Computers in Biology and Medicine 149 (2022) 106029

Contents lists available at ScienceDirect

Computers in Biology and Medicine

journal homepage: www.elsevier.com/locate/compbiomed

# Analysis of transcriptomic responses to SARS-CoV-2 reveals plausible defective pathways responsible for increased susceptibility to infection and complications and helps to develop fast-track repositioning of drugs against COVID-19

Enrique J. deAndrés-Galiana [a,c,*], Juan Luis Fernández-Martínez [a,b], Óscar Álvarez-Machancoses [a], Guillermina Bea [a,b], Carlos M. Galmarini [d], Andrzej Kloczkowski [e]

[a] Group of Inverse Problems, Optimization and Machine Learning. Department of Mathematics, University of Oviedo, C. Federico García Lorca, 18, 33007, Oviedo, Spain
[b] DeepBioInsights, Spain
[c] Department of Computer Science, University of Oviedo, C. Federico García Lorca, 18, 33007, Oviedo, Spain
[d] Topazium Artificial Intelligence, Paseo de la Castellana 40, 28046, Madrid, Spain
[e] Battelle Center for Mathematical Medicine, Nationwide Children's Hospital, and Department of Pediatrics, The Ohio State University, Columbus, OH, USA

ARTICLE INFO

*Keywords:*
*SARS-CoV-2*
*Side effects*
*Drug repositioning*
*Small scale genetic signature*
*Coronavirus*
*Machine learning*

ABSTRACT

*Background:* To understand the transcriptomic response to SARS-CoV-2 infection, is of the utmost importance to design diagnostic tools predicting the severity of the infection.
*Methods:* We have performed a deep sampling analysis of the viral transcriptomic data oriented towards drug repositioning. Using different samplers, the basic principle of this methodology the biological invariance, which means that the pathways altered by the disease, should be independent on the algorithm used to unravel them.
*Results:* The transcriptomic analysis of the altered pathways, reveals a distinctive inflammatory response and potential side effects of infection. The virus replication causes, in some cases, acute respiratory distress syndrome in the lungs, and affects other organs such as heart, brain, and kidneys. Therefore, the repositioned drugs to fight COVID-19 should, not only target the interferon signalling pathway and the control of the inflammation, but also the altered genetic pathways related to the side effects of infection. We also show via Principal Component Analysis that the transcriptome signatures are different from influenza and RSV. The gene COL1A1, which controls collagen production, seems to play a key/vital role in the regulation of the immune system. Additionally, other small-scale signature genes appear to be involved in the development of other COVID-19 comorbidities.
*Conclusions:* Transcriptome-based drug repositioning offers possible fast-track antiviral therapy for COVID-19 patients. It calls for additional clinical studies using FDA approved drugs for patients with increased susceptibility to infection and with serious medical complications.

## 1. Background

Coronaviruses are a varied group of single positive-stranded RNA viruses, with a crown-like appearance under an electron microscope, due to the presence of spike glycoproteins on the envelope. These viruses originate in bats and circulate in a wide range of hosts [1]. There are several different types of coronaviruses: Alpha-coronaviruses (alpha-CoV), Beta-coronaviruses (betaCoV), Delta-coronaviruses (deltaCoV), and Gamma-coronaviruses (gammaCoV). Most of them seem to cause colds or other mild respiratory illnesses. There have been far more deadly varieties/adaptations of coronaviruses in the past known as Severe Acute Respiratory Syndrome (SARS) and Middle East Respiratory

* Corresponding author. Group of Inverse Problems, Optimization and Machine Learning. Department of Mathematics, University of Oviedo, C. Federico García Lorca, 18, 33007, Oviedo, Spain.
*E-mail addresses:* andresenrique@uniovi.es (E.J. deAndrés-Galiana), jlfm@uniovi.es, jl.fernandez@deepbioinsights.com (J.L. Fernández-Martínez), oscar.alvarez@deepbioinsights.com (Ó. Álvarez-Machancoses), guillermina.bea@icloud.com (G. Bea), cmgalmarini@topazium.com (C.M. Galmarini), andrzej.kloczkowski@nationwidechildrens.org (A. Kloczkowski).
*URL:* https://www.deepbioinsights.com (J.L. Fernández-Martínez).

https://doi.org/10.1016/j.compbiomed.2022.106029
Received 20 April 2022; Received in revised form 8 July 2022; Accepted 20 August 2022
Available online 30 August 2022

Syndrome (MERS).

The SARS-COV outbreak was first discovered in southern China in February 2003. The outbreak lasted approximately six months and the disease propagated to more than two dozen countries in North America, South America, Europe, and Asia, before it was stopped in July 2003. It has infected more than 8,000 people and the mortality rate was around 9.6%. In late 2017, Chinese scientists traced the virus to cave-dwelling horseshoe bats in Yunnan province.

The Middle East Respiratory Syndrome Coronavirus (MERS-CoV)-related pandemic outbreak started in September 2012 in the Middle East, and it had subsided by June 2013. Around 2,500 laboratory confirmed cases were reported, and the fatality rate was around 34.4%. MERS-CoV appears to have originated from an animal source in the Arabian Peninsula. Researchers have found MERS-CoV in camels in several countries. Evidence suggests that MERS-CoV could have been also originated in bats that infected camels which in turn passed the infection on to humans.

In December 2019, a novel coronavirus, SARS-CoV2, today known as COVID-19, was identified as responsible for an outbreak of viral pneumonia focused on Wuhan, Hubei, China [2]. COVID-19 virus shares structural and sequential features with SARS-CoV and MERS-CoV viruses, causing severe respiratory complications in their hosts [3,4]. It is likewise originated in horseshoe bats found in several caves in China [5,6], which infected intermediate hosts, like cats, raccoons or dogs, transferring the infection to humans. As of today (July 27, 2020), there are over 16 million COVID-19 infection cases with approximately 650,000 worldwide. However, in some countries such as Italy or Spain, the death rate is significantly higher [7]. COVID-19 symptoms are characterized by high fever, cough and a general malaise [8]. However, severe cases experience acute respiratory distress syndrome, and lung injury leading to lung inflammation and pneumonia [9]. SARS-CoV-2 forms spherical or elliptic particles of a diameter of approximately 50–200 nm containing single-stranded RNA associated with a nucleoprotein within a capsid. There are four main structural proteins of coronaviruses: spike (S), membrane (M), envelope (E), and nucleocapsid (N) glycoproteins. SARS-CoV-2 has around 79% similarity to SARS-CoV and about 50% similarity to MERS-CoV [10].

The structural arrangements of nucleocapsid protein (N), envelope protein (E), and membrane protein (M) among SARS-CoV-2, SARS-CoV and MERS-CoV beta-coronaviruses are different. The patient's physiological response starts at the cellular level following virus replication. The spike glycoprotein (S) of the coronavirus ease the entrance into the virus' targeted cells, which depends on the binding of the virus protein to a cellular receptor [11]. Li and co-workers reported that SARS-S protein binds to angiotensin-converting enzyme 2 (ACE2) as the entry receptor and utilizes cellular serine protease (TMPRSS2) for S protein priming [12]. To detect pathogens such as bacteria and viruses, the immune system is equipped with receptors called pattern recognition receptors (PRRs). These receptors are a key element of the immune system. Subsequent to the virus cell infection, replication is performed through a Pattern Recognition Receptor (PRRs) [13]. Binding of virus specific RNA structures yields to oligomerization of these receptors and the activation of down-stream transcription factors, where interferon regulator factors (IRFs) and nuclear factors κB (NF-κB) are the most likely ones [14]. These two factors generate two antiviral responses in the host. The first antiviral response is mediated by transcriptional induction of type I and II interferons (IFN–I and IFN-II) and overexpression of IFN-stimulated genes (ISGs) [15]. The second response is carried out by leukocytes and mainly induced by chemokine secretion [16]. Furthermore, the response of the host varies depending on the underlying genetics, showing different degrees of morbidity, including mortality [17].

The current COVID-19 outbreak is acute and rapidly growing into a global crisis. Consequently, it is of utmost importance to analyze the transcriptomics data of the viral infection [18] to precisely detect the genetic pathways severely altered by the virus and to perform drug repurposing aimed at restoring homeostasis. The analysis of disease's transcriptomic response, plays a key role in the development of novel and effective therapeutic strategies [19]. Comparing the transcriptomic response on SARS-CoV-2 to other respiratory diseases, such as influenza A virus (IAV), and SARS-CoV-1, we can obtain not only a better insight on potential treatments, but also a way to repurpose existing FDA-approved drugs, as a complementary approach to the design of new drugs and novel vaccines. In terms of transcriptomics, Frieman et al. reported that these respiratory diseases involve a variety of different antagonists to IFN-I and IFN-III response [20]. More precisely, Frieman et al. and Kopecky-Bromberg et al. attributed IFN antagonism to the nucleocapsid (N) gene products, ORF3B and ORF6 [20,21]. García--Sastre et al. reported that, either IAV and SARS-CoV-1, encode the IFN-I and IFN-III antagonist non-structural protein 1 (NS1), which blocks the human body initial detection by PRRs via binding and masking RNA produced during the infection [22].

Clinicians believe that the main cause in many critically ill patients is related to the cytokine storm that take place when the immune system reacts against viral infection. These signaling proteins secreted by the immune system, act as chemical messengers against the infection, helping to regulate the immune system response by binding to the surface of the infected cells. Sepsis occurs when the immune system overreacts. Other clinicians emphasize the hyperinflammatory state generated by the infection and argue that targeting the cytokines will speed up the viral replication. New studies on the side-effects triggered by the COVID infection, show that infection may also lead to blood vessel constriction and the main risk factors seem to be associated with diabetes, obesity, and hypertension. Other organs such as the brain, the eyes, the nose, the heart and the blood vessels, the liver, the kidneys, and the intestines are affected as well. Heart damage, renal and kidney failures, neurological and gastrointestinal problems have also been reported in COVID-19 patients [23]. All these medical data indicate that varying genetic pathways are being altered by the coronavirus and that identifying these alterations is the crucial step in understanding the dynamics of the infection and finding possible treatments via drug repurposing. Varying approaches have been applied to fight COVID-19 infection, ranging from the use of natural products and human intestinal defensins, to sophisticated methods combining drug repositioning, virtual screening and molecular dynamics simulations [24–27].

In this manuscript, we aim to understand the mechanisms of the disease and repurpose existing FDA-approved drugs to treat COVID-19 by performing a retrospective analysis of SARS-CoV-2 transcriptomic data, that were made publicly available (GEO 147507) by the Icahn School of Medicine at Mount Sinai Hospital (New York). This analysis was carried out by sampling different equivalent high discriminatory genetic networks, that are related to the uncertainty space of the classifier that is used to predict the phenotype (disease versus controls). For that purpose, we have used the Holdout (HDS) and the Leave-One-Out-LOOCV Samplers, described by us in detail in earlier works [28]. The subsequent analysis entails finding the most frequently sampled genetic signatures with high validation accuracy in the phenotype prediction. We apply this knowledge of altered genetic pathways to conduct drug repositioning by using the connectivity map (CMAP) library, developed by the Broad Institute which contains over 1.5 million gene expression profiles, around 5000 small molecule compounds, and nearby 3000 genetic reagents [29]. The main philosophy of using CMAP library is to find drugs that are potentially efficient to reestablish the homeostasis [30]. From this perspective, given a set of altered genes, drugs in CMAP library are score-based on a disease-drug connectivity and rank-based on the two-sample Kolmogorov-Smirnov test to determine which ones are more likely to have either a positive effect or an amplified negative effect [31]. The algorithm samples the equivalence region of a regression problem by using bootstrapping to find different sets of equivalent predicting parameters [32,33]. Finally, these findings have been cross validated with the results obtained by analyzing independent datasets.

In conclusion the use of transcriptomic data enables a better

understanding of the disease and offers the possibility of fast-tracking antiviral therapies for COVID-19 patients via drug repositioning.

## 2. Material and methods

### 2.1. Material

To achieve drug repositioning, we perform a retrospective analysis of two transcriptomic datasets. The first one was deposited in the Gene Expression Omnibus Database https://www.ncbi.nlm.nih.gov/geo/under the accession number GEO147507.

It contains transcriptomic data from cells and animal models of SARS-CoV-2, Respiratory Syncytial Virus (RSV) and Influenza A Virus (IAV). These data has been recently studied by Blanco-Melo et al. [18], whose research revealed an unsuitable inflammatory response by low levels of type I and type III interferons juxtaposed to elevated chemokines and high expression levels of IL-6. In their studies, they used cell models of SARS-CoV-2 infection, in addition to transcriptional and serum profiling of COVID-19 patients. This dataset was the main target of this study since authors identified a muted transcriptional response to SARS-CoV-2, which supports a model in which initial failure to rapidly respond to infection results in prolonged viral replication and an influx of proinflammatory cells that induce alveolar damage and manifest in COVID-19 lung pathology.

We utilized transcriptomic data derived from independent triplicates of primary normal human bronchial epithelial (NHBE) cells, that were mock-treated (control set) or infected with SARS-CoV-2, IAV H1N1, or IAV dNS1 strains, and treated with human interferon-beta (cell-lines). Additional data were derived from independent biological triplicates of adenocarcinoma human alveolar basal epithelial cells (A549) that were mock-treated or infected with SARS-CoV-2, RSV or IAV H1N1. We have used the expression data from transformed lung alveolar (A549) transduced with a vector expressing human ACE2 and transformed lung-derived Calu-3 cells, both infected and controlled. In this paper we only perform the SARS-CoV-2 analysis (infected vs. Mock Treatment cases). The main analysis to identify drug repurposing uses only SARS-CoV-2 infected cell lines, although the additional infection datasets were used for some comparative studies. This dataset also includes 4 samples from mock-treated lung biopsies (2) and COVID-19 infected lungs (2): uninfected human lung biopsies were derived from one male (age 72) and one female (age 60) and lung samples derived from a single male COVID-19 deceased patient (age 74). These samples were used to check the findings performed with the cell-lines samples.

Additionally, we have interpreted the dataset provided by Duke University (https://www.ebi.ac.uk/arrayexpress/experiments/E-MTAB-8871/), a transcriptomic profiling conducted on blood collected from COVID-19 cases by Desai et al., [34]. The blood was collected in Tempus Blood RNA tubes, and RNA was extracted using the Tempus Spin RNA Isolation Kit. Mock-treated samples were taken from participants of a measles, mumps, and rubella re-vaccination study. In this respect, the researchers used the NanoString Human Immunology Panel, that includes the gene expression analysis of over 437 immunology genes. Panel includes major classes of cytokines and their receptors, enzymes with specific gene families such as the major chemokine ligands and receptors, interferons and their receptors, the TNF-receptor superfamily, and the KIR family genes. This panel is ideal for the study of allergy, autoimmune diseases, and infectious disease immune response. This dataset shared most of the targeted genes with the first study and allowed us to confirm the immune pathway analysis found in the first experiment and perform drug repositioning targeting the chemokines storm and the immune response.

### 2.2. Methods

To perform the retrospective cohort study, we applied our novel machine learning uncertainty-based method to sample (within the entire dataset) different combinations of highly predictive genes for a specific phenotype prediction problem. This enables us to identify the discriminatory genetic pathways for SARS-CoV-2 infected vs. Mock Treatment cases. The main idea is to identify the genes that are up- and down-regulated upon infection to establish personalized treatments.

Prediction of the transcriptomic response can be framed as a general regression model between the high discriminatory genes that characterize the infection and a set of sample cases that are part of the training data set [35]. The most important issue in the analysis of genetic data is the lack of a conceptual model that can relate the different genes/probes to the class prediction (phenotype). As a result, a classifier $\mathbf{L}^*(\mathbf{g})$ must be built as an algorithm that maps genetic signatures $\mathbf{g}$ to the set of classes into which the phenotype is divided, in our case:

$$\mathbf{L}^*(\mathbf{g}) : \mathbf{g} \in \mathbb{R}^s \rightarrow \mathrm{C} = \{\mathrm{SARS-CoV-2; Mock}\}. \quad (1)$$

To properly map the genetic signatures and the phenotype, the modeling scheme is divided in two steps: learning and validation. The learning stage is carried out with a subset of samples T, whose class vector $\boldsymbol{c}^{obs}$ is known and consists of finding the genetic signature that maximizes accuracy according to:

$$Acc(\mathbf{g}) = ||100 - \mathbf{L}^*(\mathbf{g}) - \mathbf{c}^{\mathrm{obs}}||_1. \quad (2)$$

$||100 - \mathbf{L}^*(\mathbf{g}) - \mathbf{c}^{\mathrm{obs}}||_1$ stands for the prediction error (in percents) which is determined via Leave-One-Out-Cross-Validation (LOOCV). The shortest genetic signature with the highest predictive accuracy is considered the small-scale signature and offers an idea of the complexity of the phenotype prediction. Due to the inherent uncertainty of this type of problems [36,37], the stability of the small-scale signature can be established by performing data bagging [38]. One of the novel techniques recently published in this regard was introduced by Li et al. and Yang et al. The first one proposes a methodology that identifies the proteomic signature (in this case it can also be applied to a genetic signature) of good reproducibility and aggregating them to ensemble feature ranking by ensemble learning, assessing the generalizability of ensemble feature ranking to acquire the optimal signature and indicating the phenotype association of discovered signature [39]. The second one introduce a novel feature selection strategy integrating repeated random sampling with consensus scoring and evaluating the consistency of gene rank among different datasets was constructed [40]. In our case; and similarly to Li et al. and Yang et al., we utilized a holdout sampler [41,42] in combination to the Fisher's ratio in order to filter the highest discriminatory genes according to this parameter [43].

Fig. 1 shows the algorithm workflow, that consists of three steps:

1. **Data bagging:** a set of random 75/25 data bags holdouts were generated, where 75% of the data is used for learning and 25% for validation. In the present study, we have used 1000 different bags.
2. **Gene selection:** For each data bag, the discriminatory genes are selected, and the classifier is built. The genes utilized to construct the classifier are restricted to the most discriminatory ones according to their Fisher's ratio, with a minimum cutoff of 0.5. The accuracy is calculated on the validation set. This way of proceeding is because variables with high discriminatory power span the main features of the classification, while variables with lowest discriminatory ratios account for the details in the discrimination. This method determines the minimum amount of high-frequency details (helper genes) that are needed to optimally discriminate between classes promoting the header genes, which are those that explain the phenotype in a robust way [35,36,44,45].

For clarification purposes, Fisher's ratio and the fold change of gene $j$ in a binary classification problem are defined as follows:

$$fc(\mathrm{g}_j) = log\frac{\mu_{1j}}{\mu_{2j}},$$

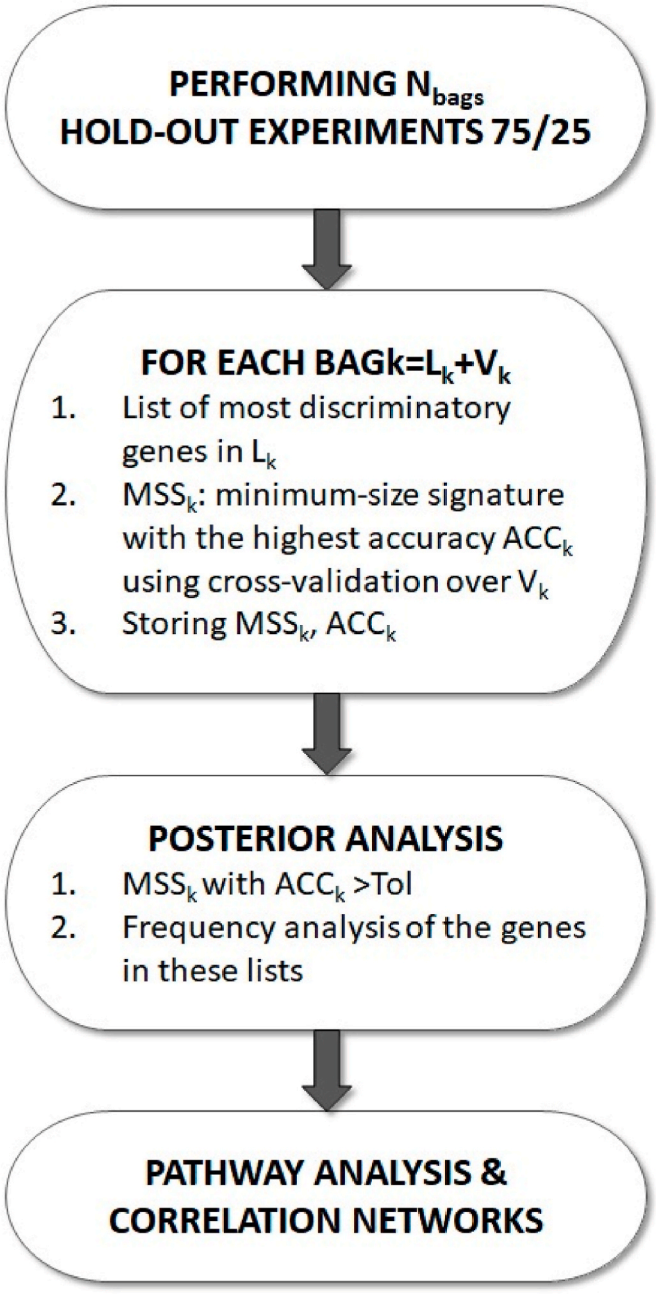


**Fig. 1.** Flowchart of the Holdout Sampler utilized to sample the transcriptomic data.

$$Fr(g_j) = \frac{(\mu_{1j} - \mu_{2j})^2}{\sigma_{1j}^2 + \sigma_{2j}^2},$$

where $\mu_{1j}, \mu_{2j}$ are the mean expressions of the gene j in classes 1 and 2 and $\sigma_{1j}^2$, $\sigma_{2j}^2$ are their respective variances. By combining both, it is possible to look for genes that are differentially expressed, and whose respective gene expression probability distribution tails do not overlap. In our analysis, we use the p-value of the T-test to decide if both gene expressions come from a distribution with the same mean or not. A small p-value indicates strong evidence for the exclusion of the null-hypothesis, that both distributions have the same mean.

**Posterior analysis:** After the data bagging simulation, the small-scale signature is found by selecting holdouts with the highest predictive accuracies. Finally, we performed a frequency analysis of the sampled genes and linked them to the potential defective pathways.

Once the different genes are ranked according to their differential gene expression with respect to the control samples, it is possible to perform CMAP-based drug repurposing with Dr. Insight package [31]. When searching for a potential drug, a gene is considered *concordantly expressed* if there is an inverse association of its expression in the disease data and in the drug-perturbed data [30,46]. This means that a gene is upregulated in the disease data but downregulated after drug perturbation, or vice-versa. Dr. Insight algorithm computes for each gene its expression rank in the disease and reference drug data according to a formulated outlier-sum based statistics that models the overall disease-drug connectivity. The algorithm computes the outlier-sum for each drug and compares it with the reference distribution outlier-sum statistics obtained from the entire CMAP database. It employs the two sample Kolmogorov-Smirnov test to determine if a specific outlier-sum for a given drug indicates a novel repurposed drug candidate [47].

## 3. Results

### 3.1. Analysis of the first data set: transcriptomic analysis in cell lines and lung biopsies

To understand the transcriptional response of SARS-CoV-2 and infer potential drug targets via drug repositioning of already FDA-approved drugs, we perform a comparative study of the altered genes and pathways in SARS-CoV-2 with respect to other respiratory diseases, including SARS-CoV-1, Respiratory Syncytial Virus (RSV) and Influenza A Virus (IAV). Fig. 2 shows the dataset, which is the transcriptomic profile of SARS-CoV2 with respect to other respiratory diseases. The colour intensity represents the expression of the genes. No visual difference can be made. Therefore, supervised methods are needed to find the genes that separate COVID infected samples versus mock treated. Although this is obvious for many modelers, some people still think that unsupervised methods could be used.

Fig. 3 shows that IAV and RSV comprises a unique cluster in the

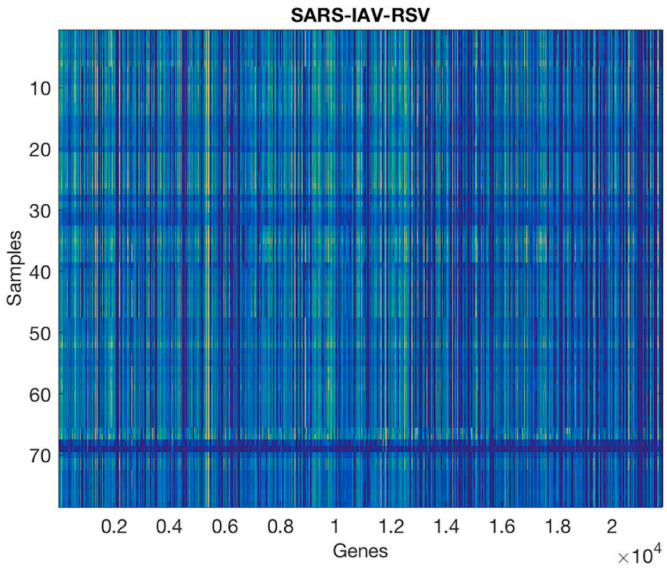


**Fig. 2.** First data set. Transcriptomic profile of SARS-CoV-2 with respect to other respiratory diseases. The x-axis represents the index of the genetic probe, and the y-axis the index of the samples. It is impossible to differentiate the SARS COVID samples from the rest. Our methodology includes cross-validations of the discriminatory genetic signatures that were sampled. In this case the image contains all the samples in this data set, not only the COVID samples. Blue color indicates lower expressions, while the red colors indicate higher expressions. It can be observed that blue colors (under expressions) are predominant in this dataset.

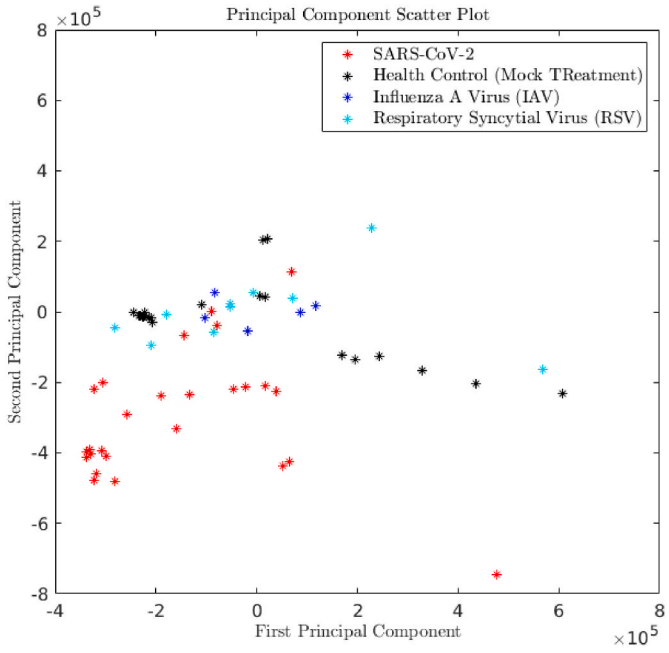


**Fig. 3.** Principal Component Representation of SARS-CoV-2 infection with respect to other respiratory diseases. The projection is done on the two first PCA terms obtained from the gene expressions of the Small-Scale Signature given in Table 1.

Principal Component Representation, which is due to a high expression of IFNs and ISGs, as suggested by Blanco-Melo and coworkers. In contrast, the infection of SARS-CoV-2 differs in this representation and based on the transcriptomic profile, it shows little expression with regards to IFN-I and IFN-III, but shows an expression in the ISGs and proinflammatory pathways (as seen later in Table 4 concerning the altered genetic pathways).

### *3.2. Genetic small-scale signature of SARS-CoV-2 shows high expression of proinflammatory pathways and infer possible secondary effects and complications*

This comparison establishes the differences between transcriptomic data for SARS treated (X samples) and mock treatment (Y samples). The small-scale signature to predict this phenotype is composed of 31 genes. We have achieved a LOOCV accuracy of 91.84% (92%). In this case, under expression indicates that mean genetic expression of these genes in SARS-CoV-2 samples, is smaller than in the mock treatment samples. Figs. 4 and 5 shows two discriminatory plots of the SARS-CoV-2 phenotype classification problem. The first one shows the fold change Fisher's ratio plot, where we observe that the genes with the highest Fisher's ratio do not coincide with the one with the highest fold change. This is because there is an overlapping between gene expressions in both classes. Fig. 5 shows the regression between the Fisher's ratio and the -$\log_{10}$ of the p-value of the T-test performed on the median expressions of the most discriminatory genes in each class (SARS and Mock-treated). It can be observed that the Fisher's ratio increases as the p-value decreases (-$\log_{10}$ increases). Therefore, the Fisher's ratio is a correct descriptor of the discriminatory power of the genes in the phenotype classification problem.

Tables 1 and 2 show the genes of the small-scale signature (over- and under expressed). We also provide the mean expression in class 1C1 (COVID) and class 2C2 (Mock-treated), the dispersions (std), the Fisher's ratio and the LOOCV accuracy of the incremented lists ranked by decreasing discriminatory power (Fisher's ratio). In this case, the minimum Fisher's ratio is $fr_{min} = 0.76$. This cut-off value makes the center of the distributions of the genes expressions in both classes to be separated:

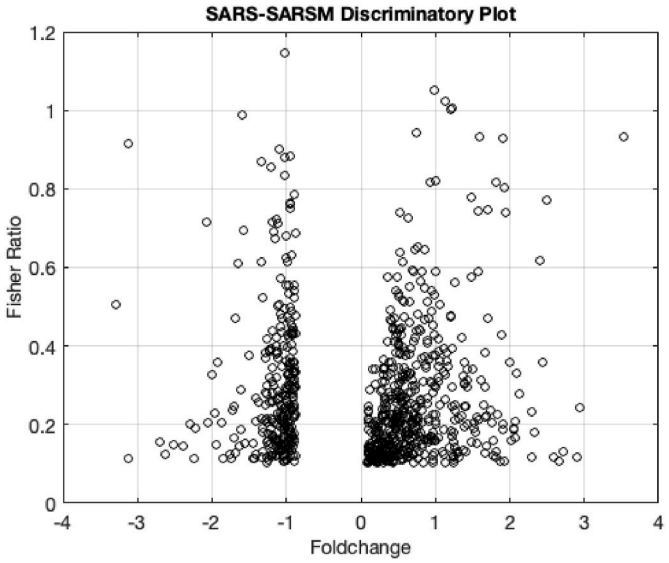


**Fig. 4.** Fisher's ratio-foldchange plot of the most discriminatory genes in the SARS discrimination.

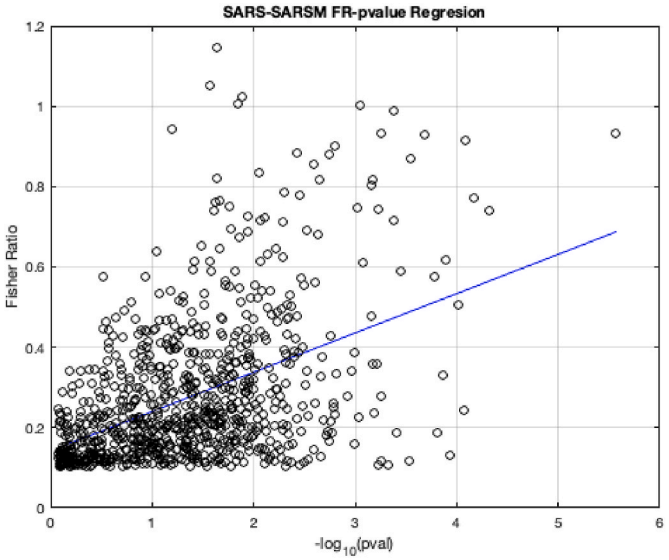


**Fig. 5.** Fisher's ratio-pvalue plot (in -log10 scale) of the most discriminatory genes in the SARS discrimination.

$$|\mu_1 - \mu_2| > \sqrt{fr_{min}}(std_1 + std_2).$$

That way, the binary classification problem becomes a linear separation problem. This was explained in Ref. [48] where the algorithm was formerly described.

It is interesting to analyze the most important pathways, diseases and drugs related to the small-scale genetic signature concerning both types of genes (under and overexpressed).

1. Under expressed genes:
   - o **Pathways:** Passive transport by aquaporins (AQP6), Platelet adhesion to exposed collagen (COL1A1), Osteoblast signaling (COL1A1), TWEAK regulation of gene expression (COL1A1), Association of TriC/CCT with target proteins during biosynthesis (FBXW9), Inflammatory response pathway (COL1A1), Glycoprotein VI-mediated activation cascade (COL1A1), Beta-3 integrin cell

**Table 1**
Small Scale Genetic Signature between SARS-CoV-2 and Health Control. This is the smallest list of genes that could expand the differences between infected and healthy patient (31 genes) with the highest discriminatory accuracy. The upper part of the table lists these 31 genes divided into underexpressed and overexpressed genes.

| Underexpressed Genes | Overexpressed Genes |
|---|---|
| *KCNU1* | *MTRNR2L1* |
| *ENO1-AS1* | *PGA5* |
| *SPOCK3* | *ATP4A* |
| *AQP6* | *DRD1* |
| *RNF126* | *SLC5A5* |
| *TMEM89* | *FKBP6* |
| *PLEKHF1* | *DNAH17* |
| *BOP1* | *KCNN1* |
| *MRPL41* | *C16orf96TNFRSF9* |
| *FBXW9* | *ZNF280A/MTRNR2L2* |
| *FSCN1* | *KCNT2/NKAIN1* |
| *COL1A1* | *HIST1H3J/ZNF627* |
| *CYB561D2* | *ANKRD33B/THAP1* |

**Table 2**
Mean expression values in each class (C1 is the SARS COVID-19 class and C2 are Mock-treated), and their mean deviations, the Fisher's ratios, and the cummulative LOOCV accuracies for each of the 31 genes given in Table 1.

| Gene Name | MeanC1 | StdC1 | MeanC2 | StdC2 | FR | LOOCV Acc |
|---|---|---|---|---|---|---|
| TMEM89 | 0.4 | 0.66 | 0.8 | 0.56 | 1.15 | 73.47 |
| HIST1H3J | 3.8 | 3.37 | 1.9 | 2.43 | 1.05 | 71.43 |
| KCNT2 | 23.5 | 19.10 | 10.7 | 15.58 | 1.02 | 71.43 |
| ZNF280A | 44.4 | 42.65 | 19.0 | 26.58 | 1.00 | 73.47 |
| MTRNR2L2 | 1.1 | 0.71 | 0.5 | 0.48 | 1.00 | 71.43 |
| ENO1-AS1 | 0.8 | 1.20 | 2.5 | 1.79 | 0.99 | 77.55 |
| ANKRD33B | 485.2 | 323.33 | 288.5 | 392.19 | 0.94 | 79.59 |
| MTRNR2L1 | 0.8 | 0.65 | 0.1 | 0.08 | 0.93 | 77.55 |
| KCNN1 | 15.6 | 11.94 | 5.2 | 7.47 | 0.93 | 77.55 |
| SLC5A5 | 8.2 | 7.06 | 2.2 | 2.77 | 0.93 | 79.59 |
| KCNU1 | 0.1 | 0.12 | 0.8 | 0.75 | 0.91 | 81.63 |
| RNF126 | 249.4 | 222.48 | 536.2 | 352.68 | 0.90 | 81.63 |
| FBXW9 | 69.4 | 75.03 | 134.2 | 73.55 | 0.88 | 81.63 |
| PLEKHF1 | 51.0 | 48.16 | 103.3 | 60.71 | 0.88 | 79.59 |
| SPOCK3 | 2.9 | 2.48 | 7.4 | 4.90 | 0.87 | 83.67 |
| AQP6 | 1.4 | 1.68 | 3.1 | 2.15 | 0.86 | 87.76 |
| BOP1 | 163.7 | 189.90 | 332.6 | 236.45 | 0.83 | 81.63 |
| NKAIN1 | 17.1 | 15.37 | 8.5 | 9.92 | 0.82 | 81.63 |
| FKBP6 | 0.7 | 0.60 | 0.2 | 0.26 | 0.82 | 85.71 |
| ZNF627 | 190.9 | 121.25 | 100.4 | 71.49 | 0.82 | 83.67 |
| DRD1 | 154.0 | 148.64 | 40.4 | 54.61 | 0.80 | 83.67 |
| CYB561D2 | 105.7 | 108.24 | 196.8 | 107.98 | 0.79 | 85.71 |
| TNFRSF9 | 197.3 | 171.47 | 70.7 | 114.07 | 0.78 | 87.76 |
| PGA5 | 0.2 | 0.21 | 0.0 | 0.06 | 0.77 | 83.67 |
| FSCN1 | 1794.1 | 2236.36 | 3454.2 | 2596.43 | 0.76 | 83.67 |
| COL1A1 | 246.6 | 292.49 | 477.2 | 385.09 | 0.76 | 87.76 |
| MRPL41 | 371.7 | 508.81 | 722.0 | 484.73 | 0.75 | 87.76 |
| DNAH17 | 133.6 | 112.37 | 41.1 | 68.15 | 0.75 | 87.76 |
| C16orf96 | 1.0 | 0.79 | 0.3 | 0.43 | 0.74 | 87.76 |
| THAP1 | 224.4 | 128.00 | 155.3 | 75.99 | 0.74 | 89.80 |
| ATP4A | 2.5 | 1.92 | 0.6 | 0.75 | 0.74 | 91.84 |

surface interactions (COL1A1), NCAM1 interactions and Syndecan 1 pathway (COL1A1).

- **Diseases:** Immune response to smallpox (secreted IL-10) (SPOCK3), Insulin resistance/response (KCNU1), Response to taxane treatment (docetaxel) (COL1A1), Lipid traits (pleiotropy) (HIPO component 1) (KCNU1), Iron status biomarkers (transferrin saturation) (KCNU1), Mild influenza (H1N1) infection (KCNU1).
- **Drugs:** dosulepin (MRPL41; COL1A1; PLEKHF1), mexiletine (MRPL41; COL1A1; PLEKHF1), bromopride (MRPL41; COL1A1; SPOCK3), fenoprofen (COL1A1; SPOCK3; FSCN1), tetrahydroalstonine (COL1A1; PLEKHF1; SPOCK3), cytisine (MRPL41; COL1A1; FSCN1).

The main pathway concerns the passive transport by aquaporins related to AQP6, that are membrane water channels proteins, which facilitates the transport of water between cells. Aquaporins are important in several tissues [49]: nervous (retina/olfactory epithelium/inner ear/brain and spinal cord), renal, cardiovascular, respiratory, reproductive, digestive and musculoskeletal (muscle and cell swelling), skin (dermis) and fat (glycerol transport). Interestingly, COL1A1 is related to the inflammatory response pathways, TWEAK proinflammatory cytokines via NF-κB and MAPK and also in glycoprotein VI-mediated activation cascade that plays a key role in some animal model of thrombosis. Moreover, COL1A1 seems to have increased expression in patients with severe COVID-19 [50–52]. Therefore, the deregulation of this gene seems to have a major role in COVID-19 infection. In the set of the related diseases, appear the immune response to smallpox and influenza infection related to the KCNU1 gene. The main drugs are antidepressant (dosulepin), dopamine antagonists (bromopride), nonsteroidal anti-inflammatory drug (frenopren), voltage-gated sodium channel blockers (mexiletine) and acetylcholine agonist (cytisine) with strong binding affinity for the nicotinic acetylcholine receptor [49].

## 4. Overexpressed genes

- **Pathways:** dopamine receptors, thyroxine biosynthesis, calcium-activated potassium channels, organic anion transporters, amine-derived hormones, 4-1BB-dependent immune response, Ck1/Cdk5 regulation by type 1 glutamate receptors, collecting duct acid secretion, hypothetical network for drug addiction, ion transport by P-type ATPases.
- **Diseases:** idiopathic dilated cardiomyopathy (DNAH17/KCNT2), dermatomyositis (ANKRD33B), cardiac effects (KCNT2), treatment response for severe sepsis (DRD1), immune response to smallpox (secreted IL-12p40) (FKBP6), age-related macular degeneration (KCNT2), asthma exacerbations in inhaled corticosteroid treatment (KCNN1).
- **Drugs:** citiolone (NKAIN1, DNAH17, KCNN1), rofecoxib (NKAIN1, DNAH17, FKBP6), estradiol (DNAH17, KCNN1, SLC5A5), ethisterone (KCNN1, SLC5A5).

The overexpressed genes seem to mainly control dopamine and glutamate receptors; thyroxine biosynthesis and amine-derived hormones; calcium activated potassium channels involved in age-related neurodegeneration (memory impairment and neurovascular disfunction); and 4-1BB-dependent immune response, which are inducible T-cell surface receptors belonging to the tumor necrosis factor receptor superfamily (48). The 4-1BB-4-1BBL pathway has been considered therapeutic in the treatment of HIV infection. Interestingly, 4-1BB-deficient mice show dysregulated immune responses and very high Ig responses to T-dependent antigens (49).

The main diseases related to the overexpressed genes in COVID-19 are cardiomyopathies (DNAH17/KCNT2) and cardiac metabolic effects (KCNT2), severe sepsis (DRD1), immune response to smallpox provoked by secreted IL-12p40 (FKBP6), and macular degeneration (KCNN1). FK506 binding protein 6 belongs to the FKBP immunophilins. These proteins act as receptors for immunosuppressive drugs such as rapamycin, cyclosporin and tacrolimus. Cyclosporin A has been proposed as a first-line therapy in COVID-19 pneumonia for different reasons (50): 1. Mitochondrial function is essential for the antiviral defense to produce ATP for the increased energetic needs of the infected cells. Mitochondrial failure has been pointed as one of the mechanisms unchaining severe forms of COVID-19 infection. Cyclosporin A seems to protect against this effect and has shown cardioprotective effects in patients with myocardial infarction; 2. It has shown remarkable antiviral activities in a variety of RNA viruses. The role of these genes in COVID should be further investigated.

The main drugs found acting on the overexpressed genes are used for liver therapy (Citiolone). COX-2 selective nonsteroidal anti-

inflammatory drug (Rofecoxib) is used to treat rheumatoid arthritis, estrogen steroid hormones (estradiol) and agonist of the progesterone receptor (Ethisterone). It has been suggested that women were less affected by COVID-19 than males.

Interestingly, when all the genes from the small scale signature are applied, the main drugs repositioned using CMAP were: sirolimus (COL1A1; PLEKHF1; KCNN1; DRD1); ivermectin (COL1A1;PLEKHF1; KCNN1;DRD1) and cytisine (MRPL41; COL1A1; HIST1H3J; FSCN1) (50) (51). Sirolimus and ivermectin have been used against COVID. Ivermectin inhibits the replication of SARS-CoV-2 in vitro. Ivermectin has been granted FDA-approval for parasitic infections. An unexpectedly low prevalence of active smokers has been observed among hospitalized patients with COVID-19. Cytisine is an alkaloid that occurs naturally in several plant genera. It has been used medically for smoking cessation since its molecular structure has some similarity to nicotine and it has similar pharmacological effects. We believe that these findings confirm the validity of the small-scale genetic signature approach.

Although we have also performed the analysis of the overexpressed and under expressed genes separately to provide insights about the pathways that are involved, all the genes of the small-scale signature work synergistically and they should be considered together to have a correct view of the ensemble. Furthermore, when all the genes are used, the analysis served to identify 3 main drugs, 2 of which are currently in trial for COVID. This fact suggests that the gene signatures, rather than individual genes, will identify disease-relevant drug findings.

### 4.1. Expression of the small-scale genetic signature in the lung biopsies

To confirm the previous findings, we have calculated the median expressions of the small-scale signature composed of 31 genes given in Table 1 in the 4 samples from healthy and infected lung biopsies. Table 3 shows the genes with the most representative mean expression differences of the genes belonging to the small-scale signature. It is very important to highlight that all these genes are downregulated in SARS-COVID biopsies, being the most significant ones: MRPLR1, COL1A1, FSCN1, RNF126 and ANKRD33B. The main pathways related to these genes are dopamine receptors, inflammatory response, and immune response. The main CMAP drugs downregulating those genes are reserpine, flurbiprofen and debrisoquine, which treat hypertension, and are nonsteroidal anti-inflammatory drugs.

An exploitative and explorative sampling of the altered genetic signature space offers a wide range of genetic pathways related to causes and plausible complications after infection/recovery.

Considering the LOOCV sampler, which is capable of sampling the uncertainty space of the phenotype prediction problem with a highly *exploitative* character, given the small-scale genetic signature, it acts as a classifier that is capable of converging rapidly to a minimum error solution. In this sense, we obtain a set of genes that are altered between the SARS-CoV-2 and the mock treatment. Similarly, to perform this analysis with the LOOCV sampler, a Holdout sampler could be used to enact the sampling of the uncertainty space of the phenotype prediction problem with a highly *explorative* character, given the small-scale genetic signature.

**Table 3**
Most representative expression differences of the genes of the Small-Scale Genetic Signature in lung biopsies (HC and SARS-COVID Infected).

| Gene-name | Mean expression in HC | Mean expression in SARS-COVID |
|---|---|---|
| MRPL41 | 1311 | 10 |
| COL1A1 | 1312 | 84 |
| FSCN1 | 745 | 20 |
| RNF126 | 729 | 12 |
| ANKRD33B | 667 | 0 |
| CYB561D2 | 232 | 0 |
| DNAH17 | 227 | 0 |
| THAP1 | 222 | 2.7 |
| TNFRSF9 | 220 | 1.9 |
| PLEKHF1 | 198 | 0 |
| ZNF627 | 198 | 0 |
| FBXW9 | 142 | 0 |
| BOP1 | 58 | 0 |
| DRD1 | 58 | 0 |
| KCNT2 | 55 | 2 |

**Table 4**
List of the most frequently sampled genes by the LOOCV sampler. C1 is the SARS COVID samples and C2 are mock treated samples (Healthy controls). We provide the mean signatures in each class, the fold change, the fisher ratio, the LOOCV accuracy of the different incremented list of genes and the sampling frequency that it is 0.58 for all of them.

| Gene Name | MeanC1 | MeanC2 | FC | FR | LOOCV Acc | Frequency |
|---|---|---|---|---|---|---|
| ENO1-AS1 | 0.84 | 2.52 | −1.59 | 0.99 | 73.47 | 0.58 |
| NKAIN1 | 17.11 | 8.54 | 1.00 | 0.82 | 73.47 | 0.58 |
| EFNA1 | 2260.14 | 805.52 | 1.49 | 0.57 | 77.55 | 0.58 |
| SNORD75 | 0.16 | 0.66 | −2.07 | 0.72 | 83.67 | 0.58 |
| SNORD74 | 3.43 | 6.98 | −1.02 | 0.49 | 89.80 | 0.58 |
| KCNT2 | 23.45 | 10.66 | 1.14 | 1.02 | 87.76 | 0.58 |
| EPC1 | 550.69 | 394.32 | 0.48 | 0.58 | 83.67 | 0.58 |
| WEE1 | 643.72 | 446.80 | 0.53 | 0.64 | 83.67 | 0.58 |
| AQP6 | 1.36 | 3.13 | −1.21 | 0.86 | 83.67 | 0.58 |
| ATP6V0D1 | 629.94 | 1256.87 | −1.00 | 0.68 | 85.71 | 0.58 |
| MTRNR2L1 | 0.75 | 0.06 | 3.54 | 0.93 | 85.71 | 0.58 |
| DNAH17 | 133.60 | 41.11 | 1.70 | 0.75 | 85.71 | 0.58 |
| RNF126 | 249.37 | 536.15 | −1.10 | 0.90 | 83.67 | 0.58 |
| NDUFS7 | 232.20 | 503.85 | −1.12 | 0.71 | 83.67 | 0.58 |
| ADAT3 | 16.44 | 37.26 | −1.18 | 0.72 | 85.71 | 0.58 |
| PPAN | 12.08 | 27.10 | −1.17 | 0.69 | 83.67 | 0.58 |
| ZNF627 | 190.88 | 100.39 | 0.93 | 0.82 | 83.67 | 0.58 |
| SLC5A5 | 8.23 | 2.20 | 1.90 | 0.93 | 85.71 | 0.58 |
| KCNN1 | 15.64 | 5.21 | 1.59 | 0.93 | 81.63 | 0.58 |
| GDF15 | 1369.24 | 928.66 | 0.56 | 0.61 | 83.67 | 0.58 |
| ZNF14 | 89.36 | 50.73 | 0.82 | 0.59 | 83.67 | 0.58 |
| PLEKHF1 | 51.00 | 103.33 | −1.02 | 0.88 | 83.67 | 0.58 |
| ATP4A | 2.46 | 0.63 | 1.95 | 0.74 | 83.67 | 0.58 |
| PLEKHG2 | 1353.67 | 826.16 | 0.71 | 0.59 | 83.67 | 0.58 |
| ZNF845 | 200.73 | 129.85 | 0.63 | 0.73 | 83.67 | 0.58 |
| USP37 | 385.81 | 227.38 | 0.76 | 0.65 | 83.67 | 0.58 |
| ZNF280A | 44.42 | 18.96 | 1.23 | 1.00 | 83.67 | 0.58 |
| CABP7 | 4.00 | 2.07 | 0.95 | 0.54 | 83.67 | 0.58 |
| TMEM89 | 0.40 | 0.80 | −1.02 | 1.15 | 83.67 | 0.58 |
| SPOCK3 | 2.93 | 7.37 | −1.33 | 0.87 | 83.67 | 0.58 |
| ANKRD33B | 485.20 | 288.52 | 0.75 | 0.94 | 83.67 | 0.58 |
| MTRNR2L2 | 1.06 | 0.46 | 1.21 | 1.00 | 81.63 | 0.58 |
| NME5 | 6.18 | 18.44 | −1.58 | 0.69 | 83.67 | 0.58 |
| DRD1 | 154.00 | 40.44 | 1.93 | 0.80 | 85.71 | 0.58 |
| HIST1H3J | 3.82 | 1.92 | 0.99 | 1.05 | 87.76 | 0.58 |
| ZNF92 | 202.38 | 111.26 | 0.86 | 0.65 | 87.76 | 0.58 |
| FKBP6 | 0.66 | 0.19 | 1.82 | 0.82 | 85.71 | 0.58 |
| CYP3A4 | 1.95 | 0.97 | 1.01 | 0.59 | 83.67 | 0.58 |
| KCNU1 | 0.09 | 0.77 | −3.13 | 0.91 | 83.67 | 0.58 |
| THAP1 | 224.43 | 155.33 | 0.53 | 0.74 | 83.67 | 0.58 |
| BOP1 | 163.67 | 332.63 | −1.02 | 0.83 | 81.63 | 0.58 |
| CBWD6 | 57.86 | 40.47 | 0.52 | 0.54 | 81.63 | 0.58 |

Table 4 shows the list with the most frequently sampled genes by LOOCV. Table 5 shows the main genes sampled by the hold out sampler.

Additionaly, Table 6 shows the altered genetic pathways obtained with the most frequently sampled genes provided by the LOOCV (exploitative) and Holdout (exploratory) samplers. The results are similar in both cases and are in line with clinical studies indicating that low levels of interferon are produced upon COVID-19 infection [18,53]. The analysis of the altered pathways supports the idea that viral replication leads to a dysregulated immune response in patients, as suggested by the alteration of Innate immune system pathways, JAK-STAT signaling pathway and antigen-activated B-cell receptor pathway. This result has been also proposed by Channappanavar et al. in the case of *SARS-CoV*-Infected mice. Based on animal models, SARS-CoV was found to induce cytokine response that generally delays IFN, leading to a paucity of immune response (with COL1A1 being a central gene in this

**Table 5**
List of the most frequently sampled genes by the Holdout sampler (HDS).

| Gene Name | MeanC1 | MeanC2 | FC | FR | LOOCV Acc | Frequency |
|---|---|---|---|---|---|---|
| ENO1-AS1 | 0.84 | 2.52 | −1.59 | 0.99 | 73.47 | 0.58 |
| NKAIN1 | 17.11 | 8.54 | 1.00 | 0.82 | 73.47 | 0.58 |
| EFNA1 | 2260.14 | 805.52 | 1.49 | 0.57 | 77.55 | 0.58 |
| SNORD75 | 0.16 | 0.66 | −2.07 | 0.72 | 83.67 | 0.58 |
| SNORD74 | 3.43 | 6.98 | −1.02 | 0.49 | 89.80 | 0.58 |
| KCNT2 | 23.45 | 10.66 | 1.14 | 1.02 | 87.76 | 0.58 |
| EPC1 | 550.69 | 394.32 | 0.48 | 0.58 | 83.67 | 0.58 |
| WEE1 | 643.72 | 446.80 | 0.53 | 0.64 | 83.67 | 0.58 |
| AQP6 | 1.36 | 3.13 | −1.21 | 0.86 | 83.67 | 0.58 |
| ATP6V0D1 | 629.94 | 1256.87 | −1.00 | 0.68 | 85.71 | 0.58 |
| MTRNR2L1 | 0.75 | 0.06 | 3.54 | 0.93 | 85.71 | 0.58 |
| DNAH17 | 133.60 | 41.11 | 1.70 | 0.75 | 85.71 | 0.58 |
| RNF126 | 249.37 | 536.15 | −1.10 | 0.90 | 83.67 | 0.58 |
| NDUFS7 | 232.20 | 503.85 | −1.12 | 0.71 | 83.67 | 0.58 |
| ADAT3 | 16.44 | 37.26 | −1.18 | 0.72 | 85.71 | 0.58 |
| PPAN | 12.08 | 27.10 | −1.17 | 0.69 | 83.67 | 0.58 |
| ZNF627 | 190.88 | 100.39 | 0.93 | 0.82 | 83.67 | 0.58 |
| SLC5A5 | 8.23 | 2.20 | 1.90 | 0.93 | 85.71 | 0.58 |
| KCNN1 | 15.64 | 5.21 | 1.59 | 0.93 | 81.63 | 0.58 |
| GDF15 | 1369.24 | 928.66 | 0.56 | 0.61 | 83.67 | 0.58 |
| ZNF14 | 89.36 | 50.73 | 0.82 | 0.59 | 83.67 | 0.58 |
| PLEKHF1 | 51.00 | 103.33 | −1.02 | 0.88 | 83.67 | 0.58 |
| ATP4A | 2.46 | 0.63 | 1.95 | 0.74 | 83.67 | 0.58 |
| PLEKHG2 | 1353.67 | 826.16 | 0.71 | 0.59 | 83.67 | 0.58 |
| ZNF845 | 200.73 | 129.85 | 0.63 | 0.73 | 83.67 | 0.58 |
| USP37 | 385.81 | 227.38 | 0.76 | 0.65 | 83.67 | 0.58 |
| ZNF280A | 44.42 | 18.96 | 1.23 | 1.00 | 83.67 | 0.58 |
| CABP7 | 4.00 | 2.07 | 0.95 | 0.54 | 83.67 | 0.58 |
| TMEM89 | 0.40 | 0.80 | −1.02 | 1.15 | 83.67 | 0.58 |
| SPOCK3 | 2.93 | 7.37 | −1.33 | 0.87 | 83.67 | 0.58 |
| ANKRD33B | 485.20 | 288.52 | 0.75 | 0.94 | 83.67 | 0.58 |
| MTRNR2L2 | 1.06 | 0.46 | 1.21 | 1.00 | 81.63 | 0.58 |
| NME5 | 6.18 | 18.44 | −1.58 | 0.69 | 83.67 | 0.58 |
| DRD1 | 154.00 | 40.44 | 1.93 | 0.80 | 85.71 | 0.58 |
| HIST1H3J | 3.82 | 1.92 | 0.99 | 1.05 | 87.76 | 0.58 |
| ZNF92 | 202.38 | 111.26 | 0.86 | 0.65 | 87.76 | 0.58 |
| FKBP6 | 0.66 | 0.19 | 1.82 | 0.82 | 85.71 | 0.58 |
| CYP3A4 | 1.95 | 0.97 | 1.01 | 0.59 | 83.67 | 0.58 |
| KCNU1 | 0.09 | 0.77 | −3.13 | 0.91 | 83.67 | 0.58 |
| THAP1 | 224.43 | 155.33 | 0.53 | 0.74 | 83.67 | 0.58 |
| BOP1 | 163.67 | 332.63 | −1.02 | 0.83 | 81.63 | 0.58 |
| CBWD6 | 57.86 | 40.47 | 0.52 | 0.54 | 81.63 | 0.58 |

**Table 6**
Altered pathways in SARS-CoV-2 transcriptomic signature obtained via Leave-One-Out Cross-validation (LOOCV) sampler (exploitative character) and Holdout Sampler (HDS) (explorative character). Both samplers converge to similar pathways suggesting plausible causes and complication of COVID19.

| Holdout Sampler (HDS) | LOOCV Sampler |
|---|---|
| Olfactory transduction | Olfactory transduction |
| Innate immune system | Innate immune system |
| Neuroactive ligand-receptor interaction | Neuroactive ligand-receptor interaction |
| Cytokine-cytokine receptor interaction | Cytokine-cytokine receptor interaction |
| Class A GPCRs (rhodopsin-like) | FSH regulation of apoptosis |
| GPCR ligand binding | Carbohydrate metabolism |
| Messenger RNA processing | Wnt signaling pathway |
| G alpha (i) signaling events | Antigen-activated B-cell receptor |
| JAK-STAT signaling pathway | Developmental biology |
| Peptide G-protein coupled receptors | Gastrin-CREB signaling pathway via PKC and MAPK |

case) [54]. Moreover, it is widely believed that the physiological reasons of a high morbidity is the selective death of pneumocytes, which leads to a lower air interchange in the lungs and filling lungs with fluid [55–58].

Due to the SARS-CoV-2 infection, the inflammatory response can lead to a sudden cytokine storm, which can result in injury of multiple organs. Various studies have shown elevated levels of proinflammatory cytokines in patients with severe course of the disease. This might be caused by DNAH17, KCNT2 and TMEM89 genes in the small genetic signature [59] and is also suggested by the sampled altered pathway of a cytokine-cytokine receptor interaction. It is possible that the SARS-CoV-2-induced immune suppression predisposes to secondary infections, especially in more severely ill patients, and it is unknown if there are any longer term effects on humoral or cell-mediated immunity following SARS-CoV-2 infection [59]. Furthermore, a meta-analysis of various publications from China, including clinical studies of 1527 patients with COVID-19, reports 9.7%, 16.4% and 17.1% prevalence of diabetes, cardio-cerebrovascular disease and hypertension, respectively [53]. This is further confirmed by our analysis of altered pathways for the class A GPCRs, GPCR ligand binding receptors and peptide G-protein coupled receptors shown in 4. The small genetic signature suggests this feature as inferred by the sampled genes KCNU1 and TMEM89.

A remarkable feature in the small genetic signature is the presence of DNAH17, a gene which is involved in kidney/liver disease. Zhang et al. reported that the incidence of hepatic abnormalities significantly increases after COVID-19 infection and during the course of the disease, which may indicate the effect of SARS-CoV-2 on the liver or side effects of the medications used by patients [60]. In addition to liver injuries, some articles have similarly reported an increased incidence of acute renal injury following COVID-19 [56,61], which could be related to the follicle stimulating hormone (FSH) regulation of apoptosis reported in Table 6 [62]. Lastly, it is worth mentioning the role of COL1A1, which seems to play a central role in SARS-CoV-2 disease and subsequent secondary effects. In this sense, it could explain several neurological disorders found after infection and recovery, as reported in recent literature [63], also supported by the fact that the neuroactive ligand-receptor interaction pathways has been pointed as altered in the sampling (see Table 6).

Gastrointestinal manifestations have also been reported in patients with SARS-CoV-2. One study in California found that a significant number of COVID-19 patients suffered likewise from a loss of appetite, nausea, vomiting and diarrhea [64]. This problem might be caused by the alteration of the gastrin-CREB signaling pathway via the PKC and MAPK pathway.

### *4.2. Connectivity mapping of two different gene samplings reveals potential therapeutic targets for COVID-19*

By identifying the overexpressed and under expressed genes from a robust transcriptomic analysis of SARS-CoV-2 data, we could create two lists of 35 and 38 drugs respectively. These drugs have the potential to reverse the transcriptomic signature linked to COVID-19. The top compounds for each sampling are listed in Table 7. However, a more comprehensive list of potential treatments is given in Supplementary Material.

Acute respiratory syndrome upon infections of influenza viruses could be potentially relieved with geldanamycin, a 1,4-benzoquinone ansamycin antitumor antibiotic, which is a family of bacterial secondary metabolites, that show antimicrobial activity against many type of bacteria. Studies with mice have shown markedly reduced production of major proinflammatory cytokines and chemokines and attenuated infiltration and activation of immune cells, but did not notice any alteration of the generation of virus-neutralizing antibodies [65]. Further studies focused on how geldanamycin binds to the heat shock protein Hsp90 inhibiting its ATPase activity and thus minimizing viral replication of Herpes Simplex Virus 2 (HSV-2) and of SARS-CoV-2 [66] suggest that Hsp90 is a potential novel target for antiviral research [67]. Tanespimycin is another Hsp90 inhibitor, that was found to be potentially beneficious in treatment filoviruses, such as Ebola virus [68], or enteroviruses, such as enterovirus 71 [69].

The Holdout sampler with a higher explorative character, found four additional drugs in addition to Trichostatin A (TSA). This sampler seems to propose drugs more orientated towards treatment of secondary effects

**Table 7**
Main drugs repositioned from CMAP via altered genetic signatures obtained with LOOCV and Holdout samplers.

| | LOOCV | | | | HOLDOUT | | |
|---|---|---|---|---|---|---|---|
| Drug | Cell Line | μM | Freq | Drug | Cell Line | μM | Freq |
| trichostatin A | MCF7 | 0.1 | 0.06 | trichostatin A | MCF7 | 0.1 | 0.06 |
| trichostatin A | MCF7 | 1 | 0.06 | trichostatin A | MCF7 | 1 | 0.06 |
| vorinostat | MCF7 | 10 | 0.06 | vorinostat | MCF7 | 10 | 0.06 |
| trichostatin A | PC3 | 0.1 | 0.05 | trichostatin A | PC3 | 0.1 | 0.05 |
| trichostatin A | PC3 | 1 | 0.05 | trichostatin A | PC3 | 1 | 0.05 |
| PNU-0230031 | MCF7 | 1 | 0.04 | PNU-0230031 | MCF7 | 1 | 0.04 |
| PNU-0230031 | MCF7 | 10 | 0.04 | PNU-0230031 | MCF7 | 10 | 0.04 |
| geldanamycin | MCF7 | 1 | 0.04 | geldanamycin | MCF7 | 1 | 0.04 |
| tanespimycin | MCF7 | 0.1 | 0.03 | tanespimycin | MCF7 | 0.1 | 0.03 |
| tanespimycin | MCF7 | 1 | 0.03 | tanespimycin | MCF7 | 1 | 0.03 |
| trichostatin A | MCF7 | 0.1 | 0.06 | trichostatin A | MCF7 | 0.1 | 0.06 |
| trichostatin A | MCF7 | 1 | 0.06 | trichostatin A | MCF7 | 1 | 0.06 |

associated with COVID-19. Due to its explorative nature, it is capable of sampling other altered pathways associated with the COVID infection. However, since the side effects of COVID-19 are still under clinical study, these results must be considered as preliminary ones, due to a lack of available data for detailed computational studies. Female hormone estradiol (estrogen) and 5-alpha-reductase-inhibitor finasteride, could potentially reduce inflammation and increase breath capacity and stability [70,71]. Another drug discovered by the Holdout Sampler, was Eldeline, which is considered an antiarrhythmic agent. This drug could be potentially used to treat the side effects of COVID-19 infection associated with arrhythmias. However, the exact contribution of COVID-19 infection to the development of arrhythmias in asymptomatic, mildly ill, critically ill, and recovered patients is still not fully confirmed [72].

Regarding TSA, it has been reported that histone deacetylases (HDCAs) inhibitors have a great potential in treating neurological disorders via binding to the CREB-binding protein (CBP) [73,74]. This suggests that TSA could be used not only to alleviate the viral infection, but also to treat viral side effects after recovery, since neurological post-COVID-19 problems have been reported in the literature [63] most likely related to the alteration of the COL1A1 gene [75]. Thalidomide was originally a sedative drug prescribed for a morning-sickness. Nevertheless, it was immediately removed from the market after it was found to cause birth defects: stunted or malformed limbs in babies born from mothers who took thalidomide during pregnancy. Currently, this drug is used to treat leprosy and multiple myeloma, and it was considered as a treatment for lung injury caused by the H1N1 flu virus when it first appeared in humans in 2009 [76]. Basically, thalidomide tamps down inflammation in the lungs, reduces scarring and rein in the overactive immune system. A clinical trial in Wenzhou, China is being conducted to elucidate whether this drug could shorten the time of recovery in patients with pneumonia caused by COVID-19, to reduce the number of patients who critically need mechanical ventilation [77].

Although our analysis has been primarily focused on finding potential drugs for COVID-19, it has not been only limited to this. A comparative analysis of the repositioned drugs obtained by using only leave-one-out cross validation (LOOCV), Holdout (HDS), and other additional samplers (FRS, RF and RS) [18,34] are given in Tables 8 and 9, and confirms what has been previously discussed. All these algorithms use different rules to sample the set of genes altered by the COVID-19 infection, that are later used in drug repositioning.

Within the drugs that reverse the disease with milder effects we have found several interesting mechanisms of action (MOAs):

1. Antibiotic protein synthesis inhibitor (puromycin or anisomycin).
2. Anthelmintics (pyrantel, niclosamide) and antibacterial (sulfaphenazole).

**Table 8**
Main repositioned drugs that reverse the disease (with stronger effects) found by different samplers: LOOCV, Holdout, Fisher's ratio (FRS), Random Forest (RF) and Random Sampler (FR). In this paper) and the Holdout (HDS) samplers are discussed in detail. Repeated drug names might correspond to different cell-lines and/or to different concentrations (see supplementary material for further details).

| LOOCV | HDS | FRS | RF | RS |
|---|---|---|---|---|
| *trichostatin A* | *trichostatin A* | *trichostatin A* | *eldeline* | *acetylsalicylic A* |
| *vorinostat* | *eldeline* | *trichostatin A* | *CP-690334-01* | *doxorubicin* |
| *trichostatin A* | *thalidomide* | *estradiol* | *CP-690334-01* | *etiocholanolone* |
| *PNU-0230031* | *estradiol* | *estradiol* | *karakoline* | *genistein* |
| *geldanamycin* | *finasteride* | *estradiol* | *thalidomide* | *genistein* |
| *tanespimycin* | *trichostatin A* | *daunorubicin* | *thalidomide* | *valproic acid* |
| *trichostatin A* | *trichostatin A* | *daunorubicin* | *trichostatin A* | *valproic acid* |
| *vorinostat* | *eldeline* | *CP-690334-01* | *trichostatin A* | *valproic acid* |
| *trichostatin A* | *thalidomide* | *CP-690334-01* | *trichostatin A* | *valproic acid* |
| *PNU-0230031* | *estradiol* | *trichostatin A* | *trichostatin A* | *valproic acid* |
| *geldanamycin* | *finasteride* | *trichostatin A* | *eldeline* | *acetylsalicylic A* |

3. Antihypertensive drugs and blood vessel conditions (bisprolol, harpagoside, hesperidin) or antiarrhythmic and anti-inflammatory (napelline, myricetin)
4. Alkaloid derivatives and anticholinergics (Hyoscyamine).
5. Steroid hormones (androsterone).

### 4.3. Testing for biological invariance

Testing for biological invariance [78] is important to assure that the drug repositioning does not depend on the algorithm that has been used. This was partially confirmed by the results presented in Tables 4 and 5 obtained by using five different types of samplers. For final confirmation, we have performed a different drug repositioning experiment by combining a robust sampling of high discriminatory genetic networks via a cross validation sampler combined with COGENA repositioning algorithm [79]. Basically, the methodology that we used is like the one that has been previously described: sampling the uncertainty space of the COVID-19 phenotype prediction problem and using the knowledge derived from the analysis of altered genetic pathways to find the desired drugs. For each learning stage of the cross validation we select genes via Random Forest [80], and support vector machine with least absolute shrinkage and selection operator (SVM LASSO) [81]. The discriminatory genes were previously selected within those that were differentially

**Table 9**
Main repositioned drugs that reverse the disease (with milder effects/lower CMAP concordance) found by different samplers: LOOCV, Holdout (HDS), Fisher's ratio (FRS), Random Forest (RF) and Random Sampler (RS). In this paper only LOOCV and HDS samplers are discussed in detail.

| LOOCV | HDS | FRS | RF | RS |
|---|---|---|---|---|
| *puromycin* | *prednisone* | *pyrantel* | *bisoprolol* | *puromycin* |
| *bisoprolol* | *androsterone* | *hyoscyamine* | *hyoscyamine* | *fendiline* |
| *napelline* | *anisomycin* | *bisoprolol* | *pentetrazol* | *pimozide* |
| *pyrantel* | *hydroflumethiazide* | *5707885* | *pyrantel* | *thioridazine* |
| *chenodeoxycholic A.* | *paroxetine* | *Prestwick-665* | *etodolac* | *thioridazine* |
| *dihydroergocristine* | *harpagoside* | *8-azaguanine* | *scoulerine* | *anisomycin* |
| *H-89* | *hesperidin* | *androsterone* | *alprostadil* | *calmidazolium* |
| *harpagoside* | *myricetin* | *cyclopentolate* | *puromycin* | *carbinoxamine* |
| *hyoscyamine* | *pyrantel* | *dexibuprofen* | *dihydroergocristine* | *carmustine* |
| *niclosamide* | *sulfaphenazole* | *hyoscyamine* | *loxapine* | *harmine* |

expressed and exhibited the highest Fisher's ratios. For each selection a group of genes was randomly selected via Metropolis-Hastings algorithm [82]. The predictive accuracy of these networks was established in the corresponding test fold via Random Forest, linear SVM and k-nearest neighbors' algorithm (k-NN). This process was repeated 500 times, providing 1500 different gene expression signatures with their corresponding predictive accuracies. Finally, a posterior frequency analysis was carried out considering those genetic signatures with a predictive accuracy higher than 80%. Finally, the drug repurposing with the first 100 genes using COGENA was performed. Previously this algorithm was used in the drug repositioning for different gene clusters. Here we have defined the sets of underregulated and overregulated genes.

The genetic pathways related to the most frequently sampled signatures revealed by this analysis were:

1. Regulation of RAS family activation.
2. The p38 MAPK signaling pathway.
3. EGFR-dependent endothelin signaling event.
4. Ras signaling in the $CD4^{+}$ TCR pathway.
5. CXCR3-mediated signaling events.
6. Regulation of cytoplasmic and nuclear SMAD2/3 signaling.
7. Cellular roles of anthrax toxin.
8. ErbB1 downstream signaling.
9. The p38 signaling mediated by MAPKAP kinases.
10. BCR signaling pathway.

This analysis points mainly to the regulation of RAS family and various signalling pathways. The RAS family of genes encode proteins that are involved in cell signalling pathways that control cell growth and cell death. Mutated forms of the RAS gene may be found in some types of cancer. The p38 mitogen-activated protein kinases are a class of mitogen-activated protein kinases (MAPKs) that are responsive to stress stimuli, such as cytokines, ultraviolet irradiation, heat and osmotic shocks. These proteins are involved in cell differentiation, apoptosis and autophagy. Oxidative stress is the most powerful stress activating p38 MAPK.

CXCR3 is a membrane receptor with affinity for certain cytokines. CXCR3 gene is located on the human X chromosome. CXCR3 is expressed primarily on activated T lymphocytes and NK cells and some epithelial cells. CXCR3 is able to regulate leukocyte trafficking and has been associated to different diseases [83–89].

The epidermal growth factor receptor is a protein coded by EGFR gene involved in signaling events. It is a receptor tyrosine kinase that is commonly upregulated in different kind of cancer. It has been found to be involved in the development of pulmonary fibrosis in patients who went through SARS coronavirus infection [90].

The role of the EGFR in viral infections has been outlined by Hondermarck, H. et al. [91], who indicated that many viruses use growth factor receptors to physically attach to the cell surface and internalize and also to divert receptor tyrosine kinase signaling to replicate. These authors propose repurposing drugs that initially have been developed to target growth factor receptors and their signalling in cancer against viral infections and particularly against COVID-19. ErbB1 downstream signaling is also related to this pathway.

The p38 MAPK pathway plays a crucial role in the release of pro-inflammatory cytokines such as IL-6 and has been implicated in acute lung injury and myocardial dysfunction. The p38 MAPK inhibition has been proposed as a promising therapeutic approach for COVID-19 [92]. It has been established that the uncontrolled inflammatory response in COVID-19 infection may be caused by disproportionately upregulated p38 activity. The idea exposed in this paper is that SARS-CoV-2 may induce overwhelming inflammation by directly activating p38 and downregulating a key inhibitory pathway, while simultaneously taking advantage of p38 activity to replicate.

Fig. 6 shows the main repositioned drugs found in this analysis. The COGENA drug repositioning method requires several clusters. We have chosen 2 clusters since the number of selected genes is low (100). The first cluster was built with 97 genes (1#97) targeting 20 drugs with a high score. The second cluster contained only 3 remaining genes gave no results (2#3). Moreover, we found that a cluster composed of 93 downregulated genes (Down#93), provided similar result of the first cluster. Finally, All#100 shows the score using all the 100 genes. The drugs with the highest scores were anisomycin and 16-dimethylprostaglandin. Interestingly, some of these drugs were discovered earlier with CMAP-based Dr. Insight package, namely: rofecoxib, cyclosporin, tanespimycin, and trichostatin A.

Our earlier CMAP-based analysis using the same genetic signatures selected the following compounds: acetylsalicylic acid, tanespimycin, riluzole, flucloxacillin, TSA, minocycline and ciclosporin, among others.

### 4.4. Analysis of the second data set: the immunological panel

The analysis of the second dataset provided the following pathways related to the immunological system. Fig. 7 shows the main pathways found using Reactome.

It can be observed that the main pathways are related to cytokine signalling, the innate and adaptive immune systems, cytokine signalling and interferon gamma-signalling that is related to the innate immune response (NK and NKT cells). Fig. 8 shows the main repositioned drugs using the main set of altered genes using the same methodology as previously.

There are different types of drugs in this list:

1. Immunosuppressors: Sirolimus
2. Antibiotics used to treat antibacterial infections: Clindamycin, Nalidixic acid, Sulfaphenazole.
3. Histone deacetylases (HDAC) inhibitors: Vorinostat, TSA.
4. Alkaloids used to treat neurological conditions: Securinine.
5. Chemotherapy drugs: Chlorambucil.
6. Antihypertensive drugs: Guanethidine.
7. Antiarrhythmic drugs: Moracizine.
8. Inhibition of thyroid hormones: Thiamazole.

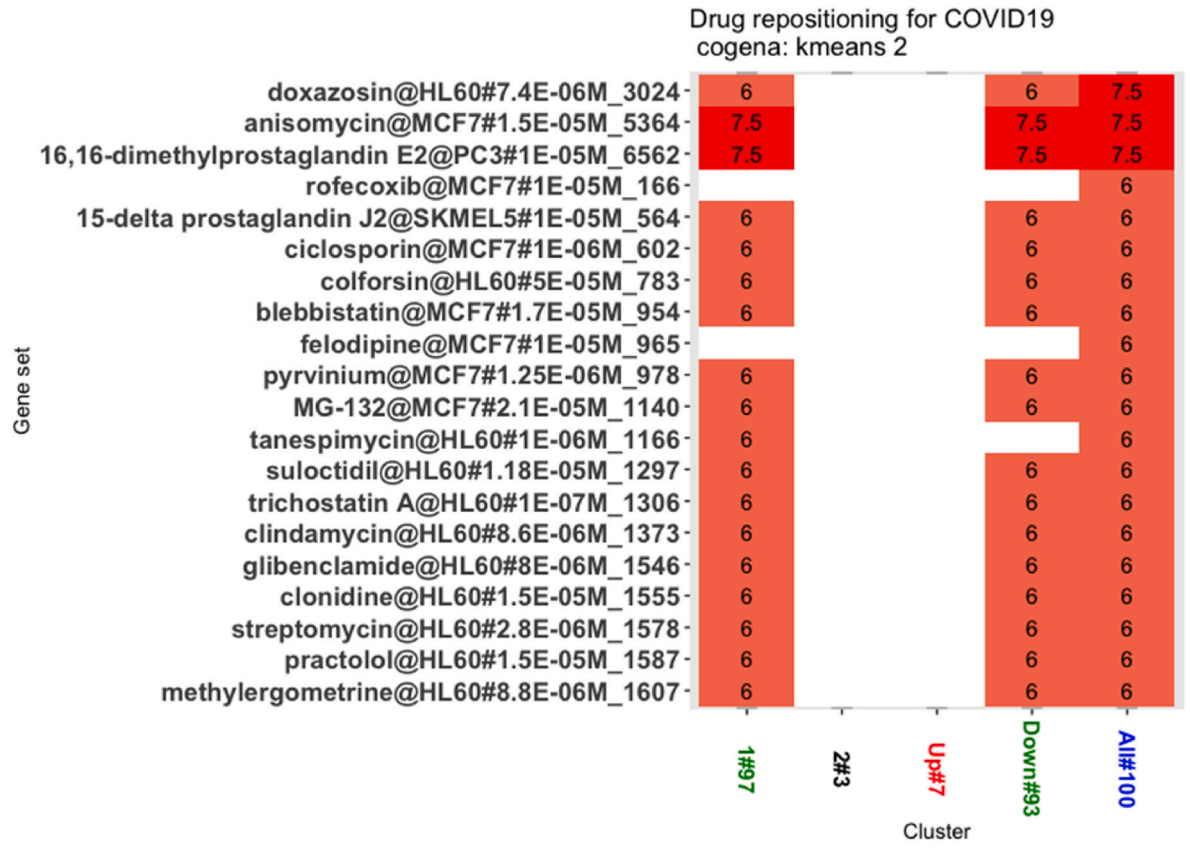


**Fig. 6.** Main drugs repositioned via COGENA. Here 1#97 and 2#3 denote that cluster number 1 has 97 genes and cluster number 2 has 3 genes. Down#93 represent a subset of 93 downregulated genes. Finally, the label All#100 shows the score for all the 100 genes.

9. Anthelmintic agents: Niridazole.
10. Antidepressant: Imipramine.

The drugs with the highest score are sirolimus, moricizine and guanethidine. Interestingly, some of these drugs have already appeared in the previous analysis presented in this paper.

## 5. Discussion

Our analysis of transcriptomic data clearly suggests that the transcriptomic response to SARS-CoV-2 infection indicates the imbalance between the control of the virus replication, the innate immunity mechanisms that provides the initial defence against infection (cellular and biochemical mechanisms that are in place before the infection and are poised to response rapidly to infections), and the activation of the adaptive immune response that develops later. Hence, COVID-19 treatments should not be targeted only on the interferon (IFN) response but also on the control of the inflammation. Our results demonstrate a variety of genes involved in chemokines and interleukins (ILs) signalling pathways, and inflammatory response. Therefore, COVID-19 treatment efforts should be focused on the standard of care drugs that can be repurposed rapidly and applied to a vast population of patients. Besides, we have shown via PCA that the transcriptomic signatures are different from RSV and influenza.

Although our research was limited to purely computational studies of transcriptomic data in cell lines, there is some clinical evidence that the identified drugs have some therapeutic potential for treating COVID-19 patients and preventing pathological cellular responses to SARS-CoV-2 infection. This is mainly inferred by clinical data showing that the repurposed drugs selected by our method for COVID-19, have been used in the past to treat similar viral diseases. Our list of possible repurposed drugs against COVID-19, needs further detailed clinical studies to elucidate whether these drugs are effective for COVID-19 treatment, either to protect against the infection by interacting with the viral mechanisms, molecular docking, etc., or to provide post-infection treatment by reversing coronavirus-induced expression signatures, captured by CMAP connectivity mapping. Additionally, our analysis suggests that COLIA1 gene plays a very important role, among other genes that have been revealed and whose roles have been analysed. This gene has been identified as a candidate for heart failure [93]. We have also provided a discriminatory small-scale signature composed of 31 genes that could be used clinically. Finally, it is plausible that the transcriptomic signature of patients with acute infection, differs from the genetic signatures that were unravelled in the present analysis. Consequently, much more publicly available transcriptional signatures data from different patients at varying disease stages, would be needed to improve our current computational limitations. However, due to the speed of the pandemic and the urgency of the problem, we believe that the proposed novel method of selecting drug candidates that could be repurposed against SARS-CoV-2, is vital and might be very useful for the whole community of virologists, medical doctors, data scientists, and pharmacologists who are currently studying COVID-19. The drug candidates were identified using a data-driven hypothesis-free approach. Although the computational approach presented in this paper is very different from the previous works by Blanco de Melo et al. [18] on the first dataset and by Desai et al. [34] on the second dataset, our results aligned with the results obtained by these two teams. The aim of our work was to find the altered genetic pathways to perform drug repositioning. Some of these drugs were reconfirm by an independent dataset and this fact becomes important to highlight. Our machine learning-based method can be significantly improved in the future by rapidly growing transcriptomic signature data available from clinical studies. This transcriptomic data should be publicly available and possibly stored in one central depository to ensure speedy research

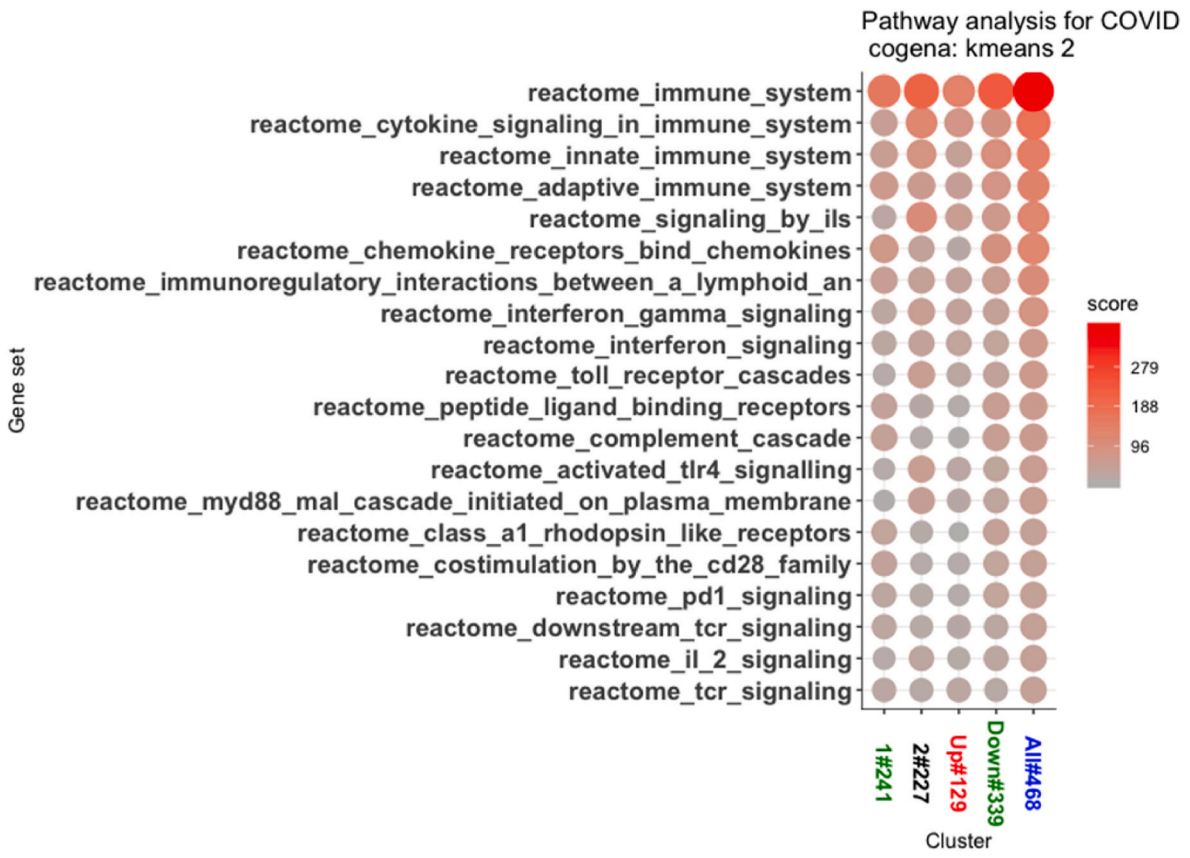


**Fig. 7.** Main altered immunological pathways. 1#241 and 2#227 denotes that cluster number 1 has 241 genes and cluster number 2 has 227 genes. Down#339 and Up#129 shows the number of genes downregulated and upregulated respectively. Finally, the label All#468 shows the score for all the 468 genes.

progress on development of new drugs or vaccines against COVID-19 or other possible future pandemics. One of the main disadvantages in performing this kind of analysis is to obtain independent and compatible datasets regarding the sequencing platform.

## 6. Conclusions

In this paper we have found the small-scale genetic signature that predicts the COVID infection with respect to healthy controls (mock treated) and we have analysed how different samplers explore the uncertainty space of the phenotype prediction problem. Based on the hypothesis of biological invariance, we were able to unravel the main disease altered pathways and to perform drug repositioning by using CMAP methodology. This methodology suggests potential COVID-19 treatments, which then need to undergo clinical testing to establish their efficacy in human patients, to be authorized as treatments by the regulatory agencies. In our opinion the most interesting finding is the role of the COL1A1 gene that plays an important role in the regulation of the immune system. The COL1A1 gene produces the alpha 1 chain of type I collagen found in different tissues (skin, tendons, corneas, lungs, and bones). Type I collagen is also a major structural protein in the lung and is stimulated during certain inflammatory reactions. It has been shown that the need for type I collagen increases substantially in adult respiratory distress syndrome [94].

A lack of glycine affects collagen production and immunity, since glycine has crucial function in cytoprotecting, immune response, growth, development, metabolism, and survival of humans [95]. In addition, the production of collagen is not the same across ethnically diverse population [96]. Around 95% of the European population corresponds to a genotype that has a lower collagen production, and therefore a higher need of both glycine and vitamin C for Type 1 collagen synthesis to maintain optimal levels of each to protect the lungs. Glycine concentrations have consistently been found to be lower in individuals with obesity, Non-Alcoholic Fatty Liver Syndrome (NAFLD), and Type 2 diabetes [97]. All three health conditions are at a higher risk for the severity of COVID-19 infection. Moreover, the decline in collagen production occurs in high-sugar diets, elderly population, and women after the menopause. Although this is just a hypothesis that could be controversial, we believe that the expression level of COL1A1could be adopted as a predictive test of the severity of COVID-19 infection.

Furthermore, we have highlighted the role of other important downregulated genes in lung biopsies, such as, MRPLR1, FSCN1, RNF126 and ANKRD33B, involved in pathways related to dopamine receptors, inflammatory response, and the immune response. Finally, we have pointed out several genes overexpressed in COVID-19 infection which are responsible for cardiomyopathies (DNAH17/KCNT2) and cardiac metabolic effects (KCNT2), severe sepsis (DRD1), immune response (FKBP6) and macular degeneration (KCNN1).

In our opinion all genes, drugs and pathways that were discussed in this paper, open new hypothesis, and possibilities of potential treatment against COVID-19 infection. Additionally, our paper can provide clinicians with useful insights, based on their prior knowledge of the disease, to test various compounds in clinical trials.

## Authors contributions

JLFM, OAM, EDAG and GB designed and implemented the algorithms of methodology used in this paper.

JLFM, OAM, CG and AK have participated in the results interpretation.

All the authors have participated in the design of the study, in the paper writing and in its critical revision.

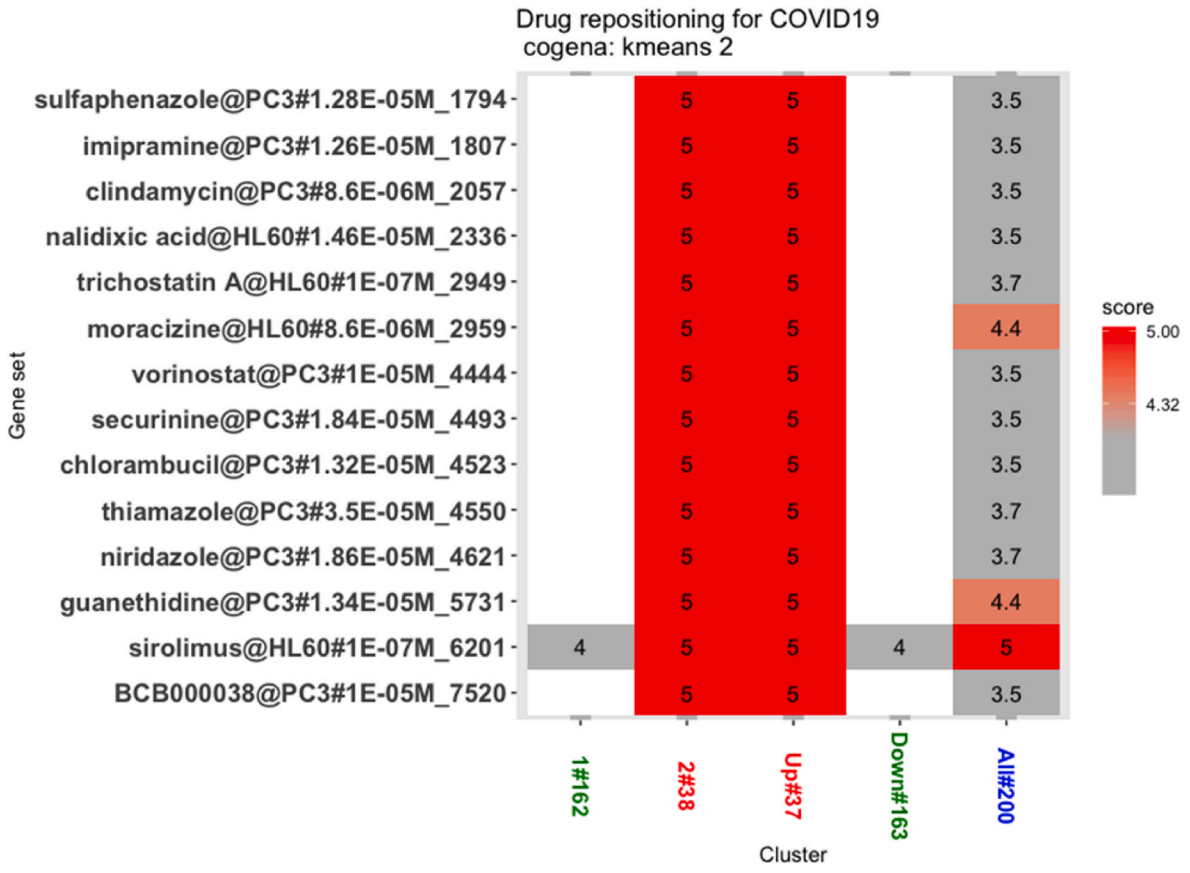


**Fig. 8.** Main drugs repositioned via COGENA for the additional data (immunological panel). 1#162 and 2#38 denotes that cluster number 1 has 162 genes and cluster number 2 has 38 genes. Down#163 and Up#37 are clusters of 163 downregulated genes and of 37 upregulated genes, respectively. Finally, the label All#200 represents the score for all the 200 genes.

## Declaration of competing interest

None Declared.

## Acknowledgments

We acknowledge financial support from NSF grant DBI 1661391, and NIH grant R01 GM127701. We would like to acknowledge Professor Robert Jernigan (Iowa State University) and Ms. Carolina deAndrés for the English revision and style corrections.

## Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.compbiomed.2022.106029.

## References

[1] J. Cui, F. Li, Z.-L. Shi, Origin and evolution of pathogenic coronaviruses, Nat. Rev. Microbiol. 17 (2019) 181–192.
[2] C. Huang, Y. Wang, X. Li, L. Ren, J. Zhao, Y. Hu, et al., Clinical features of patients infected with 2019 novel coronavirus in Wuhan, China, Lancet 395 (2020) 497–506.
[3] V.M. Corman, O. Landt, M. Kaiser, R. Molenkamp, A. Meijer, D.K.W. Chu, et al., Detection of 2019 novel coronavirus (2019-nCoV) by real-time RT-PCR, Euro Surveill. 25 (2020), 2000045.
[4] A.R. Fehr, R. Channappanavar, S. Perlman, Middle East respiratory syndrome: emergence of a pathogenic human coronavirus, Annu. Rev. Med. 68 (2017) 387–399.
[5] J.F.-W. Chan, S. Yuan, K.-H. Kok, K.K.-W. To, H. Chu, J. Yang, et al., A familiar cluster of pneumonia associated with the 2019 novel coronavirus indicating person-to-person transmission: a study of a family cluster, Lancet 395 (2020) 514–523.
[6] S.K.P. Lau, P.C.Y. Woo, K.S.M. Li, Y. Huang, H.-W. Tsoi, B.H.L. Wong, et al., Severe acute respiratory syndrome coronavirus-like virus in Chinese horseshoe bats, Proc. Natl. Acad. Sci. USA 102 (2005) 14040–14045.
[7] L.E. Miller, R. Bhattacharyya, A.L. Miller, Spatial analysis of global variability in Covid-19 burden, Risk Manag. Healthc. Pol. 13 (2020) 519.
[8] N. Chen, M. Zhou, X. Dong, J. Qu, F. Gong, Y. Han, et al., Epidemiological and clinical characteristics of 99 cases of 2019 novel coronavirus pneumonia in Wuhan, China: a descriptive study, Lancet 395 (2020) 507–513.
[9] R. Wölfel, V.M. Corman, W. Guggemos, M. Seilmaier, S. Zange, M.A. Müller, et al., Virological assessment of hospitalized patients with COVID-2019, Nature 581 (2020) 465–469.
[10] M. Fani, A. Teimoori, S. Ghafari, Comparison of the COVID-2019 (SARS-CoV-2) pathogenesis with SARS-CoV and MERS-CoV infections, Future Virol. 15 (2020) 317–323.
[11] W. Li, M.J. Moore, N. Vasilieva, J. Sui, S.K. Wong, M.A. Berne, et al., Angiotensin-converting enzyme 2 is a functional receptor for the SARS coronavirus, Nature 426 (2003) 450–454.
[12] M. Hoffmann, H. Kleine-Weber, S. Schroeder, N. Krüger, T. Herrler, S. Erichsen, et al., SARS-CoV-2 cell entry depends on ACE2 and TMPRSS2 and is blocked by a clinically proven protease inhibitor, Cell 181 (2020) 271–280.
[13] C.A. Janeway Jr., R. Medzhitov, Innate immune recognition, Annu. Rev. Immunol. 20 (2002) 197–216.
[14] S. Hur, Double-stranded RNA sensors and modulators in innate immunity, Annu. Rev. Immunol. 37 (2019) 349–375.
[15] H.M. Lazear, J.W. Schoggins, M.S. Diamond, Shared and distinct functions of type I and type III interferons, Immunity 50 (2019) 907–923.
[16] C.L. Sokol, A.D. Luster, The chemokine system in innate immunity, Cold Spring Harbor Perspect. Biol. 7 (2015) a016303.
[17] A. Garc\'\ia-Sastre, Ten strategies of interferon evasion by viruses, Cell Host \& Microbe 22 (2017) 176–184.
[18] D. Blanco-Melo, B.E. Nilsson-Payant, W.-C. Liu, S. Uhl, D. Hoagland, R. Møller, et al., Imbalanced host response to SARS-CoV-2 drives development of COVID-19, Cell 181 (2020) 1036–1045.
[19] Álvarez-Machancoses Ó, J.L. Fernández-Martínez, Using artificial intelligence methods to speed up drug discovery, Expert Opin Drug Discov 14 (2019) 769–777, https://doi.org/10.1080/17460441.2019.1621284.
[20] M.B. Frieman, J. Chen, T.E. Morrison, A. Whitmore, W. Funkhouser, J.M. Ward, et al., SARS-CoV pathogenesis is regulated by a STAT1 dependent but a type I, II and III interferon receptor independent mechanism, PLoS Pathog 6 (2010), e1000849.

[21] S.A. Kopecky-Bromberg, L. Mart\'\inez-Sobrido, M. Frieman, R.A. Baric, P. Palese, Severe acute respiratory syndrome coronavirus open reading frame (ORF) 3b, ORF 6, and nucleocapsid proteins function as interferon antagonists, J Virol 81 (2007) 548–557.

[22] A. Garc\'\ia-Sastre, A. Egorov, D. Matassov, S. Brandt, D.E. Levy, J.E. Durbin, et al., Influenza A virus lacking the NS1 gene replicates in interferon-deficient systems, Virology 252 (1998) 324–330.

[23] Meredith Wadman JC-F, Jocelyn Kaiser CM. How does coronavirus kill? Clinicians trace a ferocious rampage through the body, from brain to toes. Science (80-) 2020.

[24] A. Khan, M. Khan, S. Saleem, Z. Babar, A. Ali, A.A. Khan, et al., Phylogenetic analysis and structural perspectives of RNA-dependent RNA-polymerase inhibition from SARs-CoV-2 with natural products, Interdiscip Sci Comput Life Sci 12 (2020) 335–348.

[25] M.T. Khan, A. Ali, Q. Wang, M. Irfan, A. Khan, M.T. Zeb, et al., Marine natural compounds as potents inhibitors against the main protease of SARS-CoV-2—a molecular dynamic study, J Biomol Struct Dyn 39 (2021) 3627–3637.

[26] A. Khan, S.S. Ali, M.T. Khan, S. Saleem, A. Ali, M. Suleman, et al., Combined drug repurposing and virtual screening strategies with molecular dynamics simulation identified potent inhibitors for SARS-CoV-2 main protease (3CLpro), J Biomol Struct Dyn 39 (2021) 4659–4670.

[27] C. Wang, S. Wang, D. Li, D.-Q. Wei, J. Zhao, J. Wang, Human intestinal defensin 5 inhibits SARS-CoV-2 invasion by cloaking ACE2, Gastroenterology 159 (2020) 1145–1147.

[28] Cernea A, Fernandez-Martinez JL, deAndres-Galiana EJ, Fernandez-Ovies FJ, Fernandez-Muniz Z, Alvarez-Machancoses O, et al. Comparison of Different Sampling Algorithms for Phenotype Prediction. Lect Notes Comput Sci (Including Subser Lect Notes Artif Intell Lect Notes Bioinformatics) 2018;10814 LNBI:33-45. doi:10.1007/978-3-319-78759-6_4.

[29] X.A. Qu, D.K. Rajpal, Applications of Connectivity Map in drug discovery and development, Drug Discov Today 17 (2012) 1289–1298.

[30] J. Lamb, E.D. Crawford, D. Peck, J.W. Modell, I.C. Blat, M.J. Wrobel, et al., The Connectivity Map: using gene-expression signatures to connect small molecules, genes, and disease, Science (80-) 313 (2006), 1929–35.

[31] J. Chan, X. Wang, J.A. Turner, N.E. Baldwin, J. Gu, Breaking the paradigm: Dr Insight empowers signature-free, enhanced drug repurposing, Bioinformatics 35 (2019) 2818–2826.

[32] B. Efron, Bootstrap methods: another look at the Jackknife, Ann Stat 7 (1979) 1–26, https://doi.org/10.1214/aos/1176344552.

[33] Z. Fernández-Muñiz, K. Hassan, J.L. Fernández-Mart'\inez, Data kit inversion and uncertainty analysis, J Appl Geophys 161 (2019) 228–238, https://doi.org/10.1016/j.jappgeo.2018.12.022.

[34] N. Desai, A. Neyaz, A. Szabolcs, A.R. Shih, J.H. Chen, V. Thapar, et al., Temporal and spatial heterogeneity of host response to SARS-CoV-2 pulmonary infection, MedRxiv Prepr Serv Heal Sci (2020), https://doi.org/10.1101/2020.07.30.20165241.

[35] E.J. Deandrés-Galiana, J.L. Fernández-Martínez, S.T. Sonis, Design of biomedical robots for phenotype prediction problems, J Comput Biol 23 (2016), https://doi.org/10.1089/cmb.2016.0008.

[36] E.J. deAndrés-Galiana, J.L. Fernández-Martínez, S.T. Sonis, Sensitivity analysis of gene ranking methods in phenotype prediction, J Biomed Inform 64 (2016), https://doi.org/10.1016/j.jbi.2016.10.012.

[37] E.J. Deandrés-Galiana, J.L. Fernández-Martínez, L.N. Saligan, S.T. Sonis, Impact of microarray preprocessing techniques in unraveling biological pathways, J Comput Biol 23 (2016), https://doi.org/10.1089/cmb.2016.0042.

[38] O. Zhaxybayeva, J.P. Gogarten, Bootstrap, Bayesian probability and maximum likelihood mapping: exploring new tools for comparative genome analyses, BMC Genomics 3 (2002) 1–15.

[39] F. Li, Y. Zhou, Y. Zhang, J. Yin, Y. Qiu, J. Gao, et al., POSREG: proteomic signature discovered by simultaneously optimizing its reproducibility and generalizability, Brief Bioinform 23 (2022).

[40] Q. Yang, B. Li, J. Tang, X. Cui, Y. Wang, X. Li, et al., Consistent gene signature of schizophrenia identified by a novel feature selection strategy from comprehensive sets of transcriptomic data, Brief Bioinform 21 (2020) 1058–1068.

[41] J.L. Fernández-Martínez, A. Cernea, F.J. Fernández-Ovies, Z. Fernández-Muñiz, O. Alvarez-Machancoses, L. Saligan, et al., Sampling defective pathways in phenotype prediction problems via the Holdout sampler, Int. Conf. Bioinforma. Biomed. Eng. (2018) 24–32.

[42] A. Cernea, J.L. Fernández-Mart\'\inez, E.J. deAndrés-Galiana, F.J. Fernández-Ovies, O. Alvarez-Machancoses, Z. Fernández-Muñiz, et al., Robust pathway sampling in phenotype prediction. Application to triple negative breast cancer, BMC Bioinformatics 21 (2020) 1–13.

[43] A. Cernea, J.L. Fernández-Martínez, E.J. deAndrés-Galiana, F.J. Fernández-Ovies, Z. Fernández-Muñiz, O. Alvarez-Machancoses, et al., Sampling defective pathways in phenotype prediction problems via the Fisher's ratio sampler, Lect. Notes Comput. Sci. (including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics) 10814 (2018) 15–23, https://doi.org/10.1007/978-3-319-78759-6_2. LNBI, Springer Verlag.

[44] L.N. Saligan, J.L. Fernández-Martínez, E.J. deAndrés-Galiana, S. Sonis, Supervised classification by filter methods and recursive feature elimination predicts risk of radiotherapy-related fatigue in patients with prostate cancer, Cancer Inform 13 (2014), https://doi.org/10.4137/CIN.S19745. CIN.S19745.

[45] J.L. Fernández-Martínez, E.J. deAndrés-Galiana, S.T. Sonis, Genomic data integration in chronic lymphocytic leukemia, J Gene Med (2017), https://doi.org/10.1002/jgm.2936.

[46] A. Subramanian, R. Narayan, S.M. Corsello, D.D. Peck, T.E. Natoli, X. Lu, et al., A next generation connectivity map: L1000 platform and the first 1,000,000 profiles, Cell 171 (2017) 1437–1452.

[47] F.J. Massey Jr., The Kolmogorov-Smirnov test for goodness of fit, J Am Stat Assoc 46 (1951) 68–78.

[48] E.J. deAndrés-Galiana, J.L. Fernández-Martínez, S.T. Sonis, Design of biomedical robots for phenotype prediction problems, J Comput Biol 23 (2016) 678–692, https://doi.org/10.1089/cmb.2016.0008.

[49] R.E. Day, P. Kitchen, D.S. Owen, C. Bland, L. Marshall, A.C. Conner, et al., Human aquaporins: regulators of transcellular water flow, Biochim Biophys Acta (BBA)-General Subj 1840 (2014) 1492–1506.

[50] A. Bharat, M. Querrey, N.S. Markov, S. Kim, C. Kurihara, R. Garza-Castillon, et al., Lung transplantation for patients with severe COVID-19, Sci Transl Med 12 (2020), eabe4282.

[51] J. Bi, NK cell dysfunction in patients with COVID-19, Cell \& Mol Immunol (2022) 1–3.

[52] J.C. Melms, J. Biermann, H. Huang, Y. Wang, A. Nair, S. Tagore, et al., A molecular single-cell lung atlas of lethal COVID-19, Nature 595 (2021) 114–119.

[53] H.K.W. Law, C.Y. Cheung, H.Y. Ng, S.F. Sia, Y.O. Chan, W. Luk, et al., Chemokine up-regulation in SARS-coronavirus–infected, monocyte-derived human dendritic cells, Blood 106 (2005) 2366–2374.

[54] R. Channappanavar, A.R. Fehr, R. Vijay, M. Mack, J. Zhao, D.K. Meyerholz, et al., Dysregulated type I interferon and inflammatory monocyte-macrophage responses cause lethal pneumonia in SARS-CoV-infected mice, Cell Host \& Microbe 19 (2016) 181–193.

[55] Z. Qian, E.A. Travanty, L. Oko, K. Edeen, A. Berglund, J. Wang, et al., Innate immune response of human alveolar type ii cells infected with severe acute respiratory syndrome–coronavirus, Am J Respir Cell Mol Biol 48 (2013) 742–748.

[56] Z. Xu, L. Shi, Y. Wang, J. Zhang, L. Huang, C. Zhang, et al., Pathological findings of COVID-19 associated with acute respiratory distress syndrome, Lancet Respir Med 8 (2020) 420–422.

[57] L.S. King, S. Nielsen, P. Agre, others, Aquaporins and the respiratory system: advice for a lung investigator, J Clin Invest 105 (2000) 15–16.

[58] J. Vella, C. Zammit, G. Di Giovanni, R. Muscat, M. Valentino, The central role of aquaporins in the pathophysiology of ischemic stroke, Front Cell Neurosci 9 (2015) 108.

[59] T.-Y. Xiong, S. Redwood, B. Prendergast, M. Chen, Coronaviruses and the cardiovascular system: acute and long-term implications, Eur Heart J (2020).

[60] C. Zhang, L. Shi, F.-S. Wang, Liver injury in COVID-19: management and challenges, Lancet Gastroenterol \& Hepatol 5 (2020) 428–430.

[61] A. Rismanbaf, S. Zarei, Liver and kidney injuries in COVID-19 and their effects on drug therapy; a letter to editor, Arch Acad Emerg Med 8 (2020) e17–e17.

[62] K. Zhang, L. Kuang, F. Xia, Y. Chen, W. Zhang, H. Zhai, et al., Follicle-stimulating hormone promotes renal tubulointerstitial fibrosis in aging women via the AKT/GSK-3$β$/$β$-catenin pathway, Aging Cell 18 (2019), e12997.

[63] L. Mao, H. Jin, M. Wang, Y. Hu, S. Chen, Q. He, et al., Neurologic manifestations of hospitalized patients with coronavirus disease 2019 in Wuhan, China, JAMA Neurol 77 (2020) 683–690.

[64] G. Cholankeril, A. Podboy, V.I. Aivaliotis, others, High prevalence of concurrent gastrointestinal manifestations in patients with SARS-CoV-2: early experience from California [published online ahead of print April 10, 2020], Gastroenterology (2020) 10.

[65] C. Wang, P. Liu, J. Luo, H. Ding, Y. Gao, L. Sun, et al., Geldanamycin reduces acute respiratory distress syndrome and promotes the survival of mice infected with the highly virulent H5N1 influenza virus, Front Cell Infect Microbiol 7 (2017) 267.

[66] Sultan I, Howard S, Tbakhi A. Drug Repositioning Suggests a Role for the Heat Shock Protein 90 Inhibitor Geldanamycin in Treating COVID-19 Infection 2020.

[67] Y.-H. Li, Q.-N. Lu, H.-Q. Wang, P.-Z. Tao, J.-D. Jiang, Geldanamycin, a ligand of heat shock protein 90, inhibits herpes simplex virus type 2 replication both in vitro and in vivo, J Antibiot (Tokyo) 65 (2012) 509–512.

[68] D.R. Smith, S. McCarthy, A. Chrovian, G. Olinger, A. Stossel, T.W. Geisbert, et al., Inhibition of heat-shock protein 90 reduces Ebola virus replication, Antiviral Res 87 (2010) 187–194.

[69] L. Han, K. Li, C. Jin, J. Wang, Q. Li, Q. Zhang, et al., Human enterovirus 71 protein interaction network prompts antiviral drug repositioning, Sci Rep 7 (2017) 1–13.

[70] R. Rastogi, M.S. Badr, A. Sankari, K. Bakkila, H. Abbas, V. Mukkavilli, et al., Effect of finasteride on breathing stability during Nrem sleep, in: The Elderly. A80-A. Nov. Ther. OSA, American Thoracic Society, 2017. A2587–A2587.

[71] S. Chowdhuri, A. Bascom, D. Mohan, M.P. Diamond, M.S. Badr, Testosterone conversion blockade increases breathing stability in healthy men during NREM sleep, Sleep 36 (2013) 1793–1798.

[72] D.R. Lakkireddy, M.K. Chung, R. Gopinathannair, K.K. Patton, T.J. Gluckman, M. Turagam, et al., Guidance for cardiac electrophysiology during the COVID-19 pandemic from the heart rhythm society COVID-19 task force; electrophysiology section of the American college of cardiology; and the electrocardiography and arrhythmias committee of the council on, Circulation 141 (2020) e823–e831.

[73] A.M. Avila, B.G. Burnett, A.A. Taye, F. Gabanella, M.A. Knight, P. Hartenstein, et al., Trichostatin A increases SMN expression and survival in a mouse model of spinal muscular atrophy, J Clin Invest 117 (2007) 659–671.

[74] C. Rouaux, J.-P. Loeffler, A.-L. Boutillier, Targeting CREB-binding protein (CBP) loss of function as a therapeutic strategy in neurological disorders, Biochem Pharmacol 68 (2004) 1157–1164.

[75] S. Hou, P. Lu, Direct reprogramming of somatic cells into neural stem cells or neurons for neurological disorders, Neural Regen Res 11 (2016) 28.

[76] T.R.C. Vergara, S. Samer, J.R. Santos-Oliveira, L.B. Giron, M.S. Arif, M.L. Silva-Freitas, et al., Thalidomide is associated with increased t cell activation and

inflammation in antiretroviral-naive HIV-infected individuals in a randomised clinical trial of efficacy and safety, EBioMedicine 23 (2017) 59–67.
[77] M.P. Lythgoe, P. Middleton, Ongoing clinical trials for the management of the COVID-19 pandemic, Trends Pharmacol Sci 41 (2020) 363–382.
[78] F.M. Juan Luis, The importance of biological invariance in drug design, Biomed J Sci Tech Res 18 (2019) 13211–13212, https://doi.org/10.26717/bjstr.2019.18.003086.
[79] Z. Jia, Y. Liu, N. Guan, X. Bo, Z. Luo, M.R. Barnes, Cogena, a novel tool for co-expressed gene-set enrichment analysis, applied to drug repositioning and drug mode of action discovery, BMC Genomics 17 (2016) 1–11.
[80] T.K. Ho, Random decision forests, Proc. 3rd Int. Conf. Doc. Anal. Recognit. 1 (1995) 278–282.
[81] R. Tibshirani, Regression shrinkage and selection via the lasso, J R Stat Soc Ser B 58 (1996) 267–288.
[82] Hastings WK. Monte Carlo Sampling Methods Using Markov Chains and Their Applications 1970.
[83] F. Mach, A. Sauty, A.S. Iarossi, G.K. Sukhova, K. Neote, P. Libby, et al., Differential expression of three T lymphocyte-activating CXC chemokines by human atheroma-associated cells, J Clin Invest 104 (1999) 1041–1050.
[84] T.L. Sørensen, M. Tani, J. Jensen, V. Pierce, C. Lucchinetti, V.A. Folcik, et al., Expression of specific chemokines and chemokine receptors in the central nervous system of multiple sclerosis patients, J Clin Invest 103 (1999) 807–815.
[85] D. Jiang, J. Liang, J. Hodge, B. Lu, Z. Zhu, S. Yu, et al., Regulation of pulmonary fibrosis by chemokine receptor CXCR3, J Clin Invest 114 (2004) 291–299.
[86] S. Frigerio, T. Junt, B. Lu, C. Gerard, U. Zumsteg, G.A. Holländer, et al., $β$ cells are responsible for CXCR3-mediated T-cell infiltration in insulitis, Nat Med 8 (2002) 1414–1420.
[87] U. Panzer, O.M. Steinmetz, H.-J. Paust, C. Meyer-Schwesinger, A. Peters, J.-E. Turner, et al., Chemokine receptor CXCR3 mediates T cell recruitment and tissue injury in nephrotoxic nephritis in mice, J Am Soc Nephrol 18 (2007) 2071–2084.
[88] W.W. Hancock, B. Lu, W. Gao, V. Csizmadia, K. Faia, J.A. King, et al., Requirement of the chemokine receptor CXCR3 for acute allograft rejection, J Exp Med 192 (2000) 1515–1520.
[89] J.S. Smith, L.T. Nicholson, J. Suwanpradid, R.A. Glenn, N.M. Knape, P. Alagesan, et al., Biased agonists of the chemokine receptor CXCR3 differentially control chemotaxis and inflammation, Sci Signal 11 (2018), eaaq1075.
[90] T. Venkataraman, M.B. Frieman, The role of epidermal growth factor receptor (EGFR) signaling in SARS coronavirus-induced pulmonary fibrosis, Antiviral Res 143 (2017) 142–150.
[91] H. Hondermarck, N.W. Bartlett, V. Nurcombe, The role of growth factor receptors in viral infections: an opportunity for drug repurposing against emerging viral diseases such as COVID-19? FASEB BioAdvances 2 (2020) 296–303.
[92] J.M. Grimes, K.V. Grimes, p38 MAPK inhibition: a promising therapeutic approach for COVID-19, J Mol Cell Cardiol 144 (2020) 63–65.
[93] X. Hua, Y.-Y. Wang, P. Jia, Q. Xiong, Y. Hu, Y. Chang, et al., Multi-level transcriptome sequencing identifies COL1A1 as a candidate marker in human heart failure progression, BMC Med 18 (2020) 1–16.
[94] J.A. Last, A.D. Siefkin, K.M. Reiser, Type I collagen content is increased in lungs of patients with adult respiratory distress syndrome, Thorax 38 (1983) 364–368.
[95] M.A. Razak, P.S. Begum, B. Viswanath, S. Rajagopal, Multifarious beneficial effect of nonessential amino acid, glycine: a review, Oxid Med Cell Longev 2017 (2017).
[96] T.-F. Chan, A. Poon, A. Basu, N.R. Addleman, J. Chen, A. Phong, et al., Natural variation in four human collagen genes across an ethnically diverse population, Genomics 91 (2008) 307–314.
[97] A. Alves, A. Bassot, A.-L. Bulteau, L. Pirola, B. Morio, Glycine metabolism and its alterations in obesity and metabolic diseases, Nutrients 11 (2019) 1356.