ADVANCED BIOLOGY
www.advanced-bio.com

In the Age of Machine Learning Cryo-EM Research is Still Necessary: A Path toward Precision Medicine

Dominique C. Stephens, Amber Crabtree, Heather K. Beasley, Edgar Garza-Lopez, Margaret Mungai, Larry Vang, Kit Neikirk, Zer Vue, Neng Vue, Andrea G. Marshall, Kyrin Turner, Jian-qiang Shao, Bishnu Sarker, Sandra Murray, Jennifer A. Gaddy, Antentor O. Hinton Jr, Steven Damo,* and Jamaine Davis**

Machine learning has proven useful in analyzing complex biological data and has greatly influenced the course of research in structural biology and precision medicine. Deep neural network models oftentimes fail to predict the structure of complex proteins and are heavily dependent on experimentally determined structures for their training and validation. Single-particle cryogenic electron microscopy (cryoEM) is also advancing the understanding of biology and will be needed to complement these models by continuously supplying high-quality experimentally validated structures for improvements in prediction quality. In this perspective, the significance of structure prediction methods is highlighted, but the authors also ask, what if these programs cannot accurately predict a protein structure important for preventing disease? The role of cryoEM is discussed to help fill the gaps left by artificial intelligence predictive models in resolving targetable proteins and protein complexes that will pave the way for personalized therapeutics.

1. Machine Learning and Protein Structure Prediction

Prediction of accurate 3D protein structures has been a big challenge for structural biologists for decades. Indeed, over this timespan, it has been widely accepted that the instructions for predicting protein folds are encoded within its amino acid sequence. In recent times, however, there emerged two sides of this phenomenon, researchers that are interested in the fold, and researchers interested in the folding. Advancements in computing power and capabilities have ushered in machine learning and artificial intelligence (AI) approaches throughout research and have significantly advanced protein structure prediction.

D. C. Stephens, A. Crabtree, H. K. Beasley, L. Vang, K. Neikirk, Z. Vue, N. Vue, A. G. Marshall, A. O. Hinton Jr
Department of Molecular Physiology and Biophysics
Vanderbilt University
Nashville, TN 37232, USA
E-mail: antentor.o.hinton.jr@vanderbilt.edu

D. C. Stephens, K. Turner, S. Damo
Department of Life and Physical Sciences
Fisk University
Nashville, TN 37232, USA
E-mail: sdamo@fisk.edu

E. Garza-Lopez, M. Mungai
Department of Internal Medicine
University of Iowa
Iowa City, IA 52242, USA

The ORCID identification number(s) for the author(s) of this article can be found under https://doi.org/10.1002/adbi.202300122

DOI: 10.1002/adbi.202300122

J.-qiang Shao
Central Microscopy Research Facility
University of Iowa
Iowa City, IA 52242, USA

B. Sarker
School of Applied Computational Sciences
Meharry Medical College
Nashville, TN 37208, USA

S. Murray
Department of Cell Biology
College of Medicine
University of Pittsburgh
Pittsburgh, PA 15260, USA

J. A. Gaddy
Division of Infectious Diseases
Vanderbilt University School of Medicine
Nashville, TN 37232, USA

J. A. Gaddy
U.S. Department of Veterans Affairs
Tennessee Valley Healthcare Systems
Nashville, TN 37212, USA

J. Davis
Department of Biochemistry and Cancer Biology
Meharry Medical College
Nashville, TN 37208, USA
E-mail: jdavis@mmc.edu

These advancements are transforming the rate at which discoveries can be made that will ultimately lead to better treatments and health outcomes. Breakthroughs from the first wave of AI-based protein structure prediction methods, such as Deepmind's AlphaFold and RoseTTAFold are significantly accelerating the discovery of new mechanistic questions, and new treatments, by providing detailed knowledge of protein structures. Protein structure prediction methods were reported as The Method of the Year 2021, and each method is a constantly improving neural network trained to produce protein structures from amino acid sequences, multiple sequence alignments, and homologous proteins with unprecedented speed and accuracy.[1–3] RoseTTAFold, and the second version of AlphaFold, AlphaFold2, shifted the landscape to provide researchers, particularly those outside of structural biology, with access to 3D protein structure models.[4,5] In a short time, AlphaFold2 has created a database of over 200 million predicted protein structures allowing a better understanding of proteins that is sure to accelerate the development of new drugs to treat diseases. This tool can create accurate models and predictions for many folded proteins, as well as identify some of the dynamic behaviors within domains.[6]

With such accuracy in protein structure prediction, an ongoing debate in the field is whether experimental structural biology methods are still necessary. A variety of protein structure prediction methods exist and have long been useful tools in computational structural biology to predict a folded protein.[7–9] The ability to make these predictions have been based on the understanding of the folding process of which thermal motions cause conformational changes moving the protein toward an energetically favored, native structure, commonly known as the funnel shape energy landscape.[10] This theory helps to explain how proteins follow a path to adopt a specific, stable conformation, and why they sometimes undergo structural changes in response to certain stimuli, such as changes in temperature, pH, or the presence of ligands. This implies that proteins adopt multiple structural conformations in a variety of environments. AlphaFold alone cannot predict the conformational folding landscape of proteins.

Machine learning prediction programs rely on experimental structural biology details, whether determined by X-ray, nuclear magnectic resonance, or cryogenic electron microscopy (cryoEM) to build and validate predictions. Much of the success of AlphaFold2, RoseTTAFold, and other machine-learning approaches is attributed to the archive of 3D protein structures housed in the Protein Data Bank. The experimental data within the PDB contributes to the accuracy and reliability, however, protein prediction programs are in essence hypotheses. Experts in structural biology agree that machine learning methods are vital aspects of science and will likely help to significantly advance biology. However, the results obtained need to be taken with caution because of the limitations AlphaFold2 possesses. Terwillinger et al. (2021) compared AlphaFold predictions to experimentally determined crystal structures and found that some predictions disagree with experimental data.[11] In another example, a recent study used a dataset of cryoEM density maps of viruses to assess predictive modeling methods for resolving atomic models.[12] The biological data sets ranged from 1.8 to 4.5 Å resolution and the AlphaFold2 models were compared with the experimentally resolved models. As expected, Alpha Fold2 was able to accurately predict known areas, but struggled with novel structures.[12] In the case of the Mud Crab Reovirus model in this study, despite numerous related structures, none of the four capsid proteins predicted by AlphaFolds2 agreed with the density map or experimentally resolved models.[11,12] The authors emphasize the power of AlphaFold2 in predicting individual proteins, but difficulty in modeling protein assemblies.

At the same time as machine learning approaches are advancing and revolutionizing structural biology, the resolution revolution in cryoEM is happening. CryoEM provides high-resolution images of macromolecular structures that can reveal important details that may not be captured through sequence information alone, and this advancement also garnered it to win Method of the Year in 2015.[13] A key strength of CryoEM is the ability to resolve large proteins and assemblies in their native state, which is essential for understanding their function and behavior. X-ray crystallography is still the most prominent method for determining protein structures and for a long time, NMR has been the second. However, more researchers are resolving structures by cryoEM which has significantly increased the total number of EM structures deposited in the PDB over the years, so much that it has now surpassed NMR for second place. For example, as of April 2023, the total number of proteins and complexes resolved by cryoEM (15150), has now surpassed the total number of structures determined by NMR (13986). The revolution in CryoEM has ushered in more detailed and accurate structural information of macromolecular complexes and cellular systems. This resolution revolution has been achieved through improved sample preparation, better detectors and cameras, advanced image processing algorithms, and increased computational power.[14] This has enabled researchers to study the structures of previously intractable biospecimens, leading to a deeper understanding of cellular processes, disease mechanisms, and potential drug targets. Although CryoEM has countless advantages, this popular technique also has its own deficiencies. For example, only relatively large protein complexes can be used for cryo-EM (*i.e.*, > ≈100 kDa) because smaller protein particles are difficult to observe under the electron microscope. Also, sample homogeneity is still very important and flexible proteins make it difficult to get good results.[15] Despite these limitations, the CryoEM revolution will continue to advance our knowledge of proteins based on technological advancements and integration with machine learning.

2. Machine Learning and CryoEM: The Synergy Needed to Tackle Precision Medicine

Precision health considers a patient's unique genes, environment, and lifestyle to create a specialized treatment plan to improve health outcomes. Precision (personalized) medicine has been around for many years, but only recently has there been a shift toward enabling machine learning technology to predict and interpret patient data. Without accurate experimental data, machine learning technology would struggle to develop accurate models and structures, limiting its ability to interpret patient data for a personalized treatment plan. This is the goal of precision medicine, using the data gathered from a patient's genetics and environment to select a treatment plan that best suits their specific ailment, the right drug for the right person at the right dose at the right time. One of the biggest limitations of delivering precision medicine for everyone is that genomic medicine

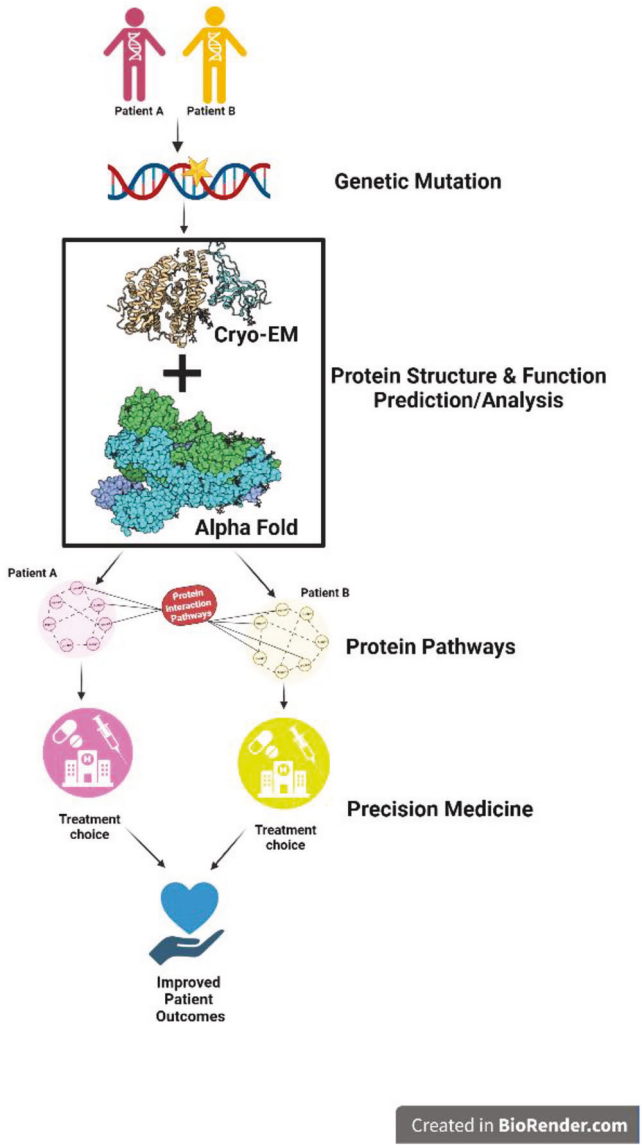

Figure 1. Structural biology guided precision medicine model. Genetic variants in humans can present the same phenotype, but display differences in disease manifestation. Dysregulated proteins from these genetic dispositions can be structurally determined computationally and experimentally. This information will capture the details about key changes in protein function, which will help target altered pathways. Precise treatments can be tailored based on the individual patient's mutation status rather than treating the disease.

does not inform scientists and clinicians of the key interactions and pathways needed for precise treatment. For the successful treatment of an individual, we must understand how genetic dispositions influence the encoded protein dynamics and behavior, of which this data can be used to unveil relevant protein interactions and pathways leading to unique characteristics of that patient (**Figure 1**). This is critically important since more than 99% of drugs target proteins. CryoEM can address the limitations of predictive tools and really shape efforts toward precision medicine. In an intriguing discovery, Yang et al., (2020) reported the cryoEM structures of isolated amyloid ß peptides (Aß 42) from human brains differed from the filaments assembled in vitro. These Aß42 filaments had two structural related S-shape folds that are categorized into two types . [16] In individuals with sporadic Alzheimer's disease, type I filaments were common, whereas in individuals with familial Alzheimer's disease and other conditions, type II was found. The impact of this work will lead to better-informed in vitro and animal models, as well as the potential for personalized treatments for sporadic versus familial disease. Precision medicine is broadly defined but approaches that isolate and characterize proteins directly from patients will have a major impact in accelerating novel treatments and ultimately improving health outcomes.

3. Precision Medicine Challenge

The majority of protein structures predicted can be reliably used for additional studies in biology. However, there are proteins that machine learning programs cannot predict very well and some that cryoEM has not been able to resolve. Key examples are proteins with multiple flexible domains and intrinsic disordered regions. One such example is the Breast Cancer Susceptibility protein 1 (BRCA1). There are extensive studies on how genetic variants of BRCA1 are associated with an increased risk of breast, ovarian, and pancreatic cancer.[17] Therapies that target these malignancies based on BRCA1 mutation status exist, however novel therapies of BRCA1-associated cancers are still urgently needed. The BRCA1 RING and BRCT domains have been structurally characterized extensively,[18,19] yet after 30 years since its discovery, no structural information for the majority of the protein, which consists of a large, disordered region. Using AlphaFold2 to predict the folded structure results in an unreliable prediction, primarily due to the intrinsically disordered region (**Figure 2**). The accuracy of this model is based on a per-residue confidence score, called the predicted local distance difference test, which uses a scale from 0–100. Values greater than 90 are designated blue and indicate high confidence, between 90 and 70 are confident, between 70 and 50 are low, while values below 50 are designated orange and represents very low confidence or unstructured. Not being able to resolve the full-length BRCA1 structure represents a major challenge in providing complete molecular details of this protein, and hence potential cancer therapies.[20] However, combing the power of structure prediction and cryoEM will help overcome this challenge and provide an opportunity for new treatments for the many BRCA1 mutation carriers. To achieve this, more experimentally determined BRCA1 structures will be needed in a variety of conditions, either alone or as part of large DNA repair or transcriptional complexes. This experimental data can then be used to train prediction methods and lead to advancements in predicting structures with comparable amino acid sequences. Combining machine learning tools with cryo-EM can help address the shortcomings of each individual technique, allowing for the identification of structural models of unknown proteins in a heterogenous endogenous mixture.[21]

There is precedence for this synergetic relationship. In recent studies, researchers have combined cryoEM and AlphaFold2 to be able to predict and refine protein structures. Terashi et al., designed a method using AlphsFold2 that refines protein structures created by cryoEM maps.[22] This synergistic relationship

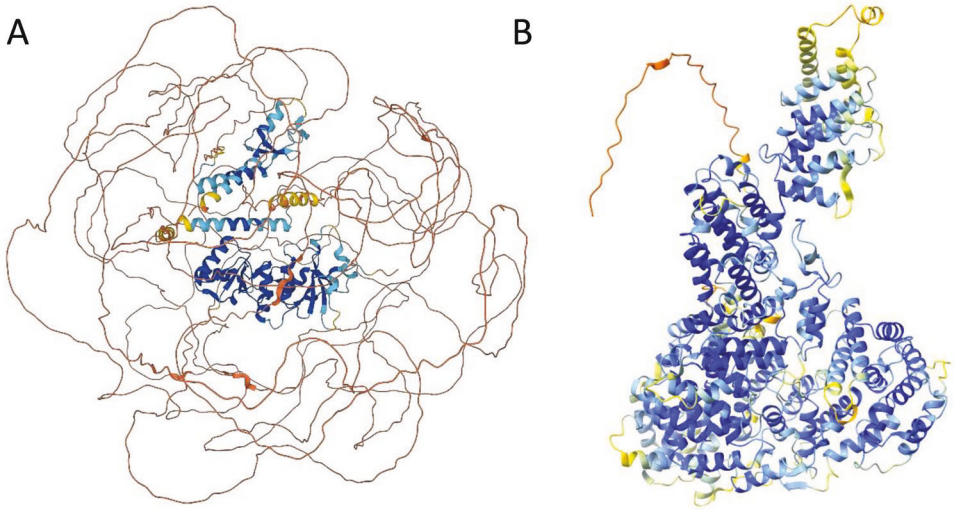

Figure 2. Challenges in AlphaFold2 predictions. A) The predicted full-length structure of the BRCA1 protein (1863 amino acids), contains a central region of disorder. Protein structure prediction methods cannot accurately model regions of large disorder. B) AlphaFold predicted the structure for Mediator Complex Subunit 23, which recently had a crystal structure uncovered to validate the precise structure.

between predictive models and cryoEM has shown to be the next step in identifying new novel interactions of the SARS-CoV-2 Omicron variant, that can push healthcare toward a precision medicine approach for the Covid-19 pandemic.[23] In this study, Mostafavi et al., describe how cryoEM is used to analyze the structure of the SARS-CoV-2 Omicron spike protein and how it forms a complex with human ACE2. They also discuss how AI, big data reservoirs, bioinformatic systems, and advanced in vitro 3D models can be employed to provide more specific therapies and better patient outcomes utilizing the experimental data from collected patient samples.[21] These powerful tools can predict, discover, and interpret unique changes in protein-protein interactions that result in altered pathways from patient samples experiencing diseases, which is critical for developing personalized and effective treatment plans. As predictive technology continues to progress, cryoEM will still be a dependable technique to experimentally confirm structural predictions. Precision medicine is rapidly advancing by integrating machine learning tools with cryoEM, allowing for accurate predictions and interpretations of patient data and human disease. This is the direction needed to provide novel strategies to find the drug at the right dose, for the right person at the right time.

Conflict of Interest

The authors declare no conflict of interest.

Author Contributions

D.C.S., A.C., and H.K.B. contributed equally to this work.

Keywords

alphaFold, cryoEM, machine Learning, precision medicine

Received: March 21, 2023
Revised: April 29, 2023
Published online: May 28, 2023

[1] J. Jumper, R. Evans, A. Pritzel, T. Green, M. Figurnov, O. Ronneberger, K. Tunyasuvunakool, R. Bates, A. Žídek, A. Potapenko, A. Bridgland, C. Meyer, S. A. A. Kohl, A. J. Ballard, A. Cowie, B. Romera-Paredes, S. Nikolov, R. Jain, J. Adler, T. Back, S. Petersen, D. Reiman, E. Clancy, M. Zielinski, M. Steinegger, M. Pacholska, T. Berghammer, D. Silver, O. Vinyals, A. W. Senior, et al., *Proteins.* **2021**, *89*, 1711.
[2] M. Baek, F. Dimaio, I. Anishchenko, J. Dauparas, S. Ovchinnikov, G. R. Lee, J. Wang, Q. Cong, L. N. Kinch, R. D. Schaeffer, C. Millán, H. Park, C. Adams, C. R. Glassman, A. Degiovanni, J. H. Pereira, A. V. Rodrigues, A. A. Van Dijk, A. C. Ebrecht, D. J. Opperman, T. Sagmeister, C. Buhlheller, T. Pavkov-Keller, M. K. Rathinaswamy, U. Dalwadi, C. K. Yip, J. E. Burke, K. C. Garcia, N. V. Grishin, P. D. Adams, et al., *Science* **2021**, *373*, 871.
[3] *Nat. Methods* **2022**, *19*, 1.
[4] M. Varadi, S. Anyango, M. Deshpande, S. Nair, C. Natassia, G. Yordanova, D. Yuan, O. Stroe, G. Wood, A. Laydon, A. Žídek, T. Green, K. Tunyasuvunakool, S. Petersen, J. Jumper, E. Clancy, R. Green, A. Vora, M. Lutfi, M. Figurnov, A. Cowie, N. Hobbs, P. Kohli, G. Kleywegt, E. Birney, D. Hassabis, S. Velankar, *Nucleic Acids Res.* **2022**, *50*, D439.
[5] J. Jumper, R. Evans, A. Pritzel, T. Green, M. Figurnov, O. Ronneberger, K. Tunyasuvunakool, R. Bates, A. Žídek, A. Potapenko, A. Bridgland, C. Meyer, S. A. A. Kohl, A. J. Ballard, A. Cowie, B. Romera-Paredes, S. Nikolov, R. Jain, J. Adler, T. Back, S. Petersen, D. Reiman, E. Clancy, M. Zielinski, M. Steinegger, M. Pacholska, T. Berghammer, S. Bodenstein, D. Silver, O. Vinyals, et al., *Nature* **2021**, *596*, 583.
[6] A. Perrakis, T. K. Sixma, *EMBO Rep.* **2021**, *22*, 54046.
[7] L. A. Kelley, S. Mezulis, C. M. Yates, M. N. Wass, M. J. E. Sternberg, *Nat. Protoc.* **2015**, *10*, 845.
[8] K. T. Simons, R. Bonneau, I. Ruczinski, D. Baker, *Proteins* **1999**, *37*, 171.
[9] Y. Zhang, *Proteins* **2007**, *69*, 108.
[10] J. N. Onuchic, Z. Luthey-Schulten, P. G. Wolynes, *Annu. Rev. Phys. Chem.* **1997**, *48*, 545.

[11] T. C. Terwilliger, D. Liebschner, T. I. Croll, C. J. Williams, A. J. McCoy, B. K. Poon, P. V. Afonine, R. D. Oeffner, J. S. Richardson, R. J. Read, P. D. Adams, *bioRxiv* **2022**.
[12] C. F. Hryc, M. L. Baker, *iScience* **2022**, *25*, 104496.
[13] Method of the Year 2015, *Nat. Methods* **2016**, *13*, 1.
[14] W. Kühlbrandt, *Science* **2014**, *343*, 1443.
[15] X.u Benjin, L. Ling, *Protein Sci.* **2020**, *29*, 872.
[16] Y. Yang, D. Arseni, W. Zhang, M. Huang, S. Lövestam, M. Schweighauser, A. Kotecha, A. G. Murzin, S. Y. Peak-Chew, J. Macdonald, I. Lavenir, H. J. Garringer, E. Gelpi, K. L. Newell, G. G. Kovacs, R. Vidal, B. Ghetti, B. Ryskeldi-Falcon, S. H. W. Scheres, M. Goedert, *Science* **2022**, *375*, 167.
[17] J. A. Clapperton, I. A. Manke, D. M. Lowery, T. Ho, L. F. Haire, M. B. Yaffe, S. J. Smerdon, *Nat. Struct. Mol. Biol.* **2004**, *11*, 512.
[18] R. S. Williams, R. Green, J. N. M. Glover, *Nat. Struct. Biol.* **2001**, *8*, 838.
[19] S. L. Clark, A. M. Rodriguez, R. R. Snyder, G. D. V. Hankins, D. Boehning, *Comput. Struct. Biotechnol. J* **2012**, *1*, 201204005.
[20] T. Gillyard, J. Davis, *Int. Rev. Cell Mol. Biol.* **2021**, *364*, 111.
[21] E. C. Nice, *Anal. Biochem.* **2022**, *644*, 113840.
[22] G. Terashi, X. Wang, D. Kihara, *Acta Crystallogr., D Struct. Biol.* **2023**, *79*, 10.
[23] S. M. Mostafavi Zadeh, F. Tajik, Y. Moradi, J. Kiani, R. Ghods, Z. Madjd, *BMJ Open* **2022**, *12*, 063748.

