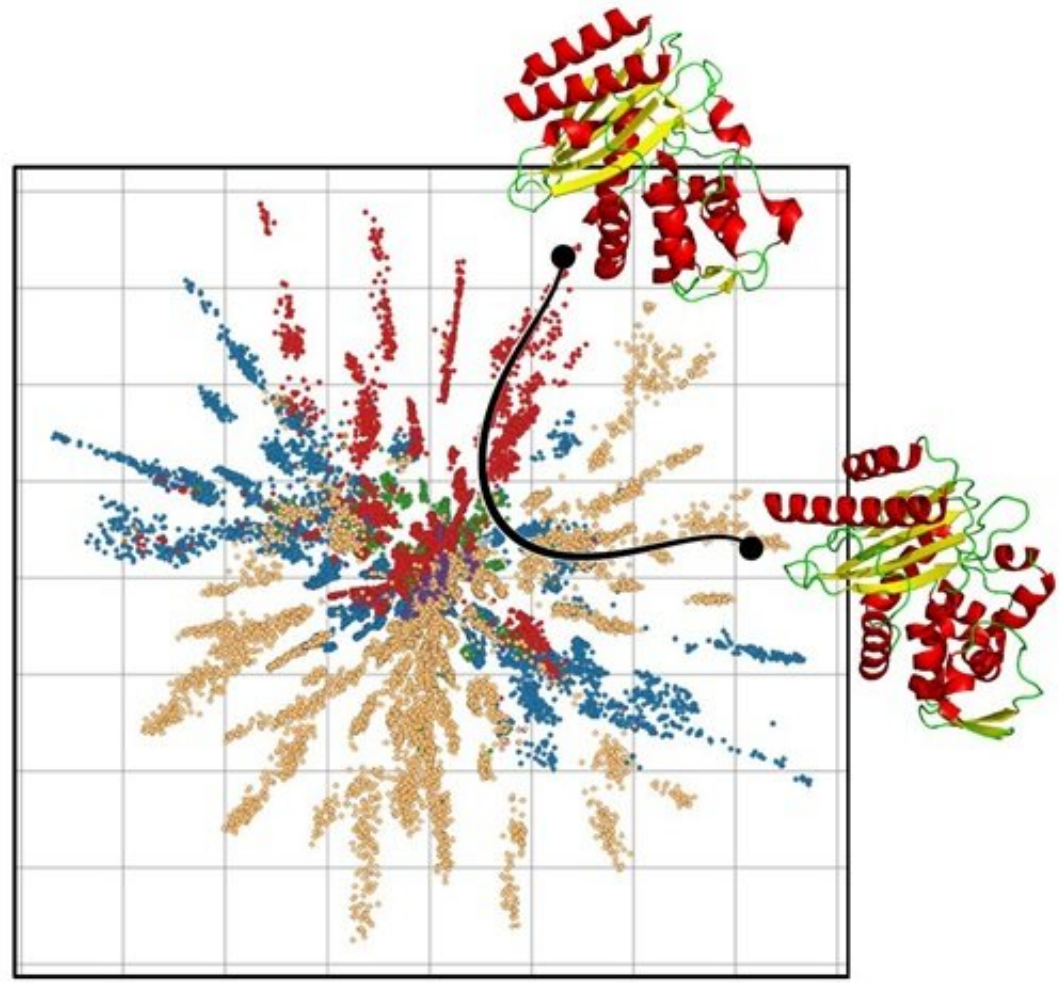
# New machine learning maps the potentials of proteins

May 4 2022, by Hanne Kokkegård



An example of the shortest path between two proteins, considering the geometry of the graphing. By defining distances in this way, it is possible to achieve biologically more precise and robust conclusions. Credit: W. Boomsma, N. S. Detlefsen, S. Hauberg.

The biotech industry is constantly searching for the perfect mutation, where properties from different proteins are synthetically combined to achieve a desired effect. It may be necessary to develop new medicaments or enzymes that prolong the shelf-life of yogurt, break down plastics in the wild, or make washing powder effective at low water temperature.

New research from DTU Compute and the Department of Computer Science at the University of Copenhagen (DIKU) can in the long term help the industry to accelerate the process. In the journal *Nature Communications*, the researchers explain how a new way of using Machine Learning (ML) draws a map of proteins, which makes it possible to appoint a candidate list of the proteins that you need to examine more closely.

In recent years, we have started to use Machine Learning to form a picture of permitted mutations in proteins. The problem is, however, that you get different images depending on what method you use, and even if you train the same model several times, it can provide different answers about how the biology is related.

"In our work, we are looking at how to make this process more robust, and we are showing that you can extract significantly more biological information than you have previously been able to. This is an important step forward in order to be able to explore the mutation landscape in the hunt for proteins with special properties," says Postdoc Nicki Skafte Detlefsen from the Cognitive Systems section at DTU Compute.

## The map of the proteins

A protein is a chain of amino acids, and a mutation occurs when just one of these amino acids in the chain is replaced with another. As there are 20 natural amino acids, this means that the number of mutations

increases so quickly that it is completely impossible to study them all. There are more possible mutations than there are atoms in the universe, even if you look at simple proteins. It is not possible to test everything in an experimental manner, so you must be selective about which proteins you want to try to produce synthetically.

The researchers from DIKU and DTU Compute have used their ML model to generate a picture of how the proteins are linked. By presenting the model for many examples of protein sequences, it learns to draw a card with a dot for each protein so that closely related proteins are placed close to each other while distantly related proteins are placed far from each other.

The ML model is based on mathematics and geometry developed to draw maps. Imagine that you must make a map of the globe. If you zoom in on Denmark, you can easily draw a map on a piece of paper that preserves the geography. But if you must draw the earth, mistakes will occur because you stretch the globe, so that the Arctic becomes a long country instead of a pole. So, on the map, the earth is distorted. For this reason, research in map-making has developed a lot of mathematics that describe the distortions and compensate for the distortions on the map.

This is exactly the theory that DIKU and DTU Compute have been able to expand to cover their Machine Learning model (deep learning) for proteins. Because they have mastered the distortion on the map, they can also compensate for it.

"It enables us to talk about what a sensible distance target is between proteins that are closely related, and then we can suddenly measure it. In this way, we can draw a path through the map of the proteins that tells us which way we expect a protein to develop from to another—i.e. mutated, since they are all related to evolution. In this way, the ML model can measure a distance between the proteins and draw optimal

paths between promising proteins," says Wouter Boomsma, Associate Professor in the section for Machine Learning at DIKU.

The researchers have tested the model on data from numerous proteins that are found in nature, where their structure is known, and they can see that the distance between proteins starts to correspond to the evolutionary development of the proteins, so that proteins that are close to each other evolutionally are placed close to each other.

"We are now able to put two proteins on the map and draw the curve between them. On the path between the two proteins are possible proteins, which have closely related properties. This is no guarantee, but it provides an opportunity to have a hypothesis about which proteins it could be that the biotech industry ought to test when new proteins are designed," says Søren Hauberg, professor in the Cognitive Systems section at DTU Compute.

The unique collaboration between DTU Compute and DIKU was established through a new center for Machine Learning in Life Sciences (MLLS), which started last year with the support of the Novo Nordisk Foundation. In the center, researchers in artificial intelligence from both universities are working together to solve the fundamental problems in Machine Learning driven by important issues within the field of biology.

The developed protein maps are part of a large-scale project that spans from basic research to industrial applications, e.g. in collaboration with Novozymes and Novo Nordisk.

**More information:** Nicki Skafte Detlefsen et al, Learning meaningful representations of protein sequences, *Nature Communications* (2022). DOI: 10.1038/s41467-022-29443-w

Provided by Technical University of Denmark