**Title:**

Drug repurposing to target COVID-19: a big data driven approach

建立用於新冠肺炎研究及藥物開發的大數據平台

**Keywords:** knowledge graph, drug repurposing, big data, pharmacy, drug interactions

**Project Duration:** 12 months

**Funding requested:** $2.5M

PC: Dr. Esther Chan

Co-PIs: Prof. Reynold CK Cheng, Prof. Ian CK Wong, Dr. Ruibang Luo, Dr. Chuan Wu, Prof. Ivan FN Hung, Prof. David CW Siu

Co-Is: Prof. Benjamin Cowling, Prof. Kevin CC Chang, Dr. Bo Tang

**ABSTRACT (400 words)**

COVID-19 has emerged as a severe global epidemic with high morbidity and mortality. In view of the possibility for viral mutation of SARS-CoV-2 and the potential for COVID-19 to become a recurrent pandemic, there is an urgent need for developing efficient tools and data platform for COVID-19 drug repurposing and research.

To date, there is no drug or vaccine clinically proven to be effective for management of COVID-19. While a number of antiviral and antimalarial agents are currently under trial for evaluating their efficacy as COVID-19 treatment, preliminary results suggest that some of these agents may not be as promising as speculated, whereas others are associated with serious adverse effects which limits their usage. Traditionally, structure-based virtual screening of chemical compound libraries and protein docking analyses are often used to search for new drug candidates for certain drug targets. However, such methods are limited in their ability to find candidates translatable to clinical use, since these methods search for candidates based on their chemical properties and affinity to bind specific drug targets, but do not take into account the “big picture” of complex intrinsic linkages between drugs, genes, diseases and viruses, which are important predictors of clinical efficacy. Hence, efficient computational methods that harness the wealth of big data, rather than structure-based screening alone, are essential to allow timely discovery of potential COVID-19 treatment.

A two-part project is proposed to address the unmet needs in COVID-19 research. First, we aim to build and refine a comprehensive COVID-19 knowledge graph by linking big data from a variety of data sources which includes information on drug, genes, viruses, proteins, diseases and symptoms. A comprehensive knowledge graph will enable the extraction of hidden linkages to generate useful insights regarding COVID-19 disease characteristics and potential treatment. The knowledge graph will be shared as an open data platform to facilitate global collaborative research. Second, we aim to utilize the COVID-19 knowledge graph to extract potential drug candidates that may be repurposable for COVID-19 treatment. Using knowledge graphs for drug repurposing enable us to consider complex intrinsic linkages between drugs, genes, diseases and viruses, where traditional structure-based virtual screening methods neglects. This will facilitate efficient discovery of safer alternatives for COVID-19 treatment, which can be tested via clinical trials and big data observational studies, and subsequently be translated to clinical use.

**1. Impact and objectives**

**(a) Project objectives**

The objective of the proposed project is to develop an open knowledge-graph-based platform to facilitate global COVID-19 research, and to utilize the platform for discovering potential drug candidates for COVID-19 treatment.

Specifically:

1. To develop and refine an open COVID-19 knowledge graph that will enable researchers worldwide to uncover hidden relationships between COVID-19, drugs, genes, proteins and diseases/symptoms, providing a platform for hypothesis generation to support future research on COVID-19.
2. To identify drugs and combinations of drugs that have potential to be repurposed for COVID-19 treatment using a COVID-19 knowledge graph.

**(b) Pathways to Impact Statement (should not exceed two A4 pages)**

COVID-19 has emerged as a severe global epidemic. By the end of May 2020, the number of confirmed cases worldwide had reached over 5 million. The lack of specific drug treatment for COVID-19 had contributed to more than 350,000 deaths worldwide with 100,000 deaths in the United States alone. While the situation in Hong Kong is gradually under control, the situations in Europe and America have not yet shown clear signs of improvement, and the infection is still actively spreading with more than 100,000 new cases per day worldwide. These countries are still hopeful for specific COVID-19 drug treatments to emerge, which could potentially save tens of thousands of lives globally, especially in the setting of potential viral mutation.

Unfortunately, although a number of antiviral and antimalarial agents are currently under trial for evaluating their efficacy as COVID-19 treatment, preliminary results suggest that some of these agents may not be as promising as speculated (1), whereas others are associated with serious adverse effects such as electrocardiographic changes which limits their usage (2). Hence, effective data platform and tools are essential to enable efficient discovery of new drug candidates and drug combinations, in a search for safer alternatives.

The proposed project aims to provide a knowledge-graph-based platform and tools that will enable researchers worldwide to efficiently extract previously less intuitive linkages between COVID-19, drugs, genes, proteins and diseases. This will further facilitate efficient discovery of potential COVID-19 treatments that can be tested via clinical trials and Big Data observational studies, and subsequently be translated to clinical use.

The proposed research is expected to have local and international impacts as illustrated below.

1. **Who are the potential beneficiaries of the proposed research in the short (1-3 years), medium (4-10 years) and long term (over 10 years)?**

*Short-term*: COVID-19 patients, their family and caregivers; healthcare professionals, government, medical researchers worldwide*. Medium-term*: COVID-19 patients, their family and caregivers; healthcare system users and the society as a whole. *Long-term*: the global population, patients with chronic diseases, especially in the setting of an ageing population who are prone to infections and chronic diseases.

1. **How will the potential beneficiaries benefit? What will be the objective demonstrable/measurable benefits beyond academia?**

In the short term, hidden or less obvious linkages between COVID-19, diseases, drugs, genes, and patient demographics, can be discovered. These hidden linkages may provide insight on previously unknown relations of how sex and age, underlying diseases or concomitant medications may influence a patient’s severity or chance of COVID-19 infection. Subsequently this would shed light on ways to prevent or reduce severity of COVID-19 in specific patient populations, through randomized clinical trials, population-based studies, *in vitro* testing on cell-based platforms and animal tests. This could potentially reduce the severity of the outbreak and hasten the resumption of normal work and economic productivity, in Hong Kong as well as Europe and America where the outbreak is still not under control. Findings will also help inform treatment decisions, prioritization of healthcare resources, and design and implementation of public health policies by the government. Furthermore, privacy-preserving data-mining strategies explored in this project will enable future incorporation of sensitive clinical data into open access data models without compromising the privacy of any individual. This will incentivize more data possessing institutions like hospitals and clinics to share data on COVID-19 and facilitate global collaboration.

In the short to medium term, new treatments for COVID-19, originating from the drug candidates extracted from our knowledge graph, which have proven to have benefit may have new added indications or approved for COVID-19 treatment and made accessible to patients worldwide. As COVID-19 could potentially be a recurrent epidemic and mutations of COVID-19 in the future may render some of the repurposed drugs ineffective, availability of new COVID-19 treatment could significantly reduce mortality and morbidity due to COVID-19 worldwide, potentially saving tens of thousands of lives, particularly in older individuals with chronic health conditions.

Furthermore, the outbreak of COVID-19 and the lack of specific drug treatment has led to an unprecedented demand on healthcare system in countries worldwide. Indeed, the saturation of healthcare system capacity due to COVID-19 had also interfered with the normal course of treatment of patients with other diseases such as psychiatry and cancer. Patients with acute conditions requiring emergency treatments, such as those with acute myocardial infarction, suffered significant delays resulting in complicated in-hospital course and worse clinical outcomes. As such, availability of new COVID-19 treatment will immensely benefit not only COVID-19 patients and their caregivers, but also other patients and users of the healthcare system, as well as the society as a whole. Such benefits may be demonstrable by shortened hospital length of stay, reduction in service delays and time from symptoms to first medical contact for emergency medical conditions, as well as reduced healthcare costs due to COVID-19 and its complications.

In the long-term, the proposed project provides powerful tools and experience for knowledge-graph-based drug repurposing. These knowledge graphs, tools and techniques could be applied to any infectious disease that may emerge in the future, and allow for quick identification and testing of new treatments. This can potentially allow early control of any potential outbreak and assist in preventing large-scale public health crisis due to new viral or infectious diseases. On the other hand, these knowledge-graph based techniques could also be extended to help identify treatments for currently incurable diseases and chronic diseases as well, which could potentially reduce the economic and societal burden associated with these diseases worldwide. These benefits could be demonstrated by a shortened time to discovery of treatments for new diseases and increased number of curative treatments for previously incurable diseases and chronic diseases, as well as reduced economic and healthcare costs associated with these diseases.

1. **What will be done during and / or after the project to increase the likelihood of achieving the identified benefit and reaching the identified beneficiaries?**

During the project, we will engage closely with collaborators to share interim findings and challenges. Findings will be disseminated at international conferences and in respectable peer-reviewed journals, as well as to the general public, policy makers, and international healthcare community through press conferences, potentially reaching >50 international media sources; and media channels including newsletters, website and social media to raise public awareness on this topic.

**2. Background of research, research plan and methodology**

**(a) Background of research and expected project commencement date**

COVID-19 has emerged as a severe global epidemic, with millions of people infected and tens of thousands of deaths worldwide. Yet to date, there is no effective drug or vaccine available to treat or prevent COVID-19 and its complications. There is an urgent need for efficient tools and data platform for COVID-19 drug repurposing and research.

We propose two interlinked projects to address these unmet needs in COVID-19 research, in the setting of potential viral mutation. These projects are expected to commence by January 2021.

**(b) Research plan and methodology**

**Project 1: Building and refinement of an open knowledge graph for COVID-19 research**

**Rationale:**

Knowledge graphs enable us to identify valuable information regarding the large-scale, complex relationships among different entities related to COVID-19, including drug, disease, virus, proteins and genes. Modern computational algorithms can be used to extract hidden linkages from knowledge graphs to generate useful insights and testable hypotheses. Example applications of knowledge graphs include drug repurposing, identifying potential adverse effect of drugs, disease subtyping (3) and herb recommendation (4).

However, to date, there is no unified knowledge graph openly available for use in COVID-19 research. While there are recent publications on knowledge graphs linking COVID-19 publications, case statistics and genes (5), they do not include essential information for drug discovery, for instance drug-protein and drug-gene relationships. Others like OpenKG contained only a set of disconnected small knowledge graphs for limited subtopics and themes, which are disjoint and could not be directly used to extract useful information for drug repurposing. Individual data sources are available providing information on drugs, drug-protein, gene-disease and virus-disease relationships (6, 7), yet these data sources are all independent and have not be linked into a unified knowledge graph.

A comprehensive, unified knowledge graph is needed to enable efficient extraction of high-level linkages, such as existing drugs repurposable for other diseases. These high-level linkages provide promising leads and enable faster discovery of potential treatments and unknown intrinsic characteristics of new diseases.

**Pilot data and knowledge graph prototype:**

As a proof-of-concept, a small-scale knowledge graph (~48K nodes and 815K edges) was built based on limited data from OpenKG and DrugBank. Using motif-discovery algorithms, a pilot list of 22 drugs with potential effects on COVID-19 was identified. The list included a variety of drugs from different drug classes, including vasodilators, antimicrobials, antimalarials, antivirals, immunosuppressants, coenzymes and amino acids, and trace elements. Some of these drugs are undergoing efficacy trials while others have not been explored as COVID-19 treatment currently. These results support the feasibility of our approach and verifies that findings from knowledge graph are in line with existing biomedical literature while also able to generate novel findings that did not exist in current literature. Through this project, we will further enrich the knowledge graph through incorporating high quality open data sources, and explore the use of private data sources, to build a large-scale comprehensive knowledge graph, which would allow us to uncover more drug candidates and safer alternatives which has not been explored currently but carry potential for COVID-19 treatment.

**Objectives:**

An overarching COVID-19 knowledge graph is a bedrock of this project. Building upon the foundation of the small-scale knowledge graph prototype described above, we aim to further enrich and develop an open COVID-19 knowledge graph which will enable researchers worldwide to uncover hidden relationships between COVID-19, drugs, genes, proteins and diseases/complications, providing a big data platform for hypothesis generation and future research on COVID-19. Specifically,

1. To enrich the coverage of the knowledge graph by incorporating open data sources, including non-knowledge-graph databases and other knowledge graphs
2. To explore ways to incorporate private data sources, such as clinical and patient data, and address the challenges involved including privacy concern.
3. To provide an open user interface for researchers and healthcare professionals to utilize the knowledge graph for their topics of interest.

**Methodology:**

1. **Enriching the knowledge graph with open data sources**

*Data sources*

To build a comprehensive knowledge graph for COVID-19 research and drug repurposing, we will include not only the databases that are directly associating diseases with drugs but also databases that provide “bridges” via chemical, structural, genetic, symptomatic, and societal interactions. The schema for this knowledge graph is shown in Figure 1, covering linkages between COVID-19, drugs, genes, proteins, diseases and symptoms. Below we listed the databases that we will be using, and they are divided into three bioinformatics categories: 1) drug-target interactions; 2) gene-gene interactome; and 3) gene-disease network. We will also incorporate 4) biomedical literature mining and 5) other knowledge graphs as appropriate.

1. Drug-target interactions:

Drug-target interactions involve drug metadata and drug-target linkages. The drug metadata consist of descriptions of the drug from DrugBank and the clinical trial information from the ClinicalTrials.gov. The DrugBank ID will be used to represent the drug in the graph. The linkages between drugs and targets (gene/protein) are collected from the PharmGKB database, BindingDB, ChEMBL (v20), IUPHAR/BPS Guide to pharmacology, Therapeutic Target Database, and DrugBank (v4.3). The drug-target interactions can be filtered by binding affinities and review status from UniProt. Descriptions for each database are as follows:

* Pharmacogenomics Knowledgebase (PharmGKB) is a publicly available, online knowledgebase responsible for the aggregation, curation, integration, and dissemination of knowledge regarding the impact of human genetic variation on drug response.
* BindingDB is a public, web-accessible database of measured binding affinities, focusing chiefly on the interactions of proteins considered to be candidate drug-targets with ligands that are small, drug-like molecules.
* 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.
* IUPHAR/BPS Guide to pharmacology is an expert-curated resource of ligand-activity-target relationships, the majority of which come from high-quality pharmacological and medicinal chemistry literature.
* Therapeutic Target Database (TTD) involves information about (i) target-regulating microRNAs and transcription factors, (ii) target-interacting proteins, and (iii) patented agents and their targets (structures and experimental activity values if available), which can be conveniently retrieved and is further enriched with regulatory mechanisms or biochemical classes.
* DrugBank database (7) is a comprehensive, freely accessible database containing information on drugs and drug targets. It currently contains 13,575 drug entries and is widely used by industry, medical practitioners and the general public. It has enabled the discovery and repurposing of a number of existing drugs to treat rare and newly identified illnesses and is a unique bioinformatics and cheminformatics resource.
* ClinicalTrials.gov is a Web-based resource that provides patients, their family members, health care professionals, researchers, and the public with easy access to information on publicly and privately supported clinical studies on a wide range of diseases and conditions.

1. Gene-gene interactome:

The gene-gene interactome involves BioGRID, Database of Interacting Proteins, HPRD, OPHID, KEGG, PathBank, and Reactome. Among them, KEGG, PathBank, and Reactome provide the gene-gene relationship based on pathway information while the others are based on the protein-protein interaction. The NCBI Entrez ID and official gene symbol are used to represent the gene while the mapping information of the gene and protein is from UniProt. Descriptions for each database are as follows:

* BioGRID is an interaction repository with data compiled through comprehensive curation efforts.
* Database of Interacting Proteins catalogs experimentally determined interactions between proteins. It combines information from a variety of sources to create a single, consistent set of protein-protein interactions.
* Human Protein Reference Database (HPRD) represents a centralized platform to visually depict and integrate information pertaining to domain architecture, post-translational modifications, interaction networks, and disease association for each protein in the human proteome.
* OPHID is integrated known, experimental, and predicted PPIs to facilitate experimentation and integrated computational analysis with model organism PPI networks.
* KEGG is a database resource for understanding high-level functions and utilities of the biological system, such as the cell, the organism and the ecosystem, from molecular-level information, especially large-scale molecular datasets generated by genome sequencing and other high-throughput experimental technologies.
* PathBank is an interactive, visual database containing more than 100 000 machine-readable pathways found in model organisms such as humans, mice, E. coli, yeast, and Arabidopsis thaliana. The majority of these pathways are not found in any other pathway database.
* Reactome is a free, open-source, curated, and peer-reviewed pathway database.
* NCBI Entrez is a molecular biology database system that provides integrated access to nucleotide and protein sequence data, gene-centered and genomic mapping information, 3D structure data, PubMed MEDLINE, and more.

1. Gene-disease network:

The gene-disease network involves CTD, Genetic Association Database, Human Disease Network, OMIM, Orphanet, and HPO. Disease mapping is based on the Disease Ontology, while MeSH ID is used to represent the disease in the graph. Descriptions for each database are as follows:

* Comparative Toxicogenomic Database (CTD) provides manually curated information about chemical–gene/protein interactions, chemical–disease, and gene-disease relationships. These data are integrated with functional and pathway data to aid in the development of hypotheses about the mechanisms underlying environmentally influenced diseases.
* Genetic Association Database is a database of genetic association data from complex diseases and disorders.
* Human Disease Network is a network in which nodes represent diseases and two diseases are connected if they share at least one gene in which mutations are associated with both diseases.
* Online Mendelian Inheritance in Man (OMIM) is a comprehensive, authoritative compendium of human genes and genetic phenotypes.
* Orphanet is a unique resource, gathering and improving knowledge on rare diseases to improve the diagnosis, care, and treatment of patients with rare diseases.
* Human Phenotype Ontology (HPO) is central in medical genetics and genomics It provides a standardized vocabulary of phenotypic abnormalities encountered in human disease and serves as a computational bridge between genome biology and clinical medicine.
* Disease Ontology (DO) provides an open-source ontology for the integration of biomedical data that is associated with human disease.
* MeSH (Medical Subject Headings) is the NLM controlled vocabulary thesaurus used for indexing articles for PubMed.

1. Biomedical literature mining:

The PubMed Central Open Access corpus released in May 2018 contains 2.2 million full-text biomedical articles. Most of the articles are under the CC0 or CC-BY license, meaning we are free to redistribute and reuse most of the articles in the corpus. The full-texts are stored in text files (.txt) and are available for bulk download. Relying on natural language processing breakthroughs, Dr. Luo and the project team have identified 300k+ new gene-disease interactions from the 2.2 million biomedical articles that were missing in the aforementioned databases. These new interactions will be added to the our COVID-19 knowledge graph to massively enhance the success rate of using our graph for repurposing drugs.

1. Other knowledge graphs

Existing publications which released integrated datasets would also be incorporated, if relevant, in building the knowledge graph.

*Data processing*

A knowledge graph is a multi-relational graph composed of entities (nodes) and relations (edges), which are triplets of facts (head entity, relation, tail entity), denoted as (h, r, t) (8). In the case of a COVID-19 knowledge graph, each node could represent a specific protein, gene, drug, virus, disease or symptom, whereas each edge represents a known existing linkage between any two nodes. Data on linkages from different data sources detailed above will be processed into the corresponding nodes and edges, which will be linked and combined into a large-scale knowledge graph.

To align the data from different sources, records from terminology databases such as HPO will be used, which provides unique identifiers for entities with different alias. Knowledge graph provides a comprehensive means to capture, represent and formalize structured information for knowledge discovery.

~~Algorithms such as Jaccard Coefficient, Katz Index and Rooted PageRank (9) will be used to predicate the potential linkages between drugs and SARS-CoV-2. Novel machine learning techniques will be used to generate latent, low-dimensional representations of the knowledge graph called embeddings, which can then be utilized for standard downstream artificial intelligence modules such as clustering, link prediction and ranking (8).~~

1. **Exploring the use of private data sources**

Private data sources refer to data sources involving sensitive data where access must be restricted. One prominent example is clinical data of COVID-19 patients. During the COVID-19 world crisis, many hospitals and institutions worldwide admitted patients and explored different treatments. Each hospital/institution stores data about patients, symptoms, drugs used during treatment, laboratory test results and clinical outcomes in the form of electronic medical records. Clinical data could provide valuable information on drug-disease linkages for the knowledge graph.

However, one key issue arising in the joint learning of knowledge graph from clinical data/electronic medical records is how to ensure privacy since information will be shared and exchanged during the joint learning process and patient information is sensitive data which should be protected from any leakage. There have been some recent studies showing that it is possible to extract sensitive information from only the parameters of a trained model (10, 11). Besides, with the General Data Protection Regulation (GDPR) put in place by European Union and the Health Insurance Portability and Accountability Act (HIPAA) being enacted by the United States, privacy preserving is not a choice but a necessity when learning with sensitive data.

As such, we will explore how information from clinical data could be utilized in building the knowledge graph without integrating these data sources directly or compromising data privacy. The following approaches will be investigated to address these challenges.

*Federated learning over knowledge graph*

Federated learning is a novel distributed learning paradigm proposed by Google (12). Instead of centralizing all the data from distributed devices to one single server, federated learning keeps the training data locally and updates are performed by local participants. Only updated weights from each device are sent to central server for aggregation in order to update the learning model. Locally stored training data is the basic need to protect privacy. Besides, there are both theoretically and empirically proved federated learning algorithms, such as FedAvg (12) and FegProx (13), to ensure successful distributed learning. Unlike non-structural data like text and images, learning over structural data, like graph, requires specific neural network architecture. Graph Neural Network (GNN) is a type of neural network architecture designed for dealing with graph data. Information from neighboring nodes and edges are gathered in the learning process and the graph structure can remain the same in GNN training.

*Differential privacy for privacy preserving*

Simply anonymizing data by removing sensitive personal information is not enough for privacy preserving (10, 11). Differential privacy is a privacy preserving technique that has been investigated for decades. The main advantage of differential privacy is that it provides principled and rigorous privacy guarantee even when the adversary knows the privacy preserving mechanism and has arbitrary-side information. The promise from differential privacy is that the learning result will not reveal whether a specific individual participates in the learning process or not with high confidence, and this is exactly the desired property when dealing with individual related sensitive data. Combined with federated learning process, we assume the central server can be trusted. Updates from distributed devices will be aggregated and well designed differential privacy mechanisms will be applied to updates at the central server so that no more sensitive information can be inferred from the learning process and the trained model.

1. **Building an open data platform**

*Data storage*

The graphs are extremely large, on the order of million to billions of nodes, and hence difficult and inefficient to process. Therefore, we will investigate novel large-scale distributed learning methodologies for efficient graph processing, such as graph neural networks (GNN). The very large graphs can be stored in graph stores distributed onto multiple storage nodes, and graph processing algorithms, e.g, motif-search algorithm, are executed over multiple computing nodes while ensuring algorithm convergence and correctness.

*Data sharing*

The COVID-19 knowledge graph generated will be made available to researchers worldwide as an open data platform to facilitate global collaborative research. An open user interface for the knowledge graph will also be built to enable researchers and healthcare professionals to explore the knowledge graph and extract useful information based on their own topics of interest.

**Application and impact:**

By linking a variety of scattered data sources into a unified knowledge graph, previously unexplored relationships between COVID-19, diseases, drugs, genes, and patient demographics, could be efficiently extracted. These hidden linkages may provide insight on previously unknown relations of how underlying diseases, concomitant medications or genetic profile may influence a patient’s severity or chance of COVID-19 infection. It will also allow an exploration of any potential differences among different sex or age groups such as elderly. Subsequently this would shed light on ways to prevent or reduce severity of COVID-19 in specific patient populations. Privacy preserving strategies explored in this project will allow the knowledge graph to be open-accessed to facilitate other applications and researches without compromising the privacy of any individual. This will incentivize more data possessing institutions like hospitals and clinics to contribute to the joint model, which can further improve the model and provide more useful insights into the possible treatment of COVID-19.

**Project 2: Graph-based drug repurposing for COVID-19 treatment (undertaken simultaneously with Project 1)**

**Rationale:**

To date, there is no specific drug or vaccine available to treat or prevent COVID-19 and its complications. As such, there is an urgent need for developing effective strategies for prevention and treatment of COVID-19. While conventional structure-based screening methods such as protein docking analyses are traditionally used for *de novo* drug discovery, repurposing of existing drugs provide a more cost and time efficient way to discover treatment for new diseases (14).

Globally, antimalarial and antiviral agents are currently under trial for treatment of COVID-19 (15, 16). While these may have shed hope on treating COVID-19, yet from experience of other diseases, only a small proportion of new chemical entities or trialed drug candidates would succeed to show efficacy and be approved for treatment. Indeed, preliminary results from some of these trials suggest that certain candidates may not be as promising as speculated. For instance, in a randomized trial from China of 237 patients with severe COVID-19, remdesivir and placebo did not show significant difference in times to clinical improvement (median 21 versus 23 days) and mortality rates (14% versus 13%) (1). On the other hand, antimalarial and antiviral agents are generally associated with undesirable adverse effect profile and drug-drug interactions. For instance, hydroxychloroquine and chloroquine are known to prolong QTc interval. In an observational study where 84 patients received hydroxychloroquine, 10% had electrocardiographic changes which required discontinuation (2). Hence, efficient discovery of more drug candidates, especially safer alternatives, is essential.

Currently, speculation on drug repurposing for COVID-19 treatment have largely focused on antiviral and antimalarial agents, which were known to exert antiviral effect on HIV, Ebola and other viruses, but not proven effective for SARS-CoV-2. Antiviral drug combinations, such as cocktail therapy using two or more antiviral agents, are also being explored. However, repurposing of non-antiviral agents had not been adequately explored, attributable to the less intuitive linkages between non-antiviral agents and COVID-19. Yet, it has been shown that non-antiviral agents could also influence host response to viral infections, by reducing the chance of viral entry into cells or suppressing overreaction of the immune system to the virus. For instance, Angiotensin II Receptor Blockers (ARBs) may antagonize the proinflammatory effects of angiotensin II which is increased due to COVID-19 infection (17). Further, cardiovascular medications may improve survival as COVID-19 have been reported to have cardiovascular sequelae (18).

Yet, with the vast number of existing drugs available, experimental approaches to find repurposable drugs would be costly and time-consuming. Structure-based virtual screening of chemical compound libraries and protein docking analyses are often used to search for new drug candidates for certain drug targets. However, such methods are limited in their ability to find candidates translatable to clinical use, since these methods search for candidates based on their chemical properties and affinity to bind specific drug targets, but do not take into account the “big picture” of complex intrinsic linkages between drugs, genes, diseases and viruses, which are important predictors of clinical efficacy. Hence, efficient computational methods that harness the wealth of big data, rather than structure-based screening alone, are essential to allow timely discovery of potential COVID-19 treatment.

**Objectives:**

This project will be carried out concurrently with project 1.

1. To identify drugs and combination of drugs that have potential to be repurposed for COVID-19 treatment using big data.
2. To evaluate the quality of the knowledge graph with expert knowledge

**Methodology**:

1. **Drug repurposing for COVID-19**

*Extraction of drug candidates from knowledge graph*

Potential linkages relevant to drug repurposing will be extracted from a large-scale, comprehensive COVID-19 knowledge graph generated from project 1, using data mining techniques. Specifically, motif discovery algorithms will be designed to discover frequent high-order patterns of interest (i.e. motifs) in knowledge graphs (19). Examples of motifs relevant to drug repurposing include drug-protein-virus, virus-symptom-gene-drug and others which will be further defined and explored. Subgraphs that matches the motifs of interest will then be extracted using specific motif-clique discovery algorithms (20). These subgraphs will provide information on drug candidates potentially repurposable for COVID-19 treatment. Combinations of all drug classes including antiviral agents, immunosuppressants and cardiovascular medications will also be explored similarly, by extracting subgraphs where drugs of different classes share a common set of linkages to genes, proteins, diseases and symptoms.

*Evaluation and ranking of drug candidates for their clinical applicability*

The potential drug and drug combinations extracted from the knowledge graph will be further ranked through computational scoring and literature review, in order to provide a way to prioritize certain drug candidates over others in future testing on cell-based platforms or clinical trials.

+Weights that can be applied on the knowledge graph nodes/edges

+Scores or other metrics that can be computed for the drug candidates found based on the knowledge graph

**Algorithms such as Jaccard Coefficient, Katz Index and Rooted PageRank (9) will be used to predicate the potential linkages between drugs and** **SARS-CoV-2. For example, Katz Index can be used to generate scores for each drug candidate, considering the number and length of the shortest paths between the drug and the COVID-19 virus on the knowledge graph. Drugs with higher scores will be recommended for clinical examination with higher priority.**

**Novel machine learning techniques will be used to generate latent, low-dimensional representations of the knowledge graph called embeddings, which can then be utilized for standard downstream artificial intelligence modules such as clustering, link prediction and ranking (8).**

**Geographic data will be used to refine the quality of the rank. Viral evolution and genome variability enable viruses to escape host immunity and to develop drug resistance. Since mutations are detected from SARS-CoV-2 stains of different cities (21), drugs to be recommended should be customized. For example, the drug candidate will be affected if its targeting virus protein is mutated. Thus, this drug candidate should not be recommend to a city where virus strains are detected with this kind of virus protein mutation.**

Literature review will be conducted for the highest ranked drug candidates to provide evidence to rank their clinical applicability for use in COVID-19 patients, with respect to the following aspects:

1. possible underlying mechanisms of action and biological pathways involved that support their potential effect on treating COVID-19
2. potential adverse effects and safety for use, considering different population including the elderly
3. potential drug-drug interactions and drug-disease interactions, which may preclude the use of the drug candidates in patients with certain underlying diseases or taking specific concomitant medications
4. **Evaluating the quality of the knowledge graph with expert knowledge**

With a vast number of data sources and great variety in data formats involved, it is essential to ensure that the combined knowledge graph is free from human error and in line with existing knowledge in the medical research field. As such, we will concurrently evaluate and refine the knowledge graph in an iterative process. To evaluate the clinical relevance of the combined knowledge graph, a selected set of known drug-disease relationships will be extracted from the knowledge graph using methods described in part A. An expert panel consisting of collaborators and experts from infectious disease and cardiology fields will review the knowledge graph outputs and provide clinical input on whether results extracted from the knowledge graph are in line with existing literature and known biological mechanisms of existing drugs or diseases. Data sources would then be added or removed in an iterative process based on these evaluations to enhance the overall quality and reduce redundancy in the knowledge graph.

**Application and impact:**

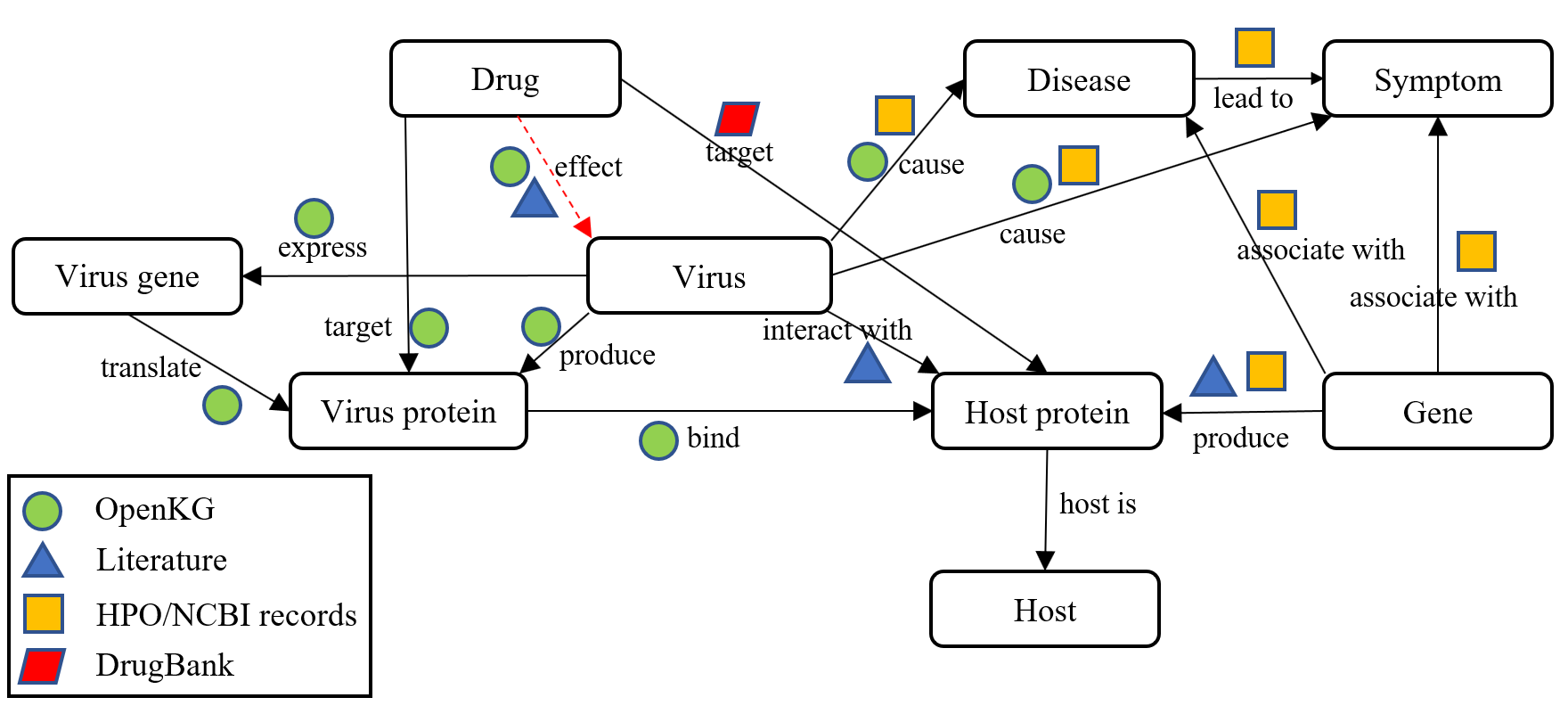
Findings will reveal novel drug candidates and combinations of drugs from different drug classes that have the potential to be repurposed for COVID-19 treatment. The efficacy of these drug candidates for COVID-19 treatment can then be further validated via randomized clinical trials and multi-centre Big Data observational studies using electronic medical records, which can be translated to clinical practice and form the treatment armamentarium for COVID-19. Findings can also contribute to testing of specific drugs and drug combinations directly using cell-based platforms.

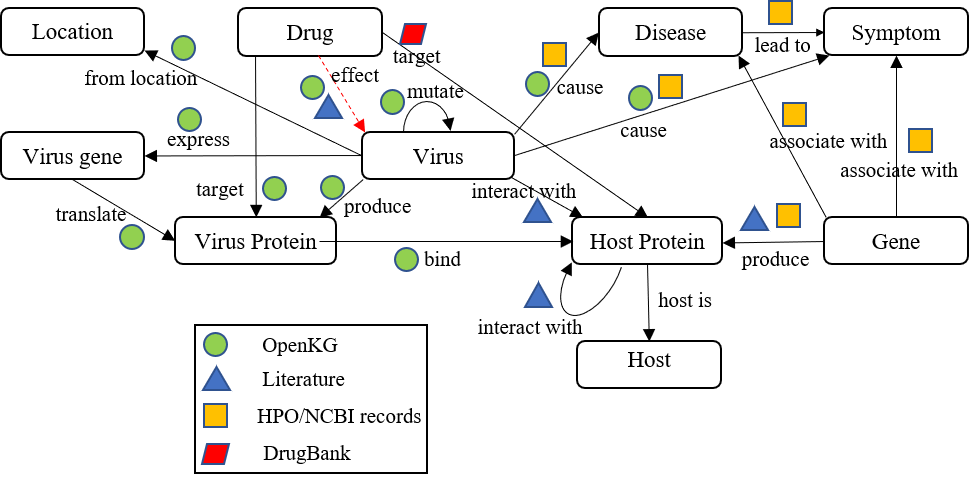
**(c) A one-page Gantt chart showing the research activities**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Activity | 2021 | | | | |
| Jan | Mar | Jun | Sep | Dec |
| Data collection |  |  |  |  |  |
| Data processing to build COVID-19 knowledge graph |  |  |  |  |  |
| *Milestone 1: 6 months from project commencement*  Completion of open COVID-19 knowledge graph |  |  |  |  |  |
| Extraction of a list of potential drug candidates for COVID-19 treatment from knowledge graph |  |  |  |  |  |
| Literature review of biological mechanisms, adverse effects and drug interactions of the drug candidates |  |  |  |  |  |
| *Milestone 2: 12 months from project commencement*  Publication of results (list of repurposable drugs for COVID-19 treatment) and submission of final report |  |  |  |  |  |
| Exploring the potential for selected drug candidates to be tested on cell-based platforms or though clinical trials |  |  |  |  |  |

**(d) Figures and tables**

**Figure 1. Sample schema of COVID-19 knowledge graph**





**(e) References**

1. Wang Y, Zhang D, Du G, Du R, Zhao J, Jin Y, Fu S, Gao L, Cheng Z, Lu Q, Hu Y, Luo G, Wang K, Lu Y, Li H, Wang S, Ruan S, Yang C, Mei C, Wang Y, Ding D, Wu F, Tang X, Ye X, Ye Y, Liu B, Yang J, Yin W, Wang A, Fan G, Zhou F, Liu Z, Gu X, Xu J, Shang L, Zhang Y, Cao L, Guo T, Wan Y, Qin H, Jiang Y, Jaki T, Hayden FG, Horby PW, Cao B, Wang C. Remdesivir in adults with severe COVID-19: a randomised, double-blind, placebo-controlled, multicentre trial. The Lancet. 2020;395(10236):1569-78.

2. Mahévas M, Tran V-T, Roumier M, Chabrol A, Paule R, Guillaud C, Fois E, Lepeule R, Szwebel T-A, Lescure F-X, Schlemmer F, Matignon M, Khellaf M, Crickx E, Terrier. Clinical efficacy of hydroxychloroquine in patients with covid-19 pneumonia who require oxygen: observational comparative study using routine care data. BMJ. 2020;369.

3. Li B, Cheng R, Hu J, Fang Y, Ou M, Luo R, Chang KCC, Lin. X. MC-Explorer: Analyzing and Visualizing Motif-Cliques on Large Networks. IEEE International Conference on Data Engineering 2020. p. 1722-5.

4. Jin Y, Zhang W, He X, Wang X, Wang X. Syndrome-aware Herb Recommendation with Multi-Graph Convolution Network2020.

5. The CovidGraph Project. CovidGraph. 2020. Available from: <https://covidgraph.org/>.

6. Köhler S, Vasilevsky NA, Engelstad M, Foster E, McMurry J, Aymé S, Baynam G, Bello SM, Boerkoel CF, Boycott KM, Brudno M, Buske OJ, Chinnery PF, Cipriani V, Connell LE, Dawkins HJS, Demare LE, Devereau AD, de Vries BBA, Firth HV, Freson K, Greene D, Hamosh A, Helbig I, Hum C, Jähn JA, James R, Krause R, F Laulederkind SJ, Lochmüller H, Lyon GJ, Ogishima S, Olry A, Ouwehand WH, Pontikos N, Rath A, Schaefer F, Scott RH, Segal M, Sergouniotis PI, Sever R, Smith CL, Straub V, Thompson R, Turner C, Turro E, Veltman MWM, Vulliamy T, Yu J, Von Ziegenweidt J, Zankl A, Züchner S, Zemojtel T, Jacobsen JOB, Groza T, Smedley D, Mungall CJ, Haendel M, Robinson PN. The Human Phenotype Ontology in 2017. Nucleic Acids Res. 2017;45(D1):D865-D76.

7. Wishart DS, Feunang YD, Guo AC, Lo EJ, Marcu A, Grant JR, Sajed T, Johnson D, Li C, Sayeeda Z, Assempour N, Iynkkaran I, Liu Y, Maciejewski A, Gale N, Wilson A, Chin L, Cummings R, Le D, Pon A, Knox C, Wilson M. DrugBank 5.0: a major update to the DrugBank database for 2018. Nucleic Acids Res. 2018;46(D1):D1074-d82.

8. Wang Z, Zhang J, Feng J, Chen Z. Knowledge Graph Embedding by Translating on Hyperplanes2014.

9. Lü L, Zhou T. Link prediction in complex networks: A survey. Physica A: Statistical Mechanics and its Applications. 2011;390(6):1150-70.

10. Wu X, Fredrikson M, Jha S, Naughton JF, editors. A Methodology for Formalizing Model-Inversion Attacks. 2016 IEEE 29th Computer Security Foundations Symposium (CSF); 2016 27 June-1 July 2016.

11. Fredrikson M, Jha S, Ristenpart T. Model Inversion Attacks that Exploit Confidence Information and Basic Countermeasures. Proceedings of the 22nd ACM SIGSAC Conference on Computer and Communications Security; Denver, Colorado, USA: Association for Computing Machinery; 2015. p. 1322–33.

12. McMahan HB, Moore E, Ramage D, Hampson S, Arcas BAy, editors. Communication-Efficient Learning of Deep Networks from Decentralized Data. AISTATS; 2017.

13. Li T, Sahu AK, Zaheer M, Sanjabi M, Talwalkar A, Smith V. Federated optimization in heterogeneous networks. arXiv preprint arXiv:181206127. 2018[Epub ahead of print].

14. Cheng F. In Silico Oncology Drug Repositioning and Polypharmacology. Methods Mol Biol. 2019;1878:243-61.

15. Sanders JM, Monogue ML, Jodlowski TZ, Cutrell JB. Pharmacologic Treatments for Coronavirus Disease 2019 (COVID-19): A Review. JAMA. 2020;323(18):1824-36.

16. Hung I, Lung K-C, Tso E, Liu R, Chung T, Chu M-Y, Ng Y-Y, Lo J, Chan J, Tam A, Shum H, Chan V, Wu A, Sin K-M, Leung W-S, Law W-L, Lung D, Sin S, Yeung P, Yuen K-Y. Triple combination of interferon beta-1b, lopinavir–ritonavir, and ribavirin in the treatment of patients admitted to hospital with COVID-19: an open-label, randomised, phase 2 trial. The Lancet. 2020[Epub ahead of print].

17. Rothlin RP, Vetulli HM, Duarte M, Pelorosso FG. Telmisartan as tentative angiotensin receptor blocker therapeutic for COVID-19. Drug development research. 2020[Epub ahead of print].

18. Zhou M, Wong C-K, Un K-C, Lau Y-M, Lee JC-Y, Tam FC-C, Lau Y-M, Lau HW, Tam A, Lam Y-Y, Pang P, Tong T, Tang M, Tse HF, HO D, NG M-Y, Chan E, Wong I, Lau C-P, HUNG I, Siu CW. Cardiovascular Sequalae in Uncomplicated COVID-19 Survivors. BMJ (under review). 2020[Epub ahead of print].

19. Li Y, Lou Z, Shi Y, Han J, editors. Temporal Motifs in Heterogeneous Information Networks. 14th International Workshop on Mining and Learning with Graphs; 2018.

20. Hu J, Cheng R, Chang KC, Sankar A, Fang Y, Lam BYH, editors. Discovering Maximal Motif Cliques in Large Heterogeneous Information Networks. 2019 IEEE 35th International Conference on Data Engineering (ICDE); 2019 8-11 April 2019.

21. Pachetti M, Marini B, Benedetti F, Giudici F, Mauro E, Storici P, Masciovecchio C, Angeletti S, Ciccozzi M, Gallo RC, Zella D. Emerging SARS-CoV-2 mutation hot spots include a novel RNA-dependent-RNA polymerase variant. Journal of Translational Medicine. 2020 Dec;18:1-9.

**7. (b) Plan(s) for collaboration in this application (max 2 pages 800 words)**

Experts from pharmacology and pharmacy, computer science, public health and medicine formed a multidisciplinary team to support this application. This collaboration combines the strengths of each discipline by applying modern data, graph and computational techniques to connect the basic sciences in pharmacy and medicine for drug repurposing, to search for potential COVID-19 treatment and provide an open platform to facilitate global collaborative research.

**Expertise in Pharmacology & Pharmacy and Big data**

The PC, Dr. Esther WY Chan is a pharmacist with expertise in medication safety and effectiveness. She has led completed multicentre randomised controlled trials in Hong Kong and Australia (ECS/RGC, NHMRC); and completed two GRF/RGC funded big data projects. She will lead and supervise the research team in all aspects of this study, including literature review, research methodology, interpretation of knowledge graph outputs in a clinical context, generation of research reports and dissemination of findings.

Co-PI, Prof. Ian CK Wong is an expert in using healthcare big data with a strong track record of leading large research programmes on medication use and safety, with success in drug repurposing for children. Together with the PC, he will refine methodology and contribute to interpretation of results and research dissemination.

**Expertise in Bioinformatics, Health informatics, Big graphs**

Co-PI, Dr Ruibang Luo, is experienced in health informatics and computational biology. His work on gene-disease associations and bioinformatic algorithms have been published in peer-reviewed journals including Nature family journals. His work involved application of deep-learning approaches to extract gene-disease associations from big data. He will advise on the use of big data information sources for building the knowledge graph and application of knowledge graph techniques for COVID-19 research and drug repurposing.

Co-PI, Dr. Chuan Wu is an expert in cloud computing and big data analytics systems / platforms. She has built large-scale distributed big data analytics/machine learning systems and published in top venues in the area. She has served as program committee members and review panels for leading database/system conferences and journals. She will assist in the implementation, storage and processing of large-scale knowledge graphs.

Co-PI, Prof. Reynold CK Cheng is an expert in data management and mining of uncertain data. He has led projects involving application of heterogeneous information networks to solve real world problems and developed efficient query algorithms for large-scale graph databases. He has served on program committees and review panels for leading database conferences and journals. He will provide database tools to build large-scale COVID-19 knowledge graph and apply computational techniques for drug repurposing and research using knowledge graphs.

**Expertise in Infectious diseases and Medicine**

Co-PI, Prof. Ivan FN Hung is a world-renowned expert in infectious diseases and a pioneer in COVID-19 drug treatment. He recently led a randomized clinical trial showing superiority of triple therapy of lopinavir/ritonavir, ribavirin and interferon beta-1b over lopinavir/ritonavir in COVID-19 patients. He will be responsible for assessing the clinical applicability of the drug candidates found from knowledge graphs and advise on the potential for selected candidates to be further validated in clinical trials.

Co-PI, Prof. David CW Siu is a leading cardiologist and stem cell biologist. His current work on COVID-19 include assessment of cardiovascular sequelae of COVID-19 patients, and disease modelling and drug screening using human induced pluripotent stem cell platform. His recent work demonstrated cytopathogenic effects of SARS-CoV-2 on human cardiomyocytes and its relationship with ACE2 expression. He will be responsible for the clinical interpretation of findings from the knowledge graph and advise on the testing of potential drug candidates using cell-based platforms.