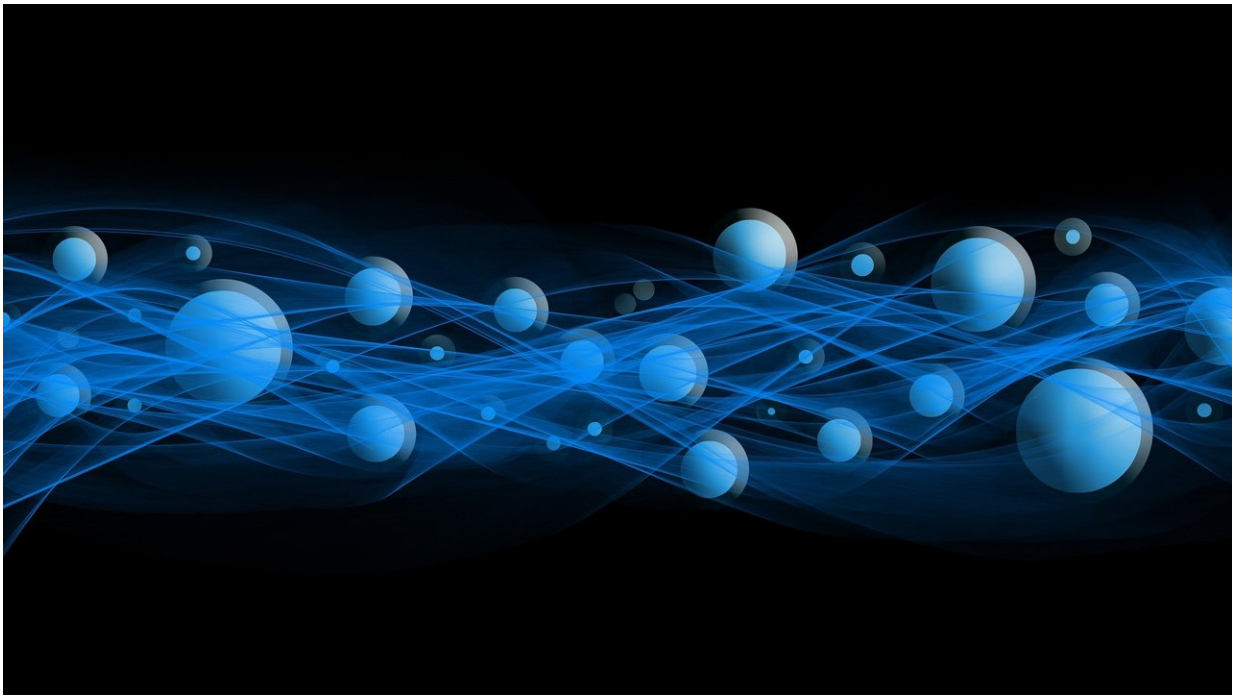
Identifying the dark matter of the molecular world

April 20 2020, by Tom Rickey



Credit: CC0 Public Domain

Imagine that your Facebook feed poses a tantalizing puzzle. You're presented with a few fragments about a person—eye color, hair color, age, and height—and have just one minute to pick out the person's name and identity from hundreds of profiles. If you do so, you win $100 million.

But you know only 10 of these people by name. For the others, you have only a paucity of data to work from. Some are young and some are not so young. Some are blond and some are brunette. Some of their names sound familiar but you can't quite pinpoint how you know them.

This type of scenario—a seemingly impossible task with an enormous payoff—confronts PNNL researchers who study metabolomics. That's the study of small molecules that underlie and inform every aspect of our lives, including energy production, the fate of the planet, and our health.

Scientists estimate that less than 1 percent of small molecules are known. A typical commercially available metabolomics library has maybe 5,000 compounds, but scientists know there are billions more.

How do they "identify" something about which they know so little? It's like asking Galileo to identify stars in deep space that were impossible to detect when he used one of the first telescopes more than 400 years ago.

Enter DarkChem, a research project funded by PNNL's Deep Learning for Scientific Discovery Agile Investment. A team led by Ryan Renslow is bringing artificial intelligence to the table to tackle the vast, unknown landscape of metabolites that bedevil researchers like Tom Metz, who leads PNNL's metabolomics effort.

"Right now, we're just skimming what is potentially knowable and saying goodbye to very interesting data because we can't identify the vast majority of metabolites that our technology detects," said Metz. "Deep learning is providing a new way to solve the puzzle."

Renslow and colleagues Sean Colby and Jamie Nunez have adopted deep learning principles commonly used in applications like language translation and applied them to this dark matter of the molecular world.

Early results are noteworthy: The team's DarkChem network can calculate a key feature of a molecule in milliseconds and with 13 percent fewer errors, compared to 40 hours on a supercomputer running PNNL's flagship quantum chemistry software, NWChem.

"We were shocked at how well DarkChem did," said Renslow.

The network isn't simply crunching through data to compile results. Rather, the network draws upon artificial intelligence. DarkChem was developed so that it can discover new things that are still unknown to humans.

Of football and collision cross-section

In this case, the team trained the program to understand and predict a chemical property known as collision cross-section (CCS). While CCS masks as an intimidating scientific acronym, anyone who has watched a football game has seen something like CCS in action.

Picture a ballcarrier smashing through opposing players. A smaller player might have fewer collisions, but when they do collide with an opponent, the effect is different than when a hulk-like Marshawn Lynch goes into beast mode and shakes off several impacts.

You learn a lot about football players by watching them crash into each other.

In the same way, tracking collisions between metabolite ions traveling through a laboratory instrument filled with gas molecules tells scientists a lot about metabolite ion structures—their size, their mass, and other features. CCS is the mathematical measure of that action, and it's central to unlocking the gas-phase chemical structure—the true "identification"—of a molecule.

Renslow and his team trained DarkChem to calculate CCS for chemical structures, then turned it loose to make the calculation for more than 50 million compounds—a portion of the library of PubChem. The program solved that task in a snap.

While that's a promising step forward, the team is more excited about the implications for all those as-yet-unidentified small molecules.

The network can run forwards as well as backwards—that is, it can solve a molecule's CCS and predict other properties, but it can also generate new chemical structures based on the properties one is looking for. For example, Renslow's team has used DarkChem to put forth several novel chemical structures that have potential for influencing the NMDA receptor, which is involved in memory and other important brain functions.

The network is not simply memorizing data. In fact, the team intentionally adds some numerical fuzziness into the challenges the network faces to keep it from memorizing.

"It's like teaching a computer to recognize a dog," said Renslow. "It could simply memorize the picture, but you want the network to be able to recognize a variety of dogs, so you might flip the picture upside down, stretch it a bit, change its colors. You perturb the image so the program is forced to generalize and rely on the knowledge and rules it has learned."

Teaching the network to learn

To create the network, the team used a form of artificial intelligence called transfer learning, where the network learns from one data set and then applies its knowledge to another data set. The training consisted mainly of three steps:

The program perused more than 50 million known molecules in PubChem, learning the basics of chemistry and how to represent chemical structures mathematically. But the database lacked information about CCS, a crucial measurement for understanding metabolites.

Then, the team exposed DarkChem to a PNNL-developed set of computational CCS data, about 700,000 molecules. This helped train the program about how to link the general information it had learned about chemical structure to CCS.

Finally, the team fine-tuned the network using a small, robust data set of about 1,000 chemical structures whose CCS measurements have been determined through painstaking work in the laboratory.

The ability to calculate CCS for unknown molecules—molecules whose only hint of existence may be one thin line from a mass-spectrometry experiment—adds an important feature to help scientists differentiate one metabolite from another. To shine a light on dark molecular matter.

"Every dimension you add gives you better resolving power," said Colby, who is helping scope out other possible molecular features for DarkChem to analyze, such as infrared spectra, fragmentation patterns, and solvent-accessible surface data.

It's analogous to honing our ability to identify thousands of acquaintances on Facebook.

"You can say someone is male and wears glasses," said Renslow. "But if you can add that he's 54 years old and drives a red Mercedes, you restrict the candidates.

"It's not that much different with metabolites. We keep adding characteristics we can measure, and eventually there is only one

molecule in the universe that fits that combination of data," he added.

Provided by Pacific Northwest National Laboratory