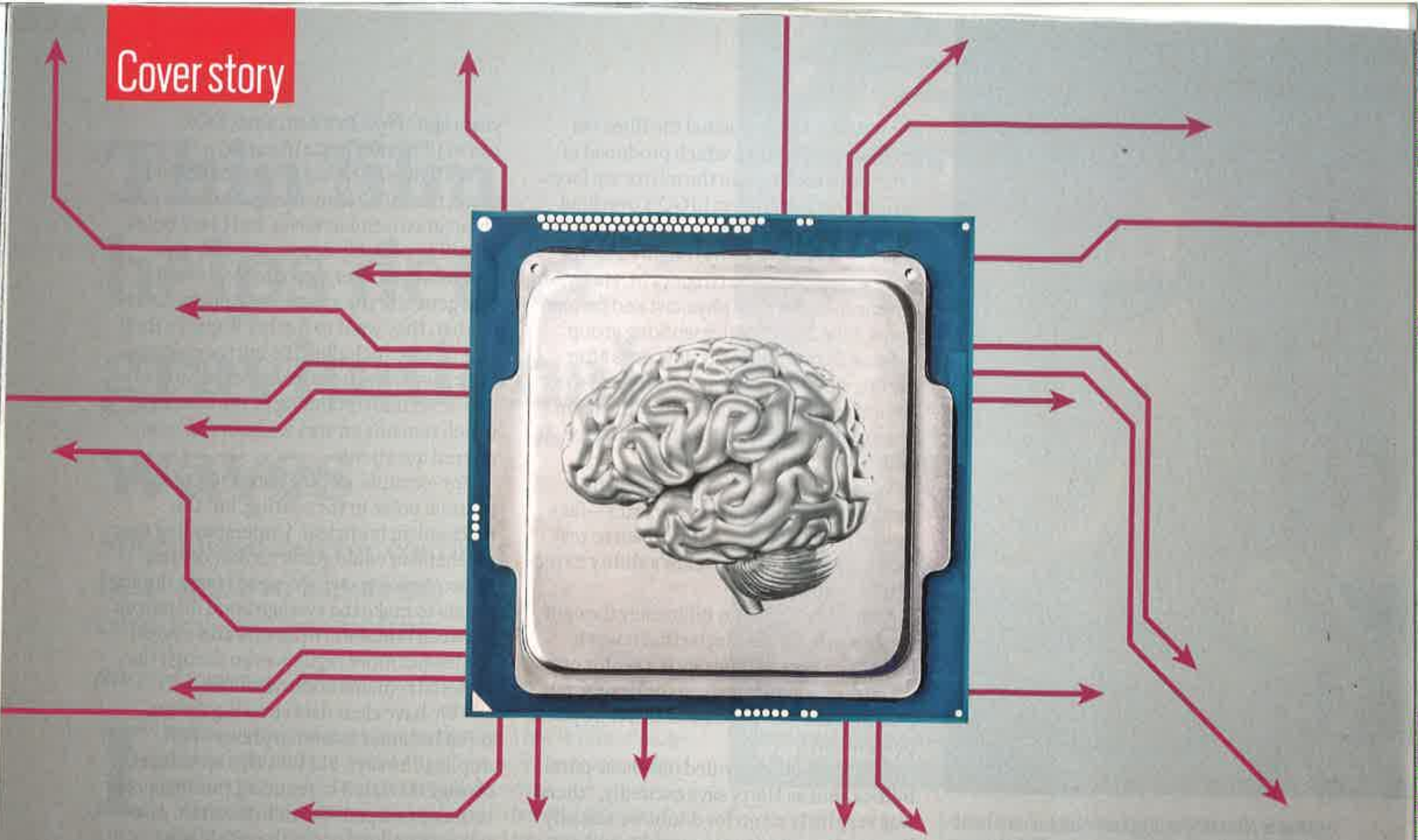


New directions for machine learning

Algorithms that have changed daily life move into the lab

Applications such as spam filtering, economic forecasting, and Netflix recommendations are powered by algorithms that allow computers to learn from and make predictions based on vast collections of data. But even as machine learning has purged e-mail inboxes of Vi@gra, the technology has been slower to infiltrate chemistry.

As described in this collection of stories, however, the gap is starting to fill. Some researchers are pushing the boundaries of computer science and chemistry, applying the most sophisticated version of machine learning yet—deep learning—to drug discovery. A poster child of this effort is a nascent partnership between pharma firm GlaxoSmithKline and a pair of government labs that seeks to use screening data on more than a million compounds to find promising therapeutic candidates (see page 31). Meanwhile, publishers and others haven't given up on more traditional machine-learning algorithms: They're using them to develop tools to capitalize on information catalogued in journals, patents, and elsewhere (see page 33).

Read on to learn how chemists are embracing the era of big data. ■

CREDIT: WILL LUDWIG/C&EN/SHUTTERSTOCK

COMPUTATIONAL CHEMISTRY

Deep learning to the rescue

Pharmaceutical chemists pin hopes on new machine-learning method for drug discovery

ELIZABETH K. WILSON, C&EN WEST COAST

A depressing trend looms large for today's pharmaceutical chemist: For the past 60 years, the cost of developing a new drug—currently about $2.5 billion—has roughly doubled every nine years. Many have placed the blame partly on cheminformatics: Even though lots of time and money has been spent on computationally searching for and docking drugs into proteins and other target molecules, the technique hasn't lived up to the hype and hasn't produced a prolific stream of good drug candidates.

But some chemists believe they've found a new strategy that may help solve their drug discovery woes. They think that a serendipitous combination of computer processing advances, access to huge chemical data sets, and a groundbreaking computational strategy called deep learning could usher in a way to quickly and efficiently teach computers to find successful drugs in ways that far surpass current computer-based methods.

Only in use for about five years, deep learning is a specialized, more sophisticated type of machine learning, a computational technique that generally uses data sets to teach itself and then applies its newly acquired "knowledge" to make predictions. Chemists have been using traditional machine learning for several decades to train computers to search libraries of compounds for possible drug candidates.

Deep learning burst on the scene in 2012 as a computer science breakthrough and has led to remarkable advances in technologies such as voice and face recognition. Once clumsy and error-prone, voice and face recognition are now accurate and ubiquitous on computers, tablets, and smartphones thanks to the computational strategy. Self-driving cars have even benefited from deep learning's ability to navigate roadways after training itself with radar, sonar, imaging and other data. As evidence, the transportation company Uber announced in December 2016 that it would begin pilot-testing its self-driving cars in San Francisco (a project later moved to Arizona).

Now, a growing number of chemists are hoping this strategy might work for them, too.

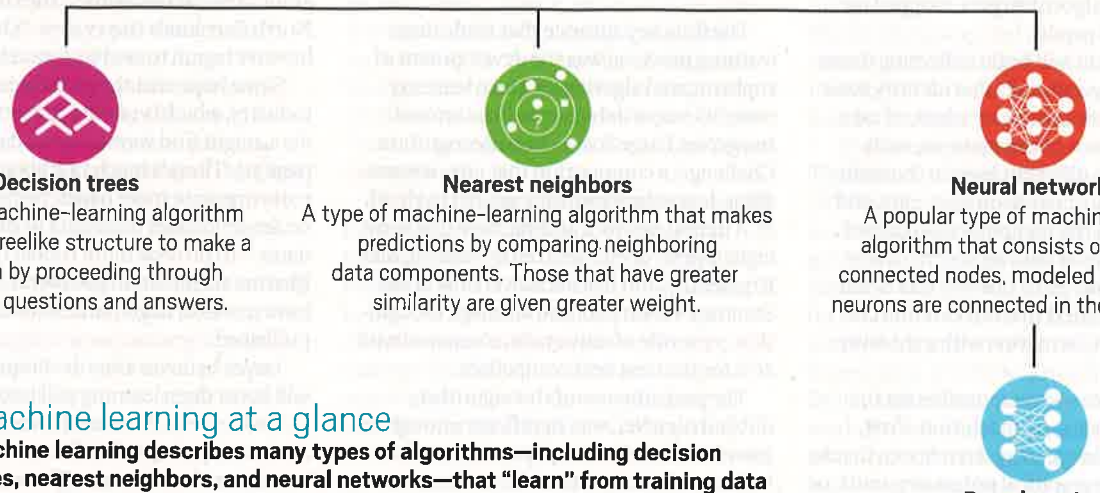
In drug discovery, the current collection of 5,000 or so molecular properties, such as aromaticity or bond strength, that computers use to find potential therapeutic compounds have been hand-selected by chemists over the years. Deep learning, however, has the potential to find, all by itself, combinations of druglike properties needed for therapeutic candidates. And some of these combinations could very well be parameters no human could discern.

Deep learning applied to chemistry has almost evangelical support in some research circles, made up of academicians and industry scientists alike. These researchers hope the new strategy will reverse the declining success of drug discovery. Money is pouring into start-ups that offer different types of deep-learning platforms. For instance, Palo Alto-based start-up TwoXAR recently raised $3.4 million in seed money, and San Francisco-based Atomwise raised $6 million.

As with any new field that generates great excitement, some drug discovery chemists warn that the enthusiasm needs to be tempered with realism. Big data sets, a requirement for good deep-learning performance, are still hard to come by in chemistry. And researchers need to develop ways to ensure the molecules "discovered" by deep-learning algorithms are compounds that chemists can realistically synthesize.

Deep learning has yet to show that it's significantly better at finding drug candidates than other machine-learning methods, maintains Mark Murcko, chief scientific officer of Relay Pharmaceuticals.

Nonetheless, an increasing number of groups that blend chemists with computer scientists are banking on continuing advances in deep learning as the drug discovery tool of the future. The use of deep learning in chemistry is less than two years old, says Olexandr Isayev, a computational chemist at the University of North Carolina, Chapel Hill, "but despite that



Decision trees
A type of machine-learning algorithm that uses a treelike structure to make a prediction by proceeding through a series of questions and answers.

Nearest neighbors
A type of machine-learning algorithm that makes predictions by comparing neighboring data components. Those that have greater similarity are given greater weight.

Neural networks
A popular type of machine-learning algorithm that consists of groups of connected nodes, modeled after the way neurons are connected in the human brain.

Deep learning
A variant of neural networks, with an architecture that uses multiple layers of nodes, each layer making clearer predictions.

Machine learning at a glance

Machine learning describes many types of algorithms—including decision trees, nearest neighbors, and neural networks—that "learn" from training data sets and then make real-world predictions. Deep learning is a sophisticated type of neural network.

COMPUTATIONAL CHEMISTRY

Deep learning to the rescue

Pharmaceutical chemists pin hopes on new machine-learning method for drug discovery

ELIZABETH K. WILSON, C&EN WEST COAST

A depressing trend looms large for today's pharmaceutical chemist: For the past 60 years, the cost of developing a new drug—currently about $2.5 billion—has roughly doubled every nine years. Many have placed the blame partly on cheminformatics: Even though lots of time and money has been spent on computationally searching for and docking drugs into proteins and other target molecules, the technique hasn't lived up to the hype and hasn't produced a prolific stream of good drug candidates.

But some chemists believe they've found a new strategy that may help solve their drug discovery woes. They think that a serendipitous combination of computer processing advances, access to huge chemical data sets, and a groundbreaking computational strategy called deep learning could usher in a way to quickly and efficiently teach computers to find successful drugs in ways that far surpass current computer-based methods.

Only in use for about five years, deep learning is a specialized, more sophisticated type of machine learning, a computational technique that generally uses data sets to teach itself and then applies its newly acquired "knowledge" to make predictions. Chemists have been using traditional machine learning for several decades to train computers to search libraries of compounds for possible drug candidates.

Deep learning burst on the scene in 2012 as a computer science breakthrough and has led to remarkable advances in technologies such as voice and face recognition. Once clumsy and error-prone, voice and face recognition are now accurate and ubiquitous on computers, tablets, and smartphones thanks to the computational strategy. Self-driving cars have even benefited from deep learning's ability to navigate roadways after training itself with radar, sonar, imaging and other data. As evidence, the transportation company Uber announced in December 2016 that it would begin pilot-testing its self-driving cars in San Francisco (a project later moved to Arizona).

Now, a growing number of chemists are hoping this strategy might work for them, too.

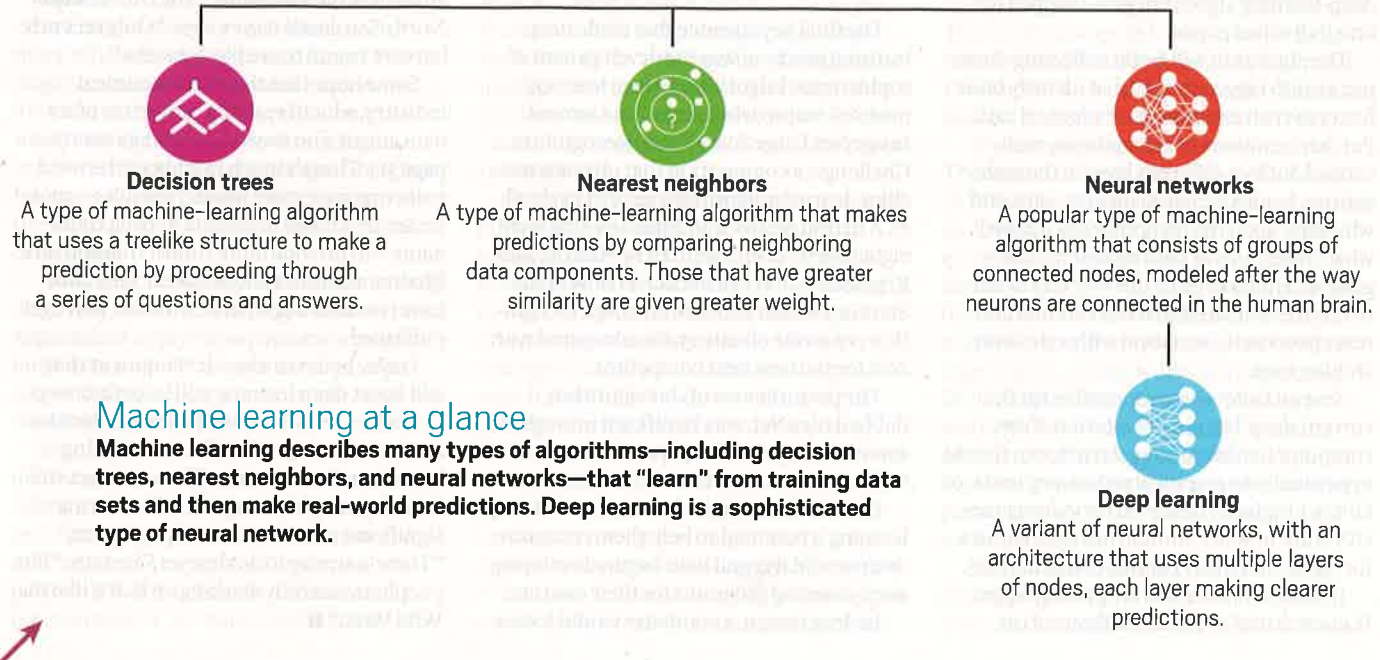
In drug discovery, the current collection of 5,000 or so molecular properties, such as aromaticity or bond strength, that computers use to find potential therapeutic compounds have been hand-selected by chemists over the years. Deep learning, however, has the potential to find, all by itself, combinations of druglike properties needed for therapeutic candidates. And some of these combinations could very well be parameters no human could discern.

Deep learning applied to chemistry has almost evangelical support in some research circles, made up of academicians and industry scientists alike. These researchers hope the new strategy will reverse the declining success of drug discovery. Money is pouring into start-ups that offer different types of deep-learning platforms. For instance, Palo Alto-based start-up TwoXAR recently raised $3.4 million in seed money, and San Francisco-based Atomwise raised $6 million.

As with any new field that generates great excitement, some drug discovery chemists warn that the enthusiasm needs to be tempered with realism. Big data sets, a requirement for good deep-learning performance, are still hard to come by in chemistry. And researchers need to develop ways to ensure the molecules "discovered" by deep-learning algorithms are compounds that chemists can realistically synthesize.

Deep learning has yet to show that it's significantly better at finding drug candidates than other machine-learning methods, maintains Mark Murcko, chief scientific officer of Relay Pharmaceuticals.

Nonetheless, an increasing number of groups that blend chemists with computer scientists are banking on continuing advances in deep learning as the drug discovery tool of the future. The use of deep learning in chemistry is less than two years old, says Olexandr Isayev, a computational chemist at the University of North Carolina, Chapel Hill, "but despite that



Decision trees
A type of machine-learning algorithm that uses a treelike structure to make a prediction by proceeding through a series of questions and answers.

Nearest neighbors
A type of machine-learning algorithm that makes predictions by comparing neighboring data components. Those that have greater similarity are given greater weight.

Neural networks
A popular type of machine-learning algorithm that consists of groups of connected nodes, modeled after the way neurons are connected in the human brain.

Deep learning
A variant of neural networks, with an architecture that uses multiple layers of nodes, each layer making clearer predictions.

Machine learning at a glance

Machine learning describes many types of algorithms—including decision trees, nearest neighbors, and neural networks—that "learn" from training data sets and then make real-world predictions. Deep learning is a sophisticated type of neural network.

we're seeing tremendous progress.”

For example, Atomwise is hoping its deep-learning architecture, known as AtomNet, will allow the firm to repurpose a drug, already evaluated by the Food & Drug Administration for a different use, to prevent Ebola infection.

AtomNet screened 7,000 already evaluated medications for their ability to bind strongly to a protein in the Ebola virus known as glycoprotein 2. This protein drives Ebola infection with a clawlike structure that tears into cell membranes, allowing the virus to slip in and infect cells. AtomNet pinpointed 17 promising small molecules that block this action, one of which prevented Ebola infection of cells in the lab, says Abraham Heifets, CEO of Atomwise. The firm is not yet disclosing this top candidate's identity.

The invention of deep learning, which aided AtomNet in finding the promising candidate, can be traced back as far as 1998. But it wasn't until 2006 that pioneering computer scientist Geoffrey Hinton, now professor emeritus at the University of Toronto, began publishing papers that laid serious groundwork for deep learning.

There are many varieties of traditional machine-learning algorithms—with names like decision trees, nearest neighbors, and neural networks. Neural networks—first described in the 1950s—model the action of neurons in the brain and are one of the most popular machine-learning methods.

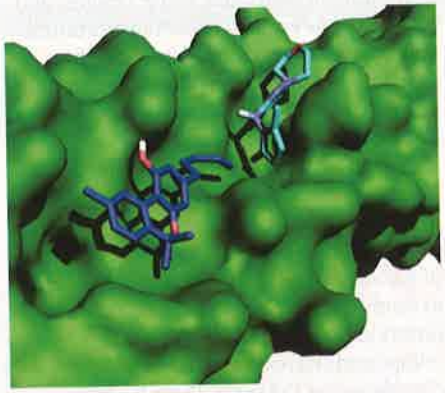
Deep learning takes the architecture of neural networks to a new level of accuracy. Instead of a process that employs one set of data that leads to an output, deep learning is based on multiple layers of calculations. For example, in training a computer to recognize images of cats, a data set fed into a deep-learning algorithm gets chopped up into individual pixels.

The algorithm will begin collecting those pieces into bigger chunks that identify basic features such as outlines, or edges, of cats. Further iterations of calculations, each carried out by a different layer in the architecture, begin to close in on eyes, ears, and whiskers. Soon the computer has learned what collections of data represent cats in general. Humans learn this way too: Scientists have estimated that our cerebral cortexes process information with a six-layer architecture.

Several factors are responsible for the current deep-learning revolution. First, computer hardware has gotten faster, thanks in particular to graphical processing units, or GPUs. Originally designed for video games, GPUs are now in common use by scientists for rapid, massively parallel computations.

In addition, data sets are getting bigger. It's now trivial to collect millions of cat images off the internet, for example, or to scan millions of tweets. In fact, without huge amounts of data, deep learning can actually perform worse than other algorithms. Because it's so good at learning to recognize patterns, if an initial training data set is too small, the algorithm will just memorize it—a problem known as overfitting.

In a classic example of this problem, in the 1980s the U.S. military wanted to train computers to recognize tanks in aerial footage. They prepared training sets, consisting of forested areas with and without tanks. The computer learned its task on that data set without a hitch. But in real life, the project failed. It turned out that the training footage with the tanks had been taken in the morning, and the footage without the tanks had been taken in the afternoon. So the computer trained itself to recognize tanks by looking for differences in light rather than identifying the tanks themselves.



AtomNet searched databases of drugs (two shown here, stick representations) for molecules that inhibit the action of an Ebola virus protein (green).

CREDIT: ATOMWISE

The final key advance that made deep learning practical was the development of sophisticated algorithms. Deep learning made its major debut at the 2012 annual ImageNet Large Scale Visual Recognition Challenge, a competition that pits new machine-learning algorithms against each other. A neural network architecture that went eight layers “deep,” written by Hinton, Alex Krizhevsky, and Ilya Sutskever now at the Stanford Vision Lab, had an image recognition error rate of only 15.4%, compared with 26% for the next best competitor.

The performance of this algorithm, dubbed AlexNet, was significant enough to astonish the field. The paper that resulted has been cited more than 8,000 times.

Chemists have quickly taken note of deep learning's potential to help them recognize drug candidates and have begun developing deep-learning programs for their own use.

In drug design, a computer would look at aspects of molecules rather than pixels in images, explains Bartosz Grzybowski, chemistry professor at Ulsan National Institute of Science & Technology. Instead of identifying humans, it would identify, say, molecules that block G-protein-coupled receptors, which are popular drug targets. “It's very difficult to wrap your head around all those data,” he says.

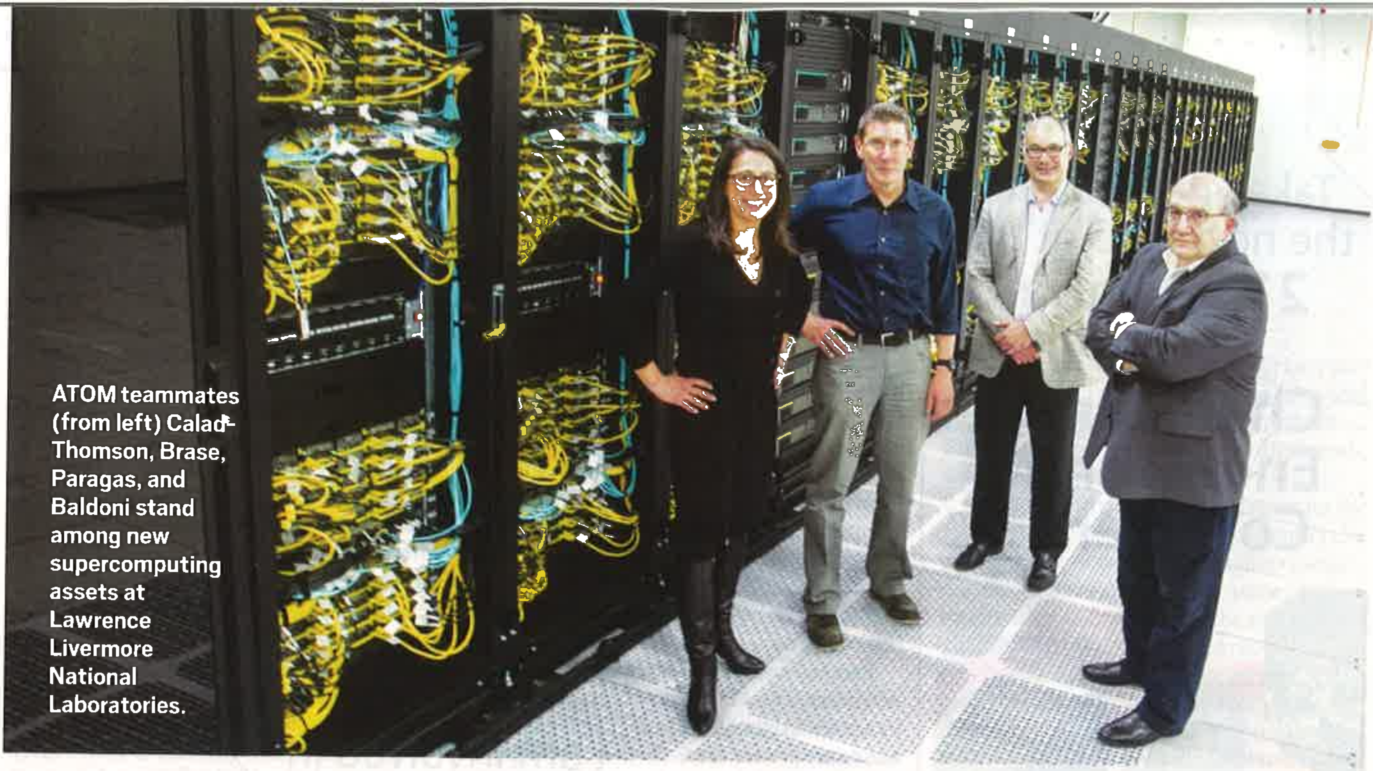
Stanford chemistry professor Vijay Pande's group is collaborating with Google, using the company's massive neural network system to search for drug candidates. The group has been training computers to model molecules that bind to and thwart the action of the enzyme β-secretase 1 (BACE1), which is believed to help produce the hallmark brain plaques of Alzheimer's disease (*J. Chem. Inf. Model.* 2016, DOI: 10.1021/acs.jcim.6b00290).

Deep-learning research is already popping up in computational chemistry journals. For instance, Insilico Medicine chemist Alexander Aliper and coworkers used the technique to predict drug pharmacological properties and whether approved compounds could be repurposed (*Mol. Pharmaceuticals* 2016, DOI: 10.1021/acs.molpharmaceut.6b00248). And a group led by Cicero Nogueira dos Santos at IBM Watson is working on the use of deep learning to improve the success of virtually screening libraries of molecules that dock strongly to their targets (*J. Chem. Inf. Model.* 2016, DOI: 10.1021/acs.jcim.6b00355).

Drug development still faces more hurdles with deep learning than image or voice recognition. Large compound libraries strain to provide the huge amounts of data required for deep learning to be effective. “The reality is that chemistry lags in the amount of data available,” the University of North Carolina's Isayev says. “Only recently have we begun to see big data sets.”

Some hope that the pharmaceutical industry, which is rabidly protective of its data, might find ways to share data sets (see page 31). There's much talk about the need to incorporate more negative results—molecules that failed to pan out as drug candidates—to provide more robust training sets. Pharma scientists and academicians alike have troves of negative results that don't get published.

Isayev believes a key development that will boost deep learning will be to find ways to feed representations of entire molecules into a computer, rather than submitting human-crafted features. The machines will then do a much better job of discovering significant combinations of properties. “There's no way to do that yet,” he says. “But people are actively working on it. It's like the Wild West.” ■



ATOM teammates (from left) Calad-Thomson, Brase, Paragas, and Baldoni stand among new supercomputing assets at Lawrence Livermore National Laboratories.

CREDIT: JULIE RUSSELL/LAWRENCE LIVERMORE NATIONAL LABORATORIES

PHARMACEUTICALS

Partnership applies deep learning to very big data

GlaxoSmithKline teams with government labs on computational models for drug discovery

RICK MULLIN, C&EN NEW YORK CITY

GlaxoSmithKline has for several years engaged in a campaign of creative destruction, acting to replace traditional methods of manufacturing and research with more efficient ways of doing things. The implementation of programs such as its Manufacturing Technology Roadmap, a manifesto against the standard pharma method of manufacturing drugs in batches, for example, has helped earn the company the reputation of risk-taker among major drug firms.

Now, GSK is turning its attention to drug discovery, where it is hoping that a nascent approach of applying supercomputing to huge data sets will allow it to move from target identification to a molecule ready for the clinic in just one year.

John Baldoni, senior vice president of platform technology and science at GSK, refers to the drastic telescoping of a process that has traditionally taken closer to a decade as the company's "moon shot" project. It will rely heavily, he says, on the replacement of the iterative chemistry process of identifying and testing molecules, with a networking approach that finds and tests huge numbers of molecules simultaneously, using gigantic stores of data. The technique GSK is exploring, deep learning, exists at the cutting edge of artificial intelligence research in which computers are developing data models that evolve, or "learn," through computational experience (see page 29).

Facets of the new technique have already been employed elsewhere in industrial research, such as in the development of self-driving cars. But it has not, according to Baldoni, been explored by any major drug company in a large-scale effort to manage the explosion of data in drug research that followed the decoding of the genome and the rise of cloud computing.

"Having spent a year looking at statistics around the drug discovery process," Baldoni says, "we began asking whether it's time to rethink how we do things." Early-stage discovery methods developed with the rise of compound libraries and high-throughput screening in the early 2000s are time-consuming, predicated on huge investments, and currently standard in the industry, he says. Meanwhile, the process of drug discovery was bogging down in data.

"A few people in my department had been talking to folks at the Department of Energy about how we might be able to take advantage of high-performance computing to replace some of the empirical work that we do in the drug discovery process," Baldoni says. Those discussions led to a partnership between GSK, DOE labs headed by Lawrence Livermore National Laboratories (LLNL), and the National Cancer Institute (NCI).

The partnership, called Accelerating Therapies for Opportunities in Medicine (ATOM), launched earlier this month, aims to develop computing models that will guide researchers based on a computer's ability to quickly vet millions of molecules for efficacy and structural relationships, models that will adapt as they are applied to new data. ATOM will begin by putting DOE's supercomputing powerhouse to work on data from GSK and NCI, taking advantage of the drug company's expertise in chemistry and biology as a framework in pioneering applications of deep learning for drug research.

The partners, having collaborated on defining ATOM's mission over the past year, are currently deciding where to locate

their central laboratory. The group is also acting to expand its membership. Baldoni says it seeks to recruit other major drug companies willing to contribute significant quantities of research data to its computer modeling program.

"What was really driving the conversation was changes in supercomputing," says Jason Paragas, director of innovation at LLNL, who was involved in early meetings with GSK. The advent of cloud computing, in which companies such as Amazon and Google were hosting huge volumes of data from many sources, and advances in supercomputing set the table for researchers to apply computational learning systems to the highly complex field of drug discovery, drawing from heretofore unapproachable banks of data.

Paragas says new supercomputers that will be used by ATOM have been purchased by a consortium of national research labs, including Oak Ridge and Argonne, both of which will provide systems and research backup to the partnership.

> "I am involved in an initiative to get companies to share their data, and I have to tell you it's extremely frustrating."
>
> **—John Baldoni,** senior vice president of platform technology and science, GlaxoSmithKline

Jim Brase, associate director of computation at LLNL, says ATOM will be taking a deep dive into big data. "We have large amounts of experimental data—genomic data, transcriptome data, assay data—on how biological systems respond to chemicals and their structures. We are seeking to understand from large data sets what particular combinations of things and standards in those data sets are important to building predictive models."

Brase says the variety of deep learning ATOM might be most interested in is unstructured or unsupervised feature learning, where the focus is on early-stage identification of data sets that go together and significant patterns without predetermined parameters or expectations. The work is comparable to what Google has accomplished with face recognition, he says, but the data sets are much larger and far more complex.

Baldoni says GSK has agreed to contribute information on 500 failed compounds, including complete toxicology and clinical testing data, in addition to 600,000 advanced compounds in screening at the company. In all, GSK will give ATOM access to more than a million compounds screened over the past 15 years, all of which have biological data associated with them.

But that isn't enough. The partnership, Baldoni says, will need to recruit other large pharmaceutical companies willing to pony up comparable stores of data. He adds that information on failed compounds will be key to achieving breakthroughs in drug discovery.

The problem is that big drug companies are not forthcoming with the data or any information on failed discovery projects. "I am involved in an initiative to get companies to share their data, and I have to tell you it's extremely frustrating," he says.

"There are hundreds of thousands of failed molecules or molecules no longer of interest with information on structures, analogs, toxicity, and structure-function relationships," Baldoni says. "Why would you not want to put them into the greater good? We feel there is an obligation that we have to patients in trials to share these failed compounds so we can develop better drugs faster."

But the partnership is hopeful that industry will come around and contribute data. Baldoni says the group is in discussion with various research entities and hopes to announce the addition of another large drug company in the coming weeks.

There has been little pushback on the changes to chemistry in research at GSK as the result of bringing in heavy computational firepower in early discovery. On the contrary, says Stacie Calad-Thomson, a GSK chemist who is coordinating the activities of ATOM laboratories and key liaison with LLNL and NCI, there is a great deal of enthusiasm in the lab.

"I think it will have incredible impact on chemistry, allowing us to do research in a more rapid and agile fashion," she says. "Everyone is very excited about that."Calad-Thomson also notes that the company has already seen the benefit accrued in tearing up traditional procedures in manufacturing. "Now we're doing it in the discovery space at the forefront of innovation."

Baldoni agrees, adding that the company needs to prepare for what's coming. "There is a recognition that the state-of-the-art computers in the national laboratories will soon be at pharma companies. We might as well start building the tools to use them. We'll get ahead by a few years, and that is critical." ■

COMPUTATIONAL CHEMISTRY

Getting the most out of chemistry data with machine learning

Publishers and others apply standard artificial intelligence techniques to synthesis planning and education

JYLLIAN KEMSLEY, C&EN WEST COAST

Although chemists are excited by the potential of so-called deep-learning computational tools to make a splash in drug discovery, publishers and others are still looking to squeeze findings out of earlier, less sophisticated versions of these tools. With machine-learning techniques that "teach" themselves with large data sets, they hope to get more out of scientific information, whether in the lab or the classroom.

"It's about how you make discoveries consumable," says Conal Thompson, chief technology officer for CAS, a division of the American Chemical Society that's looking into how to get more out of its chemistry databases. "What's going to become more valuable is insight from your data or content, rather than just the content itself." ACS publishes C&EN.

One emerging area of chemistry that is capitalizing on machine learning is computer-aided synthesis design: feeding software a target molecule and getting back possible routes chemists might use to make it.

"Eight years ago, there was a lot of skepticism and resistance of chemists to the whole notion" of artificial intelligence being used to solve chemical problems, says Orr Ravitz, product manager for Wiley's ChemPlanner, one platform offering help with synthesis design. "I think a lot has changed since then, and I think that's related to us using so many computational tools in our daily life. People are starting to expect it." Also, the falling cost of computing power has made chemistry applications faster and less expensive.

But machine learning does not help in every situation. For example, the basic reaction rules underlying ChemPlanner and a similar program developed by a start-up company, Chematica, do not come from computers automatically extracting information from journals or patents. Instead, humans are extensively involved to identify key reactions and write the reaction rules on which the programs run. This is in part because artificial intelligence programs learn best when they train on hundreds of examples.

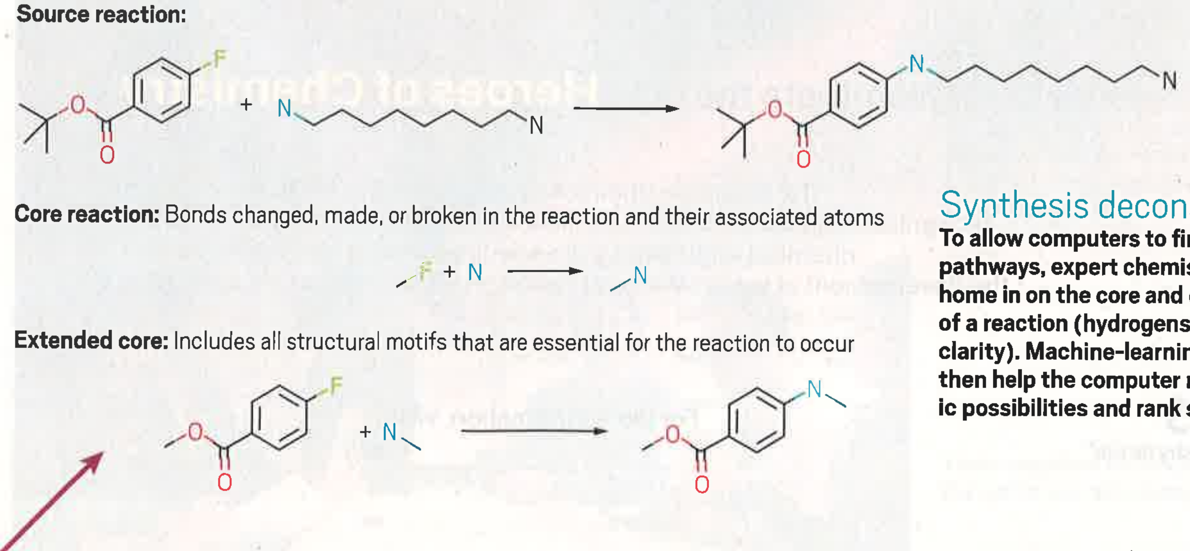
If chemists just want to use very common, well-established reactions, then machine learning likely could extract them from the literature, says Bartosz A. Grzybowski, developer of Chematica and a chemistry professor at Ulsan National Institute of Science & Technology and at the Polish Academy of Sciences. "But for complex synthetic planning, very rare reactions can be very important. A reaction that might appear in the literature only three times may be key to making a natural product," Grzybowski adds.

Consequently, expert chemists write the rules that allow the software to identify the core of a reaction—the bonds that change during the reaction and their associated atoms. Chemists also write the rules that dictate when and how to incorporate other components of reagent structures that may influence reactivity, such as aromaticity or electron-donating or -withdrawing groups.

The software then uses those rules to identify possible reactions based on whether the chemical structures of those "extended cores" share similar properties. Machine learning comes in for navigating the options among a huge network of synthetic possibilities.

Machine-learning algorithms also play a role in scoring synthetic pathways to prioritize the order in which they're shown to the user. Scoring is not a one-size-fits-all process, the software developers have found. Different chemists differently prioritize things such as cost, yield, number of steps, or use of protecting groups. "What we hope to do with machine learning in the future is to basically learn from the user's interaction with the system and try to tailor prioritization to their taste, similar to what Netflix does based on your viewing history," Ravitz says.

Researchers are also actively applying machine learning to materials science.



Source reaction:

Core reaction: Bonds changed, made, or broken in the reaction and their associated atoms

Extended core: Includes all structural motifs that are essential for the reaction to occur

Synthesis deconstruction

To allow computers to find synthetic pathways, expert chemists write rules to home in on the core and extended core of a reaction (hydrogens omitted for clarity). Machine-learning algorithms can then help the computer navigate synthetic possibilities and rank solutions.

CREDIT: CHEMATICA

Northwestern University professor Chris Wolverton and colleagues recently published a general framework for using machine-learning approaches to predict properties of inorganic materials (*npj Comput. Mater.* 2016, DOI: 10.1038/npjcompumats.2016.28). Separately, a team led by Sorelle A. Friedler, Joshua Schrier, and Alexander J. Norquist of Haverford College used machine-learning models to predict conditions for successful crystallization of inorganic-organic hybrid materials (*Nature* 2016, DOI: 10.1038/nature17439).

Grzybowski (left) and Karol Molga explore synthetic pathways using Chematica.

CREDIT: BARTOSZ A. GRZYBOWSKI

Notably, the Haverford group says in its paper that the researchers used information on "dark" reactions—failed or unsuccessful syntheses—collected from their archived laboratory notebooks to help train their machine-learning model, and they have a "Dark Reactions Project" website set up to gather similar information at darkreactions.haverford.edu. Such "dark" data will become increasingly important as people look to develop machine-learning applications, experts say.

"Nobody likes publishing negative results, but a machine and its intelligence would be much more informed by having positives and negatives," CAS's Thompson says. "It's not a mistake anymore, it's valuable information." However, making that valuable information accessible is an unsolved problem in a scientific culture that prizes positive findings and largely ignores so-called negative results in its publications.

Other areas that may benefit from machine learning include scientific education, where algorithms can potentially improve student learning outcomes. "One of our divisions creates educational materials for nurses, but many of our students get frustrated with the challenging material, drop out of the course, and never take their certification exam," Dan Olley, chief technology officer at Elsevier, told *CIO* magazine last year. "We are using algorithms that learn how students actually use the course material," he continued. "This way, we can create adaptability and personalization within the course to engage the students and drive better pass rates."

Where machine learning will take scientific learning and research in the future remains to be seen. But just as computing technology has changed daily life to incorporate activities previously only seen on "Star Trek"—"Alexa, lower the temperature to 68 degrees"—it has the potential to allow the scientific enterprise to do things researchers previously only dreamed about. ■