# DISCOVERY REPORT

An ACS member exclusive

ACS Chemistry for Life®

# AI for materials discovery

Computers can spot data patterns that humans cannot. How will chemists guide the machines?

Published by c&en

$39.99

# The future of AI for materials discovery

In the popular imagination, chemistry is an experimental science, a world filled with flasks and bubbling liquids. So what to make of news reports from 2019 about researchers using artificial intelligence (AI) to transform a brittle polymer into a supercompressible material without having to tinker in the lab before fabrication? It's enough to trigger an identity crisis.

The researchers making strides in AI and machine learning don't see such news as problematic. Discovering new compounds and materials continues to be a time-consuming, trial-and-error process based on researchers' experiences. As a result, scientists have amassed staggering amounts of data. They produce more each day—a body of knowledge even the most expert group of humans would find hard to process.

Advances in computing power and algorithms are making it possible to process decades' worth of literature, terabytes of instrument outputs, and reams of lab notebooks and use it all to offer predictions about materials' properties or get labs to run more efficiently. It's about freeing chemists' minds to think about high-level problems. Inside this Discovery Report, you'll meet specialists using AI to make the production of concrete less energy intensive, keep a lab running remotely during the COVID-19 pandemic, and more.

Contributing editor Carmen Drahl, who has covered organic chemistry and green chemistry for C&EN, edited this report. It includes a reading list of papers and patents curated by our sources, as well as by information scientists at the CAS division of the American Chemical Society.

As an ACS member, you get exclusive access to the Discovery Report, a quarterly publication bringing you cutting-edge research defining the chemical sciences and our industry. Look for the next one in the third quarter of 2021.

Amanda Yarnell

Amanda Yarnell
Editorial director, C&EN
@amandayarnell

# INSIDE

Published by
c&en

**Editors:** Carmen Drahl, Jyllian Kemsley

**Contributing writers:** Sam Lemonick, Rick Mullin, Prachi Patel

**Creative:** Robin Braverman, Rob Bryson, Chris Gash, Yang H. Ku

**Copyeditor:** Michele Arboit

**Production:** Manny I. Fox Morone, Marsha-Ann Watson

**Comments or questions?** Email cenfeedback@acs.org

CHEAT SHEET

# 5 questions and answers about AI in materials discovery

Q.

## What are artificial intelligence, machine learning, and deep learning?

» **Artificial intelligence (AI)** is the science of making machines that assess their environment and solve problems like humans do. AI is in technology that finishes words as we type or recommends products based on past purchases.

» **Machine learning is a subconcept of AI.** Machine-learning algorithms identify patterns in data, build models, and make predictions based on those patterns. They reprogram themselves as they incorporate more data.

» **Deep learning is a subconcept of machine learning.** It detects patterns and makes predictions with little human intervention. This technology comprises multilayered neural networks, which are algorithms that simulate the brain but fail to replicate it.

Q.

## How is AI being used in materials discovery and development?

» **AI is broadly applicable** to polymers, semiconductors, perovskites, catalysts, and any other class of materials with a body of experimental data for training algorithms.

» **Materials with enhanced performance** discovered with AI, such as optical lenses or displays, are in development.

» **More environmentally friendly versions** of materials such as touch sensors have been discovered with AI.

» **AI can optimize or automate** production of new materials by incorporating robotics.

» **Data from disparate sources** can be formatted and organized with AI.

Q.

## Why are researchers turning to AI?

» **Computers spot patterns more quickly** than people do, so AI can predict properties of a material with fewer experiments for background knowledge.

» **Lab and plant managers want to cut costs,** and they hope AI might cut inefficiencies at even well-oiled operations.

» **Algorithms can consolidate and triage data** to find useful experiments that haven't been capitalized on and to discard anomalous results that could be misleading.

» **The COVID-19 pandemic** made remote work common, and AI-driven robotics is one solution to maintaining labs whenever social distancing is required.

Q.

## What are the challenges of using AI in materials discovery?

» **AI relies on existing data,** so the quality of its output will match the quality of the data. Algorithms could lead to inaccurate predictions or steer researchers toward predetermined outcomes due to confirmation bias. As the saying goes: garbage in, garbage out.

» **AI costs money to implement,** and it's unclear when investments will pay off.

» **AI cannot replace humans** but can automate repetitive tasks, which means that many people performing those tasks could lose their jobs and will need to train for other types of work.

Q.

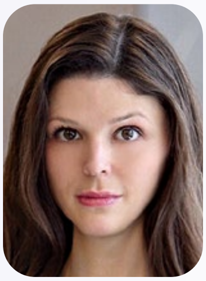
## What's next for AI in materials discovery?

» **Researchers are comparing the performance** of different algorithms for discovering materials to see which might be broadly applicable or which might be most useful for specific tasks (see page 16).

» **The 100% automated lab doesn't yet exist,** but researchers are expanding the capabilities of AI-driven robots to perform ever-longer experimental sequences (see page 8).

» **Gleaning insights faster with fewer data** is a long-term goal for developing next-generation algorithms because experimental data are sparse in some fields.

# 8 experts weigh in on the future for AI in materials



## Jill Becker

» Cofounder and CEO, Kebotix

Jill Becker is excited to spearhead the use of artificial intelligence (AI) to solve some of society's most pressing problems. Kebotix's AI-powered robotic lab is already cutting the time and cost of discovering materials that could impact agriculture, energy, medicine, and the environment.

"No other company has a self-driving lab," Becker says, referring to the closed-loop system between Kebotix's algorithms, which design new molecular combinations likely to give desired properties, and the robotic arms that make compounds, test them, and report back to the computer. This feedback loop means the system continually improves.

In a pilot project with the National Institutes of Health, the approach made high-throughput experiments five times faster and reduced costs by 80% by cutting run time and the amount of lab supplies. Kebotix is now partnering with Bayer to discover chemicals for crop protection and with Koura to develop ecofriendly high-performance materials.

The Boston start-up generates its own data for its AI models to tap into, including data from unsuccessful experiments that can still provide valuable insights.

Internally, Kebotix has its eyes set on developing organic electronic materials, such as high-efficiency and long-lasting pigments for organic light-emitting diodes. "The hope is that in the future AI could say, 'OK, this chemical lasts 3 years in a device; this one could last 30,'" Becker says. "It takes the serendipity out of science and gives you a smart way to get the material you want."

> "The biggest limitation of AI is it's an expert system within its own domain but doesn't know anything outside of that domain."

## Andrew Cooper

» Director, Materials Innovation Factory, University of Liverpool, and CEO, Mobotics

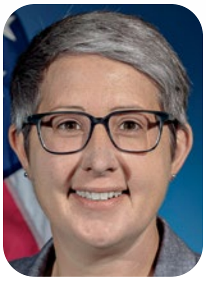
Andrew Cooper's team has built a robot chemist that gives a glimpse into the autonomous lab of the future. The one-armed robot moves independently, running reactions using equipment designed for humans. Its AI brain navigates 98 million experimental possibilities, based on variable concentrations of 10 reagents, to tell the robot which reaction to run next.

The bot can work 24/7 and do hundreds of experiments a day. But Cooper believes that AI's virtue should extend beyond speed. Brute-force searching of an experimental design space becomes limited with thousands of variables giving billions of possible test conditions. The goal should be to direct the search using a human chemist's knowledge.

"Human approaches are smart and slow," Cooper says. "Robots are fast and somewhat dumb. The big thing is for those two to converge and for us to eventually have something smart and fast. Right now with robots, it's almost like a very hardworking and intelligent baby being let loose in the lab."

What's needed to make robots smarter are AI systems that can read scientific literature and use that preexisting knowledge, Cooper says. But even then, AI trained on existing data will be unable to find a totally new idea for a material. "The biggest limitation of AI is it's an expert system within its own domain but doesn't know anything outside of that domain," he says.

CREDIT: ALL IMAGES COURTESY OF INSTITUTIONS



## Anne Fischer

**» Program manager, Defense Advanced Research Projects Agency**

"AI in chemistry is really about data in chemistry," Anne Fischer says. "The currency of AI is data." The shortfalls of that currency quickly became apparent to her in 2016, when Fischer launched the Defense Advanced Research Projects Agency's (DARPA's) Make-It program to accelerate the discovery of small molecules for pharmaceuticals, fuels, and explosives.

Published data often omit relevant experimental information or are incomplete, which can bias AI models, Fischer says. In addition, most data in scientific literature are inscrutable to computers. To extract information from published text and figures, two institutions participating in Make-It—the Massachusetts Institute of Technology and SRI International—found innovative workarounds that automate the human process of comprehending the literature, including natural language processing and computer vision algorithms. The institutions have developed automated chemical synthesis platforms that include software to design chemical routes and robotic hardware to do the syntheses.

Feeding test data from the robots back to the AI model would expand the capacity of these automated labs. So in 2019, Fischer started DARPA's Accelerated Molecular Discovery program to enable this type of closed-loop discovery. Through the program, a University of Toronto team is developing AI that can design dye molecules with desired spectral ranges, and SRI is making a model to predict antiviral compounds for diseases like COVID-19.

"We're not there yet," Fischer says of the optimal closed loop. "We'll get better and faster at these closed-loop cycles [and] have models that can predict molecules with properties applicable across many applications."

## Aaron Gilad Kusne

**» Research scientist, National Institute of Standards and Technology**

Aaron Gilad Kusne believes that AI and automation have freed materials research from the bounds of human limitations. For decades, materials scientists have had to stick to simple processing steps and compositions. "Now we have the capability to make more complex materials," he says.

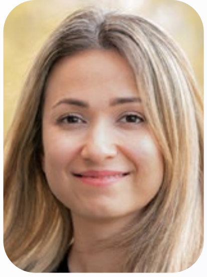
Even for substances like superconductors, for which fundamental understanding is lacking, AI can help provide a framework to optimize properties.

Kusne focuses on active learning, a type of machine learning that puts humans in the loop by periodically prompting them to provide labels—meaningful context that helps the AI learn from data. Insufficient or unlabeled data are rife in materials science, Kusne says, and active learning helps overcome those issues. "By building physical theory and human intuition into the AI, you can accelerate the discovery of better materials," he says.

He has created an algorithm, a closed-loop autonomous system for materials exploration and optimization (CAMEO), which helps scientists zero in on an ideal material using fewer experiments. It's especially helpful to reduce tinkering time when using expensive facilities such as particle accelerators or synchrotrons for analysis.

Kusne's team reported in November 2020 that CAMEO reduced by 90% the time taken for neutron-scattering experiments to determine temperature-dependent magnetic properties. The team also found a phase-change material, a germanium-antimony-tellurium alloy, that hits lab benchmarks better than similar alloys used today in electronic data storage. They are now trying to patent that alloy. With CAMEO, Kusne says, "you can do the least number of experiments to have maximal impact."

"Material discovery is like finding a needle in a haystack. But it's useless to look in a haystack that has no needles."



## Nastaran Meftahi

**» Research fellow, RMIT University**

Nastaran Meftahi thinks that AI's biggest impact is in reducing the time, energy, and other resources needed to discover materials that haven't been synthesized before. Applying machine learning to predictions about the properties of materials developed in silico "will help scientists develop new materials by only testing the best candidates out of potentially millions of candidates," she says.

But scientists face the hurdle of finding large, reputable, and—ideally—publicly available data sets needed to train machine-learning models, Meftahi says. "Otherwise it's garbage in, garbage out."

Consequently, Meftahi collaborates directly with experimentalists around the world. Using their data as input, she harnesses her chemistry background and experience in computational technique to develop machine-learning models that can determine relationships between a material's structure and its properties.

She recently showed that machine learning can use chemical fragments of materials to predict complicated photovoltaic properties of organic compounds. Her algorithm is computationally less intensive, and hence faster, than supercomputer-based methods researchers had used before, while giving predictions that are more accurate.

Reputable data is only one facet of a healthy AI ecosystem for materials discovery. The literature is full of machine-learning models, but using them requires understanding how they work and writing the code, Meftahi says. That's an accessibility barrier for materials scientists who lack coding expertise. So she made her algorithm and data sets available for free download at GitHub, a cloud-based software repository. "Anybody, with a few clicks, can access my models to make better photovoltaic materials," Meftahi says.

## Bryce Meredig

**» Cofounder and chief science officer, Citrine Informatics**

Discovering new materials might be rewarding, but a lot more must be done to get those materials into the real world. "Discovery is one step in a longer process that involves scale-up, manufacturing, and commercialization," Bryce Meredig says.

The cloud-based AI platform of his Redwood City, California, start-up is enabling companies to accelerate development across all those stages. Customers include BASF, Lanxess, and Panasonic.

AI's strength is its ability to parse enormous amounts of data to make decisions that maximize the probability of success, Meredig says. Citrine's machine-learning models are built on physical and chemical laws that are the foundation of materials research. The models train on the company's vast databases, which include both public and in-house data, to help researchers make optimal decisions.

The platform lets users customize the methods and AI process for their product line. The software's graphical interface is intuitive for scientists of all stripes, Meredig says, so they can inject their own expertise—whether in the form of physics equations or structure-property relationships.

Users can specify reasonable chemical motifs to construct new candidate materials to investigate. "This is important, because material discovery is like finding a needle in a haystack," Meredig says. "But it's useless to look in a haystack that has no needles."

## Nicola Pohl

**» Associate dean for natural and mathematical sciences and research, Indiana University**

Nicola Pohl, who is left handed, says working with equipment designed for right-handed people requires extra thought. Chemistry labs are not set up for people with differing abilities. "Most things we do require dexterity that not everyone has," she says. "Even bench height is standard."

Automated lab systems could help level the playing field. In addition to improving accessibility, Pohl says, they would facilitate reliable digital records, aiding the creation of large, rich data sets that can be mined for machine learning. She is on a mission to break down barriers to widespread adoption of automation.

For starters, the next generation of scientists must team with AI effortlessly. Pohl's group develops automated synthesis machines to make and analyze carbohydrates. Her students aim to establish reproducible procedures and generate data that machines can read. Students also learn to use robots to carry out experiments. Pohl sits on the Chemical Science Roundtable of the National Academies of Sciences, Engineering, and Medicine, which seeks to incorporate automation into chemistry and materials science curricula and is planning a lab automation workshop for November 2021.

Second, companies are improving machine interface design, which should help make the machines more intuitive for researchers to use. The machines are only useful if you can have smart people using them, she says.

Finally, the international community must agree on standards for data. "So much data we collect depends on the specific instrument," Pohl says. "Someone making compounds here in Bloomington should have the same standards as someone in Germany."

## Mardochee Reveil

**» Senior research scientist, Corning**

Machine learning excels at leveraging data to predict the performance of materials before you make and test them, Mardochee Reveil says. This is vital for glass, whose properties are very hard to predict using traditional methods—despite glass being one of the first materials humans engineered.

AI speeds the screening of potential candidates for specific applications and makes it easier to generate high-quality glass candidates, whether for vials to store vaccines or for next-generation displays. "This saves time and money, and perhaps more importantly, reduces risk to make bolder exploration feasible," he says.

At Corning, Reveil creates predictive AI tools and machine-learning models to help design new glass compositions for various products. He also creates similar tools to devise novel organic materials for applications such as cameras used in manufacturing and self-driving cars.

If AI could get beyond "empirical predictions of property values to [playing] a more central role in the actual design of new candidate materials," it could have a greater impact than it already does, Reveil says. He likens this next-level role to movie-streaming services that make relevant recommendations based on users' viewing history and preferences.

For now, the dearth of data in materials science is a serious problem, Reveil says. He is leading an effort at Corning to shape how the company manages R&D data throughout the life cycle of projects in its research portfolio. "This will help us capture more and better data that we can then leverage to innovate faster using advanced machine-learning techniques," he says. ■

CREDIT: ALL IMAGES COURTESY OF INSTITUTIONS

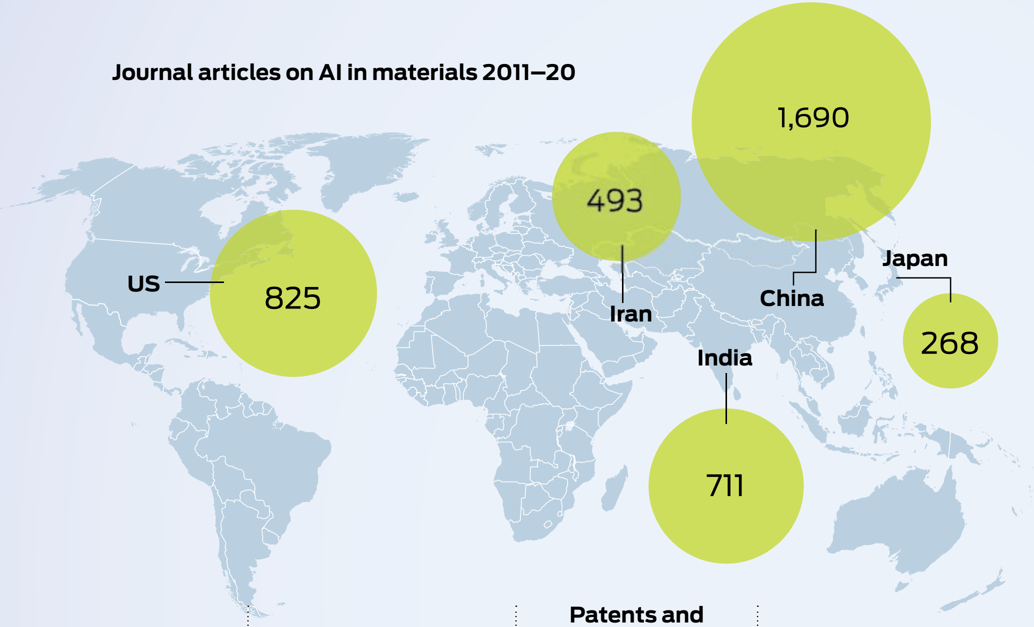
BY THE NUMBERS

# Understand trends in AI for materials

## Who's who

**The countries that publish and patent the most across all fields also lead in AI and materials. The exception is Iran, which ranks higher for publishing in AI and materials than it does for all fields.**

**Journal articles on AI in materials 2011–20**

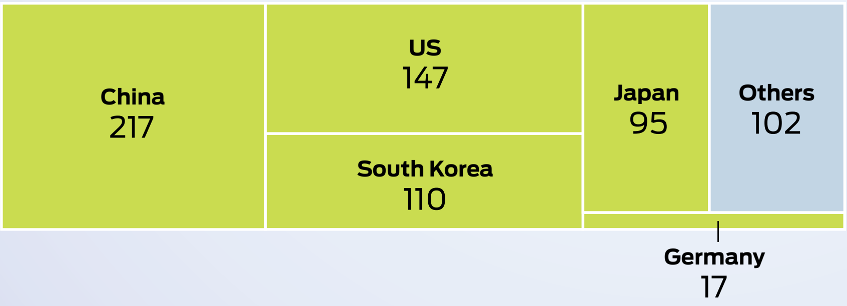


| Country | Rank by all scientific journal articles published, 2018 | Patents and applications on AI in materials 2011–20 | Rank by all patents and applications published, 2018 |
|---|---|---|---|
| **China** | 1 | 217 | 1 |
| **US** | 2 | 147 | 2 |
| **India** | 3 | 16 | 8 |
| **Iran** | 15 | 0 | does not rank |
| **Japan** | 5 | 95 | 3 |

## Concentration factor

**Five countries account for 85% of patents and patent applications in AI in materials.**

**Patents and applications on AI in materials 2011–20**



## AI stats

**Boost your intelligence with our selection of facts and figures.**

**15.8%**

Percentage of AI in materials patents held by the top 10 corporations with patents in this area, 2011–20

**1950**

Year that Alan Turing published his test of a machine's ability to exihibit intelligence

**40**

Number of US faculty from all AI disciplines who left academia for industry in 2018

**$15.7 trillion**

Potential contribution to the global economy by 2030 from all AI applications

**54%**

Percentage of executives who worry about making bad decisions based on AI recommendations

**$1.1 billion**

Amount raised by our 15 Companies to Watch (see page 13) across all funding rounds

**22%**

Percentage of AI professionals globally who are female

**Sources:** CAS (a division of the American Chemical Society); Crunchbase, Deloitte, MIND, PwC, Stanford Institute for Human-Centered Artificial Intelligence, World Bank, World Economic Forum, World Intellectual Property Organization.
**Notes:** CAS information scientists searched patents and publications containing the concepts of AI and materials from 2011 to 2020.

ANALYSIS

# THE LAB OF THE FUTURE IS NOW

**RICK MULLIN,** C&EN STAFF

The laboratory of the future is a visionary concept: a goal that evolves as technology advances. Predictions are made, skeptics respond, and the vision reboots. The future offers few guarantees. But it also tends to develop over time rather than spring up, shockingly new, overnight.

The lab of the future, Platonic ideal that it remains, has always had a foot in the door. But some say it crossed the threshold to reality entirely last year.

High-tech labs unveiled by academic researchers, a computing giant, and a major drug company all blend artificial intelligence (AI) computing and robotics in ways that may herald a new world of research. They suggest a kind of Renaissance lab for multidisciplinary scientists steeped in chemistry, biology, and data science.

Alán Aspuru-Guzik, a professor of chemistry and computer science at the University of Toronto, and colleagues reported in *Science Advances* the discovery of thin-film materials in a "self-driving laboratory" in which AI controls automated synthesis and validation in a cycle of machine-learning data analysis. Andrew I. Cooper, director of the Materials Innovation Factory at the University of Liverpool, and colleagues published results from an AI-directed robotics lab that optimized a photocatalytic process for generating hydrogen from water after running about 700 experiments in 8 days (*Nature* 2020, DOI: 10.1038/s41586-020-2442-2).

> "'Self-guided' in this area means within boundaries set by human scientists."

Meanwhile, IBM launched a self-driving—or autonomous—lab combining AI with robotics at its research facility in Zurich. And Eli Lilly and Company rang in 2020 with the debut of a self-driving lab at its biotechnology center in San Diego.

"My lab has already been able to close the loop," Aspuru-Guzik says, describing a circuit of continuous learning in which AI algorithms guide data analysis and automation toward the identification, synthesis, and validation of novel molecules. The autonomous lab accesses, produces, and reprocesses data as it goes along (*Science Advances* 2020, DOI: 10.1126/sciadv.aaz8867).

These labs aren't perfect yet, and much work and convincing are needed before research managers in the drug and chemical industries embrace them. Observations vary on where the bottlenecks lie in the clocklike loops—some point to the data and others to the robots. But the innovators are unanimous regarding the importance of the third element in the loop: the human researcher.

"This idea of the clockwork laboratory is, in the long term, not the strongest approach," Cooper says of the notion of a self-sufficient research machine. "The strongest approach is to have the clockwork laboratory with a very permeable interface so that the human knowledge can be captured as well."

## The academics

AI most likely debuted in science fiction with Samuel Butler's 1872 novel *Erehwon, or Over the Range*. It made news in the real world a quarter century ago when IBM's Deep Blue supercom-

CREDIT: ADAM COISH

**Alán Aspuru-Guzik, a professor of chemistry and computer science at the University of Toronto, stands in front of his "self-driven" laboratory. It and the rest of the facility were not operating because of the pandemic.**

puter took down Garry Kasparov, the world chess champion. Kasparov came to terms with his defeat, much as the research community has learned to stop worrying and love a machine that accelerates discovery.

"You ask me why I'm doing this; it's because the world has no time," Aspuru-Guzik says. He points to rapid design techniques in industries such as automotive that rely on advanced materials and the urgency to confront climate change with new materials for storing sun and wind energy. "We have to enter the era of rapid prototyping in materials."

Researchers say the challenge in accelerating discovery comes down to improving their data sets. That's a primary function of machine learning in an autonomous lab as it cycles and directs data from synthesis and validation, melding them with data from available published literature.

Assembling and fine-tuning a system in Toronto to demonstrate the power of closed-loop autonomous discovery took about a year and a half, Aspuru-Guzik says. Once operational, the machine was able to produce about 40 molecules in a production run, he says, "which is more than the number of published molecules in the field I am working on, organic semiconductors for laser devices."

But while the system is intended to support an around-the-clock process, it is still operating in discrete runs, each averaging a day and a half. "The biggest holdup is not AI," Aspuru-Guzik says. "Data bottlenecks? Zero. The area that is the challenge for getting the self-driving lab to work is the synthesis machinery. The robotics are a little finicky, a little hard to control."

Those robots are the source of the synthesis and validation data necessary for machine-learning algorithms to drive toward discovery. "The point is to generate data on the fly," he says. "That is what you're getting a robot for."

The University of Liverpool applied a twist to robotics in its autonomous lab. "It's ironic that Alán calls his a self-driven laboratory," Cooper quips. "Our robots actually drive." Indeed, robotic agents scoot around a traditional-looking research space in a video released by Cooper's group. "We decided to automate the chemist, not the instrument," he says.

Cooper's catalyst experiment operated nonstop for 8 days, completing about 6,500 manipulations in a complex workflow involving solid and liquid handling, some under a nitrogen-sealed, inert atmosphere, and multiple measurements. The mobile robots scurried a total of 2.2 km in a room measuring 5 by 10 m. Cooper says the system can theoretically operate for much longer. "We used 8 different experimental stations," he says. "It could have been 18 or 80."

As for bottlenecks, Cooper also points to robot mechanics. "AI is most powerful when there are multiple choices, a wide variety of measurements," he says. But the more involved the experiment, the more complex the machinery, "and every bit needs to be really reliable. Once you string 10 operations together, even a failure point of 0.1% becomes quite significant."

Regina Barzilay, a computer science professor at the Massachusetts Institute of Technology and colead of the Machine Learning for Pharmaceutical Discovery and Synthesis Consortium, notes that AI can run into problems as it operates in new areas of chemical space. But the technology works to solve the problems, partly through interaction with a researcher.

"You need to have a mechanism whereby the machine can tell you it needs to look in a particular part of chemical space where it needs to have more training," Barzilay says. "An essential component is to know when the machine is actually not confident in its own prediction, which means it needs more help."

Connor W. Coley, a professor of chemical engineering at MIT and part of the consortium, says advances made in pursuit of the autonomous laboratory reveal the critical importance of the human element. Even if the lab is a closed loop, it intersects with an adjacent loop defined by researcher interaction with the machine, starting from the word "Go."

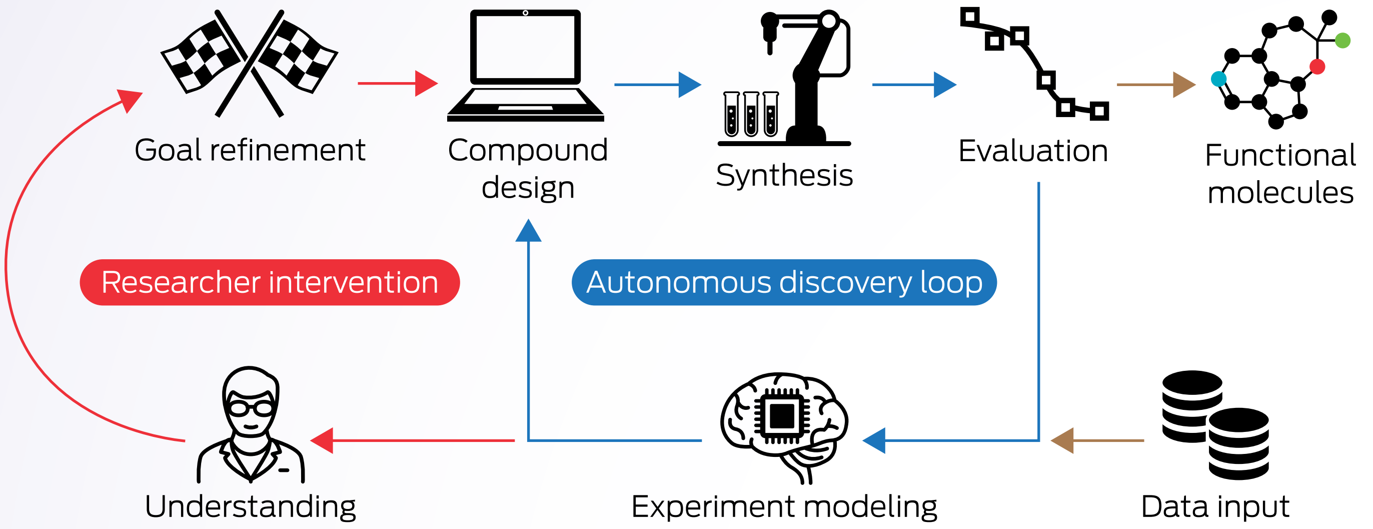
"Humans are always going to set the design objective and specify something that an algorithm can reduce to a numerical optimization," Coley says. "Humans will always be setting the big-picture goal" through rounds of interrogation.

"The big questions include which experiments are accessible to the platform compared to which are needed to prove or disprove the kinds of hypotheses we're after," he continues. "Another big question is which are the workflows that we can physically automate as steps in the process."

And there are considerations beyond academic labs like Coley's. "The interesting question is

## Looped intelligence

**An autonomous chemistry laboratory runs experimental cycles intended to yield useful molecules. In the cycle, AI models the experiment and designs a compound, robotic equipment runs the synthesis, and AI evaluates the output; researchers interpret the data and adjust experimental models or the goal definition as needed.**



**SOURCE:** CONNOR COLEY/MASSACHUSETTS INSTITUTE OF TECHNOLOGY

whether industry is actually going to bite on this or whether it's some academic bullshit," Cooper says. His laboratory is exploring this question directly: Mobotics, a company spun off from Cooper's lab, is pursuing business with materials companies.

Similarly, Aspuru-Guzik colaunched Kebotix, an AI services firm working with robotics, in 2017 when he was at Harvard University. He is also coordinating the launch of an initiative, called Acceleration Consortium, focused on autonomous technology in materials discovery. Members include industrial and academic researchers.

## The start-ups

IBM followed up its chess win with a second popular showcase for AI by introducing Watson, a question-answering computer, as a contestant on the quiz show *Jeopardy!* In 2011, Watson defeated champions Brad Rutter and Ken Jennings, winning $1 million.

IBM went on to launch Watson as a commercial product; its first application was in decision management for lung cancer treatment at Memorial Sloan Kettering Cancer Center. The firm also customized Watson for industrial research markets, including chemicals, in which companies such as the big German firm Evonik Industries have deployed it.

In 2019, IBM announced that it would shift development in its Watson Health division from drug discovery, in which the tool struggled to gain traction, to clinical development. And the company is working on a new AI architecture to support autonomous chemical discovery in both materials and drug applications.

> **"You ask me why I'm doing this; it's because the world has no time."**

IBM's RoboRXN for Chemistry combines AI algorithms, commercially available robotics, and cloud computing. The firm debuted RoboRXN in a fully autonomous lab last year at its Zurich research laboratory with experiments involving photoacid generator molecules, carbon-capture materials, and pharmaceutical compounds.

IBM began working on the system in 2017 with a project to apply natural language programming to predictive chemical synthesis problems. Philippe Schwaller, a PhD student at IBM who studied reaction prediction at the Zurich lab, says the AI component of RoboRXN digitizes chemistry through language to promote machine learning.

"Our models are trained on molecular representations where atoms are like characters, molecules are like words, and chemical reactions are like sentences," Schwaller says. Thus programmed, the model can learn in a fashion that mimics human learning.

The model is also generating data, says Teodoro Laino, a research scientist and leader of the RoboRXN project. "AI is generating instructions," he says. "The net effect of AI is to add software that is writing software for the robot."

IBM is working with industrial and institutional partners to implement the technology in research, Laino says. For example, XChem, an experimental facility at Diamond Light Source, the UK's national synchrotron facility, is testing whether RoboRXN can discover compounds from data the group generates on how small molecules bind to proteins.

IBM has competition in offering AI research services: start-ups like Insilico Medicine, Exscientia, and Citrine Informatics are landing contracts with big chemical and pharmaceutical firms. Some

CREDIT: ADAPTED FROM CONNOR W. COLEY/WILL LUDWIG/C&EN

of the start-ups are also teaming up with lab robotics providers.

Insilico, for example, announced a partnership last July with Arctoris, a supplier of automated drug discovery technology. Insilico CEO Alex Zhavoronkov says the company is looking for an optimum fit between AI and robotics in a closed-loop laboratory. "We are trying to get high-quality data from a controlled robotics environment," he says.

For Zhavoronkov, as for Aspuru-Guzik at the University of Toronto, robots are the crux. "Synthesis is the main bottleneck. Currently we outsource synthesis" to a contract research organization (CRO), he says. Insilico used AI to discover an idiopathic pulmonary fibrosis drug candidate earlier this year, then signed on the CRO WuXi AppTec to synthesize it for testing.

Exscientia launched in 2012 with a vision of automated drug design, CEO Andrew Hopkins says. The firm is continually searching for elements of the process in which AI can increase productivity, he says.

"Drug discovery is both a big data and a small data problem," Hopkins says. Databases, patents, and pertinent literature create a huge amount of data from which to build models, he says. "But whenever we look at first-in-class drug targets, chances are we know very little about them." AI allows researchers to design algorithms that explore chemical space beyond the parameters of the original experimental model. "We think of drug discovery as a learning problem, not a screening problem," he says.

The company, which counts GlaxoSmithKline as a client, is preparing to bring robots into the fold. "We are building labs in this space and hiring a director of automation," Hopkins says.

Citrine targets the materials industry and boasts the chemical giant BASF as a client. The company is keeping its focus on AI, joining where needed with partners for robotics and other components of a closed-loop discovery system. CEO Greg Mulholland describes the role of AI in an autonomous lab as a central guidance system working hand in hand with the researcher. Other elements of the loop will vary from company to company, as will the closedness of the loop, Mulholland says.

## The real-world users

"Artificial intelligence is very much on our agenda," says Henrik Hahn, Evonik's chief digital officer. The German chemical company has been working for several years to understand the implications of AI in materials discovery as part of a digitization program that has introduced AI at various stages of research. The firm is moving toward closed-loop AI and robotics, but Hahn questions whether such an environment will ever fully evolve.

"The autonomous laboratory in materials discovery appears to be some kind of holy grail as computer programs and algorithms surpass the creativity of our materials science experts," he says. "But this is really rather a vision, and a vision means it will never come true."

Evonik is using IBM's Watson AI technology in its labs and programming its own algorithms, Hahn says. And it is beginning to explore automation, looking, for example, into IBM's RoboRXN technology.

Hahn emphasizes that the autonomous lab poses a steep management challenge. "Wherever we can automate, we will try to do so," he says, though cost can be an issue. And the people who work in the lab now are no small consideration. Lab automation "is clearly disrupting classic lab work," Hahn says. "This term *autonomous lab* has to be handled with care because there will always be the connotation that we are substituting human beings."

**"This term *autonomous lab* has to be handled with care because there will always be the connotation that we are substituting human beings."**

Dow is similarly pacing itself toward achieving autonomy in the lab. "My vision of the lab of the future has been practiced at Dow for a while," says A. N. Sreeram, the firm's chief technology officer. He notes that Dow has built and programmed its own supercomputers since late in the first decade of this century and is among the industry pioneers in automated research.

These days, Sreeram adds, the firm is working with robotics, big data, high-performance computing, and now AI and machine learning. And the company has been accessing technology through partnerships. They include a 2017 pact with 1QBit to develop quantum computing applications and a deal announced last year with Microsoft to develop AI for polyurethane research.

The British chemical maker Johnson Matthey is working with Mobotics, the company launched by Cooper at the University of Liverpool, on robotics for data-driven discovery. The firm says the work builds on a tradition of enabling scientists in ma-

**A scientist at Evonik Industries prepares a high-throughput screening apparatus at the company's laboratory in Essen, Germany. Evonik is investigating the integration of automation and artificial intelligence.**

CREDIT: EVONIK INDUSTRIES

**A researcher at Exscientia employs artificial intelligence to search for potential drug compounds. CEO Andrew Hopkins describes drug discovery as "a learning problem, not a screening problem."**

CREDIT: EXSCIENTIA

terials discovery through technology that supports synthesis, characterization, and measurement.

"The self-driving or self-guided laboratory is another practical capability which our scientists will use to speed up the innovation process," Paul Collier, a research fellow at the company, says in an email. "'Self-guided' in this area means within boundaries set by human scientists." AI-directed automation provides a tool "that complements human scientists rather than replacing them," he says.

Another chemical maker, DSM, in January announced a partnership with Delft University of Technology aimed at linking AI with automation. Marcus Remmers, DSM's chief technology officer, says the project aims to supercharge aspects of industrial biotech research—automation, modeling, data management, and AI—that are already in place.

A major focus will be on the researcher, Remmers says. "In the early stages of our digital transformation in R&D, we were focused on pilots and tools, trying to work things out in small environments," he says. "As we mature, we realize the tools are not the limiting factor anymore. More so, it is the mindset of the people. The scientists will have to reinvent what they are and how they see themselves and the value that they bring to this new world. If you see yourself as just somebody setting up a machine, you may well end up missing the big picture."

In drug discovery, AI has gained traction managing data complexity. "Huge progress has been made over the last few years in comprehensively cleaning, unlocking, and harnessing the diverse and large volumes of discovery data accumulated over decades of research," Hugo Ceulemans, scientific director of discovery data sciences at Janssen Research & Development, says in an email.

Drug companies are investing in data generation and acquisition for traditional data types such as assays and chemical reactions, as well as for new types, such as high-content microscopy images. As a result, Ceulemans says, pharmaceutical research is building up an AI and automation infrastructure.

"The large data volumes flowing from automated pipelines boost AI impact on the portfolio," he says. AI-created insights are also beginning to direct the data that scientists collect, he adds.

Biotech companies have the luxury of implementing AI from the ground up. Moderna's success in arriving at a vaccine with 95% efficacy against COVID-19 in less than a year is partly attributed to AI algorithms, according to an article in Digital Initiative, published by Harvard Business School.

For established drug companies, the adoption of AI can be a more protracted journey. Some major companies, however, are making a determined effort to close the research technology loop.

Eli Lilly and Company opened a robotics center in San Diego early last year in partnership with Strateos, a developer of research-scheduling software. The Lilly Life Sciences Studio is a 1,100 $m^2$ facility that includes an autonomous lab with over 100 instruments and storage of over 5 million compounds. It is the culmination of a 6-year project, says James P. Beck, the head of medicinal chemistry at the center.

Like most drug companies, Lilly has been doing automated biology for decades and automated chemistry for more than 10 years, Beck says. The Life Sciences Studio puts the company's expertise in chemistry, in vitro biology, sample management, and analytical data acquisition in a closed loop. AI controls robots that Lilly researchers can access via the cloud, he says.

The lab, which is operated by Strateos, is also accessible to outside researchers, who can bring their own data, compounds, and experiments to the system.

Christos A. Nicolaou, head of chemical informatics at Lilly, says AI algorithms have evolved to the level where they can orchestrate automated operations in a lab. "AI is mature enough nowadays, and we have enough good data to come down to earth and design with action in mind," he says. Lilly worked with software developers, Nicolaou says, but designed the AI architecture in-house.

"The lab of the future is here today," Beck says. But closing the loop requires heavy lifting. "It is a multifactorial challenge involving science, hardware, software, and engineering," he says. "It is far more than a science story."

In fact, proponents of the autonomous lab suggest it is a human evolution story—one in which a technological environment rises around the enlightened scientist, posing little threat to the human. "Imagination and creativity will remain human for the foreseeable future," the University of Liverpool's Cooper says.

Aspuru-Guzik at the University of Toronto takes that idea one step further, quoting Jorge Luis Borges's poem "Chess":

*God moves the player as he the pieces*
*But what god behind God plots the advent*
*Of dust and time and dreams and agonies?*

That god, Aspuru-Guzik says, is human.

This article is reprinted with permission from C&EN. A version of this article was published in C&EN on March 29. 2021, on page 28.

# We choose 15 promising companies using AI to reinvent materials discovery

» **Aether Biomachines**

» aetherbio.com

» **Based:** Menlo Park, California

» **Launched:** 2017

» **Money raised in start-up funding rounds:** $12 million

» **Publicly traded:** No

» **Key partnerships:** Not disclosed

» **Strategy:** Aether runs millions of enzymatic reactions in its high-throughput robotic lab and uses machine learning to create a searchable enzyme index. The company aims to produce enzymes that manufacture sophisticated or novel materials at low cost and with minimal environmental impact.

» **Why watch:** Aether's investment partners include the 1517 Fund, which also administers the Thiel Fellowship, for those who forgo or leave school to pursue trailblazing entrepreneurial, artistic, or activist work.

» **Alcemy**

» alcemy.tech

» **Based:** Berlin

» **Launched:** 2018

» **Money raised in start-up funding rounds:** Not disclosed

» **Publicly traded:** No

» **Key partnerships:** Not disclosed

» **Strategy:** Alcemy is dedicated to making concrete sustainable. Production of the material accounts for roughly 8% of global carbon emissions. Low-carbon methods exist but are challenging to execute with reliable quality compared with traditional approaches. Alcemy's artificial intelligence (AI) helps concrete manufacturers adjust parameters while producing low-carbon concrete to ensure consistent quality.

» **Why watch:** The company is developing technology that will control the concrete-mixing process autonomously.

» **Chemspeed Technologies**

» chemspeed.com

» **Based:** Fullinsdorf, Switzerland

» **Launched:** 1997

» **Money raised in start-up funding rounds:** Not disclosed

» **Publicly traded:** No

» **Key partnerships:** University of Toronto

» **Strategy:** Chemspeed develops software and robotics that enhance research productivity during development cycles for new materials or specialty chemicals. Its customers have used the technology for a number of applications, including to boost performance of exhaust gas catalysts and automate polymer purification.

» **Why watch:** Chemspeed is an official partner in the global(see page 10), which aims to use AI and robotics to hasten the discovery of novel materials.

» **Citrine Informatics**

» citrine.io

» **Based:** Redwood City, California

» **Launched:** 2013

» **Money raised in start-up funding rounds:** $48 million

» **Publicly traded:** No

» **Key partnerships:** AGC Glass Europe, BASF, Lanxess, Siemens, UL

» **Strategy:** One of C&EN's 2017 10 Start-Ups to Watch, Citrine packs AI tools into a secure, cloud-based platform. Customers can upload proprietary synthesis and characterization data to the platform, which can then predict high-performing materials (see page 16).

» **Why watch:** The Japanese fund Universal Materials Incubator has invested in Citrine to incorporate AI into research and development at Japan's top materials and chemical firms.

» **Hue.ai**

» hueai.com

» **Based:** Vienna, Virginia

» **Launched:** 2018

» **Money raised in start-up funding rounds:** $1.2 million

» **Publicly traded:** No

» **Key partnerships:** Not disclosed

» **Strategy:** Hue.ai has adapted AI for the optical industry. The company's flagship product is lenses to correct red-green color blindness. Hue's platform requires little data to generate product insights compared with established AI technology, which could make it useful for clients that have not yet fully digitized legacy data about their materials.
» **Why watch:** Hue's next goal is improving progressive lenses, which incorporate multiple prescriptions for close-up, middle, and distance viewing. It is also planning to expand to the ink, paint, and cosmetic industries.

» **IBM RoboRXN**
» rxn.res.ibm.com/rxn/robo-rxn/welcome
» **Based:** Zurich
» **Launched:** 2020
» **Money raised in start-up funding rounds:** Not applicable
» **Publicly traded:** Yes
» **Key partnerships:** University of Pisa
» **Strategy:** IBM RoboRXN is an AI-driven lab that automates the early stages of materials discovery (see page 10). The company trained its AI-assisted synthesis planning software with a data set of approximately 1 million patents. The firm combined the software with robots that can carry out up to five consecutive synthetic steps without human intervention.
» **Why watch:** IBM RoboRXN is working on incorporating more complex purifications into the robots' capabilities, expanding their repertoire of possible reaction combinations.

» **Imubit**
» imubit.com
» **Based:** Houston
» **Launched:** 2016
» **Money raised in start-up funding rounds:** $2.3 million
» **Publicly traded:** No
» **Key partnerships:** Not disclosed
» **Strategy:** Imubit designs AI technology that enables chemical plants and refineries to more efficiently run complex processes like polymerization. The platform is designed to connect teams and information that are traditionally siloed in plant operations.
» **Why watch:** Imubit says its system is also applicable to optimizing ammonia production. Ammonia is one of the most-produced inorganic chemicals in the world, and making it consumes about 1% of global energy.

» **Kebotix**
» kebotix.com
» **Based:** Cambridge, Massachusetts
» **Launched:** 2017
» **Money raised in start-up funding rounds:** $16.4 million
» **Publicly traded:** No
» **Key partnerships:** Bayer, BP, Northeastern University, Orbia
» **Strategy:** One of C&EN's 2019 10 Start-Ups to Watch, Kebotix discovers new chemicals and materials quickly and inexpensively by using machine learning and robotics (see page 4). Its vision is a self-driving lab where human researchers specify desired properties but algorithms suggest recipes for materials and robots conduct testing.
» **Why watch:** Kebotix has expanded into a second lab at a chemistry-focused accelerator in Woburn, Massachusetts.

» **Kyulux**
» kyulux.com
» **Based:** Fukuoka, Japan
» **Launched:** 2015
» **Money raised in start-up funding rounds:** $96.1 million
» **Publicly traded:** No
» **Key partnerships:** LG Display, Nippon Soda, Samsung Display, WiseChip Semiconductor
» **Strategy:** One of C&EN's 2016 10 Start-Ups to Watch, Kyulux uses AI technology to develop cost-effective and efficient organic light-emitting diode (OLED) materials for displays and lighting panels. The company licensed its AI-driven materials screening platform from Harvard University in 2016.
» **Why watch:** In 2020, Kyulux shipped its first OLED product. WiseChip will incorporate it into medical devices now and potentially into wearable and consumer electronics in the future.

» **Materials Zone**
» materials.zone
» **Based:** Tel Aviv, Israel
» **Launched:** 2017
» **Money raised in start-up funding rounds:** $8 million
» **Publicly traded:** No
» **Key partnerships:** Not disclosed
» **Strategy:** The Materials Zone platform enables researchers to more quickly discover novel materials. The company's software standardizes data from customers' scientific instruments, manufacturing facilities, and literature libraries, and its AI analyzes the data to generate insights that save time, save money, or streamline decision-making.
» **Why watch:** Materials Zone currently focuses on customers specializing in semiconductors, perovskites, or materials for green construction or energy storage. The platform is not restricted to those materials, however, which leaves the door open for further growth.

» **Noble.AI**
» noble.ai
» **Based:** San Francisco
» **Launched:** 2017
» **Money raised in start-up funding rounds:** $3.5 million
» **Publicly traded:** No
» **Key partnerships:** Solvay
» **Strategy:** Noble offers two AI-driven software products that can be applied to advanced materials applications, such as aluminum alloy design. The first product, Blueprint, consolidates and analyzes data from instruments, handwritten documents, and third-party software to lower the costs of research and development. The second, Reactor, focuses on developing new chemicals or materials.
» **Why watch:** Data science company StartUs Insights named Noble one of its top five materials informatics start-ups in 2020.

PHASESHIFT

» **Phaseshift Technologies**
» thephaseshift.com
» **Based:** Toronto
» **Launched:** 2019
» **Money raised in start-up funding rounds:** $775,000
» **Publicly traded:** No
» **Key partnerships:** Not disclosed
» **Strategy:** Phaseshift's machine-learning algorithms specialize in metallic glasses—tough and corrosion-resistant materials with applications that include aerospace construction and surgical pins. The AI predicts which alloys are likely to form the disordered atomic arrangement typical of these materials, as well as mechanical properties.
» **Why watch:** In 2020, the company joined the Digital Media Zone at Ryerson University, one of the world's top business incubators.

STOICHEIA™

» **Stoicheia**
» stoicheia.ai
» **Based:** Skokie, Illinois
» **Launched:** 2021
» **Money raised in start-up funding rounds:** $5 million
» **Publicly traded:** No
» **Key partnerships:** Not disclosed
» **Strategy:** Stoicheia builds libraries of millions of nanomaterials on a single chip, screens them for a host of desired properties, such as catalytic activity or corrosion resistance, and uses the results to train its AI to discover novel materials quickly.
» **Why watch:** The company's lithography technology can combine elements in new ways and create architectures with as many as seven elements, whereas similar firms are focusing on one- and two-element materials.

uncountable.

» **Uncountable**
» uncountable.com
» **Based:** San Francisco
» **Launched:** 2016
» **Money raised in start-up funding rounds:** Not disclosed
» **Publicly traded:** No
» **Key partnerships:** AGC Chemicals, Showa Denko, Solidia Technologies
» **Strategy:** Uncountable's machine-learning models determine how the interplay of a formulation's components affect performance so that researchers can optimize formulations with fewer experiments, thus shortening development time and cutting research costs.
» **Why watch:** Uncountable completed a stint at the Massachusetts Institute of Technology's Startup Exchange STEX25 accelerator, which offered the opportunity to connect with over 1,900 other start-ups and over 260 member corporations.

zymergen

» **Zymergen**
» zymergen.com
» **Based:** Emeryville, California
» **Launched:** 2013
» **Money raised in start-up funding rounds:** $874.1 million
» **Publicly traded:** Yes
» **Key partnerships:** FMC, Sumitomo Chemical
» **Strategy:** Zymergen uses machine-learning algorithms to engineer microbes that manufacture materials. In collaboration with Sumitomo, the company has already developed a biobased, transparent film for displays and touch sensors.
» **Why watch:** In May 2021, Zymergen raised $500 million in an initial public offering. The company plans to use some of the funds for its automated materials discovery labs.

**Note:** Companies were included because of the novelty and potential of their methods, amount of capital raised, number of partnerships, and number and identity of investors.

**Sources:** Crunchbase (accessed May 2021), company websites, news reports.

HOT SHEET

CREDIT: MIKE REDDY

# Machine-learning performance put to the test

**SAM LEMONICK,** C&EN STAFF

Imagine you're a materials scientist and your job is to discover a new material, a combination of atoms no one has ever made. Maybe you're looking for a metal-organic framework (MOF). They have a lot of potential applications: carbon capture, drug delivery, and hydrogen fuel storage, just to name a few.

So how do you find a new MOF? You could try machine learning. People are saying good things about machine learning, especially graph neural networks (GNNs), which are designed to behave like neurons in a brain.

And lucky for you, there are already a handful of GNNs designed to predict new materials. Better yet, you can download them right now, for free, from a site like GitHub.

> "We need to be very careful about how we do this because the floor is way deeper than the ceiling is high."

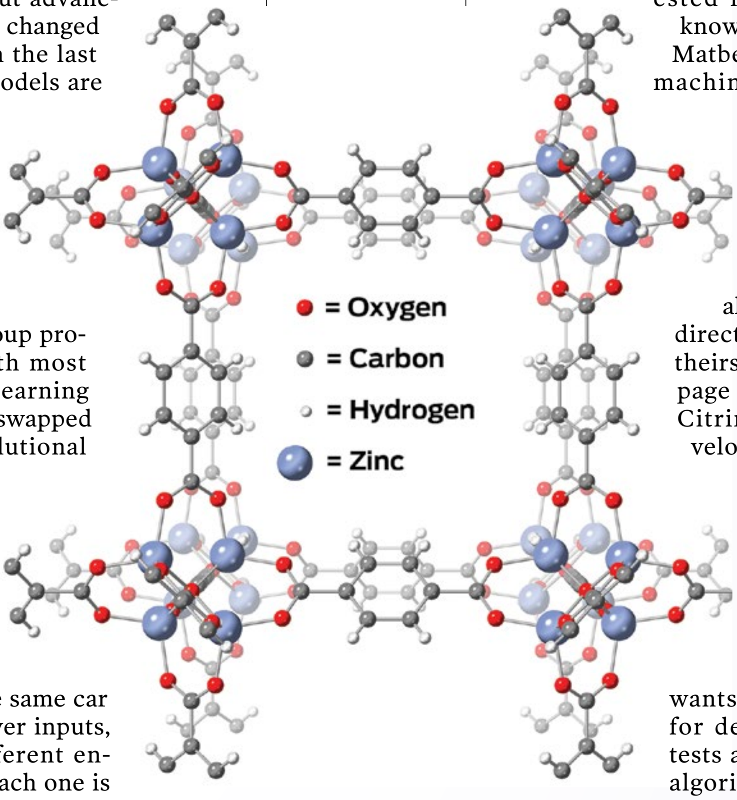
But which GNN should you use? How should you teach it to make accurate predictions? What are the optimal settings? Choose wisely; you're about to invest time and money and maybe hire a graduate student to go MOF hunting. Which GNN will find the best new MOF in the shortest time?

Those are not easy questions to answer, regardless of whether you're new to machine learning for materials discovery or an old hand. Some scientists are trying to make those decisions easier by developing methods for comparing the performance of machine-learning algorithms. These researchers say that adopting these benchmarking methods could help speed the discovery of new materials. It could also help developers of machine-learning models improve their algorithms and approaches.

The idea of benchmarking isn't new. In simple terms, it means comparing the performance of one process against a baseline to quantify how much that process helps you. Chemists have benchmarked computational approaches before—for example, comparing how well approaches to density functional theory (DFT) predict experimentally derived chemical properties. Through that benchmarking, chemists now know when they can trust DFT to make accurate predictions.

Benchmarking hasn't become widespread in the world of machine learning for materials discovery. Bobby G. Sumpter of Oak Ridge National Laboratory (ORNL) has been experimenting with machine learning for decades. He says there are many machine-learning methods available, many of which are open source, and more are appearing all the time. "People get sort of overwhelmed by what to choose," Sumpter says.

Sumpter, his ORNL colleague Victor Fung, and others developed a tool called MatDeepLearn for benchmarking GNNs in materials discovery. Fung says a few years ago he thought machine learning was probably overhyped, but advances in GNNs since then have changed his mind. He says papers in the last 1–2 years show that these models are capable of chemical accuracy, meaning their predictions match properties measured experimentally. Still, like Sumpter, Fung says choosing which one to use can "be a roll of the dice."



**Oak Ridge National Laboratory researchers benchmarking state-of-the-art graph neural networks found that they poorly predicted properties of metal-organic frameworks, like this one.**

In MatDeepLearn, the group programmed a framework with most of the steps of a machine-learning discovery process and then swapped in different models' convolutional operator, which is these algorithms' central component that processes data to make predictions. You can think of this benchmarking process like testing and comparing car engines. The researchers have built test beds with the same car body, wheels, tires, and driver inputs, and then they swap in different engines to measure how fast each one is in a race.

The team tested five GNNs in its framework in a recent preprint to see how well the algorithms predicted properties of different classes of materials (*npj Comput. Mater.* 2021, DOI: 10.1038/s41524-021-00554-0). The top four GNN models all performed about equally. Fung says these results suggest that for scientists simply looking for a model that performs well at the tasks tested, it might not make much difference which model they choose.

But he says for scientists developing new GNNs and machine-learning methods, the results raise some questions. The researchers found that MEGNet, a GNN published in 2019 (*Chem. Mater.* 2019, DOI: 10.1021/acs.chemmater.9b01294), performed about as well as SchNet, released in 2017 in a preprint (arXiv 2017, arXiv: 1706.08566). Preprints are not peer reviewed. If 2 years hasn't led to an increase in algorithm performance, "are we making progress?" Fung asks. He says his team's study points to another way that benchmarks can be useful. They help developers of models identify what they're doing right or wrong as they try to improve their methods.

Alex Dunn of the University of California, Berkeley, and Lawrence Berkeley National Laboratory says aiding developers was the motivation for a benchmarking method called Matbench that he, Anubhav Jain of Berkeley Lab, and colleagues developed (*npj Comput. Mater.* 2020, DOI: 10.1038/s41524-020-00406-3). Without a way to compare machine-learning models fairly, Dunn says, "it can be hard for someone who's interested in advancing the field to know what avenue to go down."

Matbench tests algorithms on 13 machine-learning tasks, such as predicting a material's bandgap. And the scientists created a reference algorithm for users to benchmark their algorithm against.

"Now if you have a new algorithm or method, you can directly compare your results to theirs," says Bryce Meredig (see page 6), chief science officer at Citrine Informatics, which develops machine-learning methods for materials science. Meredig says the materials science community has realized in the past few years that it lacks a set of universally acknowledged benchmarks for gauging model performance. Dunn wants it to be common practice for developers to use standard tests and data sets to benchmark algorithms and to publish those results.

Like MatDeepLearn, Matbench produced some counterintuitive results. While the GNNs that Matbench tested outperformed the reference algorithm when trained on data sets with more than 10,000 entries, the more simplistic reference algorithm did better than the GNNs on most predictions when fewer data were available. The researchers say these results suggest that researchers can predict some properties accurately without more computationally expensive algorithms like GNNs.

Another attempt at benchmarking machine-learning methods set a more holistic goal. Santosh K. Suram of Toyota Research Institute, John M. Gregoire of the California Institute of Technology, and colleagues evaluated how long it took different machine-learning methods to do three tasks: find one good catalyst in a data set, find all the good catalysts in that data set, and predict the performance of catalysts not in the training data set. “This benchmarking evaluates the impact instead of the predictive power of an algorithm,” Gregoire says. In other words, they wanted to determine not just how well a machine-learning model can predict properties but how well it can address the larger goal of accelerating materials discovery. The researchers used a data set of catalysts whose properties they had already determined experimentally.

They found that the most advanced machine-learning models they tested, called sequential learning models, can discover all the high-performing catalysts in the data set about 20 times as fast as random sampling, a less sophisticated type of machine-learning model (*Chem. Sci.* 2020, DOI: 10.1039/C9SC05999G). Sequential learning means the computer chooses which experiments to do next to improve a model’s predictive performance. But the group also found that if the model wasn’t set up optimally, sequential learning could take 1,000 times as long as random sampling.

Gregoire says the results are a good lesson about understanding what tasks different machine-learning models are good for in materials discovery. “We need to be very careful about how we do this because the floor is way deeper than the ceiling is high,” he says.

**“Now if you have a new algorithm or method, you can directly compare your results to theirs.”**

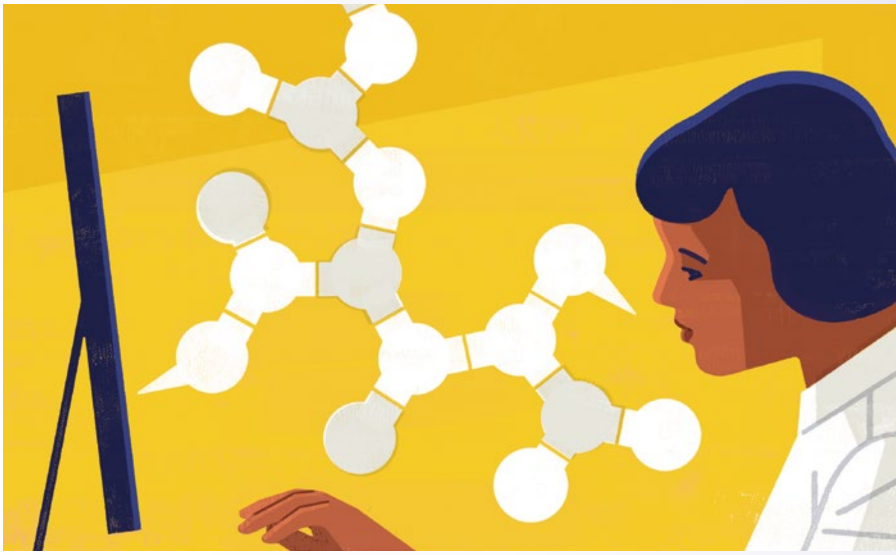
These researchers are developing benchmarking methods as stand-alone projects. Heather J. Kulik of the Massachusetts Institute of Technology and her team have been implementing benchmarking as a natural part of their materials discovery process with machine learning. “I have to be able to defend that we learned something in a way we couldn’t” without machine learning, Kulik says. She says her group typically publishes the results of its benchmarks in papers. Benchmarks also help new group members understand the strengths of different models, she says.

Kulik thinks new benchmarking tools like the ones Fung, Dunn, Gregoire, and their colleagues have developed are great if they can get people to use them. But she points out that there’s a human element that may override even the most rigorous and empirical benchmark. “Even if we know what should be the most accurate thing, we don’t always choose it, maybe because of our own biases,” she says. People might choose the machine-learning model that’s most cited or the one their graduate advisers used, she says, even against the evidence.

This article is reprinted with permission from C&EN. A version of this article was published in C&EN on April 12, 2021, on page 16.

# Teaching computers to speak chemistry

**SAM LEMONICK**, C&EN STAFF



**To help computers make more powerful chemical predictions, computational chemists hope the machines can learn to read the language of chemistry.**

CREDIT: CHRIS GASH

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. *1S* 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.

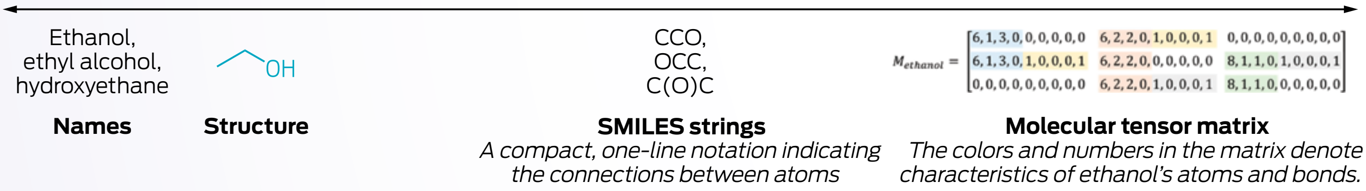
Some chemists are now using these line-notation systems to harness some of the most impressive machine-learning models to make predictions about new molecules and reactions. Natural language processing, also called machine translation, is one of the most active areas of research for computer scientists. 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 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 (*ACS Cent. Sci.* 2017, DOI: 10.1021/acscentsci.7b00512). 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 (arXiv 2016, arXiv:1612.09529). 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 (*Chem. Sci.* 2018, DOI: 10.1039/C8SC02339E). 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 its own retrosynthesis prediction algorithm (*ACS Cent. Sci.* 2017, DOI: 10.1021/acscentsci.7b00303).

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.


Ethanol,
ethyl alcohol,
hydroxyethane
OH
CCO,
OCC,
C(O)C
$M_{ethanol} = \begin{bmatrix} 6,1,3,0,0,0,0,0,0 & 6,2,2,0,1,0,0,0,1 & 0,0,0,0,0,0,0,0,0 \\ 6,1,3,0,1,0,0,0,1 & 6,2,2,0,0,0,0,0,0 & 8,1,1,0,1,0,0,0,1 \\ 0,0,0,0,0,0,0,0,0 & 6,2,2,0,1,0,0,0,1 & 8,1,1,0,0,0,0,0,0 \end{bmatrix}$
Names
Structure
SMILES strings
A compact, one-line notation indicating the connections between atoms
Molecular tensor matrix
The colors and numbers in the matrix denote characteristics of ethanol's atoms and bonds.

**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.**

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; "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 2020, 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 3D 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 Lee of the University of Cambridge and others, has adapted the transformer approach for reaction prediction (*ACS Cent. Sci.* 2019, DOI: 10.1021/acscentsci.9b00576). 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 3D space.

Lee 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

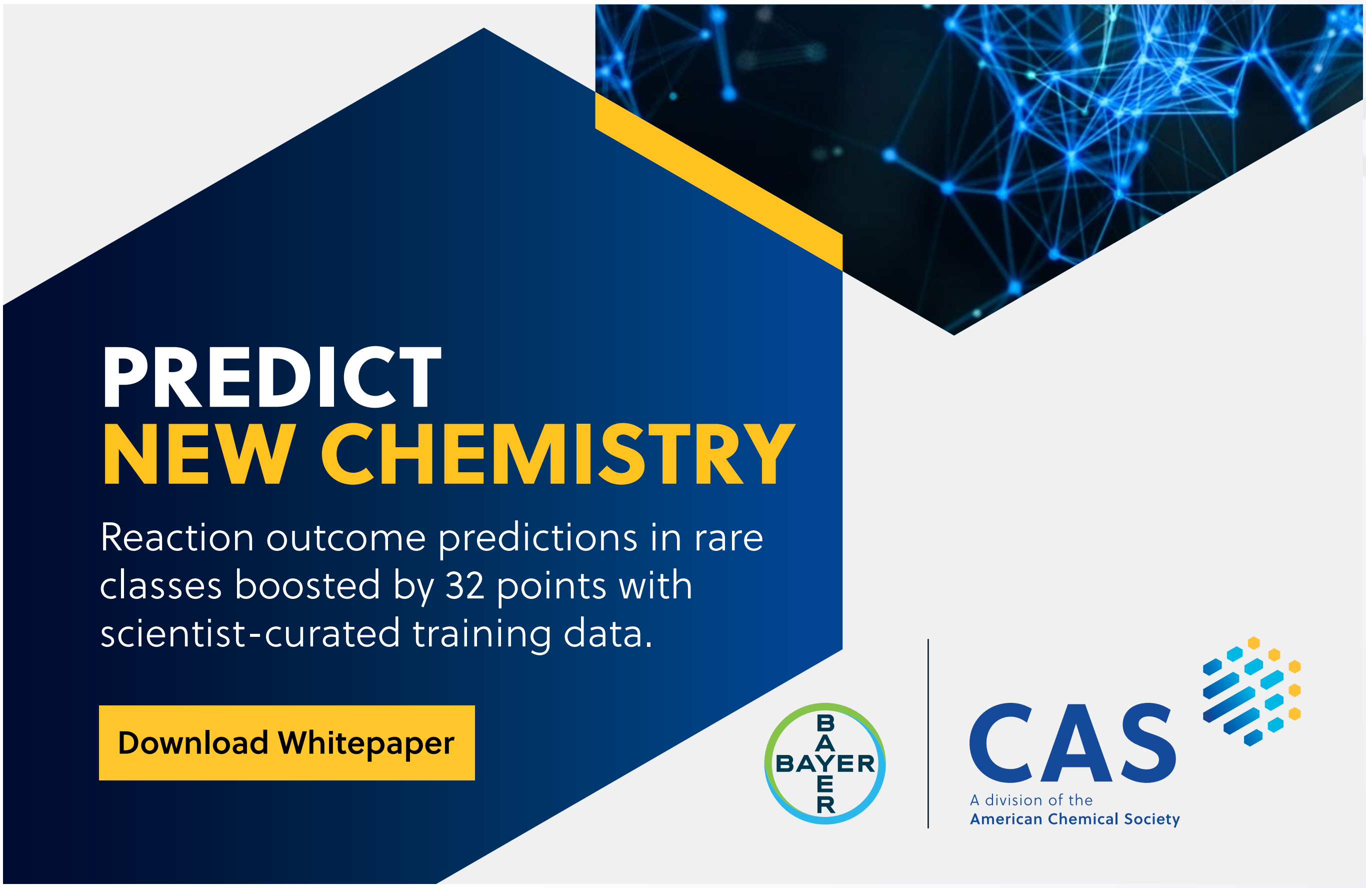
**“It will be quite a high burden to make this work decently, but the idea is that once you can do it, there are countless opportunities.”**

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 (*J. Med. Chem.* 2020, DOI: 10.1021/acs.jmedchem.9b02187). 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 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 demonstrated those differences in required computing resources. They’re developing a technique for pretraining transformer-based models with SMILES strings to make the models faster at chemistry prediction (arXiv 2020, arXiv: 2010.09885). Grand notes that a similar graph-based method from the company Tencent needs 250 processors in its calculations (arXiv 2020, arXiv: 2007.02835). The trio’s method uses just 1.

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. Still, these scientists see the field progressing because as chemists talk to computers, the computers are learning to talk back.

This article is reprinted with permission from C&EN. A version of this article was published in C&EN on February 8, 2021, on page 19.



READING LIST

# Our picks of the patent and journal literature on AI in materials discovery

## 2021

» Ziatdinov, Maxim, Ayana Ghosh, Tommy Wong, and Sergei V. Kalinin. **"AtomAI: A Deep Learning Framework for Analysis of Image and Spectroscopy Data in (Scanning) Transmission Electron Microscopy and Beyond."** arXiv (May 16, 2021). arXiv: 2105.07485.

» Duan, Chenru, Fang Liu, Aditya Nandy, and Heather J. Kulik. **"Putting Density Functional Theory to the Test in Machine-Learning-Accelerated Materials Discovery."** *J. Phys. Chem. Lett.* 12, no. 19 (May 20, 2021): 4628–37. https://doi.org/10.1021/acs.jpclett.1c00631.

» Citrine Informatics. **"Using Machine Learning to Explore Formulations Recipes with New Ingredients."** US Patent 10,984,145, filed July 21, 2020, and issued April 20, 2021.

» Maik Jablonka, Kevin, Giriprasad Melpatti Jothiappan, Shefang Wang, Berend Smit, and Brian Yoo. **"Bias Free Multiobjective Active Learning for Materials Design and Discovery."** *Nat. Commun.* 12 (April 19, 2021): 2312. https://doi.org/10.1038/s41467-021-22437-0.

» Maffettone, Phillip M., Lars Banko, Peng Cui, Yury Lysogorskiy, Marc A. Little, Daniel Olds, Alfred Ludwig, and Andrew I. Cooper. **"Crystallography Companion Agent for High-Throughput Materials Discovery."** *Nat. Comput. Sci.* 1 (April 19, 2021): 290–97. https://doi.org/10.1038/s43588-021-00059-2.

» Pollice, Robert, Gabriel dos Passos Gomes, Matteo Aldeghi, Riley J. Hickman, Mario Krenn, Cyrille Lavigne, Michael Lindner-D'Addario, et al. **"Data-Driven Strategies for Accelerated Materials Design."** *Acc. Chem. Res.* 54, no. 4 (Feb. 16, 2021): 849–60. https://doi.org/10.1021/acs.accounts.0c00785.

## 2020

» Gilad Kusne, Aaron, Heshan Yu, Changming Wu, Huairuo Zhang, Jason Hattrick-Simpers, Brian DeCost, Suchismita Sarker, et al. **"On-the-Fly Closed-Loop Materials Discovery via Bayesian Active Learning."** *Nat. Commun.* 11 (Nov. 24, 2020): 5966. https://doi.org/10.1038/s41467-020-19597-w.

» Meftahi, Nastaran, Mykhailo Klymenko, Andrew J. Christofferson, Udo Bach, David A. Winkler and Salvy P. Russo. **"Machine Learning Property Prediction for Organic Photovoltaic Devices."** *npj Comput. Mater.* 6 (Nov. 6, 2020): 166. https://doi.org/10.1038/s41524-020-00429-w.

» Corning. **"System and Method for Screening Homopolymers, Copolymers or Blends for Fabrication."** US Patent 10,790,045, filed Oct. 15, 2019, and issued Sept. 29, 2020.

» Kim, Yoolhee, Edward Kim, Erin Antono, Bryce Meredig, and Julia Ling. **"Machine-Learned Metrics for Predicting the Likelihood of Success in Materials Discovery."** *npj Comput. Mater.* 6 (Aug. 26, 2020): 131. https://doi.org/10.1038/s41524-020-00401-8.

» Jin, Wengong, Regina Barzilay, and Tommi Jaakkola. **"Hierarchical Generation of Molecular Graphs Using Structural Motifs."** In Proceedings of the 37th International Conference on Machine Learning, virtual, July 13–18, 2020. Hal Daumé III, Aarti Singh, eds. *Proceedings of Machine Learning Research,* 119: 4839–48. http://proceedings.mlr.press/v119/jin20a/jin20a.pdf.

» Burger, Benjamin, Phillip M. Maffettone, Vladimir V. Gusev, Catherine M. Aitchison, Yang Bai, Xiaoyan Wang, Xiaobo Li, et al. **"A Mobile Robotic Chemist."** *Nature* 583 (July 8, 2020): 237–41. https://doi.org/10.1038/s41586-020-2442-2.

» Citrine Informatics. **"Predictive Design Space Metrics for Materials Development."** US Patent 10,657,300, filed Oct. 2, 2019, and issued May 19, 2020.

» Yu-Tung Wang, Anthony, Ryan J. Murdock, Steven K. Kauwe, Anton O. Oliynyk, Aleksander Gurlo, Jakoah Brgoch, Kristin A. Persson, and Taylor D. Sparks. **"Machine Learning for Materials Scientists: An Introductory Guide toward Best Practices."** *Chem. Mater.* 32, no. 12 (June 23, 2020): 4954–65. https://doi.org/10.1021/acs.chemmater.0c01907.

» Zhong, Miao, Kevin Tran, Yimeng Min, Chuanhao Wang, Ziyun Wang, Cao-Thang Dinh, Phil De Luna, et al. **"Accelerated Discovery of $CO_2$ Electrocatalysts Using Active Machine Learning."** *Nature* 581 (May 14, 2020): 178–83. https://doi.org/10.1038/s41586-020-2242-8.

» MacLeod, Benjamin P., Fraser G. L. Parlane, Thomas D. Morrissey, Florian Häse, Loïc M. Roch, Kevan E. Dettelbach, Raphaell Moreira, et al. **"Self-Driving Laboratory for Accelerated Discovery of Thin-Film Materials."** *Sci. Adv.* 6, no. 20 (May 13, 2020): eaaz8867. https://doi.org/10.1126/sciadv.aaz8867.

» Morgan, Dane, and Ryan Jacobs. **"Opportunities and Challenges for Machine Learning in Materials Science."** *Annu. Rev. Mater. Res.* 50 (July, 2020): 71–103. https://doi.org/10.1146/annurev-matsci-070218-010015.

» Gongora, Aldair E., Bowen Xu, Wyatt Perry, Chika Okoye, Patrick Riley, Kristofer G. Reyes, Elise F. Morgan, and Keith A. Brown. **"A Bayesian Experimental Autonomous Researcher for Mechanical Design."** *Sci. Adv.* 6, no. 15 (April 10, 2020): eaaz1708. https://doi.org/10.1126/sciadv.aaz1708.

**Note:** This list was chosen by experts who work in the field, CAS information scientists, and C&EN editorial staff.

**Download a copy of this and all Discovery Reports at www.acs.org/discoveryreports**