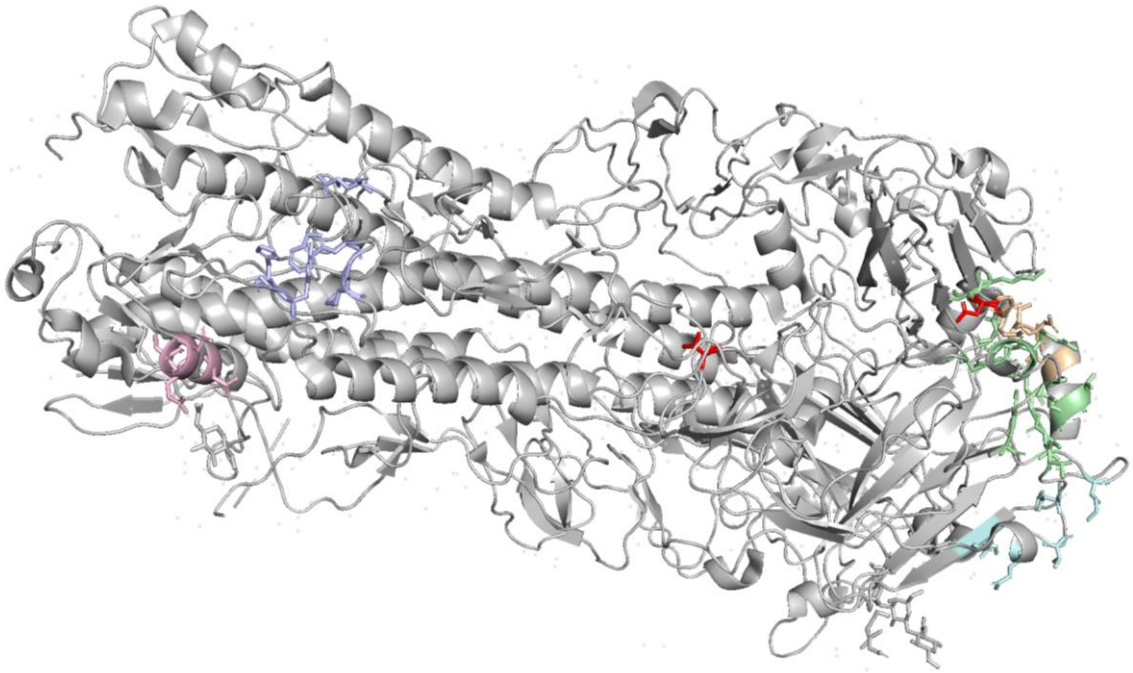
Zhou, P., Yang, X.-L. & Shi, Z.-L., 2020. A pneumonia outbreak associated with a new coronavirus of probable bat origin. *Nature,* Volume 579, pp. 270-273.

Zhu, H., Du, W., Song, M. & Liu, Q., 2021. Spontaneous binding of potential COVID-19 drugs (Camostat and Nafamostat) to human serine protease TMPRSS2. *Computational and Structural Biotechnology Journal,* Volume 19, pp. 467-476.

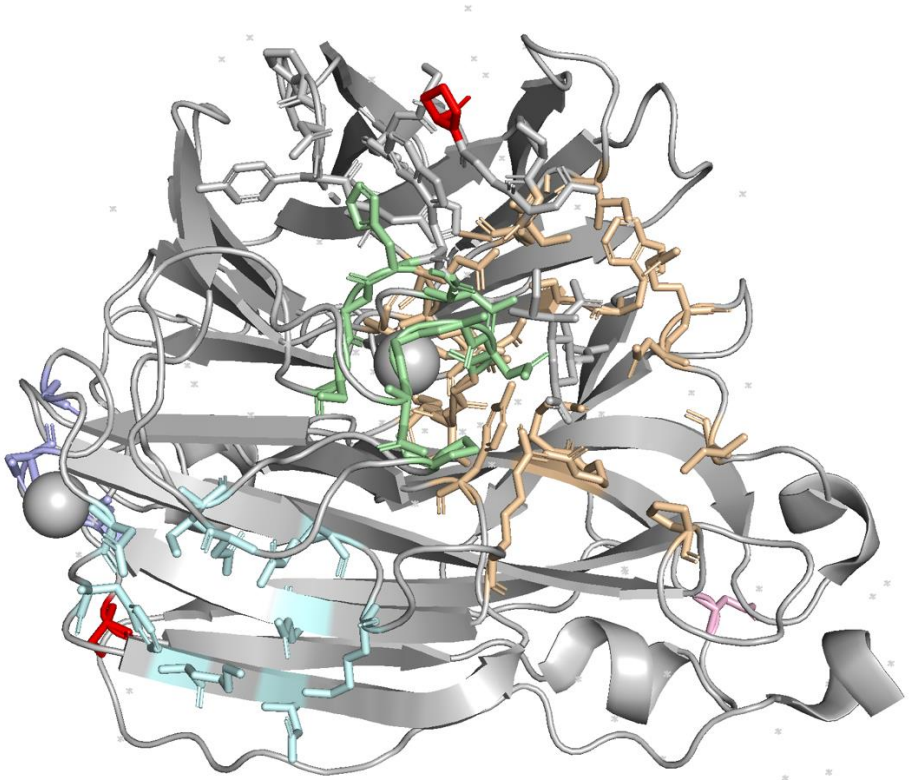
Zhu, Z. et al. 2020. From SARS and MERS to COVID-19: a brief summary and comparison of severe acute respiratory infections caused by three highly pathogenic human coronaviruses. *Respiratory Research,* Volume 21.

Appendix

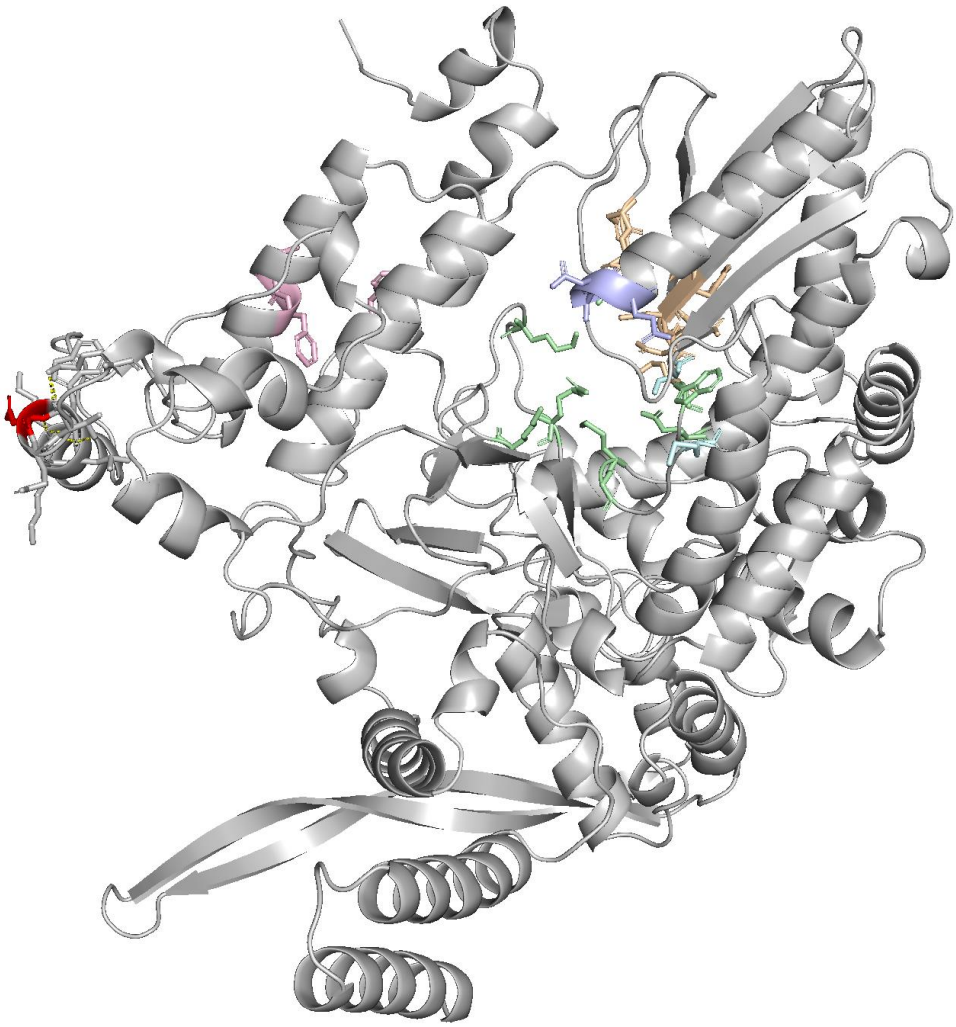
Appendix 1: Structure of influenza protein haemagglutinin (H2) with avian receptor



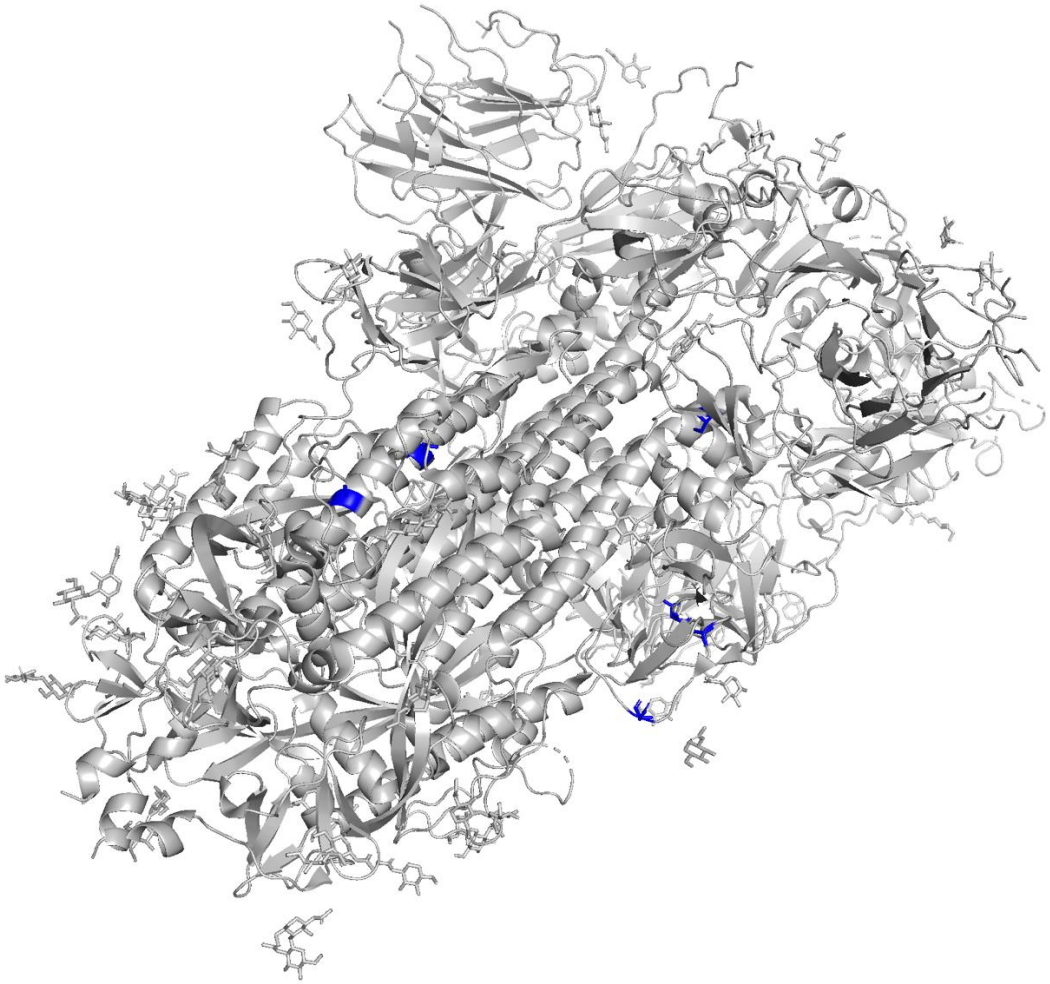
Appendix 2: Structure of influenza protein neuraminidase (N1)



Appendix 3: Structure of influenza protein RNA-directed RNA polymerase (RdRp) catalytic subunit

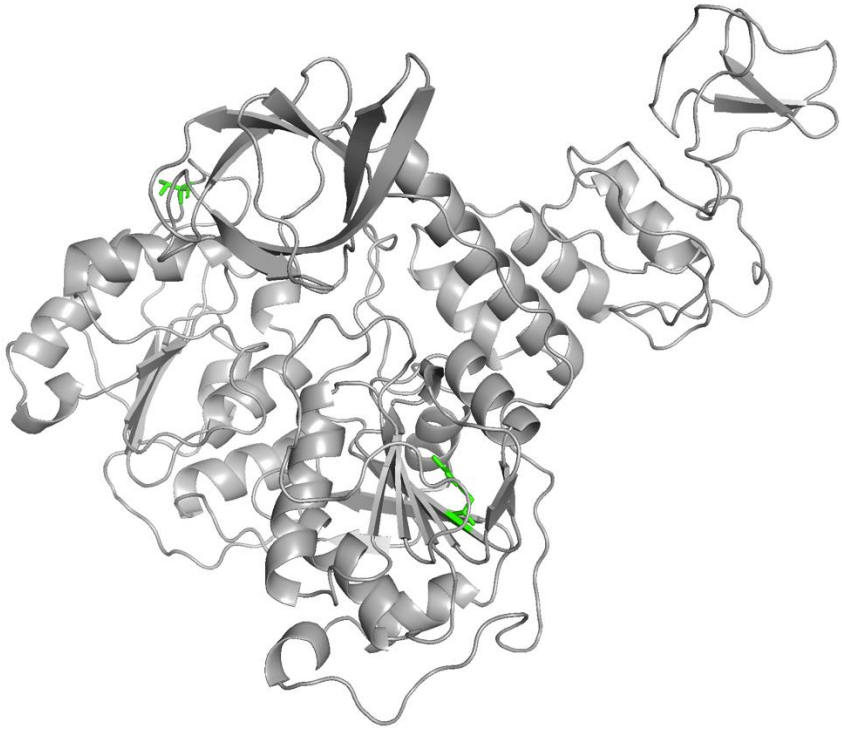


Appendix 4: Structure of SARS-CoV-2 Spike Protein



Appendix 4: Structure of SARS-CoV-2 NSP9

Appendix 5: Structure of SARS-CoV-2 NSP13



Appendix 6: Python script for flagging residues for FFM1 analysis

```
import pandas as pd
import matplotlib
import numpy as np

file = pd.read_csv("/Users/charlottethorpe/Documents/SARS-CoV-2/7-Ffm-1.camostat.var", delimiter="\t")
arr = file.values
headers = list(file.columns.values)
#open and read file in using pandas, convert to numpy array

my_dict = {"A":3, "C":4, "G":5, "T":6}
newdf = pd.DataFrame()
#create new dictionary with each of the nucletide columns as a separate key indicating the corresponding columns in the file
#create new dataframe for results to be stored in

for i in arr:
    if(i[0] != 0):
        ref = i[1]
        column = my_dict[ref]
        percent = i[column]
        if(percent < 90.0):
            sorted = pd.DataFrame(i).transpose()
            newdf = pd.concat([newdf,sorted],axis=0)
newdf.columns = headers
newdf.to_csv("camostatsorted.csv")
#for each line in the numpy array, after the first line (headings)...
#cont.. if the reference column does not match with the dictionary values or is less than 90% add to the new dataframe
#creates and stores values in new .csv file
```