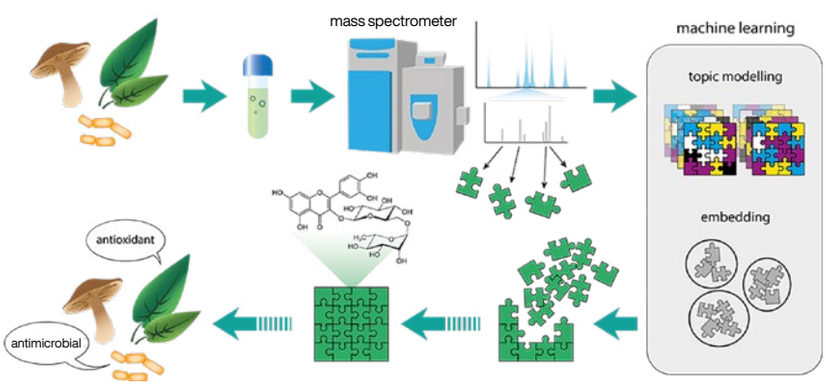
THE UNIVERSAL LANGUAGE OF THE MOLECULE

All life speaks a universal language made up of smells, colours and tastes: the language of the molecule, says Justin van der Hooft, assistant professor in the Bioinformatics group. But universal as this language may be, we don't understand it at all. Van der Hooft wants to change that through his research in metabolomics.

Text and illustration Stijn Schreven ◆ Photo Eric Scholten



What is metabolomics?

Van der Hooft: 'It is the fourth in the set of "omics", along with genomics, transcriptomics and proteomics. The first of these studies DNA, the second the RNA transcripts of DNA, and the third the proteins made that way. Metabolomics is the study of mixtures of small molecules such as glucose. Such mixtures include what is excreted by bacteria, as well as plant extracts such as coffee and tea, and urine specimens. The ultimate aim is to understand the language of these mixtures, thus learning about functions, active substances, and in the case of urine, for example, about a person's diet and health.'

What fascinates you about metabolomics?

'I am fascinated by the variety of molecular forms that exist. That such small compounds can have such a big impact. Sometimes the structure of two molecules only differs in a single group that points in a different direction, but that can make a huge difference to its effect or smell. An intriguing example is the mirror images of the menthol molecule: one smells like peppermint and the other is bitter. I would like to understand that. Another thing that makes this field interesting is that it is multidisciplinary: you are using analytical chemistry, statistics, machine learning and chemical informatics.'

You want to decode the language of small molecules. How?

'My group is working on computational metabolomics: we develop the tools for analysing metabolomics data. Those datasets come from specialized instruments such as mass spectrometers. A mixture of molecules is put into the machine, where they collide with inert gases and disintegrate into fragments. We see these fragments reflected in peaks in a spectrum. The location and height of a peak tell us something about the form of the fragment and the amount of it in the molecule. The question then is which fragments they are, and which molecules they formed between them. So we can't immediately say which molecules were in the mixture; we must first put the fragments together like a puzzle and see how they fit together. A molecule can produce lots of different fragments and the same fragment can occur as a building block in several molecules.'

How do you solve such puzzles?

'That used to be done manually. When I was doing my PhD, I figured out which

fragments belonged to which molecule. Worldwide, such studies now form a databank of 16,000 compounds that have been thoroughly studied – a collection of solved puzzles. The collection is growing steadily, but slowly, because the research is time-consuming. We have recently started making use of machine learning, thus automizing and speeding up the process. You give the computer the data and the labels saying what's what, and wish it luck. The computer learns to recognize the patterns itself. We use two methods of machine learning, both inspired by text mining.'

Text mining? Which method are you referring to?

'Topic modelling aims to extract topics from a text based on the words that occur in it most frequently. In metabolomics, the fragments of molecules are identified on the basis of the fragments that often appear together in spectra. We are also developing new techniques based on word-embedding, which looks at the context of words to decide whether sentences are similar to each other. For example: "I like coffee and cookies" and "I like a cappuccino and cake". The words are different, but the sentences are very close in meaning. Similarly, in metabolomics we try to identify chemical classes (the meaning) from the fragments (the words) without having to put together the molecules as a whole (the sentences). Examples of chemical classes are flavonoids and alkaloids. It's like finding the corner pieces and edges of the jigsaw puzzle: then you've got the structure of the molecule, which helps you solve the rest of the puzzle.'

How far have you got with developing these tools?

'At the start of this year, we used the first method to apply word-embedding in metabolomics. At the moment, machine-learning studies are popping up all over the place and a new publication comes out every month. AlphaFold 2 was launched recently for proteins. This is a machine-learning technique that can predict the 3D structures of proteins with 15 to 20 per cent more accuracy. Instead of months of laboratory work, it sometimes takes just 10 minutes to find out what a protein looks like. It is only a matter of time before there is some such breakthrough in metabolomics.'

Where do you want to go with this research, ultimately?

'My group focuses on the structures and functions of natural products – to find new antibiotics, for example. Ultimately, I want to solve those molecular puzzles in order to understand why an ecosystem works the way it does, and what language is spoken in it. For example: what functions does a plant extract with particular flavonoids have? Once we know that, we can start steering things. By introducing the right bacteria and fungi, for instance, you can make a soil tolerant of salt stress, drought or heat, so that it retains functions and plants continue to grow. Currently, that is still a long way off, though.' ■