UC SANTA BARBARA

THE Current

June 30, 2026
Cameron Walker

New tools to study protein assembly may help in the search for neurodegenerative disease therapeutics

Proteins, made of chains of amino acids, play critical roles in the body, whether building bone and muscle or supporting the immune system. While many proteins have a consistent folding pattern that keeps their 3D structure stable, others are more malleable.

In some cases, these more malleable proteins — called *intrinsically disordered proteins* — fold themselves to respond quickly to their environment, allowing them to switch rapidly between roles in the cell. Under certain conditions, however, these proteins can assemble into solid-like fibrils or liquid-like droplets. Although such assemblies can perform useful cellular functions, disruptions in their formation have been associated with conditions such as Alzheimer's disease, Parkinson's disease, and myotrophic lateral sclerosis.

These shape-shifting, clumping proteins can be challenging to untangle at a molecular level. Now, researchers at UC Santa Barbara have developed two new computational models that use artificial intelligence to predict these distinct forms of protein assembly, providing a novel approach to investigating the molecular behavior underlying neurodegenerative disease.

“One challenge in this field of aggregation-based neurodegenerative disorders is that we don’t have reliable model systems that you can test and characterize at the bench, or screen therapeutics against,” says M. Scott Shell, a professor of chemical engineering in The Robert Mehrabian College of Engineering. Designed by recent chemical engineering Ph.D. graduate Sam Lobo, in collaboration with professor of chemistry and biochemistry Joan-Emma Shea, these models are “a big advance,” Shell says, “in both the fundamental understanding of protein assembly and rapidly identifying proteins or regions that may warrant future therapeutic targets.”

Their research is published in the *Proceedings of the National Academy of Sciences.*

Predicting protein assembly

As Lobo learned about how word-based LLMs work, he became interested in applying similar approaches to proteins. A large language model, such as ChatGPT, is trained to recognize patterns in sequences of words. “Proteins are also a language,” Lobo says. “Instead of having words, it’s amino acids that come together to form a larger structure.”

Protein language models work in a similar way, but instead of learning from books or websites, they learn from enormous databases of protein sequences. By studying hundreds of millions of natural proteins, these models learn patterns of how amino acids appear together — patterns that can contain clues about protein functions and properties, including their tendency to aggregate.

Lobo used those learned patterns as the starting point for two new predictors. One, called *amyloid-predict,* estimates whether a protein sequence is likely to form amyloid fibrils, the ordered aggregates associated with some neurodegenerative diseases. The other, called *LLPS-predict,* estimates whether a protein sequence is likely to undergo liquid-liquid phase separation, a process in which proteins condense into liquid-like droplets inside cells.

Shell compares the information generated by the protein language model to a barcode that captures important features of each sequence in a compact numerical form that is easier to analyze. “These AI models can whittle all that information down into a concise format, based on looking at the spectrum of sequences and the patterns in them that you find in nature,” he says. “Then, Sam took these barcodes and figured out how to predict functional properties.”

Remarkably, the underlying language model was not originally trained to recognize either amyloid formation or liquid-liquid phase separation. Yet the patterns it learned from natural protein sequences contained enough information for Lobo's models to predict both behaviors.

The researchers used these models to scan the human proteome — the entire set of twenty thousand or so proteins that the human genetic code can produce. By doing this, they learned that intrinsically disordered proteins have different tendencies, depending on their sequence. Some have a high likelihood of forming both amyloids and droplets. Other proteins might form amyloids *or* droplets, but not both. "The interesting thing about the models," Shell says, "is that they show that these two types of assembly aren't intrinsically coupled."

The models also revealed interesting patterns in proteins that resemble *prions*, a type of misfolded protein known to cause diseases such as bovine spongiform encephalopathy (commonly known as mad cow disease), kuru, and Creutzfeldt-Jakob disease; these prion-like proteins are also linked to neurodegenerative disease. "We noticed that many prion-like proteins seemed to contain two distinct types of regions," Lobo says. "One region scored highly for amyloid aggregation, while the other scored highly for droplet formation."

That combination may help explain why prions are unusually difficult to control, he says. "So we think that maybe having both of these regions together in a single protein is an important aspect that makes these proteins so good at transmitting disease."

While their models currently allow researchers to understand the probability that a particular protein will aggregate, they do not yet explain the molecular mechanisms driving each prediction. The researchers' next goal is to use the patterns recognized by the AI system to understand why the proteins behave as they do.

"We want to understand how the signals that are coming out of the protein language model relate to the physics of aggregation at a molecular level," Shell says, "as well as how the surrounding conditions, such as pH or temperature, affect the propensity of aggregation."

Establishing those connections could eventually help researchers determine why certain proteins become associated with disease and where new interventions might be most effective.

“The human proteome contains thousands of intrinsically disordered proteins whose aggregation behavior remains uncharacterized,” Lobo says. “These models give researchers a nice way to prioritize which proteins, and which regions within them, are most likely to drive disease and most worth targeting next.”

Tags
Artificial Intelligence
Health and Medicine

Media Contact
Shelly Leachman
Editorial Director
(805) 893-2191
sleachman@ucsb.edu

About UC Santa Barbara

The University of California, Santa Barbara is a leading research institution that also provides a comprehensive liberal arts learning experience. Our academic community of faculty, students, and staff is characterized by a culture of interdisciplinary collaboration that is responsive to the needs of our multicultural and global society. All of this takes place within a living and learning environment like no other, as we draw inspiration from the beauty and resources of our extraordinary location at the edge of the Pacific Ocean.