Math and AI-based Repositioning of Existing Drugs for COVID-19

By Duc D. Nguyen and Guo-Wei Wei

Coronavirus disease 2019 (COVID-19), an infectious disease caused by severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2), was first reported in Wuhan, China, in December 2019 and has rapidly spread throughout the world. By April 2, 2020, COVID-19 had infected more than one million individuals and was responsible for over 50,000 fatalities. No specific antiviral drug for this pandemic currently exists.

Drug discovery involves target discovery, lead discovery, lead optimization, preclinical development, three phases of clinical trials, and an eventual market launch — only if everything goes well. On average, bringing a new drug to market requires about $2.6 billion dollars and more than 10 years of preparation. It typically takes researchers over a year to develop effective viral vaccines.

Drug repositioning (also known as drug repurposing)—which concerns the investigation of existing drugs for new therapeutic target indications—is one of the most feasible strategies for treating COVID-19 patients. It has emerged as a successful way to accelerate drug discovery due to the reduced costs and expedited approval procedures [7]. Several successful examples unveil its value: Nelfinavir, which was initially developed to treat the human immunodeficiency virus (HIV), is now used for cancer treatments. Researchers first designed amantadine to combat the type A influenza viral infection, but it is presently used to treat Parkinson's disease. And remdesivir, an experimental drug initially developed to inhibit Middle East Respiratory Syndrome, is being repurposed for COVID-19.

In recent years, the rapid growth of drug-related datasets and open data initiatives has inspired new developments for computational drug repositioning, particularly structural-based drug repositioning (SBDR). Machine learning (ML), network analysis, and text mining and semantic inference are three major computational approaches that researchers commonly apply to drug repurposing [7]. Rapid accumulation of genetic and structural databases, development of low-dimensional mathematical representations of complex biomolecular structures [10], and availability of advanced deep learning algorithms have made ML-based drug repositioning one of the most promising approaches for COVID-19 [7].

The first step of SBDR is the selection of one or several effective targets. Studies show that the SARS-CoV-2 genome is very close to that of the early SARS-CoV. The sequence identity percentages of the SARS-CoV-2 main protease, RNA polymerase, and spike protein with those of corresponding SARS-CoV proteins are 96.08, 96, and 76, respectively. The catalytic sites of the SARS-CoV main protease are very conservative and serve as attractive therapeutic targets. Therefore, a potent inhibitor of this protease is likely a potent inhibitor of the SARS CoV-2 main protease. Unfortunately, there is also currently no effective SARS-CoV therapy.

Nevertheless, the SARS-CoV main protease is relatively well-studied. Roughly 119 three-dimensional (3D) X-ray crystal structures of the SARS-CoV or SARS-CoV-2 main protease and its ligand complexes are in the Protein Data Bank (PDB), and the binding affinities of more than 277 potential SARS main protease inhibitors are available in the ChEMBL database. Additionally, there are 17,679 protein-ligand complexes—with binding affinities and 3D X-ray crystal structures—in the PDBbind 2019 general set. Moreover, the DrugBank[1] contains about 1,600 drugs approved by the U.S. Food and Drug Administration (FDA) and nearly 6,000 experimental drugs. The aforementioned data provides a sound foundation for an SBDR machine learning model for SARS-CoV-2 main protease inhibition.

After selecting an appropriate target for COVID-19 drug repositioning, the next challenge involves accurately screening existing drugs from the DrugBank. Over the last several decades, researchers have developed a wide variety of methods for virtual screening. It turns out that mathematics-based artificial intelligence (AI) approaches were top winners in recent years in the D3R Grand Challenge,[2] a worldwide competition series in computer-aided drug design that is funded by the National Institutes of Health [11, 13]. Essentially, although deep learning algorithms—such as convolutional neural networks (CNNs)—can automatically extract features from simple data (e.g., images), they do not work well for data with intricate internal structures. In the case of macromolecules with intrinsically complex structures and high ML dimensionalities, AI approaches must invoke descriptors or representations to simplify their structural complexity and reduce their dimensionality.

Mathematics is a natural choice for data presentation. For example, topology [6]—especially persistent homology [1, 3]—offers the so-called topological simplification that represents complex protein-drug interactions in terms of low-dimensional topological invariants or Betti numbers. Such invariants can be translational, rotational, and scale invariant, which is a requirement of ML [10]. Differential geometry, particularly differentiable manifold, provides a sophisticated abstraction of high-dimensional data [10]. The interplay among differential geometry, differential topology, and algebraic topology yields a variety of geometric, spectral, and topological representations [2].

[1] https://www.drugbank.ca/

[2] https://drugdesigndata.org/about/grand-challenge

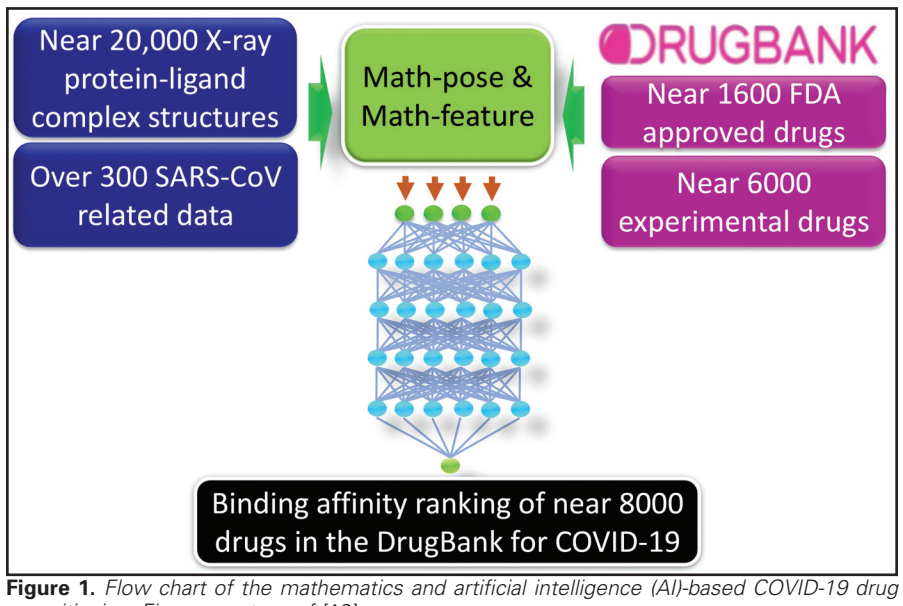

Figure 1. *Flow chart of the mathematics and artificial intelligence (AI)-based COVID-19 drug repositioning. Figure courtesy of [12].*

Discrete mathematics—such as geometric graph theory, algebraic graph theory, topological graph theory, and combinatorics—is a prominent apparatus for data representation [4, 5, 10]. The integration of multiscale analysis, spectral analysis, and topological data analysis promises some of the most powerful data representations [8, 14]. Figure 1 depicts the use of mathematical representations to (i) construct *math-poses* that recreate 3D structures of protein-ligand complexes and (ii) extract *math-features* that contain critical chemical and biological information [13]. We pair math-poses and math-features with CNN, generative network complex, and reinforcement learning algorithms for protein-ligand pose selection, binding affinity prediction, ranking, scoring, and screening [9, 13].

One can validate top-ranking existing drugs inferred from virtual screening with *in vitro* cell culture tests. The toxicities of FDA-approved drugs are known, which means that researchers can then bypass many steps in conventional drug discovery. Controlled clinical trials can further test the confirmed effective drugs to study their antiviral efficacy, dose, and frequency.

References

[1] Carlsson, G. Zomorodian, A., Collins, A., & Guibas, L.J. (2005). Persistence barcodes for shapes. *Int. J. Shape Model., 11*(02), 149-187.

[2] Chen, J., Zhao, R., Tong, Y., & Wei, G.-W. (2019). Evolutionary de Rham-Hodge method. Preprint, *arXiv:1912.12388*.

[3] Edelsbrunner, H., Letscher, D., & Zomorodian, A. (2000). Topological persistence and simplification. In *Proceedings 41st annual symposium on foundations of computer science* (pp. 454-463). Redondo Beach, CA: IEEE Computer Society.

[4] Heitsch, C., & Poznanović, S. (2014). Combinatorial insights into RNA secondary structure. In *Discrete and topological models in molecular biology* (pp. 145-166). New York, NY: Springer.

[5] Jonoska, N., & Twarock, R. (2008). Blueprints for dodecahedral DNA cages. *J. Phys. A: Math. Theor., 41*(30), 304043.

[6] Kaczynski, T., Mischaikow, K., & Mrozek, M. (2006). *Computational homology*. In *Applied mathematical sciences* (Vol. 157). New York, NY: Springer Science & Business Media.

[7] Li, J., Zheng, S., Chen, B., Butte, A.J., Swamidass, S.J., & Lu, Z. (2016). A survey of current trends in computational drug repositioning. *Brief. Bioinform., 17*(1), 2-12.

[8] Meng, Z., & Xia, K. (2020). Persistent spectral based machine learning (PerSpect ML) for drug design. Preprint, *arXiv:2002.00582*.

[9] Nguyen, D.D., & Wei, G.-W. (2019). AGL-Score: Algebraic graph learning score for protein–ligand binding scoring, ranking, docking, and screening. *J. Chem. Info. Model., 59*(7), 3291-3304.

[10] Nguyen, D.D., Cang, Z., & Wei, G.-W. (2020). A review of mathematical representations of biomolecular data. *Phys. Chem. Chemic. Phys., 22*(8), 4343-4367.

[11] Nguyen, D.D., Cang, Z., Wu, K., Wang, M., Cao, Y., & Wei, G.-W. (2019). Mathematical deep learning for pose and binding affinity prediction and ranking in D3R Grand Challenges. *J. Comp.-Aided Mol. Des., 33*(1), 71-82.

[12] Nguyen, D.D., Gao, K., Chen, J., Wang, R., & Wei, G.-W. (2020). Potentially highly potent drugs for 2019-nCoV. Preprint, *bioRxiv*.

[13] Nguyen, D.D., Gao, K., Wang, M., & Wei, G.-W. (2019). MathDL: Mathematical deep learning for D3R Grand Challenge 4. *J. Comp.-Aided Mol. Des., 34*, 131-147.

[14] Wang, R., Nguyen, D.D., & Wei, G.-W. (2019). Persistent spectral graph. Preprint, *arXiv:1912.04135*.

Duc D. Nguyen is an assistant professor of mathematics at Michigan State University. Guo-Wei Wei is a professor of mathematics at Michigan State University.