**Artificial intelligence (AI) - powered drug repositioning for treating COVID-19**

**1 Objectives/Hypotheses**

Emerging infectious diseases have been an ever-present threat to public health, and COVID-19 is a recent example. There is a urgent need to develop a robust framework to comeback the disease with safe and effective therapeutic options. Great efforts are being achieved as data continues to be generated for a better understanding of the underlying mechanisms of COVID-19. It provides a unique opportunity to implement computational drug repositioning approaches to not only accelerate the discovery of COVID-19 treatments but also prepare us for future infectious diseases. This project aims to systematically survey and prioritize approved or investigational drugs for their potential use to treat COVID-19 using computational drug repositioning principle with artificial intelligence.

This project with three specific aims:

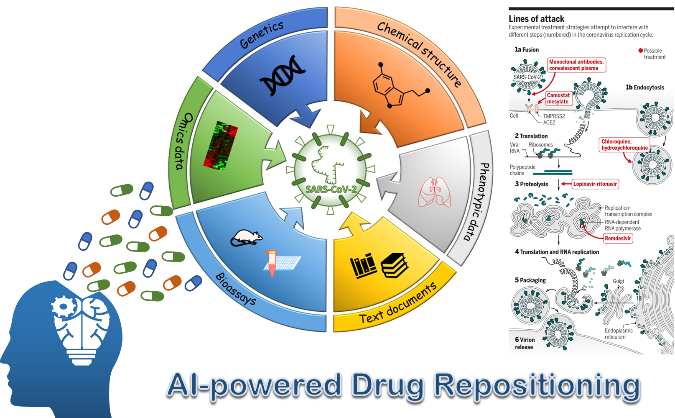
1. To build an AI-powered drug repositioning framework to prioritize approved or investigational drugs for their potential to treat COVID-19.
2. Ultimately, the goal would be to leverage the AI-powered framework for the evaluation of drugs for the treatment of future emerging infections.

**2 Background**

The novel coronavirus SARS-CoV-2, the causative agent of COVID-19 respiratory disease, has infected over 3.6 million people, killed over 250,000, and caused worldwide social and economic disruption [1,2]. There are currently no antiviral drugs with fully proven clinical efficacy, nor are there vaccines for its prevention, and these efforts are hampered by limited knowledge of the molecular details of SARS-CoV-2 infection. FDA only authorized remdesivir to treat hospitalized patients with severe Covid-19 (<https://www.fda.gov/media/137564/download>) based on results of clinical trials globally [3,4]. To accelerate and facilitate the development of effective and safer treatment regimens for COVID-19, we aim to develop an AI-powered *in silico* drug repositioning pipeline to systematically prioritize the approved or investigational drugs for their potential uses for COVID-19 treatment. The candidates will be verified by real-world evidence. The drug candidate list could potentially serve as an initial step to accelerate effective and safer COVID-19 treatment development. Meanwhile, the developed AI-powered drug repositioning framework will facilitate information sharing and communication by different stakeholders to further enhance agency capability to deal with future emerging infections.

**3 Experimental Methods**

In our previous project, we have comprehensively evaluated the computational drug repurposing approaches and proposed critical steps for implementing *in silico* drug repositioning into different domain applications, which includes repurposing with a purpose, a strategy, and with confidence. We will further apply the critical steps to develop the AI-powered drug repositioning framework for COVID-19 to treat, eliminate, and ameliorate the symptoms of COVID-19 (**Figure 1**). The study consists of the following sequential steps:



**Figure 1** Illustration of the workflow of this project

**Step 1** - Development of both COVID-19 signatures tailored to its mechanism and drug signatures associated with mode-of-actions (MoAs): the signatures will be developed based on a broad range of information including the protein target information (e.g., ACE2), pathways, drug transcriptomics data (e.g., LINCS and CMap), chemical structures, clinical notes (e.g., CURE ID) of the different stages of a COVID19-physiological response (mild, moderate, severe, death), etc.

**3.1.1. Mechanistic signatures of COVID-19**

To develop rapid medical responses for COVID-19, tremendous efforts to understand better the pathogenesis of this disease are ongoing and will undoubtedly enlighten the optimal management of the growing pandemic [5]. In this project, we will curate the public available molecular data and clinical information of COVID-19. The curated information and datasets were categorized and listed in **Table 1**.

**Table 1** Mechanistic signatures of COVID-19

|  |  |  |
| --- | --- | --- |
| **Data resources** | **Weblink** | **Notes** |
| ***Disease etiology*** | | |
| Coronavirus replicate cycle | <https://science.sciencemag.org/content/367/6485/1412/tab-pdf> | Efforts of current drug repositioning for COVID-19 |
| ***Structure information*** | | |
| Protease enzyme of COVID-19  (PDB ID: 6LU7) | <https://www.rcsb.org/structure/6LU7> | The crystal structure of COVID-19 main protease in complex with an inhibitor N3 |
| SARS-CoV S-protein in complex with ACE2  (PDB ID: 2AJF) | <https://www.rcsb.org/structure/2AJF> | Structure of SARS coronavirus spike receptor-binding domain complexed with its receptor |
| ***Protein information*** | | |
| 119 Human CoV-associated host proteins | <https://static-content.springer.com/esm/art%3A10.1038%2Fs41421-020-0153-3/MediaObjects/41421_2020_153_MOESM1_ESM.pdf> | [6] |
| 332 high-confidence SARS-CoV-2-human protein-protein interactions | <https://static-content.springer.com/esm/art%3A10.1038%2Fs41586-020-2286-9/MediaObjects/41586_2020_2286_MOESM7_ESM.xlsx> | [2] |
| ***Pathogen genome*** | | |
| Pathogen genome data of SARS-CoV-2 | <https://nextstrain.org/help/coronavirus/SARS-CoV-2> | Nextstrain is an open-source project to harness the scientific and public health potential of pathogen genome data |
| ***Ongoing clinical trials*** | | |
| Clinicaltrial.gov | <https://clinicaltrials.gov/ct2/results?cond=COVID-19> | Ongoing clinical trials for COVID-19 |
| Clinical trials from WHO | <https://clinicaltrials.gov/ct2/who_table> | COVID-19 Studies from the World Health Organization Database |
| Therapeutic Options for COVID-19 Currently Under Investigation | <https://covid19treatmentguidelines.nih.gov/therapeutic-options-under-investigation/> | NIH: Therapeutic Options for COVID-19 Currently Under Investigation. Serve |
| ***COVID-19 ontologies*** | | |
| COVID-19 ontology | <https://github.com/virtual-biohackathons/covid-19-bh20/wiki/Ontology> | Update ontologies with COVID-19 and CoV related concepts to make knowledge machine-readable and actionable for mining, logical reasoning, etc... |
| ***Literatures*** | | |
| PubMed | <https://www.kaggle.com/allen-institute-for-ai/CORD-19-research-challenge> | A collection of literatures related COVID-19 |
| ***Vulnerable population-related COVID-19*** | | |
| Vulnerable populations for COVID-19 | <https://www.cdc.gov/coronavirus/2019-ncov/need-extra-precautions/people-at-higher-risk.html> | Vulnerable population for COVID-19 from CDC |

**3.1.2 Drug-centric information**

The chemical and biological profiling data of approved drugs and investigated drugs will be collected from publicly available data resources. The database resources of approved drugs and existing drug candidates information is listed in **Table 2**.

**Table2** Drug information

|  |  |  |
| --- | --- | --- |
| **Database** | **Weblink** | **Notes** |
| ***Drug-target relationship*** | | |
| **DrugBank** | <https://www.drugbank.ca/> | The latest release of DrugBank (version 5.1.4, released) contains 13,341 drug entries including 2,596 approved small molecule drugs, 1,289 approved biotech (protein/peptide) drugs, 130 nutraceuticals and over 6,305 experimental drugs. Additionally, 5,199 non-redundant protein (i.e. drug target/enzyme/transporter/carrier) sequences are linked to these drug entries. |
| **STITCH** | <http://stitch.embl.de/> | A resource to explore known and predicted interactions of chemicals and proteins. |
| **Therapeutic Target Database (TTD)** | <http://bidd.nus.edu.sg/group/cjttd/> | Therapeutic Target Database (TTD) is a database to provide information about the known and explored therapeutic protein and nucleic acid targets, the targeted disease, pathway information, and the corresponding drugs directed at each of these targets. |
| ***Drug transcriptomic profiles*** | | |
| **GEO** | <https://www.ncbi.nlm.nih.gov/geo/> | GEO is a public functional genomics data repository supporting MIAME-compliant data submissions. Array- and sequence-based data are included. |
| **The Connectivity Map (CMap)** | <https://www.broadinstitute.org/cmap/> | Provide a comprehensive drug transcriptional responses of 1309 drugs or lead compounds in the clinical trials to six or seven different cancer cell lines |
| **Open TG-GATEs** | <http://toxico.nibio.go.jp/english/index.html> | TG-GATEs consists of the comprehensive toxicogenomic profiles of 170 compounds with four different assay types (human/rat in vitro/vivo) and multiple time and dose points in rat liver and kidney. The histopathological profiles for compounds are also available. |
| **DrugMatrix** | <https://ntp.niehs.nih.gov/drugmatrix/index.html> | DrugMatrix contains toxicogenomic profiles for 638 different compounds from both Codelink and Affymetrix platforms, which covers multiple organism including liver, kidney, heart, bone marrow, spleen and skeletal muscle. |
| **LINCS – L1000** | <http://www.lincsproject.org/> | The Library of Integrated Network-Based Cellular Signatures (LINCS) Program aims to create a network-based understanding of biology by cataloging changes in gene expression and other cellular processes that occur when cells are exposed to a variety of perturbing agents |
| **Bioassay data** | | |
| **PubChem** | <https://pubchem.ncbi.nlm.nih.gov/> | PubChem is the world's largest collection of freely accessible chemical information. Search chemicals by name, molecular formula, structure, and other identifiers. Find chemical and physical properties, biological activities, safety and toxicity information, patents, literature citations and more. |
| **ChEMBL** | <https://www.ebi.ac.uk/chembl/> | ChEMBL is a manually curated database of bioactive molecules with drug-like properties. It brings together chemical, bioactivity and genomic data to aid the translation of genomic information into effective new drugs. |
| **Drug safety data** | | |
| **SIDER 4** | <http://sideeffects.embl.de/> | SIDER contains information on marketed medicines and their recorded adverse drug reactions. The information is extracted from public documents and package inserts. |
| **FDALabelTM** | <https://nctr-crs.fda.gov/fdalabel/ui/search> | The FDALabel Database is a free web-based application used to query a database of over 100,000 drug labels |
| **FDA Adverse Event Reporting System (FAERS)** | <https://www.fda.gov/drugs/questions-and-answers-fdas-adverse-event-reporting-system-faers/fda-adverse-event-reporting-system-faers-public-dashboard> | It is a computerized information database designed to support the FDA's post-marketing safety surveillance program for all approved drug and therapeutic biologic products. |
| **PharmaPendium** | <https://www.pharmapendium.com/login/email> | PharmaPendium provides comparative regulatory document-based evidence in a single translational database for better-informed risk-benefit analyses and drug candidate assessments. |

**Step 2** - Establishment of COVID-19-drug interaction using drug repositioning principle: **a)** Determination of a fit-for-purpose principle based on different mechanisms for treating and eliminating COVID-19 related symptoms and **b)** Selection of drug candidates: Systematical *in-silico* screen of approved or investigational drugs to determine candidates for COVID-19 treatment.

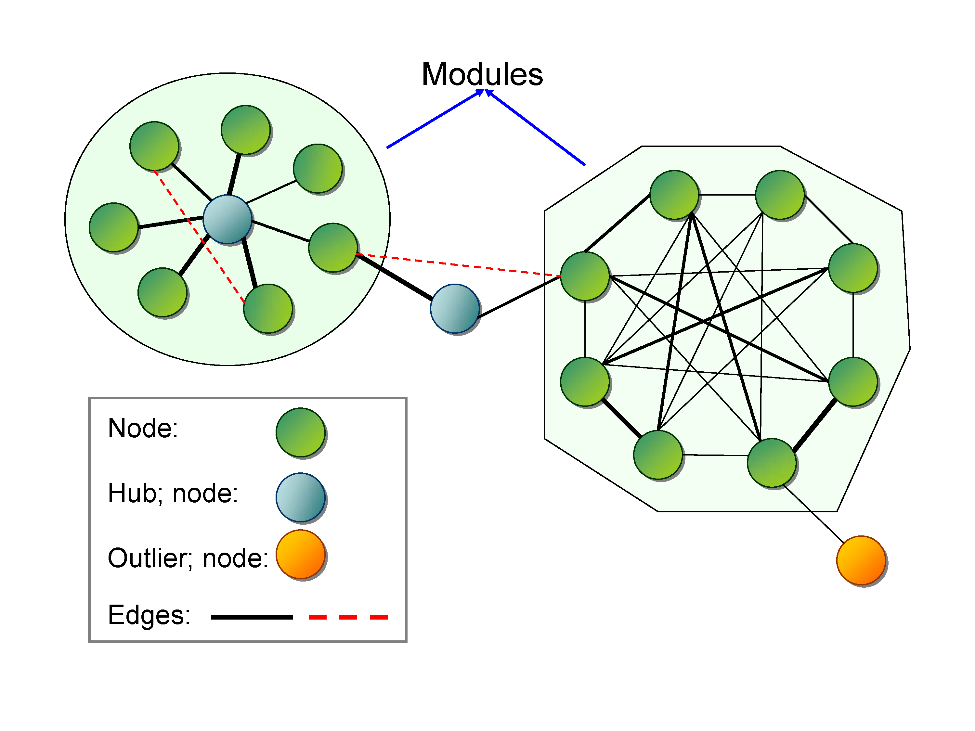
**3.2.1 Network modeling**

Network modeling is well-established in drug repositioning and has been applied in COVID-19 treatment development [6]. Although a network graph often presents challenges for visual interpretation, the simplicity of its concept inspires its persistent use in drug repositioning along with heat maps and clustering methodologies. Consequently, network modeling will be the main methodology evaluated in this project. A typical scenario of using network modeling in this project will resemble the method reported by Nacher et al.

They investigated the relationship between FDA approved drugs and human therapies by using a simple hypothesis to construct a drug-therapy network: if two disease therapies share the same drug, a link (i.e., edge) is established between them; *vice versa*, the same approach can also be applied to two drugs if they share the same therapeutic indication. Consequently, two networks were generated, one with drugs as nodes and the other using diseases as nodes. This approach will be extended to other types of data (e.g., transcriptomics, phenotypic data) to construct a network. The concept of network modeling in the context of drug repurposing and the key parameters that will be used to assess the success of an *in-silico* repurposing approach are summarized in **Box 1**.

**Box 1. Network modeling in the context of drug repositioning**

Network modeling links repositioning objects in a network format. The network consists of nodes, edges, hubs, modules, and outliers (see the illustration below), and its biological relevance is measured by the purity of identified modules and topological parameters such as betweenness centrality, closeness, etc. The network can then be used to predict novel repositioning opportunities.

****

**(a) Nodes and edges**

Nodes can be any of the repositioning objects, including drugs, diseases, targets, and mode of actions (MoAs). The functional connection between the nodes, named edges, reflect the specific physiological relationship between the nodes. For example, in a drug network where nodes are drugs, the edges represent similarity between drug parameters such as chemical structure or transcriptional responses.

**(b) Hubs and outliers**

Nodes possessing a high number of functional connections are named hubs (illustration, dark blue circles). Hubs can be inside the modules or serve to link several modules. In a disease network, for example, if one disease is a hub that connects with multiple diseases, it might indicate that the different diseases share commonality (e.g., pathogenesis, genetic mutations). In contrast, outliers are defined as nodes possessing few connections with other nodes. For instance, in a protein-protein network, a target is highly connected (a hub) and its inhibition may be involved in multiple biological processes. Thus, a drug which interacts with the target has a potential to be repurposed. A less connected target (an outlier) could be specific for a disease, and thus its corresponding drug may have the least potential to be repositioned via this target.

**(c) Modules**

A module contains a set of nodes that are highly interlinked. It is assumed that the nodes in the same module possess similar biological properties as defined by the module. For example, if a drug module is enriched for a specific therapeutic category, all the drugs in the module could be applied for this therapeutic use. The purity is a measure of enrichment.

**(d) Betweenness centrality**

This statistical measure quantifies the relative importance of a node in the network. For example, in a drug network, a drug with a high betweenness centrality value tends to be involved in multiple therapeutic usage and thus has a high potential to be repositioned.

**(e) Predict repositioning opportunities**

There are two ways to predict repositioning opportunities using network modeling. One method is to establish the previously unknown edge (new functional connection) based on the topological relationships of nodes, shown as a red dash line in Box 1. The other method is to position a new objective (e.g., drug or disease) in the network.

**3.2.2 Genomics based approaches**

The assumption of genomic-based drug repositioning is that (1) if the transcriptomic response signatures between a drug and a disease are reversely correlated, the drug has potential to treat the disease; (2) if the transcriptomic response signatures between the two drugs (drug A and drug B) are positively correlated, drug A can treat the indication of drug B, *vice versa* [7]. The Connectivity Map (CMap) [8] and LINCS project as the key sources has been widely applied to drug repositioning fields [9,10]. For example, Dudley *et al.* [11] proposed a novel approach that aims to look for inverse drug diseases relationship by comparing the disease signature generated from the Gene Expression Omnibus (GEO) databases [12] and drug signatures obtained from CMap. Besides CMap, several large toxicogenomics efforts such as TG-GATEs [13] and DrugMatrix [14] have accumulated hundreds of drugs transcriptome data profiles at multiple time/dose/assay type points. In this study, we will utilize collected COVID-related transcriptomic profiles to reversely compare with drug transcriptomic profiles to enrich the repositioning candidates.

Furthermore, some other genomics approaches will be also investigated for enriching candidates for COVID-19. For example, pathway or network approaches can be helpful in finding genes involved in general signaling networks or biological pathways, and could provide a list of proteins for therapeutic target identification [15]. Furthermore, linking the common disease with COVID-19 based on a shared gene is an idea originally proposed by Goh [16] which developed as a concept to identify the disease-disease relationship based on their shared pathways [17]. Then, the drug originally deigned to treat common diseases may potentially treat COVID-19 as well.

**3.2.3 Structure-based approaches**

Structure-based virtual screening approaches such as molecular docking have been adopted as an effective way for to prioritize the affinity of small molecules for certain disease-associated protein targets [18-20]. Molecular docking-based drug repositioning approaches comprises two underlying hypothesis: (1) if the disease-related protein target is known, then a set of small molecular or existing drugs could be docked to the known protein to prioritize the probability of the potential of drugs for treating the disease; (2) if the small molecule is synthesized by the chemist and aims to look for the potential pharmacological application, then reverse docking screening could be implemented to prioritize the probable disease-related targets that bind the small molecules [21]. As crystallization techniques advance, increasing numbers of protein structures with specific genetic mutations are available for COVID-19 in the public Protein Data Bank (**Table 1**) [22]. In this project, we will use molecule docking approaches and dynamic simulation to prioritize candidates for COVID-19 based on their affinity and PK/PD properties.

**3.2.4 Perturbation of immune system**

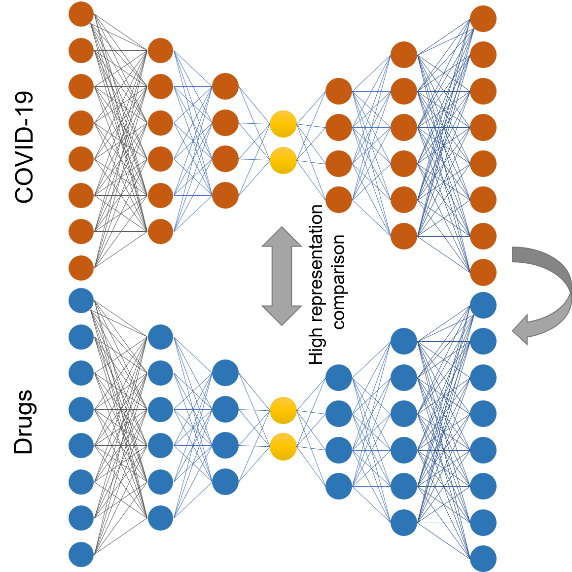
A better understanding how the existing drugs affect or trigger the immune system may pave a way to develop the treatment opinion of COVID-19 for different vulnerable populations. In the past decade, a lot of large data compendia on the immune systems have been generated and publicly available [23]. National Institute for Allergy and Infectious Disease (NIAID) in NIH implemented an ImmPort (<https://immport.niaid.nih.gov>) to share the molecular and clinical data of immune-related studies including population genetics analysis about immune systems, HLA region genomics in immune-related diseases [24]. Immunological Genome Project Consortium developed the pilot studies named ImmGen (<http://www.immgen.org>), which aims to provide a microarray gene expression data atlas of mouse immune-related cell line [25]. Those datasets could be applied to verify different hypothesis of repurposing existing drugs for COVID-19 treatment development.

**Step 3** - Prioritization of repositioning candidates with AI: prioritize candidates based on AI-driven algorithms, where drugs in current or proposed clinical trials for COVID-19 (e.g., Antivirals, Antimalarials, agents with ability to ameliorate the cytokine storm, and immunosuppressive medications) are used as positive controls.

Considering the big data curated from various resources and diverse of computational drug repostioning approaches will be employed in this project, we will develop AI-based approaches to prioritize repostioning candidates. specifically, three advanced deep learning will be adopted and investigated in this project.

**3.3.1 Autoencoder**

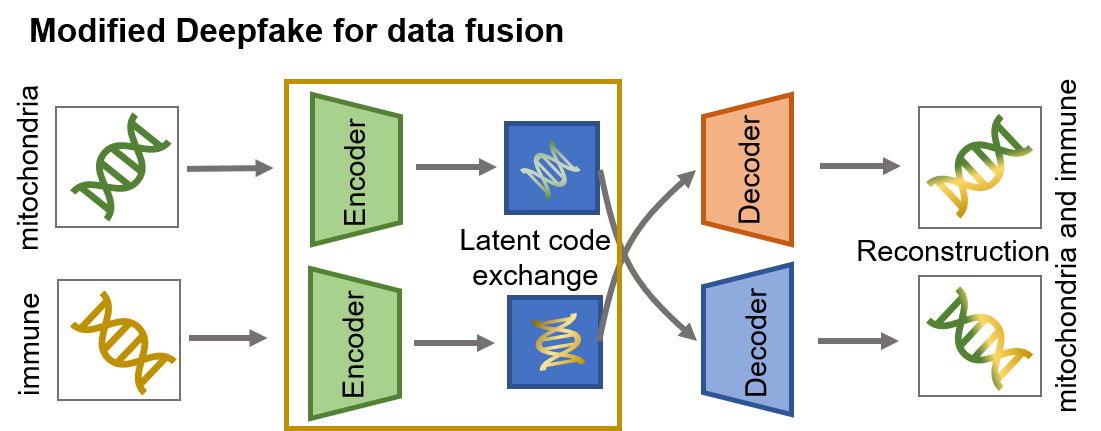
An autoencoder is a type of artificial neural network used to learn efficient data codings in an unsupervised manner [26]. The aim of an autoencoder is to learn a representation (encoding) for a set of data, typically for dimensionality reduction, by training the network to ignore signal “noise”. In this project, we will employ the Autoencoder algorithms to generated high representation for COVID-19 and drugs. If the high representation between drug and COVID-19 is high correlated, we considered the drug has potential to treat COVID-19 (**Figure 2**).



**Figure 2** proposed deep Autoencoder for COVID-19 treatment prioritization

**3.3.2 DeepFake**

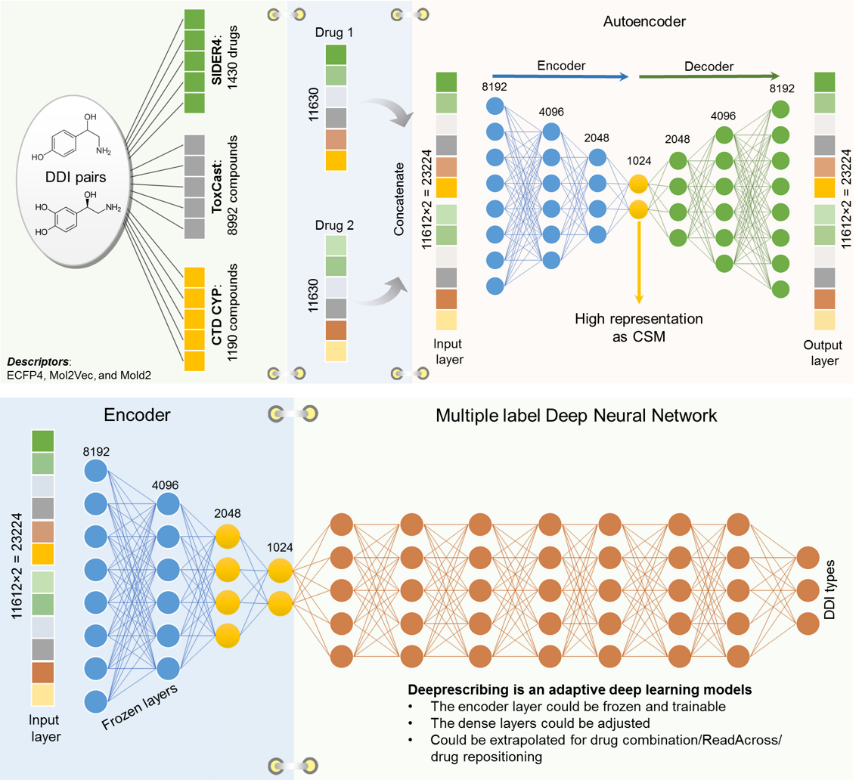
The immune systems are the key for patients conquering severe symptoms of COVID-19. Considering the immune response varies among the different vulnerable populations, a consideration should be taken to investigate the drug perturbation among the different immune-related mechanism. We will apply advanced deep learning algorithm named DeepFake [27] to investigate the energy and immune association for vulnerable patients and further prioritize the repostioning candidates for COVID-19 (**Figure 3**).



**Figure 3** Modified DeepFake for COVID-19 treatment prioritization

**3.3.3 Transfer learning**

Transfer learning is a research problem in machine learning that focuses on storing knowledge gained while solving one problem and applying it to a different but related problem [28]. Considering the preexisting condition of COVID-19 patients, they may take various of medicine. A careful check of potential drug-drug interactions (DDI) is of great importance to prioritize the repostioning candidates for patients with preexisting conditions. In this study, we will apply our developed transfer learning model titled DeepPrescribing for identifying the potential DDIs for proposed repostioning candidates. The framework of DeepPrescribing is illustrated in **Figure 4**.



**Figure 4** The framework of DeepPrescribing

**Step 4** - Evidence-based verification: shortlist the candidates by balancing safety and efficacy from a clinical perspective.

# **References**

**1** Wu, F. et al. (2020) A new coronavirus associated with human respiratory disease in China. *Nature* 579 (7798), 265-269

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

**3** Wang, Y. et al. (2020) Remdesivir in adults with severe COVID-19: a randomised, double-blind, placebo-controlled, multicentre trial. *Lancet (London, England)*

**4** Grein, J. et al. (2020) Compassionate Use of Remdesivir for Patients with Severe Covid-19. *New England Journal of Medicine*

**5** Cao, W. and Li, T. (2020) COVID-19: towards understanding of pathogenesis. *Cell Research* 30 (5), 367-369

**6** Zhou, Y. et al. (2020) Network-based drug repurposing for novel coronavirus 2019-nCoV/SARS-CoV-2. *Cell Discovery* 6 (1), 14

**7** Iorio, F. et al. (2013) Transcriptional data: a new gateway to drug repositioning? *Drug Discovery Today* 18 (7–8), 350-357

**8** Lamb, J. et al. (2006) The connectivity map: Using gene-expression signatures to connect small molecules, genes, and disease. *Science* 313 (5795), 1929-1935

**9** Qu, X.A. and Rajpal, D.K. (2012) Applications of Connectivity Map in drug discovery and development. *Drug Discovery Today* 17 (23–24), 1289-1298

**10** Iorio, F. et al. (2010) Discovery of drug mode of action and drug repositioning from transcriptional responses. *Proceedings of the National Academy of Sciences* 107 (33), 14621-14626

**11** Dudley, J.T. et al. (2011) Computational Repositioning of the Anticonvulsant Topiramate for Inflammatory Bowel Disease. *Science Translational Medicine* 3 (96), 6

**12** Barrett, T. et al. (2007) NCBI GEO: mining tens of millions of expression profiles - database and tools update. *Nucleic Acids Research* 35, D760-D765

**13** Igarashi, Y. et al. (2015) Open TG-GATEs: a large-scale toxicogenomics database. *Nucleic Acids Research* 43 (D1), D921-D927

**14** Natsoulis, G. et al. (2008) The liver pharmacological and xenobiotic gene response repertoire. *Molecular Systems Biology* 4, 12

**15** Jin, G. and Wong, S.T. (2014) Toward better drug repositioning: prioritizing and integrating existing methods into efficient pipelines. *Drug discovery today* 19 (5), 637-644

**16** Goh, K.-I. et al. (2007) The human disease network. *Proceedings of the National Academy of Sciences* 104 (21), 8685-8690

**17** Li, Y. and Agarwal, P. (2009) A Pathway-Based View of Human Diseases and Disease Relationships. *Plos One* 4 (2), 6

**18** Ekins, S. et al. (2007) In silico pharmacology for drug discovery: methods for virtual ligand screening and profiling. *British Journal of Pharmacology* 152 (1), 9-20

**19** Kolb, P. et al. (2009) Docking and chemoinformatic screens for new ligands and targets. *Current Opinion in Biotechnology* 20 (4), 429-436

**20** Kinnings, S.L. et al. (2009) Drug Discovery Using Chemical Systems Biology: Repositioning the Safe Medicine Comtan to Treat Multi-Drug and Extensively Drug Resistant Tuberculosis. *Plos Computational Biology* 5 (7), e1000423

**21** Zahler, S. et al. (2007) Inverse In Silico Screening for Identification of Kinase Inhibitor Targets. *Chemistry & Biology* 14 (11), 1207-1214

**22** Berman, H.M. et al. (2000) The Protein Data Bank. *Nucleic Acids Research* 28 (1), 235-242

**23** Sparks, R. et al. (2016) Expanding the Immunology Toolbox: Embracing Public-Data Reuse and Crowdsourcing. *Immunity* 45 (6), 1191-1204

**24** Bhattacharya, S. et al. (2014) ImmPort: disseminating data to the public for the future of immunology. *Immunologic Research* 58 (2), 234-239

**25** Heng, T.S.P. et al. (2008) The Immunological Genome Project: networks of gene expression in immune cells. *Nat Immunol* 9 (10), 1091-1094

**26** Hinton, G.E. and Salakhutdinov, R.R. (2006) Reducing the Dimensionality of Data with Neural Networks. *Science* 313 (5786), 504-507

**27** Kietzmann, J. et al. (2020) Deepfakes: Trick or treat? *Business Horizons* 63 (2), 135-146

**28** Yosinski, J. et al. (2014) How transferable are features in deep neural networks? In *Advances in neural information processing systems*, pp. 3320-3328