Common Datasets and Software Suitable for COVID-19 Modeling

Databases are recently used for COVID-19 Modeling

Top-5 commonly used databases for small-molecule, natural compounds, biologics and nucleosides.

- **ZINC:** https://zinc15.docking.org/
    - ZINC, a free database of commercially available compounds for virtual screening. ZINC contains over 230 million purchasable compounds in ready-to-dock, 3D formats. ZINC also contains over 750 million purchasable compounds you can search for analogs in under a minute.
    - Good Database for small-molecule, biologics, natural compounds, marketed drugs and nucleosides
- **PubChem**: https://pubchem.ncbi.nlm.nih.gov/
    - PubChem contains 103 M compounds.
    - Good Database for small-molecule, biologics, natural compounds, marketed drugs and nucleosides
- **SWEETLEAD:** https://simtk.org/projects/sweetlead
    - The SWEETLEAD database has been created to provide an exhaustive and highly curated resource for chemical structures of the world's approved medicines, illegal drugs, and isolates from traditional medicinal herbs
- **ChEMBL**: https://www.ebi.ac.uk/chembl/
    - Contains 2M compounds
    - Good Database for small-molecule, biologics, natural compounds, marketed drugs and nucleosides
- **DrugBank:** https://www.drugbank.ca/
    - contains 13,536 drug entries including 2,630 approved small molecule drugs, 1,372 approved biologics (proteins, peptides, vaccines, and allergenics), 131 nutraceuticals and over 6,358 experimental (discovery-phase) drugs.
    - Good Database for small-molecule, natural compounds, marketed drugs and nucleosides

Databases used for COVID-19 Modeling

DB	DB used/mentioned in the following COVID-19 articles	About the DB
ZINC	Repurposing Therapeutics for COVID-19: Supercomputer-Based Docking to the SARS-CoV-2 Viral Spike Protein and Viral Spike Protein-Human ACE2 Interface **(link)** Anti-HCV, Nucleotide Inhibitors, Repurposing Against COVID-19 (link)The potential chemical structure of anti-SARS-CoV-2 RNA-dependent RNA polymerase **(link)** Rapid Identification of Potential Inhibitors of SARS-CoV-2 Main Protease by Deep Docking of 1.3 Billion Compounds **(link)** Unrevealing Sequence and Structural Features of Novel Coronavirus Using in silico Approaches: The Main Protease as Molecular Target **(link)**	ZINC is a free Opensource database of commercially available compounds for virtual screening. ZINC contains over 230 million purchasable compounds in ready-to-dock, 3D formats. ZINC also contains over 750 million purchasable compounds you can search for analogs in under a minute.
PubChem	In Silico Screening of Chinese Herbal Medicines With the Potential to Directly Inhibit 2019 Novel Coronavirus **(link)**	PubChem is an open chemistry database (103 M compounds) at the National Institutes of Health (NIH). “Open” means that you can put your scientific data in PubChem and that others may use it. Since the launch in 2004, PubChem has become a key chemical information resource for scientists, students, and the general public. Each month our website and programmatic services provide data to several million users worldwide. PubChem mostly contains small molecules, but also larger molecules

		such as nucleotides, carbohydrates, lipids, peptides, and chemically-modified macromolecules. We collect information on chemical structures, identifiers, chemical and physical properties, biological activities, patents, health, safety, toxicity data, and many others.
SWEETLEAD	Repurposing Therapeutics for COVID-19: Supercomputer-Based Docking to the SARS-CoV-2 Viral Spike Protein and Viral Spike Protein-Human ACE2 Interface **(link)**	The SWEETLEAD database has been created to provide an exhaustive and highly curated resource for chemical structures of the world's approved medicines, illegal drugs, and isolates from traditional medicinal herbs. This database has been built using a consensus generating scheme pulling data from several public chemical databases (such as PubChem, ChemSpider, PharmGKB, etc.), as detailed in the publication.
Prestwick Chemical Library http://www.prestwickchemical.com/libraries-screening-lib-pcl.html	In vitro screening of a FDA approved chemical library reveals potential inhibitors of SARS-CoV-2 replication	A unique collection of 1520 off-patent small molecules, 99% approved drugs (FDA, EMA and other agencies)

Here is a detailed summary of additional potential molecular databases:

- **SuperDRUG2:** http://cheminfo.charite.de/superdrug2/
    - SuperDRUG2 database is a unique, one-stop resource for approved/marketed drugs, containing more than 4,600 active pharmaceutical ingredients.
- **DRUGCENTRAL:** http://drugcentral.org/
    - DrugCentral is online drug information resource created and maintained by Division of Translational Informatics at University of New Mexico in collaboration with the IDG. DrugCentral provides information on active ingredients chemical entities, pharmaceutical products, drug mode of action, indications, pharmacologic action. It contains ~250,000 compounds
- **Molport:** https://www.molport.com/shop/screeening-compound-database
    - The MolPort database contains data and prices for over 7 million compounds purchasable from stock and over 20 million made-to-order compounds.

- **SWEETLEAD:** https://simtk.org/projects/sweetlead
    - The SWEETLEAD database has been created to provide an exhaustive and highly curated resource for chemical structures of the world's approved medicines, illegal drugs, and isolates from traditional medicinal herbs
- **DRUGCENTRAL:** http://drugcentral.org/
- **Molport:** https://www.molport.com/shop/screeening-compound-database
- **All FDA approved Drugs**

Software commonly used (based on the number of publications for COVID-19 host protein interactions):

Docking:

- **AutoDock Vina:** http://vina.scripps.edu/
    - flexible ligand-receptor docking
- **Dock** http://dock.compbio.ucsf.edu
    - predict binding modes of small molecule-protein complexes
    - search databases of ligands for compounds that mimic the inhibitory binding interactions of an experimentally validated inhibitor
    - search databases of ligands for compounds that bind a particular site of a specific protein
    - search databases of ligands for compounds that bind nucleic acid targets
    - examine possible binding orientations of protein-protein and protein-DNA complexes
    - help guide synthetic efforts by examining small molecules that are computationally derivatized
- **idock:** https://github.com/HongjianLi/idock
    - idock is a standalone tool for structure-based virtual screening powered by fast and flexible ligand docking.
- **Libdock** (Discovery Studio)

MD simulation and protein modeling software:

- **GROMACS:** http://www.gromacs.org/:
    - GROMACS is a versatile package to perform molecular dynamics, i.e. simulate the Newtonian equations of motion for systems with hundreds to millions of particles. It is

designed for biochemical molecules like proteins, lipids and nucleic acids that have a lot of complicated bonded interactions.

- **AMBER:** https://ambermd.org/:
    - Amber is a suite of biomolecular MD simulation programs.
- **NAMD:** https://www.ks.uiuc.edu/Research/namd/:
    - Parallel molecular dynamics code designed for high-performance simulation of large biomolecular systems.

Modeling Software:

- **SwissModel:** https://swissmodel.expasy.org:
    - SWISS-MODEL is a fully automated protein structure homology-modelling server, accessible via the ExPASy web server, or from the program DeepView (Swiss Pdb-Viewer). The purpose of this server is to make protein modelling accessible to all life science researchers worldwide
- **Phyre2:** http://www.sbg.bio.ic.ac.uk/~phyre2/html/page.cgi?id=index
    - Phyre2 is a suite of tools available on the web to predict and analyze protein structure, function and mutations
- **ROSETTA:** https://www.rosettacommons.org/software
    - Rosetta software suite includes algorithms for computational modeling and analysis of protein structures. It has enabled notable scientific advances in comp. biology, including de novo protein design, enzyme design, ligand docking and structure prediction of biological macromolecules and macromolecular complexes.
- **I-TASSER:** https://zhanglab.ccmb.med.umich.edu/I-TASSER/
    - I-TASSER is a hierarchical approach to protein structure and function prediction. I-TASSER is the top-ranking automatic prediction software in the CASP competition.