viruses

MDPI

Article

SARSMutOnto: An Ontology for SARS-CoV-2 Lineages and Mutations

Jamal Bakkas [1], Mohamed Hanine [2], Abderrahman Chekry [1], Said Gounane [3], Isabel de la Torre Díez [4,*], Vivian Lipari [5,6,7], Nohora Milena Martínez López [5,8,9] and Imran Ashraf [10,*]

1 LAPSSII Laboratory, Graduate School of Technology, Cadi Ayyad University, Safi 46000, Morocco
2 Department of Telecommunications, Networks, and Informatics, LTI Laboratory, ENSA, Chouaib Doukkali University, Eljadida 24000, Morocco
3 MIMSC Laboratory, Graduate School of Technology, Cadi Ayyad University, Essaouira 44000, Morocco
4 Department of Signal Theory and Communications and Telematic Engineering, University of Valladolid, Paseo de Belén, 15, 47011 Valladolid, Spain
5 Research Group on Foods, Nutritional Biochemistry and Health, Universidad Europea del Atlántico, Isabel Torres 21, 39011 Santander, Spain;
6 Department of Project Management, Universidad Internacional Iberoamericana Campeche, Mexico City 24560, Mexico
7 Fundación Universitaria Internacional de Colombia Bogotá, Bogotá 11001, Colombia
8 Research Group on Foods, Nutritional Biochemistry and Health Universidad Internacional Iberoamericana, Arecibo, PR 00613, USA
9 Research Group on Foods, Nutritional Biochemistry and Health Universidade Internacional do Cuanza, Cuito EN250, Angola
10 Department of Information and Communication Engineering, Yeungnam University, Gyeongsan 38541, Republic of Korea
* Correspondence: isator@tel.uva.es (I.d.l.T.D.); imranashraf@ynu.ac.kr (I.A.)

Abstract: Mutations allow viruses to continuously evolve by changing their genetic code to adapt to the hosts they infect. It is an adaptive and evolutionary mechanism that helps viruses acquire characteristics favoring their survival and propagation. The COVID-19 pandemic declared by the WHO in March 2020 is caused by the SARS-CoV-2 virus. The non-stop adaptive mutations of this virus and the emergence of several variants over time with characteristics favoring their spread constitute one of the biggest obstacles that researchers face in controlling this pandemic. Understanding the mutation mechanism allows for the adoption of anticipatory measures and the proposal of strategies to control its propagation. In this study, we focus on the mutations of this virus, and we propose the SARSMutOnto ontology to model SARS-CoV-2 mutations reported by Pango researchers. A detailed description is given for each mutation. The genes where the mutations occur and the genomic structure of this virus are also included. The sub-lineages and the recombinant sub-lineages resulting from these mutations are additionally represented while maintaining their hierarchy. We developed a Python-based tool to automatically generate this ontology from various published Pango source files. At the end of this paper, we provide some examples of SPARQL queries that can be used to exploit this ontology. SARSMutOnto might become a 'wet bench' machine learning tool for predicting likely future mutations based on previous mutations.

Keywords: ontology; genome structure; SARS-CoV-2; mutation; lineage

check for updates

Citation: Bakkas, J.; Hanine, M.; Chekry, A.; Gounane, S.; de la Torre Díez, I.; Lipari, V.; López, N.M.M.; Ashraf, I. SARSMutOnto: An Ontology for SARS-CoV-2 Lineages and Mutations. *Viruses* **2023**, *15*, 505. https://doi.org/10.3390/v15020505

Academic Editor: Hernan Garcia-Ruiz

Received: 13 November 2022
Revised: 4 February 2023
Accepted: 8 February 2023
Published: 11 February 2023

Copyright: © 2023 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (https://creativecommons.org/licenses/by/4.0/).

1. Introduction

Severe acute respiratory syndrome coronavirus (SARS-CoV) evolved from China in 2002, and Middle East respiratory syndrome coronavirus (MERS-CoV) originated in the Middle East in 2012 [1]. They are two coronavirus-type viruses of zoonotic origin that have already sounded the alarm about the dangers that this kind of virus can generate. Unlike its predecessors from the same coronavirus family, the severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2), which was discovered at the end of 2019 in Wuhan, China [2,3],

has spread across the world with exceptional speed. The COVID-19 disease caused by the SARS-CoV-2 virus, officially declared as a pandemic by the World Health Organization (WHO) in March 2020, continues to spread and affect individuals of all ages around the world. As of 15 January 2023, the WHO had reported more than 661 million confirmed cases, including 6.7 million deaths, with incalculable health, social, and economic costs [4]. Several vaccines were developed in record time. Massive vaccination campaigns have been launched, resulting in a significant reduction in mortality and hospitalization rates, especially among the elderly [5]. Nevertheless, the virus is still present and permanently evolving, so there is a need for any method, computational technique, or new tool that can be used to provide information about its evolution. A variety of research projects have been triggered related to this pandemic and all related topics, to understand it, predict its spread, and propose measures to besiege it. Bioinformatics and ontology engineering research projects have been contributing to this movement by proposing many ontologies to model domains related to the pandemic. Most of these ontologies are available through the National Center for Biomedical Ontology (NCBO BioPortal) [6]. Among these ontologies, we cite the Gene Ontology (GO) [7] proposed in 2000 and extended over the years. Its latest release (September 2022) contains 50,977 classes. This ontology describes the knowledge of the biological domain according to three aspects: molecular function, cellular component, and biological process. GO has a general purpose dealing with the management of information regarding genes and gene products amongst different species. The ontology provides a description of the genes, their relationships, and functions. After the success of GO, several ontologies have emerged over the past years. Disease Ontology (DO) is one of them, first appearing in 2012 [8]. It is considered the core for human disease semantic integration by providing a standardized ontology for describing disease terms. Its latest release (7 January 2022) describes the complexity of more than 10,000 human diseases. Having extensible content allowing for the knowledge sharing of new discoveries, DO has been at the origin of several other new ontologies and projects. Another important and specific ontology is the Infectious Disease Ontology (IDO) [9] which is designed as a set of interoperable ontologies covering the infectious diseases domain. These ontologies are built around the core ontology (IDO-Core), which provides a set of relevant entities to describe the clinical and biomedical aspects of infectious diseases. Since COVID-19 is an infectious disease caused by the coronavirus, the Coronavirus Infectious Disease Ontology (CIDO) [10] appeared as an extension of IDO and covers everything that depends on infection with the different coronaviruses and associated diseases. A more specific ontology called IDO-COVID-19 (COVID-19 Infectious Disease Ontology) [11] extends CIDO by describing the domain of infections with SARS-CoV-2 virus strains and related COVID-19 disease. Continuing with the COVID-19 disease domain, we cite the COVID-19 ontology [12] with 2270 classes for describing molecular and cellular entities and their roles in virus–host interactions and the virus life cycle, as well as a wide range of medical and epidemiological concepts related to COVID-19. Finally, we cite the OntoRepliCov ontology [13] that describes the genomic structure of the SARS-CoV-2 virus and the different steps of its replication process.

After a survey of existing ontologies, we noted that, so far, only the last CIDO ontology update included a limited number of terms for GISAID clades, Pango lineages, and WHO variants. It provides about 39 specific classes that describe specific SARS-CoV-2 variants [14]. However, we have not found any ontology describing all Pango lineages and mutations. Information about mutations and lineages is available and accessible on the internet, but scientists sometimes need more complex information that is not explicitly available. The SARS-Cov-2 variant accumulates mutations to produce new lineages. Data on these mutations and involved genes are tracked and reported by outbreak.info [15]. In this work, we retrieve these data and additional data from other sources and restructure them into an ontology. Called SARSMutOnto, this ontology provides a detailed description of the mutations and lineages reported by Pango scientists and researchers [16]. The SARSMutOnto ontology can be found on the bioportal at https://bioportal.bioontology.

org/ontologies/SARSMUTONTO and on GitHub at https://raw.githubusercontent.com/jbakkas/SARSMutOnto/main/SARSMutOnto.owl.

2. Background

2.1. SARS-CoV-2 Virus Structure

First time occurring, SARS-CoV-2 virus has been classified in the coronavirus family. Its genome structure corresponds to the specific genetic characteristics recognized for Coronaviruses. As detailed in [2,17,18], this genome is composed of two replicate proteins ORF1a, ORF1b, and four structural proteins: the spike protein (S), the envelope protein (E), the nucleocapsid proteins (N), and the membrane glycoprotein (M). Between these proteins, nine other proteins are distributed which are called accessory proteins ORF3a, ORF3b, ORF6, ORF7a, ORF7b, ORF8, ORF9a, ORF9b, and ORF10. Figures 1 and 2 show the sequence of the genes from the 5′-UTR end to the 3′-UTR end.

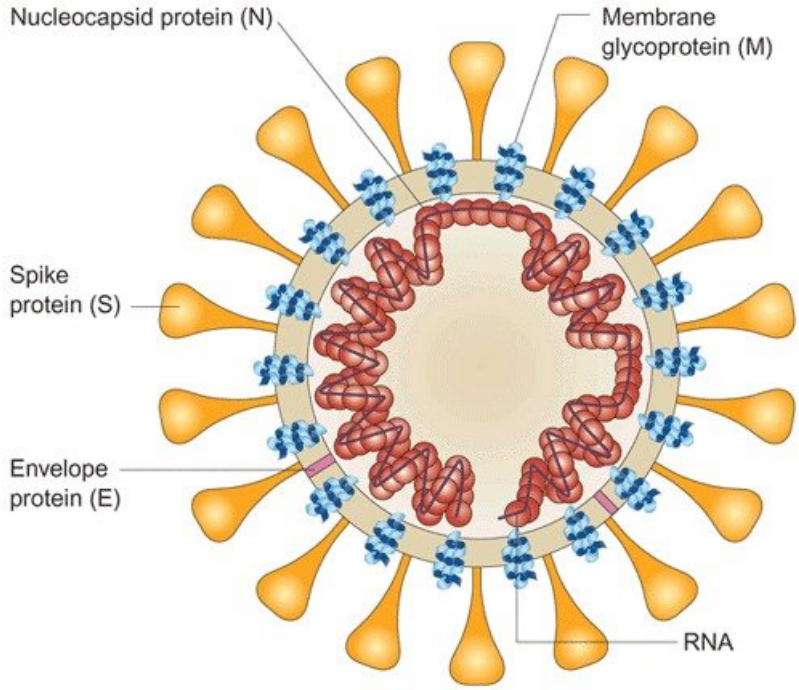

Figure 1. General structure of SARS-CoV [19].

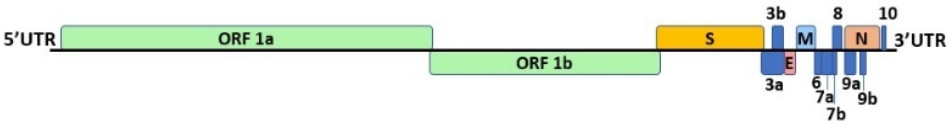

Figure 2. The set of genes making up the SARS-CoV 2 genome.

2.2. Mutation, Lineage, and Recombinant Virus

The SARS-CoV-2 virus, like any other virus, has mutated continuously since its emergence. A viral mutation is a change in the virus genome during the virus multiplication process. This change may have no impact and thus produce a neutral mutation, or it may have enough impact to develop another strain of the virus with different properties, or to a lesser extent to evolve a lineage. Lineages are related viruses descended from common ancestors. Most lineages disperse naturally. Natural selection always favors viral lineages that have acquired characteristics that allow them to survive more easily. Some mutations give the virus a certain evolutionary advantage, which may make it more contagious, fiercer, or more resistant to the immune system and vaccines. A lineage with such mutations becomes more contagious than others, and therefore more dominant. When several variants infect the same cell and, during multiplication processes, a hybrid genome results from portions

of several genomes, then a so-called 'recombinant' virus is produced. Viral recombination is favored if there is a large circulation of virus variants.

2.3. Lineage Nomenclatures

Because of its wide spread, SARS-CoV-2 mutations have given rise to thousands of lineages from which several variants have emerged worldwide, with different characteristics from one variant to another. In early 2021, The most popular variants were the British variant [20], the Indian variant [21], the South African variant [22], and the Brazilian variant [23]. In the media and among the general public, the variants were usually designated by their original country name. The WHO has renamed the most widespread variants with Greek letters [24], to have names that are easy to remember, but also to avoid the stigma of names designating the countries in which the variant first appeared. For example, the WHO gave the name 'Alpha' to the variant known in the media as the 'British variant', and 'Beta' to the variant known as 'South Africa', etc. In reality, the scientists assign names not only to variants but also to all reported lineages. The main nomenclatures available are those proposed by Pango, GISAID [25], and Nextstrain. While the GISAID [26] and Nextstrain [27] nomenclatures provide an overview of clade trends; the Pango nomenclatures offer detailed information on lineages, allowing early prediction of local lineage expansion. For example, Pango assigned 'B.1.1.7' to the 'Alpha' variant, while GISAID used 'GR/501Y.V1', and Nextstrain used the name '20I/S:501Y.V1'. Pango researchers propose a lineage naming algorithm and the Pangolin lineage naming tool that dynamically attributes names to lineages; the latter being available both as a web application and as a command line tool [28,29].

2.4. SARS-CoV-2 Mutations

During the replication process, new mutations may occur anywhere in any gene composing the virus genome. The mutation becomes interesting when it confers additional characteristics to this virus. Most researchers have linked the SARS-CoV-2 propagation speed to mutations detected in the spike protein (S) gene. Indeed, the virus relies on this protein and more specifically on the receptor-binding domain (RBD) of this protein to bind to lung cell surface receptors (ACE2) when entering the host cell [30,31]. The S protein is the main target of antibodies generated either by the natural reaction of an infected human body or by vaccination. Mutations occurring in this area influenced the virus' ability to enter cells by increasing or decreasing the efficiency of binding to the ACE2 receptor [32,33]. Lineages resulting from these mutations develop characteristics that allow increased contagiousness and evasion of cellular immunity [34]. Examples of these lineages include B.1.427/B.1.429 [35] assigned by the WHO with the Greek letter Epsilon. Other more well-known and dominant variants causing successive waves of the pandemic around the world include the Alpha, Beta, Gamma, and Delta variants. The fifth wave of the pandemic was triggered on 26 November 2021, when the WHO announced the appearance of the Omicron variant (B.1.1.529). In this variant, a very high number (32) of mutations are found in the spike protein S, compared to its devastating predecessor Delta, which has only five mutations.

Mutations in other genes, other than the spike protein, have also been the focus of study. One of these studies links mutations in the gene encoding ORF3a accessory protein to increased mortality rates [36]. We noted that in order to study and understand the characteristics of a lineage, and to predict the behavior of future variants, it is necessary to study the mutations they have undergone, especially recurrent mutations. Through this work, we present an ontology that gathers all mutations reported by researchers in great detail since the first strain of the virus was found.

3. Materials and Methods

In this study, we present a lightweight ontology specifically designed to describe lineages and associated mutations in detail. Since the SARS-CoV-2 virus continues to

evolve, the study also provides a tool to automatically regenerate updated versions. Our data sources are mainly Pango files and outbreak.info API [15]. The workflow diagram of the approach followed in this study is shown in Figure 3.

Figure 3. Workflow diagram of the steps followed in this study.

New lineages are identified around the world every day and are reported regularly by Pango. To keep our ontology up to date, we have developed the SARSMutOnto generator tool shown in Figure 4. This tool allows the automatic generation of the updated release of SARSMutOnto.

The entries of this system, as shown in the workflow diagram in Figure 3, are of two types. The Pango files and the API are provided by outbreak.info. The files are parsed to extract information about the lineages and their hierarchy. The information includes Pango-assigned lineage name, direct ancestor name or ancestor names if a recombinant, WHO-assigned name, lineage description, and alias; the aliases are retrieved using the 'pango_aliasor' Python library available via the following link: 'https://github.com/corneliusroemer/pango_aliasor'. The API is consulted to retrieve the details of the mutations for each lineage. The API provides the mutation and the gene where it occurred.

The generation process starts with the creation of the ontology's general structure. This structure is composed of the following classes: 'SARS-CoV-2', 'variant', 'lineage', 'recombinant', 'genome', 'gene', 'structural_gene', 'non_structural_gene', 'accessory_gene','mutation', and 'SNP'. Then, we create the individuals that represent the fifteen genes comprising the genome. These components are linked to each other by inheritance relationships and by object properties as shown in Figure 5. The retrieved entries are then used to generate classes, individuals representing lineages, and mutations to the ontology using the Owlready2 [37] Python package dedicated to ontology-oriented programming.

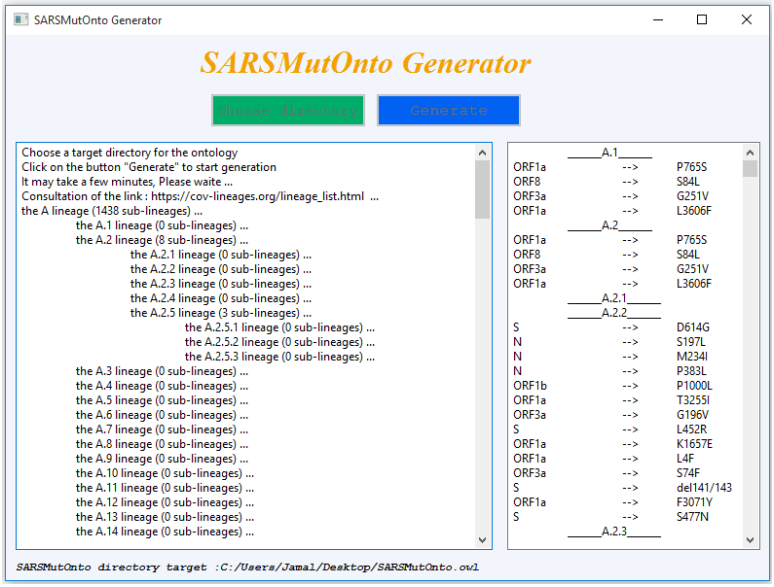

Figure 4. SARSMutOnto Generator: tool used to generate SARSMutOnto.

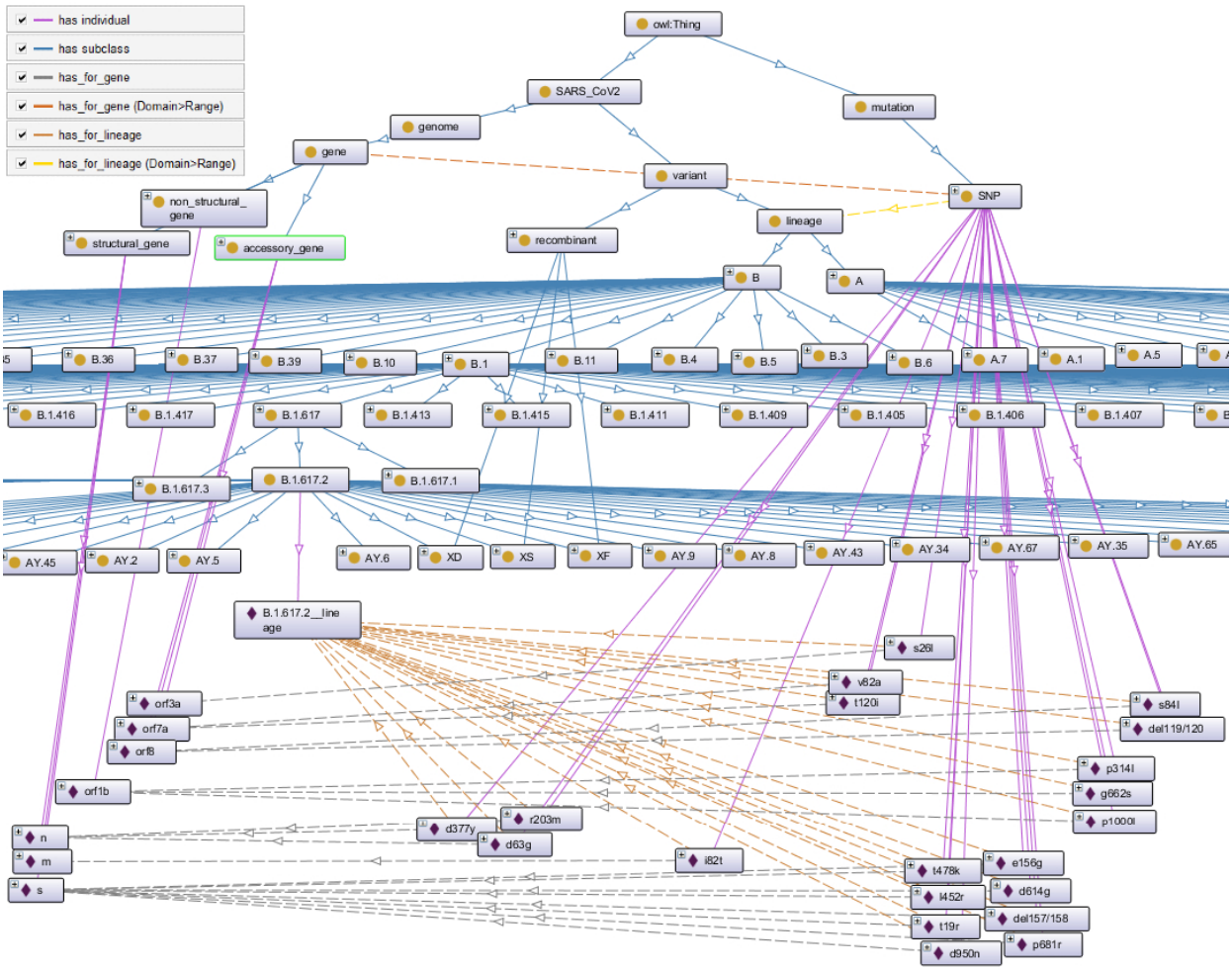

Figure 5. Portion of the ontology that describes the Delta (B.1.617.2) variant.

The steps followed by the SARSMutOnto Generator tool to generate the ontology are shown in Algorithm 1 using the pseudo-code. The graphical interface of the tool is provided in Figure 4. It is divided into two sections; the first one on the left displays the progress of the lineages, and the second one on the right displays the mutations extracted

from the outbreak.info API for each lineage. It is developed using the 3.10.0 Python version. The tool's source code is open and available on GitHub via the following link: https://github.com/jbakkas/SMOGenerator.

Algorithm 1 Ontology generation algorithm

Inputs: Pango fils and outbreak.info API

```
Ontology initialization
Retrieve lineage list from lineage.yml
Retrieve alias from alias_key.json using pango_aliasor
Retrieve descriptions from lineage_notes.txt
for Each lineage do
    Create corresponding ontology class
    Connect the class to its direct ancestor(s)
    if a recombinant lineage connect to 'recombinant' class by 'is_a' link
    Create individual from class
    Add alias, if exists, by assertion of 'has_for_description' ataTypeProperty
    Add description by assertion of 'has_for_description' ataTypeProperty
    Extract list of lineage mutations and their genes
    for Each mutation do
        Create an individual of 'SNP' class
        Connect mutation to gene by assertion of 'has_for_gene' ObjectProperty
        Connect mutation to lineage by assertion of 'has_for_lineage' ObjectProperty
    end for
end for
```

Output: SARSMutOnto ontology

4. Results

4.1. The Proposed Ontology

Biomedical ontologies, and especially those that extend the ″GO″ ontology, generally describe biological domains in three aspects; cellular component, molecular function, and biological process. The gene product encoded by a gene performs an elementary low-level activity. This activity is described by ontologies as (molecular function). This activity occurs in a specific location of the cell. This location is modeled as a (cellular component). The elementary activities cooperate to perform a more general and larger scale activity described as a (biological process). The cellular component is the most visible aspect of the proposed ontology. Indeed, SARSMutOnto describes the genome of the SARS-CoV-2 virus as well as its different component genes. The activities are not the focus of this study. According to its type, each of the 15 genes that make up the genome is presented as an individual of one of the three classes designating the 'structural_gene', 'non-structural_gene', and 'accessory_gene' gene types, as shown in Figure 5.

Lineages are represented by classes inheriting directly or indirectly from the superclass 'lineage'. The first two lineages A and B are represented by the 'A' and 'B' classes, which inherit directly from the 'lineage' class. The other lineages are linked to each other and to 'A' and 'B' by hierarchical links. The class that represents a given lineage is a subclass of the class representing the ancestor of this lineage and is the superclass of all classes representing its descendants, which allows the hierarchical relationship (lineage/sub-lineage) between lineages to be maintained. A class representing a recombinant lineage inherits the 'recombinant' class and all classes representing its parent lineages. Each class representing a lineage has a corresponding individual that provides the information characterizing this lineage by 'dataTypeProperty' assertion. This information includes a brief lineage description, an alias if available, and the first appearance date, as well as all the mutations produced with respect to the first strain of the virus. Each mutation is represented by an individual of the 'SNP' class. It is linked to the lineage, in which it occurs by the assertion of the ObjectProperty 'has_for lineage', and to the affected gene by the

assertion of the ObjectProperty 'has_for gene'. The SARSMutOnto ontology consists of 2206 classes and 2886 individuals and is available via Bioportal, the repository of biomedical ontologies.

4.2. Lineage Description

The taxonomy of classes proposed by SARSMutOnto illustrates the hierarchy of all lineages. It allows us to represent the phylogenetic tree of the Pango lineages. Thus, for a given lineage, all ancestors and descendants of this tree can be obtained and, therefore, all mutations that led to the appearance of each lineage can be obtained too. For example, the Delta variant also called B.1.617.2, triggered the fourth wave of the pandemic. This lineage emerged as a result of a succession of mutations from the B lineage, one of the two earliest observed strains of the virus: First the emergence of the lineage B.1 with the following mutations: S(d614g), ORF1b(P314L), and ORF855s84l), then B.1.617 with the mutations S(L452R, D614G, P681R), ORF1B(P314L), ORF3a(S26L), ORF7a(V82A), ORF8(S84L), and N(R203M, D377Y). Finally, B.1.617.2, which underwent mutations S(T19R, E156G, del157/158, L452R, T478K, D614G, P681R, D950N), ORF1B(P314L, G662S, P1000L), ORF3a(S26L), M(I82T), ORF7a(V82A, T120I), ORF8(S84L, del119/120), and N(D63G, R203M, D377Y) (Table 1). For this example, as we can see in the SARSMutOnto segment in Figure 5, the ancestor hierarchy of B.1.617.2 is represented by classes. Each class is linked to its direct ancestor by the 'is_a' relationship. Mutations are represented by individuals of the 'SNP' class. These individuals are linked to the lineage by assertions of the 'has_for_lineage' objectProperty, and to the concerned gene by the assertions of the 'has_for_gene' objectProperty. The classes representing the recombinant lineages inherit directly from the 'recombinant' class, in addition to the classes representing their parents. Hence, all classes representing recombinant lineages are descendants of the 'recombinant' class.

Table 1. Mutations affecting B.1.617.2 lineage [38]. ORF: open reading frames gene.

Gene	Amino Acid
Spike protein gene	T19R
Spike protein gene	E156G
Spike protein gene	del157/158
Spike protein gene	L452R
Spike protein gene	T1T478K9R
Spike protein gene	D614G
Spike protein gene	P681R
Spike protein gene	D950N
Nucleocapsid gene	D63G
Nucleocapsid gene	R203M
Nucleocapsid gene	D377Y
ORF7a	V82A
ORF8a	S84L
ORF8a	del119/120
ORF1b	P314L
ORF1b	P1000L
ORF1b	G662S

4.3. Querying SARSMutOnto

The SARSMutOnto ontology allows us to easily find the list of mutations associated with each lineage. It can be used by biologists or virologists to extract different types of information about SARS-CoV-2 mutations, for example, the list of all ancestors of a variant, the list of mutations that have been located in a given gene for all variants combined, the list of all lineages with a given mutation, the list of variants with a name assigned by the WHO, etc. Hereafter, some examples of SPARQL queries interrogating the SARSMutOnto ontology performed with the SPARQL language, using the twinkle tool (http://ldodds.com/projects/twinkle/, accessed on 20 January 2023) are given. More examples are available in Appendix A.

4.3.1. List of Lineage Mutations

The first example aims at extracting a given lineage, the list of mutations, as well as the genes concerned by these mutations. The query in Listing 1 allows the mutations of the B.1.617.2 variant and the genes where they occur to be extracted.

Listing 1. Query to extract the mutation list of B.1.617.2 variant.

```
PREFIX ns:<https://github.com/jbakkas/SARSMutOnto/blob/main/SARSMutOnto.owl#>
SELECT ?mutationName ?gene
FROM <https://raw.githubusercontent.com/jbakkas/SARSMutOnto/main/SARSMutOnto.owl>
WHERE{
  ?mutation   a   ns:SNP.
  ?lineage    a   ns:B.1.617.2.
  ?mutation    ns:has_for_lineage  ?lineage.
  ?mutation     ns:has_for_gene      ?gene .
  ?mutation   ns:mutation_name   ?mutationName
}ORDER BY DESC(?gene)
```

4.3.2. List of Lineages with a Given Mutation

A mutation can endow the virus with a specific characteristic that can cause severe forms of disease or make it more resistant to human immunity or a vaccine. To more effectively reduce the spread of such mutations, scientists need a list of all variants or lineages that carry this mutation. The following example, Listing 2, is a SPARQL query to retrieve the list of lineages containing the 'N501Y' mutation of the Omicron variant. sub-lineages with this mutation are more infectious and dangerous for patients with cancer [39].

Listing 2. Query to extract lineages with a given mutation.

```
PREFIX ns:<https://github.com/jbakkas/SARSMutOnto/blob/main/SARSMutOnto.owl#>
SELECT ?lineageName
FROM <https://raw.githubusercontent.com/jbakkas/SARSMutOnto/main/SARSMutOnto.owl>
WHERE{
  ?snp   a   ns:SNP.
  ?snp   ns:has_for_lineage   ?lineage.
  ?lineage  ns:label   ?lineageName.
  Filter(?snp=ns:N501Y)
}ORDER BY ?lineageName
```

4.3.3. List of Gene Mutations

Further useful information for biologists is the list of mutations that have occurred in a given gene since the virus first appeared, in particular, the list of mutations that have occurred in the spike protein S gene encoding the surface protein. The query in Listing 3 allows the extraction of the list of all mutations that have occurred in the spike protein gene S, all variants combined.

Listing 3. Query to extract the list of all mutations occurring in the spike (S) protein.

```
PREFIX owl:<http://www.w3.org/2002/07/owl#>
PREFIX ns:<https://github.com/jbakkas/SARSMutOnto/blob/main/SARSMutOnto.owl#>
SELECT ?mutationName ?gene
FROM <https://raw.githubusercontent.com/jbakkas/SARSMutOnto/main/SARSMutOnto.owl>
WHERE{
  ?mutation  a                  ns:SNP.
  ?mutation  a                  owl:NamedIndividual.
  ?mutation  ns:has_for_gene    ns:S.
  ?mutation  ns:mutation_name   ?mutationName.
}
```

4.3.4. List of Recombinant Lineages

The wide circulation of various variants of the virus generates recombinant lineages. Using a simple SPARQL query, we can list all Pango lineages resulting from a recombinant mutation of SARS-CoV-2. The query in Listing 4 returns a list of all Pango recombinant lineages with each lineage's parents.

Listing 4. Query to extract the list of all Pango reconbinant lineages with their parents.

```
PREFIX ns:<https://github.com/jbakkas/SARSMutOnto/blob/main/SARSMutOnto.owl#>
PREFIX rdfs:<http://www.w3.org/2000/01/rdf-schema#>
PREFIX rdf:<http://www.w3.org/2000/01/rdf-schema#>
PREFIX owl:<http://www.w3.org/2002/07/owl#>
SELECT ?l ?c
FROM <https://raw.githubusercontent.com/jbakkas/SARSMutOnto/main/SARSMutOnto.owl>

WHERE {
        ?l rdf:subClassOf  ns:recombinant.
        ?l rdf:subClassOf  ?c.
        Filter(?c!=ns:recombinant)
}
```

5. Conclusions

This paper contributes extensively to the design and implementation of a novel ontology called SARSMutOnto. This ontology is designed to describe the SARS-CoV-2 lineages and mutations. It provides the concepts and semantic entities necessary for studies and research that deal with mutations and variants of the SARS-CoV-2 virus. It is intended primarily for use by semantic interoperability approaches or for text mining in the SARS-CoV-2 domain. As this virus is in continuous mutation, lineages and even variants will constantly appear; the updated release of the ontology can be generated thanks to the aforementioned SARSMutOnto generator tool. The updated ontology release can be found at the Bioportal portal. Future research in this area will use machine learning techniques to predict possible future mutations of the SARS-CoV-2 virus based on the SARSMutOnto ontology. As part of our ongoing effort to harmonize ontologies, we will keep working to bring together different COVID-19-related ontologies. To manage the description of coronaviral variations, we will keep updating our ontology. Additionally, we will explore and design more applications that use this ontology.

Author Contributions: Conceptualization, J.B. and S.G.; Data curation, S.G. and V.L.; Formal analysis, A.C., V.L. and N.M.M.L.; Funding acquisition, I.d.l.T.D.; Investigation, J.B. and A.C.; Methodology, M.H.; Project administration, M.H., I.d.l.T.D. and N.M.M.L.; Resources, I.d.l.T.D.; Software, S.G. and A.C.; Supervision, I.A.; Validation, V.L., N.M.M.L., and I.A.; Visualization, M.H.; Writing—original draft, J.B. and M.H.; Writing—review and editing, M.H. and I.A. All authors have read and agreed to the published version of the manuscript.

Funding: This research was supported by the European University of the Atlantic.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The datasets generated during and/or analysed during the current study are publicly available at https://bioportal.bioontology.org/ontologies/SARSMUTONTO/?p=summary.

Conflicts of Interest: The authors declare no conflict of interests.

Appendix A. Additional Queries

This appendix presents some examples of SPARQL queries tested on the SARSMutOnto ontology. These queries are executed using the twinkle tool available via this link (http://ldodds.com/projects/twinkle/).

Appendix A.1. Query 1

List of all lineages with a description and date of appearance for each lineage.

Listing A1. List of all lineages.

```
PREFIX ns:<https://github.com/jbakkas/SARSMutOnto/blob/main/SARSMutOnto.owl#>
PREFIX owl:<http://www.w3.org/2002/07/owl#>
prefix  rdfs:     <http://www.w3.org/2000/01/rdf-schema#>
SELECT ?lineageName      ?date  ?description
FROM <https://raw.githubusercontent.com/jbakkas/SARSMutOnto/main/SARSMutOnto.owl>
WHERE{
  ?lineage      a                       owl:NamedIndividual.
  ?lineage      ns:label                ?lineageName.
  ?lineage      ns:appeared_on       ?date.
  ?lineage      ns:has_for_description         ?description.
}order by (?lineageName)
```

Appendix A.2. Query 2

Date of appearance of a given lineage: the date of appearance of the Delta variant (B.1.617.2) for example.

Listing A2. The appearance date of a Delta variant (B.1.617.2).

```
PREFIX ns:<https://github.com/jbakkas/SARSMutOnto/blob/main/SARSMutOnto.owl#>
PREFIX owl:<http://www.w3.org/2002/07/owl#>
prefix  rdfs:     <http://www.w3.org/2000/01/rdf-schema#>
SELECT ?lineageName     ?date
FROM <https://raw.githubusercontent.com/jbakkas/SARSMutOnto/main/SARSMutOnto.owl>
WHERE{
  ?lineageI     a                       owl:NamedIndividual.
  ?lineageI      ns:label               ?lineageName.
  ?lineageI       ns:appeared_on                ?date
  FILTER(?lineageName=``B.1.617.2'')
}
```

Appendix A.3. Query 3

Lineages with OMS-assigned names. These are generally the most common variants.

Listing A3. List of lineages with WHO-assigned names.

```
PREFIX ns:<https://github.com/jbakkas/SARSMutOnto/blob/main/SARSMutOnto.owl#>
PREFIX owl:<http://www.w3.org/2002/07/owl#>
prefix  rdfs:     <http://www.w3.org/2000/01/rdf-schema#>
SELECT ?lineageName      ?WHO_name
FROM <https://raw.githubusercontent.com/jbakkas/SARSMutOnto/main/SARSMutOnto.owl>
WHERE{
  ?lineage      a                       owl:NamedIndividual.
  ?lineage      ns:label                ?lineageName.
  ?lineage      ns:has_for_WHO_name     ?WHO_name .
  FILTER(?WHO_name!='')
}
```

Appendix A.4. Query 4

Names of all sub-lineages of a given lineage. For example, the sub-lineages of the lineage (B.1.1.529) which represents the Omicron variant.

Listing A4. Sub-lineages of the lineage (B.1.1.529).

```
PREFIX ns:<https://github.com/jbakkas/SARSMutOnto/blob/main/SARSMutOnto.owl#>
PREFIX owl:<http://www.w3.org/2002/07/owl#>
prefix  rdfs:     <http://www.w3.org/2000/01/rdf-schema#>
SELECT ?name
FROM <https://raw.githubusercontent.com/jbakkas/SARSMutOnto/main/SARSMutOnto.owl>
WHERE{
  ?subLineage   rdfs:subClassOf         ns:B.1.1.529.
  ?ind          a               owl:NamedIndividual.
  ?ind          a               ?subLineage.
  ?ind          ns:label        ?name
}
```

References

1. Cui, J.; Li, F.; Shi, Z.L. Origin and evolution of pathogenic coronaviruses. *Nat. Rev. Microbiol.* **2019**, *17*, 181–192. [CrossRef] [PubMed]
2. Wu, F.; Zhao, S.; Yu, B.; Chen, Y.M.; Wang, W.; Song, Z.G.; Hu, Y.; Tao, Z.W.; Tian, J.H.; Pei, Y.Y.; et al. A new coronavirus associated with human respiratory disease in China. *Nature* **2020**, *579*, 265–269. [CrossRef] [PubMed]
3. Lu, H.; Stratton, C.W.; Tang, Y.W. Outbreak of pneumonia of unknown etiology in Wuhan, China: The mystery and the miracle. *J. Med. Virol.* **2020**, *92*, 401. [CrossRef] [PubMed]
4. WHO. WHO Coronavirus (COVID-19) Dashboard. 2023. Available online: https://covid19.who.int/ (accessed on 15 January 2023).
5. Moghadas, S.M.; Vilches, T.N.; Zhang, K.; Wells, C.R.; Shoukat, A.; Singer, B.H.; Meyers, L.A.; Neuzil, K.M.; Langley, J.M.; Fitzpatrick, M.C.; et al. The impact of vaccination on coronavirus disease 2019 (COVID-19) outbreaks in the United States. *Clin. Infect. Dis.* **2021**, *73*, 2257–2264. [CrossRef]
6. Whetzel, P.L.; Noy, N.F.; Shah, N.H.; Alexander, P.R.; Nyulas, C.; Tudorache, T.; Musen, M.A. BioPortal: Enhanced functionality via new Web services from the National Center for Biomedical Ontology to access and use ontologies in software applications. *Nucleic Acids Res.* **2011**, *39*, W541–W545. [CrossRef]
7. Ashburner, M.; Ball, C.A.; Blake, J.A.; Botstein, D.; Butler, H.; Cherry, J.M.; Davis, A.P.; Dolinski, K.; Dwight, S.S.; Eppig, J.T.; et al. Gene ontology: Tool for the unification of biology. *Nat. Genet.* **2000**, *25*, 25–29. [CrossRef]
8. Schriml, L.M.; Arze, C.; Nadendla, S.; Chang, Y.W.W.; Mazaitis, M.; Felix, V.; Feng, G.; Kibbe, W.A. Disease Ontology: A backbone for disease semantic integration. *Nucleic Acids Res.* **2012**, *40*, D940–D946. [CrossRef]
9. Cowell, L.G.; Smith, B. Infectious disease ontology. In *Infectious Disease Informatics*; Springer: Berlin/Heidelberg, Germany, 2010; pp. 373–395.
10. He, Y.; Yu, H.; Ong, E.; Wang, Y.; Liu, Y.; Huffman, A.; Huang, H.h.; Beverley, J.; Hur, J.; Yang, X.; et al. CIDO, a community-based ontology for coronavirus disease knowledge and data integration, sharing, and analysis. *Sci. Data* **2020**, *7*, 181. [CrossRef]
11. Babcock, S.; Beverley, J.; Cowell, L.G.; Smith, B. The infectious disease ontology in the age of COVID-19. *J. Biomed. Semant.* **2021**, *12*, 13. [CrossRef]
12. Sargsyan, A.; Kodamullil, A.T.; Baksi, S.; Darms, J.; Madan, S.; Gebel, S.; Keminer, O.; Jose, G.M.; Balabin, H.; DeLong, L.N.; et al. The COVID-19 ontology. *Bioinformatics* **2020**, *36*, 5703–5705. [CrossRef]

13. Laddada, W.; Soualmia, L.F.; Zanni-Merk, C.; Ayadi, A.; Frydman, C.; Imbert, I. OntoRepliCov: An Ontology-Based Approach for Modeling the SARS-CoV-2 Replication Process. *Procedia Comput. Sci.* **2021**, *192*, 487–496. [CrossRef] [PubMed]
14. He, Y.; Yu, H.; Huffman, A.; Lin, A.Y.; Natale, D.A.; Beverley, J.; Zheng, L.; Perl, Y.; Wang, Z.; Liu, Y.; et al. A comprehensive update on CIDO: The community-based coronavirus infectious disease ontology. *J. Biomed. Semant.* **2022**, *13*, 25. [CrossRef] [PubMed]
15. Gangavarapu, K.; Latif, A.A.; Mullen, J.L.; Alkuzweny, M.; Hufbauer, E.; Tsueng, G.; Haag, E.; Zeller, M.; Aceves, C.M.; Zaiets, K.; et al. Outbreak.info genomic reports: Scalable and dynamic surveillance of SARS-CoV-2 variants and mutations. *medRxiv* **2022.** [CrossRef]
16. PANGO. PANGO Lineages. 2021. Available online: https://cov-lineages.org (accessed on 13 September 2022).
17. Zhang, Q.; Xiang, R.; Huo, S.; Zhou, Y.; Jiang, S.; Wang, Q.; Yu, F. Molecular mechanism of interaction between SARS-CoV-2 and host cells and interventional therapy. *Signal Transduct. Target. Ther.* **2021**, *6*, 233. [CrossRef]
18. Wu, A.; Peng, Y.; Huang, B.; Ding, X.; Wang, X.; Niu, P.; Meng, J.; Zhu, Z.; Zhang, Z.; Wang, J.; et al. Genome composition and divergence of the novel coronavirus (2019-nCoV) originating in China. *Cell Host Microbe* **2020**, *27*, 325–328. [CrossRef]
19. Peiris, J.S.; Guan, Y.; Yuen, K. Severe acute respiratory syndrome. *Nat. Med.* **2004**, *10*, S88–S97. [CrossRef]
20. Tang, J.W.; Tambyah, P.A.; Hui, D.S. Emergence of a new SARS-CoV-2 variant in the UK. *J. Infect.* **2021**, *82*, e27–e28. [CrossRef]
21. Kirola, L. Genetic emergence of B. 1.617. 2 in COVID-19. *New Microbes New Infect.* **2021**, *43*, 100929. [CrossRef]
22. Tegally, H.; Wilkinson, E.; Giovanetti, M.; Iranzadeh, A.; Fonseca, V.; Giandhari, J.; Doolabh, D.; Pillay, S.; San, E.J.; Msomi, N.; et al. Emergence and rapid spread of a new severe acute respiratory syndrome-related coronavirus 2 (SARS-CoV-2) lineage with multiple spike mutations in South Africa. *MedRxiv* **2020**. [CrossRef]
23. Voloch, C.M.; da Silva Francisco Jr, R.; de Almeida, L.G.; Cardoso, C.C.; Brustolini, O.J.; Gerber, A.L.; Guimarães, A.P.d.C.; Mariani, D.; da Costa, R.M.; Ferreira, O.C., Jr.; et al. Genomic characterization of a novel SARS-CoV-2 lineage from Rio de Janeiro, Brazil. *J. Virol.* **2021**, *95*, e00119-21. [CrossRef]
24. (WHO), W.H.O. Tracking SARS-CoV-2 Variants. 2020. Available online: www.who.int/en/activities/tracking-SARS-CoV-2-variants (accessed on 10 September 2022).
25. Shu, Y.; McCauley, J. GISAID: Global initiative on sharing all influenza data–from vision to reality. *Eurosurveillance* **2017**, *22*, 30494. [CrossRef] [PubMed]
26. GISAID. Clade and Lineage Nomenclature Aids in Genomic Epidemiology Studies of Active hCoV-19 Viruses. 2021. Available online: https://gisaid.org/resources/statements-clarifications/clade-and-lineage-nomenclature-aids-in-genomic-epidemiology-of-active-hcov-19-viruses/ (accessed on 20 August 2022).
27. Hodcroft, E.B.; Hadfield, J.; Neher, R.A.; Bedfor, T. Year-Letter Genetic Clade Naming for SARS-CoV-2 on Nextstrain.org. 2020. Available online: https://nextstrain.org/blog/2020-06-02-SARSCoV2-clade-naming (accessed on 25 July 2022).
28. Rambaut, A.; Holmes, E.C.; O'Toole, Á.; Hill, V.; McCrone, J.T.; Ruis, C.; du Plessis, L.; Pybus, O.G. A dynamic nomenclature proposal for SARS-CoV-2 lineages to assist genomic epidemiology. *Nat. Microbiol.* **2020**, *5*, 1403–1407. [CrossRef] [PubMed]
29. O'Toole, Á.; Scher, E.; Underwood, A.; Jackson, B.; Hill, V.; McCrone, J.T.; Colquhoun, R.; Ruis, C.; Abu-Dahab, K.; Taylor, B.; et al. Assignment of epidemiological lineages in an emerging pandemic using the pangolin tool. *Virus Evol.* **2021**, *7*, veab064. [CrossRef]
30. Wang, Q.; Zhang, Y.; Wu, L.; Niu, S.; Song, C.; Zhang, Z.; Lu, G.; Qiao, C.; Hu, Y.; Yuen, K.Y.; et al. Structural and functional basis of SARS-CoV-2 entry by using human ACE2. *Cell* **2020**, *181*, 894–904. [CrossRef]
31. Liu, Z.; Xiao, X.; Wei, X.; Li, J.; Yang, J.; Tan, H.; Zhu, J.; Zhang, Q.; Wu, J.; Liu, L. Composition and divergence of coronavirus spike proteins and host ACE2 receptors predict potential intermediate hosts of SARS-CoV-2. *J. Med. Virol.* **2020**, *92*, 595–601. [CrossRef]
32. Plante, J.A.; Liu, Y.; Liu, J.; Xia, H.; Johnson, B.A.; Lokugamage, K.G.; Zhang, X.; Muruato, A.E.; Zou, J.; Fontes-Garfias, C.R.; et al. Spike mutation D614G alters SARS-CoV-2 fitness. *Nature* **2021**, *592*, 116–121. [CrossRef]
33. Ozono, S.; Zhang, Y.; Ode, H.; Sano, K.; Tan, T.S.; Imai, K.; Miyoshi, K.; Kishigami, S.; Ueno, T.; Iwatani, Y.; et al. SARS-CoV-2 D614G spike mutation increases entry efficiency with enhanced ACE2-binding affinity. *Nat. Commun.* **2021**, *12*, 848. [CrossRef]
34. Motozono, C.; Toyoda, M.; Zahradnik, J.; Saito, A.; Nasser, H.; Tan, T.S.; Ngare, I.; Kimura, I.; Uriu, K.; Kosugi, Y.; et al. SARS-CoV-2 spike L452R variant evades cellular immunity and increases infectivity. *Cell Host Microbe* **2021**, *29*, 1124–1136. [CrossRef]
35. McCallum, M.; Bassi, J.; De Marco, A.; Chen, A.; Walls, A.C.; Di Iulio, J.; Tortorici, M.A.; Navarro, M.J.; Silacci-Fregni, C.; Saliba, C.; et al. SARS-CoV-2 immune evasion by the B. 1.427/B. 1.429 variant of concern. *Science* **2021**, *373*, 648–654. [CrossRef]
36. Majumdar, P.; Niyogi, S. ORF3a mutation associated with higher mortality rate in SARS-CoV-2 infection. *Epidemiol. Infect.* **2020**, *148*, e262. [CrossRef]
37. Lamy, J.B. Owlready: Ontology-oriented programming in Python with automatic classification and high level constructs for biomedical ontologies. *Artif. Intell. Med.* **2017**, *80*, 11–28. [CrossRef] [PubMed]

38. Gangavarapu, K.; Latif, A.A.; Mullen, J.; Alkuzweny, M.; Hufbauer, E.; Tsueng, G.; Haag, E.; Zeller, M.; Aceves, C.; Zaiet, K.; et al. B.1.617.2 Lineage Report, Outbreak.info. 2022. Available online: https://outbreak.info/situation-reports?pango=B.1.617.2 (accessed on 20 September 2022).
39. Kazybay, B.; Ahmad, A.; Mu, C.; Mengdesh, D.; Xie, Y. Omicron N501Y mutation among SARS-CoV-2 lineages: Insilico analysis of potent binding to tyrosine kinase and hypothetical repurposed medicine. *Travel Med. Infect. Dis.* **2022**, *45*, 102242. [CrossRef] [PubMed]

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.