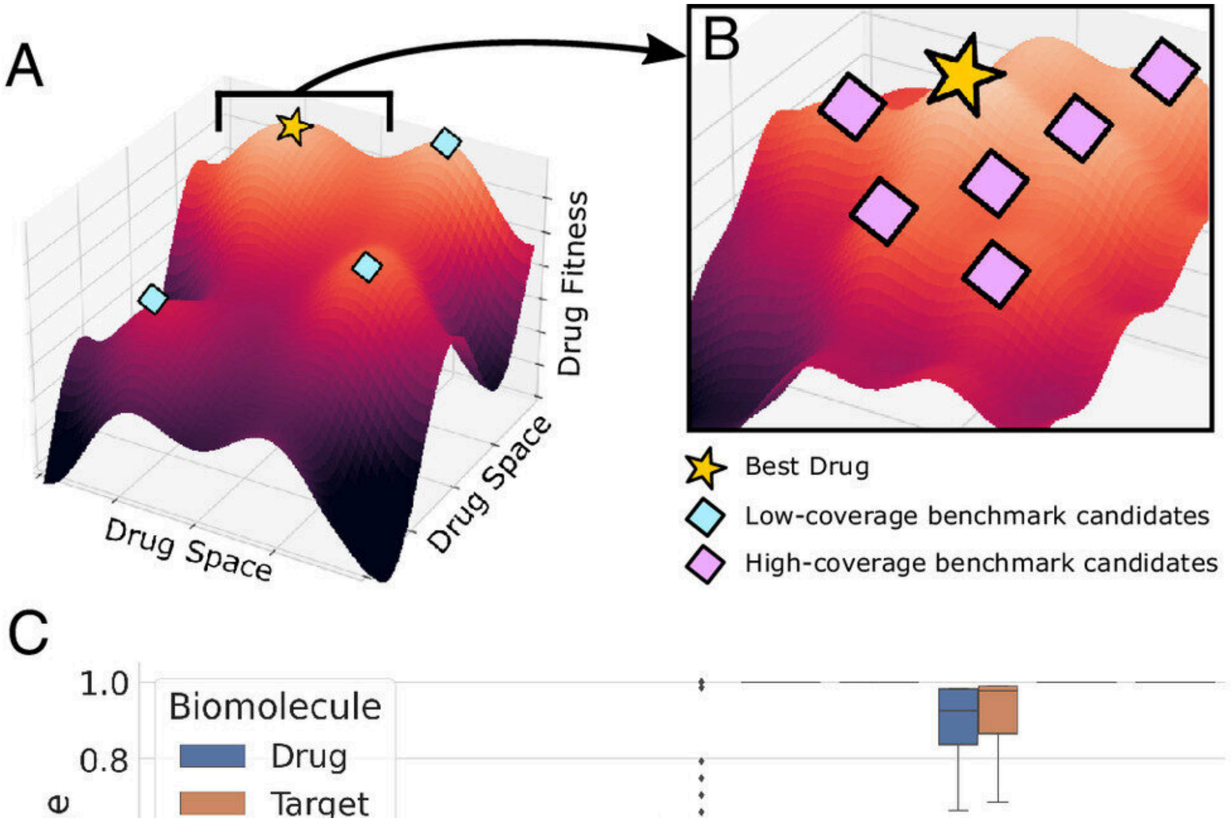
New model offers a way to speed up drug discovery

June 8 2023, by Anne Trafton



Drug-target interaction benchmarks display highly variable levels of coverage. Coverage is defined as the proportion of drugs or targets for which a data point (positive or negative) exists in that dataset. High- vs. low-coverage benchmarks tend to reward different types of model performance. (A) In this cartoon of an example low coverage dataset, drug candidates cover the full diversity of the space, and no two drugs are highly similar. A successful model can learn a coarse estimate of the fitness landscape, but must accurately model a large part of drug space to generalize to all candidates. (B) For high-coverage datasets, drugs tend to be targeted to a specific protein family. Thus, a successful model does not need to generalize nearly as widely but must be able to capture more minor variations in drug fitness to achieve high specificity and differentiate between similar drugs. (C) In a review of existing popular DTI benchmark datasets, we find widely varying coverage, from datasets with nearly zero coverage (each drug/target is represented only a few times) to nearly full coverage (all drug-by-target pairs are known in the data). Credit: *Proceedings of the National Academy of Sciences* (2023). DOI: 10.1073/pnas.2220778120

Huge libraries of drug compounds may hold potential treatments for a variety of diseases, such as cancer or heart disease. Ideally, scientists would like to experimentally test each of these compounds against all possible targets, but doing that kind of screen is prohibitively time-consuming.

In recent years, researchers have begun using computational methods to screen those libraries in hopes of speeding up drug discovery. However, many of those methods also take a long time, as most of them calculate each target protein's three-dimensional structure from its amino-acid sequence, then use those structures to predict which drug molecules it will interact with.

Researchers at MIT and Tufts University have now devised an alternative computational approach based on a type of artificial intelligence algorithm known as a large language model. These models—one well-known example is ChatGPT—can analyze huge amounts of text and figure out which words (or, in this case, amino acids) are most likely to appear together. The new model, known as ConPLex, can match target proteins with potential drug molecules without having to perform the computationally intensive step of calculating the molecules' structures.

Using this method, the researchers can screen more than 100 million compounds in a single day—much more than any existing model.

"This work addresses the need for efficient and accurate in silico screening of potential drug candidates, and the scalability of the model enables large-scale screens for assessing off-target effects, drug repurposing, and determining the impact of mutations on drug binding," says Bonnie Berger, the Simons Professor of Mathematics, head of the Computation and Biology group in MIT's Computer Science and Artificial Intelligence Laboratory (CSAIL), and one of the senior authors of the new study.

Lenore Cowen, a professor of computer science at Tufts University, is also a senior author of the paper, which appears this week in the *Proceedings of the National Academy of Sciences*. Rohit Singh, a CSAIL research scientist, and Samuel Sledzieski, an MIT graduate student, are the lead authors of the paper, and Bryan Bryson, an associate professor of biological engineering at MIT and a member of the Ragon Institute of MGH, MIT, and Harvard, is also an author. In addition to the paper, the researchers have made their model available online for other scientists to use.

Making predictions

In recent years, computational scientists have made great advances in developing models that can predict the structures of proteins based on their amino-acid sequences. However, using these models to predict how a large library of potential drugs might interact with a cancerous protein, for example, has proven challenging, mainly because calculating the three-dimensional structures of the proteins requires a great deal of time and computing power.

An additional obstacle is that these kinds of models don't have a good track record for eliminating compounds known as decoys, which are very similar to a successful drug but don't actually interact well with the target.

"One of the longstanding challenges in the field has been that these methods are fragile, in the sense that if I gave the model a drug or a small molecule that looked almost like the true thing, but it was slightly different in some subtle way, the model might still predict that they will interact, even though it should not," Singh says.

Researchers have designed models that can overcome this kind of fragility, but they are usually tailored to just one class of drug molecules, and they aren't well-suited to large-scale screens because the computations take too long.

The MIT team decided to take an alternative approach, based on a protein model they first developed in 2019. Working with a database of more than 20,000 proteins, the language model encodes this information into meaningful numerical representations of each amino-acid sequence that capture associations between sequence and structure.

"With these language models, even proteins that have very different sequences but potentially have similar structures or similar functions can be represented in a similar way in this language space, and we're able to take advantage of that to make our predictions," Sledzieski says.

In their new study, the researchers applied the protein model to the task of figuring out which protein sequences will interact with specific drug molecules, both of which have numerical representations that are transformed into a common, shared space by a neural network. They trained the network on known protein-drug interactions, which allowed it to learn to associate specific features of the proteins with drug-binding ability, without having to calculate the 3D structure of any of the molecules.

"With this high-quality numerical representation, the model can short-circuit the atomic representation entirely, and from these numbers predict whether or not this drug will bind," Singh says. "The advantage of this is that you avoid the need to go through an atomic representation, but the numbers still have all of the information that you need."

Another advantage of this approach is that it takes into account the flexibility of protein structures, which can be "wiggly" and take on slightly different shapes when interacting with a drug molecule.

High affinity

To make their model less likely to be fooled by decoy drug molecules, the researchers also incorporated a training stage based on the concept of contrastive learning. Under this approach, the researchers give the model examples of "real" drugs and imposters and teach it to distinguish between them.

The researchers then tested their model by screening a library of about 4,700 candidate drug molecules for their ability to bind to a set of 51 enzymes known as protein kinases.

From the top hits, the researchers chose 19 drug-protein pairs to test experimentally. The experiments revealed that of the 19 hits, 12 had strong binding affinity (in the nanomolar range), whereas nearly all of the many other possible drug-protein pairs would have no affinity. Four of these pairs bound with extremely high, sub-nanomolar affinity (so strong that a tiny drug concentration, on the order of parts per billion, will inhibit the protein).

While the researchers focused mainly on screening small-molecule drugs in this study, they are now working on applying this approach to other types of drugs, such as therapeutic antibodies. This kind of modeling could also prove useful for running toxicity screens of potential drug compounds, to make sure they don't have any unwanted side effects before testing them in animal models.

"Part of the reason why drug discovery is so expensive is because it has high failure rates. If we can reduce those failure rates by saying upfront that this drug is not likely to work out, that could go a long way in lowering the cost of drug discovery," Singh says.

This new approach "represents a significant breakthrough in drug-target interaction prediction and opens up additional opportunities for future research to further enhance its capabilities," says Eytan Ruppin, chief of the Cancer Data Science Laboratory at the National Cancer Institute, who was not involved in the study. "For example, incorporating structural information into the latent space or exploring molecular generation methods for generating decoys could further improve predictions."

More information: Rohit Singh et al, Contrastive learning in protein language space predicts interactions between drugs and protein targets, *Proceedings of the National Academy of Sciences* (2023). DOI: 10.1073/pnas.2220778120

This story is republished courtesy of MIT News (web.mit.edu/newsoffice/), a popular site that covers news about MIT research, innovation and teaching.

Provided by Massachusetts Institute of Technology

Citation: New model offers a way to speed up drug discovery (2023, June 8) retrieved 20 June

2024 from https://phys.org/news/2023-06-drug-discovery.html