Bielefeld University

Faculty of Technology

Scientific Report

Bioinformatics and Data Science of Pandemics

Drug Repurposing

submitted by

Marco Sohn

Matriculation Number: 3101604

Supervisor Prof. Dr. Alexander Schönhuth (Genome Data Science)

Bielefeld, 29 September 2021

Contents

1 Introduction

Drug repurposing is a strategy for identifying new therapeutic uses for existing drugs. In the past, drug repurposing happened mostly by serendipity, as exemplified by the drug sildenafil, which was originally intended to treat hypertension and angina pectoris (see Tab. 1.1). However, in recent years there have been many advances in the field of human genomics and network biology that allow a more efficient approach.

Drug repurposing is becoming an increasingly relevant topic, as it is a time- and cost-effective method compared to de novo drug development (Pushpakom et al., 2018). It is a possible alternative, especially for neglected diseases for which conventional drug development is cost- and time-intensive and not profitable due to fewer disease cases. But also for sudden challenges like the COVID-19 pandemic, drug repurposing can find a possible treatment through algorithms and the underlying data. The cost and time savings can be explained by the lower number of clinical tests required and by the pharmaceutical supply chains already in place, which facilitate the path to market release (Rosa & Santos, 2020).

This scientific report is mainly concerned with the paper "Network medicine framework for identifying drug-repurposing opportunities for COVID-19" (Morselli Gysi et al., 2021). Contributions from other scientific work are marked accordingly. In chapter Current State of Drug Repurposing, current problems regarding existing drug repurposing methods are discussed. Subsequently, chapter Methods exemplifies which input drug repurposing algorithms require in order to determine possible drug candidates and, in particular, three algorithms based on artificial intelligence, network diffusion, and network proximity are examined. In chapter Evaluation, the study design, the results of the three algorithms used, a multimodal approach, and the proposed drugs against SARS-CoV-2 are discussed. Finally, chapter Conclusion summarizes the most important aspects.

Table 1.1: Examples of Drug Repurposing.

Drug	Original indication	Repurposing
Slidenafil	Angina pectoris	Male erectile dysfunction
Phentolamine	Hypertension	Impaired night vision
Minoxidil	Hypertension	Hair loss
Finasteride	Prostate enlargement	Hair loss
Sibutramine	Depression	Obesity

2 Current State of Drug Repurposing

A current problem of drug repurposing is that there are hardly any reliable drug repurposing methods. Realizing the full potential of medical data is still a challenge (Xue et al., 2018). There are many strategies, each with its own advantages and disadvantages, but they all have in common that they have room for improvement in terms of sensitivity of the proposed drugs (Jarada et al., 2020). Since most drug repurposing methods are not yet reliable in terms of the predicted drug candidates, and since in an experimental screening never all available drugs are tested due to cost and time constraints, a winner-takes-all pattern emerges in which studies focus only on the drugs with the presumed best outcome. Therefore, other possible drug candidates could be neglected. An example of this winner-takes-all pattern are the drugs hydroxychloroquine and chloroquine, which were increasingly focused on during the COVID-19 pandemic (Horby et al., 2021, Ghazy et al., 2020), while many other potential drugs were hardly considered. For this reason, there is a need for efficient strategies that allow drug prioritization and consider as many drugs as possible to achieve the best possible outcome. In order to validate these new strategies, which rank drugs based on molecular profiles, chemical structures, or network perturbations, and to determine their true predictive power, the majority of the predicted drug candidates must be tested experimentally.

3 Methods

3.1 Input Data

In order to make drug repurposing feasible at all, a number of underlying data are needed. Besides the human interactome, which is a set of protein-protein interactions, targets of pathogens or diseases and a set of drugs with drug targets are needed.

3.2 Artificial Intelligence

The artificial intelligence approach is based on a graph neural network architecture (Zitnik, Agrawal, et al., 2018). For this, a multimodal graph $G = (V, R)$ with N nodes $v_i \in V$ of three different types representing drugs, targets and diseases. The edges consist of a triplet (v_i, r, v_j) capturing four different types r of interactions: protein-protein interaction, disease-protein association, target-drug association and drug-disease treatment. The multimodal graph consists of an encoder, which is a graph convolutional network and computes different d-dimensional $z_i \in R^d$ embeddings for drug and disease nodes on G, and a decoder, which takes the embeddings and reconstructs them into a triplet using a function g, which are then used to predict drug scores, representing how promising a given drug is. Different pipelines are possible here, which rely on different drug-disease embeddings.

3.3 Network Diffusion

The network diffusion approach is based on the graph diffusion property (Cao et al., 2013) to define a similarity metric for node pairs. Then, the similarity of the nodes is considered in terms of their impact on the network.

The vector $He(V_i) = [He(V_i, V_1), He(V_i, V_1), He(V_i, V_1), ..., He(V_i, V_n)]$ describes

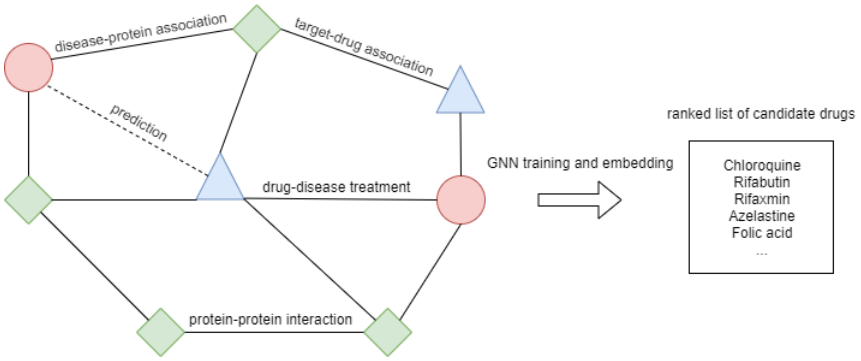

Figure 3.1: Visualisation of the artificial intelligence approach.

how a perturbation from one node impacts the other nodes in the human interactome, where $He(A, B)$ represents the expected number of random walkers starting at node A and visiting node B. Then the pairwise similarity scores between all nodes are calculated with the L1-norm $DSD(A, B) = ||He(A) - He(B)||$.

Different pipelines result from the application of distinct statistical measures such as Eq. 3.1, where $DSD(t, v)$ represents the diffusion state distance between node t and v, which indicates the impact of drug target t on the SARS-CoV-2 host protein target v. If instead the Kullback-Leibler divergence is applied to prevent a possible loss of information through the L1 norm, two further statistical measures arise. For Eq. 3.2, the minimum and for Eq. 3.3, the medium of divergence is considered. In order to consider a symmetrical measure, the Jensen-Shannon divergence is regarded (see Eq. 3.4, 3.5).

$$I \min_{DSD} = \frac{1}{||T||} \sum_{t \in T} \min_{v \in V} DSD(t, v) \tag{3.1}$$

$$I \min_{KL} = \frac{1}{|T||} \sum_{t \in T} \min_{v \in V} KL(t, v) \tag{3.2}$$

$$I \operatorname{median}_{KL} = \frac{1}{||T|} \sum_{t \in T} \operatorname{median}_{v \in V} KL(t, v) \tag{3.3}$$

$$I \min_{JS} = \frac{1}{||T||} \sum_{t \in T} \min_{v \in V} JS(t, v) \tag{3.4}$$

$$I \operatorname{median}_{JS} = \frac{1}{|T||} \sum_{t \in T} \operatorname{median}_{v \in V} JS(t, v) \tag{3.5}$$

3.4 Network Proximity

The network proximity approach is based on the distance $d(v, t)$ between host protein targets of SARS-CoV-2 V and closest protein targets of drugs T. This shortest path length is calculated by $d_c(V, T) = \frac{1}{||T||} \sum_{t \in T} \min_{v \in V} d(v, t)$.

Using degree-binning (Guney et al., 2016), expected distances between two randomly selected groups of proteins, matching the size and degree of the original V and T, are calculated. Now, using the mean μ_{dc} and the standard deviation σ_{dc}, the relative distance $Z_{dc} = \frac{d_c - \mu_{dc}(V,T)}{\sigma_{dc}(V,T)}$ to the known absolute distance d_c is calculated.

There are three different pipelines for this approach. The first pipeline contains all drug targets. The second pipeline removes all proteins that had functions in drug delivery and metabolism to see if removing them can increase predictive power. In the third pipeline, only differentially expressed genes are considered to assess whether they increase sensitivity.

4 Evaluation

4.1 Study Design and Timeline

The study design is depicted in Fig. 4.1. The human interactome was assembled from 21 public databases with a total of over 18508 proteins and 332749 protein-protein interactions and serves together with the 320 host protein targets of SARS-CoV-2 and the 7591 drugs with 4187 drug targets as the starting point for the drug repurposing predictions.

A total of 12 different pipelines were tested in the study. The four predictive ranking lists of the artificial intelligence pipelines each predicted a total of 1670 drugs, while the five pipelines of the network diffusion determined 6125 drugs each. In the network proximity approach, the three pipelines predicted different numbers of drugs, ranging from 6100 drugs for P1 to 769 drugs for P3.

Subsequently, 918 drugs were identified for which all pipelines except P3 made predictions. The effect of these drugs on the virus was then experimentally validated in VeroE6 cells isolated from African green monkey renal epithelial cells (E918). In Fig. 4.2, the dose-response curves with the four possible outputs (S,W,C,N) are shown. The drugs azelastine and ivermectin exhibited a strong response, while cinacalcet, for example, proved to have only a weak effect. Drugs that reduced cells, such as colchicine or doxorubicin, were classified as cytotoxic. Drugs that reduced neither cells nor viruses were collected under the outcome no-effect. A total of 35 drugs were classified as cytotoxic. 37 drugs had a strong effect, 40 had a weak effect on the virus and 806 had no effect at all.

For CT415, clinicaltrials.gov was scanned, finding 67 drugs in 134 clinical trials for SARS-CoV-2. These were then compared to the E918 drugs, leading to 37 drugs that showed a strong reaction being classified as positive and the others as negative. The results of these studies are mostly unknown. Therefore, CT415 will be used to

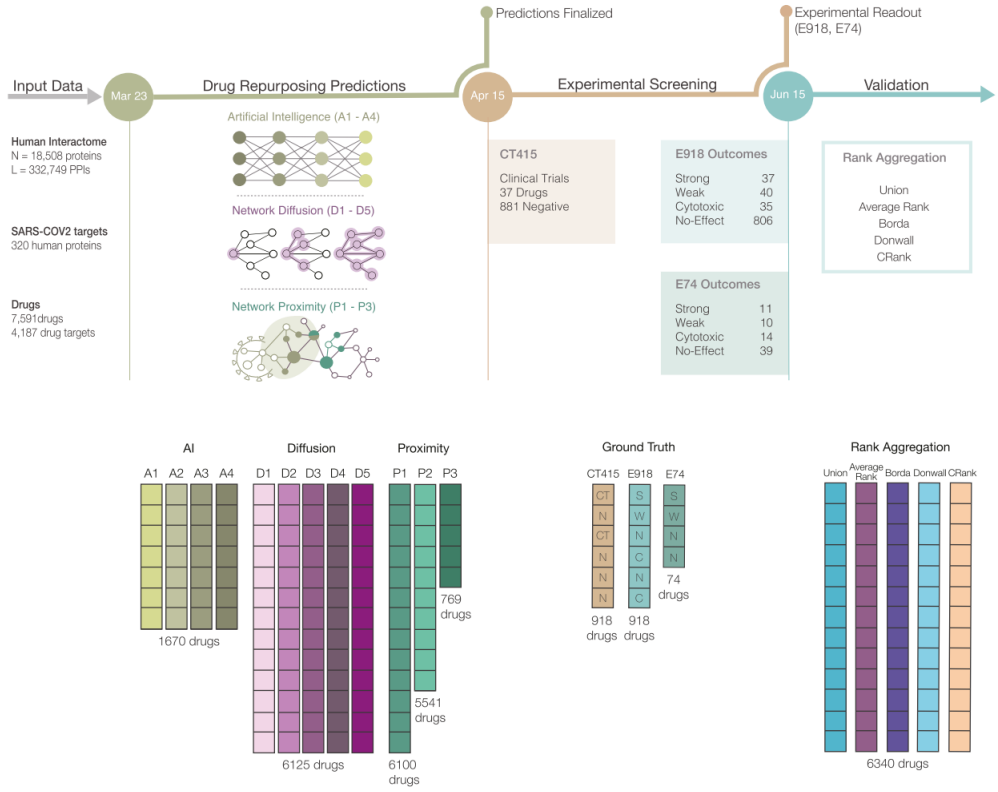

Figure 4.1: Network medicine framework for drug repurposing.

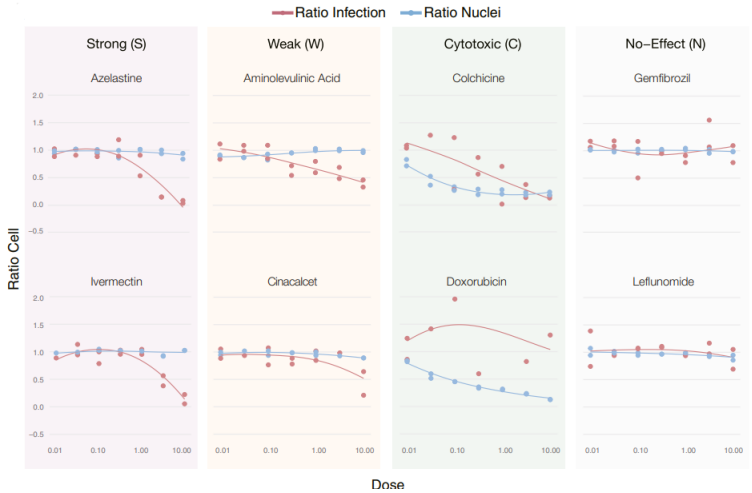

Figure 4.2: Dose–response curves for eight of the E918 drugs, depicting the four observed outcomes (Strong, Weak, Cytotoxic, and No-Effect).

test the extent to which the pipelines are consistent with the scientific consensus on options for treating SARS-CoV-2.

Furthermore, 76 of the 77 drugs that had a strong or weak effect on reducing viral activity were found not to bind the proteins targeted by SARS-CoV-2. In other words, 76 drugs that showed efficacy are network drugs that cannot be detected with docking-based strategies. Fig. 4.3 illustrates three different mechanisms of action of drugs.

4.2 Pipeline Results

The evaluation showed that the artificial intelligence approach has strong predictive power for drugs selected for clinical trials (CT415). The proximity approach, on the other hand, demonstrated strong predictive power for the E918 experimental outcome. The network diffusion approach showed weak predictive power for both CT415 and E918.

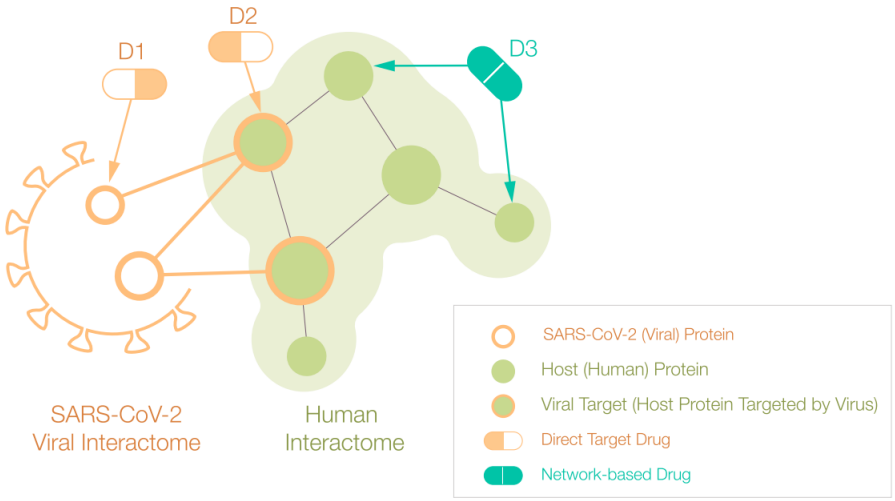

Figure 4.3: Direct target drugs bind either to a viral protein (D1) or to a host protein target of the viral proteins (D2). Network drugs (D3) bind to the host proteins and limit viral activity by perturbing the host subcellular network (Morselli Gysi et al., 2021).

Figure 4.4 shows the correlation of the different approaches. It can be seen that there is high correlation among the ranking lists predicted by the same method. However, there is no correlation between all ranking lists predicted by A, P and D.

4.3 Multimodal Approach

The aim of the multimodal approach is a consensus ranking out of several independently constructed rankings. This is expected to increase sensitivity by combining the joint predictive power of all pipelines. A classical approach would be to maximize the number of pairwise agreements between the final ranking and each input. However, this procedure is NP-hard, also known as Kemeny consensus (Dwork et al., 2001). Therefore, heuristics or approximates have to provide a solution, of which two approaches will be discussed in the following.

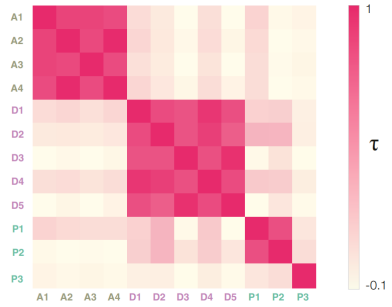

Figure 4.4: Heatmap of the Kendall τ statistic.

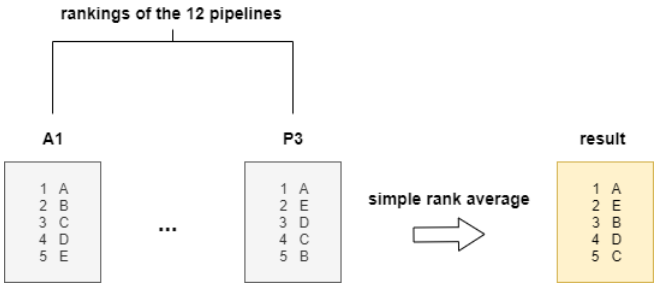

Figure 4.5: Visualization of the Average Ranking method.

4.3.1 Average Ranking Method

The average ranking method calculates a simple rank average over 12 rankings returned by the pipelines to obtain the overall ranking as visualized in Fig 4.5. Studies have shown that the average ranking method is rather a poor aggregation approach (Eilbeck et al., 2017).

4.3.2 cRank

The cRank algorithm (Zitnik, Sosič, et al., 2018) is a more sophisticated approach to determine the overall ranking. As input, it takes a ranked list of drugs, R_r, each one emerging from a different pipeline, r. These ranked lists are partitioned into equally sized bags i, where each bag i has an importance weight $K_r^{i_r}$ whose initial values are all equal. The aggregation rank R is determined as the weighted average

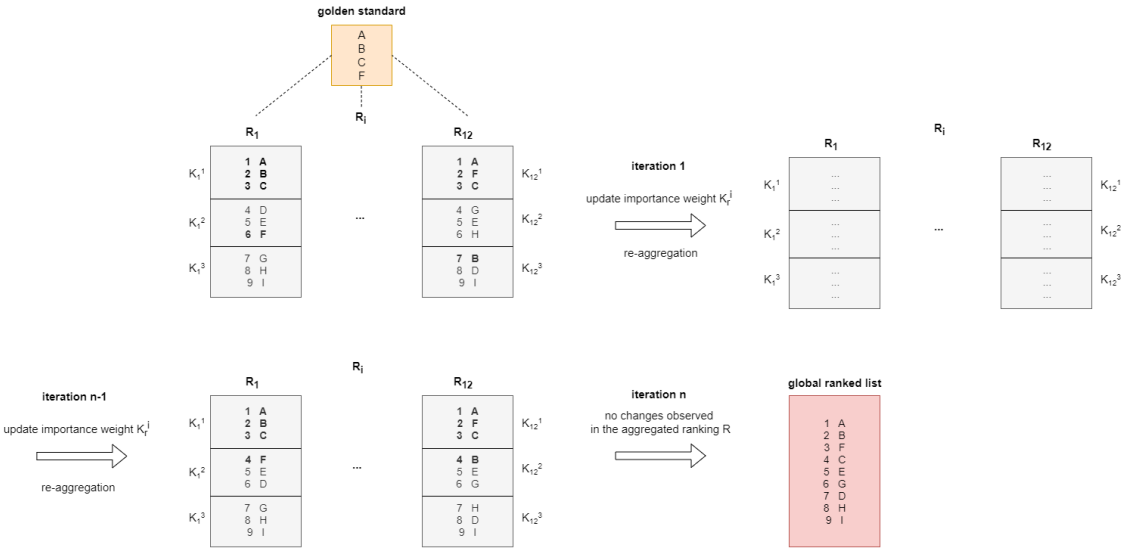

Figure 4.6: Visualization of the cRank approach.

of all ranked lists R_r. This aggregation rank is important for determining the end of the repeating two-stage procedure, as the algorithm ends when no changes in R are detected.

Fig. 4.6 depicts the algorithm. One iteration consists of two stages: In the first stage, the current aggregated ranking R is used to update the importance weights K_r^{ir} for all R_r. Here, the top-ranked drugs in R serve as the current golden standard. This golden standard is then used to update the importance weight K_r^{ir} for each bag i in ranked list R_r by regarding how many drugs of the golden standard are in bag i. In the second stage, the ranked lists are re-aggregated based on the previously calculated importance weights. For this, the new ranking $R(a)$ of a drug a of the ranking r is determined by $R(a) = \sum_r K_r^{ir(a)} R_r(a)$, where $K_r^{ir(a)}$ indicates the importance weight of bag $i_r(a)$ of a drug a in ranking r and $R_r(a)$ is the rank of drug a according to r. The output is the ranked list R of all drugs, which contains the overall assessment of all pipelines.

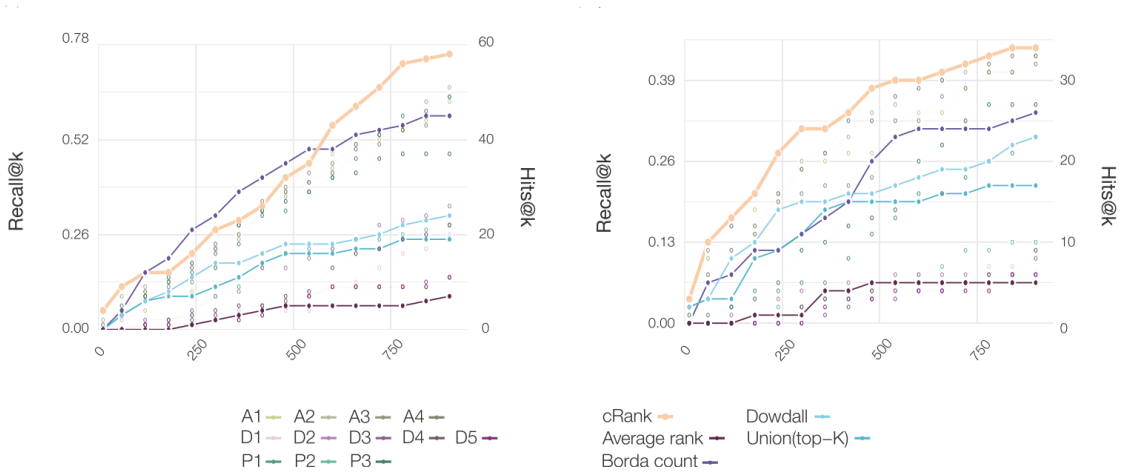

Figure 4.7: Top K precision and recall for the rank aggregation methods (connected points) and individual pipelines (empty symbols).

4.3.3 Results of Ranking Approaches

Fig. 4.7 illustrates how different rank aggregation algorithms performed in terms of precision and recall. Here, the average rank achieved the worst result in comparison and performs similarly to the network diffusion pipelines. The cRank algorithm, on the other hand, has the highest F1-score and shows consistently high predictive performance. Thus, cRank even performs better than the best individual pipelines (A1, A2, P1). This has demonstrated that a multimodal approach can be an important component of drug repurposing if one wants to combine different algorithms and their informative value.

4.4 Proposed Drugs

Tab. 4.1 shows drugs and their original indications that have been proposed by the multimodal approach. It can be seen that both hydroxychloroquine and chloroquine are listed, which have been in focus as possible drug candidates against SARS-CoV-2. However, in addition to these already known drugs, potential new candidates

Table 4.1: Proposed drugs and original indication.

Drug	Original indication
Chloroquine	Malaria
Azelastine	Allergic rhinitis
Methotrexate	Cancer, autoimmune diseases
Digoxin	Atrial fibrillation
Hydroxychloroquine	Malaria
Omeprazole	Gastroesophageal reflux disease

such as azelastine or omeprazole have also been predicted. The extent to which the suggestions actually help against SARS-CoV-2 must be tested in human studies, as is currently being done by the CARVIN study of the University Hospital of Cologne with the drug azelastine (Klußmann, 2021). The table also exibits that the original indications of the proposed drugs are very different, so drugs used in different fields of application could help in treating SARS-CoV-2.

5 Conclusion

Drug repurposing holds considerable potential to find new drug candidates time and cost-effectively due to the numerous possible implementations of algorithms and the combination of their predictive ranking lists.

This scientific report has focused on drug repurposing for the SARS-CoV-2 virus. For this purpose, algorithms based on artificial intelligence, network diffusion, and network proximity have been considered. The evaluation showed that these three approaches did not provide consistently reliable results. However, combining the 12 predictive ranking lists of these algorithms into a multimodal approach using the cRank algorithm clearly increased the accuracy. It was also observed that 76 out of 77 drugs are network drugs, i.e. they do not bind a target directly but reduce the viral load by perturbing the host subcellular network. Some of the potential drug candidates identified are already known in the scientific community or are undergoing further clinical testing. Overall, the network medicine framework offers a possibility to efficiently identify repurposable drugs for neglected diseases or future pathogens.

Further improvements are feasible if binding predictions between drug and target organism are used in addition to experimentally known bindings. The use of gene expression data sets perturbed by the drugs studied and the results of clinical trials that could be used to rank the different pipelines could also improve the sensitivity of the framework.

6 Bibliography

Cao, M., Zhang, H., Park, J., Daniels, N., Crovella, M., Cowen, L., & 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

Dwork, Cynthia, Kumar, P., Ravi, Naor, Moni, & Sivakumar, D. (2001). Rank aggregation methods for the web. *WWW*. https://doi.org/10.1145/371920.372165

Eilbeck, K., Quinlan, A., & Yandell, M. (2017). Settling the score: Variant prioritization and mendelian disease. *Nature Reviews Genetics*, *18*. https://doi.org/10.1038/nrg.2017.52

Ghazy, R., Almaghraby, A., Shaaban, R., Kamal, A., Beshir, H., Moursi, A., Ramadan, A., Hamed, S., & Taha, N. (2020). A systematic review and meta-analysis on chloroquine and hydroxychloroquine as monotherapy or combined with azithromycin in covid-19 treatment. *Scientific Reports*, *10*, 1–18. https://doi.org/10.1038/s41598-020-77748-x

Guney, E., Menche, J., Vidal, M., & Barabasi, A.-L. (2016). Network-based in silico drug efficacy screening. *Nature Communications*, *7*, 10331. https://doi.org/10.1038/ncomms10331

Horby, P., Mafham, M., & Linsell, L. (2021). Hydroxychloroquine in hospitalized patients with covid-19. *New England Journal of Medicine*, *384*(9), 881–882. https://doi.org/10.1056/nejmc2035374

Jarada, T., Rokne, J., & Alhajj, R. (2020). A review of computational drug repositioning: Strategies, approaches, opportunities, challenges, and directions. *Journal of Cheminformatics*, *12*, 1–23. https://doi.org/10.1186/s13321-020-00450-7

Klußmann, J. (2021). *COVID-19: Azelastine Reduces Virus-load In Nasal swabs*. Retrieved September 28, 2021, from https://hno.uk-koeln.de/forschung/carvin-studie/

Morselli Gysi, D., do Valle, Í., Zitnik, M., Ameli, A., Gan, X., Varol, O., Ghiassian, S. D., Patten, J. J., Davey, R. A., Loscalzo, J., & Barabási, A.-L. (2021).

Network medicine framework for identifying drug-repurposing opportunities for COVID-19. *Proceedings of the National Academy of Sciences*, *118*(19). https://doi.org/10.1073/pnas.2025581118

Pushpakom, S., Iorio, F., Eyers, P., Escott, K., Hopper, S., Wells, A., Doig, A., Guilliams, T., Latimer, J., McNamee, C., Norris, A., Sanseau, P., Cavalla, D., & Pirmohamed, M. (2018). Drug repurposing: Progress, challenges and recommendations. *Nature Reviews Drug Discovery*, *18*. https://doi.org/10.1038/nrd.2018.168

Rosa, S., & Santos, W. (2020). *Clinical trials on drug repositioning for covid-19 treatment*. https://doi.org/10.26633/RPSP.2020.40

Xue, H., Li, J., Xie, H., & Wang, Y. (2018). Review of drug repositioning approaches and resources. *International journal of biological sciences*, *14*, 1232–1244. https://doi.org/10.7150/ijbs.24612

Zitnik, M., Agrawal, M., & Leskovec, J. (2018). Modeling polypharmacy side effects with graph convolutional networks. *Bioinformatics (Oxford, England)*, *34*. https://doi.org/10.1093/bioinformatics/bty294

Zitnik, M., Sosič, R., & Leskovec, J. (2018). Prioritizing network communities. *Nature Communications*, *9*. https://doi.org/10.1038/s41467-018-04948-5