scientific reports

OPEN

A computational approach to aid clinicians in selecting anti-viral drugs for COVID-19 trials

Aanchal Mongia[1], Sanjay Kr. Saha[2], Emilie Chouzenoux[3]✉ & Angshul Majumdar[1]✉

The year 2020 witnessed a heavy death toll due to COVID-19, calling for a global emergency. The continuous ongoing research and clinical trials paved the way for vaccines. But, the vaccine efficacy in the long run is still questionable due to the mutating coronavirus, which makes drug re-positioning a reasonable alternative. COVID-19 has hence fast-paced drug re-positioning for the treatment of COVID-19 and its symptoms. This work builds computational models using matrix completion techniques to predict drug-virus association for drug re-positioning. The aim is to assist clinicians with a tool for selecting prospective antiviral treatments. Since the virus is known to mutate fast, the tool is likely to help clinicians in selecting the right set of antivirals for the mutated isolate. The main contribution of this work is a manually curated database publicly shared, comprising of existing associations between viruses and their corresponding antivirals. The database gathers similarity information using the chemical structure of drugs and the genomic structure of viruses. Along with this database, we make available a set of state-of-the-art computational drug re-positioning tools based on matrix completion. The tools are first analysed on a standard set of experimental protocols for drug target interactions. The best performing ones are applied for the task of re-positioning antivirals for COVID-19. These tools select six drugs out of which four are currently under various stages of trial, namely Remdesivir (as a cure), Ribavarin (in combination with others for cure), Umifenovir (as a prophylactic and cure) and Sofosbuvir (as a cure). Another unanimous prediction is Tenofovir alafenamide, which is a novel Tenofovir prodrug developed in order to improve renal safety when compared to its original counterpart (older version) Tenofovir disoproxil. Both are under trail, the former as a cure and the latter as a prophylactic. These results establish that the computational methods are in sync with the state-of-practice. We also demonstrate how the drugs to be used against the virus would vary as SARS-Cov-2 mutates over time by predicting the drugs for the mutated strains, suggesting the importance of such a tool in drug prediction. We believe this work would open up possibilities for applying machine learning models to clinical research for drug-virus association prediction and other similar biological problems.

There has been an exponential rise in the total active cases and deaths due to COVID-19 (COrona VIrus Disease-2019) since the first case in Wuhan, China in December, 2019[1]. The disease results in severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2), which is known to be highly transmittable and has spread across more than 100 countries. This pandemic has wreaked havoc on people's social life, the global economy, and most importantly the health of the human race. The death numbers are frightening, confirming about 467 K deaths worldwide till mid-June, 2020[1].

As medical professionals are striving to save lives, research scientists specialized in drug development, are racing against time to develop a vaccine against SARS-CoV-2[2]. The investigation involved for developing a vaccine (or even a new drug) is time consuming, requiring several phases of extensive trials. Experts believe that it is highly unlikely that a vaccine will be ready before a year or more. In such circumstances the best bet may be to re-position existing drugs for treating COVID-19. This is a well known approach where existing drugs (which have already been approved for release in the market) are investigated for new diseases[3]. Drug re-positioning is usually cost effective and fast (compared to developing a new drug/vaccine) since its effects are well studied. One classic example for drug re-positioning is Chlorocyclizine , which was initially developed as an anti-allergic but later found to act against the hepatitis C virus[4]. Another example is Imatinib mesylate (sold under the trade

[1]Department of CSE, IIIT-Delhi, New Delhi 110020, India. [2]Department of Community Medicine, IPGMER Kolkata, Kolkata, India. [3]CVN, Inria Saclay, University of Paris Saclay, 91190 Gif-sur-Yvette, France. ✉email: emilie.chouzenoux@centralesupelec.fr; angshul@iiitd.ac.in

name Gleevec), it was originally used as a treatment for leukemia but later was found to be effective against genes associated with gastrointestinal-stromal tumors[5,6].

Given the relatively large drug-virus association space, manual investigation in wet-labs is not a scalable strategy. Putting all the anti-virals in trials for treating corona is not very feasible either; especially because time is of essence. In such a scenario, computational approaches can help; they can be used to prune down the search space for the drugs to be investigated[7]. Practically, such approaches could also assist the clinicians to come up with treatments for rapidly mutating viruses by pruning the anti-viral drug space (see "Discussion" section). Specifically, a computational approach which takes into account the genomic structure of the latest viral isolate or its similarity with the previously occuring strains of viruses would be helpful in deciding the treatment. With this objective, we have manually curated a comprehensive database called DVA (Drug Virus Association), having the approved (anti-viral) drug-virus associations in the literature along with the similarity measures associated with drugs (chemical structure similarity) and viruses (genome sequence similarity). To the best of our knowledge there is no existing database for drug virus association. There are several recent studies such as[8–11] that have used docking for finding a treatment for COVID-19 infection. Our work however contrasts with the docking approach. While docking is based on biological simulation, our approach is based on machine learning.

The DVA database we propose in this work lies the foundation for further computational studies on this topic. There can be various methodologies to predict drug virus association. The prediction problem can be approached via feature-based classification models, neighborhood models, matrix completion models, network diffusion models etc. A recent empirical study on well established drug-target interaction databases exhibit the best prediction performance by matrix completion models[7]. In computer science, matrix completion is used routinely in recommendation systems . The general problem of drug-disease association can actually be thought of as a recommendation system, where drugs are being recommended for treating a disease. Given the success of matrix completion techniques in drug target interaction, we deploy state-of-the-art matrix completion techniques on our curated DVA database. We perform a thorough comparative analysis of those for predicting assessed drug-disease associations. Then, we apply the methodology for pruning the search space of potential candidates for COVID-19 trial drugs. Finally, we show how the tool helps in selecting drugs as the virus mutates.

Our objective is to make our solution user friendly for clinicians and scientists. In pursuit of this goal, we have made the solution (dataset and algorithms) available as a webserver. The webserver can be used in two ways. First, given the genome of the virus, the webserver can predict (and rank) the probable antivirals. Second, given a drug and a virus, it can output a normalized score depicting how effective the drug will be against the virus. The functionalities and usage of the webserver have been described in Supplementary sect. 3 (Supplementary Figs. 1–7).

Results

We assess the performance of different matrix completion techniques in this section. The techniques have been described in the "Methods" section. Six matrix completion methods were used, which can be categorized into three families provided below.

- Basic frameworks (MF: Matrix factorization[12] and MC: Matrix completion or Nuclear norm minimization[12]).
- Deep frameworks (DMF: Deep matrix factorization)[13].
- Graph regularized frameworks (GRMF: graph regularized matrix factorization[14], GRMC: graph regularized matrix completion[15], GRBMC: graph regularized binary matrix completion[16]).

Matrix factorization (MF) is the traditional matrix completion method which factorizes the data matrix into two latent factor matrices (tall and short) and the algorithm recovers these factor matrices to recover the original matrix. Since this problem is non-convex, it may not converge to a global minimum of the cost function. Nuclear-norm minimization based matrix completion (MC) was proposed as a (mathematically) better alternative; it directly recovers the matrix by penalising its nuclear norm (convex surrogate of rank). Deep matrix factorization (DMF) generalises MF to more than two factors. None of the techniques mentioned so far can take advantage of genomic structure of the viruses or chemical structure of the drugs. The said pieces of information can be incorporated into the graph regularized matrix completion techniques (GRMF, GRMC, GRBMC). These techniques have been explained in detail in the "Methods" section.

Overview: DVA prediction.

The typical anti-viral drug discovery process involves genomic and biophysical understanding of the virus. It aims to target the enzymes or peptides involved in the viral replication cycle and takes years for successful clinical validation. Other approaches involve screening all the broad-spectrum anti-viral drugs or chemical libraries comprising large numbers of existing compounds/databases (having information on transcriptional signatures in different cell lines) to be further evaluated by standard anti-viral assays[17]. In view to assist acceleration of this process (by pruning down the search space), we create and share a publicly available DVA database, along with a number of matrix completion techniques (mentioned above) for drug-virus association prediction.

The originality of the proposed work lies in the formalization of the drug-virus association prediction as a matrix completion problem, without the need for any anti-viral assays. Such a computational approach requires the chemical structure of the drugs and, in case of graph-regularized matrix completion techniques, the genome of the viruses, or existing associations otherwise. Figure 1 depicts the schematic flow of the proposed work involving data curation and implementation overview explicitly showing the input to the algorithm (DVA matric and the drug/virus similarity matrices) and the predicted output.

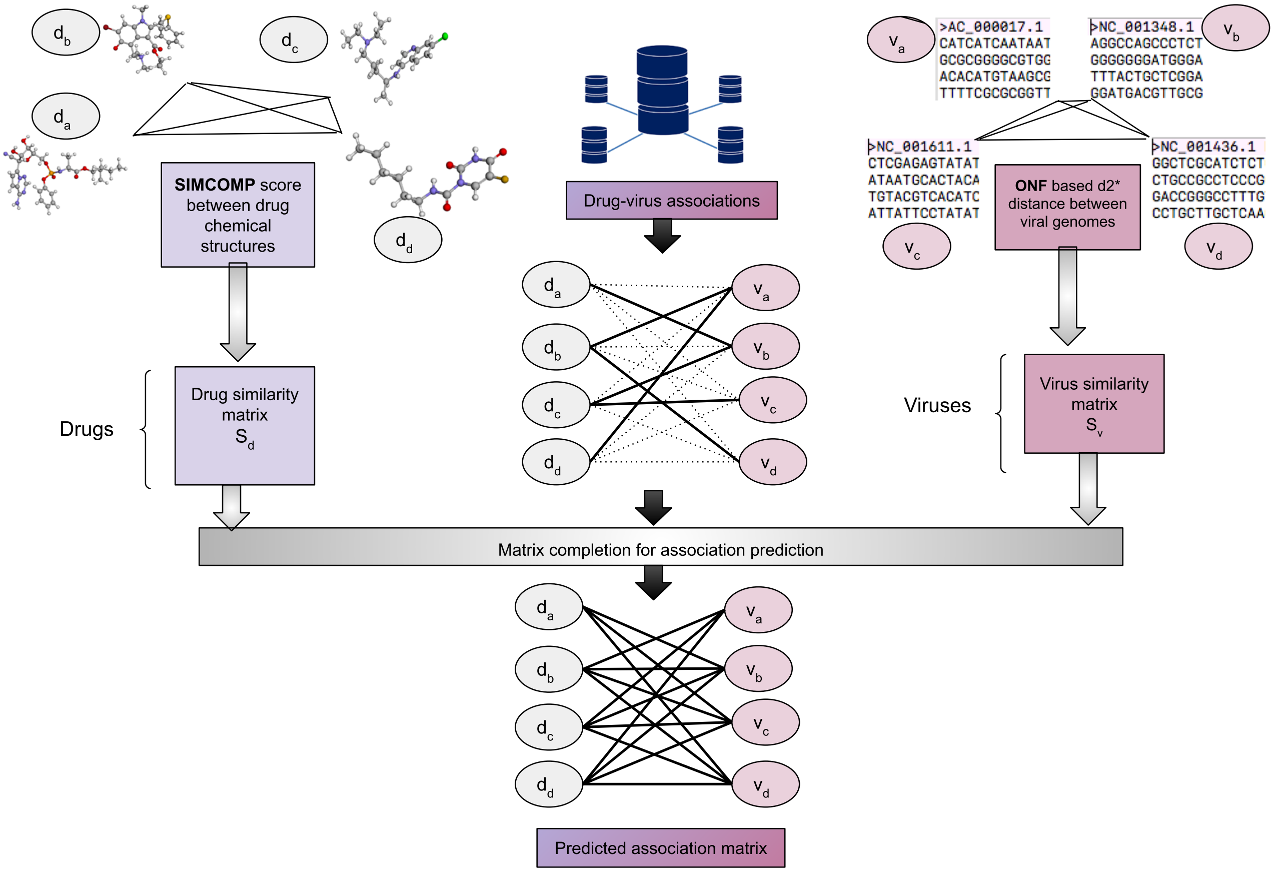


Figure 1. Schematic diagram depicting the DVA framework.

	Metric	MC	MF	DMF	GRMF	GRMC	GRBMC
CV1	AUC	0.5959	0.6753	0.6974	0.8652	0.8279	0.8834
	AUPR	0.3238	0.2656	0.2615	0.4812	0.4445	0.5220
CV2	AUC	0.4909	0.5033	0.5704	0.7346	0.6705	0.6632
	AUPR	0.1106	0.0504	0.0855	0.3112	0.2951	0.2746
CV3	AUC	0.5438	0.5215	0.4529	0.7806	0.7507	0.8181
	AUPR	0.0538	0.0637	0.0824	0.4265	0.4333	0.4383

Table 1. Results for association prediction for all techniques under the 3 cross validation settings.

Empirical evaluation. In this sub-section, we carry out extensive experimental protocol to illustrate and compare the ability of the different methods to retrieve the drug-disease associations available in our curated dataset. The protocol dictates three variants of 10-fold cross-validation setting (CV). In the first setting CV1 (cross validation 1), 10% of the associations selected at random are left out as testing set. This allows to assess each algorithm's ability to predict associations between existing drugs and viruses. To evaluate an algorithm for its ability to predict association for novel drugs and viruses i.e. those which have no association information, we use two other (more stringent) CV settings. In CV2 and CV3, 10% of the complete virus and drug entities selected at random are left out as test set respectively.

The standard metrics for evaluation are the AUC (Area under the Receiver Operating Characteristic curve) and AUPR (Area under the precision-recall curve). AUC is more common in machine learning literature, it assumes that the classes are evenly balanced. Problems in drug-disease association have highly imbalanced classes, in such a scenario the AUPR is a more appropriate metric for evaluation[14,18].

Table 1 shows how each of the 6 tested algorithms performs in retrieving the associations. A clear observation from the experiments is that the graph regularized-based matrix completion algorithms that incorporate the similarity information associated with the drugs and viruses, perform fairly well giving an AUC greater or equal than 0.83 in CV1. The best performing algorithm (GRBMC) exhibits an AUC and AUPR of 0.88 and 0.54 respectively. Predicting the associations for novel drugs and viruses also have a reasonable performance with

	MC	MF	DMF	GRMF	GRMC	GRBMC
# drugs with MPV = 1	2	4	4	26	22	8
% drugs with MPV = 1	2.6316	5.2632	5.2632	34.2105	28.9474	10.5263

Table 2. Number and percentage of drugs predicted with MPV = 1 by the matrix completion methods.

the best AUC/AUPR of 0.81/0.44 and 0.73/0.31 by GRBMC and GRMF. It can be noted that the standard matrix completion methods, which do not take into account the metadata, fail to learn from the association data giving a near-random performance as far as the prediction on novel viruses is concerned, depicting how very important the similarity information is.

Association prediction for new drugs. DVA database and its associated computational tools can also be used on new drugs without any previously known virus association information. For evaluating this ability, we identified in our database all the drugs which are known to interact with only one virus (drugs associated with a single virus only) and hide that association to the methods. This allows us to assess the performance of the algorithms in predicting viruses associated with the new drugs in the database.

We hide the only virus corresponding to each of the 76 drugs (with only a single virus associated with it) and run matrix completion to predict candidate viruses for these drugs. The drugs for which the test virus associated with it is the top-ranked virus predicted by the algorithm would have the maximum precision value (MPV) of 1. The number and percentage of drugs with a maximum precision value of 1 are reported in Table 2.

Nearly 34 % (26/76) of single association drugs with a maximum precision of 1 were predicted using GRMF. Other graph regularized frameworks show comparable performance in terms of predicting drugs with MPV of 1.

SARS-CoV-2 prediction. In this experiment, we add the SARS-CoV-2 sample in our database by providing its ONF based d2* similarity[19] in the virus similarity matrix.

We then apply the matrix completion algorithms to predict the associations and rank prediction scores corresponding to SARS-CoV-2 to predict the top ten recommended drugs.

As can be seen from the results of "Empirical evaluation" (Table 1), MC, MF and DMF often yield considerably worse results than their graph regularized counterparts (GRMF, GRMC and GRBMC). Such poor performance of non-graph regularized versions of matrix completion methods could be explained as they do not incorporate any knowledge about the genomic structure of the viruses and the chemical structure of the drugs. Since the three graph-based methods perform reasonably well in the prediction task, we consider these techniques for the drug prediction on the novel coronavirus. The top-10 drugs they predicted have been reported in Table 3 (ranked by their predicted scores). Drugs highlighted with blue text are unanimously predicted drugs by the three considered matrix completion techniques and those in red text are predicted by two methods. We also highlighted with yellow cells the drugs which are under trial/investigation as a potential cure/prophylactic against COVID-19.

It can be seen that the three techniques have consistently and unanimously selected six drugs, namely Remdesivir, Ribavarin, Sofosbuvir, Taribavirin, Tenofovir alafenamide and Vidarabine. Umifenovir has been recommended by two (GRMF and GRBMC) out of three techniques. Amongst these recommendations, Remdesivir[20], Ribavarin[21,22], Sofosbuvir[23] and Umifenovir[24] are under clinical trials. Taribavirin is similar to Ribavirin but it is not approved by the FDA. Tenofovir alafemanide (an antiretroviral for HIV-1) is on undergoing trial[25]. GRMF has additionally selected Ibuprofen which is expected to be investigated in UK[26,27]. The fact that three techniques unanimously select the aforementioned drugs make us confident about these recommendation results.

Predictions evolution with mutating novel coronavirus. In the previous sub-section, we have established that the results from our models are mostly in sync with clinical practice. In this sub-section, we will demonstrate how our proposed approach can be of help to clinicians.

All the results generated so far have been generated using the reference sequence of the SARS-Cov-2 strain (collected in December, 2019 in Wuhan). The novel coronavirus is rapidly mutating[28]. In such a scenario, it is necessary to select drugs that are effective against the mutated strain. While mutating, the virus isolates may develop resistance to previous drugs used for its treatment. Our model may be of help to clinicians in this respect. Before proposing a treatment regime (trial, for e.g.) for COVID-19 treatment, the practitioner may use our approach to check the drugs selected for the particular isolate of novel coronavirus. In Table 4, we have experimented with three isolates of the novel coronavirus (collected over an interval of 2 months), in addition to the reference sequence (collected in December 2019). Those three isolates have been collected in February (from USA), April (from Australia) and June (from India).

One can note from the Table 4 that the selected drugs change with mutations. Baloxavir marboxil was not selected even once for the reference sequence from December 2019, but has been selected by two methods for the February isolate. A recent pre-print[29] reports the results of this antiviral on COVID-19 patients. The drug Ibuprofen, was selected by one of the methods for the December reference sequence, it was not selected for the February isolate, then it was selected by two methods for the April isolate and selected by all three for the June isolate. It may be worthy to note that lipid Ibuprofen is being considered in a trial in UK from starting June, 2020[26]. Similarly, Pleconaril has been selected for by all three methods for the most recent (June) isolate, it was selected by only one of the techniques for the reference sequence (December) and was not selected for the February or June sequences. Pleconaril, although developed for treating enterovirus and rhinovirus, is not

Technique	SARS-Cov-2
GRMF	Remdesivir
	Ribavirin
	Sofosbuvir
	Umifenovir
	Taribavirin
	Tenofovir alafenamide
	Ibuprofen
	Pleconaril
	Geldanamycin
	Vidarabine
GRMC	Remdesivir
	Ribavirin
	Sofosbuvir
	Taribavirin
	Tenofovir alafenamide
	Vidarabine
	Telaprevir
	Boceprevir
	Simeprevir
	Palivizumab
GRBMC	Remdesivir
	Ribavirin
	Sofosbuvir
	Umifenovir
	Taribavirin
	Vidarabine
	Brivudine
	Tenofovir alafenamide
	Paritaprevir
	Peginterferon alfacon-1

Table 3. Top-10 drugs predicted for SARS-Cov-2 by the DVA computational methods.

FDA approved. Rilpivirine and Etravirine are two antiretrovirals developed for treating HIV positive subjects. Both of them have been predicted by all three methods in the latest isolate, but not in the previous isolates or in the reference sequence. To the best of our knowledge, this antiretroviral is not under study for COVID-19 trials. Note that Vidarabine, which was getting predicted for the reference sequence (albeit wrongly) has not been predicted from the later ones. Based on this discussion, we can see that how the mutations in genomic structure results in different predictions of drugs. Since the novel coronavirus is mutating, it may be judicious to account for the structure of the latest isolate while deciding the treatments to be put in trial. In such a case, our model may be of help to clinicians.

Execution time.

We recorded the time taken by each of the matrix completion algorithms for a single run (Table 5), on a single core machine with a clock speed of 2.8 GHz, 64 GB RAM (Intel(R) Xeon(R) CPU E5-1603 v3 processor). All the methods have relatively low computational requirements. Matrix factorization methods are faster than the nuclear norm minimization based techniques, with a difference of few seconds. Such difference may not be practically significant, given the nature of our problem as an improved anti-viral prediction in pandemic is much more crucial than the running time in the order of seconds.

Discussion

We have collected a comprehensive dataset comprising of all the anti-viral drugs which act against viruses known to infect humans, along with the similarity information associated with the drugs and the viruses (see "Methods" section). On this database, we deploy state-of-the-art drug target interaction techniques based on matrix completion.

General discussion.

The drug-virus associations and the similarity information are assembled as three matrices: drug-virus association matrix (Y), drug similarity matrix (S_d) and virus similarity matrix (S_v). Several matrix completion methods have then been implemented and compared. The matrix completion methods

Technique	SARS-Cov-2: February	SARS-Cov-2: April	SARS-Cov-2: June
GRMF	Remdesivir	Remdesivir	Remdesivir
	Ribavirin	Sofosbuvir	Umifenovir
	Umifenovir	Umifenovir	Pleconaril
	Taribavirin	Ribavirin	Ibuprofen
	Sofosbuvir	Tenofovir alafenamide	Sofosbuvir
	Baloxavir marboxil	Ibuprofen	Rilpivirine
	Geldanamycin	Pleconaril	Etravirine
	Tenofovir alafenamide	Hydroxychloroquine	Tenofovir alafenamide
	Tecovirimat	Valomaciclovir	Rimantadine
	Peramivir	Dexamethasone	Ribavirin
GRMC	Remdesivir	Remdesivir	Umifenovir
	Umifenovir	Sofosbuvir	Remdesivir
	Ribavirin	Tenofovir alafenamide	Ibuprofen
	Taribavirin	Boceprevir	Pleconaril
	Sofosbuvir	Telaprevir	Sofosbuvir
	Vidarabine	Palivizumab	Chloroquine
	Tenofovir alafenamide	Simeprevir	Etravirine
	Nelfinavir	Ribavirin	Rilpivirine
	Amprenavir	Umifenovir	Tenofovir alafenamide
	Boceprevir	Ibuprofen	Nelfinavir
GRBMC	Remdesivir	Remdesivir	Umifenovir
	Ribavirin	Umifenovir	Remdesivir
	Umifenovir	Sofosbuvir	Pleconaril
	Taribavirin	Ribavirin	Ibuprofen
	Sofosbuvir	Taribavirin	Sofosbuvir
	Paritaprevir	Paritaprevir	Rilpivirine
	Tenofovir alafenamide	Brivudine	Etravirine
	Atazanavir	Vidarabine	Ribavirin
	Baloxavir marboxil	Daclatasvir	Tenofovir alafenamide
	Favipiravir	Beclabuvir	Trifluridine

Table 4. Top-10 drugs predicted for three isolates of SARS-Cov-2 (collected at an interval of 2 months) by the DVA computational methods.

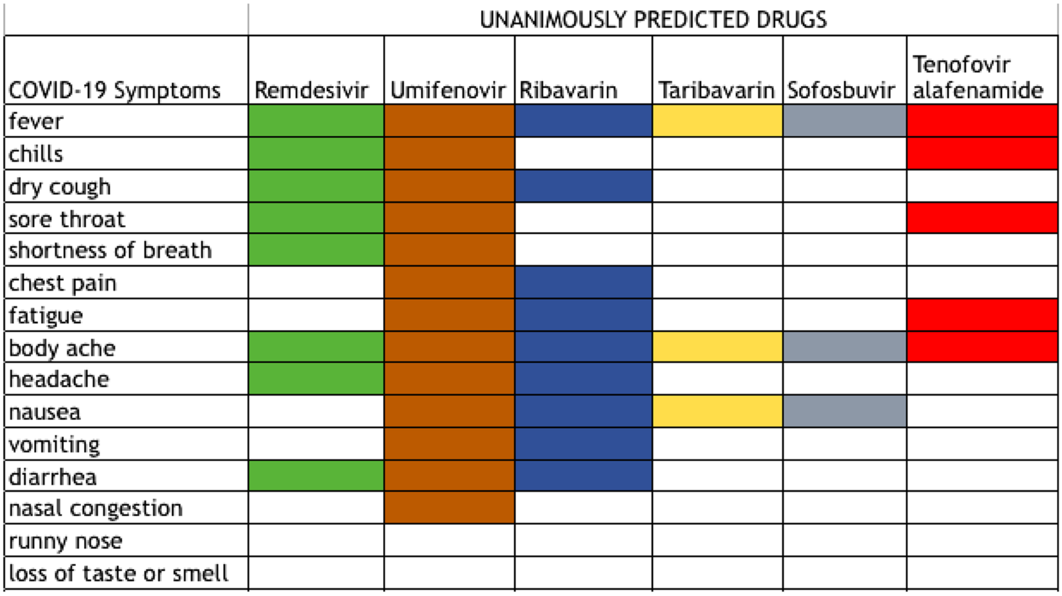
	MC	MF	DMF	GRMF	GRMC	GRBMC
Time (s)	0.0859	0.0149	0.0529	0.0457	10.55	5.22

Table 5. Running time of the DVA computational methods.

which are not designed to incorporate the similarity information take association matrix as input (assuming it to be a partially filled matrix from which the full low-rank association matrix would be recovered) along with the masking operator which stores information on the position of train and test indices. On the other hand, the graph regularized frameworks utilize similarity information and give an improved prediction performance in the cross-validation evaluation (Best AUC = 0.88, AUPR = 0.52). The graph regularized matrix completion methods are not only capable of predicting associations between existing drugs and viruses but can also take into account novel viruses and drugs for which no association information is known (as can be seen in the latter two cross-validation settings). The similarity information for such novel drugs/viruses (S_d and S_v) can be added to the metadata using the chemical structure and sequence information of the drug and virus respectively. The validity of the proposed pipeline is illustrated by the fact that 4 out of the 6 drugs unanimously predicted in top-10 prediction by the graph regularized methods are already under trial for treating SARS-Cov-2.

Drug recommendations for the novel coronavirus.

From the prediction provided by graph regularized methods, we observe a consensus over the recommendation of six drugs, namely Remdesivir, Ribavarin, Sofosbuvir, Taribavirin, Tenofovir alafenamide and Vidarabine (7th common drug being Umifenovir recommended by two models). Note that Umifenovir is under investigation as per FDA. It has been approved in countries like Russia, India and China for treating COVID-19. Remdesivir has obtained approval for emergency use by FDA. It is also been approved for treating moderate to severe COVID-19 patients in India, Russia, China



	UNANIMOUSLY PREDICTED DRUGS					
COVID-19 Symptoms	Remdesivir	Umifenovir	Ribavarin	Taribavarin	Sofosbuvir	Tenofovir alafenamide
fever	🟩	🟧	🟦	🟨	▨	🟥
chills	🟩	🟧				🟥
dry cough	🟩	🟧	🟦			
sore throat	🟩	🟧				🟥
shortness of breath	🟩	🟧				
chest pain		🟧	🟦			
fatigue		🟧	🟦			🟥
body ache	🟩	🟧	🟦	🟨	▨	🟥
headache	🟩	🟧	🟦			
nausea		🟧	🟦	🟨	▨	
vomiting		🟧	🟦			
diarrhea	🟩	🟧	🟦			
nasal congestion		🟧				
runny nose						
loss of taste or smell						

Figure 2. List of COVID-19 symptoms treated by drugs unanimously predicted by the three graph-regularized matrix completion methods.

and others. Researchers working on Ribavarin trials argue that since it is an established drug with ready availability and established supply chains, it is worthy to investigate its potency against COVID-19. Ribavirin[21,22], in combination with other antiviral drugs has recently been studied for the effectiveness and safety of different antiviral regimens (combination therapies) for the treatment of COVID-19. More recent studies on the use of this antiviral showed encouraging results in both mild and severe COVID-19 infections[30–32]. Sofosbuvir is used specifically for hepatitis C infection. Currently it is under trial for treating COVID-19 patients. This is because, superposition of the hepatitis C virus polymerase bound to sofosbuvir, with the SARS-CoV polymerase shows that the residues that bind to the drug are present in the latter[23]. There are many recent studies that have shown the efficacy of this drug in combination with others for treating moderate to severe COVID-19 patients[33–35]. The clinical trial of Tenofovir disoproxil as a prophylactic, is based on recent albeit sparse literature that shows that RNA synthesis nucleos(t)ide analogue inhibitors, acting as viral RNA chain terminators, like Tenofovir disoproxil, abacavir or lamivudine, amongst others, could have an effect against SARS-CoV-2[36]. Our algorithm selects Tenofovir alafenamide, which is less harmful to the kidneys than Tenofovir disoproxil. Tenofovir alafenamide is known to have large antiviral efficacy at ten times lesser dose than Tenofovir disoproxil. It is also under investigation as a combination therapy (emtricitabine/tenofovir-alafenamide and lopinavir/ritonavir) to treat COVID-19 patients[25]. Some variants of Tenofovir have also shown encouraging results in the treatment of COVID-19 patients[37,38].

Symptomatic treatment options from unanimously predicted drugs.

In this sub-section, our objective is to establish that the drugs selected by our algorithms for SARS-Cov-2 are clinically sensible predictions, in the sense that they are known to be effective against a significant number of the COVID-19 symptoms. The Table 3 shows that out of the top ten selections, six are common across all the three techniques and another (Umifenovir) have been predicted by two out of three. Although Vidarabine has been selected by all three models, it is an antiviral effective against DNA viruses. Since the novel coronavirus is an RNA virus this antiviral is not supposed to work. We have discussed the rest of the six antivirals in Supplementary sect. 1. The descriptions have been primarily taken from drugbank.ca.

Symptoms of different virus infections have been discussed in Supplemetary sect. 2. From a symptomatic aspect, Umifenovir, Remdesivir and Ribavarin covers most of the symptoms for COVID-19, as can be seen from Fig. 2. All of them are now under clinical trials as discussed in the previous subsection. Figure 2, hence establishes the efficacy of the model in predicting clinically sensible drugs which work against most of the symptoms of SARS-Cov-2.

Effect of viral mutations on treatments.

Furthermore, significance of such computational models for predicting anti-viral therapeutics would correlate with the rate at which a virus mutates. RNA viruses (like HIV, flu virus) are known to mutate at a much faster rate than the DNA viruses[39], helping them to evade the human immune system and develop drug resistance. SARS-CoV-2 is no exception and has been mutating over the past few months[40]. Clinically keeping up with the evolving viruses and drug resistance could be a major challenge for development of an anti-viral treatment[41]. Hence, artificial intelligence, and in particular the presented matrix completion techniques, could help the clinicians to prune down the drug space for viral strains which have been mutating rapidly and avoid unnecessary testing on drugs for the new viral strain/s.

Conclusion

Computational techniques have the inherent advantage of learning from the data (which can be huge) and scale to a large number of drugs and viruses and hence be of immense importance to the clinicians by narrowing down the search space for the clinical trials to be carried out. Through this work, we not only show how such techniques can be leveraged to infer the drugs which could act against a viral infection using the novel coronavirus as a use case but also we show how the predicted drugs (Remdesevir, Ribavirin, Umifenovir and Sofosbuvir here predicted for SARS-Cov-2) would vary as the virus mutates.

We would like to emphasize that the proposed DVA database and methods are not particular to the novel coronavirus. Such computational approaches have the general capability to help for identification of drugs which might be effective against a broad spectrum of viruses[42], or the viruses which can be targeted by multiple drugs (since many drugs could target specific elements of viral replication)[43]. We believe that the proposed work will pave the way for more scientific ideas for anti-viral drug re-positioning and assist clinicians in the process.

A limitation of our approach is that it gives a pointwise prediction but, as it is not probabilistic, it cannot give any quantitative measure of uncertainty about the prediction. However, the confidence of the models on the prediction can be assessed qualitatively by inspecting the averaged ranking of the predicted drugs, over random initializations (for example). The higher the averaged rank, the more confident is the model in the efficacy of the drug.

Methods

Drug-virus associations database.

The proposed DVA dataset aims at being exhaustive. It compiles various existing sources, housing together all the anti-viral drugs proved clinically to be effective against viruses infecting humans. We believe such resource would be highly useful for analysing and proposing anti-virals not only for the novel coronaviruses but other viruses too. Along with that, it may also be used to computationally identify viruses that a newly discovered drug may target. The associated metadata (information about the drugs and viruses) may also help clinicians in manual analysis and having a deeper insight.

All the associations corresponding to anti-viral drugs clinically shown to act against human host viruses have been assembled from standard DrugBank database[44] (https://www.drugbank.ca/categories/DBCAT000066). To ensure that the database is fully comprehensive, other literature works[45–53] and resources such as ViPR[54] were also scanned for any additional drug-viral associations. ViPR or NIAID Virus Pathogen Database and Analysis Resource (http://www.viprbrc.org/) is a repository of data and analysis tools for virology research[54] capturing various types of information derived from comparative genomics analysis and visualization tools. It has antiviral drug information (for 21 viral species) derived imported from DrugBank (https://www.drugbank.ca/)[44].

The drug-virus indications have been stored (see "Supplementary data section") and processed in a matrix form of size $m \times n$ (m being no of drugs in the database and n being the number of viruses) to be used as input for any of the 6 matrix completion algorithms we made available in our repository.

The DrugBank Identifier (DrugBank ID) of the anti-viral drugs involved is considered as the unique key for the drugs, obtained from DrugBank vocabulary (https://www.drugbank.ca/releases/latest#open-data). Along with the viral association information, we also store the target pathway and mechanism of action of each drug for quick reference in any further investigation. Apart from this, each drug is mapped to its corresponding KEGG Identifier (KEGG ID) from the KEGG Compound/KEGG Drug database (https://www.genome.jp/kegg/drug/, https://www.genome.jp/kegg/compound/) of the KEGG (Kyoto encyclopedia of genes and genomes)[55]. The KEGG IDs were taken from the linking file provided at https://www.drugbank.ca/releases/latest#external-links[44] or manually added in the case of drugs missing in the linking file.

Each virus is identified by an acronym assigned to it (in case of no acronym, full virus name is used). The viral family, genome type, transmission route and incubation period is also available in the virus metadata file along with the accession number of the complete genomic sequence of the viruses fetched from NCBI (National Center for Biotechnology Information) Viral genome browser https://www.ncbi.nlm.nih.gov/genomes/GenomesGroup.cgi[56]).

Similarity computation.

To integrate the similarity information to the drug-virus associations, we have computed similarities between the drugs based on their chemical structures and between the viruses using their complete genomic sequences.

- DRUG SIMILARITY: All the DrugBank IDs were mapped to KEGG IDs of the corresponding drug/compound in the KEGG database[55]. The chemical structure similarity was measured between the drugs by computing the SIMCOMP score[57] based on the maximum common substructures between the chemical structure of the compounds using the KEGG API page at GenomeNet (https://www.genome.jp/tools/gn_tools_api.html). The drugs for which the SIMCOMP score was less than the set cutoff (0.001) and the drugs with no KEGG IDs available were assigned a similarity score of 1 to themselves and 0 to other drugs in the dataset.
- VIRUS SIMILARITY: The $d2^*$ distance based on ONF (Oligonucleotide frequency) measure between the DNA sequences was shown to be the best amongst various other ONF metrics with several k-mers length in host prediction accuracy at the genus level[19]. Hence, we compute d2* dissimilarity/distance (at k=6) between the viral genome sequences obtained from NCBI[56]. The reference sequences of viruses were saved in FASTA format to be used by the distance computation software (https://github.com/jessieren/VirHostMatcher) proposed by[19]. The $d2^*$ distance was subtracted from 1 to obtain the similarity measure. For the viruses with

segmented structure (Influenza A virus, Influenza B virus, Influenza C virus, Lassa mammarenavirus), the coding sequence in the nucleotide sequence of each genomic segment (taken in decreasing order of length was taken) was combined to form the complete viral sequence.

Proposed method: matrix completion. In this subsection, we describe each of the matrix completion algorithms used (www.github.com/AanchalMongia/DVA), along with their mathematical formulations and resolution strategies.

Let $X_{m\times n}$ be the complete drug-virus association matrix (with m drugs and n viruses) with binary entries (1 denoting that the drug is known to act against the virus and 0 denoting no association). Here X is the matrix to be recovered from its sampled (partially known) entries in Y. Let M denote the masking operator (elementwise multiplied to X) having 1's at positions where associations are known and 0 otherwise. Then, the matrix completion problem can be formulated as searching for X satisfying:

$$Y = M(X), \tag{1}$$

under specific constraints. In particular, it is typically assumed that similar drugs act in a similar manner, hence X to be recovered (from Y and M) is of low-rank.

Matrix factorization (MF). The most straightforward technique of solving low-rank matrix completion is matrix factorization, where the data matrix $X_{m\times n}$ is decomposed into two latent factor matrices $U_{m\times k}$ and $V_{k\times n}$, where k denotes the number of latent (hidden) factors deciding if a drug is associated with a virus or not. X is recovered by solving for U and V in the following minimization problem:

$$\min_{U,V} ||Y - M(UV)||_F^2. \tag{2}$$

The above problem is solved in an alternating manner, by first decoupling the mask using a majorization-minimization technique[58,59] and then using alternating least squares method[60] to obtain U and V. The complete algorithm is described in[12].

Deep matrix factorization (DMF). An extension of matrix factorization has been proposed motivated by the success of deep dictionary learning[61], where the data matrix X is decomposed into multiple factor matrices (analog to multiple layers) to capture more complex hidden features in the data. The formulation of the minimization problem in the case of 2-layer matrix factorization is given below:

$$\min_{U_1,U_2,V} ||Y - M(U_1U_2V)||_F^2 \text{ such that } U_1 \geq 0, U_2 \geq 0. \tag{3}$$

The above problem is solve alternatively. The minimization with respect to variables U_1 and V, is done in a similar way to that of matrix factorization, while the update on U_2 can be obtained as shown in[62].

Graph regularized matrix factorization (GRMF). Another variant of Matrix factorization has been proposed to incorporate metadata associated with the row and column entities (drug and virus similarities in this case)[14]. Here, the drug and virus entities form the nodes of two separate graphs and the similarity between them is assumed to be the weights between the nodes. Regularization is imposed by adding graph Laplacian penalties to the cost function of matrix factorization as shown below:

$$\min_{U,V} ||Y - M(UV)||_F^2 + \mu_1 \mathrm{tr}(U^\top L_d U) + \mu_2 \mathrm{tr}(V L_v V^\top), \tag{4}$$

where $\mu_1 > 0$ and $\mu_2 > 0$ are coefficients penalizing the graph regularization Laplacian terms and tr denotes the trace of the matrix. $L_d = D_d - S_d$ and $L_v = D_v - S_v$ are the graph Laplacians[63] for S_d (row/drug similarity matrix) and S_v (column/virus similarity matrix), respectively, and $D_d^{ii} = \Sigma_j S_d^{ij}$ and $D_v^{ii} = \Sigma_j S_v^{ij}$ are the associated degree matrices. A resolution technique for the above formulation has been shown in[14].

Matrix completion (MC). Matrix factorization based approach leads to a non-convex minimization problem and hence rarely benefits from global convergence guarantees. To limit the space of minimizers, it may be useful to impose a low-rank constraint on the solution X. Since rank minimization is still an NP-hard problem, it was proposed to relax the above constraint to its closest convex surrogate by making use of the nuclear norm penalty[64,65]. The formulation for the resulting nuclear norm minimization problem (referred to as matrix completion by the authors) is:

$$\min_{X} ||X||_* \text{ such that } Y{=}M(X)\,. \tag{5}$$

The above problem can be solved alternatively, by invoking majorization-minimization arguments[59] to deal with the mask operator M and by applying thresholding operations on the singular values to process the nuclear norm term[12].

Graph regularized matrix completion (GRMC). Just like matrix factorization, nuclear norm minimization based matrix completion can also be graph regularized by incorporating graph Laplacian penalties to take metadata/similarity information into account. The formulation for the minimization problem is given by:

$$\min_X ||Y - M(X)||_F^2 + \lambda ||X||_* + \mu_1 \text{tr}(X^\top L_d X) + \mu_2 \text{tr}(X L_v X^\top). \quad (6)$$

The above formulation can either be solved using ADMM (alternating direction method of multipliers)[66,67] as was done in[15] (referred as GRMC here) or by explicitly taking care of the constraint that the recovered values should be in the range [0, 1]. If the latter range constraint is taken into account, we obtain then a new variant called graph regularized binary matrix completion. The minimization with respect to X can be solved by making use of the PPXA (parallel proximal algorithm)[68]. Such approach allows to decouple the constraints by introducing proxy variables and then solving each subproblem in a parallel fashion as shown in[16] (referred as GRBMC here).

Setting of hyperparameters. The stepsize, regularization parameters and latent factor dimensions, for the above techniques have been tuned using cross-validation on training set (after hiding 10 % of the data) in each of the three cross-validation settings (see "Empirical evaluation"). The parameters obtained after extensive cross-validation on the setting CV2 (randomly hiding the virus entities) have been further used in predicting drugs for SARS-Cov-2 and the corresponding isolates (see "SARS-CoV-2 prediction" and "Predictions evolution with mutating novel coronavirus"). Similarly, the parameters selected for the setting CV3 (randomly hiding drug entities) have been used to evaluate the performance of the approaches in "Association prediction for new drugs".

Data availability

The dataset along with the solution is made available publicly at https://github.com/aanchalMongia/DVA and the prediction tool named DVApred (Drug-virus association prediction server) with a user-friendly interface is available as a webserver at http://dva.salsa.iiitd.edu.in.

Received: 3 August 2020; Accepted: 1 April 2021
Published online: 27 April 2021

References

1. Coronavirus Update (Live)-Worldometer, 2019. https://www.worldometers.info/coronavirus/. (Accessed 22 June 2020).
2. Harrison, C. Coronavirus puts drug repurposing on the fast track. *Nat. Biotechnol.* **38**(4), 379–381 (2020).
3. Ashburn, T. T. & Thor, K. B. Drug repositioning: identifying and developing new uses for existing drugs. *Nat. Rev. Drug Discov.* **3**(8), 673 (2004).
4. He, S. *et al.* Repurposing of the antihistamine chlorcyclizine and related compounds for treatment of hepatitis c virus infection. *Sci. Transl. Med.* **7**(282), 282ra49 (2015).
5. Frantz, S. Drug discovery: Playing dirty (2005).
6. McLean, S. R. *et al.* Imatinib binding and ckit inhibition is abrogated by the ckit kinase domain i missense mutation val654ala. *Mol. Cancer Ther.* **4**(12), 2008–2015 (2005).
7. Ezzat, A. *et al.* Computational prediction of drug–target interactions using chemogenomic approaches: an empirical survey. *Brief. bioinf.* **20**(4), 1337–1357 (2019).
8. Beg, M. A. & Athar, F. Anti-HIV and anti-HCV drugs are the putative inhibitors of RNA-dependent-RNA polymerase activity of nsp12 of the sars cov-2 (COVID-19). *Pharm. Pharmacol. Int. J.* **8**(3), 163–172 (2020).
9. Lee, V.S., Chong, W.L., Sukumaran, S.D., Nimmanpipug, P., Letchumanan, V., Goh, B.H., Lee, L.-H., Zain, S.M., & Rahman, N.A. Computational screening and identifying binding interaction of anti-viral and anti-malarial drugs: Toward the potential cure for sars-cov-2. *Progress Drug Discov. Biomed. Sci.* **3**(1) (2020).
10. Lipsitch, M., Perlman, S. & Waldor, M. K. Testing COVID-19 therapies to prevent progression of mild disease. *Lancet Infect. Diseases* **20**(12), 1367 (2020).
11. Sahoo, S.K. & Vardhan, S. Computational evidence on repurposing the anti-inuenza drugs baloxavir acid and baloxavir marboxil against COVID-19. arXiv preprint arXiv:2009.01094 (2020).
12. Mongia, A., Sengupta, D. & Majumdar, A. Mcimpute: Matrix completion based imputation for single cell RNA-seq data. *Front. Genet.* **10**, 9 (2019).
13. Mongia, A. & Majumdar, A. Deep matrix completion on graphs: Application in drug target interaction prediction. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP 2020)*, 1324–1328. (IEEE, 2020).
14. Ezzat, A., Zhao, P., Min, Wu., Li, X.-L. & Kwoh, C.-K. Drug-target interaction prediction with graph regularized matrix factorization. *IEEE/ACM Trans. Comput. Biol. Bioinform. (TCBB)* **14**(3), 646–656 (2017).
15. Mongia, A. & Majumdar, A. Drug-target interaction prediction using multi graph regularized nuclear norm minimization. *PLoS ONE* **15**(1), e0226484 (2020).
16. Mongia, A., Chouzenoux, E., & Majumdar, A. Computational prediction of drug-disease association based on graph-regularized one bit matrix completion. *bioRxiv* (2020). https://www.biorxiv.org/content/10.1101/2020.04.02.020891v1.abstract.
17. Zumla, A., Chan, J. F. W., Azhar, E. I., Hui, D. S. C. & Yuen, K.-Y. Coronaviruses|drug discovery and therapeutic options. *Nat. Rev. Drug Discov.* **15**(5), 327–347 (2016).
18. Burez, J. & Van den Poel, D. Handling class imbalance in customer churn prediction. *Expert Syst. Appl.* **36**(3), 4626–4636 (2009).
19. Ahlgren, N. A., Ren, J., Lu, Y. Y., Fuhrman, J. A. & Sun, F. Alignment-free oligonucleotide frequency dissimilarity measure improves prediction of hosts from metagenomically-derived viral sequences. *Nucleic Acids Res.* **45**(1), 39–53 (2017).
20. Beigel, J. H., *et al.* Remdesivir for the treatment of Covid-19. *N Engl J Med* **383**(19), 1813-1826 (2020).
21. Hung, I.F.-N. *et al.* 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. *Lancet* **395**(10238), 1695–1704 (2020).
22. Zeng, Y.-M. *et al.* Comparative effectiveness and safety of ribavirin plus interferon-alpha, lopinavir/ritonavir plus interferon-alpha, and ribavirin plus lopinavir/ritonavir plus interferon-alpha in patients with mild to moderate novel coronavirus disease 2019: Study protocol. *Chin. Med. J.* **133**(9), 1132–1134 (2020).
23. Rodrigo, J. *et al.* Sofosbuvir as a potential alternative to treat the SARS-CoV-2 epidemic. *Sci. rep.* **10**(1), 1–5 (2020).
24. Wang, Z., Chen, X., Yunfei, Lu., Chen, F. & Zhang, W. Clinical characteristics and therapeutic procedure for four cases with 2019 novel coronavirus pneumonia receiving combined chinese and western medicine treatment. *Biosci. Trends* **14**(1), 64–68 (2020).
25. Duan, Y., Hai-Liang, Z., & Chongchen, Z. Advance of promising targets and agents against COVID-19 in China. *Drug discovery today* **25**(5), 810–812 (2020).
26. New trial starts in UK to see if ibuprofen can help prevent severe breathing problems in Covid-19 patients. https://www.thejournal.ie/ibuprofen-trial-coronavirus-5113390-Jun2020/.

27. Martins-Filho, Paulo Ricardo, Edmundo Marques do Nascimento-Júnior, and Victor Santana Santos. No current evidence supporting risk of using Ibuprofen in patients with COVID-19. *Int J Clin Pract* **74**(10), e13576 (2020).
28. Chatterjee, S. An overview of mutations occurring within the coronavirus-2 genome: Mutations data reporting on sars-cov-2. SSRN 3632241 (2020).
29. Lou, Yan, *et al.* Clinical outcomes and plasma concentrations of baloxavir marboxil and favipiravir in COVID-19 patients: an exploratory randomized, controlled trial. *Eur J Pharm Sci* **157**, 105631 (2021).
30. Kasgari, H. A. *et al.* Evaluation of the efficacy of sofosbuvir plus daclatasvir in combination with ribavirin for hospitalized COVID-19 patients with moderate disease compared with standard care: A single-centre, randomized controlled trial. *J. Antimicrobial Chemother.* **75**(11), 3373–3378 (2020).
31. Elalfy, H., Besheer, T., El-Mesery, A., El-Gilany, A.H., Elazez, M.S.A., Alhawarey, A., Alegezy, M., Elhadidy, T., Hewidy, A.A., Zaghloul, H., *et al.* Effect of a combination of nitazoxanide, ribavirin and ivermectin plus zinc supplement (mans. nriz study) on the clearance of mild COVID-19. *J. Med. Virol.* (2021).
32. Eslami, G. *et al.* The impact of sofosbuvir/daclatasvir or ribavirin in patients with severe COVID-19. *J. Antimicrob. Chemother.* **75**(11), 3366–3372 (2020).
33. Jácome, R., Campillo-Balderas, J. A., de León, S. P., Becerra, A. & Lazcano, A. Sofosbuvir as a potential alternative to treat the sars-cov-2 epidemic. *Sci. Rep.* **10**(1), 1–5 (2020).
34. Roozbeh, F. *et al.* Sofosbuvir and daclatasvir for the treatment of COVID-19 outpatients: A double-blind, randomized controlled trial. *J. Antimicrob. Chemother.* **76**(3), 753–757 (2021).
35. Sayad, B., Sobhani, M. & Khodarahmi, R. Sofosbuvir as repurposed antiviral drug against COVID-19: Why were we convinced to evaluate the drug in a registered/approved clinical trial?. *Arch. Med. Res.* **51**(6), 577–581 (2020).
36. Randomized Clinical Trial for the Prevention of SARS-CoV-2 Infection (COVID-19) in Healthcare Personnel (EPICOS). https://clinicaltrials.gov/ct2/show/NCT04334928.
37. Clososki, G. C. *et al.* Tenofovir disoproxil fumarate: New chemical developments and encouraging in vitro biological results for sars-cov-2. *J. Braz. Chem. Soc.* **31**(8), 1552–1556 (2020).
38. Kutlu, O. Can tenofovir diphosphate be a candidate drug for sars-cov2? First clinical perspective. *Int. J. Clin. Practice* e13792 (2021).
39. Sanjuán, R., Nebot, M. R., Chirico, N., Mansky, L. M. & Belshaw, R. Viral mutation rates. *J. Virol.* **84**(19), 9733–9748 (2010).
40. Pachetti, M. *et al.* Emerging sars-cov-2 mutation hot spots include a novel RNA-dependent-RNA polymerase variant. *J. Transl. Med.* **18**, 1–9 (2020).
41. Hussain, M., Galvin, H. D., Haw, T. Y., Nutsford, A. N. & Husain, M. Drug resistance in influenza a virus: The epidemiology and management. *Infect. Drug Resistance* **10**, 121 (2017).
42. Huggins, J. W. Prospects for treatment of viral hemorrhagic fevers with ribavirin, a broad-spectrum antiviral drug. *Rev. Infect. Diseases* **11**(Supplement 4), S750–S761 (1989).
43. Schaefer, E. A. K. & Chung, R. T. Anti-hepatitis c virus drugs in development. *Gastroenterology* **142**(6), 1340–1350 (2012).
44. Wishart, D. S. *et al.* Drugbank: A comprehensive resource for in silico drug discovery and exploration. *Nucleic Acids Res.* **34**(suppl 1), D668–D672 (2006).
45. Chopra, A., Saluja, M. & Venugopalan, A. Effectiveness of chloroquine and inflammatory cytokine response in patients with early persistent musculoskeletal pain and arthritis following chikungunya virus infection. *Arthritis Rhumatol.* **66**(2), 319–326 (2014).
46. Das, I. *et al.* Heat shock protein 90 positively regulates chikungunya virus replication by stabilizing viral non-structural protein nsp2 during infection. *PLoS ONE* **9**(6), e100531 (2014).
47. De Clercq, E. & Li, G. Approved antiviral drugs over the past 50 years. *Clin. Microbiol. Rev.* **29**(3), 695–747 (2016).
48. Gallegos, K. M., Drusano, G. L., Argenio, D. Z. D. & Brown, A. N. Chikungunya virus: In vitro response to combination therapy with ribavirin and interferon alfa 2a. *J. Infect. Diseases* **8**, 1192–1197 (2016).
49. Jin, Z., Zhao, Y., Sun, Y., Zhang, B., Wang, H., Wu, Y., Zhu, Y., Zhu, C., Hu, T., Du, X. *et al.* Structural basis for the inhibition of COVID-19 virus main protease by carmofur, an antineoplastic drug. bioRxiv (2020). https://www.biorxiv.org/content/10.1101/2020.04.09.033233v1. (**abstract**).
50. Razonable, R. R. Antiviral drugs for viruses other than human immunodeficiency virus. *Mayo Clin. Proc.* **86**, 1009–1026 (2011) (**Elsevier**).
51. Shiryaev, S. A. *et al.* Repurposing of the anti-malaria drug chloroquine for zika virus treatment and prophylaxis. *Sci. Rep.* **7**(1), 1–9 (2017).
52. Sugaya, N. & Ohashi, Y. Long-acting neuraminidase inhibitor laninamivir octanoate (cs-8958) versus oseltamivir as treatment for children with influenza virus infection. *Antimicrob. Agents Chemother.* **54**(6), 2575–2582 (2010).
53. Winther, B. & Mygind, N. Potential benefits of ibuprofen in the treatment of viral respiratory infections. *Inflammopharmacology* **11**(4), 445 (2003).
54. Pickett, B. E. *et al.* Vipr: An open bioinformatics database and analysis resource for virology research. *Nucleic Acids Res.* **40**(D1), D593–D598 (2012).
55. Kanehisa, M. *et al.* From genomics to chemical genomics: New developments in KEGG. *Nucleic Acids Res.* **34**(suppl 1), D354–D357 (2006).
56. NCBI. https://www.ncbi.nlm.nih.gov/.
57. Hattori, M., Tanaka, N., Kanehisa, M. & Goto, S. Simcomp/subcomp: Chemical structure search servers for network analyses. *Nucleic Acids Res.* **38**(suppl 2), W652–W656 (2010).
58. Chouzenoux, E., Jezierska, A., Pesquet, J. C. & Talbot, H. A majorize-minimize subspace approach for l2–l0 image regularization. *SIAM J. Imaging Sci.* **6**(1), 563–591 (2013).
59. Sun, Y., Babu, P. & Palomar, D. P. Majorization-minimization algorithms in signal processing, communications, and machine learning. *IEEE Trans. Signal Process.* **65**(3), 794–816 (2017).
60. Hastie, T., Mazumder, R., Lee, J. D. & Zadeh, R. Matrix completion and low-rank SVD via fast alternating least squares. *J. Mach. Learn. Res.* **16**(1), 3367–3402 (2015).
61. Tariyal, S., Majumdar, A., Singh, R. & Vatsa, M. Deep dictionary learning. *IEEE Access* **4**, 10096–10109 (2016).
62. Mongia, A., Debarka, S. & Angshul, M. deepmc: Deep matrix completion for imputation of single-cell rna-seqdata. *J. Comput. Biol.* **27**(7), 1011–1019 (2020).
63. Chung, Fan RK, and Fan Chung Graham. Spectral graph theory. No. 92. *American Math. Soc.*, 1997.
64. Candes, E. J. & Recht, B. Exact matrix completion via convex optimization. *Found. Comput. Math.* **9**(6), 717–772 (2009).
65. Candès, E. J. & Tao, T. The power of convex relaxation: Near-optimal matrix completion. *IEEE Trans. Inf. Theor.* **56**(5), 2053–2080 (2010).
66. Boyd, S., Parikh, N., Chu, E., Peleato, B., Eckstein, J. Distributed optimization and statistical learning via the alternating direction method of multipliers. *Found. Trends Mach. Learn.* **3**(1), (2010).
67. Komodakis, N. & Pesquet, J. Playing with duality: An overview of recent primal-dual approaches for solving large-scale optimization problems. *IEEE Signal Process. Mag.* **32**(6), 31–54 (2015).
68. Pustelnik, N., Chaux, C. & Pesquet, J.-C. Parallel proximal algorithm for image restoration using hybrid regularization. *IEEE Trans. Image Process.* **20**(9), 2450–2462 (2011).

Acknowledgements

This work has been done within the Inria associate team COMPASS (COmputational Medicine for oPtimal drug ASSociation) fostering the collaboration between the Inria project-team OPIS, France, and the SALSA laboratory in IIIT Delhi, India.

Author contributions

A.M. and E.C. led the study. A.M. contributed to the problem formulation, design of the experiments and interpretation of the results. E.C. analyzed the backend algorithms involved. A.M. (Aanchal Mongia) curated the database, implemented the proposed methodology and performed the experiments in the framework. S.K.S. helped with the manuscript structuring and gave his inputs on the medical applicability of the work. All authors read and reviewed the manuscript.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1038/s41598-021-88153-3.

Correspondence and requests for materials should be addressed to E.C. or A.M.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.