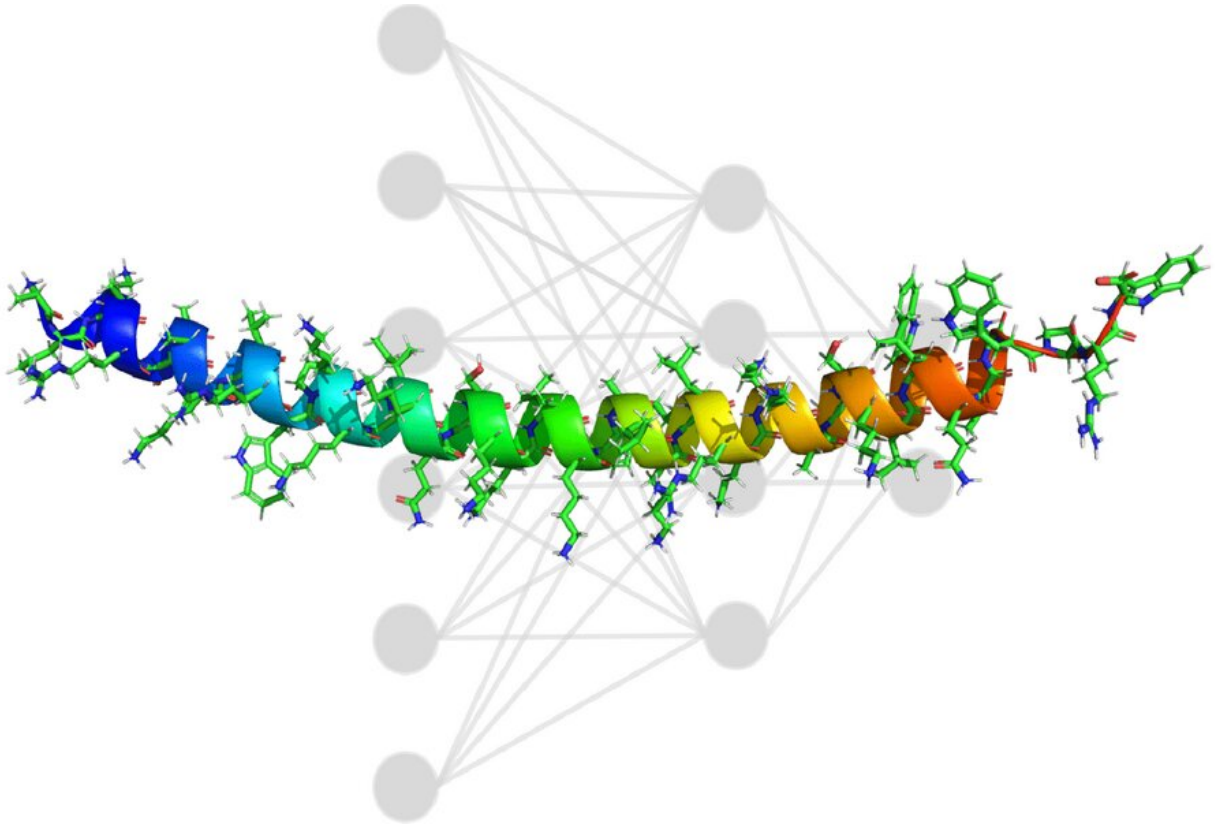
Machine learning discovers new sequences to boost drug delivery

August 10 2021



MIT researchers combined experimental chemistry with artificial intelligence to discover non-toxic, highly-active peptides that can be attached to phosphorodiamidate morpholino oligomers (PMO) to aid drug delivery. By developing these novel sequences, researchers hope to rapidly accelerate the development of gene therapies for Duchenne muscular dystrophy and other diseases. Credit: Massachusetts Institute of Technology

Duchenne muscular dystrophy (DMD), a rare genetic disease usually diagnosed in young boys, gradually weakens muscles across the body until the heart or lungs fail. Symptoms often show up by age 5; as the disease progresses, patients lose the ability to walk around age 12. Today, the average life expectancy for DMD patients hovers around 26.

It was big news, then, when Cambridge, Massachusetts-based Sarepta Therapeutics announced in 2019 a breakthrough drug that directly targets the mutated gene responsible for DMD. The therapy uses antisense phosphorodiamidate morpholino oligomers (PMO), a large synthetic molecule that permeates the cell nucleus in order to modify the dystrophin gene, allowing for production of a key protein that is normally missing in DMD patients. "But there's a problem with PMO by itself. It's not very good at entering cells," says Carly Schissel, a Ph.D. candidate in MIT's Department of Chemistry.

To boost delivery to the nucleus, researchers can affix cell-penetrating peptides (CPPs) to the drug, thereby helping it cross the cell and nuclear membranes to reach its target. Which peptide sequence is best for the job, however, has remained a looming question.

MIT researchers have now developed a systematic approach to solving this problem by combining experimental chemistry with artificial intelligence to discover nontoxic, highly-active peptides that can be attached to PMO to aid delivery. By developing these novel sequences, they hope to rapidly accelerate the development of gene therapies for DMD and other diseases.

Results of their study have now been published in the journal *Nature Chemistry* in a paper led by Schissel and Somesh Mohapatra, a Ph.D. student in the MIT Department of Materials Science and Engineering, who are the lead authors. Rafael Gomez-Bombarelli, assistant professor of materials science and engineering, and Bradley Pentelute, professor of

chemistry, are the paper's senior authors. Other authors include Justin Wolfe, Colin Fadzen, Kamela Bellovoda, Chia-Ling Wu, Jenna Wood, Annika Malmberg, and Andrei Loas.

"Proposing new peptides with a computer is not very hard. Judging if they're good or not, this is what's hard," says Gomez-Bombarelli. "The key innovation is using machine learning to connect the sequence of a peptide, particularly a peptide that includes non-natural amino acids, to experimentally-measured biological activity."

Dream data

CPPs are relatively short chains, made up of between five and 20 amino acids. While one CPP can have a positive impact on drug delivery, several linked together have a synergistic effect in carrying drugs over the finish line. These longer chains, containing 30 to 80 amino acids, are called miniproteins.

Before a model could make any worthwhile predictions, researchers on the experimental side needed to create a robust dataset. By mixing and matching 57 different peptides, Schissel and her colleagues were able to build a library of 600 miniproteins, each attached to PMO. With an assay, the team was able to quantify how well each miniprotein could move its cargo across the cell.

The decision to test the activity of each sequence, with PMO already attached, was important. Because any given drug will likely change the activity of a CPP sequence, it is difficult to repurpose existing data, and data generated in a single lab, on the same machines, by the same people, meet a gold standard for consistency in machine-learning datasets.

One goal of the project was to create a model that could work with any amino acid. While only 20 amino acids naturally occur in the human

body, hundreds more exist elsewhere—like an amino acid expansion pack for drug development. To represent them in a machine-learning model, researchers typically use one-hot encoding, a method that assigns each component to a series of binary variables. Three amino acids, for example, would be represented as 100, 010, and 001. To add new amino acids, the number of variables would need to increase, meaning researchers would be stuck having to rebuild their model with each addition.

Instead, the team opted to represent amino acids with topological fingerprinting, which is essentially creating a unique barcode for each sequence, with each line in the barcode denoting either the presence or absence of a particular molecular substructure. "Even if the model has not seen [a sequence] before, we can represent it as a barcode, which is consistent with the rules that model has seen," says Mohapatra, who led development efforts on the project. By using this system of representation, the researchers were able to expand their toolbox of possible sequences.

The team trained a convolutional neural network on the miniprotein library, with each of the 600 miniproteins labeled with its activity, indicating its ability to permeate the cell. Early on, the model proposed miniproteins laden with arginine, an amino acid that tears a hole in the cell membrane, which is not ideal to keep cells alive. To solve this issue, researchers used an optimizer to decentivize arginine, keeping the model from cheating.

In the end, the ability to interpret predictions proposed by the model was key. "It's typically not enough to have a black box, because the models could be fixating on something that is not correct, or because it could be exploiting a phenomenon imperfectly," Gomez-Bombarelli says.

In this case, researchers could overlay predictions generated by the

model with the barcode representing sequence structure. "Doing that highlights certain regions that the model thinks play the biggest role in high activity," Schissel says. "It's not perfect, but it gives you focused regions to play around with. That information would definitely help us in the future to design new sequences empirically."

Delivery boost

Ultimately, the machine-learning model proposed sequences that were more effective than any previously known variant. One in particular can boost PMO delivery by 50-fold. By injecting mice with these computer-suggested sequences, the researchers validated their predictions and demonstrated that the miniproteins are nontoxic.

It is too early to tell how this work will affect patients down the line, but better PMO delivery will be beneficial in several ways. If patients are exposed to lower levels of the drug, they may experience fewer side effects, for example, or require less-frequent doses (PMO is administered intravenously, often on a weekly basis). The treatment may also become less costly. As a testament to the concept, recent clinical trials demonstrated that a proprietary CPP from Sarepta Therapeutics could decrease exposure to PMO by 10-fold. Also, PMO is not the only drug that stands to be improved by miniproteins. In additional experiments, the model-generated miniproteins carried other functional proteins into the cell.

Noticing a disconnect between the work of machine-learning researchers and experimental chemists, Mohapatra has posted the model on GitHub, along with a tutorial for experimentalists who have their own list of sequences and activities. He notes that over a dozen people from across the world have adopted the model so far, repurposing it to make their own powerful predictions for a wide range of drugs.

More information: Carly K. Schissel et al, Deep learning to design nuclear-targeting abiotic miniproteins, *Nature Chemistry* (2021). DOI: 10.1038/s41557-021-00766-3

This story is republished courtesy of MIT News (web.mit.edu/newsoffice/), a popular site that covers news about MIT research, innovation and teaching.

Provided by Massachusetts Institute of Technology