



universität
wien

MASTERARBEIT / MASTER'S THESIS

Titel der Masterarbeit / Title of the Master's Thesis

„Data Science Approaches to Understand COVID-19
Epidemiology and Drug-repurposing Possibilities“

verfasst von / submitted by

Danilo Đokić

angestrebter akademischer Grad / in partial fulfillment of the requirements for the degree of

Master of Science (MSc)

Wien, 2021 / Vienna 2021

Studienkennzahl lt. Studienblatt /
degree programme code as it appears on
the student record sheet:

UA 066 606

Studienrichtung lt. Studienblatt /
degree programme as it appears on
the student record sheet:

Masterstudium Drug Discovery &
Development

Betreut von / Supervisor:

Mag. Dr. Barbara Zdrazil, Privatdoz.

Acknowledgment

First of all, I would like to thank my supervisor Mag. Dr. Barbara Zdrazil for giving me an opportunity to work on this scientific and challenging topic that concerns current major global health issue – COVID-19 pandemic. Besides, I would like to emphasize her never-ending support and constant patience that were essential for me. Furthermore, I want to thank my colleague and exceptional scientist Dr. Alžběta Türková for helping me during my practical work of this thesis. Many thanks also go to all my classmates who made this journey easier. Eventually, I would like to thank my family for their endless positive reinforcement, support and valuable advice all along the way.

Abstract

Since the outbreak in December 2019, the spread of novel coronavirus (SARS-CoV-2) and the disease that it causes remain the major problem of healthcare systems worldwide, causing the largest global health crisis in last 80 years. In order to discover and understand our possibilities in prevention and treatment of COVID-19, we need complete comprehension of 2019-nCoV's features (genome, replication, structure etc.), its interaction with human cells, and the response of our organism to the infection. Another important aspect of this current pandemic is its epidemiology and routes of viral transmission from human to human. The objective of this master's thesis was to use different data science approaches that will provide understanding of COVID-19 pandemic, its epidemiologic trends in diverse countries and comparison of specific strategies for the control of disease spread. Beside social and public health measures, indispensable aspect in fight against current pandemic is the development of remedies for the prevention and the therapy of the illness that SARS-CoV-2 causes. With regard to this, another aim of this thesis was to create an overview of potential COVID-19 protein targets and receptors in terms of their sequence and structure similarity, and, in the next step to produce a data set of ligands that could provide an understanding of current drug-repurposing possibilities.

Zusammenfassung

Seit dessen Ausbruch im Dezember 2019, sind die Ausbreitung des neuartigen Coronavirus (SARS-CoV-2) und die von ihm verursachte Krankheit nach wie vor das Hauptproblem der Gesundheitssysteme weltweit und verursachen die größte globale Gesundheitskrise in den letzten 80 Jahren. Um unsere Möglichkeiten zur Vorbeugung und Behandlung von COVID-19 zu erkennen und zu verstehen, müssen wir erstens die Merkmale von 2019-nCoV (Genom, Replikation, Struktur usw.), seine Interaktion mit menschlichen Zellen und die Reaktion unseres Organismus auf die Infektion vollkommen begreifen. Ein weiterer wichtiger Aspekt der gegenwärtigen Pandemie ist ihre Epidemiologie und die Wege der Virusübertragung von Mensch zu Mensch. Das Ziel dieser Masterarbeit ist anhand von verschiedenen datenwissenschaftlichen Ansätzen, das Verständnis der COVID-19-Pandemie, ihrer epidemiologischen Trends in verschiedenen Ländern sowie einen Vergleich spezifischer Strategien zur Kontrolle der Ausbreitung der Krankheit zu ermöglichen. Zusätzlich zu den Maßnahmen im Bereich der sozialen und öffentlichen Gesundheit ist die Entwicklung von Mitteln zur Vorbeugung und Therapie von der Krankheit, die SARS-CoV-2 verursachen kann, ein unverzichtbarer Aspekt im Kampf gegen die aktuelle Pandemie. In diesem Zusammenhang, ein weiteres Ziel dieser Arbeit besteht darin, einen Überblick potenzieller COVID-19-Proteinziele und Rezeptoren hinsichtlich ihrer Sequenz- und Strukturähnlichkeit zu schaffen und in weiterer Folge einen Datensatz ihrer Liganden zu erstellen, der das Verständnis der aktuellen Möglichkeiten zur Umnutzung von Arzneimitteln vermitteln soll.

List of Contents

1. Introduction	5
1.1. Coronaviruses	6
1.2. Covid-19 Pandemic	7
1.3. Source and Route of Transmission.....	8
1.4. SARS-CoV-2 Genome and Viral Cycle	11
1.5. The Structure of SARS-CoV-2	15
1.6. Pathogenesis and Clinical Manifestations of COVID-19	23
1.6.1. Immune Response to SARS-CoV-2 Infection.....	24
1.6.2. Clinical Characteristics and Potential Complications of SARS-CoV-2 Infection.....	27
1.7. Risk Factors and Diagnosis of COVID-19	29
1.8. Specific Aims of the Project	33
2. Materials and Methods	34
2.1. Databases	34
2.2. Python.....	37
2.3. R Programming Language	38
2.4. The Konstanz Information Miner (KNIME).....	39
2.5. Iroki – Phylogenetic Tree Visualization Software	41
2.6. DALI – Protein Structure Comparison Server	41
2.7. RaptorX Tool	41
3. Results	43
3.1. Data Science Approaches to Understand COVID-19 Epidemiology	43
3.2. COVID-19 Protein Targets and Drug-repurposing Possibilities.....	69

4. Discussion.....	85
4.1. State of the Art of Current Therapeutic Options for COVID-19	86
5. Conclusion	90
6. References	91
7. Appendix	104
7.1. Python Script for Retrieving a World Map of COVID-19 Total Cases.....	104
7.2. R Scripts	105

1. Introduction

The coronavirus disease or COVID-19 is caused by a novel virus from large family of *Coronaviridae* named severe acute respiratory syndrome coronavirus 2 (SARS-Cov-2). It is believed that outbreak of COVID-19 pandemic originates from Wuhan City, Hubei Province China where first case was identified in December 2019 (Shereen *et al.*, 2020). Three months later, on 11 March 2020, World Health Organization announced its characterization of COVID-19 as a pandemic (WHO, 2020). So far, this virus has been spread in more than 200 countries and has caused more than 1,900,000 deaths worldwide as of January 10, 2021 (Worldometers, 2021). Although the origin of SARS-CoV-2 hasn't still been confirmed, the analysis of the first cluster of infected patients suggest that seafood market in Wuhan could be initial point of this infectious disease (Dhama *et al.*, 2020). This means that this virus probably has the same zoonotic origin as previous two viruses from this family that also caused epidemic in 2002 (SARS – severe acute respiratory syndrome) and in 2012 (MERS-Middle East Respiratory Syndrome) (Fani, Teimoori and Ghafari, 2020). The SARS-Cov-2 has very rapidly spread from China to other countries all over the world. Italy was the first country in Europe that was strongly impacted by COVID-19. The experience of patients from this region showed that symptoms of this disease can vary from mild to life-threatening, where elderly and patients with chronic diseases are exposed to the highest risk. The COVID-19 pandemic has been shortly confirmed on all continents causing serious effect on the healthcare systems especially in North and South America. However, some countries showed particularly good response to this global health problem and proved the importance of following government-imposed measures and providing self-awareness in prevention of disease transmission (Teslya *et al.*, 2020).

Nevertheless, the most promising way to cope with current global pandemic is to understand the pathogenesis of this virus. Firstly, we need to discover the structure of SARS-Cov-2, the mechanisms of its interaction with human cells and the ways how this virus escapes our immune system by harming other cells and tissues. This will provide the development of effective therapy approaches in fight against SARS-Cov-2.

1.1. Coronaviruses

The coronaviruses (CoVs) belong to big family of single-stranded RNA viruses named *Coronaviridae*. These viruses are characterized by large size (probably one of the largest known RNA viruses) and enveloped structure. The *Coronaviridae* family is divided into two subfamilies: *Coronavirinae* and the *Torovirinae* (Payne, 2017). According to current taxonomy, CoVs belong to subfamily *Orthocoronavirinae*, order *Nidovirales* and realm *Riboviria* (Fan *et al.*, 2019). Based on genome analysis, this subfamily is further classified into four genera: *Alphacoronavirus* (α CoV), *Betacoronavirus* (β CoV), *Gammacoronavirus* (γ CoV), and *Deltacoronavirus* (δ CoV) (Payne, 2017). SARS, MERS and SARS-2, as the most significant members of this family, belong to the β CoVs (Helmy *et al.*, 2020). The name of coronaviruses originates from spike glycoproteins on the envelope of these microbes which have crown-like appearance.

Members of this family of viruses are widespread among both domestic and wild animals. α CoVs and β CoVs mainly infect mammals, while another two genera, γ CoVs and δ CoVs, predominantly infect birds (Helmy *et al.*, 2020). Bats represent large reservoir of coronaviruses, especially β CoVs. The analysis of viruses isolated from these animals has revealed that bats have numerous unique and highly specific coronaviruses (BtCoVs). This suggests exceptionally long history of evolution between these microorganisms and bats as their hosts. Although many animal species are resistant to CoVs, these microbes can be very harmful, and they are cause of diverse animal diseases. Infectious bronchitis virus (IBV) that attacks chickens was identified in 1930 and is a member of CoVs family (Payne, 2017). In 2014, a pig coronavirus was a cause of death over a million of piglets in USA. Certain strains from *Coronaviridae* generate extraordinarily complex infections and even deadly diseases in domestic cats. However, the interest for these viruses was significantly increased after the outbreak of SARS in 2002 that revealed the ability of certain viruses from this family to be pathogenic for humans as well (Payne, 2017).

To date, seven coronaviruses that infect humans have been identified (B. Chen *et al.*, 2020). Based on symptoms which they cause, these viruses can be divided into two groups. The first group is formed of following four viruses: (HCoV)-229E, HCoV-NL63, HCoV-OC43 and HCoV-HKU1. These four human viruses principally cause mild symptoms of common cold. Second group of human coronaviruses includes SARS-CoV, MERS-CoV and novel SARS-CoV-

2. Clinical features of diseases caused by these three viruses vary from asymptomatic to severe and life-threatening medical conditions (B. Chen *et al.*, 2020). Although these microorganisms show numerous distinctions concerning their genome features, molecular structure and etiology, some common clinical manifestations have been observed. The symptoms that emerge from SARS-CoV-2, MERS-CoV or SARS-CoV-2 infection include cough, fatigue, fever and pneumonia with abnormal CT chest examinations. Although some cases with upper respiratory infection symptoms have been reported (primarily sore throat and rhinorrhea), the cells in lower compartments of respiratory tract are considered to be main targets of these viruses (Zhang *et al.*, 2020). However, recent findings of those patients infected with SARS-CoV-2 are showing certain typical symptoms that are developed in early stages of disease and such as loss of smell or taste (Mullol *et al.*, 2020).

1.2. Covid-19 Pandemic

In late 2019, a novel coronavirus (SARS-CoV-2) emerged in Wuhan, Hubei province in China, causing Covid-19 (Coronavirus disease 2019) that rapidly became a global threat characterized as a pandemic. The first cluster of twenty patients with pneumonia-like symptoms was reported on December 31, 2019 (Huang *et al.*, 2020). Seven cases from this group were in severe condition with unknown etiology of disease. All patients reported certain linkage to Seafood Market in Wuhan which is believed to be a source of outbreak (Mizumoto, Kagaya and Chowell, 2020). As previously mentioned, this virus probably has the zoonotic origin. A study from July 2020 that analyzed the genome of this virus and traced its origin revealed that SARS-CoV-2 shows most similar properties with the virus that circulates in horseshoe bats (Burki, 2020b). Soon after the identification of first cluster of patients with SARS-CoV-2 infection, the new cases with the same symptoms were reported. The analysis of new cases showed that between 8,6% and 51% of patients didn't have any linkage to Seafood Market in Wuhan (Burki, 2020b). This was an obvious proof of person-to-person viral transmission. Analogous to other respiratory pathogens, the main route of human-to-human transmission is through respiratory droplets that occur during sneezing and coughing but talking and breathing as well. The size of these droplets is $> 5 \mu\text{m}$ in diameter and they are easily transmitted through the mucosa of nose and mouth and conjunctiva of eyes (Cascella *et al.*, 2020). This route of transmission also involves direct contact with these surfaces. Recent studies revealed that 2019-nCoV might be also spread by indirect contact and

oral-fecal route (Amirian, 2020). However, there remains a need for further conclusive evidence concerning these ways of infection, but also for other as vertical or sexual route of transmission.

The clinical presentation of patients infected with SARS-CoV-2 varies from asymptomatic to the conditions characterized by respiratory failure that necessitates mechanical ventilation and treatment in ICU (Intensive Care Unit) (Thevarajan, Busing and Cowie, 2020). The first reports showed that clinical spectrum of COVID-19 disease depends on age and comorbidity, but majority of patients suffer from at least one of following symptoms: dry cough, mild to high fever, sore throat, nasal congestion, headache, muscle pain and malaise (Stawicki *et al.*, 2020). Recently, another two symptoms have been reported in many SARS-CoV-2 cases and they involve loss of smell and taste (Butowt and von Bartheld, 2020). These signs are becoming a main hallmark of COVID-19 disease. It is important to underline that although clinical manifestations of this disease vary between the population, every patient regardless of the symptoms can be a carrier. The viral infection diagnosis involves viral and antibody testing (Mathuria, Yadav and Rajkumar, 2020). The assays for viral testing include detection of SARS-CoV-2 nucleic acid or antigen in the respiratory samples such as oral/nasal swab and saliva, while antibody testing requires serology samples which are tested for certain antibodies and they provide an information whether person was previously infected. The antibody testing represents a tool for the estimation and surveillance of current epidemiology situation (Mathuria, Yadav and Rajkumar, 2020).

1.3. Source and Route of Transmission

SARS-CoV-2 and its genomic sequence was very fast identified, within 28 days of the outbreak. In comparison with time taken to the identification of SARS-CoV in 2003, this was imposing result (Dhama *et al.*, 2020). As aforementioned, the origin of SARS-CoV-2 is probably the same as the origin of previous two viruses from this family that also caused serious healthcare issues – SARS-CoV in 2002 and MERS-CoV in 2012. The analysis of the genome of novel coronavirus showed high percentage of similarities with the viruses from this family found in bats. However, the route of transmission and intermediate hosts for these three outbreaks differentiate.

In the case of SARS epidemic, the virus was probably transmitted to humans from Asian palm civet (**Figure 1**) (Helmy *et al.*, 2020). The first cluster of infected patients were people employed

in food industry. This could be a point where transmission from animals to humans occurred and it is very important to consider the fact that both SARS and Covid-19 outbreaks are related to food handlers in China (Xu *et al.*, 2004).

The route of MERS-CoV transmission to humans was different. This virus emerged in Middle East and had very high mortality rate (2494 cases in total with 858 deaths). Its route of transmission was long debatable, but eventually the phylogenetic analysis revealed that MERS-CoV was transferred from dromedary camels to humans (**Figure 1**) (Fani, Teimoori and Ghafari, 2020). Nevertheless, this virus evidently originates from bats, although it had distinct spreading cycle.

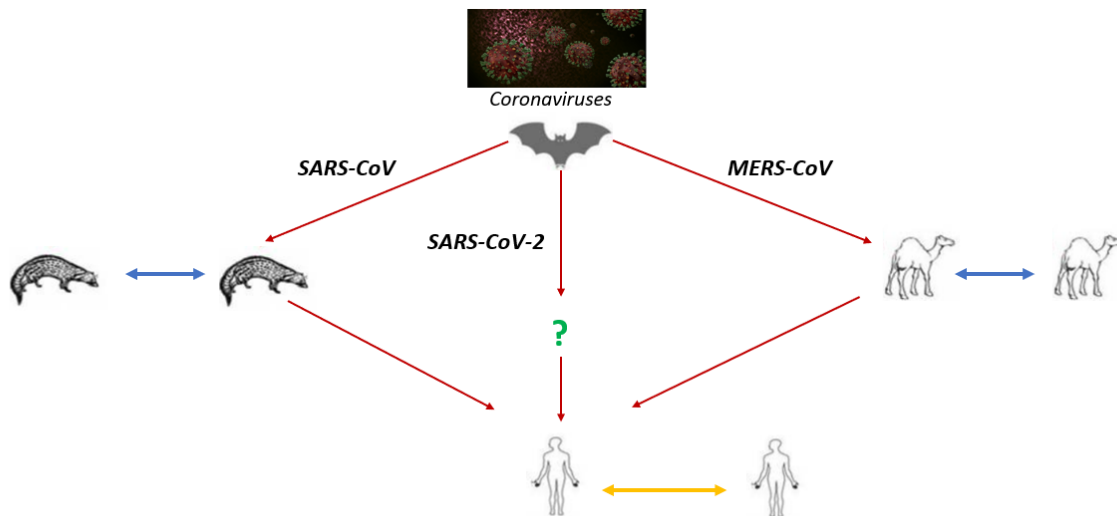


Figure 1. The viral transmission cycle of SARS-CoV, MERS-CoV and SARS-CoV-2, adapted from (Helmy *et al.*, 2020)

As previously stated, the investigation of first cluster of patients infected with SARS-CoV-2 showed common exposure to Seafood Market in Wuhan, China. These markets are famous for diverse animal species that are exposed and ready for human consumption, but also for live animals such as bats and snakes that are being sold (Dhama *et al.*, 2020). The reason why bats are such a big reservoir of different viruses (CoVs among them) is their specific metabolism and immunity that has been developed through the evolution. As in the case of SARS outbreak, inappropriate handling with animals could be one of the main causes for SARS-CoV-2 spreading from animals to humans (Nadimpalli and Pickering, 2020). The fact that all three outbreaks took

place in Asia where animal-human contact is often not controlled and where sanitary conditions are not fulfilled, should also be considered.

Although the source of transmission as a natural reservoir of a virus is important for the understanding of current global pandemic, the main reason for the rapid and unrestricted virus spreading lays in human-human transmission. It is generally considered that viral respiratory infections have two main routes of spreading: by direct contact, such as touching the fomites and surfaces that was previously infected by a carrier of infection, or by respiratory droplets (occurred during breathing, coughing, sneezing etc.) in the situation when the persons are in close proximity (Morawska and Cao, 2020). The important aspect of airborne transmission is the duration of viral particles presence in the air. Essentially, these virions are situated in air droplets that consist of liquid and gas content. After the liquid portion evaporates, the droplets become that small that they have more propensity for a transport by air currents than via gravitation. These small droplets are then free to be transported by air and thereby carry the viral particles much further from where they originate (**Figure 2**) (Morawska and Cao, 2020). The studies reported that major route of SARS-CoV-1 was by air, especially in indoor (Yu *et al.*, 2005). Considering the similarities between SARS-CoV-1 and SARS-CoV-2 and available evidence on viral transport, it is highly evident that this route of spreading is typical for novel coronavirus as well.

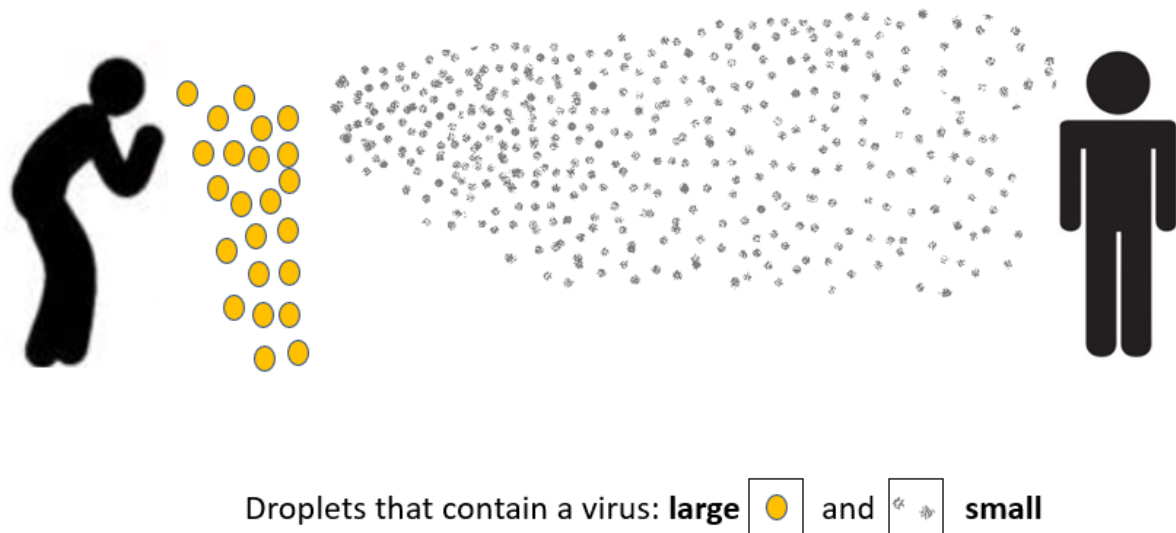


Figure 2. Airborne transmission of viral particles. Illustration adapted from (Morawska and Cao, 2020)

Another route of transmission that has been considered a possible is ocular (Dockery *et al.*, 2020). It was commonly assumed that respiratory mucosa and eye surface are the most susceptible routes of viral entry. Although the studies showed that there is low likelihood for ocular area to be the gateway of the infection, the presence of SARS-CoV-2 in this environment could cause eye issues including local infection, conjunctivitis and ocular discomfort (X. Chen *et al.*, 2020). This is the reason why eye protection should be essential precaution in the virus protection exceptionally for medical stuff.

Beside these two surfaces, that are mostly exposed to infectious respiratory droplets, other routes of viral transmission could be oral-fecal, bloodborne and vertical (mother to child) (Meyerowitz *et al.*, 2020). However, these types of infection are still uncertain and need further research. It is also important to emphasize that SARS-CoV-2 was identified in breast milk, although there is no evidence to date that mother can infect a child through this transmission route (Health Ontario and publique Ontario, 2020).

1.4. SARS-CoV-2 Genome and Viral Cycle

SARS-CoV-2 possesses the basic features characteristic for *Coronaviridae* family. This microorganism has a positive single-stranded RNA genome typical of its large size. The length of SARS-CoV-2 genome ranges from 26 to 32 kb and contains 6-11 open reading frames (ORFs) that encode 9680 amino acid proteins (Kumar *et al.*, 2020). The genome analysis revealed that novel coronavirus shares around 79% and 50% genome sequence identity with SARS-CoV and MERS-Cov, respectively (H. Wang *et al.*, 2020). On the other hand, the genome sequence similarity of the sequences responsible for the encoding of main structural entities in these viruses was over 90% (Naqvi *et al.*, 2020). It is worth of pointing out that gene sequence of SARS-CoV-2 was mostly related to BatCoV RaTG13 genome (sequence alignment 96.3%) (Paraskevis *et al.*, 2020).

The 2019-nCoV contains 4 structural proteins, 16 non-structural proteins and at least 6 accessory proteins. Structural proteins have significant role in viral pathogenesis, and they include: spike glycoprotein (S), envelope protein (E), membrane (matrix) protein (M) and nucleocapsid protein (N) (Helmy *et al.*, 2020). These proteins also possess high sequence similarity with the corresponding proteins in SARS-CoV and MERS-CoV (Naqvi *et al.*, 2020).

Novel coronavirus is monopartite (class of genome consisted of single strand of nucleic acid), and its genome consists of 29,903 nucleotides which makes largest known RNA genome sequence (Y. Wang *et al.*, 2020). This viral genome is composed of two untranslated regions (UTRs) and they are located at the 3' and 5' ends. Apart from that, SARS-CoV-2 genetic material contains 11 open reading frames (ORFs) which encode 27 proteins in total (Helmy *et al.*, 2020). The first ORF encodes 16 non-structural proteins, and this region of novel coronavirus RNA composes approximately 67% of whole genome. The rest of ORFs encode structural and accessory proteins (**Figure 3**). Non-structural proteins are predominantly involved in the processes of viral transcription and replication, while structural proteins play main role in viral entry and survival in host cells (Kumar *et al.*, 2020).

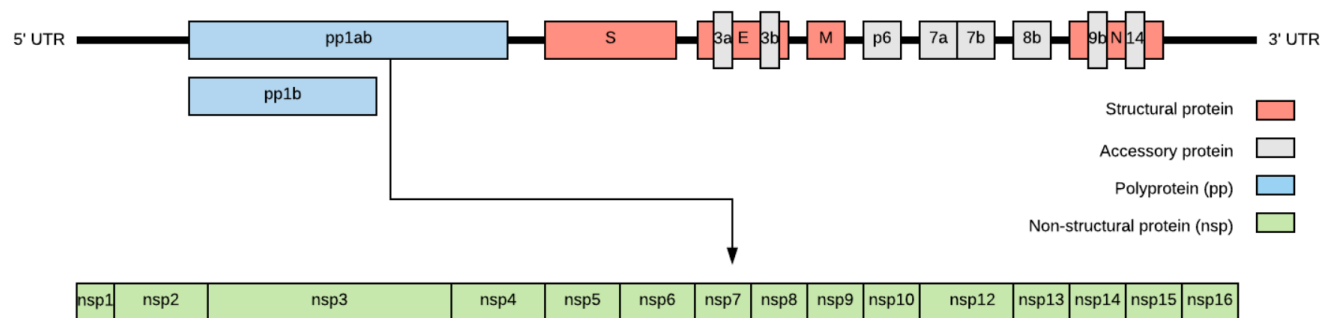


Figure 3. Genome organization of SARS-CoV-2. The structure of SARS-CoV-2 is: 5'-UTR-*orf1a-orf1ab*-S (Spike)-E (Envelope)-M (Membrane)-N (Nucleocapsid)-3'UTRpoly (A) tail. Adapted from Wikimedia.org.

The penetration of viral particle into human cell starts with the interaction between human receptors ACE2 (angiotensin-converting enzyme 2) and structural S protein of SARS-CoV-2. However, recent studies show the essential role of TMPRSS2 (transmembrane protease serine subtype 2) and its activity in viral entry and pathogenesis (Mollica, Rizzo and Massari, 2020). Several observations support the hypothesis that this segment of viral penetration is probably the same for both SARS-CoV and SARS-CoV-2 (D. Wang *et al.*, 2020). In these two related viruses, TMPRSS2 promotes viral entry by two different mechanisms. The first one involves the interaction with ACE-2 that could stimulate viral uptake, while second mechanism implies S glycoprotein cleavage that leads to priming of this protein and membrane fusion (Hoffmann *et al.*, 2020). ACE-2 and TMPRSS2 are expressed in various tissues of human body. This suggests that SARS-CoV-2 particles could attack and infect not only cells in respiratory compartments as

it was earlier thought. The investigations revealed that ACE2 expression is the highest in kidneys, heart, adipose and thyroid tissue, testis and small intestine, while the lowest expression was found in mussels, brain, blood, bone marrow and spleen (Li *et al.*, 2020). On the other hand, the highest levels of TMPRSS2 were detected in kidneys, prostate, gastrointestinal tract and pancreas (*Tissue expression of TMPRSS2 - Summary - The Human Protein Atlas*, 2020).

Moreover, a recent study showed that lung and nasal airways expression of ACE2 and TMPRSS2 can be modified by age or disease status (Saheb Sharif-Askari *et al.*, 2020). The children had significantly lower distribution of these two receptors in upper and lower respiratory compartments. Besides, the expression of these SARS-CoV-2 host cell entry genes was highly upregulated in smokers and in patients with chronic obstructive pulmonary disease (COPD) (Saheb Sharif-Askari *et al.*, 2020). Both these genes were also highly expressed in blood of those with essential hypertension. These results suggest that clinical manifestations of COVID-19 could be attributed to different expression levels of ACE2 and TMPRSS2 in airways tissue (Saheb Sharif-Askari *et al.*, 2020).

After viral attachment to cell surface, series of cascades and iterations in plasma membrane, the process of viral penetration will end with endocytosis that will pull a virus into the host cell. This is followed by the uncoating which will eventually lead to the releasing of viral RNA (V'kovski *et al.*, 2020). In the next step, this viral RNA strand will be translated by host cell machinery into couple of ORFs (ORF1a and ORF1ab). These two ORFs will be then translated into two large polyproteins – polyprotein1a (pp1a) and polyprotein1ab (pp1ab), in a process of ribosomal frame shifting (Astuti and Ysrafil, 2020).

The next stage in viral cycle is a proteolysis of these two polyproteins into 16 smaller proteins (non-structural proteins) by two proteases: 3C-like protease (3CLpro) and papain-like protease (PLpro) (InvivoGen, 2020). The newly synthesized non-structural proteins are involved in the processes of replication (replicase complex) and transcription (transcriptase complex). The replicase complex proteins, together with RNA-dependent RNA polymerase (RdRp), will interact with positive single-strand RNA of SARS-CoV-2 and facilitate viral replication that will result in production of anti-sense strand of viral RNA (Romano *et al.*, 2020). The new anti-sense strand is then processed into two different pathways. First one includes replication and production of positive (sense) RNA that is essentially identical to viral RNA that entered the host cell. The

second pathway involves process of transcription mediated also by RNA-dependent RNA polymerase (RdRp) that will result in forming of structural and accessories proteins. In this cascade anti-sense RNA is transcribed into bunch of sub-genomic RNAs (+) in the process of discontinuous transcription (a type of transcription unique among RNA viruses) (Kim *et al.*, 2020). These sub-genomic RNAs (+) are mRNAs that will be translated into structural and accessories proteins in the next step of viral cycle. After the process of replication, the synthesis of sub-genomic mRNAs and their translation into viral proteins, these newly synthesized macromolecules are transported and inserted into endoplasmic reticulum (ER) (Astuti and Ysrafil, 2020). They are subsequently moved along the signaling pathway endoplasmic reticulum-Golgi intermediate compartment (ERGIC). In this stage, previously replicated genome will directly join the N protein (nucleocapsid structural protein) to nucleocapsid and be transported to ERGIC (Cascarina and Ross, 2020). In this section, nucleocapsids will fuse with other structural proteins and form mature virions. This is followed by the SARS-CoV-2 virions exocytosis from the host cell into the surrounding environment where the infection cycle will be repeated (Astuti and Ysrafil, 2020) (**Figure 4**).

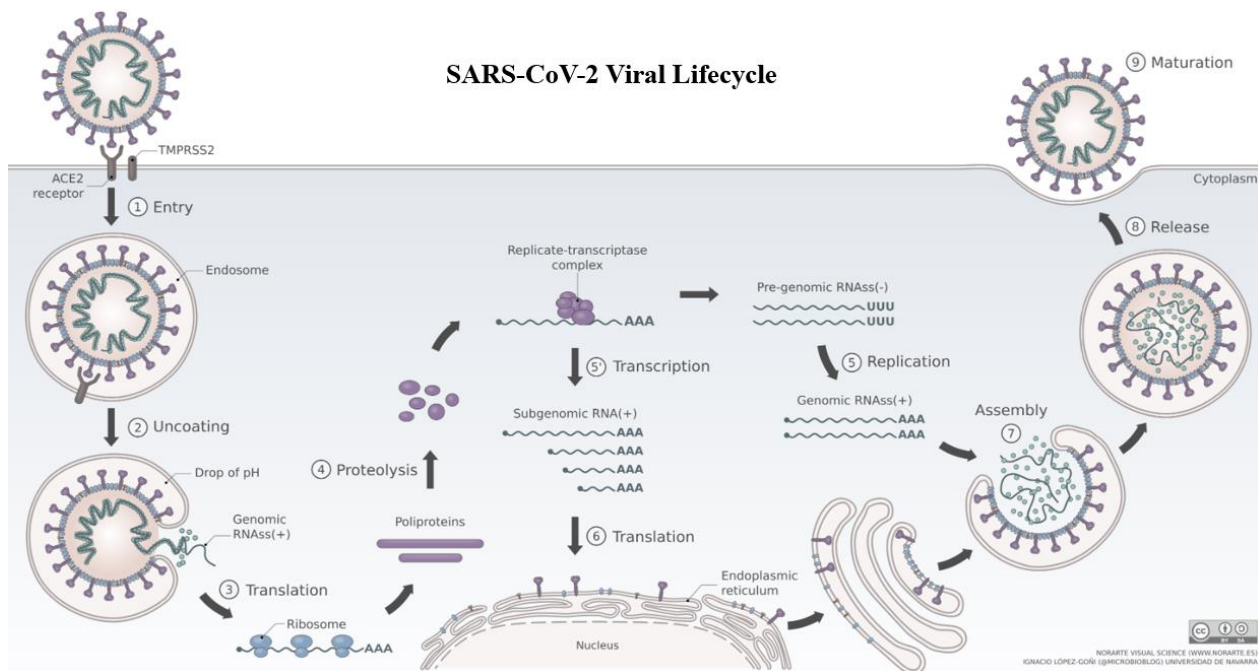


Figure 4. SARS-CoV-2 viral penetration and lifecycle in host cell. Adapted from Wikimedia.org.

1.5. The Structure of SARS-CoV-2

As previously mentioned, 2019-nCov consists of three types of proteins: structural, accessory and non-structural proteins. These macromolecules have diverse and essential roles in the viral lifecycle, so each group will be considered separately.

SARS-CoV-2 has four structural proteins: spike glycoprotein (S), envelope protein (E), membrane (matrix) protein (M) and nucleocapsid protein (N) (**Figure 5**). The N protein is structural protein of coronaviruses (SARS-CoV-2 among them) with size 43-50 kD and diverse viral functions. As already stated, structural proteins define physical properties of viral particles (such as viral shape, structure etc.), but they are also involved in viral pathogenesis. N protein has three regions: N-terminal domain (NTD) or RNA-binding domain, C-terminal domain (CTD) – dimerization domain and serine rich (SR) linker region between them (Asghari *et al.*, 2020). This is the most abundant viral protein and the only protein present in the nucleocapsid of SARS-CoV-2. The most important roles of this macromolecule involve the complex formation with viral RNA, increased efficiency of viral transcription and ease of M protein interaction necessary during the assembly of virion (Dhama *et al.*, 2020). It is suggested that complete and optimal interaction with RNA requires both domains of this protein (Fehr and Perlman, 2015).

The NTD region binds to the 3' end of the viral genome, probably via electrostatic interactions. However, the SR linker region has also an ability to interact directly with *in vitro* RNA and is involved in cell signaling. Another property of this linker domain is its modulation of the antiviral host response by acting as an antagonist for interferon (IFN) (Dhama *et al.*, 2020). Although N protein of SARS-CoV-2 shows five amino acid mutations in comparison to this the same protein in SARS-CoV, it is worth mentioning that the N protein in SARS-CoV stimulates the activation of Cyclooxygenase-2 (COX-2) causing inflammation in the lungs. The earlier investigations revealed that this macromolecule interacts with the p42 proteasome subunit, which is involved in the degradation of the viral proteins (Satarker and Nampoothiri, 2020). The N protein also interacts with nsp3, one of the crucial components of the replicase complex (Fehr and Perlman, 2015). Due to its surface location and lower affinity to variation in comparison to some other structural proteins, N protein of SARS-CoV-2 was studied as a potential vaccine candidate. *In silico* analysis revealed that effects of Glycyrrhizic acid and the phytochemical

Theaflavi on the N protein could be potentially new option in the treatment of Covid-19 (Asghari *et al.*, 2020).

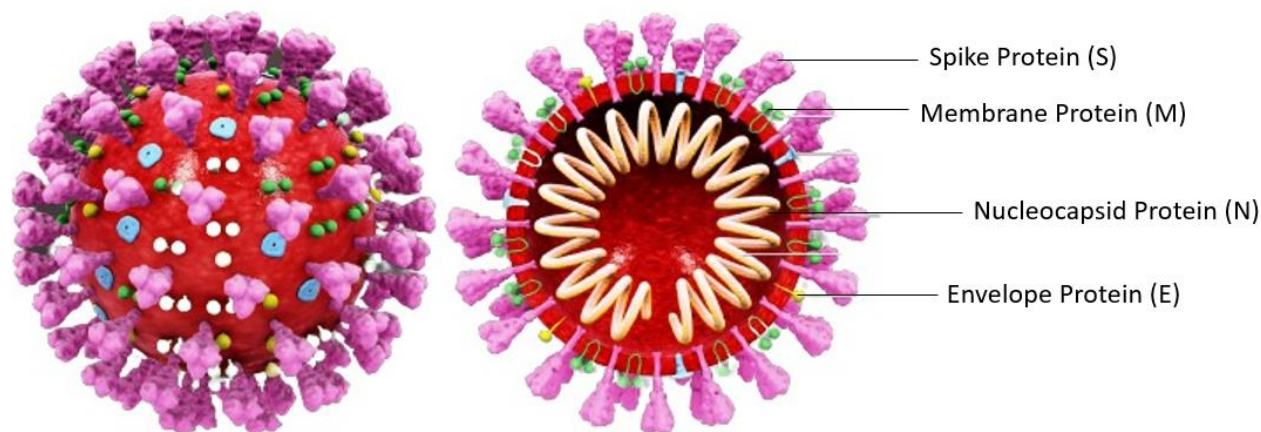


Figure 5. SARS-CoV-2 Structure. Adapted from Wikimedia.org.

The second structural protein of SARS-CoV-2 is S (spike) glycoprotein. It is a large protein (150 KDa with 1300 amino acids) that belongs to class 1 of transmembrane proteins. The main role of S protein is the mediation of viral entry into the host cell (Meirson, Bomze and Markel, 2020). The spike glycoprotein of β CoVs acts as a homotrimer on the surface of virion giving it a crown-like appearance. Due to this feature, this family is named coronaviruses (corona- a crown in Latin). This protein contains three domains: an ectodomain, transmembrane component and intracellular C fragment. The ectodomain is composed of two subunits that promote viral attachment to the surface of target cells (S1 subunit) and fusion of host membrane and viral membrane (S2 subunit) (Satarker and Nampoothiri, 2020). The S1 subunit is further split into two domains: N-terminal domain (NTD) and C-terminal domain (CTD). The receptor-binding domain (RBD) is a part of CTD (Q. Wang *et al.*, 2020). This compartment of spike protein interacts with ACE2 on the host cell. In order to bind for a host membrane receptor, S protein needs to be structurally transformed. The RBD of SARS-CoV-2 detects and binds to the human ACE2 with 10-20 times higher affinity in comparison to the SARS-CoV subdomain (Satarker and Nampoothiri, 2020). The reason for this could be a polymorphism at 501T that enhances ACE2-S protein interaction and thus increases infectivity (Satarker and Nampoothiri, 2020). In general, the spike protein amino sequence of SARS-CoV-2 shows approximately 75% similarity with the same macromolecule of SARS-CoV. On the other hand, the investigations reported an identity of

99% for S2 subunit and an identity of 71% for S1 subunit between the same protein in novel coronavirus and SARS-CoV (Satarker and Nampoothiri, 2020). Spike protein in all beta-coronaviruses is cleaved by proteases such as cathepsin (lysosomal protease) and TMPRSS2 (Transmembrane protease serine 2) during the interaction between viral and host cell. Due to role of these proteases in viral penetration into the human cells, certain inhibitors of these enzymes (MI-1851 and camostat mesylate) are being investigated as the potential agents in a treatment of SARS-CoV-2 (Asghari *et al.*, 2020). Another therapy option against novel coronavirus includes direct targeting of spike protein. Antibodies are usually used in this therapy approach and CR3022 is one of the drugs that was applied in the treatment of SARS disease. Its mechanism of action includes binding to the RPD site. However, potential use against SARS-CoV-2 needs to be confirmed (Asghari *et al.*, 2020).

The spike protein shows essential roles in the functioning and viral entry and thus is another crucial component of SARS-CoV-2 together with N protein. These structural components exhibit a promising solution in the development of effective therapeutics for the prevention and treatment of COVID-19.

The M protein (membrane protein) is small structural protein (approximately 25-30 kDa) that defines the shape of the virus. It is most abundant structural protein in SARS-CoV-2 and it is bound to viral nucleocapsid (Dhama *et al.*, 2020). This macromolecule has three transmembrane domains with short N-terminal glycosylated ectodomain and significantly larger C-terminal endodomain (Fehr and Perlman, 2015). The membrane protein of SARS-CoV-2 has an ability to bind to other structural proteins of this virus. The N protein is stabilized by the interaction with M protein and this interconnection also provides the termination of viral assembly by preserving N protein-RNA complex (Thomas, 2020). Although the investigations confirmed the binding of M protein to spike protein, the exact mechanism of this interaction is still not fully understood. As the other structural proteins, membrane protein of SARS-CoV-2 expresses high sequence homology with the same protein of SARS-CoV (with the bat SARS-CoV the similarity is 98.2%, while with the pangolin SARS-CoV is 90%) (Thomas, 2020). In comparison to the M protein of MER-CoV, the homology was lower. The recent *in silico* analysis revealed that membrane protein of SARS-CoV-2 possesses high sequence identity with prokaryote sugar transporter SemiSWEET (Thomas, 2020). This protein has multiple cell functions and this similarity with structural membrane protein could potentially explain the new roles of SARS-CoV-2 M protein

in its immune evasion and replication. However, future research is needed for comprehensive understanding.

The fourth structural protein is E protein (envelope protein) which main roles involve viral assembly and the release of virions. This macromolecule is the smallest of all structural proteins in SARS-CoV-2 (8-12 kDa) (Sarkar and Saha, 2020). The envelope protein is hydrophobic macromolecule that consists of three main domains: short N-terminus, transmembrane part with approximately 25 amino acids and C terminal domain. The transmembrane compartment of E protein constitutes pentameric ion channels that have low ion selectivity, and they might act as the viroporin. These proteins are embedded on the host membranes and thus promote viral penetration and maturation. Due to these properties, viroporins are considered as ideal therapeutic targets (Sarkar and Saha, 2020). An important region of CT domain is PDZ-binding motif (PBM). This protein compartment is considered as a one of the essential agents in viral pathogenicity of both SARS-CoV-2 and SARS-CoV (Asghari *et al.*, 2020). The studies that involved the deletion of E protein or mutations in PBM region as a potential therapeutic strategy for the effective vaccines against SARS-CoV-2, revealed the importance of this structural protein, and PBM domain as its part, for the viral infectivity. As a matter of fact, after the gene manipulations that involved PBM motif of E protein, novel coronavirus was capable of synthesizing and incorporation of novel chimeric protein that contained PBM (Sarkar and Saha, 2020). These results suggest the significance of this domain of E protein and its indispensability for SARS-CoV-2 virulence.

Beside structural and non-structural proteins, 2019-nCov also contains accessory proteins that are largely unknown as they show less homology with the same class of proteins of other coronaviruses (V'kovski *et al.*, 2020). ORFs that encode accessory proteins intercept the ORFs of structural proteins in the 3' end of SARS-CoV-2 genome. Although the number of these type of macromolecules in coronaviruses vary, the latest studies revealed that the novel coronavirus probably contains not less than 9 accessory proteins (3a, 3b, 6, 7a, 7b, 8, 9b, 9c, and 10; see **Table 1**) (Zhu *et al.*, 2020). Despite the fact that these proteins are not crucial for the viral replication, they have an important role in the pathogenicity and could modulate the immune response of the host (Michel *et al.*, 2020).

The procedure of genome replication and transcription are mediated by RNA processing complex (replicase-transcriptase complex). The components of this complex originate from 16 non-structural proteins derived from polyprotein1a (pp1a) and polyprotein1ab (pp1ab) provided by ORF1a and ORF1ab (Neuman *et al.*, 2008). The breakdown of pp1a produces non-structural proteins 1-11 (nsps 1-11), while proteolysis of pp1ab provides nsps 1-16 (Krichel *et al.*, 2020). However, these macromolecules can also modulate immune response of host and due to this features, non-structural proteins are considered as possible therapeutic targets. The complete roles of all non-structural proteins are yet not known, but their homology with the same proteins in other viruses from *Coronaviridae* suggest that their function could be similar or the same.

Nevertheless, further investigations are needed to understand all functions of the accessory proteins and their potential in therapy against SARS-CoV-2.

Table 1. Accessory proteins of SARS-CoV-2 and their role

Accessory Protein	Role of Protein
3a	Involved in the formation of viroporins (Issa <i>et al.</i> , 2020) and could modulate virus release and infectivity (UniProtKB, 2020)
3b	Small peptide, potential IFN-1 antagonist (Konno <i>et al.</i> , 2020)
6	In SARS-CoV blocks the expression of interferon stimulated genes (ISGs) (UniProtKB, 2020) and acts as IFN antagonist (one of the most potent antagonist of IFN in SARS-CoV-2) (Kopecky-Bromberg <i>et al.</i> , 2007)
7a	Plays a role as the antagonist of host tetherin (BST2), disrupting its antiviral effect (by binding to BST2 and thereby interfering with its glycosylation) (UniProtKB, 2020)
7b	In SARS-CoV, the peptide consists of 43 amino acids that overlaps the sequence of 7a accessory protein (Schaecher, Mackenzie and Pekosz, 2007)
8	SARS-CoV-2 has single ORF8 protein while SARS-CoV has two of them; It is involved in recognition by the immune system and transcription of N protein (Su <i>et al.</i> , 2020)

9b	It has a role in the inhibition of host innate immune response by targeting the mitochondrial-associated adapter (UniProtKB, 2020)
9c	The investigations revealed that this protein modulates the NF-KB pathway and IκB kinases, and interacts with other proteins in the virus (Gordon <i>et al.</i> , 2020)
10	This protein has 38 amino acids and is unique for SARS-CoV-2 (it does not exist in SARS-CoV); although its function is still not determined, this protein probably does not have essential role (Yoshimoto, 2020)

Non-structural protein 1 (NSP1) is a leader protein that suppresses protein synthesis in human cells and blocks the antiviral response of the cell infected with SARS-CoV-2 (Abbadi, 2020). It is also found in SARS-CoV and the investigations showed that NSP1 has an ability to bind for 40S ribosome subunit of the host cell in order to inactivate translation and stimulates host mRNA degradation, while SARS-CoV mRNA remains undamaged (Yoshimoto, 2020).

The second protein from this category is NSP2. The exact function of this macromolecule is yet not known. Nevertheless, the studies on this protein in SARS-CoV revealed that it could bind prohibitin 1 and 2 (PHB 1 and PHB 2) (Yoshimoto, 2020). There is high probability that NSP2 has the similar role in SARS-CoV-2 as well. These proteins are located in human cells and regulate different processes such as cell differentiation, apoptosis, cell migration etc. The binding of NSP2 for PHB1 and PHB2 could cause disrupting of these cell mechanisms and cell damaging as its consequence (Yoshimoto, 2020).

Non-structural protein 3 (NSP3) is large, trans-membrane protein. It has several domains including ubiquitin-like domain 1 and 2 (Ubl1 and Ubl2), Y1 domain, macro domain and Glu-rich domain (also called “hypervariable region”) and papain-like protease 2 (PL2^{pro}) among others (Neuman *et al.*, 2008). The PL2^{pro} domain breaks down pp1a and pp1ab as it has protease activity. As NSP3 has important role in the proteolytic degradation and release of proteins essential for the virus, it is an important target in therapy against SARS-CoV (Yoshimoto, 2020).

Non-structural protein 4 (NSP4) also contains transmembrane domain, and among 16 non-structural proteins, only NSP3, NSP4 and NSP6 possess this region within their structure. NSP4

interacts with numerous host proteins, but with NSP3 as well, and this interaction is considered as essential for viral replication (Yoshimoto, 2020).

3C-like protease, or non-structural protein 5 (NSP), also called main protease, plays an important role in viral proteins synthesis by proteolyzing pp1a and pp1ab into non-structural proteins (Asghari *et al.*, 2020). This macromolecule also inhibits interferon I signaling procedure by the decomposition of STAT 1 transcription factor and the interaction in NF- κ B cascades (Asghari *et al.*, 2020). As NSP5 has a significant role in SARS-CoV-2 life cycle, it could be a promising drug target in antiviral therapy.

As above mentioned, non-structural protein 6 is also transmembrane protein. NSP6 can induce the membrane proliferation and the process of Atg5 and LC3II-positive vesicles production normally observed in autophagy (J Alsaadi and Jones, 2019). However, these autophagosomes are quite different in size and number in comparison with those formed in starvation.

The non-structural protein 7 (NSP7) together with non-structural protein 8 (NSP8) acts as accessory subunit of RdRp of SARS-CoV-2 (Hillen *et al.*, 2020). This polymerase is composed of main subunit NSP12. The NT residue of NSP8 is essential for the interaction with NSP7 in the process of heterodimer formation (Asghari *et al.*, 2020).

The non-structural protein 9 (NSP9) is involved in SARS-CoV-2 replication as a ssRNA-binding protein (Naqvi *et al.*, 2020). NSP9 interacts with DEAD-box RNA helicase 5 (DDX5) cellular protein and this interaction is essential for the complete viral replication (Yoshimoto, 2020).

The next protein from this group of SARS-CoV-2 proteins is non-structural protein 10 (NSP10). This macromolecule is also involved in the viral replication and it contains 148 amino acids with two zinc finger domains necessary for enzymatic interaction (Asghari *et al.*, 2020). It interacts with both NSP14 and NSP16. The first interaction activates NSP14 (this non-structural protein is known to work as an S-adenosylmethionine (SAM)-dependent (guanine-N7) methyl transferase (N7-MTase)) (Yoshimoto, 2020). The second interaction stimulates the activity of NSP16 (2'-O-methyltransferase) (Naqvi *et al.*, 2020).

The role of non-structural protein 11 (NSP11) is still not completely clear. However, according to some studies this protein has two main domains, NTD and CTD, and performs endo-ribonuclease activity (Asghari *et al.*, 2020).

The 12th protein from this category, non-structural protein 12 (NSP12) is RNA-dependent RNA polymerase (RdRp). It was already mentioned that this protein needs to be in a complex with NSP7 and NSP8 in order to perform full function. This enzyme ranges from 240 to 450 kD and is a crucial enzyme for the survival of majority of RNA viruses (Aftab *et al.*, 2020). The evidence revealed that vitamin B12 (methylcobalamin) has an affinity to bind for active site of RdRp and inhibit it (Asghari *et al.*, 2020). Thus, this vitamin could be also used against SARS-CoV-2 (Asghari *et al.*, 2020). Nevertheless, it is important to emphasize that the first approved drug in the therapy against SARS-CoV-2, remdesivir, acts by affecting NSP12. Remdesivir is a member of one of the oldest classes of the drugs – nucleoside analogues (Pruijssers and Denison, 2019). More precisely, it is an adenosine analogue that was used in the treatment against SARS-CoV and MERS-CoV as well (Asghari *et al.*, 2020). The complete mechanism of action of this antiviral drug, together with other current therapy approaches against SARS-CoV-2 will be discussed hereafter.

The non-structural protein 13 is a helicase that is a member of super family 1 (SF1) (Asghari *et al.*, 2020). Helicases are the class of enzymes also essential for the viral lifecycle. They have an ability to unwind both double-stranded and single-stranded nucleic acid chains with energy derived from NTP hydrolysis (ProteoGenix, 2020). The mechanism of this reaction involves hydrolyzing of deoxyribonucleotide triphosphates (dNTPs) and ribonucleotide triphosphates (rNTPs) (Asghari *et al.*, 2020). The binding of NSP12 to this helicase can induce the activity of this enzyme (Yoshimoto, 2020). The evidence shows that NSP13 can hydrolyze different forms of NTPs and unwind NTP-dependent RNA helices via RNA helicase and NTPase activity. Some divalent ions (Ca^{2+} , Mg^{2+} , Zn^{2+}) can impair the function of this enzyme (Asghari *et al.*, 2020). The analyzes reported that bismuth salts, such as ranitidine bismuth citrate (RBC) and potassium citrate (BPC) can significantly decrease the activity of both NTPase and RNA helicase in SARS-CoV-2 (Shu *et al.*, 2020). Although further investigations are necessary to confirm these results, NSP13 could be a valuable target in a therapy against a novel coronavirus.

The non-structural protein 14 (NSp14) in coronaviruses has exoribonuclease activity in 3'-5' direction and N7-methyltransferase activity (Yoshimoto, 2020). This protein shows antagonistic activity against IFN, that could be potentially used in the therapy against coronaviruses. (Yuen *et al.*, 2020). As most non-structural proteins, NSP14 is also expressed in the MERS-CoV and SARS-CoV.

The 15th non-structural (NSP15) protein in SARS-CoV-2 is a endoribonuclease that has three domains (NT, central and CT) and structure consists of 345 amino acids (Asghari *et al.*, 2020). This enzyme breaks down double-stranded and single-stranded RNA by the identification and cleavage of uridine sites (Bhardwaj *et al.*, 2008). NSP15 has a sequence homology with SARS-CoV and other coronaviruses of 95% and 88%, respectively (Asghari *et al.*, 2020). This sequence variations could be related to SARS-CoV-2 virulence.

The last non-structural protein is numbered 16 (NSP16). It is a methyltransferase that is involved in mRNA cap 2'-O-ribose methylation to the 5'-cap structure of viral mRNAs (Naqvi *et al.*, 2020). As previously mentioned, NSP16 generates a complex with NSP10 that provides efficient viral translation and escape from host immune response. It can interfere with interferon β and also downregulate RIG1 and MDA5 proteins (Asghari *et al.*, 2020). The evidence indicates that certain drugs as Dolutegravir and Bictegravir could inhibit NSP16 and thus they might present another therapy solution against SARS-CoV-2 (Beck *et al.*, 2020).

1.6. Pathogenesis and Clinical Manifestations of COVID-19

The disease caused by SARS-CoV-2 was firstly named severe specific contagious pneumonia (SSCP), afterwards Wuhan pneumonia and finally COVID-19 (Dhama *et al.*, 2020). The clinical presentation of novel coronavirus shows lower severity level in comparison to SARS-CoV and MERS-CoV. However, the transmission rate is significantly higher in the case of SARS-CoV-2 and the association of these two factors could be main reason for the massive spreading of the virus and the global pandemic. Although the incubation period varies from 2-14 days, depending on patient's age and immunity, but also exposure conditions, according to the study, average incubation period for SARS-CoV-2 is 6 days (Wassie *et al.*, 2020). The 2019-nCoVs primarily attacks the cells in the lower and upper respiratory tract (Fani, Teimoori and Ghafari, 2020). Consequently, the most common symptoms of COVID-19 disease include fever, fatigue, cough, myalgia and from recently loss of smell, frequently followed by loss of taste (Butowt and von Bartheld, 2020). As previously mentioned, these manifestations can range from asymptomatic or mild to severe health conditions that demand urgent medical intervention. Nevertheless, it is essential to understand how certain factors (such as age, gender and comorbidity) and human immune response can determine the outcome of this disease. Moreover, the response of various

cells and tissues in human body to the SARS-CoV-2 infection also needs to be taken into consideration, especially in a manner of potential complications and consequences of the illness.

1.6.1. Immune Response to SARS-CoV-2 Infection

The control of the COVID-19 disease directly depends on patient's immune response and its dysregulation could have a crucial role in the modification of clinical outcomes (Portela Sousa and Brites, 2020). The first stage in the development of immune defense is a recognition of the virus or its surface epitopes (pathogen-associated molecular patterns (PAMPs) by pattern recognition receptors (PRRS) such as retinoic acid-inducible (RIG-I) type I receptors and toll-like receptors (TLR3, TLR7 and TLR9) (Portela Sousa and Brites, 2020). This initial response is a part of innate immunity and it represents the first line of defense. The receptor activation is followed by signal pathways triggering, such as NF- κ B, IRF3, and AP-1 which synergistically stimulate the production of type I interferon (IFN-I) (Taefehshokr et al., 2020). The IFN-I (two main subtypes are IFN- α and IFN- β) is a large subgroup of interferon proteins that is involved in the regulation of immune response. They act in both paracrine and autocrine manner by binding for IFN- α/β receptor (IFNAR) and stimulating the expression of numerous interferon-stimulated genes (ISGs) that induce antiviral activities of host cells (Lee and Shin, 2020). These genes are essential for the early antiviral response and they are mainly stimulated by the initiation of Jak/STAT pathway (mediated through IFNAR activation) that results in STAT1/2 heterodimers and STAT1 homodimers binding for the promoter domains of ISGs (Taefehshokr *et al.*, 2020). Although these interferons have systematic antiviral activity, their stimulation triggers secretion of various proinflammatory cytokines and chemokines such as IL-6, IL-1 β and TNF (Tumor Necrosis Factor). Consequently, type I IFNs are considered as one of the substantial factors for cytokine release syndrome (CRS) that characterizes fatal cases of COVID-19 (Portela Sousa and Brites, 2020). Another type of cytokines that are involved in innate immune response against SARS-CoV-2 are III IFNs. (IFN- λ s). They are secreted by different immune cells, such as neutrophils and dendritic cells, but more important, epithelial cells also known as barrier cells that are widely widespread in human body (Portela Sousa and Brites, 2020). Unlike type I IFNs, IFN- λ s possess controlled antiviral protection without following hyperinflammatory response and tissue damage (Broggi *et al.*, 2020). In the infections with lower viral load, III IFNs are sufficient to protect and control viral spreading and maintain immunity. However, the infections with higher viral loads require the activation of I IFNs that provide improved antiviral response

(Portela Sousa and Brites, 2020). The importance of innate immunity in early response to viral infections was confirmed in both SARS-CoV and MERS-CoV. In these diseases, the viruses were able to suppress the response to viral infection by interfering with signaling pathways involved in the production of IFNs (García-Salido, 2020). The non-structural proteins of these two coronaviruses were involved in these interactions and could be a main reason for the impaired function of host innate immune cells (Shah *et al.*, 2020). This dysregulation of IFNs signaling pathways in the early stages of the viral infection affects the overall cytokine productions and has a significant impact on the process of inflammation. The recent investigations revealed that SARS-CoV-2 has developed multiple mechanisms to interfere these stages of antiviral response as well, probably via similar procedures (Taefehshokr *et al.*, 2020). Thus, IFN signaling pathway is targeted by non-structural proteins 13 and 15, and ORF9b, while NF- κ B pathway is dysregulated by non-structural protein 13 and ORF9c (Azkur *et al.*, 2020).

The novel coronavirus activates not only innate, but adaptive immunity as well. According to WHO, innate immune response is responsible for early and mild symptoms of the disease such as weakness, fever and muscle aches. The adaptive immunity is triggered after 6-8 days and involves activation of both B and T cells. It is worth highlighting that weaker innate immune response (in elderly or patients with suppressed immunity) can cause delayed activation of adaptive immunity which could be important factor for disease prognosis (WHO, 2020). Once the virus is inside the host's cell, various viral particles are presented within class I major histocompatibility complex (MHC I) to CD8⁺ cytotoxic T cells (Azkur *et al.*, 2020). For a short time, these cells produce memory T cells through the process of virus-specific development and clonal expansion. The antigen-presenting cells (dendritic cells and macrophages) present virus and its peptides not only to CD8⁺ T cells, but also to CD4⁺ T cells through to class II major histocompatibility complex (MHC II) molecules (Azkur *et al.*, 2020). T immune cells are part of cell-mediated adaptive immune response and their main task is the elimination of infected cells. On the other side, B cells can directly recognize viral particles (without MHC molecules) and be activated (Azkur *et al.*, 2020). Another route for the activation of these immune cells is via CD4⁺ T cells. Essentially, in both procedures, the result is the production of specific antibodies by activated B cells. Within the first week of infection and symptoms manifestations, IgM isotype of virus specific antibodies is observed, while IgG isotype of antibodies follows the early response and provide long-lasting immunity against the virus (Azkur *et al.*, 2020). A long-term study

found that IgA type of antibodies against spike glycoprotein are also produced in early stage of infection (during the first week), and their peak in concentration is reached around 20 days after the onset of symptoms (Poland, Ovsyannikova and Kennedy, 2020). Notwithstanding, another study revealed that the patients with mild clinical manifestations of COVID-19 have characteristic and rapid decline of IgG antibodies titers within the 2-4 months. This suggests that SARS-CoV-2 induced humoral immunity in those patients with mild symptoms might not be long-lasting (Poland, Ovsyannikova and Kennedy, 2020).

As aforesaid, beside the humoral response, another component of the adaptive immunity is cell-mediated response. The main mediators of cellular immunity are T cells which also have important role in the immune system's fight against SARS-CoV-2 (Annette Plüddemann, 2020). After death of the cells infected by 2019-nCov, various molecules called damage-associated molecular patterns (DAMPs) are released. They are recognized by adjacent endothelial and epithelial cells, but by macrophages as well, and their activation and recognition will result in the production of pro-inflammatory cytokines and chemokines (Annette Plüddemann, 2020). These cytokines will engage additional macrophages, monocytes and T cells that will be recruited to the site of infection. The interferon-gamma ($\text{IFN}\gamma$) will be produced by activated T cells as a part of inflammatory response (Annette Plüddemann, 2020). As already stated, the cellular immune response involves two types of T cells: CD4^+ and CD8^+ cells. The novel coronavirus activates these cells in the first week of infection and virus-specified memory CD4^+ and CD8^+ cells reach peak of the concentration within 2 weeks, while they remain detectable at lower levels for at least 100 days (Stephens and McElrath, 2020). CD4^+ T helper (Th) cells interact with CD8^+ cells and activate cytotoxic response whose task is the elimination of the infected cells. Another subtype of CD4^+ cells are follicular helper T (TFH) cells and they interact with B cells and promote humoral response through the cell-cell interactions and release of cytokines (Annette Plüddemann, 2020). The virus specific CD4^+ and CD8^+ memory T cells were detected in all patients, from those with mild symptoms to those with severe forms of disease. It is essential to stress that these cells have significant role in the control of viral replication through the interaction and activation of other adaptive immune cells and cytotoxic activity (Stephens and McElrath, 2020). On the other hand, numerous aspects of the role of T cells in COVID-19 disease development and progression stay unclear. Do memory T cells provide the protection against SARS-CoV-2 in the absence of virus specific antibodies? Could we use these cells for the

prevention of disease? Further studies and full understanding of immune response to SARS-CoV-2 are necessary for the answers on these questions, but many others as well.

1.6.2. Clinical Characteristics and Potential Complications of SARS-CoV-2 Infection

Balanced and controlled immune response to viral infections might be difficult for certain patients. Although vast of people infected by SARS-CoV-2 have only mild or moderate clinical manifestations, significant number of patients, particularly elderly and those with the certain chronic diseases (such as diabetes, hypertension and obesity), develop severe forms of a disease with potentially fatal outcome. Even though main mechanisms and factors that determine which form of disease will individual manifest still remain not totally clear, probable cause for mortality of patients with COVID-19 disease is disbalanced immune response to the infection and acute respiratory distress syndrome (ARDS) as its consequence (Chowdhury *et al.*, 2020).

The pathological failures of patients infected with novel coronavirus resemble those observed in previous two epidemics caused by viruses from the same family: SARS-CoV and MERS-CoV. They are characterized by high serum levels of IL-10, IL-8, IL-6, IL-2R and TNF- α in patients with severe form of disease, while these cytokine levels were within normal range in mild and moderate cases (Melenotte *et al.*, 2020). Besides, in hospitalized cases with serious clinical presentations, bronchoalveolar lavages showed interstitial mononuclear infiltrates with pulmonary edema, hyaline formation, lymphocytes and pneumocytes residues (Melenotte *et al.*, 2020). The postmortem pathology of patients with diagnosed COVID-19 disease reveals extreme form of lymphopenia and hyperactivated T cells in the blood. The levels of regulatory T cells were also low and this is especially notable given their essential role in prevention of excessive inflammation and immune homeostasis (Marbán and Akhmerov, 2020). According to the earlier studies on H7N9 viral infection, the dysregulated immune response in these patients is probably resulted from over-activation of T cells (Melenotte *et al.*, 2020). The values of immunological parameters in severe cases of COVID-19 suggest that these forms of disease are dictated as a cytokine release syndrome (CRS) or cytokine storm. Based on present clinical experience, the IL-6 could be a “marker” for the disease severity as most non-survivors had very high blood levels of this inflammatory mediator (Hojyo *et al.*, 2020).

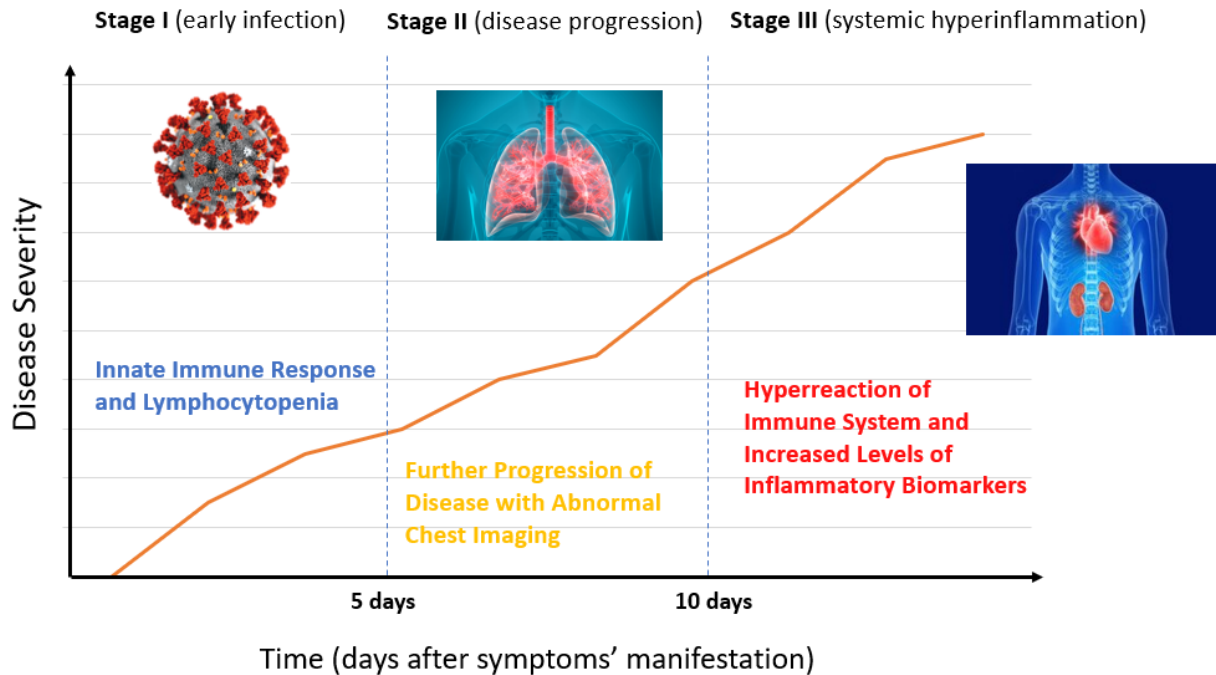


Figure 6. COVID-19 stages and progression of acute disease. Adapted from (Marbán and Akhmerov, 2020)

Acute COVID-19 disease can be divided in three phases depending on illness progression. The first stage of disease is characterized by the viral infiltration of lung parenchyma and initiation of viral replication. The innate immunity is employed and patient starts to develop mild symptoms of disease (Marbán and Akhmerov, 2020). The key laboratory parameter of this stage is lymphocytopenia. If patient's immunity does not manage to control the progression of the infection, second (pulmonary) stage is reached. The main characteristic of this phase is abnormal chest imaging (Marbán and Akhmerov, 2020). Exaggerated systemic inflammation (cytokine storm) together with ARDS and lymphocytopenia is a main hallmark of severe and fatal cases of COVID-19 disease and it represents an advanced (third) phase of illness (**Figure 6**) (Marbán and Akhmerov, 2020).

Beside ARDS, another consequence of dysregulated immune response to SARS-CoV-2 is Multi-Organ Dysfunctional Syndrome (MODS) (Shah *et al.*, 2020). This pathological state is associated with intravascular coagulation which is mainly mediated by inflammatory cytokine, especially IL-6 (Hojyo *et al.*, 2020). Although the hyper-reaction starts in the tissue of initial injury (in this case lungs), the amplified systemic inflammatory response can have devastating effects on other organs. The patients with severe form of disease exhibit multiorgan failure characterized by

increased D-dimer and lower platelets count which explain microthrombi of the lungs, heart, liver, kidneys, brain, hands and lower limbs (Hojyo *et al.*, 2020). Along with these observations, the investigations show that biomarkers of cardiac injury acquired after SARS-CoV-2 infection correlate with increased levels of inflammatory markers (Marbán and Akhmerov, 2020). Another factor for MODS is the fact that SARS-CoV-2 infection of endothelial cells also causes cellular death, which generates vascular leakage and promotes a cytopathic effect on epithelial cells in lower airways (Hojyo *et al.*, 2020).

Eventually, it seems that disease severity and mortality originate from cytokine storm and ARDS initialized by viral lung infection that leads to multiorgan failure across the human body. The systemic inflammation is also induced by vascular hyperpermeability and increased expression of ACE2 on endothelial cells of veins and arteries (Hojyo *et al.*, 2020). Due to these reasons, vast of therapeutic approaches against COVID-19 are focused on finding a solution for control and suppress of immunity hyperreaction.

1.7. Risk Factors and Diagnosis of COVID-19

The factors that affect the transmission of SARS-CoV-2 can be divided into two groups: social (demographic) and clinical factors (de Lusignan *et al.*, 2020). Besides reducing social contacts which includes limitation of traveling, closing schools and universities, hotels and restaurants and asking people to work from home whenever possible, other social aspects should also be taken into consideration. Age and gender are well-established risk factors for the COVID-19, especially for severe forms of the disease. According to *Worldometers, 2021* ([URL Worldometers](#)), as of May 13, 2020, 48,7% of total deaths in USA were patients older than 75 years, while 3,96% of deaths were people 45 years old or younger. The same source with data provided by *NYC Health, 2021* ([URL NYC Health](#)) reveals that as of April 14, 2020 male sex was associated with higher risk for death due to COVID-19 (61,8%) in comparison to female (38,2%). In Austria, the numbers are similar but not quite the same. According to our analysis, we can see that 80% of all deaths due to current the pandemic were concerning patients older than 75 years. On the other hand, the situation with gender susceptibility to fatal outcome of the disease is different. Thus, out of the total number of deaths, 51% were female patients in Austria (*Figure 7*).

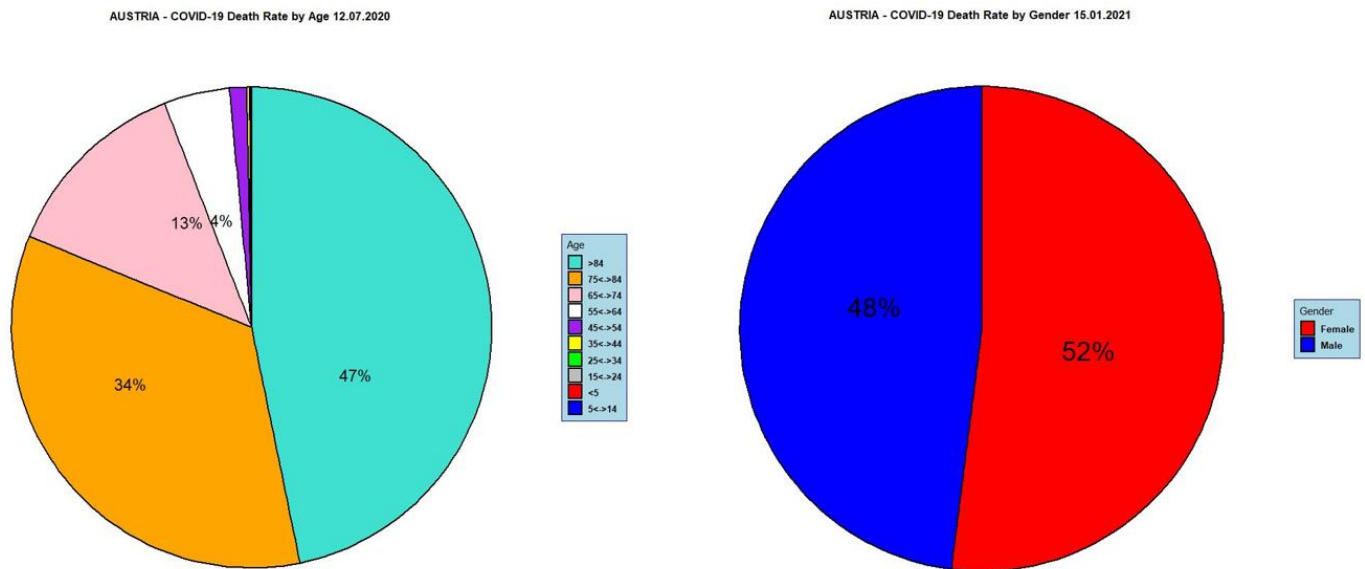


Figure 7. Age and Gender Percentage in Austria. Data Source: AGES Dashboard COVID19 (<https://covid19-dashboard.ages.at/>). Data collected on 15.01.2020.

It is important to understand how these factors (age, sex) could alter immune response and serological status during SARS-CoV-2 infection. Investigations showed that S1-specific IgG response was remarkably higher in women older than 40 years than in the other two groups (women younger than 40 years and men) which could provide them more effective antibody response in comparison to the other two groups (Poland, Ovsyannikova and Kennedy, 2020). On the other hand, a small study on convalescent individuals revealed that men had significantly higher values of anti-RBD, but also anti-spike glycoprotein IgG titers in comparison to women (Poland, Ovsyannikova and Kennedy, 2020). In regard to age, middle age (40-59 year) and older age (60-85 years) patients had notably higher titers of SARS-CoV-2 neutralizing antibodies than younger patients (Sanyaolu *et al.*, 2020). It is obvious that different manifestations of the disease originate from differences in immune responses, whether age or gender groups are compared. However, these investigations showed that both genders of older patients have higher titers of virus-specific antibodies in comparison to younger patients. Although high IgG antibody immune response should provide a long protection from new infections, according to these findings, younger patients, that mostly have mild manifestations of COVID-19 disease, could not be able to develop sufficient quantities of SARS-CoV-2 specific IgG antibodies.

Another factor that could have an impact on the infection and disease development is the blood type. While detailed mechanisms for this observation are still not known, the studies confirmed that risk for the infection was slightly increased in patients with non-O blood type (Zietz, Zucker and Tatonetti, 2020). According to the same study, the risk of death was the highest for AB blood type and lowest for blood types A and B, while Rh-negative blood type showed protective effect for all outcomes (Zietz, Zucker and Tatonetti, 2020). These findings suggest that the blood type might have an important role in COVID-19, both for its prevention and manifestations.

Finally, yet important, the presence of other medical conditions together with SARS-CoV-2 infection could have decisive influence on disease progression and outcome. From the data collected from January to April 2020, patients with comorbidities such as diabetes mellitus, cardiovascular diseases (primarily hypertension), renal disorders and cancer, have higher risk for severe forms of the disease (Sanyaolu *et al.*, 2020). These findings arise from previously explained systemic hyperinflammation and multiorgan failure. A new inflammation with dysregulated immune response that seizes the organs which are already pathologically damaged, will evidently aggravate the patient's condition. Moreover, older people, especially those over 65 years, with some of the listed medical conditions, have increased admission rate into ICU (intensive care unit) and mortality from COVID-19 disease (Sanyaolu *et al.*, 2020). Due to all of that, patients that suffer from a specific chronic disease should take additional precautions to prevent infection with SARS-CoV-2.

Common symptoms of COVID-19 include fatigue, cough, fever, muscle ache, chest pain and dyspnea (Fani, Teimoori and Ghafari, 2020). As previously mentioned, loss of smell with or without loss of taste represents one of the SARS-CoV-2 infection hallmarks. Nonetheless, besides these clinical manifestations, the diagnosis of viral infection requires confirmation by some of the approved testing methods. There are two main types of tests: diagnostic and serological (antibody) tests. According to ECDC, specimen for diagnosis tests can be sampled from lower (bronchoalveolar lavage – BAL or sputum) or upper respiratory compartments (nasal aspirate/wash, saliva or nasopharyngeal swabs). The most commonly used and probably the most reliable testing method to date is molecular real-time RT-PCR test (Dhama *et al.*, 2020). This is a nucleic acid detection technique that requires samples storage at four degrees Celsius, which is followed by amplification of genetic material extracted from samples through reverse polymerase chain reaction (RT-PCR) that includes a synthesis of double-stranded DNA from RNA mold

(Cascella *et al.*, 2020). Once the synthesized genetic material is sufficient, the analysis and search for the SARS-CoV-2 conserved genes is performed. Another type of testing method that requires samples from respiratory tract is Rapid Antigen Detection (RAD) tests. These tests detect the presence of a viral antigen, commonly a fragment of a surface protein. Although they provide rapid test results, their reliability is questionable due to potential false-negative results and these tests should be used only as adjunct to RT-PCR tests (Mak *et al.*, 2020).

Antibody (serological) tests are used for the detection of specific antibodies generated against SARS-CoV-2. As aforementioned, 7-14 days after infection activated B-cells start to produce virus-specific antibodies (IgM, IgA and IgG) (Zalzala, 2020). IgM and IgA antibodies appear first, but IgG type of antibodies provide long-lasting immune protection. Due to its immunogenicity, neutralizing antibodies against the RBD of the S protein are most commonly used (Zalzala, 2020). Although lateral flow assays (LFA) and chemoluminescence assays (CLIA) could be also used, a standard method for SARS-CoV-2 antibody test are enzyme-linked immunosorbent assays (ELISA).

It is important to emphasize the significance of other clinical tests such as laboratory examinations and Computed Tomography (CT) imaging. Particularly during the first stages of disease, laboratory values of white blood cell count (WBC), lymphocytes, C-reactive protein, muscle and liver enzymes and other indicators could provide essential information concerning immune response and disease progression and prognosis (Cascella *et al.*, 2020). Chest CT and X-ray examinations represent valuable techniques for determination of lungs' condition, but they are mostly applied after some other indicators reveal disease progression.

1.8. Specific Aims of the Project

The main goal of this project was to provide a comprehensive overview of SARS-CoV-2 and COVID-19 disease, with a specific focus on the following specific aims:

- The development of R and Python scripts for automated data visualization of trends in COVID-19 disease spread and related parameters (e.g., lethality).
- The comparison of currently available drug targets for drug repurposing strategies in terms of sequence-based and secondary-structure based similarities of protein targets.
- The Analysis of SARS-CoV-2 protein targets and their ligands obtained from different publicly available data sources.
- The investigation on potential drug-repurposing opportunities.
- The generation of a state-of-the-art summary of current treatment options for COVID-19.

2. Materials and Methods

Definitions

Epidemiological research

In this Master thesis data analyses have been performed that fall under the category of Epidemiological Research. Epidemiology is a branch of medical science that studies and investigates the determinants and the distribution of health-related conditions or events in specified populations, and the application of these studies in the control of health issues (*Principles of Epidemiology*, CDC, 2021).

Cheminformatics & Drug Discovery

As a second branch of research included in this master's thesis, we have been using Methods of Cheminformatics and Drug Discovery in order to deliver first hints for performing more extensive drug repurposing investigation. Cheminformatics is a discipline which uses informational and computational methods for the of problems in chemistry (*Cheminformatics - Latest research and news* | *Nature*, 2021). Cheminformatics techniques are widely used in efforts that include evaluation and investigation of large number of compounds such as drug discovery. Drug discovery is a complex process performed with the aim of identification small molecules or large biomolecules that could be transformed into drug candidates after series of comprehensive evaluations (Sinha and Vohora, 2017). It is followed by phases of preclinical and clinical drug development. On the other hand, drug repurposing (repositioning) is a field in drug discovery that identifies new treatment options for already existing drugs (Doan *et al.*, 2011). Many pharmaceutical companies pursue this strategy in drug discovery as it increases productivity by decreasing the time of development.

2.1. Databases

An essential prerequisite for both data visualization and protein target analysis is the availability of reliable and comprehensive data sources. The COVID-19 pandemic still represents a major health issue worldwide and thus, the significance of efforts put into obtaining, collecting and using epidemiological data as well as small molecule and protein structural data should be

stressed. At some parts of this thesis, specific information from one resource was combined with some other type of the information provided by another data supplier. It is worth mentioning that some countries do not share all their epidemiological data, so it was not always feasible to compare and analyze all epidemiological aspects for all countries.

Our World in Data (OWD) (<https://ourworldindata.org/>) was founded by economist and social historian Max Roser. This web platform provides long-term data storage that focuses on large global issues such as poverty, climate change and diseases (COVID-19 pandemic among them) (*Our World in Data - Wikipedia*). The OWD research team is based at the University of Oxford and besides providing diverse databases, this webpage presents comprehensive trends on how our world is changing through maps and graph visualizations (*About - Our World in Data*). This data source is one of the most prominent and reliable data supplier for the current pandemic and provides information for different aspects of COVI-19 epidemiology. Most of the data used for the creation of charts and diagrams in this thesis originate from this resource, and it was mainly fetched via downloading several CSV files.

Die Agentur für Gesundheit und Ernährungssicherheit GmbH (AGES) (<https://www.ages.at/>) is a company owned by the Austrian Federal Ministry of Labor, Social Affairs, Health and Consumer protection and the Austrian Federal Ministry of Sustainability and Tourism. AGES was founded in June 2002 and together with other institutions it deals with questions related to public health, animal health, consumer protection, food safety and medical and drug safety (*Agentur für Gesundheit und Ernährungssicherheit – Wikipedia*). This company provides the official COVID-19 Dashboard for Austria, with all epidemiological data. Nearly all generated graphs and diagrams related to Austria in this study were created on basis of data obtained from AGES by the download of several CSV files.

The Humanitarian Data Exchange (HDX) (<https://data.humdata.org/>) is open web platform launched in July 2014 for sharing data among the community during the crises. The main goal of this organization is to make humanitarian data easy to find, access and use for analysis (*What We Do – The Centre for Humanitarian Data*). HDX is managed by the United Nations Office for the Coordination of Humanitarian Affairs (OCHA) and it is focused on four main areas: data services, data literacy, data responsibility and predictive analysis (*Data Fellows Programme –*

The Centre for Humanitarian Data). This platform provides plenty of very useful data related to the COVID-19 pandemic, especially time series data sets.

ChEMBL (<https://www.ebi.ac.uk/chembl/>) is manually curated chemical database of bioactive compounds with drug-like properties, maintained by the European Bioinformatics Institute (EBI) of the European Molecular Biology Laboratory (EMBL) based in Hinxton, UK (*ChEMBL - Wikipedia*). This database contains 2-D structures, abstracted bioactivities (such as ADMET data, pharmacology etc.) and calculated properties (e.g. Molecular Weight, Lipinski Parameters, logP etc.) of drug-like molecules (*About - ChEMBL Interface Documentation*). It brings together genomic, chemical and bioactivity data to improve the translation of genomic information into effective new drugs. The data is curated from scientific literature and covers a significant amount of structure-activity relationship (SAR) and modern drug discovery data (Papadatos *et al.*, 2014). This open data source was used for fetching COVID-19 ligand bioactivity data via a KNIME workflow (see chapter 2.3) through Application Programming Interfaces (APIs).

PubChem (<https://pubchem.ncbi.nlm.nih.gov/>) is another data source used for retrieving COVID-19 ligand data in this thesis. It was launched in 2004 and has become world's largest public repository for information related to chemical substances and their biological properties (Wang, Cheng and Bryant, 2017). PubChem was created as a component of the Molecular Libraries Roadmap Initiatives of the US National Institutes of Health (NIH) and it provides searching and sorting of chemicals by their name, physical or chemical properties, biological activities, toxicity, molecular formula, literature citations and more (*PubChem - Wikipedia*). The same as ChEMBL, PubChem provides an API which was used for fetching its data via KNIME.

The International Union of Basic and Clinical Pharmacology (IUPHAR) / British Pharmacological Society (BPS) (<https://www.guidetopharmacology.org/>) is a resource of ligand-bioactivity-target relationships, with the majority of its data derived from high quality medicinal and pharmacological literature (*IUPHAR/BPS Guide to PHARMACOLOGY*). This platform provides pharmacological data with quantitative information on drug targets, the registered drugs that act on them, but also on experimental drugs and compounds. The web user interface of this database is organized into two levels (*IUPHAR/BPS Guide to PHARMACOLOGY*). The first one is an initial view that gives an overview on certain targets. The second one provides more details on pharmacological, physiological, genetic and other types

of target information. The data can also be accessed through an API and it is enhanced with hyperlinks to provide additional information from other databases such as Drugbank, ChEMBL and PubChem. IUPHAR was the third source for retrieving COVID-19 ligand data via a KNIME workflow used in this thesis.

UniProt (<https://www.uniprot.org/>) is also an open access database of protein sequence information and annotations for proteins across all branches of life. UniProt contains large amounts of data about biological and genetic properties of certain proteins derived from scientific literature (*UniProt - Wikipedia*). It was launched and maintained as a collaboration between the European Bioinformatics Institute (EMBL-EBI) and the SIB Swiss Institute of Bioinformatics and the Protein Information Resource (PIR) (*About UniProt*). UniProt provides a special portal for latest SARS-CoV-2 protein data with all entries and receptors that are updated and curated on a regular basis. The data from this portal was widely used in retrieving information for SARS-CoV-2 protein targets.

Protein Data Bank (PDB) (<https://www.rcsb.org/>) was launched as the first open access data resource in medicine and biology fields (*RCSB PDB: PDB History*). Today, PDB provides access to the 3D data of large biological molecules (primarily proteins and nucleic acids) through the web platform and downloadable data sets (*RCSB PDB: About RCSB PDB: Enabling Breakthroughs in Scientific and Biomedical Research and Education*). The data is mainly generated by NMR spectroscopy and X-ray crystallography (and increasingly via cryo-electron microscopy) and submitted by scientists around the world (*Protein Data Bank - Wikipedia*). In this work, PDB was used as a main sources of COVID-19 protein structures that were structurally compared.

2.2. Python

Python is an interpreted and general-purpose programming language founded by Guido van Rossum in 1990 and released in 1991 (*Python (programming language) - Wikipedia*). It is used for developing desktop GUI applications, web and software development, system scripting and more. It offers dynamic typing and binding options that makes it an attractive tool for the rapid application development (*Introduction to Python*). The main advantages of this programming language in comparison to others are improved productivity, free and open-source access and

unique simple syntax (*Python Advantages and Disadvantages - Step in the right direction - TechVidvan*). The simplicity of Python syntax rules further makes it easier for keeping the code base readable and application maintainable. It is important to highlight the compatibility of Python with major platforms and operating systems as well.

As a general-purpose programming language, Python is applied in various fields, from internet and web development to scientific computing and data science. Another important application of this programming language is in the field of deep learning and machine learning approaches, while its various packages also represent powerful instrument for drug discovery research. According to TIOBE's ranking index, Python is currently third most popular programming language right behind C and Java (*index | TIOBE - The Software Quality Company*). The script used in this work will be listed in the appendix of this thesis.

Used Python packages: pandas, pycountry and plotly.

2.3. R Programming Language

R is a free programming language designed for statistical computing and graphics. It was developed by Ross Ihaka and Robert Gentleman in 1993, and R is as a matter of fact an implementation of the S programming language (*R (programming language) - Wikipedia*). This programming language possesses a diverse spectrum of graphical and statistical methods such as linear regression, time series, machine learning algorithms and others (*R: What is R?*). R is extensively used among data scientist and statisticians as a tool for developing statistical software and data analysis. It is compatible with wide variety of platforms: Windows, MacOS and UNIX. Although R possesses a command line interface, several third-party graphical user interfaces have been developed such as RStudio (an integrated developed environment). Besides its wide application for scientific purposes, many large companies including Google, Facebook and Airbnb also use R programming language. According to TIOBE's index popularity of programming languages, R ranks 16th place (*index | TIOBE - The Software Quality Company*).

Due to its accessibility, large variety of data manipulation possibilities and numerous plotting and visualizations options, R finds large implementation in different academic fields, especially in health and life sciences. R and its packages are widely used in different stages of the pharma lifecycle, drug research and discovery, but in clinical trials data analysis as well. Other fields

where R is applied frequently are cheminformatics and bioinformatics. This programming language provides solutions for many issues related to statistical modeling of chemical and biological information.

In this work, R was used for the epidemiological analysis of COVID-19 data. Different graphs and trend plots have been generated in order to deliver a good understanding of COVID-19 epidemiology. All scripts used for visualizations in R will be listed in the appendix. The R packages used in this work are shown in **Table 2**.

Used R packages	
dplyr	gridExtra
dygraphs	lubridate
ggplot2	stringr
ggthemes	scales
grid	tidyr
tidyverse	xts

Table 2. Used R packages

2.4. The Konstanz Information Miner (KNIME)

The KNIME Analytics Platform is an open-source and comprehensive platform for the integration of different computational applications through pipelining workflows (*KNIME - Wikipedia*). This modular data pipelining provides implementations of various components of data mining and machine learning. KNIME was developed in 2004 at the University of Konstanz (Germany) and has gained popularity over the years, especially as a platform for data science (*Open Source Story | KNIME*). Moreover, this platform provides integration of a wide range of applications that can be used in a pipeline management to generate automated workflows in fields such as finances, natural sciences, mathematics, chemistry. The workflows are comprised of joined nodes and metanodes (nodes that contain subworkflows, i.e. in the workflow they look like a single node, but they are composed of many nodes or even more meta nodes) which possess data processing features enabling analysis, modeling, prediction and visualization of data without coding (*KNIME - Wikipedia*).

The KNIME analytics platform provides users to build workflows in a pipeline fashion mode using nodes and metanodes containing software algorithms for data processing and transport. The KNIME graphical interface enables node selection and implementation into workflows by drag and drop fashion. The processed data within a node is connected to the next node via input-export ports. A Traffic light system in each of them creates the visualization of every single node execution that can be tracked. The indication of each is shown in **Figure 8**.

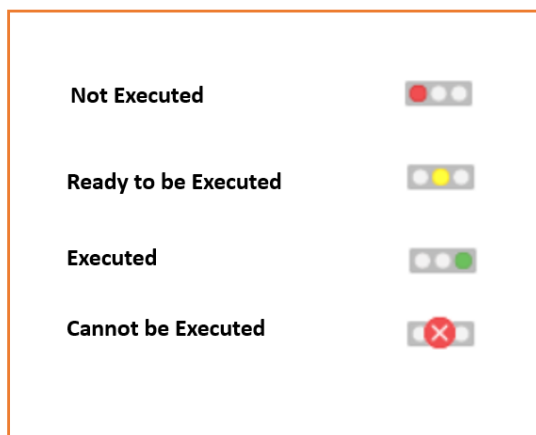


Figure 8. Node Status

KNIME nodes are classified into four main groups based on their function: 1) data import and export; 2) data manipulation and curation; 3) data balancing and filtration; 4) data standardization and transformation. There are also nodes that can be used for data visualization. In order to improve and ease the processes of data mining, nodes are organized to use different sets of machine learning approaches such as decisions trees, clustering, support vector machines and many more. The greatest feature of this platform is the ability to share and integrate self-programmed nodes which is beneficial for the whole KNIME community. A case study workflow was used for retrieving all available ligands for COVID-19 protein targets derived from UniProt protein database (Tuerkova and Zdrazil, 2020). The ligands were obtained from three sources: ChEMBL, PubChem and IUPHAR.

2.5. Iroki – Phylogenetic Tree Visualization Software

Iroki is a user-friendly web application for generating and visualizing phylogenetic and taxonomic trees. This tool provides visualizations of qualitative and quantitative metadata (Moore *et al.*, 2020). Iroki has found its application in exploring trends in biological and ecological fields through a variety of trees with numerous leaf nodes that are customized according to collected data. JavaScript and Scalable Vector Graphics are used for tree rendering in Iroki (Moore *et al.*, 2020). This online tool offers a balance between utility and flexibility by phylogenetic tree visualizations in a clean and easily accessible web interface. In this thesis, Iroki was applied for the generation of COVID-19 protein targets phylogenetic tree according to sequence similarity of these macromolecules.

2.6. DALI – Protein Structure Comparison Server

The Dali server is a web service for the 3D protein structures comparison that consists of three segments: 1) the server that compares newly solved structures against structures in the PDB database; 2) the part that provides browsing precomputed structural neighbors (multiple structural alignment); 3) the option that generates pairwise comparisons of proteins (*Dali server*). The inputs are PDB identifiers or structures uploaded by users. Dali server has a very simple interface that is easy to use. The main advantage of this type of protein alignment is the fact that structure comparison sometimes might reveal biologically important similarities that cannot be detected by sequence alignment alone. In this work, Dali server was used for multiple comparison and analysis of COVID-19 protein structures. More precisely, the 48 potential protein targets were previously clustered based on their sequence similarity and in the next step, each of these clusters was separately structurally compared and analyzed.

2.7. RaptorX Tool

RaptorX is a web server, developed by Xu group, used for protein structure prediction. This tool is mainly used for predicting 3D structures for proteins sequences without homologs in Protein Data Bank (PDB) (*About RaptorX*). For a given input sequence, RaptorX predicts both secondary and tertiary structure, but also solvent accessibility, binding sites etc. This tool also provides

scores (coefficients) that are assigned to each 3D model indicating the quality of the predicted structure: P-value for the relative global quality, GDT (global distance test) and uGDT (un-normalized GDT) for the absolute global quality (*RaptorX* - *Wikipedia*). In this work, RaptorX was used for the prediction of 13 COVID-19 protein structures which were not provided by PDB. This process and the following steps will be discussed further in the results.

3. Results

The results section of this master thesis consists of two parts. The first one discusses data visualizations and interpretations for the understanding of epidemiological aspects of the COVID-19 pandemic. The second part includes an analysis of currently available potential SARS-CoV-2 proteins targets and of their ligands to inform drug repurposing strategies.

3.1. Data Science Approaches to Understand COVID-19 Epidemiology

As previously mentioned, SARS-CoV-2 is not the first virus from *Coronaviridae* family that created serious impact on health systems by uncontrolled spreading and causing severe disease symptoms. The first epidemic outbreak associated to a virus from this family emerged at the end of February 2003 in China (WHO, 2021). The respiratory disease caused by a novel coronavirus (SARS-CoV) was named Severe Acute Respiratory Syndrome (SARS) and this microorganism also had a zoonotic origin as previously discussed. The pathogenesis of SARS is analogous to the one observed in patients infected by 2019-nCov and it is characterized by severe injury of the lower respiratory tract with a tendency of viral dissemination to other organs (Gu and Korteweg, 2007). In 2012, a new human disease caused by another coronavirus (MERS-CoV) emerged in the Middle East, more precisely in Saudi Arabia, and due to its origin, the disease was titled Middle East Respiratory Syndrome (MERS). Investigations revealed that MERS-CoV probably origins from bats, but the transmission to humans included intermediate hosts. MERS-CoV infections led to clinical presentations very similar to those observed for SARS-CoV and SARS-CoV-2.

Although the three viruses and the diseases that they are causing share related etiopathology and origin, their epidemiology widely differs (**Figure 9**).



Figure 9. Total Number of Confirmed Cases and Deaths for SARS, MERS and COVID-19. Data source: Worldometers (<https://www.worldometers.info/coronavirus/>) and WHO (https://www.who.int/csr/sars/country/2003_07_11/en/). Data collected on 17.1.2021

In **Figure 9**, one can see that both SARS and MERS epidemics had significantly lower total numbers of cases (8 437 and 2 494 respectively) in comparison to the current COVID-19 pandemic. On the day of data collection (17.1.2021), total number of patients infected with the novel coronavirus was 95 088 785. The same trend was observed with total number of deaths which was presented on the second chart of **Figure 9**. However, the total numbers of deaths for SARS and MERS were similar although their total case numbers were dissimilar. The explanation for this can be found in **Figure 10** which shows mortality rates for these three pandemics.

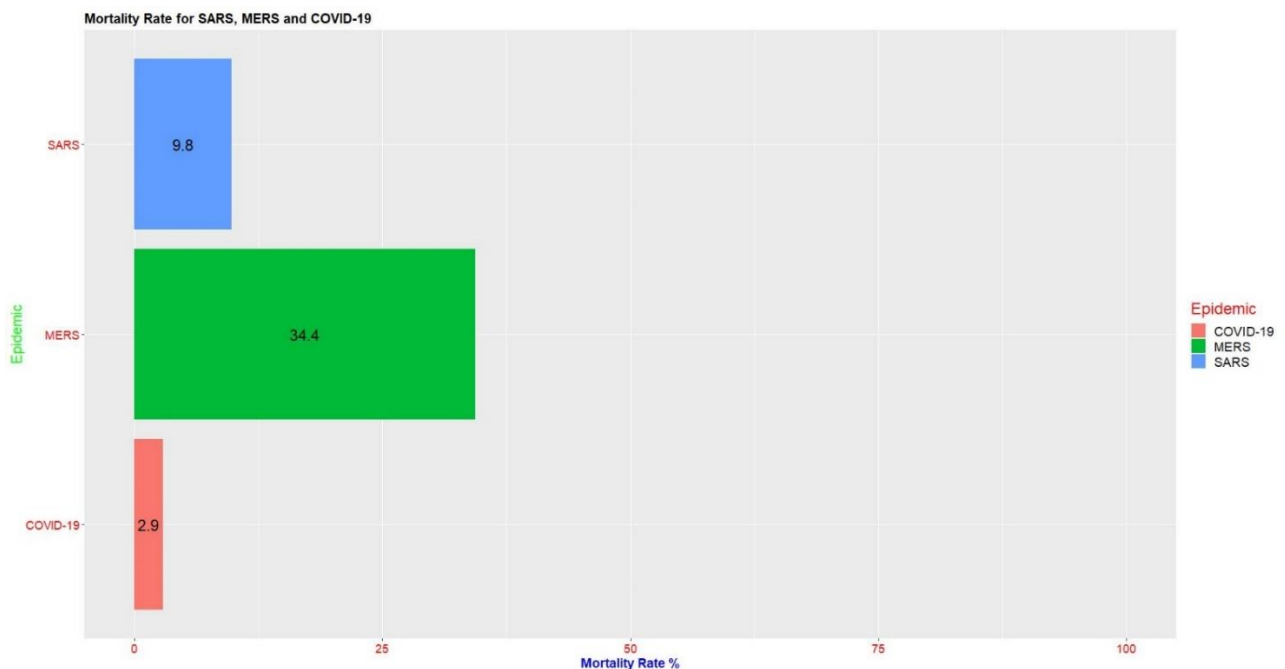


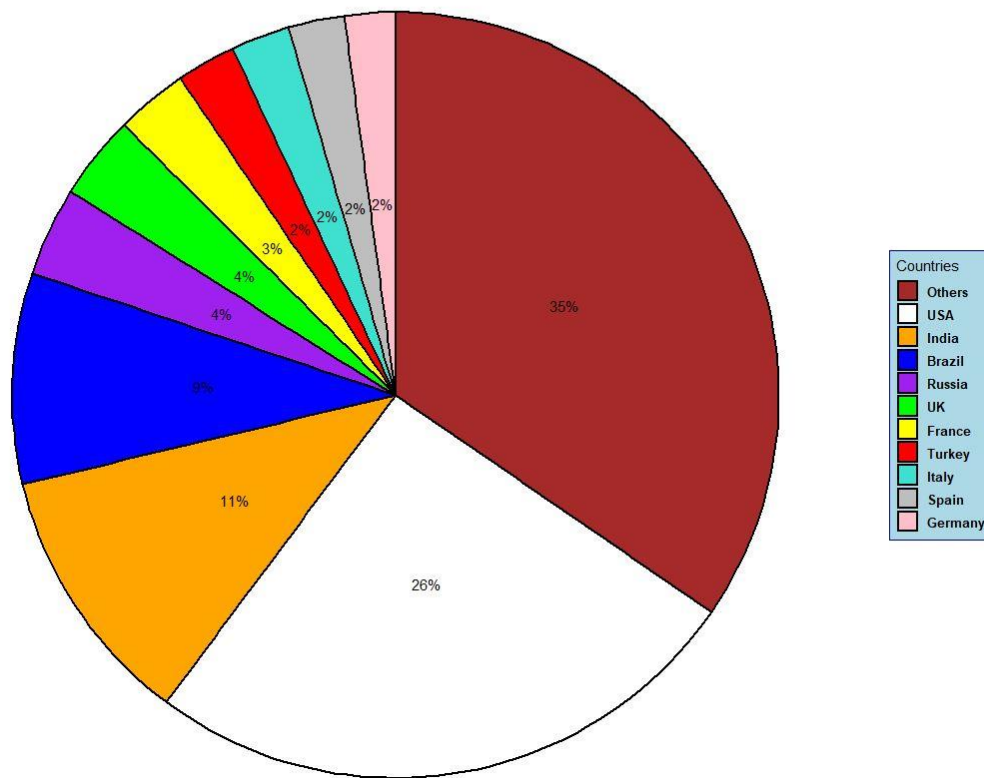
Figure 10. Mortality rates for SARS, MERS and COVID-19. Data Source: Worldometers (<https://www.worldometers.info/coronavirus/>) and WHO (https://www.who.int/csr/sars/country/2003_07_11/en/). Data Collected on 17.1.2021

The very high mortality rate of MERS led to deaths of more than one third of all infected patients. Due to this, the numbers of deaths due to the SARS and MERS epidemics were similar despite notably higher number of patients infected by SARS-CoV. On the other hand, the Figures 9 and 10 both show that COVID-19 total cases and deaths numbers are incomparable higher than those caused by SARS and MERS, but at the same time there is a significant lower mortality rate observed due to the current pandemic. Although it might sound paradox, the lower mortality rate could be one of the reasons for uncontrolled spreading of SARS-CoV-2. Recent studies revealed

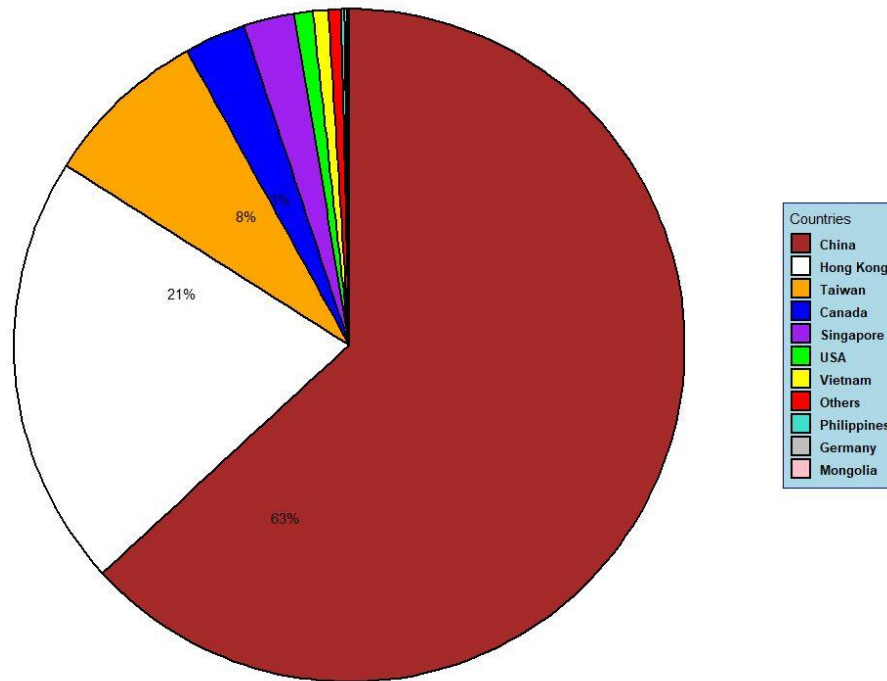
higher contagiousness and lower virulence of the novel corona virus in comparison to earlier SARS-CoV (Ferretti *et al.*, 2020). Together with low mortality, these two factors cause faster spread of this disease, with the majority of patients having asymptomatic or mild clinical manifestations but with a high potential for transmission. On the contrary, a meaningful number of patients in SARS, and especially in the MERS epidemic, developed severe forms of disease (with fatal outcome in a significant number of cases), and that provided their rapid identification and isolation.

In **Figure 11**, we can see the distribution of total COVID-19 cases registered in different countries until mid-January 2021, as well as total SARS and MERS cases in different countries. For SARS-CoV-2 almost half of all cases are concentrated on US, India, and Brazil but the remaining registered cases have been noted for a wide range of different countries. In contrast, MERS and SARS stayed local phenomena, with China registering 84% of all SARS infections (21% in Hong Kong only) and Saudi Arabia with 84% of all registered MERS cases.

COVID-19 TOP 10 Countries (Total Number of Cases -> 95 555 805)



SARS TOP 10 Countries (Total Number of Cases -> 8 432)



MERS TOP 10 Countries (Total Number of Cases -> 2 449)

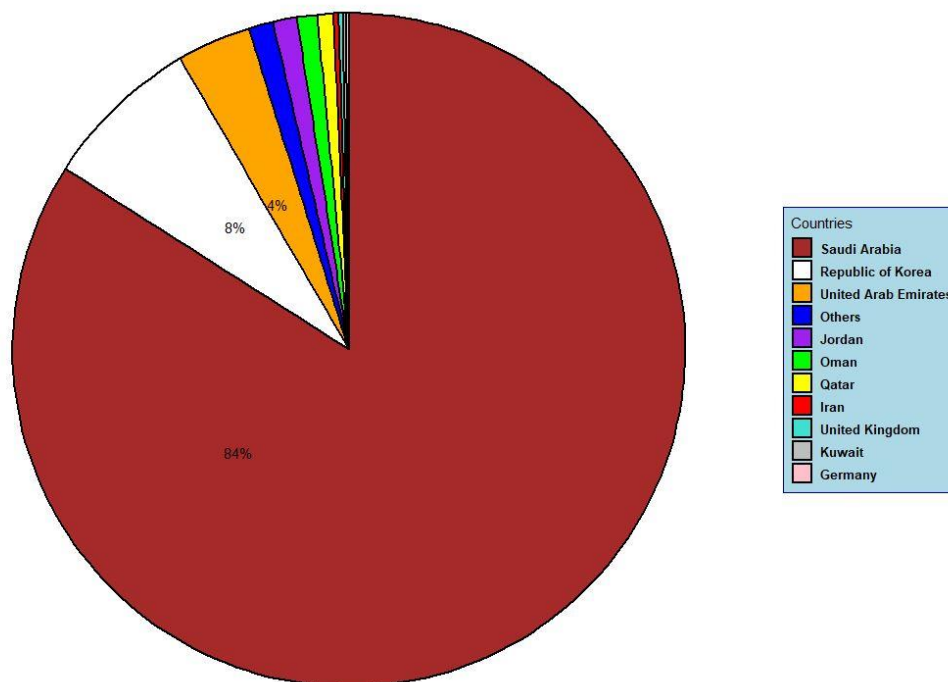
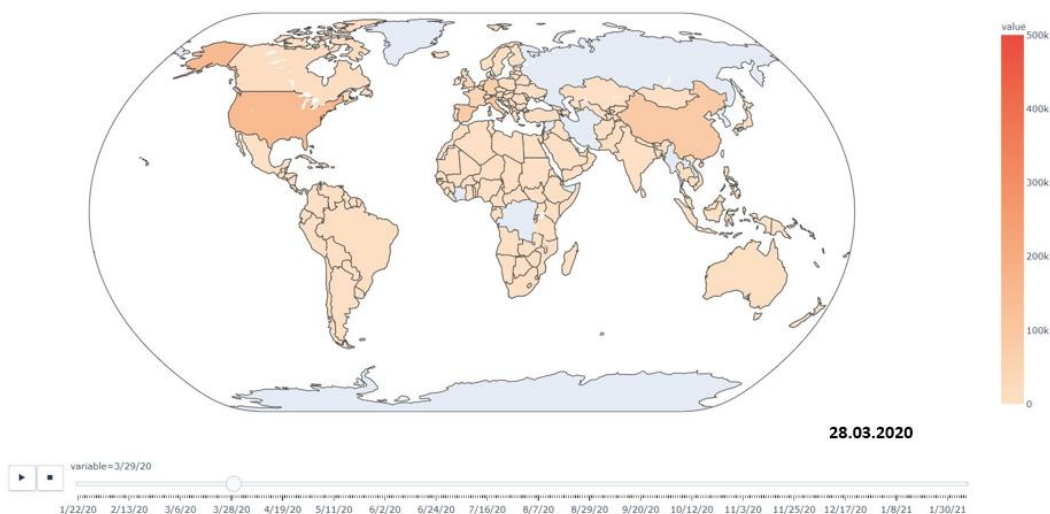


Figure 11. Top 10 Countries According to Total Number of Cases for COVID-19, SARS and MERS infections. Data Source: Worldometers (<https://www.worldometers.info/coronavirus/>) and Kaggle (<https://www.kaggle.com/>). Data collected on 18.1.2021

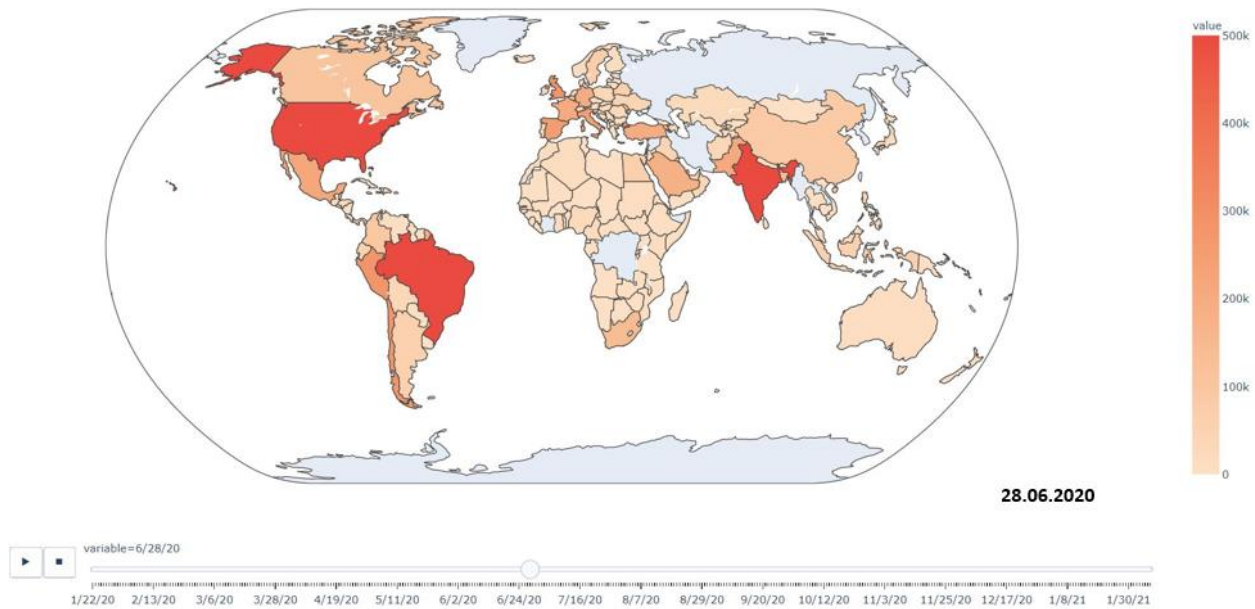
As seen from **Figure 11**, in the cases of SARS and MERS epidemics, viral spreading was prevented and restricted to nearly only one country. If we disregard already mentioned features of these three coronaviruses which influence their spread, the remaining question is how authorities in China and Saudi Arabia managed to hinder and limit virus transmission and why the same measures were not immediately implemented in the case of COVID-19 outbreak? As of February 2021, a WHO team is in China investigating the origin of SARS-CoV-2 and the circumstances leading to its spread. Delays in public release of data about the outbreak of the current global pandemic in China make it hard to estimate if initial measures were appropriate. However, the scientific community assumes that the COVID-19 pandemic could have been prevented, at least to a certain extent. However, future investigations are needed for a complete understanding of COVID-19 outbreak and virus transmission, the factors that led to worldwide spreading, but more importantly, the options for epidemics prevention.

The first COVID-19 case identified outside of China was in Thailand, in January, 2020 (WHO, 2021). Very rapidly, the virus was transmitted to other continents causing the biggest health crisis in 21st century. The time trend of global SARS-CoV-2 spread is presented in an interactive map animation in **Figure 12**.

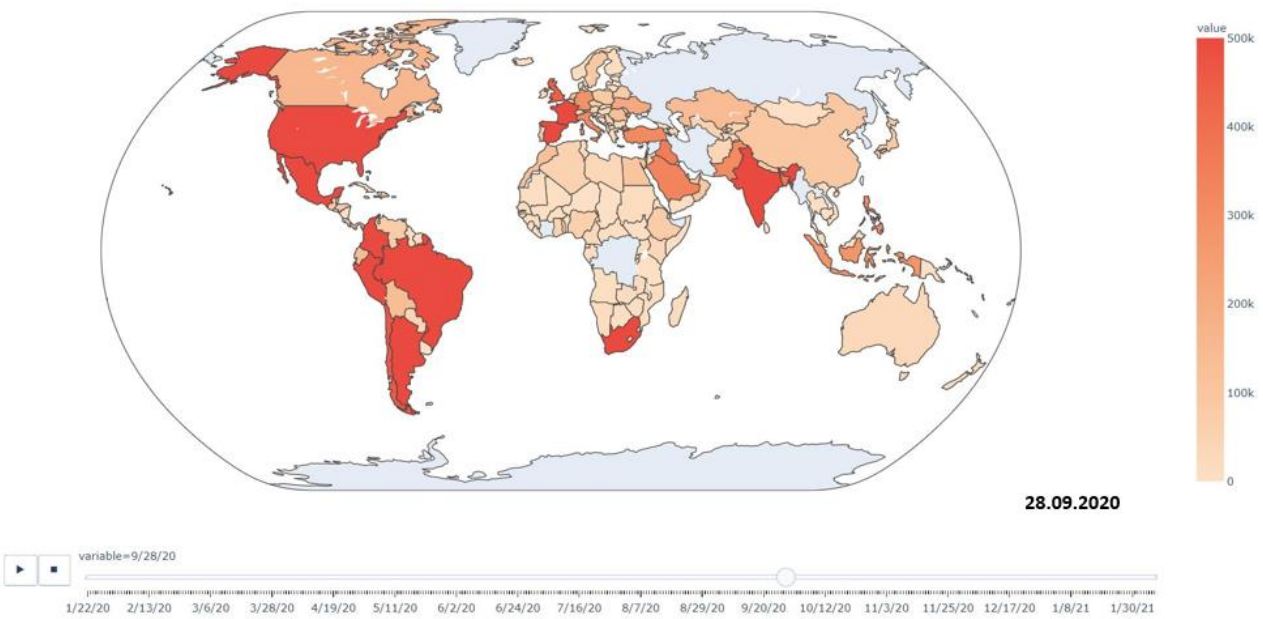
A)



B)



C)



D)

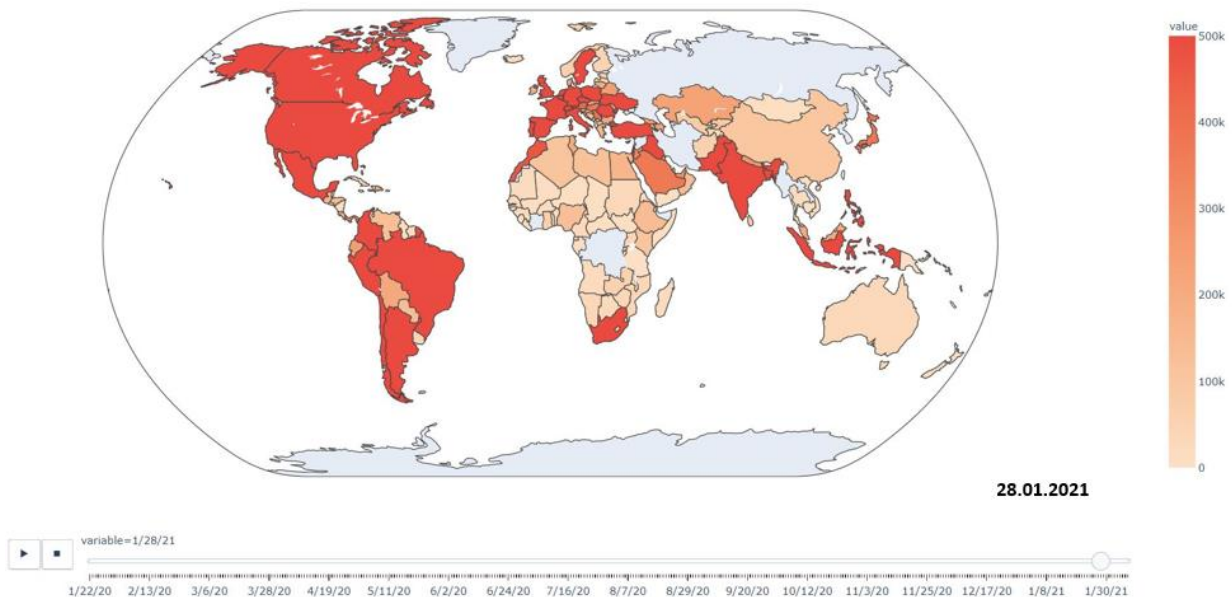


Figure 12. The World Map of COVID-19 spread at four different time points with total numbers of cases: 1) COVID-19 distribution on 28.03.2020; 2) COVID-19 distribution on 28.06.2020; 3) COVID-19 distribution on 28.09.2020; 4) COVID-19 distribution on 28.01.2021. Data source: The Humanitarian Data Exchange (<https://data.humdata.org/dataset/novel-coronavirus-2019-ncov-cases>). Data collected on: 07.02.2021.

In the first image (A) in **Figure 12** we can observe that only China and US had more than 100 000 total confirmed cases as of March 28, 2020. However, we need to consider both China and US have large numbers of inhabitants with low values of total cases per capita at this point of time. In the same period, the number of COVID-19 cases was rapidly increasing in Europe, in the first half of 2020 mainly in Italy. Although first reactions and responses to the pandemic were dramatic, based on image B of **Figure 12** which shows COVID-19 spread on June 28 there were only three countries with total numbers over 500 000 cases. Besides US, Brazil and India became new hotspots. The situation in Brazil became disastrous in this period with total COVID-19 cases of over a million. This visualization also reveals that case numbers in Europe (mainly in UK, France, Spain and Italy) were rapidly increasing. By the end of September 2020 (image C in **Figure 12**) almost all South American states became red, which means that they had more than 500 000 cases. One of explanations could be the winter season from June until August (2020) in the south hemisphere as coronaviruses have better conditions for spreading during the cold days

(Aboubakr, Sharafeldin and Goyal, 2020). Besides this region, the main impacted countries in Europe during this period also exceeded the mentioned cut-off. Image C also reveals that new red colored countries emerged on the world map of COVID-19 cases - South Africa. In comparison to the image A and B, the number of total cases in South Africa was increased almost five times in image C. The pandemic outbreak in this area is even more impactful if we know that new, more contagious and dangerous strain of SARS-CoV-2 originate from this country. Image D of Figure 12 (January 28, 2021) reveals an alarming situation in almost the entirety of Europe, North and South America. Also, we see that new regions, mostly in Asia (Indonesia, Philippines, Pakistan), reached half of a million total cases in January 2021. The overall number of confirmed COVID-19 cases exceeded 100 million by January 2021 which confirms the severeness of this public health crisis one year after its outbreak. Nevertheless, it is essential to tightly watch the epidemiological situation and responses to the global pandemic in several countries that had different strategies to fight COVID-19. The epidemiological situation over time will be presented for the following countries: Austria, Serbia, Italy, Germany, Sweden, United Kingdom (UK) and United States of America (US). Some countries unfortunately do not share their epidemiological COVID-19 data with the public. Thus, it was not possible to compare and analyze all listed countries.

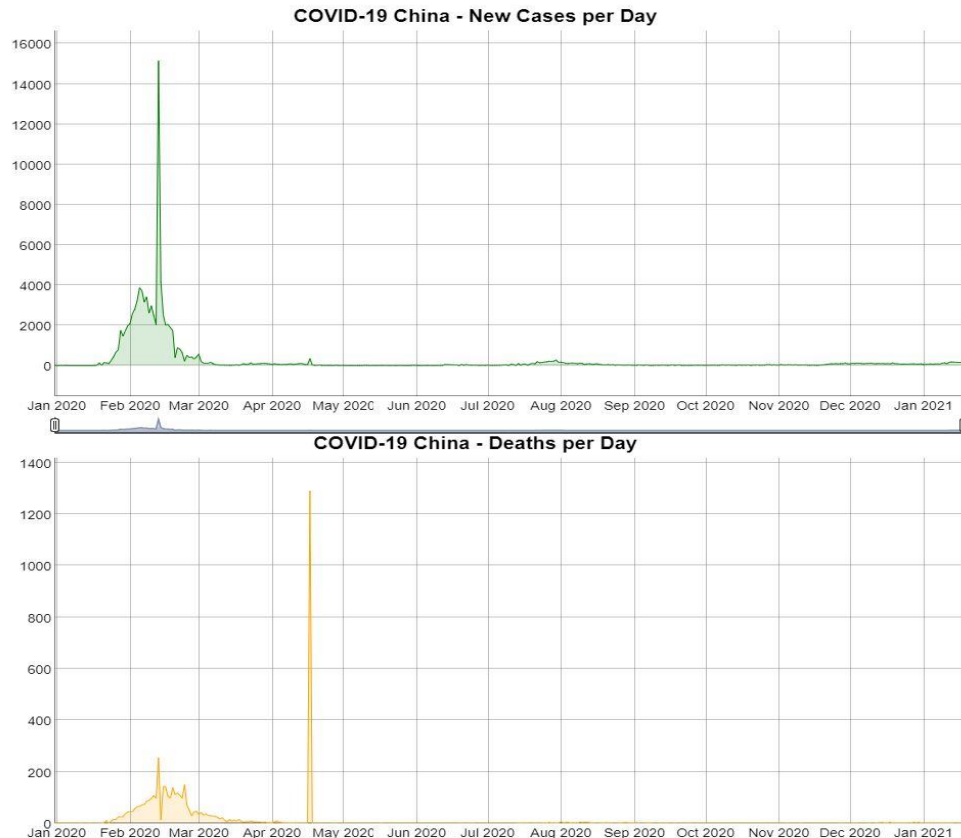


Figure 13. Time trend of New Cases and Deaths per day for China. Data source: Our World in Data. Data collected on: 18.01.2021.

As already mentioned in the introduction, the first cases of SARS-CoV-2 infection in humans showed up in Wuhan, China. Based on **Figure 13**, we can see that besides February and March 2020, when the numbers of new cases were relatively high, China established pretty good control of disease spread, without any other significant rise of numbers after the first quarter of 2020. Despite being the first country in the world hit by this novel pandemic, this country and their citizens responded responsibly to the health crisis which minimized the possibilities for viral transmission. This is a result of the centralized system response and strict measures, but also a consequence of Chinese previous experience with SARS-CoV and its high mortality (Burki, 2020a). Besides the populations' awareness of the coronaviruses' harmfulness, it is worth mentioning that certain social factors in this country could be also influential. For example, only 3% of elderly Chinese people live in care homes (mostly they live with their children or alone), which are facilities that have been identified as a major sources of viral infection and

transmission (Burki, 2020a). Unfortunately, SARS-CoV-2 rapidly spread around the world starting from 2020, primarily in Europe and North America.

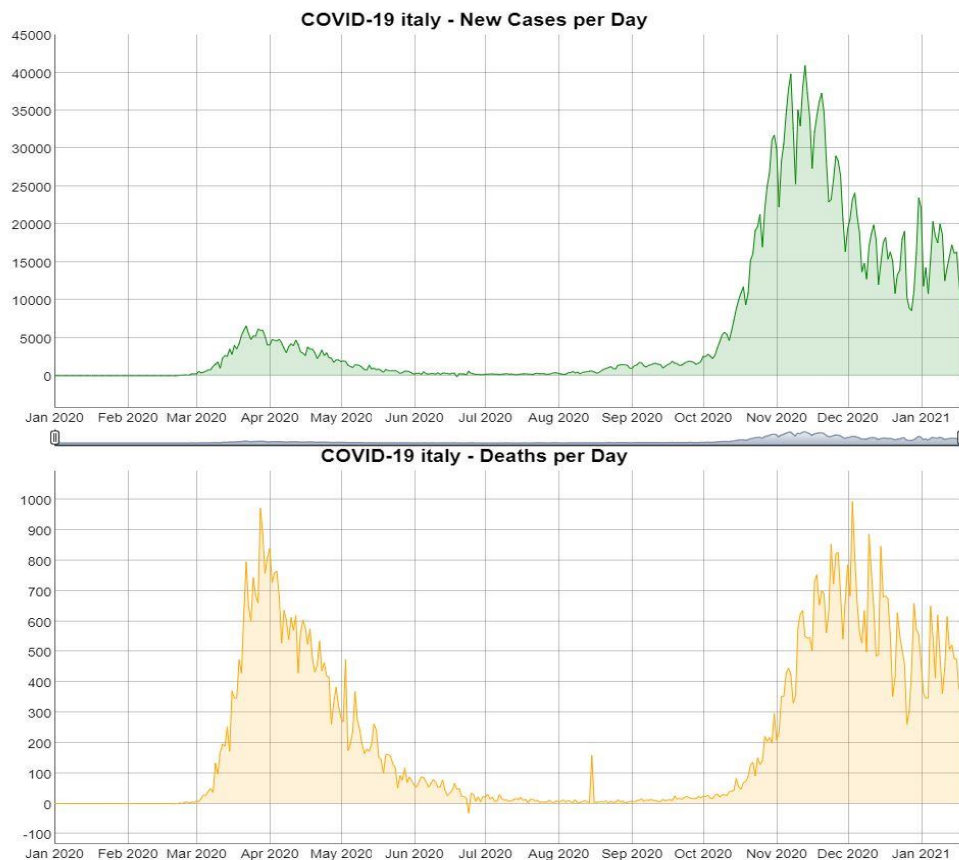
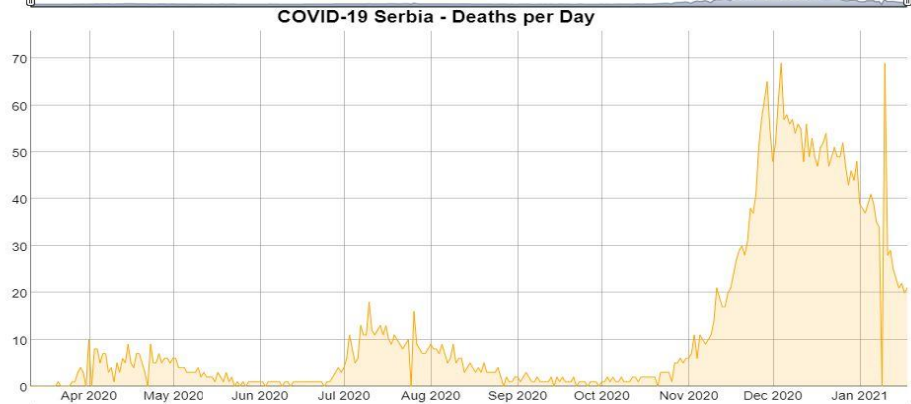
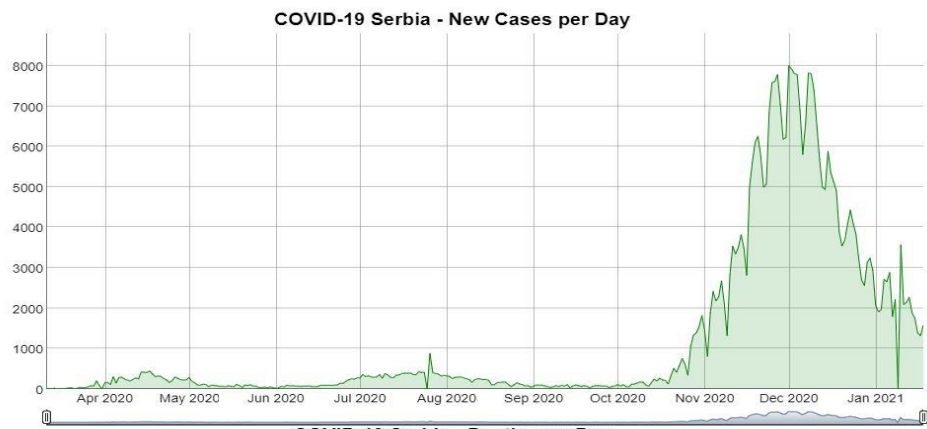
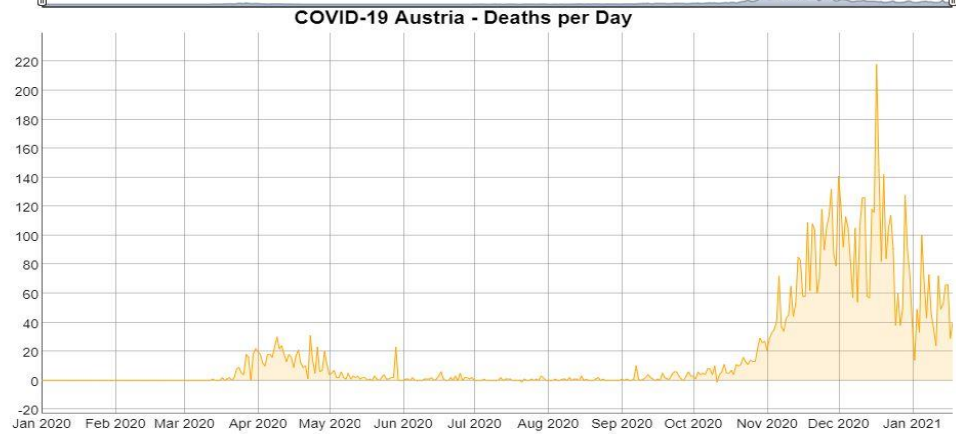
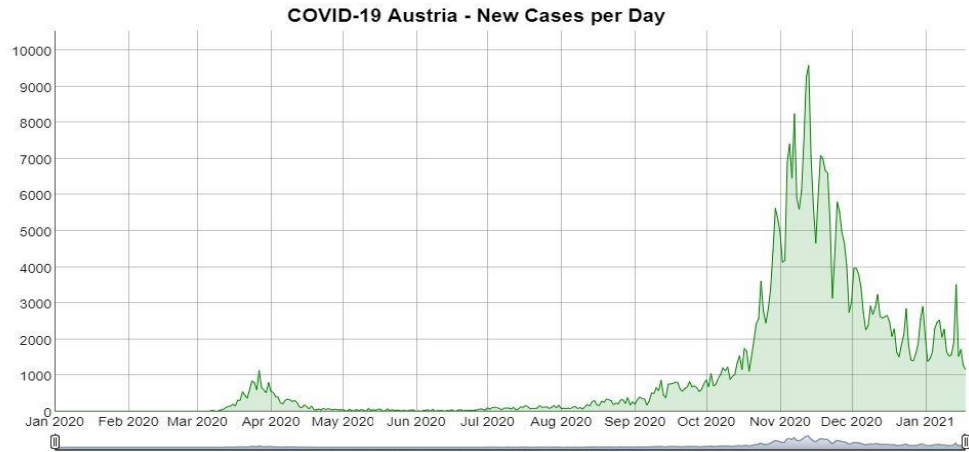


Figure 14. Time trend of New Cases and Deaths per day for Italy. Data source: Our World in Data. Data collected on: 18.01.2021.

Italy was the first country in Europe that was highly affected by the COVID-19 pandemic. The etiology and treatment options against the disease were still unknown and the reactions to the health crisis were late and inappropriate. The number of new cases was followed by increased numbers of deaths as it can be observed in **Figure 14**. However, the second wave that has started in October 2020, has brought significantly higher numbers of new infections, which is observable in **Figure 14**. The amount of daily new cases reached 40 000 during the second wave, while during the first wave, it did not exceed 8000. Similar trends were observed in Austria, Serbia and Germany and that is presented in **Figure 15**.



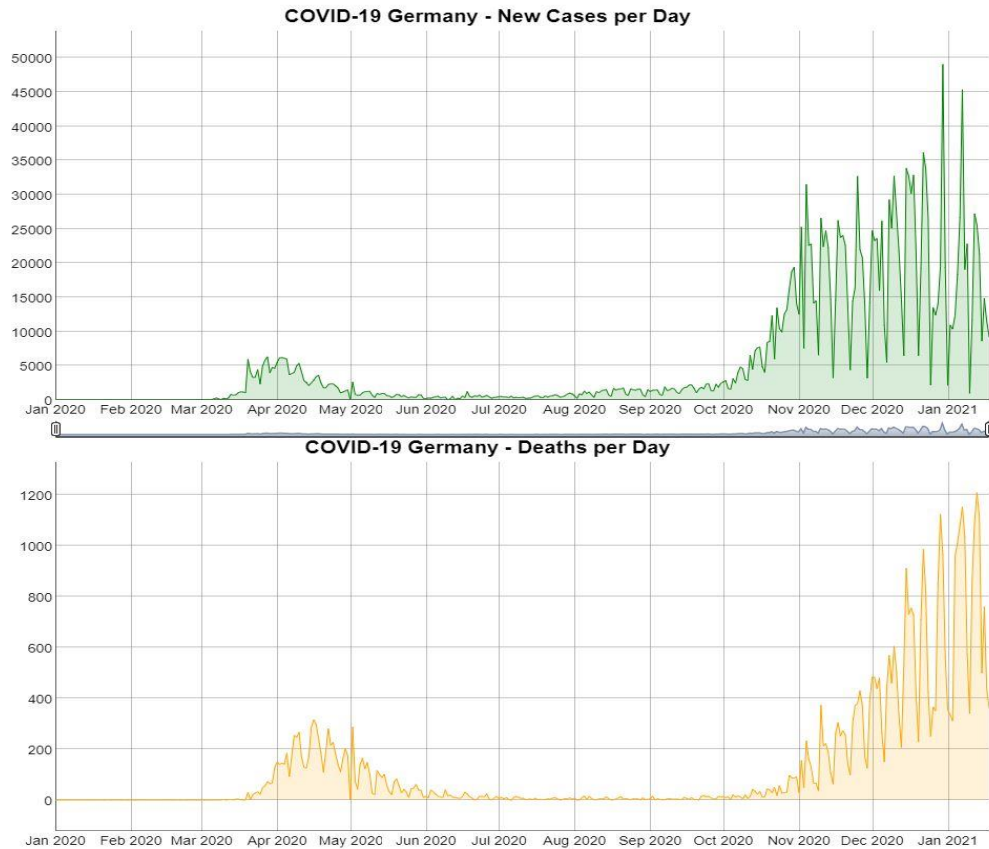
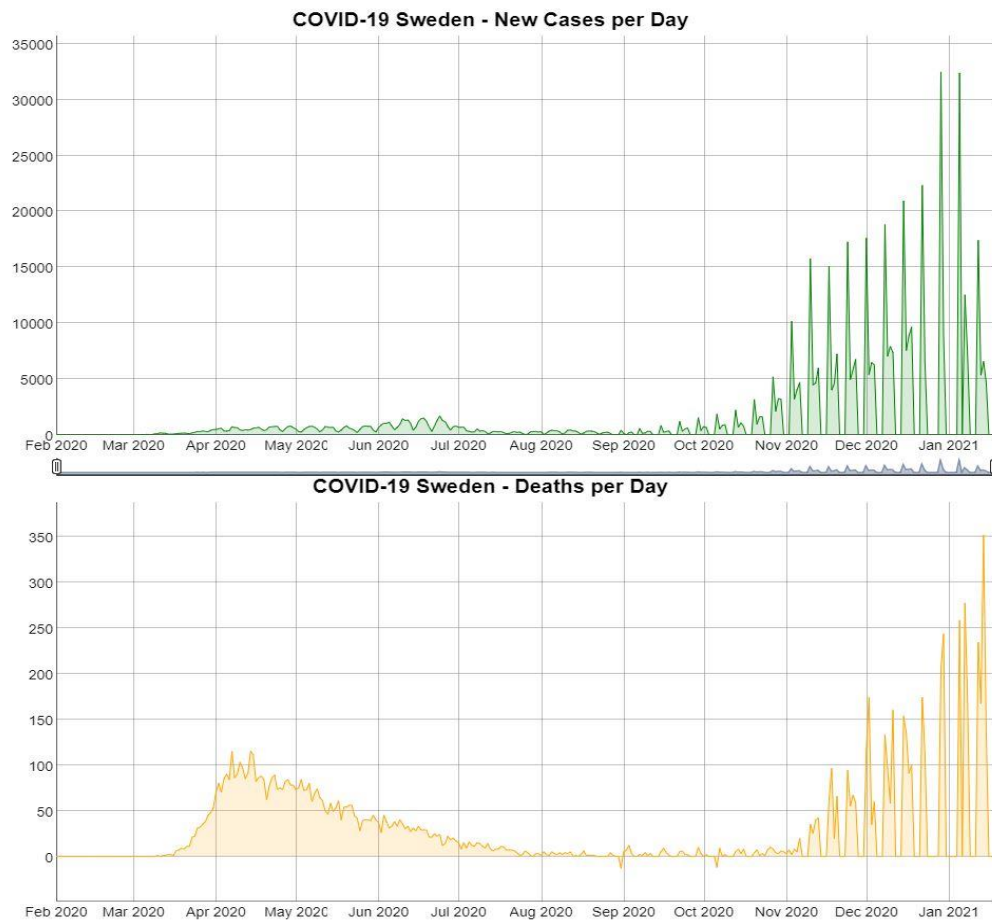


Figure 15. Time trend of New Cases and Deaths per day for Austria, Serbia and Germany. Data source: Our World in Data. Data collected on: 18.01.2021.

All three countries reached their peaks of daily new cases and deaths during the second wave of the pandemic. The third graph in Figure 15 is showing the trend for Germany, reveals fluctuations in numbers during the last four months, probably due to a different approach in COVID-19 data collection. However, the general trend and notable increase in both new cases and deaths of the second versus the first wave are obvious. The highest numbers in Austria and Germany were observed in the middle of November and December, respectively. The situation in Serbia was alike but not identical. Between the first and second wave, this Balkan country had another increase in numbers during the summer which is presented in the second visualization of this Figure. This was probably the result of lifting measures that included regular gastronomy working, organizing sports events with audience and allowing public gathering, both indoor and outdoor. Although viruses have lower tendency of transmission during the summer months, this growth in numbers demonstrate the significance of measures taken and their compliance, even during the warm days. Despite the critics, the second wave that occurred during the fall and

winter 2020/2021 in the entire northern hemisphere was inevitable from my own point of view. It is important to stress that all mentioned countries (Italy, Austria, Serbia and Germany) sooner or later managed to control the pandemic crisis and reduce the numbers and health system loads of the second wave. Countries that had different strategies for fighting COVID-19 and those that did not succeed to limit the viral spread will be discussed in the following sections.



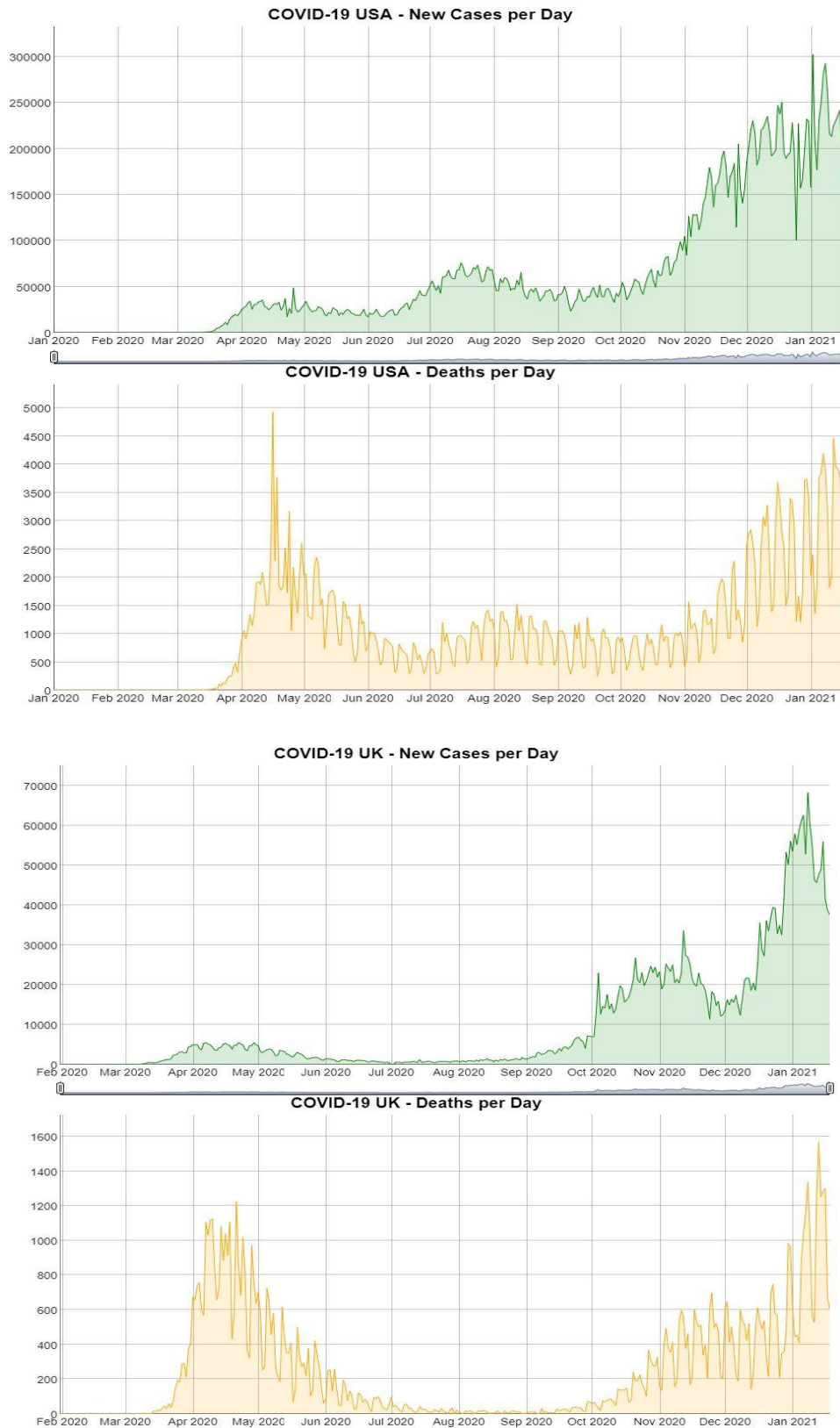
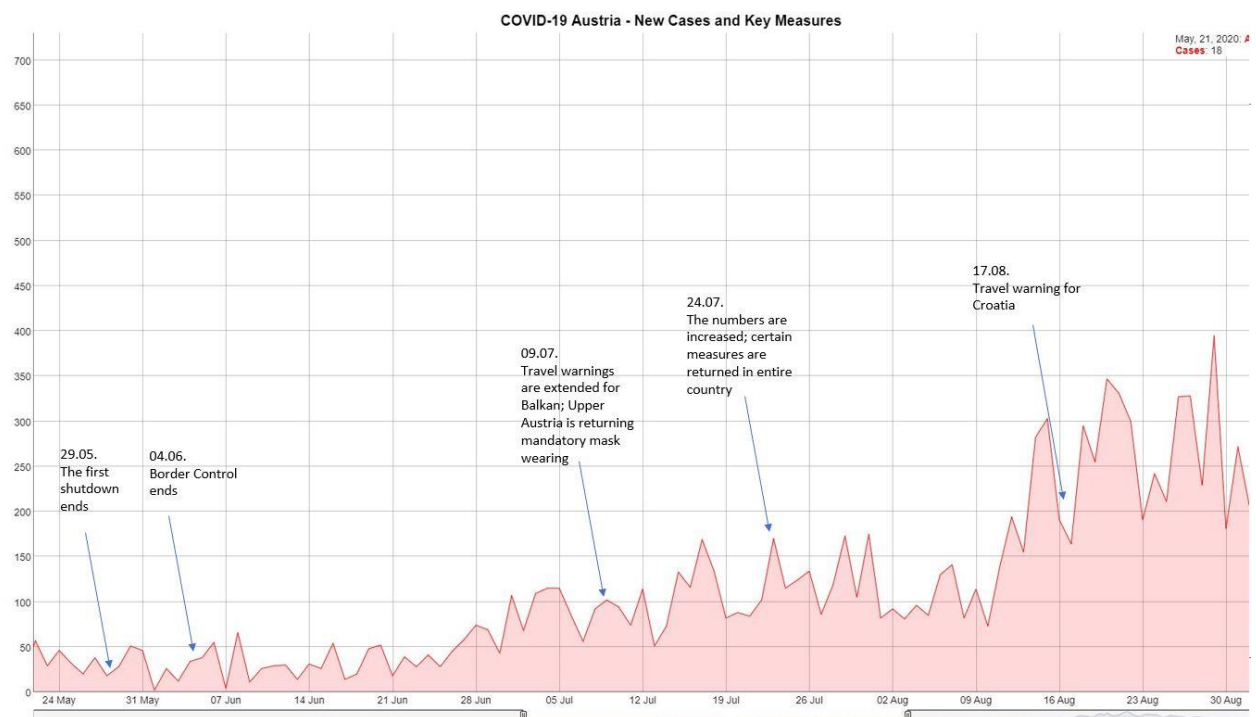
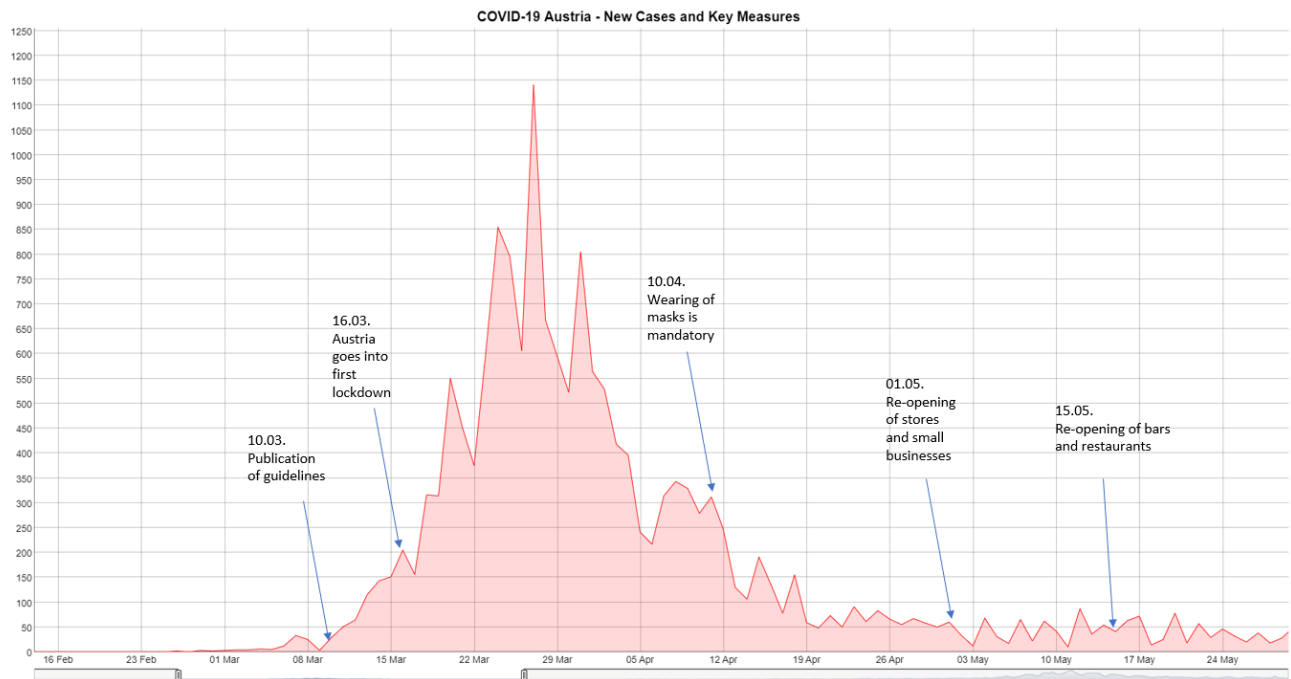


Figure 16. Time trend of New Cases and Deaths per day for Sweden, US and UK. Data source: Our World in Data. Data collected on: 18.01.2021.

In the **Figure 16**, three countries with inadequate response during the first year of the pandemic are presented. The first image of this Figure shows trends of daily numbers of cases and deaths in Sweden. The government of this Scandinavian country tried to implement a different strategy in struggle against the novel coronavirus. Their idea was to sustain normal life, without any strict measures, and thereby obtain herd immunity of the population. However, this resulted in uncontrolled spread of the disease and health system overload. Eventually, the Swedish government confessed wrong anticipation of the potential of COVID-19 and after all they applied similar measures like the majority of countries in the European Union. Besides, it seems that the Swedish data is incomplete as presented on first image of **Figure 16**. Both new cases and deaths curves show large fluctuations. The second and third image of the same Figure reveals time trends for these two parameters in the US and UK. Whereas US encountered the second wave of the COVID-19 pandemic during the summer as an extension of the first one, UK managed to keep the spread of the disease during the sunny season under control and avoided a second wave in the summer months (like most of the European countries). Probably due to same reasons which caused an explosion of case numbers in most countries worldwide by the beginning of fall 2020 (such as, indoor gatherings, opening of schools and universities etc.), the situation in UK got worse in October 2020. In UK, one of the biggest problems is definitely an overloaded health care system as a consequence of uncontrolled and devastating virus spreading. This country had very high numbers of new cases per capita for a long period of time and one of the reasons for this is probably the new strain of SARS-CoV-2 that shows higher potential for spreading and that originates from the UK (B.1.1.7). It should be mentioned that Sweden, US, and UK have highly developed health and educational systems, with enormous investments in science, research and technology, but still, they are having serious difficulties in getting control over the pandemic. This confirms the severeness and complexity of the current global health crisis.

Another aspect of current pandemic, that is frequently a subject of discussion, is the implementation of strategies and measures against COVID-19. Unfortunately, a significant percentage of people, even in the developed countries, do not understand and/or accept the need for the implemented measures to prevent the spread of the virus (such as mandatory mask wearing, reduction of social contacts, especially indoor, mass testing etc.). In order to corroborate and illustrate the significance of these measures, a time trend plot showing the numbers of new cases per day and implemented measures in Austria has been created (**Figure 17**). The time

period of February 2020 until Jan 2021 was divided into three different time stages for creating these plots. The first one shows the period from COVID-19 outbreak until the end of May 2020, the second one presents the situation and trends during the summer 2020, and the last plot shows the interval from September until January 2021.



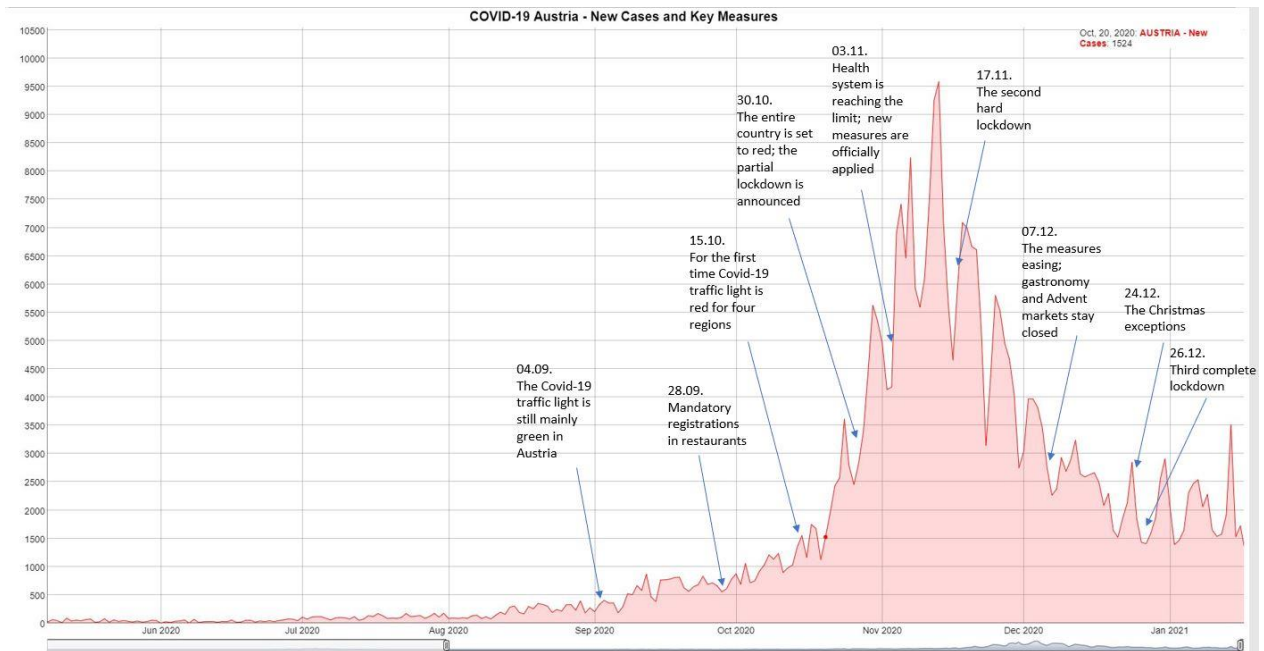


Figure 17. Key Measures implemented and New Cases per Day in Austria. Data source: AGES. Data collected on: 18.01.2021.

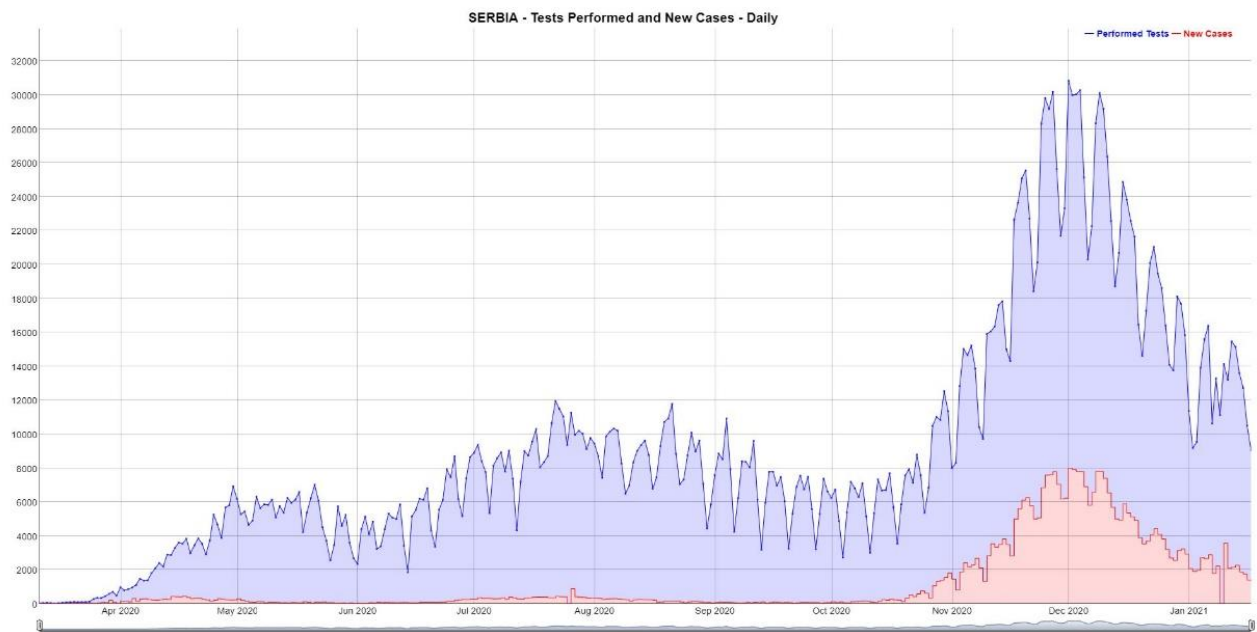
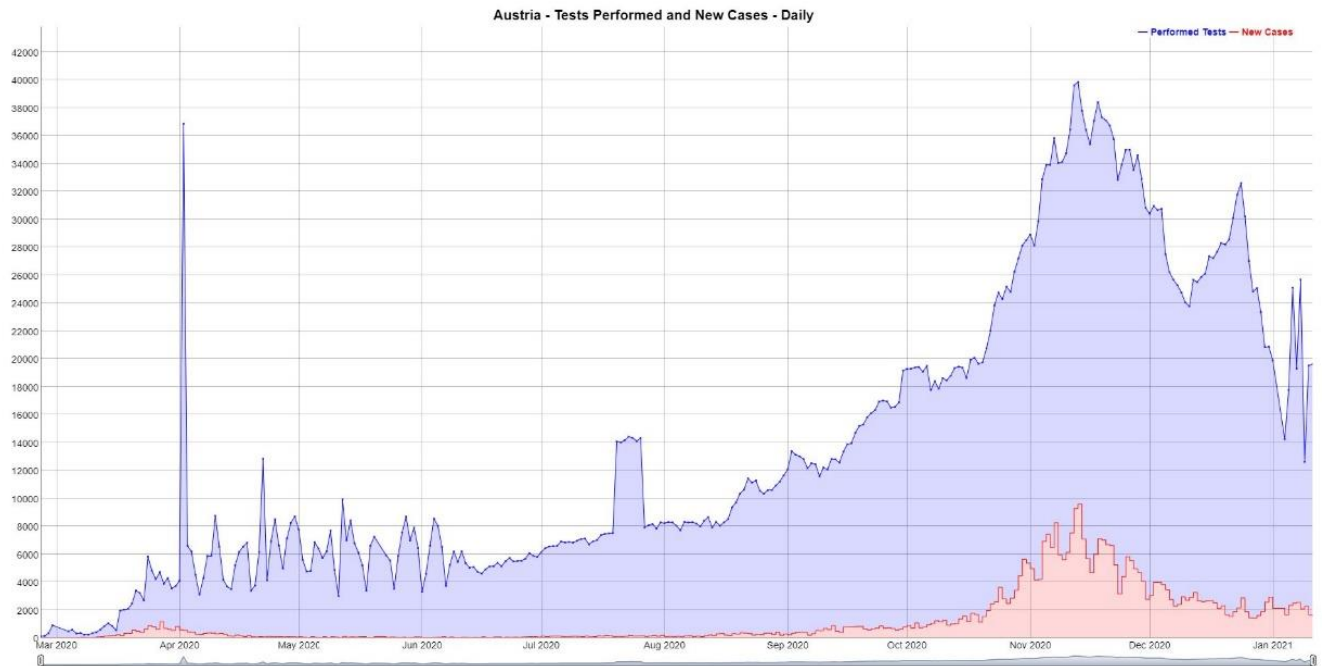
In the first image of Figure 17 shows the first wave of the pandemic in Austria. It can be observed that the guidelines for this health crisis were published six days before the official lockdown that started on 16.03.2020 (closing of schools, universities, restaurants, and all shops that are not relevant for daily supply, working from home etc.). The consequences of this first lockdown in Austria were visible three weeks after their adoption, and can be noted from the downward trend of the curve on the first image of **Figure 17**. At the beginning of May relaxation of some of the measures started, as the viral spread got better under control (observed on the last section of the first image and on the first part of second image in **Figure 17**). As expected, lifting of measures (especially the re-opening of borders) resulted in increased numbers of new infected patients that led to the publication of travel warnings for certain countries and resuming of some of the measures. Nevertheless, the daily numbers of new cases did not exceed 400 during July and August which looked like the pandemic was under control during these months (second image of **Figure 17**).

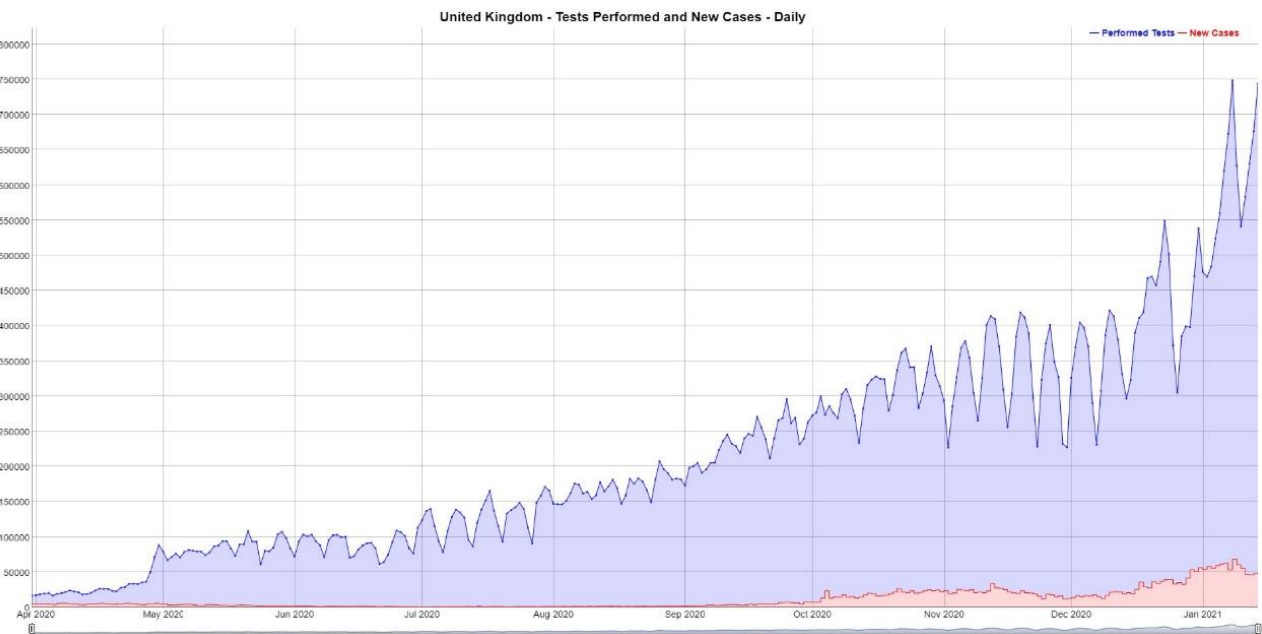
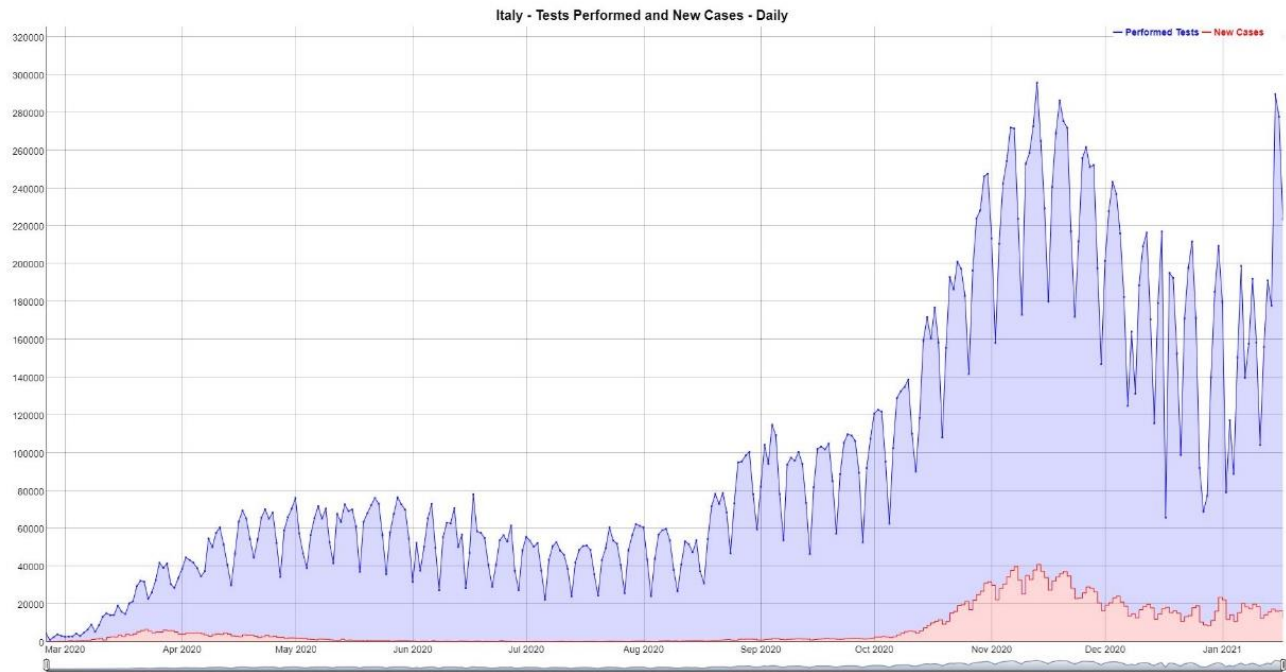
The third image of **Figure 17** shows the number of daily new cases in the period from July 2020 until January 2021 and reveals the severity of COVID-19 disease spread that obviously restarted already during the summer months. In comparison to the two previous images and periods, where maximum abscissae values were 1250 and 700 respectively, the one from the third chart reaches

10 500 new case per day. As it is observed on the third image, the numbers already started to rise in September, and this resulted in certain new measures such as a mandatory registration in restaurants. On October 15th, 2020 for first time since the beginning of pandemic, the traffic light was red in four federal states in Austria. This system was developed by Austrian authorities with the aim of estimating, managing and minimizing the risks for a new wave of COVID-19 spread and enabling its prevention. The four colors indicate levels of risk for a specific area: green (low risk), yellow (medium risk), orange (high risk) and red (acute). By the end of October, entire Austria was set to red and that led to the announcement of partial lockdown. Nonetheless, as the situation was alarming for a long period of time and Austria's health system started to reach its limits, the second hard lockdown needed to start on November 17th 2020. Although, certain measures were eased at the beginning of December and for a few days during Christmas, the second lockdown was followed by the third one (that is actually only an extension of the previous one) and it lasts until today (February 15th).

Even though the second wave of COVID-19 led to high numbers of new cases and deaths in Austria (and in many other countries worldwide as well during the same period of time), it seems that still there is not sufficient acceptance about the importance of the measures regularly taken in the fight against this pandemic. The conformation for this statement is the effect of second and the third lockdown on daily new case numbers, observed in the third image from **Figure 17**, which for some time managed to control disease spread.

Besides the implementation of measures another important approach for fighting the spread is mass testing. This procedure provides identification of symptomatic cases, but more significantly, of asymptomatic cases as well. This enables isolation of positive cases and tracking of their contacts which will result in spreading control and prevention of new infections. Owing to global collaboration of the bioscience community and huge progress of technologies in the (bio-)medical field, several methods for the rapid identification of SARS-CoV-2 have been developed since the outbreak of this pandemic. These statements are confirmed in **Figure 18**, where trends of performed tests and new cases per day are presented. It is worth mentioning that most countries do not share their testing data, so these visualizations were feasible for Austria, Serbia, Italy, UK and US.





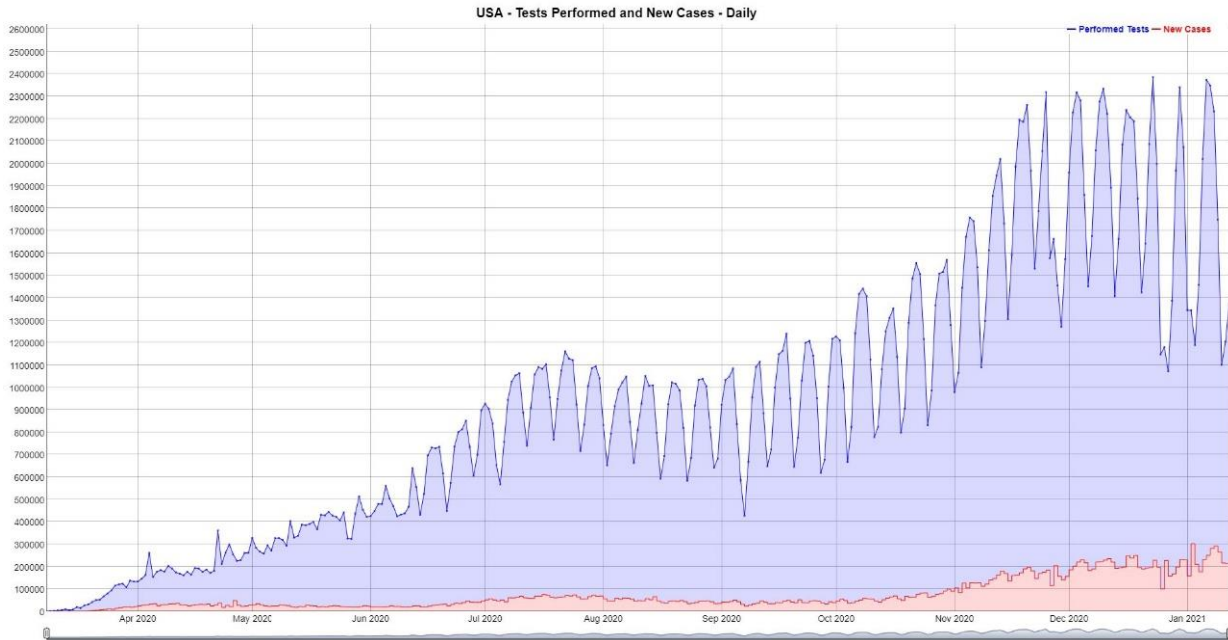


Figure 18. Performed tests and positive cases per day for: Austria, Serbia, Italy, UK and US. Data source: Our World in Data. Data collected on: 17.01.2021.

Based on **Figure 18**, we can observe the growth of daily performed tests for all five countries (blue curve). The oscillations in number of tested patients, particularly in Serbia, Italy and US, arise from different testing performances within one week. Usually, more people are tested during working days than on weekends. Although, we could expect proportionally increased numbers of positive cases with the growth of tested patients per day, the chart from **Figure 18** refutes this assumption.

Thus, it can be noted that the elevation of performed tests that has started in June for most of the studied countries, did not lead to a proportional rise of new cases per day. So, it can be concluded that the appearance of second and third wave in this pandemic occurred independently from growth in the testing performance.

Another important aspect related to testing performance is the fraction of positive cases among the total number of tests over time. **Figure 19** reveals the fluctuating of this pandemic indicator for Austria, Serbia, Italy, US, UK and Sweden. Notably, the data for Sweden seems untrustworthy (although this country shows a general growth of the fraction of positive tested patients) which is shown on the last chart on the grid from **Figure 19**.

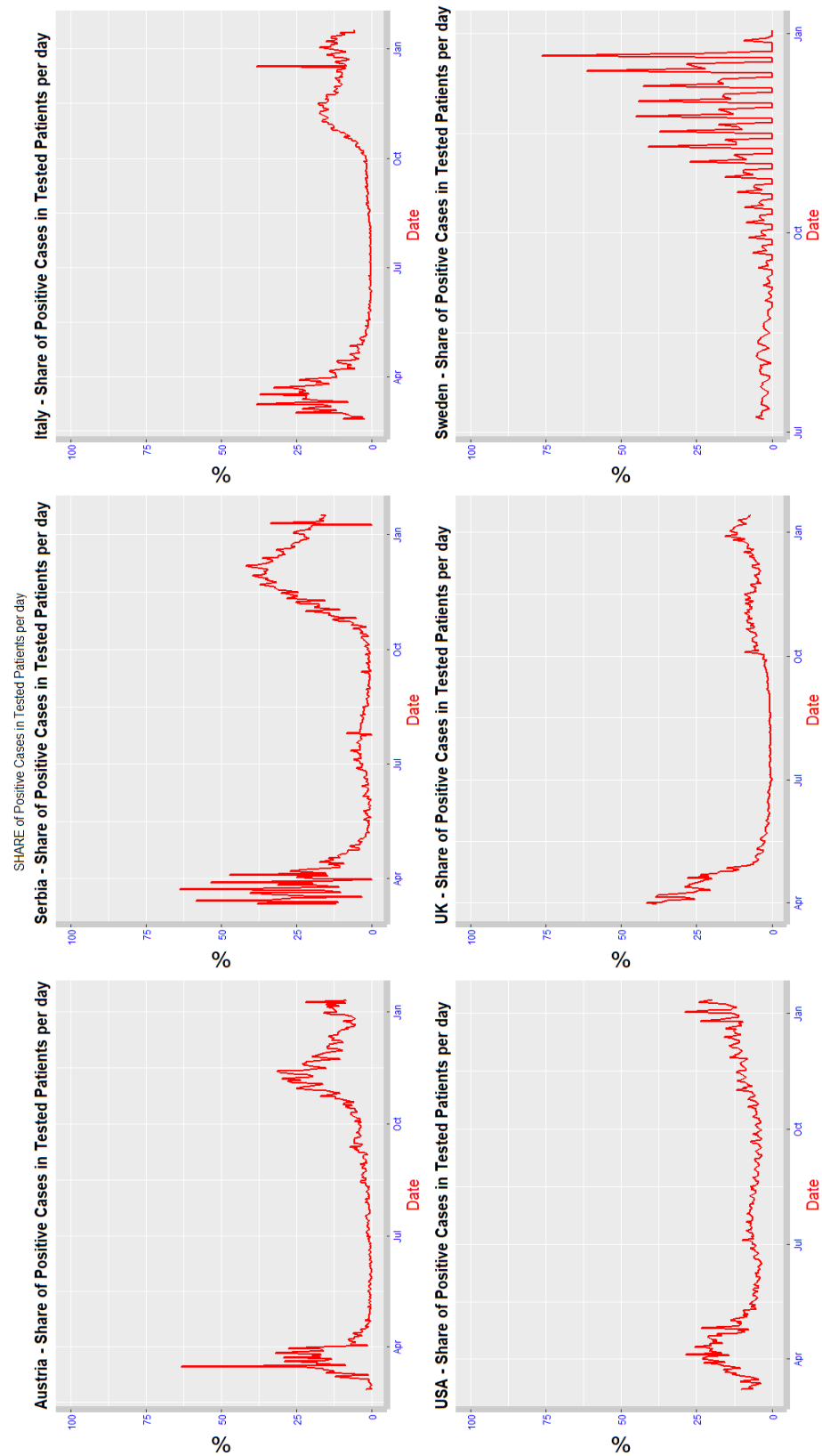
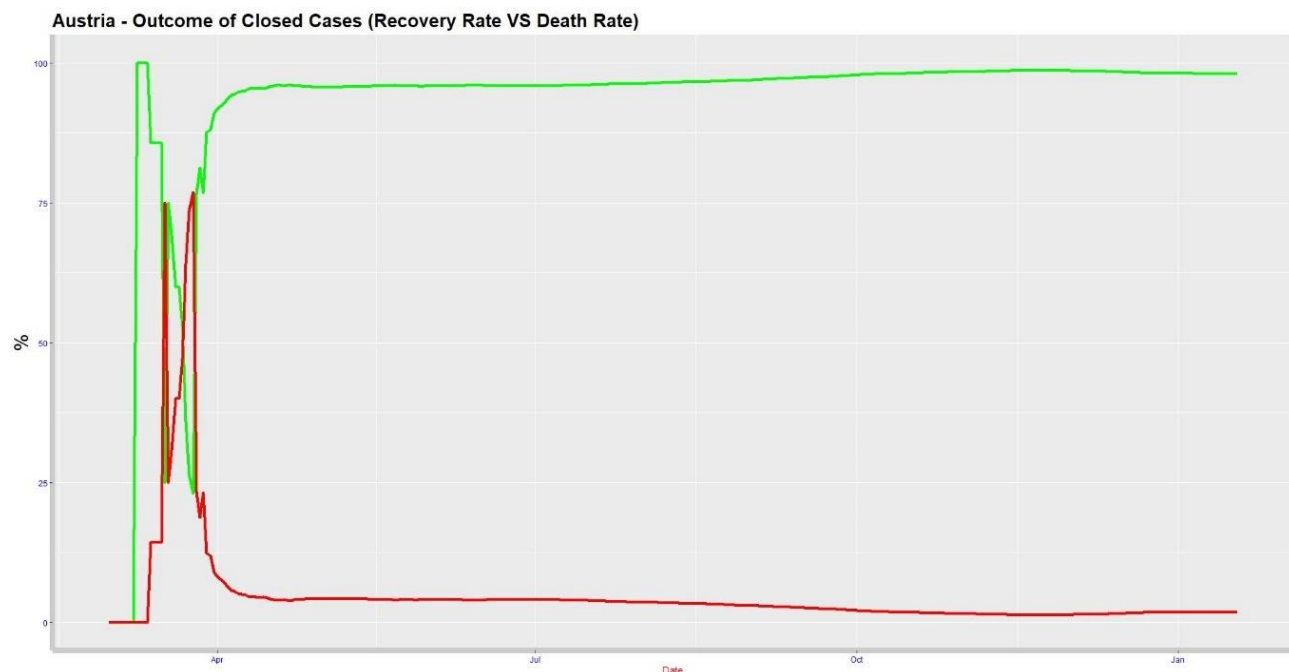


Figure 19. Percentage of Positive Cases in Tested Patients per Day for Austria, Serbia, Italy, USA, UK and Sweden. Data source: Our World in Data. Data collected on: 17.01.2021.

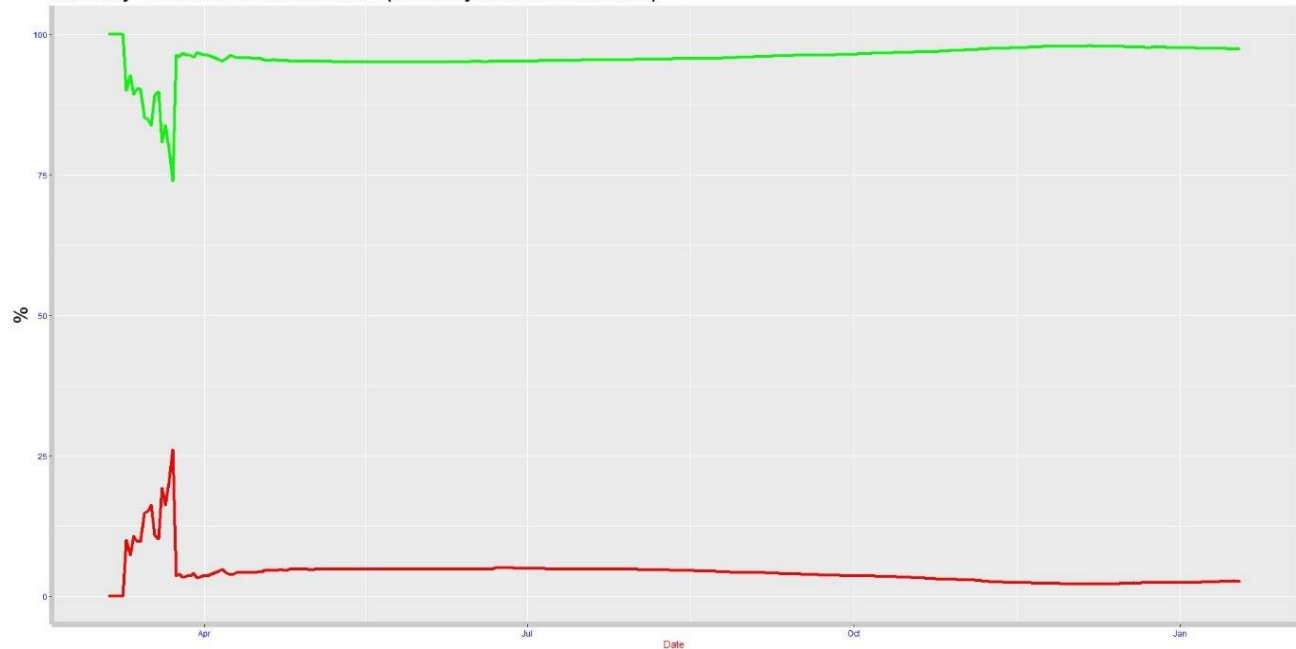
These diagrams confirm the previous statements that the trend of increased testing was not followed by growth in positive cases. However, all countries have similar curve leaning with two big increases that match two waves of this pandemic – during the COVID-19 outbreak and in fall 2020. These trends are especially observed in Austria, Serbia, Italy and USA, and to some extent in UK.

An encouraging fact for all countries from **Figure 19** is the fact that although numbers of tests have significantly increased over time, the percentage of positive cases was approximately equal or lower during the second wave of the pandemic.

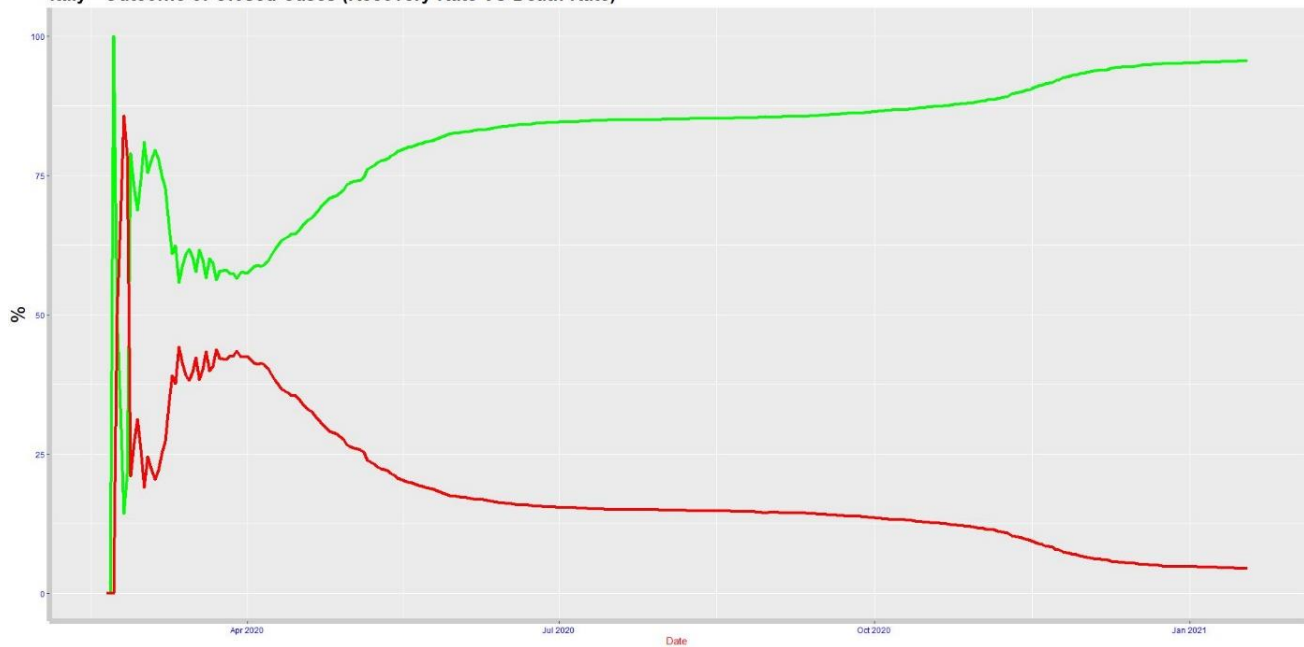
The final figure of the first part of the practical work in this thesis (**Figure 20**) represents the outcome of closed COVID-19 cases over time and thus the success of therapeutical approaches and measures in the period from February 2020-January 2021.



Germany - Outcome of Closed Cases (Recovery Rate VS Death Rate)



Italy - Outcome of Closed Cases (Recovery Rate VS Death Rate)



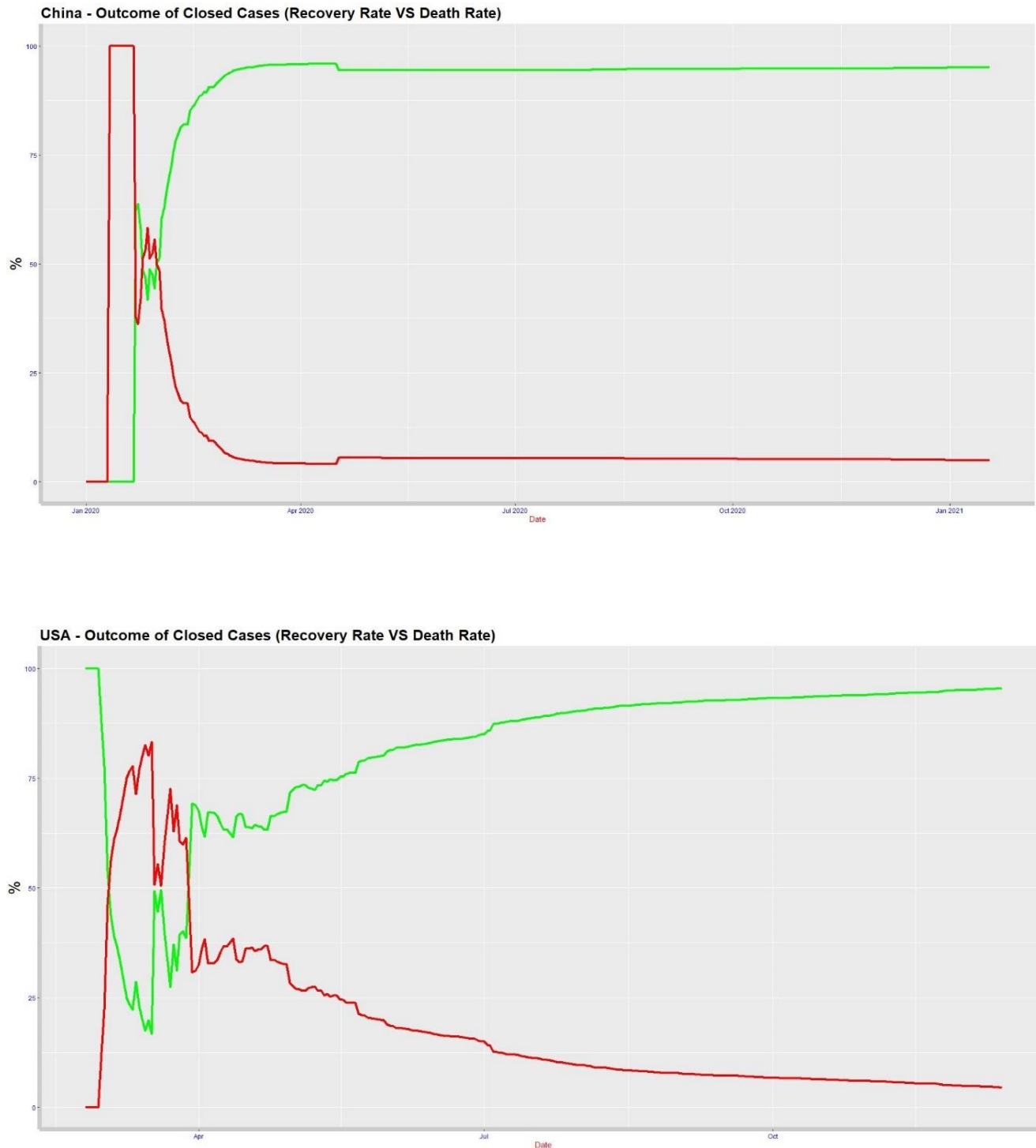


Figure 20. The Outcome of Closed Cases (Recovery Rate VS Death Rate) for Austria, Germany, Italy, China and US. Data source: Our World in Data. Data collected on: 18.01.2021.

Based on the diagrams from **Figure 20**, we can note the stability (and even constant growth for most of the studied countries) of the recovery rate (green curve) in COVID-19 epidemiology.

Although all countries except Germany from this Figure show relative difficulties and low control of disease spread during the first period of the pandemic (particularly Italy, China and USA), which is visible through high values of death rate (red curve), by the end of April 2020, all these states managed to restrain SARS-CoV-2 transmission and could significantly lower the death rate. As demonstrated, in the following months the recovery rate reached high and stable values in all countries. During the second wave of pandemic impact, these trends were slightly disrupted. However, the global rise of recovery rate and its stability, that originate from better understanding of COVID-19 disease factors, improved treatment options and strict implementation of measures, are promising and encouraging indicators in the fight against the current pandemic.

3.2. COVID-19 Protein Targets and Drug-repurposing Possibilities

As previously discussed, an indispensable aspect of SARS-CoV-2 therapy is the comprehension of the viral structure, its lifecycle and mechanisms of interaction with human cells. Therefore, the main idea behind the second part of the practical work of this thesis was to provide a comprehensive overview of SARS-CoV-2 proteins and other related targets (in the first-place human proteins) which might be of interest for drug repurposing strategies.

UniProt provides a comprehensive overview of protein targets that might be interesting in the search for therapeutic strategies against COVID-19 (available at https://covid-19.uniprot.org/uniprotkb?query=*). The list of protein targets provided by UniProt includes human protein targets, protein targets from SARS-CoV-2, and targets from SARS-CoV(-1). The high sequence identity of SARS-CoV-2 and SARS-CoV of almost 80% makes it interesting to also study the structure of SARS-CoV(-1) and include available ligands for these targets in drug repurposing strategies. We wanted to characterize and group the proposed protein targets in terms of sequence-based and structure-based similarities. We therefore first performed sequence-based clustering of the proteins and in the second step analyzed the resulting clusters in terms of their structure-based similarity. This allows to identify interesting groups of proteins that likely can be targeted by similar small molecules and likely interfere at the same point in the viral lifecycle and also points to similarities between SARS-CoV-2 and SARS-CoV(-1) protein structure. In addition, a set of potential ligands for these targets was retrieved, and it was used in the final part

of this work, to extract a set of already registered drugs that are active against aforementioned protein targets.

As a source of COVID-19 proteins and receptors for this work, a data set from UniProt (*UniProt*, 2021) was used. On the day of data collection (17.8.2020), this set contained 53 different protein structures. These COVID-19 related macromolecules were compared based on their sequence similarities using Iroki – Phylogenetic Tree Visualization Software (Moore *et al.*, 2020) (**Figure 21**). The Protein Data Bank (PDB) (Berman *et al.*, 2000) provides protein structures for 37 out of 53 protein structures (marked red in **Figure 21**). In the next step, four clusters from the sequence alignment dendrogram were generated.

Besides sequence similarity of proteins, another important aspect for their comparison is structural similarity. The proteins from four different clusters from **Figure 21** were separately aligned via the Dali Protein Structure Comparison Server. Besides pairwise protein alignment and alignment of structural queries against the PDB, this web tool offers an option for comparison of a list of input protein structures. For those structures that have a published crystal structure, the pdb input files retrieved from Protein Data Bank were used. For the generation of the 16 other structures that did not have PDB identifiers and known structure, RaptorX (Wang *et al.*, 2016) software was used. This tool allows the prediction of protein structures based on their amino acid sequence. However, for the structures with the following UniProt entries, the prediction was not feasible: A0A663DJA2, P0DTD3 and P0DTD8. Besides, certain predicted structures were not recognized by Dali Protein Structure Comparison Server and they could not be compared based on their structure (Q7TFA1 and P59633).

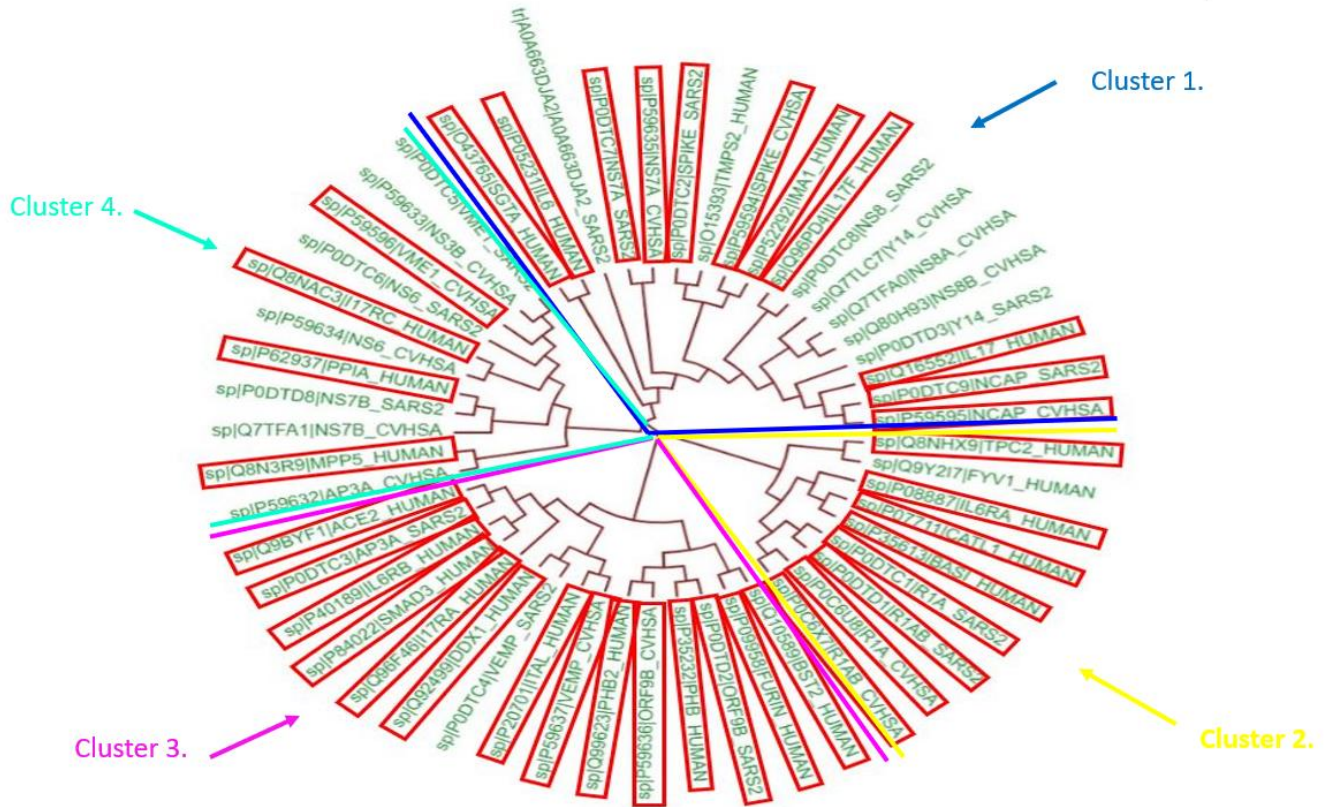


Figure 21. COVID-19 Protein Targets Sequence Alignment Phylogenetic Tree. Generated via Iroki Phylogenetic Tree Visualization Software. Proteins with an experimentally resolved structure recorded in PDB are enframed in red.

Eventually, 48 of 53 COVID-19 protein targets, divided into four clusters, were structurally aligned and analyzed. All proteins, their names, organisms and clusters to which they belong, as well as their UniProt identifiers and Structure Alignment Abbreviations are listed in **Table 3**.

Table 3. COVID-19 protein targets.

Protein Target Name	Organism	Sequence Alignment Abbreviation (UniProt ID)	Cluster Number	Structure Alignment Abbreviation
Small glutamine-rich tetratricopeptide repeat-containing protein alpha	Homo Sapiens	O43765	I	S001
Interleukin-6	Homo Sapiens	P05231	I	S002
ORF7a Protein	SARS-CoV-2	P0DTC7	I	S003
ORF7a protein	SARS-CoV	P59635	I	S004
Spike glycoprotein	SARS-CoV-2	P0DTC2	I	S005
Transmembrane protease serine 2	Homo Sapiens	O15393	I	S006
Spike glycoprotein	SARS-CoV	P59594	I	S007
Importin subunit alpha 1	Homo Sapiens	P52292	I	S008
Interleukin-17F	Homo Sapiens	Q96PD4	I	S009
ORF8 Protein	SARS-Cov-2	P0DTC8	I	S010
Uncharacterized protein 14	SARS-CoV	Q7TLC7	I	S011
ORF8a protein	SARS-CoV	Q7TFA0	I	S012
ORF8b Protein	SARS-CoV-2	Q80H93	I	S013
Interleukin-17A	Homo Sapiens	Q16552	I	S014
Nucleoprotein	SARS-CoV-2	P0DTC9	I	S015
Nucleoprotein	SARS-CoV	P59595	I	S016
Two pore calcium channel protein 2	Homo Sapiens	Q8NHX9	II	S001

1-phosphatidylinositol 3-phosphate 5-kinase	Homo Sapiens	Q9Y2I7	II	S002
Interleukin-6 receptor subunit alpha	Homo Sapiens	P08887	II	S003
Procathepsin L	Homo Sapiens	P07711	II	S004
Basigin	Homo Sapiens	P35613	II	S005
Replicase polyprotein 1a	SARS-CoV-2	P0DTC1	II	S006
Replicase polyprotein 1ab	SARS-CoV-2	P0DTD1	II	S007
Replicase polyprotein 1a	SARS-CoV	P0C6U8	II	S008
Replicase polyprotein 1ab	SARS-CoV	P0C6X7	II	S009
Bone marrow stromal antigen 2	Homo Sapiens	Q10589	III	S001
Furin	Homo Sapiens	P09958	III	S002
ORF9b protien	SARS-CoV-2	P0DTD2	III	S003
Prohibitin	Homo Sapiens	P35232	III	S004
ORF9b protein	SARS-CoV	P59636	III	S005
Prohibitin-2	Homo Sapiens	Q99623	III	S006
Envelope small membrane protein	SARS-CoV	P59637	III	S007
Integrin alpha L	Homo Sapiens	P20701	III	S008
Envelope small membrane protein	SARS-CoV-2	P0DTC4	III	S009
ATP-dependent RNA helicase DDX1	Homo Sapiens	Q92499	III	S010
Interleukin-17 receptor A	Homo Sapiens	Q96F46	III	S011

Mothers against decapentaplegic homolog 3	Homo Sapiens	P84022	III	S012
Interleukin-6 receptor subunit beta	Homo Sapiens	P40189	III	S013
ORF3a protein	SARS-CoV-2	P0DTC3	III	S014
Angiotensin-converting enzyme 2	Homo Sapiens	Q9BYF1	III	S015
ORF3a protein	SARS-CoV	P59632	IV	S001
MAGUK p55 subfamily member 5	Homo Sapiens	Q8N3R9	IV	S002
Peptidyl-prolyl cis-trans isomerase A	Homo Sapiens	P62937	IV	S003
ORF6 protein	SARS-CoV	P59634	IV	S004
Interleukin-17 receptor C	Homo Sapiens	Q8NAC3	IV	S005
ORF6 protein	SARS-CoV-2	P0DTC6	IV	S006
Membrane protein	SARS-CoV	P59596	IV	S007
Membrane protein	SARS-CoV-2	P0DTC5	IV	S008

The proteins from already mentioned four clusters were in the next step grouped in new, smaller clusters (colored red, blue, yellow and green on the following figures). Some of the new generated groups of potential targets will be analyzed in more detail based on available data related to these proteins and their significance for viral pathogenesis.

The first cluster from sequence-based alignment includes 16 structures (without the two aforementioned proteins that do not have an experimentally resolved structure). It is important to mention that the analyzed proteins have different numbers of protein chains. Therefore, in the following Figures, every single chain of a specific macromolecule was individually compared with others.

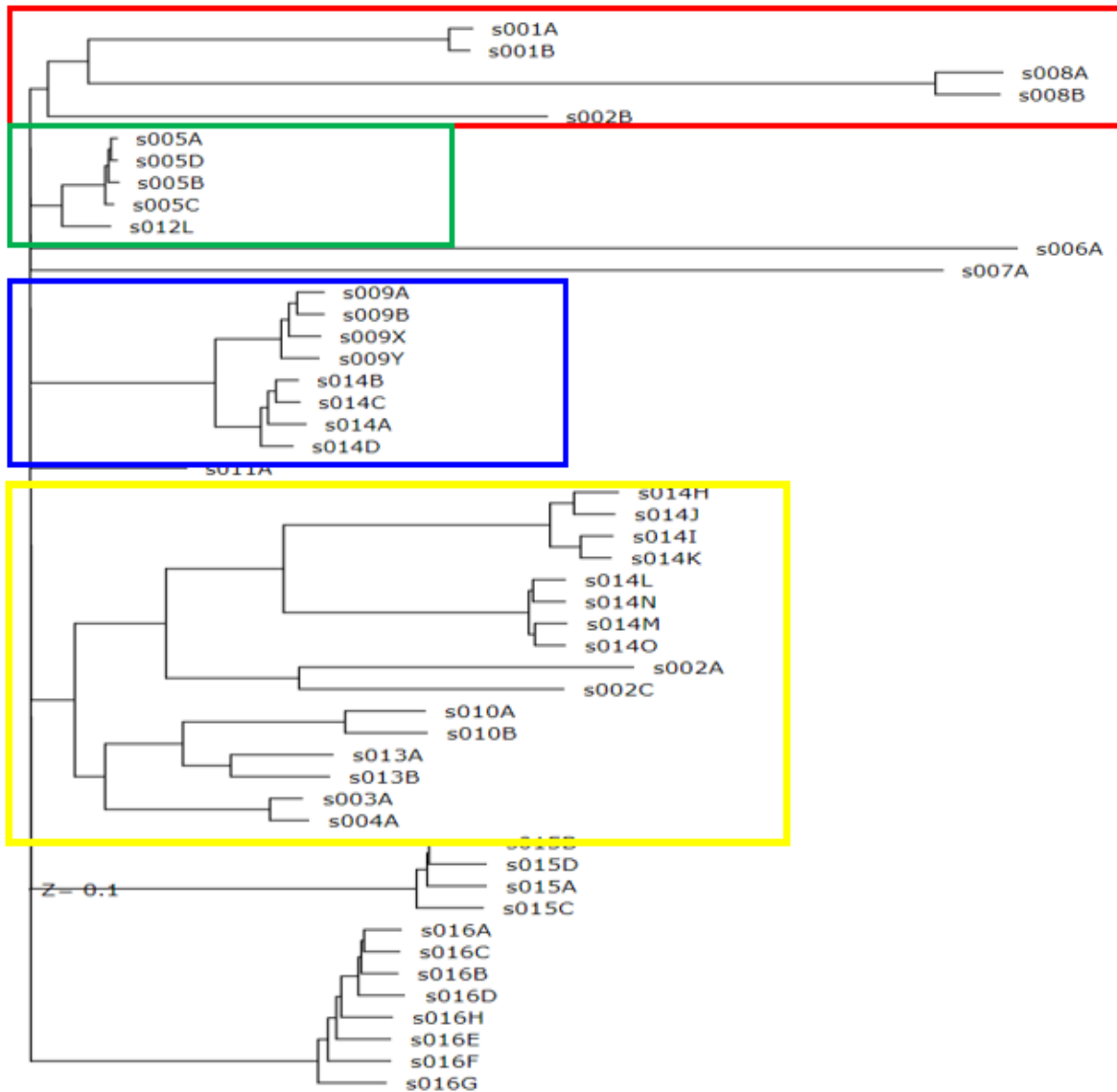


Figure 22. Structural similarity dendrogram for COVID-19 protein targets from the first cluster of the sequence-based dendrogram. The dendrogram is derived by average linkage clustering of the structural similarity matrix (Dali Z-scores).

Based on **Figure 22**, six groups of structurally similar proteins have been observed and four of them have been analyzed in more detail (colored red, green, blue and yellow). Both chains of s001 (small glutamine-rich tetratricopeptide repeat-containing protein alpha - SGTA) show high similarity with both chains of s008 (importin subunit alpha 1) and with b chain of s002 (interleukin-6). The first two proteins from this cluster, SGTA and Importin subunit alpha 1 (that both have sequence and structure similarity which was confirmed in **Figure 21** and **Figure 22**), share the same feature of interacting with BAG6 (ATP-independent chaperone, part of

BAG6/BAT3 complex, that prevents the aggregation of misfolded and hydrophobic patches-containing proteins (Leznicki *et al.*, 2010). Although this complex has a role in a variety of physiological and pathological processes (antigen presentation, T-cell response, apoptosis etc.), recent studies have shown that BAG6 could also have an important function in the control of aggregation-prone polypeptide biogenesis (Kawahara, Minami and Yokota, 2013). However, there is no data to indicate any connection between this protein cascade and human-viral interaction. On the other hand, the role of IL-6 in the development of hyper-reaction and lung damage in patients with SARS-CoV2 has been confirmed (Chen and Quach, 2021) and some therapeutic approaches that involve the blockage of this cytokine have already been applied (Cron, 2021). The following analysis of binding sites will provide more information about possible cross-interaction between these protein targets and their ligands.

The second group of two similar structures from **Figure 22** (colored green) reveals the connection between s005 (spike glycoprotein of SARS-CoV-2) and s012 (ORF8a protein of SARS-CoV). The structure and characteristics of SARS-CoV-2 spike glycoprotein have already been discussed. Besides the enhancement and initiation of viral penetration and replication, no mutual feature or interaction between these two macromolecules (s005 and s012) has been found so far.

The third cluster from **Figure 22** includes two cytokines from the same family: interleukin-17F (s009) and interleukin-17A (s014). These pro-inflammatory mediators play important roles in host defense and inflammation (Pacha, Sallman and Evans, 2020). They are involved in the production of other cytokines such as IL-6 and IL-8, they stimulate proliferation of peripheral blood mononuclear cells and T-cells, but also they play a role in the induction of neutrophilia in the lungs and in the exacerbation of antigen-induced pulmonary allergic inflammation (*IL17F - Interleukin- Uniprot.* 2021). Interleukin-17 has an obvious potential in COVID-19 treatment as this cytokine has significant contribution in pro-inflammatory response, viral persistence and lung destruction.

The fourth cluster of targets represented in **Figure 22** involves 6 different structures (marked yellow). Besides several chains of already mentioned interleukin-17A and interleukin-6, this cluster shows similarity between four additional proteins: s010 (ORF8 protein in SARS-CoV-2), s013 (ORF8b protein subunit in SARS-CoV), s003 (ORF7a protein in SARS-CoV-2) and s004 (ORF7a protein in SARS-CoV). As previously discussed, the significance of both interleukins (IL-17 and IL-6) in the development of hyper-inflammation and progression toward severe forms

of diseases has already been confirmed. The sequence-based and structure-based similarity of these two cytokines could be extremely beneficial from the COVID-19 therapy point of view. The inhibitors of IL-6 have already been effectively used in the therapy of certain diseases like rheumatoid arthritis (Hennigan and Kavanaugh, 2008). As already stated, the evidence shows that the so-called “cytokine storm” could play a pivotal role in the pathology of lung damage in patients infected by SARS-CoV-2 and its prevention could be a turning point in disease treatment. The second cytokine that also shows a potential to be used for the treatment of conditions caused by SARS-CoV-2 infection is interleukin-17. The murine viral models revealed that the increased levels of IL-6 stimulate generation of TH17 cells which results in synergistical promotion of viral persistence due to protection of virus-infected cells from apoptosis (Megna, Napolitano and Fabbrocini, 2020). Infections caused by the influenza virus, the mortalities owing to acute lung injury is linked to neutrophilia in the airspace promoted by IL-17 (Megna, Napolitano and Fabbrocini, 2020). The application of antagonists of IL-17 receptor led to reduced neutrophils in mice airspace, decreased mortality and lower immune-related lung injury. Also, a review from 2019 revealed that this cytokine is crucial at different stages of viral infections and its targeting showed big potential for treatment of conditions caused by these infections (by minimizing viral infections and suppressing tissue damage) (Ma *et al.*, 2019). Investigations show increased levels of IL-17 in hospitalized COVID-19 patients and its inhibition as a promising solution in a therapy (Megna, Napolitano and Fabbrocini, 2020).

Another four targets from this cluster belong to ORF7 and ORF8 proteins. Their function and diffusion are still not completely understood, however, based on their interaction with SARS-CoV-2 and its structural and non-structural compartments, modulation of leukocytes and disruption of antiviral effect, they might have the potential to be used as a targets in COVID-19 therapy.

The second cluster represented in **Figure 23** has three groups of related COVID-19 protein targets. The first one involves s001 (two pore calcium channel protein - TPCN2), s003 (interleukin-6 receptor subunit alpha) and s002 (1-phosphatidylinositol 3-phosphate 5-kinase). The potential use of interleukin-6 (IL-6) and its receptor in a therapy against COVID-19 has already been explained. However, in this case, the structural similarity between TPCN2 and interleukin-6 receptor (IL-6R) could provide new therapeutic approaches that might involve simultaneous inhibition of these two targets. The inhibition of IL-6R would block

proinflammatory response of activated IL6 while TPCN2's role in SARS-CoV-2 viral replication needs to be confirmed yet, although some studies showed that inhibition of these channels could be beneficial for the treatment of COVID-19 disease (Filippini *et al.*, 2020).

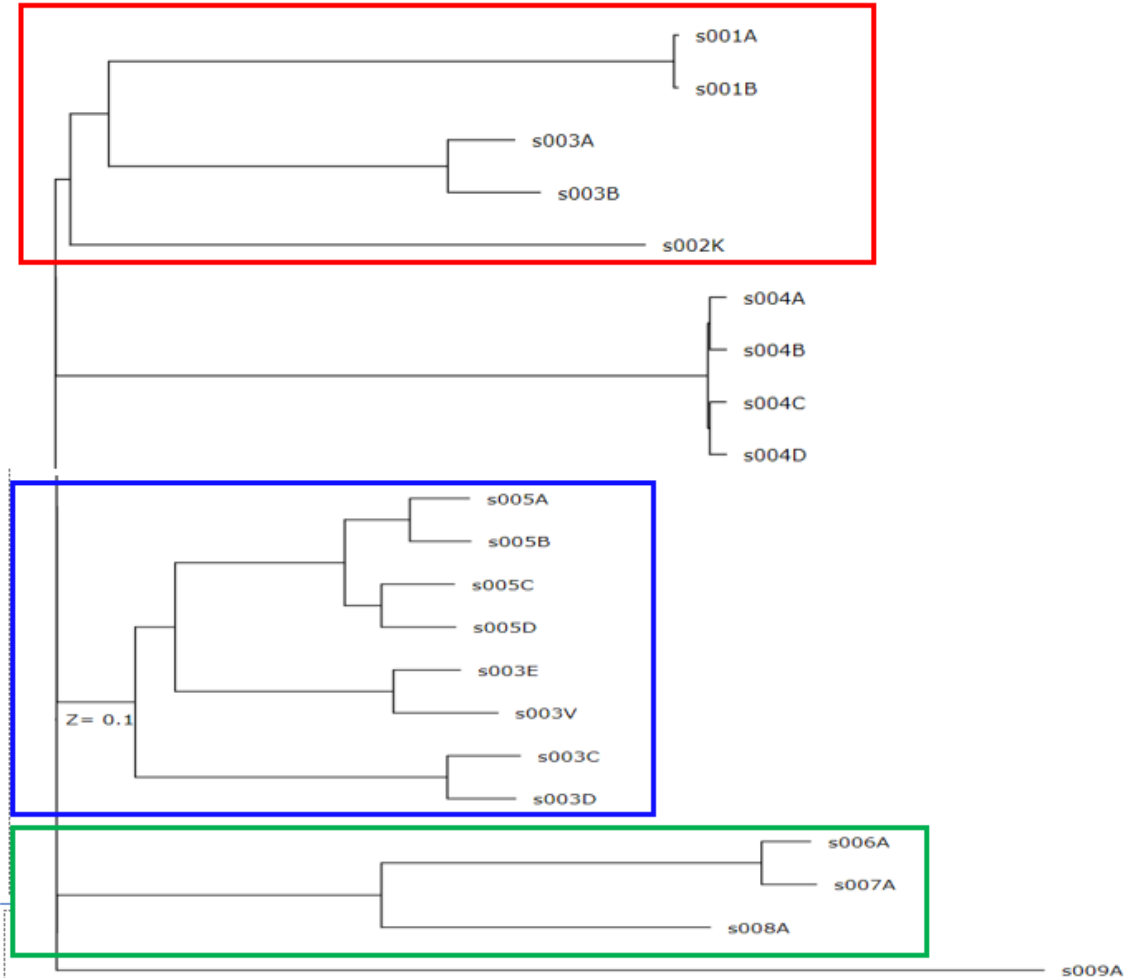


Figure 23. Structural similarity dendrogram for COVID-19 protein targets from the second cluster of the sequence-based dendrogram. The dendrogram is derived by average linkage clustering of the structural similarity matrix (Dali Z-scores).

As seen from **Figure 23**, the second group of similar structures (marked blue) includes four chains of s005 (basigin) and four chains of s003 (interleukin-6 receptor subunit alpha). Basigin is a human glycoprotein located in cell plasma membrane and melanosomes (*UniProt - BSG gene & protein*). It is involved in processes such as spermatogenesis, tumor progression and embryo implantation, but more important, isoform 1 of macromolecule may induce the production of pro-inflammatory cytokines such as IL6 in monocytes (Redzic *et al.*, 2011). Although certain studies revealed possible interaction between transmembrane glycoprotein basigin and spike protein of

SARS-CoV-2, recent investigations discarded these claims (Shilts *et al.*, 2021). While the potential use of IL-6R in a COVID-19 therapy has already been discussed, understanding of the role of basigins, interactions and treatment possibilities require further investigations.

The third group of structurally similar proteins, colored green in **Figure 23** includes three functionally related polyprotein enzymes: replicase polyprotein 1a from SARS-CoV-2 (s006), replicase polyprotein 1ab that also originates from SARS-CoV-2 (s007) and replicase polyprotein 1a from SARS-CoV (s008). If we consider their identical role in the viral life cycle and the genome sequence similarity between SARS-CoV and SARS-CoV-2, this clustering could have been anticipated. The role of both replicase polyproteins in viral replication and their significance for coronaviruses were already discussed.

The first group of COVID-19 targets from the third cluster (marked red) contains six protein structures: s007 (envelope small protein in SARS-CoV), s009 (envelope small protein in SARS-CoV-2), s012 (mothers against decapentaplegic homolog 3 – SMAD3), s014 (ORF3a protein), s015 (angiotensin-converting enzyme 2) and s004 (prohibitin). This cluster is assembled of diverse protein targets. The first two proteins, s007 and s009, are the same macromolecules according to their function and structure but located in different viruses. As aforementioned, envelope small membrane protein has an important role in the viral life cycle that involves assembly, budding, interaction with host cells and pathogenesis (Sarkar and Saha, 2020). Structure s012 (SMAD3) is a human intracellular signal transducer and transcriptional modulator activated by TGF-beta, it is involved in various functions regulation cascades and interactions, with SARS-CoV nucleoprotein, among others (*UniProt- SMAD3 gene & protein*). ORF3a is a protein located in SARS-CoV-2 that forms viroporins and has also many other interactions with human compartments, that need further investigations in order to determine potential use of ORF3a in therapy against SARS-CoV-2 (Miao *et al.*, 2021). Last two proteins are human. It has already been stated that ACE-2 is a crucial membrane molecule for the entrance of the coronavirus into human cells and due to this, ACE-2 and its modulations could be an effective therapeutic strategy in fight against COVID-19 (Hoffmann *et al.*, 2020). Prohibitin is also a human protein but its role in the interaction with SARS-CoV-2 is not clear. According to UniProt, the only known interaction is the one with nsp2.

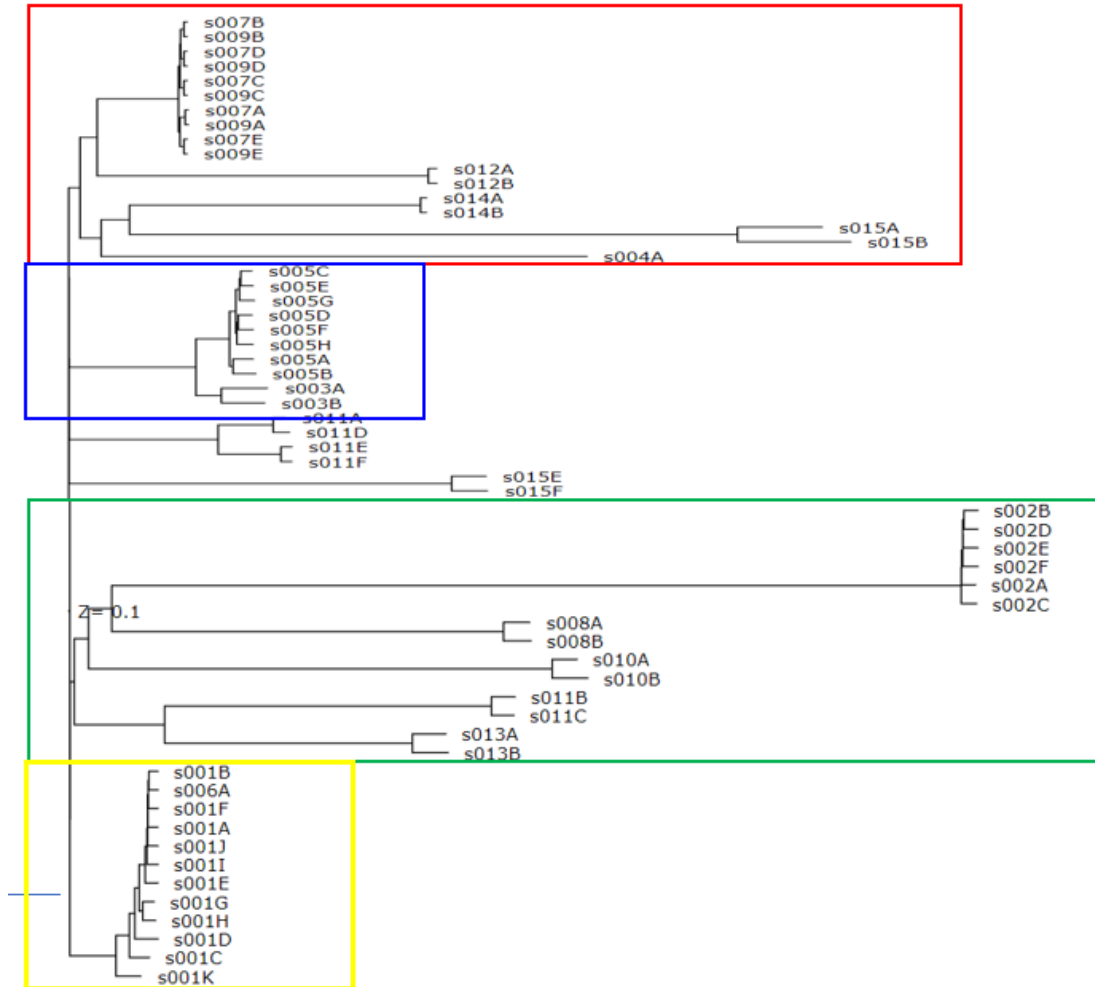


Figure 24. Structural similarity dendrogram for COVID-19 protein targets from the third cluster of the sequence-based dendrogram. The dendrogram is derived by average linkage clustering of the structural similarity matrix (Dali Z-scores).

The second cluster from **Figure 24** (colored blue) is composed of eight chains of protein target s005 (ORF9b from SARS-CoV) and two chains of s003 (ORF9b from SARS-CoV-2). In both coronaviruses, ORF9b plays a role in the inhibition of host innate immune response by targeting the mitochondrial-associated adapter MAVS (Shi *et al.*, 2014). These two macromolecules are the same proteins in different organisms that do not show any similarity with other targets.

The third group of COVID-19 targets from third cluster (in **Figure 24**, marked green) includes five different proteins: s002 (furin), s008 (integrin alpha L), s010 (ATP-dependent RNA helicase DDX1), s011 (interleukin-17 receptor A) and s013 (interleukin-6 receptor subunit beta). This cluster consists of 5 very diverse protein targets with the roles in different phases of viral infection. Furin is a crucial factor for the interaction and cleavage of spike protein, and the

modulation of its structure or function could provide huge potential as a treatment strategy. On the other hand, the integrins also show promising possibilities in COVID-19 therapy as they interact with certain parts of the spike protein and might be necessary for complete interaction between virus and human cells. The roles of IL-6 and IL-17 and their receptors are already described. However, the detected structural and sequence-based similarity of these proteins with the other protein targets from the same cluster previous two could be made use of for designing new therapy solutions which could interfere with multiple points of intervention during the viral life cycle simultaneously.

The last cluster from **Figure 24** reveals both sequence and structure similarity between s001 (Bone marrow stromal antigen 2) and s006 (prohibitin-2). Both are human macromolecules with antiviral activity. According to Uniprot, bone marrow stromal antigen 2 interacts with ORF7a protein in SARS-CoV, while prohibitin-2 interacts with nsp2 in both SARS-CoV and SARS-CoV-2. The majority of features related to these proteins have been established based on studies on SARS-CoV, and due to this, further investigations are necessary for potential adaption on SARS-CoV-2. These two structures could have prospective use in therapy against novel coronaviruses as their targeting could stimulate antiviral response by two divided routes.



Figure 25. Structural similarity dendrogram for COVID-19 protein targets from the fourth cluster of the sequence-based dendrogram. The dendrogram is derived by average linkage clustering of the structural similarity matrix (Dali Z-scores).

The protein targets from the last cluster of the sequence-based similarity dendrogram (**Figure 21**) are structurally aligned and represented in **Figure 25**. Two additional groups of related entities are generated. The first one involves s004 (ORF6 protein in SARS-CoV) and s006 (ORF6 protein in SARS-CoV-2). Once again, the performed structure alignment has confirmed the high level of similarity between the same proteins in SARS-CoV and SARS-CoV-2. According to this analysis, these targets have no similarity with any other protein structure from COVID-19 database provided by Uniprot. ORF6 has a function in viral activity and disruption of host immune response (Miorin *et al.*, 2020).

The second group from **Figure 25** includes four potential COVID-19 targets: s007 (membrane protein in SARS-CoV), s005 (interleukin-17 receptor C), s008 (membrane protein in SARS-CoV-2) and s001 (ORF3a protein in SARS-CoV). This cluster is also composed of two almost identical proteins from two different viruses (SARS-CoV and SARS-CoV-2) – membrane proteins, and additional two diverse targets (IL-17RC and ORF3a). Although they have essentially the identical role in two different coronaviruses, based on **Figure 25**, membrane proteins from SARS-CoV and SARS-CoV-2 share higher level of structural similarity with s005 and s001 than between each other. Another important aspect of their similarity could be the fact that M protein could interact with ORF3a (Yuan *et al.*, 2005). Nonetheless, further studies are necessary for the understanding of complete mechanisms of action and prospective therapy application of these entities.

As aforementioned, the last section of this work includes generation of a set of the available ligands for 53 COVID-19 protein targets. However, after the filtering of primary data and the cut off setting, the final table with unique compounds per target contains only 13 of 53 protein targets. This task was carried out in KNIME Analytics Platform and the workflow used for this work is a part of a case study published by Tuerkova & Zdrazil (Tuerkova and Zdrazil, 2020). This workflow has eight main segments. First five parts were adopted in this work, and they will be explained in the following. The first segment involves importing of specific drug targets via “File Reader” node. The file with specified COVID-19 targets is provided by UniProt (*UniProt*), and downloaded in TSV format. The second part of the used workflow encompasses retrieval of

available PDB codes for the set of the targets which is followed by the third part that includes fetching information about the protein structures of interest from PDB. Both the second and the third segment involve a creating of API to access UniProt and PDB data in XML format and for these tasks next three nodes were used: “String Manipulation”, “GET Request” and “XPath” node. In the fourth part of this workflow, the ligand data is obtained from PDB (by using the same procedure and above noted three nodes) and in the next step, fetched structures of these ligands are standardized. As already mentioned, in the fifth segment, a programmatic data access is used for retrieval of available compound data from ChEMBL, PubChem and IUPHAR data sources. In order to obtain active ligands per target, certain segments of this part of workflow were modified by setting the cutoff below 10 μ M (performed in “Row Filter” node). This part of workflow is represented in **Figure 26**.

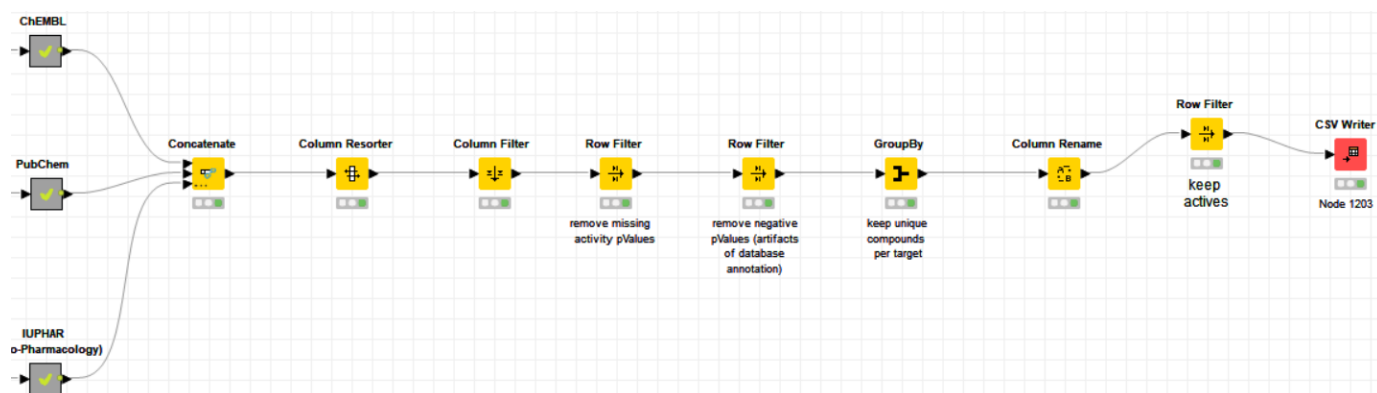


Figure 26. Part of modified KNIME Workflow from “An Integrative Drug Repurposing Pipeline using KNIME and Programmatic Data Access: A case study on COVID-19 Data case study written by Türková & Zdrzil”(Tuerkova and Zdrzil, 2020).

The final table obtained from this workflow includes 6166 ligands of 13 COVID-19 protein targets (output file was via the “CSV Writer” node).

The idea behind this work was to provide a set of ligands that are active against COVID-19 targets, which could be used for further investigations, in the first place for drug-repurposing approaches. Certain already registered drugs (such as remdesivir (Beigel *et al.*, 2020)) showed potential benefits when applied in patients infected with SARS-CoV-2. The herein retrieved set of ligands could serve as a good starting point for future studies and analysis in this field. The table that includes numbers of unique compounds retrieved for 13 potential COVID-19 protein targets is shown on **Figure 27**. The two structures that have highest number of potential ligands

are peptidyl-prolyl cis-trans isomerase A (P62937) – 2767 ligands, and procathepsin L (P07711) – 1972 ligands. These two human proteins have diverse functions in cellular life cycle. Although both macromolecules have the interactions with coronaviruses during their entry into human cells, due to role of procathepsin L in facilitation of SARS-CoV-2 infection via proteolysis of spike glycoproteins in lysosome (Ou *et al.*, 2020), this protein might be preferable target. The next three COVID-19 targets, according to number of unique ligands per target, are furin (P09958) with 649 unique ligands, integrin alpha L (P20701) – 411 ligands, and ACE2 (119) – with 119 unique ligands. The roles of these proteins, especially in the interaction with human cells have already been discussed. Nonetheless, another reason for further investigation of these targets and their ligands is the similarity between furin and integrin alpha L. According to **Figure 24** of this work, these two structures belong to the same cluster (colored green) which means that beside high number of ligands, these macromolecules possess structure and sequence-based similarity.

Table 4. Unique Compounds per COVID-19 Protein Target

Name of Protein Target	Sequence Alignment Abbreviation (UniProt ID)	Number of Unique Compounds per Target
Peptidyl-prolyl cis-trans isomerase A	P62937	2767
Procathepsin L	P07711	1972
Furin	P09958	649
Integrin alpha L	P20701	411
Angiotensin-converting enzyme 2	Q9BYF1	119
Importin subunit alpha 1	P52292	82
Replicase polyprotein 1ab	P0C6X7	64
Replicase polyprotein 1a	P0C6U8	49
Mothers against decapentaplegic	P84022	28
Interleukin-6	P05231	14
Interleukin-6 receptor subunit beta	P40189	5
Transmembrane protease serine 2	O15393	3
Two pore calcium channel protein 2	Q8NHX9	2
1-phosphatidylinositol 3-phosphate 5-kinase	Q9Y2I7	1

4. Discussion

The current COVID-19 pandemic still remains the main health issue worldwide. More than one year after the outbreak, the situation in most countries is still destructive and alarming and the majority of states, especially in European Union, report significantly worse pandemic conditions in comparison to March 2020. On the day of writing this paragraph (March 16, 2021), the number of new cases in Italy is 20 396, and the daily number of deaths is 502 (Worldometers, 2021). For comparison purposes, this Mediterranean country (one of the worst affected by COVID-19), on the same day last year had 3 235 new SARS-CoV-2 cases and 350 deaths. According to these numbers, it can obviously be noted that current situation with novel coronavirus is even more complicated and adverse. If we assume that the scientific and medical communities have more experience and knowledge about SARS-CoV-2, its features, main routes of transmission and interaction with the human immune system, the question that arises is why don't we manage to control this virus and the disease that it causes? Beside social factors that include very slow distribution of vaccines and non-compliance with COVID-19 related public health measures, the most probable explanation for this issue is the occurrence of new SARS-CoV-2 variants. According to CDC (Centers for Disease Control and Prevention) three main variants have emerged so far: B.1.1.7 (also known as UK or British variant), P.1 (Brazilian variant) and ` (strain that originates from South Africa). B.1.1.7 variants is registered in majority of European countries causing serious health issues due to induced transmission potential and probable higher risk of mortality (Challen *et al.*, 2021).

Still, simultaneously with the emergence of new SARS-CoV-2 strains, significant progress in therapy possibilities has been accomplished. Primarily, several vaccines have been developed and approved, but on the other hand, improved comprehension of viral infection and COVID-19 illness have provided better therapy opportunities and higher percentage of preferable outcome of the disease. Hereafter, the current approaches for the prevention and treatment against the disease caused by the novel coronavirus will be discussed.

4.1. State of the Art of Current Therapeutic Options for COVID-19

Besides public health and social measures that include primarily social distancing and mandatory mask wearing, most important aspect in the fight and the prevention of current global health crisis is the development and the production of therapeutic tools and weapons against SARS-CoV-2 and the COVID-19 disease. In the first place, this includes the design of vaccines (as the most powerful tool for the disease prevention) and the development of novel drugs or repurposing of already existing ones.

According to New York Times (*Covid-19 Vaccine Tracker Updates: The Latest - The New York Times*, 2021), 98 vaccines are currently in one of first three phases of clinical trials (44 in the first phase, 32 in the second and 22 in the third). If we consider that 7 vaccines are approved for full use, and only 4 were abandoned after trials, the registration and approval of new vaccines could be expected in the following months. It is important to emphasize that all accepted vaccines are actually in the IV phase of clinical trials (some of them even in phase III) also named post-marketing surveillance trial. This part of clinical testing involves pharmacovigilance (the safety surveillance of registered drug) and technical support of a new drug that already has permission to be on the market. The same rule applies for vaccines as well, and the use of drugs/vaccines that are a part of this phase of clinical trial does not diminish its safety and efficacy. This kind of safety surveillance is designed to follow and detect long-term side effects over larger patient population. In the following section, the basic features of 6 approved and mostly used vaccines will be the subject of this discussion.

The first company to report positive phase 3 clinical data was Pfizer. In cooperation with a German biotechnology company BioNTech, Pfizer developed a vaccine that was approved for emergency use on December 11, 2020 in USA and on December 21 2020 in Europe. The vaccine was named Comirnaty (also known as tozinameran or BNT162b2) and according to available data, its efficacy is 95% (*Bloomger: Covid-19 Vaccine Tracker*, 2021). BNT162b2 is administered in two doses with 3 weeks apart. This vaccine uses messenger RNA (mRNA) that encodes the SARS-CoV-2 spike protein modified by two proline mutations to sustain the prefusion conformation (Polack *et al.*, 2020). Pfizer/BioNTech vaccine has lipid nanoparticle–formulation that protects the active substance from degradation. After vaccination, the vaccine particles penetrate into human cells, this is followed by mRNA release and reading of its

sequence. In the next step, based on the mRNA sequence, spike proteins are built and presented in the complex with antigen-presenting cells. These fragments of SARS-CoV-2 spike protein are then recognized by T cells which leads to activation of B cells and the production of antibodies against the spike protein. The mRNA vaccines are a recently developed type of vaccines, that showed very high efficacy in COVID-19 prophylaxis. However, the Pfizer/BioNTech vaccine has a challenging distribution as it needs to be kept frozen at -70 °C. The reason for this is already mentioned lipid nanoparticle-formulation.

Sputnik V (also known as Gam-Covid-Vac) is a viral vector vaccine developed by the Gamaleya Research Institute in Russia. This vaccine deployed from the combination of two adenoviruses – Ad5 and Ad26 (Jones and Roy, 2021). Sputnik V vaccine also contains SARS-CoV-2 genetic instructions for building the spike protein, but unlike Pfizer/BioNTech vaccine, the genetic material is stored in double-stranded DNA. (*Covid-19 Vaccine Tracker Updates: The Latest - The New York Times*, 2021). The two mentioned adenoviruses are used as the carriers in this formulation. The spike protein genes are added to Ad5 and Ad26 which can invade human cells but cannot replicate. The vaccine efficacy of 91,6% was confirmed in phase 3 of clinical trials and officially published on February 20 in the Lancet (Logunov *et al.*, 2021). Although the development and production of Sputnik V vaccine was followed by absence of transparency and haste, the final outcome according to the investigations is clear which means that this vaccine can also be used in a fight against the incidence of COVID-19 (Jones and Roy, 2021).

The second authorized vaccine by FDA is mRNA-1273, made by Boston-based company Moderna. This vaccine is also made from mRNA, like the one produced by Pfizer/BioNTech. mRNA-1273 is administered in two doses with 28 days apart. The investigations revealed 94.1% efficacy of Moderna vaccine at preventing COVID-19 illness, including severe disease (Baden *et al.*, 2021). The patients' response to this vaccine is very satisfactory and comparable to that in patients vaccinated with Pfizer/BioNTech product. Another advantage of mRNA-1273 is its storage – this vaccine can be kept for 30 days in refrigerator and up to 6 months at the -20 °C (*Covid-19 Vaccine Tracker Updates: The Latest - The New York Times*, 2021).

The British-Swedish company AstraZeneca partnered with The University of Oxford to develop a COVID-19 vaccine known as AZD1222 or ChAdOx1. According to clinical trials, the efficacy of this vaccine is 82,4%. (Ramasamy *et al.*, 2020). AZD1222 is a type of viral vector vaccine given

in two doses with 12 weeks apart. Similar to Sputnik V vaccine, in the case of AstraZeneca/Oxford vaccine, modified version of chimpanzee adenovirus (named ChAdOx1) is used as a carrier of the SARS-CoV-2 spike protein gene. Although the studies confirmed high immune response even after first dose of ChAdOx1 vaccine, (Hung and Poland, 2021) certain countries such as Denmark, France and Ireland have temporarily suspended the use of this vaccine. The reason for this decision lies in adverse side effects that could be related to AstraZeneca/Oxford vaccine. Namely, some people developed blood clots after receiving a first dose of ChAdOx1 vaccine, and in a case of patient in Denmark, this side effect led to a critical health condition that ended lethal. Although AZD1222 vaccine definitely has high potential of immunogenicity, the incidence of side effects (in majority of cases mild ones such as fever, headache, weakness) is higher in comparison to other approved vaccines. Nevertheless, according to EMA (European Medicines Agency) benefits of AstraZeneca/Oxford vaccine still exceed the risks despite possible link to adverse side effects.

The Beijing Institute of Biological Products and state-owned Chinese company Sinopharm created an inactivated coronavirus vaccine named BBIBP-CorV. Clinical trials led by Sinopharm revealed the efficacy rate of 79%. This vaccine is administered in two doses with three weeks apart. BBIBP-CorV vaccine was created by obtaining the SARS-CoV-2 variant from hospitalized patients. After the multiplication of the virus in monkey kidney, the large stocks of 2019-nCov were deactivated via beta-propiolactone that bonded the viral genes (*Covid-19 Vaccine Tracker Updates: The Latest - The New York Times*, 2021). The virus was not able to replicate anymore, but its proteins, including spike protein, remained intact and capable of triggering human immune response. Recent studies showed that Sinopharm vaccine is effective against the new strain of SARS-CoV-2 - B.1.351 but this studies have not been published in any medical journal yet.

The last approved vaccine in certain European countries and USA is the one produced by Johnson & Johnson company, named Ad26.COV2.S. The data reveals the variable efficacy of this vaccine. According to investigations, the efficacy rate of Ad26.COV2.S was 61% in South America, 64% in South Africa and 72% in USA (*Covid-19 Vaccine Tracker Updates: The Latest - The New York Times*, no date). Johnson & Johnson's method for the development of vaccine against SARS-CoV-2 was already applied for the vaccines Ebola and it involves Adenovirus 26 (Ad26) as a carrier (the same type of vaccine as those created by Gamaleya Research Institute in Russia and AstraZeneca/Oxford). This vaccine requires only one dose and it can be stored for 3

months on refrigerator temperature and up to two years at -20°C (*Covid-19 Vaccine Tracker Updates: The Latest - The New York Times*, 2021).

Current treatment approaches for COVID-19 can be divided into two main categories: **immune modulators** which help our immune system to cope with the virus and prevent the overreaction and **antivirals** which prevent viral multiplication (*The Pharmaceutical Journal*, 2021). Certain used drugs involve multiple mechanisms. At the same time, numerous clinical trials of COVID-19 therapeutics are performed around the world. So far, remdesivir is the only drug approved by FDA against COVID-19 disease. However, the research suggests that this medication provides only modest benefits for patients (*Coronavirus Drug and Treatment Tracker - The New York Times*, 2021). Many other already registered drugs are being tested against SARS-CoV-2 looking at drug-repurposing possibilities. Some of them showed not promising results such as hydroxychloroquine and chloroquine, oleandrin, lopinavir and ritonavir, while other had tentative or mixed evidence (molnupiravir and favipiravir e.g.) (*Coronavirus Drug and Treatment Tracker - The New York Times*, 2021). The drugs that modulate human immune response to the infection also show inconsistent results. Certain corticosteroids (dexamethasone on the first place) and cytokine inhibitors (tocilizumab) revealed promising evidence when administered in early stage of disease. At the same time, the application of interferons and monoclonal antibodies resulted in indefinite immune response (*Coronavirus Drug and Treatment Tracker - The New York Times*, 2021). It is tempting to speculate that the best therapy of COVID-19 might include a combination of different types of drugs, primarily depending on patient's clinical condition and stage of the disease.

5. Conclusion

More than one year after the COVID-19 outbreak, the novel coronavirus and the disease that it causes remains the main global health issue. The current pandemic has made devastating impact not only on people's physical health, but mentally as well. Everyday life like we experienced it 1.5 years ago is at the moment inconceivable. The consequences of COVID-19 pandemic affect many other spheres of life such as business and economy, tourism, education and many other aspects of social life. Based on work in this thesis, there has been an obvious progress in the fight against this pandemic. Currently (March 2021), the global numbers of new cases per day are lower than those from the months November 2020 until Jan 2021. Another encouraging epidemiological parameter is the daily number of deaths on a global level which shows a decreasing trend since January 2021. However, new strains of SARS-CoV-2 that were also discussed in this thesis and that have higher potential of transmission are the main reasons for a currently reinforced impact of the COVID-19 disease. The most powerful mechanisms that we have in this struggle are mass vaccinations and people's compliance with disease preventive measures. Besides, the faith and the support of the scientific community and of health workers, together with patience and perseverance are essential.

6. References

3a - ORF3a protein - Severe acute respiratory syndrome coronavirus 2 (2019-nCoV) - 3a gene & protein (2020) UniProtKB. Available at: <https://www.uniprot.org/uniprot/P0DTC3> (Accessed: 14 December 2020).

6 - ORF6 protein - Severe acute respiratory syndrome coronavirus 2 (2019-nCoV) - 6 gene & protein (no date). Available at: <https://www.uniprot.org/uniprot/P0DTC6> (Accessed: 13 December 2020).

7a - ORF7a protein precursor - Severe acute respiratory syndrome coronavirus 2 (2019-nCoV) - 7a gene & protein (no date). Available at: <https://www.uniprot.org/uniprot/P0DTC7> (Accessed: 13 December 2020).

9b - ORF9b protein - Severe acute respiratory syndrome coronavirus 2 (2019-nCoV) - 9b gene & protein (no date). Available at: <https://www.uniprot.org/uniprot/P0DTD2> (Accessed: 13 December 2020).

Abbadi, O. S. (no date) *Non-Structural Proteins of SARS-CoV-2 as potential sources for vaccine synthesis*.

Aboubakr, H. A., Sharafeldin, T. A. and Goyal, S. M. (2020) ‘Stability of SARS-CoV-2 and other coronaviruses in the environment and on common touch surfaces and the influence of climatic conditions: A review’, *Transboundary and Emerging Diseases*. doi: 10.1111/tbed.13707.

About - ChEMBL Interface Documentation (no date). Available at: <https://chembl.gitbook.io/chembl-interface-documentation/about> (Accessed: 7 April 2021).

About - Our World in Data (no date). Available at: <https://ourworldindata.org/about> (Accessed: 7 April 2021).

About | IUPHAR/BPS Guide to PHARMACOLOGY (no date). Available at: <https://www.guidetopharmacology.org/about.jsp> (Accessed: 7 April 2021).

About RaptorX (no date). Available at: <http://raptorx.uchicago.edu/about/> (Accessed: 7 April 2021).

About UniProt (no date). Available at: <https://www.uniprot.org/help/about> (Accessed: 7 April 2021).

Aftab, S. O. *et al.* (2020) ‘Analysis of SARS-CoV-2 RNA-dependent RNA polymerase as a potential therapeutic drug target using a computational approach’, *Journal of Translational Medicine*. doi: 10.1186/s12967-020-02439-0.

Agentur für Gesundheit und Ernährungssicherheit – Wikipedia (no date). Available at: https://de.wikipedia.org/wiki/Agentur_für_Gesundheit_und_Ernährungssicherheit (Accessed: 7 April 2021).

2021).

Amirian, E. S. (2020) 'Potential fecal transmission of SARS-CoV-2: Current evidence and implications for public health', *International Journal of Infectious Diseases*. doi: 10.1016/j.ijid.2020.04.057.

Annette Plüddemann, J. K. A. (2020) 'What is the role of T cells in COVID-19 infection? Why immunity is about more than antibodies', *Centre for Evidence-Based Medicine*.

Asghari, A. *et al.* (2020) 'The Novel Insight of SARS-CoV-2 Molecular Biology and Pathogenesis and Therapeutic Options', *DNA and Cell Biology*. doi: 10.1089/dna.2020.5703.

Astuti, I. and Ysrafil (2020) 'Severe Acute Respiratory Syndrome Coronavirus 2 (SARS-CoV-2): An overview of viral structure and host response', *Diabetes and Metabolic Syndrome: Clinical Research and Reviews*, 14(4), pp. 407–412. doi: 10.1016/j.dsx.2020.04.020.

Azkur, A. K. *et al.* (2020) 'Immune response to SARS-CoV-2 and mechanisms of immunopathological changes in COVID-19', *Allergy: European Journal of Allergy and Clinical Immunology*. doi: 10.1111/all.14364.

Baden, L. R. *et al.* (2021) 'Efficacy and Safety of the mRNA-1273 SARS-CoV-2 Vaccine', *New England Journal of Medicine*. doi: 10.1056/nejmoa2035389.

Beck, B. R. *et al.* (2020) 'Predicting commercially available antiviral drugs that may act on the novel coronavirus (SARS-CoV-2) through a drug-target interaction deep learning model', *Computational and Structural Biotechnology Journal*. doi: 10.1016/j.csbj.2020.03.025.

Beigel, J. H. *et al.* (2020) 'Remdesivir for the Treatment of Covid-19 — Final Report', *New England Journal of Medicine*. doi: 10.1056/nejmoa2007764.

Berman, H. M. *et al.* (2000) 'The Protein Data Bank', *Nucleic Acids Research*. doi: 10.1093/nar/28.1.235.

Bhardwaj, K. *et al.* (2008) 'Structural and functional analyses of the severe acute respiratory syndrome coronavirus endoribonuclease Nsp15', *Journal of Biological Chemistry*. doi: 10.1074/jbc.M708375200.

Broggi, A. *et al.* (2020) 'Type III interferons disrupt the lung epithelial barrier upon viral recognition', *Science*. doi: 10.1126/science.abc3545.

BSG - Basigin precursor - *Homo sapiens* (Human) - BSG gene & protein (no date). Available at: <https://www.uniprot.org/uniprot/P35613> (Accessed: 11 April 2021).

Burki, T. (2020a) 'China's successful control of COVID-19', *The Lancet. Infectious diseases*. doi: 10.1016/S1473-3099(20)30800-8.

Burki, T. (2020b) ‘The origin of SARS-CoV-2’, *The Lancet. Infectious diseases*. doi: 10.1016/S1473-3099(20)30641-1.

Butowt, R. and von Bartheld, C. S. (2020) ‘Anosmia in COVID-19: Underlying Mechanisms and Assessment of an Olfactory Route to Brain Infection’, *Neuroscientist*. doi: 10.1177/1073858420956905.

Cascarina, S. M. and Ross, E. D. (2020) ‘A proposed role for the SARS-CoV-2 nucleocapsid protein in the formation and regulation of biomolecular condensates’, *FASEB Journal*. doi: 10.1096/fj.202001351.

Cascella, M. *et al.* (2020) *Features, Evaluation and Treatment Coronavirus (COVID-19)*, *StatPearls*.

Challen, R. *et al.* (2021) ‘Risk of mortality in patients infected with SARS-CoV-2 variant of concern 202012/1: matched cohort study.’, *BMJ (Clinical research ed.)*, 372, p. n579. doi: 10.1136/bmj.n579.

ChEMBL - Wikipedia (no date). Available at: <https://en.wikipedia.org/wiki/ChEMBL> (Accessed: 7 April 2021).

Cheminformatics - Latest research and news | Nature (no date). Available at: <https://www.nature.com/subjects/cheminformatics> (Accessed: 7 April 2021).

Chen, B. *et al.* (2020) ‘Overview of lethal human coronaviruses’, *Signal Transduction and Targeted Therapy*. doi: 10.1038/s41392-020-0190-2.

Chen, L. Y. C. and Quach, T. T. T. (2021) ‘COVID-19 cytokine storm syndrome: a threshold concept’, *The Lancet Microbe*. doi: 10.1016/s2666-5247(20)30223-8.

Chen, X. *et al.* (2020) ‘SARS-CoV-2 on the ocular surface: is it truly a novel transmission route?’, *British Journal of Ophthalmology*, p. bjophthalmol-2020-316263. doi: 10.1136/bjophthalmol-2020-316263.

Chowdhury, M. A. *et al.* (2020) ‘Immune response in COVID-19: A review’, *Journal of Infection and Public Health*. Elsevier Ltd, pp. 1619–1629. doi: 10.1016/j.jiph.2020.07.001.

Coronavirus disease (COVID-19) (no date). Available at: <https://www.who.int/emergencies/diseases/novel-coronavirus-2019> (Accessed: 8 December 2020).

Coronavirus Drug and Treatment Tracker - The New York Times (no date). Available at: <https://www.nytimes.com/interactive/2020/science/coronavirus-drugs-treatments.html> (Accessed: 22 March 2021).

Covid-19 Vaccine Tracker Updates: The Latest - The New York Times (no date). Available at: <https://www.nytimes.com/interactive/2020/science/coronavirus-vaccine-tracker.html> (Accessed: 20 March 2021).

Cron, R. Q. (2021) ‘COVID-19 cytokine storm: targeting the appropriate cytokine’, *The Lancet Rheumatology*. doi: 10.1016/S2665-9913(21)00011-4.

Dali server (no date). Available at: <http://ekhidna2.biocenter.helsinki.fi/dali/> (Accessed: 7 April 2021).

Data Fellows Programme – The Centre for Humanitarian Data (no date). Available at: <https://centre.humdata.org/data-fellows/> (Accessed: 7 April 2021).

Dhama, K. *et al.* (2020) ‘Coronavirus disease 2019–COVID-19’, *Clinical Microbiology Reviews*, 33(4), pp. 1–48. doi: 10.1128/CMR.00028-20.

Doan, T. L. *et al.* (2011) ‘The Future of Drug Repositioning. Old Drugs, New Opportunities’, in *Annual Reports in Medicinal Chemistry*. doi: 10.1016/B978-0-12-386009-5.00004-7.

Dockery, D. M. *et al.* (2020) ‘The Ocular Manifestations and Transmission of COVID-19: Recommendations for Prevention’, *Journal of Emergency Medicine*. doi: 10.1016/j.jemermed.2020.04.060.

Everything you need to know about the COVID-19 therapy trials - The Pharmaceutical Journal (no date). Available at: <https://pharmaceutical-journal.com/article/feature/everything-you-need-to-know-about-the-covid-19-therapy-trials> (Accessed: 22 March 2021).

Fan, Y. *et al.* (2019) ‘Bat coronaviruses in China’, *Viruses*. doi: 10.3390/v11030210.

Fani, M., Teimoori, A. and Ghafari, S. (2020) ‘Comparison of the COVID-2019 (SARS-CoV-2) pathogenesis with SARS-CoV and MERS-CoV infections’, *Future Virology*. doi: 10.2217/fvl-2020-0050.

Fehr, A. R. and Perlman, S. (2015) ‘Coronaviruses: An overview of their replication and pathogenesis’, in *Coronaviruses: Methods and Protocols*. doi: 10.1007/978-1-4939-2438-7_1.

Ferretti, L. *et al.* (2020) ‘Quantifying SARS-CoV-2 transmission suggests epidemic control with digital contact tracing’, *Science*. doi: 10.1126/science.abb6936.

Filippini, A. *et al.* (2020) ‘Could the Inhibition of Endo-Lysosomal Two-Pore Channels (TPCs) by the Natural Flavonoid Naringenin Represent an Option to Fight SARS-CoV-2 Infection?’, *Frontiers in Microbiology*. doi: 10.3389/fmicb.2020.00970.

García-Salido, A. (2020) ‘Narrative review of the immune response against coronavirus: An overview, applicability for SARS-COV-2, and therapeutic implications’, *Anales de Pediatría*. doi: 10.1016/j.anpedi.2020.04.016.

Gordon, D. E. *et al.* (2020) ‘A SARS-CoV-2 protein interaction map reveals targets for drug repurposing’,

Nature. doi: 10.1038/s41586-020-2286-9.

Gu, J. and Korteweg, C. (2007) 'Pathology and pathogenesis of severe acute respiratory syndrome', *American Journal of Pathology*. doi: 10.2353/ajpath.2007.061088.

Health Ontario, P. and publique Ontario, S. (no date) *COVID-19 Routes of Transmission – What We Know So Far*.

Helmy, Y. A. *et al.* (2020) 'The COVID-19 Pandemic: A Comprehensive Review of Taxonomy, Genetics, Epidemiology, Diagnosis, Treatment, and Control', *Journal of Clinical Medicine*. doi: 10.3390/jcm9041225.

Hennigan, S. and Kavanaugh, A. (2008) 'Interleukin-6 inhibitors in the treatment of rheumatoid arthritis', *Therapeutics and Clinical Risk Management*. doi: 10.2147/tcrm.s3470.

Hillen, H. S. *et al.* (2020) 'Structure of replicating SARS-CoV-2 polymerase', *Nature*. doi: 10.1038/s41586-020-2368-8.

Hoffmann, M. *et al.* (2020) 'SARS-CoV-2 Cell Entry Depends on ACE2 and TMPRSS2 and Is Blocked by a Clinically Proven Protease Inhibitor', *Cell*. doi: 10.1016/j.cell.2020.02.052.

Hojyo, S. *et al.* (2020) 'How COVID-19 induces cytokine storm with high mortality', *Inflammation and Regeneration*. doi: 10.1186/s41232-020-00146-3.

Huang, C. *et al.* (2020) 'Clinical features of patients infected with 2019 novel coronavirus in Wuhan, China', *The Lancet*. doi: 10.1016/S0140-6736(20)30183-5.

Hung, I. F. N. and Poland, G. A. (2021) 'Single-dose Oxford-AstraZeneca COVID-19 vaccine followed by a 12-week booster.', *Lancet (London, England)*, 397(10277), pp. 854–855. doi: 10.1016/S0140-6736(21)00528-6.

IL17F - Interleukin-17F precursor - Homo sapiens (Human) - IL17F gene & protein (no date). Available at: <https://www.uniprot.org/uniprot/Q96PD4> (Accessed: 8 April 2021).

index | TIOBE - The Software Quality Company (no date). Available at: <https://www.tiobe.com/tiobe-index/> (Accessed: 7 April 2021).

Introduction to Python (no date). Available at: https://www.w3schools.com/python/python_intro.asp (Accessed: 7 April 2021).

InvivoGen (no date) *Spotlight on COVID-19 - The infection cycle of SARS-CoV-2*. Available at: www.invivogen.com (Accessed: 13 December 2020).

- Issa, E. *et al.* (2020) ‘SARS-CoV-2 and ORF3a: Nonsynonymous Mutations, Functional Domains, and Viral Pathogenesis’, *mSystems*. doi: 10.1128/msystems.00266-20.
- J Alsaadi, E. A. and Jones, I. M. (2019) ‘Membrane binding proteins of coronaviruses’, *Future Virology*. doi: 10.2217/fvl-2018-0144.
- Jones, I. and Roy, P. (2021) ‘Sputnik V COVID-19 vaccine candidate appears safe and effective’, *The Lancet*. doi: 10.1016/s0140-6736(21)00191-4.
- Kawahara, H., Minami, R. and Yokota, N. (2013) ‘BAG6/BAT3: Emerging roles in quality control for nascent polypeptides’, *Journal of Biochemistry*. doi: 10.1093/jb/mvs149.
- Kim, D. *et al.* (2020) ‘The Architecture of SARS-CoV-2 Transcriptome’, *Cell*. doi: 10.1016/j.cell.2020.04.011.
- KNIME - Wikipedia (no date). Available at: <https://en.wikipedia.org/wiki/KNIME> (Accessed: 7 April 2021).
- Konno, Y. *et al.* (2020) ‘SARS-CoV-2 ORF3b Is a Potent Interferon Antagonist Whose Activity Is Increased by a Naturally Occurring Elongation Variant’, *Cell Reports*. doi: 10.1016/j.celrep.2020.108185.
- Kopecky-Bromberg, S. A. *et al.* (2007) ‘Severe Acute Respiratory Syndrome Coronavirus Open Reading Frame (ORF) 3b, ORF 6, and Nucleocapsid Proteins Function as Interferon Antagonists’, *Journal of Virology*. doi: 10.1128/jvi.01782-06.
- Krichel, B. *et al.* (2020) ‘Processing of the SARS-CoV pp1a/ab nsp7-10 region’, *Biochemical Journal*, 477(5), pp. 1009–1019. doi: 10.1042/BCJ20200029.
- Kumar, S. *et al.* (2020) ‘Morphology, Genome Organization, Replication, and Pathogenesis of Severe Acute Respiratory Syndrome Coronavirus 2 (SARS-CoV-2)’, in. doi: 10.1007/978-981-15-4814-7_3.
- Lee, J. S. and Shin, E. C. (2020) ‘The type I interferon response in COVID-19: implications for treatment’, *Nature Reviews Immunology*. doi: 10.1038/s41577-020-00429-3.
- Leznicki, P. *et al.* (2010) ‘Bat3 promotes the membrane integration of tail-anchored proteins’, *Journal of Cell Science*, 123(13), pp. 2170–2178. doi: 10.1242/jcs.066738.
- Li, M. Y. *et al.* (2020) ‘Expression of the SARS-CoV-2 cell receptor gene ACE2 in a wide variety of human tissues’, *Infectious Diseases of Poverty*. doi: 10.1186/s40249-020-00662-x.
- Logunov, D. Y. *et al.* (2021) ‘Safety and efficacy of an rAd26 and rAd5 vector-based heterologous prime-boost COVID-19 vaccine: an interim analysis of a randomised controlled phase 3 trial in Russia’, *The*

Lancet. doi: 10.1016/s0140-6736(21)00234-8.

de Lusignan, S. *et al.* (2020) 'Risk factors for SARS-CoV-2 among patients in the Oxford Royal College of General Practitioners Research and Surveillance Centre primary care network: a cross-sectional study', *The Lancet Infectious Diseases*. doi: 10.1016/S1473-3099(20)30371-6.

Ma, W. T. *et al.* (2019) 'The protective and pathogenic roles of IL-17 in viral infections: Friend or foe?', *Open Biology*. doi: 10.1098/rsob.190109.

Mak, G. C. *et al.* (2020) 'Evaluation of rapid antigen test for detection of SARS-CoV-2 virus', *Journal of Clinical Virology*. doi: 10.1016/j.jcv.2020.104500.

Marbán, E. and Akhmerov, A. (2020) 'COVID-19 and the Heart', *Circulation Research*, 126, pp. 1443–1455. doi: 10.1161/CIRCRESAHA.120.317055.

Mathuria, J. P., Yadav, R. and Rajkumar (2020) 'Laboratory diagnosis of SARS-CoV-2 - A review of current methods', *Journal of Infection and Public Health*. doi: 10.1016/j.jiph.2020.06.005.

Megna, M., Napolitano, M. and Fabbrocini, G. (2020) 'May IL-17 have a role in COVID-19 infection?', *Medical Hypotheses*. Churchill Livingstone, p. 109749. doi: 10.1016/j.mehy.2020.109749.

Meirson, T., Bomze, D. and Markel, G. (2020) 'Structural basis of SARS-CoV-2 spike protein induced by ACE2', *Bioinformatics*. doi: 10.1101/2020.05.24.113175.

Melenotte, C. *et al.* (2020) 'Immune responses during COVID-19 infection', *OncoImmunology*, 9(1). doi: 10.1080/2162402X.2020.1807836.

Meyerowitz, E. A. *et al.* (2020) 'Transmission of SARS-CoV-2: A Review of Viral, Host, and Environmental Factors', *Annals of Internal Medicine*. doi: 10.7326/m20-5008.

Miao, G. *et al.* (2021) 'ORF3a of the COVID-19 virus SARS-CoV-2 blocks HOPS complex-mediated assembly of the SNARE complex required for autolysosome formation', *Developmental Cell*. doi: 10.1016/j.devcel.2020.12.010.

Michel, C. J. *et al.* (2020) 'Characterization of accessory genes in coronavirus genomes', *Virology Journal*. doi: 10.1186/s12985-020-01402-1.

Miorin, L. *et al.* (2020) 'SARS-CoV-2 Orf6 hijacks Nup98 to block STAT nuclear import and antagonize interferon signaling', *Proceedings of the National Academy of Sciences of the United States of America*. doi: 10.1073/pnas.2016650117.

Mizumoto, K., Kagaya, K. and Chowell, G. (2020) 'Effect of a wet market on coronavirus disease

(COVID-19) transmission dynamics in China, 2019–2020’, *International Journal of Infectious Diseases*. doi: 10.1016/j.ijid.2020.05.091.

Mollica, V., Rizzo, A. and Massari, F. (2020) ‘The pivotal role of TMPRSS2 in coronavirus disease 2019 and prostate cancer’, *Future Oncology*. doi: 10.2217/fon-2020-0571.

Moore, R. M. *et al.* (2020) ‘Iroki: automatic customization and visualization of phylogenetic trees’, *PeerJ*. doi: 10.7717/peerj.8584.

Morawska, L. and Cao, J. (2020) ‘Airborne transmission of SARS-CoV-2: The world should face the reality’, *Environment International*. doi: 10.1016/j.envint.2020.105730.

More Than 423 Million Shots Given: Covid-19 Vaccine Tracker (no date). Available at: <https://www.bloomberg.com/graphics/covid-vaccine-tracker-global-distribution/> (Accessed: 20 March 2021).

Mullol, J. *et al.* (2020) ‘The Loss of Smell and Taste in the COVID-19 Outbreak: a Tale of Many Countries’, *Current Allergy and Asthma Reports*. doi: 10.1007/s11882-020-00961-1.

Nadimpalli, M. L. and Pickering, A. J. (2020) ‘A call for global monitoring of WASH in wet markets’, *The Lancet Planetary Health*. doi: 10.1016/S2542-5196(20)30204-7.

Naqvi, A. A. T. *et al.* (2020) ‘Insights into SARS-CoV-2 genome, structure, evolution, pathogenesis and therapies: Structural genomics approach’, *Biochimica et Biophysica Acta - Molecular Basis of Disease*. doi: 10.1016/j.bbadis.2020.165878.

Neuman, B. W. *et al.* (2008) ‘Proteomics Analysis Unravels the Functional Repertoire of Coronavirus Nonstructural Protein 3’, *Journal of Virology*. doi: 10.1128/jvi.02631-07.

NSP13 of SARS-CoV-2 (2019-nCoV) - Order easily at ProteoGenix (no date). Available at: <https://www.proteogenix.science/product/nsp13/> (Accessed: 13 December 2020).

NYC Health (2021). Available at: <https://www1.nyc.gov/site/doh/index.page> (Accessed: 5 January 2021).

Open Source Story | KNIME (no date). Available at: <https://www.knime.com/knime-open-source-story> (Accessed: 7 April 2021).

Ou, X. *et al.* (2020) ‘Characterization of spike glycoprotein of SARS-CoV-2 on virus entry and its immune cross-reactivity with SARS-CoV’, *Nature Communications*. doi: 10.1038/s41467-020-15562-9.

Our World in Data - Wikipedia (no date). Available at: https://en.wikipedia.org/wiki/Our_World_in_Data (Accessed: 7 April 2021).

- Pacha, O., Sallman, M. A. and Evans, S. E. (2020) 'COVID-19: a case for inhibiting IL-17?', *Nature Reviews Immunology*. doi: 10.1038/s41577-020-0328-z.
- Papadatos, G. *et al.* (2014) 'A document classifier for medicinal chemistry publications trained on the ChEMBL corpus', *Journal of Cheminformatics*. doi: 10.1186/s13321-014-0040-8.
- Paraskevis, D. *et al.* (2020) 'Full-genome evolutionary analysis of the novel corona virus (2019-nCoV) rejects the hypothesis of emergence as a result of a recent recombination event', *Infection, Genetics and Evolution*. doi: 10.1016/j.meegid.2020.104212.
- Payne, S. (2017a) 'Family Coronaviridae', in *Viruses*. Elsevier, pp. 149–158. doi: 10.1016/b978-0-12-803109-4.00017-9.
- Payne, S. (2017b) 'Family Coronaviridae', in *Viruses*. doi: 10.1016/b978-0-12-803109-4.00017-9.
- Polack, F. P. *et al.* (2020) 'Safety and Efficacy of the BNT162b2 mRNA Covid-19 Vaccine', *New England Journal of Medicine*. doi: 10.1056/nejmoa2034577.
- Poland, G. A., Ovsyannikova, I. G. and Kennedy, R. B. (2020) 'SARS-CoV-2 immunity: review and applications to phase 3 vaccine candidates', *The Lancet*. doi: 10.1016/S0140-6736(20)32137-1.
- Portela Sousa, C. and Brites, C. (2020) 'Immune response in SARS-CoV-2 infection: the role of interferons type I and type III', *Brazilian Journal of Infectious Diseases*. doi: 10.1016/j.bjid.2020.07.011.
- Principles of Epidemiology | Lesson 1 - Section 1* (no date). Available at: https://www.cdc.gov/csels/dsepd/ss1978/Lesson1/Section1.html#_ref1 (Accessed: 7 April 2021).
- Protein Data Bank - Wikipedia* (no date). Available at: https://en.wikipedia.org/wiki/Protein_Data_Bank (Accessed: 7 April 2021).
- Pruijssers, A. J. and Denison, M. R. (2019) 'Nucleoside analogues for the treatment of coronavirus infections', *Current Opinion in Virology*. doi: 10.1016/j.coviro.2019.04.002.
- PubChem - Wikipedia* (no date). Available at: <https://en.wikipedia.org/wiki/PubChem> (Accessed: 7 April 2021).
- Python (programming language) - Wikipedia* (no date). Available at: [https://en.wikipedia.org/wiki/Python_\(programming_language\)](https://en.wikipedia.org/wiki/Python_(programming_language)) (Accessed: 7 April 2021).
- Python Advantages and Disadvantages - Step in the right direction - TechVidvan* (no date). Available at: <https://techvidvan.com/tutorials/python-advantages-and-disadvantages/> (Accessed: 7 April 2021).
- R: What is R?* (no date). Available at: <https://www.r-project.org/about.html> (Accessed: 7 April 2021).

R (programming language) - Wikipedia (no date). Available at:

[https://en.wikipedia.org/wiki/R_\(programming_language\)](https://en.wikipedia.org/wiki/R_(programming_language)) (Accessed: 7 April 2021).

Ramasamy, M. N. *et al.* (2020) ‘Safety and immunogenicity of ChAdOx1 nCoV-19 vaccine administered in a prime-boost regimen in young and old adults (COV002): a single-blind, randomised, controlled, phase 2/3 trial’, *The Lancet*. doi: 10.1016/S0140-6736(20)32466-1.

RaptorX - Wikipedia (no date). Available at: <https://en.wikipedia.org/wiki/RaptorX> (Accessed: 7 April 2021).

RCSB PDB: About RCSB PDB: Enabling Breakthroughs in Scientific and Biomedical Research and Education (no date). Available at: <https://www.rcsb.org/pages/about-us/index> (Accessed: 7 April 2021).

RCSB PDB: PDB History (no date). Available at: <https://www.rcsb.org/pages/about-us/history> (Accessed: 7 April 2021).

Redzic, J. S. *et al.* (2011) ‘The retinal specific CD147 Ig0 domain: From molecular structure to biological activity’, *Journal of Molecular Biology*. doi: 10.1016/j.jmb.2011.04.060.

Romano, M. *et al.* (2020) ‘A Structural View of SARS-CoV-2 RNA Replication Machinery: RNA Synthesis, Proofreading and Final Capping’, *Cells*. doi: 10.3390/cells9051267.

Saheb Sharif-Askari, N. *et al.* (2020) ‘Airways Expression of SARS-CoV-2 Receptor, ACE2, and TMPRSS2 Is Lower in Children Than Adults and Increases with Smoking and COPD’, *Molecular Therapy - Methods and Clinical Development*. doi: 10.1016/j.omtm.2020.05.013.

Sanyaolu, A. *et al.* (2020) ‘Comorbidity and its Impact on Patients with COVID-19’, *SN Comprehensive Clinical Medicine*. doi: 10.1007/s42399-020-00363-4.

Sarkar, M. and Saha, S. (2020) ‘Structural insight into the role of novel SARSCoV-2 E protein: A potential target for vaccine development and other therapeutic strategies’, *PLoS ONE*. doi: 10.1371/journal.pone.0237300.

Satarker, S. and Nampoothiri, M. (2020) ‘Structural Proteins in Severe Acute Respiratory Syndrome Coronavirus-2’, *Archives of Medical Research*. doi: 10.1016/j.arcmed.2020.05.012.

Schaecher, S. R., Mackenzie, J. M. and Pekosz, A. (2007) ‘The ORF7b Protein of Severe Acute Respiratory Syndrome Coronavirus (SARS-CoV) Is Expressed in Virus-Infected Cells and Incorporated into SARS-CoV Particles’, *Journal of Virology*. doi: 10.1128/jvi.01691-06.

Severe Acute Respiratory Syndrome (SARS) (no date). Available at: <https://www.who.int/health->

topics/severe-acute-respiratory-syndrome#tab=tab_1 (Accessed: 8 February 2021).

Shah, V. K. *et al.* (2020) ‘Overview of Immune Response During SARS-CoV-2 Infection: Lessons From the Past’, *Frontiers in Immunology*. doi: 10.3389/fimmu.2020.01949.

Shereen, M. A. *et al.* (2020) ‘COVID-19 infection: Origin, transmission, and characteristics of human coronaviruses’, *Journal of Advanced Research*. doi: 10.1016/j.jare.2020.03.005.

Shi, C.-S. *et al.* (2014) ‘SARS-Coronavirus Open Reading Frame-9b Suppresses Innate Immunity by Targeting Mitochondria and the MAVS/TRAF3/TRAF6 Signalingosome’, *The Journal of Immunology*, 193(6), pp. 3080–3089. doi: 10.4049/jimmunol.1303196.

Shilts, J. *et al.* (2021) ‘No evidence for basigin/CD147 as a direct SARS-CoV-2 spike binding receptor’, *Scientific Reports*, 11(1), p. 413. doi: 10.1038/s41598-020-80464-1.

Shu, T. *et al.* (2020) ‘SARS-Coronavirus-2 Nsp13 Possesses NTPase and RNA Helicase Activities That Can Be Inhibited by Bismuth Salts’, *Virologica Sinica*. doi: 10.1007/s12250-020-00242-1.

Sinha, S. and Vohora, D. (2017) ‘Drug Discovery and Development: An Overview’, in *Pharmaceutical Medicine and Translational Clinical Research*. doi: 10.1016/B978-0-12-802103-3.00002-X.

SMAD3 - Mothers against decapentaplegic homolog 3 - Homo sapiens (Human) - SMAD3 gene & protein (no date). Available at: <https://www.uniprot.org/uniprot/P84022> (Accessed: 11 April 2021).

Stawicki, S. P. *et al.* (2020) ‘... american college of academic international medicine-world academic council of emergency medicine multidisciplinary COVID-19 working group consensus ...’, *Journal of global ...*.

Stephens, D. S. and McElrath, M. J. (2020) ‘COVID-19 and the Path to Immunity’, *JAMA - Journal of the American Medical Association*. doi: 10.1001/jama.2020.16656.

Su, Y. C. F. *et al.* (2020) ‘Discovery and genomic characterization of a 382-nucleotide deletion in ORF7B and orf8 during the early evolution of SARS-CoV-2’, *mBio*. doi: 10.1128/mBio.01610-20.

Taefehshokr, N. *et al.* (2020) ‘Covid-19: Perspectives on Innate Immune Evasion’, *Frontiers in Immunology*. Frontiers Media S.A., p. 2549. doi: 10.3389/fimmu.2020.580641.

Teslya, A. *et al.* (2020) ‘Impact of self-imposed prevention measures and short-term government-imposed social distancing on mitigating and delaying a COVID-19 epidemic: A modelling study’, *PLoS Medicine*. doi: 10.1371/journal.pmed.1003166.

Thevarajan, I., Buising, K. L. and Cowie, B. C. (2020) ‘Clinical presentation and management of COVID-

19', *Medical Journal of Australia*. doi: 10.5694/mja2.50698.

Thomas, S. (2020) 'The structure of the membrane protein of sars-cov-2 resembles the sugar transporter semisweet', *Pathogens and Immunity*. doi: 10.20411/pai.v5i1.377.

Tissue expression of TMPRSS2 - Summary - The Human Protein Atlas (2020) *The Human Protein Atlas*. Available at: <https://www.proteinatlas.org/ENSG00000184012-TMPRSS2/tissue> (Accessed: 13 December 2020).

Tuerkova, A. and Zdrazil, B. (2020) 'A ligand-based computational drug repurposing pipeline using KNIME and Programmatic Data Access: case studies for rare diseases and COVID-19', *Journal of Cheminformatics*. doi: 10.1186/s13321-020-00474-z.

UniProt (no date). Available at: https://covid-19.uniprot.org/uniprotkb?query=* (Accessed: 8 April 2021).

UniProt - Wikipedia (no date). Available at: <https://en.wikipedia.org/wiki/UniProt> (Accessed: 7 April 2021).

V'kovski, P. *et al.* (2020) 'Coronavirus biology and replication: implications for SARS-CoV-2', *Nature Reviews Microbiology*. doi: 10.1038/s41579-020-00468-6.

Wang, D. *et al.* (2020) 'Clinical Characteristics of 138 Hospitalized Patients with 2019 Novel Coronavirus-Infected Pneumonia in Wuhan, China', *JAMA - Journal of the American Medical Association*. doi: 10.1001/jama.2020.1585.

Wang, H. *et al.* (2020) 'The genetic sequence, origin, and diagnosis of SARS-CoV-2', *European Journal of Clinical Microbiology and Infectious Diseases*. doi: 10.1007/s10096-020-03899-4.

Wang, Q. *et al.* (2020) 'Structural and Functional Basis of SARS-CoV-2 Entry by Using Human ACE2', *Cell*. doi: 10.1016/j.cell.2020.03.045.

Wang, S. *et al.* (2016) 'RaptorX-Property: a web server for protein structure property prediction', *Nucleic acids research*. doi: 10.1093/nar/gkw306.

Wang, Y. *et al.* (2020) 'Human SARS-CoV-2 has evolved to reduce CG dinucleotide in its open reading frames', *Scientific Reports*. doi: 10.1038/s41598-020-69342-y.

Wang, Y., Cheng, T. and Bryant, S. H. (2017) 'PubChem BioAssay: A Decade's Development toward Open High-Throughput Screening Data Sharing', *SLAS Discovery*. doi: 10.1177/2472555216685069.

Wassie, G. T. *et al.* (2020) 'Incubation Period of Severe Acute Respiratory Syndrome Novel Coronavirus 2 that Causes Coronavirus Disease 2019: A Systematic Review and Meta-Analysis', *Current Therapeutic*

Research - Clinical and Experimental. Excerpta Medica Inc., p. 100607. doi: 10.1016/j.curtheres.2020.100607.

What We Do – The Centre for Humanitarian Data (no date). Available at: <https://centre.humdata.org/what-we-do/> (Accessed: 7 April 2021).

WHO (2020) *WHO/Europe | Coronavirus disease (COVID-19) outbreak - About the virus*. Available at: <https://www.euro.who.int/en/health-topics/health-emergencies/coronavirus-covid-19/novel-coronavirus-2019-ncov> (Accessed: 13 December 2020).

Worldometers (2020) *Coronavirus Update (Live): 72,381,223 Cases and 1,615,845 Deaths from COVID-19 Virus Pandemic - Worldometer*. Available at: <https://www.worldometers.info/coronavirus/> (Accessed: 13 December 2020).

Xu, R. H. *et al.* (2004) 'Epidemiologic clues to SARS origin in China', *Emerging Infectious Diseases*. doi: 10.3201/eid1006.030852.

Yoshimoto, F. K. (2020) 'The Proteins of Severe Acute Respiratory Syndrome Coronavirus-2 (SARS CoV-2 or n-COV19), the Cause of COVID-19', *Protein Journal*. doi: 10.1007/s10930-020-09901-4.

Yu, I. T. S. *et al.* (2005) 'Temporal-spatial analysis of severe acute respiratory syndrome among hospital inpatients', *Clinical Infectious Diseases*. doi: 10.1086/428735.

Yuan, X. *et al.* (2005) 'Subcellular localization and membrane association of SARS-CoV 3a protein', *Virus Research*. doi: 10.1016/j.virusres.2005.01.001.

Yuen, C. K. *et al.* (2020) 'SARS-CoV-2 nsp13, nsp14, nsp15 and orf6 function as potent interferon antagonists', *Emerging Microbes and Infections*. doi: 10.1080/22221751.2020.1780953.

Zalzala, H. H. (2020) 'Diagnosis of COVID-19: facts and challenges', *New Microbes and New Infections*. Elsevier Ltd, p. 100761. doi: 10.1016/j.nmni.2020.100761.

Zhang, X. Y. *et al.* (2020) 'Biological, clinical and epidemiological features of COVID-19, SARS and MERS and AutoDock simulation of ACE2', *Infectious Diseases of Poverty*. doi: 10.1186/s40249-020-00691-6.

Zhu, G. *et al.* (2020) 'Minireview of progress in the structural study of SARS-CoV-2 proteins', *Current Research in Microbial Sciences*. doi: 10.1016/j.crmicr.2020.06.003.

Zietz, M., Zucker, J. and Tatonetti, N. P. (2020) 'Associations between blood type and COVID-19 infection, intubation, and death', *Nature Communications*. doi: 10.1038/s41467-020-19623-x.

7. Appendix

7.1. Python Script for Retrieving a World Map of COVID-19 Total Cases

```
3. import pycountry
import pandas as pd

# Aggregate the dataset
df_confirm = pd.read_csv('time_series_covid19_confirmed_global.csv')
df_confirm = df_confirm.drop(columns=['Province/State', 'Lat', 'Long'])
df_confirm = df_confirm.groupby('Country/Region').agg('sum')
date_list = list(df_confirm.columns)

# Get the three-letter country codes for each country
def get_country_code(name):
    try:
        return pycountry.countries.lookup(name).alpha_3
    except:
        return None
df_confirm['country'] = df_confirm.index
df_confirm['iso_alpha_3'] =
df_confirm['country'].apply(get_country_code)

# Transform the dataset in a long format
df_long = pd.melt(df_confirm, id_vars=['country', 'iso_alpha_3'],
value_vars=date_list)

import plotly.express as px

fig = px.choropleth(df_long,                                # Input
Dataframe
                    locations="iso_alpha_3",                # identify
country code column
                    color="value",                          # identify
representing column
                    hover_name="country",                  # identify hover
name
                    animation_frame="variable",            # identify date
column
                    projection="natural earth",              # select
projection
                    color_continuous_scale = 'Peach',        # select prefer
color scale
                    range_color=[0,500000]                  # select range
of dataset
                    )
fig.show()
fig.write_html("example_map.html")                          # write the
result to HTML file
```

7.2. R Scripts

Age Distribution Script

```
library(dplyr)

library(ggplot2)

austria_age_breakage <- read.csv("C://Users//Danilo//Documents//Danilo MT//R//age breakage austria.csv")

austria_age_plot <- austria_age_breakage %>%

  group_by(Age) %>%

  summarise(volume = sum(Number.of.Cases)) %>%

  mutate(share=volume/sum(volume)*100.0) %>%

  arrange(desc(volume))

austria_age_plot$Age <- factor(austria_age_plot$Age, levels = rev(as.character(austria_age_plot$Age)))

pie_chart_age_breakage <- ggplot(austria_age_plot, aes("", share, fill = Age)) +

  geom_bar(width = 2, size = 1, color = "black", stat = "identity") +

  coord_polar("y") +

  geom_text(aes(label = paste0(round(share), "%")),

    position = position_stack(vjust = 0.5)) +

  labs(x = NULL, y = NULL, fill = "Age",

    title = "AUSTRIA - COVID-19 Death Rate by Age 12.07.2020 ") +

  guides(fill = guide_legend(reverse = TRUE)) +

  scale_fill_manual(values=c("blue", "red", "grey", "green", "yellow", "purple", "white", "pink", "orange", "turquoise")) +

  theme_classic() +

  theme(axis.line = element_blank(),
```

```
axis.text = element_blank(),

axis.ticks = element_blank(),

plot.title = element_text(hjust = 0.5, color = "black", size = 13, face = "bold"),

legend.text = element_text(colour="black", size=10,

                             face="bold"),

legend.background = element_rect(fill="lightblue",

                                  size=0.5, linetype="solid",

                                  colour ="darkblue"))

pie_chart_age_breakage
```

Gender Distribution Script

```
library(dygraphs)

library(xts)

library(htmltools)

library(dplyr)

library(ggplot2)

austria_gender_breakage <- read.csv("C://Users//Danilo//Documents//Danilo MT//R//gender breakage austria.csv")

austria_gender_plot <- austria_gender_breakage %>%

  group_by(Gender) %>%

  summarise(volume = sum(Percentage)) %>%

  mutate(share=volume/sum(volume)*100.0) %>%

  arrange(desc(volume))
```

```

austria_gender_plot$Gender <- factor(austria_gender_plot$Gender, levels = rev(as.character(austria_gender_plot$Gender)))

pie_chart_gender_breakage <- ggplot(austria_gender_plot, aes("", share, fill = Gender)) +

  geom_bar(width = 2, size = 1, color = "black", stat = "identity") +

  coord_polar("y") +

  geom_text_repel(aes(label = paste0(round(share), "%")),

    position = position_stack(vjust = 0.5), size = 10) +

  labs(x = NULL, y = NULL, fill = "Gender",

    title = "AUSTRIA - COVID-19 Death Rate by Gender 15.01.2021 ") +

  guides(fill = guide_legend(reverse = TRUE)) +

  scale_fill_manual(values=c("blue", "red")) +

  theme_classic() +

  theme(axis.line = element_blank(),

    axis.text = element_blank(),

    axis.ticks = element_blank(),

    plot.title = element_text(hjust = 0.5, color = "black", size = 13, face = "bold"),

    legend.text = element_text(colour="black", size=10,

      face="bold"),

    legend.background = element_rect(fill="lightblue",

      size=0.5, linetype="solid",

      colour ="darkblue"))

pie_chart_gender_breakage

```

```
mortality_rate_sars_mers_covid19 <- read.csv("C:\\Users\\Danilo\\Documents\\Danilo MT\\R\\SARS, MERS and
COVID19\\Mortality Rate\\Mortality Rates.csv")
```

```
chart_mortality_rates <- ggplot(mortality_rate_sars_mers_covid19, aes(x = Mortality, y = Epidemic)) +

  geom_col(aes(fill = Epidemic), width = 0.9) +

  geom_text(aes(label = Mortality), size = 6, position = position_stack(vjust = 0.5)) +

  labs(x = "Mortality Rate %", y = "Epidemic", title = "Mortality Rate for SARS, MERS and COVID-19") +

  theme(axis.text = element_text(size = rel(1.2), colour = "red"),

        axis.title.y = element_text(size = rel(1.5), angle = 90, colour = "green"), axis.title.x = element_text(size = rel(1.3), colour =
"blue", face = "bold"), plot.title = element_text(size=15, face="bold"),

        legend.title = element_text(color = "red", size = 18),

        legend.text = element_text(color = "black", size = 14)) + xlim(0,100)
```

```
chart_mortality_rates
```

```
str(total_number_of_confirmed_cases)
```

Total Number of Cases/Deaths for SARS, MERS and COVID-19 Script

```
library(scales)
```

```
total_number_of_confirmed_cases <- read.csv("C:\\Users\\Danilo\\Documents\\Danilo MT\\R\\SARS, MERS and
COVID19\\Total Number of Cases\\Total Cases.csv")
```

```
chart_total_number_of_confirmed_cases <- ggplot(total_number_of_confirmed_cases, aes(x = No..of.Confirmed.Cases, y =  
Epidemic)) +  
  
  geom_col(aes(fill = Epidemic), width = 0.7) + scale_x_log10(name = "Number of Confirmed Cases (Log Values)", labels =  
comma) +  
  
  geom_text(aes(label = No..of.Confirmed.Cases), size = 6, position = position_stack(vjust = 0.5)) +  
  
  labs(title = "Total Number of Confirmed Cases for SARS, MERS and COVID-19 - Log Scale") +  
  
  theme(axis.text = element_text(size = rel(1.2), colour = "blue"),  
  
        axis.title.y = element_text(size = rel(1.5), angle = 90), axis.title.x = element_text(size = rel(1.3), colour = "red"), plot.title =  
element_text(size=15, face="bold"),  
  
        legend.title = element_text(color = "blue", size = 18),  
  
        legend.text = element_text(color = "black", size = 14))  
  
chart_total_number_of_confirmed_cases  
  
str(total_number_of_confirmed_cases)
```

Top 10 Countries for SARS, MERS and COVID-19 Script

```
install.packages("ggrepel")  
  
library(ggplot2)  
  
library(ggrepel)  
  
library(gridExtra)  
  
library(dplyr)
```

```
covid19_top10 <- read.csv("C:\\Users\\Danilo\\Documents\\Danilo MT\\R\\SARS, MERS and COVID19\\covid19 total number
of cases top 10.csv")
```

```
covid19_f <- covid19_top10 %>%
```

```
  group_by(Country) %>%
```

```
  summarise(volume = sum(Percentage)) %>%
```

```
  mutate(share=volume/sum(volume)*100.0) %>%
```

```
  arrange(desc(volume))
```

```
covid19_f$Country <- factor(covid19_f$Country, levels = rev(as.character(covid19_f$Country)))
```

```
pie_chart_covid19 <- ggplot(covid19_f, aes("", share, fill = Country)) +
```

```
  geom_bar(width = 2, size = 1, color = "black", stat = "identity") +
```

```
  coord_polar("y") +
```

```
  geom_text_repel(aes(label = paste0(round(share), "%")),
```

```
    position = position_stack(vjust = 0.75), max.overlaps = 3) +
```

```
  labs(x = NULL, y = NULL, fill = "Countries",
```

```
    title = "COVID-19 TOP 10 Countries (Total Number of Cases -> 95 555 805)") +
```

```
  guides(fill = guide_legend(reverse = TRUE)) +
```

```
  scale_fill_manual(values=c("pink", "grey", "turquoise", "red", "yellow", "green", "purple", "blue", "orange", "white", "brown"))
```

```
+
```

```
  theme_classic() +
```

```
  theme(axis.line = element_blank(),
```

```
    axis.text = element_blank(),
```

```

axis.ticks = element_blank(),

plot.title = element_text(hjust = 0.5, color = "black", size = 13, face = "bold"),

legend.text = element_text(colour="black", size=10,

                             face="bold"),

legend.background = element_rect(fill="lightblue",

                                   size=0.5, linetype="solid",

                                   colour ="darkblue"))

pie_chart_covid19

mers_top10 <- read.csv("C:\\Users\\Danilo\\Documents\\Danilo MT\\R\\SARS, MERS and COVID19\\MERS total number.csv")

mers_f <- mers_top10 %>%

  group_by(Country) %>%

  summarise(volume = sum(Percentage)) %>%

  mutate(share=volume/sum(volume)*100.0) %>%

  arrange(desc(volume))

mers_f$Country <- factor(mers_f$Country, levels = rev(as.character(mers_f$Country)))

pie_chart_mers <- ggplot(mers_f, aes("", share, fill = Country)) +

  geom_bar(width = 2, size = 1, color = "black", stat = "identity") +

  coord_polar("y") +

  geom_text_repel(aes(label = paste0(round(share), "%")),

```

```

position = position_stack(vjust = 0.75, max.overlaps = 3) +

labs(x = NULL, y = NULL, fill = "Countries",

title = "MERS TOP 10 Countries (Total Number of Cases -> 2 449)" +

guides(fill = guide_legend(reverse = TRUE)) +

scale_fill_manual(values=c("pink", "grey", "turquoise", "red", "yellow", "green", "purple", "blue", "orange", "white", "brown"))

+

theme_classic() +

theme(axis.line = element_blank(),

axis.text = element_blank(),

axis.ticks = element_blank(),

plot.title = element_text(hjust = 0.5, color = "black", size = 13, face = "bold"),

legend.text = element_text(colour="black", size=10,

face="bold"),

legend.background = element_rect(fill="lightblue",

size=0.5, linetype="solid",

colour ="darkblue"))

pie_chart_mers

sars_top10 <- read.csv("C:\\Users\\Danilo\\Documents\\Danilo MT\\R\\SARS, MERS and COVID19\\SARS total numbers.csv")

sars_f <- sars_top10 %>%

group_by(Country) %>%

summarise(volume = sum(Percentage)) %>%

```

```
mutate(share=volume/sum(volume)*100.0) %>%
```

```
arrange(desc(volume))
```

```
sars_f$Country <- factor(sars_f$Country, levels = rev(as.character(sars_f$Country)))
```

New cases VS New Deaths per Day Script

```
library(dygraphs)
```

```
library(xts)
```

```
library(htmltools)
```

```
library(dplyr)
```

```
library(ggplot2)
```

```
austria_age_breakage <- read.csv("C://Users//Danilo//Documents//Danilo MT//R//age breakage austria.csv")
```

```
austria_age_plot <- austria_age_breakage %>%
```

```
group_by(Age) %>%
```

```
summarise(volume = sum(Number.of.Cases)) %>%
```

```
mutate(share=volume/sum(volume)*100.0) %>%
```

```
arrange(desc(volume))
```

```
austria_age_plot$Age <- factor(austria_age_plot$Age, levels = rev(as.character(austria_age_plot$Age)))
```

```
pie_chart_age_breakage <- ggplot(austria_age_plot, aes("", share, fill = Age)) +

  geom_bar(width = 2, size = 1, color = "black", stat = "identity") +

  coord_polar("y") +

  geom_text_repel(aes(label = paste0(round(share), "%")),

    position = position_stack(vjust = 0.64), max.overlaps = 10, size = 6) +

  labs(x = NULL, y = NULL, fill = "Age",

    title = "AUSTRIA - COVID-19 Death Rate by Age 12.07.2020 ") +

  guides(fill = guide_legend(reverse = TRUE)) +

  scale_fill_manual(values=c("blue", "red", "grey", "green", "yellow", "purple", "white", "pink", "orange", "turquoise")) +

  theme_classic() +

  theme(axis.line = element_blank(),

    axis.text = element_blank(),

    axis.ticks = element_blank(),

    plot.title = element_text(hjust = 0.5, color = "black", size = 13, face = "bold"),

    legend.text = element_text(colour="black", size=10,

      face="bold"),

    legend.background = element_rect(fill="lightblue",

      size=0.5, linetype="solid",

      colour ="darkblue"))

pie_chart_age_breakage
```

Key Measures and New Cases for Austria Script

```
library(dygraphs)
```

```
library(xts)
```

```
measures_new_cases <- read.csv("C:\\Users\\Danilo\\Documents\\Danilo MT\\R\\new cases and deaths austria measures.csv")
```

```
measures_new_cases$time <- as.Date(measures_new_cases$Date, format = "%d-%b-%y")
```

```
str(measures_new_cases)
```

```
plot_measures_nc <- xts(x = measures_new_cases[,1], order.by = measures_new_cases$time)
```

```
dygraph(plot_measures_nc)
```

```
dygraph(plot_measures_nc, main = "COVID-19 Austria - New Cases and Key Measures ") %>%
```

```
  dySeries("V1", label = "AUSTRIA - New Cases") %>%
```

```
  dyOptions(stackedGraph = TRUE, colors = "red") %>%
```

```
  dyRangeSelector(height = 20)
```

```
measures_deaths <- read.csv("C:\\Users\\Danilo\\Documents\\Danilo MT\\R\\new cases and deaths austria measures.csv")
```

```
measures_deaths$time <- as.Date(measures_deaths$Date, format = "%d-%b-%y")
```

```
str(measures_deaths)
```

```
plot_measures_d <- xts(x = measures_deaths[,2], order.by = measures_deaths$time)
```

```
dygraph(plot_measures_d)
```

```
dygraph(plot_measures_d, main = "COVID-19 Austria - Deaths per day and Key Measures ") %>%

dySeries("V1", label = "AUSTRIA - Deaths per day") %>%

dyOptions(stackedGraph = TRUE, colors = "purple") %>%

dyRangeSelector(height = 20)
```

Tests per Day VS New Cases per Day Script

```
library(dygraphs)

library(xts)

options(scipen = 999)

tests_vs_new_cases_daily_austria <- read.csv("C:\\Users\\Danilo\\Documents\\Danilo MT\\R\\tests per day vs new cases\\Austria
tests per day vs new cases.csv")

tests_vs_new_cases_daily_austria$time <- as.Date(tests_vs_new_cases_daily_austria$Date, format = "%d-%b-%y")

str(tests_vs_new_cases_daily_austria)

t_vs_nc_austria <- xts(x = tests_vs_new_cases_daily_austria[,1:2], order.by = tests_vs_new_cases_daily_austria$time)

dygraph(t_vs_nc_austria)

dygraph(t_vs_nc_austria, main = "Austria - Tests Performed and New Cases - Daily") %>%

dySeries("Tests", label = "Performed Tests", drawPoints = TRUE, pointShape = "square", color = "blue") %>%
```

```
dySeries("New.Cases", label = "New Cases", stepPlot = TRUE, fillGraph = TRUE, color = "red") %>%
```

```
dyOptions(stackedGraph = TRUE) %>%
```

```
dyRangeSelector(height = 20)
```

Share of Positive Cases in Tested Patients Per Day Scripts

```
library(ggplot2)
```

```
library(lubridate)
```

```
library(scales)
```

```
library(gridExtra)
```

```
library(ggthemes)
```

```
library(tidyverse)
```

```
require(gridExtra)
```

```
sweden_share<- read.csv("C:\\Users\\Danilo\\Documents\\Danilo MT\\R\\tests and new cases share\\sweden.csv")
```

```
sweden_share$new.date <- as.Date(sweden_share$Date, format = "%d-%b-%y")
```

```
str(sweden_share)
```

```
plot_sweden_share <- ggplot() + geom_line(data = sweden_share, size=1, aes(x=new.date, y=Share), group=1, color="red") +
```

```
labs(x = "Date",y = "%", title = "Sweden - Share of Positive Cases in Tested Patients per day", size=5) +
```

```
theme(axis.line = element_line(size = 3, colour = "grey80"), axis.text = element_text(colour = "blue"),
```

```
axis.title.y = element_text(size = rel(2), angle = 90), axis.title.x = element_text(colour = "red", size = 15),
```

```
plot.title = element_text(size=15, face="bold"), legend.position = c(0.95, 0.95), legend.justification = c("right", "top")) +
```

```
ylim(0,100)
```

```
plot_sweden_share
```

```
grid.arrange(plot_austria_share, plot_serbia_share, plot_italy_share, plot_usa_share, plot_uk_share, plot_sweden_share, nrow = 2,
top = "SHARE of Positive Cases in Tested Patients per day")
```

Death Rate VS Recovery Rate Script

```
library(ggplot2)
```

```
library(lubridate)
```

```
library(scales)
```

```
library(gridExtra)
```

```
library(ggthemes)
```

```
library(tidyverse)
```

```
require(gridExtra)
```

```
usa_dr_vs_rr <- read.csv("C:\\Users\\Danilo\\Documents\\Danilo MT\\R\\Death rate VS Recovery rate\\USA DR and RR.csv")
```

```
usa_dr_vs_rr$new.date <- as.Date(usa_dr_vs_rr$Date, format = "%d-%b-%y")
```

```
str(usa_dr_vs_rr)
```

```
plot_usa_dr_vs_rr <- ggplot() + geom_line(data = usa_dr_vs_rr, size=1.5, aes(x=new.date, y=Recovery.Rate), group=1,
color="green") +
```

```
geom_line(data = usa_dr_vs_rr, size = 1.5, aes(x=new.date, y=Death.Rate), group=1, color="red") +
```

```
labs(x = "Date", y = "%", title = "USA - Outcome of Closed Cases (Recovery Rate VS Death Rate)", size=5) +
```

```
theme(axis.line = element_line(size = 3, colour = "grey80"), axis.text = element_text(colour = "blue"),
```

```
axis.title.y = element_text(size = rel(2), angle = 90), axis.title.x = element_text(colour = "red"),
```

```
plot.title = element_text(size=10, face="bold"), legend.position = c(0.95, 0.95), legend.justification = c("right", "top")) +
```

```
ylim(0,100)
```

plot_usa_dr_vs_rr