The virtual cell

Virtual cells based on artificial intelligence models are on the horizon

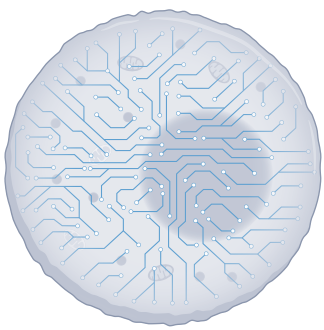
By Lin Tang

The paramount appeal of a virtual cell model cannot be overstated (*Cell* **187**, 7045–7063; 2024). From molecular cell biology to translational medicine, almost every branch of life sciences stands to benefit from it. Despite largely being a conceptual feat at present, some of our wish list items for a virtual cell model include that: (1) it will provide a holistic picture of all the molecular and cellular phenotypes of a cell; (2) it will be mechanistic and dynamic, and reveal the biological underpinnings of various cellular behaviors; and (3) it will be predictive, with the ability to generate predictions under a wide range of conditions.

With the age of big data and artificial intelligence (AI), massive volumes of biological data (especially those generated by high-throughput omics technologies) coupled with advanced machine-learning approaches represent two of the major drivers that fuel the aspiration of virtual cell models. A large number of AI models already exist for various specific biological tasks, including recently emerging foundation models that aim to be versatile performers (*Nature* **640**, 623–633; 2025). However, there is still a long way to go to meeting the challenges on our wish list.

One persistent bottleneck is data scarcity. As recently touched upon by a *Nature Methods* Editorial (*Nat. Methods* **22**, 1387; 2025), the availability of biological data – although orders of magnitude higher than decades ago – is still dwarfed as compared with areas such as language and image processing, in which powerful large language models now dominate. More critically, high-quality data that are particularly valuable for revealing biological mechanism and causality, such as multimodal, time-series or perturbation data, are still badly needed. Also, there is still no consensus on the best modeling strategy to build a virtual cell model; how to leverage existing models and biological knowledge is an open question.

There are even yet more directions to explore, such as generating virtual tissues using spatial omics technology data, which take into account intercellular interactions and communication. Although it is still early days to affirmatively answer the question of when a virtual cell model will be within reach, we are excited to witness the flourishing of methods development towards this ambitious goal.



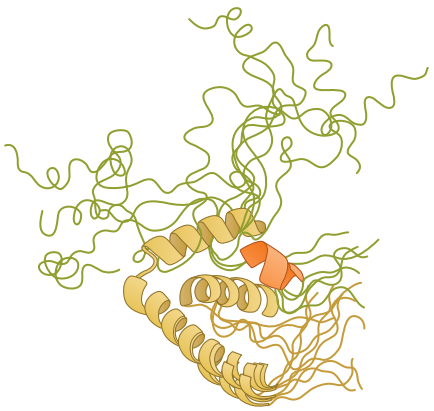
Using virtual cell models to simulate biological experiments.

Intrinsic protein disorder at scale

Methods to study the structural and functional properties of proteins that contain intrinsically disordered regions at the proteome scale are on the rise.

By Arunima Singh

Recent advances in computational protein structure prediction methods, along with high-throughput and comprehensive proteomic analysis approaches, have deepened our understanding of the relationships between protein structure and function. However, one class of proteins continues to elude more comprehensive investigation – intrinsically disordered proteins (IDPs) and proteins that contain intrinsically disordered regions (IDRs). Unlike structured proteins, IDPs lack a fixed three-dimensional fold, which makes them difficult to characterize using traditional biochemical and structural methods. Nevertheless, nearly 40% of human proteins are estimated to contain IDRs and are involved in key regulatory pathways and have been linked to various diseases, which necessitates the development of methods to characterize them.



Intrinsic disorder in proteins makes structural characterization a challenge.

Recent developments have advanced both experimental and computational methods in this field. DisP-seq makes use of disordered protein precipitation followed by DNA sequencing to map the genome-wide binding of DNA-associated proteins that contain IDRs (*Nat. Biotechnol.* **42**, 52–64; 2024). This is a departure from previously available technologies that use antibodies to pull down proteins and were restricted to well-characterized proteins.

Another method for global analysis of endogenous protein disorder uses a bifunctional chemical probe called TME to capture unfolded proteins, marked by exposed cysteine residues, in situ (*Nat. Methods* **22**, 124–134; 2025). The captured proteins are detected by a fluorescence readout or enriched and analyzed by mass spectrometry-based proteomics. The approach enables the capture of both basal disordered proteins as well as proteins whose folding status changes under stress at a cellular level.

ALBATROSS is a deep-learning-based model that combines rational sequence design and large-scale molecular simulations to predict the ensemble properties of IDPs directly from sequence (*Nat. Methods* **21**, 465–476; 2024). Another computational method to generate conformational ensemble for IDRs was used to simulate nearly all (more than 28,000) IDRs from the human proteome (*Nature* **626**, 897–904; 2024). Additionally, a strategy specifically developed to target flexible IDPs and IDRs has proven effective in generating experimentally validated binders against a variety of disordered proteins (*Nature* **644**,