frontiers | Frontiers in Bioengineering and Biotechnology

TYPE Editorial
PUBLISHED 05 January 2023
DOI 10.3389/fbioe.2022.1108962

OPEN ACCESS

EDITED AND REVIEWED BY
Manfred Zinn,
HES-SO Valais-Wallis, Switzerland

*CORRESPONDENCE
Ratul Chowdhury,
ratul@iastate.edu

SPECIALTY SECTION
This article was submitted to Bioprocess Engineering, a section of the journal Frontiers in Bioengineering and Biotechnology

RECEIVED 26 November 2022
ACCEPTED 19 December 2022
PUBLISHED 05 January 2023

CITATION
Chowdhury R (2023), Editorial: Advances in protein structure, function, and design. *Front. Bioeng. Biotechnol.* 10:1108962. doi: 10.3389/fbioe.2022.1108962

Editorial: Advances in protein structure, function, and design

Ratul Chowdhury[1,2]*

[1]Department of Chemical Engineering, Iowa State University, Ames, IA, United States, [2]Nanovaccine Institute, Ames, IA, United States

KEYWORDS

protein structure/folding, AlphaFold2, protein design and engineering, biological network, deep learning, AI-guided biology

Editorial on the Research Topic
Advances in protein structure, function, and design

Accessibility to ever improving computing infrastructure has led to a paradigm shift towards data-driven modeling in all areas of science and arts. Eponymously, data-driven modeling relies on 1) well curated, domain-knowledge-driven datasets, and 2) appropriate utilization of said data (*i.e.*, avoid overfitting, under sampling, *etc.*). The domain of protein biology has historically been on the lookout for a reliable method to discern the 3D-shape (structure) of a protein given its amino acid sequence. Precise knowledge of a protein's structure enables us to first, explain how it works as a tiny molecular machine, and then devise rules to modify existing proteins or design new ones for therapeutic and engineering applications spanning–healthcare, green chemistry, energy, and novel functional materials.

One way to accurately determine the 3D-shape of a protein is *via* experiments (spectroscopy–Nuclear Magnetic Resonance, or crystallography–X-Ray diffraction) to catalog the 3D-Cartesian coordinates of each atom that are present in the protein. Such per-atom information is stored in a PDB (Protein Data Bank) format. Set up in 1976, the PDB (Berman et al., 2000) is a publicly accessible dataset of ~198 k protein structures and has aggressively expanded at the rate of ~11 k new entries per year since 2013. While this is quite a substantial dataset, this barely scratches the surface and constitutes only a meagre ~.09% of the total set of 230M known protein sequences reported till date (UniParc dataset (Bairoch et al., 2005)). This has prompted the emergence of a gamut of data driven deep-learning techniques to reason over known sequence-structure pairs (from PDB) and create neural operators which can then predict the structure from any new protein sequence.

Two emerging, yet different schools of thought that fuel these deep-learning pipelines for structure prediction are: 1) family sequence alignment-based (FSA), and 2) single sequence-based (SS). FSA methods such as AlphaFold2 (Jumper et al., 2021) and RosettaFold (Baek et al., 2021) group all sequences a protein family (say, all amylases across species) and corresponding structures into constellations of similarity. During prediction, each input sequence is first sent through a sequence alignment pipeline to find which constellation it belongs to and use structures from the same constellation as templates to thread a possible predicted structure. Such methods, while powerful, cannot account for significant structural changes from point mutations unless such a mutant is a part of the training set (in which case it simply memorizes it). Interestingly, designed proteins with tailored function and disease-causing protein sequence variants fold into very different structures. Structural changes in these proteins are elusive to FSA structure predictors like AlphaFold2 and RosettaFold. On the other hand, SS methods (like RGN2 (Chowdhury et al., 2022)) use natural language processing to encode sequences to high-dimensional vectors and map such encodings to atomic coordinates of one (C α) or more atoms

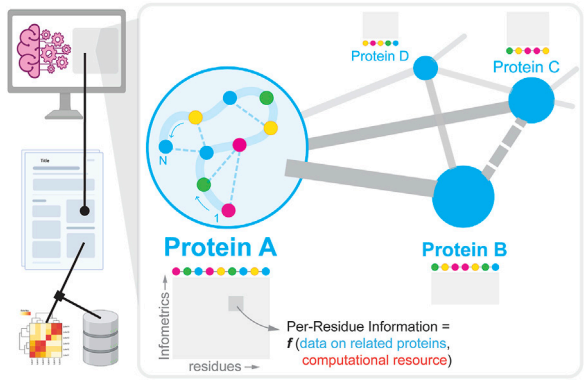

FIGURE 1
Overview of how experimental data describing a protein molecule and its interactions (inset) is reported in literature, and extracted using natural language processing to create datasets to support subsequent machine learning/AI-guided protein engineering efforts.

(N, C, and C α) of each amino acid. Since SS methods do not rely on similarity constellations, they are equipped to predict structural changes emerging from as little as a single amino acid change such as rapidly evolving viral proteins, and non-homologous *de novo* proteins. While neither FSA nor SS models are explicitly trained to be performative on structural changes arising from point mutations, at the limit of performance, SS models are better poised to capture such effects. Moreover, the biological event of protein folding upon synthesis is a molecular process which depends on the sequence of amino acids that make up the protein, not how similar proteins are folded in other species.

In this Research Topic, one of the key contributions is by Villalobos-Alva et al. which provides an extensive summary of ML/AI-based recipes and neural architectures that have been developed and deployed till date to learn, predict, and design proteins for various use cases (see Figure 1). Along the lines of structure prediction, Jin et al. show how Generative Adversarial Networks (GANs) can be utilized to predict secondary structure of proteins as a function of the inputs—1) amino acid sequence, and 2) similarity with known proteins with experimentally verified structure (i.e., structural prior). Next, Wang et al. first construe a protein structure as a network of amino acids either connected by covalent or non-covalent bonds and show how the network topology appended to node properties (of individual amino acids) can be mapped to protein function. The study reveals that DNA binding proteins tend to have buried hydrophobic pockets which are thermodynamically amenable to drug-binding. On the engineering front, Zhou et al. use information about active site of a particular enzyme (glutamate dehydrogenase) and tailors it to catalyze a stereoselective reaction with high efficiency. Pan et al. zoom out of a single protein and focuses on the entire protein-protein interaction (PPI) network in a single plant cell. In such a network, each node is a protein (and not the individual amino acids that make up the protein). A PPI dictates cellular phenotype *i.e.,* how an organism behaves as a consequence of proteins interacting with each other. They describe DWPPI–a deep walk algorithm that traverses a plant PPI to discern how information cascades through such a network.

Author contributions

The author confirms being the sole contributor of this work and has approved it for publication.

Conflict of interest

The author declares that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's note

References

Baek, M., DiMaio, F., Anishchenko, I., Dauparas, J., Ovchinnikov, S., Lee, G. R., et al. (2021). Accurate prediction of protein structures and interactions using a three-track neural network. *Science*, 373 (6557), 871–876. doi:10.1126/science.abj8754

Bairoch, A., Apweiler, R., Wu, C. H., Barker, W. C., Boeckmann, B., Ferro, S., et al. (2005). The universal protein resource (UniProt). *Nucleic Acids Res.*, 33. D154–9. doi:10.1093/nar/gki070

Berman, H. M., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T. N., Weissig, H., et al. (2000). The protein Data Bank. *Protein Data Bank Nucleic Acids Res.* 28, 235–242. doi:10.1093/nar/28.1.235

Chowdhury, R., Bouatta, N., Biswas, S., Floristean, C., Kharkar, A., Roy, K., et al. (2022). Single-sequence protein structure prediction using a language model and deep learning.Erratum in. *Nat. BiotechnolNat Biotechnol.* 40 (11), 1617–1623. doi:10.1038/s41587-022-01432-w

Jumper, J., Evans, R., Pritzel, A., Green, T., Figurnov, M., Ronneberger, O., et al. (2021). Highly accurate protein structure prediction with AlphaFold. *Nature* 596 (7873), 583–589. doi:10.1038/s41586-021-03819-2