Protein Discovery with Discrete Walk-Jump Sampling

ICLR Outstanding Paper Winner!

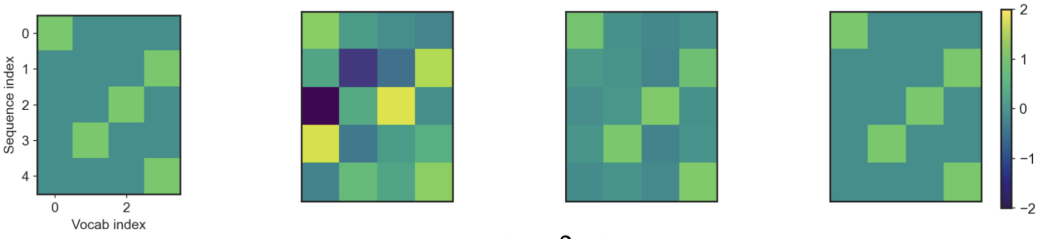
Nathan Frey is a Principal Machine Learning Scientist and Group Leader at Prescient Design, Genentech. He speaks to us fresh from winning an Outstanding Paper Award at ICLR 2024 for his pioneering work on generative modeling for drug discovery.

Nathan Frey and the team at **Prescient Design in Genentech** are pushing the boundaries of drug discovery by applying generative modeling techniques, commonly associated with **image generation**, to create surprising new proteins and discover antibodies and other molecules to cure disease.

"We have the major advantage that everything we design can be tested in the laboratory!"

However, as these are complex molecules critical for various biological functions, **how can the validity of these protein designs be verified? *"We have the major advantage that everything we design can be tested in the laboratory,"* Nathan explains. *"We work with our experimental colleagues and look at sequences together.*** *They tell us if we've done something catastrophically wrong, if there's something we haven't thought of or been aware of, or if the human body or immune system would never do this. They teach us those kinds of things, and then we can teach the model and test them."*

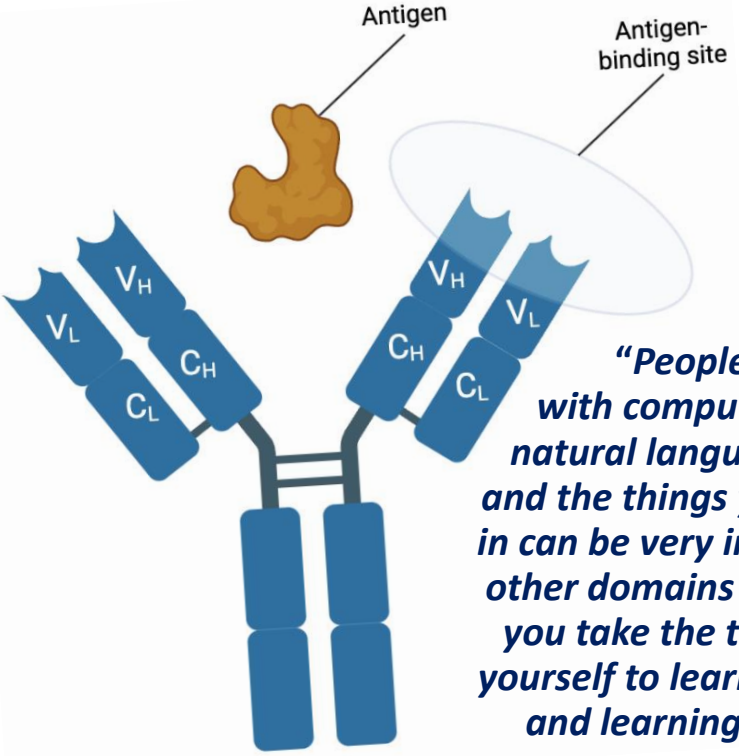
$$x = \text{EVQLV...} \qquad y = x + \varepsilon \qquad \hat{x}_\phi = y + \sigma^2 g_\phi(y) \qquad s = \operatorname{argmax} \hat{x}_\phi$$



$$\varepsilon \sim \mathcal{N}(0, \sigma^2 I_d)$$

There are many families of generative models and approaches for generative modeling. The motivation behind this project was to resolve the problems previously seen in the protein space when using energy-based and diffusion models.

Protein design is an instance of the **discrete sequence generation problem**, where amino acids are the building blocks of protein, and each amino acid sequence is a string of characters with only 20 possible choices at each position. Precision is crucial in determining the composition



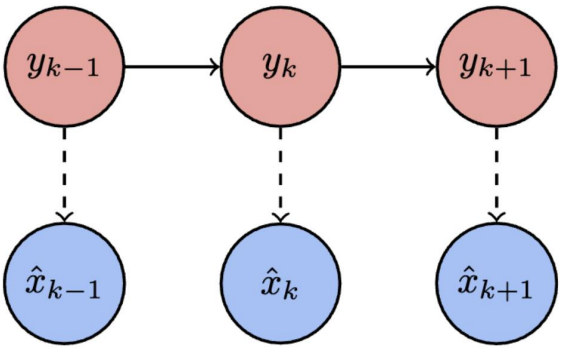
"People familiar with computer vision and natural language, your skills and the things you're interested in can be very impactful in these other domains in the sciences if you take the time to dedicate yourself to learning from people and learning those areas!"

of each position and modifying existing proteins. "*Basically, a lot of generative models are very bad at doing that,*" Nathan points out. "*Coming up with **a good, robust, sample-efficient method** for generating these discrete sequences was the problem we were interested in.*"

The paper represents these sequences as discrete characters that look like language but are one-hot encoded to make an image representation, where the pixels tell you what character is in each position. The team then applies a noising and denoising process. This approach is already used to **create new molecules, aid in sequence diversification, and hit expansion in the real world, which are critical phases in drug development**. "*What we're trying to do is take some starting protein sequences and use the generative model to explore around them to find other real proteins that are reasonable and interesting to look at,*" Nathan tells us. "*That's part one of a much bigger process called **lab-in-the-loop for protein design**.*"

Still coming down from his **Outstanding Paper Award win**, Nathan says he is excited about the **growing acceptance of machine learning in biology**. "*As far as I can tell, this is the first bio ML paper that's won an outstanding paper award at ICLR – maybe at any major machine learning conference,*" he reflects. It is a fantastic achievement with **only five Outstanding Paper winners out of 7,262 submitted and 2,260 accepted research papers**. What does he think convinced the judges that it deserved such an accolade? "*I would guess that, if anything, **it's the real-world impact!** What we did is not in the mainstream of generative modeling right now. There's an aspect of going against the grain, developing something new, and actually showing that we needed to develop something new for real-world impact.*"

This work has been a collaboration between Nathan and his co-authors, **Dan Berenberg** and **Saeed Saremi**, as well as **many other scientists and engineers across Prescient and Genentech**. "*One of many things*



that brought me to Prescient Design in Genentech was the hope of building things that are actually used in the laboratory to make real things," Nathan recalls. "*I think that's a sentiment shared by many of our scientists and engineers. For this process to work, you need all the machine learning and biology domain knowledge and great engineering. You really need all of those things together.*"

"When we solve real problems and have something interesting to say to the community, that's when we write something!"

At his lab, Nathan tells us he does not spend much time working on research papers. Instead, he is learning about and contributing to drug discovery. "*Writing papers is a byproduct of solving problems,*" he points out. "*When we solve real problems and have something interesting to say to the community, that's when we write something.*"

Looking ahead, he hints at ongoing collaborations and upcoming projects within and outside Prescient building on this work, including the **Generative and Experimental Perspectives for Biomolecular Design (GEM) workshop** that just took place at ICLR. "*Stay tuned for much, much more work!*" he reveals. "*People familiar with computer vision and natural language, your skills and the things you're interested in can be very impactful in these other domains in the sciences if you take the time to dedicate yourself to learning from people and learning those areas.*"

If, after reading this, you think you have what it takes to join Genentech, you will be pleased to hear that it is hiring. You, too, could be advancing the frontiers of science very soon! Genentech sees itself as **the first biotechnology company going back many decades** and is now trying to be **the first machine learning-based drug discovery company**! "*It's a reinvention of a company that has had massive success in drug discovery,*" Nathan adds. "*We want to continue leading on that front over the next many decades.*"