Queen Mary University of London

William Harvey Research Institute

MSc in Clinical Drug Development

Full Dissertation title: Review of AI-driven pre-clinical drug/vaccine development against COVID-19

Full Name: Hashim Khalid Abdul-Karim Taha

Student ID: 190839734

Submission Date: 22/07/2022

Programme Lead: Dr Dunja Aksentijevic

Total word count: 17,600

William Harvey Research Institute Clinical Drug Development Queen Mary University of London.

Dissertation submitted to Queen Mary University of London in partial fulfilment of the requirements for the Master of Science degree.

Abstract:

Background and aim:

COVID-19 pandemic created a demand for viable therapeutics against SARS-CoV-2. Traditional drug/vaccine development timelines and success rates are inadequate and wasteful in the pandemic response context. Data-driven approaches improve drug/vaccine development success. AI approaches, including ML and NLP are attractive avenues to explore for data-driven pre-clinical development against SARS-CoV-2. The COVID-19 pandemic serves as an excellent case study to examine AI-driven pre-clinical development.

Methods:

A review of literature since the start of the pandemic utilizing AI-driven methods for pre-clinical development was carried out.

Results:

AI was used to aid drug repurposing, de novo design and vaccine design. Out of 153 predictions from drug repurposing, 51 were tested clinically, 24 were tested in-vitro, and 78 had only in-silico evidence supporting their use. 0 studies were found supporting the use of de novo molecules clinically or in-vitro. Clinical evidence was found supporting the predictions made for AI-driven vaccine design.

Conclusions:

Variance amongst papers in their reporting methods, peer-review, model workflows, and validation methods necessitates individualised scrutiny of models as limitations vary. Ideally, further in-vitro testing of results will shed light on the accuracy of model predictions and create trust in AI-driven method predictions. Standardised benchmarks for AI-based reporting and validation will enable greater trust in AI research. Trends in data sharing and public databases serve as excellent training data for AI models, holding promise for the future of AI's use in future pandemics and in pre-clinical development in general.

Acknowledgment:

An immense thank you for the patience all the staff at QMUL have had with me as I have been working to complete this work. In particular, thank you Dr Akenstijevic, Dr Ravic, and Rahat Uddin. It has been a very testing challenge and I am pleased to present this work. Thank you to all my lecturers which were wonderful in enabling me to understand and write about many facets of drug development.

Table of contents:

1. Introduction

The World Health Organisation (WHO) declared an international health emergency in March 2020, due to the outbreak and continued spread of the SARS-CoV-2 virus across the globe (i.e., COVID-19). By October 2020, national authorities had provided the WHO with data which showed >37 million confirmed COVID-19 cases and 1 million confirmed deaths globally (Wang et al., 2020). After one year, as of 29th November 2021, there had been over 5,200,000 total confirmed global deaths associated with COVID-19. Determinations of SARS-CoV-2 reproductive number had found an R_0 that ranges from 2 to 3.5 (Wang et al., 2020b), placing COVID-19 to effectively disseminate throughout a population directly posing a grand mortality risk, and indirectly doing so via overwhelming any public health providers as critical care capacity was filled by COVID-19 patients requiring urgent critical care. As such, there was a great deal of urgency to deliver therapeutics to alleviate the mortality rate and ease the burden placed upon society by public health measures aiming to suppress spread of the virus. Thus, the COVID-19 pandemic serves as a valuable and novel case study for investigating approaches to drug/vaccine development in the modern era.

This is especially true for approaches that utilize machine learning, which has seen many breakthroughs in its long history leading up to the pandemic. Take, for instance, google's AlphaFold program, which utilized deep learning, a type of machine learning, to predict a proteins 3-d structure from its 1-d polypeptide sequence (Callaway, 2020); indeed, a historic achievement in biochemistry. Machine Learning falls within the scope of Artificial Intelligence, as it relates to the intersectional application of cognitive science, technology and programming. Other examples of artificial intelligence include natural language processing (Mohamed Zakaria Kurdi, 2017), knowledge representation and reasoning (Brachman and Levesque, 2009), and automated planning and scheduling (Ghallab et al., 2016).

In the wake of such advancements in AI tools, the catalyzing effect the pandemic has had on existing trends towards greater digitalization, and the urgency to find a drug/vaccine; the pandemic sets the stage for researchers to demonstrate the use of AI tools in drug/vaccine development.

1.1 What drives the race for COVID-19 therapeutics?

Early observations showed that infection severity varied in an apparently age dependent manner. It has been reported that there is an 8.1 times greater mortality rate amongst a 65+ years old age group versus <54 years old age group (Yanez et al., 2020). Moreover, the apparent influence of age and COVID-19 infection severity is highly associated with six age-dependant risk factors, namely; diabetes, hypertension, coronary heart disease/cerebrovascular disease, compromised immunity, previous respiratory disease and renal disease (Romero Starke et al., 2020). In one meta-analysis and meta-regression study, when adjusting for these important age-dependant cofactors, a weak influence of age on COVID-19 disease severity was seen (Romero Starke et al., 2020). Therefore, these six age-dependant risk factors play a central role in COVID-19 disease severity.

Even generally, there is a significant prevalence of these risk factors within the global population.

For one of these risk factors alone, namely diabetes, there an estimated prevalence of 1 in 11 people of the world's adult population in 2022 which equates to around 415 million people,

and is considered an epidemic (Diabetes UK, 2019). So, one can see that the spread of COVID-19 poses a grand risk to human life globally in accordance with observations of the so-called predictors of COVID-19 infection—associated mortality (Corona et al., 2021) and presence of these predictors with regards to global population demographics.

This was especially true in countries with elevated prevalence of the previously mentioned six age-dependant risk factors, particularly developed countries with long life expectancies and high median age demographics. In the UK alone, for example, in 2019 the prevalence of diabetes within the population was around double that of the global population, that is, more than 2 in 11 people (Diabetes UK, 2020) (Office for National Statistics, 2021).

In these early stages of the global COVID-19 pandemic data was scarcer about the virus and its transmission. Notably, there was uncertainty regarding the possibility of asymptomatic viral spread and the overall contribution this made to transmission of COVID-19. For example, some had misinterpreted data from a Wuhan report (Cao et al., 2020) of a city-wide COVID-19 nucleic acid screening programme between May – June 2020 that showed 0 close contacts contracted COVID-19 from 300 asymptomatic cases to mean that asymptomatic spread was not possible. Initial uncertainty about transmissibility and mortality risk affected public policy; the UK had an approach that attempted to manage the virus' spread throughout the population in hopes of achieving herd immunity (House of Commons, 2021).

However, once it had become clear that the National Health Service (NHS) could be overwhelmed, the UK public health strategy towards the virus changed to focus on suppression in the form of restrictions on mass gathering, contact tracing and testing, social distancing, shielding, self-quarantine, nation-wide lockdowns, and compulsory use of medical masks. Meanwhile, UK ministers identified that a vaccine would be the ultimate route out of the pandemic (House of Commons, 2021).

Therefore, it is clear that, firstly, lockdowns were seen as unpreferable, justified purely according to their economic complications and a now-criticized perceived futility, as detailed by the Sixth Report of the Health and Social Care Committee (House of Commons, 2021), but were ultimately deemed necessary to circumvent both a massive loss of life via covid and overwhelming the NHS. Therefore, it is evident that lockdowns amongst other public health measures were not employed as a first and final measure to stop the pandemic and deaths, but rather to supress spread until the availability of viable therapeutics which would ultimately nullify the mortality rate emerged. In this scenario, the lack of immediate SARS-CoV-2 antivirals or available vaccines placed a huge pressure on drug development efforts to alleviate the burden of the virus on human society and in the meantime, the public were suffering the consequences from the disruption of lockdowns and public health measures upon their lives, for example; more widespread depression (Dettmann, Adams and Taylor, 2022), and economic hardship (Richter, 2021).

On March 16, 2020, the National Institute of Health Research (NIHR) Clinical Research Network paused site set up for new or ongoing studies that are not nationally prioritized COVID-19 studies in the UK (NIHR, 2020). Thus, certain actions which were taken to address the immediate need for COVID-19 therapeutics into COVID-19 clinical research shows the urgency with which research into SARS-CoV-2 virus infection and treatment was conducted.

1.2. An overview of the drug/vaccine development paradigm

In its widest definition, the drug development process from bench to bedside spans across pre-clinical drug development and clinical drug development. Referring to figure 1., phases that precede clinical trials are classed as pre-clinical drug development, whereas phases including clinical trials onwards are classed as clinical drug development.

The translation of biomedical research findings to safe, efficacious, and commercially viable therapies that see use in clinical practice is classically a long and complex developmental process. The average length of a successful drug development timeline is estimated to be at least 10 years. Likewise, a typical vaccine development timeline is reported to be similar in length (John Hopkins School Of Medicine, n.d.). The total cost is estimated to be 985 million – 1.3 billion, though there is debate surrounding this figure (Wouters, McKee and Luyten, 2020).

Drug development is undertaken by the pharmaceutical industry whereby a variety of stakeholders are involved across public and private sectors, including regulatory and governmental entities, academic institutions, charities, patients, clinicians, manufacturers, distributors and more. The complexity of the process is owed not only to the multiplicity of stakeholders, but also special scientific and commercial considerations that are associated with risk of study failure and commercial viability.

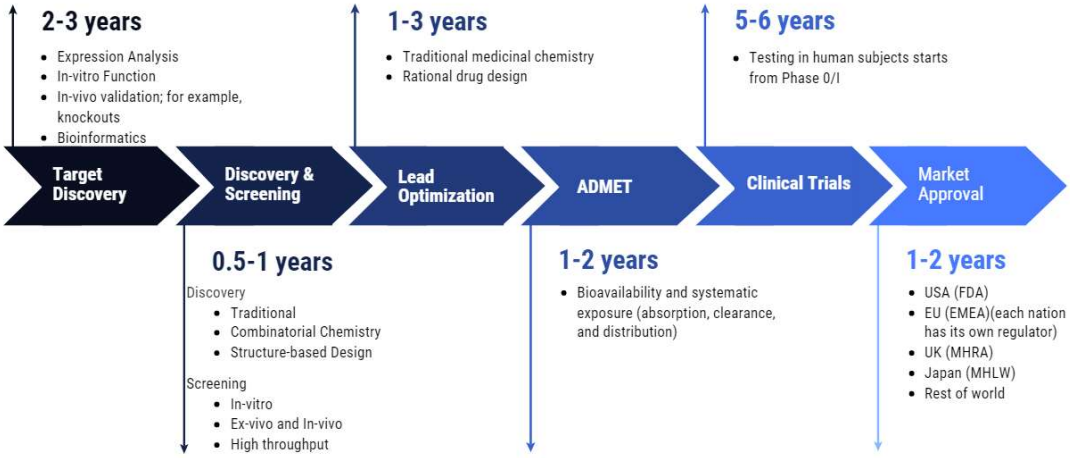


Figure 1. Made by Hashim Taha, adapted from (Ashburn and Thor, 2004) de novo drug development 10-17 year process and less than 10% success rate from discovery to approval.

Confidence in the sustainability of pharmaceutical research and development has been undermined by lower research and development (R&D) productivity, high and increasing costs, and lengthening drug development cycle times. This has been referred to as a productivity crisis. It is therefore integral to establish the challenges associated with drug development across its various stages to gain a better perspective on AI solutions and their potential pay-off if successfully incorporated.

1.2.1. Pre-clinical Development, Challenges, and the emergence of AI

In the pre-clinical side of drug development, the goal is to produce a viable compound that researchers believe can be utilized as a therapy for a specific indication. Modern approaches to

drug discovery rely upon reverse pharmacology whereby available literature is utilized to rationally guide identification of a druggable receptor target associated with the biological processes that a part of the underlying mechanism of disease (target discovery). The hypothesis formed is based on the idea that modulating the identified receptors activity will provide an observable therapeutic benefit. Lead compounds are identified that interact with the identified target (Lead discovery and screening). The lead compound is developed further by optimizing its structure to improve target selectivity, pharmacokinetics (PK), pharmacodynamics (PD), and safety (lead optimization and ADMET). To note, there are also different approaches to drug development as per the "forward" pharmacology (also referred to as phenotypic-based screening) or drug repurposing, each with their own associated timelines, costs, and attrition rate. Take drug repurposing, for example, where safety testing carries less risk as the safety profile is already well established.

The productivity crisis is observed despite an increased wealth of literature available on various pathologies.

A number of scientific and technological advancements have been made in pre-clinical drug development. A vastly greater number of drug-like molecules can be synthesized by chemists per year owing to the advent of combinatorial chemistry, where an 800-fold increase was observed over a decade from the 1980s to 1990s (Dolle, 2010). This is responsible for the expansion of chemical libraries. A tenfold reduction in the cost of screening drug-like compounds against targets was seen by 2012 from the 1990s due to the utilization of high throughput screening (HTS) (Mayr and Fuerst, 2008). The ability to identify targets associated with disease has improved due to a billion-fold increase in the ability to sequence DNA since 1995 (Mayr, L. M. & Fuerst, P. The future of high-throughput screening. J. Biomol. Screen. 13, 443–448 (2008).) and the availability of genome wide association studies (GWAS) (Cao and Moult, 2014). Advances in 3-D protein structure elucidation via x-ray crystallography have enabled at least a three-fold reduction of time investment since the 1950s to identify the 3-D structure of a protein (Van Brunt, 1986). As such, protein data banks have grown substantially and are utilized to mediate structure-led identification of potential lead compounds (Wang et al., 2018). As a greater number of targets involved in disease could be identified and a greater capacity to screen molecules for hits against targets became available, initial optimism was generated for improved productivity of pharmaceutical R&D.

Despite increased investment and innovation in the drug development process, the number of new molecular entities approved per year has not increased substantially, indicating that therapeutic innovation has become more challenging, and a number of reports have aimed to assess the contributing factors to the productivity crisis in drug development. Some reports have pointed out that drug development has become more challenging as a result of a saturation of effective drugs available on the market and that a new drug must prove to be both more effective in terms of efficacy and cost compared to existing treatment or treat disease in which drugs have not been developed and drug development is inherently more challenging (Evenson, 1993). However, it is important to note that the aforementioned developments in HTS and combinatorial chemistry alongside the genomics revolution drove R&D organizations towards a so-called "industrialization" of the drug development industry (Cook et al., 2014). Efficiency was prioritized to pump out drug candidates with quantity-based metrics mediating this push during a block-buster drug era, with a particular hypothesis in mind; the number of candidates entering clinical development would correlate to the number of approved drugs (Cook et al.,

2014). However, trends in attrition rates over the years show worsening rates of clinical study failures across all phases of drug development.

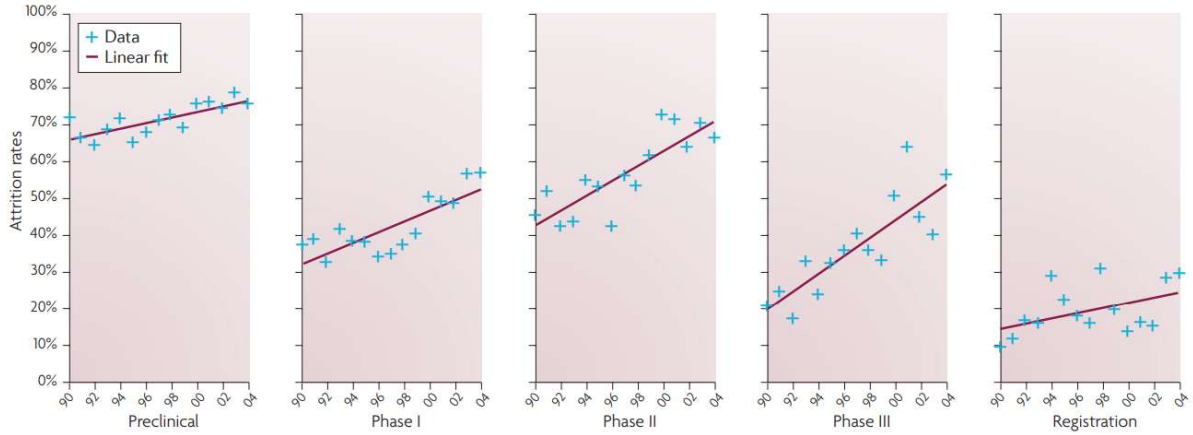


Figure 2. (Pammolli, Magazzini and Riccaboni, 2011) Attrition rates of drug development projects from 1990 to 2004 at various phases.

A landmark review of AstraZeneca's R&D pipeline acknowledges this cultural issue within its R&D pipelines and recognizes that quantity-driven metrics and volume-based goals resulted in projects that focused more-so on project milestone delivery rather than hypothesis testing and truth-seeking behaviours. They found that this had deterred from a scientific understanding of the disease and target biology, identification of ideal subject populations, and conservation of the idea that potential new medicines generated must confer an advantage over existing treatments.

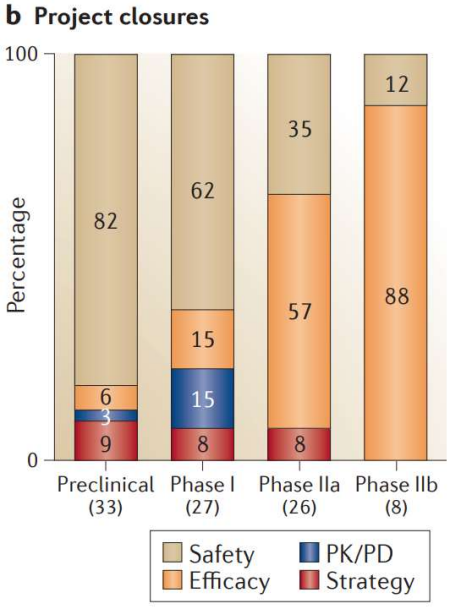


Figure 3. (Cook et al., 2014) Project closures and their related reasoning at various phases.

This report also contributes greatly to establishing critical factors for success and failure in drug development projects. To ensure maximal confidence in drug candidates entering clinical development, satisfying these factors are important.

AI techniques have had breakthrough applications in the drug pre-clinical development space. These applications have potential implications that address major factors identified by the AZ report that affect drug development project success.

The AZ report stresses the importance of selecting biological targets with a high confidence by ensuring targets have a wide range of evidence supporting their selection. AI methods have been employed to inform target selection, namely, to classify druggable targets using ML techniques. Such is case in one novel model, where a variety of genomic datasets and classification based-algorithms considered gene essentiality, mRNA expression, DNA copy number, mutation occurrence and a protein–protein interaction network topology to identify targets of interest (Jeon et al). Another example of ML based informed target selection is a classifier algorithm that utilized protein-protein, metabolic, transcriptional interaction, tissue expression and subcellular localization data to select genes strongly associated with morbidity and drugability (Costa et al.). AI methodologies that inform target selection utilize a vast array of datasets and inputs for training a model to classify new data. The AZ report found that compounds interacting with targets linked by human genetic data to the disease were more likely to pass into later stages of clinical development than targets for which there was an absence of such data. Given that a data-driven approach is stressed as a critical factor for project success, the implications of AI in this space are intriguing, as machine learning is an inherently data driven approach to classification and regression problems which can be used in models useful for target selection.

Furthermore, ML techniques have been utilized to predict on-target and off-target activities (Ma et al., 2015) using quantitative structure-activity relationship models which serve as a form of molecule-derived efficacy biomarker and have been previously used, to not only screen, but also design and optimize molecules (Neves et al., 2018) (Verma, Khedkar and Coutinho, 2010). According to the AZ report, the availability of pre-clinical efficacy biomarkers was associated with a greater success rate in projects. AI methodologies are leveraged to generate digital biomarkers, which can supplement other information sources to inform prioritization of further in-vitro experimentation. AI models automate the generation of efficacy biomarkers, albeit digital, from molecular structure input data thus may have implications for the success rate of clinical research.

Another key factor contributing to project success identified by the report is the stratification of patient cohorts into sub-populations and identification of those that are most likely to respond well to a specific intervention based on predisposing factors i.e. genes. With regards to this, pre-clinical in vitro cell line data have been used to build a selective ML prediction model of drug sensitivity which has applications in stratifying patients (Li et al., 2015), hence may impact clinical research success via this avenue.

ML techniques offer the option to model predicted ADME-tox properties in silico. Jan Wenzel et al. demonstrate the power of deep neural network approaches by developing a multitask network that is capable of modelling ADME-Tox properties including microsomal liability (Zhao, 1997), passive permeability, active transport, the partition coefficient measuring lipophilicity amongst others in eight species (including human, mouse and rat) virtually

(Wenzel, Matter and Schmidt, 2019). Given the number of projects identified in the AZ report to have failed as a result of poor safety, where the failure to detect a safety signal or inaccurate assessment of risk drove this result, access to data-driven models that generate and classify molecules based on these results may aid project success.

To conclude, the established pre-pandemic pre-clinical drug development strategies generally generate a quality of leads that assure rates of attrition which are wasteful and inadequate to promptly address the need for therapeutics against COVID-19. AI approaches show promising applications in pre-clinical development, particularly given examples of previously observed applicability in target selection, digital biomarker generation, ADME-tox predictions, and patient stratification. The successful application of AI in these areas have implications for key factors determining drug development project success, and the uncertainty around the AI methods presents itself as a potential opportunity to improve the status quo for quality and/or reliability of leads generated in pre-clinical stages. This review will reveal the success of AI methods that have been applied to the development of therapeutics against the novel coronavirus SARS-CoV-2 within a two-year timeframe; something that would indeed be indicative that these uncertain methods can be readily applied to novel real-world situations with success.

2. Methods

Research aims and objective:

The overall aim of this project is to establish a contemporary view of the real-world applications and limitations of AI technology in pre-clinical drug development. The COVID-19 pandemic is used as a case study. Each identified paper and model will be presented in the results section as a written summary. Summaries will introduce the research and then provide abridged information covering the model framework/workflow, validation, and results. The culmination of these summaries will support a discussion on the uses, limitations and future directions of AI in drug development. To this end, a review of Artificial Intelligence applications in the COVID-19 vaccine and drug pre-clinical development effort employed from the start and throughout the COVID-19 pandemic was carried out according to the following methodology.

Information Sources and Search Strategy:

A systematic search was conducted in Embase. The following search terms were used:

'covid' AND ('artificial intelligence' OR ‘machine learning’ OR ‘deep learning’ OR ‘natural language processing’) AND ('drug' OR ‘vaccine’) to find any applications of AI in the pre-clinical development of therapies including drug and vaccines for COVID-19.

Searched Period:

This systematic review was conducted to only incorporate papers published from the date of the first reported incidence of COVID-19, by filtering out any results published prior to this date. Hence, filters on search period were set to remove results published from before December 2019 (https://journals.plos.org/plospathogens/article?id=10.1371/journal.ppat.1009620).

Eligibility Criteria for selecting studies:

After the initial search, results were narrowed down via removal of duplicates.

Then, abstracts were screened to propose papers of interest for thorough full-text review.

Any review papers that show up using the search terms were also screened for references that fall within the scope of this review.

Screening of abstracts were conducted according to the following inclusion and exclusion criteria:

Inclusion Criteria - Literature was included if it demonstrated an application of AI technology in the pre-clinical development of therapies in the COVID-19 Era, including but not limited to drug and vaccines.

Studies of all languages are included in-so-far as they can be legibly translated with GoogleTranslate.

Both pre-print and peer reviewed research publications were chosen to be appropriate for inclusion, given the proximity in time between the COVID-19 pandemic case study and the execution of this review and given the time it takes to conduct research, write an article and publish (Powell, 2016).

Exclusion Criteria - Studies that discovered COVID-19 drugs or vaccines without the use of AI technology will be excluded.

Studies that use AI in the pre-clinical development of non-COVID-19 indications will also be excluded.

Studies that are outside the scope of pre-clinical development, e.g., clinical development, are excluded from this review.

Publications where full-text access cannot be achieved via QMUL institutional access within a reasonable timeframe will be excluded from this review.

3. Results

AI for COVID-19 Drug Repurposing

N = 17 studies were identified where researchers utilize AI-based models for repurposing drugs to use as potential COVID-19 therapies.

Beck et al. selected six SARS-CoV-2 replication-related viral proteins as potential targets for a hypothetical post-viral entry treatment to supress viral replication. They proposed that an AI-based model can be used for the prediction of drug-target interactions (DTIs) owing to the vast availability of complex information between molecules. To this end, a previously developed AI-based model named Molecule Transformer-Drug Target Interaction (MT-DTI) was employed to predict the binding affinity of 3,410 FDA approved drugs against six SARS-CoV-2 proteins of interest: 3C-like protease (3CLpro), RdRp, helicase, endoRNAase, 3′-5′exonuclease and 2′-O-ribose-methyltransferase.

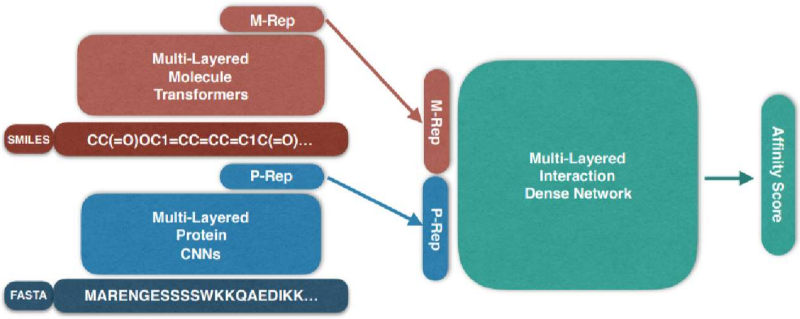
Model Framework This model utilizes a self-attention mechanism to learn high-dimensional structure of a molecule from a given raw sequence. The model generates a molecular representation (M-Rep) from an input “simplified molecular-input line-entry system” (SMILES) string using multi-layered bidirectional transformer encoders. This architecture was

used for its ability to encode the relationship among long-distance atoms. The model also generates a protein representation (P-Rep) from a given FASTA formatted string using multi-layered convolutional neural networks.

Together, the M-Rep and P-Rep of inputs are concatenated to create the input for the Interaction Denses, which is a multi-layered feed-forward network pre-trained to approximate an affinity score between a given M-Rep and P-Rep. The final layer is a regression layer as the model treats protein-ligand interaction and binding affinity prediction as a regression problem.

Pre-training of the model utilized Drug Target Common (DTC) and BindingDB databases. Efficacy values, including inhibitory constant (Ki), dissociation constant Kd, and half-maximal inhibitory concentration (IC50), for drug-target interactions were taken and fed through consistence-score-based averaging algorithm. The Interaction Denses consider features of P-Rep and M-Rep to generate predicted affinity score via a multi-layered feed-forward network with dropout regularization. The mean square error between the network output and actual affinity values determined the optimization of weights of nodes in the Interaction Denses network.

Figure showing DTI-Model Architecture (Shin et al., 2019).



Validation The choice for performance validation for the MT-DTI was an in-silico comparison with AutoDock Vina, as the 3-D structure of 3CLpro had been elucidated by X-ray crystallography (Jin et al., 2020). Beck et al. reported a Pearson correlation co-efficient R of -0.32 with a p-value < 2.2e-16 between MT-DTI predictions and AutoDock Vina predictions of binding affinities of 3CLpro and FDA approved drugs.

Prediction Results The MT-DTI model ranked FDA-approved antivirals according to their predicted affinity and Kd. The model identified eight approved antivirals with potential 3CLpro inhibition within the Kd <1000 nM range. Atazanavir was predicted to have the highest affinity score against 3CLpro. Then, Remdesivir, Efavirenz, ritonavir and Dolutegravir. Atazanavir was found to have a potential binding affinity to all six viral subunits investigated (RNA polymerase Kd 21.83 nM, helicase K_d 25.92 nM, 3′-to-5′ exonuclease K_d 82.36 nM, 2′-O-ribose methyltransferase K_d of 390.67 nM, and endoRNAse K_d 50.32 nM). Thus, Beck et al. suggest all subunits of SARS-CoV-2 replication complex may be inhibited simultaneously by Atazanavir. Researchers also found that ganciclovir had been predicted to bind to three subunits of the SARS-CoV-2 replication complex, namely, RNA-dependent RNA polymerase (K_d 11.91 nM), 3′-to-5′ exonuclease (K_d 56.29 nM), and RNA helicase (K_d 108.21 nM). Lopinavir and Ritonavir were found to have predicted binding affinity with helicase subunit

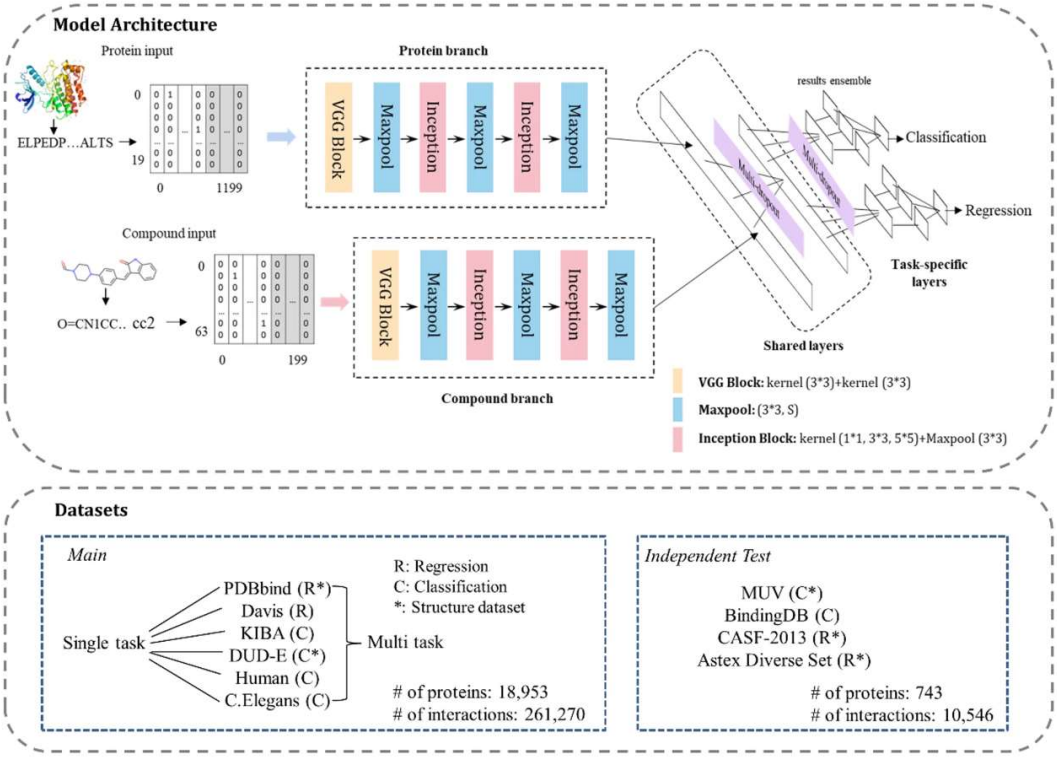
and researchers found that it had been previously identified as a potential Middle East respiratory syndrome coronavirus (MERS-CoV) therapy. Although Remdesivir was not an FDA approved drug at the time, it was included for predicted potency against SARS-CoV-2 viral proteins.

Name	MT-DTI affinity score	MT-DTI predicted Kd	AutoDockVina (kcal/mol)
Atazanavir	7.02	94.94	-7.4
Remdesivir	6.95	113.13	-6.4
Efavirenz	6.7	199.17	-5.4
Ritonavir	6.69	204.05	-6.8
S/GSK1349572	6.47	336.91	-7.2
Atazanavir sulfate (BMS-232632-05)	6.31	488.1	-8.2
Asunaprevir (BMS-650032)	6.24	581.77	-6.5
Ritonavir	6.22	609.02	-6.7
Simeprevir	6.08	826.24	-7.4

Top DTI prediction results of FDA approved antiviral drugs available on markets against SARS-CoV-2 **3CLpro** adapted from supplementary material by Beck et al. Note: higher MT-DTI affinity score reflects high affinity, lower Kd reflects higher affinity, lower AutoDock Vina score reflects higher affinity.

Hu et al. selected eight viral proteins of SARS-CoV-2 to be treated as potential targets including RNA polymerase (RdRp), 3-chymotrypsin-like (3CL) protease, papain-like protease, helicase, spike glycoprotein, exonuclease, endoRNAse, 2'-O-ribose methyltransferase and envelope protein. An AI model was generated and screened a dataset of commercially available drugs for activity against the eight selected viral proteins of SARS-CoV-2. Drugs with high predicted affinities were listed as potential inhibitors.

Model Framework Amino Acid sequences of selected viral target proteins extracted from NCBI. Virus-specific dataset ascertained from GHDDI. Samples with exact binding and affinity are saved. A multi-task deep neural network model was pre-trained by large amounts of data from various heterogenous protein-ligand datasets. The pre-trained model consists of two parts: feature extraction from proteins/ligands based on sequence/SMILES and interaction prediction by shared and task-specific layers. The defined tasks are binary classification of protein-ligand binding or not, and regression analysis to generate predicted protein-ligand binding affinity. The model is fine-tuned by the virus specific dataset to ensure more robust results. After retraining, the model predicts a binding affinity (pKa) between drug and target and screens a library of 4895 commercially available drugs for activity against selected viral proteins of SARS-CoV-2.



Validation 3-fold cross-validation was employed to evaluate the pre-trained model performance on three classification datasets, namely DUD-E, Human and C. elegans.

Prediction Results Hu et al. publish a table of 10 potential inhibitors for SARS-CoV-2 manually selected from the outputs of their model framework, excluding drugs that authors deemed may have strong side effects. These have been adapted into a table below.

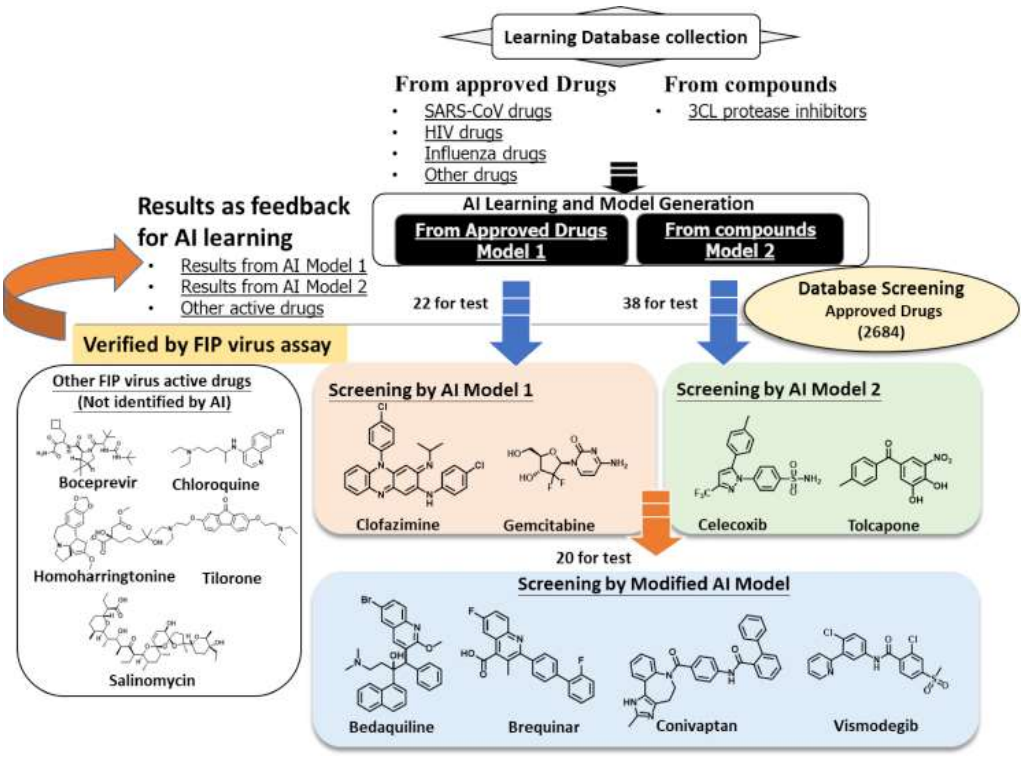
Figure Potential inhibitors for SARS-CoV-2 according to Hu et al.

Drug	**Target**	**Affinity(nM)**
Abacavir (sulfate)	3C-like proteinase Papain-like protease RdRp helicase	28.42 22.90 3.03 3.06
Darunavir	3C-like proteinase Papain-like protease RdRp	57.30 46.16 6.09
Darunavir (Ethanolate)	3C-like proteinase Papain-like protease RdRp	44.51 35.86 4.73
Itraconazole	Papain-like protease RdRp	127.98 16.90
Almitrine Mesylate	3C-like proteinase RdRp	29.31 3.12
Daclatasvir	RdRp	15.03

Daclatasvir (dihydrochloride)	RdRp	19.87
Metoprolol tartrate	Papain-like protease	153.23
Fiboflapon sodium	Papain-like protease	197.63
Roflumilast	3C-like proteinase	248.89

Ke et al. suggests that an AI platform can be applied to learn important molecular descriptors of compounds reported active against 3CLpro, and against viruses that share a level of subunit homology SARS-CoV, SARS-CoV-2, human immunodeficiency virus (HIV), and influenza virus to further screen marketed drugs for potential antiviral activity against coronaviruses.

Model Framework Two independent datasets were collated. One pertained to compounds reported or proven active against SARS-CoV, SARS-CoV-2, human immunodeficiency virus (HIV) and influenza virus. The other contained known 3CLpro inhibitors. The algorithm was trained to generate three types of molecular descriptors from the structure of a given molecule. These molecular descriptor types are extended connectivity fingerprints (ECFP), functional-class fingerprints (FCFPs), and octanol–water partition coefficients (ALogP_count). ECFP are generated from systematically recording the neighbourhood of each non-hydrogen atom into circular layers up to a given diameter of that molecule. FCFPs reports topological pharmacophore fingerprints via the pharmacophore identification of atoms. The ALogP_count is an array of 120 numbers corresponding to 120 Ghose and Crippen atom types. A total of 613 descriptors were used. A Deep Neural Network (DNN) algorithm was then used to identify the molecular descriptors with the most important weight, generating a model for each dataset. Each model was then applied to screen a database of market-approved drugs.



Validation AI-predicted drugs were tested in-vitro against feline infectious peritonitis serotype II of Taiwan isolate NTU156 strain, in feline catus whole fetus-4 (Fcwf-4) cells. An assay was carried out that allowed non-specific cytotoxic analysis and antiviral effect measurements. The drugs that showed favourable effects in this assay were fed back into the learning datasets to improve screening capacity.

Prediction Results The AI system reportedly identified 80 marketed drugs with potential. 8 of which showed favourable properties in-vitro cell-based analysis against FIP proliferation in Fcwf-4 cells. These 8 drugs are bedaquiline, brequinar, celecoxib, clofazimine, conivaptan, gemcitabine, tolcapone, and vismodegib.

Kim et al. select ACE2 and TMPRSS2 as drug targets in potential anti-COVID-19 therapy. They use a high-throughput AI-based binding affinity prediction platform to screen existing FDA approved drugs for repositioning against COVID-19.

Model Framework Researchers used Fluency models. Fluency is a single universal quantitative structure-activity relationship deep learning model. The model takes SMILES input to represent small molecule and protein amino acid sequences. It was trained using chembl databases containing experimental binding data. ACE2 was selected as the target and Fluency screened all chemicals in the Selleckchem FDA approved drug library against ACE2 and ranked them according to a pBind value. Top hits were run in fluency but in reverse to predict binding of a single small molecule to 20,206 human proteins to score specificity. Moreover, ACE2 specificity against ACE1 specificity was investigated by comparing two Fluency models that screened the aforementioned compound library against each receptor. The predicted binding of a molecule against ACE1 (as predicted by one of the models) is compared to the predicted binding of that molecule against ACE2 (as predicted by the second model).

Validation Researchers note that Fluency has been previously validated with experiments for multiple targets. Here, Fluency has been validated in silico by predict binding of Selleckchem FDA approved drugs separately to COVID-19 related prediction results.

Prediction Results Binding affinity prediction of 657 FDA approved drugs determined potential strong affinity between ACE2 and certain drugs namely Piperacillin, Fosamperanivir, Emricasan, and Glutathione.

Table 2 Top ranked fluency hits for binding to ACE2, based on a consensus ranking using the results of both models

Drug Name	Highest similarity to known binder	pBind_a	pBind_b	pBind_a_rank	pBind_b_rank	Reverse Fluency Rank (out of 20,206)	Description
Enalaprilat	0.43	7.90	8.42	4	76	17 (0.084%)	ACE inhibitor; antihypertensive drug
Orlistat	0.25	6.11	8.43	42	72	43 (0.21%)	Reversible inhibitor of lipases; obesity drug
Sotagliflozin	0.39	5.55	8.53	83	36	55 (0.27%)	Inhibits sodium–glucose co-transporters; type I diabetes drug
Tirofiban hydrochloride	0.43	8.43	8.27	1	125	34 (0.17%)	Reversible antagonist of fibrinogen binding to the GP IIb/IIIa receptor; blood thinner
Argatroban	0.54	5.99	8.40	45	83	80 (0.40%)	Inhibiting thrombin-catalyzed or induced reactions; blood thinner
Piperacillin sodium	0.54	5.51	8.44	88	67	7 (0.035%)	Binds to specific penicillin-binding proteins; antibacterial
Ramipril	0.47	6.85	8.20	17	159	3 (0.015%)	ACE inhibitor; high blood pressure
Lisinopril	0.47	7.72	8.15	7	176	16 (0.08%)	ACE inhibitor; high blood pressure
Monopril	0.47	7.84	7.98	6	240	174 (0.86%)	ACE inhibitor; high blood pressure
Captopril	0.37	7.54	7.91	9	262	5 (0.025%)	ACE inhibitor; high blood pressure
Nateglinide	0.70	7.59	7.86	8	284	69 (0.34%)	Interacts with the ATP-sensitive potassium (K+ ATP) channel on pancreatic beta-cells; anti-diabetic
R-406	0.46	8.10	7.16	2	548	3735 (18.5%)	Tyrosine-protein kinase SYK inhibitor
Emricasan	0.44	7.10	8.07	14	209	250 (1.24%)	pan-caspase inhibitor

For each version of fluency run (models a and b), the predicted binding and rank is reported. A higher "pBind" signifies a higher binding affinity. A lower "Reverse Fluency" rank signifies a higher predicted specificity to the intended target

Nguyen et al. proposes that due to 96% sequence homology between 3CLpro of SARS-CoV-2 virus and 3CLpro of SARS virus, a potent SARS 3CLpro inhibitor may also be a potent SARS-CoV-2 3CLpro inhibitor. Nguyen et al. report that despite no existence of an effective SARS therapy at time of publication, the availability of X-Ray crystal structure of SARS 3CLpro and datasets that report 115 potential SARS 3CLpro inhibitors would enable structural based drug repositioning facilitated by machine learning methods.

Model Framework Two previously developed models, MathPose and MathDL were utilized to predict three dimensional poses and protein-ligand binding affinities. The authors note that MathPose converts SMILES formatted entry into 3-D poses and has been recognized as the top performer in D3R Grand Challenge 4 in predicting the poses of 24 beta-secretase 1 binders. MathDL predicts druggable properties of 3D molecules. It converts high dimensional biomolecular data into low-resolution representations. It applies algebraic graph theory-based algorithms, differential geometry, and algebraic topology methods for the mathematical representation of data. These representations of data are then integrated with deep learning models such as gradient-boosted trees and CNNs for pose ranking and binding affinity prediction.

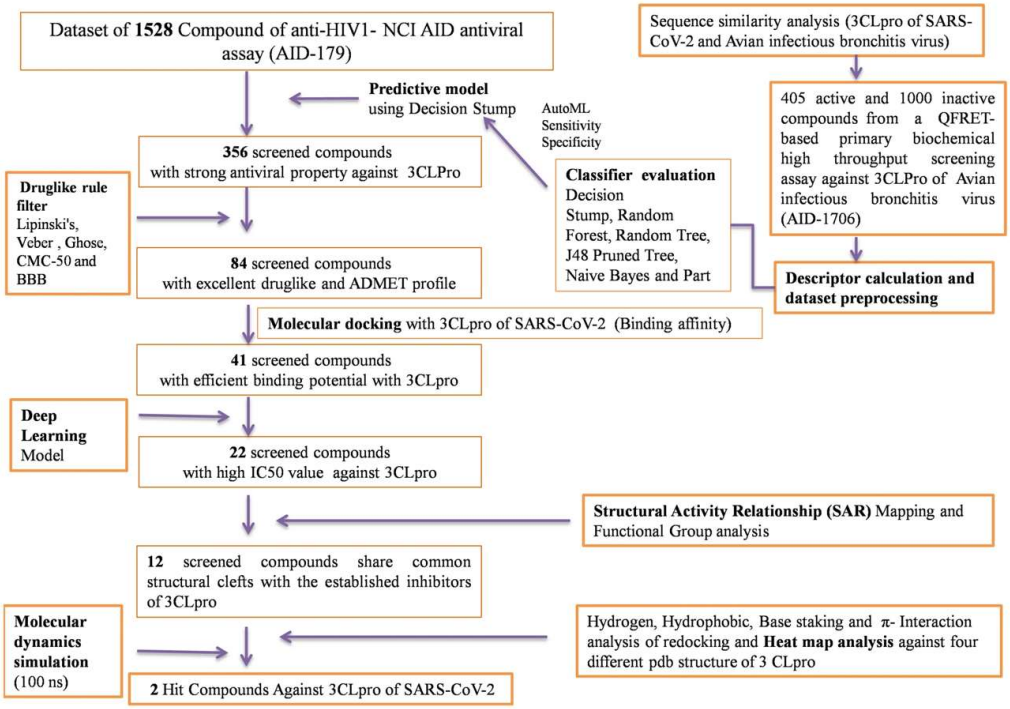
The researchers utilized MathPose to dock 115 SARS-CoV protease inhibitors from ChEMBL against a SARS-CoV-2 3CLpro. Complexes generated by MathPose used as a training dataset alongside protein-ligand complexes from the PDBind database. MathDL trained on the above datasets is used to screen 1465 approved drugs in the DrugBank against SARS-CoV-2 3CLpro.

Validation In silico.

Prediction Results The researchers identified 15 molecules and ranked them according to their predicted binding affinity with SARS-CoV-2 3CLpro. Bortezomib was the top candidate, with predicted binding affinity to SARS-CoV-2 3CLpro of -12.29kcal/mol. Second best was Flurazepam with a binding affinity of -10.37kcal/mol. Third best was Ponatinib with binding affinity of -10.29kcal/mol. 12 other molecules were alongside their binding affinities: Sorafenib (-10.01 kcal/mol), Dasatinib (-9.87 kcal/mol), Paramethasone (-9.71kcal/mol), Clocortolone (-9.58 kcal/mol), Flucloxacillin (-9.57 kcal/mol), Sertindole (-9.54 kcal/mol), Clevidipine (-9.52 kcal/mol), Aprepitant (-9.49 kcal/mol), Atorvastatin (-9.49 kcal/mol), Cinolazepam (-9.47 kcal/mol), Clofazimine (-9.43 kcal/mol), and finally Fosaprepitant (-9.39 kcal/mol).

Nand et al. selected 3CLpro as a vital molecular target against SARS-CoV-2 and screens compounds with anti-HIV1 activity against 3CLpro.

Model Framework Firstly, a predictive model utilizing a decision stump classifier screened a dataset of 1528 compounds with known anti-HIV1 compounds from NCI AID antiviral assay (AID-179) to identify compounds with likely activity against 3CLpro. Drug-likeness filters were applied to the 356 compounds generated by the predictive model to screen for compounds with excellent druglike and ADMET profiles, followed by molecular docking simulations of candidates with SARS-CoV-2 3CLpro. 41 compounds with excellent binding potential with 3CLpro were identified, to which a deep learning model was applied to generate predicted IC50 values for each compound against 3CLpro using regression analysis with efficacy measured in terms of R squared (R^2), Mean squared error (MSE), Root MSE (RMSE), and mean absolute error (MAE). 22 candidates made it past this part of the screen and further non-AI based screens were applied including SAR mapping, functional group analysis and molecular dynamics simulation to arrive at 2 hit compounds against SARS-CoV-2 3CLpro.



Validation In silico.

Prediction Results The two hit compounds identified were CID-230119 and CID-948801.

CID-230119 or 4-{[5-(2-Nitrophenyl)-2-furyl] methylene}-3-phenyl-5(4H)-isoxazolone, has been found to have activity against Dengue virus-2 strain 16,681-PDK53 and Macacamulatta polyomavirus1. Compounds containing Phenyl oxazole, as does CID-230119, have aromatic moieties with binding potential to S1 or S2 sites of SARS protease via H-bond formation and hydrophobic interactions. Moreover, researchers identify the compound has capacity to bind and inhibit His-41, of His41-Cys145 dyad of 3CLpro relating to the catalytic activity of the enzyme.

The second compound, CID-948801 or 4-Chloro-N-(1-methyl-1H-benzimidazole-5-yl) benzamide, had been identified by researchers as previously reported active in 33 PubChem bioassays. Researchers highlight that it forms a H-bond with Arg298, which is a residue critical for structural integrity of the dimer and has a role in regulating catalytic activity.

Both compounds were seen to be active against 3CLpro of Avian Coronavirus and were described as promising candidates against SARS-CoV-2.

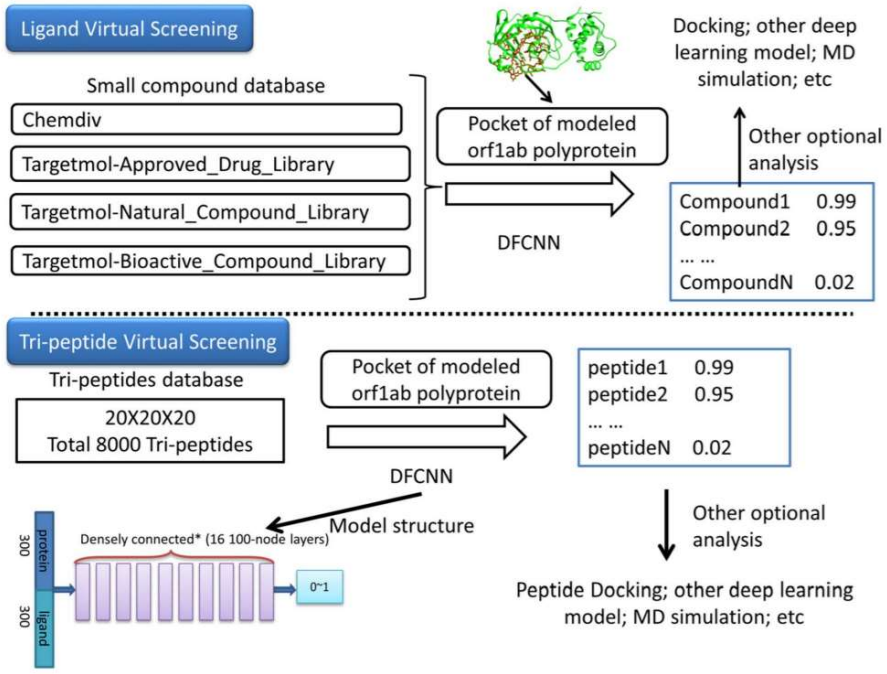
Zhang et al. selected 3CLpro as the therapeutic target for affinity virtual screening of ligands and tripeptides. A previously built model was employed.

Model Framework The model is a pre-trained Dense Fully Convolutional Neural Network which can pre-form a binary classification, sorting screened molecules against a query molecule (target) into two subsets. One set is the potential targets with high possibility to bind with query molecule and other set is molecules with a low possibility to bind to query molecule. The DFCNN is able to predict binding affinity of protein-ligand complexes by learning the effects, binding mode, and specificity implicitly by learning protein-ligand interface contact information from large protein-ligand datasets.

Ligands from Chemdiv dataset screened for activity against 3CLpro. DFCNN used for virtual screening. Top predictions by the model were chosen for further in silico testing by AutoDock Vina based docking simulation.

The Targetmol-Approved_Drug_Library, Targetmol-Natural_Compound_Library, and Targetmol-Bioactive_Compound_Library contain about 2040, 1680, and 5370 compounds, respectively. The pre-trained DFCNN was applied to screen these proteins to identify SARS-CoV-2 protease.

Tri-amino acid peptide database containing 8000 molecules. The tripeptides were converted into a molecule vector by Mol2Vec. The tripeptides vector was represented by the sum of its amino acid vector. 3CLpro protein pocket is then converted into a vector. Pocket and peptide vectors are concatenated into a one line input. Researchers note that the DFCNN should be suitable for protein-small peptide interactions because it is trained by a protein-ligand dataset from PDB bind database.



Validation Researchers used three independent databases for validation of previous model. The model was found to achieve a root mean squared error (RMSE) value of pKa about 1.6-1.8 and R value of 0.5-0.6. Researchers noted the RMSE and R values were better than AutoDock Vina which had RMSE of 2.2-2.4 and R value of 0.42-0.57.

Prediction Results Researchers found that from the targetmol-approved drugs library, Meglumine, Vidarabine, Adenosine, d-Sorbitol, d-Mannitol, Sodium_gluconate, Ganciclovir and Chlorobutanol, respectively, are top predictions according to the DFCNN score.

From the ChemDiv dataset, the top compounds identified by DFCNN are shown in the table below.

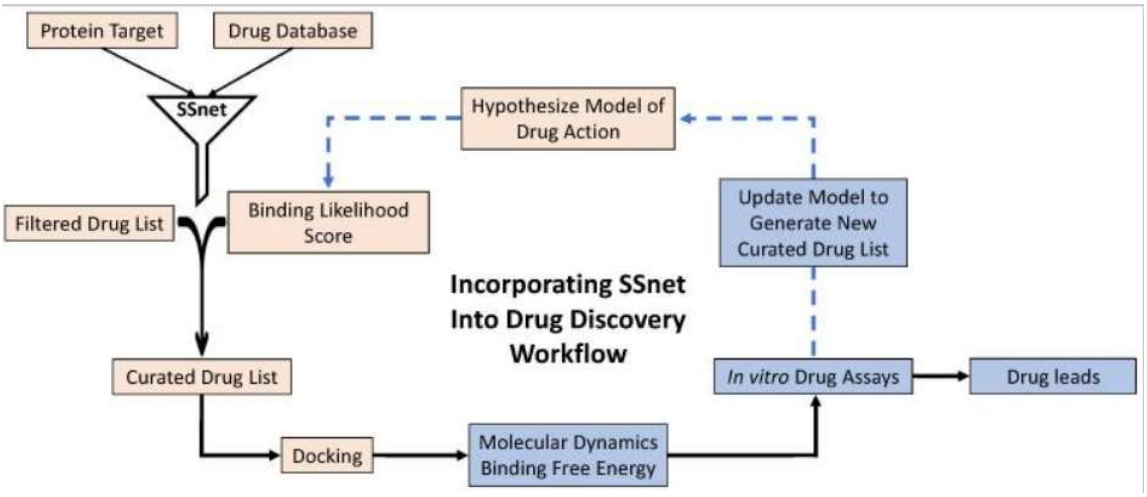
Chemdiv ID	Vina score (kcal/mol)	DeepBindVec	Recommendation
C998-0189	−8.5	>0.995	Recommended
C998-0197	−7.9	>0.995	Can try
C998-0090	−7.8	>0.995	Can try
C998-0948	−7.7	>0.995	Recommended
C998-1046	−7.6	>0.995	Recommended
D076-0195	−7.3	>0.995	Recommended

Compound C998-0189 from ChemDiv dataset had the top Vina score compared to the other six. Researchers identified this compound as N~2~-(3,5-dimethylphenyl)-N~2~-(5,5-dioxido-3a,4,6,6atetrahydrothieno[3,4-d][1,3] thiazol-2-yl)-N~ 1 ~ -[3-(trifuoromethyl)phenyl]glycinamide and noted it satisfied most drug-likeness parameters including Lipinski's filters. It was noted to have 497.6 g/mol molecular weight.

Tripeptide constituted by isoleucine, lysine and proline amino acids were predicted by DFCNN to exhibit favourable affinity for 3CLpro binding site.

Karki et al. selected the human ACE2 receptor as the drug target and employed their deep neural network SSnet alongside classical and virtual screening methods to propose a workflow for AI-integrated drug discovery and to identify drugs effective against COVID-19. They justified choice of target by citing the high affinity between ACE2 and the SARS-CoV-2 N-terminal S1 domain, allowing the virus to adhere to cell surface of human cells. They investigated three conformational states of ACE2 - open, closed, and closed in complex with the S-protein – against compound libraries.

Model Framework The authors proposed a workflow that firstly utilizes their pre-trained algorithm SSnet algorithm to identify compounds predicted to have high binding affinities. Protein structure of ACE2 in each conformation, represented in PDB format, is given to the algorithm as an input alongside ligands formatted as SMILES to predict protein-ligand interactions. The PDB formatted protein is used to extract secondary structural features such as curvature and torsion of the protein backbone. The SMILES formatted ligand is used to extract molecular fingerprints namely the Morgan Fingerprint of the ligand. SSnet is pre-trained to utilize this information and score the likelihood of ligand-target protein binding at IC50 < 10nM.



After this initial step, non-AI based methodologies are employed in a series of steps to provide drug leads. Cross-validation of results was carried out by comparison against traditional drug docking algorithm smina and its scoring function to generate a curated list of drugs.

Validation SSnet is cross validated *in silico* against Smina predicted binding affinity. No further validation was carried out.

Predicted Results Molecules with top binding probabilities were identified as fitting into three categories: antivirals, protease inhibitors, and kinase inhibitors. They publish their list of top scorers. They highlight that drugs like Venetoclax and Aliskiren have shown efficacy towards COVID-19; Pyronaridine has shown efficacy in cell based assays, and furthermore Methylprednisone, Linagliptin, and ormeloxifene are being tested clinical trials.

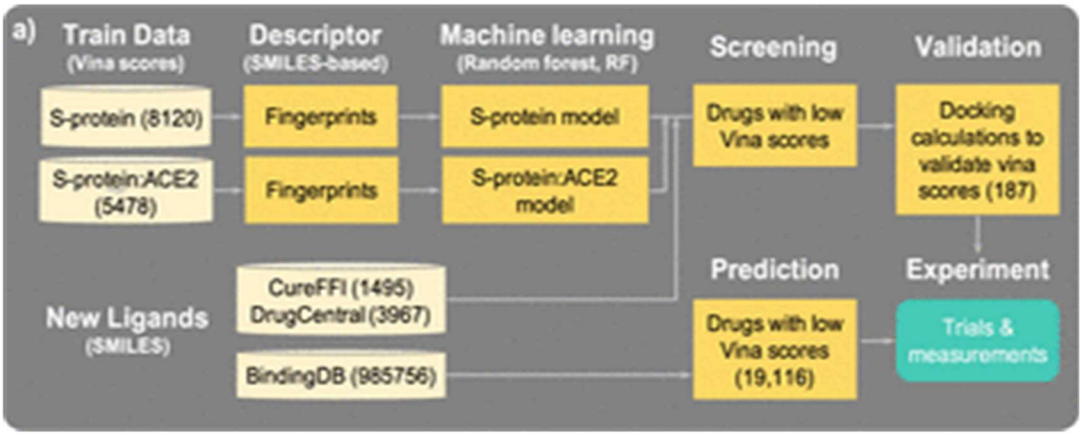
Batra et al. present a combined ML and high fidelity ensemble docking model as a computational strategy for the identification of molecules with high affinity against isolated SARS-CoV-2 S-Protein at its host receptor region or S-protein:Human ACE2 interface complex. They state that through this they aim to identify ligands that limit or disrupt host-virus interactions.

Model Framework Researchers utilized two training datasets obtained from Smith et al. One dataset containing the Vina scores of 9127 molecules alongside their SMILES representation against the isolated S-Protein, and the other against S-Protein:ACE2 interface complex.

These 9127 molecules and their SMILES representation were fed into a fingerprinting algorithm that identifies a set of hierarchical descriptors capturing various geometric and chemical information for each molecule.

Ligands from the Smith et al. datasets were curated to remove skewness in data from the final training datasets, resulting in 5478 data points against isolated S-protein and 8120 data points against s-protein ACE2 interface.

Random forest regression algorithm was used to train the two Vina score models. After training and validation of the two models, three additional drug data sets were used to make predictions of affinity against isolated S-protein and S-protein:Human ACE2 interface complex. The datasets used were FDA approved drug and CNS drug "CureFFI" dataset, common active ingredient "DrugCentral" dataset, and small molecule "BindingDB" dataset. The authors obtained SMILES representation of the molecules from each dataset, resulting in 1495, 3967 and 985756 SMILES respectively. The first two datasets were exclusively used for validation of ML models, whereas the BindingDB dataset was used for ML based prediction.

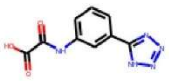


Validation Docking calculations of the top candidates were performed by the researchers as mentioned above. Furthermore, validation against in-silico vina scores from datasets excluded from training.

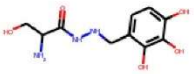
Predicted Results Batra et al. provide a list containing FDA approved and novel candidates, ordered by predicted affinity ranking. The top candidates identified from this workflow are displayed in the figure below.

Overall Top Ligands

Protirelin
Interface: -7.7, S-protein: -4.9



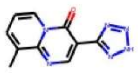
Acitazanolast
Interface: -7.6, S-protein: -5.0



Benserazide
Interface: -7.4, S-protein: -4.6

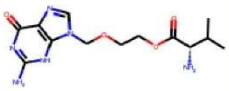
Top FDA Approved Ligands

ID	General Name	Interface Vina Score	S-protein Vina Score	Source
1	Pemirolast	-7.3	-5	CureFFI
2	Sulfamethoxazole	-7.2	-4.7	CureFFI
3	Valaciclovir	-7.2	-4.3	CureFFI
4	Sulfamerazine	-7.1	-4.8	CureFFI
5	Tazobactam	-7	-4.8	CureFFI
6	Nitrofurantoin	-7	-4.8	CureFFI



Pemirolast
Interface: -7.3, S-protein: -5.0

Sulfaperin
Interface: -7.2, S-protein: -4.8



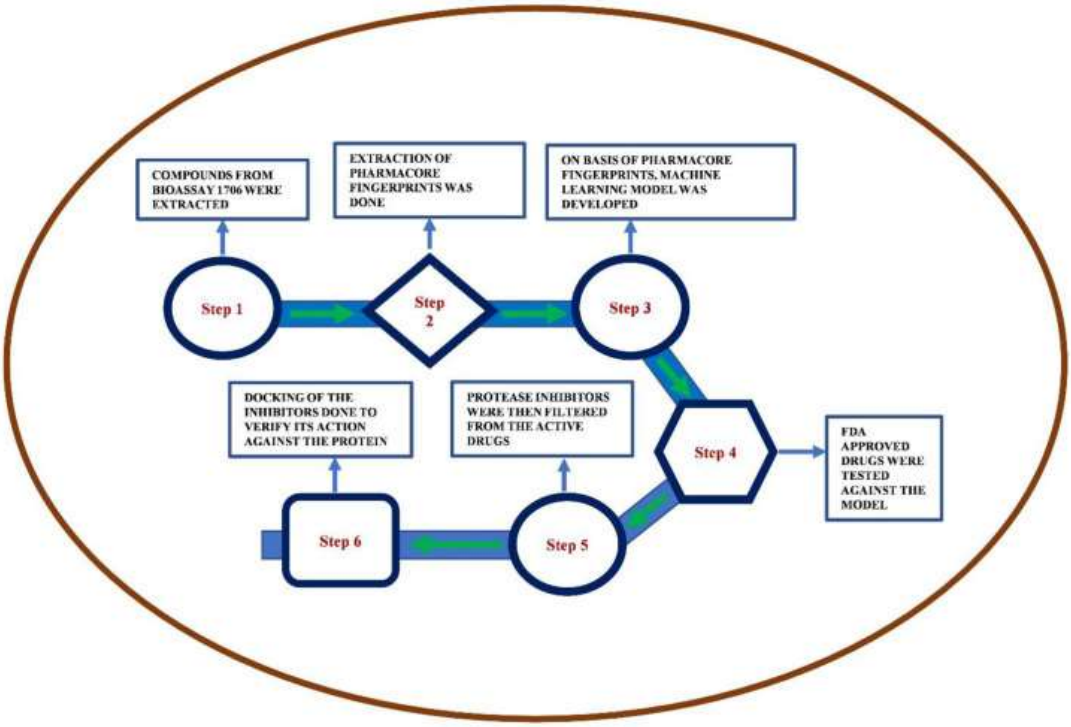
Valaciclovir
Interface: -7.2, S-protein: -4.3

Top Other Ligands

ID	General Name	Interface Vina Score	S-protein Vina Score	Source
1	Protirelin	-7.7	-4.9	DrugCentral
2	Acitazanolast	-7.6	-5	DrugCentral
3	Benserazide	-7.4	-4.6	DrugCentral
4	Sulfaperin	-7.2	-4.8	DrugCentral
5	Succinylsulfathiazole	-7.2	-4.4	DrugCentral
6	Uridine triphosphate	-7.2	-4.9	DrugCentral

Mohapatra et al. utilized a supervised ML-based rapid initial screen of FDA approved drugs with activity against SARS-CoV 3CLPro. This is a binary classifier, which when presented novel data assigns them into one of two categories (active or inactive), in hope of rapidly narrowing the number of candidates with potential anti-SARS-CoV-2 3CLPro activity to be further tested in molecular docking simulations.

Model Framework The PubChem Bioassay AID 1706 containing 290893 compounds was utilized in the preparation of a training dataset. This bioassay aims to identify compounds with inhibitory effect against SARS-3CLPro-mediated peptide cleavage. Three categories of compounds exist within the dataset namely active, inactive and inconclusive. The authors used this dataset to train a Naïve Bayesian classification algorithm, forming a ML based model that is reapplied to a dataset of FDA approved drugs. The Naïve Bayes algorithm identifies attributes of ligands within the training dataset that probabilistically relate to their active or inactive status, which when reapplied to novel datasets of ligands enable assignment of active or inactive status against 3CLpro. The dataset of FDA approved drugs was obtained by authors from DrugBank.



Validation For the validation of binary classifiers, authors investigated the true and false positive rates when applied to non-training datasets where pre-existing knowledge of activity or inactivity was available. Authors write that their model had a TP accuracy rate of 73% and FP accuracy rate of 50.7%. They note that the AUC-ROC value of the model is 0.67.

Predicted Results The ML side of the authors pipeline sorted 471 drugs into the active class from the FDA approved drugs in the DrugBank database. Authors selected drugs from those sorted into this active class with a high confidence level (>90%), resulting in 28 drugs taken further for *in silico* testing with molecular docking simulations. They publish the top 10 drugs generated from their pipeline after molecular docking simulation, namely, Amprenavir, Fosamprenavir, Indinavir, Saquinavir, Darunavir, Ritonavir, Paritaprevir, Lopinavir, Atazanavir and Tipranavir.

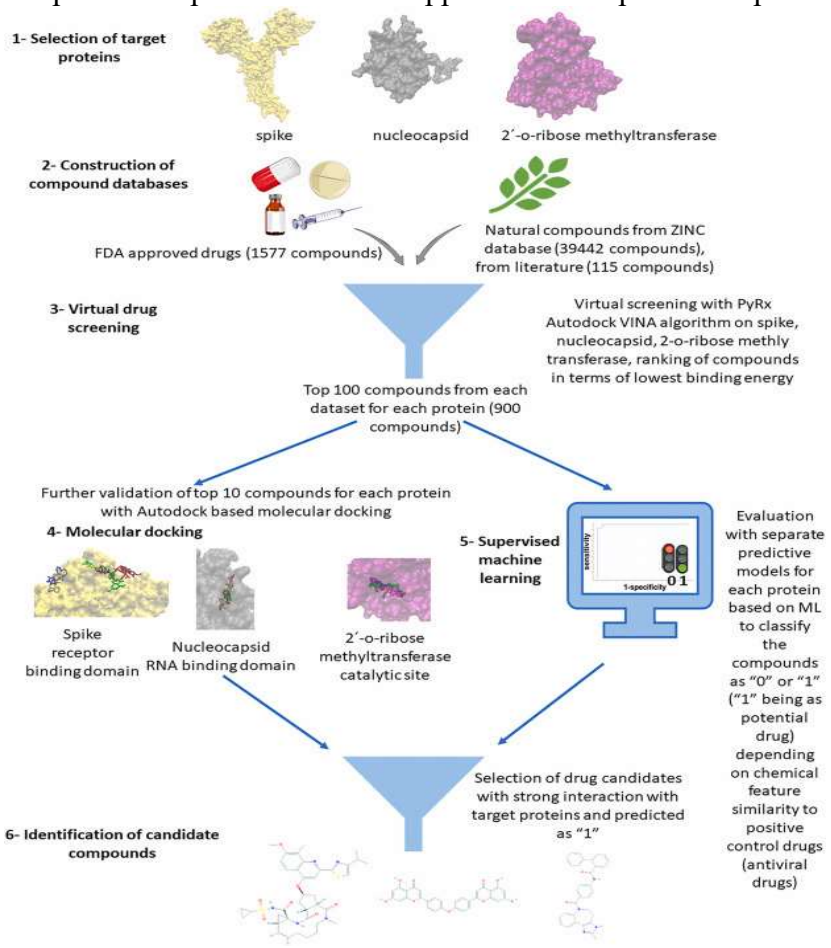
Kadioglu et al. investigated S-Protein, N-Protein and 2′-o-ribose methyltransferase protein of SARS-Cov-2 as targets in an array of virtual screening approaches, including the use of supervised ML based tools supported by the supercomputer MOGON.

Model Framework Firstly, 3D homology models were generated using 3 SARS-CoV protein structures (S-protein, N-protein, 2`-o-ribose methyltransferase protein) obtained from the Protein Data Bank. Alongside this, the available crystal structures of S-protein ACE2 binding domain, N-protein RNA binding domain, and 2`-o-ribose methyltransferase catalytic site were also used in the generation of 3D homology models of SARS-CoV-2 proteins from SARS-CoV proteins.

A compound database was constructed. It consists of FDA approved drugs from the ZINC database (1577 drugs), natural products from ZINC database (39,442 natural products), and natural products from literature (115). Authors selected clinically established anti-viral drugs as positive controls and drugs with no anti-viral activity as negative controls. All compounds were formatted as special data file inputs.

Virtual drug screening using AutoDock Vina to generate a ranking list of compounds with affinity against SARS-CoV-2 proteins. Followed by molecular docking analysis of the top 100 candidates to see whether they have desirable interactions with relevant pharmacophores.

"Drug-likeness analysis" was carried out by a supervised machine learning algorithm which was trained to generate a prediction model for anti-viral activity of test compounds based on 12 chemical descriptors. The predictions were applied to the top 100 compounds.



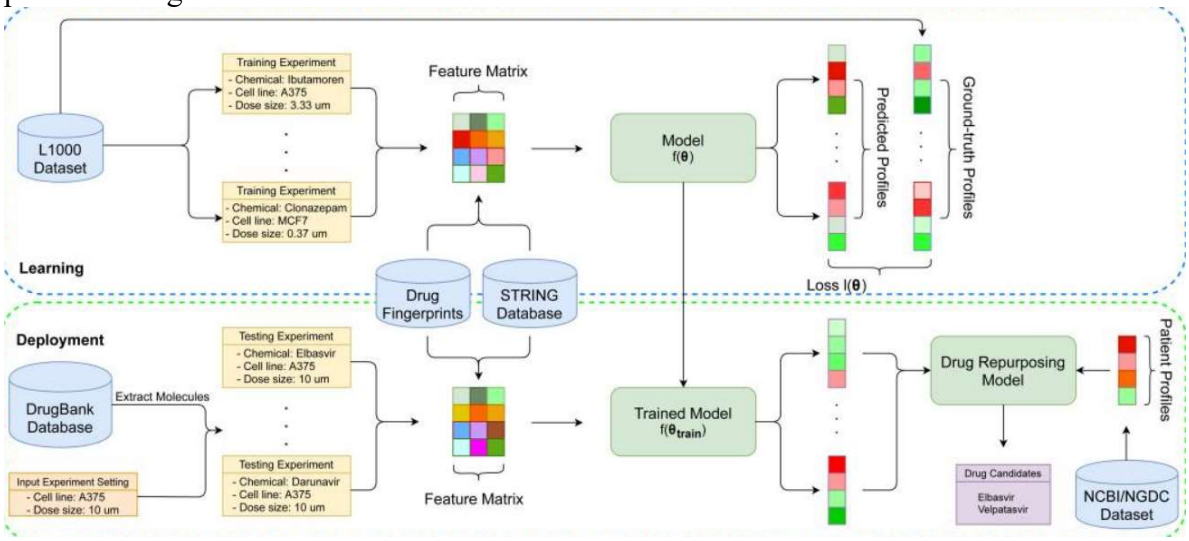
Validation For each target, a set of positive and negative control drugs were used to externally validate the ML model. The model predicted anti-viral activity and had 0 false negative and positives for all three validation sets.

Predicted Results For the S-protein, the candidates with the target interactions with the S-protein were observed to be highest for simeprevir, euphol and ZINC252515584. With the nucleocapsid protein, the top candidates were paritaprevir, ilexsaponin B1 and ZINC27215482. For 2′-o-ribose methyltransferase the top candidates were conivaptan, loniflavone and ZINC15675938.

Pham et al., 2021 use chemically induced gene expression profiles to serve as mechanistic signatures in their proposed method for mechanism-driven neural network based phenotypic screening. They refer to their model as DeepCE. Their method is applied to FDA approved drugs to predict a differential gene expression profile seen in cell lines agitated with a given drug. They apply this for COVID-19 to find drugs that reverse the differential gene expression seen in SARS-CoV-2 infected human cell lines.

Model Framework The researchers utilize the L1000 database, developed by the NIH library of integrated network-based cellular signatures (LINCS) program (Subramanian et al., 2017) to train their model. This dataset contained 1,400,000 gene expression profiles of ~50 cell lines in reaction to ~20,000 compounds across a range of concentrations. The training of the model involves taking vector representation of cell lines, chemical dose size, L1000 genes which are inputs for the "interaction component" to determine feature associations including chemical substructure:gene expression and gene expression:gene expression feature associations. Multi-head attention mechanisms are used for this feature association identification. DeepCE, once trained on high quality data ascertained from the L1000 dataset, was used to predict gene expression profiles for data points in the L1000 dataset that were considered incomplete.

The researchers obtained COVID-19 infected patient expression datasets from National Genomics Data Center (NGDC) (Zhou et al., 2020) and National Centre for Biotechnology Information (NCBI). The NGDC dataset includes 8 SARS-CoV-2 patient and 12 healthy subject samples. The NCBI data has 1 SARS-CoV-2 patient and 2 healthy subject samples. The authors considered the NGDC a population-based dataset for population-based prediction, and the NCBI dataset a patient-specific sample for patient-specific prediction. They determined differential gene expression profiles for these two datasets, and then screened drugs from Drug Bank database for drugs that could reverse the phenotypic changes seen in SARS-CoV-2 versus healthy cell lines. They utilized the Spearman's rank-order correlation coefficient which accounts for both strength and direction. Drugs with the most negative scores were selected as potential drugs.



Validation The DeepCE model achieved a pearson co-efficient of 0.4907. Authors found that simpler variants of DeepCE with removed interaction components had a reduced pearson co-efficients. Authors note that a model "TT-WOPT" that only leverages gene expression values has a pearson co-efficient of 0.0144.

Promising drug candidate identified by pipeline (Population analysis)
Faldaprevir
Alisporivir
NIM811

Ceftobiprole medocaril
Anidulafungin
Otesconazole
Voclosporin
Cyclosporine
Valspodar
Evacetrapib

Predicted Results The top drugs identified for both the patient-specific dataset and the "population"-specific dataset were published. The candidates for the population sample are presented in the table to the left.

Belyaeva et al. suggest that differential gene expression due to SARS-CoV-2 infection must be analysed in tandem with differential gene expression relating to ageing, given that infection severity is strongly associated with age. They suggest a potential hypothesis linking SARS-CoV-2 infection and ageing being greater tissue stiffness in the more elderly, citing the work of Uhler and Shivashankar (Uhler and Shivashankar, 2020). They develop a machine learning platform that uses multiple data modalities to repurpose drugs as therapies against COVID-19.

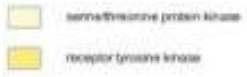
Model Framework The researchers utilized numerous data sources for training an over parameterized autoencoder. These include single-cell transcriptomic analyses data of SARS-CoV-2 versus healthy cells and young versus old lung tissue, protein-protein interaction data, drug-target data and the L1000 Connectivity Map. Researcher used an overparameterized autoencoder, a type of unsupervised neural network, to predict and obtain missing drug signatures within the L1000 database. They then used the completed database to identify the FDA-approved drugs that effectively reverse COVID-19 disease signatures. To do so, Steiner Tree analysis was employed for the identification of a minimal subnetwork that connects genes differentially expressed in normal versus SARS-CoV-2-infected cells and genes differentially expressed in young versus old individuals, revealing a disease interactome.

Validation After identifying the drugs, researchers employed causal structure discovery methods. They decided to do so as their network was undirected, which provided no *a priori* understanding of which direction nodes in the interactome affected one another. For example, it would not be clear if differentially expressed terminal nodes in the network are caused by a drug targeting a node in the network, as the target may be downstream rather than upstream from the terminal nodes.

Predicted Results Their pipeline identified 15 drugs targeting protein gene products with genes differentially expressed (either up or down-regulated) in both SARS-CoV-2 infection and ageing. The drugs are all FDA-approved, high affinity (Ki/Kd/IC50/EC50 < 10uM) and match the SARS-CoV-signature (correlation > 0.86). The authors note that drugs are mostly either Serine/Threorine protein kinases or receptor tyrosine kinases. RIPK1 was found to have the largest number of downstream nodes in the interactome (127), pointing to a highly age-dependant role in COVID-19 infection.

gene	protein name	drug	correlation	affinity
ACVR2A	Activin receptor type-2A	dasatinib	0.88	6.68
AURKC	Aurora kinase C	erlotinib	0.87	6.22
		sorafenib	0.87	6.68
		sunitinib	0.87	6.66
		pazopanib	0.87	6.12
		ruxolitinib	0.87	5.06
		axitinib	0.88	8.89
BRSK1	Serine/threonine-protein kinase BRSK1	sunitinib	0.87	5.46
CDK17	Cyclin-dependent kinase 17	sorafenib	0.87	5.8
		sunitinib	0.87	5.92
EGFR	Epidermal growth factor receptor	dasatinib	0.88	7.1
		docetaxel	0.87	9.0862
		erlotinib	0.87	9.22
		imatinib	0.87	5.12
		sunitinib	0.87	6.07
		axitinib	0.88	5.64
		afatinib	0.86	10
		bosutinib	0.86	7.74
FGFR1	Fibroblast growth factor receptor 1	dasatinib	0.88	5.43
		imatinib	0.87	5
		sorafenib	0.87	5.6
		sunitinib	0.87	6.28
		pazopanib	0.87	6
		axitinib	0.88	6.42
FGFR3	Fibroblast growth factor receptor 3	dasatinib	0.88	5.41
		sorafenib	0.87	5.38
		sunitinib	0.87	6.54
		pazopanib	0.87	6.21
		axitinib	0.88	6.68

gene	protein target	drug	correlation	affinity
HDAC1	Histone deacetylase	vorinostat	0.87	8
		belinostat	0.87	9.07
HSP90AA1	Heat shock protein HSP 90-alpha	formoterol	0.87	8.5229
		primaquine	0.87	8.2218
IRAK1	Interleukin-1 receptor-associated kinase 1	imatinib	0.87	5.92
		sunitinib	0.87	7.85
		pazopanib	0.87	5.23
		ruxolitinib	0.87	6.54
		axitinib	0.88	5.51
		afatinib	0.86	6.62
		bosutinib	0.86	6.22
PAK1	Serine/threonine-protein kinase PAK 1	bosutinib	0.86	5.64
PDE4B	Phosphodiesterase 4	milrinone	0.86	5.22
		vardenafil	0.86	5.35
RIPK1	Receptor-interacting serine/threonine-protein kinase 1	sunitinib	0.87	6.43
		pazopanib	0.87	6.59
		axitinib	0.88	5.6
RIPK2	Receptor-interacting serine/threonine-protein kinase 2	dasatinib	0.88	7.51
		erlotinib	0.87	6.39
		sorafenib	0.87	5.89
		pazopanib	0.87	6.24
		axitinib	0.88	5
		afatinib	0.86	5.57
		bosutinib	0.86	5.43
STK3	Serine/threonine-protein kinase 3	sunitinib	0.87	7.25
		axitinib	0.88	5.66
		bosutinib	0.86	6.43
		tofacitinib	0.88	5.37



Gysi et al. mapped the human and SARS-CoV-2 interactome and employed multi-modal algorithms relying on AI, network diffusion and network proximity to rank 6,430 drugs by their expected proximity against SARS-CoV-2.

Model Framework The human interactome the researchers modelled is based off a compilation of experimentally derived protein-protein interaction data from 21 databases. Authors record that there are five types of PPI data: "1) binary PPIs, derived from high-throughput yeast two-hybrid experiments, three-dimensional protein structures; 2) PPIs identified by affinity purification followed by mass spectrometry; 3) kinase substrate interactions; 4) signalling interactions; and 5) regulatory interactions". The final interactome used in the study contains 18,505 proteins with 327,924 pair-wise binding interactions. A graph convolutional neural network was designed and employed where nodes in the network represent drugs, proteins, or diseases. Edges between nodes could represent protein-protein interactions, drug-target associations, disease-protein associations, and drug-disease treatments. This was leveraged to enable identification of COVID-19 treatment suggestions. Three algorithms were employed to query the interactome for potential FDA-approved drugs for repurposing against COVID-19. An AI-based algorithm maps drug-protein targets and disease-associated protein targets to points in a low-dimensional vector space, resulting in four predictive pipelines that rely on distinct drug-disease embeddings. A diffusion algorithm ranks drugs by capturing network similarity between a drugs protein target and the SARS-CoV-2 host protein targets was also used. Five predictive pipelines were generated using the diffusion algorithm, each powered by distinct statistical methods. A proximity algorithm was employed which ranks drugs depending

on the distance between SARS-CoV-2 host protein targets and the nearest drug-protein target in the network. Three predictive pipelines were generated using the proximity algorithm where the first accounts for all network interactions indiscriminately, the second filters out protein targets involved in drug delivery and drug metabolism, and the final pipeline considers drug-induced differential gene expression in its proximity determination. A heuristic rank aggregation algorithm named CRank was utilized to incorporate the results of all pipelines. By utilizing CRank, authors found that the final pipeline matched or exceeded the predictive power of any one pipeline alone.

Validation Authors validate their work both *in silico* and *in vitro*. They used two independent datasets to quantify the predictive power of their pipeline. They found that their AI pipelines had an AUC of 0.73-0.76. *In Vitro* screening of 918 drugs that demonstrated efficacy by the predictive pipeline was carried out to experimentally test for viral replication inhibition in African green monkey VeroE6 cells, which were exposed to a drug candidate before exposure to wild-type SARS-CoV-2 strain USA-WA1/2020.

Predicted Results Gysi et al. initially found 918 compounds with predicted efficacy against SARS-CoV-2 based on graph neural network representation of the SARS-CoV-2 and human interactome. They tested these 918 compounds in-vitro in a VeroE6 assay and found that 77 had either a strong or weak effect on viral infectivity. 806 had no detectable effect (87.8% of drugs tested). They publish the list of positive in-vitro experimental outcomes. They identified that despite each pipeline mostly shows statistically significant predictive power, they also showed different performance on different ground truth datasets. This prompted the researchers to develop the multi-modal approach which incorporates CRank. 200 drugs were ranked by CRank, where 13 had previously observed positive outcomes in VeroE6 cells. The authors further experimentally validated the set of 13 results in Huh7 cell lines, in a 9-point dilution series from 25 μM to 100 nM. They identified auronofin, azelastine, digoxin, and vinblastine as exhibiting very strong anti-SARS-CoV-2 responses, and methodextrate effectiveness observed at 100 nM concentrations.

Che et al. use a knowledge-graph methodology for drug repurposing.

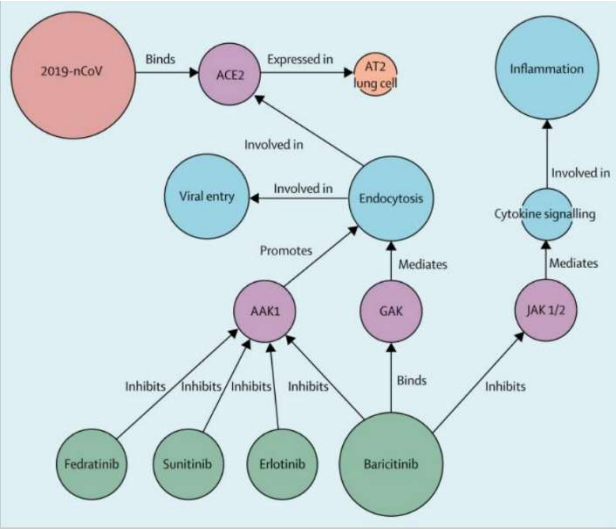
Model Framework The graphical convolutional network with over 100,000 entities and more than 670,000 relationships was employed from a previous study. The embedding of nodes is based on an attention mechanism. The 5 types of node entities included: drugs, genes, diseases, channels and side effects. COVID-19 information was acquired and incorporated into their knowledge-graph. Namely, the RdRp, ACE2, pp1ab and human immunity virus type 1 protection were selected as the COVID-19 human protein targets.

Validation The model was not validated beyond a search to see if any predicted drugs were being tested against COVID-19, of which they found 5 drugs which had already been proven viable for COVID-19 treatment.

Predicted Results Authors extracted the top 30 drugs and listed them as Efavirenz, Lamivudine, Stavudine, Abacavir, Nevirapine, Tipranavir, Roquinimex, Zalcitabine, Delavirdine, Emtricitabine, Didanosine, Tenofovir, Lopinavir, Amprenavir, Zidovudine, Saquinavir, Darunavir, Ritonavir, Atazanavir, Indinavir, Moexipril, Rilpivirine, Cefroxadine, Brecanavir, Lisinopril, Ribavirin, Pentanal, Alfaxalone, 5-[(5-fluoro-3-methyl-1H-indazol-4-yl)oxy]benzene-1,3-dicarbonitrile, SPP1148.

Richardson et al. utilize their Benevolent.AI knowledge graph to identify drugs that may block the SARS-CoV-2 viral infection process.

Model Framework Authors integrate COVID-19 data into their existing knowledge graph to search for drug recommendations. They do not refer to sources of COVID-19 data or present a methodology in their published work.



Validation No reference to validation is made in the paper.

Predicted Results Authors highlighted Baricitinib, one of the drugs identified by the Benevolent.AI knowledge graph as a promising drug to interrupt SARS-CoV-2 viral cell entry. BenevolentAI also identified fedratinib, sunitinib and erlotinib but authors cite poor tolerability, and a high dosage requirement to ascertain desirable AAK1 inhibition when stating that they do not consider these drugs as viable safe therapies for COVID-19 patients.

Han et al. use a supervised ML method to repurpose drugs for COVID-19 treatment by modelling mechanisms of action using cell image features.

Model Framework The authors employ metric learning, a type of machine learning algorithm, to pre-process input data prior to non-parametric clustering. 1105 drugs were analysed with 372 MOAs. Alongside this, drug induced cell image data included 812 data dimensions, including cell area, shape, compactness, and eccentricity. The algorithm clusters drugs based on consistent MOAs. Clusters that related to a MOA that inhibits viral entry to cells was identified.

Validation In silico.

Results authors identified cluster 21 as the cluster that would block viral entry to cells. This cluster contained Chloroquine and Clomiphene. The postulate that these drugs may be candidates as COVID-19 antivirals citing their inhibitory effect on T-cell proliferation and reduction in discharge of pro-inflammatory cytokines, thus preventing elevation of pH of endosome with subsequent endocytosis blockage.

Heiser et al. carried out phenomics analysis of human cells infected with SARS-CoV-2, and developed a chemical suppressor screen which they applied to evaluate a library of approved drugs, enabled by their deep learning neural network model.

Model Framework Researchers first infected monolayers of normal human renal cortical epithelial cells (HRCE). From images of the cell lines after infection, high-dimensional featurisation with the authors proprietary deep learning neural network was enabled the

generation of COVID-19 positive HRCE phenotypic profiles. With this phenotypic profile, an in vitro chemical suppressor screen was developed, and a 1,660 FDA approved compounds were tested in HRCE cells.

Validation There is no information available regarding validation of the author's proprietary algorithms.

Results The pipeline determined that remdesivir and its parent nucleoside GS-441524 had strong repeatable efficacy in the model. The authors highlight that chloroquine and hydroxychloroquine showed no benefit in the human cell model.

AI for COVID-19 *De Novo* Drug Discovery

N=9 papers were identified that utilized AI tools for the purpose of de novo drug discovery against COVID-19.

Ton et al. use a deep learning platform to predict the Glide docking scores of chemicals against active site of 3CLPro.

Model Framework The ZINC15 database, housing 1.36 billion compounds was used for training, validation and screening. A training set isolated and was docked in Glide (A traditional docking algorithm) and used for DD initialization. The structure of the SARS 3CLPro was obtained from the Protein Data Bank, and used in the Glide docking of compounds. The docking scores alongside the SMILES representation of each molecule is the training set for DD. DD relies on quantitative structure-activity relationship modelling to generate predicted docking scores. DD takes SMILES input representation of molecular structures computes Morgan Fingerprints and other 2D molecular descriptors and relates them to the molecule's Glide docking score. The resulting model is now trained and sets of test chemicals are given to DD which divides them into virtual hits and non-hits based on generated docking scores.

Validation The authors validated Glide prediction in a benchmarking study, using 81 known 3CLpro inhibitors ascertained from various other studies.

Results The authors used DD to rapidly estimate Glide docking scores for 1.3 billion chemical structures against 3CLpro active site. They publish a list of top 1,000 hits and describe them as chemically diverse. They note that they exhibit superior docking scores in comparison to known protease inhibitors. The list of 1,000 compounds are made publicly available by the researchers and can be accessed at http://drive.google.com/drive/folders/1xgA8ScPRqIunxEAXFrUEkavS7y3tLIMN?usp=sharing.

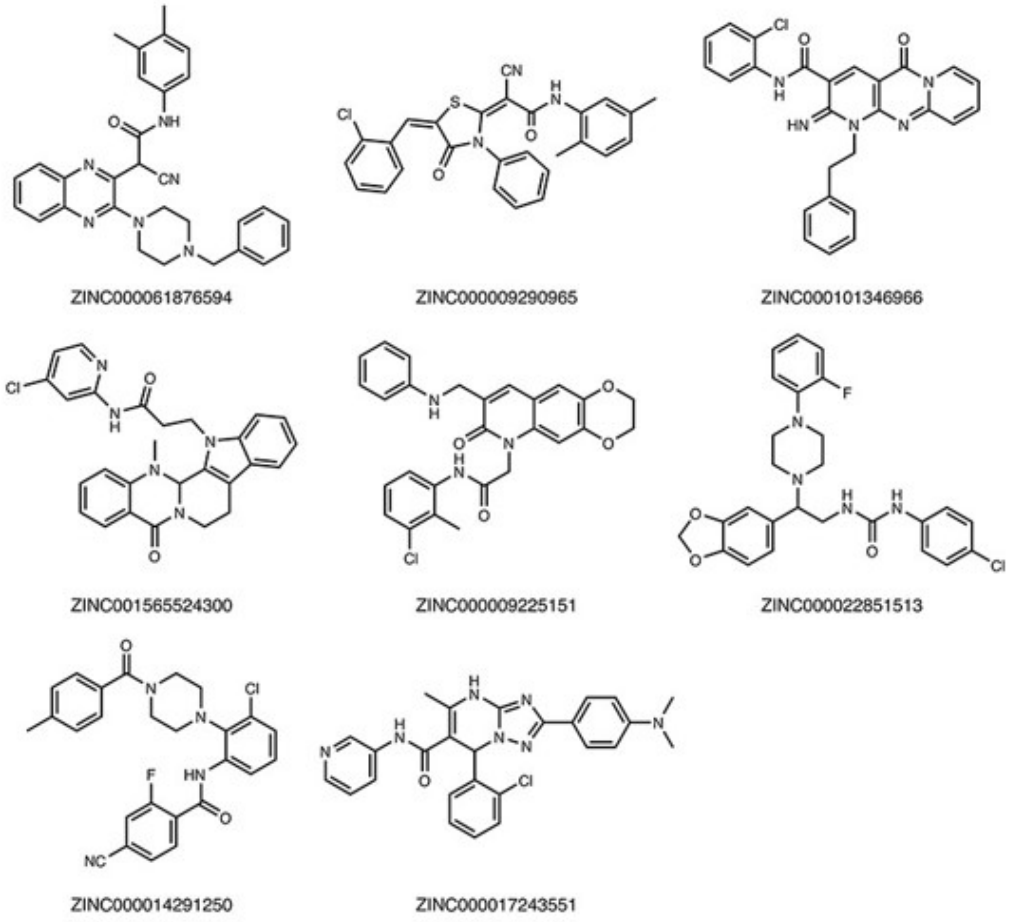
Wang et al. present their transferable deep learning approach to rapidly screen drug-like molecules for antiviral activity against SARS-CoV-2.

Model Framework To train their model in the absence of sufficient data of compounds active against SARS-CoV-2, authors accepted antiviral agents with activity against beta-coronaviruses as positive controls. They cite, amongst other justifications for this decision, the high degree of conservation in essential functional proteins between beta-coronaviruses. 90 inhibitors against HCoV-OC43, SARS-CoV and MERS-CoV (with EC50 < 10nM) from five referenced sources were ascertained. The 90 inhibitors selected as positive controls can all

inhibit at least one of the three coronaviruses. The training dataset contained 90 positive controls and 1862 negative controls. This was used as training data for a deep learning-based classification model. The authors also developed a "fine tuning" dataset containing a collection of experimentally verified active and inactive molecules against SARS-CoV-2, with 70 positives and 84 negatives, ascertained from four referenced sources. The training data sets are used to train a "COVIDVS" model which has learned the features and their weights associated with positive and negative controls to screen potential antiviral molecules from ZINC15 database. The model accounts for 200 molecular features.

Validation They validated the model using a training and test set and comparing the AUC for both. If the ROC-AUC was consistent then the model was not over-fitted. They initially saw that training set 1 had ROC-AUC of 0.99 and ROC-AUC of 0.71 for test set, which denoted overfitting. To solve this problem, they concatenated an additional vector containing the 200 molecular features computed by RDKit. The new model showed ROC-AUC of 0.97 for training set and 0.89 for test set. Furthermore, predicted positives were subjected to in silico traditional docking simulations and experimentally validated the top 40 results.

Results 4.9 million drug-like molecules from ZINC15 were screened. The screen lasted six hours and used 200 CPUs for feature generation and 4 GPUs for prediction. The model returned 3,641 molecules predicted to have SARS-CoV-2 antiviral activity with a high score. Authors noted that 94.6% of these molecules had a maximum similarity of <0.4 to the positive training data. They conducted a clustering analysis to group the predicted molecules based on similarity. Below are the top eight ZINC15 molecules predicted to have SARS-CoV-2 antiviral activity, each one being the top scoring in their respective cluster. Authors suggest these may serve as initial scaffolds for drug design.



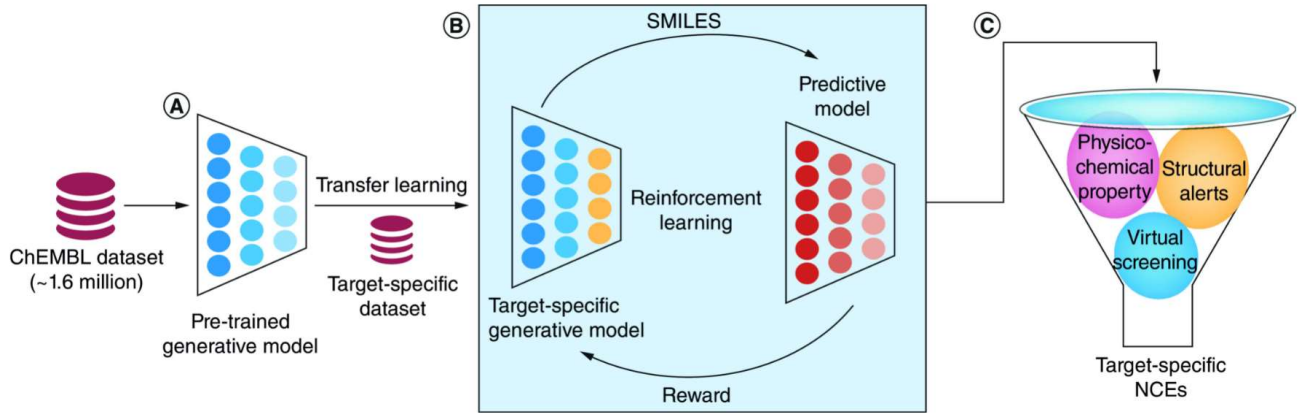
Bung et al. utilized deep neural network based generative and predictive models for de novo design of new chemical entities for SARS-CoV-2 therapy, targeting the 3CLPro site.

Model Framework 1.6 million drug like molecules were ascertained from the ChEMBL database, and were used to pre-train the generative model in SMILES format. The pre-trained generative model generates novel drug-like small molecules without any target specific information. Transfer learning is then utilized to incorporate 3CLPro target specific information. Due to a lack of abundance of 3CLpro inhibitors, experimentally verified viral protease inhibitors were collected from the ChEMBL database. 7,665 viral protease inhibitors were ascertained in total, and they were docked against the active site of SARS-CoV-2 3CLPro using AutoDock Vina. 2,515 passed this screening test to be applied to the generative model via transfer learning.

In transfer learning, the final few layers have an adjusted probability distribution which biases the model in favour of a presented chemical space (in this case, the 7,665 viral proteases used for transfer learning). After transfer learning, the authors refer to the model as “target specific”.

This deep learning platform applies reinforcement learning to iteratively optimize the generation of “target-specific” new chemicals.

Filters are applied to the generated small molecules including synthetic accessibility score ≤5.0, quantitative estimate of drug-likeness >0.4, octanol–water partition coefficient (logP) <6.0, predicted pChEMBL score (bioactivity) >6.0 and molecular weight 400–800 Da. Furthermore, the generated small molecules that passed the first set of filters were subject to the Pan Assay Interference Compounds filter, BRENK filter, NIH filter and ZINC filter.

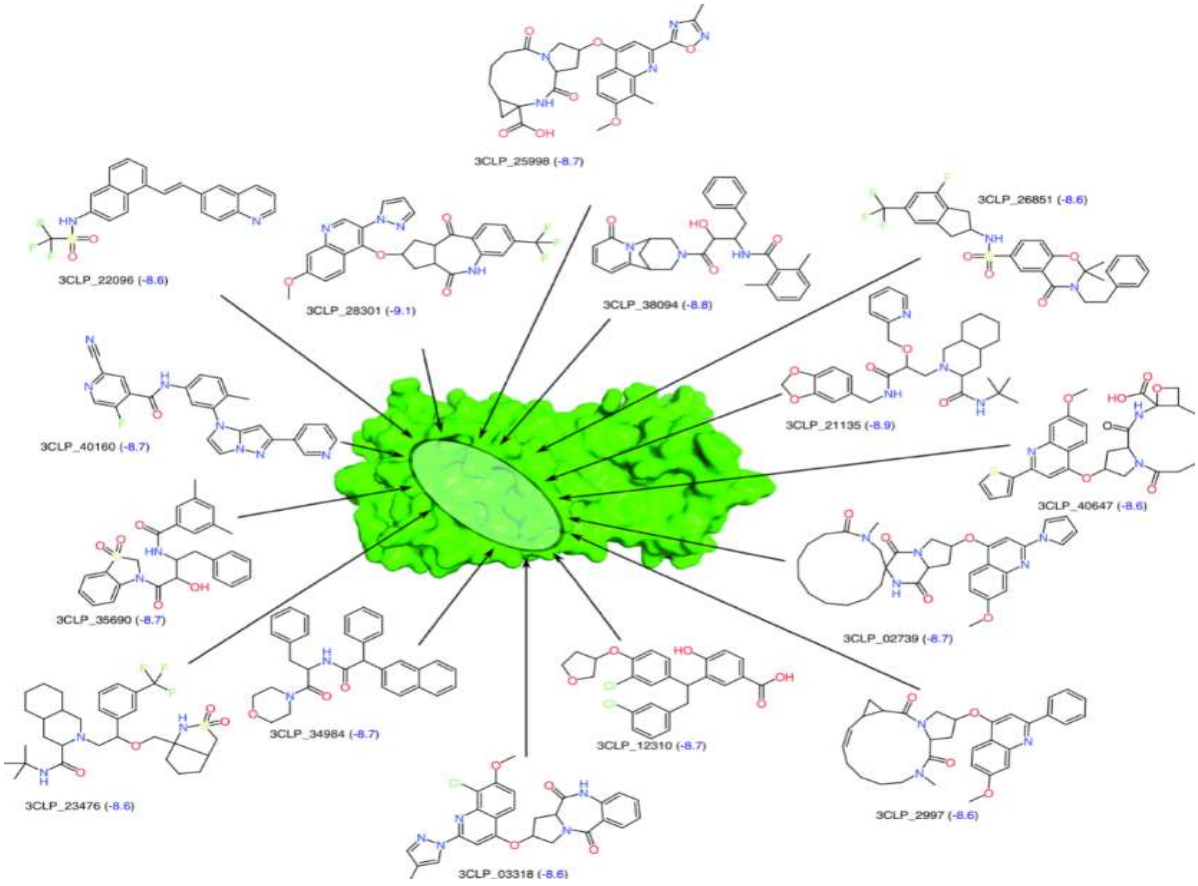


***De novo* drug design pipeline for generating small molecules against a target of interest.**

(A) Pretrained generative model. **(B)** Transfer learning (TL) to learn the features of small molecules specific to the target protein and reinforcement learning (RL) to optimize the property of interest. **(C)** Different physico-chemical property filters, structural alerts and virtual screening score were used for the final screening.

Validation Final results were validated in silico using AutoDock Vina against 3CLpro.

Results 33 potential compounds were generated by the pipeline as ideal potential candidates for further testing against SARS-CoV-2. The new chemical entities with the highest virtual screening scores are shown in the figure below.



Hofmarcher et al. employs a deep neural network coined "ChemAI" for large-scale ligand-based virtual screening of SARS-CoV-2 3CLPro and PLPro inhibitors.

Model Framework The ChemAI network is pre-trained to concurrently predict an array of biological effects. Authors note that it has been trained on more than 220 million data points across 3.6 million molecules from three public drug discovery databases, namely ChEMBL, ZINC and PubChem. It takes inputs in SMILES format and predicts 6,269 biological outcomes including binding to target, inhibitory effects, toxic effects and more.

Authors screen ~900 million molecules from the ZINC database to obtain predicted activity against 3CLPro and PLPro. More detailed information regarding the model framework / workflow is not given. It is implied that the 3CLPro and PLPro activity positive controls are for the SARS-CoV-1 version of the proteins, but results can be applied to SARS-CoV-2 due to conservation across the two coronaviruses.

Validation In silico.

Results The authors provide a list of 30,000 compounds to be used as a screening library. They provide the top compounds in the table below and make the screening library available for access at https://github.com/ml-jku/sars-cov-inhibitors-chemai.

ZINC ID	Canonical SMILES	dist	score	tox	ct
ZINC000254565785	CNC(=S)NN=Cc1c2ccccc2c(Cl)c2ccccc12	0.5455	0.8244	8	0.06
ZINC000726422572	C=C(Cl)COc1ccc(C(C)=NNC(=S)NCCc2ccccn2)cc1	0.5333	0.8232	7	0.05
ZINC000916265995	CNC(=S)NN=Cc1cc2cccc(C)c2nc1Cl	0.6111	0.8230	5	0.08
ZINC000916356873	N#CCCn1cc(C=NNC(=S)NCCc2ccc(Cl)cc2)c2ccccc21	0.6377	0.8221	17	0.07
ZINC000806591744	O=c1c(Br)nn(Cc2cnc3ccccc3c2)c2ccccc12	0.7258	0.8215	11	0.16
ZINC000178971373	O=c1c(Br)nn(Cc2nc3ccccc3s2)c2ccccc12	0.7288	0.8211	8	0.05
ZINC000000155607	CSC(=S)N/N=C/c1ccc2cc3ccccc3cc2c1	0.3902	0.8204	4	0.05
ZINC000016317677	C=CCNC(=S)NNC(=O)Cn1c(COc2ccc(Cl)cc2)nc2ccccc21	0.7000	0.8197	4	0.07
ZINC000193073749	O=C(Cn1cccc(Br)c1=O)c1ccc2ccccc2c1	0.6667	0.8197	1	0.13
ZINC000769846795	O=c1c(Br)nn(Cc2ccc3nccc3c2)c2ccccc12	0.6949	0.8195	9	0.14
ZINC000755523869	CN(N=Cc1nc2ccccn2c1Br)C(=S)NCc1ccccc1	0.6452	0.8194	4	0.05
ZINC000763345954	C=CCNC(=S)NN=Cc1nc2ccc(Cl)cc2n1C	0.6508	0.8194	5	0.07
ZINC000001448699	CSC(=S)N/N=C/c1nc(-c2ccc(Cl)cc2)n2ccccc12	0.6866	0.8192	4	0.13
ZINC000016940508	C/C(=N\NC(=S)NNC(=S)N(C)c1ccccc1)c1nccc2ccccc12	0.6721	0.8191	11	0.06
ZINC000005486767	C/C(=N/NC(=S)NNC(=S)N(C)c1ccccc1)c1nccc2ccccc12	0.6721	0.8191	11	0.06
ZINC000005527649	CSC(=S)N/N=C/c1ccc2ccccc2n1	0.6327	0.8187	6	0.05
ZINC000755497029	C=CCNC(=S)NN=Cc1nc2cc(Cl)ccc2n1C	0.6719	0.8186	5	0.06
ZINC000746495682	FC(F)(F)CNC(=S)NN=Cc1cn(Cc2ccccc2)c2ccccc12	0.5690	0.8186	15	0.07
ZINC000005719506	CN(/N=C/c1ccc(Cl)cc1)C(=S)c1ccccc1	0.6818	0.8178	4	0.05
ZINC000002149503	S=C(NCc1ccccc1)N/N=C/c1cn(CCOc2ccc(Br)cc2)c2ccccc12	0.5625	0.8175	13	0.21

Top-ranked molecules by ChemAI. All compounds have a high activity predicted on all four assays (column "score") and are relatively distant (column "dist") to current known inhibitors. The distance measure is the Jaccard distance based on binary ECFP4 fingerprints and resides in the interval [0, 1]. Some of the presented molecules might exhibit a number of toxic effects (column "tox"). Here the number of models indicating a toxic effect is reported, where the total number of toxicity models was 75. We also report the estimated probability to exhibit clinical toxicity (column "ct").

Tang et al. use AI to aid the design of targeted covalent inhibitors against 3CLPro for anti-COVID-19 drug discovery.

Model Framework Obtained and compiled SARS-CoV 3CLPro inhibitors (284 molecules). A fragment library (fragments < 200 daltons) was generated from the aforementioned molecules. A Q-Learning network was employed to generate new potential lead compounds. The platform uses reinforcement learning to tackle the problem of chemical structure generation, where according to authors the agent performs chemical fragment addition or removal in "a chemistry aware" Markov Decision Process.

Validation Results with high reinforcement learning scores were further validated in docking and covalent docking studies.

Results The authors publish 4,922 unique valid structures, of which 47 had high reinforcement learning scores and had been validated. They provide all generated molecular structures at https://github.com/tbwxmu/2019-nCov.

Verma et al. also used a deep q-learning network for fragment based de novo drug discovery. The target was 3CLPro.

Model Framework Training of Variational Encoder on database of drug-like molecules from ChEMBL. Transfer Learning on inhibitors of 3CLpro to create model that generates potential inhibitors of 3CLpro. Re-enforcement learning applied to the to optimize molecules generated by the molecule generation model, which iteratively adds fragments one by one.

Validation They validated their results using AutoDock Vina

Results They generated 9 novel molecules.

Chenthamarakshan et al. present their end-to-end framework "CogMol" for drug-like small molecule design targeting the SARS-CoV-2 viral proteins 3CLpro, receptor binding domain of S-protein and nsp9 replicase.

Model Framework A pre-trained variational autoencoder and attribute regressors are used for molecule generation. The pre-trained model predicts latent features such as a drug-likeness, synthetic accessibility etc. from molecular SMILES. The latent feature prediction is applied to mediate controlled generation of drug-like molecules. For target-specific compound design, they curated IC50-labelled data from BindingDB. A selectivity model is applied in the final stages of the pipeline.

$$Sel_{T,m} = BA(T,m) - \frac{1}{k}\sum_{i=1}^{k} BA(T_i,m)$$

They defined selectivity within their model as the above function. Selectivity (SelT,M) is "the excess of binding affinity(BA) of a molecule(M) to a target(T) of interest over its average binding affinities to a random selection of targets(K)". Their generative model leverages conditional latent space sampling to concurrently satisfy multiple conflicting objectives when generating molecules to achieve a high binding affinity to a selected SARS-CoV-2 target, high drug-likeness, and high off-target selectivity.

Finally, their pipeline incorporates in silico toxicity prediction, using a multi-task neural network for binary toxicity classification (yes-toxic, no-not toxic) as an early screening tool.

Validation They suggest that next step of their research is wet-lab validation of results, they have not conducted any validation themselves.

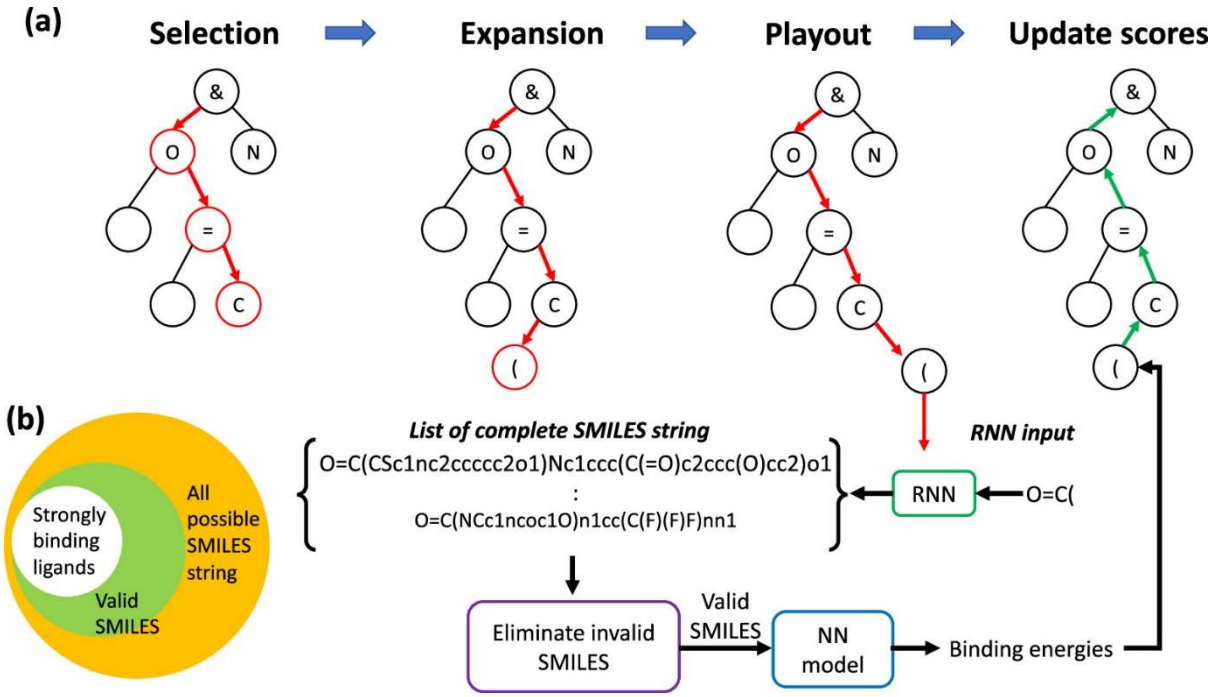
Results 1000 novel drug candidates were generated with potentially inhibitory effect against NSP9, 3CLpro and rbd of the S-protein. A list of the top 97 candidates is made available in their supplementary materials. Notably, the generative model produced SMILES with existing reports of biological activity in PubChem, these compounds are presented the table below.

CogMol-generated SMILES found in PubChem and their predicted affinity (pIC50), lowest docking free energy (kcal/mol), PubChem Compound ID (CID), and reported biological activity.

Target	Pred. Affinity	Docking Energy	CID	Biological Activity
NSP9 Dimer	6.51	−7.7	12042753	Antagonist of rat mGluR
	7.06	−5.6	44397285	Active to human S6 kinase
	7.18	−6.4	10570770	Matrix metalloproteinase inhibitor
Main Protease	7.24	−6.1	10608757	Dihydrofolate reductase inhibitor
	6.91	−6.9	872399	Shiga toxin inhibitor
RBD	7.82	−7.5	76332092	Plasmepsin inhibitor

Srinivasan et al. utilize AI in their de novo drug design method for generating new therapeutic agents targeting SARS-CoV-2 S-protein.

Model Framework De novo design is made possible by leveraging multi-task neural network as a rapid and cheap alternative to docking simulations, as well as Monte Carlo Tree search with rollouts using a recurrent neural network to explore the SMILES chemical space. SMILES strings are generated via the Monte Carlo Tree search, whereby a path from the head node to a terminal node represents a unique SMILES string. The multi-task neural network ensures that SMILES strings are generated in a controlled manner, whereby a reward function that relates to affinity with S-protein is maximized.

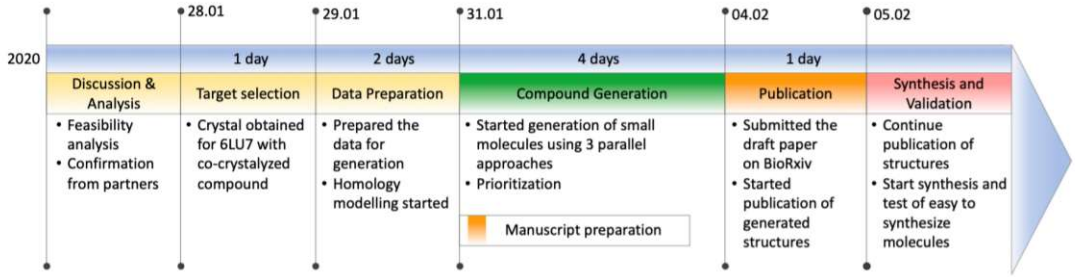


Validation In-silico validation using validation and test sets. Further validation using AutoDock Vina software of top results.

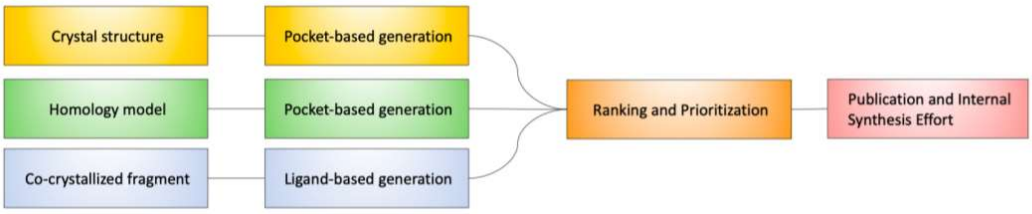
Results The authors discovered 97,973 unique molecules with a high predicted affinity to S-protein of the virus. They present the top 200 candidates with their SMILES, and validated Vina scored. Notably, authors show that a large fraction of newly discovered molecules have higher Vina scores than top FDA approved drugs identified by Batra et al.'s drug repurposing pipeline.

Zhavoronkov et al. utilize their generative deep learning approach for 3CLpro inhibitor drug design.

Model Framework A crystal structure of the SARS-CoV-2 3CLpro bound to a ligand was obtained. The ligand was extracted from the crystal and used in ligand-based generation. The binding site was annotated by the authors proprietary software to map amino acid residues to be used as input data for pocket-based generation. A homology model of 3CLpro in complex with a non-covalent ligand was determined by authors for pocket-based generation.

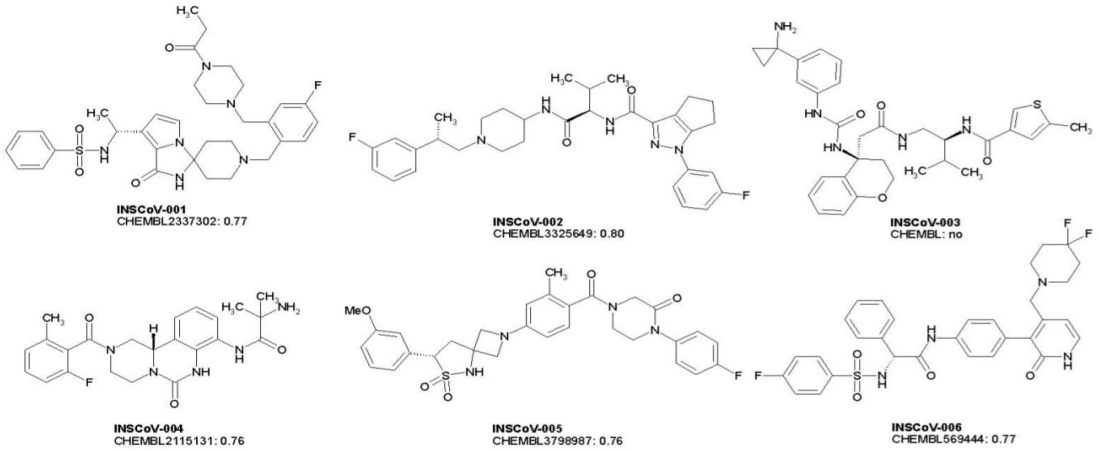


During the compound generation phase, 28 ML models generated molecular structures, and using a reward function, optimized them in a form of re-enforcement learning. The authors state their model architecture utilizes generative autoencoders, generative adversarial networks, genetic algorithms and language models. They also state that models handled an array of molecular representations, such as molecular fingerprints, string representation, and graphs. They describe the reward function as a weighted sum of multiple awards including medicinal chemistry, drug-likeness, active chemistry, structural, novelty and diversity scores.



Validation The authors state their model is previously validated but do not refer to the validation methods in this paper. They state that their team is seeking to synthesise, test and if needed optimize the generated molecules.

Results The authors make results available at https://insilico.com/ncov-sprint. They provide some representative examples of the chemical space covered by the generated molecules in the paper. They note that generated structures display high 3-D complexity, are highly novel with high medicinal chemistry evolution 2018 (MCE-18) novelty scores, and highlighted that some compounds contained stereo- and/or spiro centres commonly seen amongst peptidomimetics and proton pump inhibitors. Authors saw that amongst the molecules, none shared high (>0.7) similarity scores.



AI for COVID-19 Vaccine Design

N=5 studies were identified that utilized AI to aid COVID-19 Vaccine Design

Malone et al. utilize the NEC Immune Profiler software, which is based on machine learning algorithms, to predict which antigens have essential features for CD8, CD4 and HLA-binding, processing, presentation to the cell surface, and potential for recognition by T cells. A Monte Carlo simulation was employed following the prediction of the immunogenicity to identify statistically significant "epitope hotspot" regions in SARS-CoV-2 sequence that have a high likelihood to be immunogenic across a broad spectrum of HLA types.

Model Framework The NEC Immune Profiler has three main component modules:

HLA Binding module – the model, based on three different binding affinity predictors, generates predicted IC50 binding affinity scores (nM) between a peptide and an HLA allele input.

Processing Module – To bind to HLA and be presented on cell surface, antigens must be produced via proteasomal cleavage from a parent protein in cytosol and be transported into endoplasmic reticulum via TAP transporters. To predict processability, a series of Support Vector Machines, pre-trained on large mass spectrometry immunopeptidome data, functions in an ensemble ML layer of 13 models to generate predictions of processability in antigen presenting cells. A consensus score is generated from across the 13 models within the ensemble layer and the HLA binding module, giving a score ranging from 0 to 1. Where, 0 is a poor processability prediction, and 1 is a represents that the antigen is likely to be efficiently processed. This is referred to as an Antigen Presentation (AP) score.

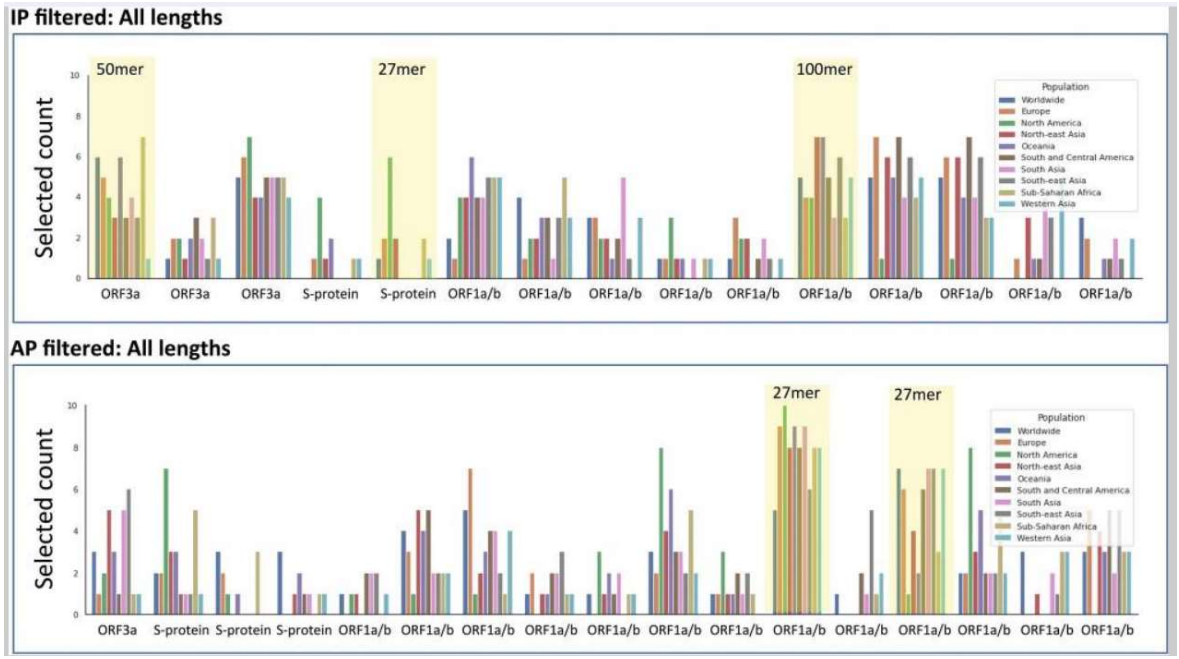
Immune Presentation score – an immune presentation (IP) score between 0 and 1 is generated from AP scores. This is determined using a "immune presentation method" which generates a score representing relative uniqueness between candidate antigens.

The overall statistical question that the researchers aim to answer using these scores is "are specific regions in a given viral protein enriched with higher immunogenic scores, with respect to a given set of HLA types, more than expected by chance?" To answer this, the epitope maps were utilized as inputs for a Monte Carlo simulation in search of the regions of viral protein that demonstrate statistically significant immunogenicity over an array of HLA types. Epitope hotspots with significant homology to proteins in human proteome were removed to account

for unintended off-target autoimmune response induction. Conservation analysis of viral mutations across 3,400 different SARS-CoV-2 sequences was utilized to remove epitope hotspots that occurred in less-conserved regions of the viral proteome. A database of 22,000 individuals with various HLA haplotypes served as a population for a graph based "digital twin" simulation, modelling the effectiveness of different combinations of hotspots in a diverse human population.

Validation Pre-validated the model in-silico by predicting AP and IP scores of identified Class I epitopes SARS-COV epitopes. The model successfully predicted 7 out of 8 epitopes as positive (IP>0.5) demonstrating an accuracy of 87%. The authors acknowledge the very small sample dataset for validation, though decide that it provides a degree of confidence that NEC Immune Profiler prediction pipeline can correctly identify good immunogenic candidates.

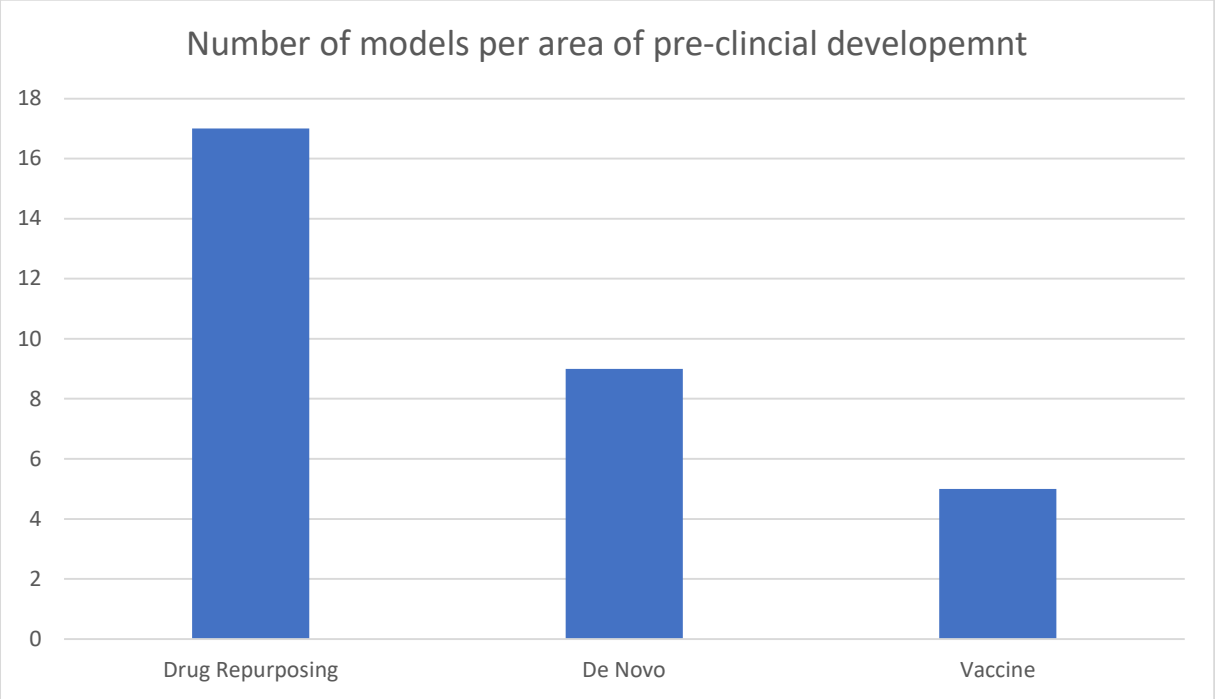
Results 100 different hotspots were identified by Monte Carlo simulation. After filtering for conservation and self-similarity, it was reduced to 50 different hotspots. Ten simulations were run for each region illustrated, where each simulation utilized a population set of 10,000 digital twin citizens, and a vaccine expressed set of vaccine elements. The goal was to select a finite set of vaccine elements i.e., epitopes, which maximize the likelihood of positive response and minimize the probability of no response among the population. The below plot details the frequency (y-axis) a particular hotspot (x-axis) was selected in each of the 10 simulations. Each bar corresponds to a different region-specific simulation setting.



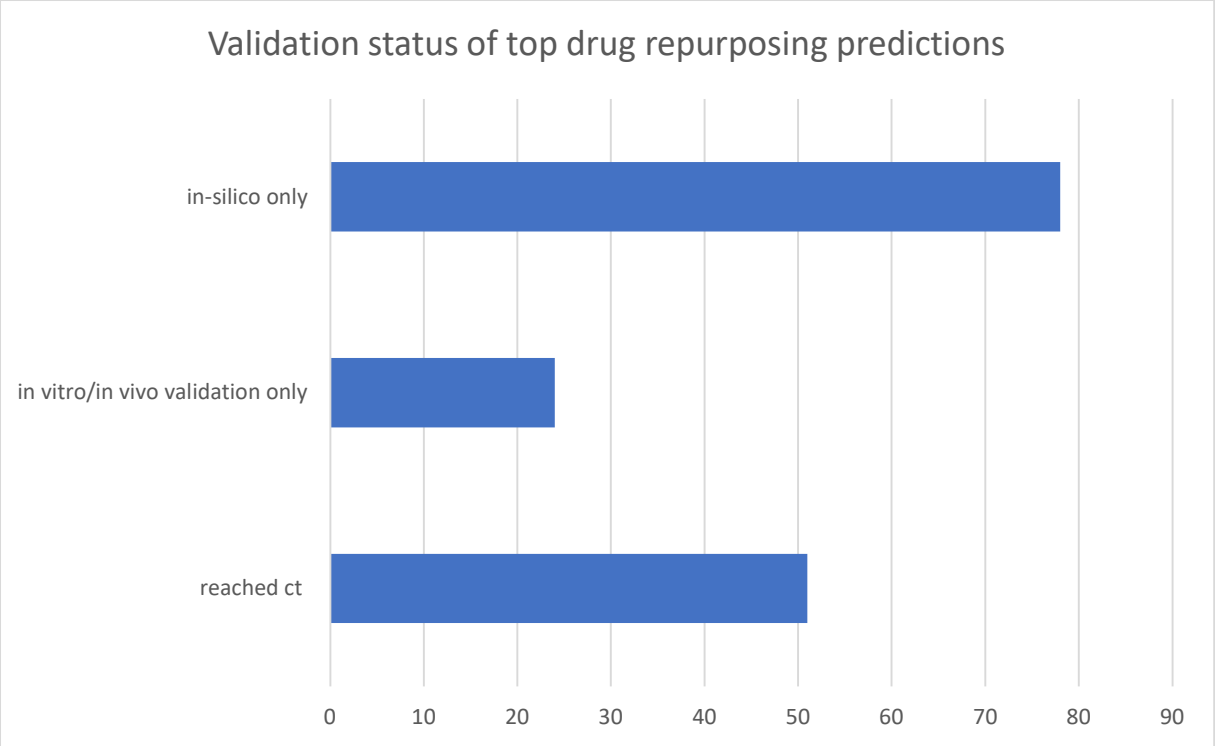
In order to offer coverage for more than 90% of the worldwide population, a subset of five hotspots was chosen based on their profile in the AP and IP digital twin investigations (highlighted in yellow).

Remaining AI-driven vaccine development related studies presented in table below.

Author(s)	Framework	Validation	Results
(Ong et al., 2020)	**Machine learning model applied to predict viral protein candidates for vaccine development. Pre-trained to identify protective antigens from data obtained in animal model.**	In-silico validation	S-protein, nsp3 and nsp8
(Prachar et al., 2020)	**Pre-trained NetMHC suite of tools machine learning based model employed to predict best epitopes for HLA-binding, testing against 10 HLA alleles.**	In-vitro MHC-peptide complex stability assay	NetMHC suite of tools was found to have low performance for epitope prediction, as validation proved predicted HLA-binding epitopes do not bind stably
(Magar, Yadav and Farimani, 2020)	**Use ML model to predict possible inhibitory synthetic antibodies. Used XGBoost, Random Forest, Multilayered Peceptron, Support vector machine and logistic regression for screening thousands of antibody sequences to find stable antibodies that potentially inhibit SARS-CoV-2**	In-silico molecular dynamics based validation	9 stable antibodies found
(Yang, Bogdan and Nazarian, 2021)	**Deep learning based model to predict and design T-cell multi-epitope vaccine.**	In Silico validation of training using test set.	26 potential vaccine subunits predicted all from spike protein



A bar chart depicting the frequencies of publications per each area of pre-clinical development.



Based on a retrospective search to see the status of the top 153 candidates extracted from the drug repurposing papers found herein. These "validations" are independent from the publications where the predictions were initially made.

4. Discussion

Following a high frequency of COVID-19 cases, hospitalizations, deaths, and the emergence of international demand for anti-COVID-19 therapeutics - this review identifies

unique AI-based models employed during the pandemic relating to *therapeutic development* against COVID-19.

There are six applications that have been identified through this review for which AI based models were utilized for COVID-19 therapeutic development. They were; drug repurposing (17 papers), *de novo* drug discovery (9 papers) vaccine design (5 papers)

When looking at the six identified application groups, there is both inter-group and intra-group variations. These exist in reference to five points, namely; defining the problem task, inputs and their digital format, the model architecture, model training and finally to what extent the model has been validated. These five points play a major role in determining model effectiveness and validity of results. The various specific approaches to each and their implications will be detailed under each subsection of this discussion.

Data-driven decision making is increasingly deemed necessary for successful drug development. The distribution of application groups reveal that AI-based models are not limited in applicability to any one single area of COVID-19 drug development but, in actuality, applications exist across pre-clinical stages. This demonstrates that ML-based models can be tailored to a variety of differing and unique problem tasks within drug/vaccine development, which given that ML is inherently a data-driven tool, represents an opportunity to improve success rates for drug/vaccine development. However, while these models may certainly be proof of application, the extent to which they have been individually validated ultimately determines whether they are appropriate for use in real-world situations.

Models can be limited in their ability to carry out their task reliably in real world scenarios given inadequate training data. This occurs as machine learning models infer rules regarding relationships that exists between data inputs. These relationships may be skewed within the training data and do not reflect the actual deterministic rules that exist in the real world – so when applying rules learnt from training data to novel world scenarios there is a risk that the model will not be reliable in carrying out its task. Generally, validation of models establishes the level of confidence in a model's ability to carry out its task in the real world. The extent to which a model is validated gives varying amount of confidence. For example, in vitro validation of a model may be more convincing than in-silico validation.

4.1 AI and Drug Repurposing

The second largest group is the drug repurposing group, following the diagnosis group. Given that COVID-19 AI models require relevant training data, the high prevalence of drug repurposing approaches can be explained in-part by the great deal of information available for use in AI-based model training. Approved drugs have been accepted by regulatory agencies as safe based on a wealth of data ascertained through pre-clinical and clinical stages. As such, a lot of drug-focused and disease focused data exists for approved drugs. Drug-focused data

includes binding affinity data between existing approved drugs and various ligands (table), or drug-induced cell morphology images. Disease-focused databases include cell-image, proteomic, genetic and epigenetic profiles of disease.
Our review identifies that these databases were imperative for the training of AI-based drug repurposing models for COVID-19, particularly PubChem, ChEMBL, LINCS, CMap and more (see figure). Furthermore, there is a growing collection of databases housing drug-focused and disease-focused data that have enabled the training of drug repurposing AI-based models (Cha et al., 2017). Given this trend, and the importance of these databases for the models found in response to the COVID-19 pandemic, ML based repurposing strategies may become a staple in future pandemic responses as information critical to successful repurposing is readily available for the train.ing of new models.

Name of the database	Description
PubChem (Bolton *et al.*, 2008)	Biological activity of >60 million unique compounds
ChEMBL (Gaulton *et al.*, 2012)	Curated database with compound activity against target genes
LINCS (Vidović *et al.*, 2014)	Follow-up project to CMap, L1000-based expression profiles of drug-treated cancer cell lines
Project Achilles (Cowley *et al.*, 2014)	Cancer cell line RNAi screen to ID genes relevant to cell survival
CMap (Lamb *et al.*, 2006)	Expression profiles of drug-treated cancer cell lines
CTRP (Basu *et al.*, 2013; Seashore-Ludlow *et al.*, 2015; Rees *et al.*, 2016)	Screen of 860 cancer cell lines for sensitivity to 480 drugs and probes. Portal allows comparison to mutation, expression, CNV data
ImmPort (Bhattacharya *et al.*, 2014)	Portal containing 222 studies with 37k subjects includes ELISA, ELISPOT and flow cytometry data
PharmGKB (Hewett, 2002)	Expert curated gene–drug genotype–phenotype connections, dosing guidelines and drug labels
e-Drug3D (Pihan *et al.*, 2012)	Mirrors US pharmacopoeia of small drugs, 1822 molecular structures
DailyMED	Catalogue of drug listings/drug label information
Comparative Toxicogenomics Database (Mattingly *et al.*, 2006; Pihan *et al.*, 2012)	~1.5 M chemical–gene, ~2 M chemical–disease and ~20 M gene–disease interactions

Figure 5. (Cha et al., 2017) databases used in machine learning model training

The use of pre-trained models was seen amongst the papers identified by this review. Nine workflows utilized pre-trained model components for drug repurposing strategies, namely Beck et al, (Hu, Jiang and Yin, 2021), (Zhang et al., 2020), (Kim et al., 2020), (Karki et al., 2021), (Ge et al., 2021), (Han et al., 2021), (Moskal et al., 2020), (Pham et al., 2021). The pre-training often related to input-reading; VAE, Transformer, or CNN components were pre-trained to "read" 1-D strings of text to extrapolate molecular features for use in regression or classification-based analysis. Thus, the growth of databases housing drug-focused and disease-focused data will enable the construction of models that can be re-applied readily in new workflows for drug repurposing in future pandemics or disease areas of interest.

There are 3 general categories of drug repositioning using AI identified herein; structure-based, network-pharmacology based and phenomics-based approaches.

4.1.1 Structure-based repurposing approach

A majority of drug repurposing papers utilized a structure-based repurposing approach, constituting 65% (11 of 17) of repurposing studies. Structure-based approaches are categorized by the input of structural information data for model computations. Structural input data,

historically, has been utilized in two kinds of approaches, namely; molecular docking (Kumar and Kumar, 2019) and similarly ensemble approaches (Keiser et al., 2007).

4.1.1.1 Binary Classification

Within the group of 11 structure-based approaches, 6 studies aimed to screen FDA drugs using ML defining the problem as a binary classification task, where the ML model is trained on dataset of molecules and their structures, sorted into positive and negative controls. If the model is supervised, then the researchers must deduce the most important molecular descriptors that correlate to whether a molecular structure is in the positive or negative group. If it is unsupervised, then the model determines the most important molecular descriptors using a reward function. Regardless, the model is reapplied to a novel dataset, to group the drugs into predicted positive and negative groups.

The effectiveness of these binary classification supervised models is in-part determined by what data are chosen to serve as positive control versus negative control training data and how well justified these choices are. As the emergence of the SARS-CoV-2 virus was novel, there was a lack of anti-viral drugs with established efficacy against COVID-19 and as such, the selection of negative and positive controls was a challenge that researchers had to find ways to overcome when constructing their binary classification-based models.

Kadiologlu et al. demonstrate that this is possible by taking the structures of clinically established anti-viral drugs as a positive control group to train their model for the binary classification task. The binary classification can then be employed as a screen to predict whether a molecule falls into the antiviral category given its molecular structure based on 12 molecular descriptors deduced from the training set as being correlated with the positive antiviral training group. These molecular descriptors were H-acceptors, H-donors, total surface area, relative PSA, molecular complexity, rotatable bonds, ring closures, aromatic atoms, sp3 atoms, symmetric atoms, amides, and aromatic nitrogens. They initially ranked 1577 FDA-approved drugs and 39,557 natural products using PyRx AutoDock VINA generating lowest binding energy (LBE) values against three SARS-CoV-2 target proteins. Then, after taking the top 100 compounds for each target protein, the supervised ML model was employed to further screen the results. Remaining compounds with both LBE < -7 kcal/mol (from molecular docking) and probability values of $R > 0.955$ (from binary classification) were proposed as candidate compounds. They validated their model's ability to predict antiviral drugs using a test set of known antivirals and non-antiviral drugs, achieving very few false positives and negatives, recording overall predictive accuracies ranging from 0.895 – 0.972, sensitivity ranging 0.889-1.000, specificity 0.900-1.000, precision 0.889-1.000 and AUC 0.994-0.997 showcasing a high degree of accuracy *in silico*. Demonstrating, that by utilising the binary classification approach, they can further filter out redundant compounds in combination with an initial non-ML based molecular docking screen despite a lack of known COVID-19 antivirals, adding an *in silico* validated filter layer to molecular docking-based screening. When directly assessing the effectiveness of the ML binary classification component, larger test sets for *in silico* validation would assure greater confidence in the accuracy of the models' ability to classify compounds as antiviral, as test sets for *in-silico* validation only contained 18-38 compounds. In addition to the limited test sets, another potential issue with Kadiologlu et al's work is that the training sets are small, consisting of 27 positive controls and 30 negative controls, relative to the libraries consisting of thousands of molecules being screened. The chemical space covered in these libraries is much vaster than the training data and this potentially could lead to an issue with generalizability. However, it is unclear to what extent

the molecular docking screen mitigates this potential issue, as the chemical space is theoretically reduced by the docking screen when obtaining the top 100 compounds. Ultimately, further testing using in-vitro assays for assessing the inhibitory activity of predicted compounds against the three viral proteins investigated by Kadiologlu et al. will provide greater insight into the overall effectiveness of the workflow (Glaab, Manoharan and Abankwa, 2021).

Besides the approach taken by Kadiologlu et al., there were other sorts of approaches taken in binary classification-based structural screening that were more specific. For instance, Nand et al. used positive inhibitors of the 3CLpro from IBV as positive controls in training data. Mohapatra et al use a PubChem Bioassay AID 1706 containing 290,893 compounds tested for activity against SARS Coronavirus 3Clpro in the cell-based assay, using active compounds as the positive training data. So, circumventing the lack of known positive and negative controls for SARS-CoV-2 is possible by selecting inhibitors against other viruses as positive controls. However, how valid these choices are rests upon the level of sequence homology between the mRNA of alternative viruses and SARS-CoV-2. For the case of SARS-CoV, it has been revealed through genome sequencing that SARS-CoV-2 sequence is 79.6% identical (Zhou et al., 2020a). Structural comparisons between the 3CLpro of SARS-CoV and SARS-CoV-2 have also been made, revealing 97% sequence homology using multiple sequence alignment method (Parmar et al., 2021), moreover 12 residues were found to be distinct between the SARS-CoV and SARS-CoV-2 3CLpro mRNA sequences, though none of them are identified as present within the catalytic and substrate binding site. This high level of sequence alignment supports the use of SARS-CoV 3CLpro inhibitors as an attractive positive control group. However, the 12 divergent residues have been found in molecular dynamics simulations to impact the overall molecular environment of 3CLpro, whereby intra-molecular interactions between these divergent residues affected key residues in the monomer and active dimer form of 3CLpro (Parmar et al., 2021). Molecular dynamics simulations revealed changes in the microenvironment of the active-site residues at the entrance (T25, T26, M49 and Q189), near the catalytic region (F140, H163, H164, M165 and H172) and other residues in substrate binding site (V35T, N65S, K88R and N180K) and suggests that variation at residues F140, E166 and H172 are likely responsible for the more stable SARS-CoV-2 dimer versus monomer as these residues are implicated in dimerization. Therefore, inhibitor discovery would require consideration of the more stable dimer form. Ultimately, it is well established that the primary sequence structure implicates the tertiary structure of a protein through a complex array of interactions, hence small changes in primary structure can result in significant changes the 3-d protein. Ultimately, the use of inhibitors for even mostly identical viruses as positive controls for training binary classification models will have discrepancies that ultimately bias the model away from molecules that could be more suitable and represents a challenge for AI-based drug repurposing for COVID-19. The extent to which this affects accuracy is unclear, and perhaps comparative studies between ML models trained on SARS-CoV versus SARS-CoV-2 inhibitors, or inhibitors of viruses with varying degrees of homology to SARS-CoV-2 will be able to explore to what extent this affects the accuracy of predictions. Namely, whether the molecular descriptors that correlate with positive control groups relating to the different viruses vary significantly between one another. For all intents and purposes, when information is scarce this approach seems to remain an attractive and inexpensive option for initial screens, though may be wasteful if valid candidates are thrown out due to discrepancies between the chosen positive control group and the COVID-19 virus.

4.1.1.2 Regression

Beyond the binary classification approach, 5 studies aimed to predict protein-ligand binding affinities of FDA drugs against SARS-CoV-2 targets using structural input data – treating the problem as a regression task by predicting a continuous data variable, that is, binding affinity. Here, regression is an appropriate approach because it models binding affinity in a way that is true to the reality of binding, which can occur at a range of strengths. Hence, regression is a higher resolution way of looking at the problem as opposed to binary classification, which in comparison gives lower resolution results by sorting results into two groups (active/inactive) with no discrepancy between actives. Though, ML-based binding affinity prediction models found herein also have their own advantages and limitations.

Linear regression models found herein determine a number of affinity related scores (often Kd, IC50, etc.) derived from molecular descriptors. Molecular descriptors are extracted from an input. Thus, the integrity of linear regression-based models is dependent upon a high-resolution representation of a target protein and test molecule, and secondly upon the selection of molecular descriptors most relevant to the determination of binding affinity-based scores.

Molecular data is high-dimensional, i.e., a large number of molecular descriptors can exist for any given molecule. This presents a challenge for machine learning based affinity prediction, as allowing the mathematical operation of linear regression models to process an over-dimensional numerical space without establishing predictive and reliable power affects the reliability of the model. This may be the case due to the inclusion of redundant and irrelevant descriptors as inputs for the linear regression operation. Thus, the utilisation of variable selection techniques is a point of interest as a determining factor in the accuracy and predictive power of regression models.

In a supervised approach, molecular descriptors are selected manually prior to training. Training data is labelled according to the molecular descriptors selected *a priori*. Batra et al. 2020 approach this challenge for COVID-19 by considering three hierarchical levels of features for molecular descriptor selection: at the atomic scale, a so-called “larger length-scale”, and so-called “morphological descriptors” were used. The specific descriptors selected are reported. The selection of these molecular descriptors was reported by Batra et al to be “based on past experiences of fingerprinting organic materials”, seemingly without determining any correlative significance between their selected descriptors and DTI affinity scores via any specified statistical technique (Batra et al., 2020).

In previous works, it has been demonstrated that ML scoring functions failed virtual screening and docking tests because they were overtrained on descriptors that do not determine DTIs but are interaction-independent (Gabel, Desaphy and Rognan, 2014). Further, it was determined that ML based scoring functions are also totally insensitive to docking poses up to 10 Å root-mean square deviations and just describe atomic element counts (Gabel, Desaphy and Rognan, 2014) and subsequently it has been established that to determine the effectiveness of descriptors, benchmarking tests of 1) sensitivity to docking pose accuracy 2) the model’s ability to enrich hit lists in true actives upon structure based virtual screening of reference datasets is required. The model by Batra et al was validated against a test set of Vina scores and achieved a RMSE 0.18 and MAE 0.13 for training data and RMSE 0.29 and MAE 0.21 for validation set fulfilling the hit list enrichment benchmark, however no exploration of sensitivity to binding pose accuracy was conducted. An issue identified by this review is the reporting of descriptor selection methodology for supervised machine learning models, as without specifying a particular analysis for selection the evidence supporting the descriptors selected cannot be scrutinized and critical analysis of a model relies purely on the validation of results. Further,

without reporting an exhaustive list of molecular descriptors used, the reproducibility of the work is greatly limited. In order for greater trust in machine learning methods reliability and accuracy to be built, detailed reporting of molecular descriptors, descriptor selection techniques, and validation benchmarks is necessary.

Given the novelty of the virus, a lack of structural information was available for researchers to work with. The first available structure of SARS-CoV-2 available was the full genome sequence (Garcés-Ayala et al., 2020). As such, 1-d structural information was available, but 2-d and 3-d structural information was not available. Early attempts to repurpose drugs against SARS-CoV-2, such as works by Beck et al which utilised a regression-based model, utilised SMILES and FASTA 1-D string inputs to represent molecules, which enabled the workflow without experimentally confirmed 3-d crystal structures of the virus. Their MT-DTI model was validated in a comparison with AutoDock Vina for two test sets, analysis showed $p = 0.0071$ and $p < 2.2e\text{-}16$. Thus, AI based methods for linear regression-based drug-target interaction predictions can be employed without the use of 3-d crystal structure of a virus. Hence, AI is an attractive methodology to employ in early pandemic stage with limited information about viral structure.

A model's effectiveness is predicated upon an interplay between how researchers decide to define the binding problem (regression versus binary classification) and decisions that relate to model training. Structure-based approaches for COVID-19 drug repurposing require well-established hypothesis regarding validity of training data particularly in the absence of known inhibitors against SARS-CoV-2. Generally, models rely upon high resolution input format, and then the deduction of valid descriptors that enable accurate predictions. Validation is extremely important to verify the integrity of a model and a majority of models reported herein had weak validation approaches that do not explore the validity of the model in full. To improve this, incorporating in-vitro assays to validate all predictions will speak to the reliability of a model more so than in-vitro analysis of cherry-picked results. Comprehensive in-vitro analysis of results will allow greater transparency about the limitations of a model and its accuracy, but currently most validations methods are comparisons with non-ML based docking programmes like autodock vina.

4.1.2 Biomedical Knowledge Graphs for Drug Repurposing

Three studies utilize AI to construct and query biomedical knowledge graphs. These 3 models fall under network-pharmacology based drug repurposing. AI-driven network-based approaches have proven to be effective computational method for drug repurposing against COVID-19 elucidating the mechanism of action and therapeutic effect of drugs. The effect of drugs and disease on various biological networks is modelled by network-based approaches, enabling the interaction networks between disease perturbations and drug targets to be interpreted and explained. Knowledge graph construction is a process by which interactions between various biological entities is identified in literature and aggregated into a network of interactions. This process is typically tedious and time-consuming, and AI is leveraged here for data mining literature enabling the construction of these networks with more ease. Richardson et al.'s knowledge graph suggested baricitinib for repurposing against COVID-19. The basis of this prediction was modulation of JAK/STAT signalling leading to inhibition of hyperinflammation, as well as modulation of GAK and AAK1 which are involved in cell mediated endocytosis in theory affecting viral entry.

One of the papers, by Richardson et al. is especially critical for establishing the power of AI based research given that their knowledge graph predicted baricitinib for repurposing and the mechanistic predictions were validated after publication of this research (Han et al., 2021), demonstrating inhibition of cytokines linked to hyperinflammation in COVID-19. Reduction of SARS-CoV-2 infection of human liver cells in super resolution microscopy via Baricitinib was also demonstrated providing in-vitro evidence of antiviral activity against SARS-CoV-2 (Stebbing et al., 2021). Additionally, Baricitinib was seen to cause reduction in expression of interferon-stimulated genes linked with platelet activation, hence potentially affecting microthrombosis seen in SARS-CoV-2 infection (Guo et al., 2020)(Eskandarian Boroujeni et al., 2022). Baricitinib is taken orally and is a once per day dose conferring advantage practically for patients. It is cleared predominantly renally with low plasma protein binding. Other antivirals evaluated at the start of the pandemic were also predominantly cleared renally, hence these properties allowed baricitinib to be readily dosed with the other antivirals being tested at the start of the pandemic. Observations in small scale clinical studies across Europe demonstrated that baricitinib reduced COVID-19 mortality and improvements in lung infection (Rodriguez-Garcia et al., 2020) (Bronte et al., 2020) (Titanji et al., 2021). Ultimately, randomised clinical trial ACTT-2 with over 1000 patients showed significant reduction in mortality and faster recovery for patients treated with baricitinib and remdesivir compared with remdesivir alone (Kalil et al., 2020). Based on this result, FDA issued emergency use authorisation for this combination therapy. Beneficial effects of baricinitib have been observed both in with or without remdesivir in phase 3 trial (Marconi et al., 2021). This particular research is critical because it exemplifies AI's ability to aid the search for pharmaceuticals and serves as a paradigm for the quick identification of potentially effective treatments in future pandemics. Repurposing drug via this method demonstrates how quickly repurposing is compared to traditional drug discovery.

4.2 AI and De Novo Drug Discovery

none of the de novo-identified molecules have been reported to have entered clinical trials. This result agrees with other reviews of AI for de-novo design (Floresta et al., 2022). This may be the case given that De novo drug development is not a viable option in an emergency circumstance like Covid-19 since there is an urgent need to identify medicines that can be provided promptly to reduce death. There is also a great deal of uncertainty when testing de novo molecules and repurposed drugs are more attractive candidates. This is an issue as all validation of de novo molecules remains in silico, with no in vitro data to support their use, so AI-driven models are difficult to appraise given the lack of validation data.

4.2.1 protein-ligand binding prediction based to screen drug-like molecule libraries

AI-based de novo design of COVID-19 therapies is directed by information based on ligands and receptor structures. In structure-based de novo drug development, model training and inpit data relates to knowledge of the three-dimensional structure of the receptor and its active site in order to make predictions about the binding energies and other steric interactions that exist between a ligand and a receptor. They are an approach that is very similar to that of regression task structural-based drug repurposing; the only difference is that these models were used to screen drug-like molecule libraries rather than approved drug libraries.

4.2.2 Generating novel drug-like molecules

Alternatively, ligand-based approaches are, in a way, target-agnostic in so far that models do not receive the target structure as an input, but rather identify data patterns in datasets of known inhibitors of the target of interest to identify key structural features and generate novel molecules. Both methods have been applied to *de novo* COVID-19 drug discovery efforts as reported in the results for identification of potential initial molecular scaffolds for further optimization. Given the estimated chemical space of drug-like compounds predicted to be on the range of 10^{60}-10^{100}, one of the key problems of AI driven de novo drug discovery is the generation of new molecular entities that are target specific.

Generating novel molecular entities in a target specific manner, given the vast chemical space of drug-like molecules estimated to be in the order of 10^{60}-10^{100} (Dobson, 2004), is thus considered one of the primary challenges of AI mediated *de novo* drug discovery. Methods that do not constrain the chemical space face issues with combinatory explosion, thus AI-mediated de novo molecule generation relies on the principle of local optimization (Schneider and Fechner, 2005).

4.3 AI and Vaccine Design

4.3.1 Vaccine epitope design

Aside from the very important role artificial intelligence plays in aiding drug design and drug repurposing against COVID-19, efforts to use AI programmes to help make more effective COVID-19 vaccines have been identified. Through B-cells (humoral immunity) and T-cells (cellular immunity), the body makes antibodies that are required in processes to ultimately kill viruses. Memory cells are responsible for recognising an antigen of a pathogen that has been eliminated, and any kind of re-exposure to the pathogen quickly turns on more effector T-cells. Vaccines are made by taking advantage of these processes. The major histocompatibility complex proteins (MHC I and MHC II proteins) are the helper proteins that show the binding regions of the antigens, called epitopes, to the antibodies, B- or T-cells so that they can bind to them and neutralise them. AI programmes can find epitopes, which are the viral antigenic parts that are more likely to be exposed on the surface of an infected cell and can bind to antibodies (Bali and Bali, 2022).

A Monte Carlo-based ML simulation has been employed to suggest viable epitopes for SARS-CoV-2 vaccines blueprinting (Malone et al., 2020), including ORF3a and ORF1a/b.

The neural network approach known as NetMHC has been shown to accurately predict which peptides would attach to MHC proteins and, as a result, identify epitopes for the SARS-CoV-2 vaccination (Prachar et al., 2020), identifying namely S, nsp3 and nsp8 as highly antigenic T cell epitopes.

Several in-vitro based studies proport data that corroborates these AI-driven computational analyses, suggesting the immunological presentation of SARS-CoV-2 non-structural protein peptides (Gangaev et al., 2021) (Saini et al., 2021) (Ferretti et al., 2020) (Grau-Expósito et al., 2021). In a cohort investigation, SARS-CoV-2 immunodominant peptides were mostly from ORF1ab and 10% of HLA epitopes were spike protein (Ferretti et al., 2020). 18 COVID-19 patients in a Denmark cohort study had 27% SARS-CoV-2-reactive CD8+ T cells (Saini et al., 2021). Immunodominant SARS-CoV-2 peptides were from ORF3 and ORF1ab, not the spike

protein (Saini et al., 2021). In another cohort investigation, ORF1ab CD8+ T cell epitopes lasted longer than spike protein epitopes (Gangaev et al., 2021). SARS-CoV-2 specific CD4+ and CD8+ T lymphocytes in the respiratory tract indicated restricting illness progression and re-infections (Grau-Expósito et al., 2021). CD4+ and CD8+ SARS-CoV-2 T cells expressed interferon, CD107a, interleukin-4, and interleukin-10 (Grau-Expósito et al., 2021). Airway infections and systemic inflammation are prevented by CD8+ T lymphocytes (Grau-Expósito et al., 2021). Thus, AI-based methods suggesting the use of non-structural proteins have been supported by a body of non-AI based evidence to support their predictions. These predictions can be used to create vaccines that cover a population with variation in their genetics and maximize the immunity gained in response to the vaccine.

4.7 Future perspective

AI-driven drug repurposing, de novo development, and vaccine design have been published during the COVID-19. Importantly, drug repurposing applications have shown to accelerate therapy development and latest applications in covid-19 demonstrate that knowledge graph-based computational paradigms have real-world implications. The interpretability of this approach is one of its great strengths as the suggested mechanism of action can be validated in-vitro to support the knowledge graph network. It also reveals the antiviral capacity of kinase inhibitors via their affect on cell mediated endocytosis (García-Cárceles et al., 2021). Hence, looking at the interactome network with regards to baricitinib, richardson et al. also proport the use of baricitinib against flavivirus infection, namely dengue fever. So, this knowledge network may be used in future for drug repurposing other indications/drugs.

On the other hand, it is concerning that many COVID-19 AI publications have not been peer-reviewed and remain published in https://arxiv.org/ or https://www.biorxiv.org/. Without adequate peer review, data is not scrutinized and can be misleading. Misleading data negatively affects drug development efforts, confuses policy makers and compromises clinical practice if implemented (Levin et al., 2020). Furthermore, review papers investigating the use of AI in COVID-19 therapeutic development fail to mention the lack of peer-review as an issue. Without proven reproducibility, data should not be utilised to inform drug development, as it has not been appropriately scrutinized. Efforts to establish validation benchmarks, transparent methodology reporting and robust data sharing will be critical for establishing standard level of trust in AI reports. Though, currently there exists a great deal of variation in the quality of reporting and so in-general AI-driven methods remain risky.

Reference list

Asaad, M., Habibullah, N.K. and Butler, C.E. (2020). The Impact of COVID-19 on Clinical Trials. *Annals of Surgery*, Publish Ahead of Print. doi:10.1097/sla.0000000000004113.

Ashburn, T.T. and Thor, K.B. (2004). Drug repositioning: identifying and developing new uses for existing drugs. *Nature Reviews Drug Discovery*, [online] 3(8), pp.673–683. doi:10.1038/nrd1468.

Bali, A. and Bali, N. (2022). Role of artificial intelligence in fast-track drug discovery and vaccine development for COVID-19. *Novel AI and Data Science Advancements for Sustainability in the Era of COVID-19*, pp.201–229. doi:10.1016/b978-0-323-90054-6.00006-4.

Batra, R., Chan, H., Kamath, G., Ramprasad, R., Cherukara, M.J. and Sankaranarayanan, S.K.R.S. (2020). Screening of Therapeutic Agents for COVID-19 Using Machine Learning and Ensemble Docking Studies. *The Journal of Physical Chemistry Letters*, [online] 11(17), pp.7058–7065. doi:10.1021/acs.jpclett.0c02278.

Beck, B.R., Shin, B., Choi, Y., Park, S. and Kang, K. (2020). Predicting commercially available antiviral drugs that may act on the novel coronavirus (SARS-CoV-2) through a drug-target interaction deep learning model. *Computational and Structural Biotechnology Journal*, 18, pp.784–790. doi:10.1016/j.csbj.2020.03.025.

Belyaeva, A., Cammarata, L., Radhakrishnan, A., Squires, C., Yang, K.D., Shivashankar, G.V. and Uhler, C. (2021). Causal network models of SARS-CoV-2 expression and aging to identify candidates for drug repurposing. *Nature Communications*, 12(1). doi:10.1038/s41467-021-21056-z.

Blasiak, A., Lim, J.J., Seah, S.G.K., Kee, T., Remus, A., Chye, D.H., Wong, P.S., Hooi, L., Truong, A.T.L., Le, N., Chan, C.E.Z., Desai, R., Ding, X., Hanson, B.J., Chow, E.K.-H. and Ho, D. (2021). IDentif.AI: Rapidly optimizing combination therapy design against severe Acute Respiratory Syndrome Coronavirus 2 (SARS-Cov-2) with digital drug development. *Bioengineering & Translational Medicine*, [online] 6(1), p.e10196. doi:10.1002/btm2.10196.

Brachman, R.J. and Levesque, H.J. (2009). *Knowledge representation and reasoning*. Amsterdam: Elsevier.

Bronte, V., Ugel, S., Tinazzi, E., Vella, A., De Sanctis, F., Canè, S., Batani, V., Trovato, R., Fiore, A., Petrova, V., Hofer, F., Barouni, R.M., Musiu, C., Caligola, S., Pinton, L., Torroni, L., Polati, E., Donadello, K., Friso, S. and Pizzolo, F. (2020). Baricitinib restrains the immune dysregulation in patients with severe COVID-19. *The Journal of Clinical Investigation*, [online] 130(12), pp.6409–6416. doi:10.1172/JCI141772.

Bung, N., Krishnan, S.R., Bulusu, G. and Roy, A. (2021). De novo design of new chemical entities for SARS-CoV-2 using artificial intelligence. *Future Medicinal Chemistry*, 13(6), pp.575–585. doi:10.4155/fmc-2020-0262.

Callaway, E. (2020). 'It will change everything': DeepMind's AI makes gigantic leap in solving protein structures. *Nature*, [online] 588. doi:10.1038/d41586-020-03348-4.

Cao, C. and Moult, J. (2014). GWAS and drug targets. *BMC Genomics*, 15(S4). doi:10.1186/1471-2164-15-s4-s5.

Cao, S., Gan, Y., Wang, C., Bachmann, M., Wei, S., Gong, J., Huang, Y., Wang, T., Li, L., Lu, K., Jiang, H., Gong, Y., Xu, H., Shen, X., Tian, Q., Lv, C., Song, F., Yin, X. and Lu, Z. (2020). Post-lockdown SARS-CoV-2 nucleic acid screening in nearly ten million residents of Wuhan, China. *Nature Communications*, 11(1). doi:10.1038/s41467-020-19802-w.

Cha, Y., Erez, T., Reynolds, I.J., Kumar, D., Ross, J., Koytiger, G., Kusko, R., Zeskind, B., Risso, S., Kagan, E., Papapetropoulos, S., Grossman, I. and Laifenfeld, D. (2017). Drug repurposing from the perspective of pharmaceutical companies. *British Journal of Pharmacology*, 175(2), pp.168–180. doi:10.1111/bph.13798.

Che, M., Yao, K., Che, C., Cao, Z. and Kong, F. (2021). Knowledge-Graph-Based Drug Repositioning against COVID-19 by Graph Convolutional Network with Attention Mechanism. *Future Internet*, 13(1), p.13. doi:10.3390/fi13010013.

Cook, D., Brown, D., Alexander, R., March, R., Morgan, P., Satterthwaite, G. and Pangalos, M.N. (2014). Lessons learned from the fate of AstraZeneca's drug pipeline: a five-dimensional framework. *Nature Reviews Drug Discovery*, [online] 13(6), pp.419–431. doi:10.1038/nrd4309.

Corona, G., Pizzocaro, A., Vena, W., Rastrelli, G., Semeraro, F., Isidori, A.M., Pivonello, R., Salonia, A., Sforza, A. and Maggi, M. (2021). Diabetes is most important cause for mortality in COVID-19 hospitalized patients: Systematic review and meta-analysis. *Reviews in Endocrine and Metabolic Disorders*, 22(2), pp.275–296. doi:10.1007/s11154-021-09630-8.

Dettmann, L.M., Adams, S. and Taylor, G. (2022). Investigating the prevalence of anxiety and depression during the first COVID-19 lockdown in the United Kingdom: Systematic review and meta-analyses. *British Journal of Clinical Psychology*. doi:10.1111/bjc.12360.

Diabetes UK (2020). *Diabetes Prevalence 2019*. [online] Diabetes UK. Available at: https://www.diabetes.org.uk/professionals/position-statements-reports/statistics/diabetes-prevalence-2019.

Dobson, C.M. (2004). Chemical space and biology. *Nature*, 432(7019), pp.824–828. doi:10.1038/nature03192.

Dolle, R.E. (2010). Historical Overview of Chemical Library Design. *Methods in Molecular Biology*, pp.3–25. doi:10.1007/978-1-60761-931-4_1.

Editor (2019). *Since 1996, the number of people with diabetes in the UK has risen from 1.4 million to 3.5 million. Diabetes prevalence is estimated to rise to 5 million by 2025.* [online] Diabetes. Available at: https://www.diabetes.co.uk/diabetes-prevalence.html#:~:text=World%20diabetes%20prevalence.

Eskandarian Boroujeni, M., Sekrecka, A., Antonczyk, A., Hassani, S., Sekrecki, M., Nowicka, H., Lopacinska, N., Olya, A., Kluzek, K., Wesoly, J. and Bluyssen, H.A.R. (2022). Dysregulated Interferon Response and Immune Hyperactivation in Severe COVID-19: Targeting STATs as a Novel Therapeutic Strategy. *Frontiers in Immunology*, 13. doi:10.3389/fimmu.2022.888897.

Evenson, R.E. (1993). Patents, R&D, and Invention Potential: International Evidence. *The American Economic Review*, [online] 83(2), pp.463–468. Available at: http://www.jstor.org/stable/2117709 [Accessed 22 Jul. 2022].

Fast, E., Altman, R.B. and Chen, B. (2020). Potential T-cell and B-cell Epitopes of 2019-nCoV. doi:10.1101/2020.02.19.955484.

Ferretti, A.P., Kula, T., Wang, Y., Nguyen, D.M.V., Weinheimer, A., Dunlap, G.S., Xu, Q., Nabilsi, N., Perullo, C.R., Cristofaro, A.W., Whitton, H.J., Virbasius, A., Olivier, K.J., Buckner, L.R., Alistar, A.T., Whitman, E.D., Bertino, S.A., Chattopadhyay, S. and MacBeath, G. (2020). Unbiased Screens Show CD8+ T Cells of COVID-19 Patients Recognize Shared Epitopes in SARS-CoV-2 that Largely Reside outside the Spike Protein. *Immunity*, [online] 53(5), pp.1095-1107.e3. doi:10.1016/j.immuni.2020.10.006.

Floresta, G., Zagni, C., Gentile, D., Patamia, V. and Rescifina, A. (2022). Artificial Intelligence Technologies for COVID-19 De Novo Drug Design. *International Journal of Molecular Sciences*, 23(6), p.3261. doi:10.3390/ijms23063261.

Gabel, J., Desaphy, J. and Rognan, D. (2014). Beware of Machine Learning-Based Scoring Functions—On the Danger of Developing Black Boxes. *Journal of Chemical Information and Modeling*, 54(10), pp.2807–2815. doi:10.1021/ci500406k.

Gangaev, A., Ketelaars, S.L.C., Isaeva, O.I., Patiwael, S., Dopler, A., Hoefakker, K., De Biasi, S., Gibellini, L., Mussini, C., Guaraldi, G., Girardis, M., Ormeno, C.M.P.T., Hekking, P.J.M., Lardy, N.M., Toebes, M., Balderas, R., Schumacher, T.N., Ovaa, H., Cossarizza, A. and Kvistborg, P. (2021). Identification and characterization of a SARS-CoV-2 specific CD8+ T cell response with immunodominant features. *Nature Communications*, [online] 12(1). doi:10.1038/s41467-021-22811-y.

Garcés-Ayala, F., Araiza-Rodríguez, A., Mendieta-Condado, E., Rodríguez-Maldonado, A.P., Wong-Arámbula, C., Landa-Flores, M., del Mazo-López, J.C., González-Villa, M., Escobar-Escamilla, N., Fragoso-Fonseca, D.E., Esteban-Valencia, M. del C., Lloret-Sánchez, L., Arellano-Suarez, D.S., Nuñez-García, T.E., Contreras-González, N.B., Cruz-Ortiz, N., Ruiz-López, A., Fierro-Valdez, M.Á., Regalado-Santiago, D. and Martínez-Velázquez, N. (2020). Full genome sequence of the first SARS-CoV-2 detected in Mexico. *Archives of Virology*, 165(9), pp.2095–2098. doi:10.1007/s00705-020-04695-3.

García-Cárceles, J., Caballero, E., Gil, C. and Martínez, A. (2021). Kinase Inhibitors as Underexplored Antiviral Agents. *Journal of Medicinal Chemistry*, 65(2), pp.935–954. doi:10.1021/acs.jmedchem.1c00302.

Ge, Y., Tian, T., Huang, S., Wan, F., Li, J., Li, S., Wang, X., Yang, H., Hong, L., Wu, N., Yuan, E., Luo, Y., Cheng, L., Hu, C., Lei, Y., Shu, H., Feng, X., Jiang, Z., Wu, Y. and Chi,

Y. (2021). An integrative drug repositioning framework discovered a potential therapeutic agent targeting COVID-19. *Signal Transduction and Targeted Therapy*, [online] 6(1). doi:10.1038/s41392-021-00568-6.

Ghallab, M., Nau, D.S., Traverso, P. and Cambridge University Press (2016). *Automated planning and acting*. New York: Cambridge University Press, Cop.

Glaab, E., Manoharan, G.B. and Abankwa, D. (2021). Pharmacophore Model for SARS-CoV-2 3CLpro Small-Molecule Inhibitors and *in Vitro* Experimental Validation of Computationally Screened Inhibitors. *Journal of Chemical Information and Modeling*, 61(8), pp.4082–4096. doi:10.1021/acs.jcim.1c00258.

Grau-Expósito, J., Sánchez-Gaona, N., Massana, N., Suppi, M., Astorga-Gamaza, A., Perea, D., Rosado, J., Falcó, A., Kirkegaard, C., Torrella, A., Planas, B., Navarro, J., Suanzes, P., Álvarez-Sierra, D., Ayora, A., Sansano, I., Esperalba, J., Andrés, C., Antón, A. and Ramón y Cajal, S. (2021). Peripheral and lung resident memory T cell responses against SARS-CoV-2. *Nature Communications*, [online] 12(1), p.3010. doi:10.1038/s41467-021-23333-3.

Guo, T., Fan, Y., Chen, M., Wu, X., Zhang, L., He, T., Wang, H., Wan, J., Wang, X. and Lu, Z. (2020). Cardiovascular Implications of Fatal Outcomes of Patients With Coronavirus Disease 2019 (COVID-19). *JAMA Cardiology*, 5(7). doi:10.1001/jamacardio.2020.1017.

Gysi, D.M., Valle, Í. do, Zitnik, M., Ameli, A., Gan, X., Varol, O., Ghiassian, S.D., Patten, J.J., Davey, R.A., Loscalzo, J. and Barabási, A.-L. (2021). Network medicine framework for identifying drug-repurposing opportunities for COVID-19. *Proceedings of the National Academy of Sciences*, [online] 118(19). doi:10.1073/pnas.2025581118.

Han, L., Shan, G., Chu, B., Wang, H., Wang, Z., Gao, S. and Zhou, W. (2021a). Accelerating drug repurposing for COVID-19 treatment by modeling mechanisms of action using cell image features and machine learning. *Cognitive Neurodynamics*. doi:10.1007/s11571-021-09727-5.

Han, L., Shan, G.C., Chu, B.F., Wang, H.Y., Wang, Z.J., Gao, S.Q. and Zhou, W.X. (2021b). Accelerating drug repurposing for COVID-19 via modeling drug mechanism of action with large scale gene-expression profiles. *arXiv:2005.07567 [cs, q-bio, stat]*. [online] Available at: https://arxiv.org/abs/2005.07567.

Heiser, K., McLean, P.F., Davis, C.T., Fogelson, B., Gordon, H.B., Jacobson, P., Hurst, B., Miller, B., Alfa, R.W., Earnshaw, B.A., Victors, M.L., Chong, Y.T., Haque, I.S., Low, A.S. and Gibson, C.C. (2020). Identification of potential treatments for COVID-19 through artificial intelligence-enabled phenomic analysis of human cells infected with SARS-CoV-2. doi:10.1101/2020.04.21.054387.

Hofmarcher, M., Mayr, A., Rumetshofer, E., Ruch, P., Renz, P., Schimunek, J., Seidl, P., Vall, A., Widrich, M., Hochreiter, S. and Klambauer, G. (2020). Large-Scale Ligand-Based Virtual Screening for SARS-CoV-2 Inhibitors Using Deep Neural Networks. *SSRN Electronic Journal*. doi:10.2139/ssrn.3561442.

House of Commons (2021). *Coronavirus: Lessons Learned to Date Sixth Report of the Health and Social Care Committee and Third Report of the Science and Technology Committee of Session 2021-22 HC 92*. [online] Available at: https://committees.parliament.uk/publications/7496/documents/78687/default/.

Hu, F., Jiang, J. and Yin, P. (2021). Prediction of potential commercially inhibitors against SARS-CoV-2 by multi-task deep model. *arXiv:2003.00728 [q-bio]*. [online] Available at: https://arxiv.org/abs/2003.00728 [Accessed 24 Jul. 2022].

Jin, Z., Du, X., Xu, Y., Deng, Y., Liu, M., Zhao, Y., Zhang, B., Li, X., Zhang, L., Peng, C., Duan, Y., Yu, J., Wang, L., Yang, K., Liu, F., Jiang, R., Yang, X., You, T., Liu, X. and Yang, X. (2020). Structure of Mpro from COVID-19 virus and discovery of its inhibitors. *Nature*, 582, pp.289–293. doi:10.1038/s41586-020-2223-y.

John Hopkins School Of Medicine (n.d.). *Vaccine Research & Development*. [online] Johns Hopkins Coronavirus Resource Center. Available at: https://coronavirus.jhu.edu/vaccines/timeline.

Kadioglu, O., Saeed, M., Greten, H.J. and Efferth, T. (2021). Identification of novel compounds against three targets of SARS CoV-2 coronavirus by combined virtual screening and supervised machine learning. *Computers in Biology and Medicine*, [online] 133, p.104359. doi:10.1016/j.compbiomed.2021.104359.

Kalil, A.C., Patterson, T.F., Mehta, A.K., Tomashek, K.M., Wolfe, C.R., Ghazaryan, V., Marconi, V.C., Ruiz-Palacios, G.M., Hsieh, L., Kline, S., Tapson, V., Iovine, N.M., Jain,

M.K., Sweeney, D.A., El Sahly, H.M., Branche, A.R., Regalado Pineda, J., Lye, D.C., Sandkovsky, U. and Luetkemeyer, A.F. (2020). Baricitinib plus Remdesivir for Hospitalized Adults with Covid-19. *New England Journal of Medicine*. doi:10.1056/nejmoa2031994.

Karki, N., Verma, N., Trozzi, F., Tao, P., Kraka, E. and Zoltowski, B. (2021). Predicting Potential SARS-COV-2 Drugs-In Depth Drug Database Screening Using Deep Neural Network Framework SSnet, Classical Virtual Screening and Docking. *International Journal of Molecular Sciences*, [online] 22(4), p.1573. doi:10.3390/ijms22041573.

Ke, Y.-Y., Peng, T.-T., Yeh, T.-K., Huang, W.-Z., Chang, S.-E., Wu, S.-H., Hung, H.-C., Hsu, T.-A., Lee, S.-J., Song, J.-S., Lin, W.-H., Chiang, T.-J., Lin, J.-H., Sytwu, H.-K. and Chen, C.-T. (2020). Artificial intelligence approach fighting COVID-19 with repurposing drugs. *Biomedical Journal*. [online] doi:10.1016/j.bj.2020.05.001.

Keiser, M.J., Roth, B.L., Armbruster, B.N., Ernsberger, P., Irwin, J.J. and Shoichet, B.K. (2007). Relating protein pharmacology by ligand chemistry. *Nature Biotechnology*, 25(2), pp.197–206. doi:10.1038/nbt1284.

Kim, J., Zhang, J., Cha, Y., Kolitz, S., Funt, J., Escalante Chong, R., Barrett, S., Kusko, R., Zeskind, B. and Kaufman, H. (2020). Advanced bioinformatics rapidly identifies existing therapeutics for patients with coronavirus disease-2019 (COVID-19). *Journal of Translational Medicine*, 18(1). doi:10.1186/s12967-020-02430-9.

Kumar, S. and Kumar, S. (2019). Molecular Docking: A Structure-Based Approach for Drug Repurposing. *In Silico Drug Design*, pp.161–189. doi:10.1016/b978-0-12-816125-8.00006-7.

Levin, J.M., Oprea, T.I., Davidovich, S., Clozel, T., Overington, J.P., Vanhaelen, Q., Cantor, C.R., Bischof, E. and Zhavoronkov, A. (2020). Artificial intelligence, drug repurposing and peer review. *Nature Biotechnology*, 38(10), pp.1127–1131. doi:10.1038/s41587-020-0686-x.

Li, B., Shin, H., Gulbekyan, G., Pustovalova, O., Nikolsky, Y., Hope, A., Bessarabova, M., Schu, M., Kolpakova-Hart, E., Merberg, D., Dorner, A. and Trepicchio, W.L. (2015). Development of a Drug-Response Modeling Framework to Identify Cell Line Derived Translational Biomarkers That Can Predict Treatment Outcome to Erlotinib or Sorafenib. *PLOS ONE*, 10(6), p.e0130700. doi:10.1371/journal.pone.0130700.

Liu, H.H., Ezekowitz, M.D., Columbo, M., Khan, O., Martin, J., Spahr, J., Yaron, D.,

Cushinotto, L. and Kapelusznik, L. (2021). The future is now: our experience starting a remote clinical trial during the beginning of the COVID-19 pandemic. *Trials*, 22(1). doi:10.1186/s13063-021-05537-6.

Liu, Z., Chen, X., Carter, W., Moruf, A., Komatsu, T.E., Pahwa, S., Chan-Tack, K., Snyder, K., Petrick, N., Cha, K., Lal-Nag, M., Hatim, Q., Thakkar, S., Lin, Y., Huang, R., Wang, D., Patterson, T.A. and Tong, W. (2022). AI-powered drug repurposing for developing COVID-19 treatments. *Reference Module in Biomedical Sciences*. doi:10.1016/b978-0-12-824010-6.00005-8.

Ma, J., Sheridan, R.P., Liaw, A., Dahl, G.E. and Svetnik, V. (2015). Deep Neural Nets as a Method for Quantitative Structure–Activity Relationships. *Journal of Chemical Information and Modeling*, 55(2), pp.263–274. doi:10.1021/ci500747n.

Magar, R., Yadav, P. and Barati Farimani, A. (2021). Potential neutralizing antibodies discovered for novel corona virus using machine learning. *Scientific Reports*, [online] 11(1). doi:10.1038/s41598-021-84637-4.

Malone, B., Simovski, B., Moliné, C., Cheng, J., Gheorghe, M., Fontenelle, H., Vardaxis, I., Tennøe, S., Malmberg, J.-A., Stratford, R. and Clancy, T. (2020). Artificial intelligence predicts the immunogenic landscape of SARS-CoV-2 leading to universal blueprints for vaccine designs. *Scientific Reports*, 10(1). doi:10.1038/s41598-020-78758-5.

Marconi, V.C., Ramanan, A.V., de Bono, S., Kartman, C.E., Krishnan, V., Liao, R., Piruzeli, M.L.B., Goldman, J.D., Alatorre-Alexander, J., de Cassia Pellegrini, R., Estrada, V., Som, M., Cardoso, A., Chakladar, S., Crowe, B., Reis, P., Zhang, X., Adams, D.H., Ely, E.W. and Ahn, M.-Y. (2021). Efficacy and safety of baricitinib for the treatment of hospitalised adults with COVID-19 (COV-BARRIER): a randomised, double-blind, parallel-group, placebo-controlled phase 3 trial. *The Lancet Respiratory Medicine*. doi:10.1016/s2213-2600(21)00331-3.

Mayr, L.M. and Fuerst, P. (2008). The future of high-throughput screening. *Journal of Biomolecular Screening*, [online] 13(6), pp.443–448. doi:10.1177/1087057108319644.

McDermott, M.M. and Newman, A.B. (2020). Preserving Clinical Trial Integrity During the Coronavirus Pandemic. *JAMA*. doi:10.1001/jama.2020.4689.

MHRA (n.d.). *MHRA guidance on coronavirus (COVID-19)*. [online] GOV.UK. Available at: https://www.gov.uk/government/collections/mhra-guidance-on-coronavirus-covid-19#clinical-trials [Accessed 22 Jul. 2022].

Mohamed Zakaria Kurdi (2017). *Natural Language Processing and Computational Linguistics 2: Semantics, Discourse and Applications*. Wiley-Blackwell.

Mohapatra, S., Nath, P., Chatterjee, M., Das, N., Kalita, D., Roy, P. and Satapathi, S. (2020). Repurposing therapeutics for COVID-19: Rapid prediction of commercially available drugs through machine learning and docking. *PLOS ONE*, 15(11), p.e0241543. doi:10.1371/journal.pone.0241543.

Moskal, M., Beker, W., Roszak, R., Gajewska, E.P., Wołos, A., Molga, K., Szymkuć, S., Grynkiewicz, G. and Grzybowski, B. (2020). Suggestions for second-pass anti-COVID-19 drugs based on the Artificial Intelligence measures of molecular similarity, shape and pharmacophore distribution. *chemrxiv.org*. [online] doi:10.26434/chemrxiv.12084690.v2.

Nand, M., Maiti, P., Joshi, T., Chandra, S., Pande, V., Kuniyal, J.C. and Ramakrishnan, M.A. (2020). Virtual screening of anti-HIV1 compounds against SARS-CoV-2: machine learning modeling, chemoinformatics and molecular dynamics simulation based analysis. *Scientific Reports*, [online] 10(1), p.20397. doi:10.1038/s41598-020-77524-x.

Neves, B.J., Braga, R.C., Melo-Filho, C.C., Moreira-Filho, J.T., Muratov, E.N. and Andrade, C.H. (2018). QSAR-Based Virtual Screening: Advances and Applications in Drug Discovery. *Frontiers in Pharmacology*, 9. doi:10.3389/fphar.2018.01275.

Nguyen, D.D., Gao, K., Chen, J., Wang, R. and Wei, G.-W. (2020). Potentially highly potent drugs for 2019-nCoV. doi:10.1101/2020.02.05.936013.

NIHR (2020). *DHSC issues guidance on the impact of COVID-19 on research funded or supported by NIHR*. [online] www.nihr.ac.uk. Available at: https://www.nihr.ac.uk/news/dhsc-issues-guidance-on-the-impact-on-covid-19-on-research-funded-or-supported-by-nihr/24469.

Office for National Statistics (2021). *Population Estimates for the UK, England and Wales, Scotland and Northern Ireland - Office for National Statistics*. [online] www.ons.gov.uk. Available at:

https://www.ons.gov.uk/peoplepopulationandcommunity/populationandmigration/populationestimates/bulletins/annualmidyearpopulationestimates/mid2020.

Ong, E., Wong, M.U., Huffman, A. and He, Y. (2020). COVID-19 Coronavirus Vaccine Design Using Reverse Vaccinology and Machine Learning. *Frontiers in Immunology*, 11(1581). doi:10.3389/fimmu.2020.01581.

Pammolli, F., Magazzini, L. and Riccaboni, M. (2011). The productivity crisis in pharmaceutical R&D. *Nature Reviews Drug Discovery*, [online] 10(6), pp.428–438. doi:10.1038/nrd3405.

Parmar, M., Thumar, R., Patel, B., Athar, M., Jha, P.C. and Patel, D. (2021). Structural Differences In 3C-like protease (Mpro) From SARS-CoV and SARS-CoV-2: Molecular Insights For Drug Repurposing Against COVID-19 Revealed by Molecular Dynamics Simulations. doi:10.1101/2021.08.11.455903.

Pham, T.-H., Qiu, Y., Zeng, J., Xie, L. and Zhang, P. (2021). A deep learning framework for high-throughput mechanism-driven phenotype compound screening and its application to COVID-19 drug repurposing. *Nature Machine Intelligence*, 3(3), pp.247–257. doi:10.1038/s42256-020-00285-9.

Powell, K. (2016). Does it take too long to publish research? *Nature News*, [online] 530(7589), p.148. doi:10.1038/530148a.

Prachar, M., Justesen, S., Steen-Jensen, D.B., Thorgrimsen, S., Jurgons, E., Winther, O. and Bagger, F.O. (2020). Identification and validation of 174 COVID-19 vaccine candidate epitopes reveals low performance of common epitope prediction tools. *Scientific Reports*, [online] 10(1), p.20465. doi:10.1038/s41598-020-77466-4.

Richardson, P., Griffin, I., Tucker, C., Smith, D., Oechsle, O., Phelan, A. and Stebbing, J. (2020). Baricitinib as potential treatment for 2019-nCoV acute respiratory disease. *The Lancet*, [online] 395(10223), pp.e30–e31. doi:10.1016/S0140-6736(20)30304-4.

Richter, F. (2021). *COVID-19 has caused a huge amount of lost working hours*. [online] World Economic Forum. Available at: https://www.weforum.org/agenda/2021/02/covid-employment-global-job-loss/.

Rodriguez-Garcia, J.L., Sanchez-Nievas, G., Arevalo-Serrano, J., Garcia-Gomez, C., Jimenez-Vizuete, J.M. and Martinez-Alfaro, E. (2020). Baricitinib improves respiratory function in patients treated with corticosteroids for SARS-CoV-2 pneumonia: an observational cohort study. *Rheumatology*, 60(1), pp.399–407. doi:10.1093/rheumatology/keaa587.

Romero Starke, K., Petereit-Haack, G., Schubert, M., Kämpf, D., Schliebner, A., Hegewald, J. and Seidler, A. (2020). The Age-Related Risk of Severe Outcomes Due to COVID-19 Infection: A Rapid Review, Meta-Analysis, and Meta-Regression. *International Journal of Environmental Research and Public Health*, [online] 17(16). doi:10.3390/ijerph17165974.

Saini, S.K., Hersby, D.S., Tamhane, T., Povlsen, H.R., Hernandez, S.P.A., Nielsen, M., Gang, A.O. and Hadrup, S.R. (2021). SARS-CoV-2 genome-wide T cell epitope mapping reveals immunodominance and substantial CD8+ T cell activation in COVID-19 patients. *Science Immunology*, [online] 6(58). doi:10.1126/sciimmunol.abf7550.

Savioli, N. (2020). One-shot screening of potential peptide ligands on HR1 domain in COVID-19 glycosylated spike (S) protein with deep siamese network. *arXiv:2004.02136 [cs, q-bio]*. [online] Available at: https://arxiv.org/abs/2004.02136.

Schneider, G. and Fechner, U. (2005). Computer-based de novo design of drug-like molecules. *Nature Reviews Drug Discovery*, 4(8), pp.649–663. doi:10.1038/nrd1799.

Shin, B., Park, S., Kang, K. and Ho, J.C. (2019). Self-Attention Based Molecule Representation for Predicting Drug-Target Interaction. *arXiv:1908.06760 [cs, stat]*. [online] Available at: https://arxiv.org/abs/1908.06760 [Accessed 22 Jul. 2022].

Sostero, M., Milasi, S., Hurley, J., Fernández-Macías, E. and Bisello, M. (2020). *Teleworkability and the COVID-19 crisis: a new digital divide?* [online] www.econstor.eu. Available at: http://hdl.handle.net/10419/231337.

Srinivasan, S., Batra, R., Chan, H., Kamath, G., Cherukara, M.J. and Sankaranarayanan, S.K.R.S. (2021). Artificial Intelligence-Guided De Novo Molecular Design Targeting COVID-19. *ACS Omega*, 6(19), pp.12557–12566. doi:10.1021/acsomega.1c00477.

Stebbing, J., Sánchez Nievas, G., Falcone, M., Youhanna, S., Richardson, P., Ottaviani, S., Shen, J.X., Sommerauer, C., Tiseo, G., Ghiadoni, L., Virdis, A., Monzani, F., Rizos, L.R.,

Forfori, F., Avendaño Céspedes, A., De Marco, S., Carrozzi, L., Lena, F., Sánchez-Jurado, P.M. and Lacerenza, L.G. (2021). JAK inhibition reduces SARS-CoV-2 liver infectivity and modulates inflammatory responses to reduce morbidity and mortality. *Science Advances*, [online] 7(1), p.eabe4724. doi:10.1126/sciadv.abe4724.

Subramanian, A., Narayan, R., Corsello, S.M., Peck, D.D., Natoli, T.E., Lu, X., Gould, J., Davis, J.F., Tubelli, A.A., Asiedu, J.K., Lahr, D.L., Hirschman, J.E., Liu, Z., Donahue, M., Julian, B., Khan, M., Wadden, D., Smith, I.C., Lam, D. and Liberzon, A. (2017). A Next Generation Connectivity Map: L1000 Platform and the First 1,000,000 Profiles. *Cell*, [online] 171(6), pp.1437-1452.e17. doi:10.1016/j.cell.2017.10.049.

Tang, B., He, F., Liu, D., He, F., Wu, T., Fang, M., Niu, Z., Wu, Z. and Xu, D. (2022). AI-Aided Design of Novel Targeted Covalent Inhibitors against SARS-CoV-2. *Biomolecules*, 12(6), p.746. doi:10.3390/biom12060746.

Titanji, B.K., Farley, M.M., Mehta, A., Connor-Schuler, R., Moanna, A., Cribbs, S.K., O'Shea, J., DeSilva, K., Chan, B., Edwards, A., Gavegnano, C., Schinazi, R.F. and Marconi, V.C. (2021). Use of Baricitinib in Patients With Moderate to Severe Coronavirus Disease 2019. *Clinical Infectious Diseases: An Official Publication of the Infectious Diseases Society of America*, [online] 72(7), pp.1247–1250. doi:10.1093/cid/ciaa879.

Ton, A.-T., Gentile, F., Hsing, M., Ban, F. and Cherkasov, A. (2020). Rapid Identification of Potential Inhibitors of SARS-CoV-2 Main Protease by Deep Docking of 1.3 Billion Compounds. *Molecular Informatics*. doi:10.1002/minf.202000028.

Uhler, C. and Shivashankar, G.V. (2020). Mechano-genomic regulation of coronaviruses and its interplay with ageing. *Nature Reviews Molecular Cell Biology*, 21(5), pp.247–248. doi:10.1038/s41580-020-0242-z.

Van Brunt, J. (1986). Protein Architecture: Designing from the Ground Up. *Nature Biotechnology*, 4(4), pp.277–283. doi:10.1038/nbt0486-277.

Verma, J., Khedkar, V.M. and Coutinho, E.C. (2010). 3D-QSAR in drug design--a review. *Current Topics in Medicinal Chemistry*, [online] 10(1), pp.95–115. doi:10.2174/156802610790232260.

Verma, M. and Bansal, D. (2020). Novel Potential Inhibitors Against SARS-CoV-2 Using

Artificial Intelligence. doi:10.26434/chemrxiv.12228362.v3.

Wang, M.-Y., Zhao, R., Gao, L.-J., Gao, X.-F., Wang, D.-P. and Cao, J.-M. (2020a). SARS-CoV-2: Structure, Biology, and Structure-Based Therapeutics Development. *Frontiers in Cellular and Infection Microbiology*, 10(587269). doi:10.3389/fcimb.2020.587269.

Wang, S., Sun, Q., Xu, Y., Pei, J. and Lai, L. (2021). A transferable deep learning approach to fast screen potential antiviral drugs against SARS-CoV-2. *Briefings in Bioinformatics*, [online] 22(6). doi:10.1093/bib/bbab211.

Wang, X., Song, K., Li, L. and Chen, L. (2018). Structure-Based Drug Design Strategies and Challenges. *Current Topics in Medicinal Chemistry*, 18(12), pp.998–1006. doi:10.2174/1568026618666180813152921.

Wang, Y., Wang, Y., Chen, Y. and Qin, Q. (2020b). Unique epidemiological and clinical features of the emerging 2019 novel coronavirus pneumonia (COVID-19) implicate special control measures. *Journal of Medical Virology*, [online] 92(6). doi:10.1002/jmv.25748.

Wenzel, J., Matter, H. and Schmidt, F. (2019). Predictive Multitask Deep Neural Network Models for ADME-Tox Properties: Learning from Large Data Sets. *Journal of Chemical Information and Modeling*, 59(3), pp.1253–1268. doi:10.1021/acs.jcim.8b00785.

Wouters, O.J., McKee, M. and Luyten, J. (2020). Estimated Research and Development Investment Needed to Bring a New Medicine to Market, 2009-2018. *JAMA*, [online] 323(9), p.844. doi:10.1001/jama.2020.1166.

Yanez, N.D., Weiss, N.S., Romand, J.-A. and Treggiari, M.M. (2020). COVID-19 mortality risk for older men and women. *BMC Public Health*, [online] 20(1). doi:10.1186/s12889-020-09826-8.

Yang, Z., Bogdan, P. and Nazarian, S. (2021). An in silico deep learning approach to multi-epitope vaccine design: a SARS-CoV-2 case study. *Scientific Reports*, [online] 11(1), p.3238. doi:10.1038/s41598-021-81749-9.

Zhang, H., Saravanan, K.M., Yang, Y., Hossain, Md.T., Li, J., Ren, X., Pan, Y. and Wei, Y. (2020). Deep Learning Based Drug Screening for Novel Coronavirus 2019-nCov. *Interdisciplinary Sciences, Computational Life Sciences*, [online] pp.1–9.

doi:10.1007/s12539-020-00376-6.

Zhao, J. (1997). A liability theory of disease: the foundation of cell population pathology. *Medical Hypotheses*, 48(4), pp.341–346. doi:10.1016/s0306-9877(97)90104-3.

Zhavoronkov, A., Aladinskiy, V., Zhebrak, A., Zagribelnyy, B., Terentiev, V., Bezrukov, D.S., Polykovskiy, D., Shayakhmetov, R., Filimonov, A., Orekhov, P., Yan, Y., Popova, O., Vanhaelen, Q., Aliper, A. and Ivanenkov, Y. (2020). Potential COVID-2019 3C-like Protease Inhibitors Designed Using Generative Deep Learning Approaches. doi:10.26434/chemrxiv.11829102.v2.

Zhou, P., Yang, X.-L., Wang, X.-G., Hu, B., Zhang, L., Zhang, W., Si, H.-R., Zhu, Y., Li, B., Huang, C.-L., Chen, H.-D., Chen, J., Luo, Y., Guo, H., Jiang, R.-D., Liu, M.-Q., Chen, Y., Shen, X.-R., Wang, X. and Zheng, X.-S. (2020a). A pneumonia outbreak associated with a new coronavirus of probable bat origin. *Nature*, 579(7798). doi:10.1038/s41586-020-2012-7.

Zhou, Z., Ren, L., Zhang, L., Zhong, J., Xiao, Y., Jia, Z., Guo, L., Yang, J., Wang, C., Jiang, S., Yang, D., Zhang, G., Li, H., Chen, F., Xu, Y., Chen, M., Gao, Z., Yang, J., Dong, J. and Liu, B. (2020b). Heightened Innate Immune Responses in the Respiratory Tract of COVID-19 Patients. *Cell Host & Microbe*, [online] 27(6), pp.883-890.e2. doi:10.1016/j.chom.2020.04.017.

Zhu, J., Deng, Y.-Q., Wang, X., Li, X.-F., Zhang, N.-N., Liu, Z., Zhang, B., Qin, C.-F. and Xie, Z. (2020). An artificial intelligence system reveals liquiritin inhibits SARS-CoV-2 by mimicking type I interferon. doi:10.1101/2020.05.02.074021.