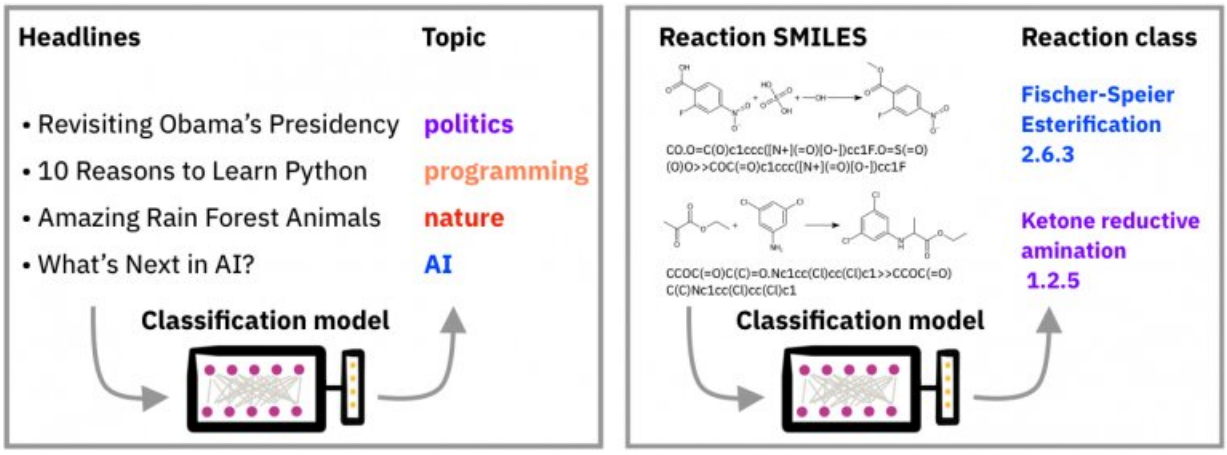
IBM RXN: New AI model boosts mapping of chemical reactions

February 8 2021, by Philippe Schwaller, Teodoro Laino, and Alain Vaucher



Classification models: Analogy between blog headline topics and chemical reaction classes. The chemical reactions are represented as text using SMILES.

Just like an astronomer investigates outer space, a chemist explores chemical space—a theoretical territory with all possible known (and unknown) chemical compounds. Researchers estimate chemical space to contain up to 10^{180} compounds—more than twice the magnitude of the number of atoms in the universe. Currently, the largest public database of molecules synthesized so far called PubChem contains just over 100 million, or roughly 10^{8}. Throw in the chemical reactions between molecules, and you've got an even larger chemical reaction space.

It's easy to see why the vastness of chemical reaction space is overwhelming to even the most seasoned chemists.

While astronomers have powerful telescopes to help them, chemists often rely on just their own experience and intuition. This is partly why it takes months or even years to discover new drugs and materials—there's simply nothing guiding them through the chemical galaxy.

Recently, though, chemists have been relying more and more on artificial intelligence (AI) as a navigation tool. AI has the potential to lead chemists to new frontiers, as it can explore chemical space (and chemical reaction space) faster than humans and help them not only to find molecules that might otherwise be overlooked but also to better understand the transformations they undergo.

Our recent Nature Machine Intelligence paper "Mapping the Space of Chemical Reactions Using Attention-Based Neural Networks", by IBM Research Europe and the University of Bern, investigates deep learning models to classify chemical reactions and visualizes the chemical reaction space. With this mapping, chemists should be able to query large datasets based on common features, retrieve similar chemical reaction entries, and explore new chemistry based on what is known on large datasets of chemical reactivity.

Chemistry as a language

In the paper, we detail our web-based app called IBM RXN for Chemistry. To create it, we were inspired by… language. Indeed, organic chemistry and language have much in common. For example, the smallest change to the syntax or tense of a word can give a phrase a whole new meaning, similar to how stereochemistry can turn thalidomide into either a medication or a deadly poison.

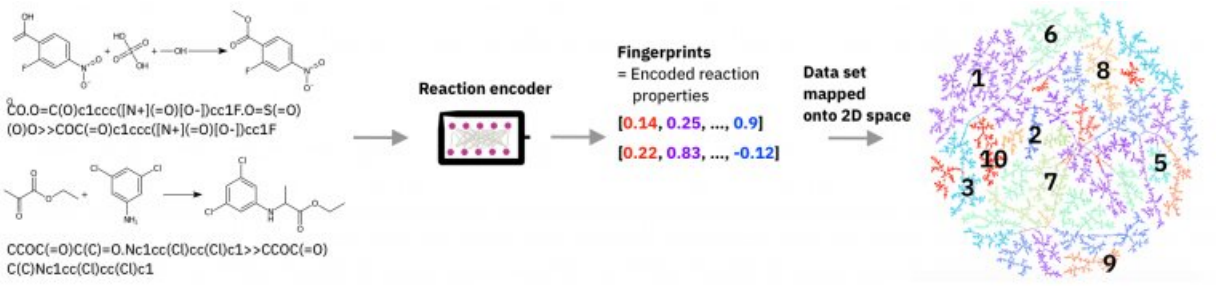
The app is based on the notion of chemistry as a language. It applies sequence-to-sequence models, used for translations from one language to another, to do reaction outcome predictions and synthesis planning. To achieve this, the molecules are encoded as sequences of characters called Simplified Molecular Input Line Entry System (SMILES) notations.

While the reaction prediction and synthesis planning models behind RXN for Chemistry prove beneficial to the process of drug and material discovery, they are typically black box models. We wanted to make the predicted chemical reactions more explainable and accessible to chemists, so we began experimenting with attention-based neural networks with the aim to map the space of chemical reactions.

Automating classification

Organic reactions are usually assigned to classes containing reactions with similar reagents and mechanisms. Reaction classes enable efficient communication between chemists. However, the classification process in a large data set is a tedious and time-consuming task. It requires the identification of the reacting atoms and the distinction between reactants and reagents.

Our idea was to automate the classification of reaction data sets using neural networks, which would allow chemists to explore reactions and detect patterns that lead to new discoveries. So, continuing to treat organic chemistry like a language, we used a text-based representation for the chemical reactions, and language AI models like BERT—a transformer-based machine learning technique for natural language processing—and trained them to classify the reactions.



Reaction fingerprints. Encoding of chemical reactions to create visualizations of chemical reaction space. Credit: IBM

What makes our deep learning models unique is that they do not rely on the formulation of specific rules, which require every reaction to be properly atom-mapped. Instead, they learn the atomic motifs that differentiate reactions from different classes, starting from raw reaction SMILES without reactant-reagent role annotations.

This may still sound a little complicated, so let's use an analogy. Say you have a headline, but no article. This headline gives you a sense of the topic, but you can only make a general assumption about the message of the content. And while you may have access to other headlines, finding the message you're looking for in this abyss of headlines seems impossible. But if an AI model classifies it with other similar headlines, you could begin to detect a common subject matter such as politics, sports, or fashion.

Reaction fingerprints

Later, we realized that we could use the embedded information from our AI classification models to create "reaction fingerprints." Basically, our model transforms any chemical reaction into a continuous vector, which

gives chemists the possibility to map chemical reaction space and allows them to easily inquire about similar reactions. These data-driven reaction fingerprints unlock the possibility of mapping the reaction space without knowing the reaction centers or the reactant-reagent split. They also enable efficient searches on the nearest neighboring reaction data sets containing millions of reactions.

Coming back to the headline analogy, the embedded information (the fingerprint) that comes out of grouping like headlines together is presented as a graph embedded in two-dimensional space, which would allow you to look deeper into the specifics, such as which sport the original headline refers to. With this information, you could easily find other headlines related to the one you're about to write. Along the same lines, chemists could use this information to find related reactions that might serve as a starting point for their next experiment.

Our models reached a classification accuracy of 98.9 percent on two different reaction data sets. And our reaction fingerprints can be used to almost perfectly cluster chemical reaction space. Essentially, we have developed a new way of exploring chemical reaction data, opening a chemical galaxy highway. Let the expedition begin!

Access the interactive reaction atlas at RXN4Chemistry on github.

More information: Philippe Schwaller et al. Mapping the space of chemical reactions using attention-based neural networks, *Nature Machine Intelligence* (2021). DOI: 10.1038/s42256-020-00284-w

This story is part of Science X Dialog*, where researchers can report findings from their published research articles.* Visit this page *for information about ScienceX Dialog and how to participate.*

Provided by IBM

Citation: IBM RXN: New AI model boosts mapping of chemical reactions (2021, February 8) retrieved 9 July 2025 from https://sciencex.com/news/2021-02-ibm-rxn-ai-boosts-chemical.html