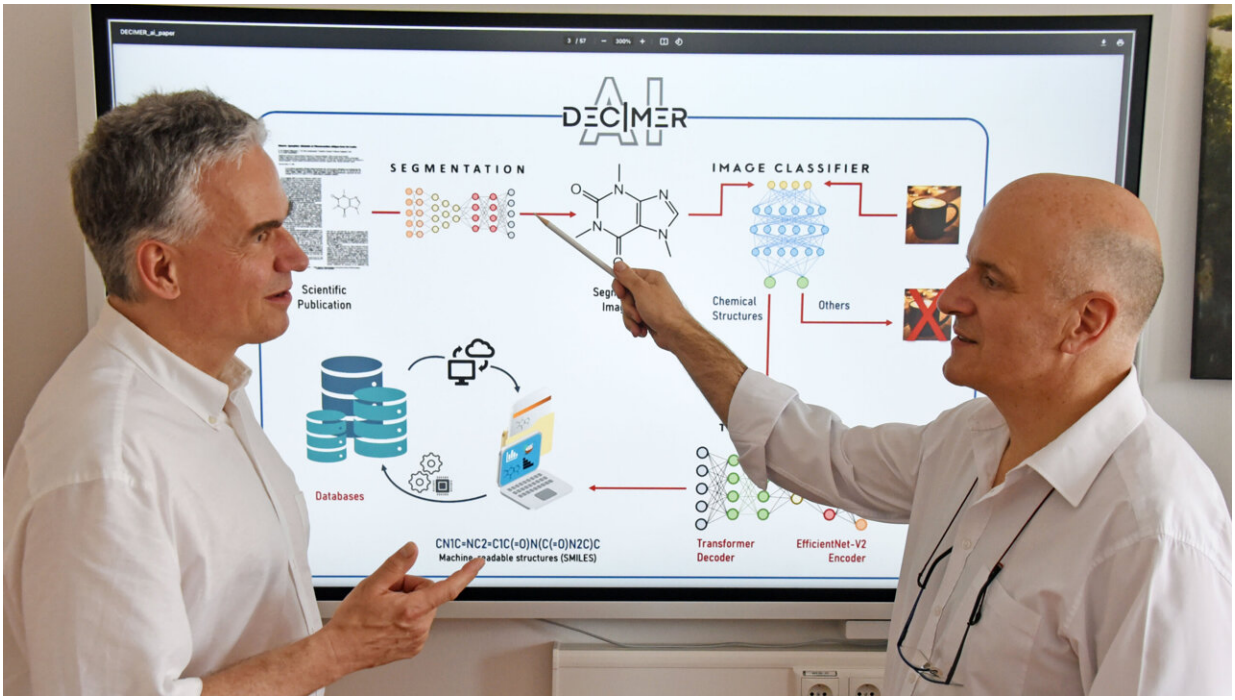
# Sharing chemical knowledge between human and machine

August 22 2023, by Ute Schönfelder




The team led by Prof. Dr Christoph Steinbeck (r.) and Prof. Dr Achim Zielesny has developed the AI tool DECIMER.ai, which researchers can use worldwide. Credit: Anne Günther/Uni Jena

Researchers from the University of Jena, the Westphalian University of Applied Sciences and the University of Chemistry and Technology Prague have developed a platform that uses artificial neural networks to translate chemical structural formulae into machine-readable form.

With this platform, they have created a tool with which information from scientific publications can be automatically fed into databases. Until now, this had to be done literally by hand and was time-consuming. In the current issue of *Nature Communications*, the team led by Prof. Christoph Steinbeck and Prof. Achim Zielesny presents the latest version of their tool, DECIMER.ai, which researchers can use worldwide.

Structural formulae show how chemical compounds are constructed, i.e., which atoms they consist of, how these are arranged spatially and how they are connected. Chemists can deduce from a structural formula, among other things, which molecules can react with each other and which cannot, how complex compounds can be synthesized or which natural substances could have a therapeutic effect because they fit together with target molecules in cells.

Developed in the 19th century, the representation of molecules as structural formulae has stood the test of time and is still used in every chemistry textbook. But what makes the chemical world intuitively comprehensible for humans is just a collection of black and white pixels for software. "To make the information from structural formulae usable in databases that can be searched automatically, they have to be translated into a machine-readable code," explains Steinbeck, professor for analytical chemistry, cheminformatics and chemometrics at the University of Jena.

And that is precisely what can be done using the artificial intelligence tool DECIMER, developed by the team led by Steinbeck and his colleague Zielesny from the Westphalian University of Applied Sciences. DECIMER stands for "deep learning for chemical image recognition." It is an open-source platform that is freely available to everyone on the Internet and can be used in a standard web browser. Scientific articles containing chemical structural formulae can be

uploaded there simply by dragging and dropping, and the AI tool will immediately get to work.

"First, the entire document is searched for images," explains Steinbeck. The algorithm then identifies the image information contained and classifies it according to whether it is a chemical structural formula or some other image. Finally, the structural formulae recognized are translated into the chemical structure code or displayed in a structure editor, so that they can be further processed. "This step is the core of the project and the real achievement," adds Steinbeck.

In this way, the chemical structural formula for the caffeine molecule becomes the machine-readable structure code CN1C=NC2=C1C(=O)N(C(=O)N2C)C. This can then be uploaded directly into a database and linked to further information on the molecule.

To develop DECIMER, the researchers used modern AI methods that have only recently become established and are also used, for example, in the Large Language Models (such as ChatGPT) that are currently the subject of much discussion. To train its AI tool, the team generated structural formulas from the existing machine-readable databases and used them as training data—some 450 million structural formulas to date. In addition to researchers, companies are also already using the AI tool, for example to transfer structural formulae from patent specifications into databases.

Steinbeck and Zielesny came up with the idea of developing an AI tool for decoding chemical images a few years ago. The two chemists were interested the development of AI methods in connection with the millennia-old Asian board game Go. In 2016, together with millions of people around the world, they watched the spectacular tournament between the best Go player at the time, the South Korean Lee Sedol, and

the computer software AlphaGo, which the machine won 4:1.

"It was a bolt from the blue that showed us how powerful AI can be," Steinbeck recalls. Until then, it had been considered practically unthinkable that an algorithm could rival human creativity and intuition in this game.

"When, a little later, an AI tool developed quasi-superhuman playing strength by not being trained laboriously through countless sessions of human games—as was still the case with AlphaGo—but simply through the process of the system playing against itself again and again, and optimizing its playing style as it did so, we realized that these new methods could also solve other very complex problems with enough training data. We wanted to use that for our research area."

## Making scientific information sustainably usable

With DECIMER, Steinbeck and his team hope at some point to be able to machine-read all chemical literature of interest to them, going back to the 1950s, and translate it into open databases.

After all, a key concern for Steinbeck, also the coordinator of the National Research Data Infrastructure for Chemistry in Germany, is to sustainably secure existing knowledge and make it available to the global scientific community.

The DECIMER AI tool is available online: https://decimer.ai

**More information:** Kohulan Rajan et al, DECIMER.ai: an open platform for automated optical chemical structure identification, segmentation and recognition in scientific publications, *Nature Communications* (2023). DOI: 10.1038/s41467-023-40782-0

Provided by Friedrich-Schiller-Universität Jena