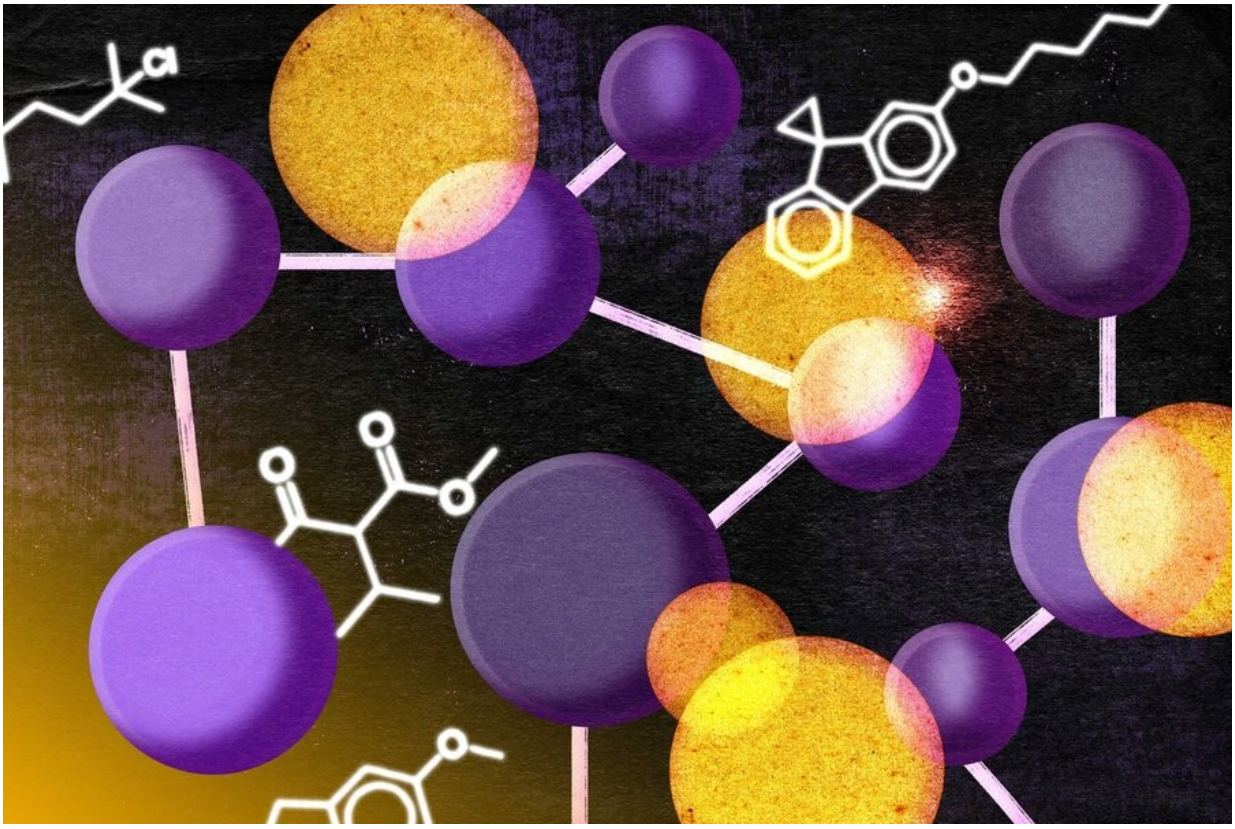
This AI system only needs a small amount of data to predict molecular properties

July 7 2023, by Adam Zewe



Researchers from MIT and the MIT-Watson AI Lab have developed a unified framework that uses machine learning to simultaneously predict molecular properties and generate new molecules using only a small amount of data for training. Credit: Jose-Luis Olivares/MIT

Discovering new materials and drugs typically involves a manual, trial-

and-error process that can take decades and cost millions of dollars. To streamline this process, scientists often use machine learning to predict molecular properties and narrow down the molecules they need to synthesize and test in the lab.

Researchers from MIT and the MIT-Watson AI Lab have developed a new, unified framework that can simultaneously predict molecular properties and generate new molecules much more efficiently than these popular deep-learning approaches.

To teach a machine-learning model to predict a molecule's biological or mechanical properties, researchers must show it millions of labeled molecular structures—a process known as training. Due to the expense of discovering molecules and the challenges of hand-labeling millions of structures, large training datasets are often hard to come by, which limits the effectiveness of machine-learning approaches.

By contrast, the system created by the MIT researchers can effectively predict molecular properties using only a small amount of data. Their system has an underlying understanding of the rules that dictate how building blocks combine to produce valid molecules. These rules capture the similarities between molecular structures, which helps the system generate new molecules and predict their properties in a data-efficient manner.

This method outperformed other machine-learning approaches on both small and large datasets, and was able to accurately predict molecular properties and generate viable molecules when given a dataset with fewer than 100 samples.

"Our goal with this project is to use some data-driven methods to speed up the discovery of new molecules, so you can train a model to do the prediction without all of these cost-heavy experiments," says lead author

Minghao Guo, a computer science and electrical engineering (EECS) graduate student.

Guo's co-authors include MIT-IBM Watson AI Lab research staff members Veronika Thost, Payel Das, and Jie Chen; recent MIT graduates Samuel Song '23 and Adithya Balachandran '23; and senior author Wojciech Matusik, a professor of electrical engineering and computer science and a member of the MIT-IBM Watson AI Lab, who leads the Computational Design and Fabrication Group within the MIT Computer Science and Artificial Intelligence Laboratory (CSAIL). The research will be presented at the International Conference for Machine Learning.

Learning the language of molecules

To achieve the best results with machine-learning models, scientists need training datasets with millions of molecules that have similar properties to those they hope to discover. In reality, these domain-specific datasets are usually very small. So, researchers use models that have been pretrained on large datasets of general molecules, which they apply to a much smaller, targeted dataset. However, because these models haven't acquired much domain-specific knowledge, they tend to perform poorly.

The MIT team took a different approach. They created a machine-learning system that automatically learns the "language" of molecules—what is known as a molecular grammar—using only a small, domain-specific dataset. It uses this grammar to construct viable molecules and predict their properties.

In language theory, one generates words, sentences, or paragraphs based on a set of grammar rules. You can think of a molecular grammar the same way. It is a set of production rules that dictate how to generate molecules or polymers by combining atoms and substructures.

Just like a language grammar, which can generate a plethora of sentences using the same rules, one molecular grammar can represent a vast number of molecules. Molecules with similar structures use the same grammar production rules, and the system learns to understand these similarities.

Since structurally similar molecules often have similar properties, the system uses its underlying knowledge of molecular similarity to predict properties of new molecules more efficiently.

"Once we have this grammar as a representation for all the different molecules, we can use it to boost the process of property prediction," Guo says.

The system learns the production rules for a molecular grammar using reinforcement learning—a trial-and-error process where the model is rewarded for behavior that gets it closer to achieving a goal.

But because there could be billions of ways to combine atoms and substructures, the process to learn grammar production rules would be too computationally expensive for anything but the tiniest dataset.

The researchers decoupled the molecular grammar into two parts. The first part, called a metagrammar, is a general, widely applicable grammar they design manually and give the system at the outset. Then it only needs to learn a much smaller, molecule-specific grammar from the domain dataset. This hierarchical approach speeds up the learning process.

Big results, small datasets

In experiments, the researchers' new system simultaneously generated viable molecules and polymers, and predicted their properties more

accurately than several popular machine-learning approaches, even when the domain-specific datasets had only a few hundred samples. Some other methods also required a costly pretraining step that the new system avoids.

The technique was especially effective at predicting physical properties of polymers, such as the glass transition temperature, which is the temperature required for a material to transition from solid to liquid. Obtaining this information manually is often extremely costly because the experiments require extremely high temperatures and pressures.

To push their approach further, the researchers cut one training set down by more than half—to just 94 samples. Their model still achieved results that were on par with methods trained using the entire dataset.

"This grammar-based representation is very powerful. And because the grammar itself is a very general representation, it can be deployed to different kinds of graph-form data. We are trying to identify other applications beyond chemistry or material science," Guo says.

In the future, they also want to extend their current molecular grammar to include the 3D geometry of molecules and polymers, which is key to understanding the interactions between polymer chains. They are also developing an interface that would show a user the learned grammar production rules and solicit feedback to correct rules that may be wrong, boosting the accuracy of the system.

More information: Paper: "Grammar-Induced Geometry for Data-Efficient Molecular Property Prediction" openreview.net/pdf?id=SGQi3LgFnqj

Provided by Massachusetts Institute of Technology