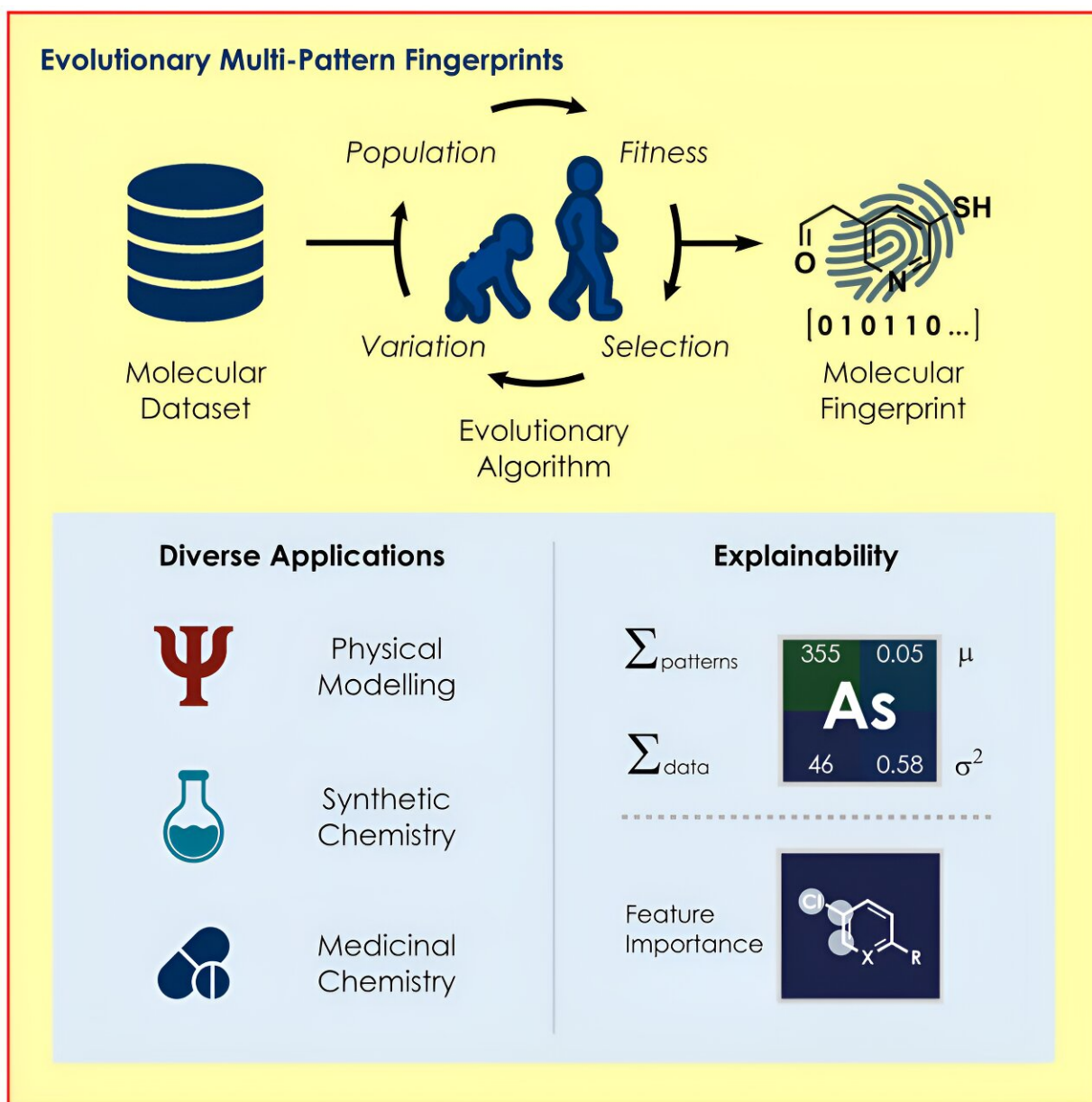# Evolutionary algorithm generates tailored 'molecular fingerprints'

May 10 2024, by Christina Hoppenbrock

A team led by Prof Frank Glorius from the Institute of Organic Chemistry at the University of Münster has developed an evolutionary algorithm that identifies the structures in a molecule that are particularly relevant for a respective question and uses them to encode the properties of the molecules for various machine-learning models.

The method is also suitable for the machine prediction of quantum chemical properties and the toxicity of molecules. It can be applied to any molecular data set and does not require expert knowledge of the underlying relationships.

Artificial intelligence and machine learning are becoming more and more relevant in everyday life—and the same goes for chemistry. Organic chemists, for example, are interested in how machine learning can help discover and synthesize new molecules that are effective against diseases or are useful in other ways.

The new algorithm developed by Glorius's team searches for optimal molecular representations based on the principles of evolution, using mechanisms such as reproduction, mutation and selection. Depending on the model and the given question, customized "molecular fingerprints" are created, which the chemists used in their study to predict chemical reactions with surprising accuracy.

The method, published in the journal *Chem*, is also suitable for predicting quantum chemical properties and the toxicity of molecules.

In order to use machine learning, researchers must first convert the molecules into a computer-readable form. Many research groups have

already tackled this problem, and consequently, there are various ways of performing this task. However, it is difficult to predict which of the available methods is best suited to answer a specific question—for example, to determine whether a chemical compound is harmful to humans.

The new algorithm is designed to help find the optimal molecular fingerprint in each case. To do this, the algorithm gradually selects the molecular fingerprints that achieve the best results in the prediction from many randomly generated molecular fingerprints.

"Following the example of nature, we use mutations, i.e. random changes to individual components of the fingerprints, or recombine components of two fingerprints," explains doctoral student Felix Katzenburg.

"In other studies, molecules are often described by quantifiable properties that have been selected and calculated by humans," adds Glorius.

"Since the algorithm we developed automatically identifies the relevant molecular structures, there are no systematic biases caused by human experts."

Another advantage is that the method of encoding makes it possible to understand why a model makes a certain prediction. For example, it is possible to draw conclusions about which parts of a molecule positively or negatively impact the prediction of how a reaction would play out, allowing researchers to change the relevant structures in a targeted manner.

The Münster team found that their new method did not always achieve the most optimal results.

"When considerable human expertise has gone into selecting particularly relevant molecular properties or very large amounts of data are available, other methods such as neural networks sometimes have the edge," says Katzenburg.

However, one of the study's primary goals was to develop a method for encoding molecules that can be applied to any molecular data set and does not require expert knowledge of the underlying relationships.

Provided by University of Münster