New statistical-modeling workflow may help advance drug discovery and synthetic chemistry

April 8 2024, by Theresa Duque

Berkeley Lab scientists have developed a new automated workflow that applies statistical analysis to process data from nuclear magnetic resonance (NMR) spectroscopy. The advance could help speed the discovery of new pharmaceutical drugs and accelerate the development of new chemical reactions. Credit: Jenny Nuss/Berkeley Lab

A new automated workflow developed by scientists at Lawrence Berkeley National Laboratory (Berkeley Lab) has the potential to allow researchers to analyze the products of their reaction experiments in real time, a key capability needed for future automated chemical processes.

The developed workflow—which applies statistical analysis to process data from nuclear magnetic resonance (NMR) spectroscopy—could help speed the discovery of new pharmaceutical drugs, and accelerate the development of new chemical reactions.

The Berkeley Lab scientists who developed the groundbreaking technique say that the workflow can quickly identify the molecular structure of products formed by chemical reactions that have never been studied before. They recently reported their findings in the *Journal of Chemical Information and Modeling*.

In addition to drug discovery and chemical reaction development, the workflow could also help researchers who are developing new catalysts. Catalysts are substances that facilitate a chemical reaction in the production of useful new products like renewable fuels or biodegradable plastics.

"What excites people the most about this technique is its potential for real-time reaction analysis, which is an integral part of automated chemistry," said first author Maxwell C. Venetos, a former researcher in Berkeley Lab's Materials Sciences Division and former graduate student researcher in materials sciences at UC Berkeley. He completed his doctoral studies last year.

"Our workflow really allows you to start pursuing the unknown. You are no longer constrained by things that you already know the answer to."

The new workflow can also identify isomers, which are molecules with

the same chemical formula but different atomic arrangements. This could greatly accelerate synthetic chemistry processes in pharmaceutical research, for example.

"This workflow is the first of its kind where users can generate their own library and tune it to the quality of that library without relying on an external database," Venetos said.

Advancing new applications

In the pharmaceutical industry, drug developers currently use machine-learning algorithms to virtually screen hundreds of chemical compounds to identify potential new drug candidates that are more likely to be effective against specific cancers and other diseases. These screening methods comb through online libraries or databases of known compounds (or reaction products) and match them with likely drug "targets" in cell walls.

But if a drug researcher is experimenting with molecules so new that their chemical structures don't yet exist in a database, they must typically spend days in the lab to sort out the mixture's molecular makeup. First, by running the reaction products through a purification machine and then using one of the most useful characterization tools in a synthetic chemist's arsenal, an NMR spectrometer, to identify and measure the molecules in the mixture one at a time.

"But with our new workflow, you could feasibly do all of that work within a couple of hours," Venetos said. The time savings come from the workflow's ability to rapidly and accurately analyze the NMR spectra of unpurified reaction mixtures that contain multiple compounds, a task that is impossible through conventional NMR spectral analysis methods.

"I'm very excited about this work as it applies novel data-driven methods

to the age-old problem of accelerating synthesis and characterization," said senior author Kristin Persson, a faculty senior scientist in Berkeley Lab's Materials Sciences Division and UC Berkeley professor of materials science and engineering who also leads the Materials Project.

Experimental results

In addition to being much faster than benchtop purification methods, the new workflow has the potential to be just as accurate. NMR simulation experiments performed using the National Energy Research Scientific Computing Center (NERSC) at Berkeley Lab with support from the Materials Project showed that the new workflow can correctly identify compound molecules in reaction mixtures that produce isomers and also predict the relative concentrations of those compounds.

To ensure high statistical accuracy, the research team used a sophisticated algorithm known as Hamiltonian Monte Carlo Markov Chain (HMCMC) to analyze the NMR spectra. They also performed advanced theoretical calculations based on a method called density-functional theory.

Venetos designed the automated workflow as open source so that users can run it on an ordinary desktop computer. That convenience will come in handy for anyone from industry or academia.

The technique sprouted from conversations between the Persson group and experimental collaborators Masha Elkin and Connor Delaney, former postdoctoral researchers in the John Hartwig group at UC Berkeley. Elkin is now a professor of chemistry at the Massachusetts Institute of Technology, and Delaney a professor of chemistry at the University of Texas at Dallas.

"In chemistry reaction development, we are constantly spending time to

figure out what a reaction made and in what ratio," said John Hartwig, a senior faculty scientist in Berkeley Lab's Chemical Sciences Division and UC Berkeley professor of chemistry.

"Certain NMR spectrometry methods are precise, but if one is deciphering the contents of a crude reaction mixture containing a bunch of unknown potential products, those methods are far too slow to have as part of a high-throughput experimental or automated workflow. And that's where this new capability to predict the NMR spectrum could help," he said.

Now that they've demonstrated the automated workflow's potential, Persson and team hope to incorporate it into an automated laboratory that analyzes the NMR data of thousands or even millions of new chemical reactions at a time.

Provided by Lawrence Berkeley National Laboratory