Machine learning uncovers potential new TB drugs

15 October 2020, by Anne Trafton

Using a machine-learning approach that incorporates uncertainty, MIT researchers identified several promising compounds that target a protein required for the survival of the bacteria that cause tuberculosis. Credit: MIT News

Machine learning is a computational tool used by many biologists to analyze huge amounts of data, helping them to identify potential new drugs. MIT researchers have now incorporated a new feature into these types of machine-learning algorithms, improving their prediction-making ability.

Using this new approach, which allows computer models to account for uncertainty in the data they're analyzing, the MIT team identified several promising compounds that target a protein required by the bacteria that cause tuberculosis.

This method, which has previously been used by computer scientists but has not taken off in biology, could also prove useful in protein design and many other fields of biology, says Bonnie Berger, the Simons Professor of Mathematics and head of the Computation and Biology group in MIT's Computer Science and Artificial Intelligence Laboratory (CSAIL).

"This technique is part of a known subfield of machine learning, but people have not brought it to biology," Berger says. "This is a paradigm shift, and is absolutely how biological exploration should be done."

Berger and Bryan Bryson, an assistant professor of biological engineering at MIT and a member of the Ragon Institute of MGH, MIT, and Harvard, are the senior authors of the study, which appears today in *Cell Systems*. MIT graduate student Brian Hie is the paper's lead author.

Better predictions

Machine learning is a type of computer modeling in which an algorithm learns to make predictions based on data that it has already seen. In recent years, biologists have begun using machine learning to scour huge databases of potential drug compounds to find molecules that interact with particular targets.

One limitation of this method is that while the algorithms perform well when the data they're analyzing are similar to the data they were trained on, they're not very good at evaluating molecules that are very different from the ones they have already seen.

To overcome that, the researchers used a technique called Gaussian process to assign uncertainty values to the data that the algorithms are trained on. That way, when the models are analyzing the training data, they also take into account how reliable those predictions are.

For example, if the data going into the model predict how strongly a particular molecule binds to a target protein, as well as the uncertainty of those predictions, the model can use that information to make predictions for protein-target interactions that it hasn't seen before. The model also estimates the certainty of its own predictions. When analyzing

new data, the model's predictions may have lower certainty for molecules that are very different from the training data. Researchers can use that information to help them decide which molecules to test experimentally.

Another advantage of this approach is that the algorithm requires only a small amount of training data. In this study, the MIT team trained the model with a dataset of 72 small molecules and their interactions with more than 400 proteins called protein kinases. They were then able to use this algorithm to analyze nearly 11,000 small molecules, which they took from the ZINC database, a publicly available repository that contains millions of chemical compounds. Many of these molecules were very different from those in the training data.

Using this approach, the researchers were able to identify molecules with very strong predicted binding affinities for the protein kinases they put into the model. These included three human kinases, as well as one kinase found in Mycobacterium tuberculosis. That kinase, PknB, is critical for the bacteria to survive, but is not targeted by any frontline TB antibiotics.

The researchers then experimentally tested some of their top hits to see how well they actually bind to their targets, and found that the model's predictions were very accurate. Among the molecules that the model assigned the highest certainty, about 90 percent proved to be true hits—much higher than the 30 to 40 percent hit rate of existing machine learning models used for drug screens.

The researchers also used the same training data to train a traditional machine-learning algorithm, which does not incorporate uncertainty, and then had it analyze the same 11,000 molecule library. "Without uncertainty, the model just gets horribly confused and it proposes very weird chemical structures as interacting with the kinases," Hie says.

The researchers then took some of their most promising PknB inhibitors and tested them against Mycobacterium tuberculosis grown in bacterial culture media, and found that they inhibited bacterial growth. The inhibitors also worked in human immune cells infected with the bacterium.

A good starting point

Another important element of this approach is that once the researchers get additional experimental data, they can add it to the model and retrain it, further improving the predictions. Even a small amount of data can help the model get better, the researchers say.

"You don't really need very large data sets on each iteration," Hie says. "You can just retrain the model with maybe 10 new examples, which is something that a biologist can easily generate."

This study is the first in many years to propose new molecules that can target PknB, and should give drug developers a good starting point to try to develop drugs that target the kinase, Bryson says. "We've now provided them with some new leads beyond what has been already published," he says.

The researchers also showed that they could use this same type of machine learning to boost the fluorescent output of a green fluorescent protein, which is commonly used to label molecules inside living cells. It could also be applied to many other types of biological studies, says Berger, who is now using it to analyze mutations that drive tumor development.

More information: Brian Hie et al. Leveraging Uncertainty in Machine Learning Accelerates Biological Discovery and Design. *Cell Systems*. Published:October 15, 2020. DOI: 10.1016/j.cels.2020.09.007