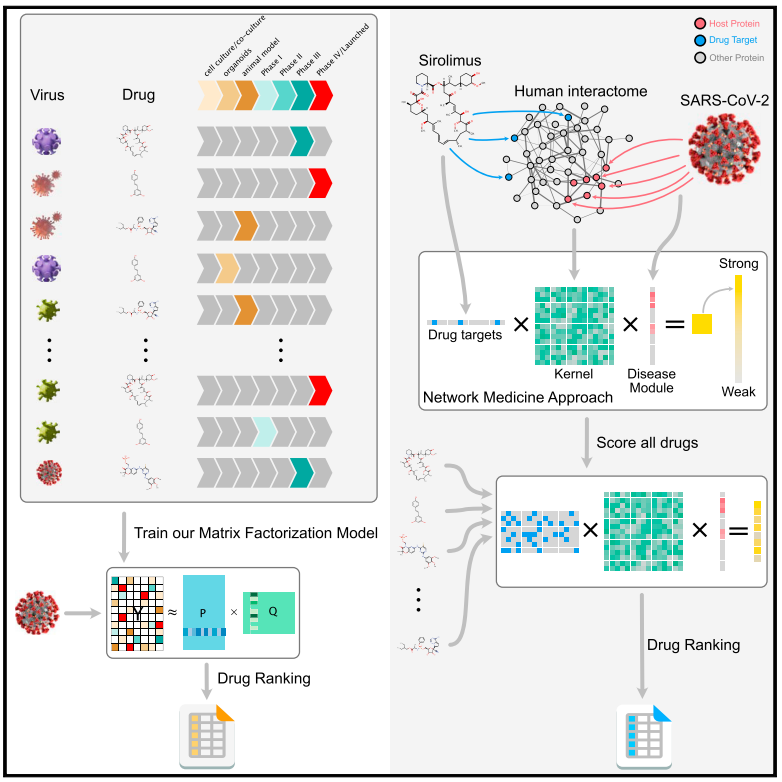
Article

Patterns

Machine learning and network medicine approaches for drug repositioning for COVID-19

Graphical abstract



Highlights

- A matrix decomposition model for repurposing broad-spectrum antivirals
- A graph kernel approach to model perturbations induced by drugs on the interactome
- Graph kernels can integrate transcriptomics data to improve drug repurposing
- CoREx: a free online tool to formulate hypothesis for drug repurposing for COVID-19

Authors

Suzana de Siqueira Santos, Mateo Torres, Diego Galeano, María del Mar Sánchez, Luca Cernuzzi, Alberto Paccanaro

Correspondence

alberto.paccanaro@rhul.ac.uk

In brief

We present two complementary machine learning approaches for drug repositioning against COVID-19 that target SARS-CoV-2 and its cellular processes in the host, respectively. Our matrix decomposition approach exploits drug developmental information to predict broad-spectrum antivirals; our graph kernel-based approach, rooted in ideas from network medicine, predicts which FDA-approved drugs are more likely to perturb the human subnetwork that is crucial for SARS-CoV-2 infection/replication. We also introduce CoREx, a freely available online tool to reason and formulate hypothesis about drug repurposing in the context of biological networks and pharmacological information.

Santos et al., 2022, Patterns *3*, 100396
January 14, 2022 © 2021
https://doi.org/10.1016/j.patter.2021.100396

CellPress

Please cite this article in press as: Santos et al., Machine learning and network medicine approaches for drug repositioning for COVID-19, Patterns (2021), https://doi.org/10.1016/j.patter.2021.100396

Patterns

CellPress
OPEN ACCESS

Article

Machine learning and network medicine approaches for drug repositioning for COVID-19

Suzana de Siqueira Santos,[1,5,6] Mateo Torres,[1,5,6] Diego Galeano,[1,3,5,6] María del Mar Sánchez,[4] Luca Cernuzzi,[4] and Alberto Paccanaro[1,2,5,6,7,*]

[1]Escola de Matemática Aplicada, Fundação Getulio Vargas, Rio de Janeiro 22250-900, Brazil
[2]Department of Computer Science, Centre for Systems and Synthetic Biology, Royal Holloway, University of London, Egham Hill, Egham TW20 0EX, UK
[3]Facultad de Ingenieria, Universidad Nacional de Asunción, Luque 110948, Paraguay
[4]Universidad Católica "Nuestra Señora de la Asunción", Asunción C.C. 1683, Paraguay
[5]COVID-19 International Research Team
[6]These authors contributed equally
[7]Lead contact
*Correspondence: alberto.paccanaro@rhul.ac.uk
https://doi.org/10.1016/j.patter.2021.100396

THE BIGGER PICTURE The development timeline for treatments against emergent viral diseases can be significantly reduced by re-using drugs already available on the market—a concept known as drug repositioning. We present two complementary machine learning approaches for drug repositioning that target SARS- CoV-2 and host factors, respectively. Our matrix decomposition approach exploits drug developmental information to predict the effectiveness of broad-spectrum antiviral drugs. Our graph kernel-based approach, rooted in ideas from network medicine, predicts which FDA-approved drugs are more likely to perturb the human subnetwork that is crucial for SARS-CoV-2 infection/replication. We also introduce CoREx, a freely available online tool that enables scientists to reason and formulate hypotheses about drug repurposing in the context of biological networks and pharmacological information.
While we have developed these methodologies for COVID-19, our approaches can be applied to any viral disease.

Development/Pre-production: Data science output has been rolled out/validated across multiple domains/problems

SUMMARY

We present two machine learning approaches for drug repurposing. While we have developed them for COVID-19, they are disease-agnostic. The two methodologies are complementary, targeting SARS-CoV-2 and host factors, respectively. Our first approach consists of a matrix factorization algorithm to rank broad-spectrum antivirals. Our second approach, based on network medicine, uses graph kernels to rank drugs according to the perturbation they induce on a subnetwork of the human interactome that is crucial for SARS-CoV-2 infection/replication. Our experiments show that our top predicted broad-spectrum antivirals include drugs indicated for compassionate use in COVID-19 patients; and that the ranking obtained by our kernel-based approach aligns with experimental data. Finally, we present the COVID-19 repositioning explorer (CoREx), an interactive online tool to explore the interplay between drugs and SARS-CoV-2 host proteins in the context of biological networks, protein function, drug clinical use, and Connectivity Map. CoREx is freely available at: https://paccanarolab.org/corex/.

INTRODUCTION

Drug discovery and development present several challenges, including high attrition rates, long development times, and substantial costs.[1] Drug repositioning involves the use of de-risked compounds in humans, which translates into lower costs and shorter development times.[2] Computational methods can assist drug repurposing research projects by providing rankings of

This is an open access article under the CC BY-NC-ND license (http://creativecommons.org/licenses/by-nc-nd/4.0/).

Please cite this article in press as: Santos et al., Machine learning and network medicine approaches for drug repositioning for COVID-19, Patterns (2021), https://doi.org/10.1016/j.patter.2021.100396

drugs based on predicted therapeutic efficacy, as well as tools to help scientists reason about drug effectiveness by integrating diverse available biomedical knowledge.

Coronaviruses are notoriously difficult to manage, as there is no specific antiviral treatment that has been proven effective against the infections they induce.[3] Identifying commercially available drugs with therapeutic effects for COVID-19 could provide early treatment options until effective therapies become widely available. A growing corpus of literature identifies several categories of treatment that revolves around the use of drugs with a mode of action that targets the molecular structure of the virus (*virally targeted agents*), or its cellular processes in the host (*host-targeted agents*), or those based on combinatorial therapies.[4–7]

In this paper, we present two different machine learning approaches, and a webtool, for drug repurposing for COVID-19. Our first machine learning approach focuses on virally targeted agents and aims at ranking broad-spectrum antiviral (BSA) drugs. Given a small number of drugs associated with a virus, and their stage in the drug development process, our matrix decomposition algorithm assigns scores to a larger group of drugs with previously unknown associations with the virus. Our method predicts BSAs against SARS-CoV-2 by exploiting information about stages of drug development that are interpreted as probabilities of drug approval. To our knowledge, our matrix decomposition model is the first that integrates developmental-stage information to predict the efficacy of drugs against viral diseases, and we show that this is crucial to obtain better predictions.

Our second machine learning approach focuses on host-targeted agents, and prioritizes FDA-approved drugs based on ideas from network medicine.[8] In particular, it exploits the concept of a disease module, which has been instrumental in the prediction of disease genes for hereditary diseases.[9–12] For a virus, a disease module can be defined as the set of human proteins (hereafter, host proteins) that interact with viral proteins, allowing the infection and replication processes. Recently, Gysi et al.[13] have shown that, for SARS-CoV-2, most of the experimentally identified human host proteins[14] form a distinct COVID-19 disease module in the interactome. Our network medicine-based approach is based on the idea that the binding of drugs to their protein targets causes a perturbation that propagates through the interactome. By quantifying this perturbation, it is possible to calculate the extent of the effect that a drug induces on the COVID-19 disease module. Our method ranks FDA-approved drugs based on this effect, which is estimated using graph kernels. An important aspect of our method is that it offers a natural way to model the relative importance of host proteins for the disease, and we show that our network medicine approach benefits from this prioritization of host proteins.

Finally, we present the COVID-19 repositioning explorer (CoREx), an online tool that enables scientists to analyze and reason about drug repurposing in a functional context on the interactome and thus allows the exploration of our results as well as the formulation of novel repurposing hypotheses. CoREx integrates 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 could reverse the gene expression signature of SARS-CoV-2 infection based on the Connectivity Map (CMAP).[15,16]

RESULTS

A matrix decomposition model for antiviral discovery

Recently, Andersen et al.[17] published a dataset containing 850 associations between 126 BSA drugs and 80 viruses for which they have been approved or are under development. Importantly, each drug-virus association was manually curated and is annotated with its stage in the drug development process.

Figure 1 shows the number of drug-virus associations that corresponds to each developmental stage, as well as histograms of the associations grouped per drug and per virus. We notice that the associations are not uniformly distributed for viruses or drugs (Figure 1, left and right panels). This type of long-tailed distribution of entries has been previously observed in datasets that appear in the recommender system literature, such as Netflix or Movielens,[18] and we have recently exploited this property to build a recommender system based on matrix factorization for predicting drug side effect frequencies.[19]

Various types of recommender systems have recently been developed for different settings of the drug repositioning problem. A few methods are based on variations of the non-negative matrix factorization (NMF) algorithm,[20,21] such as the NMF with L2 regularization by Bakal et al.,[22] the TriFactor NMF by Ceddia et al.,[23] and the indicator-regularized non-negative matrix factorization (IRNMF) method by Tang et al.,[24] which was developed to repurpose drugs for COVID-19. Our aim is also to build a recommender system that recommends BSA drugs to viruses and the novelty of our approach lies in the realization that the stages of drug development for drug-virus associations can be related to the probability of reaching the final stage of drug development (hereafter, probability of success). This observation is motivated by the empirical evidence (e.g., Dowden and Munro[25]) that the probability of success of a candidate drug increases as the candidate drug moves to the next developmental stage in the drug development process. This led us to develop a novel objective function that models the probabilities of success of drug-virus associations using their stage in the drug development process. In this paper, we show how the integration of this type of information greatly improves prediction performance.

In recommender systems based on matrix decomposition, the fundamental assumption is that users and movies can be represented as latent feature vectors in a low-dimensional space, and that a rating value for a specific user-movie pair is obtained by the dot product of the corresponding feature vectors. In our context, each drug and each virus can be represented as low-dimensional feature vectors in a latent space such that the dot product between the vectors model effective drug-virus associations. Having collected all the associations in a binary matrix Y, where each entry $y_{ij} = 1$ if and only if drug i is associated to virus j in the Andersen et al.[17] dataset ($y_{ij} = 0$ otherwise), for each drug i we learn a low-dimensional feature vector $p_i \in \mathbb{R}^k$ (the *drug signature*) and for each virus j a low-dimensional feature vector $q_j \in \mathbb{R}^k$ (the *virus signature*), such that $y_{ij} \approx p_i^T q_j$. Therefore, our algorithm amounts to decomposing the $n \times m$ matrix Y into the product of two matrices $P \in \mathbb{R}^{n \times k}$ in which each row is a drug signature p_i^T,

Please cite this article in press as: Santos et al., Machine learning and network medicine approaches for drug repositioning for COVID-19, Patterns (2021), https://doi.org/10.1016/j.patter.2021.100396

CellPress
OPEN ACCESS

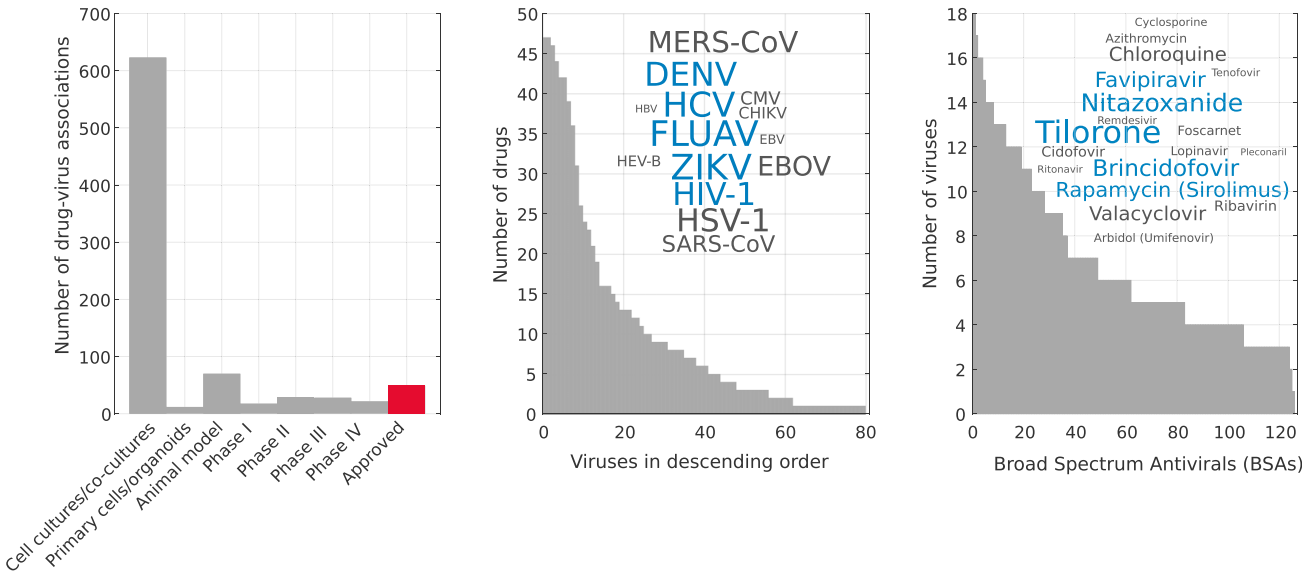


Figure 1. Drug-virus dataset statistics

We used the dataset manually curated by Andersen et al.[17] (Left) Number of drug-virus associations grouped 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 to clinical trials in humans (phases I–IV). It terminates with the approval of the drug for commercial use (in red). (Middle) Number of drugs (BSAs) associated to each virus in the dataset. Inset: the word cloud shows the 14 viruses with most associations. The size of the word is proportional to its number of associations and the five most popular viruses among drugs are colored blue. (Right) Number of viruses associated to each drug in the dataset. Inset: the word cloud shows the 18 drugs with most associations and the five most popular drugs among viruses are colored blue.

and $Q \in \mathbb{R}^{k \times m}$ in which each column is a virus signature q_j, and $k \ll \min(n, m)$. Indicating their product with $\widehat{Y}$, we have $Y \approx PQ = \widehat{Y}$. Matrices P and Q are learned by minimizing the following cost function:

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

$$\text{subject to non} - \text{negative constraints } P, Q \geq 0, \qquad \text{(Equation 1)}$$

where $\|.\|_F$ is the Frobenius norm of a matrix, $\circ$ is the element-wise (Hadamard) product, and the letters A, B, C, D, E indicate disjoint subsets of entries in Y that are defined according to the known developmental stages of drug-virus associations, as explained below. Let us now analyze Equation 1 to understand how the information about drug developmental stages is integrated into our system to model probabilities of success of drug-virus associations.

During learning, the drug-virus associations are divided into groups according to their stage of development. The first term in Equation 1 is the fitting constraint on the approved and phase IV drug-virus associations (set A). Matrix M^A is used to apply the summation only to entries in Y belonging to the set of approved associations A, being defined as: $M^A_{ij} = 1$ if drug i was approved or is in phase IV for virus j, or 0 otherwise. Thus, the first term of Equation 1 is attempting to find a decomposition PQ to reconstruct the associations in set A exactly. The second term in Equation 1 has an equivalent role for the remaining known associations in Y, corresponding to earlier stages in the drug development process—sets B, C, and D contain entries in clinical trials phases I, II, and III, respectively, while set E contains associations in *in vitro* and animal model stages. Here the corresponding M^s matrices are used to apply the summations only to entries belonging to the corresponding sets ($M^s_{ij} = 1$ if the entry $y_{i,j}$ belongs to set s). However, for these sets, their contributions to the loss are weighted differently using the parameters $\alpha_s \in [0, 1]$. These parameters have the key role of downweighting these terms in the minimization, in a way that reflects their higher uncertainty of success due to their earlier stage of drug development, thus effectively coding probabilities of success for each subset. Similarly, the third term in Equation 1 is used to down-weight the importance of the zero entries of Y while also serving

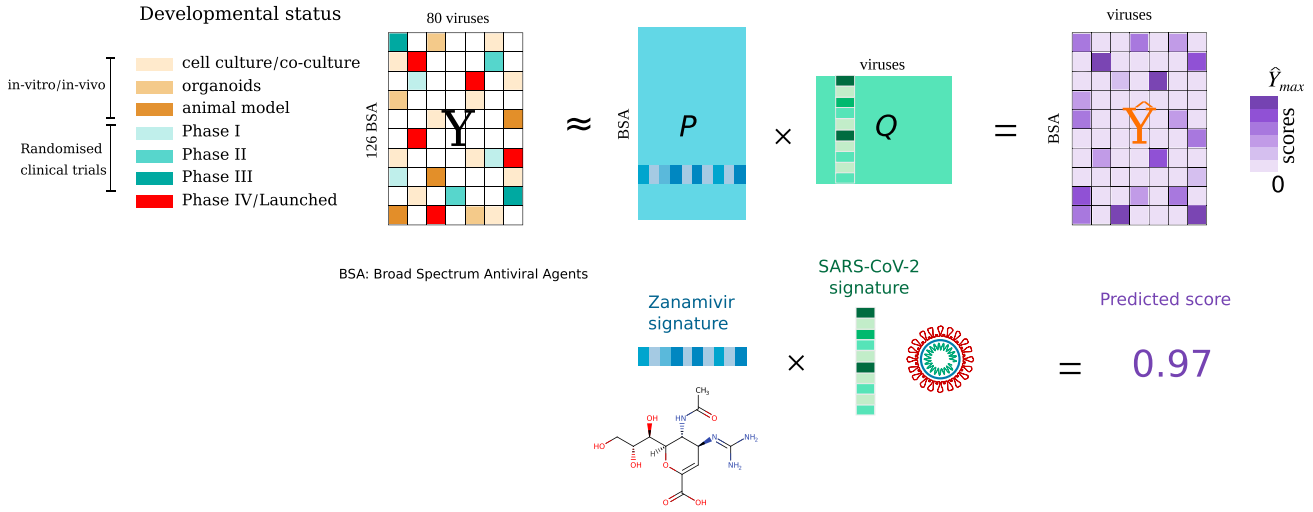


Figure 2. Overview of our matrix decomposition model for predicting effective drug-virus associations

Totals of 850 associations for $n = 126$ different BSAs and $m = 80$ distinct viruses were collected from the Andersen et al.[17] database. The observed associations were arranged into an $n \times m$ matrix Y by setting $y_{ij} = 1$. Unobserved associations were encoded with zeros. 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 $\hat{Y}$, which models Y, where all the 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. The dot product of these two signatures is the predicted efficacy of zanamivir against SARS-CoV-2.

as a regularization term.[19] Finally, we impose non-negative constraints on P and Q to favor the interpretability of the learned representations.[19,20]

Thus, our model is closely related to NMF.[20] Both models seek to decompose a data matrix Y into the product of two non-negative matrices P and Q. However, the NMF model considers all the entries in Y equally during the learning—this works well when entries have the same meaning, e.g., pixels in an image.[20] Instead, in our approach, we assign different levels of importance to subsets of entries to reflect the drug stages of development, thus coding the probability of drug success, which is what we are trying to predict. This gives rise to a loss function in Equation 1 that is different from NMF. Finally, notice that, in our model, if we set the values of all the α parameters to 1—which amounts to discarding the role of probabilities of success—we obtain the original NMF model.

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, which minimize the loss function in Equation 1, by employing an iterative algorithm that uses a simple multiplicative update rule (see the Experimental procedures). Our algorithm, inspired by the diagonally rescaled principle of NMF,[20] is fast, it does not require setting a learning rate or applying a projection function and it satisfies the Karush-Kuhn-Tucker (KKT) complementary conditions of convergence (see the Experimental procedures). Having learned P and Q such that $Y \approx PQ$, we calculate the matrix $\hat{Y} = PQ$. Note that, while Y contains binary entries, $\hat{Y}$ contains real positive numbers that are our predicted scores.

Predicting effective BSA drugs against viruses

To perform an *in silico* evaluation of the performance of our model, we formulated a matrix completion task under a leave-one-out cross-validation (LOOCV) procedure using the 49 BSA drugs that have been approved for use, and the 22 that reached phase IV of clinical trials for 28 viruses. To prevent overfitting and biases during hyperparameter tuning, we performed a different LOOCV by using clinical trials associations from phases I, II, and III to set the model parameters. Our final model parameters were: $k = 5$, $\alpha_B = 0.16$, $\alpha_C = 0.27$, $\alpha_D = 0.71$, $\alpha_E = 0.01$, and $\alpha_z = 2$ (see the Experimental procedures).

We compared the performance of our algorithm with the other drug-repurposing approaches that we mentioned earlier, namely the NMF with L2 regularization,[22] the TriFactor NMF,[23,26] and the IRNMF,[24] which was also developed for COVID-19. Moreover, we also included standard NMF and truncated singular value decomposition (tSVD)[18] as baselines. The relation between previous NMF-based drug repositioning methods and our model is explained in Note S1.

Following other works that used LOOCV evaluations,[9,27,28] 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. Here, it is important to remind ourselves that our model takes as input an incomplete sparse drug-virus matrix, with only 8.43% non-zero entries, and outputs predicted scores for all the entries in the matrix. In the evaluation presented here, we focus on validating predictions corresponding to the interesting case where drug-virus associations are not yet under development (see Note S2 for the case of predicting drugs already under development, but not approved, for specific viruses). Therefore, in our LOOCV procedure, one drug-virus association (approved or phase IV) was removed at a time from the drug-virus matrix Y (by setting the corresponding entry to zero). We then trained the model, and scores were predicted for all drugs. Finally, we ranked drugs that had no known association with that virus and checked the percentage of cases in

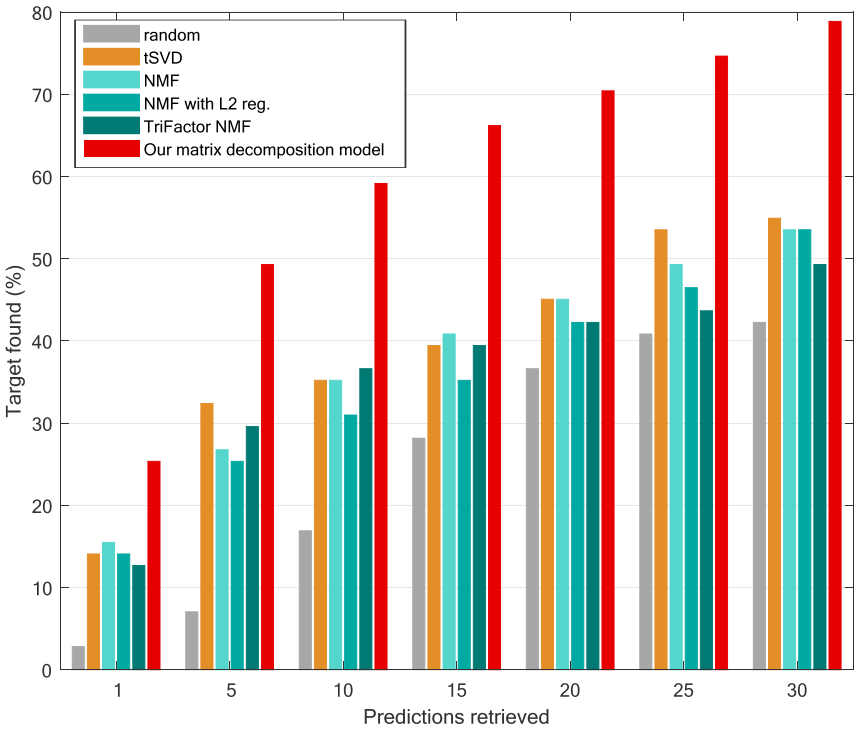


Figure 3. Performance at predicting approved/phase IV BSAs for 28 viruses

Percentage of approved or phase IV BSA drugs found for a specific virus in the top *K* predictions retrieved. The performance of our method is compared with different matrix decomposition algorithms in a leave-one-out fashion. NMF, non-negative matrix factorization; tSVD, truncated singular value decomposition. A baseline based on random scores sampled from a uniform distribution is also included.

which the correct (effective) drug for the virus was found among the top *K* predictions.

Figure 3 shows the performance of the methods at predicting effective (approved/phase IV) BSA drugs against specific viruses. Our model outperforms the competitors for each number of predictions retrieved: by 9.8%–22.5% in the top 1, by 16.9%–42.2% in the top 5, by 22.5%–42.2% in the top 10, and by 25.3%–38% in the top 20. Overall, our method could recover 70% of the phase IV/approved BSA drugs for 28 distinct viruses in the top 20 predictions. We also observed that, in some cases, tSVD and TriFactor NMF perform slightly better than NMF. The comparison of our method's performance with IRNMF was performed in a smaller subset of the matrix *Y* (see Figure S1 in Note S1).

The good prediction performance of our model prompted us to ask how much of the prediction power could be attributed to the integration of developmental stages information in our cost function (see Equation 1). Our model significantly improves performance over two control baselines that: (1) randomize developmental stage information in the training set and (2) remove developmental stage information from our cost function (see a detailed discussion in Note S3).

The analysis of our predictions for SARS-CoV-2 is presented in detail in the "Evaluation" section, together with the results of our network medicine approach.

Repositioning FDA-approved drugs with network medicine

The majority of BSAs considered previously target viral proteins. In our work, we also explored approaches that consider drugs targeting human proteins. Human proteins interact with each other, forming a protein-protein interaction (PPI) network. This and other biological networks have been explored in relation to disease—this area of research has often been called network medicine. It has been shown that proteins associated with specific hereditary diseases tend to cluster in neighborhoods of the interactome (the disease module),[8,29,30] and successful applications of molecular network analysis have been reported for the identification of disease genes,[9] drug development,[10] and drug efficacy prediction.[29]

The use of network medicine for assisting drug repositioning was originally applied to genetic diseases.[29] A drug induces its effects on a human PPI subnetwork by binding to its target proteins,[31,32] and this causes a perturbation in the interactome that is then propagated. Thus, drug efficacy for a genetic disease can be associated to how likely the drug is to affect its disease module through the perturbations propagated in the human PPI network.[29] To implement this idea, Guney et al.[29] proposed a distance (hereafter, the Guney distance) based on the shortest path length between the disease module and the drug targets.

Recent studies suggest that an analogous approach can be useful for infectious diseases such as COVID-19.[13,33] Viruses hijack host proteins to facilitate their replication, and hence the inhibition or knockdown of such host proteins can block viral replication.[34] Gysi et al.[13] have shown that, for SARS-CoV-2, most of the experimentally identified host proteins[14] group together in a large connected component, forming a COVID-19 disease module, as illustrated in Figure 4A with red nodes (host protein subnetwork). Therefore, the idea here is to find drugs that, by binding to their targets (blue nodes in Figure 4A), are likely to perturb this module.

We can think of the perturbation caused by a drug as a process in which the effect of the drug *diffuses* on the PPI network starting from its targets. Thus, our drug repurposing problem translates into the problem of the diffusion between drug targets and the set of host proteins. Gysi et al.[13] implemented this idea for COVID-19 by using the diffusion state distance (DSD).[35]

Kernels on graphs are appealing for modeling a diffusion process on a network. They are theoretically well founded in statistical learning theory,[36,37] and have shown good empirical results in many applications.[35,38,39] Graph kernels can be interpreted as measures of similarity between nodes in a network. There are different types of kernels. The *p*-step random walk kernel, for example, is directly associated to the number of times a random

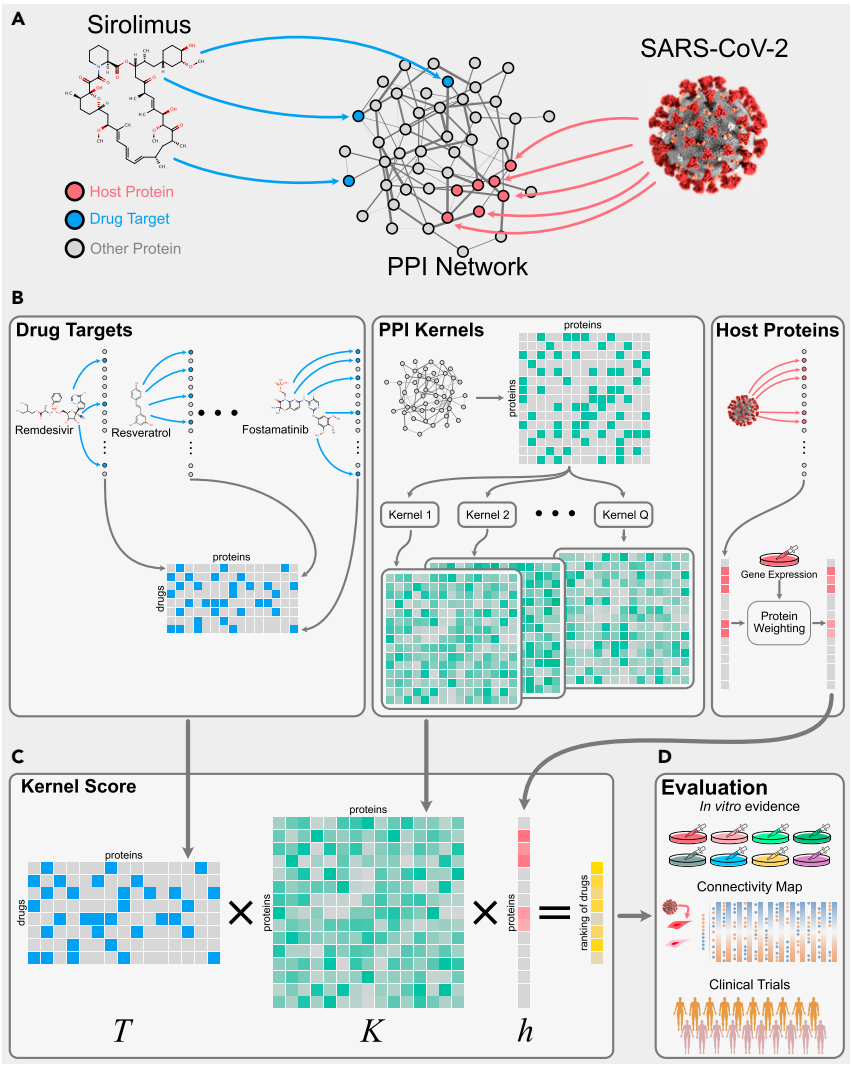


Figure 4. Overview of our network medicine approach

(A) The human interactome containing both host proteins (red) and drug targets (blue).

(B) The totals of 14,941 drug target associations between N = 2,197 FDA-approved drugs and n_V = 18,505 proteins are represented by a binary matrix T (blue matrix). Multiple graph kernels are calculated on the interactome, resulting in $n_V \times n_V$ matrices (green matrices). The host proteins are represented by a vector h of size n_V (red vector) indicating their weights (based on gene expression data).

(C) Our kernel score is calculated using a matrix multiplication to obtain a prediction score for each drug.

(D) The obtained ranking is evaluated using different types of evidence: *in vitro* efficacy against SARS-CoV-2, Connectivity Map, and clinical trials.

walker starting from a node i visits a node j after p steps.[36] Another example is the diffusion kernel (or heat kernel), which can be thought of as a random walk with an infinite number of infinitesimally small steps. An alternative interpretation is that this kernel corresponds to the amount of heat that reaches a node j after diffusing an initial heat from node i.[36]

Importantly, kernels on graphs can be applied in a natural way to nodes with weights. This property can be particularly useful for our problem: we can assign weights to the host proteins to model the different roles that they have for the infection/replication of the virus. For example, it has been shown that the ACE2 protein receptor is the viral entry factor of SARS-CoV-2.[40] Another study based on gene expression experiments on infected SARS-CoV-2 cell lines suggests that certain protein-coding genes play a key role during the infection process.[41] Then, the amount of change in gene expression after SARS-CoV-2 infection may be associated with the level of importance of the protein for the infection. In addition, perturbing host proteins whose expression levels change the most may be important for reverting the effect that the infection causes in gene expression. Predicting drugs that might revert this effect has been shown to be a good strategy for assisting in the discovery of effective small molecules for different diseases.[15,42]

To assist the repositioning of drugs for COVID-19, we used five different kernels on graphs and weighted the host proteins with differential gene expression data (absolute value of the log fold change between the gene expression levels of COVID-19 patients, and controls—see experimental procedures for details on the RNA-seq data). We used the interactome assembled by Gysi et al.,[13] and a set of 336 human proteins that were identified as hosts of SARS-CoV-2 (see experimental procedures). Every FDA-approved drug with known targets in this interactome was ranked by each of the kernels in our approach (see experimental procedures). The selected kernels are defined in terms of the graph Laplacian (see experimental procedures), as shown in Table 1. For each drug, we obtain the graph kernel-based similarities between each of its targets and each of the host proteins. The final score of a drug is the sum of these similarities weighted by the amount of change in the host protein expression levels after infection. Drug scores are then ordered, obtaining a drug ranking which is evaluated. We also calculated an aggregated ranking, which we called *avgRank*, where the ordinal position of each drug was obtained by simply averaging the ranking position that the drug had obtained in each of the kernels.

The mathematical formulation of this approach turns out to be quite simple. Let n_V the number of proteins in the PPI network, N the number of FDA-approved drugs, and T an $N \times n_V$ matrix of drug target associations, where $T_{ij} = 1$ if j is a target of drug i, and 0 otherwise (see drug targets box in Figure 4B). Let K be a square matrix of dimensions $n_V \times n_V$ representing kernel-based similarities between proteins on the PPI network (see PPI kernels box in Figure 4B). Let h be an n_V-dimensional column vector containing weights related to the differential expression data of the host proteins and zeros for the remaining proteins (see

Please cite this article in press as: Santos et al., Machine learning and network medicine approaches for drug repositioning for COVID-19, Patterns (2021), https://doi.org/10.1016/j.patter.2021.100396

Table 1. Graph kernels

Kernel	Formula
p-step random walk	$K = (aI - \tilde{L})^p$
Diffusion process	$K = \exp(-\sigma^2/2\tilde{L})$
Regularized Laplacian	$K = (I + \sigma^2\tilde{L})^{-1}$
Commute time kernel	$K = L^+$
Inverse cosine	$\cos\tilde{L}\pi/4$

Definition of graph kernels based on the normalized Laplacian ($\tilde{L}$), and pseudoinverse of the Laplacian (L^+), where a, p, and σ are given parameters.

host proteins box in Figure 4B). We obtain prediction scores simultaneously for all drugs with the following matrix multiplication $S_d = TKh$ (also illustrated in Figure 4C), resulting in a vector of drug scores, S_d.

Evaluation

To evaluate the performance of our methods, we used three different sources of evidence from ongoing research: *in vitro* experiments, clinical trials, and CMAP scores. These sources are independent of each other; hence they can be used to provide an independent evaluation of the efficacy of repurposing methods. Note that none of these three sources of evidence can be considered a gold standard, as none of them can ensure therapeutic effects for COVID-19 patients. Yet, they represent a proxy of effectiveness of drugs for COVID-19.

In vitro experiments involving drugs with antiviral efficacy indicate their potential to be effective at reducing viral infection and replication in the host cell. Evaluating our models with this kind of evidence allows us to assess whether they prioritize drugs with molecular antiviral efficacy versus other drugs.

Clinical trial studies are used to assess pharmacokinetics, dosage, therapeutic efficacy, and safety of drugs.[43] Each phase in clinical trials involves an increasing number of patients, thus achieving higher statistical significance while minimizing the number of patients that risk developing side effects.[44] Indicating a drug in a clinical trial requires satisfying several conditions set by biologists and medics, and arguments of why it might be effective. This suggests the investigators believe that the drug is safe and a potential candidate to treat the disease. Evaluating our models with clinical trial evidence allows us to determine if they prioritize drugs that would be included in such trials.

We use the CMAP[15,16] to contrast changes in gene expression levels caused by a drug (drug expression profile) with changes induced by SARS-CoV-2 infection (disease expression profile). The hypothesis is that, if a drug expression profile is opposite to a disease expression profile, then it could potentially "revert" the disease signature and have therapeutic effects—this idea has already been used before[15,42] to predict new therapeutic indications for drugs and has also been applied to COVID-19.[45,46] Therefore, evaluating our models with this source of evidence allows us to assess whether they prioritize drugs with potentially therapeutic effects.

For the matrix decomposition approach, the evaluation was carried out using the 126 BSAs in the drug-virus dataset.[17] For the network medicine approach, the evaluation was done on an interactome of 18,505 proteins with 327,924 interactions.[13] With this approach we ranked 2,197 approved drugs from DrugBank.[47]

We used the types of evidence described above to create three datasets where drugs were classified as either effective or non-effective for COVID-19 (see experimental procedures). This allowed us to assess the performance of a prediction method by formulating a binary classification problem, where the task is to discriminate the two sets of drugs, and then calculating binary classification metrics based on the analysis of the confusion matrix.

However, we note that the lack of a set of drugs with proven therapeutic effect against COVID-19 (i.e., a gold standard), poses a challenge for this type of evaluation—this problem has also been described before, e.g., in Zhou et al.[48] and Gysi et al.[13] We hypothesized that drugs with evidence against COVID-19 should behave differently from the remaining drugs. This hypothesis has an actionable consequence: a method can be evaluated by assessing whether it can discriminate between the two groups of drugs (effective and non-effective)—if it can, this is an indication that we can possibly trust the predictions it makes. Therefore, together with traditional metrics for binary classification, we also assessed whether prediction methods provided scores that were statistically different for the two classes of drugs. Our results (Figures 5A, 5B, and 5F–5H) show that the differences between the scores are significant for our matrix decomposition approach as well as our kernel methods across several evaluation settings. We observe that other network-based methods do not pass this test with such consistency (see Notes S5–S7). In the following, we present the results for each type of evidence, separately.

In vitro evaluation

Of the 126 BSAs in the drug-virus dataset, 10 have shown *in vitro* efficacy against SARS-CoV-2.[13,49] In our evaluation, these drugs were removed one at a time from the drug-virus matrix Y (by setting the corresponding entry to zero). We then trained our matrix decomposition model, and scores were predicted for all the drugs. We used the Wilcoxon-Mann-Whitney p value to assess the difference between the scores obtained for those 10 drugs and the rest of the drugs. Figure 5A shows that our matrix decomposition method significantly assigns higher scores to BSAs with *in vitro* efficacy (Wilcoxon-Mann-Whitney p value = 4.92×10^{-7}). Precision and recall are shown in Figure S2 (Note S1).

Scores predicted by the kernel-based methods are shown in Figure 5F. Of the 2,197 FDA-approved drugs considered by our network medicine approach, 81 have shown *in vitro* efficacy against SARS-CoV-2.[13,49] We observed that the scores of drugs with *in vitro* efficacy against SARS-CoV-2 are significantly higher than those of the remaining drugs for all kernels and the average ranking (avgRank).

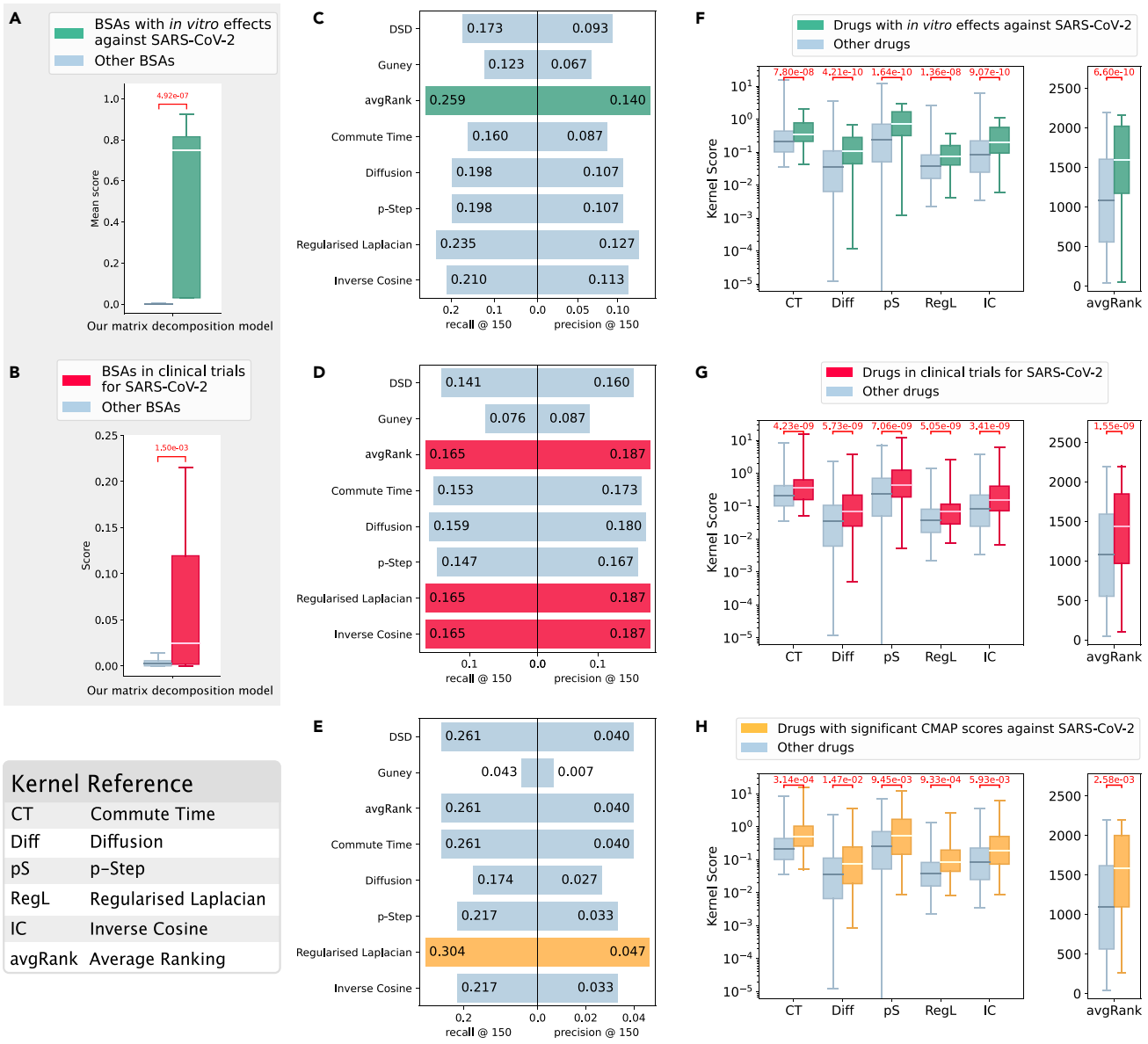
In Figure 5C, we show that the kernel-based methods performed better than the competitors for the *in vitro* evaluation. The recall@150 of the average ranking is 49.71% higher than DSD, and 110.57% higher than the Guney distance. The precision@150 of the average ranking is 50.54% higher than DSD, and 108.96% higher than the Guney distance.

Clinical trial evaluation

Of the 126 BSAs in the drug-virus dataset, 28 are in clinical trials (see experimental procedures). Figure 5C shows that prediction scores by our matrix decomposition method are significantly

CellPress
OPEN ACCESS

Patterns
Article



Kernel Reference	
CT	Commute Time
Diff	Diffusion
pS	p-Step
RegL	Regularised Laplacian
IC	Inverse Cosine
avgRank	Average Ranking

Figure 5. Analysis of the predictions for COVID-19

We used three different sources of evidence: *in vitro* (A, C, and F), clinical trials (B, D, and G), and CMAP (E and H). We compared scores for drugs with evidence of efficacy against SARS-CoV2 versus scores for the remaining drugs. Our matrix factorization model (A and B) and kernel-based methods (F, G, and H) provide scores that are significantly different between the two groups of drugs in every case (Wilcoxon-Mann-Whitney $p < 0.05$). We formulated a binary classification problem to discriminate between drugs with evidence of efficacy against SARS-CoV2 and the remaining drugs. (C, D, and F) Comparison of precision and recall at top 150 for our kernel-based methods (commute time, diffusion, *p*-step, regularized Laplacian, and inverse cosine kernels, and avgRank), DSD, and Guney's distance. The highest values are colored.

higher for drugs in clinical trials (Wilcoxon-Mann-Whitney p value = 1.5×10^{-3}). Our method can recover 50% of the correct BSAs in the top-20 predictions retrieved (see Figure S2 in Note S1).

Scores predicted by the kernel-based methods are shown in Figure 5G. Of the 2,197 FDA-approved drugs considered by our network medicine approach, 170 are in clinical trials. We observed that the scores of drugs in clinical trials for COVID-19 are significantly higher than those of the remaining drugs for all kernels and the average ranking (avgRank).

In Figure 5D, we show that the kernel-based methods performed better than the competitors for the clinical trials evaluation. The recall@150 of the average ranking is 17.02% higher than DSD, and 117.11% higher than the Guney distance. The precision@150 of the average ranking is 16.88% higher than DSD, and 114.94% higher than the Guney distance.

CMAP evaluation

We queried CMAP[15,16] obtaining a list of 23 FDA-approved drugs that present an expression profile opposite to the one expressed by SARS-CoV-2 infected cells with a τ score between

Please cite this article in press as: Santos et al., Machine learning and network medicine approaches for drug repositioning for COVID-19, Patterns (2021), https://doi.org/10.1016/j.patter.2021.100396

−90 and −100 (see experimental procedures). Figure 5H shows that the scores of FDA-approved drugs with strongly negative CMAP correlation are significantly higher than those of the remaining drugs for all kernels and the average ranking (avgRank).

In Figure 5E, we compared the performance of the kernel-based methods and competitors for the CMAP evaluation. Our average ranking has the same performance as DSD and better performance than Guney's distance (recall@150 is 50.72% higher, and precision@150 is 47.43% higher). The regularized Laplacian kernel had the best performance, with recall@150 16.48% higher than DSD, and 60.7% higher than Guney's distance, and precision@150 17.5% higher than DSD, and 57.14% higher than Guney's distance.

On the importance of integrating transcriptomics data. An interesting question is whether weighting host proteins by differential expression improves our network medicine approach. To answer this, we compared results based on weighted host proteins and unweighted/binary host proteins. For *in vitro* and clinical trials evidence, we observed that the Wilcoxon-Mann-Whitney p values are smaller (more significant) when using weighted host proteins when compared with considering all host proteins equally. The recall@150 and precision@150 are consistently higher when we use weights for the three types of evidence. These results are presented in Note S8.

Our results hold for different PPI networks and evaluation settings. An important question is whether results are consistent across different interactomes and how sensitive they are to different choices of the PPI network. We re-computed the kernel-based scores using the recently released HuRI PPI[50] as well as the interactome compiled by Cheng et al.[30] Results are presented in Note S7. For most of the kernels, FDA-approved drugs with *in vitro*, and clinical trials evidence have a significantly higher prediction score than the remaining drugs. For the three sources of evidence, the kernel-based methods have the higher recall@150 and precision@150 when compared with competitors. This indicates that our results have a high consistency across different interactomes.

Comparison with the approaches by Gysi et al.. We also extensively compared our kernel methods with the methods recently proposed by Gysi et al.,[13] although the comparison could only be carried on the Gysi et al. dataset—this consists of 918 drugs including approved, investigational, experimental, nutraceutical, and withdrawn drugs. Overall, our kernel methods perform better with respect to *in vitro* and CMAP evidence—note that, in several cases, the scores obtained by the Gysi et al. methods for sets of effective and ineffective drugs are not significantly different. GNN methods perform better than kernel methods only with respect to clinical trial evidence. A summary of the different datasets used can be found in Note S4. A detailed description of all the experiments comparing our approaches with those from Gysi et al. is presented in Note S6.

CoREx

As a further way to evaluate drug repurposing against SARS-CoV-2, we developed CoREx, a web-based tool that enables scientists to study drug repurposing in a functional context on the interactome. 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, PPIs, clinical trial status, CMAP scores, and drug's anatomical therapeutic chemical [ATC] categories). Our goal is to assist researchers to reason about drug alternatives, drug combinations, and mechanisms of actions by analyzing the interplay between drug targets and host proteins in these different contexts.

Centered around ideas from network medicine, CoREx provides two different tools: the *functional analysis tool* and *interactome analysis tool*. The functional analysis tool allows the user to study the relationships between drug targets and host proteins. A functional interactome is built by integrating protein networks available in the STRING database[51] in a way that maximizes the probability that two interacting proteins share functional characteristics (see Note S10 for details on the network combination). Then, we use the ClusterONE algorithm[52] to identify functionally similar groups of proteins, and filter those that contain at least one SARS-CoV-2 host protein, and at least one drug target. The functional enrichment of these groups is then analyzed using Enrichr.[53] 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. All of these results are presented to the user in a user-friendly interactive graphical interface, as shown in Figure 6.

The interactome analysis tool allows the user to visualize the perturbation caused by a drug on the SARS-CoV-2 host protein subnetwork. When a drug is selected, each node (host protein) is colored based on the strength of the resulting kernel score. This tool complements CoVex, by Sadegh et al.,[54] which analyzes the interplay within the virus-host-drug triad using paths on the interactome. Instead, CoREx calculates the effects that drugs have on individual host proteins through the different graph kernels. We have preloaded our interactome analysis tool with those FDA-approved drugs that have available drug targets from DrugBank.[47] Users can also submit a list of drug targets, and visualize the perturbation that a hypothetical drug (or drug combination) with those targets would have on the host proteins subnetwork.

CoREx is available at https://paccanarolab.org/corex and supporting datasets are updated every 2 weeks. The project is also open-source, and the repository is publicly available at https://github.com/paccanarolab/corex.

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.[1–3,6,48] Drug repositioning, the re-use of drugs already on the market, can help to speed up the development of such treatments by prioritizing known safe-in-human drugs for clinical trials involving COVID-19 patients. In this paper, we proposed two machine learning approaches that can assist in the prioritization of drugs, together with a human-in-the-loop website tool, CoREx, to assist current research efforts for finding drugs with therapeutic efficacy against SARS-CoV-2.

Li and De Clercq[4] indicated that finding potential repositioning candidates for COVID-19 should be focused on two main strategies: virally targeted agents and host-targeted agents. Our matrix decomposition approach is aimed at the first repositioning strategy, whereas our network medicine approach, together with CoREx, is aimed at the second one. Our first approach ranks 126 BSAs by their predicted efficacy against SARS-CoV-2, and

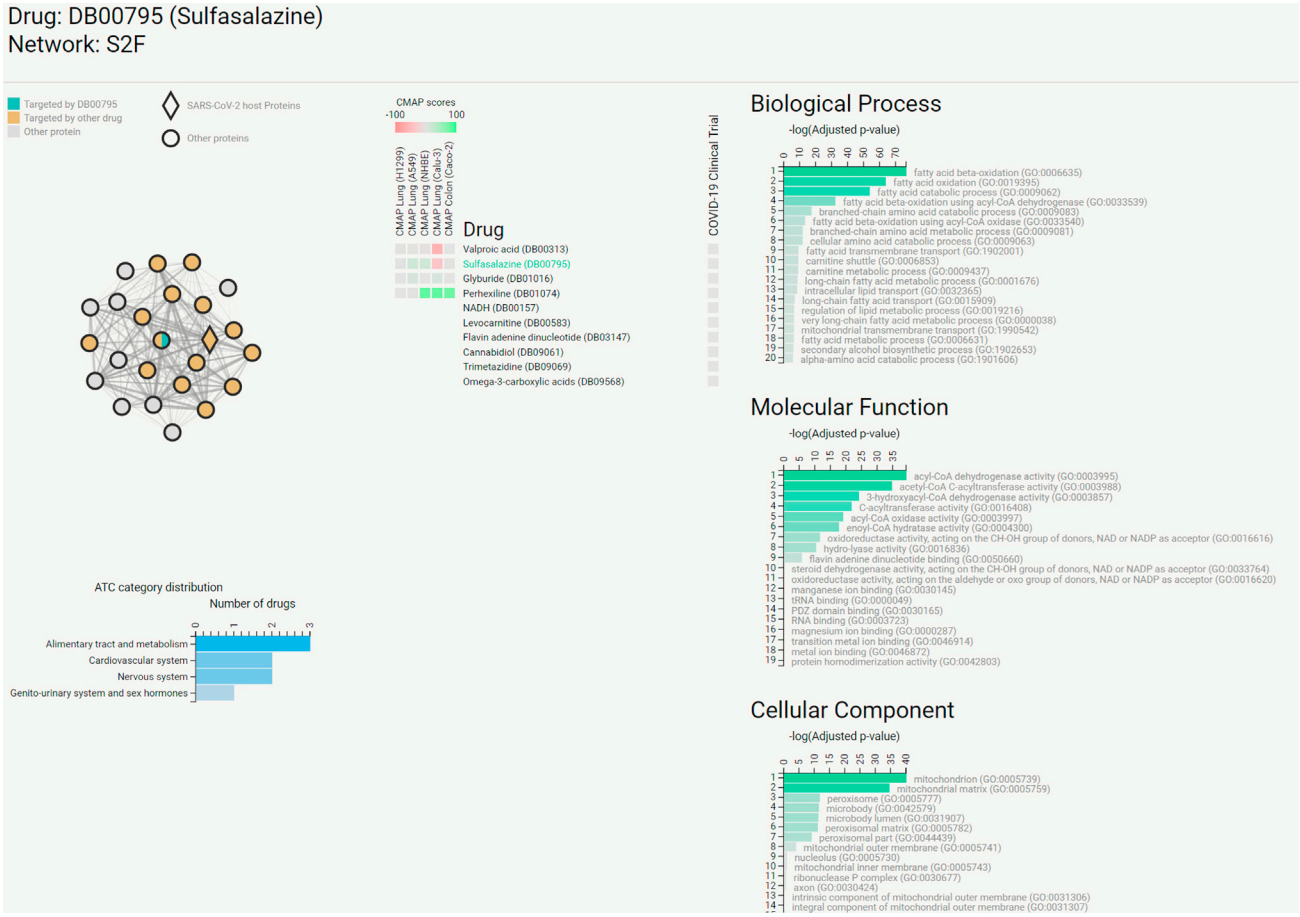


Figure 6. Screenshot of CoREx displaying a functional module for Sulfasalazine (highlighted in green in the "Drug" list)

The module is depicted as a network on the top left where nodes represent proteins, edges represent shared functional characteristics, and the thickness of the edges represents the strength of such functional similarity. Host proteins are depicted as diamonds, drug targets are colored. The list of drugs with at least one target in this functional module is presented in the center, alongside CMAP scores for five cell lines (on the left), and an indicator of whether the drug is currently in clinical trials (on the right). The bar plots on the right part correspond to the functional enrichment scores for each GO domain. The bar plot on the bottom left section of the image summarizes the ATC categories of the drugs targeting this functional module.

our second approach ranks 2,197 therapeutically diverse FDA-approved drugs by their predicted ability to perturb the COVID-19 disease module.

The objective function of our matrix decomposition approach in Equation 1 is inspired by our recent work to predict the frequencies of drug side effects.[55] The main feature of this new model is that it can account for varying levels of uncertainties in the data. We realized that different levels of drug developmental evidence can be thought of as indicating different levels of confidence in drug-virus associations and can be interpreted as probabilities. Our new model exploits the richness of this information and its outputs can be interpreted as probabilities of drug approval. Experiments in which we randomized or removed information about drug developmental stages show that such information is key to achieve a good performance (see Note S3). The implementation of our algorithm is freely available: https://github.com/paccanarolab/DrugRepoCOVID.

Our network medicine approach aims at prioritizing FDA-approved candidates based on their network-modulated effects on the COVID-19 disease protein module. In contrast to our first approach, our network medicine approach does not explicitly model the clinical efficacy of drugs, but rather their mechanistic effects on the protein interaction network. This means that a high score points to a high probability for the drug to perturb the disease module. Note, however, that our kernel methods, like most network-based approaches,[13] can quantify the perturbation on the interactome, but cannot predict in which way the host will ultimately be affected by such perturbations (see Note S11.2).

An important advantage of our kernel approaches is that they offer a natural way to integrate gene expression data and thus allow us to focus the models on particular proteins that play a key role in the infection. Our experiments show that the integration of transcriptomics data improves the results (see Note S8). Furthermore, we have shown that our kernels have similar performance across multiple interactomes (see Note S7).

We have shown that our predictions from both approaches are aligned to ongoing *in vitro* experiments and clinical trial studies. An interesting question is whether there is additional biological evidence of efficacy for the best scoring drugs from our

Please cite this article in press as: Santos et al., Machine learning and network medicine approaches for drug repositioning for COVID-19, Patterns (2021), https://doi.org/10.1016/j.patter.2021.100396

approaches. We manually curated the top 20 predicted drugs obtained from each approach. Our analysis reveals that many of these drugs are linked to ongoing efforts against COVID-19: several top-ranked BSAs from our matrix decomposition model are part of ongoing clinical trials for COVID-19, or are even already approved for compassionate use in COVID-19 patients[56–59]; several top-ranked drugs from our network medicine approach have also shown efficacy either as therapeutic alternatives or as instruments for reducing risk of infection and transmission.[60–63] An in-depth analysis of the top 20 predictions, including an analysis of their ATC classification and references to the literature, is presented in Note S11. A comparison with the set of drugs predicted by Gordon et al.[14] is also provided in Note S9. Finally, while the datasets that we used in our two approaches are different, a few drugs could be predicted by both methodologies—these are analyzed in Note S12.

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. Recently, COVID-19 atlases have been published, including single-cell transcriptomics data[64,65] that could be exploited with our approaches.

We also point out that, while we have developed and tested our two approaches for COVID-19, both of them are disease agnostic. The general principles underlying our matrix decomposition and network medicine approaches will remain valid for any other viral disease, and therefore our methods could be applied for drug repurposing in these scenarios, as long as the data are available.

Finally, the integration of heterogeneous sources of omics 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 (https://paccanarolab.org/corex) with the goal of providing the research community with a tool for the analysis and the formulation of hypothesis about drugs that can be repurposed for COVID-19. CoREx combines transcriptomics, proteomics, and functional information about the human genome together with knowledge about drugs and their protein targets, and we make it available for the scientific community.

Limitations of the study

Our matrix decomposition approach is applicable to any drug for which the developmental stage associating it to a viral disease is known. The drug may or may not be virally targeted, and the model itself will not impose such a restriction. The main limitation of the method is that it relies on drug-virus associations annotated with their stage of development, and publicly available data of this type is currently scarce—we only found this type of information in the manually curated dataset by Andersen et al.[17] that we used in our study. The main limitation of our network medicine approach is that it can only be applied to drugs with known protein targets on the host interactome.

EXPERIMENTAL PROCEDURES

Resource availability

Lead contact

The lead contact for this work is Alberto Paccanaro (alberto.paccanaro@rhul.ac.uk).

Materials availability

This study did not generate new unique reagents.

Data and code availability

Original data have been deposited to Mendeley Data: https://doi.org/10.17632/p7y5wmschg.1. The implementation of our matrix factorization model can be found at https://github.com/paccanarolab/DrugRepoCOVID. CoREx is available at https://paccanarolab.org/corex, and the source code is publicly available at https://github.com/paccanarolab/corex.

Datasets

- *The drug-virus dataset.* We used the dataset curated by Andersen et al.[17] (downloaded April 6, 2020). Drugs were mapped to DrugBank IDs, when available. Each drug-virus association was 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 BSAs and 80 viruses.
- *Protein interaction network.* The PPI network was obtained from Gysi et al.,[13] which contains 18,505 human proteins, and 327,924 interactions.
- *FDA-approved drugs and drug targets.* FDA-approved drugs and their drug targets were retrieved from DrugBank[47] and Gysi et al.[13] Our set of drugs consisted of 2,197 FDA-approved drugs consisted. Our set of drug target associations consisted of 14,941 pairs of drug and targets.
- *Host proteins.* Our COVID-19 disease module consisted of 336 host proteins. It includes 332 host proteins reported by Gordon et al.,[14] the entry receptor ACE2,[66] and three SARS-CoV-2 entry-associated proteases TMPRSS2,[67] CTSB, and CTSL.[68]
- *Gene expression data.* To weight the host proteins in our kernel-based methods, we used gene expression data from 430 COVID-19 patients, and 54 controls, collected from nasopharyngeal swabs.[69] The RNA-seq raw counts are available in the Gene Expression Omnibus (GEO),[70,71] with accession number GSE152075. We processed the data using the edgeR package,[72] and obtained the absolute value of the log fold change comparing the expression levels between COVID-19 patients and controls. For 47 host proteins with missing mRNA levels, we assigned the minimum absolute value of the log fold change. The final weights of the host proteins are available from Mendeley Data (see Table S7).
- *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.[49] and Gysi et al.[13] Eighty-one FDA-approved drugs show *in vitro* effects (see Table S7).
- *Clinical trials data.* We built a binary dataset and assigned positive labels to drugs that are involved in clinical trial studies, and negative labels to all other drugs. Information for clinical trials studies was downloaded from ClinicalTrials.gov on December 1, 2020.[73] Drugs were mapped to the DrugBank database[47] by matching their names (see Table S7).
- *CMAP data.* For the CMAP query, we used a COVID-19 signature by Ghandikota et al.[74] This gave us a list of 106 genes upregulated and 41 genes downregulated in three different models of SARS-CoV-2 infection from transcriptomics data. Two are models *in vitro* (Calu-3 and Vero E6 cells), and one model is *in vivo* (Ad5-hACE2-sensitized mice). The query with these data resulted in 30 drugs with significant negative τ score ($\tau < -90$) that were mapped to DrugBank. Twenty-three of these 30 drugs are FDA approved and have targets in the Gysi et al. interactome. The list of 30 drugs with CMAP evidence is available from Mendeley Data (see Table S7). All supplementary files available from Mendeley data and external data sources are listed in Table S7 (Note S13).

The multiplicative learning algorithm for the matrix decomposition model

To minimize Equation 1 subject to non-negative constraints, we developed an efficient multiplicative learning algorithm inspired by the diagonally rescaled

principle of NMF.[21] The algorithm consists of iteratively applying the following multiplicative update rules:

$$P_{ia} \leftarrow P_{ia} \frac{\left(\left[M^A \circ Y + \sum_{s\in\{B,C,D,E\}} \alpha_s (M^s \circ Y)\right] Q^T\right)_{ia}}{\left(\left[M^A \circ (PQ) + \sum_{s\in\{B,C,D,E\}} \alpha_s M^s \circ (PQ) + \alpha_z M^z \circ (PQ)\right] Q^T\right)_{ia}}$$

$$Q_{aj} \leftarrow Q_{aj} \frac{\left(P^T \left[M^A \circ Y + \sum_{s\in\{B,C,D,E\}} \alpha_s (M^s \circ Y)\right]\right)_{aj}}{\left(P^T \left[M^A \circ (PQ) + \sum_{s\in\{B,C,D,E\}} \alpha_s M^s \circ (PQ) + \alpha_z M^z \circ (PQ)\right]\right)_{aj}}. \quad \text{(Equation 2)}$$

Following the guidelines to implement NMF,[75] a small number $\varepsilon = 10^{-8}$ was added to the denominators in Equation 2 to prevent division by zero, and we initialized P and Q as random dense matrices uniformly distributed in the range $[0, 0.1]$. Furthermore, to avoid the well-known degeneracy[20] associated with the invariance PQ under the transformation $P \rightarrow P\Lambda$ and $Q \rightarrow \Lambda^{-1}Q$, for a diagonal matrix Λ, we normalized P at each iteration as follows:

$$Q_{aj} \leftarrow \frac{Q_{aj}}{q_a}, \quad \text{(Equation 3)}$$

where 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 1,000 iterations for $k = 5$.

Using a similar procedure to Galeano and Paccanaro,[55] it can be easily shown that our algorithm in Equation 2 satisfies the KKT conditions of convergence.

Cross-validation procedure and model selection for the matrix decomposition approach

We used a LOOCV procedure to evaluate the performance of our matrix decomposition model. To set the model hyperparameters: k, α_E and α_Z, we performed LOOCV on the drug-virus associations with clinical trials developmental stages (validation set). We performed a grid-search and selected the hyperparameters that maximize 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[25] for anti-infective 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 an LOOCV on the test set corresponding to drug-virus associations that have been approved or are in phase IV of clinical trials. The model selection for the competitors was performed on the same validation sets (see details in Note S1).

The trained model that we used in the "Evaluation" section was obtained by training the model 1,000 times using all the available data with optimal hyperparameters. We then selected the solution that gave the lowest value in the loss function.

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 the 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 $\mathcal{H} \in V$ denote the set of host proteins. Our goal is then to perturb the subnetwork induced by $\mathcal{H}$, i.e., the host protein subnetwork.

Here, we rely on different graph kernels described in the literature.[35,36,76] In the following, graph kernels and their properties are defined as in Kondor and Vert.[77] 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 or similarities on a latent feature space. More specifically, there exists the feature mapping $\varphi : V \mapsto \mathcal{F}$ such that $k(i,j) = \langle \varphi(i), \varphi(j) \rangle$ 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 normalized 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 normalized 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 five graph kernels[35,36,76]: regularized Laplacian, diffusion process, and p-step random walk in terms of the normalized Laplacian[36] (see Table 1).

In the p-step random walk, $p \geq 1$ and $a \geq 2$ are given parameters.[36] 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 generalize 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 regularized Laplacian kernel can be thought of as the convergence of an iterative process in which nodes spread information to their neighbors at each step.

We used different kernels from Smola and Kondor,[36] Cao et al.,[35] and Zhou et al.,[76] which are implemented in the R package *diffuStats*[78] for the commute time, diffusion, p-step, regularized Laplacian, and inverse cosine kernels. We set the parameter p to 2 for the p-step kernel. For the remaining kernels, we used the default parameters in *diffuStats*.

CMAP evaluation details

We consider that a drug has CMAP evidence against COVID-19 if the changes that it causes to gene expression are opposite to the ones caused by the disease.[15] To build the CMAP evaluation set, we used the CMAP pipeline[15,16] to measure how similar or opposite the drug and COVID-19 expression profiles are. We used version 1.0 of the CMAP L1000 dataset[16] available on clue.io website (https://clue.io/).

We began by obtaining a list of up-/downregulated genes in COVID-19 (genes that have higher/lower expression levels in SARS-CoV-2 infected cells compared with 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 with the expression levels with no treatment). If the upregulated 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 downregulated 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 (upregulated genes on top, and downregulated 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 with a reference database and normalized between -100 and 100.

SUPPLEMENTAL INFORMATION

Supplemental information can be found online at https://doi.org/10.1016/j.patter.2021.100396.

ACKNOWLEDGMENTS

We thank Gloria Aguilar, Martin Aguero, Rafael Adorno, Aldo Galeano, Diego Stalder, Justin Reese, Afshin Beheshti, Thiago Moreno L. Souza, Carolina de Queiroz Sacramento, Claudio Struchiner, Valdiléa Veloso, Beatriz Grinsztejn, Irina Yuri Kawashima, Teresa Gamarra, Ruben Jimenez, Santiago Noto, and Philip Ovington for useful discussions. We also thank all members of the COVID-19 International Research Team (www.cov-irt.org) for their helpful feedback during weekly meetings. A.P. was supported by Biotechnology

Please cite this article in press as: Santos et al., Machine learning and network medicine approaches for drug repositioning for COVID-19, Patterns (2021), https://doi.org/10.1016/j.patter.2021.100396

and Biological Sciences Research Council (https://bbsrc.ukri.org/) grants BB/K004131/1, BB/F00964X/1, and BB/M025047/1; Medical Research Council (https://mrc.ukri.org/) grant MR/T001070/1; Consejo Nacional de Ciencia y Tecnología Paraguay - CONACyT (http://www.conacyt.gov.py/) grants 14-INV-088 and PINV15-315; National Science Foundation Advances in Bio Informatics (https://www.nsf.gov/) grant 1660648; Fundação de Amparo à Pesquisa do Estado do Rio de Janeiro grant E-26/201.079/2021 (260380); and the School of Applied Mathematics (EMAp), Fundação Getulio Vargas. S.d.S.S., M.T., and D.G. were supported by the School of Applied Mathematics (EMAp), Fundação Getulio Vargas.

AUTHOR CONTRIBUTIONS

S.d.S.S., M.T., D.G., L.C., and A.P. conceived the study, designed the methods, and analyzed the results. S.d.S.S., M.T., and D.G. implemented and conducted the experiments. M.T. and M.d.M.S. implemented CoREx. S.d.S.S., M.T., D.G., and A.P. wrote the manuscript. All authors reviewed the manuscript. A.P. supervised the project.

DECLARATION OF INTERESTS

The authors declare no competing interests.

Received: May 18, 2021
Revised: June 21, 2021
Accepted: November 1, 2021
Published: November 3, 2021

REFERENCES

1. Ashburn, T.T., and Thor, K.B. (2004). Drug repositioning: identifying and developing new uses for existing drugs. Nat. Rev. Drug Discov. *3*, 673–683.
2. Pushpakom, S., Iorio, F., Eyers, P.A., Escott, K.J., Hopper, S., Wells, A., Doig, A., Guilliams, T., Latimer, J., McNamee, C., et al. (2018). Drug repurposing: progress, challenges and recommendations. Nat. Rev. Drug Discov. *18*, 41–58.
3. Zumla, A., Chan, J.F.W., Azhar, E.I., Hui, D.S.C., and Yuen, K.-Y. (2016). Coronaviruses —drug discovery and therapeutic options. Nat. Rev. Drug Discov. *15*, 327–347. https://doi.org/10.1038/nrd.2015.37. https://www.nature.com/articles/nrd.2015.37.
4. Li, G., and De Clercq, E. (2020). Therapeutic options for the 2019 novel coronavirus (2019-nCoV). Nat. Rev. Drug Discov. *19*, 149–150.
5. Sanders, J.M., Monogue, M.L., Jodlowski, T.Z., and Cutrell, J.B. (2020). Pharmacologic treatments for coronavirus disease 2019 (COVID-19): a review. JAMA *323*, 1824–1836. https://doi.org/10.1001/jama.2020.6019.
6. Mei, M., and Tan, X. (2021). Current strategies of antiviral drug discovery for COVID-19. Front. Mol. Biosci. *8*, 310. https://doi.org/10.3389/fmolb.2021.671263. https://www.frontiersin.org/article/10.3389/fmolb.2021.671263.
7. Dolgin, E. (2021). The race for antiviral drugs to beat COVID—and the next pandemic. Nature *592*, 340–343. https://doi.org/10.1038/d41586-021-00958-4. https://www.nature.com/articles/d41586-021-00958-4.
8. Barabási, A.-L., Gulbahce, N., Loscalzo, J., and Network Medicine. (2011). A network-based approach to human disease. Nat. Rev. Genet. *12*, 56–68. https://doi.org/10.1038/nrg2918. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3140052/.
9. Cáceres, J.J., and Paccanaro, A. (2019). Disease gene prediction for molecularly uncharacterized diseases. PLoS Comput. Biol. *15*, e1007078.
10. Silverman, E.K., Schmidt, H.H.H.W., Anastasiadou, E., Altucci, L., Angelini, M., Badimon, L., Balligand, J.-L., Benincasa, G., Capasso, G., Conte, F., et al. (2020). Molecular networks in network medicine: development and applications. WIREs Syst. Biol. Med. *12*, e1489. https://doi.org/10.1002/wsbm.1489. https://onlinelibrary.wiley.com/doi/abs/10.1002/wsbm.1489.
11. Sharma, A., Menche, J., Huang, C.C., Ort, T., Zhou, X., Kitsak, M., Sahni, N., Thibault, D., Voung, L., Guo, F., et al. (2015). A disease module in the interactome explains disease heterogeneity, drug response and captures novel pathways and genes in asthma. Hum. Mol. Genet. *24*, 3005–3020. https://doi.org/10.1093/hmg/ddv001.
12. Wang, R.-S., and Loscalzo, J. (2018). Network-based disease module discovery by a novel seed connector algorithm with pathobiological implications. J. Mol. Biol. *430*, 2939–2950. https://doi.org/10.1016/j.jmb.2018.05.016. https://www.sciencedirect.com/science/article/pii/S0022283618304273.
13. Gysi, D.M., Valle, I.d., Zitnik, M., Ameli, A., Gan, X., Varol, O., Ghiassian, S.D., Patten, J.J., Davey, R.A., Loscalzo, J., and Barabási, A.-L. (2021). Network medicine framework for identifying drug-repurposing opportunities for COVID-19. PNAS *118*. e2025581118. https://doi.org/10.1073/pnas.2025581118. https://www.pnas.org/content/118/19/e2025581118.
14. Gordon, D.E., Jang, G.M., Bouhaddou, M., Xu, J., Obernier, K., White, K.M., O'Meara, M.J., Rezelj, V.V., Guo, J.Z., Swaney, D.L., Tummino, T.A., et al. (2020). A SARS-CoV-2 protein interaction map reveals targets for drug repurposing. Nature *583*, 459–468. https://doi.org/10.1038/s41586-020-2286-9. https://www.nature.com/articles/s41586-020-2286-9.
15. Lamb, J., Crawford, E.D., Peck, D., Modell, J.W., Blat, I.C., Wrobel, M.J., Lerner, J., Brunet, J.-P., Subramanian, A., Ross, K.N., et al. (2006). The Connectivity Map: using gene-expression signatures to connect small molecules, genes, and disease. Science *313*, 1929–1935. https://www.science.org/doi/abs/10.1126/science.1132939.
16. Subramanian, A., Narayan, R., Corsello, S.M., Peck, D.D., Natoli, T.E., Lu, X., Gould, J., Davis, J.F., Tubelli, A.A., Asiedu, J.K., et al. (2017). A next generation Connectivity Map: L1000 platform and the first 1,000,000 profiles. Cell *171*, 1437–1452.e17. https://doi.org/10.1016/j.cell.2017.10.049.
17. Andersen, P.I., Ianevski, A., Lysvand, H., Vitkauskiene, A., Oksenych, V., Bjørås, M., Telling, K., Lutsar, I., Dampis, U., Irie, Y., et al. (2020). Discovery and development of safe-in-man broad-spectrum antiviral agents. Int. J. Infect. Dis *93*, 268–276.
18. Cremonesi, P., Koren, Y., and Turrin, R. (2010). Performance of recommender algorithms on top-n recommendation tasks. In Proceedings of the Fourth ACM Conference on Recommender Systems, pp. 39–46.
19. Galeano, D., Li, S., Gerstein, M., and Paccanaro, A. (2020). Predicting the frequencies of drug side effects. Nat. Commun. *11*, 1–14.
20. Lee, D.D., and Seung, H.S. (1999). Learning the parts of objects by non-negative matrix factorization. Nature *401*, 788–791.
21. Lee, D.D., and Seung, H.S. (2001). Algorithms for non-negative matrix factorization. In Advances in Neural Information Processing Systems, T. Leen, T. Dietterich, and V. Tresp, eds. (Advances in Neural Information Processing Systems), pp. 556–562.
22. Bakal, G., Kilicoglu, H., and Kavuluru, R. (2019). Non-negative matrix factorization for drug repositioning: experiments with the repoDB dataset. In AMIA Annual Symposium Proceedings, *2019* (American Medical Informatics Association), p. 238.
23. Ceddia, G., Pinoli, P., Ceri, S., and Masseroli, M. (2020). Matrix factorization-based technique for drug repurposing predictions. IEEE J. Biomed. Health Inform. *24*, 3162–3172.
24. Tang, X., Cai, L., Meng, Y., Xu, J., Lu, C., and Yang, J. (2021). Indicator regularized non-negative matrix factorization method-based drug repurposing for COVID-19. Front. Immunol. *11*, 3824.
25. Dowden, H., and Munro, J. (2019). Trends in clinical success rates and therapeutic focus. Nat. Rev. Drug Discov. *18*, 495.
26. Li, T., and Ding, C. (2006). The relationships among various nonnegative matrix factorization methods for clustering. In Sixth International Conference on Data Mining (ICDM'06) (IEEE), pp. 362–371.
27. Vanunu, O., Magger, O., Ruppin, E., Shlomi, T., and Sharan, R. (2010). Associating genes and protein complexes with disease via network propagation. Plos Comput. Biol. *6*, e1000641.

28. Mordelet, F., and Vert, J.-P. (2011). Prodige: prioritization of disease genes with multitask machine learning from positive and unlabeled examples. BMC Bioinformatics *12*, 1–15.

29. Guney, E., Menche, J., Vidal, M., and Barábasi, A.-L. (2016). Network-based in silico drug efficacy screening. Nat. Commun. *7*, 10331. https://doi.org/10.1038/ncomms10331. https://www.nature.com/articles/ncomms10331.

30. Cheng, F., Desai, R.J., Handy, D.E., Wang, R., Schneeweiss, S., Barabási, A.-L., and Loscalzo, J. (2018). Network-based approach to prediction and population-based validation of in silico drug repurposing. Nat. Commun. *9*, 1–12. https://doi.org/10.1038/s41467-018-05116-5. https://www.nature.com/articles/s41467-018-05116-5.

31. Yıldırım, M.A., Goh, K.-I., Cusick, M.E., Barabási, A.-L., and Vidal, M. (2007). Drug-target network. Nat. Biotechnol. *25*, 1119–1126.

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

33. Zhou, Y., Hou, Y., Shen, J., Huang, Y., Martin, W., and Cheng, F. (2020). Network-based drug repurposing for novel coronavirus 2019-nCoV/SARS-CoV-2. Cell Discov. *6*, 1–18.

34. Ji, X., and Li, Z. (2020). Medicinal chemistry strategies toward host targeting antiviral agents. Med. Res. Rev. *40*, 1519–1557. https://doi.org/10.1002/med.21664. https://onlinelibrary.wiley.com/doi/abs/10.1002/med.21664.

35. Cao, M., Zhang, H., Park, J., Daniels, N.M., Crovella, M.E., Cowen, L.J., and Hescott, B. (2013). Going the distance for protein function prediction: a new distance metric for protein interaction networks. PLoS One *8*, e76339. https://doi.org/10.1371/journal.pone.0076339. https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0076339.

36. Smola, A.J., and Kondor, R. (2003). Kernels and regularization on graphs. In Learning Theory and Kernel Machines, Lecture Notes in Computer Science, B. Schölkopf and M.K. Warmuth, eds. (Springer), pp. 144–158. https://doi.org/10.1007/978-3-540-45167-9_12.

37. Vishwanathan, S.V.N., Schraudolph, N.N., Kondor, R., and Borgwardt, K.M. (2010). Graph kernels. J. Mach. Learn. Res. *11*, 1201–1242. http://jmlr.org/papers/v11/vishwanathan10a.html.

38. Re, M., and Valentini, G. (2012). Cancer module genes ranking using kernelized score functions. BMC Bioinformatics *13*, S3. https://doi.org/10.1186/1471-2105-13-S14-S3.

39. Re, M., Mesiti, M., and Valentini, G. (2012). A fast ranking algorithm for predicting gene functions in biomolecular networks. IEEE/ACM Trans. Comput. Biol. Bioinform. *9*, 1812–1818. https://doi.org/10.1109/TCBB.2012.114.

40. Yan, R., Zhang, Y., Li, Y., Xia, L., Guo, Y., and Zhou, Q. (2020). Structural basis for the recognition of SARS-CoV-2 by full-length human ACE2. Science *367*, 1444–1448.

41. Blanco-Melo, D., Nilsson-Payant, B.E., Liu, W.-C., Uhl, S., Hoagland, D., Møller, R., Jordan, T.X., Oishi, K., Panis, M., Sachs, D., et al. (2020). Imbalanced host response to SARS-CoV-2 drives development of COVID-19. Cell *181*, 1036–1045.e9. https://doi.org/10.1016/j.cell.2020.04.026. https://www.cell.com/cell/abstract/S0092-8674(20)30489-X.

42. Sirota, M., Dudley, J.T., Kim, J., Chiang, A.P., Morgan, A.A., Sweet-Cordero, A., Sage, J., and Butte, A.J. (2011). Discovery and preclinical validation of drug indications using compendia of public gene expression data. Sci. Transl Med. *3*, 96ra77. https://doi.org/10.1126/scitranslmed.3001318.

43. Food and Drug Administration (2020). Clinical Research. https://www.fda.gov/patients/drug-development-process/step-3-clinical-research#Clinical_Research_Phase_Studies.

44. Food and Drug Administration (2020). Drug Development Process. https://www.fda.gov/patients/learn-about-drug-and-device-approvals/drug-development-process.

45. Li, Z., and Yang, L. (2020). Underlying mechanisms and candidate drugs for COVID-19 based on the Connectivity Map database. Front. Genet. *11*, 1168. https://doi.org/10.3389/fgene.2020.558557. https://www.frontiersin.org/article/10.3389/fgene.2020.558557.

46. Sendama, W. (2020). L1000 Connectivity Map interrogation identifies candidate drugs for repurposing as SARS-CoV-2 antiviral therapies. Comput. Struct. Biotechnol. J. *18*, 3947–3949. https://doi.org/10.1016/j.csbj.2020.11.054. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7719280/.

47. Wishart, D.S., Feunang, Y.D., Guo, A.C., Lo, E.J., Marcu, A., Grant, J.R., Sajed, T., Johnson, D., Li, C., Sayeeda, Z., et al. (2018). DrugBank 5.0: a major update to the DrugBank database for 2018. Nucleic Acids Res. *46*, D1074–D1082. https://doi.org/10.1093/nar/gkx1037.

48. Zhou, Y., Wang, F., Tang, J., Nussinov, R., and Cheng, F. (2020). Artificial intelligence in COVID-19 drug repurposing. Lancet Digit. Health *2* (12), e667–e676. https://doi.org/10.1016/S2589-7500(20)30192-8. https://www.thelancet.com/journals/landig/article/PIIS2589-7500(20)30192-8/abstract.

49. Riva, L., Yuan, S., Yin, X., Martin-Sancho, L., Matsunaga, N., Pache, L., Burgstaller-Muehlbacher, S., De Jesus, P.D., Teriete, P., Hull, M.V., et al. (2020). Discovery of SARS-CoV-2 antiviral drugs through large-scale compound repurposing. Nature *586*, 1–11. https://doi.org/10.1038/s41586-020-2577-1. https://www.nature.com/articles/s41586-020-2577-1.

50. Luck, K., Kim, D.-K., Lambourne, L., Spirohn, K., Begg, B.E., Bian, W., Brignall, R., Cafarelli, T., Campos-Laborie, F.J., Charloteaux, B., et al. (2020). A reference map of the human binary protein interactome. Nature *580*, 402–408. https://doi.org/10.1038/s41586-020-2188-x. https://www.nature.com/articles/s41586-020-2188-x.

51. Szklarczyk, D., Gable, A.L., Lyon, D., Junge, A., Wyder, S., Huerta-Cepas, J., Simonovic, M., Doncheva, N.T., Morris, J.H., Bork, P., et al. (2019). STRING v11: protein–protein association networks with increased coverage, supporting functional discovery in genome-wide experimental datasets. Nucleic Acids Res. *47*, D607–D613. https://doi.org/10.1093/nar/gky1131. https://academic.oup.com/nar/article/47/D1/D607/5198476.

52. Nepusz, T., Yu, H., and Paccanaro, A. (2012). Detecting overlapping protein complexes in protein-protein interaction networks. Nat. Methods *9*, 471–472. https://doi.org/10.1038/nmeth.1938. https://www.nature.com/articles/nmeth.1938.

53. Kuleshov, M.V., Jones, M.R., Rouillard, A.D., Fernandez, N.F., Duan, Q., Wang, Z., Koplev, S., Jenkins, S.L., Jagodnik, K.M., Lachmann, A., et al. (2016). Enrichr: a comprehensive gene set enrichment analysis web server 2016 update. Nucleic Acids Res. *44*, W90–W97. https://doi.org/10.1093/nar/gkw377. https://academic.oup.com/nar/article/44/W1/W90/2499357.

54. Sadegh, S., Matschinske, J., Blumenthal, D.B., Galindez, G., Kacprowski, T., List, M., Nasirigerdeh, R., Oubounyt, M., Pichlmair, A., Rose, T.D., et al. (2020). Exploring the SARS-CoV-2 virus-host-drug interactome for drug repurposing. Nat. Commun. *11*, 3518. https://doi.org/10.1038/s41467-020-17189-2. https://www.nature.com/articles/s41467-020-17189-2.

55. Galeano, D., Li, S., Gerstein, M., and Paccanaro, A. (2020). Predicting the frequency of drug side effects. Nat Commun *11*, 4575. https://doi.org/10.1038/s41467-020-18305-y.

56. Hassanipour, S., Arab-Zozani, M., Amani, B., Heidarzad, F., Fathalipour, M., and Martinez-de Hoyo, R. (2021). The efficacy and safety of Favipiravir in treatment of COVID-19: a systematic review and meta-analysis of clinical trials. Sci. Rep. *11*, 1–11.

57. Consortium, W.S.T. (2021). Repurposed antiviral drugs for Covid-19—interim WHO solidarity trial results. New Engl. J. Med. *384*, 497–511.

58. Kalil, A.C. (2020). Treating COVID-19—off-label drug use, compassionate use, and randomized clinical trials during pandemics. JAMA *323*, 1897–1898.

59. Grein, J., Ohmagari, N., Shin, D., Diaz, G., Asperges, E., Castagna, A., Feldt, T., Green, G., Green, M.L., Lescure, F.-X., et al. (2020). Compassionate use of remdesivir for patients with severe COVID-19. New Engl. J. Med. *382*, 2327–2336.

60. Kost-Alimova, M., Sidhom, E.-H., Satyam, A., Chamberlain, B.T., Dvela-Levitt, M., Melanson, M., Alper, S.L., Santos, J., Gutierrez, J., Subramanian, A., et al. (2020). A high-content screen for mucin-1-reducing compounds identifies fostamatinib as a candidate for rapid repurposing

Please cite this article in press as: Santos et al., Machine learning and network medicine approaches for drug repositioning for COVID-19, Patterns (2021), https://doi.org/10.1016/j.patter.2021.100396

for acute lung injury. Cell Rep. Med. *1*, 100137. https://doi.org/10.1016/j.xcrm.2020.100137. https://www.sciencedirect.com/science/article/pii/S2666379120301816.

61. Radulesco, T., Lechien, J.R., Saussez, S., Hopkins, C., and Michel, J. (2021). Safety and impact of nasal lavages during viral infections such as SARS-CoV-2. Ear Nose Throat J. *100*, 188S–191S. https://doi.org/10.1177/0145561320950491.

62. Wang, B., Kovalchuk, A., Li, D., Rodriguez-Juarez, R., Ilnytskyy, Y., Kovalchuk, I., and Kovalchuk, O. (2020). In search of preventive strategies: novel high-CBD *Cannabis sativa* extracts modulate ACE2 expression in COVID-19 gateway tissues. Aging *12*, 22425–22444. https://doi.org/10.18632/aging.202225.

63. Suba, Z. (2020). Prevention and therapy of COVID-19 via exogenous estrogen treatment for both male and female patients: prevention and therapy of COVID-19. J. Pharm. Pharm. Sci. *23*, 75–85. https://doi.org/10.18433/jpps31069. https://journals.library.ualberta.ca/jpps/index.php/JPPS/article/view/31069.

64. Melms, J.C., Biermann, J., Huang, H., Wang, Y., Nair, A., Tagore, S., Katsyv, I., Rendeiro, A.F., Amin, A.D., Schapiro, D., et al. (2021). A molecular single-cell lung atlas of lethal COVID-19. Nature *595*, 114–119. https://doi.org/10.1038/s41586-021-03569-1. https://www.nature.com/articles/s41586-021-03569-1.

65. Delorey, T.M., Ziegler, C.G.K., Heimberg, G., Normand, R., Yang, Y., Segerstolpe, A., Abbondanza, D., Fleming, S.J., Subramanian, A., Montoro, D.T., et al. (2021). COVID-19 tissue atlases reveal SARS-CoV-2 pathology and cellular targets. Nature *595*, 107–113. https://doi.org/10.1038/s41586-021-03570-8. https://www.nature.com/articles/s41586-021-03570-8.

66. Zhou, P., Yang, X.-L., Wang, X.-G., Hu, B., Zhang, L., Zhang, W., Si, H.-R., Zhu, Y., Li, B., Huang, C.-L., et al. (2020). A pneumonia outbreak associated with a new coronavirus of probable bat origin. Nature *579*, 270–273. https://doi.org/10.1038/s41586-020-2012-7. https://www.nature.com/articles/s41586-020-2012-7.

67. Hoffmann, M., Kleine-Weber, H., Schroeder, S., Krüger, N., Herrler, T., Erichsen, S., Schiergens, T.S., Herrler, G., Wu, N.-H., Nitsche, A., et al. (2020). SARS-CoV-2 cell entry depends on ACE2 and TMPRSS2 and is blocked by a clinically proven protease inhibitor. Cell *181*, 271–280.e8. https://doi.org/10.1016/j.cell.2020.02.052. https://www.cell.com/cell/abstract/S0092-8674(20)30229-4.

68. Zhao, M.-M., Yang, W.-L., Yang, F.-Y., Zhang, L., Huang, W.-J., Hou, W., Fan, C.-F., Jin, R.-H., Feng, Y.-M., Wang, Y.-C., and Yang, J.-K. (2021). Cathepsin L plays a key role in SARS-CoV-2 infection in humans and humanized mice and is a promising target for new drug development. Signal Transduct. Targeted Ther. *6*, 1–12. https://doi.org/10.1038/s41392-021-00558-8. https://www.nature.com/articles/s41392-021-00558-8.

69. Lieberman, N.A.P., Peddu, V., Xie, H., Shrestha, L., Huang, M.-L., Mears, M.C., Cajimat, M.N., Bente, D.A., Shi, P.-Y., Bovier, F., et al. (2020). In vivo antiviral host transcriptional response to SARS-CoV-2 by viral load, sex, and age. PLoS Biol. *18*, 1–17. https://doi.org/10.1371/journal.pbio.3000849.

70. Edgar, R., Domrachev, M., and Lash, A.E. (2002). Gene Expression Omnibus: NCBI gene expression and hybridization array data repository. Nucleic Acids Res. *30*, 207–210. https://doi.org/10.1093/nar/30.1.207.

71. Barrett, T., Wilhite, S.E., Ledoux, P., Evangelista, C., Kim, I.F., Tomashevsky, M., Marshall, K.A., Phillippy, K.H., Sherman, P.M., Holko, M., et al. (2013). NCBI GEO: archive for functional genomics data sets—update. Nucleic Acids Res. *41*, D991–D995. https://doi.org/10.1093/nar/gks1193.

72. Robinson, M.D., McCarthy, D.J., and Smyth, G.K. (2010). edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. Bioinformatics *26*, 139–140. https://doi.org/10.1093/bioinformatics/btp616. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2796818/.

73. U. N. L. of Medicine (2020). Listed Clinical Studies Related to the Coronavirus Disease (Covid-19). https://clinicaltrials.gov/ct2/results?cond=COVID-19.

74. Ghandikota, S., Sharma, M., and Jegga, A.G. (2021). Secondary analysis of transcriptomes of SARS-CoV-2 infection models to characterize COVID-19. Patterns *2*, 100247. https://doi.org/10.1016/j.patter.2021.100247. https://www.sciencedirect.com/science/article/pii/S2666389921000672.

75. Berry, M.W., Browne, M., Langville, A.N., Pauca, V.P., and Plemmons, R.J. (2007). Algorithms and applications for approximate nonnegative matrix factorization. Comput. Stat. Data Anal. *52*, 155–173.

76. Zhou, D., Bousquet, O., Lal, T.N., Weston, J., and Schölkopf, B. (2004). Learning with Local and Global Consistency (Max Planck Institute for Biological Cybernetics), p. 8.

77. Kondor, R., and Vert, J.-P. (2004). Diffusion kernels. In Kernel Methods in Computational Biology, B. Schölkopf, K. Tsuda, and J.-P. Vert, eds. (The MIT Press). https://doi.org/10.7551/mitpress/4057.003.0011 https://direct.mit.edu/books/book/3898/chapter/163650/di-usion-kernels.

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