COVID-19 drug and vaccine r&d with AI & Machine Learning

**# Download or try**

**1. [Cheminformatics in Python: Predicting Solubility of Molecules | End-to-End Data Science** Project](<https://www.kaggle.com/makrovh/cheminformatics-in-python-predicting-solubility-o>"Cheminformatics in Python: Predicting Solubility of Molecules | End-to-End Data Science Project")

**2. [Subtyping COVID-19 Therapeutic Research** Findings](<https://www.kaggle.com/makrovh/subtyping-covid-19-therapeutic-research-findings> "Subtyping COVID-19 Therapeutic Research Findings")

Introductory documents:

**1. [DRUG & VACCINE R&D WITH AI & MACHINE** LEARNING-1](<https://docs.google.com/document/d/1E0Sgd4YU4WtgYkTHcYG8x30oq8VMs8X2uYqWmLGMwBI/edit?usp=sharing> "DRUG & VACCINE R&D WITH AI & MACHINE LEARNING-1")

**2. [Reinforcement learning-based drug research & discovery for the** COVID-19](<https://drive.google.com/open?id=1UHXSPd_8RHgjoWfYzumXZnygPshjv3Kp>"Reinforcement learning-based drug research & discovery for the COVID-19")

# INTRODUCTION

The way we discover drugs & vaccines is EXTREMELY inefficient. Something must have to be done here to decrease the future impacts of any kind of pandemic or biological warfare.

We are accelerating drug discovery by leveraging machine learning algorithms to generate and create retro-synthesis pathways for drugs molecule design and development, calculating minimum energy conformations of a potential candidate, calculating descriptors and shortlisting them based on the correlation coefficient, cross-correlation coefficient, dissimilarity distances, cluster analysis & genetic function approach, calculating drug-likeness property, ADME, applying Lipinski’s rule for shortlisting.

# 1. Cheminformatics in Python: Predicting Solubility of Molecules | End-to-End Data Science Project

In this kaggle notebook, we will dive into the world of Cheminformatics which lies at the interface of Informatics and Chemistry. We will be reproducing a research article (by John S. Delaney

1) by applying Linear Regression to predict the solubility of molecules (i.e. solubility of drugs is an important physicochemical property in Drug discovery, design, and development).

This idea for this notebook was inspired by the excellent blog post by Pat Walters where he reproduced the linear regression model with a similar degree of performance as that of Delaney. This example is also briefly described in the book Deep Learning for the Life Sciences: Applying Deep Learning to Genomics, Microscopy, Drug Discovery, and More.

# 2. Subtyping, COVID-19 Therapeutic Research Findings

The goal of this exercise is to study this literature provided by the Kaggle COVID-19 challenge organizing team and to subtype the COVID-19 therapeutic research findings. Specifically, we carried out the following four parts of analyses:

Part A. Drugs that have been used in clinical trials for COVID-19. We identified and characterized the drugs in clinical trials by integrating the FDA drug database and PubChem repository. We hand-curated and summarized the reported effectiveness of each drug. We presented the mutual similarity of chemical structures across the drugs used in clinical trials.

We categorized the drugs based on their molecular mechanisms, which can facilitate the discovery of related drugs of similar mechanisms and the creation of an effective cocktail treatment.

Category 1. RNA mutagens

Category 2. Protease inhibitors

Category 3. Virus-entry blockers

Category 4. Virus-release blockers

Category 5. Monoclonal antibodies

Part B. Drugs that have been proposed by computational works. We identified the computational publications for COVID-19 drugs, categorized their approaches into the following categories, and listed their previous applications in other disease domains, and potential limitations.

Category 1. Gene-gene network-based algorithms.

Category 2. Expression-based algorithms

Category 3. Docking simulation of protein structure-based for

Category 3.a. Small molecules

Category 3.b. Monoclonal antibodies

Part C. Drugs that have been proposed by in vitro experiments of COVID-19 invading human cells. We characterized the chemical structures and analyzed the chemical similarity for this group.

For this list, other than literature mining, we carried out a machine learning experiment to prioritize previously unexplored FDA-approved drugs (to circumvent ADMET evaluation) for repurposing. After hand-removing the contaminations, we identified the following top candidates for repurposing: OLUMIANT(Baricitinib) used to treat rheumatoid arthritis, BRIMONIDINE, used to treat glaucoma, EDURANT(rilpivirine) used to treat Human Immunodeficiency Virus-1 (HIV-1), MARPLAN used to treat depression, Corlanor (ivabradine) used to reduce the spontaneous pacemaker activity of the cardiac sinus node. We listed the potential contaminations/biases in this and relevant protein binding-associated approaches.

Part D. Epitope study for vaccines We categorized vaccine studies by their approaches and discussed the background and limitations concerning evolution:

Approach 1. Homology-based with SARS-COV (the 2003 version of SARS), other coronavirus or Ebola.

Approach 2. Immunoinformatics including docking/molecular dynamics/protein structures/antigenicity predictions.

We hand-curated a list of 147 epitopes from these publications and their supplementary materials, grouped them by the source virus proteins, human T-cell/B-cell targets and MHC class. We merged all published epitopes into 124 consolidated groups by partial sub-sequence search and 91 unique virus protein sequence regions by BFS search algorithms. We hope the above lists will serve as the 'wisdom-of-the-crowd' reference for vaccine development.

Summary points and future recommended research topics for Phase 2.

Conclusion 1. There is not a single drug for which consistent positive response has been reported.

Conclusion 2. There are overlaps between the drugs in clinical trials, proposed by computational analysis and proposed by in vitro experiments. However, some of the overlaps, especially those with computational analysis may come from circularity in the methods.

Conclusion 3. Drug candidates proposed by computation and in vitro screening could be biased towards cancer-related targeted therapy and substantially contaminated by existing literature or sometimes anecdotes. This bias/contamination may affect a significant number of computation-based drug-repurposing studies including our work, and certainly not limited to COVID-19.

Future direction 1. Disagreement in the reported drug response can root from differences in dosage, baseline biometrics, and population groups. With more clinical trial results coming in, the next step is to carry out a meta-analysis to stratify these variables.

Future direction 2. Analyzing vaccine findings at this stage is premature as there is no clinical effectiveness study yet. It will be meaningful to make genome variation and vaccines (or maybe antibodies as well) into the same topic, therefore allowing connecting the genome variations to what fraction of the virus strains that a vaccine could cover.

Future direction 3. We suggest a topic on news (e.g., google news) retrieval for therapeutic development, as many (if not most) treatment responses may not first appear in manuscripts.

Finally, we would like to take this opportunity to make one comment: Literature tends to be biassed towards reporting positive results，known biology (e.g., cancer and immune- drugs), and anecdotes, and we should take the results of this exercise and other documents critically.

# DISCUSSION LIST

Join our discussion forum on [DRUG & VACCINE R&D WITH AI & MACHINE LEARNING](<https://castme.life/forums/topic/drug-vaccine-rd-with-ai-machine-learning/> "forum topic")

# INSTALLATION AND DATA REQUIREMENTS

Check the following datasets, these have been made public by our r&d department and some are already. Feel free to explore the data and augment it. These all data are either have already been included in the kaggle kernels or gets downloaded in that.

1. [Delaney's solubility dataset](<https://raw.githubusercontent.com/dataprofessor/data/master/delaney.csv> "Delaney's solubility dataset")

2. [COVID-19 Open Research Dataset Challenge (CORD-19)](<https://www.kaggle.com/allen-institute-for-ai/CORD-19-research-challenge>"COVID-19 Open Research Dataset Challenge (CORD-19)")

3. [drugbank.ca dataset](<https://www.kaggle.com/makrovh/drugbankca> "drugbank.ca dataset")

4. [drugbank.ca-chunk dataset](<https://www.kaggle.com/makrovh/drugbankcachunnk> "drugbank.ca-chunk dataset")

5. [drugbank.ca-csv-chunk dataset]([www.kaggle.com/makrovh/drugbankcacsvchunks](http://www.kaggle.com/makrovh/drugbankcacsvchunks) "drugbank.ca-chunk dataset")

# ACKNOWLEDGMENTS

# 1. Cheminformatics in Python: Predicting Solubility of Molecules | End-to-End Data Science Project

In this kaggle notebook, we will dive into the world of Cheminformatics which lies at the interface of Informatics and Chemistry. We will be reproducing a research article (by John S. Delaney

1) by applying Linear Regression to predict the solubility of molecules (i.e. solubility of drugs is an important physicochemical property in Drug discovery, design, and development).

This idea for this notebook was inspired by the excellent blog post by Pat Walters where he reproduced the linear regression model with a similar degree of performance as that of Delaney. This example is also briefly described in the book Deep Learning for the Life Sciences: Applying Deep Learning to Genomics, Microscopy, Drug Discovery, and More.

# 2. Subtyping, COVID-19 Therapeutic Research Findings

The goal of this exercise is to study this literature provided by the Kaggle COVID-19 challenge organizing team and to subtype the COVID-19 therapeutic research findings. Specifically, we carried out the following four parts of analyses:

Part A. Drugs that have been used in clinical trials for COVID-19. We identified and characterized the drugs in clinical trials by integrating the FDA drug database and PubChem repository. We hand-curated and summarized the reported effectiveness of each drug. We presented the mutual similarity of chemical structures across the drugs used in clinical trials.

We categorized the drugs based on their molecular mechanisms, which can facilitate the discovery of related drugs of similar mechanisms and the creation of an effective cocktail treatment:

Category 1. RNA mutagens

Category 2. Protease inhibitors

Category 3. Virus-entry blockers

Category 4. Virus-release blockers

Category 5. Monoclonal antibodies

Part B. Drugs that have been proposed by computational works. We identified the computational publications for COVID-19 drugs, categorized their approaches into the following categories, and listed their previous applications in other disease domains, and potential limitations.

Category 1. Gene-gene network-based algorithms.

Category 2. Expression-based algorithms

Category 3. Docking simulation of protein structure-based for

Category 3.a. Small molecules

Category 3.b. Monoclonal antibodies

Part C. Drugs that have been proposed by in vitro experiments of COVID-19 invading human cells. We characterized the chemical structures and analyzed the chemical similarity for this group.

For this list, other than literature mining, we carried out a machine learning experiment to prioritize previously unexplored FDA-approved drugs (to circumvent ADMET evaluation) for repurposing. After hand-removing the contaminations, we identified the following top candidates for repurposing: OLUMIANT(Baricitinib) used to treat rheumatoid arthritis, BRIMONIDINE, used to treat glaucoma, EDURANT(rilpivirine) used to treat Human Immunodeficiency Virus-1 (HIV-1), MARPLAN used to treat depression, Corlanor (ivabradine) used to reduce the spontaneous pacemaker activity of the cardiac sinus node. We listed the potential contaminations/biases in this and relevant protein binding-associated approaches.

Part D. Epitope study for vaccines We categorized vaccine studies by their approaches and discussed the background and limitations concerning evolution:

Approach 1. Homology-based with SARS-COV (the 2003 version of SARS), other coronavirus or Ebola.

Approach 2. Immunoinformatics including docking/molecular dynamics/protein structures/antigenicity predictions.

We hand-curated a list of 147 epitopes from these publications and their supplementary materials, grouped them by the source virus proteins, human T-cell/B-cell targets and MHC class. We merged all published epitopes into 124 consolidated groups by partial sub-sequence search and 91 unique virus protein sequence regions by BFS search algorithms. We hope the above lists will serve as the 'wisdom-of-the-crowd' reference for vaccine development.

Summary points and future recommended research topics for Phase 2.

Conclusion 1. There is not a single drug for which consistent positive response has been reported.

Conclusion 2. There are overlaps between the drugs in clinical trials, proposed by computational analysis and proposed by in vitro experiments. However, some of the overlaps, especially those with computational analysis may come from circularity in the methods.

Conclusion 3. Drug candidates proposed by computation and in vitro screening could be biased towards cancer-related targeted therapy and substantially contaminated by existing literature or sometimes anecdotes. This bias/contamination may affect a significant number of computation-based drug-repurposing studies including our work, and certainly not limited to COVID-19.

Future direction 1. Disagreement in the reported drug response can root from differences in dosage, baseline biometrics, and population groups. With more clinical trial results coming in, the next step is to carry out a meta-analysis to stratify these variables.

Future direction 2. Analyzing vaccine findings at this stage is premature as there is no clinical effectiveness study yet. It will be meaningful to make genome variation and vaccines (or maybe antibodies as well) into the same topic, therefore allowing connecting the genome variations to what fraction of the virus strains that a vaccine could cover.

Future direction 3. We suggest a topic on news (e.g., google news) retrieval for therapeutic development, as many (if not most) treatment responses may not first appear in manuscripts.

Finally, we would like to take this opportunity to make one comment: Literature tends to be biassed towards reporting positive results，known biology (e.g., cancer and immune- drugs), and anecdotes, and we should take the results of this exercise and other documents critically.

# DISCUSSION LIST

# INSTALLATION AND REQUIREMENTS

# ACKNOWLEDGMENTS

Part A Subtyping drugs currently in clinical trial

##### *A.1 Methods:*

We first counted how many times each FDA drug occured in the documents provided by Kaggle:

###### **A.2.1 The number of publications each drug appeared, top ones, >=100 times, are (full list in sorted\_alresult):**

* 103 hydrocortisone
* 106 ritonavir
* 111 prednisolone
* 113 dv
* 118 ciprofloxacin
* 119 cyclosporine
* 127 acyclovir
* 134 azithromycin
* 141 amoxicillin
* 155 doxycycline
* 159 dexamethasone
* 166 triad
* 177 chloramphenicol
* 177 kanamycin
* 238 isoflurane
* 248 gentamicin
* 370 bal
* 383 adenosine
* 436 insulin
* 480 ribavirin
* 1767 penicillin

###### **A.2.2 the drugs that have been related to coronavirus in literature, and the top ones, >10 times, are (full list in sorted\_alresult.coronavirus):**

* 10 times: amoxicillin
* 10 times: fluorouracil
* 10 times: kanamycin
* 12 times: azithromycin
* 12 times: hydrocortisone
* 13 times: doxycycline
* 13 times: levofloxacin
* 14 dexamethasone
* 14 isoflurane
* 15 dv
* 15 kaletra
* 15 prednisolone
* 15 tamiflu
* 16 cyclosporine
* 16 gentamicin
* 18 tao
* 19 acyclovir
* 24 triad
* 25 insulin
* 35 remdesivir
* 41 adenosine
* 60 ritonavir
* 66 bal
* 86 penicillin
* 150 ribavirin

###### **A.2.3 The drugs specifically related to COVID-19 in literature (sorted\_alresult.covid19)**

* 1 acetaminophen
* 1 acyclovir
* 1 amoxicillin
* 1 antitussive
* 1 azithromycin
* 1 bal
* 1 ceftriaxone
* 1 chloramphenicol
* 1 digoxin
* 1 doxycycline
* 1 fluorouracil
* 1 ganciclovir
* 1 ibuprofen
* 1 iclusig
* 1 insulin
* 1 levofloxacin
* 1 penicillin
* 1 sulfasalazine
* 1 tigecycline
* 2 adenosine
* 2 triad
* 3 darunavir
* 4 tao
* 7 kaletra
* 12 ribavirin
* 17 remdesivir
* 22 ritonavir

Now we analyze the chemical similarities of these drugs.

###### **A.2.4 Literature summary**

After hand-removing the irrelevant ones, the drugs can be roughly categorized by their effective mechanisms into:

|  |  |
| --- | --- |
| Group and Mechanism | Popular Drugs in Trials |
| RNA mutagens that stop the copying of the virus | Remdesivir, Favipiravir, Fluorouracil, Ribavirin, Acyclovir |
| Protease inhibitors that block the multiplication of the virus | Ritonavir, Lopinavir, Kaletra, Darunavir |
| Stopping the entry of the virus into the host cell | Arbidol, Hydroxychloroquine, Chloroquine phosphate |
| Stopping the release of the virus from the host cell | Oseltamivir |
| Monoclonal antibodies targeting a virus protein/epitope | IL-6 monoclonal antibody, Spike (S) protein antibody |

**A.2.4.1 RNA mutagens**

Viruses need to copy themselves in order to invade the host and transmit (like cancer cells), thus it makes sense that mutagens that block the copying can be used as drugs.

**Remdesivir**: It was studied in many publications related to coronavirus. It was suggested to be highly effective in the control of 2019-nCoV infection in vitro, while their cytotoxicity remains in control (0562f70516579d557cd1486000bb7aac5ccec2a1.json, 95cc4248c19a3cc9a54ebcfa09fc7c80518dac5d.json). It was also reported to significantly reduce lung viral load in mice and with successful clinical cases (0562f70516579d557cd1486000bb7aac5ccec2a1.json, 49ac69f362c27acbc6de0c5cbb640267e7a1e797.json). In clinical settings, it has been used as compassionate treatment. Other papers, e.g., 3e9ae5329eecab16d7c39f1f6dc778cf4a53ee0d.json, suggest the effect is still to be verified.

**Favipiravir**: It was suggested to be a good candidate (58be092086c74c58e9067121a6ba4836468e7ec3.json). It has been used in trials to treat SARS-CoV-2 infections, while the scores of favipiravir docking with the targets in some virtual screenings are relatively low (based on a computation study 95cc4248c19a3cc9a54ebcfa09fc7c80518dac5d.json)

**Fluorouracil**: The RNA mutagen 5-fluorouracil (5-FU) treatment will also increase the U:C and A:G transitions.

**Ribavirin**: It was suggested to be useful for MERS (e5f19b6daf956e815c779228cc0cad1293d65bbb.json). It has been reported to reduce death rate in COVID-19 patients: f294f0df7468a8ac9e27776cc15fa20297a9f040.json.

**Acyclovir**: No statistical difference in treatment effect (baabfb35a321ea12028160e0d2c1552a2fda2dd5.json)

**A.2.4.2 Protease inhibitors**

**Ritonavir**: It was suggested to inhibit proteases and thus block multiplication of the virus. It was reported to deliver a substantial clinical benefit for COVID-19 patients (0562f70516579d557cd1486000bb7aac5ccec2a1.json, and its effectiveness is suggested by *computational* docking studies (9e94f9379fd74fcacc4f3a57e03cbe9035efee8e.json), while others clinical studies showed no effect at all or 'failed' treatment (24e17488d399c436305c819953beae2961214771.json, 8349823092836fe397a59e38615d1491423dbe70.json,8349823092836fe397a59e38615d1491423dbe70.json, ). Previously, it was shown to be beneficial for treating SARS and MERS (3afd5fba7dc182ddfa769c0d766134b525581005.json ).

**Lopinavir**: Lopinavir is a protease inhibitor. It was reported with substantial benefit for treating COVID-10 patients (0562f70516579d557cd1486000bb7aac5ccec2a1.json). Most studies consider Lopinavir as a potential candidate.

**Kaletra**: It is the combination of Ritonavir and Lopinavar.

**Darunavir**: The drug was suggested to be potentially beneficial by *computational* docking experiments (9e94f9379fd74fcacc4f3a57e03cbe9035efee8e.json), and in vitro studies (95cc4248c19a3cc9a54ebcfa09fc7c80518dac5d.json).

**A.2.4.3 By stopping the entry of the virus into the host cell**

**Arbidol**: It inhibits membrane fusion between virus particles and plasma membranes, but it shows no statistical difference in treating COVID-19 patients (baabfb35a321ea12028160e0d2c1552a2fda2dd5.json)

**Hydroxychloroquine, Chloroquine phosphate**: Some studies also suggest that hydroxychloroquine is working by blocking the entry of the virus, though the exact mechanism is unknown (<https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7102587/>). Chloroquine effectively inhibited SARS-CoV-2 in vitro (58be092086c74c58e9067121a6ba4836468e7ec3.json). Chloroquine phosphate was reported to have apparent efficacy and acceptable safety against COVID-19 in a multicenter clinical trials (462cbb326ccd8587cae7a3538c8c6712d9013698.json, b70d27459fd8143edf76721da40cdbca399c9fb1.json).Chloroquine has been recently written into official recommendation for empirical therapy of COVID-19 for its adequate safety data in human (0562f70516579d557cd1486000bb7aac5ccec2a1.json)

**A.2.4.4 By stopping the release of the virus from the host cell**

**Oseltamivir**: Tamiflu, inhibitors of the neuraminidase enzyme, no statistical difference in treating COVID-19 (baabfb35a321ea12028160e0d2c1552a2fda2dd5.json)

The other drugs in the list are irrelevant in this context of effectiveness. Some are related to test of toxicity

**A.2.4.5 By generating monoclonal antibodies targeting certain proteins of the virus**

**IL-6 monoclonal antibody**: the IL-6 monoclonal antibody-directed COVID-19 therapy has been used in clinical trial in China (No.ChiCTR2000029765) (7852aafdfb9e59e6af78a47af796325434f8922a.json, c8d206a4f9af0709b6e9ee90c4d854d482cb0784.json), and IL-6 level was suggested to serve as an indicator of poor prognosis, and was suggest to be used for these patients (c8437a45bfb84fb206fe03fd18d28858bae32651.json).

**Spike (S) protein antibody**: It was suggested that monoclonal antibody against the S protein may 231 efficiently block the virus from entering the host (c8437a45bfb84fb206fe03fd18d28858bae32651.json).

Note: some other drugs, though used to treat COVID-19, are not relevant to the discussion. For example, broad-spectrum antibiotics or fever reducers are often used in the control arm.

##### *A.3. Limitations*

The above analysis has the following limitations:

1. We used a rather earlier version of the literature set (because the searching step took quite a long time), and some popular drugs, e.g. hydroxychloroquine are only discussed but without clear clinical conclusion yet.
2. Literature could be substantially biased towards positive results and by computational methods (discussed below).

## Part B Subtyping computational approaches that are used to propose drug candidates

We then subtyped computational methods developed to repurposing drugs for COVID-19.

##### *B.1 Methods*

During reading the literature curated in Part A, we came across computational studies that focus on predicting drugs suitable for repurposing for COVID-19. These works tend to propose many drugs.

##### *B.2 Results*

##### *B.2.1 Gene-gene network-based approaches*

**Example**: <https://www.nature.com/articles/s41421-020-0153-3> repurposed drugs by network approaches based on homology analysis to other viruses. The authors proposed 16 potential drugs: Irbesartan, Torernifene, Camphor, Equilin, Mesalazine, Mercaptopurine, Paroxetine, Sirolimus, Carvedilol, Colchicine, Dactinomycin, Melatonin, Quinacrine, Eplerenone, Emodin, Oxymetholone.

**Background**: Network-based drug response has been intensively used in the cancer area and was shown to excel in several benchmarks.

##### *B.2.2 Expression-based approaches*

**Example**: <https://arxiv.org/abs/2003.14333> repurposed drugs for treating lung injury in COVID-19 by 'could best reverse abnormal gene expression caused by (SARS)-CoV-2-induced inhibition of ACE2 in lung cells,' an effective drug treatment is one that reverts the aberrant gene expression back to the normal levels'. The authors proposed the following drugs': geldanamycin, panobinostat, trichostatin A, narciclasine, COL-3 and CGP-60474.

##### *B.2.3 Docking or structural-based approaches*

**B.2.3.1 Small molecule prediction**

**Example 1**: <https://www.biorxiv.org/content/10.1101/2020.03.03.972133v1.full> 'a novel advanced deep Q-learning network with the fragment-based drug design (ADQN-FBDD) for generating potential lead compounds targeting SARS-CoV-2 3CLpro' Prioritized 48 candidates by docking (supplement Table S1).

**Example 2**: <https://www.sciencedirect.com/science/article/pii/S2211383520302999> studied the proteins encoded by SARS-CoV-2 genes, compared them with proteins from other coronaviruses, predicted their structures, and built 19 structures that could be done by homology modeling, Library of ZINC drug database, natural products, 78 anti-viral drugs were screened against these targets plus human ACE2. Prioritized the hundreds of drugs, ranked by docking scores: e.g., Ribavirin, alganciclovir, β-Thymidine, Platycodin D, Chrysin,Neohesperidin, Lymecycline, Chlorhexidine, Alfuzosin, Betulonal, Valganciclovir, Chlorhexidine, Betulonal, Gnidicin.

**B.2.3.2 Monoclonal antibody prediction**

**Example 1: docking-based proposal of antibodies** <https://www.biorxiv.org/content/10.1101/2020.02.22.951178v1.full.pdf> The neutralizing antibodies are proposed by computationally docking to the S protein of COVID-19 by docking simulation.

**Example 2: ACE2 pathway-based proposal of antibodies** <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7079879/> Potential therapeutic approaches include a SARS-CoV-2 spike protein-based vaccine; a transmembrane protease serine 2 (TMPRSS2) inhibitor to block the priming of the spike protein; blocking the surface ACE2 receptor by using anti-ACE2 antibody or peptides; and a soluble form of ACE2 which should slow viral entry into cells through competitively binding with SARS-CoV-2 and hence decrease viral spread as well as protecting the lung from injury through its unique enzymatic function. MasR-mitochondrial assembly receptor, AT1R-Ang II type 1 receptor.

**Background**: Docking has been used intensively in drug discovery in areas such as cancers.

##### *B.3 Limitations*

* Computationally proposed drugs tend to be a lot in a single piece of article, sometimes, hundreds of drugs in a single study.
* Most of the works adopted methods from other pharmacogenomics field that were previously developed for cancers.
* We are not aware these approaches have generated hypotheses that are used in real-world clinical trials even in popular fields, e.g. cancer, Alzheimer's. Thus, use them with cautions.

## Part C. Drugs proposed by in vitro experiments

#### C.1 Methods

###### **C.1.1 Data curation**

Other than the drugs used in clinical trials and computational methods, we found an interesting study that carried out genome-wide in vitro binding screening of the virus proteins and human proteins, and proposed 37 drugs that directly target these proteins in the supplementary table 6 of Gordon et al (<https://www.biorxiv.org/content/10.1101/2020.03.22.002386v1.supplementary-material?versioned=true>). These drugs are currently being screened by the authors: Loratadine, Daunorubicin, Midostaurin, Ponatinib, Silmitasertib, Valproic Acid, Haloperidol, Metformin, Migalastat, S-verapamil, Indomethacin, Ruxolitinib, Mycophenolic acid, Entacapone, Ribavirin, E-52862, Merimepodib, RVX-208, XL413, AC-55541, Apicidin, AZ3451, AZ8838, Bafilomycin A1, CCT365623, GB110, H-89, JQ1, PB28, PD-144418, RS-PPCC, TMCB, UCPH-101, ZINC1775962367, ZINC4326719, ZINC4511851, ZINC95559591.

###### **C.1.2 Construction of training set**

We carried out a machine learning exercise, with the hypothesis that the drugs that will be potentially effective should overlap globally in function of these drug targets. We could extract the chemical structure of 34 of the 37 drugs proposed by the authors, which are used as positive examples. The second positive set is the combination of the first positive set and four other drugs that are currently under clinical trial and whose chemical structure can be extracted: remdesivir, hydroxychloroquine, favipiravir and Vitamin C, and thus 38 in total.

The negative training set, which is also the candidate set, is constructed using the FDA approved list, which was downloaded in Oct 2019 from <https://www.accessdata.fda.gov/scripts/cder/daf/index.cfm>. This list has a total of 7305 drugs, 5596 of which we could obtain the fingerprinting structure.

###### **C.1.3 Nested CV to prioritize drug candidates**

For each round, we randomly selected 80% of the example as training, and 20% as testing. The prediction scores for the test set are recorded in each round. We repeated this process for 20 times ensuring all examples occurred in the test set (100 experiments in total). Then the average of each example was taken as the final prediction score.

#### C.2 Results

###### **C.2.1 Top candidates in FDA approved drugs**

Among the FDA approved drugs, we identified some top candidates that do not exist in the training gold standard. We hand-searched in literature for each of the top candidates with a probability >0.05 (55 in total). Most of them come from contaminations, i.e., overlapping with an example in the training set even though the drug appears with a different name.

Cleaned-up list:

|  |  |  |
| --- | --- | --- |
| Drug name | Original usage | Potential issues in the candidate |
| OLUMIANT(Baricitinib) | Janus kinase (JAK) inhibitor |  |
| MEKTOVI | Targeted therapy to treat BRAF V600E or V600K cancers | May come from bias in cancer targeted therapy/screening |
| BRIMONIDINE | Treating glaucoma |  |
| CAPRELSA | kinase inhibitor, medullary thyroid cancer (MTC) | May come from bias in cancer targeted therapy/screening |
| EDURANT(rilpivirine) | Treating Human Immunodeficiency Virus-1 (HIV-1) |  |
| MARPLAN | Treating depression | Some schizophrenia drugs are used in the protein interaction training set, and might result in an implicit contamination here |
| Corlanor (ivabradine) | reduces the spontaneous pacemaker activity of the cardiac sinus node |  |
| LORBRENA | kinase inhibitor, ALK mutant cancer | May come from bias in cancer targeted therapy/screening |
| BRAFTOVI | kinase inhibitor, Metastatic Melanoma | May come from bias in cancer targeted therapy/screening |
| TAVALISSE | kinase inhibitor indicated for the treatment of thrombocytopenia | May come from bias in cancer targeted therapy/screening |

#### C.3 Limitations and biases in the finding

Drugs proposed by in vitro or computational protein targets/gene-gene network approaches are definitely biased towards targeted therapies in cancers, because these drugs were intensively screened in cell line experiments. This is true for both the above list and probably the original list proposed through the binding experiments, and certainly other studies.

Second, low scores only mean the drugs are not similar to others that are being investigated in the study, rather than they are not useful. Remdesivir had a high score of 0.09 (we are not sure if this is an implicit contamination from the training set), the others had low scores, including Vitamin C, hydroxychloroquine and favipiravir.

Part D. Epitope study for vaccines

#### D.1 Methods

We identied all paragraphs that contain the word vaccine and COVID-19/SARS-COV-2. Then, we looked through each of the abstract. If deemed relevant, we go to the original paper and record down their methods and proposed epitopes

#### D.2 Results

##### *D.2.1 Subtyping major approaches in vaccine research*

###### **D.2.1.1 Homology-based approach**

**Example 1.** 181b7b57851e6f58a601b68e613d10c10616f774.json. used conserved sequence with SARS-COV (2003 version of SARS), which already have experimentally validated antigenic sequences.

**Example 2.** a2a6e262098539eb875a26800d9f6d3d0d5d1875.json. tested an epitope of Ebola in mouse, and suggested that this epitope is conserved in COVID-19.

**Example 3** 74b00f19c3af87d1081644f02490ba250f57b7ca.json used conserved sequences between COVID-19 and human Coronavirus (HCov-HKU1) to identify epitope.

###### **D.2.1.1 Immunoinformatics**

Docking/molecular dynamics/protein structures and immunoinformatics such as antigenticity

**Example 1** 73c8af41cfdbf52c0dfba37727e3b94cb56b495e.json used antigenicity Prediction, Docking simulation structural prediction

**Example 2** b38ed62b303eaa444d188deb2ab0b23bbdb79211.json used structure prediction.

Note: Many studies use a combination of the above approaches.

##### *D.2.2 Compiled list of epitopes across the above publications*

|  |  |  |  |
| --- | --- | --- | --- |
| Epitope | Protein | T/B cell | MHC class |
| ILLNKHID | N | T cell | I |
| AFFGMSRIGMEVTPSGTW | N | T cell | NA |
| MEVTPSGTWL | N | T cell | I |
| GMSRIGMEV | N | T cell | I |
| ILLNKHIDA | N | T cell | I |
| ALNTPKDHI | N | T cell | I |
| IRQGTDYKHWPQIAQFA | N | T cell | NA |
| KHWPQIAQFAPSASAFF | N | T cell | NA |
| LALLLLDRL | N | T cell | I |
| LLLDRLNQL | N | T cell | I |
| LLNKHIDAYKTFPPTEPK | N | T cell | NA |
| LQLPQGTTL | N | T cell | I |
| AQFAPSASAFFGMSR | N | T cell | II |
| AQFAPSASAFFGMSRIGM | N | T cell | NA |
| RRPQGLPNNTASWFT | N | T cell | I |
| YKTFPPTEPKKDKKKK | N | T cell | NA |
| GAALQIPFAMQMAYRF | S | T cell | II |
| MAYRFNGIGVTQNVLY | S | T cell | II |
| QLIRAAEIRASANLAATK | S | T cell | II |
| FIAGLIAIV | S | T cell | I |
| ALNTLVKQL | S | T cell | I |
| LITGRLQSL | S | T cell | I |
| NLNESLIDL | S | T cell | I |
| QALNTLVKQLSSNFGAI | S | T cell | II |
| RLNEVAKNL | S | T cell | I |
| VLNDILSRL | S | T cell | I |
| VVFLHVTYV | S | T cell | I |
| DVVNQNAQALNTLVKQL | S | B cell |  |
| EAEVQIDRLITGRLQSL | S | B cell |  |
| EIDRLNEVAKNLNESLIDLQELGKYEQY | S | B cell |  |
| EVAKNLNESLIDLQELG | S | B cell |  |
| GAALQIPFAMQMAYRFN | S | B cell |  |
| GAGICASY | S | B cell |  |
| AISSVLNDILSRLDKVE | S | B cell |  |
| GSFCTQLN | S | B cell |  |
| ILSRLDKVEAEVQIDRL | S | B cell |  |
| KGIYQTSN | S | B cell |  |
| AMQMAYRF | S | B cell |  |
| KNHTSPDVDLGDISGIN | S | B cell |  |
| MAYRFNGIGVTQNVLYE | S | B cell |  |
| AATKMSECVLGQSKRVD | S | B cell |  |
| PFAMQMAYRFNGIGVTQ | S | B cell |  |
| QALNTLVKQLSSNFGAI | S | B cell |  |
| QLIRAAEIRASANLAAT | S | B cell |  |
| QQFGRD | S | B cell |  |
| RASANLAATKMSECVLG | S | B cell |  |
| RLITGRLQSLQTYVTQQ | S | B cell |  |
| EIDRLNEVAKNLNESLIDLQELGKYEQY | S | B cell |  |
| SLQTYVTQQLIRAAEIR | S | B cell |  |
| DLGDISGINASVVNIQK | S | B cell |  |
| FFGMSRIGMEVTPSGTW | N | B cell |  |
| GLPNNTASWFTALTQHGK | N | B cell |  |
| GTTLPK | N | B cell |  |
| IRQGTDYKHWPQIAQFA | N | B cell |  |
| KHIDAYKTFPPTEPKKDKKK | N | B cell |  |
| KHWPQIAQFAPSASAFF | N | B cell |  |
| YNVTQAFGRRGPEQTQGNF | N | B cell |  |
| KTFPPTEPKKDKKKK | N | B cell |  |
| LLPAAD | N | B cell |  |
| LNKHIDAYKTFPPTEPK | N | B cell |  |
| LPQGTTLPKG | N | B cell |  |
| LPQRQKKQ | N | B cell |  |
| PKGFYAEGSRGGSQASSR | N | B cell |  |
| QFAPSASAFFGMSRIGM | N | B cell |  |
| QGTDYKHW | N | B cell |  |
| QLPQGTTLPKGFYAE | N | B cell |  |
| QLPQGTTLPKGFYAEGSR | N | B cell |  |
| QLPQGTTLPKGFYAEGSRGGSQ | N | B cell |  |
| TFPPTEPK | N | B cell |  |
| RRPQGLPNNTASWFT | N | B cell |  |
| SQASSRSS | N | B cell |  |
| SRGGSQASSRSSSRSR | N | B cell |  |
| AGLPYGANK | N | T cell |  |
| AADLDDFSK | N | T cell |  |
| QLESKMSGK | N | T cell |  |
| QELIRQGTDYKH | N | T cell |  |
| LIRQGTDYKHWP | N | T cell |  |
| RLNQLESKMSGK | N | T cell |  |
| LNQLESKMSGKG | N | T cell |  |
| LDRLNQLESKMS | N | T cell |  |
| SVLNDILSR | S | T cell |  |
| GVLTESNKK | S | T cell |  |
| RLFRKSNLK | S | T cell |  |
| QIAPGQTGK | S | T cell |  |
| TSNFRVQPTESI | S | T cell |  |
| SNFRVQPTESIV | S | T cell |  |
| LLIVNNATNVVI | S | T cell |  |
| MSDNGPQNQRNAPRITFGGPSDSTGSNQNGERSGARSKQRRPQGLPNNTAS | N | B cell |  |
| RIRGGDGKMKDL | N | B cell |  |
| TGPEAGLPYGANK | N | B cell |  |
| GTTLPKGFYAEGSRGGSQASSRSSSRSRNSSRNSTPGSSRGTSPARMAGNGGD | N | B cell |  |
| SKMSGKGQQQQGQTVTKKSAAEASKKPRQKRTATKAYN | N | B cell |  |
| KTFPPTEPKKDKKKKADETQALPQRQKKQQ | N | B cell |  |
| LTPGDSSSGWTAG | S | B cell |  |
| VRQIAPGQTGKIAD | S | B cell |  |
| YQAGSTPCNGV | S | B cell |  |
| QTQTNSPRRARSV | S | B cell |  |
| VYQVNNLEEIC |  |  |  |
| SMATYYLFDESGEFK | orf1ab |  |  |
| MATYYLFDESGEFKL | orf1ab |  |  |
| ATYYLFDESGEFKLA | orf1ab |  |  |
| DSATLVSDIDITFLK | orf1ab |  |  |
| SNPTTFHLDGEVITF | orf1ab |  |  |
| NPTTFHLDGEVITFD | orf1ab |  |  |
| PTTFHLDGEVITFDN | orf1ab |  |  |
| DGEVITFDNLKTLLS | orf1ab |  |  |
| EVRTIKVFTTVDNIN | orf1ab |  |  |
| VRTIKVFTTVDNINL | orf1ab |  |  |
| RTIKVFTTVDNINLH | orf1ab |  |  |
| HEGKTFYVLPNDDTL | orf1ab |  |  |
| EGKTFYVLPNDDTLR | orf1ab |  |  |
| GKTFYVLPNDDTLRV | orf1ab |  |  |
| KTFYVLPNDDTLRVE | orf1ab |  |  |
| DLMAAYVDNSSLTIK | orf1ab |  |  |
| LMAAYVDNSSLTIKK | orf1ab |  |  |
| MAAYVDNSSLTIKKP | orf1ab |  |  |
| AAYVDNSSLTIKKPN | orf1ab |  |  |
| YREGYLNSTNVTIAT | orf1ab |  |  |
| REGYLNSTNVTIATY | orf1ab |  |  |
| IINLVQMAPISAMVR | orf1ab |  |  |
| VAAIFYLITPVHVMS | orf1ab |  |  |
| AAIFYLITPVHVMSK | orf1ab |  |  |
| PDTRYVLMDGSIIQF | orf1ab |  |  |
| DTRYVLMDGSIIQFP | orf1ab |  |  |
| TRYVLMDGSIIQFPN | orf1ab |  |  |
| RLTKYTMADLVYALR | orf1ab |  |  |
| TMADLVYALRHFDEG | orf1ab |  |  |
| TKRNVIPTITQMNLK | orf1ab |  |  |
| YEAMYTPHTVLQAVG | orf1ab |  |  |
| YDHVISTSHKLVLSV | orf1ab |  |  |
| SQSIIAYTMSLGAEN | S |  |  |
| SNNSIAIPTNFTISV | S |  |  |
| AIPTNFTISVTTEIL | S |  |  |
| IPTNFTISVTTEILP | S |  |  |
| PTNFTISVTTEILPV | S |  |  |
| TNFTISVTTEILPVS | S |  |  |
| VKPSFYVYSRVKNLN | E |  |  |
| KPSFYVYSRVKNLNS | E |  |  |
| PSFYVYSRVKNLNSS | E |  |  |
| ATKAYNVTQAFGRRG | N |  |  |
| KAYNVTQAFGRRGPE | N |  |  |
| YTGAIKLDDKDPNFK | N |  |  |

[Go Top](https://www.kaggle.com/makrovh/subtyping-covid-19-therapeutic-research-findings#top)

##### *D.2.3 Compiling list of consolidated epitope groups by partial subsequence overlap*

Now we consolidate the epitopes from various publications by partial sub-sequence overlap. The rationale is to identify the cases where one epitope is a subsequence of another, or if one overlaps to the other by more than 5 consequtive amino acides, and only the outside flanking ragion is not overlapping. In these cases, two epitodes would be considered in the same group.

These are the consolidated epitope groups that have so far been published:

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Group | Epitope 1 | Epitope 2 | Epitope 3 | Epitope 4 | Epitope 5 | Epitope 6 | Epitope 7 | Epitope 8 | Epitope 9 |
| 1 | AADLDDFSK |  |  |  |  |  |  |  |  |
| 2 | AAIFYLITPVHVMSK |  |  |  |  |  |  |  |  |
| 3 | AATKMSECVLGQSKRVD |  |  |  |  |  |  |  |  |
| 4 | AAYVDNSSLTIKKPN |  |  |  |  |  |  |  |  |
| 5 | AGLPYGANK | TGPEAGLPYGANK |  |  |  |  |  |  |  |
| 6 | AIPTNFTISVTTEIL |  |  |  |  |  |  |  |  |
| 7 | ALNTLVKQL | DVVNQNAQALNTLVKQL |  |  |  |  |  |  |  |
| 8 | ALNTLVKQL | QALNTLVKQLSSNFGAI | DVVNQNAQALNTLVKQL |  |  |  |  |  |  |
| 9 | ALNTLVKQL | QALNTLVKQLSSNFGAI |  |  |  |  |  |  |  |
| 10 | ALNTPKDHI |  |  |  |  |  |  |  |  |
| 11 | AMQMAYRF | GAALQIPFAMQMAYRF | GAALQIPFAMQMAYRFN | PFAMQMAYRFNGIGVTQ |  |  |  |  |  |
| 12 | AMQMAYRF | GAALQIPFAMQMAYRF | GAALQIPFAMQMAYRFN |  |  |  |  |  |  |
| 13 | AMQMAYRF | MAYRFNGIGVTQNVLY | PFAMQMAYRFNGIGVTQ |  |  |  |  |  |  |
| 14 | AQFAPSASAFFGMSR | KHWPQIAQFAPSASAFF | QFAPSASAFFGMSRIGM | AQFAPSASAFFGMSRIGM |  |  |  |  |  |
| 15 | AQFAPSASAFFGMSR | KHWPQIAQFAPSASAFF |  |  |  |  |  |  |  |
| 16 | ATKAYNVTQAFGRRG | KAYNVTQAFGRRGPE | YNVTQAFGRRGPEQTQGNF |  |  |  |  |  |  |
| 17 | ATKAYNVTQAFGRRG | YNVTQAFGRRGPEQTQGNF |  |  |  |  |  |  |  |
| 18 | ATYYLFDESGEFKLA |  |  |  |  |  |  |  |  |
| 19 | DGEVITFDNLKTLLS |  |  |  |  |  |  |  |  |
| 20 | DLGDISGINASVVNIQK |  |  |  |  |  |  |  |  |
| 21 | DLMAAYVDNSSLTIK |  |  |  |  |  |  |  |  |
| 22 | DSATLVSDIDITFLK |  |  |  |  |  |  |  |  |
| 23 | DTRYVLMDGSIIQFP |  |  |  |  |  |  |  |  |
| 24 | EGKTFYVLPNDDTLR |  |  |  |  |  |  |  |  |
| 25 | EVRTIKVFTTVDNIN |  |  |  |  |  |  |  |  |
| 26 | FIAGLIAIV |  |  |  |  |  |  |  |  |
| 27 | GAGICASY |  |  |  |  |  |  |  |  |
| 28 | GKTFYVLPNDDTLRV |  |  |  |  |  |  |  |  |
| 29 | GMSRIGMEV | AQFAPSASAFFGMSR | QFAPSASAFFGMSRIGM | AQFAPSASAFFGMSRIGM |  |  |  |  |  |
| 30 | GMSRIGMEV | FFGMSRIGMEVTPSGTW | QFAPSASAFFGMSRIGM | AFFGMSRIGMEVTPSGTW | AQFAPSASAFFGMSRIGM |  |  |  |  |
| 31 | GMSRIGMEV | MEVTPSGTWL | FFGMSRIGMEVTPSGTW | AFFGMSRIGMEVTPSGTW |  |  |  |  |  |
| 32 | GSFCTQLN |  |  |  |  |  |  |  |  |
| 33 | GTTLPK | LPQGTTLPKG | QLPQGTTLPKGFYAE | QLPQGTTLPKGFYAEGSR | QLPQGTTLPKGFYAEGSRGGSQ | GTTLPKGFYAEGSRGGSQASSRSSSRSRNSSRNSTPGSSRGTSPARMAGNGGD |  |  |  |
| 34 | GTTLPK | LQLPQGTTL | LPQGTTLPKG | QLPQGTTLPKGFYAE | PKGFYAEGSRGGSQASSR | QLPQGTTLPKGFYAEGSR | QLPQGTTLPKGFYAEGSRGGSQ | GTTLPKGFYAEGSRGGSQASSRSSSRSRNSSRNSTPGSSRGTSPARMAGNGGD |  |
| 35 | GTTLPK | LQLPQGTTL | LPQGTTLPKG | QLPQGTTLPKGFYAE | QLPQGTTLPKGFYAEGSR | QLPQGTTLPKGFYAEGSRGGSQ | GTTLPKGFYAEGSRGGSQASSRSSSRSRNSSRNSTPGSSRGTSPARMAGNGGD |  |  |
| 36 | GTTLPK | SQASSRSS | LPQGTTLPKG | QLPQGTTLPKGFYAE | SRGGSQASSRSSSRSR | PKGFYAEGSRGGSQASSR | QLPQGTTLPKGFYAEGSR | QLPQGTTLPKGFYAEGSRGGSQ | GTTLPKGFYAEGSRGGSQASSRSSSRSRNSSRNSTPGSSRGTSPARMAGNGGD |
| 37 | GVLTESNKK |  |  |  |  |  |  |  |  |
| 38 | HEGKTFYVLPNDDTL |  |  |  |  |  |  |  |  |
| 39 | IINLVQMAPISAMVR |  |  |  |  |  |  |  |  |
| 40 | ILLNKHID | ILLNKHIDA | LNKHIDAYKTFPPTEPK | LLNKHIDAYKTFPPTEPK |  |  |  |  |  |
| 41 | ILLNKHID | TFPPTEPK | ILLNKHIDA | LNKHIDAYKTFPPTEPK | LLNKHIDAYKTFPPTEPK | KHIDAYKTFPPTEPKKDKKK |  |  |  |
| 42 | ILSRLDKVEAEVQIDRL |  |  |  |  |  |  |  |  |
| 43 | IPTNFTISVTTEILP |  |  |  |  |  |  |  |  |
| 44 | KAYNVTQAFGRRGPE | YNVTQAFGRRGPEQTQGNF |  |  |  |  |  |  |  |
| 45 | KGIYQTSN |  |  |  |  |  |  |  |  |
| 46 | KNHTSPDVDLGDISGIN |  |  |  |  |  |  |  |  |
| 47 | KPSFYVYSRVKNLNS |  |  |  |  |  |  |  |  |
| 48 | KTFYVLPNDDTLRVE |  |  |  |  |  |  |  |  |
| 49 | LALLLLDRL |  |  |  |  |  |  |  |  |
| 50 | LITGRLQSL | EAEVQIDRLITGRLQSL | RLITGRLQSLQTYVTQQ |  |  |  |  |  |  |
| 51 | LITGRLQSL | EAEVQIDRLITGRLQSL |  |  |  |  |  |  |  |
| 52 | LITGRLQSL | RLITGRLQSLQTYVTQQ |  |  |  |  |  |  |  |
| 53 | LLIVNNATNVVI |  |  |  |  |  |  |  |  |
| 54 | LLLDRLNQL | LDRLNQLESKMS |  |  |  |  |  |  |  |
| 55 | LLLDRLNQL | QLESKMSGK | LDRLNQLESKMS |  |  |  |  |  |  |
| 56 | LLPAAD |  |  |  |  |  |  |  |  |
| 57 | LMAAYVDNSSLTIKK |  |  |  |  |  |  |  |  |
| 58 | LPQRQKKQ | KTFPPTEPKKDKKKKADETQALPQRQKKQQ |  |  |  |  |  |  |  |
| 59 | LPQRQKKQ | TFPPTEPK | KTFPPTEPKKDKKKK | YKTFPPTEPKKDKKKK | KTFPPTEPKKDKKKKADETQALPQRQKKQQ |  |  |  |  |
| 60 | LQLPQGTTL | LPQGTTLPKG | QLPQGTTLPKGFYAE | QLPQGTTLPKGFYAEGSR | QLPQGTTLPKGFYAEGSRGGSQ |  |  |  |  |
| 61 | LTPGDSSSGWTAG |  |  |  |  |  |  |  |  |
| 62 | MAAYVDNSSLTIKKP |  |  |  |  |  |  |  |  |
| 63 | MATYYLFDESGEFKL |  |  |  |  |  |  |  |  |
| 64 | MAYRFNGIGVTQNVLY | MAYRFNGIGVTQNVLYE | PFAMQMAYRFNGIGVTQ |  |  |  |  |  |  |
| 65 | MAYRFNGIGVTQNVLY | MAYRFNGIGVTQNVLYE |  |  |  |  |  |  |  |
| 66 | MEVTPSGTWL | FFGMSRIGMEVTPSGTW | AFFGMSRIGMEVTPSGTW |  |  |  |  |  |  |
| 67 | NLNESLIDL | EVAKNLNESLIDLQELG | EIDRLNEVAKNLNESLIDLQELGKYEQY |  |  |  |  |  |  |
| 68 | NLNESLIDL | RLNEVAKNL | EVAKNLNESLIDLQELG | EIDRLNEVAKNLNESLIDLQELGKYEQY |  |  |  |  |  |
| 69 | NPTTFHLDGEVITFD |  |  |  |  |  |  |  |  |
| 70 | PDTRYVLMDGSIIQF |  |  |  |  |  |  |  |  |
| 71 | PSFYVYSRVKNLNSS |  |  |  |  |  |  |  |  |
| 72 | PTNFTISVTTEILPV |  |  |  |  |  |  |  |  |
| 73 | PTTFHLDGEVITFDN |  |  |  |  |  |  |  |  |
| 74 | QGTDYKHW | LIRQGTDYKHWP | IRQGTDYKHWPQIAQFA |  |  |  |  |  |  |
| 75 | QGTDYKHW | QELIRQGTDYKH | IRQGTDYKHWPQIAQFA |  |  |  |  |  |  |
| 76 | QGTDYKHW | QELIRQGTDYKH | LIRQGTDYKHWP | IRQGTDYKHWPQIAQFA |  |  |  |  |  |
| 77 | QIAPGQTGK | VRQIAPGQTGKIAD |  |  |  |  |  |  |  |
| 78 | QLESKMSGK | LNQLESKMSGKG |  |  |  |  |  |  |  |
| 79 | QLESKMSGK | RLNQLESKMSGK | LNQLESKMSGKG | LDRLNQLESKMS | SKMSGKGQQQQGQTVTKKSAAEASKKPRQKRTATKAYN |  |  |  |  |
| 80 | QLESKMSGK | RLNQLESKMSGK |  |  |  |  |  |  |  |
| 81 | QLESKMSGK | SKMSGKGQQQQGQTVTKKSAAEASKKPRQKRTATKAYN |  |  |  |  |  |  |  |
| 82 | QLIRAAEIRASANLAAT | QLIRAAEIRASANLAATK |  |  |  |  |  |  |  |
| 83 | QQFGRD |  |  |  |  |  |  |  |  |
| 84 | QTQTNSPRRARSV |  |  |  |  |  |  |  |  |
| 85 | RASANLAATKMSECVLG |  |  |  |  |  |  |  |  |
| 86 | REGYLNSTNVTIATY |  |  |  |  |  |  |  |  |
| 87 | RIRGGDGKMKDL |  |  |  |  |  |  |  |  |
| 88 | RLFRKSNLK |  |  |  |  |  |  |  |  |
| 89 | RLNEVAKNL | EVAKNLNESLIDLQELG | EIDRLNEVAKNLNESLIDLQELGKYEQY |  |  |  |  |  |  |
| 90 | RLTKYTMADLVYALR |  |  |  |  |  |  |  |  |
| 91 | RRPQGLPNNTASWFT | GLPNNTASWFTALTQHGK | MSDNGPQNQRNAPRITFGGPSDSTGSNQNGERSGARSKQRRPQGLPNNTAS |  |  |  |  |  |  |
| 92 | RRPQGLPNNTASWFT | GLPNNTASWFTALTQHGK |  |  |  |  |  |  |  |
| 93 | RRPQGLPNNTASWFT | MSDNGPQNQRNAPRITFGGPSDSTGSNQNGERSGARSKQRRPQGLPNNTAS |  |  |  |  |  |  |  |
| 94 | RTIKVFTTVDNINLH |  |  |  |  |  |  |  |  |
| 95 | SLQTYVTQQLIRAAEIR |  |  |  |  |  |  |  |  |
| 96 | SMATYYLFDESGEFK |  |  |  |  |  |  |  |  |
| 97 | SNFRVQPTESIV |  |  |  |  |  |  |  |  |
| 98 | SNNSIAIPTNFTISV |  |  |  |  |  |  |  |  |
| 99 | SNPTTFHLDGEVITF |  |  |  |  |  |  |  |  |
| 100 | SQASSRSS | PKGFYAEGSRGGSQASSR | QLPQGTTLPKGFYAEGSRGGSQ | GTTLPKGFYAEGSRGGSQASSRSSSRSRNSSRNSTPGSSRGTSPARMAGNGGD |  |  |  |  |  |
| 101 | SQASSRSS | SRGGSQASSRSSSRSR | GTTLPKGFYAEGSRGGSQASSRSSSRSRNSSRNSTPGSSRGTSPARMAGNGGD |  |  |  |  |  |  |
| 102 | SQASSRSS | SRGGSQASSRSSSRSR | PKGFYAEGSRGGSQASSR | GTTLPKGFYAEGSRGGSQASSRSSSRSRNSSRNSTPGSSRGTSPARMAGNGGD |  |  |  |  |  |
| 103 | SQSIIAYTMSLGAEN |  |  |  |  |  |  |  |  |
| 104 | SVLNDILSR | AISSVLNDILSRLDKVE |  |  |  |  |  |  |  |
| 105 | TFPPTEPK | KTFPPTEPKKDKKKK | YKTFPPTEPKKDKKKK | KHIDAYKTFPPTEPKKDKKK | KTFPPTEPKKDKKKKADETQALPQRQKKQQ |  |  |  |  |
| 106 | TFPPTEPK | KTFPPTEPKKDKKKK | YKTFPPTEPKKDKKKK | LNKHIDAYKTFPPTEPK | LLNKHIDAYKTFPPTEPK | KHIDAYKTFPPTEPKKDKKK | KTFPPTEPKKDKKKKADETQALPQRQKKQQ |  |  |
| 107 | TFPPTEPK | KTFPPTEPKKDKKKK | YKTFPPTEPKKDKKKK | LNKHIDAYKTFPPTEPK | LLNKHIDAYKTFPPTEPK | KHIDAYKTFPPTEPKKDKKK |  |  |  |
| 108 | TKRNVIPTITQMNLK |  |  |  |  |  |  |  |  |
| 109 | TMADLVYALRHFDEG |  |  |  |  |  |  |  |  |
| 110 | TNFTISVTTEILPVS |  |  |  |  |  |  |  |  |
| 111 | TRYVLMDGSIIQFPN |  |  |  |  |  |  |  |  |
| 112 | TSNFRVQPTESI |  |  |  |  |  |  |  |  |
| 113 | VAAIFYLITPVHVMS |  |  |  |  |  |  |  |  |
| 114 | VKPSFYVYSRVKNLN |  |  |  |  |  |  |  |  |
| 115 | VLNDILSRL | AISSVLNDILSRLDKVE |  |  |  |  |  |  |  |
| 116 | VLNDILSRL | SVLNDILSR | AISSVLNDILSRLDKVE |  |  |  |  |  |  |
| 117 | VRTIKVFTTVDNINL |  |  |  |  |  |  |  |  |
| 118 | VVFLHVTYV |  |  |  |  |  |  |  |  |
| 119 | VYQVNNLEEIC |  |  |  |  |  |  |  |  |
| 120 | YDHVISTSHKLVLSV |  |  |  |  |  |  |  |  |
| 121 | YEAMYTPHTVLQAVG |  |  |  |  |  |  |  |  |
| 122 | YQAGSTPCNGV |  |  |  |  |  |  |  |  |
| 123 | YREGYLNSTNVTIAT |  |  |  |  |  |  |  |  |
| 124 | YTGAIKLDDKDPNFK |  |  |  |  |  |  |  |  |

Code for compiling list of consolidated epitope groups by partial subsequence overlap.

##### *D.2.4 Compiling the unique protein regions where epitopes have been identified from various publications by BFS*

Now let us find out the unique virus protein regions where epitopes have been identified from various publications by partial sub-sequence overlap. The difference between this section and D.2.3 is the following: When epitode A and B overlap, and B and C overlap, but A and C do not overlap substantially, in the previous section, they are considered as separate groups as we were trying to find out non-overlapping peptides, while in this section, they are considered to be in the same group as they are in the same protein regions.

These are the unique groups of protein regions where epitopes have so far been identified:

Group 1 'ILLNKHID', 'ILLNKHIDA', 'LLNKHIDAYKTFPPTEPK'

Group 2 'AFFGMSRIGMEVTPSGTW', 'MEVTPSGTWL', 'GMSRIGMEV', 'FFGMSRIGMEVTPSGTW'

Group 3 'ALNTPKDHI'

Group 4 'IRQGTDYKHWPQIAQFA', 'QGTDYKHW', 'QELIRQGTDYKH', 'LIRQGTDYKHWP'

Group 5 'KHWPQIAQFAPSASAFF', 'AQFAPSASAFFGMSR', 'AQFAPSASAFFGMSRIGM', 'QFAPSASAFFGMSRIGM'

Group 6 'LALLLLDRL'

Group 7 'LLLDRLNQL'

Group 8 'LQLPQGTTL'

Group 9 'RRPQGLPNNTASWFT', 'GLPNNTASWFTALTQHGK', 'MSDNGPQNQRNAPRITFGGPSDSTGSNQNGERSGARSKQRRPQGLPNNTAS'

Group 10 'YKTFPPTEPKKDKKKK', 'KHIDAYKTFPPTEPKKDKKK', 'KTFPPTEPKKDKKKK', 'LNKHIDAYKTFPPTEPK', 'TFPPTEPK', 'KTFPPTEPKKDKKKKADETQALPQRQKKQQ'

Group 11 'GAALQIPFAMQMAYRF', 'GAALQIPFAMQMAYRFN', 'AMQMAYRF', 'PFAMQMAYRFNGIGVTQ'

Group 12 'MAYRFNGIGVTQNVLY', 'MAYRFNGIGVTQNVLYE'

Group 13 'QLIRAAEIRASANLAATK', 'QLIRAAEIRASANLAAT'

Group 14 'FIAGLIAIV'

Group 15 'ALNTLVKQL', 'QALNTLVKQLSSNFGAI', 'DVVNQNAQALNTLVKQL'

Group 16 'LITGRLQSL', 'EAEVQIDRLITGRLQSL', 'RLITGRLQSLQTYVTQQ'

Group 17 'NLNESLIDL'

Group 18 'RLNEVAKNL', 'EIDRLNEVAKNLNESLIDLQELGKYEQY', 'EVAKNLNESLIDLQELG'

Group 19 'VLNDILSRL', 'AISSVLNDILSRLDKVE', 'SVLNDILSR'

Group 20 'VVFLHVTYV'

Group 21 'GAGICASY'

Group 22 'GSFCTQLN'

Group 23 'ILSRLDKVEAEVQIDRL'

Group 24 'KGIYQTSN'

Group 25 'KNHTSPDVDLGDISGIN'

Group 26 'AATKMSECVLGQSKRVD'

Group 27 'QQFGRD'

Group 28 'RASANLAATKMSECVLG'

Group 29 'SLQTYVTQQLIRAAEIR'

Group 30 'DLGDISGINASVVNIQK'

Group 31 'GTTLPK', 'LPQGTTLPKG', 'QLPQGTTLPKGFYAE', 'QLPQGTTLPKGFYAEGSR', 'QLPQGTTLPKGFYAEGSRGGSQ'

Group 32 'YNVTQAFGRRGPEQTQGNF', 'ATKAYNVTQAFGRRG', 'KAYNVTQAFGRRGPE'

Group 33 'LLPAAD'

Group 34 'LPQRQKKQ'

Group 35 'PKGFYAEGSRGGSQASSR', 'SQASSRSS', 'SRGGSQASSRSSSRSR','GTTLPKGFYAEGSRGGSQASSRSSSRSRNSSRNSTPGSSRGTSPARMAGNGGD'

Group 36 'AGLPYGANK', 'TGPEAGLPYGANK'

Group 37 'AADLDDFSK'

Group 38 'QLESKMSGK', 'RLNQLESKMSGK', 'LNQLESKMSGKG', 'LDRLNQLESKMS', 'SKMSGKGQQQQGQTVTKKSAAEASKKPRQKRTATKAYN'

Group 39 'GVLTESNKK'

Group 40 'RLFRKSNLK'

Group 41 'QIAPGQTGK', 'VRQIAPGQTGKIAD'

Group 42 'TSNFRVQPTESI'

Group 43 'SNFRVQPTESIV'

Group 44 'LLIVNNATNVVI'

Group 45 'RIRGGDGKMKDL'

Group 46 'LTPGDSSSGWTAG'

Group 47 'YQAGSTPCNGV'

Group 48 'QTQTNSPRRARSV'

Group 49 'VYQVNNLEEIC'

Group 50 'SMATYYLFDESGEFK'

Group 51 'MATYYLFDESGEFKL'

Group 52 'ATYYLFDESGEFKLA'

Group 53 'DSATLVSDIDITFLK'

Group 54 'SNPTTFHLDGEVITF'

Group 55 'NPTTFHLDGEVITFD'

Group 56 'PTTFHLDGEVITFDN'

Group 57 'DGEVITFDNLKTLLS'

Group 58 'EVRTIKVFTTVDNIN'

Group 59 'VRTIKVFTTVDNINL'

Group 60 'RTIKVFTTVDNINLH'

Group 61 'HEGKTFYVLPNDDTL'

Group 62 'EGKTFYVLPNDDTLR'

Group 63 'GKTFYVLPNDDTLRV'

Group 64 'KTFYVLPNDDTLRVE'

Group 65 'DLMAAYVDNSSLTIK'

Group 66 'LMAAYVDNSSLTIKK'

Group 67 'MAAYVDNSSLTIKKP'

Group 68 'AAYVDNSSLTIKKPN'

Group 69 'YREGYLNSTNVTIAT'

Group 70 'REGYLNSTNVTIATY'

Group 71 'IINLVQMAPISAMVR'

Group 72 'VAAIFYLITPVHVMS'

Group 73 'AAIFYLITPVHVMSK'

Group 74 'PDTRYVLMDGSIIQF'

Group 75 'DTRYVLMDGSIIQFP'

Group 76 'TRYVLMDGSIIQFPN'

Group 77 'RLTKYTMADLVYALR'

Group 78 'TMADLVYALRHFDEG'

Group 79 'TKRNVIPTITQMNLK'

Group 80 'YEAMYTPHTVLQAVG'

Group 81 'YDHVISTSHKLVLSV'

Group 82 'SQSIIAYTMSLGAEN'

Group 83 'SNNSIAIPTNFTISV'

Group 84 'AIPTNFTISVTTEIL'

Group 85 'IPTNFTISVTTEILP'

Group 86 'PTNFTISVTTEILPV'

Group 87 'TNFTISVTTEILPVS'

Group 88 'VKPSFYVYSRVKNLN'

Group 89 'KPSFYVYSRVKNLNS'

Group 90 'PSFYVYSRVKNLNSS'

Group 91 'YTGAIKLDDKDPNFK'

Code for compiling the unique virus protein regions where epitopes have been identified from various publications by BFS:

#### D.3 Limitations

There are numerous existing software for epitope prediction, including whether the epitope is on the surface, their docking score and MHC classes. However, studies using homology should be taken carefully, as the overall sequence similarity (based on our other studies) between COVID-19 and SARS-COV (the 2003 version ~82), Ebola (~40%) and Human Coronavirus (65-70% depending on the exact strain) is very limited. It should be taken critically that an exactly conserved epitope (~10 amino acids) indicates same effectiveness across these species.