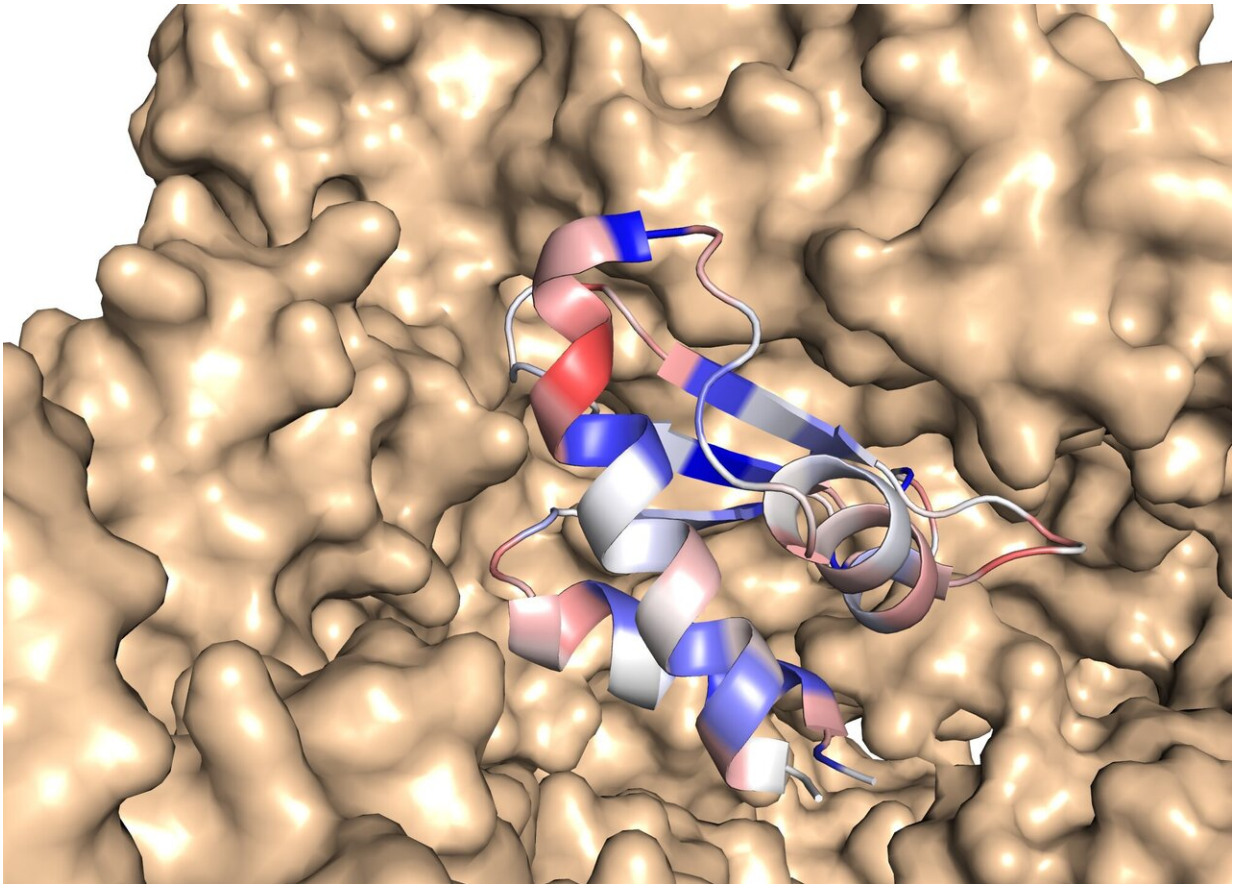# What artificial intelligence can teach us about proteins

May 15 2019

The DeeProtein Algorithm indicates protein regions which are functionally important (red) or less important (blue). Researchers use DeeProtein for targeted modification of proteins. The image shows AcrIIA4, a protein capable to stop the gene scissors CRISPR-Cas9.

Proteins are vital parts of all living organisms and perform essential tasks in our bodies. They build and repair tissues, supply components of the immune and hormone systems, regulate metabolism, and transmit signals. Researchers in Berlin and Heidelberg have now developed an intelligent neural network that can predict the functions of proteins in the human body. The team used a "trick" to observe how the network makes it predictions. The insights gained from this research could help in the search for new targeted drugs.

Intelligent virtual companions like Alexa, Siri, and Google Assistant have long been integrated into our everyday lives. And intelligent computational programs, so-called algorithms, have also evolved as an integral tool in scientific research. The huge amounts of data generated in life science research can be efficiently examined for recurring patterns with the aid of algorithms. Certain programs are able to spot recurring structures in large protein molecules and then use this information to draw conclusions about what cellular tasks these molecules perform—for example, whether they function as gene switches, molecular motors, or signaling molecules. The predictions made by such algorithms on the basis of protein sequences—which consist of a series of protein building blocks strung together like a pearl necklace—are now incredibly precise.

However, a major disadvantage of previous techniques is that users are kept completely in the dark as to why the algorithm assigns a particular function to certain protein sequences. The computer's precise knowledge about proteins is not directly available, despite the fact that such knowledge could prove invaluable in advancing the research and development of new agents.

A student team, jointly led by Roland Eils and Irina Lehmann from the Berlin Institute of Health (BIH) and Charité - Universitätsmedizin Berlin, in collaboration with Dominik Niopek from the Institute of

Pharmacy and Molecular Biotechnology (IPMB) at Heidelberg University, set itself the goal of unlocking this knowledge from the computer. It began working on this topic in 2017, and has developed an algorithm called "DeeProtein," a comprehensive and intelligent neural network that can predict the functions of proteins based on the sequence of individual protein building blocks, the amino acids. Like most learning algorithms, DeeProtein is a "black box," which means how they work remains a mystery to the programmers as well as the users. But the students have now used a "trick" to unravel the secret of this network.
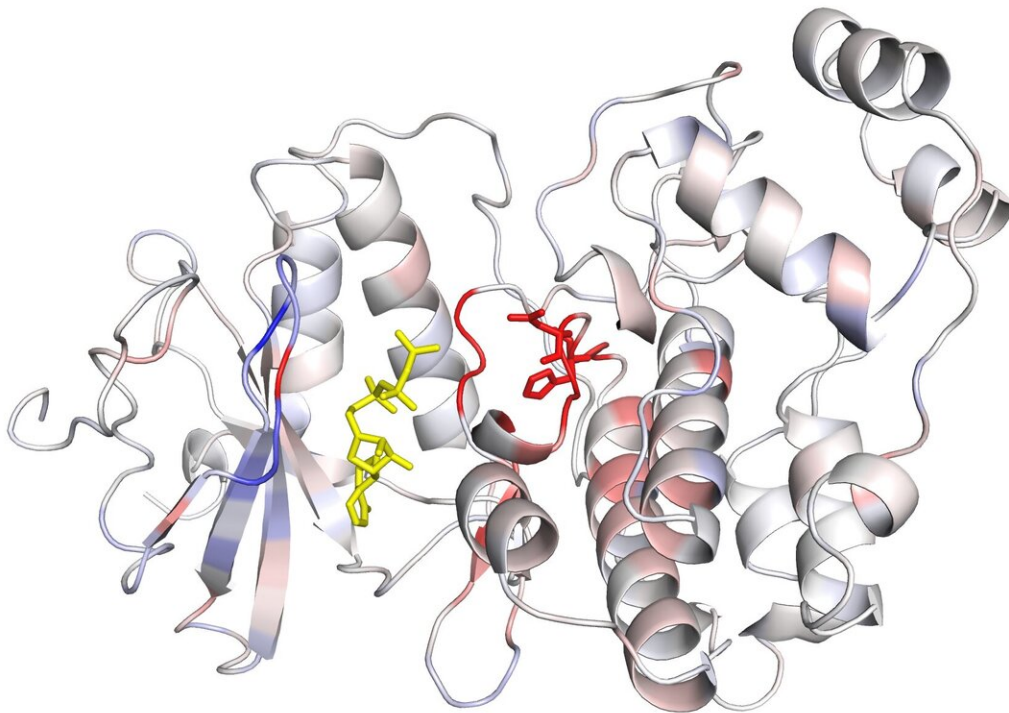
The DeeProtein Algorithm indicates protein regions which are functionally important (red) or less important (blue). Researchers use DeeProtein for targeted modification of proteins. The image shows ERK2, a protein often mutated in cancer patients.

The young scientists started by developing a way to figuratively look over the shoulder of the program as it does its work. "In the sensitivity analysis we successively mask each position in the protein sequence and let DeeProtein calculate, or rather predict, the function of the protein from this incomplete information," explains Julius Upmeier zu Belzen. He is a student in the master's program in molecular biotechnology at the IPMB and the lead author of the paper, which was just published in the journal *Nature Machine Intelligence*. "Next we give DeeProtein the complete sequence information and compare the two sets of predictions," adds Upmeier zu Belzen. "In this way we calculate, for each position in the protein sequence, how important this position is for predicting the correct function. This means that we give each position or amino acid in the protein chain a sensitivity value for the protein function."

The scientists then use the new analytical technique to identify the regions of the proteins that are vital to their function. This technique works for signaling proteins that play a role during carcinogenesis as well for the CRISPR-Cas9 gene-editing tool, which has already been tested in a large number of preclinical and clinical studies. "The sensitivity analysis enables us to identify protein regions that tolerate changes well or not so well," says Dominik Niopek. "This is an important first step if we want to make targeted changes to proteins, so as to equip them with new functions or to 'switch off' undesirable properties."

"With this work we show that not only can the predictions of neural networks be helpful, but that we can also now for the first time use this implicit knowledge for practical ends," explains Roland Eils. This approach is relevant for many issues in molecular biology and medicine. "If, for example, we want to develop targeted drugs or gene therapies, we need to know exactly where to focus our attention," adds Eils.

"DeeProtein can now help us do that."

More information: Julius Upmeier zu Belzen et al, Leveraging implicit knowledge in neural networks for functional dissection and engineering of proteins, *Nature Machine Intelligence* (2019). DOI: 10.1038/s42256-019-0049-9

Provided by Berlin Institute of Health