individual researchers, the plug-in could help scholarly publishers to keep tabs on their own journals, Meyler says, because it allows users to filter by publication.

In its 'paper scorecard', the tool also flags any papers in the three generations of referenced studies in which more than 25% of papers in the bibliography are self-citations – references by authors to their previous works.

Meyler says that RedacTek is currently in talks with the scholarly-services firm Cabell's International in Beaumont, Texas, which maintains pay-to-view lists of suspected predatory journals. These publish articles without running proper quality checks for issues such as plagiarism but still collect fees from authors.

The plan is to use these lists to improve the tool so that it can also automatically flag any cited papers that are published in such journals.

'CHATGPT FOR CRISPR' CREATES NEW GENE-EDITING TOOLS

Some of the AI-designed gene editors could be more versatile than those found in nature.

By Ewen Callaway

In the never-ending quest to discover previously unknown CRISPR gene-editing systems, researchers have scoured microbes in everything from hot springs and peat bogs to poo and even yogurt. Now, thanks to advances in generative artificial intelligence (AI), they might be able to design these systems with the push of a button.

This week, researchers published details of how they used a generative AI tool called a protein language model – a neural network trained on millions of protein sequences – to design CRISPR gene-editing proteins. They also showed that some of these systems work as expected in the laboratory[1]. In February, another team announced that it had developed a model trained on microbial genomes, and used it to design fresh CRISPR systems, which are composed of a DNA or RNA-cutting enzyme and RNA molecules that direct the molecular scissors as to where to cut[2].

"It's really just scratching the surface. It's showing that it's possible to design these complex systems with machine-learning models," says Ali Madani, a machine-learning scientist and chief executive of the biotechnology firm Profluent, based in Berkeley, California. Madani's team reported what it says is "the first successful editing of the human genome by proteins designed entirely with machine learning" in a 22 April preprint[1] on bioRxiv (which hasn't been peer reviewed).
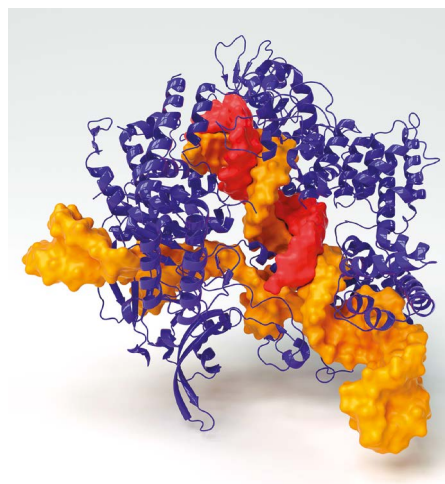
Alan Wong, a synthetic biologist at the University of Hong Kong, whose team has used machine learning to optimize CRISPR[3], says that naturally occurring gene-editing systems have limitations in terms of the sequences that they can target and the sort of changes that they can make. For some applications, therefore, it can be a challenge to find the right CRISPR. "Expanding the repertoire of editors, using AI, could help," he says.

Trained on genomes

Whereas chatbots such as ChatGPT train on text, the CRISPR-designing AIs were instead trained on vast troves of protein or genome sequences. The goal of this pre-training step was to imbue the models with insight into naturally occurring genetic sequences, such as which amino acids tend to go together. This information can then be applied to tasks such as the creation of totally new sequences.

Madani's team previously used a protein language model it developed, called ProGen, to come up with new antibacterial proteins[4]. To devise fresh CRISPRs, Madani's team retrained an updated version of ProGen with millions of

A bacterial CRISPR–Cas9 complex.

diverse CRISPR systems, which bacteria and other single-celled microbes called archaea use to fend off viruses. Because CRISPR systems comprise not only proteins, but also RNA molecules that specify their target, Madani's team developed another AI model to design these 'guide RNAs'.

The team then used the neural network to design millions of CRISPR protein sequences that belong to dozens of different families of such proteins found in nature. To see whether AI-designed CRISPRs were bona fide gene editors, Madani's team synthesized DNA sequences corresponding to more than 200 protein designs belonging to the CRISPR–Cas9 system that is now widely used in the lab. When the researchers inserted these sequences into human cells, many of the gene editors were able to precisely cut their intended targets in the genome.

The most promising Cas9 protein – a molecule they've named OpenCRISPR-1 – was just as efficient at cutting targeted DNA sequences as a widely used bacterial CRISPR–Cas9 enzyme, and it made many fewer cuts in the wrong place. The researchers also used the OpenCRISPR-1 design to create a base editor – a precision gene-editing tool that changes individual DNA 'letters' – and found that it, too, was as efficient as other base-editing systems, as well as less prone to errors.

Another team used an AI model capable of generating both protein and RNA sequences. This model, called EVO, was trained on 80,000 microbial genomes, and has not been yet tested in the lab. But predicted structures of some of the CRISPR–Cas9 systems it designed resemble those of natural proteins. The work was described in a preprint[2] posted on bioRxiv, and has not been peer reviewed.

Precision medicine

"This is amazing," says Noelia Ferruz Capapey, a computational biologist at the Molecular Biology Institute of Barcelona in Spain. She's impressed by the fact that researchers can use the OpenCRISPR-1 molecule without restriction, unlike with some patented gene-editing tools. The ProGen2 model and atlas of CRISPR sequences used to fine-tune it are also freely available.

The hope is that these tools could be better suited to medical applications than are existing CRISPRs. Profluent is hoping to partner with companies that are developing gene-editing therapies. "It really necessitates precision and a bespoke design. And I think that just can't be done by copying and pasting" from naturally occurring CRISPR, says Madani.

1. Ruffolo, J. A. et al. Preprint at bioRxiv https://doi.org/10.1101/2024.04.22.590591 (2024).
2. Nguyen, E. et al. Preprint at bioRxiv https://doi.org/10.1101/2024.02.27.582234 (2024).
3. Thean, D. G. L. et al. Nature Commun. 13, 2219 (2022).
4. Madani, A. et al. Nature Biotechnol. 41, 1099–1106 (2023).