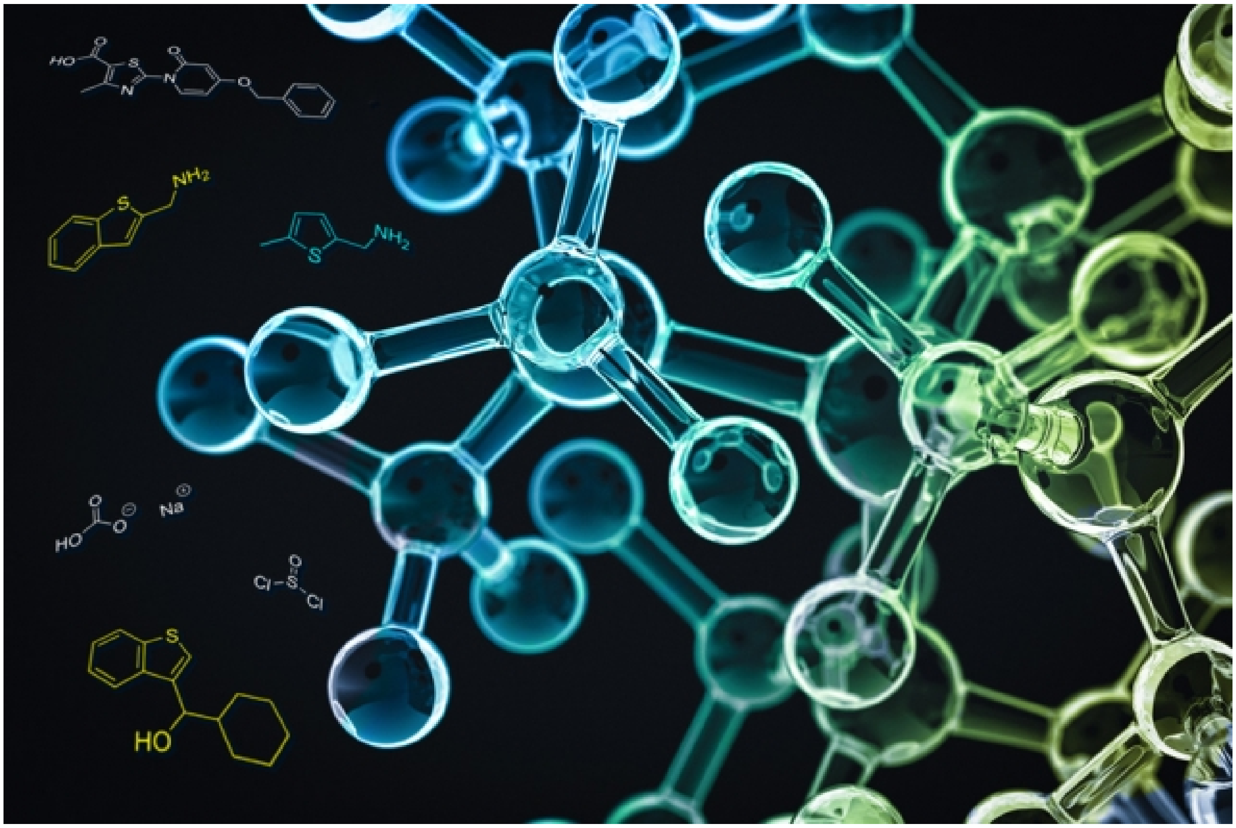
Machine learning approach could aid the design of industrial processes for drug manufacturing

June 27 2017, by Larry Hardesty



A new computer system predicts the products of chemical reactions. “The vision is that you’ll be able to walk up to a system and say, ‘I want to make this molecule.’ The software will tell you the route you should make it from, and the machine will make it,” says professor Klays Jensen. Credit: MIT News

When organic chemists identify a useful chemical compound—a new drug, for instance—it's up to chemical engineers to determine how to mass-produce it.

There could be 100 different sequences of reactions that yield the same end product. But some of them use cheaper reagents and lower temperatures than others, and perhaps most importantly, some are much easier to run continuously, with technicians occasionally topping up reagents in different reaction chambers.

Historically, determining the most efficient and cost-effective way to produce a given molecule has been as much art as science. But MIT researchers are trying to put this process on a more secure empirical footing, with a computer system that's trained on thousands of examples of experimental reactions and that learns to predict what a reaction's major products will be.

The researchers' work appears in the American Chemical Society's journal *ACS Central Science*. Like all machine-learning systems, theirs presents its results in terms of probabilities. In tests, the system was able to predict a reaction's major product 72 percent of the time; 87 percent of the time, it ranked the major product among its three most likely results.

"There's clearly a lot understood about reactions today," says Klavs Jensen, the Warren K. Lewis Professor of Chemical Engineering at MIT and one of four senior authors on the paper, "but it's a highly evolved, acquired skill to look at a molecule and decide how you're going to synthesize it from starting materials."

With the new work, Jensen says, "the vision is that you'll be able to walk up to a system and say, 'I want to make this molecule.' The software will tell you the route you should make it from, and the machine will make it."

With a 72 percent chance of identifying a reaction's chief product, the system is not yet ready to anchor the type of completely automated chemical synthesis that Jensen envisions. But it could help chemical engineers more quickly converge on the best sequence of reactions—and possibly suggest sequences that they might not otherwise have investigated.

Jensen is joined on the paper by first author Connor Coley, a graduate student in chemical engineering; William Green, the Hoyt C. Hottel Professor of Chemical Engineering, who, with Jensen, co-advises Coley; Regina Barzilay, the Delta Electronics Professor of Electrical Engineering and Computer Science; and Tommi Jaakkola, the Thomas Siebel Professor of Electrical Engineering and Computer Science.

Acting locally

A single organic molecule can consist of dozens and even hundreds of atoms. But a reaction between two such molecules might involve only two or three atoms, which break their existing chemical bonds and form new ones. Thousands of reactions between hundreds of different reagents will often boil down to a single, shared reaction between the same pair of "reaction sites."

A large organic molecule, however, might have multiple reaction sites, and when it meets another large organic molecule, only one of the several possible reactions between them will actually take place. This is what makes automatic reaction-prediction so tricky.

In the past, chemists have built computer models that characterize reactions in terms of interactions at reaction sites. But they frequently require the enumeration of exceptions, which have to be researched independently and coded by hand. The model might declare, for instance, that if molecule A has reaction site X, and molecule B has reaction site Y, then X and Y will react to form group Z—unless molecule A also has reaction sites P, Q, R, S, T, U, or V.

It's not uncommon for a single model to require more than a dozen enumerated exceptions. And discovering these exceptions in the scientific literature and adding them to the models is a laborious task, which has limited the models' utility.

One of the chief goals of the MIT researchers' new system is to circumvent this arduous process. Coley and his co-authors began with 15,000 empirically observed reactions reported in U.S. patent filings. However, because the machine-learning system had to learn what reactions wouldn't occur, as well as those that would, examples of successful reactions weren't enough.

Negative examples

So for every pair of molecules in one of the listed reactions, Coley also generated a battery of additional possible products, based on the molecules' reaction sites. He then fed descriptions of reactions, together with his artificially expanded lists of possible products, to an artificial intelligence system known as a neural network, which was tasked with ranking the possible products in order of likelihood.

From this training, the network essentially learned a hierarchy of reactions—which interactions at what reaction sites tend to take precedence over which others—without the laborious human annotation.

Other characteristics of a molecule can affect its reactivity. The atoms at a given reaction site may, for instance, have different charge distributions, depending on what other atoms are around them. And the physical shape of a molecule can render a reaction site difficult to access. So the MIT researchers' model also includes numerical measures of both these features.

According to Richard Robinson, a chemical-technologies researcher at the drug company Novartis, the MIT researchers' system "offers a different approach to machine learning within the field of targeted synthesis, which in the future could transform the practice of experimental design to targeted molecules."

"Currently we rely heavily on our own retrosynthetic training, which is aligned with our own personal experiences and augmented with reaction-database search engines," Robinson says. "This serves us well but often still results in a significant failure rate. Even highly experienced chemists are often surprised. If you were to add up all the cumulative synthesis failures as an industry, this would likely relate to a significant time and cost investment. What if we could improve our success rate?"

The MIT researchers, Robinson says, "have cleverly demonstrated a novel approach to achieve higher predictive reaction performance over conventional approaches. By augmenting the reported literature with negative reaction examples, the data set has more value."

More information: Connor W. Coley et al. Prediction of Organic Reaction Outcomes Using Machine Learning, *ACS Central Science* (2017). DOI: 10.1021/acscentsci.7b00064

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: Machine learning approach could aid the design of industrial processes for drug manufacturing (2017, June 27) retrieved 26 April 2024 from https://phys.org/news/2017-06-machine-approach-aid-industrial-drug.html

This document is subject to copyright. Apart from any fair dealing for the purpose of private study or research, no part may be reproduced without the written permission. The content is provided for information purposes only.