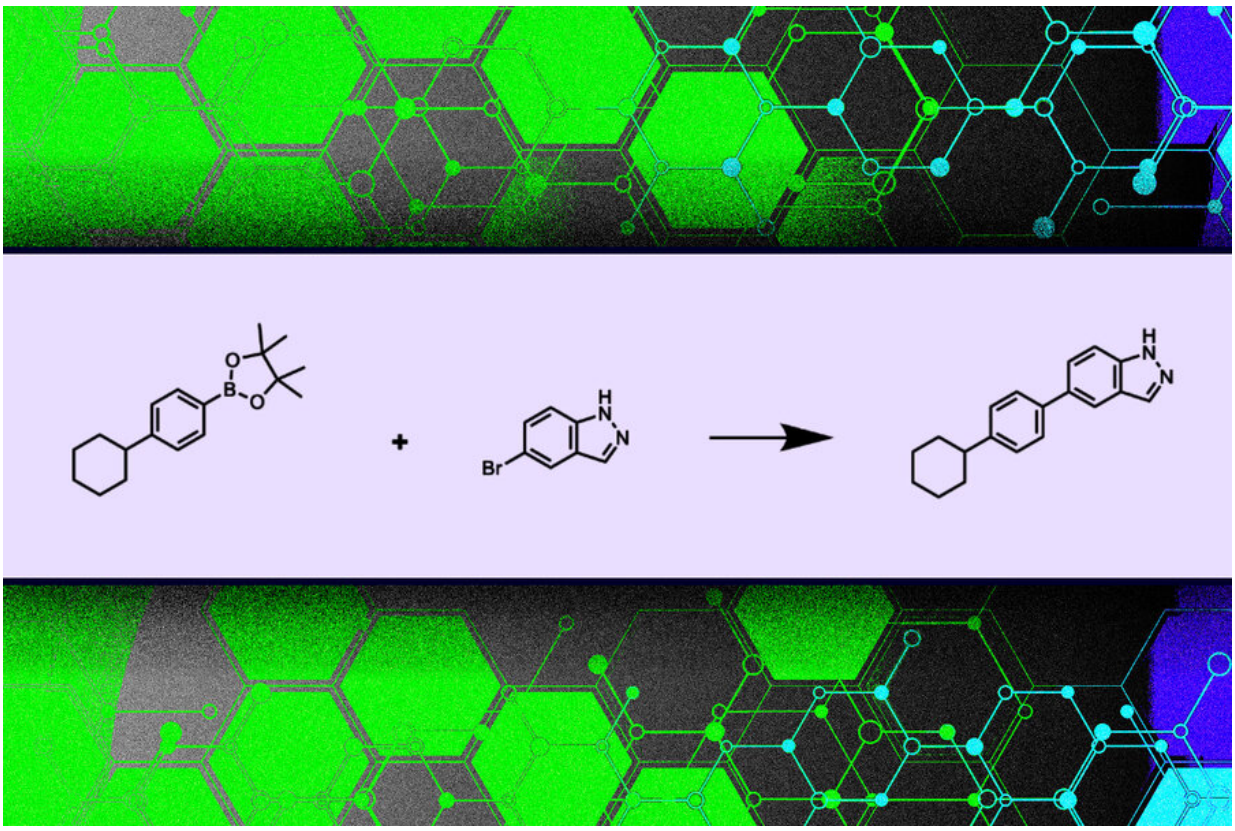
AI technique narrowed to only propose candidate molecules that can be produced in a lab

April 26 2022, by Adam Zewe



MIT researchers have developed a machine learning model that proposes new molecules for the drug discovery process, while ensuring the molecules it suggests can actually be synthesized in a laboratory. Credit: MIT News

Pharmaceutical companies are using artificial intelligence to streamline the process of discovering new medicines. Machine-learning models can propose new molecules that have specific properties which could fight certain diseases, doing in minutes what might take humans months to achieve manually.

But there's a major hurdle that holds these systems back: The models often suggest new molecular structures that are difficult or impossible to produce in a laboratory. If a chemist can't actually make the molecule, its disease-fighting properties can't be tested.

A new approach from MIT researchers constrains a machine-learning model so it only suggests molecular structures that can be synthesized. The method guarantees that molecules are composed of materials that can be purchased and that the chemical reactions that occur between those materials follow the laws of chemistry.

When compared to other methods, their model proposed molecular structures that scored as high and sometimes better using popular evaluations, but were guaranteed to be synthesizable. Their system also takes less than one second to propose a synthetic pathway, while other methods that separately propose molecules and then evaluate their synthesizability can take several minutes. In a search space that can include billions of potential molecules, those time savings add up.

"This process reformulates how we ask these models to generate new molecular structures. Many of these models think about building new molecular structures atom by atom or bond by bond. Instead, we are building new molecules building block by building block and reaction by reaction," says Connor Coley, the Henri Slezynger Career Development Assistant Professor in the MIT departments of Chemical Engineering and Electrical Engineering and Computer Science, and senior author of the paper.

Joining Coley on the paper are first author Wenhao Gao, a graduate student, and Rocío Mercado, a postdoc. The research is being presented this week at the International Conference on Learning Representations.

Building blocks

To create a molecular structure, the model simulates the process of synthesizing a molecule to ensure it can be produced.

The model is given a set of viable building blocks, which are chemicals that can be purchased, and a list of valid chemical reactions to work with. These chemical reaction templates are hand-made by experts. Controlling these inputs by only allowing certain chemicals or specific reactions enables the researchers to limit how large the search space can be for a new molecule.

The model uses these inputs to build a tree by selecting building blocks and linking them through chemical reactions, one at a time, to build the final molecule. At each step, the molecule becomes more complex as additional chemicals and reactions are added.

It outputs both the final molecular structure and the tree of chemicals and reactions that would synthesize it.

"Instead of directly designing the product molecule itself, we design an action sequence to obtain that molecule. This allows us to guarantee the quality of the structure," Gao says.

To train their model, the researchers input a complete molecular structure and a set of building blocks and chemical reactions, and the model learns to create a tree that synthesizes the molecule. After seeing hundreds of thousands of examples, the model learns to come up with these synthetic pathways on its own.

Molecule optimization

The trained model can be used for optimization. Researchers define certain properties they want to achieve in a final molecule, given certain building blocks and chemical reaction templates, and the model proposes a synthesizable molecular structure.

"What was surprising is what a large fraction of molecules you can actually reproduce with such a small template set. You don't need that many building blocks to generate a large amount of available chemical space for the model to search," says Mercado.

They tested the model by evaluating how well it could reconstruct synthesizable molecules. It was able to reproduce 51 percent of these molecules, and took less than a second to recreate each one.

Their technique is faster than some other methods because the model isn't searching through all the options for each step in the tree. It has a defined set of chemicals and reactions to work with, Gao explains.

When they used their model to propose molecules with specific properties, their method suggested higher quality molecular structures that had stronger binding affinities than those from other methods. This means the molecules would be better able to attach to a protein and block a certain activity, like stopping a virus from replicating.

For instance, when proposing a molecule that could dock with SARS-Cov-2, their model suggested several molecular structures that may be better able to bind with viral proteins than existing inhibitors. As the authors acknowledge, however, these are only computational predictions.

"There are so many diseases to tackle," Gao says. "I hope that our method can accelerate this process so we don't have to screen billions of

molecules each time for a disease target. Instead, we can just specify the properties we want and it can accelerate the process of finding that drug candidate."

Their model could also improve existing drug discovery pipelines. If a company has identified a particular molecule that has desired properties, but can't be produced, they could use this model to propose synthesizable molecules that closely resemble it, Mercado says.

Now that they have validated their approach, the team plans to continue improving the chemical reaction templates to further enhance the model's performance. With additional templates, they can run more tests on certain disease targets and, eventually, apply the model to the drug discovery process.

"Ideally, we want algorithms that automatically design molecules and give us the synthesis tree at the same time, quickly," says Marwin Segler, who leads a team working on machine learning for drug discovery at Microsoft Research Cambridge (UK), and was not involved with this work. "This elegant approach by Prof. Coley and team is a major step forward to tackle this problem. While there are earlier proof-of-concept works for molecule design via synthesis tree generation, this team really made it work. For the first time, they demonstrated excellent performance on a meaningful scale, so it can have practical impact in computer-aided molecular discovery.

The work is also very exciting because it could eventually enable a new paradigm for computer-aided synthesis planning. It will likely be a huge inspiration for future research in the field."

More information: Wenhao Gao, Rocío Mercado, Connor W. Coley, Amortized Tree Generation for Bottom-up Synthesis Planning and Synthesizable Molecular Design. arXiv:2110.06389v2 [cs.LG],

arxiv.org/abs/2110.06389

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