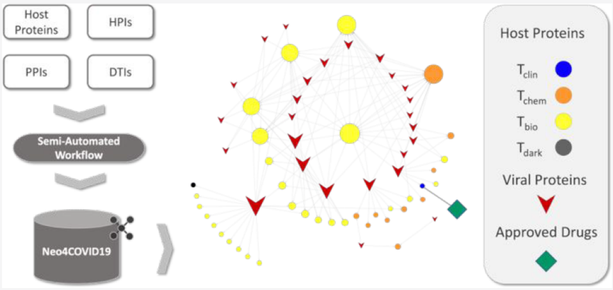
JCIM JOURNAL OF CHEMICAL INFORMATION AND MODELING

# A Workflow of Integrated Resources to Catalyze Network Pharmacology Driven COVID-19 Research

Gergely Zahoránszky-Kőhalmi,* Vishal B. Siramshetty, Praveen Kumar, Manideep Gurumurthy, Busola Grillo, Biju Mathew, Dimitrios Metaxatos, Mark Backus, Tim Mierzwa, Reid Simon, Ivan Grishagin, Laura Brovold, Ewy A. Mathé, Matthew D. Hall, Samuel G. Michael, Alexander G. Godfrey, Jordi Mestres, Lars J. Jensen, and Tudor I. Oprea*

Cite This: *J. Chem. Inf. Model.* 2022, 62, 718−729 Read Online

ACCESS | Metrics & More | Article Recommendations | Supporting Information



**ABSTRACT:** In the event of an outbreak due to an emerging pathogen, time is of the essence to contain or to mitigate the spread of the disease. Drug repositioning is one of the strategies that has the potential to deliver therapeutics relatively quickly. The SARS-CoV-2 pandemic has shown that integrating critical data resources to drive drug-repositioning studies, involving host−host, host−pathogen, and drug−target interactions, remains a time-consuming effort that translates to a delay in the development and delivery of a life-saving therapy. Here, we describe a workflow we designed for a semiautomated integration of rapidly emerging data sets that can be generally adopted in a broad network pharmacology research setting. The workflow was used to construct a COVID-19 focused multimodal network that integrates 487 host−pathogen, 63 278 host−host protein, and 1221 drug−target interactions. The resultant Neo4j graph database named "Neo4COVID19" is made publicly accessible via a web interface and via API calls based on the Bolt protocol. Details for accessing the database are provided on a landing page (https://neo4covid19.ncats.io/). We believe that our Neo4COVID19 database will be a valuable asset to the research community and will catalyze the discovery of therapeutics to fight COVID-19.

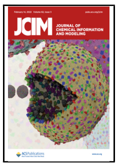
## INTRODUCTION

The pandemic of the SARS-CoV-2 virus (also commonly referred to as COVID-19 pandemic by the name of the disease it induces) put a spotlight on the need for mechanisms that can rapidly identify and integrate relevant information to give a fighting chance to the medical and research community. The Ebola and Zika outbreaks presented similar challenges, and significant advances have been made in the past years in terms of digital, laboratory, and epidemiological techniques.[1−3] Although numerous resources covering many aspects of pertinent biomedical research have been developed and made publicly available, for example, ChEMBL,[4,5] Reactome,[6] Pharos,[7] PathwayCommons,[8] BioPlanet,[9] and DrugCentral,[10] to name a few, their on-demand integration has been a translational bottleneck to date.

In the case of the outbreak of an unknown pathogen, such as SARS-CoV-2, the first line of defense (absent viable therapeutic options) is containment. Should this fail, we have to resort to mitigation to slow down the spread of the pathogen. Delays of mere days in the early stages of containment and mitigation can lead to catastrophic outcomes regarding the number of infections and death toll.[11−13] Furthermore, the COVID-19 pandemic was reported to have a dramatic impact on primary healthcare providers, limiting them to essential clinical services, which eventually led to unforeseen delays in diagnosing highly critical diseases such as cancers.[14,15] Therefore, it is imperative that we have computational workflows in place that can help researchers to connect and navigate the relevant information very fast, much faster than today. Multiple pertinent data sets subjected to such a workflow would produce a condensed, enriched starting point for timely hypothesis generation that would drive a successful containment or mitigation strategy, such as drug repositioning.[16]

Indeed, a small number of publicly available databases and knowledge graphs have been reported that connect various types of information related to COVID-19. Such resources are primarily limited to data extracted via text mining from literature and patents, reports, and experimental data.[17−27] While these resources are valuable for advancing our efforts

**Received:** April 22, 2021
**Published:** January 20, 2022



https://doi.org/10.1021/acs.jcim.1c00431

toward a possible therapy for COVID-19, they suffer from certain limitations from a translational point of view.

First, hypothesis generation in the current drug discovery paradigm, that is, network pharmacology,[28−30] can be enhanced by normalizing the nature of interactions between protein target pairs and between compounds and targets to reflect whether they are engaged in a stimulatory or an inhibitory relationship. Although necessary information might be present in existing data sources, the interaction categories are typically encoded as separate relationships. This makes it difficult to readily assess the inhibitory and stimulatory relationships. For instance, both "antagonist" and "ion channel blocker" actions can be further reduced to an "inhibitory" relationship to aid the analysis of network perturbation. Naturally, the original relationships can also be preserved to avoid loss of information. A normalized relationship of protein−protein and drug−target interactions would consist of only three values: "stimulates", "inhibits", and "undefined" (or the equivalent of these phrases).

Next, the recent concept of "target development level (TDL)"[7] of protein targets is not captured in existing knowledge graphs related to COVID-19. The annotation of TDL category of targets makes it easy to identify those whose activity can be modulated by FDA-approved drugs or by small molecules. Such information therefore facilitates drug repositioning oriented hypothesis generation.

Another category of limitations pertains to translational aspects: knowledge dissemination and real-time data integration. To our knowledge, no data source related to COVID-19 to date has been equipped with a mechanism to facilitate the data exploration and analysis for those without and with substantial bioinformatics background at the same time. However, in the case of a pandemic, it is imperative to disseminate data sources and data analysis tools to as broad a scientific community as possible and as soon as possible.

Finally, from a technical standpoint, it is of paramount importance that we have publicly available mechanisms and workflows for real-time integration of heterogeneous information. Typically, careful integration of databases can take months and even years, and oftentimes the integration workflow is quite specific for the knowledge base at hand.[31,32]

In this study, we describe such a workflow and utilize it to assemble, from several pertinent sources, a knowledge network aimed at defeating COVID-19 via enhanced hypothesis generation. The first building block is a list of pathogen−host protein interactions that was published in a preprint on March 27, 2020, by Gordon et al.[33,34] within two months of the disclosure of the SARS-CoV-2 genetic sequence and within 2 weeks of the WHO declaring COVID-19 a pandemic.[35,36] Further building blocks encompass interactions between FDA approved drugs and host protein targets (DTIs), host protein−protein and host−pathogen protein−protein interactions (PPIs and HPIs, respectively), and predicted drugs and host targets that will be introduced in this study. While this particular resultant Neo4j database is intended to catalyze COVID-19 research, the process of its creation can serve as a blueprint for inevitable future outbreaks, translating to saving precious time and, consequently, lives.

**Related Work.** The workflow presented in this study was inspired by earlier works, namely, SmartGraph[37] and Hetionet,[38] both of which are so-called multimodal networks designed to aid drug discovery efforts in a network pharmacology setting. SmartGraph is a computational platform that consists of a knowledge base and a web-based user interface. The knowledge base of SmartGraph integrates drug−target and protein−protein interactions. The user interface integrates complex but easy-to-execute workflows for the analysis of network perturbation and for bioactivity prediction and drug repositioning. Relationships between proteins are labeled as stimulatory or inhibitory to reflect the nature of the interactions. Hetionet is an open-source resource that is of notable importance and relevance. A wide variety of publicly available biomedical, disease-specific, and pharmacological databases were integrated into a large network. Unlike SmartGraph, Hetionet annotates drug−target interactions in terms of pharmacological action. Nevertheless, neither SmartGraph nor Hetionet provide a normalized annotation of the stimulatory and inhibitory relationships between proteins or between drugs and targets.

In a recent effort, researchers from University of Minnesota, Hunan University, and Amazon Web Service (AWS) artificial intelligence laboratories have collectively built the COVID-19-related Drug Repurposing Knowledge Graph (DRKG)[25] that integrates Hetionet among other data sources. Although DRKG addresses, to some extent, the issue of knowledge dissemination, this particular resource is available only as data tables, a format that requires advanced bioinformatics skills for analyzing the data in terms of network perturbation caused by drugs. Furthermore, the workflow to assemble the DRKG is not publicly available. COVID-KG[24] is another resource that focuses on extracting multimedia knowledge elements from scientific literature to be used as a knowledge graph for querying and report generation. While researchers developing both DRKG and COVID-KG have addressed the issue of data dissemination to some extent, the format they chose for distribution (tab-separated flat files) restricts network-based data analytics absent advanced bioinformatics skills. Furthermore, neither DRKG nor COVID-KG are associated with a publicly available implementation reflective of their underlying data integration workflows.

Another COVID-19 related resource is the recent "COVID-19 Disease Map" that was created by means of a collaborative effort.[39,40] While this study addresses the importance of the standardization of file formats, it does not provide a flexible workflow (and implementation as source code) that could serve as the basis for an automated or semiautomated data integration process. Although at the time writing of this manuscript the COVID-19 Disease Map is not yet accessible, the description of the integration process indicates that much emphasis was put on the use of curated data sources. This study exemplifies that a collaborative data integration strategy can yield high quality data sources, however, at the sacrifice of time.

Apart from KGs and databases, several other studies identified potential targets and contributed useful data sets that can guide drug discovery efforts. For instance, Gil et al.[17] provided their perspective on the main targets that are involved in viral replication and control of host cellular processes. In parallel, structure-based efforts[19,20] have been reported where the authors performed molecular docking simulations and virtual high-throughput screening to prioritize candidates for drug repurposing. Experimental high-throughput screening data have been made available by the National Center for Advancing Translational Sciences (NCATS/NIH) on the "OpenData Portal".[21]

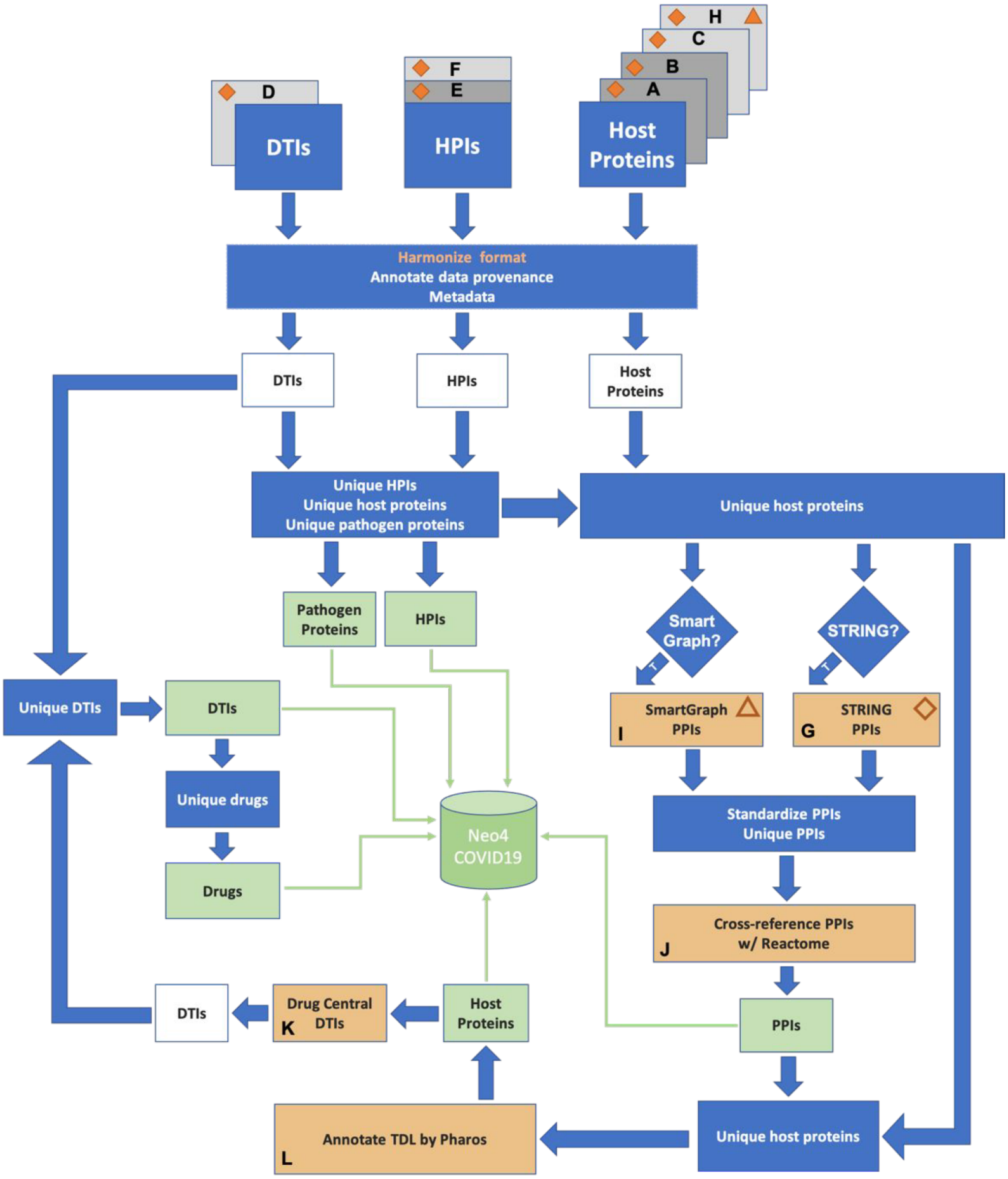


**Figure 1.** Resource integration logic. The schema highlights the most important steps of data processing. Individual inputs are labeled with letters. PPIs, host−host protein interactions; HPIs, host−pathogen (here: SARS-CoV-2) protein interactions; DTIs, drug−target interactions; TDLs, target development levels; T, True. Blue, data processing step; white, not yet aggregated data; orange, input-independent data source and processing step; dark gray, experimental data set; light gray, predicted or hypothesized data set; green, final data set. Solid triangle symbol indicates which proteins should be used as starting nodes in a SmartGraph analysis, whereas the solid diamond symbol indicates which host proteins should be used for integrating relevant PPIs from the STRING DB. The open triangle and diamond symbols indicate the destination of human proteins routed toward a specific resource, that is, to SmartGraph as starting nodes and to STRING for PPI expansion, respectively.

## ■ RESULTS AND DISCUSSION

**Semiautomated Data Integration Workflow.** In order to build a COVID-19 focused network, we needed to integrate data from multiple, diverse data sources. The bottleneck of the integration proved to be the consolidation of the differing data structures, exemplified by the lack of standardized data categories, preference of one protein identifier over another, and aggregating protein identifiers as delimiter separated list inside a column, to name a few. Also, some of the data originate from experiments whereas others are from predictions. This made it necessary to keep track of this information as well as data provenance in a transparent manner. Here, we present a rigorous workflow that addresses the above challenges and can serve as a template for semiautomated integration of future databases inspired by network pharmacology. Having such an integration workflow in place is key to the timely assembly of a network that is focused on a certain biological aspect, for example, an infectious disease caused by an emerging pathogen.

The assembly of a COVID-19 focused network involved data sources that emerged over a matter of weeks since the start of the pandemic and others that were well-established long before. Considering that new information surfaced relatively quickly, we had to ensure that the workflow we created was sufficiently flexible to accommodate new data. Here, we describe such a workflow (see Figure 1) and a COVID-19 focused Neo4j database that was produced by it. Information regarding the reproduction of the workflow is provided in the "Reproducing the Integration Workflow" section in SI.

The building blocks of the COVID-19 focused network represent experimentally determined as well as predicted HPI, PPI, and DTI data and, additionally, prioritized host targets. In these data sources, host targets are typically identified by their gene name, with some exceptions where UniProt ACs are used, such as TDLs from Pharos DB and PPIs from SmartGraph. The names of several viral proteins in the two HPI data sets were slightly different. Such differences were manually reconciled (see "Mapping of Viral Protein Names in HPIs" section in SI).

The next stage required harmonization of the input data structure. For each input type (HPI, DTI, and host protein), a data structure was defined that can accommodate all data of the respective type. Naturally, each data set needed to be tailored individually to fit the respective data structure, which is the main reason why the overall workflow is called semiautomated instead of automated. Nonetheless, the workflow was equipped with a data registry mechanism that will allow the integration of additional HPI, DTI, and host protein data in a robust and facile manner.

The data registry mechanism consists of configuring a registry of input files with focus on the type of the input (HPI/PPI/host protein), data provenance, and whether proteins extracted from the input should be subject to STRING and SmartGraph analysis. Additionally, what information, that is, fields, of a given input file should be conserved as metadata, what should be extracted as associated score, where applicable, and how a given resource should be referenced in database for the sake of data filtering can be configured.

Recognizing that various resources come in various data structures, the data registry mechanism enables the facile extension of workflow with code snippets tasked with internal data structure harmonization. This is achieved by first providing the name of the data harmonization method to be applied in the registry file. Subsequently, this name needs to be used to create new a condition in a particular function of workflow, which in turn will branch off to new function that the investigator needs to provide. This function needs to be added to a specific source code file, and it will implement the harmonization logic. The logic will convert the data structure into a predefined internal standard that we provide. For a detailed description of the data registry mechanism and how data converter functions can be added to the workflow please refer to section "Integration of Additional Datasets via Data Registry Mechanism", and Tables S2,S3 in SI.

With the help of the data registry mechanism, each resource is first harmonized to an internal standard format, see: Table S3 in SI. In this process individual inputs are deduplicated, and certain derivative data types are also created, such as pathogen proteins and drugs.

The final stage of the workflow integrates all internally harmonized inputs with the help of a well-defined aggregation strategy. This strategy focuses on preserving all the relevant data while tracking the data provenance in a transparent manner. Therefore, it gives the choice to the investigator on how to prioritize data in the light of a particular research setting, for example, by prioritizing experimental vs predicted data or by prioritizing a certain data source over another if conflicting data is found.

The aggregation process employs concatenation, which assures that data entries of entities occurring in multiple resources is tracked in a transparent and obvious manner. For example, a particular host protein might appear in multiple inputs so that it is associated with different metadata in each of these inputs. During the aggregation process, the metadata and data source information will be stored in a delimiter separated string in their respective data fields, so that the order of individual metadata and data source records are kept in sync.

The workflow of this study was designed to allow for the flexible data extension of data for each of the interaction types while providing an option to the investigator to restrict which data segment is subject to the extension. Briefly, Figure 1 depicts the following mechanism. Host proteins of potential importance were collected from resources A, B, C, D, E, F, and H. The data structures of these input data sets were harmonized with respect to the data type. In this process, information specific to individual data sets, as well as metadata and data provenance, is maintained in a transparent manner. After merging and deduplicating each data type, a set of unique host proteins was extracted from them.

The data registry file requires one to indicate whether host proteins extracted from any given resource should be used to assemble PPI with the help of STRING and stringApp APIs[41−43] (resource G) and SmartGraph[37] (resource I). In the workflow, respective subsets of host proteins are routed to SmartGraph or STRING analysis. Note that host proteins extracted from SmartGraph were not subject to the further STRING expansion and *vice versa*, unless there was an overlap between the proteins of these PPIs and the set of unique proteins extracted upstream in the workflow. Once the induced PPI subnetworks were assembled by the STRING and SmartGraph analysis, these data sets were also subject to data structure harmonization and deduplication.

One of the final challenges that we had to address in the workflow was integration of PPIs from different sources. Typically, this procedure is rather challenging due to potentially conflicting and missing PPI information. For instance, PPIs from the STRING API are not annotated with the mode of regulation (whether a target up-regulates or down-regulates its interacting partner).

A viable strategy to resolve such issues is to utilize a comprehensive and preferably curated PPI. To this end, we decided to use the "functional interactions" subset of Reactome (RFI) to fulfill this role. Each PPI was cross-referenced to the RFI subset. This process resulted in the assignment of both the direction and mechanism of regulation for each PPI in the integrated database, as well as the respective confidence score for each RFI. If a given PPI was missing from the RFI subset or the mechanism or the direction of regulation was undefined, then the corresponding properties of that PPI were set to "unknown" and "undefined", respectively.

Each host protein target was annotated by the respective TDL category extracted from the Pharos[7] database. However, as proteins are encoded with UniProt ACs[44,45] both in Pharos and in SmartGraph, we had to resolve them to gene names and *vice versa*. We used a UniProt resource[45−48] to retrieve UniProt to gene name mapping. Considering that this mapping is a many-to-many relationship, the 1:1 mapping was achieved by retaining the highest TDL of any of the UniProt ACs in the case of Pharos data. The TDL annotation of targets enables investigators to instantly identify targets for which FDA-approved drugs exist. This information combined with pathway analysis provide the foundation for the formulation of drug repositioning hypotheses. Such hypotheses can mean the first steps toward the discovery of therapeutics.

Once the final set of unique host proteins is identified: the workflow identifies DTIs with the help of the DrugCentral[49] (resource K) that involve any of these host proteins. These DTIs are harmonized and merged with the other DTIs provided as input to the workflow. After aggregating the DTIs, the set of unique drugs is identified by the workflow.

Finally, the Neo4COVID19 database is built by integrating the unique set of HPIs, PPIs, DTIs, host and viral (pathogen) protein targets, and drugs into a Neo4j database.[50]

**Database Deployment and Dissemination.** The data integration scheme discussed above was implemented in Python[51] and is available as a source code repository at https://github.com/ncats/neo4covid19.[52] The data integration script builds a Neo4j database,[50] which can be accessed publicly via Neo4j Browser, a web-based graphical user interface (GUI) provided at https://aspire.covid19.ncats.io:7473 by Neo4j. When prompted for login credentials, please select "No authentication" from the "Authentication type" drop-down list. Programmatic access is also provided to the same database via various Neo4j API interfaces, for example, "py2neo",[53] which make use of the Neo4j Bolt protocol.[50] For further information, please refer to section "Sample Python Code Snippet to Access Neo4COVID19 Database via API" in SI. Detailed information on accessing the database is available on the Neo4COVID19 Web site at https://neo4covid19.ncats.io under the "ACCESS" tab.

As Neo4COVID19 is a graph database, fundamentally, it is a collection of nodes and edges. In the database, there are 903 "HostProtein" nodes, 55 "PathogenProtein" nodes, and 635 "Drug" nodes. Three types of relationships are contained by the database: HPI, PPI, and DTI defined between two "HostProtein", a "PathogenProtein" and a "HostProtein", and a "Drug" and a "HostProtein" nodes, respectively. The Neo4COVID19 database contains 487 HPI, 63 278 PPI, and 1221 DTI relationships. Further information regarding the node and edge types extracted from each data resource are provided in Table 1. Node and edge attributes shown in Table S1 in SI can be used to fine-tune the network in a way that matches the needs of the analysis at hand. For instance, one might decide to only consider nodes and edges of the COVID-19 focused network that represent experimental data. This can be achieved with ease by filtering the data with the help of Boolean fields, each representing a specific data source. Another example is application of a confidence threshold to PPI edges retrieved from stringApp API using the "source_specific_score" field value to filter the data. This data structure facilitates the versatile use of the Neo4COVID19 database in many research settings and the corroboration of information pertaining to various interactions.

**Table 1. COVID-19 Focused Network Statistics[a]**

| data set | host targets | viral targets | drugs | HPIs | PPIs | DTIs |
|---|---|---|---|---|---|---|
| proteomics study | 102 | | | | | |
| CRISPR | 105 | | | | | |
| Meta Path AI/ML | 185 | | | | | |
| STRING | 743 | | | | 63076 | |
| SmartGraph/HATs | 148 | | | | 225 | |
| interactome study | 332 | 27 | | 332 | | |
| P-HIPSter | 38 | 28 | | 155 | | |
| predicted DTIs | 46 | | 31 | | | 86 |
| DrugCentral | 127 | | 619 | | | 1163 |

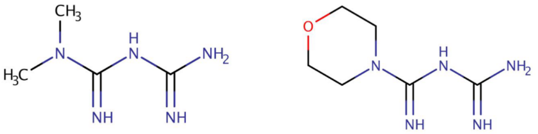
[a] Shown are summary of individual data types integrated into the Neo4COVID-19 Neo4j database. Of note, overlap may exist between data types associated with the original data sources. PPIs, host−host protein interactions; HPIs, host−pathogen (here SARS-CoV-2) protein interactions; DTIs, drug−target interactions.

**Importing Neo4COVID-19 Database into Cytoscape.** In order to facilitate the translational impact via dissemination[54] and flexible downstream analysis of the Neo4COVID19 focused network in the bioinformatics community, here we describe a simple procedure to import the database into the widely utilized Cytoscape application.[55]

Importing the Neo4COVID19 database into Cytoscape v3.8.2 requires the installation of the "Cytoscape Neo4j Plugin" v0.4,[56] which can be easily achieved from within the Cytoscape application (see Figure S1a in SI). Once the plugin is installed, a connection to the Neo4j database has to be established via the Bolt protocol, as shown in Figure S1b in SI.

After the successful establishment of database connection, the entire Neo4COVID19 database can be imported into Cytoscape with only a basic query statement written in the Cypher[50] language as shown in Figure S1c in SI. The statement is actually identical to the query provided by the plugin when removing the "LIMIT" clause from the default statement. Finally, a custom visualization style can be applied (see Figure S2 and "Applying Custom Visual Style to the Imported Network in Cytoscape", SI). The resultant network is shown in Figure S1d in SI.

**Use Cases.** Here, we describe use cases to demonstrate how one can use the Neo4COVID19 for hypothesis generation in a network pharmacology setting. First, we examined the $T_{clin}$-designated HPI subset: $T_{clin}$ designated proteins (according to TDL) from the Gordon data set that have a fold change of 10 or higher following SARS-CoV-2 exposure. Out of 166 such HPIs, at least 63 occur between 27 human and 11 viral proteins. These 27 proteins are integral components of the mitochondrial respiratory chain complex I (GO:0005747) and are targeted by the antidiabetic drug metformin. Initially, we were excited to note that metformin (**1**) shares chemical similarity with an old antiviral drug moroxydine (**2**) (see Figure 2). Shortly thereafter, as type 2 diabetes was identified as risk factor for severe COVID-19,[57] we had to briefly suspend further metformin research as we suspected its effect on the $T_{clin}$-designated HPI subset was indirect. However, recent reports show that metformin treatment is actually independently associated with a significant reduction in mortality in subjects with diabetes and COVID-19.[58] Thus,



**1**
**Metformin**

**2**
**Moroxydine**

**Figure 2.** Molecular structures of metformin and moroxydin. Molecules were depicted with the help of ChemAxon's MarvinSketch v17.15.0.[59]

while we cannot assume that the antiviral activity of moroxydine is similar, we are quite confident that metformin acts as a significant HPI perturbagen during SARS-CoV-2 infection and could serve as an adjuvant antiviral therapy under appropriate conditions.

Next, we investigated the interactions between host and virus proteins (the HPI relationships). With the help of node and edge attributes, it is possible to construct a subnetwork of the Neo4COVID19 network by retaining only virus and host targets and the edges between them. This process gives rise to a bipartite network, in which host nodes are only connected to virus nodes and *vice versa*. In this network, a natural clustering emerges where several connected components exist involving many human proteins centered around a single virus protein (see Figure 3A). The network topology also reveals that certain host targets might be thought of as the "Achilles' heel" of the virus due to their connection structure. The peculiarity of these host proteins is that they are connected to many virus proteins; hence they were named as "virus hubs". An ideal strategy would be to target such a virus hub that affects multiple biological processes of the virus while only causing a small perturbation in the regulatory network of the host.

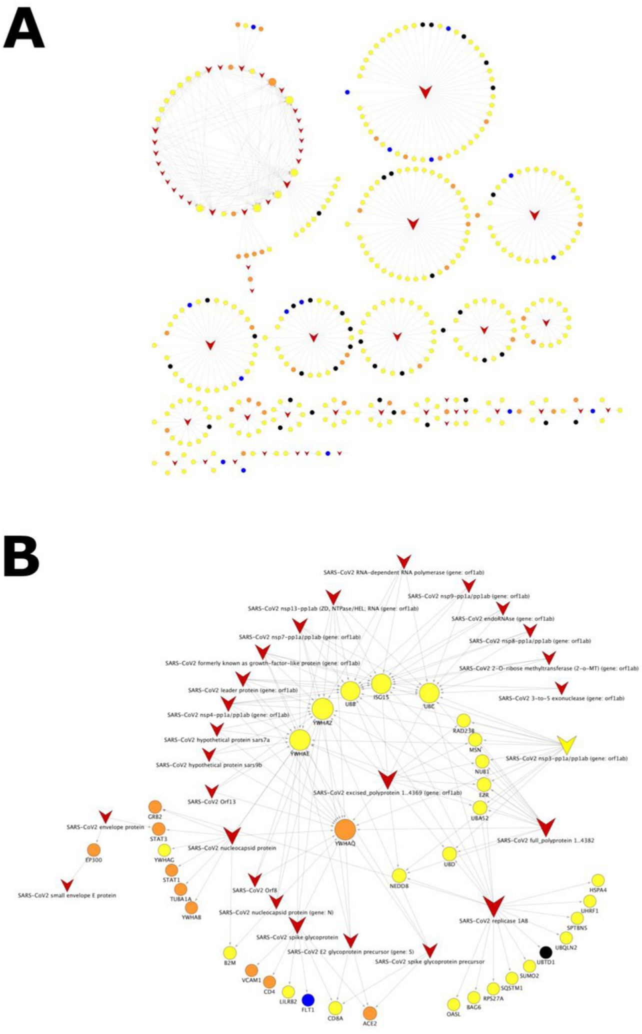

A
B
SARS-CoV2 RNA-dependent RNA polymerase (gene: orf1ab)
SARS-CoV2 nsp9-pp1a/pp1ab (gene: orf1ab)
SARS-CoV2 nsp13-pp1ab (ZD, NTPase/HEL; RNA (gene: orf1ab)
SARS-CoV2 endoRNAse (gene: orf1ab)
SARS-CoV2 nsp7-pp1a/pp1ab (gene: orf1ab)
SARS-CoV2 nsp8-pp1a/pp1ab (gene: orf1ab)
SARS-CoV2 formerly known as growth-factor-like protein (gene: orf1ab)
SARS-CoV2 2-O-ribose methyltransferase (2-o-MT) (gene: orf1ab)
SARS-CoV2 leader protein (gene: orf1ab)
UBB
ISG15
UBC
SARS-CoV2 3-to-5 exonuclease (gene: orf1ab)
SARS-CoV2 nsp4-pp1a/pp1ab (gene: orf1ab)
YWHAZ
RAD23B
SARS-CoV2 hypothetical protein sars7a
YWHAE
MSN
SARS-CoV2 nsp3-pp1a/pp1ab (gene: orf1ab)
SARS-CoV2 hypothetical protein sars9b
NUB1
SARS-CoV2 excised_polyprotein 1..4369 (gene: orf1ab)
EZR
SARS-CoV2 Orf13
UBA52
GRB2
SARS-CoV2 envelope protein
STAT3
YWHAQ
SARS-CoV2 full_polyprotein 1..4382
EP300
YWHAG
SARS-CoV2 nucleocapsid protein
UBD
STAT1
NEDD8
SARS-CoV2 small envelope E protein
TUBA1A
SARS-CoV2 Orf8
HSPA4
YWHAB
SARS-CoV2 nucleocapsid protein (gene: N)
UHRF1
SARS-CoV2 replicase 1AB
SPTBN5
SARS-CoV2 spike glycoprotein
UBQLN2
SARS-CoV2 E2 glycoprotein precursor (gene: S)
UBTD1
B2M
SARS-CoV2 spike glycoprotein precursor
SUMO2
VCAM1
SQSTM1
CD4
RPS27A
LILRB2
FLT1
BAG6
CD8A
ACE2
OASL

**Figure 3.** Bipartite network of HPIs. Human and virus proteins are depicted by circles and v-like shapes, respectively. The larger the node size, the higher the degree of node connectivity. Color of the human proteins indicate their TDL annotation: blue, $T_{clin}$; orange, $T_{chem}$; yellow, $T_{bio}$; gray, $T_{dark}$. (A) Complete HPI bipartite network visualized in a "yFiles Circular" layout. (B) Subnetwork of the largest connected component centered around the virus hub YWHAQ visualized in a "yFiles Radial" layout. network layouts were generated with the help of Cytoscape v3.8.2[60] and yFiles modules.[61]

As shown in Figure 3B, a few host targets are connected to a substantially higher number of nodes than others, as indicated by a larger size of the corresponding nodes. One such target, YWHAQ (UniProt AC P27348), is associated with the $T_{chem}$ TDL category, which means that there is a small molecule modulator for that target. Therefore, one might hypothesize that YWHAQ might be a potential host target to develop a drug against based on the existing modulator(s) as a seed active molecule(s). However, the next step will require the analysis of the role of YWHAQ in the signal transduction network of the host (human). While such analysis is outside the scope of this work, it should be noted that the Pharos database indicates 260 PPIs associated with YWHAQ, which suggests that this particular target is involved in many biological processes of the host; hence its modulation will likely considerably perturb the network. It could be still possible that a redundant (parallel) path exists in the signal transduction network, which might mitigate the potential adverse effects of the perturbation.

While the aforementioned strategy might identify promising virus hubs, unfortunately, there are no $T_{clin}$ targets among them. This indicates that at the time of this study we could not apply drug repositioning to target virus hubs. An alternative strategy could be to target multiple $T_{clin}$ host targets (blue circles on Figure 3A) that appear in separate clusters. Drug compounds associated with $T_{clin}$ targets can be easily imported into the network. This approach might form the basis for the engineering of a multiagent therapy ideally employing approved drugs if they can effectively interfere with the implicated pathogen−host interactions. The detailed procedure to replicate the use cases described above is provided in "Reproducing the Use Cases" section in SI.

**Considerations Regarding the Continuous Integration and the Validity of the Data.** The data integration workflow presented in this study was designed to be widely adoptable in a network pharmacology research setting. In this sense, our workflow can facilitate any semiautomated integration of constantly evolving data, which happens to be the nature of many modern data sets, such as Reactome,[6] DrugCentral,[10] and Pharos,[7] to name a few. However, the unprecedented pace of data influx that we have witnessed since the beginning of the COVID-19 pandemic highlighted challenges with regards to continuous data integration and data quality.

The workflow presented in this study demonstrates how essential data types can be integrated quickly with the help of the data registry mechanism. However, the rate limiting step of the integration remains the conversion and curation of data. While it would be desirable to integrate all emerging COVID-19 related HPI, PPI, and DTI data[62−69] appropriately, this remains a significant challenge in the light of these considerations even with the help of the workflow of this study. Nonetheless, we are hopeful that the workflow can be the first step toward such a research effort, which may be realized in the form of a consortium. Also, we hope that our work will promote the effort of releasing data sets, even supplementary data, in a standardized format or via an API to facilitate the continuous integration of relevant data.

While we are hopeful that data integration efforts can be solved with the help of automated workflows and collaborative research, it is of outmost importance that the quality of source and integrated data is constantly monitored. The importance of this aspect is reflected in the context of some early COVID-19 related results that have since been questioned and for which manuscripts have been retracted.[70,71] A specific controversy that emerged over time is associated with the efficacy or the lack thereof of hydroxychloroquine. Although the Neo4COVID-19 database includes data related to hydroxychloroquine, we intended to keep those records to serve as example for historically data that needs to be considered with care. Nevertheless, with the data structure of Neo4COVID-19, it is easy to filter out hydroxychloroquine related records.

We contemplate that the dichotomy between the need for relevant and bleeding edge information and the quality of the data will recreate this scenario in the case of a similar event in the future. Therefore, any data processing workflow should be dynamic, transparent, regularly reviewed, and constantly updated.

## ■ CONCLUSIONS

Here, we describe a semiautomated workflow for the integration of data sources to produce a COVID-19 focused graph database. The workflow can be easily generalized to other drug discovery scenarios, which can save precious time in the case of a pathogen outbreak. The workflow makes use of the state-of-the-art network pharmacology approaches and yields an interconnected network of host and viral protein targets and drugs, containing information on HPIs, PPIs, and DTIs. The workflow is flexible, which makes it possible to replace data sources or add new ones to it. The layered structure of the network and the underlying data schema allow researchers to filter data sources that they find relevant in their investigation. During the development of this workflow, we came across known bottlenecks related to data integration that we believe could be ameliorated to a great extent by following certain practices. For instance, the interaction types in a network pharmacology setting are well-defined (PPI, HPI, DTI), and therefore, such data should be made available in a few well-established formats. This would allow for seamless integration of already existing and emerging data sets. Furthermore, providing a robust API for a data set facilitates its integration and allows for programmatic updates. For instance, equipping SmartGraph with an API would allow for automating a significant part of the current workflow. If this API existed, then all manual intervention would be limited to configuring the data registry file and writing the data specific standardization snippets, allowing the rest of the workflow to be executed entirely automatically.

The Neo4j database generated by this workflow can be accessed via a web interface at https://aspire.covid19.ncats.io:7473/ to enable the exploration of data without much expertise in the bioinformatics field. In addition, it takes advantage of the Neo4j Bolt protocol and provides an API to facilitate the integration of the COVID-19 focused network into virtually any bioinformatics workflow. The Neo4COVID19 landing page (https://neo4covid19.ncats.io) provides detailed information under the "ACCESS" tab for those who would like to connect to the database in a programmatic manner. In this study, we provided use cases to show how the Neo4COVID-19 network can be utilized to generate hypotheses with focus on drug repositioning.

We believe that our Neo4COVID19 database will be a valuable asset to the research community and will catalyze the discovery of therapeutics to defeat COVID-19. Furthermore, the underlying flexible workflow can serve as a starting point for the integration of critical knowledge in the event of a potential future outbreak, which we all hope will never happen.

## ■ METHODS

**Assembly of a Multimodal Network.** In a drug discovery setting driven by network pharmacology, a multimodal network[37,38,72] needs to be assembled and tailored according to the disease context. In this study, we set forth to create a multimodal network focused on COVID-19 by integrating host and pathogen protein targets, drugs, and relations defined between them, such as PPIs, HPIs, and DTIs. The respective information was derived from various data sources described in detail below. Furthermore, we use a "resource" alias (see Figure 1) for each data set throughout the text in order to concisely and unambiguously refer to individual input data sets. A pseudocode of the data integration workflow is provided in the "Pseudo-Code of the Data Integration Workflow" section in Supporting Information (SI).

**Host Targets Implicated to Play Important Roles in Pathogenesis.** Host cell translation changes[73] following SARS-CoV-2 infection were studied in Caco-2 cells using translatome[74] and proteome proteomics at four different time points (2, 6, 10, and 24 h, respectively) after infection. This unbiased profile of the cellular response to SARS-CoV-2 infection was used to identify key determinants of the host cell response to infection. Extensive proteome modulation occurs 24 h postinfection, for example, reduced expression of cholesterol metabolism proteins and increased expression profile for carbon metabolism proteins and spliceosome components. Some pathways appear amenable to therapeutic intervention, for example, along the proteostasis and nucleotide biosynthesis pathways. Given quantified translation data for 2715 proteins (as documented in the Supporting Information), we selected proteins with $P$ values below 0.05 at 24 h (virus exposure compared to control), as follows: 75 proteins having lower translation values across all 4 time points (of these, 38 are involved in acetylation according to STRING); 23 proteins having positive virus-induced translation values at 24 h (of which 12 are also involved in acetylation). These 98 host proteins were subject to further processing: one (UniProt AC Q9N2J8) was removed as it cannot be mapped to a gene name, and four (UniProt ACs P84243, Q8IZP9, Q8IXH7, P63302) were mapped to multiple gene names; thus the respective records were replicated. This gave rise to 102 proteins in total, which were denoted as resource "A".

Host genes essential for cell survival in response to SARS-CoV-2 infection were identified using two Cas9 Vero-E6 cell line constructs, using a genome-wide pooled CRISPR (clustered regularly interspaced short palindromic repeats) library.[75,76] This CRISPR-Cas9 screen identified "proviral" and "antiviral" genes, as follows. "Proviral" genes are involved in resistance, and their knockout confers resistance to virus-induced cell death. These genes include the ACE2 viral entry receptor, 11 genes from the SWI/SNF (SWItch/Sucrose Non-Fermentable) chromatin remodeling complex,[77] and 7 genes associated with CDKN1A transcription upregulation via

RUNX3.[78] "Antiviral" genes are involved in sensitization, and their knockout sensitizes a cell to virus-induced cell death. These genes include HIRA (a subunit of the H3.3 histone chaperone complex), a set of 6 genes involved in viral translation, 8 genes associated with the SMN (survival of motor neurons) complex, and 5 components of the NURF (nucleosome remodeling factor) complex. A set of 53 proviral and 52 antiviral host genes detected via this CRISPR-Cas9 screen[76] were incorporated and assigned a resource label "B".

Next, we describe an AI/ML framework that was utilized to predict host proteins that potentially interact with viral proteins. The Target Central Relational Database (TCRD)[7] aggregates protein-specific data from different sources (e.g., GTEx,[79] LINCS,[80] STRING,[42] Reactome[6,81]), and it was used to build the TCRD-KG knowledge graph. The TCRD-KG nodes can be proteins, diseases, or phenotypes, and edges can be pathways, protein−protein interactions, or other biological relationships among proteins and diseases. A machine learning (ML) framework based on the TCRD-KG metapaths[82] and XGBoost[83] classification algorithm was developed to predict disease-associated genes (proteins). The metapaths specify network paths that connect proteins to specific diseases in the TCRD-KG. The degree-weighted path count (DWPC)[84] metric is used to quantify the metapath prevalence, transforming TCRD-KG data to feature vectors for ML model by metapath matching, based on an input disease. The Python package to build TCRD-KG using TCRD database and XGBoost ML model is available on GitHub (https://github.com/unmtransinfo/ProteinGraphML).[85]

A training data set comprising 104 proteins as positive labels (known to be associated with SARS-CoV-2) and 114 proteins as negative labels (not associated with SARS-CoV-2) was used to train the ML model. A total of six different models were built, using slight variations of the input data (e.g., inclusion/deletion of human proteins identified by P-HIPSTer[86,88] to interact with SARS-CoV-2 proteins and inclusion or absence of the LINCS[80] descriptors). Using 5-fold cross-validation on the training data, the area under the curve (AUC), accuracy, and Matthews correlation coefficient (MCC) were computed for each model. The ML models trained on 218 proteins were used to predict the association between 20 029 proteins and SARS-CoV-2. Using XGBoost feature importance output, metapaths were sorted in decreasing order of "gain" score to understand node interactions. The features with higher gain were more dominant in predicting SARS-CoV-2 associated proteins. A total of 986 proteins were predicted with "high confidence" by the 6 models; of the 136 predicted by 3 or more models, 99 were not part of the "understudied" proteins[57] and were given preference for this study (denoted as resource "C").

**Virus-Implicated Host Proteins.** Host proteins that were either identified as interacting partners of SARS-CoV-2 virus proteins in experimental studies or predicted to be of potential importance in terms of pathogenesis or therapy are referred to as virus-implicated host proteins (VIHPs) in this study. VIHPs were derived from experimentally determined HPIs (resource "E"), and from host proteins implicated in experimental studies (resources A and B). This initial set of VIHPs was extended by predicted HPIs (resource "F"), DTIs (resource "D"), and host proteins of potential importance (resource C).

**Host−Pathogen and Host−Host Interactions.** An experimentally determined host−pathogen interactome was published by Gordon et al.[33,34] They identified 332 HPIs defined between 26 viral proteins and 332 host proteins or host factors (resource E) that were determined via affinity purification mass-spectroscopy (AP-MS).[87] A systematic analysis revealed 67 druggable human proteins and 69 compounds (including FDA-approved drugs and investigational drugs currently tested in clinical trials and preclinical studies) that were proposed to be evaluated for efficacy against SARS-CoV-2. While the host−pathogen interactome suggests potential pharmacological targets and possible interventions, the outcomes should be cautiously interpreted as it was mentioned that the identified agents could have either beneficial or detrimental effects (e.g., HDAC2 inhibitors).

Potential HPIs were predicted using the P-HIPSTer algorithm[86,88] giving rise to 155 HPIs involving 28 SARS-CoV-2 and 38 human proteins (resource F). P-HIPSTer offers a computational framework that utilizes sequence and structural information to infer HPIs, currently covering a total of 28 viral families (>900 human viruses) and >5000 human proteins accounting for 282 528 HPIs.[88]

**Experimentally Determined PPIs.** We collected PPIs from various data sources. The network assembly process involved a network expansion step with the help of the STRING and stringApp APIs (resource "G").[42,43] In this step, experimentally determined and inferred PPIs were imported with the following parameter settings: maximum interactors = 100, alpha = 0.5. For more details on STRING expansion of the network please refer to section "Expansion of PPIs via StringApp API" in SI. Considering that interactions returned by the stringApp API are undirected, these edges were introduced into the network in both directions.

Due to the implication of histone acetyltransferases (HATs) in the pathogenesis, a set of HATs was prioritized (resource "H"), and in conjunction with virus-implicated host proteins (VIHPs), it was utilized to construct a subnetwork with the help of SmartGraph (resource "I"). In the SmartGraph analysis, the HATs and VIHPs were used as starting and end nodes, and *vice versa*. In either case, the maximum length of shortest paths between starting and end nodes was limited to 3. For more details on the assembly of the SmartGraph subnetwork please refer to section "Assembly of the Smart-Graph Subnetwork", SI. This input data set contains 248 PPIs involving 108 host targets. The "Functional interactions" data set (resource "J"), derived from the Reactome database (version 2019)[6,81] was used to cross-reference PPIs originating from various data sources.

**Drug−Target Interactions.** Pertinent DTIs were primarily extracted from the DrugCentral database (version 2020, resource "K").[10] The DrugCentral database includes 4642 drugs, of which 2549 have regulatory approval dates, 3082 have ATC codes,[89] and 1884 have INN stem annotations.[90] These drugs are associated with 110 577 formulations (drug labels) in total. Furthermore, the DrugCentral database also provides information on DTIs: 15 397 human DTIs and 4910 nonhuman DTIs; of these 2752 (2328 human) are MoA drug−target associations.

For over 90% of the DTIs, the DTI therapeutic consequence is documented. We also collected the pharmacological action for each DTI that provides additional information about the potential intervention. The DTIs were originally extracted from scientific literature, drug labels, and other data sources such as, ChEMBL,[4] IUPHAR Guide2Pharmacology,[91] WOMBAT-PK,[92] DrugBank,[93] and KEGG Drug.[94]

To further enrich the network, we included both known and predicted DTIs for drugs and human endogenous metabolites found in the vicinities of the chemical space defined by three small molecules being investigated at the time for their antiviral activity against SARS-CoV-2 assays, namely, *N*4-hydroxycytidine (NHC),[95] hydroxychloroquine (HCQ),[70] and camostat (CAM).[96] A set of 25 compounds was compiled around NHC[97] and 5 around the chemical neighborhood of HCQ, and camostat. Processing those 31 small molecules against the CLARITY platform[98] returned a total of 86 DTIs, known (from public sources[5]) or predicted[99] to have activity against 46 unique protein targets (resource D).

**Target Development Level Information.** Target development level (TDL) information for host proteins were imported from the Pharos portal (data track "L").[7,100]

## ASSOCIATED CONTENT

### Supporting Information

The Supporting Information is available free of charge at https://pubs.acs.org/doi/10.1021/acs.jcim.1c00431.

> Python code to access Neo4COVID19 database via API, overview of data integration workflow, instructions for reproducing integration workflow, detailed steps for assembling SmartGraph subnetwork, expansion of host—host interactions via StringApp API, applying custom visual style to Cytoscape network, instructions to reproduce the use cases, importing Neo4COVID19 graph database into cytoscape, customizing network visualization, node and edge attributes of the Neo4-COVID19 graph database, structure of data registry file, and definition of mandatory and optional fields of internal data structure (PDF)

## AUTHOR INFORMATION

### Corresponding Authors

**Gergely Zahoránszky-Kőhalmi** − *National Center for Advancing Translational Sciences, Rockville, Maryland 20850, United States;* orcid.org/0000-0002-2534-8770; Email: gergely.zahoranszky-kohalmi@nih.gov

**Tudor I. Oprea** − *Department of Internal Medicine, University of New Mexico School of Medicine, Albuquerque, New Mexico 87131, United States; Novo Nordisk Foundation Center for Protein Research, Faculty of Health and Medical Sciences, University of Copenhagen, 2200 Copenhagen N, Denmark; UNM Comprehensive Cancer Center, Albuquerque, New Mexico 87102, United States; Department of Rheumatology and Inflammation Research, Institute of Medicine, Sahlgrenska Academy at University of Gothenburg, 40530 Gothenburg, Sweden;* orcid.org/0000-0002-6195-6976; Email: toprea@salud.unm.edu

### Authors

**Vishal B. Siramshetty** − *National Center for Advancing Translational Sciences, Rockville, Maryland 20850, United States*

**Praveen Kumar** − *Department of Internal Medicine, University of New Mexico School of Medicine, Albuquerque, New Mexico 87131, United States; Department of Computer Science, University of New Mexico, Albuquerque, New Mexico 87131, United States*

**Manideep Gurumurthy** − *National Center for Advancing Translational Sciences, Rockville, Maryland 20850, United States*

**Busola Grillo** − *National Center for Advancing Translational Sciences, Rockville, Maryland 20850, United States*

**Biju Mathew** − *National Center for Advancing Translational Sciences, Rockville, Maryland 20850, United States*

**Dimitrios Metaxatos** − *National Center for Advancing Translational Sciences, Rockville, Maryland 20850, United States*

**Mark Backus** − *National Center for Advancing Translational Sciences, Rockville, Maryland 20850, United States*

**Tim Mierzwa** − *National Center for Advancing Translational Sciences, Rockville, Maryland 20850, United States*

**Reid Simon** − *National Center for Advancing Translational Sciences, Rockville, Maryland 20850, United States*

**Ivan Grishagin** − *National Center for Advancing Translational Sciences, Rockville, Maryland 20850, United States; Rancho BioSciences LLC., San Diego, California 92127, United States*

**Laura Brovold** − *Rancho BioSciences LLC., San Diego, California 92127, United States*

**Ewy A. Mathé** − *National Center for Advancing Translational Sciences, Rockville, Maryland 20850, United States;* orcid.org/0000-0003-4491-8107

**Matthew D. Hall** − *National Center for Advancing Translational Sciences, Rockville, Maryland 20850, United States;* orcid.org/0000-0002-5073-442X

**Samuel G. Michael** − *National Center for Advancing Translational Sciences, Rockville, Maryland 20850, United States*

**Alexander G. Godfrey** − *National Center for Advancing Translational Sciences, Rockville, Maryland 20850, United States*

**Jordi Mestres** − *Research Group on Systems Pharmacology, Research Program on Biomedical Informatics (GRIB), IMIM Hospital del Mar Medical Research Institute and University Pompeu Fabra, 08003 Barcelona, Catalonia, Spain;* orcid.org/0000-0002-5202-4501

**Lars J. Jensen** − *Novo Nordisk Foundation Center for Protein Research, Faculty of Health and Medical Sciences, University of Copenhagen, 2200 Copenhagen N, Denmark;* orcid.org/0000-0001-7885-715X

Complete contact information is available at: https://pubs.acs.org/10.1021/acs.jcim.1c00431

### Author Contributions

This research study was initiated by T.I.O. and G.Z.K. The workflow and the Neo4j database was designed and built by G.Z.K., M.G., B.G., and B.M. designed and configured the computational infrastructure to provide public access to the Neo4j database. A.G.G., S.G.M., E.M., D.M., and M.D.H. provided inspiration and feedback for the study. T.I.O., P.K., J.M., and L.J.J. provided predictions for HPIs, host and viral targets, and drugs. T.I.O., P.K., L.J.J., and V.B.S. contributed to data analysis. G.Z.K. and V.B.S. wrote the majority of the text, T.I.O., L.J.J., J.M., V.B.S., P.K., M.G., I.G., L.B., M.D.H., and A.G.G. provided edits to the manuscript, and the others contributed to the study. All authors read and approved the manuscript.

### Notes

The authors declare the following competing financial interest(s): L.J.J. is cofounder and scientific advisory board member of Intomics A/S.

The Neo4COVID19 database is publicly available at https://aspire.covid19.ncats.io:7473. Of note, "No authentication" needs to be selected from the drop-down list "Authentication type". We provide further information at the Neo4COVID19 Web site (https://neo4covid19.ncats.io) under "ACCESS" tab regarding how the Neo4COVID19 graph database can be accessed programmatically via the Neo4j Bolt protocol using API. Moreover, we provide detailed instruction how the database can be imported into Cytoscape in a user-friendly manner, see: "Reproducing the Use Cases" in SI. The source code repository of the workflow utilized to construct the Neo4COVID19 database is publicly available at https://github.com/ncats/neo4covid19. In the same repository, we provide detailed instructions on how the workflow can be replicated to build a replica of the Neo4COVID19 graph database. The instructions can be found in this file: https://github.com/ncats/neo4covid19/blob/master/README.md. All other data sets and computational tools utilized in this study were described in detail in the Methods section.

## ■ ACKNOWLEDGMENTS

The authors are thankful for the devoted work of Henderson Tozer, Kevin Diaz, James McWilliams, and all members of the IT support team of the Information Resources Technology Branch of NCATS/NIH who enabled us to continue our research in unprecedented times. Furthermore, we would like to acknowledge Tim Willson, Dac-Trung Nguyen, Noel T. Southall, Cullen Kelin, and Dave Calabrese for their fruitful discussions. This research was supported in part by the Intramural research program of the NCATS, NIH, and the Illuminating the Druggable Genome−Knowledge Management Center, NIH-U54 Grant (1U54CA189205-01, PI Tudor I. Oprea). This research was partly supported by a project from the Spanish Ministerio de Ciencia, Innovación y Universidades (SAF2017-83614-R, PI Jordi Mestres). Furthermore, this research in part supported by the Novo Nordisk Foundation (NNF14CC0001).

## ■ ABBREVIATIONS

API: application programming interface
ATC: anatomical therapeutic chemical
INN: international nonproprietary names
MoA: mechanism-of-action
FDA: U.S. Food and Drug Administration

## ■ REFERENCES

(1) Houlihan, C. F.; Whitworth, J. A. Outbreak Science: Recent Progress in the Detection and Response to Outbreaks of Infectious Diseases. *Clin. Med. (Northfield. Il)* **2019**, *19* (2), 140−144.

(2) Ravi, S. J.; Meyer, D.; Cameron, E.; Nalabandian, M.; Pervaiz, B.; Nuzzo, J. B. Establishing a Theoretical Foundation for Measuring Global Health Security: A Scoping Review. *BMC Public Health* **2019**, *19* (1), 954.

(3) Berger, K.; et al. Policy and Science for Global Health Security: Shaping the Course of International Health. *Trop. Med. Infect. Dis.* **2019**, *4* (2), 60.

(4) Gaulton, A.; et al. ChEMBL: A Large-Scale Bioactivity Database for Drug Discovery. *Nucleic Acids Res.* **2012**, *40* (D1), D1100−D1107.

(5) Bento, A. P.; et al. The ChEMBL Bioactivity Database: an Update. *Nucleic Acids Res.* **2014**, *42*, D1083−90.

(6) Croft, D.; et al. Reactome: a Database of Reactions, Pathways and Biological Processes. *Nucleic Acids Res.* **2011**, *39*, D691−D697.

(7) Nguyen, D.-T.; et al. Pharos: Collating Protein Information to Shed Light on the Druggable Genome. *Nucleic Acids Res.* **2017**, *45* (D1), D995−D1002.

(8) Cerami, E. G.; et al. Pathway Commons, a Web Resource for Biological Pathway Data. *Nucleic Acids Res.* **2011**, *39*, D685−90.

(9) Huang, R.; et al. The NCATS BioPlanet − An Integrated Platform for Exploring the Universe of Cellular Signaling Pathways for Toxicology, Systems Biology, and Chemical Genomics. *Front. Pharmacol.* **2019**, *10*, 445.

(10) Avram, S.; et al. DrugCentral 2021 Supports Drug Discovery and Repositioning. *Nucleic Acids Res.* **2021**, *49*, D1160.

(11) Holshue, M. L.; et al. First Case of 2019 Novel Coronavirus in the United States. *N. Engl. J. Med.* **2020**, *382* (10), 929−936.

(12) Li, R.; et al. Substantial Undocumented Infection Facilitates the Rapid Dissemination of Novel Coronavirus (SARS-CoV-2). *Science (Washington, DC, U. S.)* **2020**, *368* (6490), 489−493.

(13) Pei, S.; Kandula, S.; Shaman, J. Differential Effects of Intervention Timing on COVID-19 Spread in the United States. *medRxiv* **2020**, DOI: 10.1101/2020.05.15.20103655.

(14) Jones, D.; Neal, R. D.; Duffy, S. R. G.; Scott, S. E.; Whitaker, K. L.; Brain, K. Impact of the COVID-19 Pandemic on the Symptomatic Diagnosis of Cancer: The View From Primary Care. *Lancet Oncol.* **2020**, *21* (6), 748−750.

(15) Yang, Y.; Shen, C.; Hu, C. Effect of COVID-19 Epidemic on Delay of Diagnosis and Treatment Path for Patients With Nasopharyngeal Carcinoma. *Cancer Manage. Res.* **2020**, *12*, 3859−3864.

(16) Madhusoodanan, J. News Feature: to Counter the Pandemic, Clinicians Bank on Repurposed Drugs. *Proc. Natl. Acad. Sci. U. S. A.* **2020**, *117* (20), 10616−10620.

(17) Gil, C.; et al. COVID-19: Drug Targets and Potential Treatments. *J. Med. Chem.* **2020**, *63*, 12359.

(18) Operational Database Management Systems. *Knowledge Graph on COVID-19*, 2020. https://web.archive.org/web/20200715070446/http://www.odbms.org/2020/03/we-build-a-knowledge-graph-on-covid-19/ (accessed 2022-01-20).

(19) Smith, M.; Smith, J. C. Repurposing Therapeutics for COVID-19: Supercomputer-Based Docking to the SARS-CoV-2 Viral Spike Protein and Viral Spike Protein-Human ACE2 Interface. *chemRxiv* **2020**, DOI: 10.26434/chemrxiv.11871402.v4.

(20) Batra, R.; Chan, H.; Kamath, G.; Ramprasad, R.; Cherukara, M. J.; Sankaranarayanan, S. Screening of Therapeutic Agents for COVID-19 using Machine Learning and Ensemble Docking Simulations *arXiv*, http://arxiv.org/abs/2004.03766 accessed Apr. 2020.

(21) Brimacombe, K. R. An OpenData Portal to Share COVID-19 Drug Repurposing Data in Real Time. *bioRxiv* **2020**, DOI: 10.1101/2020.06.04.135046.

(22) *COVID-19 Disease Portal.* https://rgd.mcw.edu/rgdweb/portal/home.jsp?p=14.

(23) Kuleshov, M. V.; et al. The COVID-19 Drug and Gene Set Library. *Patterns* **2020**, *1* (6), 100090.

(24) Wang, Q.; et al. COVID-19 Literature Knowledge Graph Construction and Drug Repurposing Report Generation *arXiv*, http://arxiv.org/abs/2007.00576, accessed Jul. 2020.

(25) Ioannidis, V. N.; et al. DRKG - Drug Repurposing Knowledge Graph for Covid-19, https://github.com/gnn4dr/DRKG/, 2020.

(26) Haendel, M. A.; Chute, C. G.; Gersing, K. The National COVID Cohort Collaborative (N3C): Rationale, Design, Infrastructure, and Deployment. *J. Am. Med. Informatics Assoc.* **2021**, *28*, 427.

(27) *Wayback Machine - Internet Archive* http://web.archive.org/ (accessed 2022-01-20).

(28) Hopkins, A. L. Network Pharmacology. *Nat. Biotechnol.* **2007**, *25* (10), 1110−1.

(29) Maron, B. A.; et al. A Global Network for Network Medicine. *npj Syst. Biol. Appl.* **2020**, *6* (1), 29.

(30) Barabási, A.-L.; Gulbahce, N.; Loscalzo, J. Network Medicine: A Network-Based Approach to Human Disease. *Nat. Rev. Genet.* **2011**, *12* (1), 56−68.

(31) Womack, F.; McClelland, J.; Koslicki, D. Leveraging Distributed Biomedical Knowledge Sources to Discover Novel Uses for Known Drugs. *bioRxiv* **2019**, DOI: 10.1101/765305.

(32) Chen, H.; Cheng, F.; Li, J. iDrug: Integration of Drug Repositioning and Drug-Target Prediction via Cross-Network Embedding. *PLoS Comput. Biol.* **2020**, *16* (7), e1008040.

(33) Gordon, D. E. *bioRXiv* **2020**, DOI: 10.1101/2020.03.22.002386.

(34) Gordon, D. E.; et al. A SARS-CoV-2 Protein Interaction Map Reveals Targets for Drug Repurposing. *Nature* **2020**, *583* (7816), 459−468.

(35) Archived: WHO Timeline - COVID-19. https://www.who.int/news-room/detail/27-04-2020-who-timeline---covid-19.

(36) Update - Timeline of WHO's Response to COVID-19. https://www.who.int/news-room/detail/29-06-2020-covidtimeline.

(37) Zahoránszky-Kőhalmi, G.; Sheils, T.; Oprea, T. I. SmartGraph: A Network Pharmacology Investigation Platform. *J. Cheminf.* **2020**, *12* (1), 5.

(38) Himmelstein, D. S.; et al. Systematic Integration of Biomedical Knowledge Prioritizes Drugs for Repurposing. *eLife* **2017**, *6*, e26726.

(39) Ostaszewski, M.; Niarakis, A.; Mazein, A.; Kuperstein, I.; Phair, R.; Orta-Resendiz, A.; Singh, V.; Aghamiri, S. S.; Acencio, M. L.; Glaab, E.; et al. COVID-19 Disease Map, a computational repository of SARS-CoV-2 virus-host interaction mechanisms. *Mol. Syst. Biol.* **2021**, *17* (10), e10387.

(40) Ostaszewski, M.; Mazein, A.; Gillespie, M. E.; Kuperstein, I.; Niarakis, A.; Hermjakob, H.; Pico, A. R.; Willighagen, E. L.; Evelo, C. T.; Hasenauer, J.; Schreiber, F.; Drager, A.; Demir, E.; Wolkenhauer, O.; Furlong, L. I.; Barillot, E.; Dopazo, J.; Orta-Resendiz, A.; Messina, F.; Valencia, A.; Funahashi, A.; Kitano, H.; Auffray, C.; Balling, R.; Schneider, R. COVID-19 Disease Map, building a computational repository of SARS-CoV-2 virus-host interaction mechanisms. *Sci. Data.* **2020**, 7 (1), 136.

(41) Mering, C. v. STRING: a Database of Predicted Functional Associations Between Proteins. *Nucleic Acids Res.* **2003**, *31* (1), 258−261.

(42) Szklarczyk, D.; et al. STRING v11: Protein−Protein Association Networks With Increased Coverage, Supporting Functional Discovery in Genome-Wide Experimental Datasets. *Nucleic Acids Res.* **2019**, *47* (D1), D607−D613.

(43) Doncheva, N. T.; Morris, J. H.; Gorodkin, J.; Jensen, L. J. Cytoscape StringApp: Network Analysis and Visualization of Proteomics Data. *J. Proteome Res.* **2019**, *18* (2), 623−632.

(44) Apweiler, R. UniProt: The Universal Protein Knowledgebase. *Nucleic Acids Res.* **2004**, *32* (90001), 115D−119.

(45) The UniProt Consortium. UniProt: A Worldwide Hub of Protein Knowledge. *Nucleic Acids Res.* **2019**, *47* (D1), D506−D515.

(46) Patient, S.; Wieser, D.; Kleen, M.; Kretschmann, E.; Jesus Martin, M.; Apweiler, R. UniProtJAPI: A remote API for Accessing UniProt Data. *Bioinformatics* **2008**, *24* (10), 1321−1322.

(47) The UniProt Consortium. UniProt: The Universal Protein Knowledgebase. *Nucleic Acids Res.* **2017**, *45* (D1), D158−D169.

(48) UniProt Proggrammatic Service for ID Mapping, https://ftp.uniprot.org/pub/databases/uniprot/current_release/knowledgebase/idmapping/by_organism/HUMAN_9606_idmapping.dat.gz.

(49) Ursu, O.; et al. DrugCentral 2018: An Update. *Nucleic Acids Res.* **2019**, *47* (D1), D963−D970.

(50) Neo4j Graph Database. https://neo4j.com/.

(51) Python Core Team. *Python: A Dynamic, Open Source Programming Language. Python Software Foundation.* https://www.python.org/.

(52) Code Repository neo4covid19. https://github.com/ncats/neo4covid19.git.

(53) Python Library py2neo v4, https://py2neo.org/v4/.

(54) Colvis, C. M.; Austin, C. P. Innovation in Therapeutics Development at the NCATS. *Neuropsychopharmacology* **2014**, *39* (1), 230−232.

(55) Shannon, P.; et al. Cytoscape: a Software Environment for Integrated Models of Biomolecular Interaction Networks. *Genome Res.* **2003**, *13* (11), 2498−504.

(56) Warris, S.; Dijkxhoorn, S.; van Sloten, T.; van de Vossenberg, B. Mining Functional Annotations Across Species. *bioRxiv* **2018**, DOI: 10.1101/369785.

(57) CDC - Coronavirus Disease 2019 (COVID-19). https://www.cdc.gov/coronavirus/2019-ncov/need-extra-precautions/people-with-medical-conditions.html#diabetes, 2021.

(58) Crouse, A.; Grimes, T.; Li, P.; Might, M.; Ovalle, F.; Shalev, A. Metformin Use Is Associated With Reduced Mortality in a Diverse Population With Covid-19 and Diabetes. *medRxiv* **2020**, DOI: 10.1101/2020.07.29.20164020.

(59) Marvin Suite. ChemAxon Ltd., Molecules were depicted with ChemAxon's MarvinSketch 17.15.0, https://chemaxon.com/products/marvin.

(60) Shannon, P.; et al. Cytoscape: A Software Environment for Integrated Models of Biomolecular Interaction Networks. *Genome Res.* **2003**, *13* (11), 2498−2504.

(61) Wiese, R.; Eiglsperger, M.; Kaufmann, M. yFiles: Visualization and Automatic Layout of Graphs. *Proceedings of the 9th International Symposium on Graph Drawing (GD 2001)* **2002**, *2265*, 453.

(62) Daniloski, Z.; et al. Identification of Required Host Factors for SARS-CoV-2 Infection in Human Cells. *Cell* **2021**, *184* (1), 92−105.

(63) Abbott, T. R.; et al. Development of CRISPR as an Antiviral Strategy to Combat SARS-CoV-2 and Influenza. *Cell* **2020**, *181* (4), 865−876.

(64) Wei, J.; et al. Genome-wide CRISPR Screens Reveal Host Factors Critical for SARS-CoV-2 Infection. *Cell* **2021**, *184* (1), 76−91.

(65) Schneider, W. M.; et al. Genome-Scale Identification of SARS-CoV-2 and Pan-coronavirus Host Factor Networks. *Cell* **2021**, *184* (1), 120−132.

(66) Zhou, Y.; Hou, Y.; Shen, J.; Huang, Y.; Martin, W.; Cheng, F. Network-Based Drug Repurposing for Novel Coronavirus 2019-nCoV/SARS-CoV-2. *Cell Discovery* **2020**, *6* (1), 14.

(67) Zhou, Y.; et al. A Network Medicine Approach to Investigation and Population-Based Validation of Disease Manifestations and Drug Repurposing for COVID-19. *PLoS Biol.* **2020**, *18* (11), e3000970.

(68) Zhou, Y.; Wang, F.; Tang, J.; Nussinov, R.; Cheng, F. Artificial Intelligence in COVID-19 Drug Repurposing. *Lancet Digit. Heal.* **2020**, *2* (12), e667−e676.

(69) Zeng, X.; et al. Repurpose Open Data to Discover Therapeutics for COVID-19 Using Deep Learning. *J. Proteome Res.* **2020**, *19* (11), 4624−4636.

(70) Mehra, M. R.; Ruschitzka, F.; Patel, A. N. Retraction—Hydroxychloroquine or Chloroquine with or without a Macrolide for Treatment of COVID-19: A Multinational Registry Analysis. *Lancet* **2020**, *395* (10240), 1820.

(71) Mehra, M. R.; Desai, S. S.; Kuy, S.; Henry, T. D.; Patel, A. N. Cardiovascular Disease, Drug Therapy, and Mortality in Covid-19. *N. Engl. J. Med.* **2020**, *382* (25), e102.

(72) Heath, L. S.; Sioson, A. A. Multimodal Networks: Structure and Operations. *IEEE/ACM Trans. Comput. Biol. Bioinf.* **2009**, *6* (2), 321−32.

(73) Bojkova, D.; et al. Proteomics of SARS-CoV-2-Infected Host Cells Reveals Therapy Targets. *Nature* **2020**, *583* (7816), 469−472.

(74) Klann, K.; Tascher, G.; Münch, C. Functional Translatome Proteomics Reveal Converging and Dose-Dependent Regulation by mTORC1 and eIF2$\alpha$. *Mol. Cell* **2020**, *77* (4), 913−925.

(75) Hsu, P. D.; Lander, E. S.; Zhang, F. Development and Applications of CRISPR-Cas9 for Genome Engineering. *Cell* **2014**, *157* (6), 1262−1278.

(76) Wei, J.; et al. Genome-wide CRISPR Screen Reveals Host Genes that Regulate SARS-CoV-2 Infection. *bioRxiv* **2020**, DOI: 10.1101/2020.06.16.155101.

(77) Stern, M.; Jensen, R.; Herskowitz, I. Five SWI Genes are Required for Expression of the HO Gene in Yeast. *J. Mol. Biol.* **1984**, *178* (4), 853−868.

(78) Reactome, RUNX3 Regulates CDKN1A Transcription. https://reactome.org/content/detail/R-HSA-8941855.

(79) The GTEx Consortium. The GTEx Consortium Atlas of Genetic Regulatory Effects Across Human Tissues. *Science (Washington, DC, U. S.)* **2020**, *369* (6509), 1318−1330.

(80) Stathias, V.; et al. LINCS Data Portal 2.0: Next Generation Access Point for Perturbation-Response Signatures. *Nucleic Acids Res.* **2020**, *48* (D1), D431−D439.

(81) Fabregat, A.; et al. The Reactome Pathway Knowledgebase. *Nucleic Acids Res.* **2016**, *44* (D1), D481−D487.

(82) Oprea, T. I.; Yang, J. J.; Byrd, D. R.; Deretic, V. Autophagy Dark Genes: Can We Find Them With Machine Learning? *bioRxiv* **2019**, DOI: 10.1101/715037.

(83) Chen, T.; Guestrin, C. XGBoost. *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* **2016**, 785−794.

(84) Himmelstein, D. S.; Baranzini, S. E. Heterogeneous Network Edge Prediction: A Data Integration Approach to Prioritize Disease-Associated Genes. *PLoS Comput. Biol.* **2015**, *11* (7), e1004259.

(85) Code Repository ProteinGraphML. https://github.com/unmtransinfo/ProteinGraphML.

(86) Lasso, G.; et al. A Structure-Informed Atlas of Human-Virus Interactions. *Cell* **2019**, *178* (6), 1526−1541.

(87) Dunham, W. H.; Mullin, M.; Gingras, A.-C. Affinity-Purification Coupled to Mass Spectrometry: Basic Principles and Strategies. *Proteomics* **2012**, *12* (10), 1576−1590.

(88) P-HIPSTer. http://phipster.org/.

(89) Anatomical Therapeutic Chemical (ATC) Classification (WHO). https://www.who.int/medicines/regulation/medicines-safety/toolkit_atc/en/.

(90) International Nonproprietary Names (WHO).

(91) Armstrong, J. F. The IUPHAR/BPS Guide to PHARMACOLOGY in 2020: Extending Immunopharmacology Content and Introducing the IUPHAR/MMV Guide to Malaria Pharmacology. *Nucleic Acids Res.* **2019**, *48*, D1006.

(92) Olah, M.; Rad, R.; Ostopovici, L.; Bora, A.; Hadaruga, N.; Hadaruga, D.; Moldovan, R.; Fulias, A.; Mracec, M.; Oprea, T. I. WOMBAT and WOMBAT-PK: Bioactivity Databases for Lead and Drug Discovery. In *Chemical Biology: From Small Molecules to Systems Biology and Drug Design*; Schreiber, S. L., Kapoor, T. M., Wess, G., Eds.; Wiley-VCH: New York, 2007.

(93) Wishart, D. S.; et al. DrugBank 5.0: A Major Update to the DrugBank Database for 2018. *Nucleic Acids Res.* **2018**, *46* (D1), D1074−D1082.

(94) Kanehisa, M.; Goto, S.; Sato, Y.; Kawashima, M.; Furumichi, M.; Tanabe, M. Data, Information, Knowledge and Principle: Back to Metabolism in KEGG. *Nucleic Acids Res.* **2014**, *42*, D199−205.

(95) Sheahan, T. P. An Orally Bioavailable Broad-Spectrum Antiviral Inhibits SARS-CoV-2 and Multiple Endemic, Epidemic and Bat Coronavirus. *bioRxiv* **2020**, DOI: 10.1101/2020.03.19.997890.

(96) Hoffmann, M.; et al. SARS-CoV-2 Cell Entry Depends on ACE2 and TMPRSS2 and Is Blocked by a Clinically Proven Protease Inhibitor. *Cell* **2020**, *181* (2), 271−280.

(97) Mestres, J. The Target Landscape of N4-Hydroxycytidine Based on its Chemical Neighborhood. *bioRxiv* **2020**, DOI: 10.1101/2020.03.30.016485.

(98) *CLARITY* v4; Chemotargets S.L., Barcelona, https://www.chemotargets.com/PRODUCTS/CLARITY-v4, 2019.

(99) Garcia-Serna, R.; Vidal, D.; Remez, N.; Mestres, J. Large-Scale Predictive Drug Safety: From Structural Alerts to Biological Mechanisms. *Chem. Res. Toxicol.* **2015**, *28* (10), 1875−1887.

(100) Oprea, T. I. Unexplored Therapeutic Opportunities in the Human Genome. *Nat. Rev. Drug Discovery* **2018**, *17*, 317.