Bioactivity profile similarities to expand the repertoire of COVID-19 drugs

Miquel Duran-Frigola[1,†], Martino Bertoni[1], Eduardo Pauls[1], Víctor Alcalde[1], Gemma Turon[1], Núria Villegas[1], Adrià Fernández-Torras[1], Carles Pons[1], Lídia Mateo[1], Oriol Guitart-Pla[1], Pau Badia-i-Mompel[1], Aleix Gimeno[1], Nicolas Soler[1], Isabelle Brun-Heath[1], and Patrick Aloy[1,2,†]

1. Joint IRB-BSC-CRG Program in Computational Biology, Institute for Research in Biomedicine (IRB Barcelona), The Barcelona Institute of Science and Technology, Barcelona, Catalonia, Spain
2. Institució Catalana de Recerca i Estudis Avançats (ICREA), Barcelona, Catalonia, Spain

† Corresponding authors: miquel.duran@irbbarcelona.org; patrick.aloy@irbbarcelona.org

Abstract

Until a vaccine becomes available, the current repertoire of drugs is our only therapeutic asset to fight the SARS-CoV-2 outbreak. Indeed, emergency clinical trials have been launched to assess the effectiveness of many marketed drugs, tackling the decrease of viral load through several mechanisms. Here, we present an online resource, based on small-molecule bioactivity signatures and natural language processing, to expand the portfolio of compounds with potential to treat COVID-19. By comparing the set of drugs reported to be potentially active against SARS-CoV-2 to a universe of 1M bioactive molecules, we identify compounds that display analogous chemical and functional features to the current COVID-19 candidates. Searches can be filtered by level of evidence and mechanism of action, and results can be restricted to drug molecules or include the much broader space of bioactive compounds. Moreover, we allow users to contribute COVID-19 drug candidates, which are automatically incorporated to the pipeline once per day. The computational platform, as well as the source code, is available at https://sbnb.irbbarcelona.org/covid19.

Introduction

A new coronavirus, named SARS-CoV-2, is the responsible agent for the current 2019-2020 viral pneumonia (COVID-19) outbreak[1,2], which is already affecting millions of people worldwide and causing hundreds of thousands of deaths. The COVID-19 pandemic has prompted an unprecedented effort by the scientific community to understand its molecular constituents and find an effective treatment to mitigate viral infectiveness and symptoms. This is reflected in the over 6,000 COVID-related publications that appeared in the last weeks[3]. Huge efforts are being invested in the discovery of an effective vaccine, but even the most optimistic scenarios suggest that it will not be available until 2021. Other drug discovery projects have been launched to target specific viral proteins, particularly its main protease (Mpro)[4]. However, these initiatives, even if successful, could take even longer to deliver an approved drug. Thus, the repurposing of existing drugs is our best chance to face the current outbreak therapeutically, since approved drugs have known safety profiles and are ready to be tested in humans. For instance, several compounds initially developed to treat HIV (e.g. lopinavir/ritonavir)[5] or Ebola (e.g. remdesivir)[6], as well as antimalarial drugs (e.g. hydroxychloroquine)[7], are being tested against COVID-19. Indeed, we conducted a limited review of the most relevant scientific literature and identified over 200 compounds that are potentially active against COVID-19 with different levels of experimental support; from purely computational predictions, to preclinical and drugs already in clinical trials.

We now exploit this literature mining effort to identify other compounds with the potential to be effective against COVID-19. To this aim, we use the Chemical Checker (CC), a resource that provides processed, harmonized and integrated bioactivity data for about 1M small molecules[8]. In the CC, bioactivity data are expressed in a vector format, which naturally extends the notion of chemical similarity between compounds to similarities between bioactivity profiles. The CC organizes data into five levels of increasing complexity, ranging from drug binding profiles to clinical outcomes, and thus enables similarity searches that should be mechanistically and clinically relevant.

In the current resource, we use CC signatures to identify similarities between bioactive compounds and the list of current COVID-19 drug candidates (i.e. *bait* compounds). The similarity search is performed systematically across the large chemical space encompassed by the CC, thereby substantially expanding the portfolio of potential molecules effective against SARS-CoV-2. Results are stratified between drug molecules and a broader medicinal chemistry space, thus offering ranked lists of compounds that should be of value for drug repurposing endeavours as well as preclinical screening campaigns.

Methodological strategy

Our resource capitalizes on an ongoing literature curation effort made by our group. Additionally, we welcome contributions from the broader scientific community via web form, allowing users to include compounds under investigation in their labs, or to update the evidence level as new COVID-19 experiments accumulate. The scientific evidence supporting COVID-19 drug candidates is variable: some compounds come from computational predictions, some have proven their value in pre-clinical tests, others are approved drugs with a therapeutic indication unrelated to infectious diseases and, finally,

some are drugs currently used to fight SARS-CoV-2-related pathogens. The mechanisms of action (MoA) suggested to confer efficacy are also variable, ranging from immunomodulators to protease inhibitors. During curation, we classify literature COVID-19 candidates by their level of evidence and MoA (Figure 1). By 18th of April (2020), we have found that 230 small-molecules have been suggested as potential treatments for COVID-19.

Starting from the SMILES representation of a compound, we derive CC bioactivity signatures for each COVID-19 literature bait compound. We then run bioactivity similarity searches against the ~1M bioactive molecules characterized in the CC, and keep the top 10,000 most similar compounds for each search type. Likewise, we conduct conventional similarity searches solely based on 2D representations of the compounds (2048-bit Morgan fingerprints, radius 2). Similarities are expressed as empirical P-values (-log10 scale) derived from the expected similarity distribution across the full search space. A simple *support* measure is provided for each compound by adding up the number of similar COVID-19 drugs (weighted by -log10 P-value and level of evidence, as shown in Figure 1).

Besides, we complement our literature curation effort with a further level of evidence, namely text-mining, based on the automatic detection of experiments (bioassays) that could be relevant to COVID-19. More specifically, we process the text description of the ~1,2M bioassays catalogued in the ChEMBL database, and rank them according to their relevance to the current corpus of about 30,000 articles related to COVID-19 and other coronavirus infections[9]. ChEMBL bioassays[10] are ranked using two complementary approaches. (i) We construct a retrieval query from the bioassay descriptions, and use it to score each of the paragraphs and abstracts contained in the articles collection. We then use statistics of the score distribution of top scoring documents to rank the bioassays. (ii) We manually labeled a set of (*seed*) entities that tested positive in ~100 bioassays relevant to COVID-19. We then identify automatically entities from all the bioassay descriptions, and compute their contextual embeddings. Finally, we rank the bioassays according to their cosine similarity to the seed entities. After completion of (i) and (ii), we identify those bioactive molecules in the CC that tested positive (<10 μM) in at least one of the top 1,000 COVID-19 literature bioassays, in either text-mining approach. We then cross these results with the 10,000 compounds obtained from the similarity searches described above, and assign an extra literature-evidence level (text-mining) to those in common, which are then used as bait compounds.

The pipeline runs automatically every day, so that we always provide the most updated results. Searches are pre-computed for each evidence strength and MoA.

The resource

Results of the large-scale similarity search are made available as a web-resource at https://sbnb.irbbarcelona.org/covid19. The interface contains five tabs:

Candidates. We provide the 10,000 molecules, within the CC universe of 1M bioactive compounds, that are more similar to the COVID-19 bait compounds collected from the literature (Figure 2). The pre-computed similarity matrix can be queried to extract candidates that fulfil properties of interest by selecting amongst the levels of evidence for the bait compounds as well as their MoA. Besides, the resulting list of molecules can be sorted

following different criteria, including whether they are approved/experimental drugs, the cumulative level of support, or their similarity to specific COVID-19 literature drugs. Full and partial tables can be downloaded and exported to several formats.

Literature. This tab lists the COVID-19 bait compounds extracted from the literature, together with their level of experimental evidence and, if known, the MoA that confers efficacy against SARS-CoV-2.

Documentation. Here we present a brief description of the methodological strategy and, more importantly, we offer updated statistics and benchmarks of the resource. In particular, we quantify the number of literature bait compounds available at each level of evidence and MoA (Figure 3A-B), and project CC signatures on a 2D plane to offer a global view of the chemical space explored by our resource (Figure 3C-D). We see that, while significantly diverse, COVID-19 bait compounds cluster in certain regions of the chemical space, and we find new candidate molecules in their vicinity. Reassuringly, when we analyse the therapeutic categories of the top-ranked candidates, as expected, we retrieve a significant number of antiinfectives and antiparasitic drugs (Figure 4A). Other therapeutic categories such as hormonal treatments are enriched after the highest-ranking compounds. Note that, for this enrichment analysis, only drug molecules could be considered since ATC annotations are not available for most of the compounds in the CC. Finally, we perform a leave-one-out cross-validation to assess whether bait compounds can be retrieved by our similarity search. Figure 4B shows that known COVID-19 drugs are significantly up-ranked when using and evaluating all levels of evidence (Figure 4B).

Contribute. Through this form, users can contribute to the resource by including their molecules of interest. We require the name and SMILES representation of the molecules as well as their level of experimental evidence, MoA and references, if available.

Code. Link to the Gitlab repository containing the complete code to run the pipeline and analyse results.

Overall, we believe that the tool presented herein explores regions of the bioactive chemical space that could be relevant to COVID-19 treatment. Our web-based resource is updated daily and can be used to dynamically search for candidates related to COVID-19 drugs with varying levels of evidence and MoA. Therefore, our resource will be useful to a broad range of COVID-19 drug discovery approaches, ranging from those seeking a repurposing opportunity to those departing from the *in vitro* screening of compounds.

Acknowledgements

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 101003633 (RiPCoN). We would like to thank Roi Blanco, Victor Martinez and Hugo Zaragoza for their contribution to the mining of the COVID-19 literature.

References

1. Zhu, N.; Zhang, D.; Wang, W.; Li, X.; Yang, B.; Song, J.; Zhao, X.; Huang, B.; Shi, W.; Lu, R.; Niu, P.; Zhan, F.; Ma, X.; Wang, D.; Xu, W.; Wu, G.; Gao, G. F.; Tan, W.; China Novel Coronavirus, I.; Research, T., A Novel Coronavirus from Patients with Pneumonia in China, 2019. *N Engl J Med* **2020**, 382, 727-733.

2. Wu, F.; Zhao, S.; Yu, B.; Chen, Y. M.; Wang, W.; Song, Z. G.; Hu, Y.; Tao, Z. W.; Tian, J. H.; Pei, Y. Y.; Yuan, M. L.; Zhang, Y. L.; Dai, F. H.; Liu, Y.; Wang, Q. M.; Zheng, J. J.; Xu, L.; Holmes, E. C.; Zhang, Y. Z., A new coronavirus associated with human respiratory disease in China. *Nature* **2020**, 579, 265-269.

3. https://search.bvsalud.org/global-research-on-novel-coronavirus-2019-ncov/.

4. Jin, Z.; Du, X.; Xu, Y.; Deng, Y.; Liu, M.; Zhao, Y.; Zhang, B.; Li, X.; Zhang, L.; Peng, C.; Duan, Y.; Yu, J.; Wang, L.; Yang, K.; Liu, F.; Jiang, R.; Yang, X.; You, T.; Liu, X.; Yang, X.; Bai, F.; Liu, H.; Liu, X.; Guddat, L. W.; Xu, W.; Xiao, G.; Qin, C.; Shi, Z.; Jiang, H.; Rao, Z.; Yang, H., Structure of M(pro) from COVID-19 virus and discovery of its inhibitors. *Nature* **2020**.

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

6. Grein, J.; Ohmagari, N.; Shin, D.; Diaz, G.; Asperges, E., Compassionate Use of Remdesivir for Patients with Severe Covid-19. *N Engl J Med* **2020**.

7. Lover, A. A., Quantifying treatment effects of hydroxychoroquine and azithromycin for COVID-19: a secondary analysis of an open label non-randomized clinical trial. *medRxiv* **2020**.

8. Duran-Frigola, M.; Pauls, E.; Guitart-Pla, O.; Bertoni, M.; Alcalde, V.; Amat, D.; Juan-Blanco, T.; Aloy, P., Extending the small molecule similarity principle to all levels of biology. *Nat Biotechnol* **2020**, In press.

9. https://allenai.org/data/cord-19.

10. Mendez, D.; Gaulton, A.; Bento, A. P.; Chambers, J.; De Veij, M.; Felix, E.; Magarinos, M. P.; Mosquera, J. F.; Mutowo, P.; Nowotka, M.; Gordillo-Maranon, M.; Hunter, F.; Junco, L.; Mugumbate, G.; Rodriguez-Lopez, M.; Atkinson, F.; Bosc, N.; Radoux, C. J.; Segura-Cabrera, A.; Hersey, A.; Leach, A. R., ChEMBL: towards direct deposition of bioassay data. *Nucleic Acids Res* **2019**, 47, D930-D940.

Figures

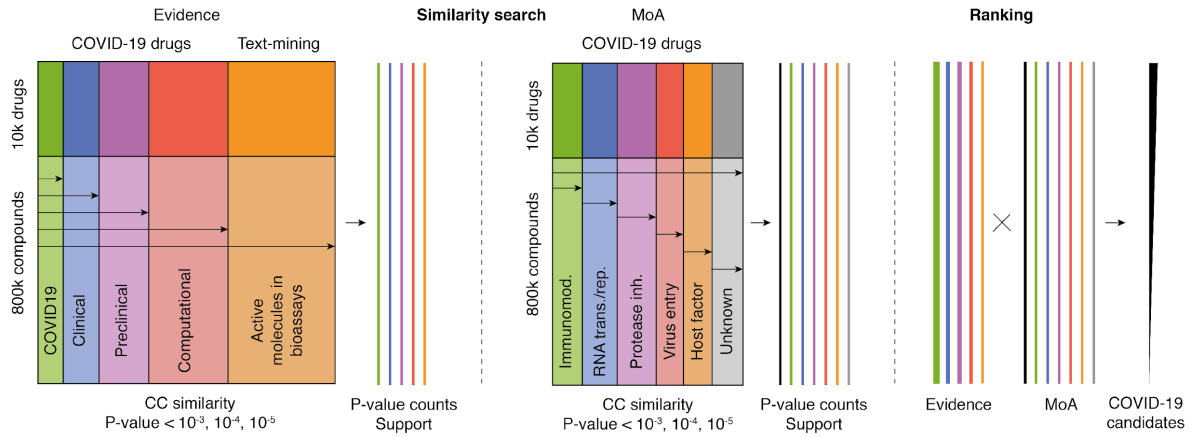

Figure 1. *Methodological strategy*

We use the list of COVID-19 compounds extracted from the literature, with different levels of experimental evidence, as bait to search for compounds with similar bioactivity or chemical features among the 800k molecules contained in the CC. We keep and rank the top 10,000 most similar molecules to the array of bait compounds. Ranking is given by a "support" score that results from the weighted sum (count) of the number of similar baits to the query molecule. Evidence weights (w_e): COVID-19=5, Clinical=4, Preclinical=3, Computational=2, Text-mining=1. Similarity weights (w_s): P-value 10^{-5}=3, 10^{-4}=2, 10^{-3}=1.

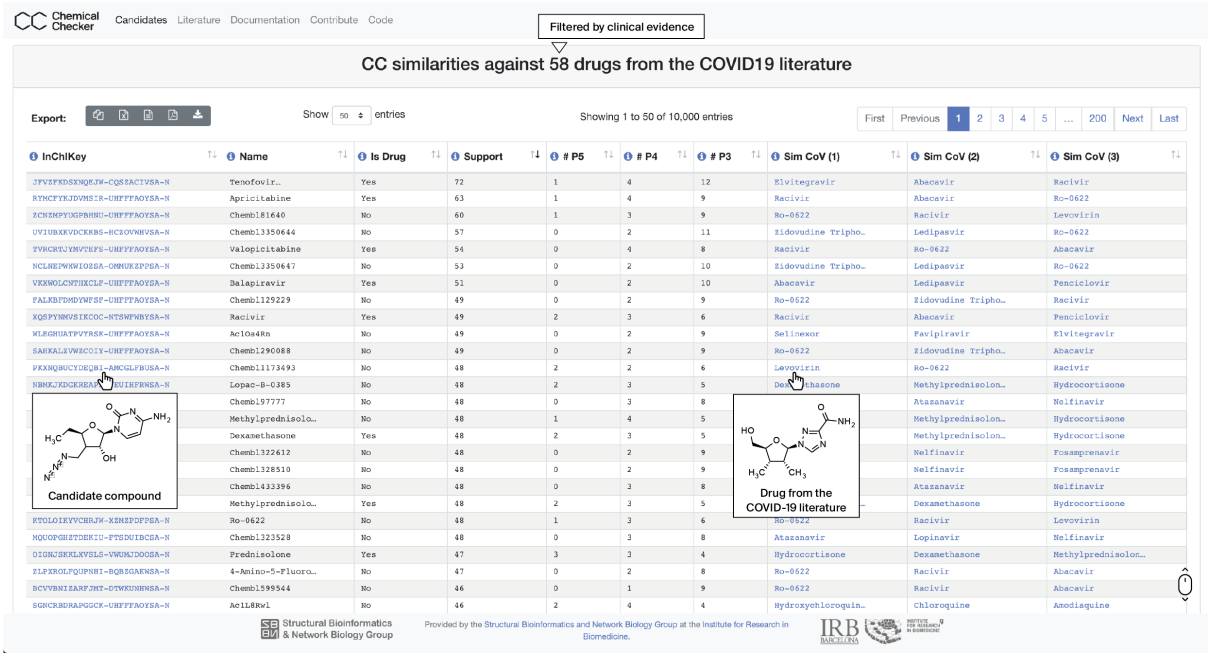

Figure 2. *Querying the compound similarity matrix*

Pre-computed similarity matrices can be queried to extract candidates with the properties of interest. The dynamic tables show information about each candidate compound: InChIKey, name, whether it is a known drug, its level of support, number of similar COVID-19 bait

compounds (P-values 10^{-5}, 10^{-4}, 10^{-3}) and, of those, the name of the three most similar ones. Additionally, for each molecule, we provide its structure and links to the corresponding CC page.

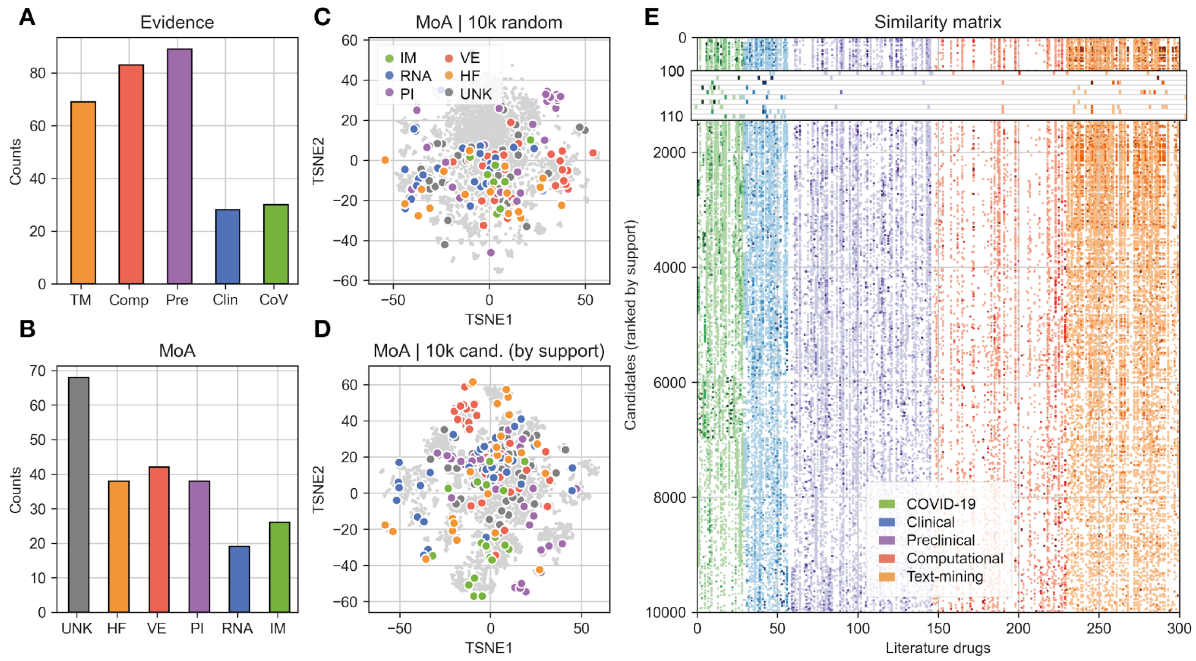

Figure 3. *COVID-19 literature bait compounds composition and functional diversity*

Number of literature bait compounds split according to (A) their level of experimental evidence or (B) MoA. (C) t-SNE projections of the bait compounds on the global space of bioactive CC molecules and on the top 10k candidate compounds (D); bait compounds are coloured by MoA. (E) A global view on the similarity matrix, stratified by level of evidence. The inset zooms into ten exemplary rows (ranking 100-110).

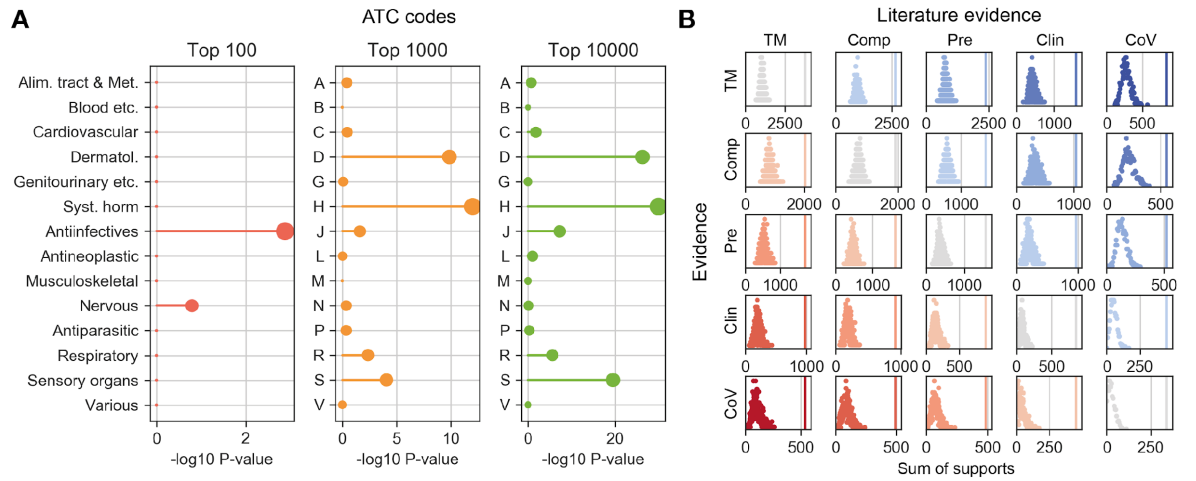

Figure 4. *Benchmark of the strategy*

(A) Enrichment analysis of therapeutic areas (ATC categories) among the top ranked candidate compounds (top 100, 1,000 and 10,000). (B) Leave-one-out cross-validation to assess whether compounds at different levels of evidence (rows) are retrieved by our

similarity search using the COVID-19 bait literature drugs (columns). The vertical line indicates the sum of support for observed candidates, and distributions represent the background expectation of the search.