

Generating new molecules with graph grammar

An efficient machine-learning method uses chemical knowledge to create a learnable grammar with production rules to build synthesizable monomers and polymers.

Lauren Hinkel | MIT-IBM Watson AI Lab

April 1, 2022



MIT and IBM researchers have use a generative model with a graph grammar to create new molecules belonging to the same class of compound as the training set.

Chemical engineers and materials scientists are constantly looking for the next revolutionary material, chemical, and drug. The rise of machine-learning approaches is expediting the discovery process, which could otherwise take years. “Ideally, the goal is to train a machine-learning model on a few existing chemical samples and then allow it to produce as many manufacturable molecules of the same class as possible, with

predictable physical properties,” says Wojciech Matusik, professor of electrical engineering and computer science at MIT. “If you have all these components, you can build new molecules with optimal properties, and you also know how to synthesize them. That’s the overall vision that people in that space want to achieve”

However, current techniques, mainly deep learning, require extensive datasets for training models, and many class-specific chemical datasets contain a handful of example compounds, limiting their ability to generalize and generate physical molecules that could be created in the real world.

Now, a new paper from researchers at MIT and IBM tackles this problem using a generative graph model to build new synthesizable molecules within the same chemical class as their training data. To do this, they treat the formation of atoms and chemical bonds as a graph and develop a graph grammar — a linguistics analogy of systems and structures for word ordering — that contains a sequence of rules for building molecules, such as monomers and polymers. Using the grammar and production rules that were inferred from the training set, the model can not only reverse engineer its examples, but can create new compounds in a systematic and data-efficient way. “We basically built a language for creating molecules,” says Matusik “This grammar essentially is the generative model.”

Matusik’s co-authors include MIT graduate students Minghao Guo, who is the lead author, and Beichen Li as well as Veronika Thost, Payal Das, and Jie Chen, research staff members with IBM Research. Matusik, Thost, and Chen are affiliated with the MIT-IBM Watson AI Lab. Their method, which they’ve called data-efficient graph grammar (DEG), will be presented at the International Conference on Learning Representations.

“We want to use this grammar representation for monomer and polymer generation, because this grammar is explainable and expressive,” says Guo. “With only a few number of the production rules, we can generate many kinds of structures.”

A molecular structure can be thought of as a symbolic representation in a graph — a string of atoms (nodes) joined together by chemical bonds (edges). In this method, the researchers allow the model to take the chemical structure and collapse a substructure of the molecule down to one node; this may be two atoms connected by a bond, a short sequence of bonded atoms, or a ring of atoms. This is done repeatedly, creating the production rules as it goes, until a single node remains. The rules and grammar then could

be applied in the reverse order to recreate the training set from scratch or combined in different combinations to produce new molecules of the same chemical class.

“Existing graph generation methods would produce one node or one edge sequentially at a time, but we are looking at higher-level structures and, specifically, exploiting chemistry knowledge, so that we don't treat the individual atoms and bonds as the unit. This simplifies the generation process and also makes it more data-efficient to learn,” says Chen.

Further, the researchers optimized the technique so that the bottom-up grammar was relatively simple and straightforward, such that it fabricated molecules that could be made.

“If we switch the order of applying these production rules, we would get another molecule; what's more, we can enumerate all the possibilities and generate tons of them,” says Chen. “Some of these molecules are valid and some of them not, so the learning of the grammar itself is actually to figure out a minimal collection of production rules, such that the percentage of molecules that can actually be synthesized is maximized.” While the researchers concentrated on three training sets of less than 33 samples each — acrylates, chain extenders, and isocyanates — they note that the process could be applied to any chemical class.


To see how their method performed, the researchers tested DEG against other state-of-the-art models and techniques, looking at percentages of chemically valid and unique molecules, diversity of those created, success rate of retrosynthesis, and percentage of molecules belonging to the training data's monomer class.

“We clearly show that, for the synthesizability and membership, our algorithm outperforms all the existing methods by a very large margin, while it's comparable for some other widely-used metrics,” says Guo. Further, “what is amazing about our algorithm is that we only need about 0.15 percent of the original dataset to achieve very similar results compared to state-of-the-art approaches that train on tens of thousands of samples. Our algorithm can specifically handle the problem of data sparsity.”

In the immediate future, the team plans to address scaling up this grammar learning process to be able to generate large graphs, as well as produce and identify chemicals with desired properties.

Down the road, the researchers see many applications for the DEG method, as it's adaptable beyond generating new chemical structures, the team points out. A graph is a very flexible representation, and many entities can be symbolized in this form — robots, vehicles, buildings, and electronic circuits, for example. “Essentially, our goal is to build up our grammar, so that our graphic representation can be widely used across many different domains,” says Guo, as “DEG can automate the design of novel entities and structures,” says Chen.

This research was supported, in part, by the MIT-IBM Watson AI Lab and Evonik.

 [Paper: "Data-Efficient Graph Grammar Learning for Molecular Generation"](#)