CLINICAL IMPLICATIONS OF BASIC RESEARCH

Elizabeth G. Phimister, Ph.D., *Editor*

A Holy Grail — The Prediction of Protein Structure

Russ B. Altman, M.D., Ph.D.

This year's Lasker Basic Medical Research Award recognizes the contributions of Demis Hassabis and John Jumper for their invention of the AlphaFold artificial intelligence (AI) system, which predicts the three-dimensional (3D) structure of proteins from the one-dimensional (1D) sequence of their amino acids.[1] Their solution of this long-standing problem provides a path to accelerated discoveries across biomedical science.

Proteins have a pivotal role in disease: they misfold and aggregate in Alzheimer's disease, they become unregulated in cancers, they are dysfunctional in inborn errors of metabolism, and they traffic to the wrong compartment in cystic fibrosis. These are but a few among many mechanisms. Detailed structural models of proteins accelerate drug discovery by providing atomic configurations that drive the design or selection of high-affinity molecules.

Protein structures are determined experimentally with the use of x-ray crystallography, nuclear magnetic resonance, and cryoelectron microscopy. These methods are expensive and time-consuming. As a result, the database of 3D protein structures has approximately 200,000 structures, whereas DNA sequencing technology has produced more than 8 million protein sequences. In the 1960s, Anfinsen et al. showed that the 1D sequence of amino acids can fold spontaneously and reproducibly into the functional 3D conformation (Fig. 1A and video, available with the full text of this article at NEJM.org).[2] Molecular "chaperones" can accelerate and facilitate this process, but these observations created a 60-year challenge for molecular biology: predict the 3D structure of a protein from its 1D sequence of amino acids. This challenge became more pressing as our ability to obtain 1D sequences exploded with the success of the Human Genome Project.

The prediction of protein structure is difficult for a few reasons. First, there is a huge search space — every possible 3D position for each atom in each amino acid. Second, proteins chemically maximize complementary interactions to configure atoms efficiently; because proteins typically have hundreds of hydrogen bond "donors" (often oxygen) that should be close to hydrogen bond "acceptors" (often nitrogen bound to hydrogen), finding configurations such that nearly every donor is near an acceptor can be exponentially difficult. Third, there are limited examples for training from experimental methods, and therefore evolutionary information from related proteins must be harnessed to understand potential 3D interactions between amino acids on the basis of the 1D sequence. In one approach to predicting protein structure, physics is used to model the interactions of atoms as they seek optimal configurations. Karplus, Levitt, and Warshel were awarded the Nobel Prize in Chemistry in 2013 for their contributions to the computational simulation of proteins. However, physics-based methods are computationally expensive and require approximations, and therefore they have not predicted accurate 3D structures at scale. In the other, "knowledge-based" approach, databases of known structures and sequences are used to train models with the use of AI and machine learning (AI–ML). Hassabis and Jumper included elements of both physics and AI–ML, but the AI–ML provided most of the novelty and leap in performance; the two researchers creatively combined large public data repositories with industry-level computational resources to build AlphaFold.

A video describing the AlphaFold system is available at NEJM.org

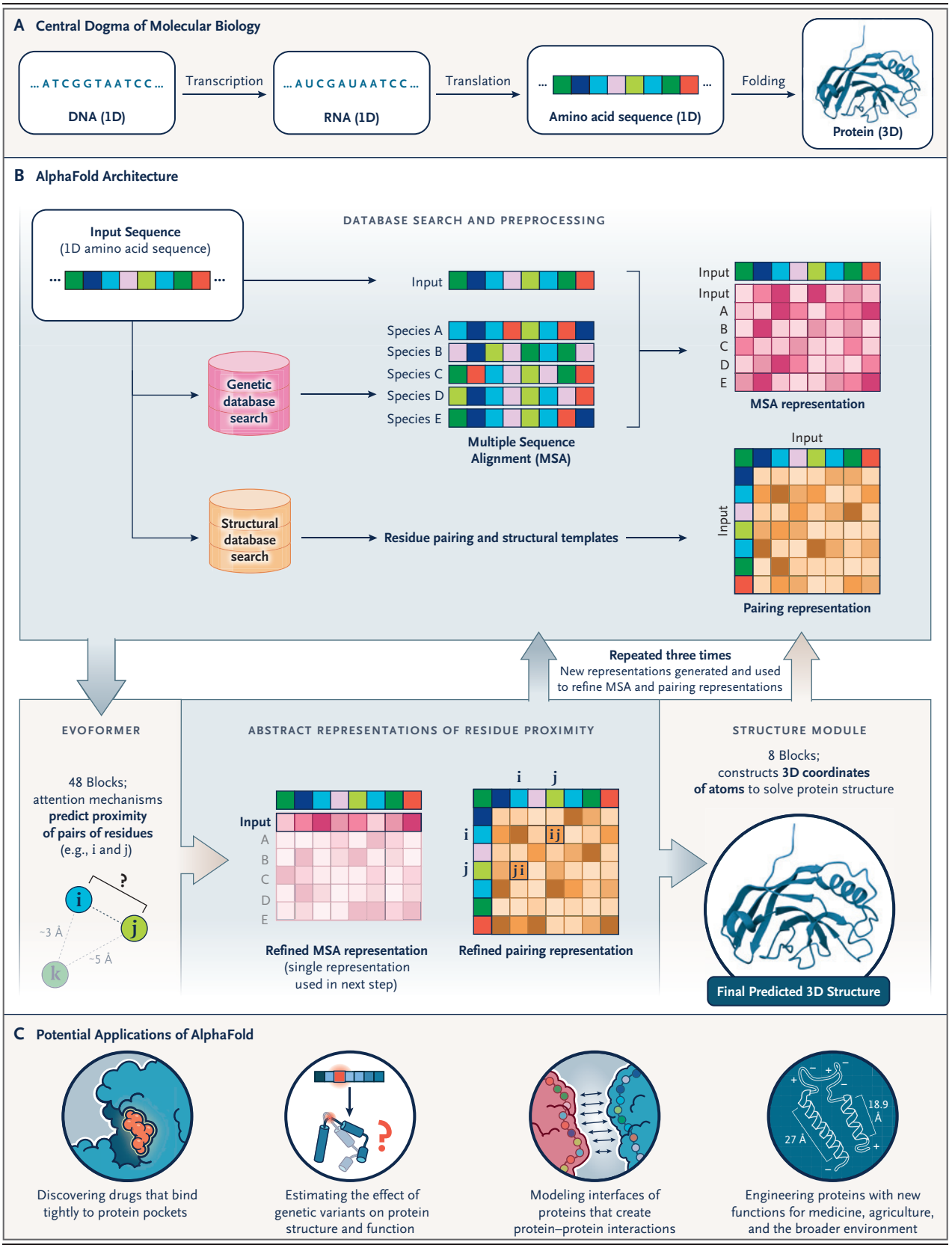

Figure 1 (facing page). Predicting Protein Structure with AlphaFold.

The central dogma of molecular biology is that the one-dimensional (1D) sequence of DNA bases (A, T, C, and G) is transcribed into mRNA (the RNA bases are A, U, C, and G), which is a working copy of the 1D sequence (Panel A). The mRNA is translated at the ribosome into a 1D protein sequence. A folded protein is the final product of the information flow that begins with genomic DNA. Anfinsen et al.[2] showed that the protein 1D sequence often has all the information required for it to reproducibly fold into the complex three-dimensional (3D) protein structures that have been solved at great expense experimentally. AlphaFold (Panel B) takes as input the 1D sequence of a protein of unknown structure and a multiple sequence alignment (MSA) of many similar proteins found in different species and tissues. It creates a deep neural-network representation of the relationship between amino acids (e.g., the pair i and j) within the protein, as well as the relationship of these two positions across the evolutionary space represented in the MSA. These representations are linked to one another and "communicate" within the Evoformer, which uses knowledge of known pairs of 1D sequence and 3D structure to infer which amino acids are proximal. The Evoformer sends this information to the structure module, which models the position of the atoms within an amino acid in 3D and seeks configurations of the atoms that are compatible with the proximities provided by the Evoformer, as well as the physical and chemical constraints on valid values for atomic bonds, angles, and torsional angles. For each modeled structure there is a multitude of potential applications (Panel C), including the design of drugs that bind tightly to protein pockets, estimation of the effect that genetic mutations have on protein structure and function, modeling of (and potentially interference with) the interfaces of proteins that create (perhaps unwanted) protein–protein interactions, and the design of new protein structures for engineering purposes.

How do we know that they "solved" the problem of structure prediction? In 1994, the Critical Assessment of Structure Prediction (CASP), which holds a meeting every 2 years, was established to track progress in structure prediction. In CASP, experimentalists disclose the 1D sequence of new unpublished structures that they have recently solved. Predictors use this 1D sequence to predict 3D structures, and assessors independently judge the quality of the predictions by comparing them with the 3D structures provided (to them alone) by the experimentalists. CASP provides truly blinded assessments and has documented periodic jumps in performance associated with methodologic innovations. However, the predictions by AlphaFold that were presented at the fourteenth CASP meeting, in 2020, showed a leap in performance so great that the organizers declared the problem of 3D structure prediction to be solved: most of the predictions had accuracy similar to that of experimental determinations.

How did it work so well? Hassabis and Jumper made an array of technical innovations, including an end-to-end (1D-to-3D) prediction pipeline that is differentiable, so that all parameters in their model can be simultaneously optimized; encodings of both the input 1D sequence and a sequence alignment of its evolutionary neighbors that collaborate to predict the relative proximities of amino acids (Fig. 1B); AI–ML computational "attention" mechanisms to simplify the search space by detecting which (among many) interactions were the most important in predicting 3D proximity; and a module for refining proximity predictions into detailed 3D atomic configurations. The ideas they introduced have, as expected, spurred a flurry of innovations borrowing and extending their ideas.[3]

AlphaFold is already being used to drive the creation of new drugs (Fig. 1C). It is helping to illuminate the "dark matter" of the proteome; previously unseen structures can now be modeled in search of their function. Protein designers are using it to refine their designs for engineered proteins. AlphaFold is also providing initial structures that are refined with experimental data, enabling models for large cellular machines that implement transcription, translation, replication, force generation, degradation, recycling, and other processes. AlphaFold does leave some important elements of 3D structure unsolved, including the modeling of variant (mutant) proteins, the path of folding from 1D to 3D, the temporal dynamics of proteins, and the way in which structure links to experimentally measured function. Nevertheless, it provides a starting point for progress on each of these fronts.

More generally, the work of Hassabis and Jumper provides a compelling demonstration of how AI–ML is transforming science. They showed that AI–ML can build complex scientific hypotheses from multiple sources of data, that attention mechanisms (similar to those in ChatGPT) can discover the key dependencies and correlations within data sources, and that AI–ML can

introspectively judge the quality of its outputs. AI–ML is essentially doing science.

Disclosure forms provided by the author are available with the full text of this article at NEJM.org.

From the Departments of Bioengineering, Genetics, Medicine, and Biomedical Data Science, Stanford University, Stanford, CA.

This article was published on September 21, 2023, at NEJM.org.

1. Jumper J, Evans R, Pritzel A, et al. Highly accurate protein structure prediction with AlphaFold. Nature 2021;596:583-9.
2. Anfinsen CB, Haber E, Sela M, White FH Jr. The kinetics of formation of native ribonuclease during oxidation of the reduced polypeptide chain. Proc Natl Acad Sci U S A 1961;47:1309-14.
3. Lin Z, Akin H, Rao R, et al. Evolutionary-scale prediction of atomic-level protein structure with a language model. Science 2023;379:1123-30.

DOI: 10.1056/NEJMcibr2307735