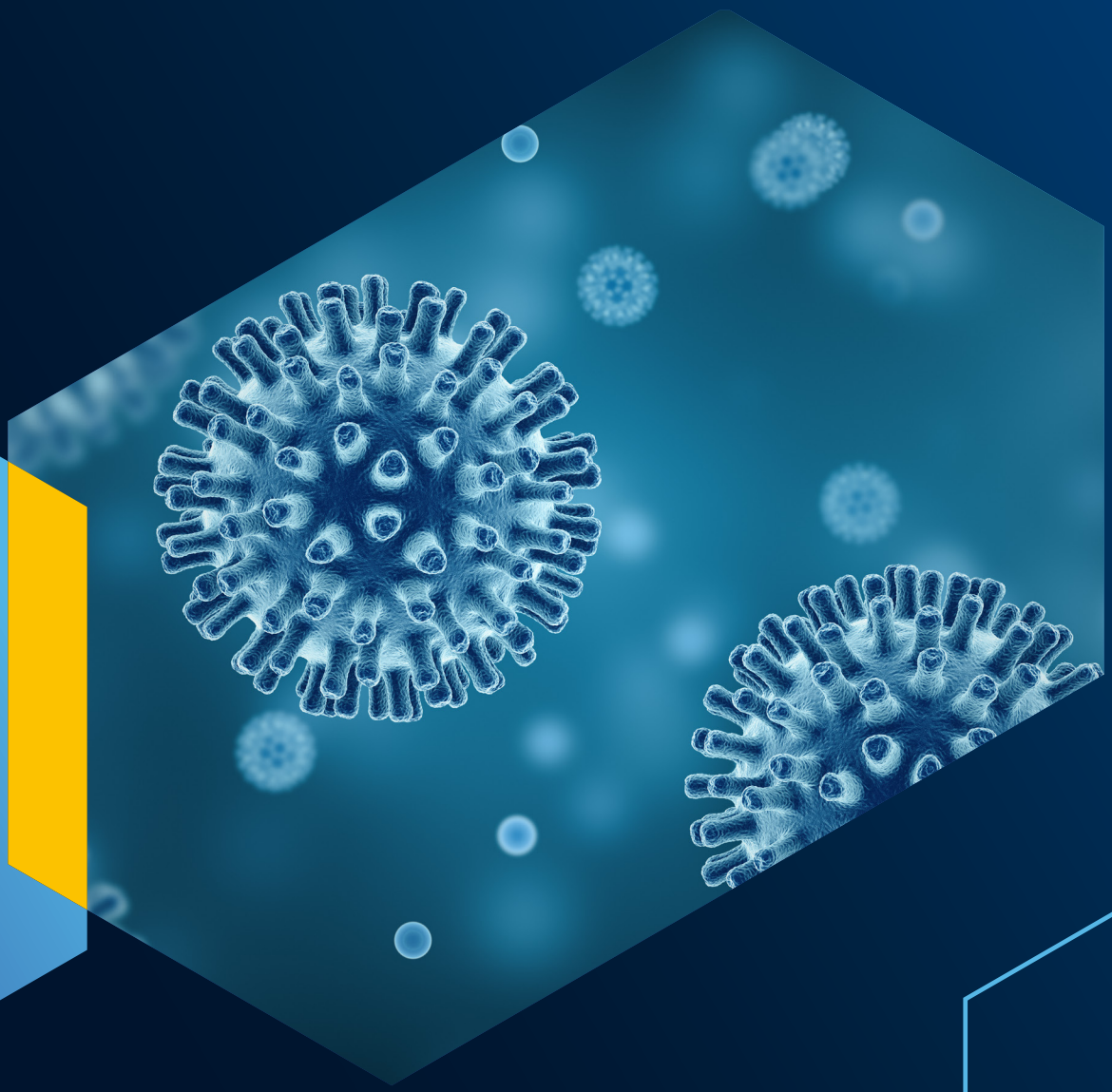
CAS BIOMEDICAL KNOWLEDGE GRAPH

Identifying candidate drugs for repurposing as COVID-19 therapeutics

CAS
A division of the
American Chemical Society

Introduction

When pandemics strike, drug repurposing becomes critical for faster development of therapies. However, assembling all of the critical information and connections around new proteins, viruses, targets, pathways, and clinical information can be challenging. CAS leverages their unique connections across the world's science for novel knowledge graphs that identify top clinical candidates to repurpose for COVID-19 therapies.



Curating and interpreting data: What are knowledge graphs?

Knowledge graphs represent a collection of data, descriptions, and relationships from disparate sources to model a particular area of interest. These graphs integrate massive amounts of data and their connections reveal insights that accelerate scientific progress. Each unit of data can be thought of as a dot (or node) connected to other units by lines (or edges) that represent the relationships between the nodes. This approach places as much importance on the relationships that connect data as on the data itself. While knowledge graphs are not necessarily new, their value has continually grown as scientific publications have grown in volume, complexity, and across disciplines to enable better discovery.

Figure 1 below provides a small, highly simplified example of using a knowledge graph to predict which drugs might inhibit vascular inflammation. Traditional databases may only show direct inhibitors of transcription factor STAT3, but a knowledge graph's ability to show deeper data connections will present with inhibitors that might act further down the pathway, such as Alpelisib.

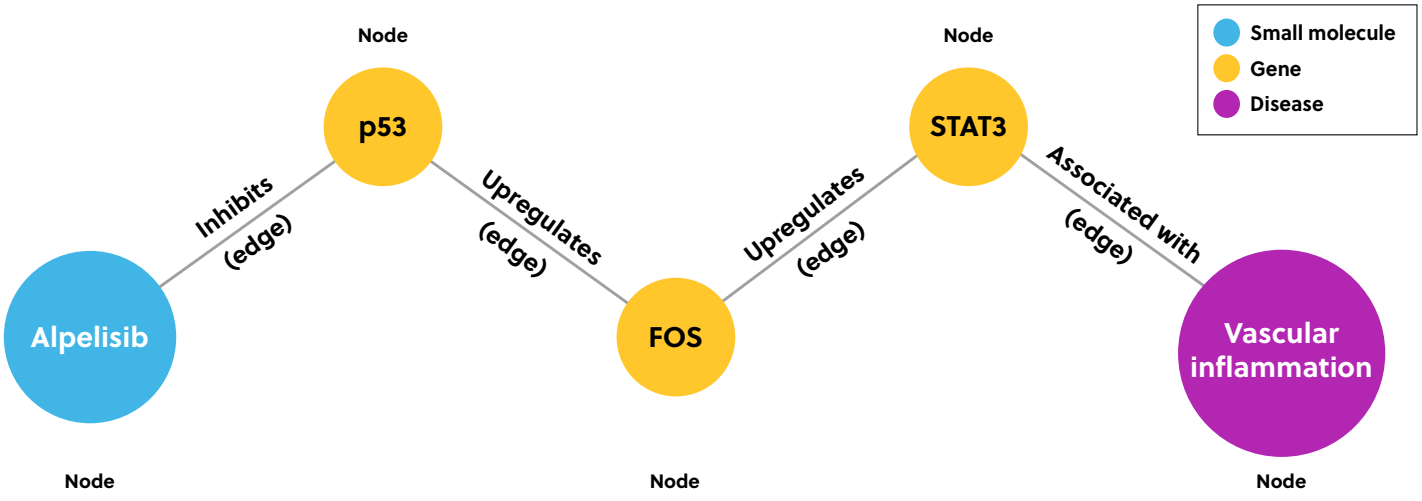


Figure 1: Visual depiction of a simplified knowledge graph relating Alpelisib to vascular inflammation

A common method in developing knowledge graphs is to use machines that ingest and connect data for faster creation. However, there is a trade-off in a fully automated approach, which can limit the connections and interpretation of scientific content. For instance, the name and taxonomy of a drug can change over time, thus scientific expertise and an established tracking structure is required to ensure that connections from disparate information sources can still be made with that particular drug. Moreover, machines alone can easily misunderstand the interpretation of formulas into molecular representation. With a lot of data being ingested by machines and optical character recognition (OCR) algorithms in knowledge graphs, there is also potential for scientific insights to be missed.

The power of a knowledge graph depends on the quality and comprehensiveness of the data sources used to build it. Like most data management systems, they must be maintained and updated periodically. Only an approach that leverages both human expertise and machines can create the most meaningful knowledge graphs with high quality data and connections.

Introducing the CAS Biomedical Knowledge Graph

CAS scientists have constructed a biomedical knowledge graph that combines both human-curated substance data in CAS REGISTRY® and publicly available biomedical data. The CAS Content Collection™ is the world's foremost collection of chemical information and is curated, connected, and analyzed by hundreds of expert scientists and specialists.

A simple visual schema of how the knowledge graph has been constructed is shown below in Figure 2. In this article, we describe how the CAS Biomedical Knowledge Graph was used for identifying small molecules that show potential for repurposing to treat one of the greatest challenges facing us today - COVID-19. This approach combined data from over 6 million small molecules, 24,000 human diseases, and 26,000 human and viral genes.

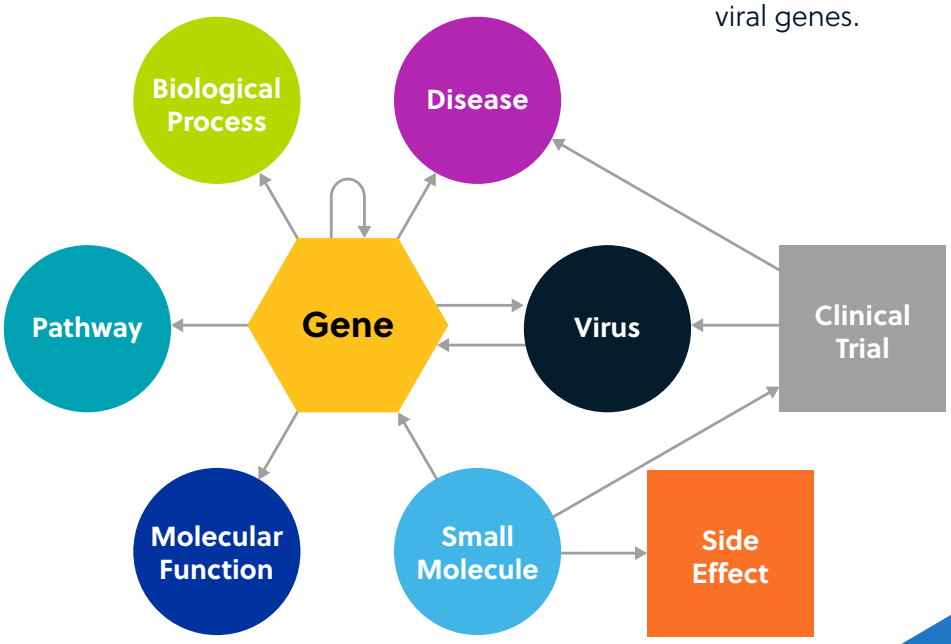


Figure 2: A simple schematic diagram of the CAS Biomedical Knowledge Graph

How CAS Biomedical Knowledge Graph is different:

- **Contains over 6 million nodes and 18 million relationships, combining biological pathways and chemical entity data in this space**
- **Connects external sources and the CAS Content Collection for unique views into the COVID-19 landscape**
- **The CAS Content Collection is curated, connected, and analyzed by scientists and includes data from sources such as CAS REGISTRY and CAS Reactions**

The CAS Biomedical Knowledge Graph accelerates research and drug discovery by:

- **Connecting disparate data sources (internal and external to CAS) for a unique view of today's current COVID-19 therapeutic landscape**
- **Identifying connections between the biological systems involved in COVID-19 disease development and the cellular pathways that underlie these systems**
- **Reducing research time and costs**

How can the CAS Biomedical Knowledge Graph be used to answer real-world challenges?

COVID-19 is one of the biggest threats facing the world today. Healthcare bodies and scientists around the world have been engaging in a global effort to find possible therapeutics. However, only a limited number of treatments have been approved by the FDA or EMA, and as SARS-CoV-2 variants like Delta continue to evolve, the need for new innovative approaches is critical.

Research has already been carried out on SARS-CoV-2 to identify how it affects the body and which biological processes are involved in COVID-19, including the cellular pathways that underlie these systems. This COVID-19- and SARS-CoV-2-specific data has been incorporated into the CAS Biomedical Knowledge Graph to identify small molecules for repurposing as COVID-19 therapeutics. A two-component approach was taken to identify these small molecules, as seen in Figure 3:

1. **CAS scientist-identified COVID-19 biological processes**: The genes and gene products (proteins) connected to these 20 biological processes were examined to specify protein targets for inhibition or activation. The inhibiting or activating relationships of the small molecules in the biomedical knowledge graph were identified and evaluated by CAS scientists.

2. **Literature-derived changes in gene expression**: Genes significantly upregulated (>2-fold) by SARS-CoV-2 infection as described by Blanco-Melo et al. (2020) were used to identify relevant biological processes. These biological processes were ranked based on how many upregulated genes they were connected to, with the top 16 processes being used to search for small molecules.

A combination of the resulting small molecules from both components were then ranked by CAS scientists using a novel algorithmic method as described in the full report to identify the most promising drug candidates, prioritizing drugs that target the proteins involved in COVID-19 disease pathways with minimal projected side effects.

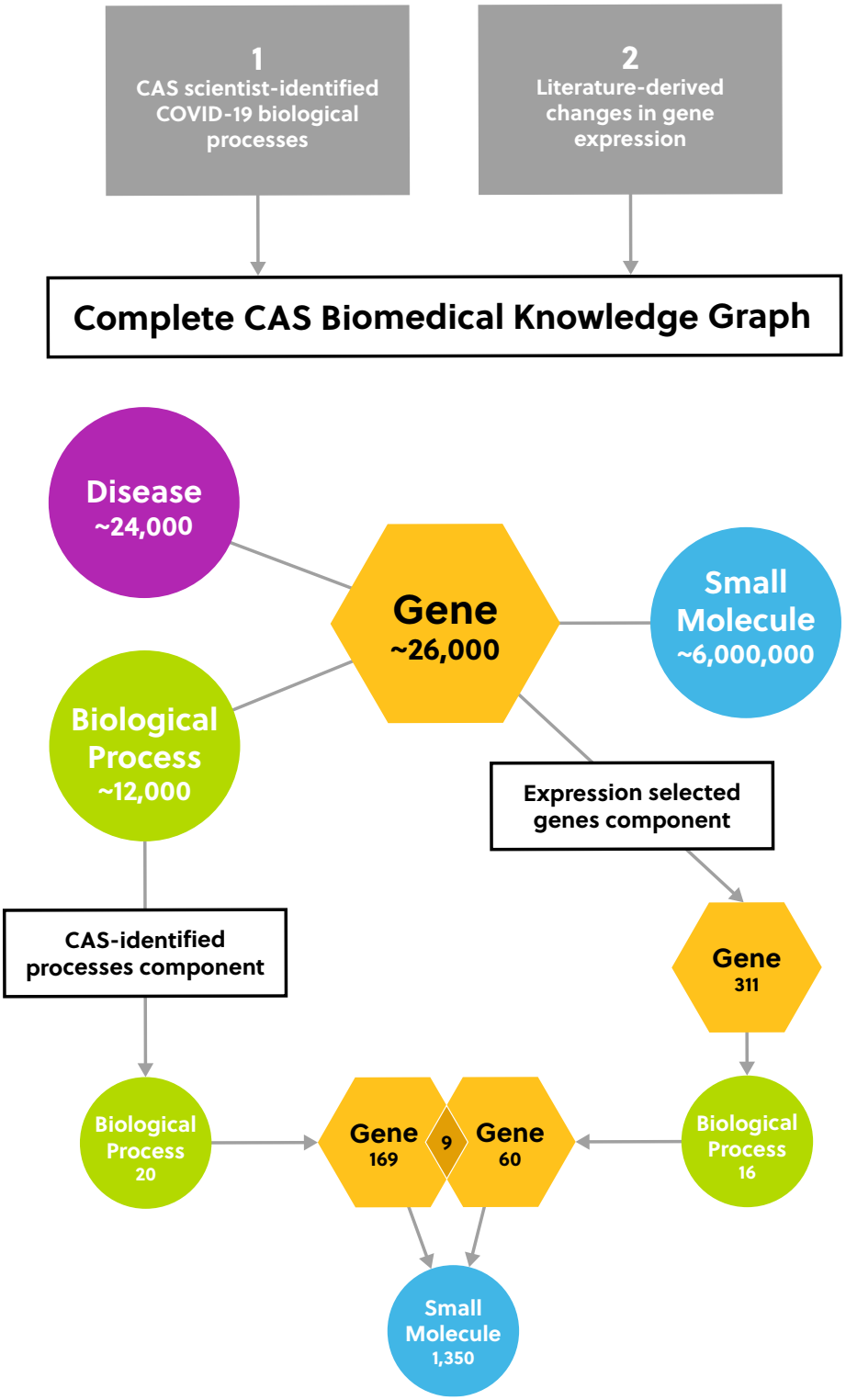


Figure 3: Diagram outlining the two-component approach to identify potential drug candidates for COVID-19 therapeutics

Identifying novel therapeutics for use in COVID-19

Using this approach, hundreds of drug repurposing candidates were identified, and the top 50 are shown in Table 1. A network diagram of the top 10 of these is shown in Figure 4.

Table 1. The top 50 drug repurposing candidates with CAS REGISTRY(R) Number, drug name, and drug class, while drugs that were difficult to classify are listed as Other

Rank	CAS REGISTRY Number	Drug Name	Drug Class
1	**149647-78-9**	**Vorinostat**	**HDAC inhibitors**
2	**179324-69-7**	**Bortezomib**	**Protease inhibitors**
3	**23214-92-8**	**Doxorubicin**	**DNA metabolism-related**
4	**284461-73-0**	**Sorafenib**	**Kinase inhibitors**
5	**183321-74-6**	**Erlotinib**	**Kinase inhibitors**
6	**231277-92-2**	**Lapatinib**	**Kinase inhibitors**
7	**114977-28-5**	**Docetaxel**	**Microtubule-regulating agents**
8	**667463-62-9**	**MLS 2052**	**Kinase inhibitors**
9	**404950-80-7**	**Panobinostat**	**HDAC inhibitors**
10	**152459-95-5**	**Imatinib**	**Kinase inhibitors**
11	56-65-5	Adenosine 5' triphosphate	Other
12	872511-34-7	BGJ 398	Kinase inhibitors
13	2447-54-3	Sanguinarine	Other
14	1339928-25-4	Fimepinostat	Other
15	183506-66-3	Apicidin	HDAC inhibitors
16	58880-19-6	Trichostatin A	HDAC inhibitors
17	943319-70-8	Ponatinib	Kinase inhibitors
18	112953-11-4	7-Hydroxystaurosporine	Kinase inhibitors
19	1256448-47-1	Nanatinostat	HDAC inhibitors
20	287383-59-9	Scriptaid	HDAC inhibitors

Rank	CAS REGISTRY Number	Drug Name	Drug Class
21	1210608-43-7	PIM 447	Kinase inhibitors
22	477600-75-2	Tofacitinib	Kinase inhibitors
23	868540-17-4	Carfilzomib	Protease inhibitors
24	989-51-5	Epigallocatechin gallate	DNA metabolism-related inhibitors
25	23541-50-6	Daunorubicin hydrochloride	DNA metabolism-related inhibitors
26	870262-90-1	Letaxaban	Coagulation factor Xa inhibitors
27	1195765-45-7	Dabrafenib	Kinase inhibitors
28	25316-40-9	Doxorubicin hydrochloride	DNA metabolism-related inhibitors
29	491-80-5	Biochanin	Other
30	405169-16-6	Dovitinib	Kinase inhibitors
31	50-65-7	Niclosamide	Other
32	957054-30-7	Pictilisib	Kinase inhibitors
33	1108743-60-7	Entrectinib	Kinase inhibitors
34	97-77-8	Tetraethylthiuram disulfide	Other
35	75706-12-6	Leflunomide	Other
36	726169-73-9	Mocetinostat	HDAC inhibitors
37	637-03-6	Phenylarsine oxide	Other
38	1951-25-3	Amiodarone	Other
39	630-60-4	Ouabain	Other
40	58-00-4	(-)-Apomorphine	Other
41	64-86-8	Colchicine	Microtubule-regulating agents
42	90-34-6	Primaquine	Other
43	936563-96-1	Ibrutinib	Kinase inhibitors
44	31431-39-7	Mebendazole	Microtubule-regulating agents
45	361442-04-8	Saxagliptin	Protease inhibitors
46	1032900-25-6	Ceritinib	Kinase inhibitors
47	446-72-0	Genistein	Kinase inhibitors
48	20830-81-3	Daunorubicin	DNA metabolism-related
49	480449-70-5	Edoxaban	Coagulation factor Xa inhibitors
50	153436-53-4	Tyrphostin AG 1478	Kinase inhibitors

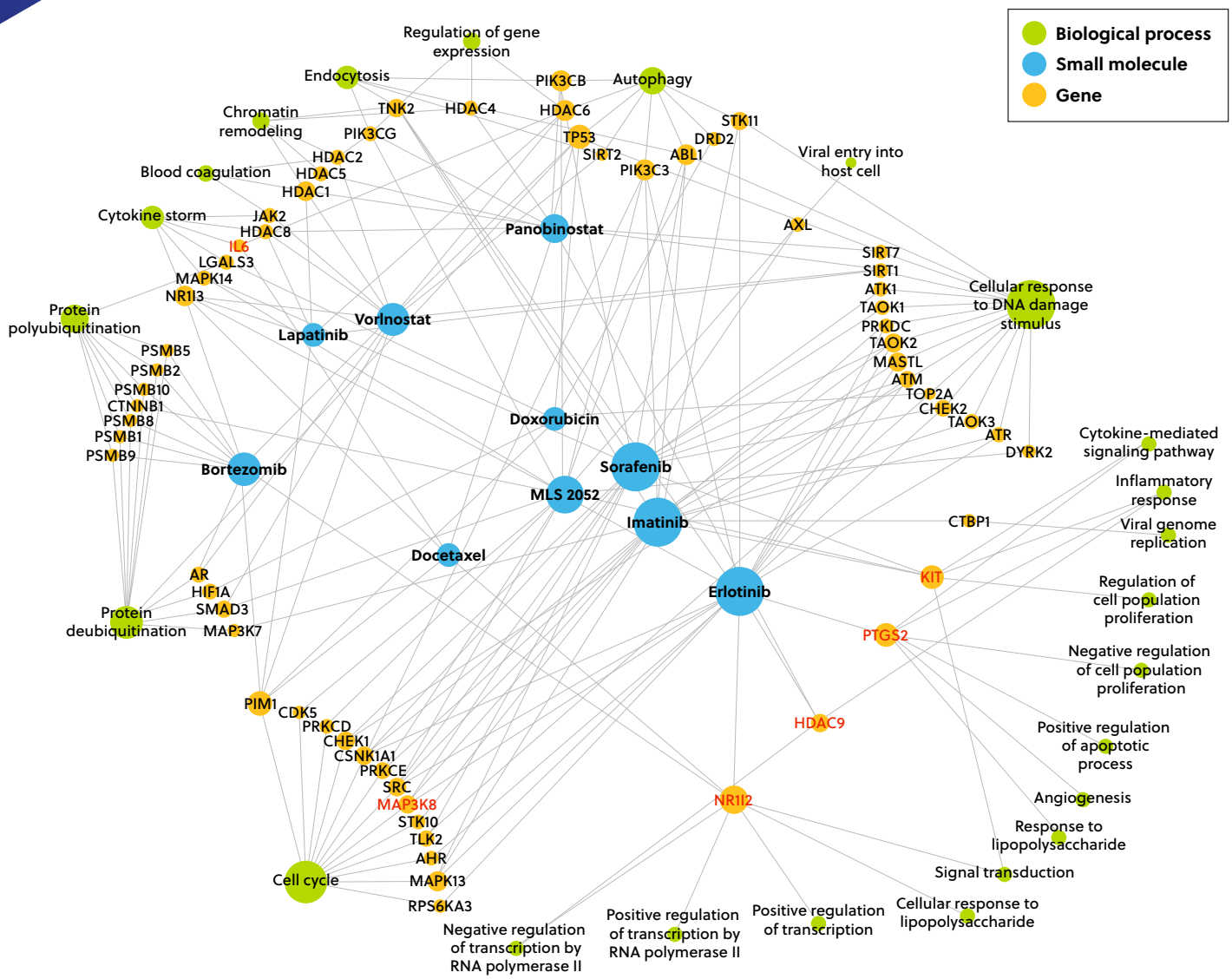


Figure 4: A network showing the connection of the top 10 scoring drugs from the results, with the size of the nodes corresponding to the number of connections to other nodes

The main class of drugs identified was kinase inhibitors. Kinases are involved in almost all biological processes. Their activities are dysregulated in many diseases, which makes them one of the most studied drug classes. They have been shown to be involved in the viral infection process, including in coronavirus infections. The types of kinase inhibitors identified by the CAS Biomedical Knowledge Graph targeted:

1. **Receptor tyrosine kinases** (RTKs) and **non-receptor tyrosine kinases** (nTRKs). RTKs are involved in the cell entry of many viruses. The kinase inhibitors identified included those affecting RTKs such as EGF, FGF, PDGF, and ALK receptors as well as nRTKs such as Bruton tyrosine kinase.

2. **Serine-threonine kinases** (STKs). The kinase inhibitors identified affected STKs such as B-RAF, PKC, PIM, and GSK-2beta.

Four of the tyrosine kinase inhibitors: imatinib, tofacitinib, ibrutinib, and genistein have been or are currently in clinical trials for COVID-19.

Three additional classes of drugs from the top 50 results were identified:

3. **Histone deacetylase inhibitors**. Histone deacetylase inhibitors (HDIs) regulate gene expression and have been shown to downregulate ACE2, the main cellular receptor for SARS-CoV-2, and ABO glycosyltransferase, an enzyme that helps regulate blood type – a known risk factor for COVID-19.

4. **Microtubule-regulating agents**. Microtubules are filaments composed of tubulin subunits, and studies have shown that COVID-19 proteins interact with microtubules. As such, microtubule-regulating agents, such as colchicine, docetaxel, and mebendazole, may offer significant opportunities for drug repurposing.

5. **Protease inhibitors**. Of the protease inhibitors identified, most were proteasome inhibitors. It has been previously shown that the ubiquitin-proteasome system (UPS) is involved in viral replication and the cytokine storm, including in coronavirus-associated diseases, so it seems rational that proteasome inhibitors may be of value in treating COVID-19. Several such inhibitors are already being investigated as COVID-19 therapeutics, and some were found in our results, such as bortezomib, carfilzomib, and saxagliptin.

The processes these small molecules target are well known to play important roles in viral infections, including in coronavirus infections. The validity of the CAS Biomedical Knowledge Graph and its novel ranking method is supported by the fact that 11 of the top 50 results are currently in clinical trials for treating COVID-19 disease.

Accelerating discovery beyond COVID-19

We have shown that the CAS Biomedical Knowledge Graph can be an important tool in addressing real-world challenges, specifically in drug discovery for potential COVID-19 treatments. The wealth of data included in the graph – from the CAS Registry file, human-curated CAS scientific data and external databases of human genes, molecular processes, pathways and diseases - has allowed the identification of 1,350 small molecules as potential drugs for treating COVID-19. By using a novel ranking method, CAS scientists identified the top 50 most promising repurposing candidates.

This method can prove to be of great value for other diseases beyond COVID-19, such as Alzheimer's disease, Parkinson's disease, autoimmune diseases, cancer, and even rare diseases. One of the benefits of using a knowledge graph is that scientists are able to identify novel interactions between two proteins to quickly pinpoint which pathways, biological processes, or diseases these interactions could alter. They can visually navigate through the pathways connecting these data points to see how modules, including non-adjacent ones, can affect each other – allowing a different perspective on a research problem. The CAS Biomedical Knowledge Graph can support innovative research in drug discovery by reducing the risk of overlooking biological targets involved in a disease process. The wider, more comprehensive view provided by the CAS Biomedical Knowledge Graph can lead to cost and timesavings in the initial drug screening process.

Knowledge graphs are both scalable and modular, and this application to COVID-19 is just one example of the broad array of possible uses for CAS-powered knowledge graphs. This method can be applied across various fields of scientific research, including other areas of chemistry and materials science, food science, energy, and environmental research. The opportunities are vast.

For more details on the approach, methodologies, and applications see our recent publication in the *Journal of Chemical Information and Modeling*

CAS is a leader in scientific information solutions, partnering with innovators around the world to accelerate scientific breakthroughs. CAS employs over 1,400 experts who curate, connect, and analyze scientific knowledge to reveal unseen connections. For over 100 years, scientists, patent professionals, and business leaders have relied on CAS solutions and expertise to provide the hindsight, insight, and foresight they need so they can build upon the learnings of the past to discover a better future. CAS is a division of the American Chemical Society.

Connect with us at cas.org

CAS

A division of the
American Chemical Society

CASGENENGWHP100190210803