THE DRUG-MAKER'S GUIDE TO THE GALAXY

HOW MACHINE LEARNING AND BIG DATA ARE HELPING CHEMISTS SEARCH THE VAST CHEMICAL UNIVERSE FOR BETTER MEDICINES.

BY ASHER MULLARD

In 2016, the pharmaceutical firm Sunovion gave a group of seasoned employees an unusual assignment. At the firm's headquarters in Marlborough, Massachusetts, the chemists were all asked to play a game to see who could discover the best leads for new drugs. On their workstations was a grid of hundreds of chemical structures, just ten of which were labelled with information on their biological effects. The experts had to select other molecules that could turn out to be drug candidates, using their hard-earned knowledge of chemical structure and biology. Of the 11 players, 10 struggled through the task for several hours. But one breezed through in milliseconds — because it was an algorithm.

That computer program was the brainchild of Willem van Hoorn, head of chemoinformatics at Exscientia, a start-up that uses artificial intelligence (AI) to design drugs. The firm, based in Dundee, UK, wanted to extend a nascent partnership with Sunovion, so the stakes were high. "My credibility was on the line," says van Hoorn. Twenty rounds of gameplay later, he tallied up the points. Relief swept over him. His algorithm had mastered at least some of the dark arts of chemistry; only one drug-hunting expert had beaten the machine.

ILLUSTRATION BY VASAVA

Exscientia and Sunovion have continued to work together to discover psychiatric drugs ever since. "This competition really helped to get buy-in from the people who make the chemistry research decisions," says Scott Brown, Sunovion's director of computational chemistry.

Exscientia is just one of a growing number of groups in industry and academia that are turning to computers to explore the mind-bogglingly large chemical universe. Chemists estimate that 10^{60} compounds with drug-like characteristics could be made — that's more small molecules than there are atoms in the Solar System. The hope is that algorithms will catalogue, characterize and compare the properties of millions of compounds *in silico* to help researchers quickly and affordably find the best drug candidates for a target. Proponents argue that these strategies could make medicines safer, ensure that fewer drugs fail in clinical trials and enable the discovery of new classes of therapeutics. They could also help to open up areas of chemical space left unexplored or assumed to be barren.

But many medicinal chemists remain sceptical of the hype, unconvinced that the ineffable complexity of chemistry can be reduced to mere lines of code. Even advocates of AI acknowledge that many attempts have fallen flat: computer-generated compounds can be riddled with components that are difficult to make, such as 3- or 4-atom rings, and infested with reactive groups that would set off safety alarms. "The execution of some computational approaches can suffer badly when researchers just don't know the field," says van Hoorn. "The compounds they come up with are just laughable." But he says that an expert human touch could yet tame these overzealous digital designers. "I think some of these ideas could work if the computer scientists would just collaborate with people who actually breathe chemistry."

"TOGETHER THE HUMAN AND AI CAN OUTPERFORM ANY HUMAN, BUT THEY CAN ALSO OUTPERFORM ANY ALGORITHM."

SPACE EXPLORATION

To navigate the chemical universe, it helps to have a map. In 2001, chemist Jean-Louis Reymond, at the University of Berne in Switzerland, started using computers to chart as much of the massive space as possible. Sixteen years on, he has amassed the largest database of small molecules in the world, a gigantic virtual collection of 166 billion compounds. The database, called GDB-17, includes all the chemically feasible organic molecules made of up to 17 atoms — as many as Reymond's computers could cope with. "Just for a computer to compile a list of the compounds in the database would now take over 10 hours," says Reymond.

To make sense of this plethora of possible drug starting points, Reymond has come up with a way to organize his chemical universe. Taking inspiration from the periodic table, he has grouped compounds in a multidimensional space in which neighbouring compounds have related properties. Positions are assigned according to 42 characteristics, such as how many carbon atoms each compound has.

For each drug that has made it to market, there are millions of compounds that are chemically almost identical to it — just sporting an extra hydrogen here or double bond there. And some of these will work better than the drug that was approved. Chemists couldn't possibly conceive of all of these variations unaided. "There is no way you can get at these isomers using a pen and a piece of paper," says Reymond.

But Reymond and his team can identify therapeutically promising 'near neighbours' of proven drugs by searching for similarities between compounds. By using a particular drug as a starting point, the team can comb through all 166 billion compounds in the database for compelling follow-on candidates in just 3 minutes. In a proof-of-principle experiment, Reymond started with a known molecule that binds the nicotinic acetylcholine receptor, a useful target for disorders involving the nervous system or muscle function, and compiled a shortlist of 344 related compounds. The team synthesized three, and found that two could activate the receptor potently, and could be useful for treating muscular atrophy in ageing[1]. The approach is like using a geological map to work out where to dig for gold, Reymond says. "You need some way to choose where you are going to dig," he says.

An alternative approach uses computers to pan lots of locations for gold without worrying too much about the starting location. In drug-hunting terms, this means screening vast chemical libraries *in silico* to find small molecules that bind to a given protein. First, researchers have to take a snapshot of a protein using X-ray crystallography to determine the shape of its binding site. Then, using molecular-docking algorithms, computational chemists can chug through compound collections to find the best fits for any given site.

As computing power has exploded, the capabilities of these algorithms have improved. Chemists at the University of California, San Francisco, led by Brian Shoichet, showcased the potential of this approach in 2016 in a search for a new class of painkiller. The team screened more than 3 million commercially available compounds to find candidates that would selectively activate μ-opioid receptor signalling to relieve pain without disturbing the closely related β-arrestin signalling pathway — which is thought to be associated with opioid side effects including a lowered breathing rate and constipation. The researchers quickly whittled down a massive compound library to just 23 highly ranked compounds for follow-up[2].

In a test tube, seven of the candidates had the desired activity. Further development turned one of these into PZM21, a compound that acts on the μ-opioid receptor without activating β-arrestin. The biotechnology firm Epiodyne, based in San Francisco, California, and co-founded by Shoichet, is now trying to develop a safer painkiller based on the findings. Shoichet plans to use the same approach to find compounds that modulate other G-protein-coupled receptors (GPCRs), a family of proteins that accounts for an estimated 40% of drug targets.

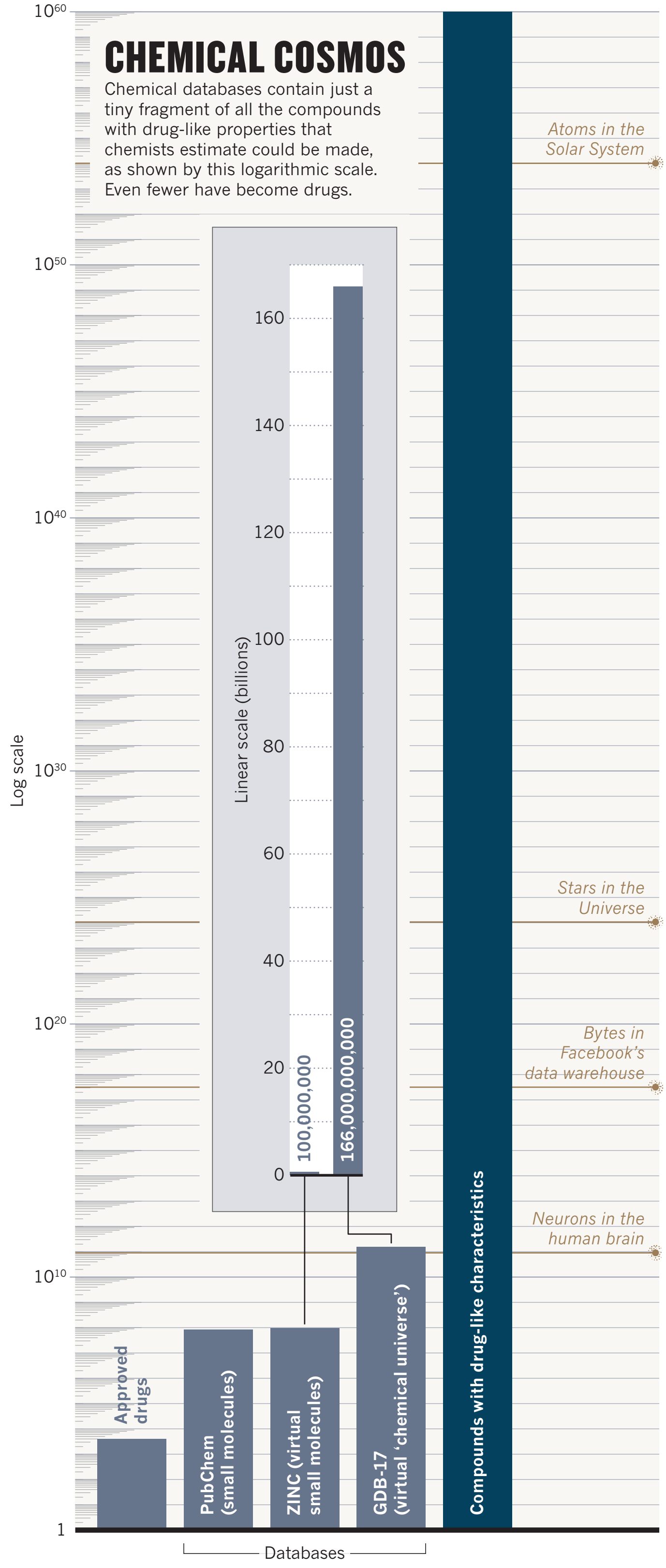
His team is also running similar experiments with a virtual nebula of 100 million compounds that have never been made before but that should be easy to synthesize. Industry drug developers are also testing out this approach: the biotech firm Nimbus Therapeutics, based in Cambridge, Massachusetts, incorporates into its docking screens virtual compounds with characteristics of naturally occurring chemicals that usually have to be laboriously sourced from natural environments such as soil. The jury is still out on whether these will lead to drugs, but Don Nicholson, chief executive of the company, says that for at least one drug-design programme, "this is where all our hits are coming from".

Preliminary results from such virtual screens are shaking one of Shoichet's core assumptions about chemical space: that it's only worth looking in established, drug-rich regions. Well-characterized galaxies of molecules are so awash with biologically active compounds that some argue it is a waste of time searching elsewhere. "Throughout my career I have believed that line of reasoning. It just made sense, even if there wasn't that much evidence to support it," says Shoichet. But unpublished results from his screens of 100 million compounds are stoking his interest in the less-explored regions of chemical space. "I'm starting to think that those galaxies are full of gold."

IN SILICO INSIGHT

These data-searching approaches are tried and tested, but the computers involved can follow only scripted instructions. The latest frontier in computational drug discovery is machine learning, in which algorithms use data and experience to teach themselves which compounds bind to which targets, finding patterns that are invisible to the human eye. Around a dozen firms have sprung up to create drug-hunting algorithms that they can test in partnership with large pharmaceutical companies.

Andrew Hopkins, chief executive of Exscientia, makes a strong case for the power of these approaches. It takes on average 4.5 years to discover



CHEMICAL COSMOS

Chemical databases contain just a tiny fragment of all the compounds with drug-like properties that chemists estimate could be made, as shown by this logarithmic scale. Even fewer have become drugs.

and optimize candidates for preclinical testing[3], and chemists often synthesize thousands of compounds to get to a promising lead (which even then has only a slim chance of making it to market). Exscientia's approach — which uses various algorithms, including the one that impressed Sunovion's research and development executives — may be able to reduce this timeline to just one year, and shrink the number of compounds that a drug-discovery campaign needs to consider.

In 2015, Exscientia finished a 12-month campaign for Sumitomo Dainippon Pharma, which owns Sunovion and is based in Osaka, Japan. The researchers trained their AI tools to find small molecules that modulate two GPCRs at the same time, and found they needed to synthesize fewer than 400 compounds in order to identify a good candidate. The drug that emerged is now moving towards clinical trials for psychiatric disease, says Hopkins. Since May, the company has inked deals worth hundreds of millions of dollars with Sanofi, based in Paris, and GlaxoSmithKline, based in Brentford, UK.

In addition to identifying leads, machine-learning algorithms can also help drug developers to decide early on which compounds to kill, says Brandon Allgood, chief technology officer of Numerate, an AI drug-design firm based in San Bruno, California. There's no point in making and testing a compound if it's going to fail on toxicity or absorption testing a few months later, he says. With AI, "it takes just a millisecond to rule it in or out", says Allgood, who trained as a cosmologist before he started using AI tools to study the chemical cosmos. Numerate has struck two deals with pharmaceutical companies this year, including one with Servier, based in Suresnes, France, to put AI-discovered drugs through clinical trials for heart failure and arrhythmias.

Industry investment is blossoming, but computational approaches still have a lot to prove. Reymond's collection is gigantic compared with other libraries, but it covers the minutest fraction of the chemical universe (see 'Chemical cosmos'). Despite the 166 billion compounds in his database, he still has further to go in his quest than an astronomer who is trying to count all the stars in the night sky but has only managed to record one. Screens that rely on matching proteins with drugs need accurate crystal structures to yield the best results, and these data take time, money and expertise to generate. These methods also struggle to cope with proteins in motion and they cannot rank their suggestions very well. Machine-learning algorithms, for their part, are only as good as the training data sets that they are based on, performing particularly poorly when they encounter compounds that look unlike molecules they have seen before. What's more, the programs run as black boxes, and cannot indicate why they predict a compound will be a good fit.

Many computational approaches also have an annoying habit of suggesting candidates that are nightmares to cook up in a lab. Chemists must then laboriously figure out a recipe for the suggested compound, which can take months or more. Even then, there is no guarantee that the molecule will work once it is made. Reymond's approach predicts a compound's activity profile correctly only 5–10% of the time, and that means chemists have to toil away on up to 20 compounds to find one that acts as expected. "I would say the bottleneck in our exploration of chemical space is the ability to dare to make compounds," says Reymond. To this end, he recently shaved his chemical universe down to a shortlist of 10 million molecules that are easy to make, and yet still cover a broad range of properties.

Mark Murcko, chief scientific officer at Relay Therapeutics in Cambridge, Massachusetts, thinks computational chemists should focus less on coming up with new algorithmic strategies, and more on improving the data sets they learn from. "One of the best ways that I know of to make a predictive model better is to keep feeding it more and more, and better and better, data," he says. Relay and others have bench chemists working closely with computational scientists, synthesizing compounds proposed by both humans and algorithms and using the resulting findings to inform future decisions.

For Hopkins, such collaborations are key. It took decades for computer scientists to write programs that could compete with chess grandmasters. Then, in 1997, IBM's Deep Blue beat Garry Kasparov. But the loss did not mark the end of chess. Instead, Kasparov created a doubles version in which each team consists of a human player and an AI. "Together the human and AI can outperform any human, but they can also outperform any algorithm," says Hopkins. He wants the same mix of data-crunching, creativity and common sense to transform drug discovery. "I believe we are at the Kasparov–Deep Blue moment." ■

Asher Mullard *is a journalist based in Ottawa, Canada.*

1. Reymond, J.-L. *Acc. Chem. Res.* **48,** 722–730 (2015).
2. Manglik, A. *et al. Nature* **537,** 185–190 (2016).
3. Paul, S. M. *et al. Nature Rev. Drug Discov.* **9,** 203–214 (2010).