

# An Integrative Drug Repurposing Pipeline using KNIME and Programmatic Data Access: A case study on COVID-19 Data

Alzbeta Tuerkova<sup>1,\*</sup>, Barbara Zdrazil<sup>1,\*</sup>

<sup>1</sup> University of Vienna, Department of Pharmaceutical Chemistry, Division of Drug Design and Medicinal Chemistry, Althanstraße 14, A-1090 Vienna, Austria

\* Corresponding authors: [alzbeta.tuerkova@univie.ac.at](mailto:alzbeta.tuerkova@univie.ac.at); [barbara.zdrazil@univie.ac.at](mailto:barbara.zdrazil@univie.ac.at)

## Abstract

Biomedical information mining is increasingly recognized as a promising technique to accelerate drug discovery and development. Especially, integrative approaches which mine data from several (open) data sources have become more attractive with the increasing possibilities to programmatically access data through Application Programming Interfaces. The use of open data in conjunction with free, platform-independent analytic tools provides the additional advantage of flexibility, re-usability, and transparency. Here, we present a strategy for performing *in silico* drug repurposing with the analytics platform KNIME, using data for 38 suggested COVID-19 drug targets as a timely use case. The workflow includes a targeted download of data through web services, data curation (including chemical structure standardization), detection of enriched structural patterns, as well as substructure searches in DrugBank and a recently deposited dataset of antiviral drugs provided by Chemical Abstracts Service. Developed workflows, tutorials with detailed step-by-step instructions, and the information gained by the analysis of COVID-19 data are made freely available to the scientific community. The provided framework can be reused by researchers for other *in silico* drug repurposing projects, and it should serve as a valuable teaching resource for conveying integrative data mining strategies.

## Keywords

drug repurposing; data integration; data mining; data access; application programming interface; structure standardization; maximum common substructure; substructure search; drug repurposing; KNIME workflow; COVID-19; SARS-CoV-2

## List of Abbreviations

API	Application Programming Interface
KNIME	Konstanz Information Miner
CDK	Chemistry Development Kit
UniProtKB	The Universal Protein Resource KnowledgeBase
COVID-19	Coronavirus Disease 2019
SARS-CoV-2	Severe Acute Respiratory Syndrome Coronavirus 2
PDB	Protein Data Bank
NMR	Nuclear Magnetic Resonance
Cryo-EM	Cryo-Electron Microscopy
RCSB	Research Collaboratory for Structural Bioinformatics
AID	Assay ID
CID	Compound ID
CAS	Chemical Abstract Service
MCS	Maximum Common Substructure
PCA	Principal Component Analysis
ACE2_HUMAN	Angiotensin-Converting Enzyme 2 in Human
R1AB_CVHSA	Replicase polyprotein 1ab in SARS-COV
R1AB_SARS2	Replicase polyprotein 1ab in SARS-COV
R1A_CVHSA	Replicase polyprotein 1a in SARS-COV
ITAL_HUMAN	Integrin L-Alpha in Human
LabuteASA	Labute's Accessible Surface Area
SMR	Molecular Refractivity
TPSA	Topological Polar Surface Area

## Background

Computer-aided mining of biomedical data is an emerging field in cheminformatics and drug design which has reshaped current drug development. (1–3) Open access to various life-science repositories, such as ChEMBL (4), PubChem (5), UniProt (6), or DrugBank (7), has provided a competitive advantage when using data-driven drug discovery approaches as opposed to non-integrative approaches (8). Furthermore, many databases enable programmatic access of the stored data through an Application Programming Interface (API). Consequently, it is of importance to find appropriate tools to analyze gathered data in an automated way. The Konstanz Integration Miner (KNIME) is an open-source data pipelining and analytics platform which enables the creation of (semi)automated workflows to process, transform, analyze, and visualize the data as well as the generation and deployment of approximative mathematical models. (9) In the recent past, the KNIME community has released a plethora of cheminformatics extensions, such as the RDKit, (10) Chemistry Development Kit (CDK), (11) Indigo, (12) or Vernalis (13) toolkits.

In this study, we are providing a general strategy and a step-by-step tutorial for automated data access and data integration from multiple open data sources (which are providing an API), along with extensive data curation and cheminformatics data analysis by using the pipelining tool KNIME. Individual operations, such as the specification and execution of API requests, extraction of properties through JSON/XPath queries, structural data standardization, identification of enriched structural fragments, and substructure searches in external data sources, are thoroughly described and demonstrated herein.

Large-scale data fusion supplied with cheminformatics data analyses can uncover underlying patterns within the data and can pave the way for the development of novel medicine. Such a strategy can be leveraged for drug repurposing (also known as drug repositioning) strategies, in which a re-evaluation of an already approved drug can lead to a treatment for another disease. (14) This approach is particularly useful to, e.g., discover a cure for orphan diseases, (15) or to find a rapid solution for an ongoing pandemic, such as Coronavirus disease 2019 (COVID-19). (16) COVID-19 is a viral disease caused by the Severe Acute Respiratory Syndrome Coronavirus 2 (SARS-CoV-2) which currently represents a global health threat. (17) Up to now, no efficient treatment has been unraveled to combat COVID-19. In addition to the other research initiatives, such as the development of a vaccine, (18) or convalescent plasma therapy, (19) drug repurposing is a way to investigate already known drugs for treating novel diseases.

In this study, protein and ligand information related to potential treatment options for COVID-19 as a use case are gathered and analyzed to demonstrate the usefulness of automated workflows for data integration and cheminformatics analysis using programmatic data access from multiple open data sources. Just recently, about 66 druggable protein targets of SARS-CoV-2 have been

reported. (20) Here, proteins listed in the UniProtKB pre-release web page (available at <https://covid-19.uniprot.org/uniprotkb>) are used as a starting point. API calls are specified to map UniProt IDs of COVID-19 targets to available structural data in the Protein Data Bank (PDB). (21) Ligands co-resolved with a protein structure are extracted as a separate entity. For sake of data augmentation, ligand bioactivity measurements (such as  $K_i$ , IC50, or  $K_m$  end-points) for the protein targets under study are retrieved from ChEMBL (4), PubChem (5), and Guide-to-Pharmacology (IUPHAR). (22) After data cleaning and chemical structure standardization, Murcko scaffolds (23) are being extracted from the ligands in the dataset and grouped by similarity into structural queries for subsequent substructure searches in DrugBank (7) and the CAS COVID-19 antiviral candidate compounds dataset (available upon request at <https://www.cas.org/covid-19-antiviral-compounds-dataset>). These searches allow for the identification of structurally analogous compounds which could potentially show similar pharmacological action at suggested COVID-19 drug targets. A list of identified hits, along with a detailed analysis of COVID-19 data, is provided as an output of the workflow. A schematic overview of the whole data-driven drug repurposing workflow is depicted in Figure 1.

Taken together, the developed data mining pipeline is a useful resource for any *in silico* drug repurposing project and is exemplified on basis of a drug repositioning strategy for the Coronavirus Disease 2019 (COVID-19) which is currently representing a health threat of enormous worldwide impact. The step-by-step instructions allow for an easy implementation for other drug discovery projects along these lines and they shall give especially guidance to students or researchers new to the field of data-driven drug discovery. All workflows can be accessed via an open GitHub Repository (available at <https://github.com/AlzbetaTuerkova/Drug-Repurposing-in-KNIME>).

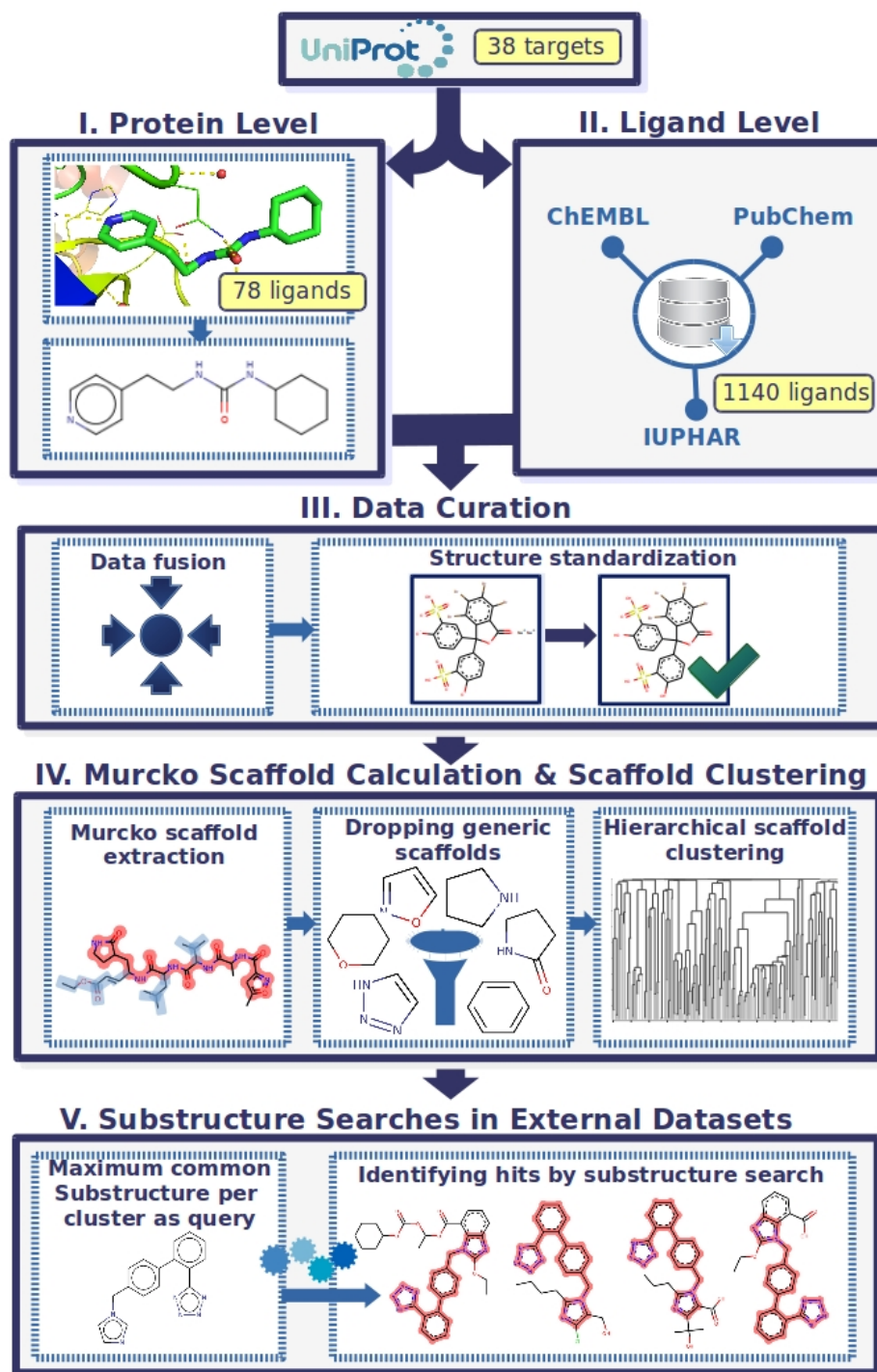


Figure 1: Schematic overview of the data-driven drug-repurposing workflow.

## Methods

### 1. Step: Programmatic access to UniProt and cross-referencing to retrieve structural data from the Protein Data Bank

UniProt IDs are used to retrieve available protein structures stored in the Protein Data Bank in Europe (PDBe). (21) When integrating data from diverse sources, it becomes beneficial to query databases programmatically, i.e., without the need of laborious manual data download and data integration. UniProtKB and other databases used in this example enable targeted access of the stored data through an Application Programming Interface (API). UniProt entries are returned in different file formats (.txt, .xml, .rdf, .fasta, .gff).

In the KNIME workflow discussed herein, a triad of KNIME nodes is consecutively executed (1) to specify the API request (via the ‘String Manipulation’ node), (2) to retrieve data from web services (via the ‘GET request’ node), and (3) to perform XPath queries to extract useful properties for a given protein (via the ‘XPath’ node). The respective part of the KNIME workflow is depicted in Figure 2.

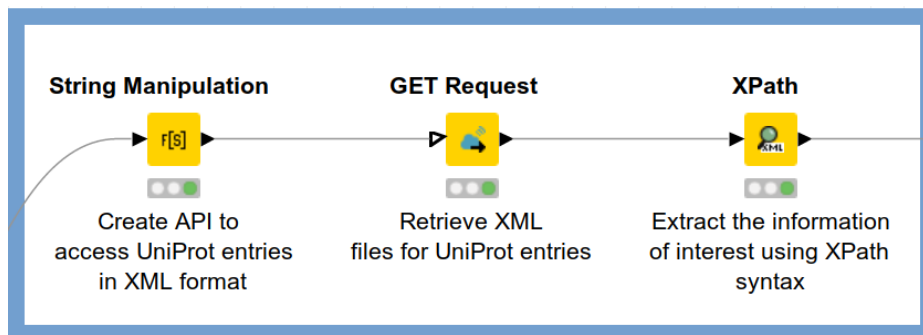


Figure 2: Three nodes for creation, execution, and post-processing API requests.

The input data is a list of UniProt IDs – in this use case for proteins that are listed to be of potential interest for treating COVID-19 (38 entries) – that is read in by a ‘File Reader’ node. Next, the ‘String Manipulation’ node is used to generate the API request for every UniProt ID from the input table. The `join()` function in the ‘String Manipulation’ node is used and the corresponding UniProt ID is forwarded to the string as a variable (`$UniProt ID$` column). The `strip()` function removes leading and trailing blanks from the `$UniProt ID$` column:

```
join("https://www.ebi.ac.uk/uniprot/api/covid-19/uniprotkb/accession/",strip($UniProt ID$),".xml")
```

As an output of the ‘String Manipulation’ node, a column with the respective API requests is appended to the output table, such as

```
https://www.ebi.ac.uk/uniprot/api/covid-
```

[19/uniprotkb/accession/O15393.xml](http://19/uniprotkb/accession/O15393.xml)

By executing the API requests (via the 'GET Request' node), the XML file is downloaded from UniProt and appended to the output table as XML cell. Additionally, columns reporting the content type, and the HTTP status code are appended (Figure 3). There exist five classes of HTTP status codes: (1) Informational responses (100-199), (2) Successful responses (200-299), (3) Redirects (300-399), (4) Client errors (400-499), and (5) Server errors (500-599). The information provided about the status of the request can be used to filter out any useless data entries. It is recommended to increase the timeout in the 'GET Request' configuration as the default specification (2 sec) is usually insufficient to receive all requested data.

S Uniprot ID	I Status	S Content type	XML XML
P59596	200	application/xml;charset=UTF-8	<pre>&lt;?xml version="1.0" encoding="UTF-8"?&gt; &lt;uniprot xmlns="http://uniprot.org/uniprot" xml &lt;entry created="2003-04-23" dataset="Swiss- &lt;accession&gt;P59596&lt;/accession&gt; &lt;accession&gt;Q658E5&lt;/accession&gt; &lt;accession&gt;Q7T6Q8&lt;/accession&gt; &lt;accession&gt;Q7T6R1&lt;/accession&gt; &lt;accession&gt;Q7T6R5&lt;/accession&gt; &lt;accession&gt;Q7T6S0&lt;/accession&gt; &lt;accession&gt;Q7T6S3&lt;/accession&gt; &lt;accession&gt;Q7T726&lt;/accession&gt; &lt;accession&gt;Q7T7P6&lt;/accession&gt; &lt;accession&gt;Q7T7S6&lt;/accession&gt; &lt;accession&gt;Q7TA12&lt;/accession&gt; &lt;name&gt;VME1_CVHSA&lt;/name&gt; &lt;protein&gt;</pre>

Figure 3: An example of the output table generated after the execution of the 'GET Request' node. Status, content type, and XML file are appended to the table as separate columns.

Subsequently, the 'XPath' node (XPath 1.0 version) is used to extract the information of interest on basis of querying different XML elements and the associated XML attributes. One can either define a XPath query within the 'XPath' node from scratch. Another way is to perform a double-click on a specific section in the XML-Cell Preview table and the XPath query is generated automatically. The XPath query below is used to retrieve all available PDB IDs for a given UniProt ID:

```
/dns:uniprot/dns0:entry/dns0:dbReference[@type='PDB']/@id
```

The 'dns0' prefix corresponds to the namespace used in the XPath query. Here, <http://uniprot.org/uniprot> is used as a namespace. Namespaces are defined automatically and are listed in the node configuration.

The example XPath query shows that PDB IDs are integrated within the <dbReference> XML element. However, UniProt entries consist of multiple <dbReference> elements which are pointing to different data sources, such as PubMed, GO, InterPro, Pfam, or PDB:

```
<dbReference type="PubMed" id="12730500"/>
<dbReference type="GO" id="GO:0039579">
<dbReference type="InterPro" id="IPR036333">
```

```
<dbReference type="Pfam" id="PF06478">  
<dbReference type="PDB" id="6NUR">
```

A key task is to query data from XML elements which do possess the ‘PDB’ attribute exclusively. The ‘@’ character is used to specify certain XML attributes in the XPath query. Therefore, `dbReference[@type='PDB']` is forwarded to the XPath query to get all PDB IDs by querying the `@id` attribute.

Programmatic access to the COVID-19 Data Portal is available through PDBe graph APIs. The PDB entities are returned in JSON format by default. Below an example is provided for a request to fetch protein structures for the ACE2 receptor (UniProt ID: Q9BYF1) from the COVID-19 Data Portal:

[https://www.ebi.ac.uk/pdbe/graph-api/mappings/best\\_structures/Q9BYF1](https://www.ebi.ac.uk/pdbe/graph-api/mappings/best_structures/Q9BYF1)

Similar to the ‘XPath’ node for processing XML documents, KNIME also provides the ‘JSON Path’ node which is used to process JSON data. The ‘JSON Path’ node enables to create JSON Path queries in both dot notation and bracket notation (depending on how the properties on an object are specified in the syntax). In the discussed KNIME workflow herein, the bracket notation is applied to extract the PDB IDs:

```
$..[*].['pdb_id']
```

Since the data are listed as a collection column type, the ‘JSON Path’ node is followed by the ‘UnGroup’ node to list multiple PDB IDs per protein target into separate rows. After concatenating data (‘Concatenate’ node) retrieved from PDB and the COVID-19 Data Portal, duplicates for a respective target were removed by grouping the data by target UniProt ID and PDB IDs (‘GroupBy’ node). The ‘PDB ID’ column is used to create the URL path to extract different properties by using the same strategy as shown in Figure 2. An example of such URL is given below:

<https://files.rcsb.org/view/2VYI.pdb>

The ‘PDB Loader’ and the ‘PDB Property Extractor’ nodes are available from the KNIME repository (created by Vernalis, Cambridge, UK) to facilitate analysis of PDB data in KNIME (Figure 4). These nodes were employed in order to explore properties of the PDB files, such as the experimental method used (X-ray diffraction, solution NMR, cryo-EM, theoretical models), the number of stored models, the resolution of structures, Space groups, R-factor, and so on.



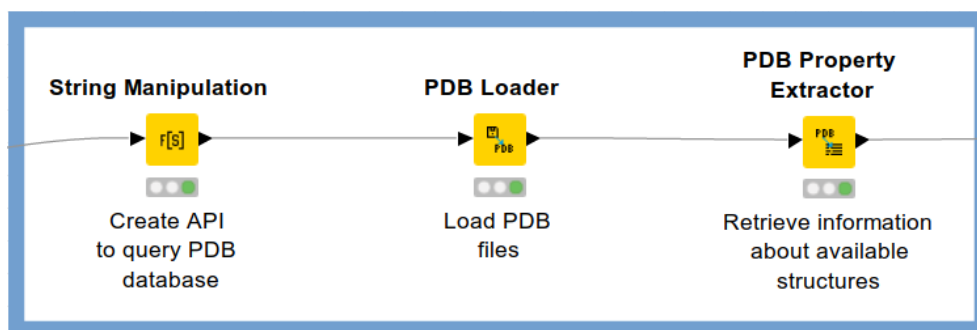


Figure 4: PDB nodes which enable to fetch and extract various properties of the deposited PDB structures.

Next, the available PDB structures were examined for their availability of co-resolved ligands. Ligand information can be received through the RCSB PDB RESTful Web services by creating the following request:

<https://www.rcsb.org/pdb/rest/ligandInfo?structureId=2VYI>

The XML column is processed by the 'XPath' node by using the following XPath queries:

```

/structureId/ligandInfo/ligand/chemicalName
/structureId/ligandInfo/ligand/@chemicalID
/structureId/ligandInfo/ligand/@molecularWeight
/structureId/ligandInfo/ligand/@structureId
/structureId/ligandInfo/ligand/@type
/structureId/ligandInfo/ligand/formula
/structureId/ligandInfo/ligand/InChIKey
/structureId/ligandInfo/ligand/InChI
  
```

Subsequently, PDB entries without a co-resolved ligand are filtered out (by applying the 'RowFilter' node). Chemical structures of the ligands can be displayed by converting the string format into a structural format (such as the SMILES notation) via the 'Molecule Type Cast' node. The 'GroupBy' node is used to keep unique ligand structures per protein target (grouping by UniProt ID and smiles string).

## 2. Step: Fetching ligand bioactivity data from open bioactivity data sources via programmatic data access

Orthogonal to fetching ligand data for potential COVID-19 targets from their protein structures, ligands and their bioactivities can also be collected from open pharmacological databases. In this example, data is retrieved from ChEMBLdb (version 26), (4) PubChem, (5) and IUPHAR (also known as Guide-to-Pharmacology, version 2020.2) (22) by using the respective web services via the ‘Get Request’ and ‘XPath’ nodes in KNIME. Automated data access can be achieved by using predefined identifiers for targets, ligands (such as ligand structure, available bioactivities, or molecule names), biochemical assays, and so on.

The KNIME workflow for fetching ChEMBL data allows to map UniProt IDs of COVID-19 drug targets to target ChEMBL IDs and subsequent retrieval of ligand bioactivities and their respective structural information (here: canonical smiles), document ChEMBL IDs, and Pubmed IDs for the primary publication. A major challenge is the limited number of bioactivities (up to 1,000 bioactivities) that are being fetched per single call. The KNIME workflow therefore has to be adopted to fetch all available data without manual intervention. The metanode that does the trick (termed ‘Get bioactivities per target’) works as follows:

1. A single XML file per target is downloaded and the number of bioactivities integrated within the `<total_count>` XML element is extracted.
2. The number of iterations needed to fetch all available bioactivities per target is calculated by dividing the number of bioactivities by 1,000 and then rounding the result up (`ceil()` function in the ‘Math Formula’ node).
3. A recursive loop is used in order to process protein targets one-by-one.
4. A nested loop is used within a recursive loop where the API call is modified in a way that it dynamically changes the ‘off-set’ parameter per each iteration; the ‘off-set’ parameter determines which is the number of bioactivities that should be skipped before downloading the next portion of bioactivities for a given target. After the loop ends, all information needed is extracted from the collected XML files by the ‘XPath’ node.

On basis of an example, this procedure shall be illustrated: There are 2,410 bioactivities for protein X available. Thus, three iterations are needed to fetch all data available for protein X. Within each iteration, a column is appended to the table containing the API call with the corresponding off-set parameter, i.e.

[https://www.ebi.ac.uk/chembl/api/data/activity?target\\_chembl\\_id=CHEMBL5118&limit=1000&offset=0](https://www.ebi.ac.uk/chembl/api/data/activity?target_chembl_id=CHEMBL5118&limit=1000&offset=0) (iteration#1)

[https://www.ebi.ac.uk/chembl/api/data/activity?target\\_chembl\\_id=CHEMBL5118&limit=1000&offset=1000](https://www.ebi.ac.uk/chembl/api/data/activity?target_chembl_id=CHEMBL5118&limit=1000&offset=1000) (iteration#2)

[https://www.ebi.ac.uk/chembl/api/data/activity?target\\_chembl\\_id=CHEMBL5118&limit=1000&offset=2000](https://www.ebi.ac.uk/chembl/api/data/activity?target_chembl_id=CHEMBL5118&limit=1000&offset=2000) (iteration#3)

At the end of the loop, 2,410 bioactivities have been collected for protein X and these are processed as indicated in the description above.

Step 3 and 4 from the workflow described above are visually depicted in Figure 5.

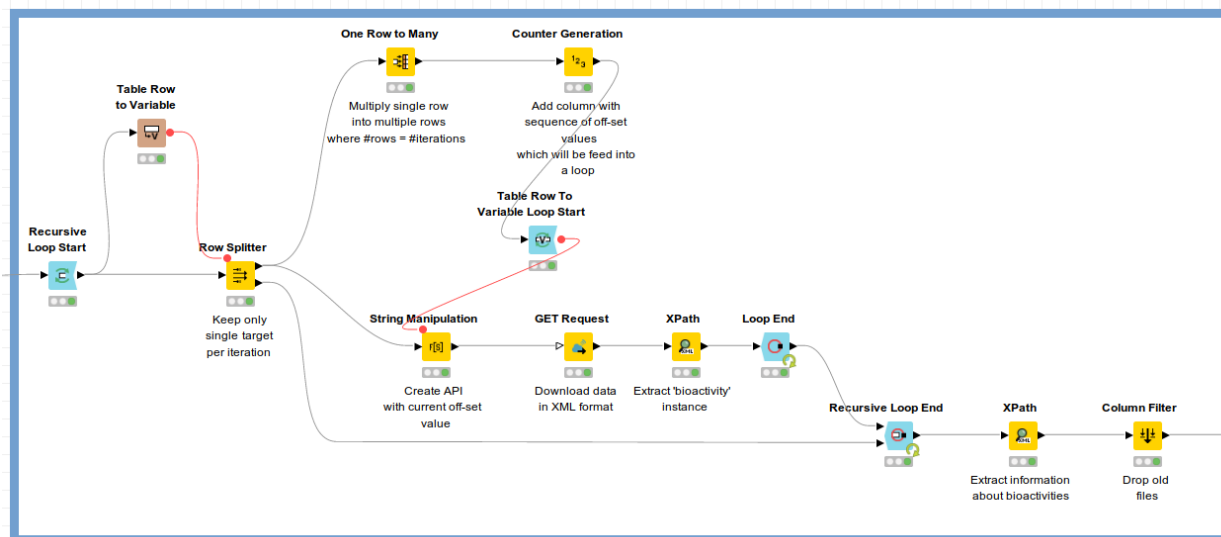


Figure 5: Nested recursive loops used to fetch the bioactivity data from ChEMBL.

In case of PubChem, UniProt IDs are mapped to ‘PubChem Assay IDs’ (AID) in the first step. Further, AIDs are mapped to available compounds by ‘PubChem Compound ID’ (CID), including bioactivity measurements and associated PubMed IDs. Compound structures and names are retrieved in the next step. In some cases, compound names in PubChem are included in the form of molecule ChEMBL IDs. If this condition is true, the ChEMBLdb is additionally queried to download a compound name, if available.

In order to query IUPHAR data, the UniProt ID is mapped to the IUPHAR target ID. API calls have a specific syntax for accessing substrates, e.g.:

<http://www.guidetopharmacology.org/services/targets/2421/substrates>

and for accessing inhibitors, e.g.:

<http://www.guidetopharmacology.org/services/targets/2421/interactions>

where “2421” is an identifier for a specific target ID. Compound ID, PubMed ID, affinity, affinity type (corresponding to a certain end-point), and action (corresponding to a certain activity annotation) were retrieved by using the ‘JSON Path’ node. Retrieval of the ligand

structural format is done by an additional API call on basis of the respective ligand ID.

Bioactivity values are converted to their negative logarithmic representation and binary labels ('1' for active and '0' for inactive) are assigned on basis of an activity cut-off. In this example, all compounds possessing a negative logarithmic value greater than 9 (i.e., < 1 nM) were labeled as '1', while the rest was labeled as '0'.

After merging the output tables from ChEMBL, PubChem, and IUPHAR, data are grouped to keep unique ligands per target and median values for binary activity labels ( by using the 'GroupBy' node). In addition, only active ligands per target (label '1') are kept and the final table is concatenated with ligand structures from PDB entries.

A prerequisite for merging ligand data from diverse sources is the standardization of the molecular representation. A similar curation strategy like the one published by Gadaleta et al. [\(24\)](#) was applied:

1. Characters encoding stereoisomerism in SMILES format (@; \; /) are removed by using the 'String Replacer' node since for the subsequent operations this information is not needed.
2. Salts are stripped by using the 'RDkit Salt Stripper' node. (This node works with pre-defined sets of different salts/salt mixtures by default. If requested, additional salt definitions can be forwarded to the node.)
3. Salt components are listed in the output table using the 'Connectivity' node (CDK plugin) followed by the 'Split Collection Column' node.
4. The 'RDKit Structure Normalizer' node neutralizes charges and checks for atomic clashes, etc. Additional criteria for compound quality check can be adjusted in the 'Advanced' section of the node configuration.
5. The 'Element Filter' node keeps compounds containing the following elements only: H,C,N,O,F,Br,I,Cl,P,S).
6. InChI, InChiKey, and Canonical smiles formats are finally created from the standardized compounds.

Steps 2-4 are visually depicted in Figure 6.

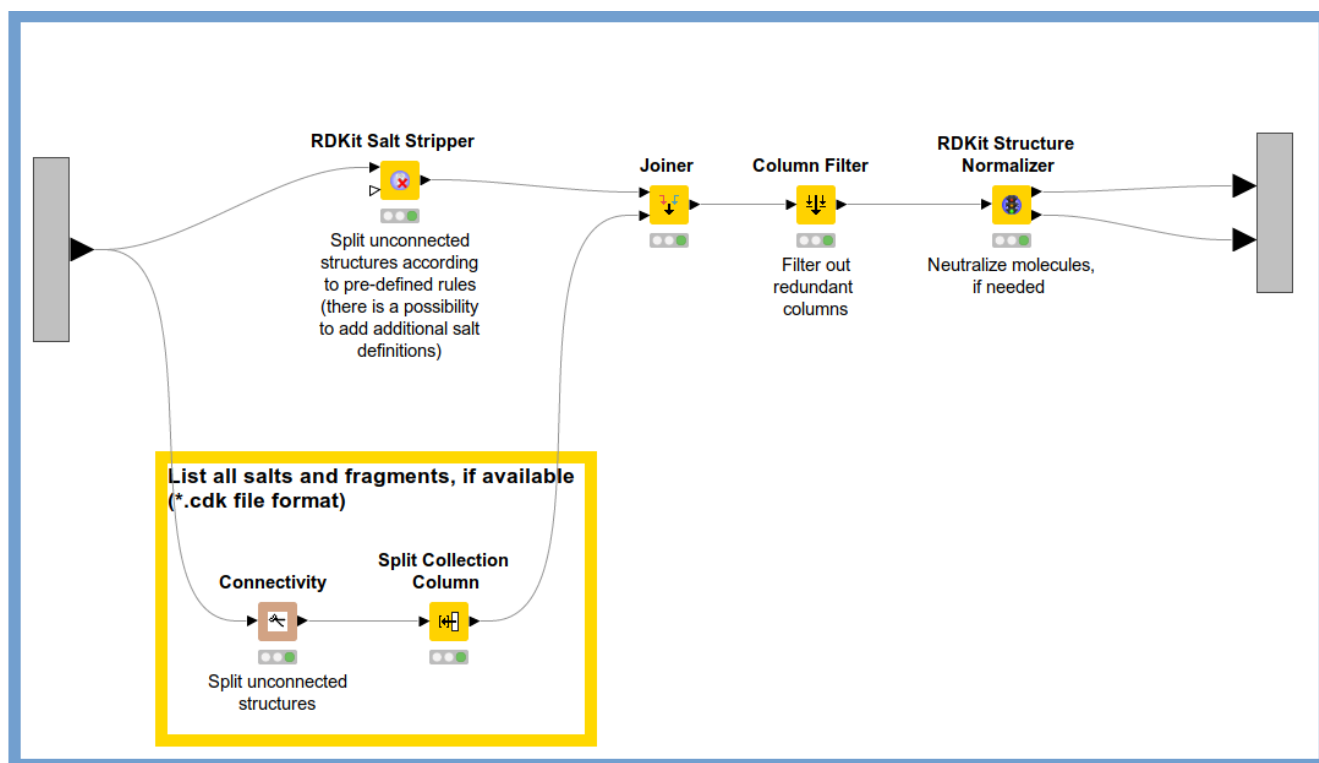


Figure 6: A part of the standardization workflow used to strip salts and neutralize charges.

### 3. Step: Substructure searches to identify potentially interesting compounds for drug repurposing

Finally, the merged datasets are used to generate structural queries in SMARTS format in order to perform substructure searches in DrugBank (version 5.1.6, approx. 10,000 compounds) and in the COVID-19 antiviral candidate compound dataset provided by the Chemical Abstracts Service (approx. 50,000 compounds, available upon request at <https://www.cas.org/covid-19-antiviral-compounds-dataset>).

Murcko scaffolds are extracted ('RDKit Find Murcko Scaffolds' node) in order to get a quick overview of the structural diversity of the curated dataset. Scaffolds possessing too generic structures (i.e., a single aromatic ring) can be filtered out (by using the 'RDKit Descriptors Calculator' node in conjunction with the 'Row Filter' node) and remaining ones can be explored with respect to their structural similarity in the context of a certain target. This step is done by (1) calculating molecular distances using the maximum common substructure as a metric of similarity ('MoSS MCSS Molecule Similarity' node), (2) hierarchical clustering (the 'Hierarchical Clustering [DistMatrix]' node), and (3) assigning a threshold (here: distance threshold = 0.5) for cluster assignment (the 'Hierarchical Cluster Assigner' node). The respective

part of the workflow is depicted in Figure 7.

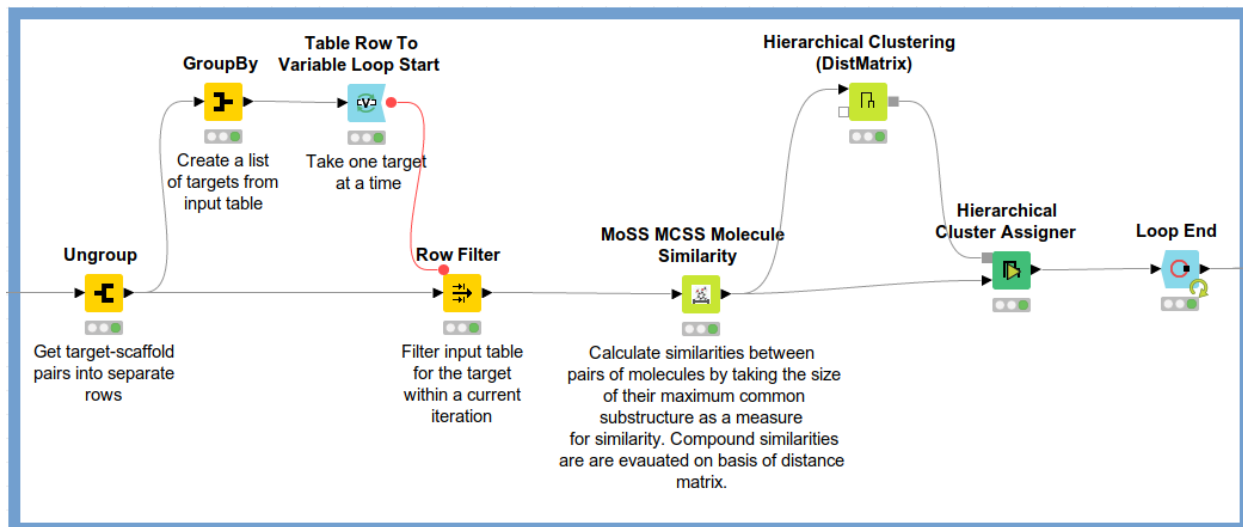


Figure 7: Hierarchical scaffold clustering in KNIME.

Next, looping over distinct clusters of associated Murcko scaffolds for a certain target is done in order to create a maximum common substructure (the ‘RDKit MCS’ node) from all associated Murcko scaffolds belonging to a respective cluster. Recursive loops are extensions to regular loops which can be used in conjunction with a ‘Row Splitter’ node to separate the current row from the rest of the table. After termination of the current iteration, the rest of the table is forwarded to the loop start and the next row is used for the subsequent iteration (see Figure 8). Generated substructures for a certain target are appended to the output table in SMARTS format.

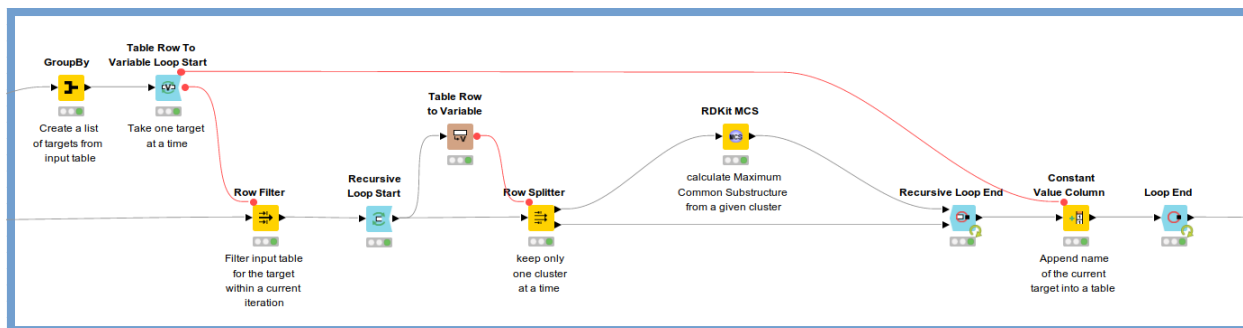


Figure 8: Looping through the scaffold clusters and generating a maximum common substructure for a given cluster.

Also for the substructure searches in DrugBank and the CAS dataset loops are being used (Figure 9). The ‘Table Row To Variable Loop Start’ forwards each substructure as a query to the ‘RDKit Substructure Filter’ node as a flow variable which then examines whether a particular substructure is contained in the data sets from DrugBank or CAS. Extracted compounds are being forwarded to the ‘RDKit molecule highlighting’ node which visualizes the highlighted

substructure within the respective compounds.

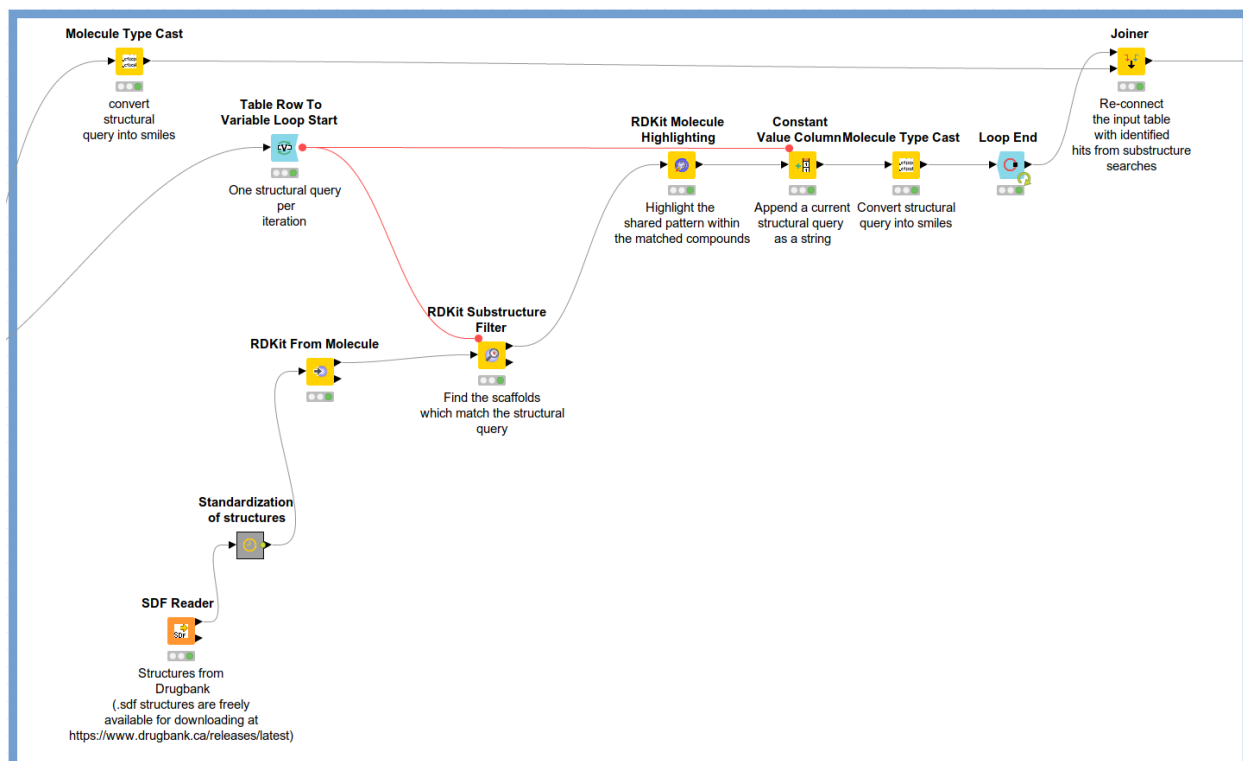


Figure 9: Automated substructure searches in KNIME.

## Software

KNIME workflows were built in KNIME version 4.1. Box plots, cumulative distribution function, and principal component analysis, and all data visualizations were performed in R version 3.4.4. (25) The KNIME workflows are freely available from GitHub (<https://github.com/AlzbetaTuerkova/Drug-Repurposing-in-KNIME>). The published workflow can be either used as a single pipeline, or as multiple stand-alone workflows (1) to gather data from PDB, (2) to retrieve ligand bioactivities from ChEMBL, PubChem, and IUPHAR, and (3) to perform substructure searches, by providing the needed data input, respectively.

## Results and discussion

In this contribution, a semiautomatic KNIME workflow for drug repurposing based on publicly available ligand data is presented. The pipeline includes automatic mapping of UniProtKB entries and PDB via cross-referencing, programmatic data access via the data sources' web services (exemplified for ChEMBL, PubChem, and IUPHAR), fully automatic data curation

(including chemical data standardization, removal of duplicates, and cut off setting for assigning activity labels), the identification of common structural patterns in SMARTS format, and substructure searches (here in DrugBank and the CAS dataset of antiviral drugs) in order to identify interesting compounds for further investigations.

## Data retrieval

The Universal Protein Resource KnowledgeBase (UniProtKB) is a freely accessible database for protein sequence and annotation data. The UniProt ID (e.g, P59596, P59637, P0C6X7) is a protein identifier which can be used to retrieve comprehensive information about a given protein, including protein names and synonyms, taxonomy, function, cellular localization, available three-dimensional structures, as well as cross-references to other databases. Cross-referenced databases include (but are not limited to) sequence databases (e.g., EMBL (26), GenBank (27), CCDS (28)), 3D structure databases (e.g, Protein Data Bank (29), ModBase (30), SWISS-MODEL-Workspace (31)), protein-protein interaction databases (e.g, Biogrid (32), IntAct (33), STRING (34)), and chemistry databases (e.g., BindingDB, (35) ChEMBL, (4) DrugBank (7)). In the framework of this case study, content from a pre-release UniProt web page (available at <https://covid-19.uniprot.org/uniprotkb>) was used as an input for the data mining pipeline to gather and analyze data for proteins potentially interesting for the treatment of infections with human SARS-CoV2 (38 proteins; see Table 1). As seen from Table 1, available protein templates include 14 SARS-CoV2, 15 SARS-CoV, and 9 structures with origin Homo Sapiens. Listed UniProt IDs were used to retrieve protein structures stored in PDB (428 structures, 386 unique structures) and the COVID-19 Data Portal (421 structures, 404 unique structures) which has been launched just recently as a response to the COVID-19 pandemic (available from <https://www.ebi.ac.uk/pdbe/covid-19>), as well as available ligand bioactivities from ChEMBL, PubChem, and IUPHAR (3,113 bioactivities).

Table 1: Drug targets with potential interest for treatment of COVID-19 (available from <https://www.ebi.ac.uk/pdbe/covid-19>).

UniProt ID	Target Name	Organism	Target Shortcut
O15393	Transmembrane protease serine 2	Homo sapiens	TMPS2_HUMAN
Q92499	ATP-dependent RNA helicase DDX1	Homo sapiens	DDX1_HUMAN
Q9BYF1	Angiotensin-converting enzyme 2	Homo sapiens	ACE2_HUMAN
O43765	Small glutamine-rich tetratricopeptide repeat-	Homo sapiens	SGTA_HUMAN
P20701	containing protein alpha	Homo sapiens	ITAL_HUMAN
P35232	Integrin alpha-L	Homo sapiens	PHB_HUMAN



P84022	Prohibitin	Homo sapiens	SMAD3_HUMAN
Q8N3R9	Mothers against decapentaplegic homolog 3	Homo sapiens	MPP5_HUMAN
Q99623	MAGUK p55 subfamily member 5	Homo sapiens	PHB2_HUMAN
P0C6U8	Prohibitin-2	SARS COV	R1A_CVHSA
P0C6X7	Replicase polyprotein 1a	SARS COV	R1AB_CVHSA
P0DTC1	Replicase polyprotein 1ab	SARS COV-2	R1A_SARS2
P0DTD1	Replicase polyprotein 1a	SARS COV-2	R1AB_SARS2
P0DTC2	Replicase polyprotein 1ab	SARS COV-2	SPIKE_SARS2
P59594	Spike glycoprotein	SARS COV	SPIKE_CVHSA
P59595	Spike glycoprotein	SARS COV	NCAP_CVHSA
P59632	Nucleoprotein	SARS COV	AP3A_CVHSA
P59635	Protein 3a	SARS COV	NS7A_CVHSA
P59637	Protein 7a	SARS COV	VEMP_CVHSA
P59596	Envelope small membrane protein	SARS COV	VME1_CVHSA
P59633	Membrane protein	SARS COV	NS3B_CVHSA
P0DTC3	Non-structural protein 3b	SARS COV-2	AP3A_SARS2
P0DTC5	Protein 3a	SARS COV-2	VME1_SARS2
P0DTC7	Membrane protein	SARS COV-2	NS7A_SARS2
P0DTC9	Protein 7a	SARS COV-2	NCAP_SARS2
P59634	Nucleoprotein	SARS COV	NS6_CVHSA
P59636	Non-structural protein 6	SARS COV	ORF9B_CVHSA
P0DTC4	Protein 9b	SARS COV-2	VEMP_SARS2
P0DTC6	Envelope small membrane protein	SARS COV-2	NS6_SARS2
P0DTD2	Non-structural protein 6	SARS COV-2	ORF9B_SARS2
Q7TFA1	Protein 9b	SARS COV	NS7B_CVHSA
Q80H93	Protein non-structural 7b	SARS COV	NS8B_CVHSA
P0DTC8	Non-structural protein 8b	SARS COV-2	NS8_SARS2

P0DTD3	Non-structural protein 8	SARS COV-2	Y14_SARS2
P0DTD8	Uncharacterized protein 14	SARS COV-2	NS7B_SARS2
Q7TFA0	Protein non-structural 7b	SARS COV	NS8A_CVHSA
Q7TLC7	Protein non-structural 8a	SARS COV	Y14_CVHSA
A0A663DJA2	Uncharacterized protein 14 ORF10 protein	SARS COV-2	A0A663DJA2_SARS2

In total, 429 unique protein structures were retrieved, with 362 of the structures being listed in both sources, 24 in PDB only, and 43 in the COVID-19 Data Portal only. From these sources, 78 unique ligands could be extracted, yielding 47 unique Murcko scaffolds.

From the orthogonal approach – the automatic gathering of ligand bioactivity data from ChEMBL, PubChem, and IUPHAR via its webservices - 1,114 unique ligands with (median) activity value <1 nM were identified (522 unique Murcko scaffolds).

### Analysis of compiled datasets

The final dataset used for generating structural queries for substructure searches is composed of 1,181 unique compounds. Numbers of unique compounds per individual COVID-19 drug target and data source are listed in Table 2. As visible from the Venn diagram in Figure 10, PubChem is the predominant source of ligands (912 unique compounds corresponding to 77%). At the other end of the scale, IUPHAR provides only nine unique compounds.

Table 2: Number of unique ligands gathered from PDB, ChEMBL, PubChem, and IUPHAR.

Target shortcut	PDB	ChEMBL	IUPHAR	PubChem	# Unique compounds
ITAL_HUMAN	13	94	2	550	564
R1AB_CVHSA	37	187	0	47	227
ACE2_HUMAN	4	65	3	161	172
R1A_CVHSA	35	92	0	79	141
SMAD3_HUMAN	3	64	0	65	71

DDX1_HUMAN	0	7	0	7	14
R1AB_SARS2	14	0	0	0	9
TMPS2_HUMAN	2	3	4	3	7
SPIKE_SARS2	5	0	0	0	5
SPIKE_CVHSA	5	0	0	0	5
SGTA_HUMAN	2	0	0	0	2
R1A_SARS2	2	0	0	0	2
VME1_CVHSA	2	0	0	0	2
MPP5_HUMAN	1	0	0	0	1
ORF9B_CVHSA	1	0	0	0	1

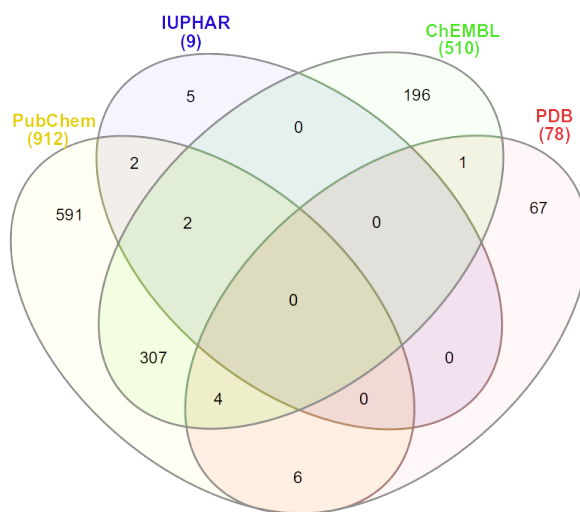


Figure 10: Constitution of the final dataset: Venn diagram showing compound overlap across different data sources.

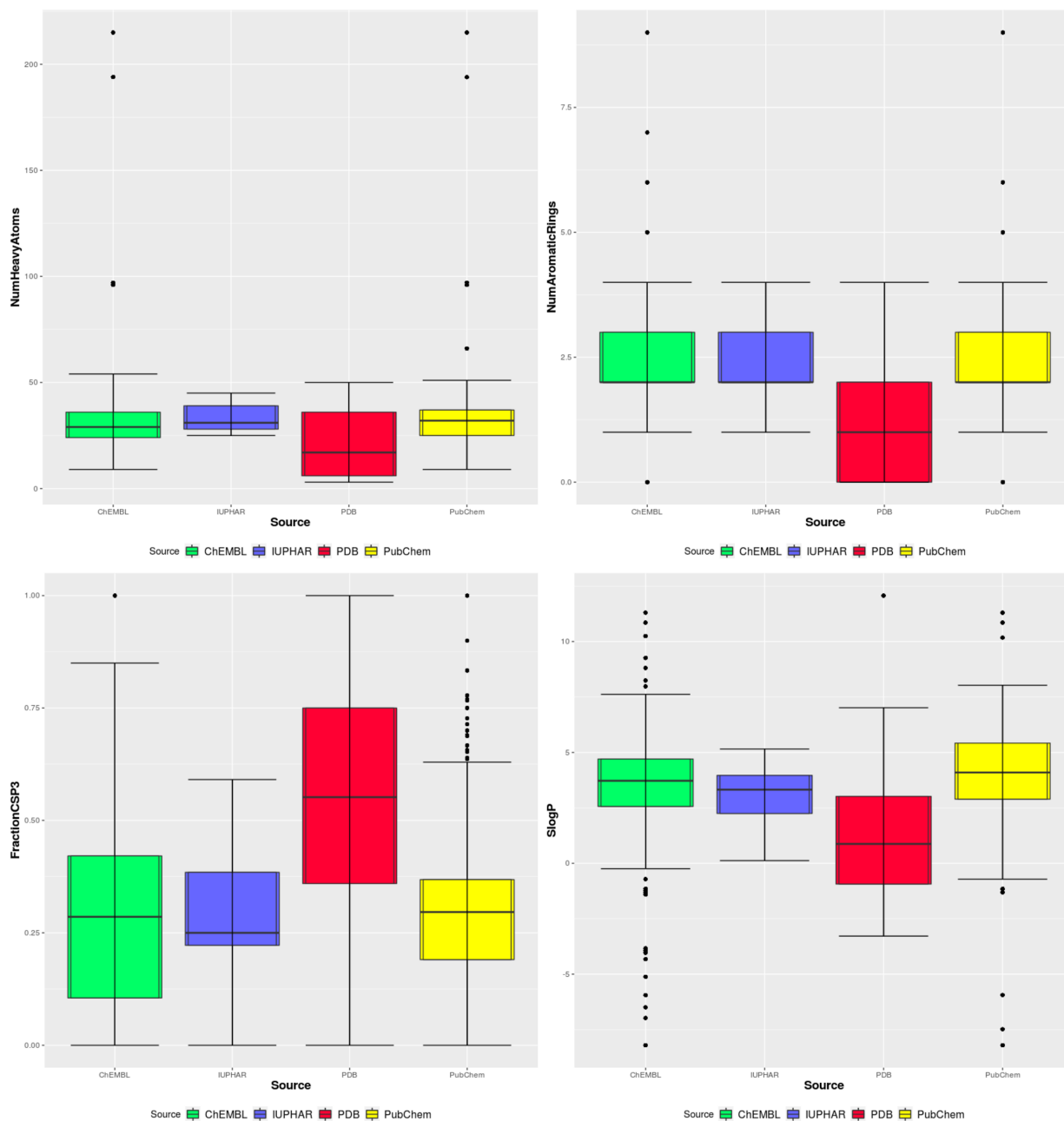


Figure 11: Box- and whisker plots showing the range of values for selected physicochemical properties in different data source: ChEMBL ... green, IUPHAR ... blue, PDB ... red, PubChem ... yellow.

Examining the distribution of different drug discovery relevant physicochemical properties, compound data extracted from PDB appears to possess compound structures chemically most dissimilar from all other data sources (Figure 11). Co-resolved ligands from PDB structures are in general smaller, less lipophilic, less aromatic, and less planar. Many PDB structures contain endogenous ligands which are generally of smaller size. For example, spike glycoprotein (UniProt ID P59594, PDB IDs 2AJF, 6CRV, 6CRW, 6CRX, 6CRZ, 6CS0, 6CS1) contains beta-D-mannose (12 heavy atoms) as co-resolved ligand (whereas 50% of the ligand data coming from other sources than PDB do possess more than 25 heavy atoms; Figure 11).

Inspecting the origin of data for the respective protein targets, it becomes apparent that ligand information for human SARS-CoV2 solely originates from PDB structures (see Table 2, entries ending with “\_SARS2”). Notably, the majority of structures for SARS-CoV2 - such as PDB IDs 6W4B [to be published], 6Y2E, or 6Y2G for replicase polyprotein 1a (36) were refined via molecular replacement based on the homology to SARS-COV. It therefore seems to be beneficial to integrate data from diverse sources, especially including PDB as a source for most up-to-date compound information.

Across all data sources, the largest number of ligand bioactivity measurements was gathered for human integrin alpha-L (UniProt ID P20701, 564 unique compounds), followed by SARS replicase polyprotein 1ab (UniProt ID P0C6X7, 227 unique compounds), human angiotensin-converting enzyme 2 (ACE2; UniProt ID Q9BYF1, 172 unique compounds), SARS replicase polyprotein 1a (UniProt ID P0C6U8, 141 unique compounds), and human mothers against decapentaplegic homolog 3 (UniProt ID P84022, 71 unique compounds). ACE2 receptor is considered a relevant therapeutic target due to its interaction with spike glycoprotein of coronaviruses when entering host cells. (37) Replicase polyproteins 1a and 1ab are attractive targets to treat COVID-19 given their crucial role in replication and transcription of viral RNAs. (38) Surprisingly, integrin alpha-L protein ranked as a target with the most data does not belong to the notoriously debated targets of COVID-19. A current study has suggested a potential role of integrins as alternative receptors for SARS-CoV-2, as the spike glycoprotein contains an integrin-binding motif. (39)

### **Substructure searches in external datasets**

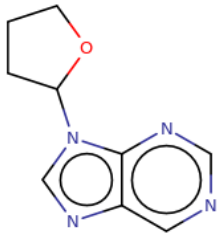
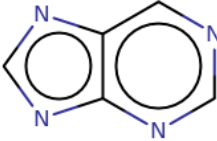
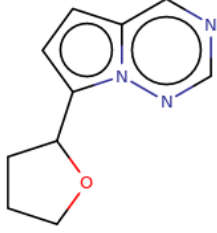
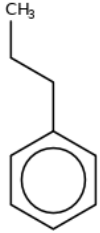
Chemical (molecular) similarity is a traditional concept in the field of cheminformatics. (40) It is

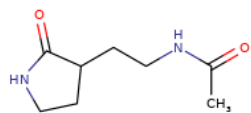
used to identify structural analogs which might exert similar biological action on similar biological targets. (41) Common cheminformatics similarity approaches are based on the global similarity of a molecule. For example, fingerprint-based descriptors are used to evaluate compound similarity by quantifying the presence/absence of the specific structural features (e.g., distinct functional groups in a molecule). On the contrary, molecular graph-based methods do capture a specific molecular topology and hence account for the local similarity of molecules. (42) Graph-based methods are therefore a robust tool to, e.g., distinguish between different structural isomers (such as n-pentane and dimethylpropane). Here, Maximum Common Substructures (MCS) of a compound collection were used as structural keys for detecting new potential drug candidates. Such substructure searches are especially useful for drug repositioning strategies, since they capture more the local similarity of chemical compounds and therefore allow for more flexibility than global similarity measures (especially if there are large size differences of the chemical compounds that are compared).

In a first instance, a Murcko scaffold for identified ligands was calculated. For each target, Murcko scaffolds were grouped into hierarchical clusters by considering their Maximum Common Substructure (MCS) as a measure of similarity. Afterwards, looping in KNIME was applied to generate a MCS (in SMARTS) per cluster (and target). For details see the Methods Section. In total, 91 distinct MCSs of a variable number of atoms (from 9 to 46) and bonds (from 9 to 50) were calculated. A complete list of MCSs can be found in Supplementary File S1.

Structural queries generated in the previous step have identified 3,102 compounds from DrugBank and 18,135 compounds from the CAS dataset. A complete list of hits found by the substructure searches is provided in Supplementary File S2 (DrugBank) and S3 (CAS dataset). Out of those hits, 128 compounds were retrieved from both DrugBank and the CAS dataset (Supplementary File S4) and were identified on basis of 11 distinct MCSs which can be combined into six separate clusters (Table 3): (1) Nucleoside/nucleotide analogs (81 hits), (2) Miscellaneous, which contain ubiquitous substructures which partly overlap with each other (30 hits), (3) Peptide-based analogs (7 hits), (4) Biphenyl analogs (5 hits), (7) Indole derivatives (3 hit), (6) Statin analogs (1 hit),

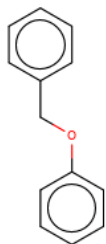
Table 3: Six clusters of MCSs (11 in total) which were retrieved from DrugBank and the CAS dataset. The structural fragment, SMARTS string, the number of identified hits, and the protein target(s) for which these hits have been found, are given.

Cluster number	Structural Fragment	SMARTS String	# Hits	Targets
1		<chem>[*6]-[*6]-[*8]-[*6](-[*6]-1)-n1cnc2cncnc12</chem>	59	R1A_CVHSA, R1AB_CVHSA, R1AB_SARS2
		<chem>C1nc2cncnc2n1</chem>	21	R1AB_SARS2, ACE2_HUMAN
		<chem>[*6]-[*6]-[*8]-[*6](-[*6]-1)-c1ccc2cncnn12</chem>	1	R1AB_SARS2
2		<chem>[*6]-[*6]-[*6]-c1ccccc1</chem>	27	ACE2_HUMAN



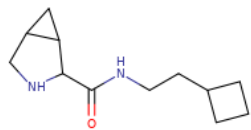
[#6]-[#6](=O)-[#7]-[#6]-[#6]-  
[#6]-1-[#6]-[#6]-[#7]-[#6]-1=O 2

R1A\_CVHSA,  
R1AB\_CVHSA,  
ACE2\_HUMAN



[#6](-[#8]-c1ccccc1)-c1ccccc1 1

ACE2\_HUMAN



O=[#6](-[#7]-[#6]-[#6]-[#6]-1-  
[#6]-[#6]-[#6]-1)-[#6]-1-[#7]-  
[#6]-[#6]-2-[#6]-[#6]-1-2 1

R1AB\_SARS2

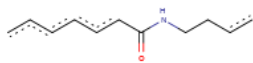


3

[#6][#6][#6][#6][#6][#6]-[#6]  
 (=O)-[#7]-[#6]-[#6]-[#6][#6]

7

ACE2\_HUMAN

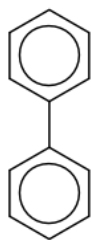


4

c1ccc(cc1)-c1ccccc1

5

ACE2\_HUMAN

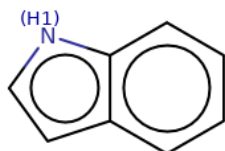


5

c1cc2ccccc2[nH1]1

3

ACE2\_HUMAN

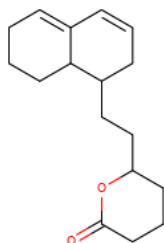


6

O=[#6]-1-[#6]-[#6]-[#6]-[#6](-  
 [#6]-[#6]-[#6]-2-[#6]-  
 [#6]=[#6]-[#6]-3=[#6]-[#6]-  
 [#6]-[#6]-[#6]-2-3)-[#8]-1

1

ITAL\_HUMAN



Using the automated data mining and integration pipeline we were able to pinpoint drugs which are currently under clinical trials and/or investigation in relation to COVID-19. Some of the hits found are described below. These findings approve our methodology for finding therapeutics for drug repositioning. In addition, we are providing a list of drugs to the research community that could be interesting to investigate further in the framework of COVID-19 drug repurposing strategies. Figure 14 shows examples of identified hits for the most pronounced clusters.

Nucleoside/nucleotide analog inhibitors are a class of medicals which mimic nucleotide substrates. Such therapeutics are having crucial implications in the treatment of viral infections, as they interact with polymerase and thus induce the termination of replication or transcription of viral RNAs. Nucleoside/nucleotide analogs were found for SARS-Cov/SARS-Cov2 proteases (Table 3). One of the identified hits is Remdesivir (DrugBank ID DB14761) which is considered a top candidate for the treatment of COVID-19 to date. (38) Delavirdine ( DrugBank ID DB00705) was identified as a sole representative of indole derivatives. Delavirdine belongs to the class of non-nucleoside reverse transcriptase inhibitors. Multiple computer-based studies were performed to exemplify its effect on the COVID-19 drug targets. (43) Next, Dasabuvir (DrugBank ID DB09183) is an antiviral drug used to treat hepatitis C type-1. It has been identified due to the presence of a biphenyl scaffold. Molecular modeling approaches have been used to elucidate the potential role of Dasabuvir to combat COVID-19. (44–46) However, these studies require additional investigations to validate the results. Lovastatin (DrugBank ID DB00227) was identified as the only analog of statins, suggesting the interaction with Integrin alpha-L (UniProt ID P20701). Interestingly, statin therapy was suggested for patients with COVID-19 just recently. (47) In addition, a cluster of compounds showing macrocyclic structures was identified from ChEMBL and PubChem. These compounds are showing a pronounced activity profile against ACE2 receptors and/or integrin alpha-L receptors. To the best of our knowledge, macrocyclic-like compounds have not been investigated in relation to COVID-19 so far. A cluster of miscellaneous substructures has delivered drugs which are currently under experimental investigations. For example, Ritonavir (DrugBank ID DB00503) is a HIV protease inhibitor used in combination with other drugs. (48) However, the effectiveness of Ritonavir/Lopinavir in the treatment of COVID-19 is still debated, as the current study shows that there is no significant improvement observed compared to the patients who were not treated by this drug combination. (49) Other potential candidates under investigation are Darunavir (DrugBank ID DB01264) or Rupintrivir (DrugBank ID DB05102). (56) Rupintrivir is commonly known protease inhibitor belonging to the class of antiviral agents. Rupintrivir has been predicted to be an interesting candidate not only for SARS-Cov/SARS-Cov2 proteases, but also for ACE2 receptor (Table 3). These findings suggest Rupintrivir as a promising candidate for experimental testing, since it could potentially possess the ability to interact with more than a single COVID-19 target which could potentially lead to a higher efficacy of that drug.

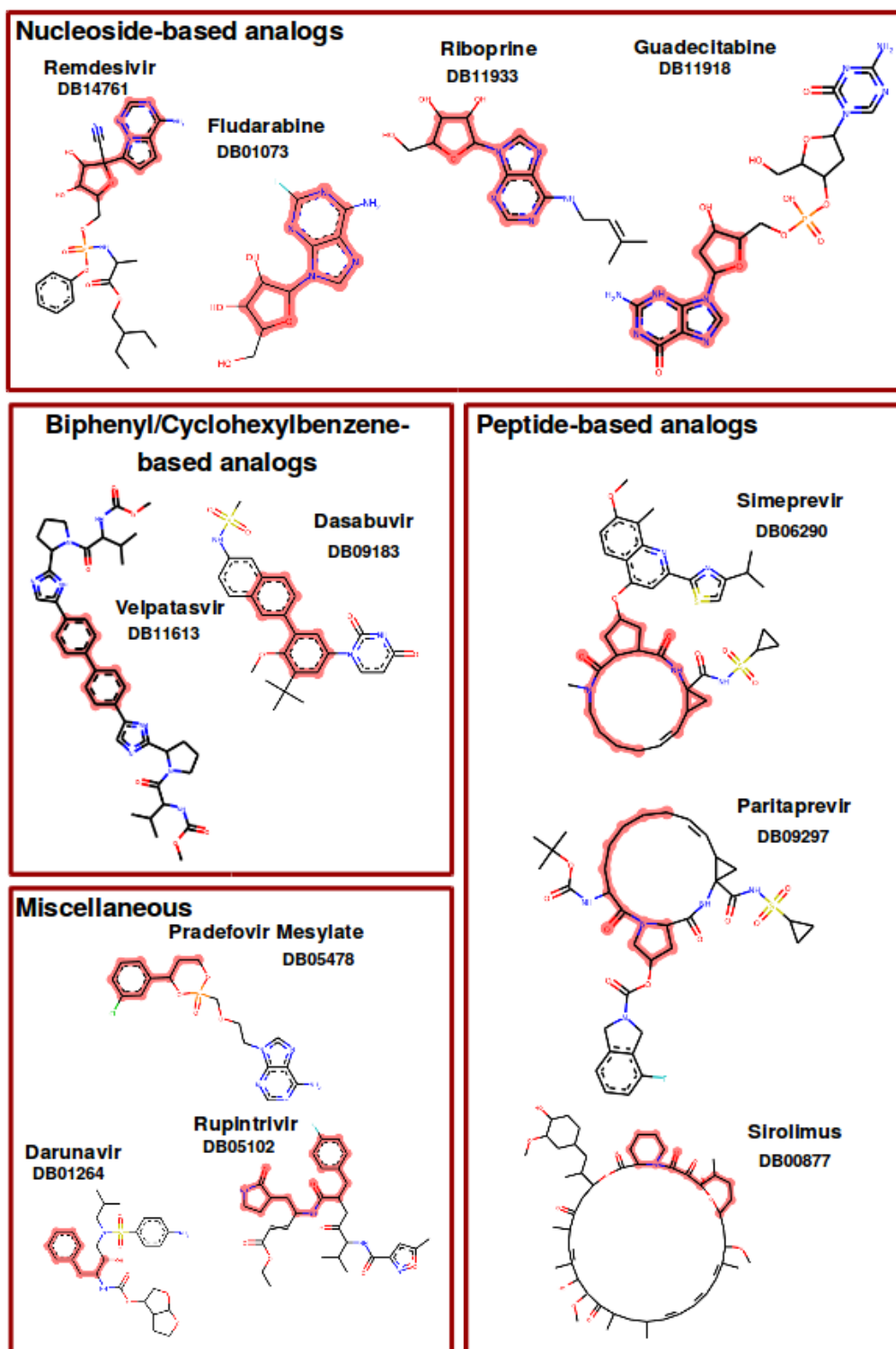


Figure 14: Examples of identified drugs with the highlighted structural query.

## Summary and conclusions

In this educational paper, we are describing a semi-automatic KNIME workflow for *in silico* drug repurposing. The consecutive data mining steps include integration, curation, and analysis of bioassay data from the open domain for specific targets of interest, as well as the generation of structural queries for automated substructure searches in collections of approved, withdrawn, and/or experimental drugs. Targeted access of data through APIs has been implemented at several stages of the KNIME workflow. Incorporation of API calls into KNIME allows repeating the whole procedure in an automated fashion, e.g., when new data is becoming available. As a consequence of the current COVID-19 pandemic, the cheminformatics analyses performed as a use case herein was tailored to ligand and protein data currently available for drug repurposing strategies in the framework of this life-threatening disease. As a side effect of analyzing the data, we are providing insights into enriched chemical substructures for proposed drug targets of SARS-CoV-2. The material has been used successfully for teaching undergraduate students the use of programmatic data access via KNIME workflows and subsequent data analyses steps. The workflows, tutorials, and the information gained on COVID-19 data are freely available to the scientific community for follow-up studies or may be tailored to specific needs of other use cases (available at <https://github.com/AlzbetaTuerkova/Drug-Repurposing-in-KNIME>).

## Authors' contributions

AT and BZ conceptualized and designed the study. AT generated the KNIME workflows, performed the data integration, processing and analyses. BZ provided advice. The manuscript was written through contributions of both authors. Both authors read and approved the final manuscript.

## Funding

No funding was received for the present study.

## Acknowledgements

The authors acknowledge active involvement of undergraduate students participating in the course “Experimental Methods in Drug Discovery and Preclinical Drug Development” at the University of Vienna in testing and applying the developed tutorial and KNIME workflow.

## Competing interests

The authors declare no competing interests.

## Footnotes

Not applicable.

## Additional information

The following additional data are available on GitHub (available at <https://github.com/AlzbetaTuerkova/Drug-Repurposing-in-KNIME>):

- a .csv Supplementary file 1 with maximum common substructures in SMARTS detected via hierarchical scaffold clustering
- a .csv Supplementary file 2 with identified hits from DrugBank
- a .csv Supplementary file 3 with identified hits from CAS Dataset
- a .csv Supplementary file 4 with identified hits by both DrugBank and CAS Dataset
- a .knwf Supplementary file with drug repurposing workflow
- a .pdf Tutorial file “Part 1: Programmatic access to UniProt database using KNIME”
- a .pdf Tutorial file “Part 2: Using cross-references to retrieve structural data from the Protein Data Bank (PDB)”
- a .pdf Tutorial file “Part 3: Integrative data mining of ligand bioactivity data from ChEMBL and PubChem”
- a .pdf Tutorial file “Part 4: Substructure searches in DrugBank”
- a .pdf file with Supplementary Information

## References

- [1. Karaman B, Sippl W. Computational Drug Repurposing: Current Trends. Curr Med Chem. 2019;26\(28\):5389-409.](#)

2. [Bajorath J. Compound Data Mining for Drug Discovery. In: Keith JM, editor. Bioinformatics: Volume II: Structure, Function, and Applications. New York, NY: Springer; 2017. p. 247–56.](#)
3. [Agatonovic-Kustrin S, Morton D. Chapter 9 - Data Mining in Drug Discovery and Design. In: Puri M, Pathak Y, Sutariya VK, Tipparaju S, Moreno W, editors. Artificial Neural Network for Drug Design, Delivery and Disposition. Boston: Academic Press; 2016. p. 181–93.](#)
4. [Gaulton A, Hersey A, Nowotka M, Bento AP, Chambers J, Mendez D, et al. The ChEMBL database in 2017. Nucleic Acids Res. 2017 Jan 4;45\(D1\):D945–54.](#)
5. [Kim S, Thiessen PA, Bolton EE, Chen J, Fu G, Gindulyte A, et al. PubChem Substance and Compound databases. Nucleic Acids Res. 2016 Jan 4;44\(D1\):D1202–13.](#)
6. [Bateman A, Martin MJ, O'Donovan C, Magrane M, Alpi E, Antunes R, et al. UniProt: the universal protein knowledgebase. Nucleic Acids Res. 2017 Jan 4;45\(D1\):D158–69.](#)
7. [Wishart DS, Feunang YD, Guo AC, Lo EJ, Marcu A, Grant JR, et al. DrugBank 5.0: a major update to the DrugBank database for 2018. Nucleic Acids Res. 2018 Jan 4;46\(D1\):D1074–82.](#)
8. [Qian T, Zhu S, Hoshida Y. Use of big data in drug development for precision medicine: an update. Expert Rev Precis Med Drug Dev. 2019;4\(3\):189–200.](#)
9. [Berthold MR, Cebron N, Dill F, Gabriel TR, Kötter T, Meinl T, et al. KNIME - the Konstanz information miner: version 2.0 and beyond. ACM SIGKDD Explor Newsl. 2009 Nov 16;11\(1\):26–31.](#)
10. [Landrum G. RDKit Documentation. :159.](#)
11. [Beisken S, Meinl T, Wiswedel B, de Figueiredo LF, Berthold M, Steinbeck C. KNIME-CDK: Workflow-driven cheminformatics. BMC Bioinformatics. 2013 Aug 22;14\(1\):257.](#)
12. [Pavlov D, Rybalkin M, Karulin B, Kozhevnikov M, Savelyev A, Churinov A. Indigo: universal cheminformatics API. J Cheminformatics. 2011 Apr 19;3\(Suppl 1\):P4.](#)
13. [Roughley S. Five Years of the KNIME Vernalis Cheminformatics Community Contribution. Curr Med Chem. 2018 Sep 3;](#)
14. [Pushpakom S, Iorio F, Eyers PA, Escott KJ, Hopper S, Wells A, et al. Drug repurposing: progress, challenges and recommendations. Nat Rev Drug Discov. 2019 Jan;18\(1\):41–58.](#)
15. [Fetro C, Scherman D. Drug repurposing in rare diseases: Myths and reality. Therapies. 2020 Apr 1;75\(2\):157–60.](#)
16. [Hui DS, I Azhar E, Madani TA, Ntoumi F, Kock R, Dar O, et al. The continuing 2019-nCoV epidemic threat of novel coronaviruses to global health - The latest 2019 novel coronavirus outbreak in Wuhan, China. Int J Infect Dis IJID Off Publ Int Soc Infect Dis. 2020;91:264–6.](#)
17. [The species Severe acute respiratory syndrome-related coronavirus : classifying 2019-nCoV and naming it SARS-CoV-2. Nat Microbiol. 2020 Apr;5\(4\):536–44.](#)
18. [Newer Vaccine Technologies Deployed to Develop COVID-19 Shot. The Scientist Magazine®](#)
19. [Duan K, Liu B, Li C, Zhang H, Yu T, Qu J, et al. Effectiveness of convalescent plasma therapy in severe COVID-19 patients. Proc Natl Acad Sci. 2020 Apr 28;117\(17\):9490–6.](#)
20. [Gordon DE, Jang GM, Bouhaddou M, Xu J, Obernier K, White KM, et al. A SARS-CoV-2 protein interaction map reveals targets for drug repurposing. Nature. 2020 Apr 30;1–13.](#)
21. [Gutmanas A, Alhroub Y, Battle GM, Berrisford JM, Bochet E, Conroy MJ, et al. PDBe: Protein Data Bank in Europe. Nucleic Acids Res. 2014 Jan 1;42\(D1\):D285–91.](#)
22. [Pawson AJ, Sharman JL, Benson HE, Faccenda E, Alexander SPH, Buneman OP, et al. The IUPHAR/BPS Guide to PHARMACOLOGY: an expert-driven knowledgebase of drug targets and their ligands. Nucleic Acids Res. 2014 Jan 1;42\(D1\):D1098–106.](#)
23. [Bemis GW, Murcko MA. The Properties of Known Drugs. 1. Molecular Frameworks. J Med Chem. 1996 Jan 1;39\(15\):2887–93.](#)
24. [Gadaleta D, Lombardo A, Toma C, Benfenati E. A new semi-automated workflow for chemical data retrieval and quality checking for modeling applications. J Cheminformatics. 2018 Dec;10\(1\):1–13.](#)

25. Team RC. R: A language and environment for statistical computing. Vienna, Austria; 2013.
26. Kanz C, Aldebert P, Althorpe N, Baker W, Baldwin A, Bates K, et al. The EMBL Nucleotide Sequence Database. *Nucleic Acids Res.* 2005 Jan 1;33(suppl 1):D29–33.
27. Benson DA, Karsch-Mizrachi I, Lipman DJ, Ostell J, Sayers EW. GenBank. *Nucleic Acids Res.* 2011 Jan;39(Database issue):D32–7.
28. Pujar S, O’Leary NA, Farrell CM, Loveland JE, Mudge JM, Wallin C, et al. Consensus coding sequence (CCDS) database: a standardized set of human and mouse protein-coding regions supported by expert curation. *Nucleic Acids Res.* 2018 Jan 4;46(D1):D221–8.
29. Goodsell DS, Zardecki C, Costanzo LD, Duarte JM, Hudson BP, Persikova I, et al. RCSB Protein Data Bank: Enabling biomedical research and drug discovery. *Protein Sci.* 2020;29(1):52–65.
30. Pieper U, Webb BM, Dong GQ, Schneidman-Duhovny D, Fan H, Kim SJ, et al. ModBase, a database of annotated comparative protein structure models and associated resources. *Nucleic Acids Res.* 2014 Jan 1;42(D1):D336–46.
31. Kiefer F, Arnold K, Künzli M, Bordoli L, Schwede T. The SWISS-MODEL Repository and associated resources. *Nucleic Acids Res.* 2009 Jan 1;37(suppl 1):D387–92.
32. Oughtred R, Chatr-aryamontri A, Breitkreutz B-J, Chang CS, Rust JM, Theesfeld CL, et al. Use of the BioGRID Database for Analysis of Yeast Protein and Genetic Interactions. *Cold Spring Harb Protoc.* 2016 Jan 1;2016(1):pdb.prot088880.
33. Kerrien S, Aranda B, Breuza L, Bridge A, Broackes-Carter F, Chen C, et al. The IntAct molecular interaction database in 2012. *Nucleic Acids Res.* 2012 Jan 1;40(D1):D841–6.
34. Szklarczyk D, Morris JH, Cook H, Kuhn M, Wyder S, Simonovic M, et al. The STRING database in 2017: quality-controlled protein–protein association networks, made broadly accessible. *Nucleic Acids Res.* 2017 Jan 4;45(D1):D362–8.
35. Gilson MK, Liu T, Baitaluk M, Nicola G, Hwang L, Chong J. BindingDB in 2015: A public database for medicinal chemistry, computational chemistry and systems pharmacology. *Nucleic Acids Res.* 2016 Jan 4;44(D1):D1045–53.
36. Zhang L, Lin D, Sun X, Curth U, Drosten C, Sauerhering L, et al. Crystal structure of SARS-CoV-2 main protease provides a basis for design of improved  $\alpha$ -ketoamide inhibitors. *Science.* 2020 Apr 24;368(6489):409–12.
37. Coutard B, Valle C, de Lamballerie X, Canard B, Seidah NG, Decroly E. The spike glycoprotein of the new coronavirus 2019-nCoV contains a furin-like cleavage site absent in CoV of the same clade. *Antiviral Res.* 2020;176:104742.
38. Yin W, Mao C, Luan X, Shen D-D, Shen Q, Su H, et al. Structural basis for inhibition of the RNA-dependent RNA polymerase from SARS-CoV-2 by remdesivir. *Science.* 2020 May 1
39. Sigrist CJ, Bridge A, Le Mercier P. A potential role for integrins in host cell entry by SARS-CoV-2. *Antiviral Res.* 2020;177:104759.
40. Klopmand G. Concepts and applications of molecular similarity, by Mark A. Johnson and Gerald M. Maggiora, eds., John Wiley & Sons, New York, 1990, 393 pp. Price: \$65.00. *J Comput Chem.* 1992 May 1;13(4):539–40.
41. Martin YC, Kofron JL, Traphagen LM. Do structurally similar molecules have similar biological activity? *J Med Chem.* 2002 Sep 12;45(19):4350–8.
42. Cao Y, Jiang T, Girke T. A maximum common substructure-based algorithm for searching and predicting drug-like compounds. *Bioinformatics.* 2008 Jul 1;24(13):i366–74.
43. Alabboud M, Javadmanesh A. In silico study of various antiviral drugs, vitamins, and natural substances as potential binding compounds with SARS-CoV-2 main protease. *DYSONA - Life Sci.* 2020 Jul 1;0:44–63.
44. Dutta K, Shityakov S, Morozova O, Khalifa I, Zhang J, Panda A, et al. Beclabuvir can Inhibit the RNA-dependent RNA Polymerase of Newly Emerged Novel Coronavirus (SARS-CoV-2). 2020 Mar 26

- [45. Lee VS, Chong WL, Sukumaran SD, Nimmanpipug P, Letchumanan V, Goh BH, et al. Computational screening and identifying binding interaction of anti-viral and anti-malarial drugs: Toward the potential cure for SARS-CoV-2. Prog Drug Discov Biomed Sci. 2020 Mar 26;3\(1\)](#)
- [46. Seo S, Park JW, An D, Yoon J, Paik H, Hwang S. Supercomputer-aided Drug Repositioning at Scale: Virtual Screening for SARS-CoV-2 Protease Inhibitor. 2020 Apr 10](#)
- [47. Dashti-Khavidaki S, Khalili H. Considerations for Statin Therapy in Patients with COVID-19. Pharmacother J Hum Pharmacol Drug Ther.](#)
- [48. Hull MW, Montaner JSG. Ritonavir-boosted protease inhibitors in HIV therapy. Ann Med. 2011 Aug;43\(5\):375–88.](#)
- [49. Cao B, Wang Y, Wen D, Liu W, Wang J, Fan G, et al. A Trial of Lopinavir–Ritonavir in Adults Hospitalized with Severe Covid-19. N Engl J Med. 2020 May 7;382\(19\):1787–99.](#)