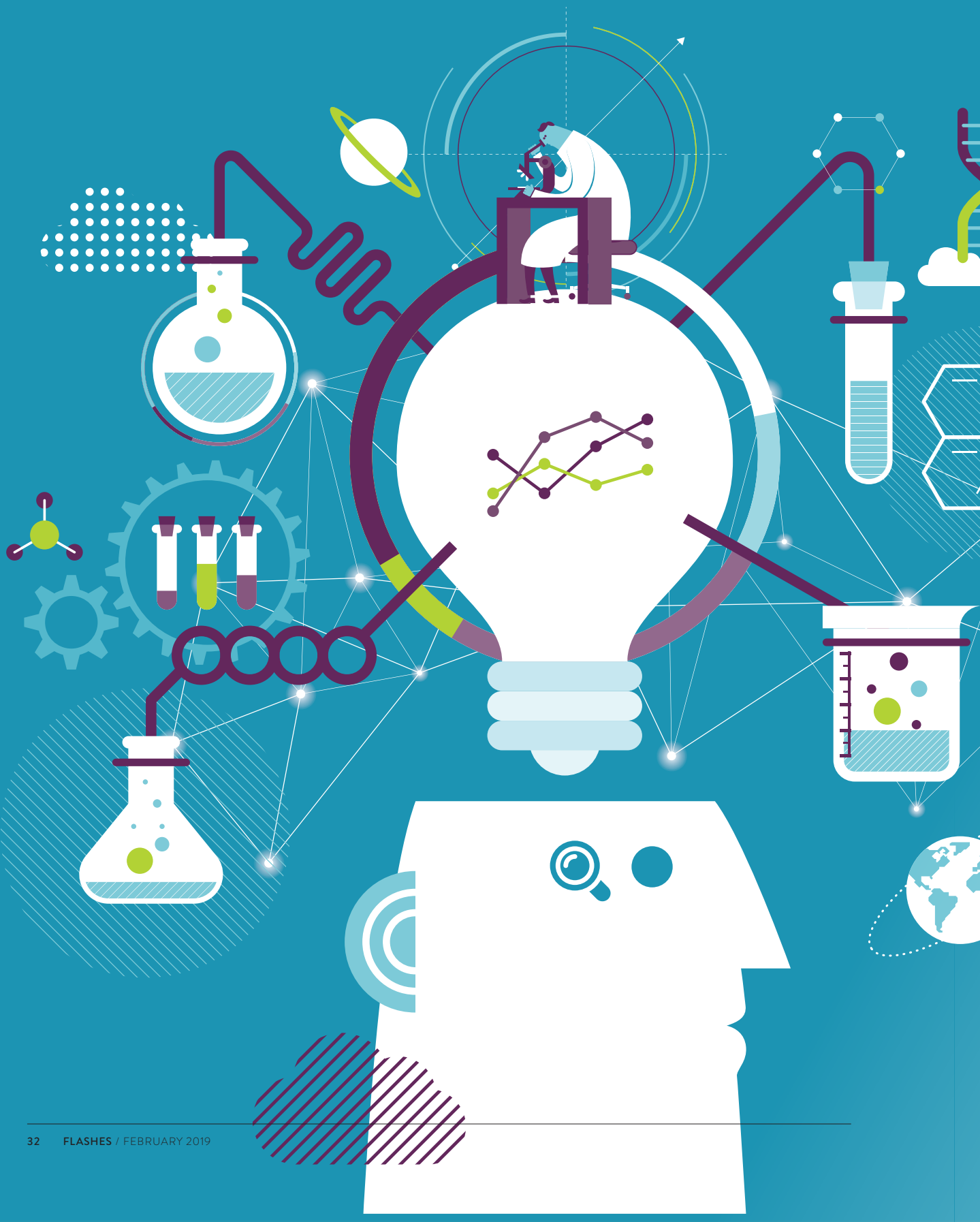


REWRITING THE RULES OF CHEMISTRY

Machine learning and artificial intelligence are revolutionising the processes of organic chemistry.

For more than 200 years, the synthesis of organic molecules has remained one of the most important tasks in organic chemistry. The work of chemists has scientific and commercial implications that range from the production of aspirin to that of nylon. Unfortunately, it is a complex and time-consuming process to find success.

Synthetic organic chemistry is the science of building desired chemical structures from simpler parts. In order to achieve that aim, organic chemists often work by thinking backwards as much as they do forwards when designing a synthetic route. The concept of retrosynthesis, introduced by E. J. Corey in the 1960s and for which he was awarded the Nobel Prize in Chemistry in 1990, codified the way in which many chemists think.

Generally, the chemists look at a target molecule and try to identify its composition, and question which bonds could have been formed, and which atoms or chemical groups could have been added or transformed? Then, the process starts again, as researchers try to determine the reactions that could have led to the precursor molecule. The aim is to work back to easily available starting compounds, while balancing the factors that make a good synthesis, including the number of steps involved, the probable product yields of those steps, and how easy it is to use the chemistry involved. Organic chemists deal constantly with such questions, especially when making compounds for testing in drug-discovery programmes.

The challenge for organic chemists in fields such as chemistry, materials science, oil and gas, and life sciences is that there are hundreds of thousands of reactions and, while it is manageable to remember a few dozen in a narrow specialist's field, it's impossible to be an expert generalist. Designing materials for a specific demand is a complex task; a random mix-and-match of atomic building blocks could yield any one of an infinite number of possible compounds. Historically, the discovery of materials has involved a combination of chance, intuition and trial and error – but this could all be set to change thanks to artificial intelligence.

Since Corey's work in the 1960s, chemists have believed that a large and well-curated database of chemical transformations could be used as the basis for a programme that not only finds reactions, but also arranges them into plausible synthetic plans.

This dream has been frustrated by two fundamental problems. Firstly, computing hardware simply could not tackle the scale of the challenge. Secondly, the chemical literature is hard to define in terms that a software programme using 1s and 0s can understand:

›

Above: Bartosz Grzybowski, a chemist at the Ulsan National Institute of Science and Technology, in South Korea, and his team spent 15 years inputting more than 50,000 rules of organic chemistry into the system for the programme to draw on

given reactions will work for the type of compound for which they were claimed to work (most of the time), but only under certain conditions. In other words, discerning between terms such as 'may', 'might' or 'will' in scientific papers is as critically important as the temperature or other parameters of the reaction.

This is where machine learning and artificial intelligence enters the picture as it offers the possibility of training computers by using the properties of materials that we already know. Plus, artificial intelligence approaches consider all available data equally and find trends that a human researcher may miss due to bias towards a given interpretation.

A new AI tool developed by Marwin Segler, an organic chemist and artificial-intelligence researcher at the University of Münster, in Germany, and his colleagues, uses deep-learning neural networks to assimilate essentially all known single-step organic-chemistry reactions – about 12.4 million of them. This enables it to predict the chemical reactions that can be used in any single step. The tool repeatedly applies these neural networks in planning a multi-step synthesis, deconstructing the desired molecule until it ends up with the available starting reagents.

Segler and his team tested the pathways that the programme threw up in a double-blind trial, to see whether experienced chemists could tell the AI's synthesis pathways from those devised by humans. They showed 45 organic chemists from two institutes in China and Germany potential synthesis routes for nine molecules: one pathway suggested by the system and another devised by humans. The chemists had no preference for which was best.

Doing retrosynthesis, Grzybowski explains, is like playing chess: there are a number of basic moves. Yet during a game, each move opens up a new branch to a different outcome

Researchers have been trying to use computing power to plan organic chemical synthesis since the 1960s, with only limited success. But Segler's tool is one of several programmes developed in recent years that use AI to flag up potential reaction routes.

Chematica, the most well-known, was acquired by German pharmaceutical company Merck in May 2017. Bartosz Grzybowski, a chemist at the Ulsan National Institute of Science and Technology, in South Korea, and his team spent 15 years inputting more than 50,000 rules of organic chemistry into the system for the programme to draw on.

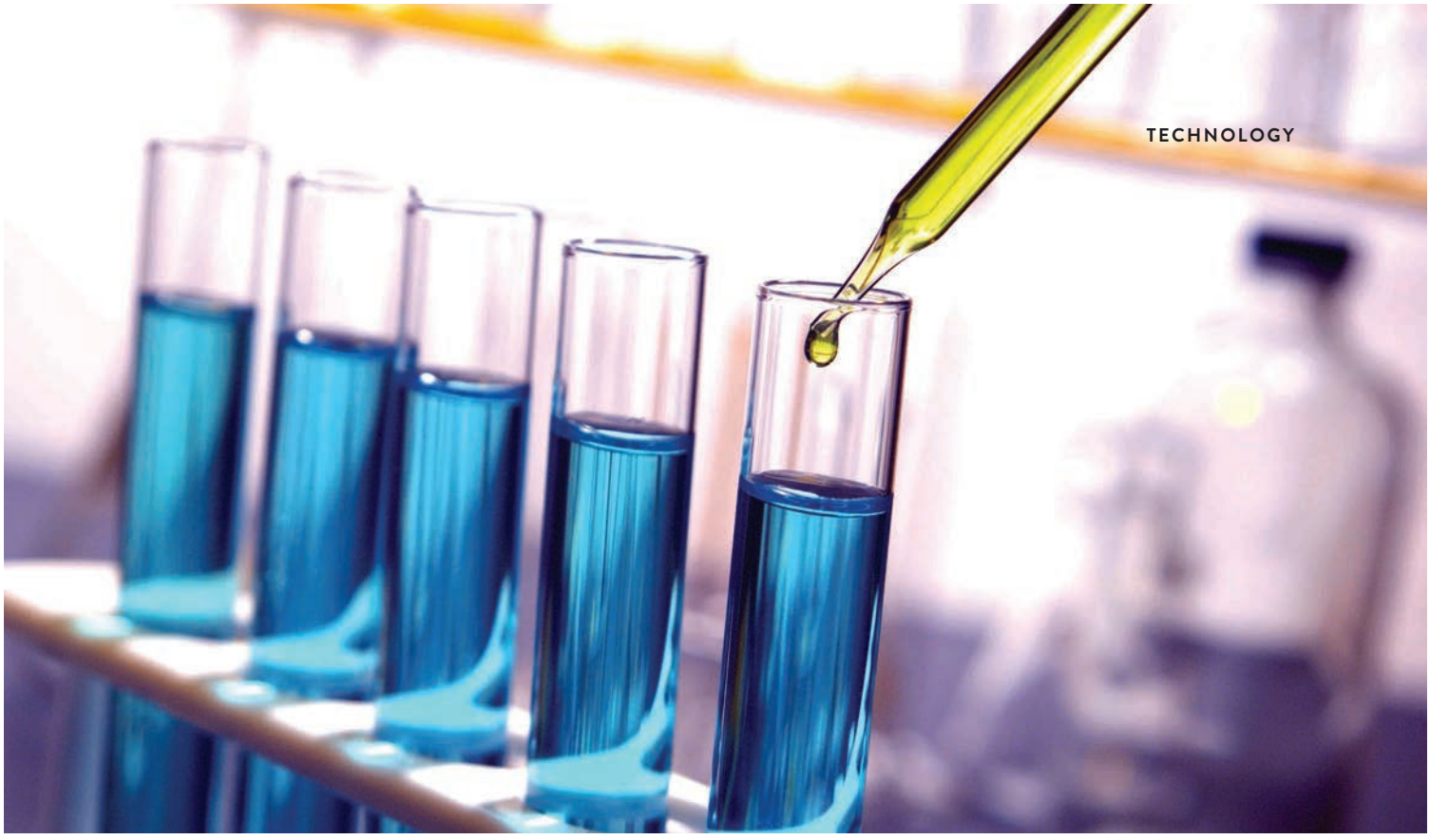
In December, Grzybowski reported that he had tested eight of his algorithm's suggested pathways in the laboratory, and that they all worked. "I'm very glad there is this revival of retrosynthesis, and welcome different approaches," he says.

Doing retrosynthesis, Grzybowski explains, is like playing chess: there are a number of basic moves. Yet during a game, each move opens up a new branch to a different outcome. After both players move, 400 possible chess board set-ups exist. After the second pair of turns, there are 197,742 possible games, and after three moves, 121 million.

However, in organic synthesis "the number of basic moves – basic reaction types – is just ginormous, in the tens of thousands", he says. After each synthetic step around 100 possible next steps become available, meaning the longer a route is the more enormous the number of possibilities becomes.

As Chematica doesn't give precise conditions for each reaction, there is still some trial and error when it comes to optimisation. However, to reflect the time and financial constraints of industry, the team limited itself to five attempts on each reaction and a maximum of 70 hours to complete each route.

The implications of using of machine learning and artificial intelligence in synthetic organic



chemistry are staggering. Currently, pharmaceutical companies spend around $2.6 billion on developing a treatment, and nine out of ten candidate therapies fail somewhere between phase one trials and regulatory approval.

Yet change is happening fast. Pfizer is using IBM Watson, a system that uses machine learning, to power its search for immuno-oncology drugs. Sanofi has signed a deal to use UK start-up Exscientia's artificial intelligence platform to hunt for metabolic disease therapies, and Roche subsidiary Genentech is using an AI system from GNS Healthcare in Cambridge, Massachusetts, to help drive the multinational company's search for cancer treatments.

This does not necessarily mean that all machine-suggested routes will work in the laboratory – but, as organic chemists know to their sorrow, many routes designed by humans fail there, too.

Above: In organic synthesis the number of basic moves – basic reaction types – is in the tens of thousands

Left: The concept of retrosynthesis was introduced by E. J. Corey in the 1960s. He was awarded the Nobel Prize in Chemistry in 1990