Teaching Computers to Read and Speak Chemistry

Sam Lemonick

Chemists are adapting powerful machine-learning tools used in voice assistants to solve chemistry problems.

Chemistry is a language. Some organic chemistry professors drop that analogy on their students, hoping to get them to see how learning symbols, rules, and suffixes can lead to a broader understanding of the field.

Now chemists are demonstrating that the analogy extends beyond the classroom. By treating molecules and reactions like words and sentences, they have found ways to get the potent machine-learning tools that let Alexa or Siri understand your questions to instead learn chemistry. The scientists hope that these algorithms can then predict molecules that can hit a specific drug target or propose new synthetic routes to compounds. The field shows a lot of potential in its infancy, but the future is hazy because scientists still aren’t sure they know the best ways to talk to computers.

As early as the 1930s, when the first computers were developed, chemists realized they needed ways to communicate chemical information to a machine. Their solution was a line notation: a series of letters, numbers, or other characters that describe or identify a chemical compound.

Simplified molecular-input line-entry system (SMILES) strings and International Chemical Identifiers (InChIs) are among the best-known forms of these line notations today. The former is a sequence of characters that describes a molecule’s atoms and the connections between them, similar to an International Union of Pure and Applied Chemistry name. Ethanol, for instance, can be written CCO, indicating the basic backbone of the molecule minus any hydrogens: a carbon bonded with a carbon bonded with an oxygen. But like chemical names, more than one SMILES string can describe the same molecule. OCC and C(O)C are also valid strings for ethanol.

InChIs contain more information. The notations are composed of different layers separated by slash marks that indicate different types of information. For example, the InChI for ethanol reads InChI=1S/C2H6O/c1-2-3/h3H,2H2,1H3. IS indicates the software version used to encode the molecule, followed by the molecule’s chemical formula, the connections between its atoms, and the number and location of hydrogens. Charge, stereochemistry, and other information can be added in subsequent layers. Unlike with SMILES, each molecule has one unique InChI.

Published: February 12, 2021
Humans usually use words or pictures like those on the left side of this spectrum to talk about molecules, while computers can work with more complex representations, like the matrix on the right. Line notations, like simplified molecular-input line-entry system (SMILES) strings, are accessible to both. Credit: C&EN.

Some chemists are now using these line-notation systems to harness some of the most impressive machine-learning models to make predictions about new molecules and reactions. Natural language processing, also called machine translation, is one of the most active areas of research for computer scientists. These algorithms back voice assistants and allow them to understand written or spoken human language and respond to them. Models like OpenAI’s GPT-3 can synthesize written information and produce lengthy news stories that you wouldn’t immediately recognize as something that a computer wrote.

A handful of chemists reasoned that if chemistry is a language and machine-learning algorithms can understand and produce language-based information, these models might be useful for making chemistry predictions. For computers to make useful predictions about molecules, such as how one might bind to a certain protein or how it might play a role in a multistep synthesis, machines have to know certain rules, like how many bonds a carbon atom can form. Natural language processing models proved the algorithms could learn the rules of spelling, grammar, and syntax. So why not the rules of chemistry?

Several different groups have demonstrated that these algorithms can learn these rules. Marwin H. S. Segler, now of the University of Münster, Mark P. Waller of the tech company Pending.AI, and colleagues developed a strategy for proposing new drug molecules using a type of machine-learning algorithm called a recurrent neural network. Your phone may translate foreign languages using a recurrent neural network. Juno Nam and Jurae Kim, students at Seoul Science High School, used the same kind of algorithm to predict the products of organic chemistry reactions. This paper and others in this story that were published on the arXiv preprint server have not been peer-reviewed.

Philippe Schwaller and colleagues at IBM Research-Zurich developed their own method for predicting reaction products in which they gave the algorithms more context about each atom or group in the molecule. This work laid the groundwork for IBM’s RXN retrosynthesis prediction software, released the same year. And at Stanford University, Vijay Pande’s group demonstrated their own retrosynthesis prediction algorithm.

These groups all used similar sequence-based machine-learning strategies, meaning the algorithm considers each item—whether a word in a sentence or an atom in a molecule notation—only in the context of the words that precede it. So when reading line notations for molecules, these algorithms go atom by atom to learn rules about how molecules are built and how those molecules may react.

It was a powerful method in natural language processing for some time, but it has limitations. For one, word order doesn’t always matter in human languages; a “not” can impart the same meaning whether it’s at the beginning, middle, or end of a sentence. As a result, an algorithm marching through a sentence may interpret meaning where there is none because of the order of the words. Also, as a sentence grows longer, these algorithms can start to forget the beginning, losing important context needed to understand the meaning.

The same problems also apply to reading molecules. SMILES strings—which several of the groups used to train their algorithms—don’t require that atom letters be in the same order to represent the same molecule. And the notations can easily grow to dozens of characters for a complex molecule.

In 2017, a new tool got around some of these limitations in natural language processing. Known as transformers, these algorithms are sequence agnostic, meaning they can understand each word or atom in relation to every other word or atom in a sentence or molecule at the same time. That ability proved to be very powerful in machine translation—it’s the foundation of the prose-writing GPT-3 algorithm—and transformers have rapidly caught on with chemists.

One of the most impressive demonstrations of transformers in chemistry came in December, when researchers at the company DeepMind announced that their AlphaFold 2 algorithm handily won a protein structure prediction competition. Their model can accurately predict a protein’s folded 3-D structure from the sequence of its amino acids in two-thirds of cases tested. The company hasn’t published details of its methods but has said the model uses transformers.

Schwaller, with Alpha A. Lee of the University of Cambridge and others, has adapted the transformer approach for reaction
prediction. Others, including Segler, have looked at ways to use transformers for drug discovery.

As with the earlier sequence-based approaches, most of these transformer-based approaches rely on SMILES strings or similar notations. But not everyone is convinced that’s the right approach for representing molecules. “A string representation is a very, very simple—even naive—representation of molecules,” Lee says. Strings don’t typically capture important information that can explain how a molecule behaves, such as its bond angles or the relationship of different atoms in 3-D space.

He and others are interested in using graph representations to notate molecules. These representations contain information—implicit in line notations—about which atoms are connected to others. Any drawn structure of a molecule is a kind of graph representation, and these can be converted to matrix notations for computers to understand. Evan N. Feinberg, a former student of Pande’s and now CEO of drug development start-up Genesis Therapeutics, and colleagues use a graph-based approach to predict drug molecule properties. They say that their method, compared with other machine-learning approaches, better predicts absorption, distribution, metabolism, elimination, and toxicity properties of potential drug molecules. Théophile Gaudin of IBM Research-Zurich and the University of Toronto is also exploring graph-based transformer models for retrosynthesis planning.

But performance is not the only consideration for computational chemists. As punch-card users back in the 1930s realized, storage space matters too. Graph representations of molecules take up more memory than line notations, which means researchers will need more computer power and more time to run data through machine-learning algorithms. Seyone Chithrananda of the University of Toronto, Gabriel Grand of Reverie Labs, and Bharath Ramsundar of DeepChem recently demonstrated those differences in required computing resources. They’re developing a technique for pre-training transformer-based models with SMILES strings to make the models faster at chemistry prediction. Grand notes that a similar graph-based method from the company Tencent needs 250 processors in its calculations. The trio’s method uses just one.

Given the advantages and disadvantages of these different methods for representing molecules, no one in this newborn field seems sure which approach, if any, will eventually win. Many say it will likely be a combination for the foreseeable future or until another revolutionary idea, like transformers, appears.

Also, these language-based algorithms are just one way that chemists are exploring machine learning. Some researchers have experimented with adapting machine-learning approaches originally developed to understand images. Others use less complex machine-learning algorithms to complement chemistry rules written by experts. All three approaches are finding their way into the scientific literature and the marketplace, and none has yet shown clear advantages over the others.

So chemists continue to ask basic questions about how to best represent molecules and communicate them to machines. And now as chemists talk to computers, the computers are learning to talk back.

Sam Lemonick is an associate editor at Chemical & Engineering News, the weekly newsmagazine of the American Chemical Society.