CREDIT: LAURENT HRYBYK

Downloaded via UNIV OF CINCINNATI on April 16, 2023 at 02:50:24 (UTC).
See https://pubs.acs.org/sharingguidelines for options on how to legitimately share published articles.

COMPUTATIONAL CHEMISTRY

Chemists enlist computers to analyze tricky visual data

Machine learning algorithms could help scientists quickly tackle torrents of data

ARIANA REMMEL, C&EN STAFF

Every rock on Earth is a time capsule, holding traces of our planet from when the minerals first formed. Consider a smidge of 164-million-year-old clay, says M. Joseph Pasterski, an organic geochemist at the University of Illinois Chicago. A sample just about 2 cm long contains complex layers of mixed materials composed of myriad molecules from Earth's distant past. Researchers like Pasterski want to use today's state-of-the-art analytical tools to sort through the mélange of minerals to reveal chemical signs of ancient life.

For example, using time-of-flight secondary ion mass spectrometry (TOF-SIMS), Pasterski can map elements in 500 by 500 µm patches within the ancient clay sample, generating around 62,000 spectra per patch. Pasterski then uses those data to look for organic molecules, such as steranes, that could be signs of life. The spectra capture not only these biosignatures but also all the other grit and debris that one might expect to find in an old clod of mud. "There's so much data that isn't noise but isn't necessarily [relevant to] the question that you're asking," Pasterski says. Manually processing the spectra from one patch to identify the signals of interest is tedious enough, he says. Now imagine processing spectra for the entire clay fragment or an entire library of specimens.

That's an intimidating chunk of data for a scientist to parse, but what about for a computer? Pasterski wondered. He's not the only researcher in the chemical sciences pondering if computers could be data analysis assistants.

Scientists are awash in seas of data. In 2022, the International System of Units gained the prefixes ronna for 10^{27} and quetta for 10^{30} to help measure massive data sets as the digitized collection of human knowledge pushes the limits of our comprehension. Advances in analytical instrumentation and methods, and the growing number of open-science repositories for chemical information, have afforded researchers a glut of data to explore questions across the biological and physical sciences.

Some scientists want to enlist computers to help them make sense of all that information. When it comes to visual data such as spectra and microscopy images, machines are well positioned to assist. Machine learning algorithms are already good at identifying patterns and creating images. For example, the Dall-E 2 program from tech company OpenAI can generate an image from a text prompt. And AlphaFold, a program created by the tech company DeepMind, has demonstrated that

machines can learn chemical concepts: it has already predicted the 3D structures of over 200 million proteins from more than 10 million species.

Now computational scientists are designing algorithms to automate the processing of molecular data sets such as multidimensional nuclear magnetic resonance spectra, complex mass spectrometry data, and micrographs. To do that, researchers are teaching computers to approach visual data as a human chemist would. In this approach, scientists need to ask, What does an expert look at in these spectra? And how do we train a machine to look at the same things? says Connor Coley, a computational chemist at the Massachusetts Institute of Technology. The resulting programs could speed up experiments, process large volumes of data, and allow researchers to study short-lived molecular systems that were previously too difficult to observe.

Wading through spectra

Biomedical researchers trying to understand a disease's molecular basis within clinical tissue samples must sift through a chemical soup. One method clinicians use to assess tissue samples' chemical composition is mass spectrometry. Mass spectrometers ionize the material to measure the mass-to-charge ratio of each compound, allowing scientists to distinguish chemicals by their molecular weight. Additional methods can break these molecules into smaller fragments whose mass spectra offer further clues to how atoms are arranged within the compound. Combining these techniques helps scientists describe molecules by their structure as well as molecular weight. "You kind of see a jigsaw of your molecule, and you have to piece it back together," says Sam Goldman, a computational biology PhD student working with Coley at MIT. Mass spectrometry is a powerful tool for chemically analyzing biological samples, but current methods—both with and without machine assistance—struggle to connect fragmentation patterns to the molecules that make them, Coley says. Goldman and Coley realized that machine learning could help biologists more easily identify unknowns in their samples by addressing this bottleneck.

Computational chemists have already trained computers to analyze spectra by considering the collection of peaks as a fingerprint of the molecule that produced the data. These programs are proficient at finding patterns in the molecular weights of whole molecules and the fragments they produce, but a lack of chemistry-specific knowledge limits their utility, Coley says. A human scientist looking at these spectra would start by identifying the signals for whole molecules and then look for the pattern created by the fragments' peaks. Researchers with chemistry expertise can quickly ascribe differences in mass between sets of fragments to the loss of functional groups. They can then intuit how those functional groups fit together within the whole molecule. Goldman and his colleagues wanted to improve on existing computational methods by coaxing a machine learning algorithm to think like a researcher. Their algorithm interprets peaks within a mass spectrum as chemical formulas and understands that fragments' peaks in the same spectrum relate to one another according to molecular bonding principles. "There's a ton of patterns you can learn from, and we want to give the model the best chance at picking the right patterns," Goldman says.

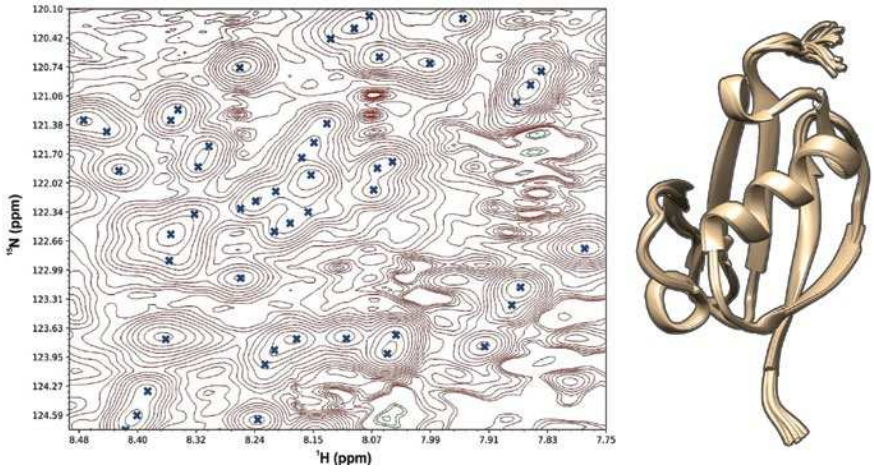
Goldman and his colleagues trained their program, called Metabolite Inference with Spectrum Transformers (MIST), with spectra from more than 27,000 molecules from public-access databases such as those from the National Institute of Standards and Technology and the Global Natural Products Social Molecular Networking. MIST also learned from spectra simulated by other machine learning algorithms. In benchmark experiments, MIST successfully identified the structures of more than 66% of the test molecules from their mass spectra, a clear improvement over alternative programs. Next, Goldman and his colleagues tested MIST on spectra taken from tissue samples from people with inflammatory bowel disease. MIST pointed the team toward novel dipeptides and alkaloid molecules associated with more severe symptoms. The results were published on a preprint server while the manuscript awaits peer review (bioRxiv 2022, DOI: 10.1101/2022.12.30.522318).

Goldman says MIST is designed to work alongside other computational tools that researchers use to process and interpret high-throughput mass spectrometry data. He hopes that using machine learning to perform computationally cumbersome structure identification will free biologists to spend more time thinking about how metabolites function within complex biological systems.

Machine learning is also helping researchers analyze data to elucidate protein structures. For example, Artificial Intelligence for NMR Applications (ARTINA) can solve a protein structure directly from multidimensional NMR spectra (*Nat. Commun.* 2022, DOI: 10.1038/s41467-022-33879-5).

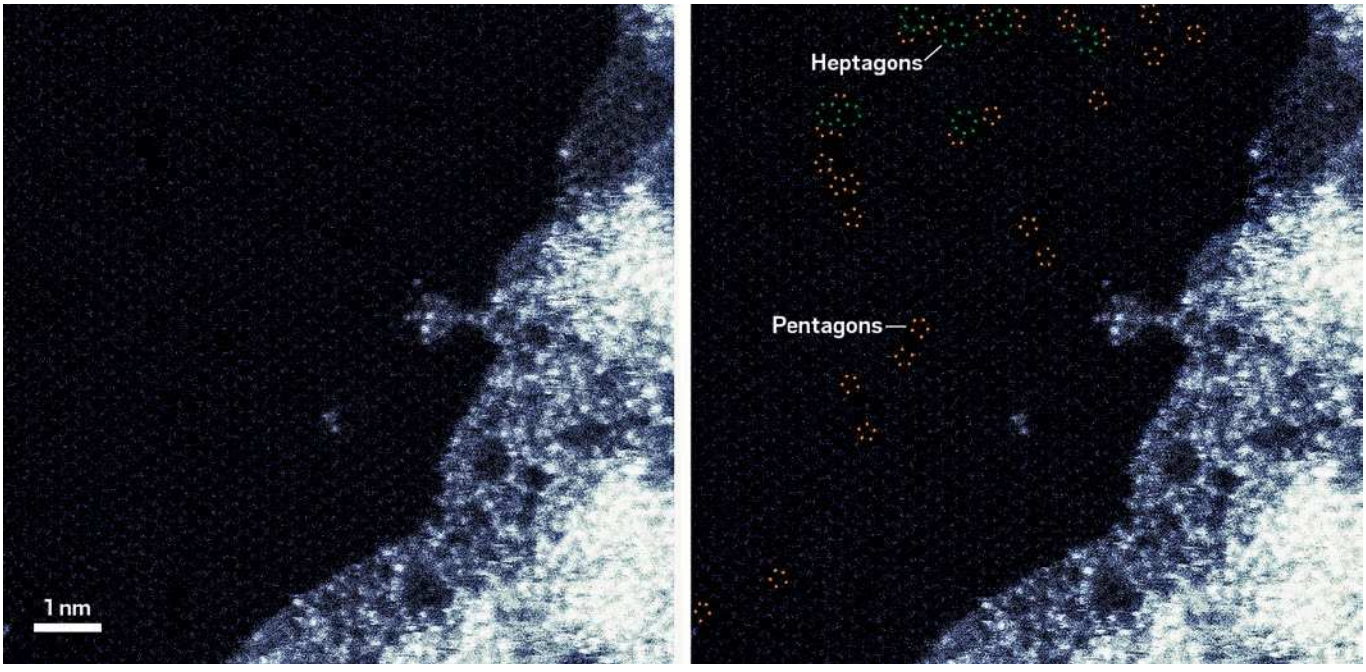
NMR is a useful tool for scientists interested in studying protein dynamics in solution or even within whole cells. Protein NMR spectra help scientists probe the molecular composition of a protein and how it interacts with other molecules. But the resulting spectra can be unwieldy. While some spectra are represented as a series

The machine learning algorithm Artificial Intelligence for NMR Applications can decode complex, multidimensional nuclear magnetic resonance spectra (left) to propose 3D structures for a given protein (right).



> "You kind of see a jigsaw of your molecule, and you have to piece it back together."
>
> **—Sam Goldman,** PhD student, MIT

CREDIT: PETER GÜNTERT



This scanning transmission electron microscopy image of graphene contains thousands of atoms within one field of view alone, mostly arranged in hexagons. AtomAI can quickly sniff out spots where the carbon atoms are arranged in pentagons and heptagons.

of peaks along a single axis, others look more like topographical maps: clusters of irregular, multidimensional peaks sprawling across a grid. Even a specialist might spend weeks or months picking out all the relevant signals, assigning them to amino acids in the protein, and assembling those amino acids into a suitable 3D protein structure. Peter Güntert, a structural biologist at Goethe University Frankfurt and the Swiss Federal Institute of Technology (ETH), Zurich, and his colleagues wanted to make this process more manageable with ARTINA.

To begin, Güntert and his colleagues compiled a training set of experimental and simulated protein NMR spectra in which each peak was labeled and assigned to a feature within the protein of interest. Using this data set, the researchers taught ARTINA to visually inspect spectral data so that it could automatically annotate the peaks and propose a 3D structure to explain the patterns it detected. The researchers tested ARTINA on NMR spectra from 100 proteins 35 to 175 amino acids long. The program accurately solved the structures for these proteins and correctly assigned 91% of spectral peaks to features within those structures. Much like human specialists, the program was better at predicting how the protein backbones folded than how the amino acid side chains arranged themselves. The most prominent errors arose in proteins containing disordered regions or certain secondary structures, such as floppy helices. The results are sufficient to show that ARTINA is no worse than the average spectroscopist, Güntert says. And while researchers should be cautious when using the program, errors are usually easy to spot. Sometimes the program clearly messes up and doesn't output results, or the structure is obviously meaningless. "It's actually usually quite difficult to get nice-looking but significantly wrong results," Güntert says.

Now researchers can upload data to a web server called NMRtist, which can perform all the steps in protein NMR analysis—such as annotating the spectra and producing a full protein structure—with zero intervention. The uploaded spectra are automatically added to new training sets, which will be used to improve ARTINA in future iterations. Because ARTINA doesn't require specialized training, Güntert hopes his team's efforts will help clinicians and biomedical scientists who are unfamiliar with protein NMR incorporate the methods in their research.

"I like to say that artificial intelligence is actually augmented intelligence."

—Maxim Ziatdinov, research scientist, Oak Ridge National Laboratory

AI assistant

With both MIST and ARTINA, a scientist inputs spectra into the program and waits for it to return a result. Maxim Ziatdinov, a research scientist at Oak Ridge National Laboratory, is developing a program that will work alongside a materials scientist as they run microscopy experiments in real time. Scientists interested in designing new materials often use electron and scanning probe microscopes to investigate atomic and molecular features within a sample. These instruments can also manipulate structures within a sample. The way the material responds can help researchers elucidate structure-function relationships, a key to understanding why the material behaves as it does.

Electron and scanning probe microscopy experiments can be time consuming. First, these microscopes record images frame by frame across the sample within fields of view that usually contain several thousand atoms and hundreds of features. Upon closer inspection, these features could reveal interesting properties, such as electrical conductivity and energy storage capabilities, Ziatdinov says. So researchers then use advanced microscopy techniques, such as laser pulses, to manipulate individual atoms within these features to gather that key structure-function information. "There's no way to analyze all of [the data] manually," Ziatdinov says.

Because exposure to the microscopes' harsh conditions can degrade samples, the

CREDIT: MAXIM ZIATDINOV

experiments have a time limit. Some materials break down under the instruments' strong vacuum and high-energy beams before the whole sample can be processed. So scientists have to be judicious in the number of experiments they run on a sample or avoid unstable materials altogether.

Ziatdinov and his colleagues saw an opportunity to apply machine learning to optimize the imaging and material manipulation process. The researchers designed a program called AtomAI, which can identify every atom and its position within a scan of each frame. The algorithm then predicts which regions are most likely to yield a given functional behavior (*Nat. Mach. Intell.* 2022, DOI: 10.1038/s42256-022-00555-8).

"The general idea is that you get this easy-to-acquire structural image, make several spectral measurements within that image, then use that information to predict how the spectra would likely look in the remaining portion of that image," Ziatdinov says. With this information, the researcher can decide what to do next, such as manipulate the structures with laser pulses or take more extensive measurements, without ever leaving the microscope. "It's an assistant [that] allows you to make decisions faster because it gives you an idea of what's going on in your system while the experiment is still running," he says. That improvement cuts experiment times down from weeks to days, he says.

So AtomAI could allow researchers to study samples that would be too unstable under the traditional, slower experimental workflows, Ziatdinov says. And he believes self-driving microscopes could make it easier for scientists to not only discover new materials but also facilitate the fabrication of atomically precise devices, such as the qubit components needed for quantum information technologies.

Back in Illinois, Pasterski has been using a TOF-SIMS and a machine learning algorithm to study those ancient clay samples. This combination can determine with greater than 80% accuracy if a section of clay contains primarily organic or inorganic material. It also successfully recognized sterane-based biosignatures with greater than 95% accuracy, according to research he presented at the 2022 Fall Meeting of the American Geophysical Union in Chicago. Pasterski believes these preliminary results show that further development could make machine learning approaches a powerful tool for seeking out signs of life in geological samples.

In particular, he thinks machine learning methods would be helpful in an extreme experimental setting: another planet. When probes like those on NASA's *Perseverance* rover study samples on Mars, the extreme distance from Earth slows data transmission, making it difficult for scientists on Earth to guide the machine on how to analyze materials. Also, while these probes carry powerful instruments, engineers can fit only so much on these remote labs because of weight limitations during launch. A machine learning program directing a mass spectrometer on a rover could help make sample analysis more efficient with the onboard equipment and less reliant on human guidance, Pasterski thinks.

But computational scientists caution that machine learning won't solve all of chemists' experimental challenges. Scientists shouldn't reach for machine learning just to use machine learning, MIT's Coley says. "We're really trying to identify the places where it excels and where it provides benefits over the existing techniques," he says. And humans are still better at many tasks.

"I like to say that artificial intelligence is actually augmented intelligence," Ziatdinov says. For as good as computers may be at crunching numbers at dizzying speeds, human scientists will always be the ones asking the questions. ■