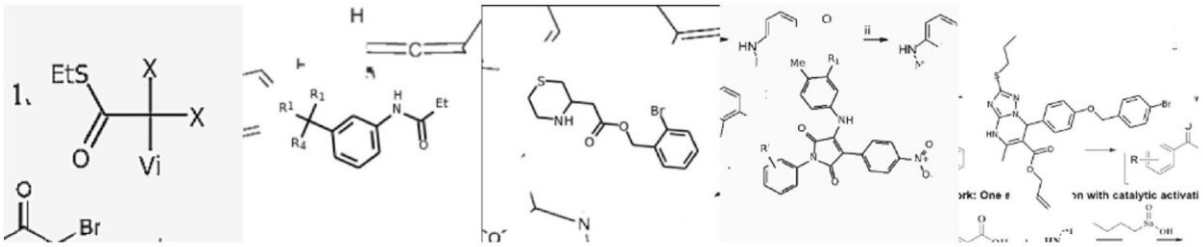
Researchers train neural network to recognize chemical formulas from research papers

February 14 2022



Examples of artificially generated templates for training neural networks to recognize actual chemical formulas. Credit: Ivan Khokhlov et al./Chemistry Methods

Researchers from Syntelly—a startup that originated at Skoltech—Lomonosov Moscow State University, and Sirius University have developed a neural network-based solution for automated recognition of chemical formulas on research paper scans. The study was published in *Chemistry–Methods*, a scientific journal of the European Chemical Society.

Humanity is entering the age of artificial intelligence. Chemistry, too, will be transformed by the modern methods of deep learning, which invariably require large amounts of qualitative data for neural network

training.

The good news is that chemical data "age well." Even if a certain compound was originally synthesized 100 years ago, information about its structure, properties and ways of synthesis remains relevant to this day. Even in our time of universal digitalization, it may well happen that an organic chemist turns to an original journal paper or thesis from a library collection—published as far back as early 20th century, say, in German—for information about a poorly studied molecule.

The bad news is there is no accepted standard way for presenting chemical formulas. Chemists customarily use many tricks in the way of shorthand notation for familiar chemical groups. The possible stand-ins for a tert-butyl group, for example, include "tBu," "t-Bu," and "tert-Bu." To make matters worse, chemists often use one template with different "placeholders" (R1, R2, etc.) to refer to many similar compounds, but those placeholder symbols might be defined anywhere: in the figure itself, in the running text of the article or supplements. Not to mention that drawing styles vary between journals and evolve with time, the personal habits of chemists differ, and conventions change. As a result, even an expert chemist at times finds themselves at a loss trying to make sense of a "puzzle" they found in some article. For a computer algorithm, the task appears insurmountable.

As they approached it, though, the researchers already had experience tackling similar problems using Transformer—a neural network originally proposed by Google for machine translation. Rather than translate text between languages, the team used this powerful tool to convert the image of a molecule or a molecular template to its textual representation. Such a representation is called Functional-Group-SMILES.

To the researchers' genuine surprise, the neural network proved capable

of learning nearly anything provided that the relevant depiction style was represented in the training data. That said, Transformer requires tens of millions of examples to train on, and collecting that many chemical formulas from research papers by hand is impossible. So instead of that, the team adopted another approach and created a data generator that produces examples of molecular templates by combining randomly selected molecule fragments and depiction styles.

"Our study is a good demonstration of the ongoing paradigm shift in the optical recognition of chemical structures. While prior research focused on molecular structure recognition per se, now that we have the unique capacities of Transformer and similar networks, we can instead dedicate ourselves to creating artificial sample generators that would imitate most of the existing styles of molecular template depiction. Our algorithm combines molecules, functional groups, fonts, styles, even printing defects, it introduces bits of additional molecules, abstract fragments, etc. Even a chemist has a hard time telling if the molecule came straight out of a real paper or from the generator," said the study's principal investigator Sergey Sosnin, who is the CEO of Syntelly, a startup founded at Skoltech.

The authors of the study hope that their method will constitute an important step toward an artificial intelligence system that would be capable of "reading" and "understanding" research papers to the extent that a highly qualified chemist would.

More information: Ivan Khokhlov et al, Image2SMILES: Transformer-Based Molecular Optical Recognition Engine, *Chemistry–Methods* (2022). DOI: 10.1002/cmtd.202100069

Provided by Skolkovo Institute of Science and Technology