Computational drug repositioning for COVID-19

Mateo Torres[1,+], Suzana de Siqueira Santos[1,+], Diego Galeano[1,+], María del Mar Sánchez[2], Luca Cernuzzi[2], and Alberto Paccanaro[1,3,+,*]

[1]Fundação Getúlio Vargas, Escola de Matemática Aplicada, Rio de Janeiro, Brazil
[2]Universidad Católica "Nuestra Señora de la Asunción", Asunción, Paraguay
[3]Royal Holloway, University of London, Department of Computer Science, Egham, United Kingdom
[*]alberto.paccanaro@rhul.ac.uk
[+]these authors contributed equally to this work

ABSTRACT

COVID-19, caused by the SARS-CoV-2 virus, has become a leading cause of morbidity and mortality worldwide. Drug repositioning, the process of finding new therapeutic indications for marketed drugs, is a promising alternative to new drug development, with lower costs and shorter development times. Here, we present two computational approaches for the repositioning of drugs for COVID-19. Our first approach consists of a novel non-negative matrix factorisation algorithm to computationally reposition 126 broad spectrum antivirals that have been approved, or are being developed, for 80 viruses. Our second approach, based on ideas from network medicine, aims at predicting drugs that are effective against COVID-19 by estimating the perturbation each drug induces on a subnetwork of the human interactome that is crucial for SARS-CoV-2 infection/replication. Using graph kernels and host proteins weighted by gene expression data from SARS-CoV-2 infected cell lines, we obtained a ranking of 1853 therapeutically diverse FDA-approved drugs. Our experiments show that nine out of 10 of our top predicted broad spectrum antivirals are already indicated for compassionate use in COVID-19 patients worldwide; and that the ranking obtained by our perturbation analysis approach aligns with independent experimental data on *in vitro* screenings. Finally, to assist scientists working in drug repositioning options for COVID-19, we present the COVID-19 Repositioning Explorer (CoREx), an interactive online tool for exploring the interplay between approved drugs and SARS-CoV-2 host proteins in the context of biological networks, protein function, drug clinical use, and the connectivity map. CoREx is freely available at: https://paccanarolab.org/corex/.

Introduction

The coronavirus disease COVID-19, caused by the novel coronavirus SARS-CoV-2, has been declared a pandemic by the World Health Organization (WHO) in March 11th 2020, and as of February 10th 2021 has infected more than 105 million people, resulting in 2,302,614 deaths worldwide[1,2]. So far, the Food and Drug Administration (FDA) has issued emergency use authorizations of drug combinations for hospitalised patients with severe COVID-19 requiring assisted oxygenation[3]. Furthermore, the manufacturing, and distribution of vaccines against SARS-CoV-2 is estimated to require 12 to 18 months[4].

Drug discovery and development present several challenges including high attrition rates, long development times, and substantial costs[5]. Drug repositioning, the process of finding new therapeutic indications for already marketed drugs, has emerged as a promising alternative to new drug development. It involves the use of de-risked compounds in human, which translates to lower costs and shorter development times[6]. Identifying commercially available drugs with therapeutic effects for COVID-19 could provide early treatment options until effective therapies and vaccination schemes become widely available and, more generally, an alternative option for COVID-19 treatment. Computational prediction models can assist in this situation by prioritising drugs based on their therapeutic potential assigned based on the available biomedical knowledge.

We present two computational approaches for the repositioning of marketed drugs for COVID-19. Our first approach focuses on predicting Broad Spectrum Antivirals (BSAs) that might be effective against SARS-CoV-2. Given a small number of drugs associated to a virus and their current stage in the approval process, our model assigns scores to a broader range of drugs with previously unknown association to the virus. Our idea is to develop a recommender system that recommends BSAs to SARS-CoV-2 using a matrix decomposition algorithm. Our method assigns low-dimensional feature vectors to each BSA and each virus so that the dot product between the vectors models the efficacy of the BSA against the virus, encoding the biological interplay between BSAs and viruses. This model is inspired by our previous work, where we showed that using a matrix decomposition model can effectively predict the frequency of drug side-effects, while being biologically interpretable[7]. We are aware of only one other recommender systems-based study focused on the prediction of unknown antivirals for viruses[8]. In relation to this previous work, this paper makes two primary contributions. First, while our work focuses on computational repositioning of approved broad spectrum antivirals, the work of Sosnina et al.[8] focuses on predicting small-molecule antiviral

activity that are not restricted to approved drugs. Second, and perhaps more importantly, our work is the first to exploit the developmental status of drug-virus associations for drug repositioning, effectively accounting for varying levels of uncertainty. We show the usefulness of our approach at predicting whether a drug will be effective for a virus, an application that becomes critical for computationally assisted clinical trials for infectious diseases such as COVID-19.

Our second approach is a network medicine approach. It follows the fundamental idea of applying network science towards studying and treating human diseases in the context of biological networks, such as the human protein-protein interaction (PPI) network (or human interactome)[9]. In this network, the nodes correspond to proteins and the links represent interactions between them. It has been shown that proteins associated with specific hereditary diseases tend to cluster in neighbourhoods of the interactome (the disease module), and network-based approaches have shown that drugs for genetic diseases can be found by measuring the interactome-based proximity between drug protein targets and the disease module[10,11]. Although the concept of network medicine has been originally applied to the study of genetic diseases, recent work shows that they can be useful for infectious diseases such as COVID-19[12]. For COVID-19, a disease module can be defined as the set of human proteins (the host proteins) that interact with SARS-CoV-2 proteins, allowing the infection and replication processes. In particular, Gysi et al.[13] have shown that, for SARS-CoV-2, most of the experimentally identified human host proteins[14] are not randomly placed in the interactome, and do indeed group together in a large connected component, forming a COVID-19 disease module. Taking this into consideration, we designed our network medicine approach to rank FDA-approved drugs based on the perturbation that each drug induces on the COVID-19 module through the interactome. To calculate this effect, we measured network-based proximities between drug targets and proteins using graph kernels. An important aspect of our method is that it models the relative importance of host proteins for the disease. Recent network medicine-based approaches for repositioning of drugs for COVID-19[12,13] consider all host proteins equally. It has been shown, however, that some of the host proteins are instrumental in the interplay with SARS-CoV-2 (e.g. ACE2 and TMPRSS2[15]). We model this by weighting the importance of each host protein based on complementary information about gene expression from SARS-CoV-2 infected cell lines. We show that our network medicine approach benefits from this prioritisation of host proteins.

Computationally identifying repositionable drug candidates presents several challenges when it comes to evaluating the results, especially for emerging diseases such as COVID-19[16]. Such diseases are missing high quality experimental evidence, preventing researchers from relying on traditional performance metrics to assess their predictions. Here, we propose studying the interplay between promising drug candidates and the disease by putting together several sources of evidence that independently support the efficacy of the drugs. Furthermore, we have developed the COVID-19 Repositioning Explorer (CoREx), an online tool that allows the exploration of our results in the context of network biology. In CoREx we integrate several sources of information, connecting functional protein modules with drug targets and host proteins. CoREx also provides additional evidence for a drug of interest, such as whether the drug is on clinical trials for COVID-19, or whether the drug reverses the gene expression signature of SARS-CoV-2 infected cell lines according to the Connectivity Map (CMAP)[17]. Our two computational approaches, together with CoREx, are intended to assist doctors, biologists, and pharmacologists working on drug repositioning hypotheses for COVID-19.

Results

A matrix decomposition model for repositioning of broad-spectrum antivirals

Our starting point was the drug-virus association dataset manually curated by Andersen et al.[18]. 850 drug-virus associations were found between $n = 126$ unique Broad-Spectrum Antiviral (BSA) drugs and $m = 80$ distinct viruses, including SARS-CoV-2. We represented each drug-virus association in a $n \times m$ matrix Y, where $y_{ij} = 1$ if the *ith* drug was associated with the *jth* virus. The remaining associations in Y were filled in with zeros, i.e. $y_{ij} = 0$.

We observed that the entries in Y have different meanings. Each known drug-virus association in Y was annotated with its current developmental status by Andersen et al.[18] – see a breakdown in Figure 1a. 73.3% of the known drug-virus associations were found by culture/co-culture experiments; the first stage in the drug development process. 11.3% of the known associations reached clinical trials in patients (phases I-IV), and only 5.7% were approved for commercial use. The observed trend makes sense because typically, the development of a new BSA drug starts by screening a large number of compounds in cell-culture/co-culture or organoid experiments, a few hundred then progress into animal models, and a few dozen end up being tested in humans (clinical trials phases I-IV), until a single candidate gets approved for commercial use[18]. It is also well known that the probability of success of a given drug, i.e. its probability of being approved, increases as the candidate drug moves to the next stage in the drug development process. Our goal is to model the developmental status of subsets of entries in Y by associating them with probabilities of success[19]. In addition, given that the ultimate goal of the drug development process is to get approved drugs for clinical use in patients, we focus on predicting missing drug-virus associations in Y that are likely to get approved or to reach clinical trials phase IV.

The distribution of the known associations in Y is not uniformly distributed for every virus. We observed that there are certain viruses that have the largest amount of associated drugs (see Figure 1b). The viruses with most associations are

influenza A (FLUAV) and zika (ZIKV), with 47 associations; hepatitis C (HCV) with 46 associations, dengue (DENV) with 44 associations and the human immunodeficiency virus (HIV) with 42 associations. Similarly, there are certain drugs, such as Tilorone, that have been found with a broader spectrum of antiviral activity against multiple viruses (see Fig. 1c).

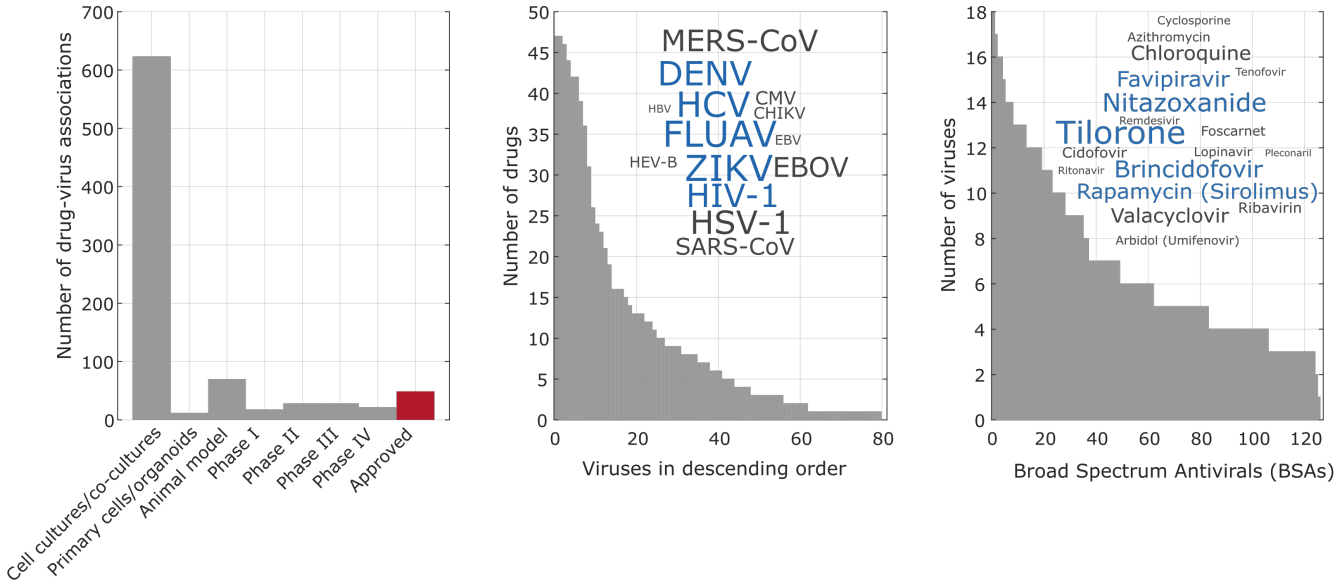

Figure 1. Drug-virus dataset statistics. We used the dataset manually curated by Andersen et al.[18]. **(a)** Number of drug-virus associations divided by their known developmental status. The development of Broad-Spectrum Antivirals (BSA) starts with *in vitro* experiments (e.g. cell culture), moves to animal models and then clinical trials in humans (phases I-IV). It finishes with the approval of the drug for commercial use (in red). **(b)** Number of drugs (BSAs) associated to each virus in the dataset. *Inset*. A word cloud shows the fourteen viruses with more associations. The size of the word is proportional to its number of associations and the most popular viruses among drugs are coloured in blue. **(c)** Number of viruses associated to each drug in the dataset. *Inset*. A word cloud shows the eighteen drugs with more associations. The size of the word is proportional to its number of associations and the most popular drugs among viruses are colored in blue.

These distributions resemble long-tailed distributions that we found previously in a dataset of drug-side effect associations, for which we developed a matrix decomposition model[20]. The fundamental assumption of this model is that each drug and each side effect can be represented as a low-dimensional feature vectors in a latent space such as the dot product between the vectors modelled the drug-side effect association. We realised that this assumption is also reasonable for our task: drugs and viruses can be represented as latent feature vectors in a low-dimensional space where the latent features might capture specific molecular mechanism of antiviral activity. Therefore, in our approach, each drug i is assigned to a low-dimensional feature vector $p_i^T \in \mathbb{R}^k$ (drug signature) and, similarly, each virus j is assigned to a low-dimensional feature vector $q_j \in \mathbb{R}^k$ (virus signature) such that an entry in y_{ij} is modelled by the dot product between the feature vectors, i.e. $\hat{y}_{ij} = p_i^T q_j$. Our model consists of decomposing the matrix Y by the product of two matrices $\hat{Y} = PQ$, where $P \in \mathbb{R}^{n\times k}$ (each row is a drug feature vector p_i), $Q \in \mathbb{R}^{k\times m}$ (each column is a virus feature vector q_j), and $k << \min(n,m)$ is the number of features that is assigned to each drug and each virus; k is also known as the rank of $\hat{Y}$. Our matrix decomposition model learns P and Q by minimising the following cost function:

$$\min_{P,Q\geq 0} \mathscr{L}(P,Q) = \underbrace{\frac{1}{2}\left\|M^A \circ (Y - PQ)\right\|_F^2}_{\text{approved, phase IV}} + \underbrace{\frac{1}{2}\sum_{s\in\{B,C,D,E\}} \alpha_s \left\|M^s \circ (Y - PQ)\right\|_F^2}_{\textit{In vitro}\text{, animal model, clinical trials}} + \underbrace{\frac{\alpha_z}{2}\left\|M^z \circ (PQ)\right\|_F^2}_{\text{zero-driven regularisation}}$$

$$\text{subject to non-negative constraints } P, Q \geq 0 \tag{1}$$

where $\|.\|_F$ is the Frobenius norm of a matrix and $\circ$ is the Hadamard or element-wise product. The first term in our model is the fitting constraint on the approved and phase IV drug-virus associations (set A). The projection matrix M^A is used to apply the summation only to entries in Y belonging to the set of approved associations A, that is, $M^A_{ij} = 1$ if drug i was approved or is in phase IV for virus j, or 0 otherwise. The second term in Equation (1) is the fitting constraint on the remaining known associations in Y that correspond to earlier stages in the drug development process. Sets B,C and D contains entries in Y belonging to clinical trials phases I, II and III, respectively. Set E contains associations in the *in vitro* and animal model stages. This disjoint split of the input data was introduced here because our dataset is fundamentally different from the drug side

effects dataset. While in the drug side effect association matrix known entries have the same meaning, in the drug-virus dataset, known associations in Y can be associated to probabilities of success. This means that drug-virus associations in early stages of development (e.g. set E) have a lower probability of being approved than those that are already on clinical trials. We introduced the parameters $\alpha_s \in [0,1]$ to control the probabilities of success for each subset of associations during learning. The third term in Equation (1) works as a regularisation term on the zero entries in Y, and it was introduced in our previous work[20]. Finally, to favour interpretability on the learned representations, we impose non-negative constraints on P and Q[21].

An overview of our matrix decomposition model is illustrated in Figure 2. Our starting point is the matrix Y containing binary drug-virus associations. We learn the matrices P and Q that minimise the loss function in Equation (1), by employing an iterative algorithm that uses a simple multiplicative update rule (see Methods). Our algorithm, inspired by the diagonally rescaled principle of non-negative matrix factorisation[21], is fast, it does not require setting a learning rate nor applying a projection function and it satisfies the Karush–Kuhn–Tucker (KKT) complementary conditions of convergence (see Methods). Having learned P and Q such that $Y \simeq PQ$, we calculate the matrix $\hat{Y} = PQ$. Note that, while Y contains binary numbers $[0,1]$ that represent our original data, $\hat{Y}$ contains real positive numbers that are our predicted scores. The methodological details are given in the next sections.

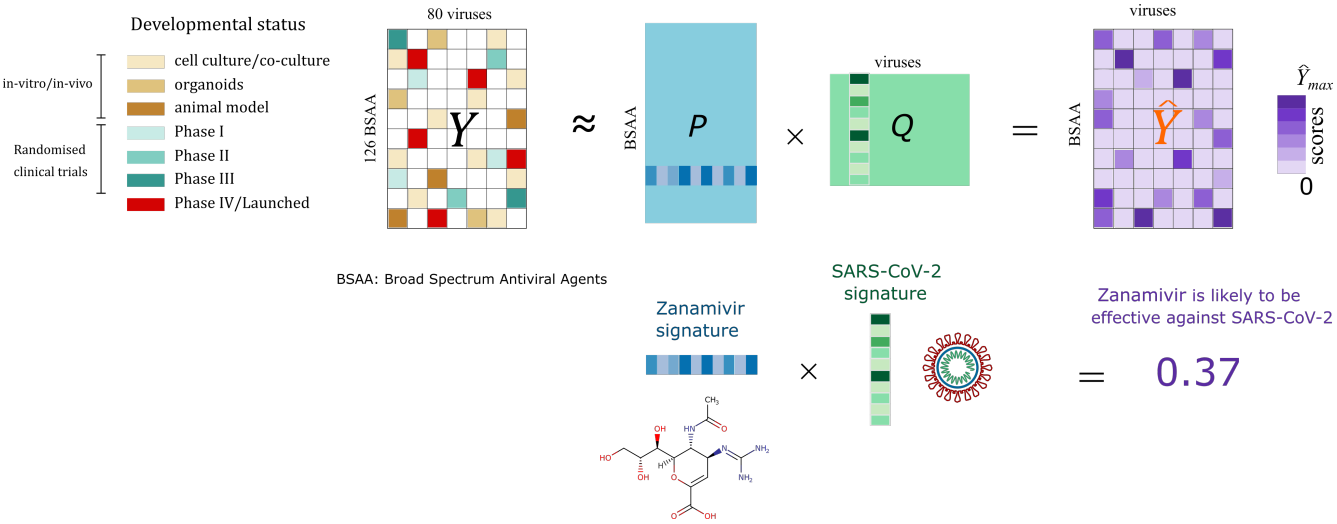

Figure 2. Overview of our matrix decomposition model for predicting effective drug-virus associations. 850 associations for $n = 126$ different Broad-Spectrum Antivirals (BSA) and $m = 80$ distinct viruses were collected from the Andersen et al.[18] database. The observed associations were arranged into an $n \times m$ matrix Y by setting $y_{ij} = 1$. Unobserved associations were encoded with zeroes. Our algorithm decomposes the matrix Y into the product of two matrices, P (of size $n \times k$) and Q (of size $k \times m$). By multiplying the matrices P and Q, we obtain $\hat{Y}$, that models Y, where all the zero entries are replaced with real numbers – these correspond to our predicted scores. Rows of P are the BSA feature vectors (or BSA signature); columns of Q are the virus feature vectors (virus signature). The lower illustration depicts how our model discovers a low-dimensional signature vector for the antiviral drug Zanamivir, and a low-dimensional signature vector for SARS-CoV-2, such that the dot-product of these two signatures models the predicted efficacy of the drug against SARS-CoV-2.

Predicting effective drugs against viruses

Our main goal is to predict drugs that might be effective against SARS-CoV-2. The drug-virus dataset contains five distinct strains of Human Coronaviruses (HCoV), including SARS-CoV-2. Its corresponding column in $\hat{Y}$ contains predicted scores that will be used to rank candidate drugs. Given that the validation of these predictions is not straightforward, we will dedicate a separate section to analyse specific SARS-CoV-2 predictions generated with our approach.

An interesting question is how well our model performs as a general repositioning method of broad-spectrum antiviral drugs for specific viruses. To assess this, we framed a matrix completion task where the goal is to predict missing drug-virus associations in Y that are likely to be approved or reach phase IV of clinical trials. There were 71 phase IV/approved associations (22 phase IV, 49 approved) between 26 BSA drugs and 28 distinct viruses. We performed a leave-one-out cross validation (LOOCV) procedure in which a single drug-virus association was removed from Y. We then measured how well our system was able to retrieve it. For every effective drug-virus pair (i, j) that was placed on a test set, we set $y_{ij} = 0$, and then trained the model using $k = 5, \alpha_B = 0.16, \alpha_C = 0.27, \alpha_D = 0.71, \alpha_E = 0.01$ and $\alpha_z = 2$ (our hyperparameter tuning procedure is explained in Methods). For the virus in the test set, we ranked its predicted drugs based on the predicted scores in the corresponding

column of $\hat{Y}$. Only drugs that were not associated to the virus in the training matrix were ranked. Drugs with higher values in the corresponding column of $\hat{Y}$ were ranked top in the list of predictions, and those with low values were ranked at the end of the list.

Sosnina et al.[8] have shown in their recent work that standard matrix factorization models such as truncated Singular Value Decomposition (tSVD)[22], and Non-Negative Matrix Factorization (NMF)[23] can be effectively used for predicting small molecule-antiviral activity in cell-assay derived interactions. Therefore, we compared the performance of our method against tSVD and NMF. Furthermore, following other methods that used LOOCV evaluations[24], we evaluated the performance at predicting one drug at a time, measuring how often that drug was found within the first 1, 5, 10, 15, 20, 25, and 30 drugs predicted by the different algorithms.

Figure 3 shows the performance of the methods at predicting effective BSA drugs against specific viruses. Our model outperforms the competitors for each number of predictions retrieved: by 4.2-15.5% in the top-1, by 14-40% in the top-5, by 25-47% in the top-10 and by 30-59% in the top-20. Overall, our method could recover 80% of the phase IV/approved BSA drugs for 28 distinct viruses in the top-25 predictions. We also observed that, in some cases, tSVD performs slightly better than NMF.

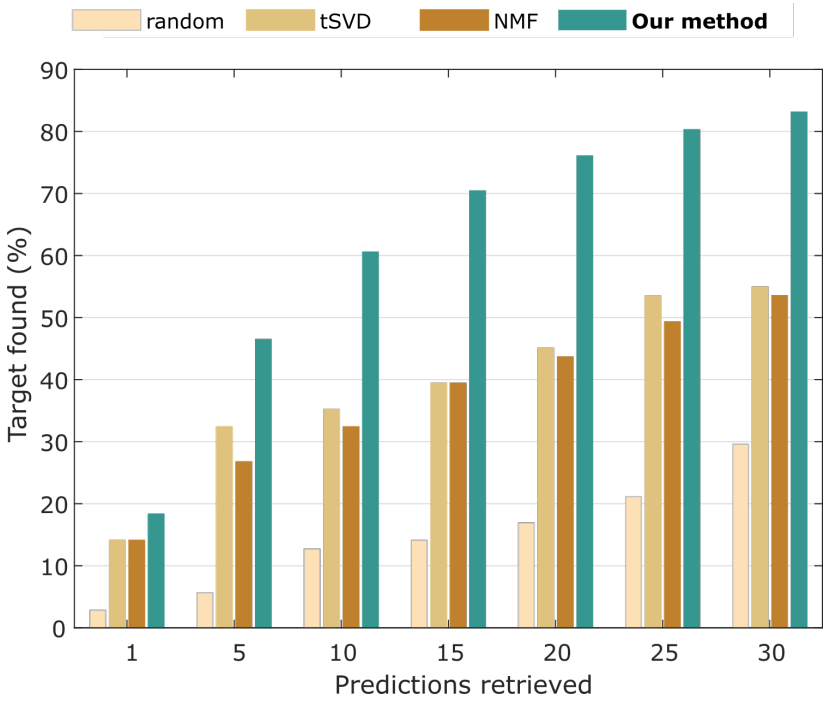

Figure 3. Performance at predicting approved/phase IV Broad Spectrum Antivirals (BSA) for 28 viruses. Percentage of approved BSA found for a specific viral disease in the predictions vs. the number of predictions retrieved. The performance of our method is compared to different matrix decomposition algorithms in a leave-one-out testing for viruses with approved BSAs or with BSAs in phase IV of drug development. NMF stands for Non-negative Matrix Factorisation (NMF) and tSVD for truncated Singular Value Decomposition (tSVD). Random is also included as a baseline.

Repositioning FDA-approved drugs with Network Medicine

Network medicine approaches have been largely successful at studying molecularly characterized diseases[9,24]. They rely on functional interdependencies between the molecular components in a human cell[9]. The central assumption is that a disease is rarely a consequence of an abnormality in a single gene, but a result of perturbations of the complex intracellular network[9]. For several diseases, it was observed that these perturbations are not randomly located in a molecular network, but they occur in nodes close to each other, forming a disease module[9]. Successful applications of molecular network analysis have been reported in the identification of disease genes[24] (e.g. coronary heart disease, diabetes mellitus, chronic lung diseases), and in drug development[25].

The use of network medicine for predicting drug efficacy was originally applied to genetic diseases. Guney et al.[10] proposed to predict drugs according to how likely they are to affect the disease module in the human PPI. It is based on the idea that a drug induces its effects on a human PPI subnetwork by binding to its target proteins[26,27] (blue nodes in Figure 4a). This causes

a perturbation in the interactome that is then propagated. The goal of their method is then to rank drugs according to how close the drug targets are from the disease module in the interactome. To measure this distance, Guney et al. use a z-score (hereafter, the Guney distance) based on the distribution of the shortest path length across random resamplings of nodes.

Recently, Gysi et al. applied an analogous idea for repositioning drugs for COVID-19[13]. It is based on the fact that viruses replicate themselves and infect the human cell by hijacking human translation mechanisms[13]. They observed that human proteins involved in SARS-CoV-2 replication and infection (the host proteins) are not randomly located in the PPI: they form a subnetwork (host protein subnetwork) analogous to a disease module[13], in which nodes tend to be within the same connected component, as illustrated in Figure 4a with red nodes. Gysi et al.[13] suggest that we can disrupt the molecular mechanisms of infection/replication by perturbing this subnetwork. They then rank the drugs according to how close the drug targets are from the host protein subnetwork. In particular, they use a method proposed by Cao et al.[28], called the diffusion state distance (DSD), which measures distances on graph in terms of random walks. It represents each node in the network as a vector indicating the number of times a p-step random walk starting from node i visits each node j. The DSD distance between two nodes is obtained by the L_1-norm between their vector representations.

Predictions for COVID-19 by Gysi et al.[13] are aligned with *in vitro* experiments in SARS-CoV-2 infected cell lines. Then, following the same assumptions than Gysi et al., we expect to predict drugs with *in vitro* efficacy against SARS-CoV-2 by using other network-based distances/proximities between nodes, such as the Guney distance, and graph kernels. In particular, we rely our predictions on graph kernels, whose theoretical properties and applications have been studied for more than 10 years[29,30]. They were shown to be successful at ranking genes with respect to cancer modules[31], predict gene function[32], and protein function[28].

Our network medicine approach relies on gene expression data for prioritising host proteins. Both works by Gysi et al[13], and Guney et al[10] consider all proteins/genes equally important for the infecction/disease. However host proteins may have different roles in the infection and replication of the virus. For example, it has been discovered that the ACE2 protein receptor is the viral entry factor of SARS-CoV-2[33]. Furthermore, recent gene expression experiments on infected SARS-CoV-2 cell lines also suggest that there are certain protein-coding genes that might play a key role during the infection process[34].

According to Sirota et al[35], a drug that reverts the effects that a disease causes in gene expression is a potential candidate for treating the disease. Analogously, a drug that reverts the SARS-CoV-2 infection gene expression profile can have therapeutic effects for COVID-19. Then, if we prioritise host targets whose expression levels change the most in SARS-CoV-2 infection, we are more likely to predict candidates for COVID-19 that align with the approach by Sirota et al[35]. We propose to use gene expression data from SARS-CoV-2 infected and mock treated cell lines to weight the host proteins in our network medicine approach. The final score of each drug is obtained by a weighted sum of graph kernel-based similarities between drug targets and host proteins, as described in the next Section.

Kernel-based scores for drugs

We obtain a vector of drug scores by computing the following matrix multiplication (illustrated in Figure 4c).

$$\text{scores} = TK\mathbf{h}. \tag{2}$$

where $T \in \mathbb{R}^{N \times n_V}$ is a matrix containing drug target associations for $N = 1853$ drugs and n_V proteins. For each drug i, and each protein j, we assign $T_{i,j} = 1$ if j is a target of i, and $T_{i,j} = 0$, otherwise (Drug Targets box in Figure 4b). $K \in \mathbb{R}^{n_V \times n_V}$ represents one of five different kernels (see Methods). Given a graph kernel and a drug i, we score i by summing the graph kernel proximities between the drug target proteins and the host proteins (see PPI Kernels box in Figure 4b). h is a column vector indicating the weights of the host proteins. For each host protein, we assign the log fold change between the gene expression levels of A549 cell lines infected with SARS-CoV-2, and mock-treated cell lines (See Methods for details on the RNAseq data). This measures how many times the expression levels of the protein-encoding gene increase or decrease in cells infected with SARS-CoV-2 compared to the non-infected ones. We assign zero to the remaining proteins (see Figure 4b and 4c). (see Host Proteins box in Figure 4b). The vector of scores is then used as a ranking of drugs for COVID-19.

Evaluation

We rely on three different sources of evidence from ongoing research that point to effectiveness of drugs for COVID-19. These independent sources of evidence indicate drug efficacy according to different criteria: *in vitro* experiments, clinical trials, and Connectivity Map scores. First, *in vitro* experiments show the potential for compounds to be effective at reducing viral infection and replication. Using data from *in vitro* experiments, we frame a binary classification problem, where a drug is assigned a label 1 if it shows effectiveness *in vitro* against SARS-CoV-2, or 0 otherwise. Evaluating our models with this kind of evidence allows us to assess whether drugs with molecular antiviral efficacy are prioritised vs. other drugs.

Second, clinical trial studies make up several crucial steps in the drug approval process. These trials are used to assess pharmacokinetics, dosage, therapeutic efficacy, and safety of drugs[36]. Each phase in clinical trials involves an increasing

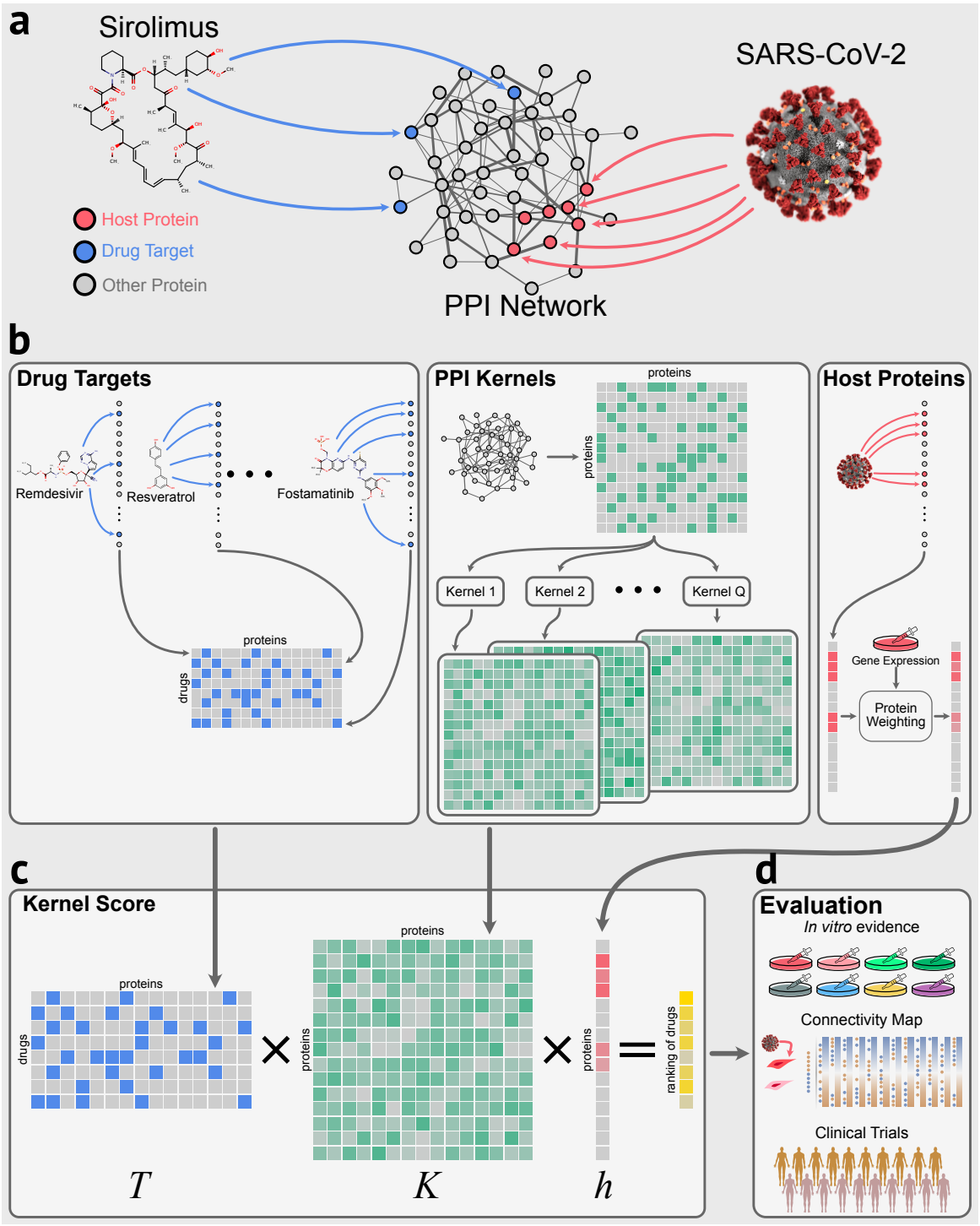

Figure 4. Overview of our Network Medicine approach. In **(a)**, we illustrate the human interactome containing both host proteins (red) and drug targets (blue). In **(b)**, we show the pipeline of our approach. 8129 drug target associations between $N = 1853$ FDA-approved drugs and $n_V = 15646$ proteins are represented by a binary matrix T (blue matrix), multiple graph kernels are calculated in the interaction on $n_V \times n_V$ matrices (green matrices), and a vector h of size n_V indicating the host proteins (red vector). In **(c)** our Kernel score is calculated using a matrix multiplication to obtain a prediction score for each drug. Finally in **(d)**, the resulting ranking is evaluated using different types of evidence: *in vitro* efficacy against SARS-CoV-2, Connectivity Map, and clinical trials.

number of patients, thus achieving higher statistical significance while minimising the number of patients that risk developing adverse side effects[37]. Due to the risks and expenses involved in clinical trials, a lot of conditions set by biologists and medics must be met before they are allowed to be carried out. We frame a second binary classification problem, where a drug is assigned a label 1 if it is involved in a clinical trial, or 0 otherwise. Evaluating our models with clinical trial evidence allows us to determine if they prioritise drugs that would be included in such trials.

Third, following the seminal work of Sirota et al.[35] (see Methods), we use the Connectivity Map (CMAP)[17] to contrast changes in gene expression levels caused by a drug (drug expression profile) against changes induced by SARS-CoV-2 infection (disease expression profile). The fundamental idea is that if a drug expression profile is opposite to a disease expression profile, then it is likely to "revert" the disease signature and to have therapeutic effects[35]. This idea is mostly applicable to host targeting drugs. Thus, we show this evaluation only for the network medicine approach. Again, we frame a binary classification problem where a drug is assigned a label 1 if it reverts the COVID-19 disease signature, or 0 otherwise. It allows us to assess whether our approach prioritises drugs with potentially therapeutic effects.

Together, these sources of evidence provide support for the efficacy of a particular drug against COVID-19. They are also aligned in terms of their Anatomical Therapeutic Chemical classification, which suggests that there is an agreement supported by these independent sources of evidence. Table 1 shows the most frequent ATC categories for drugs showing any of the three types of evidence. Nervous System is the top category for drugs with *in vitro* efficacy, as well as the group of drugs with significant CMAP scores. This is interesting, considering widely reported neurological symptoms such as loss of smell and taste[38]. Cardiovascular System is present in all three groups of drugs, and there is research linking COVID-19 with cardiovascular complications, which are among the top comorbidities for COVID-19[39].

Source of evidence	Top category	number of drugs	% of category
in vitro	Nervous System	23	7.77%
	Cardiovascular System	16	7.11%
	Antineoplastic and immunomodulating agents	13	5.49%
Clinical Trials	Antineoplastic and immunomodulating agents	33	13.92%
	Cardiovascular System	28	12.44%
	Blood and blood forming organs	25	25.00%
CMAP	Nervous System	9	3.04%
	Cardiovascular System	5	2.22%
	Antineoplastic and immunomodulating agents	5	2.10%

Table 1. Top ATC categories for FDA approved drugs showing efficacy *in vitro*, involved in clinical trials, and reversing COVID-19 gene expression profiles. Interestingly, Nervous System is consistently represented throughout these sources of evidence. This is consistent with widely reported neurological symptoms such as loss of smell and taste[38]. Similarly, the consistent inclusion of Cardiovascular System as a top category is consistent with research linking COVID-19 with cardiovascular complications, as well as cardiovascular diseases being among the top comorbidities for COVID-19[39].

None of these three sources of evidence can be considered a gold standard, as neither of them can ensure therapeutic effects for COVID-19 patients. Therefore, traditional binary classification metrics might not be enough to properly evaluate the usefulness of these methods. Instead, our main evaluation strategy is to quantify the alignment between our predictions and the three sources of evidence. By evaluating in this way, we avoid selection bias, that may assign an unfair advantage to drugs that might appear promising simply because they are the first to be included in clinical trials, or are less complex/expensive to test *in vitro*. We use the Wilcoxon-Mann-Whitney test, comparing prediction scores between drugs with some type of evidence and the remaining ones. The boxplots comparing the predictions scores are shown in Figure 5. The DSD-based method[13], and the Guney distance[10] are included for comparison. We report the results of traditional classification metrics on the Supplementary Material.

In vitro *evaluation*

We started by analysing whether predicted drugs with our matrix factorization approach aligns to in-vitro experimental evidence. Of the 126 BSAs in the drug-virus dataset, 10 are included in the set of drugs that have shown *in vitro* efficacy against SARS-CoV-2[13,40]. Our matrix decomposition method significantly assigns higher scores to BSAs with *in vitro* efficacy (Wilcoxon-Mann-Whitney p-value 5e−3) as illustrated in Figure 5a.

Of the 1853 FDA-approved drugs considered by our network medicine approach, 78 are included in the set of drugs that have shown *in vitro* efficacy against SARS-CoV-2[13,40]. We observed that the scores of drugs with *in vitro* efficacy against SARS-CoV-2 are significantly higher than the remaining drugs for all kernels (Wilcoxon–Mann–Whitney p-values 4.11e−3 for the Commute Time kernel, 2.91e−3 for the Diffusion Kernel, 3.62e−4 for the p-Step kernel, 4.39e−3 for the Regularised

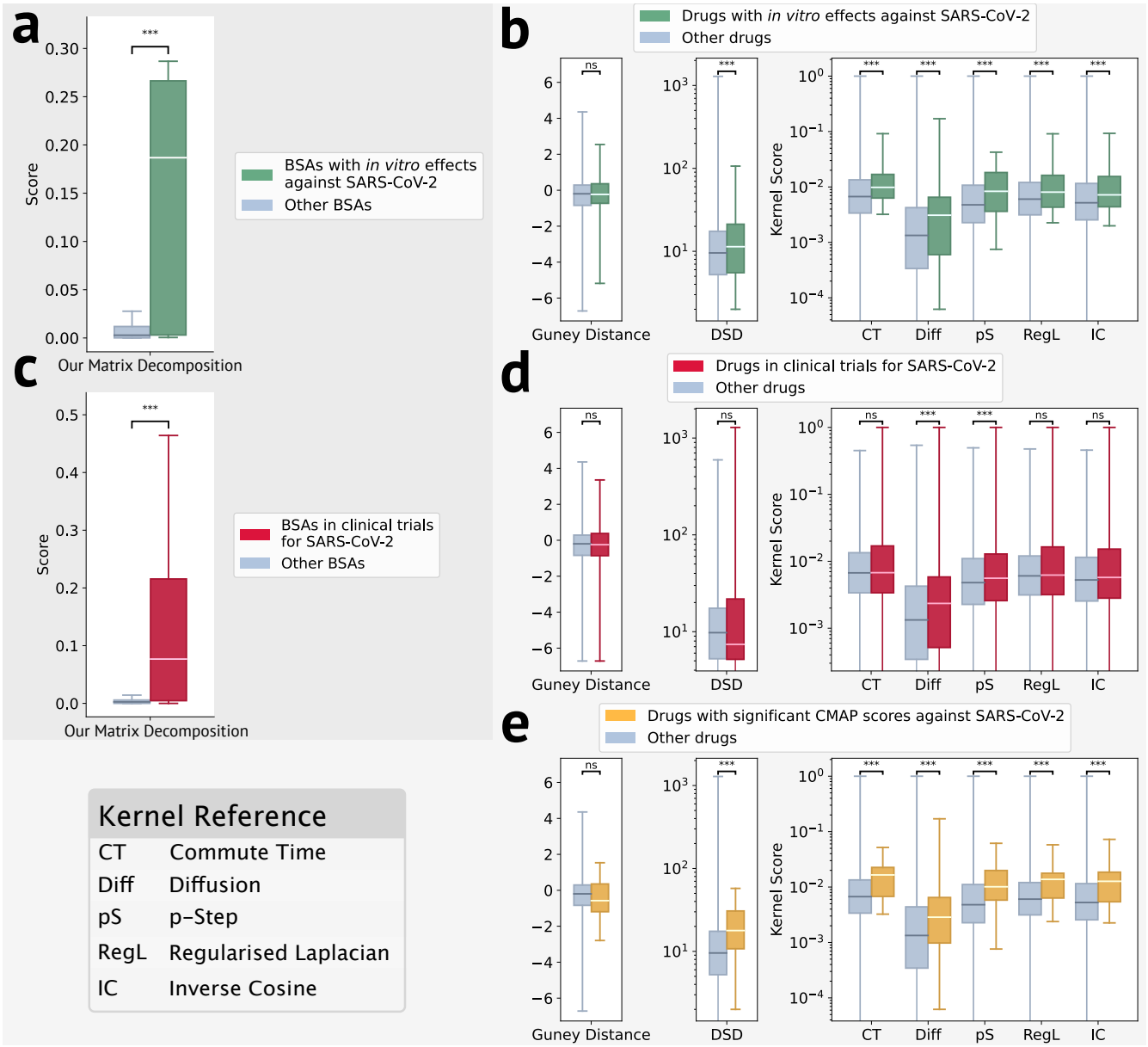

Figure 5. Evaluation of predictions for COVID-19. Alignment to multiple sources of evidence. Boxplots showing the distribution of scores of our two approaches and baseline methods are shown, together with Wilcoxon-Mann-Whitney p-values (significant below 0.05). **(a)** Our matrix decomposition approach that predicts scores for 126 Broad-Spectrum Antivirals significantly assigns higher scores to BSAs with *in vitro* effects against SARS-CoV-2 (p-value 5e−3); **(b)** Alignment of network based methods for 1,853 FDA-approved drugs. Two baseline methods are compared against our kernel based methods using five kernels. From left to right, p-values are 4.72e−1 for the Guney Distance, 1.60e−2 for DSD, 3.30e−3 for the Commute Time kernel, 1.93e−3 for the Diffusion Kernel, 4.14e−4 for the p-Step kernel, 3.58e−3 for the Regularised Laplacian kernel, and 1.03e−2 for the Inverse Cosine kernel. **(c)** Our matrix decomposition approach assigns higher scores to BSAs involved in clinical trials (p-value 1.22e−5); **(d)** Alignment of network based methods to FDA approved drugs involved in clinical trials. From left to right, p-values are 3.29e−1, 4.80e−1, 1.98e−1, 7.89e−3, 2.86e−2, 1.70e−1, and 1.07e−1. Only our network medicine approach, using the Diffusion and p-Step kernels, achieves significant p-values **(e)** Alignment of network based methods to FDA approved drugs involved in clinical trials. From left to right, the p-values are 1.35e−1, 1.36e−3, 1.65e−3, 2.43e−2, 2.48e−3, 1.97e−3, and 2.82e−3. Note that for the Guney Distance, the more negative the score, the better[10].

Laplacian kernel, and 2.70e−2 for the Inverse Cosine kernel), as illustrated in Figure 5b. Differences obtained by DSD-based scores[13] are also significant (p-value 0.0160). In contrast, scores obtained by the method of Guney et al[10] are not significantly different between the two groups of drugs.

Clinical Trial evaluation

Of the 126 BSAs in the drug-virus dataset, 28 are in clinical trials (see Methods). Prediction scores by our matrix decomposition method are significantly higher for drugs in clinical trials (Wilcoxon–Mann–Whitney p-value = 1.22e−5, see Figure 5c).

Of the 1853 FDA approved drugs considered by our network medicine approach, 153 are in clinical trials. For most of the methods (including all the baseline approaches), drugs in clinical trials do not have prediction scores significantly different than the remaining drugs. The kernel-based scores are significantly higher only for the p-Step and diffusion kernels (Wilcoxon–Mann–Whitney p-values 2.86e−2, and 7.89e−3 respectively, see Figure 5d).

Connectivity Map evaluation

For the network medicine approach, we queried CMAP[17] to obtain a list of 23 FDA-approved drugs that show a reverse expression profile than the one expressed by SARS-CoV-2 infected cells with a τ score between -90 and -100 (see Methods). We observe that the scores assigned by the kernel to FDA-approved drugs with strongly negative CMAP correlation are significantly different than those assigned to the remaining set of FDA-approved drugs (Wilcoxon–Mann–Whitney p-values 1.65e−3 for the Commute Time kernel, 2.43e−2 for the Diffusion Kernel, 2.48e−3 for the p-Step kernel, 1.97e−3 for the Regularised Laplacian kernel, and 2.82e−3 for the Inverse Cosine kernel), as illustrated in Figure 5e. Scores assigned by DSD also achieve a significant difference between the two groups (1.36e−3). However, scores obtained using the Guney Distance are not significantly different between the two groups (1.35e−1).

Assessing the importance of host proteins

To evaluate the effect of prioritising host proteins in our network medicine approach, we compared results based on weighted host proteins to unweighted/binary host proteins. For most kernels, we observe that the Wilcoxon-Mann-Whitney p-values are smaller (more significant) when using weighted host proteins when compared to considering all host proteins equally. This result is consistent for the three sources of evidence we consider (see Supplementary Tables 1, 2, and 3). Furthermore, to check whether our network approach has consistent results using an alternative interactome, we re-computed the kernel-based scores using the recently released HuRI PPI[41]. For every kernel, FDA-approved drugs with *in vitro*, clinical trials, and CMAP evidence have a significantly higher prediction score than the remaining drugs (see Supplementary Notes 1.1, 1.2, and 1.3).

We searched the literature for other *in silico*, *in vitro*, and clinical trials evidences for COVID-19 for the top 20 predictions. The *in silico* evidences include findings in the literature and CMAP scores. Interestingly, 9 out of the top 10 BSAs ranked by our first approach (Fapiravir, Arbidol, Lopinavir, Ritonavir, Tenofovir, Lamivudine, Remdesivir, and Hydroxychloroquine) are in clinical trials, and 3 of those (Fapiravir, Arbidol, and Hydroxychloroquine) have evidence from *in vitro* experiments that point toward efficacy against SARS-CoV-2. Also, 4 out of the top 10 FDA approved drugs ranked by our network medicine approach (Fostamatinib, Copper, Zinc, and Resveratrol) are in clinical trials for COVID-19 (see Supplementary Tables 4 and 5).

CoREx: the COVID-19 Repositioning Explorer

The volume of research and information related to COVID-19 is rapidly changing as more data is collected and analysed. To help researchers evaluate drug repositioning hypotheses, we have developed the COVID-19 Repositioning Explorer (CoREx). Given a set of drug targets, CoREx offers the users a panoramic point of view that puts together several biologically relevant contexts (i.e. functional relationships, protein-protein interactions, clinical trial status, CMAP scores, and ATC categories). Our goal is to assist researchers to quickly observe interesting drug alternatives, combinations, and mechanisms of actions that might play a key role in understanding and fighting COVID-19 by analysing the interplay between drug targets and host proteins in these different contexts.

Centred around ideas from network medicine, CoREx provides two different tools. First, an interactome analysis tool, which highlights the impact that a drug has on the host proteins. Second, a functional analysis tool, which shows the interplay in the context of functional modules.

In the interactome tool, the SARS-CoV-2 host protein subnetwork is used to visualise the impact of a ranked list of drugs. When a drug is selected, each node (host protein) is coloured based on the strength of the resulting kernel score. This tool is inspired on CoVex, by Sadegh et al.[42], where they study the interplay between the virus-host-drug triad using paths on the interactome. Our interactome tool complements this by calculating the effects that drugs have on individual host proteins through the different graph kernels. We have preloaded this tool with FDA approved drug with available drug targets in DrugBank[43]. An important feature of this tool is that users can submit a list of drug targets, and visualise the impact that a drug (or drug combination) with those targets will have on the host proteins subnetwork.

The functional tool puts together several interrelated datasets. A functional interactome is built by integrating protein-protein networks available in the STRING database[44] in a way that maximises the probability that two interacting proteins share

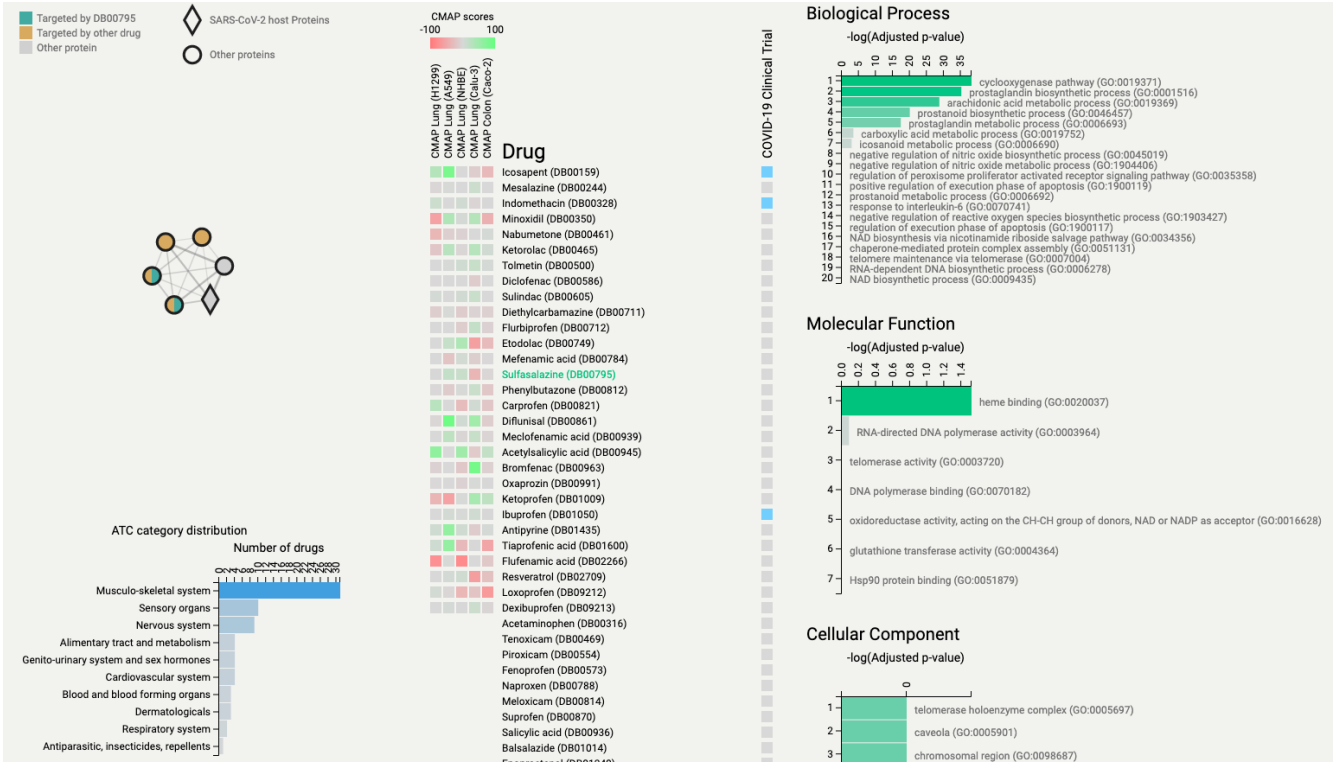

Figure 6. Screenshot of a CoREx functional module for Sulfasalazine (highlighted in green in the "Drug" list). The module is depicted as a network on the top left. Host proteins are depicted with diamonds, and drug targets are coloured. A list of drugs with at least one target in the functional module is located in the centre of the image, alongside CMAP scores for 5 cell lines on the left, and an indicator of whether the drug is in clinical trials on the right. The barplots on the far right part of the image correspond to the functional enrichment scores for each GO domain. Finally, the barplot on the bottom left section of the image gives information on the ATC categories of the involved drugs.

functional characteristics (see Supplementary Note 3 for details on the network combination). Then, we use the ClusterONE algorithm[45] to identify highly cohesive groups of proteins that contain at least one host protein, and at least one drug target. These functional modules are then enriched with protein function using Enrichr[46], and all the drugs that interact with the module through their targets are enriched with their ATC categories, CMAP evidence, and clinical trial status against COVID-19 (see Figure 6).

CoREx is available at `https://paccanarolab.org/corex` and supporting datasets are updated every 2 weeks.

Discussion

The development of computational approaches that can assist in the rational and fast discovery of treatments is critical for emergent infectious diseases such as COVID-19. Drug repositioning, the re-use of old drugs on the market, can help to speed up the development of such treatments by prioritising known-safe-in-human approved drugs for clinical trials involving COVID-19 patients. Here we proposed two computational approaches that prioritise drugs based on their efficacy against SARS-CoV-2 together with a human-in-the-loop website tool, CoREx, to assist current research efforts of finding suitable drugs with therapeutic efficacy against SARS-CoV-2.

The objective function of our first approach in Equation (1) is inspired on our recent model to predict the frequencies of drug side effects[7]. The main feature of this model is that it can account for varying levels of uncertainties in the data. We realised that the different levels of drug developmental evidence provided in the drug-virus dataset of Andersen et al.[18] inherently introduces varying levels of confidence in the information provided in the drug-virus associations. This prior knowledge allow us to weight the importance of each set of entries in the drug-virus matrix Y during the learning by setting different α_s. Moreover, because associations in Y represent different stages of drug efficacy, the output of our matrix decomposition method can be interpreted as modelling the efficacy of BSAs against viruses.

It is important to notice that the majority of the 126 BSAs target viral proteins. In contrast, in our network medicine

approach, we focus on host-targeting drugs. As host proteins are not controlled by the viral genome, targeting them can be a good strategy[47]. Also, drugs that target host proteins tend to exhibit broad-spectrum antiviral activities. Examples of FDA approved host targeting antivirals include Tromantadine, and Perginterferon alfa-2b.

Our network medicine approach is able to prioritise FDA approved candidates based on their network-based effects on the COVID-19 protein module. Results are aligned to recently reported to *in vitro* effects[13,40]. Our kernel scores show very similar performance patterns when tested with multiple kernels and across multiple interactomes (see Supplementary Note 1). In contrast to our first approach, our network medicine approach does not explicitly model the therapeutic efficacy of drugs, but their mechanistic effects on the protein interaction network. This means that a high score does not directly imply therapeutic effects, but only a high probability of molecular interplay between the drug and the SARS-CoV-2 host protein subnetwork. Our experiments with weighting this subnetwork in h show that we can focus our model to particular proteins that play a key tole in the infection, potentially identifying therapy related host proteins.

We have shown that both of our approaches are aligned to ongoing *in vitro* experiments and clinical trial studies. It is important to note that these two areas allow us to place our predictions in the light of ongoing research, but can not ultimately determine if a BSA or approved drug will have therapeutic effects without further experimentation. Our third area of evaluation involving CMAP can be integrated into the prediction pipeline to further refine the results.

Our computational approaches leverage available data to produce the predictions. As more reliable data becomes available, we expect the performance of our models to increase accordingly. We see a lot of opportunity for the integration of the expertise of experimentalists and these computational methods as accelerators of drug repositioning for infectious diseases. We believe that human-in-the-loop systems, where computational tools provide researchers with evidence driven prioritisation of repositioning candidates are already accelerating the development of effective and safe therapies[11,42,48].

Finally, the integration of heterogeneous sources of information with multiple layers of interconnection is a challenge in itself. Prime examples of such complex data are the molecular datasets involved in drug repositioning. We built CoREx with the goal of providing the research community with a tool that highlights potential repositioning candidates based on multiple entities that ultimately impact the human interactome.

Methods

Datasets

- *The drug-virus dataset*. We used the dataset curated by Andersen et al.[18]. Drugs are mapped to DrugBank IDs, when available. Each drug-virus association is annotated with their developmental status/stage. There are eight stages of development in the dataset, namely: cell culture/co-culture, primary cells/organoids, animal model, clinical trials phase I, phase II, phase III, phase IV, and approved. In total, our dataset contains 850 associations between 126 broad-spectrum antivirals and 80 viruses.

- *Protein-interaction network*. The PPI network was obtained from Cheng et al.[11], which contains 15,646 human proteins, and 218,015 interactions. Only reviewed proteins included in UniProtKB/Swiss-Prot[49] were used in our analysis.

- *FDA approved drugs and drug targets*. FDA approved drugs and their drug targets were retrieved from DrugBank[43]. 1853 FDA approved drugs with targets in UniProtKB/Swiss-Prot constitute our set of drugs.

- *Host proteins*. 332 host proteins reported by Gordon et al.[14] constitute our set of SARS-CoV-2 host proteins.

- *SARS-CoV-2 gene expression*. The SARS-CoV-2 gene expression data used to weight the host proteins, and to query on CMAP was obtained from Blanco-Melo et al[34]. It corresponds to RNAseq data from independent biological triplicates of transformed lung alveolar (A549) cells that were mock treated or infected with SARS-CoV-2. We measure the differential expression by the log fold change between the expression levels of SARS-CoV-2 infected, and mock treated cell lines. The results from our query to CMAP was mapped to DrugBank[43] (see Supplementary Data 1).

- *In vitro data*. We built a binary dataset, assigning positive labels to drugs that were reported to show efficacy against SARS-CoV-2 infection *in vitro*, and negative labels to all other drugs. Data for drug efficacy *in vitro* was built as the union of experiments reported by Riva et al.[40] and Gysi et al.[13]. 78 FDA approved drugs show *in vitro* effects (see Supplementary Data 2).

- *Clinical trials data*. We built a binary dataset and assigned positive labels to drugs that are involved in clinical trial studies on phase 2 or later, and negative labels to all other drugs. Information for clinical trials studies was retrieved from ClinicalTrials.gov[50]. Drugs were mapped to the DrugBank database[43] by matching their names (see Supplementary Data 3)

Matrix decomposition model

We modelled the estimation of the efficacy of a drug-virus association as a linear combination of drugs and virus feature vectors, that we called signatures. To predict a drug-virus score, each of the k components in the drug signature is multiplied by the corresponding component in the virus signatures, and then the products are summed together. Thus the efficacy score of a drug i against a virus j can be expressed as a combination of k components, as follow,

$$\hat{y}_{ij} = \sum_{\mu=1}^{k} P_{i\mu} Q_{\mu j} = \mathbf{p}_i \cdot \mathbf{q}_j \tag{3}$$

where $P_{i\mu}$ indicates the (i,μ) element of matrix P whose row i, $\mathbf{p}_i$, is the signature for drug i and $Q_{\mu j}$ indicates the (μ, j) element of matrix Q whose column j, $\mathbf{q}_j$, is the signature for virus j.

The objective function and decomposition algorithm

Let Y denote our $n \times m$ drug-virus matrix for n drugs and n viruses, where $Y_{ij} = 1$ if the drug i is associated with virus j or $Y_{ij} = 0$ otherwise. Our matrix decomposition model approximates the data matrix Y by the product of two matrices, as follow:

$$\hat{Y} = PQ \tag{4}$$

where P is the $n \times k$ matrix of drug signatures (each row contains a drug feature vector) and Q is the $k \times m$ matrix of virus signatures (each column contains a virus feature vector). k is the number of features assigned to each drug and each virus, also known as the rank of $\hat{Y}$. To learn the matrices P and Q in Eq. (4), we minimise the following loss:

$$\min_{P,Q\geq 0} \mathscr{L}(P,Q) = \frac{1}{2} \sum_{s\in\{A,B,C,D,E,Z\}} \alpha_s \left\| M^s \circ (Y - PQ) \right\|_F^2 \tag{5}$$

subject to non-negative constraints $P, Q \geq 0$.

where $\|.\|_F^2$ is the Frobenius norm of a matrix, $\circ$ indicates element-wise matrix multiplication, $\{A,B,C,D,E,Z\}$ are a partition of the set of entries in Y, A corresponds to approved and phase IV drug-virus associations, B,C and D to clinical trials phase I, II and III, E to evidence on animal models, cell culture/co-culture or organoids experiments, and Z represents the subset of zero entries in Y. α_s are constant values to assign distinct levels of confidence or probability of success to each subset in the partition. α_A was set to 1. M^s is $n \times m$ projection matrix defined for each subset in the partition $\{A,B,C,D,E,Z\}$ where, for instance, $M^A_{ij} = 1$ if $(i,j) \in A$ or $M^A_{ij} = 0$ otherwise.

To minimise Equation (5) subject to non-negative constraints, we developed an efficient multiplicative learning algorithm inspired by the diagonally rescaled principle of non-negative matrix factorization[23]. The algorithm consists of iteratively applying the following multiplicative update rules:

$$\begin{aligned}
P_{ia} &\leftarrow P_{ia} \frac{\left(\left[M^A \circ Y + \sum_{s\in\{B,C,D,E\}} \alpha_s (M^s \circ Y) \right] Q^T \right)_{ia}}{\left(\left[M^A \circ (PQ) + \sum_{s\in\{B,C,D,E\}} M^s \circ (PQ) + \alpha_z M^z \circ (PQ) \right] Q^T \right)_{ia}} \\
Q_{aj} &\leftarrow Q_{aj} \frac{\left(P^T \left[M^A \circ Y + \sum_{s\in\{B,C,D,E\}} \alpha_s (M^s \circ Y) \right] \right)_{aj}}{\left(P^T \left[M^A \circ (PQ) + \sum_{s\in\{B,C,D,E\}} M^s \circ (PQ) + \alpha_z M^z \circ (PQ) \right] \right)_{aj}}
\end{aligned} \tag{6}$$

Following the guidelines to implement NMF[51], a small number $\varepsilon = 10^{-8}$ was added to the denominators in Eq. 6 to prevent division by zero, and we initialised P and Q as random dense matrices uniformly distributed in the range $[0, 0.1]$. Furthermore, to avoid the well-known degeneracy[21] associated with the invariance PQ under the transformation $P \to P\Lambda$ and $Q \to \Lambda^{-1}Q$, for a diagonal matrix Λ, we normalised P at each iteration as follows:

$$Q_{aj} \leftarrow \frac{Q_{aj}}{\|\mathbf{q}_a\|} \tag{7}$$

where $\mathbf{q}_a$ denotes the ath row vector of Q.

The stopping criteria of our algorithm was based on the maximum tolerance of the relative change in the elements of P and Q. The default value was $tolX < 10^{-3}$, which occurred typically in about 1000 iterations for $k = 5$.

Using a similar procedure to[7], it can be easily shown that our algorithm in Equation (6) satisfies the KKT conditions of convergence.

Cross-validation procedure and model selection

We used leave-one-out cross validation (LOOCV) procedure to evaluate the performance of the model. To set the model hyperparameters: k, α_E and α_Z, we performed LOOCV on the clinical trials associations phase I, II and III (validation set). We performed a grid-search and selected the set of hyperparameters that maximise the mean recall across the top $1,5,10,15,20,25$ and 30 predictions retrieved. We found that $k = 5$, $\alpha_E = 0.01$ and $\alpha_Z = 2$ provided a good performance. The other hyperparameters of our model were set based on the probabilities of success reported by Dowden and Munro[19] for antiinfective drugs on distinct phases of clinical trials, i.e., $\alpha_B = 0.16$ (phase I), $\alpha_C = 0.27$ (phase II) and $\alpha_D = 0.71$ (phase III). Having set all these hyperparameters, we performed another LOOCV on the approved and phase IV associations, corresponding to effective drug-virus associations.

Graph kernels

A PPI network is represented by a graph $G = (V,E)$, in which $V = \{1,2,\ldots,n_V\}$ is the set of nodes (proteins), and E is a set of links connecting the nodes (protein interactions). If the graph is weighted, then for each edge $e \in E$, we associate a non-negative real value $w(e)$. Let $\mathscr{H} \in V$ denote the set of host proteins. Our goal is then to perturb the sub-network induced by $\mathscr{H}$, i.e. the host protein sub-network.

Here we rely on different graph kernels described in the literature[28,29,52]. In the following, graph kernels and their properties are defined as in Kondor and Vert[53]. A graph kernel $k : V \times V \mapsto \mathbb{R}$ provides a similarity metric on the set of nodes V based on the graph structure. It is positive definite, that is, for any $i, j \in V$ and any $c_i, c_j \in \mathbb{R}$, we have that $\sum_{i=1}^{n_V} \sum_{j=1}^{n_V} c_i c_j k(i,j) \geq 0$.

We can use it to define distances/similarities on a latent feature space. More specifically, there exists the feature mapping $\phi : V \mapsto \mathscr{F}$ such that $k(i,j) = \langle \phi(i), \phi(j) \rangle$, for all $i, j \in V$.

A graph kernel can be represented by an $n_V \times n_V$ matrix K whose elements correspond to $K_{i,j} = k(i,j)$ for every $i, j \in V$. It is usually defined in terms of the Normalised Laplacian, which we explain below.

Let W be an $n_V \times n_V$ matrix denoting the weighted adjacency matrix of G. That is, $W_{i,j} = w(e)$, if there is an edge e connecting i and j, and $W_{i,j} = 0$, otherwise. If G is unweighted, we assume that $w(e) = 1$ for every edge $e \in E$. Let D denote an $n_V \times n_V$ diagonal matrix in which each diagonal element corresponds to the node degree, that is, $D_{i,i} = \sum_{j=1}^{n_V} W_{i,j}$ for every $i \in V$. The Laplacian is defined as $D - A$, and its pseudoinverse (Moore–Penrose inverse) is denoted by L^+. The Normalised Laplacian is defined as $\tilde{L} := I - D^{-\frac{1}{2}} W D^{-\frac{1}{2}}$, where I denotes the identity matrix.

There are different ways to define K and we focus on three popular graph kernels[28,29,52]. Table 2 shows the definitions of the kernels used: Regularised Laplacian, Diffusion Process, and p-Step Random Walk in terms of the normalised Laplacian[29].

Kernel	Formula
p-Step Random Walk	$K = (aI - \tilde{L})^p$
Diffusion Process	$K = \exp(-\sigma^2/2\tilde{L})$
Regularised Laplacian	$K = (I + \sigma^2\tilde{L})^{-1}$
Commute time kernel	$K = L^+$
Inverse cosine	$\cos \tilde{L}\pi/4$

Table 2. Graph Kernels. Definition Graph Kernels based on the normalised Laplacian.

In the p-Step Random Walk, $p \geq 1$ and $a \geq 2$ are given parameters[29]. The element $K_{i,j}$ measures how likely it is to go from node i to node j after p steps in a random walk. If we generalise it to a continuous time (infinitesimally small steps) and take an infinite number of steps, we have the Diffusion Process $K = \exp(-\sigma^2/2\tilde{L})$, where σ is a parameter controlling the diffusion. Finally, the Regularised Laplacian Kernel can be thought as the convergence of an iterative process in which nodes spread information to their neighbours at each step.

We used different kernels from[28,29,52], which are implemented in the R package *diffuStats*[54].

Connectivity Map Evaluation details

Following the suggestions by Sirota et al[35], we consider that a drug is a good candidate for COVID-19 if the changes that it causes to gene expression are opposite to the ones caused by the disease. To measure how similar or opposite the drug and

COVID-19 expression profiles are, we used the Connectivity Map (CMAP) pipeline[17].

We began by obtaining a list of up/down-regulated genes in COVID-19 (genes that have higher/lower expression levels in SARS-CoV-2 infected cells compared to non-infected cells). Then, we queried the COVID-19 signature in CMAP. For each drug, CMAP has a list of genes ordered from the most expressed to the least expressed after treatment (in comparison to the expression levels with no treatment). If the up-regulated genes in COVID-19 are located on the bottom of the list (that is, if they have low expression levels in cells treated with the drug), and the down-regulated genes are located on the top (that is, they have high expression levels in cells treated with the drug), we say that the drug and disease signatures have a strong negative correlation. If we observe the opposite (up-regulated genes on top, and down-regulated genes on bottom), we say that they have a strong positive correlation.

For each drug, CMAP outputs an enrichment score that is positive when the correlation between the drug and disease signatures is positive (the drug mimics the disease), and negative when the correlation is negative (the drug reverses the disease). The final values (denoted by τ) are compared to a reference database and normalised between -100 and 100.

Author contributions statement

M.T., S.S., D.G., L.C., and A.P. conceived the study, designed the methods, and analysed the results. M.T., S.S., and D.G. implemented and conducted the experiments. M.S. implemented the dynamic components of CoREx. All authors reviewed the manuscript.

References

1. Wu, F. *et al.* A new coronavirus associated with human respiratory disease in China. *Nature* **579**, 265–269, DOI: 10.1038/s41586-020-2008-3 (2020). Number: 7798 Publisher: Nature Publishing Group.
2. World Health Organization. Coronavirus disease (covid-2019) situation reports (2020). [Online; accessed 10-December-2020].
3. Food and Drug Administration. Emergency use authorization | fda (2021). [Online; accessed 12-February-2021].
4. Chen, W.-H., Strych, U., Hotez, P. J. & Bottazzi, M. E. The SARS-CoV-2 Vaccine Pipeline: an Overview. *Curr. Trop. Medicine Reports* **7**, 61–64, DOI: 10.1007/s40475-020-00201-6 (2020).
5. Ashburn, T. T. & Thor, K. B. Drug repositioning: Identifying and developing new uses for existing drugs. *Nat. Rev. Drug Discov.* **3**, 673–683 (2004).
6. Pushpakom, S. *et al.* Drug repurposing: progress, challenges and recommendations. *Nat. Rev. Drug Discov.* **18**, 41–58 (2018).
7. Galeano, D. & Paccanaro, A. Predicting the frequency of drug side effects. *bioRxiv* 594465 (2019).
8. Sosnina, E. A. *et al.* Recommender systems in antiviral drug discovery. *ACS omega* **5**, 15039–15051 (2020).
9. Barabási, A.-L., Gulbahce, N. & Loscalzo, J. Network Medicine: A Network-based Approach to Human Disease. *Nat Rev Genet.* **12**, 56–68, DOI: 10.1038/nrg2918 (2011).
10. Guney, E., Menche, J., Vidal, M. & Barábasi, A.-L. Network-based in silico drug efficacy screening. *Nat. Commun.* **7**, 10331, DOI: 10.1038/ncomms10331 (2016).
11. Cheng, F. *et al.* Network-based approach to prediction and population-based validation of in silico drug repurposing. *Nat. Commun.* **9**, 1–12, DOI: 10.1038/s41467-018-05116-5 (2018). Number: 1 Publisher: Nature Publishing Group.
12. Zhou, Y. *et al.* Network-based drug repurposing for novel coronavirus 2019-ncov/sars-cov-2. *Cell discovery* **6**, 1–18 (2020).
13. Gysi, D. M. *et al.* Network medicine framework for identifying drug repurposing opportunities for covid-19. *arXiv preprint arXiv:2004.07229* (2020).
14. Gordon, D. E. *et al.* A SARS-CoV-2 protein interaction map reveals targets for drug repurposing. *Nature* **583**, 459–468, DOI: 10.1038/s41586-020-2286-9 (2020).
15. Hoffmann, M. *et al.* SARS-CoV-2 Cell Entry Depends on ACE2 and TMPRSS2 and Is Blocked by a Clinically Proven Protease Inhibitor. *Cell* **181**, 271–280.e8, DOI: 10.1016/j.cell.2020.02.052 (2020). Publisher: Elsevier.
16. Zhou, Y., Wang, F., Tang, J., Nussinov, R. & Cheng, F. Artificial intelligence in COVID-19 drug repurposing. *The Lancet Digit. Heal.* **0**, DOI: 10.1016/S2589-7500(20)30192-8 (2020). Publisher: Elsevier.

17. Subramanian, A. *et al.* A Next Generation Connectivity Map: L1000 Platform and the First 1,000,000 Profiles. *Cell* **171**, 1437–1452.e17, DOI: 10.1016/j.cell.2017.10.049 (2017).

18. Andersen, P. I. *et al.* Discovery and development of safe-in-man broad-spectrum antiviral agents. *Int. J. Infect. Dis.* (2020).

19. Dowden, H. & Munro, J. Trends in clinical success rates and therapeutic focus. *Nat. reviews. Drug discovery* **18**, 495 (2019).

20. Galeano, D., Li, S., Gerstein, M. & Paccanaro, A. Predicting the frequencies of drug side effects. *Nat. Commun.* **11**, 1–14 (2020).

21. Lee, D. D. & Seung, H. S. Learning the parts of objects by non-negative matrix factorization. *Nature* **401**, 788–791 (1999).

22. Cremonesi, P., Koren, Y. & Turrin, R. Performance of recommender algorithms on top-n recommendation tasks. In *Proceedings of the fourth ACM conference on Recommender systems*, 39–46 (2010).

23. Lee, D. D. & Seung, H. S. Algorithms for non-negative matrix factorization. In *Advances in neural information processing systems*, 556–562 (2001).

24. Cáceres, J. J. & Paccanaro, A. Disease gene prediction for molecularly uncharacterized diseases. *PLoS computational biology* **15**, e1007078 (2019).

25. Silverman, E. K. *et al.* Molecular networks in Network Medicine: Development and applications. *WIREs Syst. Biol. Medicine* **12**, e1489, DOI: https://doi.org/10.1002/wsbm.1489 (2020).

26. Yıldırım, M. A., Goh, K.-I., Cusick, M. E., Barabási, A.-L. & Vidal, M. Drug—target network. *Nat. biotechnology* **25**, 1119–1126 (2007).

27. Hopkins, A. L. Network pharmacology. *Nat. biotechnology* **25**, 1110–1111 (2007).

28. Cao, M. *et al.* Going the Distance for Protein Function Prediction: A New Distance Metric for Protein Interaction Networks. *PLOS ONE* **8**, e76339, DOI: 10.1371/journal.pone.0076339 (2013). Publisher: Public Library of Science.

29. Smola, A. J. & Kondor, R. Kernels and Regularization on Graphs. In Schölkopf, B. & Warmuth, M. K. (eds.) *Learning Theory and Kernel Machines*, Lecture Notes in Computer Science, 144–158, DOI: 10.1007/978-3-540-45167-9_12 (Springer, Berlin, Heidelberg, 2003).

30. Vishwanathan, S. V. N., Schraudolph, N. N., Kondor, R. & Borgwardt, K. M. Graph Kernels. *J. Mach. Learn. Res.* **11**, 1201–1242 (2010).

31. Re, M. & Valentini, G. Cancer module genes ranking using kernelized score functions. *BMC Bioinforma.* **13**, S3, DOI: 10.1186/1471-2105-13-S14-S3 (2012).

32. Re, M., Mesiti, M. & Valentini, G. A Fast Ranking Algorithm for Predicting Gene Functions in Biomolecular Networks. *IEEE/ACM Transactions on Comput. Biol. Bioinforma.* **9**, 1812–1818, DOI: 10.1109/TCBB.2012.114 (2012).

33. Yan, R. *et al.* Structural basis for the recognition of sars-cov-2 by full-length human ace2. *Science* **367**, 1444–1448 (2020).

34. Blanco-Melo, D. *et al.* Imbalanced Host Response to SARS-CoV-2 Drives Development of COVID-19. *Cell* **181**, 1036–1045.e9, DOI: 10.1016/j.cell.2020.04.026 (2020).

35. Sirota, M. *et al.* Discovery and preclinical validation of drug indications using compendia of public gene expression data. *Sci Transl Med* **3**, 96ra77, DOI: 10.1126/scitranslmed.3001318 (2011).

36. Food and Drug Administration. Clinical research (2020). [Online; accessed 25-July-2020].

37. Food and Drug Administration. Drug development process (2020). [Online; accessed 25-July-2020].

38. Mullol, J. *et al.* The Loss of Smell and Taste in the COVID-19 Outbreak: a Tale of Many Countries. *Curr. Allergy Asthma Reports* **20**, 61, DOI: 10.1007/s11882-020-00961-1 (2020).

39. Bansal, M. Cardiovascular disease and COVID-19. *Diabetes & Metab. Syndr. Clin. Res. & Rev.* **14**, 247–250, DOI: 10.1016/j.dsx.2020.03.013 (2020).

40. Riva, L. *et al.* Discovery of SARS-CoV-2 antiviral drugs through large-scale compound repurposing. *Nature* 1–11, DOI: 10.1038/s41586-020-2577-1 (2020). Publisher: Nature Publishing Group.

41. Luck, K. *et al.* A reference map of the human binary protein interactome. *Nature* **580**, 402–408, DOI: 10.1038/s41586-020-2188-x (2020).

42. Sadegh, S. *et al.* Exploring the SARS-CoV-2 virus-host-drug interactome for drug repurposing. *Nat. Commun.* **11**, 3518, DOI: 10.1038/s41467-020-17189-2 (2020). Number: 1 Publisher: Nature Publishing Group.

43. Wishart, D. S. *et al.* DrugBank 5.0: a major update to the DrugBank database for 2018. *Nucleic Acids Res.* **46**, D1074–D1082, DOI: 10.1093/nar/gkx1037 (2018).

44. Szklarczyk, D. *et al.* STRING v11: protein–protein association networks with increased coverage, supporting functional discovery in genome-wide experimental datasets. *Nucleic Acids Res.* **47**, D607–D613, DOI: 10.1093/nar/gky1131 (2019). Publisher: Oxford Academic.

45. Nepusz, T., Yu, H. & Paccanaro, A. Detecting overlapping protein complexes in protein-protein interaction networks. *Nat. Methods* **9**, 471–472, DOI: 10.1038/nmeth.1938 (2012).

46. Kuleshov, M. V. *et al.* Enrichr: a comprehensive gene set enrichment analysis web server 2016 update. *Nucleic Acids Res.* **44**, W90–W97, DOI: 10.1093/nar/gkw377 (2016).

47. Ji, X. & Li, Z. Medicinal chemistry strategies toward host targeting antiviral agents. *Medicinal Res. Rev.* **40**, 1519–1557, DOI: 10.1002/med.21664 (2020).

48. Zhou, Y. *et al.* Network-based drug repurposing for novel coronavirus 2019-nCoV/SARS-CoV-2. *Cell Discov.* **6**, 1–18, DOI: 10.1038/s41421-020-0153-3 (2020). Number: 1 Publisher: Nature Publishing Group.

49. Consortium, T. U. UniProt: a worldwide hub of protein knowledge. *Nucleic Acids Res.* **47**, D506–D515, DOI: 10.1093/nar/gky1049 (2019).

50. of Medicine, U. N. L. listed clinical studies related to the coronavirus disease (covid-19) (2020). [Online; accessed 1-December-2020].

51. Berry, M. W., Browne, M., Langville, A. N., Pauca, V. P. & Plemmons, R. J. Algorithms and applications for approximate nonnegative matrix factorization. *Comput. statistics & data analysis* **52**, 155–173 (2007).

52. Zhou, D., Bousquet, O., Lal, T. N., Weston, J. & Schölkopf, B. Learning with Local and Global Consistency. 8 (2004).

53. Kondor, R. & Vert, J.-P. Diffusion Kernels. In Schölkopf, B., Tsuda, K. & Vert, J.-P. (eds.) *Kernel Methods in Computational Biology*, DOI: 10.7551/mitpress/4057.003.0011 (The MIT Press, 2004).

54. Picart-Armada, S., Thompson, W. K., Buil, A. & Perera-Lluna, A. diffuStats: an R package to compute diffusion-based scores on biological networks. *Bioinformatics* **34**, 533–534, DOI: 10.1093/bioinformatics/btx632 (2018).