

Protein Scientist Jan Velecký: Machine Learning Can Shine Where Human Imagination Falls Short

Jan Velecký is a researcher at the RECETOX center. In his dissertation, he focuses on machine learning methods for protein engineering. He works at the center within Loschmidt Laboratories, specifically on Stanislav Mazurenko's team. In the interview, he explains what protein engineering is about, why it is important to study protein solubility, and how machine learning can be crucial for human knowledge.

11 Dec 2024 Sabina Vojtěchová Interview Research Personality



How did you get into protein engineering?

Since childhood, I had a curious relationship with computers, which led me to a technical high school in Zlín, where I studied information technology. Then, I continued at the Faculty of Information Technology at Brno University of Technology. After my bachelor's degree, I realized I would like to combine computer science with something more tangible, not just living inside the computer. One option was bioinformatics, which I started to pursue during my master's studies. Even then, I began to focus on protein research, specifically choosing the topic of protein solubility for my diploma thesis. And that's what I still do.

Could you explain a bit about what protein engineering actually focuses on?

Protein engineering is, simply put, the design or modification of proteins. Sometimes we create completely new proteins, or more commonly, we modify existing ones. We often proceed based on mutations, i.e., targeted changes that a scientist makes in the protein. To be able to change and modify proteins purposefully, we first need to study them and find out how they work. In its early days, the field aimed at creating mutations that would "damage" the function of a protein. If we manage to disable the protein through mutation, we actually reveal which

part of it is essential for its function. That's why protein engineering was initially dubbed protein terrorism.

How can I practically imagine such deliberate protein mutating?

A protein is a linear molecule, a chain of amino acids encoded by a gene. An organism produces it according to the gene and subsequently the protein folds, gaining a shape. The shape then determines the function of the protein. We don't need to know the shape, but we can play with the composition of amino acids. Typically, we replace an arbitrary amino acid with another. We have 20 types of amino acids, so I can, for example, put an alanine, which is very small and neutral, instead of a cysteine, making it ideal for our purposes. If the protein loses its function after such modification, we know that this cysteine is essential for its function or some key property.

Why do you focus on solubility particularly?

Solubility is one of the basic properties; it is crucial for producing new proteins. It determines whether it will be possible to produce the protein in sufficient quantity or quality and at an acceptable cost. It has direct implications in medicine and industry; drugs for stroke in development, for example, are protein-based, so are ingredients contained in washing detergents. Decent solubility is important for yield and thus more efficient production. I specifically deal with methods to estimate changes in protein solubility. This means that where we would usually need a person and material to measure solubility, we could use a computer and estimate solubility in silico, saving both money and time. For example, with the mentioned therapeutics for stroke, we could determine solubility in advance and avoid experiments that wouldn't make sense. To sum up, I try to measure and predict changes in protein properties, especially solubility, after introducing a mutation.

Do you have a personal connection to the field?

Bioinformatics interested me mainly because I saw its potential for the future; it simply seemed exciting to me. By choosing my then-teacher Jiří Hon, who was doing his doctorate at the faculty, as my diploma thesis supervisor at BUT, I also became part of the Loschmidt Laboratories, where I subsequently started working on my own doctorate under the supervision of Dr. Stanislav Mazurenko.

Can you explain a bit about what the Loschmidt Laboratories are?

The full name is the Loschmidt Laboratories of Protein Engineering, which is our main research focus. We are named after Johann Josef Loschmidt, a Bohemian-Austrian scientist of the 19th century, a founder of modern chemistry. We consist of four teams, two more theoretical, more focused on computational methods, and two experimental, working in the laboratory. What is specific about our work is that although we have specific protein targets we focus on, they serve more as case studies for the methods we create. We develop several software tools for protein engineers, biochemists, or clinical doctors to analyze or modify proteins for their applications. For example, one of our successful and widely used tools is EnzymeMiner, which can find more proteins with the desired function based on the well-studied protein. Thanks to

it, we can recycle and use what nature has already invented in the course of evolution. Our work is mainly about developing methods and tools.

Where does your work fall within the focus of the Loschmidt Laboratories?

Researching solubility is about protein optimization. If we develop a new drug, but the yield is so low that it is not enough even for laboratory experiments, we can increase it thanks to protein engineering. I mentioned the drug for stroke; another good example is insulin, which had low solubility earlier and originally used to be derived from pigs, but today, we can produce it artificially in cells and then extract it, making it financially accessible. Solubility prediction is also important in diagnostics. Many diseases are caused by mutation in an individual's genetic code that significantly reduces the solubility of a particular protein. And since insoluble proteins can even be toxic, there is a potential use in personalized medicine too.

If I understand correctly, the role of IT in your field is primarily to save time and money. But does machine learning have any other specific advantages?

Certainly. Let's say that all the problems a person solves can be well managed if they are imaginable in two to four dimensions. However, once there are, say, a hundred dimensions, finding interdependencies and explaining the phenomena and behaviours that lead to them is definitely not trivial. This is where machine learning can shine, as it has no limitations in the possible number of dimensions it can think in and can find what is hidden from us.

How can you be sure that the information you obtain this way is reliable?

Naturally, we have methods to verify reliability. The basis is that we train the machine learning model from data with known outcomes. However, we do not show the model all the data we have. The data we keep aside then serves to test the model. Machine learning is also about experimentation. It's not like you design a model and it works straight on.

And thanks to today's availability of computational hardware, we are putting into practice what was previously unimaginable. After all, the cost per unit of computation has dropped a quadrillion times since the 1950s when the foundations of machine learning were laid.

Where is the field heading and what are its ambitions?

There is a whole array of unresolved problems. The latest major challenge solved was the protein structure prediction problem. Now we know what the shape of a protein molecule will be based on its genetic code. This is an important advancement because, as I said, the function depends on a shape. Thus, I think we will now see many problems in molecular biology solved in quick succession. If I were to put it on a timeline, we used to mutate proteins to find out how they work. Today, we are redesigning them. In the future, we will be ordinarily creating completely new proteins, supported not only by the development of machine learning but also by robotics and other high-throughput methods that will enable us to collect data for machine learning.

Do you think the development of AI could threaten some jobs in science?

In my opinion, the development of AI will improve the lives of scientists because it will make repetitive and boring work easier. I don't think AI is targeting any specific positions. Proofreaders and translators might be needed less. In other positions, it will actually increase efficiency and thus the amount of work completed. Someone who now characterizes five protein variants a week in the lab might one day do five thousand.

What are your personal vision and scientific ambitions?

I would like my work to advance the field of protein engineering so that the field can continue to build on it. Right now, I want to focus mainly on completing my doctorate. In the future, I would like to stay in research. However, I want to experiment a bit at the same time. It could be an experience from research in the commercial sphere or a voyage into a new domain where I could apply my machine-learning skills. Not stopping broadening my horizons.