Allowing machine learning to ask questions can make it smarter

July 27 2023, by Michaela Kane

Credit: AI-generated image (disclaimer)

Biomedical engineers at Duke University have demonstrated a new method to significantly improve the effectiveness of machine learning models searching for new molecular therapeutics when using just a fraction of the available data. By working with an algorithm that actively identifies gaps in datasets, researchers can, in some cases, more than

double their accuracy.

This new approach could make it easier for scientists to identify and classify molecules with characteristics that could be useful for the development of new drug candidates and other materials.

This work appeared in the journal *Digital Discovery* published by the Royal Society of Chemistry on June 23.

Machine learning algorithms are increasingly used to identify and predict the properties of small molecules such as drug candidates and other compounds. While there have been significant advances in both computing power and machine learning algorithms, their abilities are currently limited by the existing datasets used to train them, which are far from perfect.

One of the main issues involves bias in the data. This occurs when there is a significant number of datapoints that showcase one property far more than another, like a molecule's potential ability to inhibit a specific protein or characteristics about its structure.

"It's like if you trained an algorithm to distinguish pictures of dogs and cats, but you gave it one billion photos of dogs to learn from and only one hundred photos of cats," explained Daniel Reker, an assistant professor of biomedical engineering at Duke University. "The algorithm will get so good at identifying dogs that everything will start to look like a dog, and it will forget everything else in the world."

This is a particularly problematic issue for drug discovery and development, where scientists often deal with datasets where more than 99% of the tested compounds are "ineffective," and only a small fraction of molecules are labeled as potentially useful.

To counter this issue, researchers use a process known as data subsampling, where their algorithm learns from a small but (hopefully) representative subset of the data. While this process can remove bias by giving the model an equal number of examples to learn from, it can also cut out key data points and negatively impact an algorithm's overall accuracy. To compensate, researchers have developed hundreds of subsampling techniques to limit the amount of lost information.

But Reker and his collaborators wanted to explore if a technique known as active machine learning could resolve this long-standing issue.

"With active machine learning, the algorithm is essentially able to ask questions or request more information if it is confused or senses a gap in the data, rather than passively sifting through it," said Reker. "This makes active-learning models very efficient at predicting performance."

Typically, Reker and other researchers apply active learning algorithms to generate new data for example to identify new drugs, but Reker and his team wanted to explore what happens if the algorithm is let loose on existing datasets. While this subsampling application of active machine learning had been explored in other research, Reker and his team were the first to test the algorithm on molecular biology and drug development.

To test the efficiency of their active subsampling approach, the team compiled datasets of molecules with different characteristics, including molecules that could cross the blood-brain barrier, molecules that could inhibit a protein associated with Alzheimer's disease, and compounds that have been shown to inhibit HIV replication. They then tested their active-learning algorithm against models that learned from the full dataset and against 16 state-of-the-art subsampling strategies.

The team showed that active subsampling was able to identify and

predict molecular characteristics more accurately than each of the standard subsampling strategies, and, most significantly, was up to 139 percent more effective than the algorithm that trained on the full dataset in some cases. Their model was also able to accurately adjust to mistakes in the data, indicating that it could be especially useful for low-quality datasets.

But most surprisingly, the team discovered the ideal amount of data to use was much lower than expected, in some cases requiring only 10% of the available data.

"There is a point where the active-subsampling model collects all the information it needs, and if you add more data, it's detrimental to performance," explained Reker. "That problem was especially interesting to us, because it hints that there's an inflection point where more information is no longer helpful, even in a subsample."

While Reker and his team hope to examine this inflection point in future work, they also plan to use this new approach to identify new molecules for potential therapeutic targets. Because active machine learning is becoming popular in many different research areas, the team is optimistic their work will help scientists better understand this algorithm and its robustness to errors in the data.

"Not only does this approach boost machine learning performance, but it can also reduce data storage needs and costs because it's working with a more refined dataset," said Reker. "This makes machine learning more reproducible, accessible and powerful for everyone."

More information: Yujing Wen et al, Improving molecular machine learning through adaptive subsampling with active learning, *Digital Discovery* (2023). DOI: 10.1039/D3DD00037K

Provided by Duke University