 doi:10.20944/preprints202012.0281.v1

A COVID-19 Drug Repurposing Strategy Through Quantitative Homological Similarities by using a Topological Data Analysis based formalism

Raúl Pérez-Moraga[a,b,1], Jaume Forés-Martos[a,b,1], Beatriz Suay[a,b], Jean-Louis Duval[c], Antonio Falcó[a,b,*], Joan Climent[a,c,*]

[a] *ESI International Chair@CEU-UCH, Universidad Cardenal Herrera-CEU, CEU Universities, San Bartolomé 55, 46115 Alfara del Patriarca (Valencia), Spain*
[b] *Departamento de Matemáticas, Física y Ciencias Tecnológicas, Universidad Cardenal Herrera-CEU, CEU Universities, San Bartolomé 55, 46115 Alfara del Patriarca (Valencia), Spain*
[c] *Departamento de Producción y Sanidad Animal, Salud Pública Veterinaria y Ciencia y Tecnología de los Alimentos, Universidad Cardenal Herrera-CEU, CEU Universities, San Bartolomé 55, 46115 Alfara del Patriarca (Valencia), Spain*
[d] *ESI Group, 3bis rue Saarinen, 94528 Rungis CEDEX, France*

Abstract

Since its emergence in March 2020, the SARS-CoV-2 global pandemic has produced more than 65 million cases and one point five million deaths worldwide. Despite the enormous efforts carried out by the scientific community, no effective treatments have been developed to date. We created a novel computational pipeline aimed to speed up the process of repurposable candidate drug identification. Compared with current drug repurposing methodologies, our strategy is centered on filtering the best candidate among all selected targets focused on the introduction of a mathematical formalism motivated by recent advances in the fields of algebraic topology and topological data analysis (TDA). This formalism allows us to compare three-dimensional protein structures. Its use in conjunction with two in silico validation strategies (molecular docking and transcriptomic analyses) allowed us to identify a set of potential drug repurposing candidates targeting three viral proteins (3CL viral protease, NSP15 endoribonuclease, and NSP12 RNA-dependent RNA polymerase), which included rutin, dexamethasone, and vemurafenib among others. To our knowledge, it is the first time that a TDA based strategy has been used to compare a massive amount of protein structures with the final objective of performing drug repurposing..

Keywords: *covid-19, drug repurposing, topological data analysis, persistent homology*

*Corresponding authors, afalco@uchceu.es , joan.climentbataller@uchceu.es
[1]Equally contributed.

 doi:10.20944/preprints202012.0281.v1

1. Introduction

On March 11, 2020, the World Health Organization (WHO) declared the Coronavirus Disease 2019 (COVID-19) outbreak, produced by the novel SARS-CoV-2 virus, a global pandemic [1]. So far, three previously approved antiviral drugs and one antimalarial medication (remdesevir, lopinavir, Interferon-β1, and hydroxychloroquine) have been tested for efficacy against SARS-CoV-2 infection by the WHO SOLIDARITY consortium in a large multicentric study. The results of the trial suggested that these treatments had little or no effect in a set of clinical outcomes which included overall mortality, time to initiation of mechanical ventilation, and duration of hospital stay [2].

With the second wave ongoing in many countries, herd immunity far down the road, and no date scheduled for the release of an effective vaccine, it is still a pressing need to find adequate treatments for the disease. De novo drug development and testing, including preclinical research and clinical trials, is a slow process that could take more than 12 years ([3],[4]). However, the current sanitary emergency makes it imperative to shorten this time frame. Therefore, sustained efforts to identify potential candidates for drug repurposing are necessary.

In the context of COVID-19, Kumar and co-workers [5] compiled sets of genes linked to the disorder and studied their distribution in the human interactome. They first identified the interactome subnetworks' hub genes in which the disease-related genes were placed. Then, they queried the drug-gene interaction database [6] [7] to identify FDA approved drugs which had the hub genes as their target (i.e., chloroquine, lenalidomide, pentoxifylline). Zhou and collaborators compiled a list of human proteins that physically interact with four previous human coronaviruses (SARS-CoV, MERS-CoV, HCoV-229E, and HCoV-NL63) and used network proximity measures to prioritize 16 potential anti-human coronavirus repurposable drugs including melatonin, mercaptopurine, and sirolimus [8]. Virtual screening studies based on molecular docking approaches have also been reported. To cite an example, Kerestsu et al. used a protease inhibitors database (MEROSP) and the geometric structure of the 3C-Like virus protease ($3CL^{pro}$) to identified 15 potential inhibitors using the surflex-Dock software [9].

Here we present a general-purpose drug repositioning workflow and its application to the specific case of COVID-19. Our procedure is based on recent developments in the field of Topological Data Analysis and its use in the study of biological geometric structures [10].

Particularly, our method relies on the idea that drugs that are known to target a specific protein would likely target other proteins that present high degrees of topological similarities with the first. Therefore, the accumulated knowledge of drug-protein interactions available in public repositories such as DrugBank in combination with the information about protein geometric structures found in the Protein Data Bank (PDB) can be used to predict new potential drug protein targets based on the computation of protein-protein topological similarities. Figure 1 contains a brief summary of the general methodology.

Following this principle, we aimed to identify candidate repurposable drugs to target SARS-CoV-2 proteins. To this end, first, we retrieved information about all FDA approved drugs and their protein targets from DrugBank [11] and the available geometric structures of the target proteins, as well as the SARS-CoV-2 protein structures, were obtained from the Protein Data Bank (PDB).

Second, for each protein geometric structure, the coordinates of alpha carbons were selected. This is often referred to as the coarse-grained representation of the protein. Then, using persistent homology theory, we generated the barcodes of each protein's coarse representation through the construction of Vietoris-Rips complexes for the first three persistent similarity measures that provide information about the protein's shape. In short, the initial point cloud used to represent a protein in a three-dimensional space is transformed into a set of three numbers, which include information about the topology of the object. A given pair of barcodes can be tested for similarity using appropriate metrics.

Persistent similarity measures were then computed between the SARS-CoV-2 protein structures and the whole set of protein structures which were known targets of FDA approved drugs. Drugs targeting proteins presenting large topological similarity values with SARS-CoV-2 proteins were selected as potential repurposing candidates and tested in a further twofold *in silico* validation step.

To validate the findings *in silico*, first, drugs selected in the previous step were subjected to blind docking using both the predicted target protein and drug three-dimensional structures and the binding energies of the multiple detected pockets were computed. Second, we carried out searches for transcriptomic studies including samples infected with SARS-CoV-2 and uninfected controls, and obtained the gene expression signatures of SARS-CoV-2 infection by differential expression analysis. Then, the SARS-CoV-2 infection signatures were compared to the transcriptomic profiles produced by treatment with FDA approved drugs generated by the LINCS L1000 team.

2. Results

2.1. Validation of the persistence similarity function

Before applying our pipeline to the identification of potentially repurposable drugs for the treatment of COVID-19, we tested the capacity of the persistent similarity measure to identify proteins with closely related three-dimensional structures. To this end, we retrieved information from two curated protein classifications (i.e., the Skolnick's dataset and a random sampling of 500 protein structures derived from ten different subfamilies of the SCOPe database). Then, we tested the capacity of our method to reproduce them correctly by computing all possible pairwise persistence similarity measures between the structures included in each dataset.

In Skolnicks's dataset, the pairwise similarity matrix generated by our TDA workflow was able to cluster together those proteins belonging to the same structural families. Figure 2 A and B provide a principal component and heatmap representation of the analysis, respectively. In addition, the persistence similarity measures were also found to accurately group the proteins from the SCOPe dataset based on their protein superfamily membership (Figure 2 E and F). The similarity values of protein pairs belonging to the same superfamily tended to be significantly higher than those belonging to different superfamilies (t-test p-value $< 2.2e-16$) (Figure 2 D). We evaluated the protein superfamily classification performance by computing the receiver operating characteristic curve (ROC) and the area under the ROC curve (AUC). An AUC value of $0.951(0.949-0.953)$ was observed (Figure 2 H), indicating that the persistence similarity-based metric employed is a reliable measure to determine similarities between proteic three-dimensional structures.

Furthermore, we tested the ability of our method to identify potential drugs for the treatment of the Cytomegalovirus (CMV) infection using as an input the three-dimensional structure of the viral DNA polymerase of the CMV UL44 (PDB ID: 1T6L). After computing persistent similarity measures between 1T6L and all the protein structures present in our database, we identified only one additional structure (PDB ID: 1YYP) displaying a mean similarity score higher than 0.9 (Figure 3 A). 1YYP was found to contain the three-dimensional structure of the DNA polymerase of the cytomegalovirus UL44 bound to the C-terminal peptide from the CMV UL54. The persistent similarity mean between 1T6L and 1YYP was found to be 0.905, with k-persistent similarity measure presenting values of 0.926, 0.965, and 0.823, for $k = 0, 1, 2$ respectively, (Figure 3 B). 1YYP was linked to two FDA approved drugs in our database (Table 1), Cidofovir and Foscarnet. Cidofovir acts as a viral DNA polymerase inhibitor and is used for the treatment of CMV infection [12], whereas Foscarnet is indicated to treat the retinitis produced by CMV infection in VIH patients as well as the herpes simplex virus infection of the skin and the mucosal membranes. Therefore, our method allows the indirect identification of potential drug candidates that could target specific protein structure by the comparison of the barcodes obtained from the three-dimensional structure of the problem protein to a database of barcodes derived from the three-dimensional structures of a large array of proteins that are known targets of FDA approved drugs.

2.2. Transcriptomic data analysis results

Differential gene expression analyses were carried out with the three identified datasets including samples infected with SARS-CoV-2 and uninfected controls, and were followed by Enrichment and LINCS L1000 analyses. Differential gene expression analysis of **DS1** yielded 451 deregulated genes (DEGs), of which 213 were found to be upregulated, and 238 were downregulated in SARS-CoV-2 infected samples compared to controls. The top upregulated genes were derived from the virus open reading frames. Gene Set Enrichment Analysis (GSEA) showed that pathways linked to the immune response were heavily upregulated in SARS-CoV-2 infected samples. Instances of such pathways included immune response mediated by circulating immunoglobulin (p-adj $= 1.8e-25$), B-cell mediated immunity, (p-adj $= 3.2e-22$), and adaptive immune response (p-adj $= 2.0e-20$). The FDA approved drugs showing the strongest negative correlation in LINCS L1000 analysis were niclosamide, bisacodyl, and perhexiline (r $= -0,21$, -0.19, -0.18). GSEA analysis of the transcriptomic signatures produced by those medications suggested that they induce significant gene expression changes in pathways linked to interleukin signaling and NF-kB activation. Genes included in the set of potential therapeutics for SARS were also found to be upregulated in the bisacodyl signature (NES $=$ 1.61, p-adj $= 2.19e-02$) . The JAK-STAT complex and the TCF dependent signaling pathways were found to be downregulated in the perhexiline and niclosamide signatures, respectively.

Eight thousand three hundred and eighty DEGs were identified in the **DS2** analysis. Four thousand six hundred and six genes were found to be upregulated, and 3744 were found to be downregulated in SARS-CoV-2 infected samples compared to uninfected controls. Upregulated genes were enriched in components of the humoral immune response, epidermis development, keratinization, and B-cell mediated immunity among others (p-adj $= 1.1e-20, 8.2e-20, 1.3e-18, 2.5e-10$). The top negatively correlated drugs included instances of several different compound families such as anti-inflammatories (phenylbutazone, r $= -0.21$), antidiabetics (troglitazone, r $= -0.20$), antimalarials (chloroquine, r $= -0.20$), and other compounds such as nicotine (r $= -0.17$). Treatment with phenylbutazone was found to upregulate the gene expression of genes included in the interleukin-12 and 17 signaling pathways. In contrast, interleukin-4 and 13 signaling related genes tended to be downregulated by chloroquine treatment (NES $= -1.45$, p-adj $= 4.30e-02$). Genes involved in the viral mRNA translation and the ISG15 antiviral mechanism were also upregulated in the gene expression profiles induced by treatment with chloroquine, phenylbutazone, and troglitazone. In addition, the SARS-CoV infection pathway was found to be upregulated in samples treated by chloroquine and troglitazone. ADORA2B mediated anti-inflammatory cytokine production-related genes was downregulated by the treatment of the three top negatively correlated drugs.

DS3 presented the lowest yield in terms of differentially expressed genes. One hundred and eighty-eight genes were found to be upregulated to controls, whereas 31 genes were found to be downregulated in infected samples compared to controls. Twenty-nine biological processes were found to be significantly upregulated and were mainly linked to mechanisms aimed to fight the viral infection and immune system-related processes including, defense response to virus (p-adj $= 7.2e-13$), myeloid leukocyte mediated immunity (p-adj $= 8.8e-15$), regulation of cytokine production (p-adj $= 1.5e-08$), and response to interferon-gamma (p-adj $= 1.9e-08$) among others. Chloroquine was found to be the top negatively correlated drug (r $= -0.11$), followed by others such as pazopanib, spectinomycin, and troglitazone (r $= -0.11$, -0.11, -0.10). The correlations observed in this dataset tended to be weaker than those computed for **DS1** and **DS2**. GSEA analyses of the drug signatures showed that troglitazone increased the expression of genes classified as potential therapeutics for SARS (NES $= 1.46$, p-adj $= 4.65e-02$), as well as antiviral pathways such as the ISG15 and IFN-stimulated antiviral mechanisms. Spectinomycin was found to reduce the expression of interferon-gamma signaling and interleukin 2, 3, and 5 pathways related genes, whereas pazopanib was found to upregulate viral related pathways such as viral mRNA translation influenza and SARS-CoV-2 infection. Supplementary File 1 includes the complete differential gene expression and enrichment analysis results for transcriptomic datasets 1, 2, and 3, whereas Supplementary File 2 contains the full LINCS L1000 analysis information.

2.3. Drugs, protein targets, and PDB structures included in this study

DrugBank queries yielded 1825 medications approved by the American Food and Drug Administration (FDA). The identified drugs had 1821 known unique protein targets, for which 27839 tridimensional structures were available in the protein databank. Barcodes associated with the first three persistent similarity measures were successfully calculated for 25800 out of the 27839 structures, whereas computational limitations prevented us from estimating the remaining 1622 structures' barcodes. We also retrieved multiple protein structures from SARS-CoV-2 that were available in PDB, including the Spike protein receptor-binding domain, the RNA-dependent RNA polymerase (NSP12), the endoribonuclease (NSP15), the ADP ribose phosphatase (NSP3), the RNA binding protein (NSP9), the 3C-like protease, and the NSP 8 and 7. In total, we calculated the barcodes of 23 viral protein structures. Table 2 shows the complete information regarding the included SARS-CoV-2 protein structures.

2.4. TDA results, viral proteins above 0.9 similarity with structures with known FDA-approved drugs.

We compared twenty-three PDB structures derived from SARS-CoV-2 with 25800 structures belonging to proteins that are known targets of FDA approved drugs through the computation of 593400 persistent similarity measures. Based on the results of the Skolnick's, SCOPe, and cytomegalovirus analyses, we selected a stringent mean similarity persistent threshold of 0.9 in order to call two protein structures similar. Three viral structures, the 3CL protease (6M2Q), the RNA-dependent RNA polymerase (6M71), and the NSP15 endoribonuclease (6W01), presented persistent similarity mean values higher than the selected threshold with proteins known to be targeted by approved drugs. The 3CL protease was found to be associated with 284 PDB structures (Supplementary table 1), most of them classified as Aldo/Keto reductases and protein kinases, which were targeted by 55 different pharmacological compounds (Supplementary Table 2). The RNA-dependent RNA polymerase was found to be significantly associated with 361 PDB structures (Supplementary tables 3), which in many cases belonged to the protein kinase and flavin-containing oxidoreductase families, and that were found to be targeted by 204 unique drugs (Supplementary Table 4). Finally, the viral NSP15 endoribonuclease presented topological similarity values higher than 0.9 with 13 PDB structures (Supplementary table 5), where the most abundant group was the poly(Adp-RIbose) Polymerase Catalytic Domain. These structures were targeted by 45 drugs (Supplementary table 6).

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 11 December 2020 doi:10.20944/preprints202012.0281.v1

Drugs known to target those proteins presenting high topological similarity values with the SARS-CoV-2 structures were subjected to blind docking with the viral proteins. A set of potential repurposable candidates was then selected based on the topological similarity criteria, the transcriptomic effects that exert, and the binding energies derived from the blind docking analyses. Therefore, the selected candidates are known to target proteins with high topological similarity with a specific viral protein, present high affinities with the viral structures, and have the capacity to partially revert the transcriptomic effects induced by the viral infection. The full description of the candidates can be consulted in Table 3.

We identified six repurposable candidates to target the 3CL viral protease. Cholic acid, an amphipathic sterol, presented the strongest binding energies (BE = -15.06 $kcal/mol$), and was found to negatively correlate with transcriptomic dataset 2 (**DS2** r $= -0.11$). Rutin (BE = -14.52 $kcal/mol$, **DS2** r $= -0.184$ DS3 r $= -0.1$), a flavonoid-3-o-glycoside with known antioxidant and cytoprotective activity was also selected ([13], [14]). Two non-steroidal anti-inflammatory drugs , indomethacin (BE = -13.31 $kcal/mol$, **DS2** r $= -0.12$) and sulindac (BE = -13.14 $kcal/mol$, **DS2** r $= -0.12$) were also identified. Whereas indomethacin presents antipyretic and analgesic properties ([15]), sulindac is used to treat conditions that involve chronic inflammation, such as arthritis [16]. Finally, sulfisoxazole (BE = -11,59 $kcal/mol$ **DS2** r $= -0.13$), a sulfanilamide used as a broad-spectrum antibiotic, and dasatinib (BE = -10.94 $kcal/mol$ **DS2** r $= -0.15$), a tyrosine kinase inhibitor indicated for the treatment of chronic myeloid leukaemia [17], were also identified as drugs with the potential of targeting the viral 3CL protease.

Five compounds were found to be candidates to target the SARS-CoV-2 NSP15 endoribonuclease, which included two corticosteroids, dexamethasone (BE = -11.42 $kcal/mol$, DS2 r $= -0.15$) and spironolactone (BE = -10.99 $kcal/mol$, **DS1** r $= -0.12$ and **DS2** r $= -0.1$) which are indicated for the treatment of allergies and asthma [18] and resistant hypertension [16] [19] respectively. Phenolphthalein (BE = -11.15 $kcal/mol$, **DS1** r $= -0.13$), a compound historically used as a laxative [20]. Mifepristone (BE = -10.04 $kcal/mol$, **DS1** r $= -0.13$, **DS2** r $= -0.14$) ,a synthetic steroid progesterone antagonist drug that is indicated for Cushing's syndrome [21] and is also used as an emergency contraceptive pill [22]. Finally, Carbamazepine (BE = -9.66 $kcal/mol$, **DS2** r $= -0.147$) is a pharmacologically active molecule related to the group of tricyclic antidepressants, mainly used as anticonvulsant. ([16],[23]).

Lastly, the analysis of the NSP12 RNA-dependent RNA polymerase yielded multiple antineoplastic drugs as possible repurposing candidates: Vemurafenib (BE = -8.09 $kcal/mol$ **DS2** r $= -0.16$) , a BRAF inhibitor ([24], [25]), Sorafenib (BE = -7.34 $kcal/mol$ **DS1** r $= -0.11$, **DS2** r $= -0.15$), a multitarget protein kinase inhibitor [26], Levonorgestrel (BE = -7,21 $kcal/mol$, **DS2** r $= -0,14$), a synthetic progestogen used as a first-line oral emergency contractive pill [16], The opioid antagonist naloxone (BE = -7.07 $kcal/mol$, **DS2** r $= -0.11$) and, raloxifene (BE = -7.05 $kcal/mol$, **DS1** r $= -0.13$ and **DS2** r $= -0.17$), a selective estrogen receptor modulator mainly used to treat osteoporosis in postmenopausal women and avoid bone loss [27].

 doi:10.20944/preprints202012.0281.v1

2.5. GSEA analysis of the repurposing candidates

We determined the transcriptomic impact of the treatment with the selected candidates on two sets of biological processes linked to COVID-19, viral infections, and immune-related pathways by performing Gene Set Enrichment Analysis (GSEA) of their gene expression signatures derived from LINCS L1000. The transcriptomic profiles generated by cholic acid, rutin, sulfafurazole and sulindac treatment (candidates to target the 3CL protease) were found to be enriched in the ISG15 antiviral mechanism. Furthermore, genes related to interleukin-1 and 12 signaling tended to be upregulated in rutin's signature, as well as genes belonging to the Potential therapeutics for SARS gene set (NES = 1.51, p-adj = $3.85e-2$) whereas WNT ligen biogenesis and trafficking (NES) genes were found to be downregulated by rutin treatment (NES = -1.99 , p-adj = $2.12e-3$) (Supplementary table 7). RNA-dependent RNA polymerase drug candidates, levonorgestrel and raloxifene, were found to be enriched in pathways related to antiviral processes such as ISG15 antiviral mechanism (levonorgestrel , NES = 2.08, p-adj = $9.95e-4$, raloxifene, NES = 2.06, p-adj = $8.13e-4$) and antiviral mechanism by IFN-stimulated genes (levonorgestrel , NES = 1.95, p-adj = $1.22e-3$, raloxifene, NES = 1.94, p-adj = $1.12e-3$). In addition interferon alpha/beta signaling was observed to be depleted in raloxifene treated cells (NES = -1.52, p-adj = $4.59e-2$) (Supplementary table 8). Finally, in the case of NSP15 endoribonuclease candidate drugs, dexamethasone produced gene expression signatures upregulated in pathways associated with viral infection response, such as ISG15 antiviral mechanism (NES = 1.82, p-adj = $3.17e-3$) and the antiviral mechanism by IFN-stimulated genes (NES = 1.59, p-adj = $1.20e-2$). This pathway was also found to be upregulated in the gene expression profiles of carbamazepine and mifepristone. Finally, interleukin-7 signaling (NES = -1.64, p-adj = $3.47e-2$) and interferon alpha/beta signaling (NES = -1.68, p-adj = $5.48e-3$) were downregulated by dexamethasone treatment (Supplementary table 9). Figure 4 shows a dot plot representation of the GSEA analysis results. Supplementary tables 14, 15 and 16 shown the GSEA results in detail.

3. Discussion

On 31st December 2019, the World Health Organization (WHO) was officially notified about several cases of pneumonia in Wuhan City, China, caused by the COVID-19, a disease with no effective treatment nor specific vaccine. A disease, which history and quest for a cure is a daily struggle and is constantly being rewritten. Because specific antiviral treatments and vaccines are still under development, drug repurposing strategies suggesting the use of FDA approved drugs for other conditions quickly became the only option to treat COVID-19. However, to date no therapeutic agents have yet been proven effective. Several treatments have been currently reported under investigation specifically to treat COVID-19 as the result of drug repurposing strategies [28] [29] [28] and, as this draft is being written, up to 700 research papers have already

 doi:10.20944/preprints202012.0281.v1

been published. The number of clinical trials using repurposed drugs such as hydroxychloroquine, remdesivir, lopinavir/ritonavir among others, alone or in combination, is also exponentially growing, although in most cases unfortunately the results are not as good as initially expected [30, 31, 32].

Here, we report a based TDA novel strategy for drug repurposing in combination with current methodologies of molecular docking, differential expression analysis of SARS-CoV-2 infected cells and correlation with FDA approved drugs transcriptomic profiles. Our results indicate that the proposed TDA based formalism is an excellent tool to address biological problems from a dual perspective. In the first place, from a structural biology point of view, we use of Vietoris-Rips complex to compute the barcodes encoding the shape of each protein structure. Next in combination with its Persistent Betti Functions, we transform individual barcodes into one-dimensional functions to measure a degree of similarity between proteins. It allowed us to classify proteins based solely in the Cα atomic coordinates. Persistent homology has been previously proposed as a method to study the topological invariants of the three-dimensional structure of biomolecules. Several studies have employed use TDA-based methods to classify of protein structures using only the three-dimensional coordinates of the atoms from crystallographic resolved proteins. For instance, Xia and collaborators [33] performed persistence homology analysis of three-dimensional biomolecular structures in order to study their structural characteristics, flexibility prediction, and folding properties. Hence, they define the molecular topological fingerprints (MFTs) to extract the topological information from protein structures using persistent Betti numbers [34]. K. Dey and colleagues proposed another topology-based method to create protein signatures to create a fast domain classifier using a support vector machine [35]. Interestingly, our mean persistence similarity metric was able to achieve results comparable to those obtained by the state of the art structural alignment method, DALI [36], and presented a high predictive power clustering proteins in terms of external classifications.

Molecular docking simulation is a rapid screening method to test compounds binding activity and transcriptomic data represent a very rich alternative resource for inferring non-obvious relationships between drugs and genes. Previous *in silico* molecular docking studies have highlighted the potential of repurposed drugs for the treatment of COVID-19 [37, 38, 39, 40, 41, 42, 43]. Yet, here we used “in silico” molecular docking combined with transcriptomic small molecule treatment data from LINCS L1000 to determine which FDA approved drugs may reverse the effects or SARS-CoV-2 infection. The gene expression profiles in response to the identified drugs support the docking results and offer a plausible perspective on the pathways associated with protein responses to drug binding SARS-CoV-2 proteins. To our knowledge, this is the first time that an application of barcode-based similarity measures is used for the analysis of large datasets of PDB structures.The generation of barcodes depends upon the previous construction of Vietoris-Ripps complexes, which have a computational cost that scales exponentially with the number of points defining a particular structure. Although our analyses were carried out in a cluster with 32 cores and up to 500 GB of

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 11 December 2020 doi:10.20944/preprints202012.0281.v1

RAM, the computational cost of the barcode generation of the excluded 1622 exceeded the available amount of RAM or was not possible to end finite time.

Among all the SARS-Cov-2 proteins analyzed (n=23, Table 2), only three showed a persistent similarity score above 0.9 against other protein structures targeted with known drugs. Interestingly, these proteins are key components in coronavirus replication and structural assembly: The Viral 3CL protease (6M2Q),a chymotrypsin-like protease that is essential for the production of non-structural proteins [44]; the nsp12 RNA-dependent RNA polymerase (6M71), the main component of coronavirus replication and transcription machinery, and because of that an excellent target for new therapeutics [45] and the nsp15 endoribonuclease (6W01), a protein a with a poorly defined role in SARS-CoV-2 infection but has been described to be linked to pRB downregulation affecting host cell cycle division and coronavirus infection [46] in other coronaviruses (SARS-CoV) but also with a role as an antagonist of host dsRNA sensors during coronavirus infection in macrophages to evade innate immune system defenses [47]. Hence, in this study, we select three proteins from the SARS-CoV-2 coronavirus as the best candidates to find repurposed drugs to combat the disease.

Our differential expression analyses revealed that troglitazone, niclosamide and chloroquine, among multiple candidates, were the top negatively correlated drugs that may revert the effects of the SARS-CoV-2 infection to the cell transcriptome. Moreover, chloroquine is already under study in several clinical trials, although recent results reported by the WHO SOLIDARITY study stated that chloroquine has no significant effect on hospitalized COVID-19 patients, in an overall mortality level [2]. Niclosamide is also being evaluated under a Phase 2 clinical trial [48]. In addition, the antiviral activity of the niclosamide has been demonstrated against SARS-CoV *in vitro* studies [49] and recent investigations against SARS-CoV-2 [50], and also previously against other MERS coronavirus [51].

Our more promising candidates arise from the combination of molecular docking and transcriptomic results, and the cornerstone of our work, the TDA-based formalism. Among the 16 compounds related to the three SARS-CoV-2 proteins analyzed, 9 have been described as possible candidates from other repurposing studies and 5 of them have already shown antiviral activity or already have been described as possible COVID-19 treatments (Supplementary table 10) although preclinical studies will be required to determine their efficacy. In this direction three out of the sixteen compounds are being evaluated under different clinical trials (Indomethacin (n=2), Dexamethasone (n=40) and Spironolactone(n=4)).

Rutin and Indomethacin were amongst the notable compounds selected from 3CL main protease. Besides, they have been proven as good candidates in other studies. Rutin is a polyphenolic flavonoid that has shown a wide range of pharmacological applications due to its significant antioxidant properties [52]. Our results from GSEA analyses revealed that rutin might act in early stages of SARS-COV-2 infection by activating the Interferon-induced ISG15 pathway. ISG15 is an interferon-induced protein that has been implicated as a central player in the host antiviral response, and it is the key element for the innate immune response against viral infection [53]. Besides, ISG15 modulates the immune system stimulating the IFN-gamma production by NK cells that lead to the promotion of early viral response [54]. Although the result of the possible interaction between rutin and 3CL protease has been reported by other studies using an in silico approach [55], our results provide a transcriptomic dimension to the possible effect of rutin during infection with SARS-COV-2. Moreover, to our knowledge this is the first time the natural compound rutin is related with the antiviral activity induced by the protein ISG15.

Dexamethasone, a corticosteroid used in a wide range of conditions for its anti-inflammatory and immunosuppressive effects, could be one of the most promising repurposed drugs chosen to treat COVID-19 disease, based on some results that proven a decrease on the incidence of death versus the usual care group among patients receiving invasive mechanical ventilation [56]. This compound was chosen because of its properties as immunosuppressant to treat the cytokine storm induced by the immune response to coronavirus infection in late stages of the disease. Nonetheless, our results, indicated that dexamethasone could also be a good candidate to target nsp15 endoribonuclease, although some repurposed works also suggested it as the target of the main protease [57]. That data could support the idea of giving corticosteroids not just in the advanced infection stage but also at the beginning, however a recent study [58] tested multiple pharmacological compounds derived from the steroids in vitro and demonstrated that dexamethasone has no antiviral activity against SARS-CoV-2. Nevertheless, other corticosteroids we also found that could interact with nsp15 protein, such as mifepristone suppressed viral growth conferring more than 95% of cell survival rate after viral infection and drug administration in vitro [58].

Lastly, the RNA-dependent RNA polymerase nsp12 of SARS-CoV-2 is a protein that performs essential functions in the coronavirus life cycle with no host cell homolog. This gives an advantage for antiviral drug development, reducing the risk of affecting any protein in human cells as it has been proven by many drug repurposing studies directed against nsp12 RdRP [59, 60, 61, 62]. Vemurafenib, sorafenib and raloxifene may be potential candidates against nsp12 RdRP. Vemurafenib can disturb the cellular Raf/MEK/ERK signaling cascade via binding in the ATP-binding site of BRAF(V600E) kinase and inhibiting its function [63], whereas sorafenib is another kinase inhibitor that targets VEGFR, PDGFR, and RAF kinases [64]. Interestingly, SARS-CoV-1 uses Raf/MEK/ERK signaling pathways to promote its replication via various mechanisms [65, 66, 67] presenting this signaling cascade as a critical therapeutic target for host-directed SARS-CoV-2 antivirals.

In conclusion, our strategy on Quantitative Homological Similarities through TDA based formalism would allow researchers and clinicians to select optimal candidates from drug repurposing to hit in the bullseye not only of SARS-CoV-2 coronavirus but also any new viruses that may appear in the future, by choosing the best targets among all virus proteins. In this specific case, by targeting nsp15 endonuclease and nsp12 RNA polymerase, in addition to other promising drug targets of the 3CL main protease, could support the development of a cocktail of anti-coronavirus treatments that could also be potentially used for the discovery of broad-spectrum antivirals. Furthermore, by choosing a precision multidrug treatment, we could rescue any specific drug failure or avoid any future drug resistance due to possible acquired mutation in any of the proteins as a consequence of continuous virus replication and spreading, since the virus will be attacked from different fronts. Nevertheless, our results based on multidrug combinations should be validated both in vitro and in vivo experiments not just to prove the effectiveness of the treatment but also to select the best combination against SARS-Cov-2 infection and consequent disease symptoms.

4. Material and Methods

4.1. Data obtention

DrugBank queries were carried out [11] to retrieve the information regarding medications with known protein targets. In short, the DrugBank database version 5.1.5 was downloaded in XML format, and the dbparser package [68] and custom R scripts were employed to extract the relevant information. We only selected drugs approved by the American Food and Drug Administration (FDA) and retrieved the names and UniProt identifiers of their protein targets. Then, UniProt IDs were mapped to their respective Protein Data Bank (PDB) structures using the Retrieve/ID mapping tool available at uniprot. All the PDB structures targeted by FDA approved drugs were downloaded in PDB format and stored for downstream analysis. Protein Data Bank queries were also performed to identify the three-dimensional structures of SARS-Cov-2 proteins.

4.2. Data preparation and barcodes computation

All protein structures in PDB format were loaded into the R's environment using the bio3d package [69]. Then, the coarse-grain representation of each structure was generated by selecting only the tridimensional atomic coordinates of the alpha-carbons of the amino acids. Two main reasons compelled us to work with this reduced representation. First, the construction of Vietoris-Rips complexes scales exponentially with the number of initial points present in the point cloud. Therefore, structures defined by a very large amount of points are not computationally tractable even in state of the art computers. Second, all-atom models present a high degree of detail that could mask the general structure of the protein. Cang and co-workers compared the all-atom and the coarse-grain model representations of the M2 channel protein of the influenza A virus and observed that the all-atom model contained too many details which masked useful information, such as, the Betti 1 barcodes indicating alpha helix structures [70]. Barcodes were constructed using the R package of TDAstats [71]. TDAstats makes use internally of Ripser C++ library [72], an optimized fast software for Vietoris-Rips computation and barcode construction.

4.3. Validation of the Betti persistence similarity function

Two independent datasets were used to test the ability of the persistence Betty function to cluster protein based on previous structural classifications. The First, termed as the Skolnick dataset, includes a collection of manually curated domains from the Catalytic Site Atlas (CSA) [73]. In particular, 40 proteins are classified into four structural catalytic families, including Flavodoxin-like fold CheY-related, Plastocyanin, TIM Barrel, and Ferratin [74] [75]. To construct the second dataset we carried out a random sampling of 500 protein structures derived from 10 different protein superfamilies from SCOPe [76]. Supplementary Table (PONER TABLA) contains the information regarding the superfamilies and protein structures selected for SCOPe analysis. In short, barcodes were computed for all the involved protein structures and pairwise similarity matrices were computed using the Betti persistent similarity function. The Human cytomegalovirus (HCMV) is a widespread pathogen that belongs to the subfamily of the beta-herpesvirus. Whereas HCMV often generates persistent asymptomatic infections in healthy people, it can produce severe complications in immunosuppressed individuals [77]. Cidofovir, a nucleotide analogue that inhibits the viral DNA polymerase, is used to treat patients with severe HCMV infection [12]. To test if our TDA-based drug repurposing strategy was able to identify potential medications for the treatment of HCMV infection, we retrieved the crystallographic structure of the HCMV DNA polymerase from PDB (PDB ID: 1T6L [78]) and computed persistent Betti similarities to all other protein structures that are known targets of FDA approved drugs.

4.4. Protein-ligand binding with autodock 4.2

Ligand preparation was carried out as follows: First, the FDA-approved drugs in SDF format were retrieved from DrugBank. A custom R script and Open Babel v.3.0.0 [79] were used to transform the SDF into the mol2 format. Then, the MGLTools v.1.5.7 toolkit was employed to add the polar hydrogens and protonation at pH 7.4. mol2 drug structures were converted into PDBQT format, and their stereochemical properties were computed using Autodock 4.2 [80]. A virtual screening library was then constructed using the preprocessed drug structures. Drugs containing atoms different from the ones included in the following list (H, C, N, O, F, Mg, P, S, Cl, Ca, Mn, Fe, Zn, Br, I) were discarded from the subsequent analyses since Autodock does not include the values of their atomic force fields, and it is, therefore, unable to perform molecular docking using them. Polar hydrogens were also added to the SARS-CoV-2 protein pdb structures which were also transformed to the PDBQT format. Docking was carried out using Autodock 4.2 [80], a molecular docking software developed by the Scripps Research Institute. A grid box spanning the whole protein structure was set in order to perform blind docking. Autodock was configured following the manual recommendations [81]. We increased the parameter ga_runs from 10 to 150 to improve the accuracy of the results.

4.5. Differential gene expression analyses of SARS-CoV-2 infected human samples and cell lines and uninfected controls.

We carried out searches for transcriptomic datasets of patients and human-derived cell lines including samples infected with SARS-CoV-2 and uninfected controls. At the time the searches were carried out, three datasets were identified. Dataset 1 (**DS1**) was found in gene expression omnibus (GEO) under ID GSE150316 [82]. It includes formalin-fixed paraffin-embedded samples from multiple tissues (i.e., lung, jejunum, heart) derived from SARS-CoV-2 infected individuals and uninfected controls obtained in autopsies. We restricted our analysis to lung samples. Twenty-one samples (16 cases and five controls) were selected for downstream analysis.

Dataset 2 (**DS2**) [83] gathers samples derived from bronchoalveolar lavage fluids (BALF) of SARS-CoV-2 infected patients (four samples derived from two patients with two technical replicates) and three healthy controls. Samples derived from infected patients were stored at National Genomics Data Center under accession number CRA002390, whereas control samples were downloaded from the NCBI SRA database and were available under the following identifiers SRR10571724, SRR10571730, and SRR10571732. Sequence alignment using the human reference genome hGR38 and count extraction were carried out using the Rsubread package [84].

Finally, the third dataset (**DS3**) was available in GEO under accession ID GSE147507 [85]. It presented a complex design including both primary cell lines derived from the human lung epithelium and transformed lung alveolar which were either mocked treated or infected with different viruses including the influenza A virus (IAV), the respiratory syncytial virus (RSV), and SARS-CoV-2, as well as, samples derived from infected ferrets and two technical replicates of a lung sample derived from a SARS-CoV-2 infected human patient. We restricted our analysis to the cell lines NHBE, A549, and Calu-3, which were either infected with SARS-CoV-2 or were mock treated. The infected human lung samples and the healthy lung biopsies were also included. Overall, twenty-eight samples were analyzed in this dataset.

For each dataset, differential gene expression analysis between SARS-CoV-2 infected samples and uninfected controls carried out using the DESeq2 package [86].

4.6. Identification of LINCS L1000 signatures negatively correlated with the SARS-CoV-2 differential gene expression profiles

LINCS L1000 [87] contains an extensive collection of gene expression profiles generated using thousands of perturbagens (i.e., small molecules, ligands, micro-environments, CRISPR gene over-expression, and knockdown perturbations) and different cell lines, doses, and exposure times. In particular, LINCS L1000 Level 5 data includes differential gene expression signatures computed by comparing three technical replicates of the same perturbation to appropriate controls. Level 5 LINCS L1000 phases I (GSE92742) and II (GSE70138) datasets were downloaded from GEO. Signatures involving FDA approved drugs were identified with the help of the information contained in file *repurposing_drugs_20180907.txt* and *repurposing_samples_20180907.txt* available at LINCS L1000 repurposing hub [87]. Drugbank and LINCS 1000 data were merged based on Pubchem compound identifiers. Then, the subset of signatures corresponding to FDR approved medications with known Pubchem identifiers were selected. Overall we obtained 52144 expression signatures generated using 1313 approved drugs. To identify drugs with the potential of reverting the differential expression profiles generated by SARS-CoV-2 infection, we computed Pearson's correlations between each expression signature derived from LINCS L1000 and the differential expression profiles from **DS1**, **DS2**, and **DS3**, and picked those drugs exhibiting the most negative correlations.

4.7. Gene Set Enrichment Analysis (GSEA)

Dysregulated biological processes were identified for each transcriptomic dataset using the pre-ranked Gene Set Enrichment Analysis (GSEA) implementation of fgsea package [88]. The C5 molecular signatures collection, which contains gene sets derived from the three branches of Gene Ontology (GO) was used as a source of functional information. GO terms including more than 500 or less than 15 genes were filtered out. GSEA analyses were also performed for those LINCS L1000 level 5 expression signatures negatively correlated with the differential gene expression profiles generated by the SARS-CoV-2 infection to determine their effect in specific pathways and biological processes. Reactome (version 73) was used as a source of pathway information and analyses were carried out using the clusterprofiler package [89]. Biological processes and pathways presenting false discovery rate (FDR) adjusted p-values were called to be significantly dysregulated.

4.8. A Topological Data Analysis based formalism to compare, at quantitative level, the homological similarities of pairwise three-dimensional molecules considered as surfaces

In this section we borrow well-known concepts for Algebraic Topology [90] and Persistent Homology [91] (see also [92] and references therein) to introduce a mathematical formalism that allows us to compare three-dimensional protein structures considered as surfaces. This fact implies that only consider a shape composed by a union of two-dimensional faces each of them composed by one-dimensional segments, that are constructed by a finite set of zero-dimensional points contained in a three-dimensional space. Intuitively, we assume that a molecule is a kind of graph embedded in a three-dimensional space and we take into account the path following by the molecule to be folded.

Simplicial complexes for three-dimensional molecules considered as surfaces

Throughout this paper we will identify a molecule $\mathcal{M}$ with a finite set of 3-dimensional data points denoted by

$$\mathbb{M} = \{x_1, x_2, \ldots, x_M\} \subset \mathbb{R}^3,$$

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 11 December 2020 doi:10.20944/preprints202012.0281.v1

where we assume that M is a high natural number, together a set, denoted by $\mathcal{S}(\mathcal{M})$ containing the information of the molecular combinatorial structures at different scales and its relationships. To describe this kind of molecular geometry we use the so-called k-simplexes ($0 \leq k \leq 2$) defined from the data set $\mathbb{M}$ as follows.

- A 0-simplex, also called vertex, is generated by an individual point $x_0 \in \mathbb{M}$ and we will denote its associated 0-simplex by $[x_0]$.
- A 1-simplex is generated by two different vertices $[x_0]$ and $[x_1]$ and it is defined as the set

$$[x_0, x_1] := \left\{ z \in \mathbb{R}^3 : z = \lambda x_0 + (1-\lambda) x_1, 0 \leq \lambda \leq 1 \right\},$$

  that is, the linear segment that joins the corresponding two 0-simplexes $[x_0]$ and $[x_1]$

- Finally a 2-simplex is generated by three different vertices $[x_0]$, $[x_1]$ and $[x_2]$ such that $x_1 - x_0$ and $x_2 - x_0$ are linearly independent vectors and it is defined by

$$[x_0, x_1, x_2] := \left\{ z \in \mathbb{R}^3 : z = \lambda_0\, x_0 + \lambda_1\, x_2 + (1 - \lambda_1 - \lambda_2)\, x_3 \text{ where } 0 \leq \lambda_1, \lambda_2 \leq 1 \right\}.$$

Given $k = 1, 2$ and for $0 \leq i \leq k$ we can define the operator R_i that for each k-simplex removes the i-th position:

$$R([x_0, x_1 \ldots, x_k]) := [x_0, \ldots, x_{i-1}, x_{i+1}, \ldots, x_k].$$

In order to have a well-define map from a set of k-simplexes to a set of $k-1$-simplexes, we construct a finite sequence of sets of k-simplexes, denoted by $\mathcal{S}_k(\mathcal{M})$ ($k \in \mathbb{Z}$), obtained as follows.

- $\mathcal{S}_k(\mathcal{M}) = \emptyset$ for every integer number $k \neq 0, 1, 2$; otherwise
- Put $\mathcal{S}_0(\mathcal{M}) := \{[x] : x \in \mathbb{M}\}$, then the simplexes in $\mathcal{S}_k(\mathcal{M})$ for each $k = 1, 2$ are obtained from the simplexes in $\mathcal{S}_{k-1}(\mathcal{M})$ taking into account the following two properties:
  (P1) for every $\sigma \in \mathcal{S}_k(\mathcal{M})$ it holds that $R_i(\sigma) \in \mathcal{S}_{k-1}(\mathcal{M})$ for $0 \leq i \leq k$, and
  (P2) if σ, γ are in $\mathcal{S}_k(\mathcal{M})$ and $\sigma \cap \gamma \neq \emptyset$, then there exists $1 \leq \ell \leq k$ such that $\sigma \cap \gamma \in \mathcal{S}_{k-\ell}(\mathcal{M})$.

As consequence of the above inductive construction of the simplexes we provide the following definition of the face of a simplex.

Definition 4.1. *Let be $\sigma \in \mathcal{S}_k(\mathcal{M})$ for $k = 1, 2$. Then we will say that $\gamma \in \mathcal{S}_{k-\ell}(\mathcal{M})$ for some $1 \leq \ell \leq k$ is a face of σ if there exists a finite sequence $\alpha_0\alpha_1 \cdots \alpha_{\ell-1} \in \{0, 1, \ldots, k\}^\ell$ where $\alpha_i \neq \alpha_j$ such that*

$$\gamma = \left(R_{\alpha_0} \circ \cdots \circ R_{\alpha_{\ell-1}} \right)(\sigma).$$

 doi:10.20944/preprints202012.0281.v1

Since we consider that $x_i \in \mathbb{R}^3$ for $1 \leq i \leq M$, and that $\mathbb{M}$ is a surface, we associate to our molecule $\mathbb{M}$ three non-empty sets of simplexes:

$$\mathcal{S}(\mathcal{M}) = \{\mathcal{S}_0(\mathcal{M}), \mathcal{S}_1(\mathcal{M}), \mathcal{S}_2(\mathcal{M})\},$$

recall that $\mathcal{S}_k(\mathcal{M}) = \emptyset \subset \mathcal{S}(\mathcal{M})$, for every integer number $k \in \mathbb{Z}$, $k \neq 0, 1, 2$. As we will see below it will have some consequences. From the construction of $\mathcal{S}(\mathcal{M})$ the following two properties holds.

(P1) Every face of a simplex $\sigma \in \mathcal{S}(\mathcal{M})$ is also in $\mathcal{S}(\mathcal{M})$.

(P2) Given $\sigma, \gamma \in \mathcal{S}(\mathcal{M})$ either $\sigma \cap \gamma = \emptyset$ or $\sigma \cap \gamma$ is a face of σ and γ.

Definition 4.2. *We say that a three dimensional molecule $\mathcal{M}$ considered as surface is a pair* $(\mathbb{M}, \mathcal{S}(\mathcal{M}))$ *where $\mathbb{M}$ is a set of points of $\mathbb{R}^3$ and $\mathcal{S}(\mathcal{M}) = \{\mathcal{S}_0(\mathcal{M}), \mathcal{S}_1(\mathcal{M}), \mathcal{S}_2(\mathcal{M})\}$ is a collection of simplexes satisfying (P1) and (P2).*

Simplicial homology for three-dimensional molecules considered as surfaces

In order to perform geometric operations (like unions of simplexes) at each k-level and to describe the relationship between two consecutive levels (like cut the faces of a simplex) we endow to each $\mathcal{S}_k(\mathcal{M})$ with an algebraic structure of vector space over a finite field of scalars. To this end, we consider the finite field $\mathbb{Z}_2 = \{\mathbf{0}, \mathbf{1}\}$. We recall that the two operations over $\mathbb{Z}_2$ are the sum and the multiplication:

$+$	**0**	**1**
0	**0**	**1**
1	**1**	**0**

$\cdot$	**0**	**1**
0	**0**	**0**
1	**0**	**1**

Now, for $0 \leq k \leq 3$ we introduce the vector space of formal series of k-simplexes with coefficients over the finite field $\mathbb{Z}_2$ as

$$\mathbb{Z}_2[\mathcal{S}_k(\mathcal{M})] := \left\{ \sigma : \sigma = \sum_{i=1}^{\ell} \eta_i \, \sigma_i \text{ where } \eta_1, \ldots, \eta_\ell \in \mathbb{Z}_2 \text{ and } \sigma_1, \ldots, \sigma_\ell \in \mathcal{S}_k(\mathcal{M}) \right\}.$$

Observe that if $\sigma \in \mathcal{S}_k(\mathcal{M})$ then we can identify this simplex with the formal series also denoted by $\sigma = \mathbf{1}\, \sigma \in \mathbb{Z}_2[\mathcal{S}_k(\mathcal{M})]$. In consequence,

$$\sigma + \sigma = \mathbf{1}\, \sigma + \mathbf{1}\, \sigma = 0,$$

because $\mathbf{1} + \mathbf{1} = \mathbf{0}$ in $\mathbb{Z}_2$. Thus, for a given $\sigma_1, \ldots, \sigma_\ell \in \mathcal{S}_k(\mathcal{M})$ the formal series $\sigma = \sum_{i=1}^{\ell} \eta_i \, \sigma_i$ represents a union or a "packet" of k-simplexes where $\eta_i = \mathbf{1}$ if the k-simplex σ_i is in σ and $\eta_i = \mathbf{0}$ otherwise. For $k \in \mathbb{Z}$, $k \neq 0, 1, 2, 3$ we have $\mathbb{Z}_2[\mathcal{S}_k(\mathcal{M})] = \mathbb{Z}_2[\emptyset] = \{0\}$ is the trivial vector space.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 11 December 2020 doi:10.20944/preprints202012.0281.v1

Example 4.3. *Consider a molecule $\mathcal{M} = (\mathbb{M}, \mathcal{S}(\mathcal{M}))$ given by the surface of a tetrahedron defined by a set of points $\mathbb{M} = \{x_0, x_1, x_2, x_3\} \subset \mathbb{R}^3$ (see Figure 5) and where*

$$\mathcal{S}_2(\mathcal{M}) = \{[x_1, x_2, x_3], [x_0, x_2, x_3], [x_0, x_1, x_3], [x_0, x_1, x_2]\},$$
$$\mathcal{S}_1(\mathcal{M}) = \{[x_2, x_3], [x_1, x_3], [x_1, x_2], [x_0, x_3], [x_0, x_2], [x_0, x_1]\},$$
$$\mathcal{S}_0(\mathcal{M}) = \{[x_0], [x_1], [x_2], [x_3]\}.$$

Now, we can identify $\mathbb{Z}_2[\mathcal{S}_0(\mathcal{M})] \equiv \mathbb{Z}_2^4$, by using

$$[x_0] \equiv (\mathbf{1}, \mathbf{0}, \mathbf{0}, \mathbf{0}),\ [x_1] \equiv (\mathbf{0}, \mathbf{1}, \mathbf{0}, \mathbf{0}),\ [x_2] \equiv (\mathbf{0}, \mathbf{0}, \mathbf{1}, \mathbf{0}),\ [x_3] \equiv (\mathbf{0}, \mathbf{0}, \mathbf{0}, \mathbf{1}).$$

Now, we can identify $\mathbb{Z}_2[\mathcal{S}_1(\mathcal{M})] \equiv \mathbb{Z}_2^6$, where we identify

$$[x_0, x_1] \equiv (\mathbf{1}, \mathbf{0}, \mathbf{0}, \mathbf{0}, \mathbf{0}, \mathbf{0}),\ [x_0, x_2] \equiv (\mathbf{0}, \mathbf{1}, \mathbf{0}, \mathbf{0}, \mathbf{0}, \mathbf{0}),\ [x_0, x_3] \equiv (\mathbf{0}, \mathbf{0}, \mathbf{1}, \mathbf{0}, \mathbf{0}, \mathbf{0}),\ [x_1, x_2] \equiv (\mathbf{0}, \mathbf{0}, \mathbf{0}, \mathbf{1}, \mathbf{0}, \mathbf{0}),$$
$$[x_1, x_3] \equiv (\mathbf{0}, \mathbf{0}, \mathbf{0}, \mathbf{0}, \mathbf{1}, \mathbf{0}),\ [x_2, x_3] \equiv (\mathbf{0}, \mathbf{0}, \mathbf{0}, \mathbf{0}, \mathbf{0}, \mathbf{1}),$$

Finally, we have $\mathbb{Z}_2[\mathcal{S}_2(\mathcal{M})] \equiv \mathbb{Z}_2^4$.

From now on, we identify $\mathbb{Z}_2[\mathcal{S}_k(\mathcal{M})] \equiv \mathbb{Z}_2^{\ell_k(\mathcal{M})}$ where $\ell_k(\mathcal{M})$ is the number of elements in $\mathcal{S}_k(\mathcal{M})$ for each $0 \leq k \leq 2$. In particular, $\ell_0(\mathcal{M}) = M$. Moreover

- For any integer number k we can endow each vector space $\mathbb{Z}_2^{\ell_k(\mathcal{M})}$ with a total ordering. To this end for a given $[x_{i_0}, x_{i_1}, \ldots, x_{i_k}] \in \mathbb{Z}_2^{\ell_k(\mathcal{M})}$ where $i_0 < i_1 < \ldots < i_k$ we consider the lexicographical order of its index set $i_0 i_1 \cdots i_k \in \{0, 1, \ldots, M\}^k$. For example, consider the surface of a tetrahedron $\mathbb{M} = \{x_0, x_1, x_2, x_3\} \subset \mathbb{R}^3$ as in Example 4.3, then in $\mathcal{S}_2(\mathcal{M})$ we have

$$[x_0, x_1] < [x_0, x_2] < [x_0, x_3] < [x_1, x_2] < [x_1, x_3] < [x_2, x_3].$$

- If $k = 1, 2$ then we can extend the map $R_i : \mathbb{Z}_2^{\ell_k(\mathcal{M})} \longrightarrow \mathbb{Z}_2^{\ell_{k-1}(\mathcal{M})}$ defined as $R_i\left(\sum_{j=1}^{\ell} \eta_j\, \sigma_j\right) = \sum_{j=1}^{\ell} \eta_j\, R_i(\sigma_j)$. Now, R_i is a linear map between vectors spaces for $0 \leq i \leq k$.

- Assume that $\sigma, \gamma \in \mathbb{Z}_2^{\ell_k(\mathcal{M})}$ and $\sigma \cap \gamma = R_i(\sigma) = R_i(\gamma)$. Then

$$R_i(\mathbf{1}\,\sigma + \mathbf{1}\,\gamma) = \mathbf{1}\,R_i(\sigma) + \mathbf{1}\,R_i(\gamma) = 0,$$

  because $\mathbf{1} + \mathbf{1} = \mathbf{0}$ in $\mathbb{Z}_2$.

Homology groups and features for three-dimensional molecules considered as surfaces

Next, we associate an incidence matrix (defined by a linear map between vector spaces) to each pair of consecutive levels $(k-1, k)$ as we show in the next example.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 11 December 2020 doi:10.20944/preprints202012.0281.v1

Example 4.4. *Consider the molecule $\mathcal{M} = (\mathbb{M}, \mathcal{S}(\mathcal{M}))$ given in the Example 4.3. We can relate the different molecular levels by using matrices following the next strategy. In Table 4 the columns are in correspondence with the 1-simplexes of $\mathcal{M}$, the rows are in correspondence with the 0-simplexes and the entries are determined by incidence of a 1-simplex with its 0-simplex face. The result is a matrix 4×6 matrix with $\mathbb{Z}_2$ entries, that is, a $\mathbb{Z}_2^{4\times 6}$-matrix. For the incidence matrix of 2-simplexes against its 1-simplexes considered as faces the construction of the corresponding $\mathbb{Z}_2^{6\times 4}$-matrix is explained in Table 5.*

The formal way to introduce the above matrices is the following. For $1 \leq k \leq 2$ we can define the following linear map between vector spaces:

$$\partial_{k-1,k} : \mathbb{Z}_2^{\ell_k(\mathcal{M})} \longrightarrow \mathbb{Z}_2^{\ell_{k-1}(\mathcal{M})}, \quad \sigma \mapsto \partial_{k-1,k}\left(\sum_{i=1}^{\ell} \eta_i\,\sigma_i\right) = \sum_{j=0}^{k} R_j\left(\sum_{i=1}^{\ell} \eta_i\,\sigma_i\right).$$

This map uses the whole set $\{R_0, R_1, \ldots, R_k\}$ of remove the i-th position linear map for $0 \leq i \leq k$. Observe that for $k = 0$ we have

$$\partial_{-1,0} : \mathbb{Z}_2^M \longrightarrow \mathbb{Z}_{-1}[\mathcal{S}_0(\mathcal{M})] = \{0\}, \quad [x] \mapsto 0,$$

the zero map, and also for $k = 3$

$$\partial_{2,3} : \mathbb{Z}_2[\mathcal{S}_3(\mathcal{M})] = \{0\} \longrightarrow \mathbb{Z}_2^{\ell_2(\mathcal{M})}, \quad 0 \mapsto 0,$$

we obtain the 0-map. Finally, we consider that $\partial_{k-1,k} = 0$ for all integer k such that $k \neq 0, 1, 2$.

To better understand the role of these maps, observe that if we have two simplexes $[x_0, x_1, x_2]$ and $[x_3, x_4, x_5]$ in $\mathbb{Z}_2^{\ell_2(\mathcal{M})}$ without common faces then the union of both is described under this algebraic framework by the sum $[x_0, x_1, x_2] + [x_3, x_4, x_5]$ (see Figure 6(a)). By using the map $\partial_{1,2}$ we obtain its description in $\mathbb{Z}_2^{\ell_0(\mathcal{M})}$ as

$$\partial_{1,2}([x_0, x_1, x_2] + [x_3, x_4, x_5]) = [x_1, x_2] + [x_0, x_2] + [x_0, x_1] + [x_4, x_5] + [x_3, x_5] + [x_3, x_4],$$

that is, the sum of the six faces of the two simplexes (see Figure 6(a)). They represent the total number of faces in their union. However, if we consider the union of two simplexes $[x_0, x_1, x_2]$ and $[x_1, x_2, x_4]$ with a common face, $R_0([x_0, x_1, x_2]) = R_2([x_1, x_2, x_3]) = [x_1, x_2]$ (see Figure 6(b)), then its description in $\mathbb{Z}_2^{\ell_0(\mathcal{M})}$ is now

$$\begin{aligned}\partial_{1,2}([x_0, x_1, x_2] + [x_1, x_2, x_3]) &= [x_1, x_2] + [x_0, x_2] + [x_0, x_1] + [x_2, x_3] + [x_1, x_2] + [x_1, x_3] \\ &= [x_0, x_2] + [x_0, x_1] + [x_2, x_3] + [x_1, x_3],\end{aligned}$$

because $[x_1, x_2] + [x_1, x_2] = 0$. This fact implies that the union of both is now described by the non-common four faces by forgetting the inner common face (see Figure 6(b)).

Example 4.5. *Consider the molecule $\mathcal{M} = (\mathbb{M}, \mathcal{S}(\mathcal{M}))$ given in the Example 4.3. Then the operators $\partial_{k-1,k}$ acts over the set of k-simplexes $(0 \leq k \leq 3)$ as follows,*

$$\begin{aligned} \partial_{2,3} &= 0, \\ \partial_{1,2}([x_0, x_1, x_2]) &= [x_1, x_2] + [x_0, x_2] + [x_0, x_1], \\ \partial_{0,1}([x_0, x_1]) &= [x_0] + [x_1], \\ \partial_{-1,0} &= 0. \end{aligned}$$

Moreover, the matrices associated to the linear maps $\partial_{0,1}$ and $\partial_{1,2}$ are

$$\partial_{0,1} = \begin{bmatrix} \mathbf{1} & \mathbf{1} & \mathbf{1} & \mathbf{0} & \mathbf{0} & \mathbf{0} \\ \mathbf{1} & \mathbf{0} & \mathbf{0} & \mathbf{1} & \mathbf{1} & \mathbf{0} \\ \mathbf{0} & \mathbf{1} & \mathbf{0} & \mathbf{1} & \mathbf{0} & \mathbf{1} \\ \mathbf{0} & \mathbf{0} & \mathbf{1} & \mathbf{0} & \mathbf{1} & \mathbf{1} \end{bmatrix} \text{ and } \partial_{1,2} = \begin{bmatrix} \mathbf{1} & \mathbf{1} & \mathbf{0} & \mathbf{0} \\ \mathbf{1} & \mathbf{0} & \mathbf{1} & \mathbf{0} \\ \mathbf{0} & \mathbf{1} & \mathbf{1} & \mathbf{0} \\ \mathbf{1} & \mathbf{0} & \mathbf{0} & \mathbf{1} \\ \mathbf{0} & \mathbf{1} & \mathbf{0} & \mathbf{1} \\ \mathbf{0} & \mathbf{0} & \mathbf{1} & \mathbf{1} \end{bmatrix}.$$

It is not difficult to see that the matrix product $\partial_{0,1} \cdot \partial_{1,2}$ is the zero matrix. Then the vector space generated by the columns of the matrix $\partial_{1,2}$, denoted by $\operatorname{Col} \partial_{1,2}$, is contained in the vector space, denoted by $\operatorname{Nul} \partial_{0,1}$, defined by the solutions of the homogeneous linear system

$$\begin{bmatrix} \mathbf{1} & \mathbf{1} & \mathbf{1} & \mathbf{0} & \mathbf{0} & \mathbf{0} \\ \mathbf{1} & \mathbf{0} & \mathbf{0} & \mathbf{1} & \mathbf{1} & \mathbf{0} \\ \mathbf{0} & \mathbf{1} & \mathbf{0} & \mathbf{1} & \mathbf{0} & \mathbf{1} \\ \mathbf{0} & \mathbf{0} & \mathbf{1} & \mathbf{0} & \mathbf{1} & \mathbf{1} \end{bmatrix} \begin{bmatrix} \alpha \\ \beta \\ \gamma \\ \delta \\ \nu \\ \eta \end{bmatrix} = \begin{bmatrix} \mathbf{0} \\ \mathbf{0} \\ \mathbf{0} \\ \mathbf{0} \end{bmatrix}.$$

Thus, we have the following subspaces $\operatorname{Col} \partial_{1,2} \subset \operatorname{Nul} \partial_{0,1} \subset \mathbb{Z}_2^6$. In a simlar way we have that $\operatorname{Col} \partial_{3,2} = \{\mathbf{0}\} \subset \operatorname{Nul} \partial_{1,2} \subset \mathbb{Z}_2^4$ and $\operatorname{Col} \partial_{0,1} \subset \operatorname{Nul} \partial_{-1,0} = \mathbb{Z}_2^4$.

It is possible to prove that $\partial_{k-1,k} \cdot \partial_{k,k+1} = 0$ holds for all $0 \leq k \leq 3$ (indeed, it is true for all integer number k) if introduce the 0 map as

$$\partial_{2,3} : \mathbb{Z}_2[\mathcal{S}_3(\mathcal{M})] = \{0\} \longrightarrow \mathbb{Z}_2[\mathcal{S}_2(\mathcal{M})]$$

This property means that the linear subspace

$$\operatorname{Col} \partial_{k,k+1} \subset \mathbb{Z}_2^{\ell_k(\mathcal{M})}$$

 doi:10.20944/preprints202012.0281.v1

generated by the columns of the matrix $\partial_{k,k+1}$ is contained in the linear subspace

$$\operatorname{Nul}\partial_{k-1,k} := \left\{\sigma \in \mathbb{Z}_2^{\ell_k(\mathcal{M})} : \partial_{k-1,k}\,\sigma = \mathbf{0}\right\}.$$

of the solution of the homogeneous linear system with matrix $\partial_{k-1,k}$. From the rank-nullity theorem, we known that

$$\ell_k(\mathcal{M}) = \dim \operatorname{Nul}\partial_{k-1,k} + \dim \operatorname{Col}\partial_{k-1,k}.$$

In particular, $\operatorname{Nul}\partial_{-1,0} = \mathbb{Z}_2^M$ and $\operatorname{Col}\partial_{2,3} = \operatorname{Col} 0 = \{\mathbf{0}\}$ is the trivial subspace.

It allows us to introduce the vector space of k-features (known as the k-th homology group) of $\mathbb{M}$ by

$$H_k(\mathcal{M}) := \operatorname{Nul}\partial_{k-1,k}/\operatorname{Col}\partial_{k,k+1},$$

and where

$$\begin{aligned}\dim H_k(\mathcal{M}) &= \dim \operatorname{Nul}\partial_{k-1,k} - \dim \operatorname{Col}\partial_{k,k+1}\\ &= \ell_k(\mathcal{M}) - \dim \operatorname{Col}\partial_{k-1,k} - \dim \operatorname{Col}\partial_{k,k+1}.\end{aligned}$$

In particular,

$$H_0(\mathcal{M}) = \operatorname{Nul}\partial_{-1,0}/\operatorname{Col}\partial_{0,1} = \mathbb{Z}_2^M/\operatorname{Col}\partial_{0,1},$$

and

$$H_2(\mathcal{M}) = \operatorname{Nul}\partial_{1,2}/\{\mathbf{0}\} = \operatorname{Nul}\partial_{1,2}.$$

Moreover, $H_k(\mathcal{M}) = \{0\}$ for all integer $k \neq 0,1,2$.

The elements in $H_k(\mathcal{M})$ are equivalence classes $\boldsymbol{\sigma}$ obtained from elements $\sigma \in \mathbb{Z}_2^{\ell_k(\mathcal{M})}$ satisfying that $\partial_{k-1,k}\,\sigma = \mathbf{0}$ and where each equivalence class is defined by a set

$$\boldsymbol{\sigma} := \left\{\gamma \in \operatorname{Nul}\partial_{k-1,k} : \gamma = \sigma + \partial_{k,k+1}(\delta) \text{ for some } \delta \in \mathbb{Z}_2^{\ell_{k+1}(\mathcal{M})}\right\}.$$

Observe, that a k-feature $\boldsymbol{\sigma} \in H_k(\mathcal{M})$ is a "packet" of simplexes σ that share its faces, that is $\partial_{k-1,k}(\sigma) = 0$ (and it can be seen as a connected component of the intersection of $\mathbb{M}$ with a k-dimensional linear subspace of $\mathbb{R}^3$), plus the vector subspace $\operatorname{Col}\partial_{k,k+1}$. Since $H_k(\mathcal{M})$ is a vector space it is possible to find a basis of $\beta_k(\mathcal{M}) := \dim H_k(\mathcal{M})$-vectors (or k-features). Hence there exists $\boldsymbol{\sigma}_1, \ldots, \boldsymbol{\sigma}_{\beta_k(\mathcal{M})}$ linear independent vectors in $H_k(\mathcal{M})$ such that it generates the whole vector space, that is,

$$H_k(\mathcal{M}) = \operatorname{span}\{\boldsymbol{\sigma}_1, \ldots, \boldsymbol{\sigma}_{\beta_k(\mathcal{M})}\}.$$

Moreover, we can easily extend the order relation of the vectors is $\mathbb{Z}_2^{\ell_k(\mathcal{M})}$ to the vectors in $H_k(\mathcal{M})$, and hence we will assume that the basis vector is ordered as follows $\boldsymbol{\sigma}_1 < \cdots < \boldsymbol{\sigma}_{\beta_k(\mathcal{M})}$. This basis vector (or k-features) characterizes the k-th homological group associated to $\mathbb{M}$. More precisely, each $\boldsymbol{\gamma} \in H_k(\mathcal{M})$ can be written as $\boldsymbol{\gamma} = \sum_{\ell=1}^{\beta_k(\mathcal{M})} \xi_\ell \boldsymbol{\sigma}_\ell$ where $\xi_\ell \in \mathbb{Z}_2$ for $1 \leq \ell \leq \beta_k(\mathcal{M})$.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 11 December 2020 doi:10.20944/preprints202012.0281.v1

Definition 4.6. *Let be a molecule $\mathcal{M} = (\mathbb{M}, \mathcal{S}(\mathcal{M}))$ described by a set of 3-dimensional points. The number $\beta_k(\mathcal{M}) = \dim H_k(\mathcal{M})$ is called de k-th Betti number for $0 \leq k \leq 2$.*

Definition 4.7. *Let be two molecules $\mathcal{M} = (\mathbb{M}, \mathcal{S}(\mathcal{M}))$ and $\mathcal{N} = (\mathbb{N}, \mathcal{S}(\mathcal{N}))$ considered as surfaces. We will say that $\mathcal{M}$ is homological to $\mathcal{N}$ if $\beta_k(\mathcal{M}) = \beta_k(\mathcal{N})$ for $0 \leq k \leq 2$.*

We recall that two vector spaces are linearly isomorphic if and only if they have the same dimension.

Persistent homology for three-dimensional molecules considered as surfaces

To describe the evolution of the features of a three-dimensional molecule $\mathcal{M}$ we can use the so-called *Vietoris-Rips complex at scale ε*. It is constructed as an approximation of $\mathcal{S}(\mathcal{M})$ which usually is computationally intractable. To define it fix some real number $\varepsilon > 0$ and construct a finite sequence of sets

$$S_\varepsilon(\mathcal{M}) := \{S_\varepsilon^{(0)}(\mathcal{M}), S_\varepsilon^{(1)}(\mathcal{M}), S_\varepsilon^{(2)}(\mathcal{M})\} \subset \mathcal{S}(\mathcal{M})$$

where the elements of each set are defined as follows:

1. $S_\varepsilon^{(0)}(\mathcal{M}) = \mathcal{S}_0(\mathcal{M})$.
2. For $k = 1, 2$ we have that $[x_0, x_1, \ldots, x_k] \in S_\varepsilon^{(k)}(\mathcal{M})$ if and only if $[x_0, x_1, \ldots, x_k] \in S_k(\mathcal{M})$ and
$$\|x_i - x_j\| < 2\varepsilon$$
holds for all $i \neq j$.

Proceeding in a similar way as above we can construct vector spaces over the finite field $\mathbb{Z}_2$ obtaining

$$\mathbb{Z}_2[S_\varepsilon^{(0)}(\mathcal{M})] = \mathbb{Z}_2[S_0(\mathcal{M})] \equiv \mathbb{Z}_2^M,$$

and for $k \geq 1$ we have that $\mathbb{Z}_2[S_\varepsilon^{(k)}(\mathcal{M})] \equiv \mathbb{Z}_2^{\ell_{k,\varepsilon}(\mathcal{M})} \subset \mathbb{Z}_2[S_k(\mathcal{M})] \equiv \mathbb{Z}_2^{\ell_k(\mathcal{M})}$ is a linear subspace that depends on $\varepsilon > 0$, for $k \geq 1$. Moreover, we have a vector space

$$H_{k,\varepsilon}(\mathcal{M}) := \mathrm{Nul}\,\partial_{k-1,k}^\varepsilon / \mathrm{Col}\,\partial_{k,k+1}^\varepsilon$$

of dimension $\beta_{k,\varepsilon}(\mathcal{M}) := \dim H_{k,\varepsilon}(\mathcal{M})$, for each integer number k. In particular, $H_{0,\varepsilon}(\mathcal{M}) = \mathbb{Z}_2^M / \mathrm{Col}\,\partial_{0,1}^\varepsilon$, and $H_{2,\varepsilon}(\mathcal{M}) = \mathrm{Ker}\partial_{1,2}^\varepsilon$ Also, $H_{k,\varepsilon}(\mathcal{M}) = \{0\}$ for every integer number $k \neq 0, 1, 2, 3$. Thus, we only need to compute

$$\left(H_{0,\varepsilon}(\mathcal{M}) = \mathbb{Z}_2^M / \mathrm{Col}\,\partial_{0,1}^\varepsilon, H_{1,\varepsilon}(\mathcal{M}) = \mathrm{Nul}\,\partial_{0,1}^\varepsilon / \mathrm{Col}\,\partial_{1,2}^\varepsilon, H_{2,\varepsilon}(\mathcal{M}) = \mathrm{Ker}\partial_{1,2}^\varepsilon\right)$$

for each $\varepsilon > 0$. In a similar way as above, we have a basis for each of this three vector spaces (representative k-features at ε-scale), namely

$$H_{k,\varepsilon}(\mathcal{M}) = \mathrm{span}\left\{\boldsymbol{\sigma}_1, \ldots, \boldsymbol{\sigma}_{\beta_{k,\varepsilon}(\mathcal{M})}\right\},$$

for $k = 0, 1, 2$. It allows us to introduce the following definition

 doi:10.20944/preprints202012.0281.v1

Definition 4.8. *Let be two molecules $\mathcal{M} = (\mathbb{M}, \mathcal{S}(\mathcal{M}))$ and $\mathcal{N} = (\mathbb{N}, \mathcal{S}(\mathcal{N}))$ considered as surfaces. We will say that $\mathcal{M}$ is ε-homological persistent to $\mathcal{N}$ if for all $\varepsilon > 0$ it holds that $\beta_{k,\varepsilon}(\mathcal{M}) = \beta_{k,\varepsilon}(\mathcal{N})$ for $0 \le k \le 2$.*

To determine if two molecules are ε-homological persistent, we can study the behaviour of the basis functions of the vector spaces $H_{k,\varepsilon}(\mathcal{M})$ depending on ε. To this end, we introduce the notion of *birth* and *death* point of a k-basis vector at ε-scale $\boldsymbol{\gamma} \in H_{k,\varepsilon}(\mathcal{M})$ as follows. The birth point of a k-feature $\boldsymbol{\gamma}$ is defined by

$$a_k(\boldsymbol{\gamma}) = \inf \{\varepsilon > 0 : \boldsymbol{\gamma} \in H_{k,\varepsilon}(\mathcal{M})\},$$

and the corresponding death point

$$b_k(\boldsymbol{\gamma}) = \sup \{\varepsilon > 0 : \boldsymbol{\gamma} \in H_{k,\varepsilon}(\mathcal{M})\}.$$

Since $a_k(\boldsymbol{\gamma}) \le b_k(\boldsymbol{\gamma})$ holds we will call the interval $[a_k(\boldsymbol{\gamma}), b_k(\boldsymbol{\gamma})]$ *the barcode of the feature* $\boldsymbol{\gamma}$.

In order to implement in practice the comparison between two molecules a range represented by an interval $[\varepsilon_{\min}, \varepsilon_{\max}]$ of real numbers is chosen. This interval reflects the smallest and largest features scales that we will consider. A maximal choice is to take $\varepsilon_{\min} = 0$ and $\varepsilon_{\max}$ the farthest distance between points of $\mathbb{M}$. We can take values

$$\varepsilon_j := \varepsilon_{\min} + j\left(\frac{\varepsilon_{\max} - \varepsilon_{\min}}{m}\right)$$

For each $k = 0, 1, \ldots$ and $j = 1, 2, \ldots, m$ fixed we have

$$H_{k,\varepsilon_j}(\mathcal{M}) = \operatorname{span}\left\{\boldsymbol{\gamma}_1, \ldots, \boldsymbol{\gamma}_{\beta_k^{(j)}}\right\},$$

Let N_k $(0 \le k \le 2)$ be the number of vectors in the set $\left\{\boldsymbol{\gamma} \in H_{k,\varepsilon_j}(\mathcal{M}) : \text{ for some } 1 \le j \le m\right\}$ that we can ordered as

$$\boldsymbol{\gamma}_1 < \boldsymbol{\gamma}_2 < \cdots < \boldsymbol{\gamma}_{N_k}.$$

For each k-feature $\boldsymbol{\gamma}_\upsilon$ $(1 \le \upsilon \le N_k)$ we have its barcode $\mathbb{I}_{k,\nu} := [a_k(\boldsymbol{\gamma}_\upsilon), b_k(\boldsymbol{\gamma}_\upsilon)]$, where

$$a_k(\boldsymbol{\gamma}_\upsilon) = \min_{0 \le j \le m} \{\varepsilon_j : \boldsymbol{\gamma}_\upsilon \in H_{k,\varepsilon_j}(\mathcal{M})\},$$

and

$$b_k(\boldsymbol{\gamma}_\upsilon) = \max_{0 \le j \le m} \{\varepsilon_j : \boldsymbol{\gamma}_\upsilon \in H_{k,\varepsilon_j}(\mathcal{M})\}$$

that we represent graphically as we show in Figure 7.

Thus, we have that the k-level barcodes associated with the partition $\mathcal{P} := \{\varepsilon_j\}_{j=1}^m$ are given by $\mathcal{B}_k(\mathbb{M}, \mathcal{P}) := \{\mathbb{I}_{k,\nu} : 1 \le \upsilon \le N_k\}$. To simplify notation, we will write

$$\mathbb{I}_{k,\nu} := [a_k^{(\upsilon)}, b_k^{(\upsilon)}] \text{ for } 1 \le \upsilon \le N_k \text{ and } k = 0, 1, 2, 3;$$

where the $a_k^{(\upsilon)} := a_k(\boldsymbol{\gamma}_\upsilon)$ and $b_k^{(\upsilon)} := b_k(\boldsymbol{\gamma}_\upsilon)$. Then $\mathcal{B}_k = \mathcal{B}_k(\mathbb{M}, \mathcal{P}) = \{\mathbb{I}_{k,1}, \mathbb{I}_{k,2}, \ldots, \mathbb{I}_{k,N_k}\}$ for $0 \le k \le 2$ are the barcodes of $\mathbb{M}$ under partition $\mathcal{P}$.

 doi:10.20944/preprints202012.0281.v1

Persistent similarity using barcodes between molecules considered as surfaces

In order to determine the grade of similarity between two barcodes from proteins we need to set a similarity metric. Based on the barcodes concept, it is possible to build a model that enables us to study the structure of the proteins in different k-scales. To this end we introduce the so-called k-th Persistent Betti Function (PBF) [10] by:

$$f(x; \mathcal{B}_k) := \sum_{j=1}^{N_k} w_k^{(j)} \exp\left(-\left[\frac{x - \frac{1}{2}(b_k^{(j)} + a_k^{(j)})}{\sigma(b_k^{(j)} - a_k^{(j)})}\right]^{2\kappa}\right)$$

for $\kappa > 0$ and $k = 0, 1, 2$. Here

- $w_k^{(j)}$ is the weight for the j-th associated k-feature. Usually is considered that $w_k^{(j)} = 1$ for all k and j.
- σ is the resolution parameter. Usually is considered $\sigma = 1$, otherwise we can change the resolution parameter to observe variations of structure properties from various scales.
- Finally, κ is the kernel scale parameter.

In this way, it is possible to transform the complexity of the three-dimensional protein structure and the barcodes into unidimensional continuous functions. Therefore, only 3 Persistent Betti Functions (PBFs) (one of each feature level) are needed to represent a protein tertiary structure. To compare two protein structures, namely $\mathcal{M}$ and $\mathcal{N}$, we construct its corresponding family of PFBs denoted by

$$\mathcal{F}_{\mathcal{M}} := \{f_{\mathcal{M}}(x, \mathcal{B}_k) : 0 \le k \le 2\} \text{ and } \mathcal{F}_{\mathcal{N}} := \{f_{\mathcal{N}}(x, \mathcal{B}_k) : 0 \le k \le 2\}.$$

We implemented in R the Persistent Betti Functions (see [10] and references therein). Then, by the help of the family of PFBs, we introduce the following k-similarity measure between two molecules $\mathcal{M}$ and $\mathcal{N}$ $(0 \le k \le 2)$.

Definition 4.9. *We called k-persistent similarity measure* $(0 \le k \le 3)$ *between two molecules $\mathcal{M}$ and $\mathcal{N}$ to the number*

$$PS_k(\mathcal{M}, \mathcal{N}) := \frac{\int_{\mathcal{R}} \min(f_{\mathcal{M}}(x, \mathcal{B}_k), f_{\mathcal{N}}(x, \mathcal{B}_k))dx}{\int_{\mathcal{R}} \max(f_{\mathcal{M}}(x, \mathcal{B}_k), f_{\mathcal{N}}(x, \mathcal{B}_k))dx}.$$

Observe that since

$$0 < \min(f_{\mathcal{M}}(x, \mathcal{B}_k), f_{\mathcal{N}}(x, \mathcal{B}_k)) \le \max(f_{\mathcal{M}}(x, \mathcal{B}_k), f_{\mathcal{N}}(x, \mathcal{B}_k))$$

holds for all $x \in \mathbb{R}$ and $\max(f_{\mathcal{M}}(x, \mathcal{B}_k), f_{\mathcal{N}}(x, \mathcal{B}_k))$ decreases fast to 0 as $|x|$ increases to ∞, we obtain that

$$0 < PS_k(\mathcal{M}, \mathcal{N}) \le 1$$

holds for all $k = 0, 1, 2$. Clearly, if $f_{\mathcal{M}}(x, \mathcal{B}_k) \approx f_{\mathcal{N}}(x, \mathcal{B}_k)$ holds, that is, the k features obtained from $\mathcal{M}$ and $\mathcal{N}$ have a similar homological structure, and then $PS_k(\mathcal{M}, \mathcal{N}) \approx 1$.

Definition 4.10. *We define the persistent similarity mean between between two molecules $\mathcal{M}$ and $\mathcal{N}$ by*

$$\overline{PS(\mathcal{M},\mathcal{N})} := \frac{1}{3}\sum_{k=0}^{2} PS_k(\mathcal{M},\mathcal{N}).$$

In this paper, we computed the k-Persistent Similarity ($0 \leq k \leq 2$) of PDB structures from a Drugbank against SARS-CoV-2 proteins. We calculated the mean of the Persistent similarity for each protein comparison, drugs $\mathcal{M}$ with structures satisfying $\overline{PS(\mathcal{M}, \text{SARS-CoV-2})} \geq 0.9$, would be candidates for transcriptomics validation and molecular docking by autodock 4.

Acknowledgements *This work is partially supported by grants FONDOS SUPERA COVID-19, 2020-2021 and Fundación BBVA a equipos de investigación científica SARS-CoV-2 y COVID-19, IA4COVID19 2020-2022.*

References

[1] A. Al-Mandhari, D. Samhouri, A. Abubakar, R. Brennan, Coronavirus Disease 2019 outbreak: preparedness and readiness of countries in the Eastern Mediterranean Region, East Mediterr Health J 26 (2) (2020) 136–137.

[2] H. and Pan, R. Peto, Q. A. Karim, M. Alejandria, A. M. Henao-Restrepo, C. H. García, M.-P. Kieny, R. Malekzadeh, S. Murthy, M.-P. Preziosi, S. Reddy, M. R. Periago, V. Sathiyamoorthy, J.-A. Røttingen, S. a. Swaminathan, Repurposed antiviral drugs for covid-19-interim who solidarity trial results, medRxivarXiv:https://www.medrxiv.org/content/early/2020/10/15/2020.10.15.20209817.full.pdf, doi:10.1101/2020.10.15.20209817.
URL https://www.medrxiv.org/content/early/2020/10/15/2020.10.15.20209817

[3] R. C. Mohs, N. H. Greig, Drug discovery and development: Role of basic biological research, Alzheimers Dement (N Y) 3 (4) (2017) 651–657.

[4] J. A. DiMasi, L. Feldman, A. Seckler, A. Wilson, Trends in risks associated with new drug development: success rates for investigational drugs, Clin Pharmacol Ther 87 (3) (2010) 272–277.

[5] S. Kumar, Covid-19: A drug repurposing and biomarker identification by using comprehensive gene-disease associations through protein-protein interaction network analysisdoi:10.20944/preprints202003.0440.v1.

[6] M. Griffith, O. L. Griffith, A. C. Coffman, J. V. Weible, J. F. McMichael, N. C. Spies, J. Koval, I. Das, M. B. Callaway, J. M. Eldred, C. A. Miller, J. Subramanian, R. Govindan, R. D. Kumar, R. Bose, L. Ding, J. R. Walker, D. E. Larson, D. J. Dooling, S. M. Smith, T. J. Ley, E. R. Mardis, R. K. Wilson, DGIdb: mining the druggable genome, Nat Methods 10 (12) (2013) 1209–1210.

[7] S. Freshour, S. Kiwala, K. C. Cotto, A. C. Coffman, J. F. McMichael, J. Song, M. Griffith, O. L. Griffith, A. H. Wagner, Integration of the drug-gene interaction database (dgidb) with open crowdsource effortsdoi:10.1101/2020.09.18.301721.

[8] Y. Zhou, Y. Hou, J. Shen, Y. Huang, W. Martin, F. Cheng, Network-based drug repurposing for novel coronavirus 2019-nCoV/SARS-CoV-2, Cell Discov 6 (2020) 14.

[9] S. Keretsu, S. P. Bhujbal, S. J. Cho, Rational approach toward COVID-19 main protease inhibitors via molecular docking, molecular dynamics simulation and free energy calculation, Sci Rep 10 (1) (2020) 17716.

[10] K. Xia, G.-W. Wei, Persistent homology analysis of protein structure, flexibility, and folding, International Journal for Numerical Methods in Biomedical Engineering 30 (8) (2014) 814–844. doi:10.1002/cnm.2655.

[11] D. S. Wishart, Drugbank: a comprehensive resource for in silico drug discovery and exploration, Nucleic Acids Research 34 (90001) (2006) D668–D672. doi:10.1093/nar/gkj067.

[12] X. Xiong, J. L. Smith, M. S. Chen, Effect of incorporation of cidofovir into DNA by human cytomegalovirus DNA polymerase on DNA elongation, Antimicrob Agents Chemother 41 (3) (1997) 594–599.

[13] H. Javed, M. M. Khan, A. Ahmad, K. Vaibhav, M. E. Ahmad, A. Khan, M. Ashafaq, F. Islam, M. S. Siddiqui, M. M. Safhi, F. Islam, Rutin prevents cognitive impairments by ameliorating oxidative stress and neuroinflammation in rat model of sporadic dementia of Alzheimer type, Neuroscience 210 (2012) 340–352.

[14] S. K. Richetti, M. Blank, K. M. Capiotti, A. L. Piato, M. R. Bogo, M. R. Vianna, C. D. Bonan, Quercetin and rutin prevent scopolamine-induced memory impairment in zebrafish, Behav Brain Res 217 (1) (2011) 10–15.

[15] S. Lucas, The Pharmacology of Indomethacin, Headache 56 (2) (2016) 436–446.

[16] W. R. S. Munjal A, In: Statpearls [internet]., Treasure Island (FL): StatPearls Publishing.
URL https://www.ncbi.nlm.nih.gov/books/NBK556107/

[17] E. A. Keskin D, Sadri S, Dasatinib for the treatment of chronic myeloid leukemia: patient selection and special considerations., Drug Des Devel Ther. 10:3355-3361.

[18] A. E. Shefrin, R. D. Goldman, Use of dexamethasone and prednisone in acute asthma exacerbations in pediatric patients, Can Fam Physician 55 (7) (2009) 704–706.

[19] K. N. Y. K. e. a. Nakano, S., Cardioprotective mechanisms of spironolactone associated with the angiotensin-converting enzyme/epidermal growth factor receptor/extracellular signal-regulated kinases, nad(p)h oxidase/lectin-like oxidized low-density lipoprotein receptor-1, and rho-kinase pathways in aldosterone/salt-induced hypertensive rats., Hypertens 28 (7) (2005) 925–936.

[20] N. C. for Biotechnology Information (2020)., Pubchem compound summary for cid 4764, phenolphthalein. (2020).
URL https://pubchem.ncbi.nlm.nih.gov/compound/4764

[21] F. Díaz-Castro, M. Monsalves-Álvarez, L. E. Rojo, A. del Campo, R. Troncoso, Mifepristone for treatment of metabolic syndrome: Beyond cushing's syndrome, Frontiers in Pharmacology 11 (2020) 429. doi:10.3389/fphar.2020.00429.
URL https://www.frontiersin.org/article/10.3389/fphar.2020.00429

[22] L. Silvestre, C. Dubois, M. Renault, Y. Rezvani, E.-E. Baulieu, A. Ulmann, Voluntary interruption of pregnancy with mifepristone (ru 486) and a prostaglandin analogue, New England Journal of Medicine 322 (10) (1990) 645–648, pMID: 2304490. arXiv:https://doi.org/10.1056/NEJM199003083221001, doi:10.1056/NEJM199003083221001.
URL https://doi.org/10.1056/NEJM199003083221001

[23] K. W. Al-Quliti, Update on neuropathic pain treatment for trigeminal neuralgia. The pharmacological and surgical options, Neurosciences (Riyadh) 20 (2) (2015) 107–114.

[24] N. C. for Biotechnology Information (2020)., Pubchem compound summary for cid 42611257, vemurafenib. (2020).
URL https://pubchem.ncbi.nlm.nih.gov/compound/4764

[25] J. A. Sosman, K. B. Kim, L. Schuchter, R. Gonzalez, A. C. Pavlick, J. S. Weber, G. A. McArthur, T. E. Hutson, S. J. Moschos, K. T. Flaherty, P. Hersey, R. Kefford, D. Lawrence, I. Puzanov, K. D. Lewis, R. K. Amaravadi, B. Chmielowski, H. J. Lawrence, Y. Shyr, F. Ye, J. Li, K. B. Nolop, R. J. Lee, A. K. Joe, A. Ribas, Survival in BRAF V600-mutant advanced melanoma treated with vemurafenib, N Engl J Med 366 (8) (2012) 707–714.

[26] L. Liu, Y. Cao, C. Chen, X. Zhang, A. McNabola, D. Wilkie, S. Wilhelm, M. Lynch, C. Carter, Sorafenib blocks the RAF/MEK/ERK pathway, inhibits tumor angiogenesis, and induces tumor cell apoptosis in hepatocellular carcinoma model PLC/PRF/5, Cancer Res 66 (24) (2006) 11851–11858.

[27] N. C. for Biotechnology Information (2020)., Pubchem compound summary for cid 5035, raloxifene.

(2020).
URL https://pubchem.ncbi.nlm.nih.gov/compound/5035

[28] X. Wang, Y. Guan, COVID-19 drug repurposing: A review of computational screening methods, clinical trials, and protein interaction assays, Med Res Rev.

[29] L. Riva, S. Yuan, X. Yin, L. Martin-Sancho, N. Matsunaga, L. Pache, S. Burgstaller-Muehlbacher, P. D. De Jesus, P. Teriete, M. V. Hull, M. W. Chang, J. F. Chan, J. Cao, V. K. Poon, K. M. Herbert, K. Cheng, T. H. Nguyen, A. Rubanov, Y. Pu, C. Nguyen, A. Choi, R. Rathnasinghe, M. Schotsaert, L. Miorin, M. Dejosez, T. P. Zwaka, K. Y. Sit, L. Martinez-Sobrido, W. C. Liu, K. M. White, M. E. Chapman, E. K. Lendy, R. J. Glynne, R. Albrecht, E. Ruppin, A. D. Mesecar, J. R. Johnson, C. Benner, R. Sun, P. G. Schultz, A. I. Su, A. Garc?a-Sastre, A. K. Chatterjee, K. Y. Yuen, S. K. Chanda, Discovery of SARS-CoV-2 antiviral drugs through large-scale compound repurposing, Nature 586 (7827) (2020) 113–119.

[30] J. H. Beigel, K. M. Tomashek, L. E. Dodd, A. K. Mehta, B. S. Zingman, A. C. Kalil, E. Hohmann, H. Y. Chu, A. Luetkemeyer, S. Kline, D. Lopez de Castilla, R. W. Finberg, K. Dierberg, V. Tapson, L. Hsieh, T. F. Patterson, R. Paredes, D. A. Sweeney, W. R. Short, G. Touloumi, D. C. Lye, N. Ohmagari, M. D. Oh, G. M. Ruiz-Palacios, T. Benfield, G. F?tkenheuer, M. G. Kortepeter, R. L. Atmar, C. B. Creech, J. Lundgren, A. G. Babiker, S. Pett, J. D. Neaton, T. H. Burgess, T. Bonnett, M. Green, M. Makowski, A. Osinusi, S. Nayak, H. C. Lane, A. K. Mehta, N. G. Rouphael, J. J. Traenkner, V. D. Cantos, G. Alaaeddine, B. S. Zingman, R. Grossberg, P. F. Riska, E. Hohmann, M. Torres-Soto, N. Jilg, H. Y. Chu, A. Wald, M. Green, A. Luetkemeyer, P. B. Crouch, H. Jang, S. Kline, J. Billings, B. Noren, D. Lopez de Castilla, J. W. Van Winkle, F. X. Riedo, R. W. Finberg, J. P. Wang, M. Wessolossky, K. Dierberg, B. Eckhardt, H. J. Neumann, V. Tapson, J. Grein, F. Sutterwala, L. Hsieh, A. N. Amin, T. F. Patterson, H. Javeri, T. Vu, R. Paredes, L. Mateu, D. A. Sweeney, C. A. Benson, F. Ali, W. R. Short, P. Tebas, J. Torgersen, G. Touloumi, V. Gioukari, D. C. Lye, S. W. X. Ong, N. Ohmagari, A. Mikami, G. F?tkenheuer, J. J. Malin, P. Koehler, A. C. Kalil, L. Larson, A. Hewlett, M. G. Kortepeter, C. B. Creech, I. Thomsen, T. W. Rice, B. Taiwo, K. Krueger, S. H. Cohen, G. R. Thompson, C. Wolfe, E. B. Walter, M. Frank, H. Young, A. R. Falsey, A. R. Branche, P. Goepfert, N. Erdmann, O. O. Yang, J. Ahn, A. Goodman, B. Merrick, R. M. Novak, A. Wendrow, H. Arguinchona, C. Arguinchona, S. L. George, J. Tennant, R. L. Atmar, H. M. El Sahly, J. Whitaker, D. A. Price, C. J. A. Duncan, S. Metallidis, T. Chrysanthidis, S. L. F. McLellan, M. D. Oh, W. B. Park, E. S. Kim, J. Jung, J. R. Ortiz, K. L. Kotloff, B. Angus, J. D. Germain Seymour, N. A. Hynes, L. M. Sauer, N. Ahuja, K. Nadeau, P. E. H. Jackson, T. D. Bell, A. Antoniadou, K. Protopapas, R. T. Davey, J. D. Voell, J. Mu?oz, M. Roldan, I. Kalomenidis, S. G. Zakynthinos, C. I. Paules, F. McGill,

J. Minton, N. Koulouris, Z. Barmparessou, E. Swiatlo, K. Widmer, N. Huprikar, A. Ganesan, G. M. Ruiz-Palacios, A. Ponce de Le?n, S. Rajme, J. Regalado Pineda, J. A. Martinez-Orozco, M. Holodniy, A. Chary, T. Wolf, C. Stephan, J. C. Wasmuth, C. Boesecke, M. Llewelyn, B. Philips, C. J. Colombo, R. E. Colombo, D. A. Lindholm, K. Mende, T. Lee, T. Lalani, R. C. Maves, G. C. Utz, J. Lundgren, M. Helleberg, J. Gerstoft, T. Benfield, T. Jensen, B. Lindegaard, L. Weise, L. Knudsen, I. Johansen, L. W. Madsen, L. ?stergaard, N. St?rke, H. Nielsen, A. G. Babiker, S. Pett, J. D. Neaton, D. S. Stephens, T. H. Burgess, T. M. Uyeki, R. Walker, G. L. Marks, A. Osinusi, H. Cao, K. K. Chung, S. E. Chambers, M. Green, M. Makowski, J. L. Ferreira, M. R. Wierzbicki, T. Bonnett, N. Gettinger, T. Engel, J. Wang, J. H. Beigel, K. M. Tomashek, S. Nayak, L. E. Dodd, W. Dempsey, E. Nomicos, M. Lee, P. Wolff, R. Pikaart-Tautges, M. Elsafy, R. Jurao, H. Koo, M. Proschan, D. Follmann, H. C. Lane, Remdesivir for the Treatment of Covid-19 - Final Report, N Engl J Med 383 (19) (2020) 1813–1826.

[31] Y. Wang, D. Zhang, G. Du, R. Du, J. Zhao, Y. Jin, S. Fu, L. Gao, Z. Cheng, Q. Lu, Y. Hu, G. Luo, K. Wang, Y. Lu, H. Li, S. Wang, S. Ruan, C. Yang, C. Mei, Y. Wang, D. Ding, F. Wu, X. Tang, X. Ye, Y. Ye, B. Liu, J. Yang, W. Yin, A. Wang, G. Fan, F. Zhou, Z. Liu, X. Gu, J. Xu, L. Shang, Y. Zhang, L. Cao, T. Guo, Y. Wan, H. Qin, Y. Jiang, T. Jaki, F. G. Hayden, P. W. Horby, B. Cao, C. Wang, Remdesivir in adults with severe COVID-19: a randomised, double-blind, placebo-controlled, multicentre trial, Lancet 395 (10236) (2020) 1569–1578.

[32] B. Cao, Y. Wang, D. Wen, W. Liu, J. Wang, G. Fan, L. Ruan, B. Song, Y. Cai, M. Wei, X. Li, J. Xia, N. Chen, J. Xiang, T. Yu, T. Bai, X. Xie, L. Zhang, C. Li, Y. Yuan, H. Chen, H. Li, H. Huang, S. Tu, F. Gong, Y. Liu, Y. Wei, C. Dong, F. Zhou, X. Gu, J. Xu, Z. Liu, Y. Zhang, H. Li, L. Shang, K. Wang, K. Li, X. Zhou, X. Dong, Z. Qu, S. Lu, X. Hu, S. Ruan, S. Luo, J. Wu, L. Peng, F. Cheng, L. Pan, J. Zou, C. Jia, J. Wang, X. Liu, S. Wang, X. Wu, Q. Ge, J. He, H. Zhan, F. Qiu, L. Guo, C. Huang, T. Jaki, F. G. Hayden, P. W. Horby, D. Zhang, C. Wang, A Trial of Lopinavir-Ritonavir in Adults Hospitalized with Severe Covid-19, N Engl J Med 382 (19) (2020) 1787–1799.

[33] K. Xia, G. W. Wei, Persistent homology analysis of protein structure, flexibility, and folding, Int J Numer Method Biomed Eng 30 (8) (2014) 814–844.

[34] Z. Cang, L. Mu, K. Wu, K. Opron, K. Xia, G.-W. Wei, A topological approach for protein classification (2015). `arXiv:1510.00953`.

[35] T. K. Dey, S. Mandal, Protein classification with improved topological data analysis 113 (2018) 6:1–6:13. `doi:10.4230/LIPIcs.WABI.2018.6`.
URL `http://drops.dagstuhl.de/opus/volltexte/2018/9308`

[36] L. Holm, Dali and the persistence of protein shape, Protein Science 29 (1) (2020) 128–140. arXiv: https://onlinelibrary.wiley.com/doi/pdf/10.1002/pro.3749, doi:https://doi.org/10.1002/pro.3749.
URL https://onlinelibrary.wiley.com/doi/abs/10.1002/pro.3749

[37] K. Baby, S. Maity, C. H. Mehta, A. Suresh, U. Y. Nayak, Y. Nayak, Targeting SARS-CoV-2 RNA-dependent RNA polymerase: An in silico drug repurposing for COVID-19, F1000Res 9 (2020) 1166.

[38] A. Acharya, R. Agarwal, M. Baker, J. Baudry, D. Bhowmik, S. Boehm, K. Byler, L. Coates, S. Y. Chen, C. J. Cooper, O. Demerdash, I. Daidone, J. Eblen, S. R. Ellingson, S. Forli, J. Glaser, J. C. Gumbart, J. Gunnels, O. Hernandez, S. Irle, J. Larkin, T. J. Lawrence, S. LeGrand, S. H. Liu, J. C. Mitchell, G. Park, J. M. Parks, A. Pavlova, L. Petridis, D. Poole, L. Pouchard, A. Ramanathan, D. Rogers, D. Santos-Martins, A. Scheinberg, A. Sedova, S. Shen, J. C. Smith, M. Smith, C. Soto, A. Tsaris, M. Thavappiragasam, A. F. Tillack, J. V. Vermaas, V. Q. Vuong, J. Yin, S. Yoo, M. Zahran, L. Zanetti-Polzi, Supercomputer-Based Ensemble Docking Drug Discovery Pipeline with Application to Covid-19, ChemRxiv.

[39] B. N. Marak, J. Dowarah, L. Khiangte, V. P. Singh, Step toward repurposing drug discovery for COVID-19 therapeutics through in silico approach, Drug Dev Res.

[40] A. Trezza, D. Iovinelli, A. Santucci, F. Prischi, O. Spiga, An integrated drug repurposing strategy for the rapid identification of potential SARS-CoV-2 viral inhibitors, Sci Rep 10 (1) (2020) 13866.

[41] Z. Jia, X. Song, J. Shi, W. Wang, K. He, Transcriptome-based drug repositioning for coronavirus disease 2019 (COVID-19), Pathog Dis 78 (4).

[42] Y. Kumar, H. Singh, C. N. Patel, In silico prediction of potential inhibitors for the main protease of SARS-CoV-2 using molecular docking and dynamics simulation based drug-repurposing, J Infect Public Health 13 (9) (2020) 1210–1223.

[43] A. D. Elmezayen, A. Al-Obaidi, A. T. ?ahin, K. Yelek?i, Drug repurposing for coronavirus (COVID-19): in silico screening of known drugs against coronavirus 3CL hydrolase and protease enzymes, J Biomol Struct Dyn (2020) 1–13.

[44] D. W. Kneller, G. Phillips, H. M. O'Neill, R. Jedrzejczak, L. Stols, P. Langan, A. Joachimiak, L. Coates, A. Kovalevsky, Structural plasticity of SARS-CoV-2 3CL Mpro active site cavity revealed by room temperature X-ray crystallography, Nat Commun 11 (1) (2020) 3202.

[45] Y. Gao, L. Yan, Y. Huang, F. Liu, Y. Zhao, L. Cao, T. Wang, Q. Sun, Z. Ming, L. Zhang, J. Ge, L. Zheng, Y. Zhang, H. Wang, Y. Zhu, C. Zhu, T. Hu, T. Hua, B. Zhang, X. Yang, J. Li, H. Yang,

Z. Liu, W. Xu, L. W. Guddat, Q. Wang, Z. Lou, Z. Rao, Structure of the RNA-dependent RNA polymerase from COVID-19 virus, Science 368 (6492) (2020) 779–782.

[46] K. Bhardwaj, P. Liu, J. L. Leibowitz, C. C. Kao, The coronavirus endoribonuclease Nsp15 interacts with retinoblastoma tumor suppressor protein, J Virol 86 (8) (2012) 4294–4304.

[47] X. Deng, M. Hackbart, R. C. Mettelman, O.

[48] C. I. B. M. N. L. of Medicine (US). 2000 Feb 29, Identifier: Nct04542434 ,niclosamide in covid-19.

[49] S. Jeon, M. Ko, J. Lee, I. Choi, S. Y. Byun, S. Park, D. Shum, S. Kim, Identification of Antiviral Drug Candidates against SARS-CoV-2 from FDA-Approved Drugs, Antimicrob Agents Chemother 64 (7).

[50] C. J. Wu, J. T. Jan, C. M. Chen, H. P. Hsieh, D. R. Hwang, H. W. Liu, C. Y. Liu, H. W. Huang, S. C. Chen, C. F. Hong, R. K. Lin, Y. S. Chao, J. T. Hsu, Inhibition of severe acute respiratory syndrome coronavirus replication by niclosamide, Antimicrob Agents Chemother 48 (7) (2004) 2693–2696.

[51] N. C. Gassen, D. Niemeyer, D. Muth, V. M. Corman, S. Martinelli, A. Gassen, K. Hafner, J. Papies, K. Mösbauer, A. Zellner, A. S. Zannas, A. Herrmann, F. Holsboer, R. Brack-Werner, M. Boshart, B. Müller-Myhsok, C. Drosten, M. A. Müller, T. Rein, Skp2 attenuates autophagy through beclin1-ubiquitination and its inhibition reduces mers-coronavirus infection, Nature Communications 10 (1) (2019) 5770. doi:10.1038/s41467-019-13659-4.
URL https://doi.org/10.1038/s41467-019-13659-4

[52] S. Sharma, A. Ali, J. Ali, J. K. Sahni, S. Baboota, Rutin: therapeutic potential and recent advances in drug delivery, Expert Opinion on Investigational Drugs 22 (8) (2013) 1063–1079, pMID: 23795677. arXiv:https://doi.org/10.1517/13543784.2013.805744, doi:10.1517/13543784.2013.805744.
URL https://doi.org/10.1517/13543784.2013.805744

[53] Y.-C. Perng, D. J. Lenschow, Isg15 in antiviral immunity and beyond, Nature Reviews Microbiology 16 (7) (2018) 423–439. doi:10.1038/s41579-018-0020-5.
URL https://doi.org/10.1038/s41579-018-0020-5

[54] S. Kang, H. M. Brown, S. Hwang, Direct Antiviral Mechanisms of Interferon-Gamma, Immune Netw 18 (5) (2018) e33.

[55] X. Hu, X. Cai, X. Song, C. Li, J. Zhao, W. Luo, Q. Zhang, I. O. Ekumi, Z. He, Possible sars-coronavirus 2 inhibitor revealed by simulated molecular docking to viral main protease and host toll-like receptor, Future Virology (2020) 10.2217/fvl–2020–0099doi:10.2217/fvl-2020-0099.
URL https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7295248/

[56] Dexamethasone in hospitalized patients with covid-19 — preliminary report, New England Journal of Medicine 0 (0) (0) null. arXiv:https://doi.org/10.1056/NEJMoa2021436, doi:10.1056/NEJMoa2021436.
URL https://doi.org/10.1056/NEJMoa2021436

[57] I. Sarkar, A. Sen, In silico screening predicts common cold drug Dextromethorphan along with Prednisolone and Dexamethasone can be effective against novel Coronavirus disease (COVID-19), J Biomol Struct Dyn (2020) 1–5.

[58] S. Matsuyama, M. Kawase, N. Nao, K. Shirato, M. Ujike, W. Kamitani, M. Shimojima, S. Fukushi, The inhaled steroid ciclesonide blocks sars-cov-2 rna replication by targeting the viral replication-transcription complex in cultured cells, Journal of VirologyarXiv:https://jvi.asm.org/content/early/2020/10/09/JVI.01648-20.full.pdf, doi:10.1128/JVI.01648-20.
URL https://jvi.asm.org/content/early/2020/10/09/JVI.01648-20

[59] G. Ribaudo, A. Ongaro, E. Oselladore, G. Zagotto, M. Memo, A. Gianoncelli, A computational approach to drug repurposing against SARS-CoV-2 RNA dependent RNA polymerase (RdRp), J Biomol Struct Dyn (2020) 1–8.

[60] M. S. A. Parvez, M. A. Karim, M. Hasan, J. Jaman, Z. Karim, T. Tahsin, M. N. Hasan, M. J. Hosen, Prediction of potential inhibitors for RNA-dependent RNA polymerase of SARS-CoV-2 using comprehensive drug repurposing and molecular docking approach, Int J Biol Macromol 163 (2020) 1787–1797.

[61] J. Ahmad, S. Ikram, F. Ahmad, I. U. Rehman, M. Mushtaq, SARS-CoV-2 RNA Dependent RNA polymerase (RdRp) - A drug repurposing study, Heliyon 6 (7) (2020) e04502.

[62] R. Pokhrel, P. Chapagain, J. Siltberg-Liberles, Potential RNA-dependent RNA polymerase inhibitors as prospective therapeutics against SARS-CoV-2, J Med Microbiol 69 (6) (2020) 864–873.

[63] S. Pleschka, T. Wolff, C. Ehrhardt, G. Hobom, O. Planz, U. R. Rapp, S. Ludwig, Influenza virus propagation is impaired by inhibition of the Raf/MEK/ERK signalling cascade, Nat Cell Biol 3 (3) (2001) 301–305.

[64] L. Adnane, P. A. Trail, I. Taylor, S. M. Wilhelm, Sorafenib (BAY 43-9006, Nexavar), a dual-action inhibitor that targets RAF/MEK/ERK pathway in tumor cells and tyrosine kinases VEGFR/PDGFR in tumor vasculature, Methods Enzymol 407 (2006) 597–612.

[65] S. Pleschka, RNA viruses and the mitogenic Raf/MEK/ERK signal transduction cascade, Biol Chem 389 (10) (2008) 1273–1282.

[66] Y. Cai, Y. Liu, X. Zhang, Suppression of coronavirus replication by inhibition of the MEK signaling pathway, J Virol 81 (2) (2007) 446–456.

[67] M. Ghasemnejad-Berenji, S. Pashapour, SARS-CoV-2 and the Possible Role of Raf/MEK/ERK Pathway in Viral Survival: Is This a Potential Therapeutic Strategy for COVID-19?, Pharmacology (2020) 1–3.

[68] M. Ali, A. Ezzat, DrugBank Database XML Parser, Dainanahan, r package version 1.2.0 (2020).
URL https://CRAN.R-project.org/package=dbparser

[69] G. B.J., R. A.P.C., E. K.M., M. J.A., C. L.S.D., Bio3d: An r package for the comparative analysis of protein structures., Bioinformatics 22 (2006) 2695–2696.

[70] Z. Cang, L. Mu, K. Wu, K. Opron, K. Xia, G.-W. Wei, A topological approach for protein classification, Computational and Mathematical Biophysics (1). doi:https://doi.org/10.1515/mlbmb-2015-0009.
URL https://www.degruyter.com/view/journals/cmb/open-issue/article-10.1515-mlbmb-2015-0009/article-10.1515-mlbmb-2015-0009.xml

[71] R. R. Wadhwa, D. F. Williamson, A. Dhawan, J. G. Scott, Tdastats: R pipeline for computing persistent homology in topological data analysis, Journal of Open Source Software 3 (28) (2018) 860. doi:10.21105/joss.00860.
URL https://doi.org/10.21105/joss.00860

[72] U. Bauer, Ripser: efficient computation of vietoris-rips persistence barcodes (2019). arXiv:1908.02518.

[73] C. T. Porter, G. J. Bartlett, J. M. Thornton, The catalytic site atlas: a resource of catalytic sites and residues identified in enzymes using structural data, Nucleic Acids Res 32 (Database issue) (2004) D129–133.

[74] B. J. Jain, M. Lappe, Joining softassign and dynamic programming for the contact map overlap problem (2007) 410–423.

[75] G. Lancia, R. Carr, B. Walenz, S. Istrail, 101 optimal pdb structure alignments: A branch-and-cut algorithm for the maximum contact map overlap problem (2001) 193–202doi:10.1145/369133.369199.
URL https://doi.org/10.1145/369133.369199

[76] N. K. Fox, S. E. Brenner, J. M. Chandonia, SCOPe: Structural Classification of Proteins–extended, integrating SCOP and ASTRAL data and classification of new structures, Nucleic Acids Res 42 (Database issue) (2014) D304–309.

[77] S. Landolfo, M. Gariglio, G. Gribaudo, D. Lembo, The human cytomegalovirus, Pharmacology & therapeutics 98 (3) (2003) 269—297. doi:10.1016/s0163-7258(03)00034-2.
URL https://doi.org/10.1016/s0163-7258(03)00034-2

[78] B. A. Appleton, A. Loregian, D. J. Filman, D. M. Coen, J. M. Hogle, The cytomegalovirus DNA polymerase subunit UL44 forms a C clamp-shaped dimer, Mol Cell 15 (2) (2004) 233–244.

[79] N. M. O'Boyle, M. Banck, C. A. James, C. Morley, T. Vandermeersch, G. R. Hutchison, Open Babel: An open chemical toolbox, J Cheminform 3 (2011) 33.

[80] G. M. Morris, R. Huey, W. Lindstrom, M. F. Sanner, R. K. Belew, D. S. Goodsell, A. J. Olson, AutoDock4 and AutoDockTools4: Automated docking with selective receptor flexibility, J Comput Chem 30 (16) (2009) 2785–2791.

[81] S. Forli, R. Huey, M. Pique, M. Sanner, D. Goodsell, A. Olson, Computational protein-ligand docking and virtual drug screening with the autodock suite, Nature protocols 11 (2016) 905–919. doi:10.1038/nprot.2016.051.

[82] N. Desai, A. Neyaz, A. Szabolcs, A. R. Shih, J. H. Chen, V. Thapar, L. T. Nieman, A. Solovyov, A. Mehta, D. J. Lieb, A. S. Kulkarni, C. Jaicks, C. J. Pinto, D. Juric, I. Chebib, R. B. Colvin, A. Y. Kim, R. Monroe, S. E. Warren, P. Danaher, J. W. Reeves, J. Gong, E. H. Rueckert, B. D. Greenbaum, N. Hacohen, S. M. Lagana, M. N. Rivera, L. M. Sholl, J. R. Stone, D. T. Ting, V. Deshpande, Temporal and spatial heterogeneity of host response to sars-cov-2 pulmonary infection, medRxivarXiv:https://www.medrxiv.org/content/early/2020/08/02/2020.07.30.20165241.full.pdf, doi:10.1101/2020.07.30.20165241.
URL https://www.medrxiv.org/content/early/2020/08/02/2020.07.30.20165241

[83] Y. Xiong, Y. Liu, L. Cao, D. Wang, M. Guo, A. Jiang, D. Guo, W. Hu, J. Yang, Z. Tang, H. Wu, Y. Lin, M. Zhang, Q. Zhang, M. Shi, Y. Liu, Y. Zhou, K. Lan, Y. Chen, Transcriptomic characteristics of bronchoalveolar lavage fluid and peripheral blood mononuclear cells in covid-19 patients, Emerging Microbes & Infections 9 (1) (2020) 761–770, pMID: 32228226. arXiv:https://doi.org/10.1080/22221751.2020.1747363, doi:10.1080/22221751.2020.1747363.
URL https://doi.org/10.1080/22221751.2020.1747363

[84] Y. Liao, G. K. Smyth, W. Shi, The r package rsubread is easier, faster, cheaper and better for alignment and quantification of rna sequencing reads, Nucleic Acids Research 47 (8) (2019) e47–e47. arXiv:https://academic.oup.com/nar/article-pdf/47/8/e47/28534862/gkz114.pdf, doi:10.1093/nar/gkz114.
URL https://doi.org/10.1093/nar/gkz114

[85] D. Blanco-Melo, B. E. Nilsson-Payant, W.-C. Liu, S. Uhl, D. Hoagland, R. Møller, T. X. Jordan, K. Oishi, M. Panis, D. Sachs, T. T. Wang, R. E. Schwartz, J. K. Lim, R. A. Albrecht, B. R. tenOever, Imbalanced host response to sars-cov-2 drives development of covid-19, Cell 181 (5) (2020) 1036–1045.e9. doi:10.1016/j.cell.2020.04.026.
URL https://doi.org/10.1016/j.cell.2020.04.026

[86] M. I. Love, W. Huber, S. Anders, Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2, Genome Biol 15 (12) (2014) 550.

[87] A. Subramanian, R. Narayan, S. M. Corsello, D. D. Peck, T. E. Natoli, X. Lu, J. Gould, J. F. Davis, A. A. Tubelli, J. K. Asiedu, D. L. Lahr, J. E. Hirschman, Z. Liu, M. Donahue, B. Julian, M. Khan, D. Wadden, I. C. Smith, D. Lam, A. Liberzon, C. Toder, M. Bagul, M. Orzechowski, O. M. Enache, F. Piccioni, S. A. Johnson, N. J. Lyons, A. H. Berger, A. F. Shamji, A. N. Brooks, A. Vrcic, C. Flynn, J. Rosains, D. Y. Takeda, R. Hu, D. Davison, J. Lamb, K. Ardlie, L. Hogstrom, P. Greenside, N. S. Gray, P. A. Clemons, S. Silver, X. Wu, W. N. Zhao, W. Read-Button, X. Wu, S. J. Haggarty, L. V. Ronco, J. S. Boehm, S. L. Schreiber, J. G. Doench, J. A. Bittker, D. E. Root, B. Wong, T. R. Golub, A Next Generation Connectivity Map: L1000 Platform and the First 1,000,000 Profiles, Cell 171 (6) (2017) 1437–1452.

[88] G. Korotkevich, V. Sukhov, A. Sergushichev, Fast gene set enrichment analysis, bioRxivarXiv:https://www.biorxiv.org/content/early/2019/10/22/060012.full.pdf, doi:10.1101/060012.
URL https://www.biorxiv.org/content/early/2019/10/22/060012

[89] G. Yu, L. G. Wang, Y. Han, Q. Y. He, clusterProfiler: an R package for comparing biological themes among gene clusters, OMICS 16 (5) (2012) 284–287.

[90] J. Munkres, Elements of Algebraic Topology, Perseus Publishing, Cambridge, Massachusetts, 1984.

[91] V. Robins, Towards computing homology from finite approximations, in: Topology Proceedings 24 (1), 503–532, 1999.

[92] G. Carlsson, Topological pattern recognition for point cloud data, Acta Numerica 23 (2014) 289–368. doi:10.1017/S0962492914000051.

Tables and Figures

 doi:10.20944/preprints202012.0281.v1

PDB ID	drugbank id	Type	Drug name	Target id	Target name	Protein id	PS_0	PS_1	PS_2	PS Mean
1YYP	DB00369	small molecule	Cidofovir	BE0000343	DNA polymerase catalytic subunit	P08546	0.9265627	0.965132	0.823785072	0.9051599
1YYP	DB00529	small molecule	Foscarnet	BE0000343	DNA polymerase catalytic subunit	P08546	0.9265627	0.965132	0.823785072	0.9051599

Table 1: Proteins with known drugs with a mean similarity > 0.9. $PS_k = k$-Persistent Similarity.

 doi:10.20944/preprints202012.0281.v1

Entry ID	Release Date	Structure Title	Macromolecule Name	Chain ID
6LVN	2020-02-26	Structure of the 2019-nCoV HR2 Domain	Spike protein S2	A, B, C, D
6YI3	2020-04-08	The N-terminal RNA-binding domain of the SARS-CoV-2 nucleocapsid phosphoprotein	Nucleoprotein	A
6M3M	2020-03-18	Crystal structure of SARS-CoV-2 nucleocapsid protein N-terminal RNA binding domain	SARS-CoV-2 nucleocapsid protein	A, B, C, D
6VYO	2020-03-11	Crystal structure of RNA binding domain of nucleocapsid phosphoprotein from SARS coronavirus 2	Nucleoprotein	A, B, C, D
6WJI	2020-04-22	2.05 Angstrom Resolution Crystal Structure of C-terminal Dimerization Domain of Nucleocapsid Phosphoprotein from SARS-CoV-2	SARS-CoV-2 nucleocapsid protein	A, B, C, D, E, F
6LXT	2020-02-26	Structure of post fusion core of 2019-nCoV S2 subunit	Spike protein S2. Spike protein S2	A, B, C, D, E, F
6VSB	2020-02-26	Prefusion 2019-nCoV spike glycoprotein with a single receptor-binding domain up	SARS-CoV-2 spike glycoprotein	A, B, C
6VYB	2020-03-11	SARS-CoV-2 spike ectodomain structure (open state)	Spike glycoprotein	A, B, C
6W41	2020-03-25	Crystal structure of SARS-CoV-2 receptor binding domain in complex with human antibody CR3022	CR3022 Fab heavy chain	H
			CR3022 Fab light chain	L
			Spike protein S1	C
6YLA	2020-04-15	Crystal structure of the SARS-CoV-2 receptor binding domain in complex with CR3022 Fab	Spike glycoprotein	A, E
			Heavy Chain	B, H
			Light chain	C, L

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 11 December 2020 doi:10.20944/preprints202012.0281.v1

Entry ID	Release Date	Structure Title	Macromolecule Name	Chain ID
6M0J	2020-03-18	Crystal structure of SARS-CoV-2 spike receptor-binding domain bound with ACE2	Angiotensin-converting enzyme 2	A
			Spike receptor binding domain	E
6M17	2020-03-11	The 2019-nCoV RBD/ACE2-B0AT1 complex	Sodium-dependent neutral amino acid transporter B(0)AT1	A, C
			Angiotensin-converting enzyme 2	B, D
			SARS-coV-2 Receptor Binding Domain	E, F
6M2Q	2020-04-15	SARS-CoV-2 3CL protease (3CL pro) apo structure (space group C21)	SARS-CoV-2 3CL protease	A
6W4B	2020-03-18	The crystal structure of Nsp9 RNA binding protein of SARS CoV-2	Non-structural protein 9	A, B
6W9Q	2020-04-08	Peptide-bound SARS-CoV-2 Nsp9 RNA-replicase	3C-like proteinase peptide, Non-structural protein 9 fusion	A
6VXS	2020-03-04	Crystal Structure of ADP ribose phosphatase of NSP3 from SARS CoV-2	Non-structural protein 3	A, B
6W9C	2020-04-01	The crystal structure of papain-like protease of SARS CoV-2	Papain-like proteinase	A, B, C
6WCF	2020-04-15	Crystal Structure of ADP ribose phosphatase of NSP3 from SARS-CoV-2 in complex with MES	Non-structural protein 3	A
6WEN	2020-04-15	Crystal Structure of ADP ribose phosphatase of NSP3 from SARS-CoV-2 in the apo form	Non-structural protein 3	A

Entry ID	Release Date	Structure Title	Macromolecule Name	Chain ID
6WIQ	2020-04-22	Crystal structure of the co-factor complex of NSP7 and the C-terminal domain of NSP8 from SARS CoV-2	SARS-CoV-2 NSP7	A
			SARS-CoV-2 NSP8	B
6M71	2020-04-01	SARS-Cov-2 RNA-dependent RNA polymerase in complex with cofactors	SARS-Cov-2 NSP 12	A
			SARS-Cov-2 NSP 8	C
			SARS-Cov-2 NSP 7	B, D
6W01	2020-03-11	The 1.9 A Crystal Structure of NSP15 Endoribonuclease from SARS CoV-2 in the Complex with a Citrate	Uridylate-specific endoribonuclease	A, B
6VWW	2020-03-04	Crystal Structure of NSP15 Endoribonuclease from SARS CoV-2.	Uridylate-specific endoribonuclease	A, B

Table 2: PDB structures of SARS-CoV-2 analyzed in the study.

Drug name	Drug id	PC DS1 (GSE150316)	PC DS2 (CRA002390)	PC DS3 (GSE147507)	Autodock LE (kcal/mol)	Estimate Inhibition Constant (Ki)	Autodock cluster
6M2Q							
Cholic Acid	DB02659	-0.08805090329	-0.1122360022	-0.08069352402	-15.06	9,15 pM	74
Rutin	DB01698	-0.07057525739	-0.1847669519	-0.1024903315	-14.52	22.63 pM	149
Indomethacin	DB00328	-0.06872835755	-0.1234924169	-0.05359016853	-13.31	174.8 pM	146
Sulindac	DB00605	-0.06768358761	-0.1193287796	-0.06631246593	-13.14	231.58 pM	73
Sulfisoxazole	DB00263	-0.05362836277	-0.1325276338	-0.08602283924	-11.59	3.20 nM	77
Dasatinib	DB01254	-0.03959075682	-0.1464677954	-0.09185735897	-10.94	9.53 nM	43
6W01							
Dexamethasone	DB01234	-0.07176892746	-0.1529509474	-0.07847304485	-11.42	4.29 nM	49
Phenolphthalein	DB04824	-0.1299268982	-0.09894230167	-0.03725859228	-11.15	6.74 nM	101
Spironolactone	DB00421	-0.1200039446	-0.1046496564	-0.08907503297	-10.99	8.76 nM	110
Mifepristone	DB00834	-0.1336064601	-0.1414272927	-0.06288002847	-10.04	47.58 nM	28
Carbamazepine	DB00564	-0.08046486733	-0.1427960558	-0.07447221905	-9,66	83.48 nM	86
6M71							
Vemurafenib	DB08881	-0.09056824953	-0.1586436186	-0.07981160154	-8,09	1.17 uM	13
Sorafenib	DB00398	-0.1139503326	-0.1488412632	-0.0528631983	-7,34	4.14 uM	30
Levonorgestrel	DB00367	-0.08286521718	-0.1387530571	-0.07915695554	-7,21	5.20 uM	89
Naloxone	DB01183	-0.05572216872	-0.1159793987	-0.08947993744	-7,07	6.57 uM	69
Raloxifene	DB00481	-0.1278104657	-0.1700041351	-0.0726603481	-7,05	6.76 uM	6

Table 3: Crossover of the transcriptomics and molecular docking results of 6M2Q, 6W01 and 6M71 viral proteins and most promising candidates. PC: Pearson correlation. LE: Lowest energy conformation in the cluster. Candidates with a PC < -0.1 may revert the transcriptomic effects of SARS-CoV-2 infection. Maximum number of the autodock cluster is 150.

 doi:10.20944/preprints202012.0281.v1

	$[x_0, x_1]$	$[x_0, x_2]$	$[x_0, x_3]$	$[x_1, x_2]$	$[x_1, x_3]$	$[x_2, x_3]$
$[x_0]$	**1**	**1**	**1**	**0**	**0**	**0**
$[x_1]$	**1**	**0**	**0**	**1**	**1**	**0**
$[x_2]$	**0**	**1**	**0**	**1**	**0**	**1**
$[x_3]$	**0**	**0**	**1**	**0**	**1**	**1**

Table 4: The incidence matrix between the 1-level and the 0-level of $\mathcal{M}$.

	$[x_0, x_1, x_2]$	$[x_0, x_1, x_3]$	$[x_0, x_2, x_3]$	$[x_1, x_2, x_3]$
$[x_0, x_1]$	**1**	**1**	**0**	**0**
$[x_0, x_2]$	**1**	**0**	**1**	**0**
$[x_0, x_3]$	**0**	**1**	**1**	**0**
$[x_1, x_2]$	**1**	**0**	**0**	**1**
$[x_1, x_3]$	**0**	**1**	**0**	**1**
$[x_2, x_3]$	**0**	**0**	**1**	**1**

Table 5: The incidence matrix between the 2-level and the 1-level of $\mathcal{M}$.

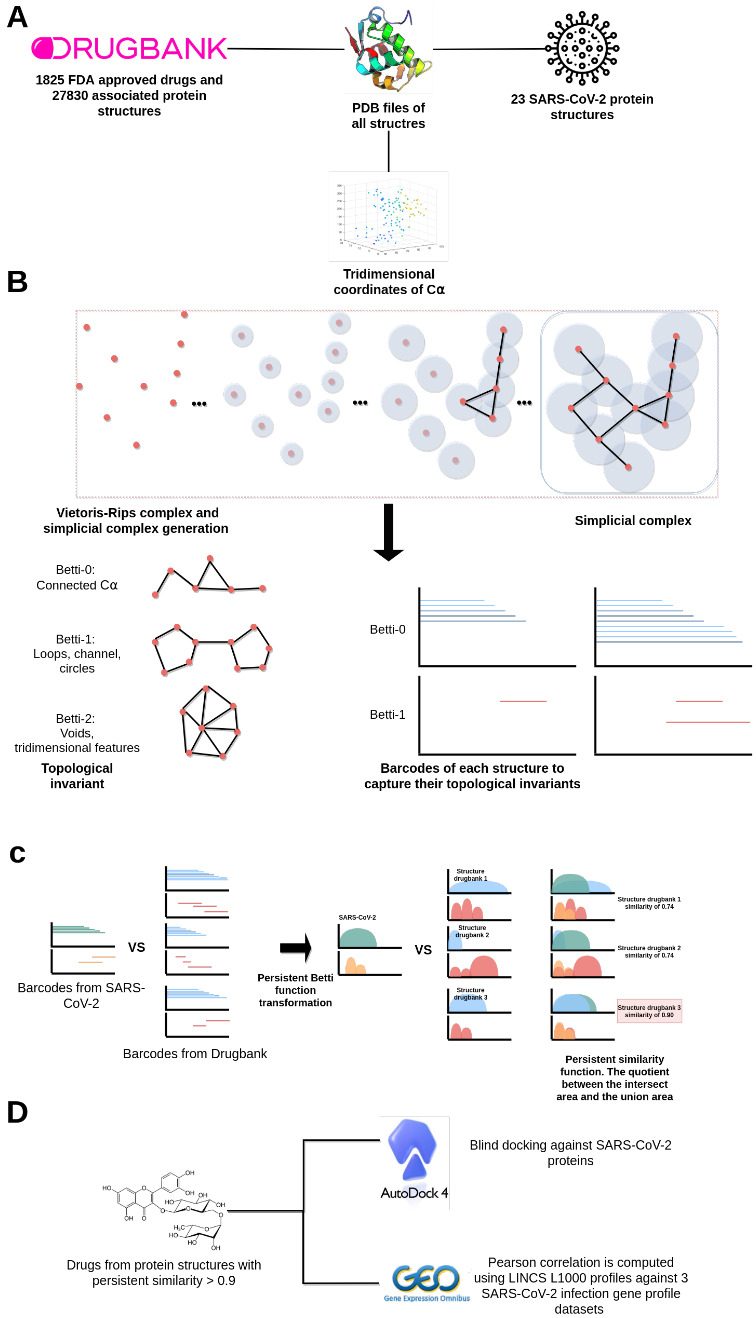

Figure 1: Bioinformatic work-flow used. (**A**) Data preprocessing and acquisition (**B**) Topological data analysis phase, Vietoris-Rips complexes at scale ε are computed to generate the barcodes. Each ε-associated Betti number captures a unique topological feature of the protein. (**C**) In order to compare barcodes of viral proteins against structures with known drugs is necessary to transform barcodes into comparable curves using PBF. (**D**) Candidate drugs from proteins with a mean Persistent Similarity score above 0.9 will be validated by a dual in-silico strategy. We use autodock 4 to analyse the capacity of the drug to bind against viral proteins. Transcriptomics analysis is performed to test the capacity of the candidate drugs to revert the transcriptomics effect induced by the COVID-19.

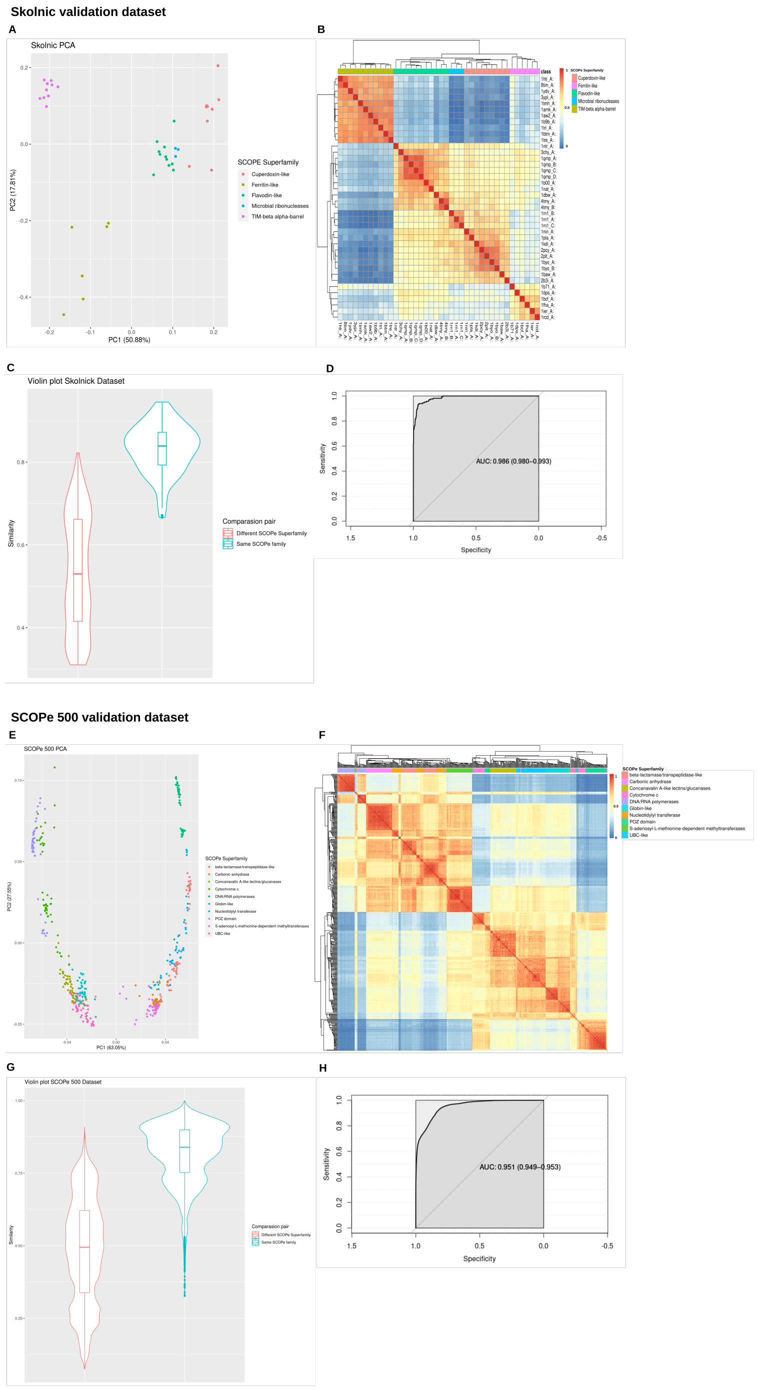

Figure 2: Results of validation of the Betti persistence Betti function with Skolnick and SCOPe datasets (**A**) Clustering results of persistent Betti function and TDA protocol of Skolnick dataset. (**B**) Principal component analysis of the Skolnick dataset. (**C**) Violin plot of the protein structures pairs that share the same super-family classification against protein pairs of different super-families. (**D**) ROC curve between same super-families pairs against different super-families protein pairs in the Skolnick dataset. (**E**) Principal component analysis of the pairwise matrix of SCOPe dataset. (**F**) Heat-map of the SCOPe subset dataset, warm colours are protein-protein pairs that have a high degree of structural similarity, while cool colours represent pairs with low degree of similarity. (**G**) Violin plot of the protein structures pairs that share the same super-family SCOPe classification against protein pairs of different super-families (t-test p-value $< 9.484e-6$) . (**H**) ROC curve between same super-families pairs against different super-families protein pairs (AUC: 0.986 (0.980-0.993)).

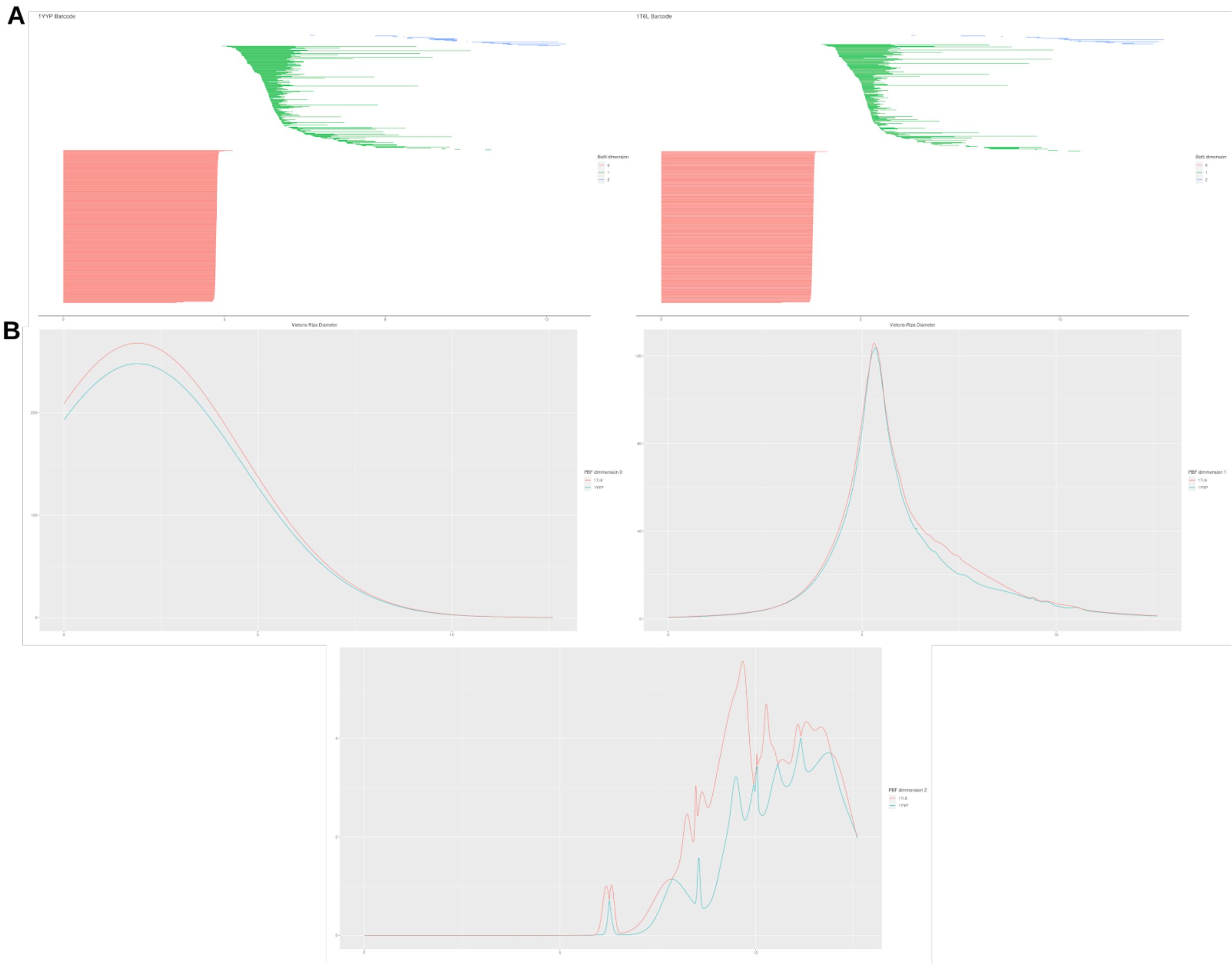

Figure 3: Cytomegalovirus DNA polymerase TDA results. (**A**) Multiple Betti dimension Barcodes of the proteins structures 1TL6 (positive control) and 1YYP. (**B**) Persistent Similarity comparison between 1TL6 and 1YYP structures. The 0-Persistent Similarity is 0.926, 1-Persistent Similarity is 0.965,2-Persistent Similarity is 0.823.

doi:10.20944/preprints202012.0281.v1

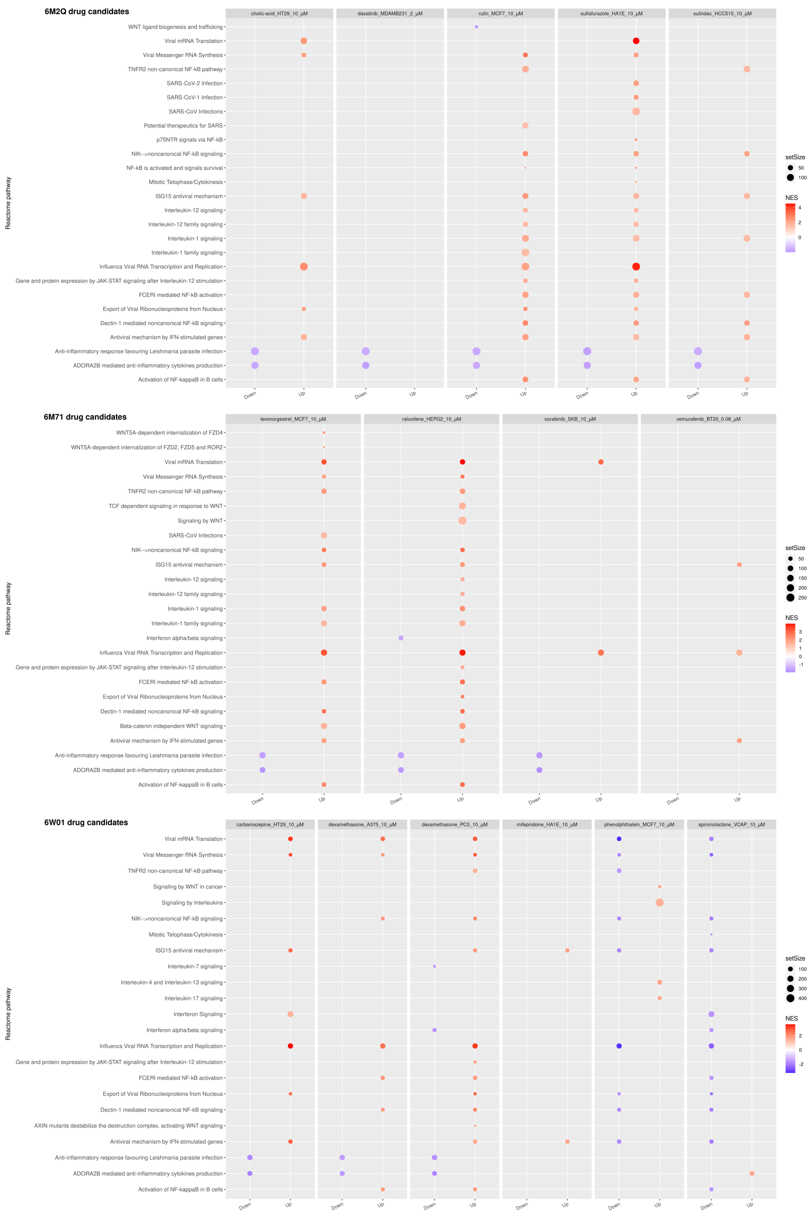

Figure 4: GSEA results for candidate drugs for 6M2Q, 6M71 and 6W01 SARS-CoV-2 structures with the expression signatures yields from correlation analyses from DS2. Reactome pathways related to the immune system and viral infections. Only drugs with at least one pathway with a p-value adjusted < 0.05 are displayed. GSEA table with the results is available in supplementary table XX

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 11 December 2020 doi:10.20944/preprints202012.0281.v1

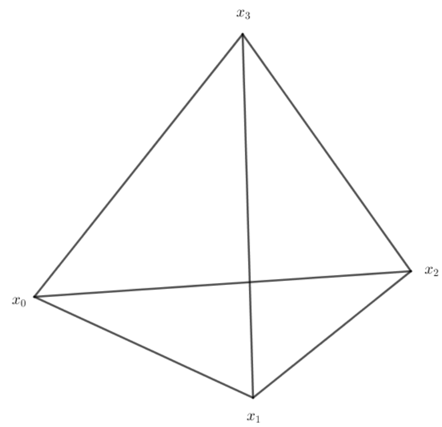

Figure 5: A tetrahedron defined by a set of points $\mathbb{M} = \{x_0, x_1, x_2, x_3\} \subset \mathbb{R}^3$.

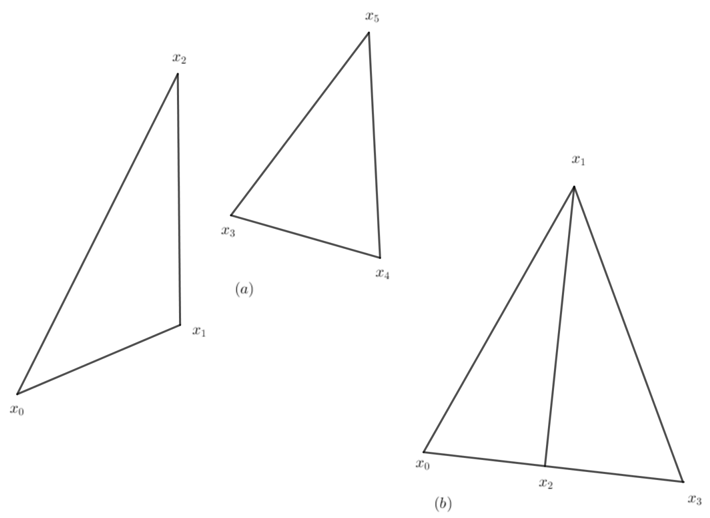

Figure 6: In (a) we have the union $[x_0, x_1, x_2] + [x_3, x_4, x_5]$ and in (b) $[x_0, x_1, x_2] + [x_1, x_2, x_3]$.

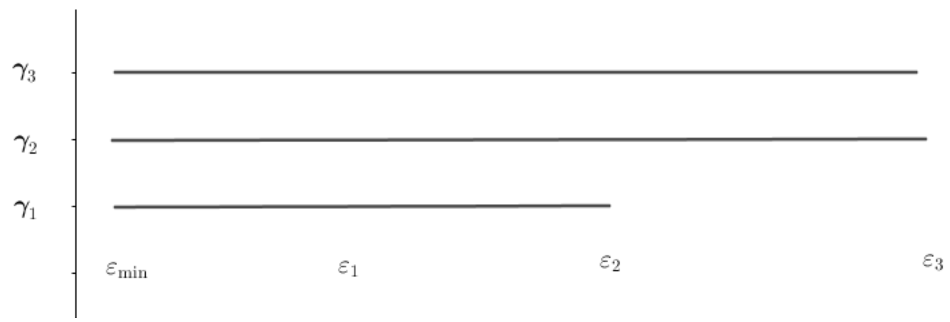

Figure 7: A graphical representation of the barcodes for a set of k-features $\gamma_1 < \gamma_2 < \gamma_3$ for a given partition $\mathcal{P} = \{\varepsilon_j\}_{j=0}^m$.