COV-DrugX Pipeline for Covid-19 Drug Repurposing

Kamal Rawal[#1], Prashant Singh[1], Robin Sinha[1], Swarsat Kaushik Nath[1], Preeti P.[1], Priya Kumari[1], Sukriti Sahai[1], Ridhima[1], Sweety Dattatraya Shinde[1], Nikita Garg[1] , Trapti Sharma[1]

1. Amity Institute of Biotechnology, Amity University, Uttar Pradesh, India

#Corresponding Author
Email ID: kamal.rawal@gmail.com
Centre for Computational Biology and Bioinformatics, AIB
Amity University, Noida.

Keywords: Bioinformatics, drug repurposing, artificial intelligence, COVID-19

Supplementary Data Website: https://sites.google.com/view/drugx-supplementary

COV-DRUGX Software Pipeline : http://drugx.kamalrawal.in/drugx/

BACKGROUND: The outbreak of the novel coronavirus disease COVID-19, caused by the SARS-CoV-2 virus has killed over 5 million people to date. Despite the introduction of population-wide vaccination drives, countries such as Austria and Germany are witnessing the re-emergence of infections and deaths. Scientists, administrators and clinicians are scrambling to find solutions that include vaccines, and active therapeutic agents. So, there is an urgent requirement for new and effective medications that can treat the disease caused by SARS-CoV-2. Artificial intelligence (AI) enabled drug repurposing, has the potential to shorten the time and reduce the cost compared to de novo drug discovery.

Methods: 1.1 Datasets of Human Interactome, SARS-CoV-2, and Drug Targets.

We compiled datasets from over 50 public databases which include network data, therapeutic data, side effects data, drug targets, gene expression, clinical features, pathways data, and structure data. A detailed description of the datasets can be found in the supplementary website (**Supplementary Table 1**). For example, the human interactome was assembled from 21 public databases that compile experimentally derived protein-protein interactions (PPI) data. The final interactome used in our study contains 19000 proteins and 320000 interactions between them. We retrieved interactions between SARS-CoV-2 human proteins detected by Gordon et al, (**Gordon et al., 2020**), and drug–target information from the DrugBank database.

1.2 Preprocessing and integration of dataset for machine learning: We processed the extracted datasets from the heterogeneous dataset using a very large team of data curators to

make them ready for machine learning by removing the redundant information, formatting the data, etc.

1.3 Multimodal Platform: We have created a multi-modal system for drug repurposing for covid19 (**see Figure 1)**. The system consists of several standalone modules which rely on multiple streams of information, such as molecular profiles (**Dudley et al., 2011**), chemical structures (**Keiser et al., 2009**), adverse profiles (**Campillos et al., 2008**), molecular docking (**Dakshanamurthy et al., 2012**), electronic health records (**Paik et al., 2015**), pathway analysis (**Greene and Voight, 2016**), phenotype information (**Casas et al., 2019**), genome-wide association studies (**Cheng et al., 2018**), and network perturbations (**Cheng et al., 2019; Guney et al., 2016; Sadegh et al., 2020; Zhou et al., 2019; Zitnik et al., 2018; Zitnik et al., 2019**). We are describing information on one of the modules for the sake of brevity (see **Supplementary File - "Module Descriptions" on the website**).

1.3.1 Drug Feature Module (DF1 and DF2): The computation of drug ADME (absorption, distribution, metabolism, excretion) properties helps pharmaceutical companies to discard compounds that are not drug-like in the discovery phase (**Lin et al., 2003**). Module 1 (DF1) and module 2 (DF2) are deep learning (DL) based modules that compute 11 biological properties such as mutagenicity and drug-likeness, and 200 cheminformatics properties such as logP value and molecular weight of drug candidates. Each of the modules takes the input of the drug candidate in SMILES format and predicts whether the candidate could be repurposed against COVID-19.

In these modules, we first retrieved different drug molecules from various research papers which had shown some effect on SARS-CoV-2 or COVID19 having some evidence in terms of wet lab studies, animal studies or clinical trials (**Supplementary Table 2**). We labelled this dataset as "P". In addition, we compiled the equivalent of drugs from different sources as control. The rationale for selecting control datasets has been explained in **Supplementary Table 3**. Next, we compiled a list of drug properties from literature such as GI absorption, binding affinity, pIC50, AlogP, molecular weight, mutagenicity etc. (**Supplementary Table 4**). Over 200 properties were compiled (**Supplementary Table 5**). Using open-source tools such as RDKIT, we characterized all the drugs in class P as well as class N. We also surveyed the literature to find the importance of these properties and evidence from literature where these properties are studied for analysis (**Supplementary Table 6**). Thereafter, we evaluated the distribution of properties in positive and negative/control datasets. We used 636 drug examples for training the system (349 samples belong to the negative dataset and 287 examples constitute positive datasets as potential covid-19 drugs). For testing purposes, we used 191 examples (105 negative and 86 positive). Further, we scaled the datasets using Standard Scaler which subtracted mean from data and scaled to unit variance. We converted our smiles structures into one-hot encoded vectors for each component in that SMILES. We padded those smile structures which are shorter than the maximum size available in our

dataset (409) with all zero vectors. Thus, our final vector has dimensions (Number of Samples * 409 * 39 * 1), where 1 is required for the convolutional neural networks which represents the number of channels in image classification. These encoded SMILES were used to train convolutional neural networks along with 11 properties data to solve classification problems. Our model consists of 2 input layers in which one input layer takes up the one hot encoded SMILES and the second input layer takes up the 200 drug properties. Our model uses convolutional neural networks to identify patterns within the one hot encoded SMILES data and generates 128 elements long feature vector corresponding to that SMILE. Further, we combine SMILE data with 11 properties and pass it into 2 deep fully connected layers for further processing. Our output layer consists of a fully connected layer with 2 nodes representing two target classes (i.e. covid-19 drug or not a covid-19 drug). We trained our model using adam optimizer and sparse cross entropy loss for 1000 epochs in the batch of 32 samples per epoch. Next, along with these we also provided class weight to balance our data. The training was stopped early at 163 epoch (using the early stopping callback option). This feature enables the monitoring of the validation accuracy. The initial results are as following: Sensitivity: 0.99; Specificity: 0.98; Precision: 0.98; Recall - 0.99; Accuracy - 0.98.

RESULT

Here, we deployed a multi-modal pipeline relying on artificial intelligence, network based system and clinical information, tasking each of them to rank 9000 drugs for their expected efficacy against SARS-CoV-2. We used ground truth data consisting of over 2000 drugs which were screened using experimental evidence such as VeroE6 cells screening or clinical trials, to test our predictions. We find that a combined approach provides consistently reliable outcomes across all datasets and metrics when compared to a single algorithm. Hence, we developed a multimodal technology that fuses the predictions of all algorithms. As an outcome, we offer a list of important molecules which could be used for the treatment of COVID19. Further, we also generated a comprehensive resource base of molecular, pathway and clinical information related to COVID19. We believe that these advances offer a new AI-based platform to identify repurposable drugs for future pathogens and neglected diseases underserved by the costs and extended timeline of de novo drug development.

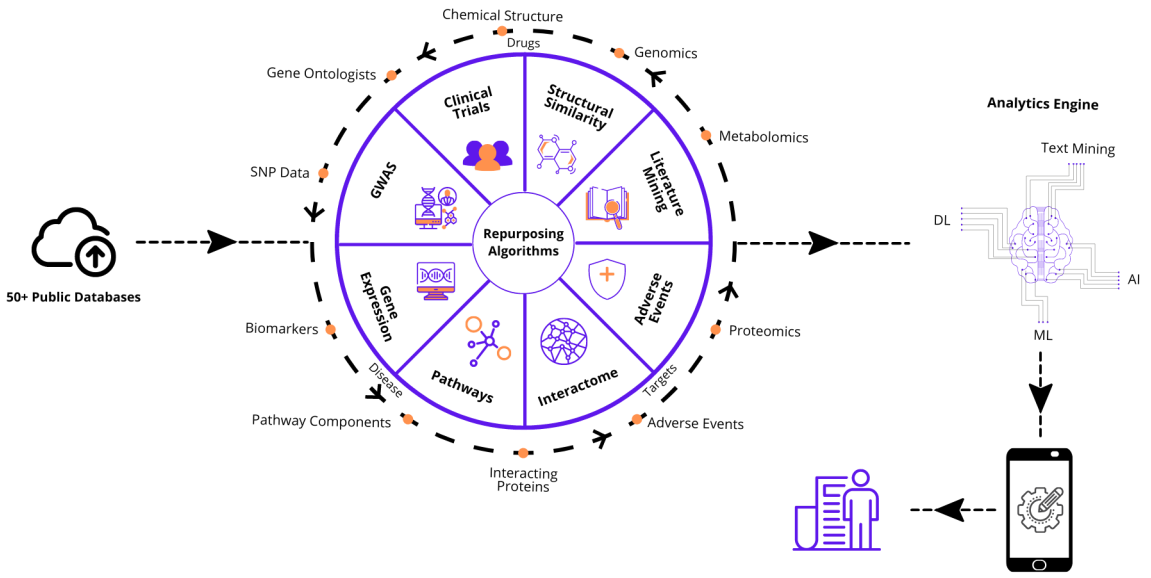

Figure 1: *Multimodal system for drug repurposing of Covid-19:* We compiled data from over 50 public databases for the implementation of various drug repurposing algorithms to compute biological/cheminformatics properties and medical manifestations of drugs such as adverse effects, pathways, gene expressions, etc. through the application of text mining, deep learning, and machine learning. The system is also capable of generating results in the form of custom visualisations and reports. The results could be further analysed by researchers.

References: See Supplementary Website: https://sites.google.com/view/drugx-supplementary

Contribution of Authors

This study was conducted under the overall guidance of KR, who contributed in protocol, critical evaluation of data and manuscript. The pipeline was designed, constructed and validated by RS and PS. Manuscript writing was done by SKN, PP, and PK. All the authors are responsible for the content of the manuscript.

Acknowledgement

We extend our sincere gratitude to Amity University for providing administrative and technical support required in the conduct of this study.

Financial Support and Sponsorship

Dr. Kamal Rawal acknowledges the support provided by SERB, Department of Science and Technology (Grant ID: CVD/2020/000842). The project involved usage of computational infrastructure (server etc) provided by the Department of Biotechnology (DBT), Ministry of Science and Technology Government of India (Grant ID: BT/PRI7252/BID/7/708/2016). SKN, PP, R, SS, SDS, NG, and TS have received financial support from grants obtained from Robert J. Kleberg Jr. and Helen C. Kleberg Foundation and Baylor College of Medicine, Houston, Texas, USA. We are also thankful to Amity University for the support provided during the conduct of this study.