Can AI transform the way we discover new drugs?

November 16 2022, by Ekaterina Pesheva

Credit: Unsplash/CC0 Public Domain

The world's aging population, the growing burden of chronic and infectious diseases, and the emergence of novel pathogens have made the need for new treatments more urgent than ever. Yet, discovering a

new drug and bringing it to market is a long, arduous, and expensive journey marked by many failures and few successes.

Artificial intelligence has been long deemed the answer to overcoming some of these hurdles due to its ability to analyze vast reams of data, uncover patterns and relationships, and predict effects.

But despite its enormous potential, AI has yet to deliver on the promise of transforming drug discovery.

Now a multi-institutional team led by Harvard Medical School biomedical informatician Marinka Zitnik has launched a platform that aims to optimize AI-driven drug discovery by developing more realistic data sets and higher-fidelity algorithms.

The Therapeutics Data Commons, described in a recent commentary in *Nature Chemical Biology*, is an open-access platform that serves as a bridge between computer scientists and machine-learning researchers on one end and biomedical researchers, biochemists, clinical researchers, and drug designers on the other end—communities that traditionally have worked in isolation from one another.

The platform offers both data set curation and algorithm design and performance evaluation for multiple treatment modalities—including small-molecule drugs, antibodies, and cell and gene therapies—at all stages of drug development, from chemical compound identification to clinical trial drug performance.

Zitnik, an assistant professor of biomedical informatics in the Blavatnik Institute at HMS, conceptualized the platform and now leads the work in collaboration with researchers at MIT, Stanford University, Carnegie Mellon University, Georgia Tech, University of Illinois-Urbana Champaign, and Cornell University.

She recently discussed the Therapeutics Data Commons with Harvard Medicine News.

HMNews: What are the central challenges in drug discovery and how can AI help solve these?

Zitnik: Developing a drug from scratch that is both safe and effective is incredibly challenging. On average, it takes anywhere between 11 and 16 years and between $1 billion to $2 billion to do so. Why is that?

It's very difficult to figure out early on whether an initially promising chemical compound would produce results in human patients consistent with the results it shows in the laboratory. The number of small molecule compounds is 10 to the power of 60—yet only a tiny fraction of this astronomically large chemical space has been canvassed for molecules with medicinal properties. Despite that, the impact of existing therapies on treating disease has been astounding. We believe that novel algorithms coupled with automation and new data sets can find many more molecules that can be translated into improving human health.

AI algorithms can help us determine which among these molecules are most likely to be safe and effective human therapies. That's the ultimate problem that drug discovery development is suffering from. Our vision is that machine learning models can help sift through and integrate vast amounts of biochemical data that we can more directly connect with molecular and genetic information, and ultimately to individualized patient outcomes.

HMNews: How close is AI to making this promise a reality?

Zitnik: We are not there yet. There are a number of challenges, but I'd

say the biggest one is understanding how well our current algorithms work and whether their performance translates to real-world problems.

When we evaluate new AI models through computer modeling, we are testing them on benchmark data sets. Increasingly, we see in publications that those models are achieving near-perfect accuracy. If that's the case, why aren't we seeing widespread adoption of machine learning in drug discovery?

This is because there is a big gap between performing well on a benchmark data set and being ready to transition to real-world implementation in a biomedical or clinical setting. The data on which these models are trained and tested are not indicative of the kind of challenges these models are exposed to when they're used in real practice, so closing this gap is really important.

HMNews: Where does the Therapeutics Data Commons platform come into this?

Zitnik: The goal of the Therapeutics Data Commons is to address precisely such challenges. It serves as a meeting point between the machine-learning community at one end and the biomedical community on the other end. It can help the machine-learning community with algorithmic innovation and make these models more translatable into real-world scenarios.

HMNews: Could you explain how it actually works?

Zitnik: First of all, keep in mind that the process of drug discovery spans the gamut from initial drug design based on data from chemistry and chemical biology, through preclinical research based on data from animal studies, and all the way to clinical research in human patients.

The machine-learning models that we train and evaluate as part of the platform use different kinds of data to support the development process at all these different stages.

For example, the machine-learning models that support the design of small-molecule drugs typically rely on large data sets of molecular graphs—structures of chemical compounds and their molecular properties. These models find patterns in the known chemical space that relate parts of the chemical structure with chemical properties necessary for drug safety and efficacy.

Once an AI model is trained to identify these tell-tale patterns in the known subset of chemicals, it can be deployed and can look for the same patterns in the vast data sets of yet-untested chemicals and make predictions about how these chemicals would perform.

To design models that can help with late-stage drug discovery, we train them on data from animal studies. These models are trained to look for patterns that relate biological data to likely clinical outcomes in humans.

We can also ask whether a model can look for molecular signatures in chemical compounds that correlate with patient information to identify which subset of patients is most likely to respond to a chemical compound.

HMNews: Who are the contributors and end users of this platform?

Zitnik: We have a team of students, scientists, and expert volunteers who come from partner universities and from industry, including small start-ups in the Boston area as well as some large pharmaceutical companies in the United States and Europe. Computer scientists and biomedical

[researchers](#) contribute their expertise in the form of state-of-the-art machine-learning models and pre-processed and curated data sets, which are standardized in a way that can be released and ready for use by others.

So, the platform contains both data sets ready for analysis and machine-learning algorithms, along with robust measures that tell us how well a machine-learning model performs on a specific data set.

Our end users are researchers from around the world. We organize webinars to present any new features, to receive feedback, and to answer questions. We offer tutorials. This ongoing training and feedback is really crucial.

We have 4,000 to 5,000 active users every month, most of them from the U.S., Europe, and Asia. Overall, we have seen over 65,000 downloads of our machine-learning algorithm/data set package. We have seen over 160,000 downloads of harmonized, standardized data sets. The numbers are growing, and we hope they will continue to grow.

HMNews: What are the long-range goals for the Therapeutics Data Commons?

Zitnik: Our mission is to support AI drug discovery on two fronts. First, in the design and testing of machine-learning methods across all stages of [drug discovery](#) and development, from chemical compound identification and drug design to clinical research.

Second, to support the design and validation of machine-learning algorithms across multiple therapeutic modalities, especially the newer ones, including biologic products, vaccines, antibodies, mRNA medicines, protein therapies, and gene therapies.

There is tremendous opportunity for machine learning to contribute to those novel therapies, and we have not yet seen the use of AI in those areas to the extent we have seen in small-molecule research, where much of the focus is today. This gap is mostly due to a dearth of standardized AI-ready data sets for those novel therapeutic modalities, which we hope to address with the Therapeutics Data Commons.

HMNews: What ignited your interest in this work?

Zitnik: I have always been interested in understanding and modeling interactions across complex systems, which are systems with multiple components that interact with one another in a nondependent manner. As it turns out, many problems in therapeutic science are, by definition, precisely such complex systems.

We have a protein target that is a complex three-dimensional structure, we have a small-molecule compound that is a complex graph of atoms and bonds between those atoms, and then we have a patient, whose description and health status are given in the form of a multiscale representation. This is a classic complex-system problem, and I really love to look and find ways to standardize and "tame" those complex interactions.

Therapeutic science is full of those kinds of problems that are ripe to benefit from machine learning. That's what we're chasing, that's what we're after.