 doi:10.20944/preprints202008.0320.v1

Data Science in Unveiling COVID-19 Pathogenesis and Diagnosis: Evolutionary Origin to Drug Repurposing

Jayanta Kumar Das[a], Giuseppe Tradigo[b], Pierangelo Veltri[c], Pietro H Guzzi[c,*], Swarup Roy[d,*],

[a]*Department of Pediatrics, Johns Hopkins University School of Medicine, Maryland, USA*
[b]*eCampus University, Via Isimbardi 10, 22060 Novedrate, CO, Italy*
[c]*Department of Surgical and Medical Sciences, Magna Graecia University, Catanzaro, 88100, Italy*
[d]*Network Reconstruction & Analysis (NetRA) Lab, Department of Computer Applications, Sikkim University, Gangtok, India*

Abstract

The outbreak of novel Coronavirus (SARS-COV-2) disease (COVID-19) in Wuhan has attracted worldwide attention. SARS-COV-2 known to share a similar clinical manifestation that includes various symptoms such as pneumonia, fever, breathing difficulty, and in particular, SARS-COV-2 also causes a severe inflammation state that leads to death.

Consequently, massive and rapid research growth has been observed across the globe to elucidate the mechanisms of infections and disease progression in genotype and phenotype scale. Data Science is playing a pivotal role in in-silico analysis to draw hidden and novel insights about the SARS-COV-2 origin, pathogenesis, COVID-19 outbreak forecasting, medical diagnosis, and drug discovery. With the availability of multi-omics, radiological, biomolecular, and medical data urges to develop novel exploratory and predictive models or customise exiting learning models to fit the current problem domain. The presence of many approaches generates the need for the systematic surveys to guide both data scientists and medical practitioners.

We perform an elaborate study on the state-of-the-art data science method-

*Corresponding Author
Email addresses: hguzzi@unicz.it (Pietro H Guzzi), sroy01@cus.ac.in (Swarup Roy)

 doi:10.20944/preprints202008.0320.v1

ologies in action to tackle the current pandemic scenario. We consider various active COVID-19 data analytics domains such as phylogeny analysis, SARS-COV-2 genome identification, protein structure prediction, host-viral protein interactomics, clinical imaging, epidemiological analysis, and most importantly (existing) drug discovery. We highlight types of data, their generation pipeline, and the data science models in use. We believe that the current study will give a detailed sketch of the road map towards handling COVID-19 like situation by leveraging data science in the future. We summarise our review focusing on prime challenges and possible future research directions .

1. Introduction

The massive outbreak of SARS-COV-2 (Severe Acute Respiratory Syndrome CoronaVirus) viral infections in the world lead to a life-threatening pathogenic disease, WHO (World Health Organisation) named it as COVID-19 (coronavirus disease) [1]. The surprisingly rapid human-to-human transmission created an alert with the increasing number of cases [1] [2]. Since December 2019, the novel coronavirus had a surprisingly high spreading rate among humans, and the WHO declared it a pandemic in March 2020. It already infected more than 17 million people and has spread to over 213 countries so far in less than six months and its continuing. Unlike other severe acute respiratory syndrome (SARS) coronaviruses (order *Nidovirales*, family *Coronaviridae*, subfamily *Coronavirinae*) like SARS-CoV or MERS (Middle East respiratory syndrome) coronavirus, SARS-COV-2 spreading is massive and so far unsolved battle. The unprecedented scenario of COVID-19 pandemic forces the scientific communities for the rapid development of vaccines and drugs to control the outbreak by understanding the disease pathogenesis.

The explosion of COVID-19 related published research confirms that scientific communities are actively contributing to understanding the pathogenicity and control of SARS-COV-2 . A growth trend is shown in Figure 2. According to Nature Index updates on 27 June 2020[2], 67,753+ papers on COVID-19 have published so far (see also [3]).

[1] https://www.who.int/emergencies/diseases/novel-coronavirus-2019/situation-reports

[2] https://www.natureindex.com/news-blog/the-top-coronavirus-research-articles-by-metrics

 doi:10.20944/preprints202008.0320.v1

Undoubtedly, a significant fraction of the scientific community, comprising almost all the disciplines, is working on developing vaccines, therapies, as well as patient and resource management, on helping the fight against the virus. It is causing the rapid availability of free to access COVID-19 related omics and clinical data allowing further research results to be released almost instantaneously. For instance, focusing only on genomic data, the GISAID database[3] has collected more than 67,000 viral genomic sequences that have been collected and shared with unprecedented speed [4]. The Johns Hopkins dashboard [4] has rapidly become one of the primary data sources for monitoring disease from an epidemiological perspective. The rapid accumulation of data and the need to support wet-lab investigation also motivated the introduction of many data/computational based approaches (e.g., deep learning, artificial intelligence, network medicine) [5]

It helped to understand the pathogenesis of the disease and the rapid development of vaccine/drugs, but conversely, this has resulted in an accumulation of data, algorithms, software, and tools that need to be categorized and organized. The whole categorization of all the approaches is undoubtedly yet impossible due to the constant rate of the introduction of novel tools.

Therefore, we aim to present the main characteristics of the current landscape, as depicted in Figure 1. We categorized them into five levels: data source, repositories, data science models, decision making, and interpretation. Some of the outcomes (say phylogeny or interactomes) are looped back, for other data science analysis, as input. The integration and analysis step also involves data science models to infer more meta-knowledge from the intermediate decision or outcome. For instance, the drug-disease association needs a network biology approach to determine the optimized relationship between drug molecules and their impact during the disease. Many laboratories are producing a massive amount of heterogeneous data, considering both format and content. Viral sequences are represented as strings. Raw clinical data are largely heterogeneous, while medical images are more standardized data. Such data are accumulated into public databases or websites (e.g., GISAID database, Johns Hopkins public repository) that may be integrated with other existing databases (e.g., virus-host interaction databases, clinical and epidemiological databases). It is the starting point for any data-

[3] https://www.gisaid.org/

[4] https://coronavirus.jhu.edu/map.html

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 14 August 2020 doi:10.20944/preprints202008.0320.v1

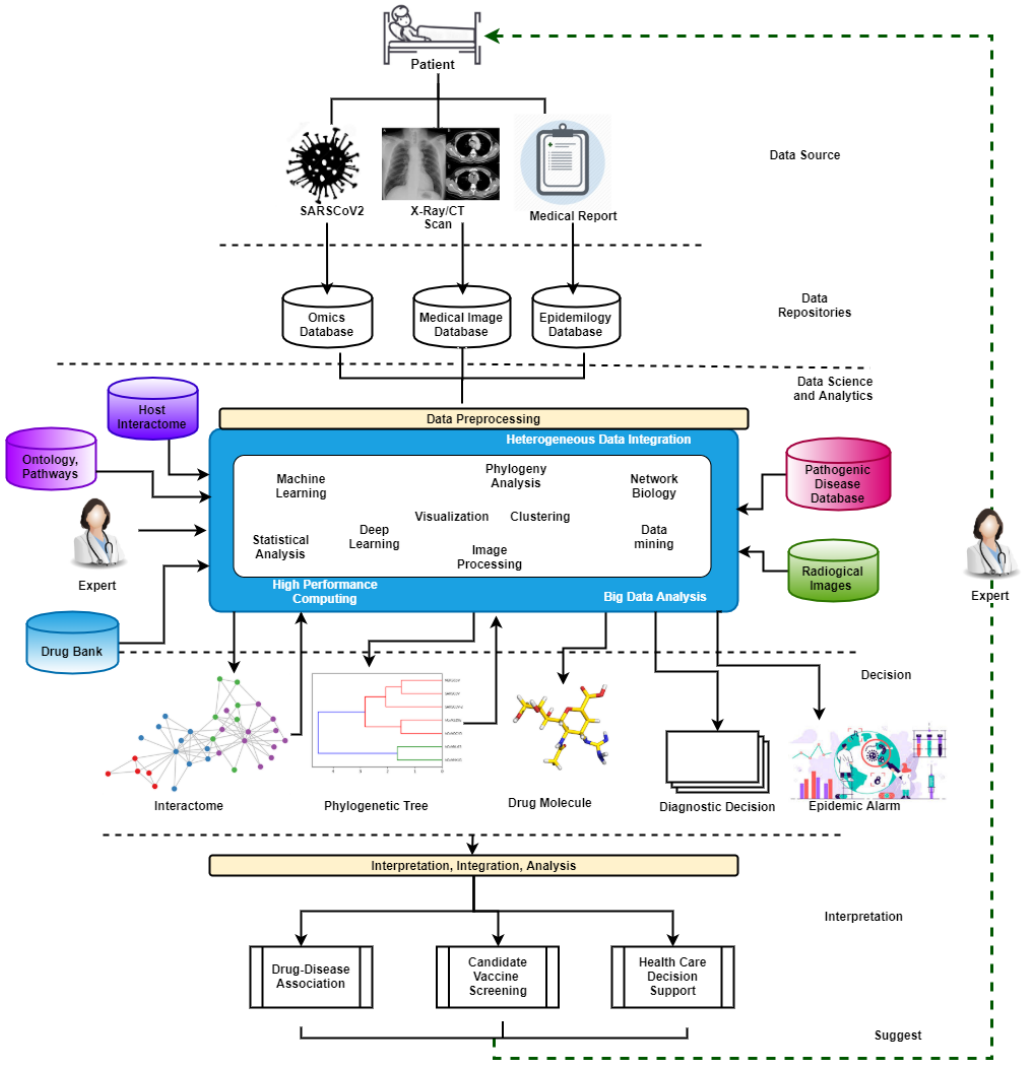

Figure 1: A Data-Science landscape for SARS-COV-2 and COVID-19 studies. Many different technologies produce a large quantity of data related to patients at different scales (e.g., molecular data, medical images, and clinical data, and epidemiological data). The accumulation of these data is the pre-requisite for a substantial rise of data-science approaches (e.g., deep-learning and classical data mining) that often integrate existing data stored in databases or apriori knowledge (e.g., domain experts or ontologies). Such approaches produce new information about molecular interactions, phylogenetic analysis, in-silico design of drugs, or healthcare management decisions. The output of this information may guide the execution of novel experiments closing the loop of the whole process.

 doi:10.20944/preprints202008.0320.v1

science based approach. Such approaches may be categorized considering the methodological approach or the aim, such as studying drug-disease association, screening possible candidate vaccines, or supporting healthcare decisions (e.g., management of resources such as Intensive Care Units). Finally, these analysis results may guide the execution of novel experiments to confirm in-silico findings (e.g., novel hypothesis).

Main contributions of this review are:

- we present a comprehensive review on the in-silico approaches adopted so far to handle COVID-19 in genomics, proteomics, interactomics, epidemics, clinical imaging, and drug identification level.
- for the first time ever we present the systematic categorization of data analytics tasks related to COVID-19;
- we present an overall landscape suggesting the judicious integration of heterogeneous data sources;
- we report the data resources, data science models, and tools used to analyze SARS-COV-2 and COVID-19 .
- in a nutshell, our current documentation will give the glance of computational biology and bioinformatics approaches available in all spare of biological data analysis;
- we aim to offer to the data-scientists, medical doctors, healthcare advisers, drug/vaccine designers, a guide to finding a suitable vademecum to select the right data, models, approaches, and tools.

2. Virology and Data Science: Background

Recent pandemic forces different research communities (e.g., virologists, computational biologists, medicine specialists, data scientists, etc.) to collaborate for the rapid understanding of the pathogenesis and diagnosis of COVID-19. Large number of recent researches generating bulk experimental data. It needs to be analyzed in-silico using data science technologies to unveil hidden and novel knowledge. It is crucial to have a basic understanding of the biology behind SARS-COV-2 and data science and its steps. Here, we try to briefly introduce the SARS-COV-2 virus and how omics, images data, and epidemiological data related to COVID-19 are generated and stored in

 doi:10.20944/preprints202008.0320.v1

publicly available data repositories. We briefly introduce the concept of data science used as a tool to unleash new insights from the available datasets to handle COVID-19.

2.1. Virus Biology of SARS-COV-2

Viruses are small microorganisms that use living cells to replicate themselves. Viruses cause many infectious diseases that are responsible for millions of death every year [6]. They exist in the form of small independent particles named virions. Each virion consists of two main components: (*i*) the genetic information, encoded as DNA or RNA, and (*ii*) a protein coat, named capsid, which wraps the genetic material. Sometimes the capsid is surrounded by an envelope of lipids. Virions have different shapes that are used for classification of themselves [7]. Viruses are not able to replicate themselves alone. Therefore they must use the metabolism of an **host** organism to reproduce themselves. The virus replication cycle may be summarised into six main steps [8] :

1. **Attachment**. First, viruses bind the surface of host cells.
2. **Penetration**. Viruses enter the host cell through receptor-mediated endocytosis or membrane fusion.
3. **Uncoating**. The viral capsid is removed, and virus genomic materials are released.
4. **Replication**: Viruses use the host cells to replicate their genomic information. In this step, viral proteins are synthesized and possibly assembled. Viral proteins may interact among them and host proteins to perform their function (e.g., regulate the protein expression).
5. **Assembly**. Following the structure-mediated self-assembly of the virus particles, some modifications of the proteins often occur.
6. **Release:** Viruses can be released from the host cell by lysis, a process that kills the cell.

There exist many different viruses classified into major classes by phenotypic characteristics, such as morphology, nucleic acid type (e.g., RNA or DNA), and more. The taxonomy of viruses is in charge of the International Committee on Taxonomy of Viruses (ICTV). In parallel, the Baltimore classification system is also used, and viruses are grouped into seven groups based on mRNA synthesis.

SARS-COV-2 belongs to β-coronaviruses that are a subgroup of the coronavirus family. Such viruses are giant enveloped positive-stranded RNA

 doi:10.20944/preprints202008.0320.v1

viruses that are usually able to infect a wide variety of mammals and avian species. All the members of the family cause respiratory or digestive and enteric diseases [9]. The infection mechanism is based on the action of surface spikes constituted by glycoproteins (named S or spike proteins) that are responsible for binding host cell receptors.

The literature contains seven β-coronaviruses that are responsible for causing disease in humans. Four strains cause mild infection of respiratory apparatus that are treated without lethal consequences (HCoV 229E, HKU1, NL63, and OC43). More recently, three strains of betacoronavirus have caused severe and lethal diseases: SARS-CoV, MERS-CoV, and SARS-COV-2 [10].

SARS-CoV (i.e., Severe Acute Respiratory Syndrome CoronaVirus) was responsible for a severe respiratory syndrome outbreak in China in 2002. MERS-CoV (i.e., the Middle East respiratory CoronaVirus) caused an outbreak in the middle east in 2012. Both viruses had a similar manifestation: patients manifested pneumonia. MERS-CoV infected patients also presented gastrointestinal complications and kidney failures.

The third member of the family, SARS-COV-2 , appeared in December 2019 in Wuhan, Hubei Province, China [11]. From the initial steps, it presented a surprisingly rapid diffusion rate. Until now, COVID-19 has killed more people than SARS, and MERS combined, despite having lower fatality rate [12]. By the end of April 2020, the COVID-19 virus caused over $1,500,000$ confirmed cases around the world, of which around $350,000$ recovered, and over 94,000 patients died. In China, more than $80,000$ have been the confirmed cases with more than $3,000$ deaths.

The sequence and structural analysis revealed a high similarity between SARS-CoV and SARS-COV-2 , as confirmed by the evidence that the new coronavirus binds with the ACE2 receptor. Unfortunately, it presents a higher affinity than the previous virus [12]. Moreover, the pattern of expression of ACE2 in human respiratory epithelia and oral mucosa may represent the cause of human-human transmission. Clinical manifestation of COVID-19 are large and severe since it seems to impact all the tissues and organs that express ACE2 receptor: severe pneumonia, kidney failure, anemia, neurological problems, cardiovascular complication and a large state of inflammation (know as cytokine storm) in more severe causes [13, 14, 15].

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 14 August 2020 doi:10.20944/preprints202008.0320.v1

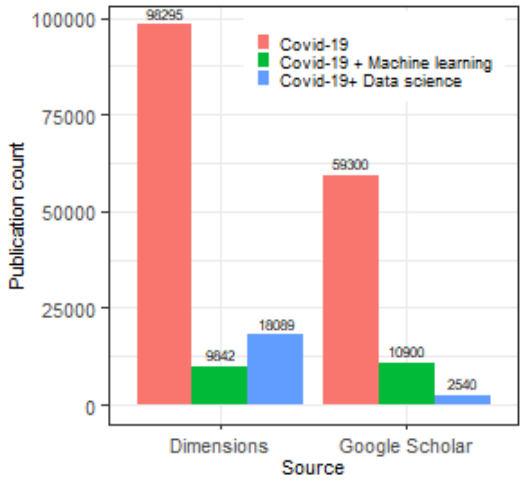

Figure 2: The trends of COVID-19 related research publications from two sources: Dimensions [7] and Google scholar [8] as of 08 August, 2020. We searched by three main keywords: COVID-19 , COVID-19 and Machine learning, and COVID-19 and Data science. The search hits includes publications in Published Articles, Preprints, Edited Books, Monographs, Proceedings, and Chapters.

2.2. COVID-19 Data Generation and Sources

On a positive note, COVID-19 pandemic is experiencing massive, unprecedented, and rapid growth in the data generation and research publications across the world (Figure 2). As reported in the academic search engine, Dimension [9], a total of 730 datasets related to COVID-19 are available publicly. Deposited data are primary data (sequence, clinical images, medical reports) or secondary data (Proteins structures, interactomes, epidemiology). Primary data are generated from the virus or the patients. Based on primary data, a more refined, summarised, and inferred outputs are again stored as secondary data. During COVID-19 data analysis and inferences, existing knowledge bases play a significant role as a reference set for drawing a new knowledge. Next, discuss briefly the data generation pipeline and publicly available repositories for COVID-19 data mining.

[9] `https://app.dimensions.ai/discover/publication`

 doi:10.20944/preprints202008.0320.v1

2.2.1. Omics Data

High-throughput omics technologies[10] use biochemical assays that assess comprehensively and simultaneously set of molecules in the biological samples. Omics data are usually categorized into Genomics, Transcriptomics, Proteomics, and Metabolomics type. To understand the severity of the COVID-19 disease, the first and foremost necessary task for scientists to sequencing the genome of SARS-COV-2 . RNA sequencing of SARS-COV-2 is essential to elucidate how this coronavirus grows, mutates, and replicates. Consequently, the rapid development of therapeutic vaccines and drugs use such sequence and protein data in the downstream pipeline.

Blood and/or throat swab specimens are collected from the suspected patients and RNA extracted using an RNA extraction kit. The extracted RNA further sequenced using Next Generation Sequencing (NGS). The principle behind NGS is capillary electrophoresis. At first, the genomic strands are fragmented, and the bases are identified in each fragment by ligation with custom linkers or template strand [16]. The NGS uses an array-based sequencing method to process millions of reactions in parallel at a very high speed at a reduced cost. There are three general steps involved in NGS: (i) Library Preparation, (ii) Amplification, and (iii) Sequencing. Before sequencing, the isolated and purified RNA must be converted to double-stranded complementary DNA (cDNA) or cDNA library. Sequencing platform-specific adapters are then added to each end of the fragments. Adding adapters and amplification of DNA to make a cDNA is the steps in library preparation. Finally, the cDNA library is sequenced using NGS. Reference-based mapping also performed to retrieve the sequence of the SARS-COV-2 before depositing into public repositories (see Table 2). Sequence data are available in FASTA or FASTQ format. Apart from these two, other sequencing data formats are Sequence Alignment Map (SAM) and BAM (a compressed binary format for SAM). A CoDing Sequence (CDS) [11] is a region of DNA or RNA whose sequence determines the sequence of amino acids in a protein. Protein sequences are derived from translations of Coding Sequence (CDS). CDS is not the same as Open Reading Frame (ORF), which is a series of DNA codons that do not contain any STOP codons. We report available SARS-COV-2 nucleotide and protein sequence repositories in the Table 2.

[10] `http://omics.org`
[11] `https://www.uniprot.org/`

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 14 August 2020 doi:10.20944/preprints202008.0320.v1

2.2.2. Interactome Data

Interactomics data refer to the set of molecules (e.g., proteins, transcription factors, small-molecules) and their biochemical interactions. Since these interactions are the elementary building blocks of almost all cellular processes, their elucidation does appear as an essential step in describing SARS-COV-2 mechanisms of infections and replications. There exist many experimental platforms for deriving such physical interactions [17], such as Affinity purification mass-spectrometry (AP-MS) and yeast-two-hybrid (Y2H) that enable the accurate identification of interactions with a relatively long time. Gordon et al. [18] expressed 26 out of 29 SARS-COV-2 proteins and used an affinity-purification followed by mass spectrometry to identify 332 human proteins to which the viral proteins bind. In parallel, there exist many bioinformatics tools that enable the prediction of interactions using biological pieces of information (e.g., homology modeling) coupled to network science (e.g., network alignment or link prediction). Table 1 summarises these approaches (see for instance [19] for a more detailed comparison). Nowadays, thanks to the use of different proteomic technologies, the complete set of interactions are available for many viruses [20, 21, 22]. At the same time, for SARS-COV-2 there exist some projects that have elucidated an almost complete map of interactions. Such interactions are usually modelled by using graphs and stored in a growing number of databases such as: Virus Mint [23], String Viruses [24], HpiDB [25], Virus Mentha [26], and VirHostNet [27].

Table 1: Technologies and Methodologies for Interactome Data Production

Class	Technique	Pros and Cons
Experimental	Mass Spectrometry - Yeast Two Hybrid	Low Throughput High Accuracy
Computational	Homology Modelling - Docking - Network Alignment	High Throughput Low Accuracy

Despite the existence of such platforms, the scenario in SARS-COV-2 studies is different due to the diffusion of the virus that imposes to the researcher to determine quickly such interactions. Consequently, the first studies used mainly the prediction of interactions made with bioinformatics tools. For instance, in [28], the homology among SARS-COV-2 and other

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 14 August 2020 doi:10.20944/preprints202008.0320.v1

coronaviruses is used to infer putative interactions among viral proteins and host-viral proteins. Differently, a wet lab approach[18] is used where SARS-COV-2 proteome is cloned and used AP-MS to identify 332 protein interactions between SARS-COV-2 and humans. More recently, some innovative projects integrate data from literature and other databases. For instance, in [29], the development of a curated dataset of SARS-COV-2 interactions extracted by the IMEx consortium is presented. Authors integrated proteins and interactions from SARS-COV-2 , SARS-CoV-1, and other members of the Coronaviridae family. The databases stores more than 2,200 interactions extracted from more than 80 publications. Following other standard initiatives, the dataset can be accessed in some standard format, and it is updated regularly.

2.2.3. Clinical Image Data

During COVID-19 clinical diagnosis, X-Ray and Computer Tomography (CT) technologies are of great use to diagnose infected lungs and respiratory tracks. X-Ray, an age-old technique appears to be effective in generating a 2D image of bones and organs. During imaging, X-ray beams are passed through the body, and detectors or a film capture the attenuated X-rays, resulting in a clinical image. CT is a relatively more advanced, powerful, and sophisticated 3D imaging technology that takes a 360-degree image of the bones and internal organs. Ground glass (GGO) pattern in an infected chest CT image is the most common finding in COVID-19 infections. Patterns are usually multifocal, peripheral, and bilateral. However, during COVID-19, GGO may appear as a unifocal lesion, most commonly located in the inferior lobe of the right lung [30]. Chest X-Ray images observed to be insensitive early in the disease. However, X-Ray may be useful in tracking disease progression.

The urgent need for an automatic diagnostic tool for the rapid detection of COVID-19 patients forces the data science communities to develop a machine learning-based diagnostic framework. A good number of free repositories and platforms are now available where labeled. X-ray and CT images can be found for building machine learning models. *TrainingData.io* [12] is one of the platforms offering a free collaborative tool. It is seeded with free image datasets. Tools allow data-scientists and radiologists to share annotations of

[12] `https://www.trainingdata.io/`

 doi:10.20944/preprints202008.0320.v1

training data to be used for training machine learning models for COVID-19. We report more such repository of annotated COVID-19 infected chest images in the Table 2.

2.2.4. Epidemiological Data

Epidemiological data is a collection of non-experimental observations deriving from field investigations, statistics, or other health-related sensors collected by domain experts. They are used to study data distributions and infer knowledge on epidemiological phenomena.

From the beginning, the availability of epidemiological data (possibly coupled to clinical and laboratory data), is essential. The relevance of such data is related to the possibility of a better understanding of the transmissibility rate, the pattern of the geographic spread, as well as associated risks or co-morbidities.

Consequently, many independent groups started to collect epidemiological data produced and made available from healthcare providers. Dong et al., [31] designed and developed the first dashboard hosted at Johns Hopkins University, providing free access to health data collected from almost all the nations in the world. Data are related to reported cases of COVID-19 , i.e., infected, dead, and recovered patients. In parallel, the Italian government provided a similar dashboard, and related data, through an interactive, web-based system [32]. Similarly, Xu et al., [33], realized an open-access database for storing patient information produced in laboratories. Stored data are related to movements (for retrieving travel history), symptoms, and demographics.

All these projects share some common characteristics: (i) the use of simple formats (e.g., tabular formats), (ii) the possibility of exportation in common data sharing format (e.g., comma-separated values or JSON), (iii) simple query interfaces, and (iv) the integration of geographic data. Moreover, some of them [33] also include demographic information. We report a few more repositories in Table 2.

2.2.5. Drug-target databases

The drug is an organic small molecule [34] that activates or inhibits the function of a therapeutically important protein in disease development. New drug molecule development (termed as drug discovery) and approval process by the FDA (Food and Drug Administration) are a complex, expensive, and time-consuming process. Two significant steps are involved during any drug

 doi:10.20944/preprints202008.0320.v1

discovery, (i) drug target identification that are critical proteins involved in a particular metabolic or signaling pathways in a specific disease condition, and (ii) developing small molecules that interact with the targets. To speed-up the process, computer-aided methods are introduced for the automated drug discovery [35], which is fast and accurate. It involves steps like hit identification using virtual screening, hit-to-lead optimization of affinity, selectivity, and lead optimization of other pharmaceutical properties while maintaining affinity. The approved drug molecules and targets are stored in a publicly available database for drug repurposing or commercial development of other drug molecules. Drug molecules are usually stored as SMILES (Simplified Molecular Input Line Entry System) format.

The **DrugBank** is one of the popular repositories of drug and drug target information. The latest release of DrugBank (version 5.1.7, released 2020-07-02) contains 13,596 drug entries, including 2,640 approved small molecule drugs, 1,389 approved biologics (proteins, peptides, vaccines, and allergenic), 131 nutraceuticals and over 6,377 experimental (discovery-phase) drugs. Additionally, 5,225 non-redundant protein (i.e. drug target/enzyme/transporter/carrier) sequences are linked to these drug entries. **PubChem** is another world's largest collection of freely accessible chemical information. It stores chemical and physical properties, biological activities, safety and toxicity information, patents, literature citations, and more about any drug.

Excelra COVID-19 Drug Repurposing Database is an open-access database which is a compilation of 'Approved' small molecules and biologics, that may rapidly enter into Phase 2 or 3. **PHARMACEUTICAL** is a coronavirus drug tracker that lists drugs in all stages of preclinical and clinical development (from discovery through to preregistration) for COVID-19. This list is updated dynamically, based on the GlobalData Pharma Intelligence Center Drugs database. **CAS** contains a connection of nearly 50,000 chemical substances, along with related metadata such as CAS Registry Number and physical properties for each element. It is available in the SD file format (.sdf). More drug databases are reported in Table 2. .

2.3. The Data Science Pipeline

Data science is a growing interdisciplinary research field that leverages methods, processes, and algorithms to support the extraction of relevant knowledge from (big)-data. The term Data science was coined in 2008 by DJ Patil and Jeff Hammerbacher [37]. Data science starts with raw data and ends with decision or meta (description or summarization) data that

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 14 August 2020 doi:10.20944/preprints202008.0320.v1

Table 2: COVID-19 Public Data Repositories.

Data Type	Repositories	Description	Source
Omics			
Nucleotide/ Protein	GISAID	more than 75,000 viral genomic sequences of hCoV-19 ((updating))	https://www.gisaid.org/
Nucleotide/ Protein	NCBI	more than 11969 nucleotide, 124288 protein (updating)	https://www.ncbi.nlm.nih.gov/labs/virus/vssi/#/
Structure	RCSB PDB	SARS-COV-2 proteins (updating)	https://www.rcsb.org/covid19
Structure	SWISS-MODEL	SARS-COV-2 proteins (updating)	https://swissmodel.expasy.org/repository/species/2697049
Heterogeneous	Covid-19 hg		https://www.covid19hg.org
Interactomics			
Interactions, network	BioGRID	more than 800 interacting proteins (updating)	https://thebiogrid.org/
Interaction, graph	IntAct	4,479 binary interactions (updating)	https://www.ebi.ac.uk/intact/
Interacting protein	HPA	more than 200 interacting human proteins with SARS-COV-2 (updating)	https://www.proteinatlas.org/humanproteome/SARS-COV-2~
Imaging			
X-Ray	github	more than 800 images (updating)	https://github.com/ieee8023/covid-chestxray-dataset
CT	github	349 images from 216 patients	https://github.com/UCSD-AI4H/COVID-CT
	github	63849 images from 377 patients	https://github.com/mr7495/COVID-CTset
	github	104,009 CT images from 1,489 patients	https://github.com/lindawangg/COVID-Net/
	MosMED	15589 and 48260 CT scan images belonging to 95 Covid-19 and 282 normal persons	https://mosmed.ai/en/
Both	BIMCV-COVID19+	X-ray images CXR (CR, DX), 1380 CX, 885 DX and 163 CT	https://osf.io/nh7g8/
Epidemiological			
Information	CIDRAP	Cases of coronavirus disease, situation report, epidemiology, virology, clinical features	https://www.cidrap.umn.edu/covid-19/epidemiology
	WHO	Information regarding covid19	https://covid19.who.int/
	Italian Civil Protection		https://github.com/pcm-dpc/COVID-19
	SCIENTIFIC DATA [36]	Curated individual-level data from national, provincial, and municipal health reports and online reports	https://doi.org/10.6084/m9.figshare.11974344
Drug repurposing			
Molecule	Drugbank	Contains around 13,606 drug entries	https://www.drugbank.ca/covid-19
	PubChem	World largest database: more than 350 million Compounds, Substances, BioAssay	https://pubchem.ncbi.nlm.nih.gov/
	ChEMBL	SARS-COV-2 related bioactive molecules with drug-like properties	https://www.ebi.ac.uk/chembl/
	Excelra	COVID19 related drugs that are 'clinical, pre-clinical and experimental' stage.	https://www.excelra.com/covid-19-drug-repurposing-database/
	CAS	Anti-viral drugs and related chemical compounds for COVID19 disease	https://www.cas.org/covid-19-antiviral-compounds-dataset
	Pharmaceutical	Drugs in all stages of preclinical and clinical development for COVID-19 indication	https://www.pharmaceutical-technology.com/coronavirus-drug-trials-studies/

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 14 August 2020 doi:10.20944/preprints202008.0320.v1

needs to be interpreted before transforming it into knowledge. Broadly data science pipeline passes through four major phases: (i) Raw data collection, (ii) Preprocessing (iii) Descriptive or Predictive modeling (iv) Interpretation. An illustrative representation of the data science workflow for COVID-19 management is shown in Figure 3.

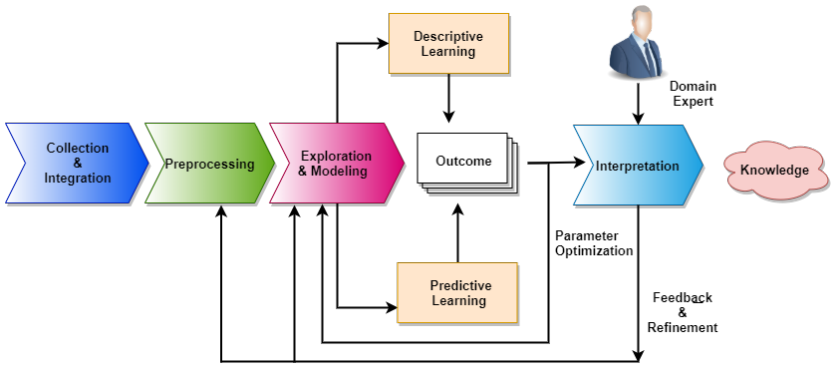

Figure 3: Major phases of Data Science pipeline towards decision making and analysis. Data initially collected and integrated from many sources. Then they need to be preprocessed to filter uninformative or possibly misleading values (e.g. outliers or noise). Then existing models are used to explain data or to extract relevant patterns describing data or to predict associations. Finally results need to be interpreted and explained by domain experts. Each step of analysis may generate corrections (or refinements) that are applied to precedent steps.

Single source data input always are not conclusive in optimized decision making. Heterogeneous data integration and fusion are vital, while any data science model is considered for analysis. The success of any data science model largely depends on the quality of data in hand. Due to flaws in data generation, storage, and transfer of data (molecular, imagery, and epidemic data for COVID-19), noise may be attributed, which may deviate the true outcome. It is crucial to apply different scrubbing and cleaning process on the data. Standardization and transformation of data are required if data that are to be consulted are generated from multiple and varying sources. Collectively, all the above steps are labelled as *preprocessing* step (more details available in [38]). Once target data is ready after scrubbing and cleaning, data exploration can be performed before feeding the data to the computing model for decisive or descriptive outcomes. To understand data better, it is suggestive of exploring the type, distribution, significant features (attributes), and relationship among the data variables. Feature selection is

 doi:10.20944/preprints202008.0320.v1

one of the steps that help in identifying important attributes. Visualization is a great way to explore the intimate relationship within the target data. Feeding the most relevant subset of data through dimensionality and data reduction helps to learn models to generate optimal outcomes. Descriptive (unsupervised) and predictive (supervised) are the two major learning paradigms through which data can be modeled and described for knowledge generation. An intermediate model, termed a semi-supervised model, culminating unsupervised learning followed by supervised learning, is another popular model. Ensemble decision making that combines the decision of multiple learning models (both supervised and unsupervised) is observed to be a more effective alternative [39]. Most popular supervised models such as machine learning and deep (machine) learning models are in great use for handling and analyzing COVID-19 pandemic. It is worth mentioning a few deep frameworks which are of great use during COVID-19 predictive data analysis. Convolutional Neural Network (CNN) [40] is used extensively in radiological chest images of COVID-19 patients. It is a modified version of traditional neural network architecture that uses convolution (linear operation) in place of general matrix multiplication in at least one of their layers. Other than input-output layers, it involves multiple hidden layers such as convolution, activation, pooling, fully connected, and softmax layers. A more evolved and powerful CNN model specific to graph or network is Graph Convolution Network (GCN) [41] that appears to be very useful in predicting interactomes or COVID-19 drug-target associations that rely largely on multiple network integration. It considers two types of input, the adjacency matrix and the feature vector of each node in the graph. Due to the lack of adequate training samples during COVID-19 to train deep models, the synthetic data play a significant role [42, 43]. A recent breakthrough in deep model architecture is the Generative Adversarial Networks (GAN) [44] for generating realistic synthetic data. A GAN is made of two simultaneously trained neural networks (Generator and Discriminator). Discriminator recognizes training samples, whereas Generator is to create fake instances to throw challenges to Discriminator. Both the networks tune themselves by conflicting each other's performance.

To a large extent, the performance of a descriptive or predictive data science model depends on the optimized feature engineering representing the true distribution of data in hand. However, with the arrival of deep learning models, life becomes easy. Deep models extract relevant features of its own, and relative performance is far superior to the shallow neural networks

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 14 August 2020 doi:10.20944/preprints202008.0320.v1

or other classification models. On the contrary, deep models are resource-intensive and work well when the sample size is large.

Regression analysis is another supervised model used to predict pandemic trends [45]. Clustering [46] is a well-known unsupervised learning model that describes and summarises the hidden pattern inside data based on certain proximity analysis. Investigating the evolutionary origin of SARS-COV-2 virus is the most highlighted and debated issue that the scientific community is trying to address with the help of clustering analysis. Decision and description offered by different employed learning models are not meaningful until it is interpreted by the domain expert to consider it as a novel knowledge. Feedback of the expert may also be propagated iteratively to refine the phases of the data science pipeline (see Figure 1).

Next, we discuss various data science tasks involved in COVID-19 data analysis. We compartmentalized our discussion into five different sections based on the type of significant activities related to understanding disease pathogenesis, diagnosis, spreading, and therapeutics. We primarily highlight data handling, learning models employed, methodologies, and software tools used.

3. Omics Data Analysis: DNA and Protein

The main objective of the omics study is to discover the proximal origin of SARS-COV-2 , its mutational variants, and developing a predictive model for identifying SARS-COV-2 from an isolated strain or sequence. In addition to that, nucleotide sequences are used to determine the SARS-COV-2 viral genome and 3D protein structure prediction.

3.1. Phylogeny and Mutant Variation Analysis

Discovering the evolutionary origin of SARS-COV-2 is the most focused research right now, trying to generate an evolutionary tree between its nearest species [47]. Clustering is the most popular technique applied to create a phylogeny among the coronavirus family-like SARS-COV, MERS-COV [48, 49, 50]. The evolution of viruses is mapped in silico using phylogenetic analysis. In general, the phylogenetic tree creation method relies primarily on either sequence alignment or without alignment (alignment-free). In bioinformatics, sequence alignment is a way of re-organizing given sequences (DNA, RNA, or protein) to identify functional and structural conserved regions within the sequences that provide a hint for the evolutionary trends.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 14 August 2020 doi:10.20944/preprints202008.0320.v1

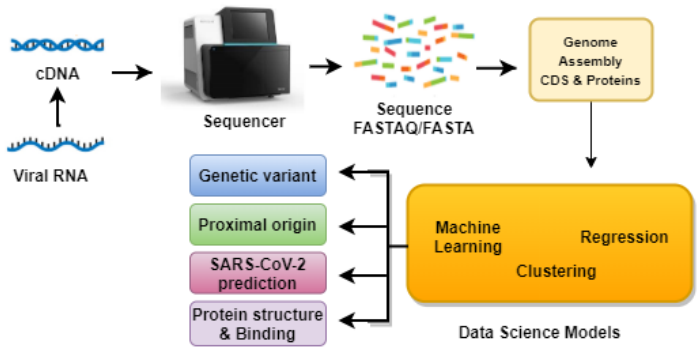

Figure 4: Omics data generation and workflow of analysis. Fragments of virus nucleic acids are extracted from host organisms (patients or other species). Then resulting in different algorithms process data. Many goals of these analyses are: (i) analysis of genetic variants, (ii) analysis of genomes of viruses infecting different species, (iii) prediction of protein interactions and interactome, and (iv) inferring structure and dynamics of viral proteins.

When sequences are lengthy, highly variable, or extremely numerous, it is almost impossible to align solely by human effort. Several algorithms have been developed to produce high-quality sequence alignments both for global alignments (focusing on the whole length) and local alignments (interest on specific region). Few common established methods and software are used for SARS-COV-2 genome alignment such as DNAMAN [13], ClustalW [51], MUSCLE [52], Jalview [53], and MAFFT [54]. From alignment data, the evolutionary history is inferred using the Neighbour-Joining method or maximum likelihood method [51, 55]. Alignment approaches are prominent in particular for identifying the SNP (Single Nucleotide Polymorphism) variation and lucidity of proximal origin of rapidly evolving viruses like SARS-COV-2 [56, 10, 10]. On the other hand, alignment-free methods are independent of the length and volume of the sequence data. It uses feature-based methods generated from sequence data and then compares the sequences by the derived features. Due to the low mutation rate and a high degree of similarity among SARS-COV-2 genome, very few studies have been performed using alignment-free methods. Correlation and partial information correlation (PIC) [57], optimal word (k-mer) to construct continuous distributed

[13] https://www.lynnon.com/dnaman.html

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 14 August 2020 doi:10.20944/preprints202008.0320.v1

representations for protein sequences, to predict MHC class I and II binding affinity [58], combined k-mer and n-gram techniques [59], chemical properties (charge, hydropathy, side chain) of region-specific amino acid, mutations are few important features used during analysis [60, 61]. These features are utilized in several studies attempting to identify the cluster of SARS-COV-2 in various ways, and its results demonstrated that this virus has multiple origins with different degrees. Studies reveal that origin of SARS-COV-2 genome is most likely similar with SARS-related coronaviruses [14] found in Pangolin [62] or Bat [63, 64]. Scientists are studying the variants of novel coronavirus to understand its mutant variants across the world. Despite of high degree of similarity among SARS-COV-2 genomes collected from the strains across the world [65, 66], significant variations are also reported[67, 68]. These studies are important to understand the clinical presentation and spread of the disease and useful for antiviral drug design [69].

3.2. SARS-COV-2 genome prediction

Detection of highly divergent viruses is a primary challenging task given its clinical importance. SARS-COV-2 genome prediction is equally essential for the rapid classification of novel pathogens as the virus is mutating to create divergent variants worldwide and differentiate with other coronaviruses. Machine learning models coupled with comparative genomics, are in extensive use to predict SARS-COV-2 . Both the alignment and alignment-free methods are applied to generate features. Some significant genomic features considered are k-mer and N-gram, amino acid chemical properties, substitution/mutation information by alignment method are mostly utilized for training machine learning models.

A combination of five well-known classification models [70], Naïve Bayes, K-Nearest Neighbors (KNN), Artificial Neural Networks (ANN), Decision tree and Support Vector Machine(SVM) are applied on N-gram features derived from 4432 genomes belonging to 93 families of coronavirus and 1869 genomes of SARS-COV-2 for novel SARS-COV-2 detection. In another attempt, both k-mer and n-gram sequence motifs are used as a feature for classification of SARS-COV-2 virus using Naïve Bayes classifier[59]. Integrated comparative genomics and machine learning applied to identify key genomic features that differentiate SARS-COV-2 from SARS-COV and MERS-

[14] https://www.ecohealthalliance.org/2020/01/

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 14 August 2020 doi:10.20944/preprints202008.0320.v1

COV [60]. Using multiple sequence alignment (MSA), high-confidence alignment blocks (regions longer than 15 nucleotides (nt)) are identified and trained SVMs with a linear kernel function on all 5-nt sliding windows in the identified high-confidence alignment regions. Concept of digital signal processing [71] coupled with decision tree also applied for sequence classification based on over 5000 unique viral genomic sequences, totaling 61.8 million bp that includes 29 COVID-19 virus sequences.

A good attempt has made leveraging the deep learning paradigm without using any pre-selected genomic features. A general Convolution Neural Networks (CNN) [72] made up of a convolutions layer with max pooling, a fully connected layer, and a final softmax layer[73] used to classify six different human coronaviruses (SARS-COV-2 MERS-CoV, HCoV-NL63,HCoV-OC43, HCoV-229E, HCoV-HKU1, and SARS-CoV). It used various encoding schemes (C=0.25, T=0.50, G=0.75, A=1.0) to encode the cDNA data into an input tensor for the CNN. A prediction model has been proposed to detect what kind of target or host a virus can infect [74]. A Bi-path CNN (Bi-PathCNN) [75] is used where each viral sequence is represented by a one-hot matrix for nucleotide bases and codons separately. Experimental outcome reports six genomes of SARS-COV-2 having high infection possibility (p-value¡0.05) to humans.

3.3. Protein Structure Prediction

Non-synonymous mutation [69] alters the amino acid that can influence the structure and function of the protein. It is of utmost importance to understand the viral protein structures for identifying functional motifs towards understanding the possible binding mechanisms with the host proteins and antiviral-drug discovery [76, 77]. Experimentally, it is difficult to determine the 3D structure of a protein. However, computationally it is possible to predict the structure of a protein from the amino acid sequence. Utilizing the inherent vital genetic information inside the amino acid sequence, it is now possible to infer a protein more accurately with the help of several machine learning models, specifically with the help of advanced deep learning models. Recently, SARS-COV-2 proteins are predicted [15] by the AlphaFold and the structures are deposited in Protein Data Bank [16]. AlphaFold [78] is

[15] https://deepmind.com/
[16] https://www.rcsb.org/

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 14 August 2020 doi:10.20944/preprints202008.0320.v1

a deep two-dimensional dilated convolutional residual network that predicts the inter-residue distances between pairs of amino acids and the angles between chemical bonds that connect those amino acids. *trRosetta* [79] that uses inter-residue distance and orientation distributions, predicted by a deep residual neural network and its modification [80] are also used to predict SARS-COV-2 protein structures. In addition to that other existing protein structure and homology modelling tools like COMPOSER [81], SWISS-MODEL [82], PyMOL c [83], and I-Tasser [84] are used for rapid prediction and comparison of Spike (S) protein [85, 86], Envelope (E) protein [2] and *ab initio* homology modelling [84].

4. Interactome Network Inference and Analysis

Interactomics research related to SARS-COV-2 has two main goals: (i) development of possible therapies for helping affected people, and (ii) introduction of a novel vaccine for blocking the spreading. Despite the existence of many different laboratories that have sequenced the whole genome and the availability of such data, the above-described point may be successfully addressed, looking at a molecular scale through the elucidation of interactions of viral proteins concerning the host proteins. As introduced before, during the replication step, the virus's proteins use the host environment, interact among themselves and with the host proteins, causing loss of function or even the death of the cells. Therefore the complete elucidation of the whole set of such interactions is a crucial step for the comprehension of viruses. Understanding the interplay between host and virus proteins is relevant since it may help identify virus-related diseases and potential targets for therapeutic strategies. The interactomics information would help one to understand the global mechanistic processes of the viral molecular machinery during the viral infection, survival within the host, and replication. With this knowledge, one could discern the protein interactions that are crucial for transmission and replication. These interactions could then be potential candidates for inhibitory drugs. Furthermore, using the network biology approach, one could identify hubs and bottlenecks, new to SARS-COV-2 , that could again be targeted by the antiviral drugs [28].

Experimentally validated physical interactions between intra-viral [87] and human-SARS-COV-2 proteins usually determined using Affinity Purification Mass Spectrometry (AP-MS) [18], Affinity Capture-Western [88], Reconstituted Complex [89] and Yeast Two-Hybrid (Y2H) methods [87]. For

 doi:10.20944/preprints202008.0320.v1

instance, AP-MS based SARS-COV-2 -host interactomes are reported for 26 SARS-COV-2 proteins with 332 host proteins [18]. The study aimed to identify possible drug targets. Therefore they isolate 66 possible drug targets in human proteins suggesting potential 69 compounds (of which, 29 drugs are approved by the US Food and Drug Administration, 12 are in clinical trials and 28 are pre-clinical compounds).

It is a time-consuming and expensive task for elucidating experimentally validated complete host protein interactions with viral proteins. Hence, the in-silico prediction is the only alternative. Some studies presented the investigation of virus-host interactomes using tools and methodologies coming from graph theory [90, 21], demonstrating the importance of studying virus-host interplay at network level [91, 92, 93, 94, 95]. Data related to the interactions (or functional associations) among biologically relevant macromolecules (e.g., proteins, genes, etc.) are usually modeled using graph theory and its related formalism [96, 97]. Consequently, biological entities are represented as nodes, while edges models their associations [98]. Such networks may contain a single kind of molecule, such as protein-protein interactions (PPI), or gene-gene interactions [17, 99]. More recently, it has been shown that biological processes are constituted by the synergistic interplay of different molecules (i.e., genes, non-coding RNA, proteins, mi-RNA, etc.) [100].

Consequently, novel models that integrate such different aspects and describe the interplay of the heterogeneous actor have been introduced. Accordingly, the use of more complex network models comprising different nodes and various associations among them is growing [101, 102]. The SARS-COV-2 scenario, as we describe in the follows, also contains such models (e.g., [8]). We represent a summarised view on in-silico interactome graph inference workflow that involves different available interactome, and omics data sources are shown in Figure 5.

From a bioinformatics point of view, few important queries exist: whether the proteins infected by viruses central or peripheral (i.e., are the infected proteins hub or not)? Do all of the viruses attach to similar proteins (from a network point of view)?. What happens in an infected host interactome? [7]. This consideration guided the first attempts to produce an interactome wide map of SARS-COV-2 proteins and their interaction with human proteins enabling scientists to answer the above questions. Interactions of viral proteins may be categorized in two main classes:(i) **Intra-Viral Interactions**, i.e., interactions that occur among viral proteins that are in general limited and easy to determine; **Host-Virus Interactions**, i.e., interactions that occur

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 14 August 2020 doi:10.20944/preprints202008.0320.v1

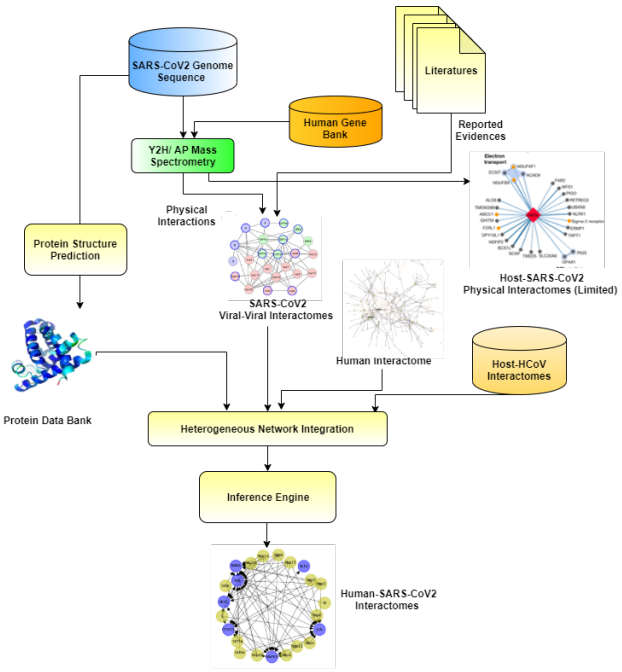

Figure 5: Data integration process to build a Host-SARS-CoV-2 Interactome graph. The building of the integrated host-viral interactome starts with the analysis of the viral genome. Then viral proteins and the interactions with host protein are determined. The determination of all the interactions is often performed by integrating experimental evidence and literature. In parallel, the structure of proteins is also predicted. All this information (structure, virus-host interactions, and viral interaction) is integrated using heterogeneous networks. The final product of the process is the desired interactome.

among viral and host proteins that may be potentially a large number. The main challenges in this area are represented by the different speeds existing between the spreading of SARS-COV-2 and the time needed for the wet-lab experiments. Therefore, all the approaches we discuss below integrate both in-silico and wet-lab experiments.

One of the first approaches of building a SARS-COV-2 interactome is described in [28]. The authors derived the first map of both intra-viral and host-viral proteins using a bioinformatics approach based on the homology among SARS-COV-2 and the previous 2002 SARS-COV virus. The hypothesis underlying the work is that the similarity among two viruses is also preserved at the interactome level. Thus many interactions of 2002 SARS-COV may be preserved in the SARS-COV-2 . Consequently, they derived a whole

 doi:10.20944/preprints202008.0320.v1

SARS-COV-2 interactome containing both intra-viral and virus-host interactions. The authors derived a 2002 SARS-COV interactome by analyzing public literature data. Such data are integrated with a genome-wide analysis through Y2H on SARS-COV ORFeome, obtaining a resulting intra-viral interaction network consisted of 31 proteins and 86 unique interactions. Then the authors used both Y2H interaction data and literature mining to derive the viral-host interactions. The final virus-host interaction network consisted of 118 proteins, 93 host proteins, and 114 unique virus-host interactions.

Multiple interactome analysis is another popular way to integrate the knowledge derived from heterogeneous protein or gene networks. In a similar attempt [103] integrated protein-protein interactions and gene expression data are derived from literature and public databases. It started with data related to three existing viruses (SARS-COV, MERS-COV, HCOV-229E) to infer the interactome of SARS-COV-2 . It also integrated an additional PPI database to reconstruct the action of SARS-COV-2 on the proteome level, obtaining a network consisting of 13,020 nodes and 71,496 interactions. In parallel, the authors inferred a gene co-expression network using Random walk with restart (RWR) algorithm and using S-glycoproteins of SARS-COV, MERS-COV, and HCOV-229E as seeds. Similarly, to build the SARS-COV-2 interactome phylogenetically close HCOV–host interactome network was built by assembling four known HCOVs (SARS-COV, MERS-COV, HCOV-229E, and HCOV-NL63), one mouse MHV, and one avian IBV (N protein).

As a novel attempt [104], the codon usage pattern is used to infer possible interactions between 26 SARS-COV-2 proteins and selective host proteins involved in 17 major cell signaling pathways. They used the RSCU score as a measure of codon usage bias to assess proximity between a pair of host and viral proteins. MAPK pathway is highlighted as the worst affected pathways during COVID-19 .

5. Chest Image Analysis

Image analysis transforms digital images into measurements that describe the meaningful information from every cell of images. SARS is a respiratory virus and found to infect primarily lungs leads to death. Chest image analysis may help in the early diagnosis of COVID-19 patient. Two kinds of chest radiography image obtained through X-ray and CT (advanced X-ray machine) scanners is usually effective in diagnosing pneumonia is now recommended for COVID-19 patients. Both technologies are significantly useful

 doi:10.20944/preprints202008.0320.v1

in large-scale screening and disease diagnosis. It is widely available and provides images for diagnosis quickly. Data science is an indispensable tool for the automatic (or semi-automatic) analysis of large amounts of data that require substantial quantitative assessment and computation. The chest image data are utilized for multiple possible use cases such as predicting the need for the ICU, predicting patient survival, and understanding a patient's trajectory during treatment and so on [105, 1]. Imaging is fast, non-invasive, relatively cheap, and potentially bedside to monitor the progression of the disease [106, 105]. The ultimate goals are many folds such as improving patient health care, biomarker design for the COVID-19 , and, most importantly, early and automatic detection of COVID-19 .

Continuous development of machine learning-based tools is in progress for the detection, quantification, and monitoring of COVID-19 disease and distinguishing non-infected individuals [107]. The majority of the study explores the power of deep convolutional neural networks (CNN). COVID-Net [106] is a CNN based first open-source framework designed based on deep learning techniques for the detection of COVID-19 by analyzing chest X-ray images, where authors developed COVIDx, an open-access benchmark dataset comprising of 13,975 CXR images across 13,870 patient cases. Model performance is evaluated with deep neural network architectures (VGG-19 and ResNet-50) for comparative purposes. The model predicts three possible outcomes from input image data: a) no infection (normal), b) non-COVID19 infection (e.g., viral, bacterial, etc.), and c) COVID-19 viral infection.

A transfer learning-based convolution neural network [1, 108], also applied for detecting various abnormalities in small medical image datasets. The authors collected 1427 X-rays images (224 COVID-19 , 700 common pneumonia, and 504 normal cases) from several sources like Cohen [17], Radiological Society of North America (RSNA), Radiopaedia, and Italian Society of Medical and Interventional Radiology (SIRM) [18]. The CNN models employed for the task are VGG19, Mobile Net, Inception, Xception, and Inception ResNet v2.

Pulmonary CT images and deep learning techniques aimed to establish an early screening model to distinguish COVID-19 pneumonia and Influenza-A viral pneumonia from healthy cases [109]. At first, candidate infection regions

[17] https://github.com/ieee8023/covid-chestxray-dataset
[18] https://www.kaggle.com/andrewmvd/convid19-xrays

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 14 August 2020 doi:10.20944/preprints202008.0320.v1

are isolated using Residual CNN (ResNet-23) from the pulmonary CT image set. ResNet-18 architecture is used for image feature extraction. It uses a location-attention classification model to categorize COVID-19, Influenza-A, and non-symptomatic groups. Infection type with probability scores is calculated using Noisy-or Bayesian function.

A combined CNN-LSTM architecture is also explored to detect infected patients X-ray images [43]. CNN is used for the purpose of feature extraction, whereas LSTM is used for prediction.

Inf-Net [110] is a CNN based lung segmentation technique for automatic segmentation of COVID-19 effected lung CT images. The high resolution and low-level features are extracted with the help of two-level convolution layers, which again passed through the next three convolution layers for extracting high-level features. All the features are then aggregated using a parallel partial decoder (PPD) to generate a global feature set. Implicit reverse attention (RA) and explicit edge-attention are used for the enhancement of the representations. The sigmoid activation function is used on the RA output for final segmentation. A semi-supervised framework is also proposed for dealing with small, manually segmented labeled images.

Due to the unavailability of adequate sample images for training, synthetically chest X-ray (CXR) images are generated using Auxiliary Classifier Generative Adversarial Network (ACGAN) based model, called CovidGAN [42].

Statistical analysis is performed on the CT image [111], based on the number of key imaging features like affected lobes, the presence of ground-glass nodules (GGO), patchy/punctate ground-glass opacities, patchy consolidation, fibrous stripes, and irregular solid nodules in each patient's chest image. The study reveals that manifestations of COVID-19 are diverse (imaging features) and changing rapidly.

A binary classifier has been designed to segregate COVID-19 affected chest X-ray images[112]. It uses Manta-Ray Foraging Optimization (MRFO) for feature extraction and K Nearest Neighbours (KNN) for classification.

6. Epidemiological Data Analysis

From the beginning of the outbreak, a significant data source was related to the observation of novel cases to predict the evolution of the diffusion first. Such data were used to run both existing, and ad hoc developed models [113, 33].

 doi:10.20944/preprints202008.0320.v1

The main aims of these approaches are: (i) controlling the diffusion of the SARS-COV-2 ; (ii) supporting healthcare providers to allocate resources (e.g., planning intensive care units places); (iii) evaluating the impact of containment measures.

From a data science point of view, almost all of these works use real data to build and fits predictive or observatory models. These works were published mainly in the first weeks after the outbreak in Wuhan. They usually use deterministic models based on classical epidemiological studies. Consequently, real data are used to derive parameters of models based on ordinary differential equations [114, 115]. The diffusion of such works has been very high; for instance, simple queries on Google Scholar or on preprint servers will return more than 2,000 papers.

In [116] authors integrated information of existing data sources provided by the Johns Hopkins University, World Health Organization, Chinese Center for Disease Control and Prevention, National Health Commission, and Dingxiangyuan (DXY, a Chinese epidemiological database). They perform an exploratory data analysis, using mainly visualization techniques to highlight (and stimulating discussion) differences of different reported cases (e.g., infected, dead, and recovered), in many countries.

Moreover, more sophisticated models tried to integrate epidemiological data with other data to study the impact of other variables (e.g., environmental/geographical variables and patient-related variables). In such cases, authors use sometime simple descriptive and inferential statistics (e.g. [117]) as well as complex data mining techniques [118].

A COVID-19 outbreak forecasting model is developed using Long short-term memory (LSTM) networks [119]. John Hopkins University and Canadian health data are used to extract key features to predict the trends and possible stopping time of the current COVID-19 outbreak within the world.

Based on the context of lockdown duration and social distancing and its impact in controlling COVID-19 spread in India, a statistical analysis is performed based on an age-structured SIR model with social contact matrices obtained from surveys and Bayesian imputation [120].

7. Drug Repurposing and Target Prediction

Drug discovery aims to identify new small molecules that potentially modulate the functions of target proteins. The development of a new drug

 doi:10.20944/preprints202008.0320.v1

molecule for COVID-19 is a time consuming and costly affair. In the COVID-19 era, the long process for the determination of a novel drug is not suitable, looking into the rapid spread of the virus. It is of utmost importance to identify in a faster way the anti-viral drugs that may control the adverse effect of COVID-19, thereby reducing the mortality rate. The best alternative is to look for already (FDA-Food and Drug Administration) approved drugs that may bind with the therapeutic target proteins (viral or host). Identifying the therapeutic target responsible for the observed phenotype is equally important [121] for the same. Data analysis for discovering possible drug candidates from the existing drugs is a well-know process referred to as *drug repurposing*. It involves the identification of new uses for approved (or experimental) drugs for novel pathologies. The process, as depicted in Figure 6, is based on the integration of molecular data (e.g., interactomes, co-expression networks), concerning the existing drug-disease association.

Availability high-resolution proteomics, interactomics, drug-target association data, it is now possible to search quickly for a suitable small molecule in-silico with the help of advanced (deep) neural network models. A good number of deep learning-based drug-target association and repurposing tools are available for other viral diseases that can be used for COVID-19 data analysis (Table **??**).

DeepPurpose [122] is one such tool applied to COVID-19 drug prioritization by leveraging the multiple deep neural network models. DeepPurpose uses SMILES string for drugs and amino acid sequences for the target as input. It uses different convolutional neural network models, namely the convolutional recurrent neural network (CNN-RNN), Transformer encoders, and Message-Passing Neural Network for encoding input strings. It ensembles seven encodings for proteins and eight encodings for drugs. The output of DeepPurpose is a score that measures the binding activity of the input drug target pair.

Recent trends adopt different network-embedding techniques [123, 5] and the help of deep learning networks producing a list of possible candidate drugs as output confirmed in wet-lab experiments or clinical trials. It should be noted that after the *in-silico* identification, the process of drug repurposing requires time and funds for the subsequent clinical trials, but the overall time is lower than the development of a new chemical [124].

Gordon et al. .[18] used their experimentally validated host-viral network consisting of 26 viral protein and 332 host proteins and identified 66 human proteins targeted by 69 existing drug compounds constituting potential drug

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 14 August 2020 doi:10.20944/preprints202008.0320.v1

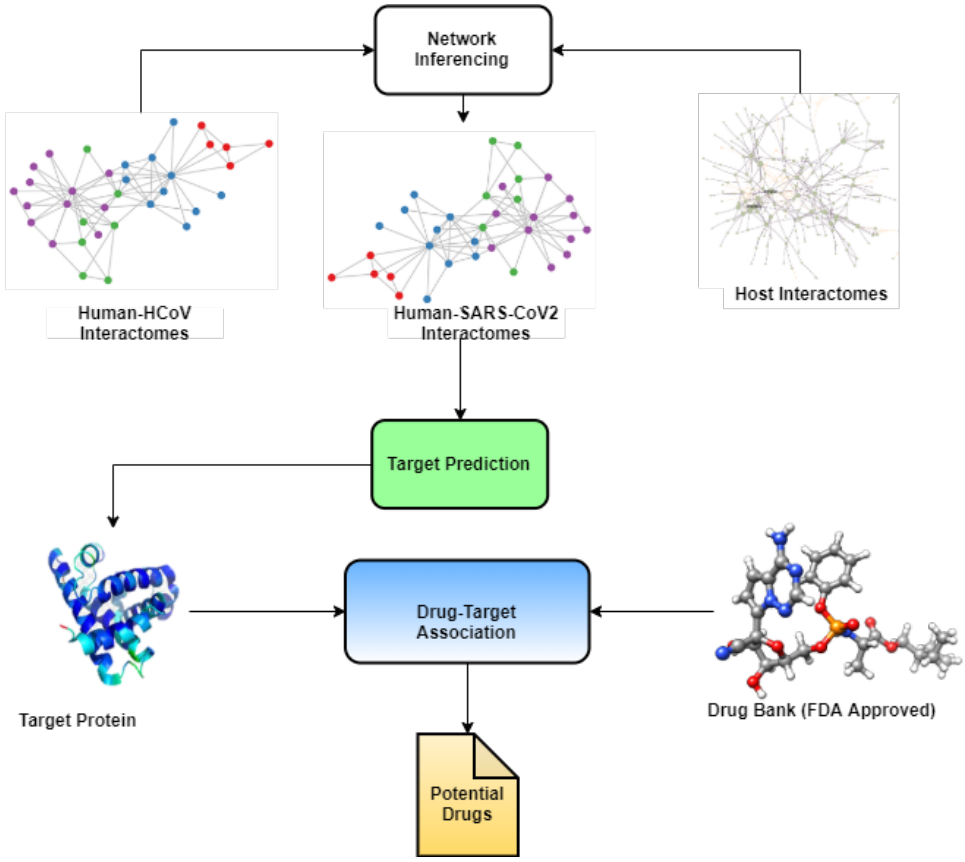

Figure 6: Drug Repurposing Process. The process, is based on the integration of molecular data and drug-disease association. The analysis is often performed by deep-learning or network embedding. The output list of candidate drugs is then confirmed in wet lab experiments or clinical trials.

targets to treat COVID-19 .

Multiple network-based strategies [5] coupled with Graph Convolution Network (GCN) is explored to rank drug repurposing candidates. At first, COVID-19 interactome modules are identified, considering 56 different human tissues. Existing drug molecules are then prioritized using a proximity measure based on their ability to interact with their protein targets present in the modules. A GCN is used to combine multiple sources of evidence for drug repurposing. The manifestation of COVID-19 in other tissues, such as the reproductive system, brain regions, and neurological co-morbidities, are also predicted during the study. In a similar heterogeneous network-

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 14 August 2020 doi:10.20944/preprints202008.0320.v1

integrated drug repurposing [123] approach, network proximity analyses performed on drug targets and HCOV–host interactions and prioritized 16 potential anti-HCOV repurposable drugs. They used host proteins from four known HCOVs (SARS-COV, MERS-COV, HCOV-229E, and HCOV-NL63) based on phylogeny analysis and performed functional enrichment followed by drug association analysis.

Another network-based approach for deriving possible drug targets is attempted [8], where both protein interaction and gene coexpression networks are used to identify master regulator [125] involved during SARS-COV-2 infection. Physical interactions of proteins were extracted from [28]. Co-expression network is generated using SARS-COV-2 -human interactome proteins derived from [126] and the largest human lung RNA-Seq dataset available from the GTEX [19] consortium. The authors identified some key proteins involved during an infection such as ACE2, TMPRSS, and MOCK. They suggested to them as potential therapeutics or vaccine targets and highlighted the evidence that COVID-19 is characterized by a large inflammation process that is not limited to respiratory apparatus only.

8. Summary and Challenges

The potential applications for data-science (and deep-learning and artificial intelligence) to improve the research related to SARS-COV-2 outbreak are countless. However, due to the tremendous spread of the virus and the birth of novel approaches worldwide, it seems that the application of data science may fail or that results are far from successful. Therefore the need for a comprehensive vademecum for practitioners as well as of a landscape for researchers is high. Consequently, we provided an in-depth overview of the data sources and methods that are currently used to elucidate the primary mechanism of pathogenesis and development of COVID-19 . We included almost all data types, from the molecular scale to patient (medical imaging) and population-scale (epidemiological data). We discussed the main approaches for modeling SARS-COV-2 infection, drug repurposing, population surveillance, disease, and treatment. An overall summary of data science models, types of tasks and data, and various software tools are reported in Table **??**. We also delineated the important challenges for data science applications in

[19] `www.gtexportal.org`

Table 3: Data Science tools and techniques for SARS-COV-2 data analysis.

Task	Data Type	Data Science Models	Available Tools
Phylogeny/ alignment	Nucleotide/ Protein sequence	UPGMA, WPGMA, Neighbour-joining, Maximum likelihood, Fitch–Margoliash method, Maximum parsimony, Bayesian inference	ClustalW, Clustalω, MAFFT, MUSCLE, T-Coffee [20] DNAMAN [21]
Structure Prediction	Protein sequence	Deep neural-network (NeBcon, ResPRE, ResTriplet, and TripletRes), QSQE, Supervised machine learning (SVM), Multiple regression	SWISS-MODEL [82], PyMOL [83], I-Tasser [84], COMPOSER [81]
SARS-COV-2 predictor	Nucleotide sequence	Conventional models (Naïve Bayes, K-Nearest Neighbors, Artificial Neural Networks, Decision tree and Support Vector Machine), Deep models (Convolutional Neural Networks (CNN), Bi-path CNN (BiPathCNN)	COVID-Predictor [59]-
Protein Interactions	Protein sequence, PPI Networks, Protein structure	Graph analysis	Cytoscape [22]
Chest imaging analysis	Chest x-ray or CT image	Deep learning model (VGG19, Mobile Net, Inception, Xception and Inception ResNet (v2,18,23,50), GAN, Dice similarity coefficient (DSC)	TrainingData.io [23]
Epidemic Trend Analysis	Experimental and observational	LSTM Statistical models (SIR, Bayesian imputation, linear and polynomial regression)	Worldometers-coronavirus [24] WHO-COVID19-report [25] COVID-19 Projections [26]
Drug Interaction & Repurposing	Protein sequence, Drug molecules, Protein Structure	Graph Analysis, Graphical Convolution Network	DeepDR [127], DeepPurpose [122], Deepdta [128], kGCN [129], DeepChem [130], Graphdta [131] D3Targets-2019-nCoV [132], CoVex [133]

cardiovascular care, including the need to introduce more standards and the integration of data. If data science is successfully implemented in this area, it will fulfill its potential as an essential component of fighting COVID-19 .

8.1. Current Challenges

- **Ontology-based Federation of Data:** The current scenario is characterized by many data formats that differ both in schemas and use; therefore the introduction of a novel mechanism of federation able to integrate data both horizontally and vertically, possibly mediated by ad hoc ontologies;

- **Development of graph-based models:** The integration of many data into a single model (comprising patients, molecular and clinical characteristics) may support the individuation of spreading diffusion and the realization of more *models*;

 doi:10.20944/preprints202008.0320.v1

- **Leveraging the use of efficient and high-throughput analysis workflows:** The rapid spreading of virus and the unprecedented production of data need the introduction of novel efficient and high-throughput based analysis environments, possibly structured as virtual laboratories federated by cloud infrastructures;

- **Need attention worldwide mutation case for structure prediction:** Due to rapid mutations (non-synonymous) in SARS-COV-2 or any other viruses, it may alter their protein structures. Structure-based drug development depends on the structural coherence between the drug molecule and target proteins, hence the dynamic variation of viral structures is essential for stable anti-viral drug development. Machine learning-based possible strain variation prediction may help early decision in designing or reusing effective anti-viral drugs.

- **Low data deep models for drug discovery:** Often the success of deep networks relies on large sample data for training. In reality, it is always not possible to generate such large scale true samples. Synthetic sample data (X-ray) are indeed generated using Generative Adversarial Networks. However, for network-based drug discovery, it may not be feasible to generate such synthetic data. In a recent attempt one-shot LSTM framework [134] has been proposed [135] for repurposed drug discovery in presence low data [136]. A similar method is yet to develop for COVID-19 .

- **Explainable AI support for the more reliable diagnostic system:** Diagnosis and drug discovery are the most sensitive task and demand very high accuracy. Due to a similar phenotype of COVID-19 infection with pneumonia, it is challenging to differentiate early symptoms of COVID-19 chest infection. Explainable AI [137] is a new concept where the reliability of a learning system can be interpreted and visualized for confidence generation about the outcome instead of treating it as a black box. Explainable AI may be incorporated with COVID-19 image-based clinical diagnostic system for better reality and confidence for early prediction. The same idea may be propagated towards drawing accurate drug-target associations.

Acknowledgements

The work is partially carried out at NetRA Lab, Sikkim University with the support of Department of Science & Technology (DST), Govt. of India under DST-ICPS Data Science program [DST/ICPS/Cluster/Data Science/General],

References

[1] Ioannis D Apostolopoulos and Tzani A Mpesiana. Covid-19: automatic detection from x-ray images utilizing transfer learning with convolutional neural networks. *Physical and Engineering Sciences in Medicine*, page 1, 2020.

[2] Martina Bianchi, Domenico Benvenuto, Marta Giovanetti, Silvia Angeletti, Massimo Ciccozzi, and Stefano Pascarella. Sars-cov-2 envelope and membrane proteins: Structural differences linked to virus characteristics? *BioMed Research International*, 2020, 2020.

[3] Maria Effenberger, Andreas Kronbichler, Jae Il Shin, Gert Mayer, Herbert Tilg, and Paul Perco. Association of the covid-19 pandemic with internet search volumes: a google trendstm analysis. *International Journal of Infectious Diseases*, 2020.

[4] Pietro H Guzzi, Marco Mina, Concettina Guerra, and Mario Cannataro. Semantic similarity analysis of protein data: assessment with biological features and issues. *Briefings in bioinformatics*, 13(5):569–585, 2012.

[5] Deisy Morselli Gysi, Ítalo Do Valle, Marinka Zitnik, Asher Ameli, Xiao Gan, Onur Varol, Helia Sanchez, Rebecca Marlene Baron, Dina Ghiassian, Joseph Loscalzo, et al. Network medicine framework for identifying drug repurposing opportunities for covid-19. *arXiv preprint arXiv:2004.07229*, 2020.

[6] World Health Organization et al. *Human viruses in water, wastewater and soil. Report of a WHO Scientific Group.* Geneva, Switzerland, 1979.

[7] Mario Cannataro, Pietro Hiram Guzzi, and Alessia Sarica. Data mining and life sciences applications on the grid. *WIREs Data Mining and Knowledge Discovery*, 3(3):216–238, 2013.

[8] Pietro H Guzzi, Daniele Mercatelli, Carmine Ceraolo, and Federico M Giorgi. Master regulator analysis of the sars-cov-2/human interactome. *Journal of clinical medicine*, 9(4):982, 2020.

[9] Yuntao Wu, Wenzhe Ho, Yaowei Huang, Dong-Yan Jin, Shiyue Li, Shan-Lu Liu, Xuefeng Liu, Jianming Qiu, Yongming Sang, Qiuhong Wang, et al. Sars-cov-2 is an appropriate name for the new coronavirus. *The Lancet*, 395(10228):949–950, 2020.

[10] Kristian G Andersen, Andrew Rambaut, W Ian Lipkin, Edward C Holmes, and Robert F Garry. The proximal origin of sars-cov-2. *Nature medicine*, 26(4):450–452, 2020.

[11] Alexandra L Phelan, Rebecca Katz, and Lawrence O Gostin. The novel coronavirus originating in wuhan, china: challenges for global health governance. *Jama*, 323(8):709–710, 2020.

[12] Zunyou Wu and Jennifer M McGoogan. Characteristics of and important lessons from the coronavirus disease 2019 (covid-19) outbreak in china: summary of a report of 72 314 cases from the chinese center for disease control and prevention. *Jama*, 323(13):1239–1242, 2020.

[13] Luca Steardo, Luca Steardo Jr, Robert Zorec, and Alexei Verkhratsky. Neuroinfection may contribute to pathophysiology and clinical manifestations of covid-19. *Acta Physiologica*, page e13473, 2020.

[14] Ying-Ying Zheng, Yi-Tong Ma, Jin-Ying Zhang, and Xiang Xie. Covid-19 and the cardiovascular system. *Nature Reviews Cardiology*, 17(5):259–260, 2020.

[15] Deepak Atri, Hasan K Siddiqi, Joshua Lang, Victor Nauffal, David A Morrow, and Erin A Bohula. Covid-19 for the cardiologist: a current review of the virology, clinical epidemiology, cardiac and other clinical manifestations and potential therapeutic strategies. *JACC: Basic to Translational Science*, 2020.

[16] Pragya D Yadav, César G Albariño, Dimpal A Nyayanit, Lisa Guerrero, M Harley Jenks, Prasad Sarkale, Stuart T Nichol, and Devendra T Mourya. Equine encephalosis virus in india, 2008. *Emerging infectious diseases*, 24(5):898, 2018.

[17] Mario Cannataro, Pietro H Guzzi, and Pierangelo Veltri. Protein-to-protein interactions: Technologies, databases, and algorithms. *ACM Computing Surveys (CSUR)*, 43(1):1–36, 2010.

[18] David E Gordon, Gwendolyn M Jang, Mehdi Bouhaddou, Jiewei Xu, Kirsten Obernier, Kris M White, Matthew J O'Meara, Veronica V Rezelj, Jeffrey Z Guo, Danielle L Swaney, et al. A sars-cov-2 protein interaction map reveals targets for drug repurposing. *Nature*, pages 1–13, 2020.

[19] Pietro Hiram Guzzi and Swarup Roy. *Biological Network Analysis: Trends, Approaches, Graph Theory, and Algorithms*. Academic Press, 2020.

[20] B De Chassey, V Navratil, Lionel Tafforeau, MS Hiet, A Aublin-Gex, S Agaugue, G Meiffren, F Pradezynski, BF Faria, T Chantier, et al. Hepatitis c virus infection protein network. *Molecular systems biology*, 4(1):230, 2008.

[21] Michael A Calderwood, Kavitha Venkatesan, Li Xing, Michael R Chase, Alexei Vazquez, Amy M Holthaus, Alexandra E Ewence, Ning Li, Tomoko Hirozane-Kishikawa, David E Hill, et al. Epstein–barr virus and virus human protein interaction maps. *Proceedings of the National Academy of Sciences*, 104(18):7606–7611, 2007.

[22] Caroline C Friedel and Jurgen Haas. Virus–host interactomes and global models of virus-infected cells. *Trends in microbiology*, 19(10):501–508, 2011.

[23] Andrew Chatr-Aryamontri, Arnaud Ceol, Daniele Peluso, Aurelio Nardozza, Simona Panni, Francesca Sacco, Michele Tinti, Alex Smolyar, Luisa Castagnoli, Marc Vidal, et al. Virusmint: a viral protein interaction database. *Nucleic acids research*, 37(suppl_1):D669–D673, 2008.

[24] Helen Cook, Nadezhda Doncheva, Damian Szklarczyk, Christian von Mering, and Lars Jensen. Viruses. string: A virus-host protein-protein interaction database. *Viruses*, 10(10):519, 2018.

[25] Mais G. Ammari, Cathy R. Gresham, Fiona M. McCarthy, and Bindu Nanduri. Hpidb 2.0: a curated database for host–pathogen interactions. *Database*, 2016, 07 2016.

[26] Alberto Calderone, Luana Licata, and Gianni Cesareni. Virusmentha: a new resource for virus-host protein interactions. *Nucleic acids research*, 43(D1):D588–D592, 2014.

[27] Thibaut Guirimand, Stéphane Delmotte, and Vincent Navratil. Virhostnet 2.0: surfing on the web of virus/host molecular interactions data. *Nucleic acids research*, 43(D1):D583–D587, 2014.

[28] Suhas Srinivasan, Hongzhu Cui, Ziyang Gao, Ming Liu, Senbao Lu, Winnie Mkandawire, Oleksandr Narykov, Mo Sun, and Dmitry Korkin. Structural genomics of sars-cov-2 indicates evolutionary conserved functional regions of viral proteins. *Viruses*, 12(4):360, 2020.

[29] Livia Perfetto, Chiara Pastrello, Noemi Del-Toro, Margaret Duesbury, Marta Iannuccelli, Max Kotlyar, Luana Licata, Birgit Meldal, Kalpana Panneerselvam, Simona Panni, et al. The imex coronavirus interactome: an evolving map of coronaviridae-host molecular interactions. *BioRxiv*, 2020.

[30] Shuchang Zhou, Yujin Wang, Tingting Zhu, and Liming Xia. Ct features of coronavirus disease 2019 (covid-19) pneumonia in 62 patients in wuhan, china. *American Journal of Roentgenology*, 214(6):1287–1294, 2020.

[31] Ensheng Dong, Hongru Du, and Lauren Gardner. An interactive web-based dashboard to track covid-19 in real time. *The Lancet infectious diseases*, 20(5):533–534, 2020.

[32] Italian Data. `https://github.com/pcm-dpc/COVID-19`.

[33] Bo Xu, Moritz UG Kraemer, Bernardo Gutierrez, Sumiko Mekaru, Kara Sewalk, Alyssa Loskill, Lin Wang, Emily Cohn, Sarah Hill,

Alexander Zarebski, et al. Open access epidemiological data from the covid-19 outbreak. *The Lancet Infectious Diseases*, 20(5):534, 2020.

[34] Peter Imming, Christian Sinning, and Achim Meyer. Drugs, their targets and the nature and number of drug targets. *Nature reviews Drug discovery*, 5(10):821–834, 2006.

[35] Clémence Réda, Emilie Kaufmann, and Andrée Delahaye-Duriez. Machine learning applications in drug development. *Computational and Structural Biotechnology Journal*, 18:241–252, 2020.

[36] Bo Xu, Bernardo Gutierrez, Sumiko Mekaru, Kara Sewalk, Lauren Goodwin, Alyssa Loskill, Emily L Cohn, Yulin Hswen, Sarah C Hill, Maria M Cobo, et al. Epidemiological data from the covid-19 outbreak, real-time case information. *Scientific data*, 7(1):1–6, 2020.

[37] Bradley Voytek. Social media, open science, and data science are inextricably linked. *Neuron*, 96(6):1219–1222, 2017.

[38] Swarup Roy, Pooja Sharma, Keshab Nath, Dhruba K Bhattacharyya, and Jugal K Kalita. Pre-processing: A data preparation step. *Encyclopedia of Bioinformatics and Computational Biology: ABC of Bioinformatics*, page 463, 2018.

[39] Monica Jha, Pietro Hiram Guzzi, Pierangelo Veltri, and Swarup Roy. Functional module extraction by ensembling the ensembles of selective module detectors. *International Journal of Computational Biology and Drug Design*, 12(4):345–361, 2019.

[40] Jürgen Schmidhuber. Deep learning in neural networks: An overview. *Neural networks*, 61:85–117, 2015.

[41] Thomas N Kipf and Max Welling. Semi-supervised classification with graph convolutional networks. *arXiv preprint arXiv:1609.02907*, 2016.

[42] Abdul Waheed, Muskan Goyal, Deepak Gupta, Ashish Khanna, Fadi Al-Turjman, and Plácido Rogerio Pinheiro. Covidgan: Data augmentation using auxiliary classifier gan for improved covid-19 detection. *IEEE Access*, 8:91916–91923, 2020.

[43] Md Zabirul Islam, Md Milon Islam, and Amanullah Asraf. A combined deep cnn-lstm network for the detection of novel coronavirus (covid-19) using x-ray images. *medRxiv*, 2020.

[44] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In *Advances in neural information processing systems*, pages 2672–2680, 2014.

[45] Samit Ghosal, Sumit Sengupta, Milan Majumder, and Binayak Sinha. Linear regression analysis to predict the number of deaths in india due to sars-cov-2 at 6 weeks from day 0 (100 cases - march 14th 2020). *Diabetes Metabolic Syndrome: Clinical Research Reviews*, 14(4):311 – 315, 2020.

[46] Swarup Roy and Dhruba K Bhattacharyya. An approach to find embedded clusters using density based techniques. In *International Conference on Distributed Computing and Internet Technology*, pages 523–535. Springer, 2005.

[47] Carmine Ceraolo and Federico M. Giorgi. Genomic variance of the 2019-ncov coronavirus. *Journal of Medical Virology*, 92(5):522–528, 2020.

[48] JM Gonzalez, P Gomez-Puertas, D Cavanagh, AE Gorbalenya, and Luis Enjuanes. A comparative sequence analysis to revise the current taxonomy of the family coronaviridae. *Archives of virology*, 148(11):2207–2235, 2003.

[49] Jila Yavarian, Farshid Rezaei, Azadeh Shadab, Mahmood Soroush, Mohammad Mehdi Gooya, and Talat Mokhtari Azad. Cluster of middle east respiratory syndrome coronavirus infections in iran, 2014. *Emerging infectious diseases*, 21(2):362, 2015.

[50] Zoltan Penzes, Jose M González, Enrique Calvo, Ander Izeta, Cristian Smerdou, Ana Méndez, Carlos M Sánchez, Isabel Sola, Fernando Almazán, and Luis Enjuanes. Complete genome sequence of transmissible gastroenteritis coronavirus pur46-mad clone and evolution of the purdue virus cluster. *Virus genes*, 23(1):105–118, 2001.

[51] Canrong Wu, Yang Liu, Yueying Yang, Peng Zhang, Wu Zhong, Yali Wang, Qiqi Wang, Yang Xu, Mingxue Li, Xingzhou Li, et al. Analysis of therapeutic targets for sars-cov-2 and discovery of potential drugs by computational methods. *Acta Pharmaceutica Sinica B*, 2020.

[52] Tao Zhang, Qunfu Wu, and Zhigang Zhang. Probable pangolin origin of sars-cov-2 associated with the covid-19 outbreak. *Current Biology*, 2020.

[53] Andrew M Waterhouse, James B Procter, David MA Martin, Michèle Clamp, and Geoffrey J Barton. Jalview version 2—a multiple sequence alignment editor and analysis workbench. *Bioinformatics*, 25(9):1189–1191, 2009.

[54] Kazutaka Katoh and Daron M Standley. Mafft multiple sequence alignment software version 7: improvements in performance and usability. *Molecular biology and evolution*, 30(4):772–780, 2013.

[55] DDR Turista, A Islamy, VD Kharisma, and ANM Ansori. Distribution of covid-19 and phylogenetic tree construction of sars-cov-2 in indonesia. *J Pure Appl Microbiol*, 14(suppl 1):1035–1042, 2020.

[56] Changchuan Yin. Genotyping coronavirus sars-cov-2: methods and implications. *Genomics*, 2020.

[57] Yang Gao, Tao Li, and Liaofu Luo. Phylogenetic study of 2019-ncov by using alignment-free method. *arXiv preprint arXiv:2003.01324*, 2020.

[58] Francesco Ballesio, Ali Haider Bangash, Didier Barradas-Bautista, Justin Barton, Andrea Guarracino, Lukas Heumos, Aneesh Panoli, Marco Pietrosanto, Anastasios Togkousidis, Phillip Davis, et al. Determining a novel feature-space for sars-cov2 sequence data.

[59] Jnanendra Prasad Sarkar, Indrajit Saha, Arijit Seal, and Debasree Maity. Covid-predictor: Rna sequence based prediction of coronavirus. 2020.

[60] Ayal B Gussow, Noam Auslander, Guilhem Faure, Yuri I Wolf, Feng Zhang, and Eugene V Koonin. Genomic determinants of pathogenicity in sars-cov-2 and other human coronaviruses. *Proceedings of the National Academy of Sciences*, 2020.

[61] Jayanta Das and Swarup Roy. Comparative analysis of human coronaviruses focusing on nucleotide variability and synonymous codon usage pattern. *BioRxiv*, 2020.

[62] Luciano Rodrigo Lopes, Giancarlo de Mattos Cardillo, and Paulo Bandiera Paiva. Molecular evolution and phylogenetic analysis of sars-cov-2 and hosts ace2 protein suggest malayan pangolin as intermediary host. *Brazilian Journal of Microbiology*, pages 1–7, 2020.

[63] Mohammed Uddin, Farah Mustafa, Tahir A Rizvi, Tom Loney, Hanan Al Suwaidi, Ahmed H Hassan Al-Marzouqi, Afaf Kamal Eldin, Nabeel Alsabeeha, Thomas E Adrian, Cesare Stefanini, et al. Sars-cov-2/covid-19: Viral genomics, epidemiology, vaccines, and therapeutic interventions. *Viruses*, 12(5):526, 2020.

[64] Xiang Li, Yuhe Song, Gary Wong, and Jie Cui. Bat origin of a new human coronavirus: there and back again. *Science China Life Sciences*, 63(3):461–462, 2020.

[65] Alireza Tabibzadeh, Farhad Zamani, Azadeh Laali, Maryam Esghaei, Fahimeh Safarnezhad Tameshkel, Hossein Keyvani, Mahin Jamshidi Makiani, Mahshid Panahi, Nima Motamed, Dhayaneethie Perumal, et al. Sars-cov-2 molecular and phylogenetic analysis in covid-19 patients: a preliminary report from iran. *Infection, Genetics and Evolution*, page 104387, 2020.

[66] Peter Forster, Lucy Forster, Colin Renfrew, and Michael Forster. Phylogenetic network analysis of sars-cov-2 genomes. *Proceedings of the National Academy of Sciences*, 117(17):9241–9243, 2020.

[67] Aditi Joshi and Sushmita Paul. Phylogenetic analysis of the novel coronavirus reveals important variants in indian strains. *BioRxiv*, 2020.

[68] Rahila Sardar, Deepshikha Satish, Shweta Birla, and Dinesh Gupta. Comparative analyses of sar-cov2 genomes from different geographical locations and other coronavirus family genomes reveals unique features potentially consequential to host-virus interaction and pathogenesis. *bioRxiv*, 2020.

[69] Manish Tiwari and Divya Mishra. Investigating the genomic landscape of novel coronavirus (2019-ncov) to identify non-synonymous mutations

for use in diagnosis and drug design. *Journal of Clinical Virology*, page 104441, 2020.

[70] Mohamed El Boujnouni. A study and identification of covid-19 viruses using n-grams with naïve bayes, k-nearest neighbors, artificial neural networks, decision tree and support vector machine. 2020.

[71] Gurjit S Randhawa, Maximillian PM Soltysiak, Hadi El Roz, Camila PE de Souza, Kathleen A Hill, and Lila Kari. Machine learning using intrinsic genomic signatures for rapid classification of novel pathogens: Covid-19 case study. *Plos one*, 15(4):e0232391, 2020.

[72] Ngoc Giang Nguyen, Vu Anh Tran, Duc Luu Ngo, Dau Phan, Favorisen Rosyking Lumbanraja, Mohammad Reza Faisal, Bahriddin Abapihi, Mamoru Kubo, Kenji Satou, et al. Dna sequence classification by convolutional neural network. *Journal of Biomedical Science and Engineering*, 9(05):280, 2016.

[73] Alejandro Lopez-Rincon, Alberto Tonda, Lucero Mendoza-Maldonado, Eric Claassen, Johan Garssen, and Aletta D Kraneveld. Accurate identification of sars-cov-2 from viral genome sequences using deep learning. *bioRxiv*, 2020.

[74] Huaiqiu Zhu, Qian Guo, Mo Li, Chunhui Wang, Zhengcheng Fang, Peihong Wang, Jie Tan, Shufang Wu, and Yonghong Xiao. Host and infectivity prediction of wuhan 2019 novel coronavirus using deep learning algorithm. *BioRxiv*, 2020.

[75] Zhencheng Fang, Jie Tan, Shufang Wu, Mo Li, Congmin Xu, Zhongjie Xie, and Huaiqiu Zhu. Ppr-meta: a tool for identifying phages and plasmids from metagenomic fragments using deep learning. *GigaScience*, 8(6):giz066, 2019.

[76] Ruchir Gupta, Jacob Charron, Cynthia Stenger, Jared Painter, Hunter Steward, Taylor Cook, William Faber, Austin Frisch, Eric Lind, Jacob Bauss, et al. Sars-cov2 (covid-19) structural/evolution dynamicome: Insights into functional evolution and human genomics. *bioRxiv*, 2020.

[77] Zhixin Liu, Xiao Xiao, Xiuli Wei, Jian Li, Jing Yang, Huabing Tan, Jianyong Zhu, Qiwei Zhang, Jianguo Wu, and Long Liu. Composition

and divergence of coronavirus spike proteins and host ace2 receptors predict potential intermediate hosts of sars-cov-2. *Journal of medical virology*, 92(6):595–601, 2020.

[78] Andrew W Senior, Richard Evans, John Jumper, James Kirkpatrick, Laurent Sifre, Tim Green, Chongli Qin, Augustin Žídek, Alexander WR Nelson, Alex Bridgland, et al. Improved protein structure prediction using potentials from deep learning. *Nature*, 577(7792):706–710, 2020.

[79] Jianyi Yang, Ivan Anishchenko, Hahnbeom Park, Zhenling Peng, Sergey Ovchinnikov, and David Baker. Improved protein structure prediction using predicted interresidue orientations. *Proceedings of the National Academy of Sciences*, 117(3):1496–1503, 2020.

[80] Lim Heo and Michael Feig. Modeling of severe acute respiratory syndrome coronavirus 2 (sars-cov-2) proteins by machine learning and physics-based refinement. *bioRxiv*, 2020.

[81] Michael J Sutcliffe, I Haneef, D Carney, and TL Blundell. Knowledge based modelling of homologous proteins, part i: Three-dimensional frameworks derived from the simultaneous superposition of multiple structures. *Protein Engineering, Design and Selection*, 1(5):377–384, 1987.

[82] Andrew Waterhouse, Martino Bertoni, Stefan Bienert, Gabriel Studer, Gerardo Tauriello, Rafal Gumienny, Florian T Heer, Tjaart A P de Beer, Christine Rempfer, Lorenza Bordoli, et al. Swiss-model: homology modelling of protein structures and complexes. *Nucleic acids research*, 46(W1):W296–W303, 2018.

[83] Giacomo Janson, Chengxin Zhang, Maria Giulia Prado, and Alessandro Paiardini. Pymod 2.0: improvements in protein sequence-structure analysis and homology modeling within pymol. *Bioinformatics*, 33(3):444–446, 2017.

[84] Jianyi Yang, Renxiang Yan, Ambrish Roy, Dong Xu, Jonathan Poisson, and Yang Zhang. The i-tasser suite: protein structure and function prediction. *Nature methods*, 12(1):7–8, 2015.

[85] Alexandra C Walls, Young-Jun Park, M Alejandra Tortorici, Abigail Wall, Andrew T McGuire, and David Veesler. Structure, function, and antigenicity of the sars-cov-2 spike glycoprotein. *Cell*, 2020.

[86] JP Romero-López, M Carnalla-Cortés, DL Pacheco-Olvera, M Ocampo, J Oliva-Ramírez, J Moreno-Manjón, B Bernal-Alferes, E García-Latorre, ML Domínguez-López, Arturo Reyes-Sandoval, et al. Prediction of sars-cov2 spike protein epitopes reveals hla-associated susceptibility. 2020.

[87] Qiming Liang, Jingjiao Li, Mingquan Guo, Xiaoxu Tian, Chengrong Liu, Xin Wang, Xing Yang, Ping Wu, Zixuan Xiao, Yafei Qu, et al. Virus-host interactome and proteomic survey of pmbcs from covid-19 patients reveal potential virulence factors influencing sars-cov-2 pathogenesis. *bioRxiv*, 2020.

[88] Ke Wang, Wei Chen, Yu-Sen Zhou, Jian-Qi Lian, Zheng Zhang, Peng Du, Li Gong, Yang Zhang, Hong-Yong Cui, Jie-Jie Geng, et al. Sars-cov-2 invades host cells via a novel route: Cd147-spike protein. *BioRxiv*, 2020.

[89] Xiaojing Chi, Xiuying Liu, Conghui Wang, Xinhui Zhang, Lili Ren, Qi Jin, Jianwei Wang, and Wei Yang. Humanized single domain antibodies neutralize sars-cov-2 by targeting spike receptor binding domain. *bioRxiv*, 2020.

[90] P. Uetz, Y. Dong, C. Zeretzke, C. Atzler, A. Baiker, B. Berger, S. Rajagopala, M. Roupelieva, D. Rose, E. Fossum, and J. Haas. Herpesviral protein networks and their interaction with the human proteome. *Science*, 311:239–242, 2006.

[91] M.D. Dyer, T.M. Murali, and B.W. Sobral. The landscape of human proteins interacting with viruses and other pathogens. *PLOS Pathog*, 4(2):e32+, 2008.

[92] Peter Uetz, Yu-An Dong, Christine Zeretzke, Christine Atzler, Armin Baiker, Bonnie Berger, Seesandra V Rajagopala, Maria Roupelieva, Dietlind Rose, Even Fossum, et al. Herpesviral protein networks and their interaction with the human proteome. *Science*, 311(5758):239–242, 2006.

[93] Marc Vidal, Michael E Cusick, and Albert-László Barabási. Interactome networks and human disease. *Cell*, 144(6):986–998, 2011.

[94] Mario Cannataro, Pietro Hiram Guzzi, Tommaso Mazza, Giuseppe Tradigo, and Pierangelo Veltri. Preprocessing of mass spectrometry proteomics data on the grid. In *18th IEEE Symposium on Computer-Based Medical Systems (CBMS'05)*, pages 549–554. IEEE, 2005.

[95] Pietro H Guzzi and Mario Cannataro. μ-cs: An extension of the tm4 platform to manage affymetrix binary data. *BMC bioinformatics*, 11(1):315, 2010.

[96] L. Cowen, T. Ideker, B. J. Raphael, and R. Sharan. Network propagation: a universal amplifier of genetic associations. *Nature Reviews Genetics*, 18:551–562, 2017.

[97] Vipin Vijayan and Tijana Milenković. Multiple network alignment via multimagna++. *IEEE/ACM transactions on computational biology and bioinformatics*, 2017.

[98] Peter J Carrington, John Scott, and Stanley Wasserman. *Models and methods in social network analysis*, volume 28. Cambridge university press, 2005.

[99] Joseph Crawford and Tijana Milenković. Cluenet: Clustering a temporal network based on topological similarity rather than denseness. *PloS one*, 13(5):e0195993, 2018.

[100] Maria Teresa Di Martino, Pietro Hiram Guzzi, Daniele Caracciolo, Luca Agnelli, Antonino Neri, Brian A Walker, Gareth J Morgan, Mario Cannataro, Pierfrancesco Tassone, and Pierosandro Tagliaferri. Integrated analysis of micrornas, transcription factors and target genes expression discloses a specific molecular architecture of hyperdiploid multiple myeloma. *Oncotarget*, 6(22):19132, 2015.

[101] Carmen Navarro, Victor Martínez, Armando Blanco, and Carlos Cano. ProphTools: general prioritization tools for heterogeneous biological networks. *GigaScience*, 6(12):1–8, 2017.

[102] Vladimir Gligorijevic, Noel Malod-Dognin, and Natasa Przulj. Integrative methods for analyzing big data in precision medicine. *Proteomics*, 16(5):741–758, 2016.

[103] Francesco Messina, Emanuela Giombini, Chiara Agrati, Francesco Vairo, Tommaso Ascoli Bartoli, Samir Al Moghazi, Mauro Piacentini, Franco Locatelli, Gary Kobinger, Markus Maeurer, et al. Covid-19: viral–host interactome analyzed by network based-approach model to study pathogenesis of sars-cov-2 infection. *Journal of Translational Medicine*, 18(1):1–10, 2020.

[104] Jayanta Das, Subhadip Chakrobarty, and Swarup Roy. Impact analysis of sars-cov2 on signaling pathways during covid19 pathogenesis using codon usage assisted host-viral protein interactions. *bioRxiv*, 2020.

[105] Joseph Paul Cohen, Paul Morrison, Lan Dao, Karsten Roth, Tim Q Duong, and Marzyeh Ghassemi. Covid-19 image data collection: Prospective predictions are the future. *arXiv preprint arXiv:2006.11988*, 2020.

[106] Linda Wang and Alexander Wong. Covid-net: A tailored deep convolutional neural network design for detection of covid-19 cases from chest x-ray images. *arXiv preprint arXiv:2003.09871*, 2020.

[107] Ophir Gozes, Maayan Frid-Adar, Hayit Greenspan, Patrick D Browning, Huangqi Zhang, Wenbin Ji, Adam Bernheim, and Eliot Siegel. Rapid ai development cycle for the coronavirus (covid-19) pandemic: Initial results for automated detection & patient monitoring using deep learning ct image analysis. *arXiv preprint arXiv:2003.05037*, 2020.

[108] Giuseppe Agapito, Mario Cannataro, Pietro Hiram Guzzi, Fabrizio Marozzo, Domenico Talia, and Paolo Trunfio. Cloud4snp: distributed analysis of snp microarray data on the cloud. In *Proceedings of the International Conference on Bioinformatics, Computational Biology and Biomedical Informatics*, pages 468–475, 2013.

[109] Charmaine Butt, Jagpal Gill, David Chun, and Benson A Babu. Deep learning system to screen coronavirus disease 2019 pneumonia. *Applied Intelligence*, page 1, 2020.

[110] Deng-Ping Fan, Tao Zhou, Ge-Peng Ji, Yi Zhou, Geng Chen, Huazhu Fu, Jianbing Shen, and Ling Shao. Inf-net: Automatic covid-19 lung infection segmentation from ct images. *IEEE Transactions on Medical Imaging*, 2020.

[111] Yueying Pan, Hanxiong Guan, Shuchang Zhou, Yujin Wang, Qian Li, Tingting Zhu, Qiongjie Hu, and Liming Xia. Initial ct findings and temporal changes in patients with the novel coronavirus pneumonia (2019-ncov): a study of 63 patients in wuhan, china. *European radiology*, pages 1–4, 2020.

[112] Mohamed Abd Elaziz, Khalid M Hosny, Ahmad Salah, Mohamed M Darwish, Songfeng Lu, and Ahmed T Sahlol. New machine learning method for image-based diagnosis of covid-19. *Plos one*, 15(6):e0235187, 2020.

[113] Giulia Giordano, Franco Blanchini, Raffaele Bruno, Patrizio Colaneri, Alessandro Di Filippo, Angela Di Matteo, and Marta Colaneri. Modelling the covid-19 epidemic and implementation of population-wide interventions in italy. *Nature Medicine*, pages 1–6, 2020.

[114] Ying Liu, Albert A Gayle, Annelies Wilder-Smith, and Joacim Rocklöv. The reproductive number of covid-19 is higher compared to sars coronavirus. *Journal of travel medicine*, 2020.

[115] Giacomo Grasselli, Antonio Pesenti, and Maurizio Cecconi. Critical care utilization for the covid-19 outbreak in lombardy, italy: early experience and forecast during an emergency response. *Jama*, 323(16):1545–1546, 2020.

[116] Samrat K Dey, Md Mahbubur Rahman, Umme R Siddiqi, and Arpita Howlader. Analyzing the epidemiological outbreak of covid-19: A visual exploratory data analysis approach. *Journal of medical virology*, 92(6):632–638, 2020.

[117] Graziano Onder, Giovanni Rezza, and Silvio Brusaferro. Case-fatality rate and characteristics of patients dying in relation to covid-19 in italy. *Jama*, 323(18):1775–1776, 2020.

[118] Malahat Khalili, Mohammad Karamouzian, Naser Nasiri, Sara Javadi, Ali Mirzazadeh, and Hamid Sharifi. Epidemiological characteristics of covid-19: A systemic review and meta-analysis. *MedRxiv*, 2020.

[119] Vinay Kumar Reddy Chimmula and Lei Zhang. Time series forecasting of covid-19 transmission in canada using lstm networks. *Chaos, Solitons & Fractals*, page 109864, 2020.

[120] Rajesh Singh and Ronojoy Adhikari. Age-structured impact of social distancing on the covid-19 epidemic in india. *arXiv preprint arXiv:2003.12055*, 2020.

[121] Monica Schenone, Vlado Dančík, Bridget K Wagner, and Paul A Clemons. Target identification and mechanism of action in chemical biology and drug discovery. *Nature chemical biology*, 9(4):232, 2013.

[122] Kexin Huang, Tianfan Fu, Cao Xiao, Lucas Glass, and Jimeng Sun. Deeppurpose: a deep learning based drug repurposing toolkit. *arXiv preprint arXiv:2004.08919*, 2020.

[123] Yadi Zhou, Yuan Hou, Jiayu Shen, Yin Huang, William Martin, and Feixiong Cheng. Network-based drug repurposing for novel coronavirus 2019-ncov/sars-cov-2. *Cell discovery*, 6(1):1–18, 2020.

[124] Charlotte Harrison. Coronavirus puts drug repurposing on the fast track. *Nature biotechnology*, 38(4):379–381, 2020.

[125] Wei Keat Lim, Eugenia Lyashenko, and Andrea Califano. Master regulators used as breast cancer metastasis classifier. In *Biocomputing 2009*, pages 504–515. World Scientific, 2009.

[126] Yan Li, Yuhua Wan, Peipei Liu, Jincun Zhao, Guangwen Lu, Jianxun Qi, Qihui Wang, Xuancheng Lu, Ying Wu, Wenjun Liu, et al. A humanized neutralizing antibody against mers-cov targeting the receptor-binding domain of the spike protein. *Cell research*, 25(11):1237–1249, 2015.

[127] Xiangxiang Zeng, Siyi Zhu, Xiangrong Liu, Yadi Zhou, Ruth Nussinov, and Feixiong Cheng. deepdr: a network-based deep learning approach to in silico drug repositioning. *Bioinformatics*, 35(24):5191–5198, 2019.

[128] Hakime Öztürk, Arzucan Özgür, and Elif Ozkirimli. Deepdta: deep drug–target binding affinity prediction. *Bioinformatics*, 34(17):i821–i829, 2018.

[129] Ryosuke Kojima, Shoichi Ishida, Masateru Ohta, Hiroaki Iwata, Teruki Honma, and Yasushi Okuno. kgcn: a graph-based deep learning framework for chemical structures. *Journal of Cheminformatics*, 12:1–10, 2020.

[130] Bharath Ramsundar, Peter Eastman, Patrick Walters, and Vijay Pande. *Deep learning for the life sciences: applying deep learning to genomics, microscopy, drug discovery, and more.* " O'Reilly Media, Inc.", 2019.

[131] Thin Nguyen, Hang Le, and Svetha Venkatesh. Graphdta: prediction of drug–target binding affinity using graph convolutional networks. *BioRxiv*, page 684662, 2019.

[132] Yulong Shi, Xinben Zhang, Kaijie Mu, Cheng Peng, Zhengdan Zhu, Xiaoyu Wang, Yanqing Yang, Zhijian Xu, and Weiliang Zhu. D3targets-2019-ncov: a webserver for predicting drug targets and for multi-target and multi-site based virtual screening against covid-19. *Acta Pharmaceutica Sinica B*, 2020.

[133] Sepideh Sadegh, Julian Matschinske, David B. Blumenthal, Gihanna Galindez, Tim Kacprowski, Markus List, Reza Nasirigerdeh, Mhaned Oubounyt, Andreas Pichlmair, Tim Daniel Rose, Marisol Salgado-Albarr'an, Julian Sp"ath, Alexey Stukalov, Nina K. Wenke, Kevin Yuan, Josch K. Pauling, and Jan Baumbach. Exploring the sars-cov-2 virus-host-drug interactome for drug repurposing. *Nature Communications*, 11(1):3518, Jul 2020.

[134] Igor I Baskin. Is one-shot learning a viable option in drug discovery? *Expert Opinion on Drug Discovery*, 2019.

[135] Han Altae-Tran, Bharath Ramsundar, Aneesh S Pappu, and Vijay Pande. Low data drug discovery with one-shot learning. *ACS central science*, 3(4):283–293, 2017.

[136] Karim Abbasi, Antti Poso, Jahanbakhsh Ghasemi, Massoud Amanlou, and Ali Masoudi-Nejad. Deep transferable compound representation

 doi:10.20944/preprints202008.0320.v1

across domains and tasks for low data drug discovery. *Journal of chemical information and modeling*, 59(11):4528–4539, 2019.

[137] David Gunning. Explainable artificial intelligence (xai). *Defense Advanced Research Projects Agency (DARPA), and Web*, 2:2, 2017.