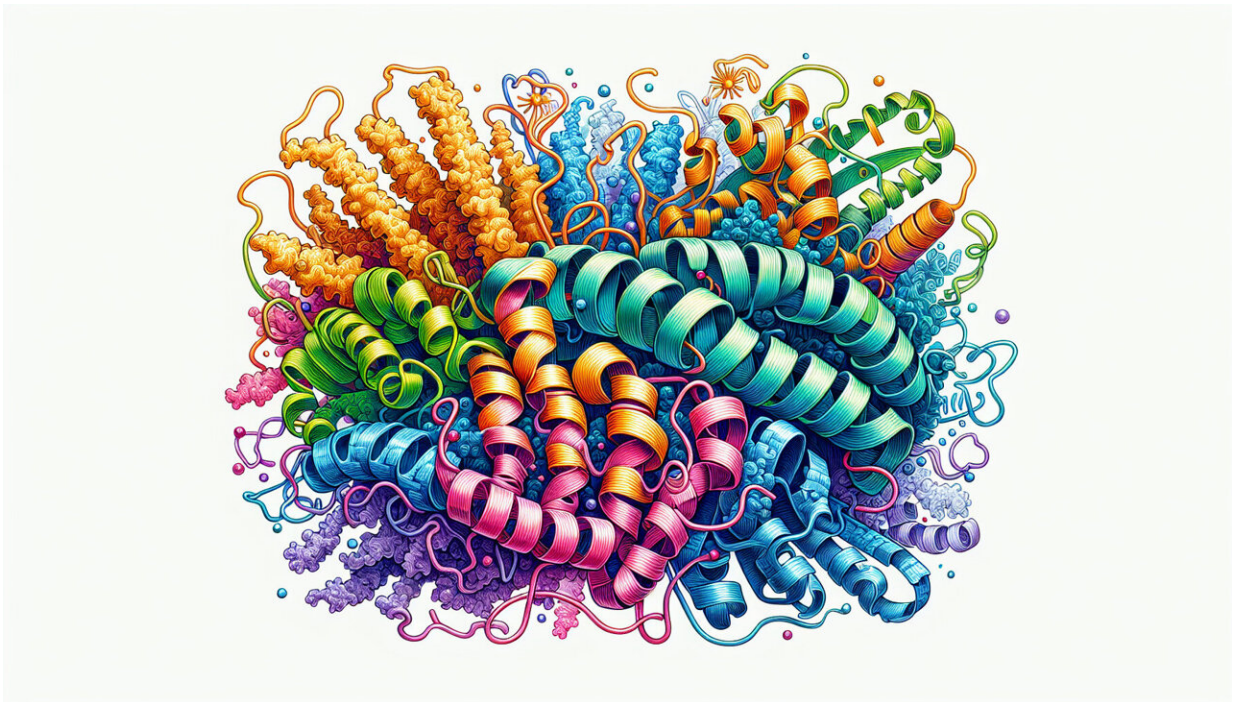# New AI tool captures how proteins behave in context

August 17 2024, by Mark Gaige

Credit: AI-generated image

A fish on land still waves its fins, but the results are markedly different when that fish is in water. Attributed to renowned computer scientist Alan Kay, the analogy is used to illustrate the power of context in illuminating questions under investigation.

In a first for the field of artificial intelligence (AI), a tool called PINNACLE embodies Kay's insight when it comes to understanding the behavior of proteins in their proper context as determined by the tissues and cells in which these proteins act and with which they interact. Notably, PINNACLE overcomes some of the limitations of current AI models, which tend to analyze how proteins function and malfunction but do so in isolation, one cell and tissue type at a time.

The development of the new AI model, described in *Nature Methods*, was led by researchers at Harvard Medical School.

"The natural world is interconnected, and PINNACLE helps identify these linkages, which we can use to gain more detailed knowledge about proteins and safer, more effective medications," said study senior author Marinka Zitnik, assistant professor of biomedical informatics in the Blavatnik Institute at HMS. "It overcomes the limitations of current, context-free models and suggests the future direction for enhancing analyses of protein interactions."

This advance, the researchers note, could propel current understanding of the role of proteins in health and disease and illuminate new drug targets for designing more precise, better tailored therapies.

PINNACLE is freely available to scientists everywhere.

A major step forward

Untangling the interactions across proteins and the effects of their contiguous biologic neighbors is tricky. Current analytic tools serve a crucial purpose by providing information on the structural properties and shapes of individual proteins. These tools, however, aren't designed to tackle the contextual nuances of the overall protein environment. Instead, they produce protein representations that are context-free,

meaning that they lack cell-type and tissue-type contextual information.

Yet proteins play different roles in the different cellular and tissue contexts in which they find themselves and also depending on whether the same tissue or cell is healthy or diseased. Single-protein representation models can't identify protein functions that vary across the multitude of contexts.

When it comes to protein behavior, it's location, location, location

Composed of twenty different amino acids, proteins form the building blocks of cells and tissues and are indispensable for a range of life-sustaining biologic functions—from transporting oxygen throughout the body to contracting muscles for breathing and walking to enabling digestion and fighting off infection, among many others.

Scientists estimate that the number of proteins in the human body ranges from 20,000 to hundreds of thousands.

Proteins interact with one another but also with other molecules, such as DNA and RNA.

The complex interplay between and across proteins creates convoluted networks of protein interaction. Situated in and among other cells, these networks engage in many complex cross talks with other proteins and protein networks.

PINNACLE's advantage stems from its ability to recognize that protein behavior can vary by cell and by tissue type. The same protein may have a different function in a healthy lung cell than it has in a healthy kidney cell or in a diseased colon cell.

PINNACLE sheds light on how these cells and tissues influence the same proteins differently, something not possible with current models. Depending on the specific cell type in which a protein network resides, PINNACLE can determine which proteins engage in certain conversations and which ones remain silent. This helps PINNACLE better decode the protein cross talk and the type of behavior and, ultimately, allows it to predict narrowly tailored drug targets for malfunctioning proteins that give rise to disease.

PINNACLE does not obviate but complements single-representation models, the researchers noted, in that it can analyze protein interactions within various cellular contexts.

Thus, PINNACLE could enable researchers to better understand and predict protein function and help elucidate vital cellular processes and disease mechanisms.

This ability can help pinpoint "druggable" proteins to serve as targets for individual medications as well as forecast the effects of various drugs in different cell types. For that reason, PINNACLE could become a valuable tool for scientists and drug developers to home in on potential targets much more efficiently.

Such optimization of the drug discovery process is sorely needed, said Zitnik, who is also an associate faculty member at the Kempner Institute for the Study of Natural and Artificial Intelligence at Harvard University.

It can take 10-15 years and cost as much as one billion dollars to bring a new drug to market, and the road from discovery to drug is notoriously bumpy with the end result often unpredictable. Indeed, nearly 90 percent of drug candidates do not become medicines.

Building and training PINNACLE

Using human cell data from a comprehensive multiorgan atlas, combined with multiple networks of protein–protein interactions, cell type-to-cell type interactions, and tissues, the researchers trained PINNACLE to produce panoramic graphic protein representations that encompass 156 cell types and 62 tissues and organs.

PINNACLE has generated nearly 395,000 multidimensional representations to date, compared to about 22,000 possible representations under current single-protein models. Each of its 156 cell types includes context-rich protein interaction networks of about 2,500 proteins.

The current numbers of cell types, tissues, and organs are not the upper limits of the model. The assessed cell types to date have come from living human donors and cover most, but not all, cell types of the human body. Moreover, many cell types haven't been identified yet, while others are rare or hard to probe, such as neurons in the brain.

To diversify the cellular repertoire of PINNACLE, Zitnik plans to make use of a data platform that includes tens of millions of cells sampled from the entire human body.

Provided by Harvard Medical School