Madelyn Aggouras
maggouras@fordham.edu

Linda Chuong
lchuong1@fordham.edu

Mohammed Uddin
mu4@fordham.edu

Samiha Chowdhury ·
schowdhury90@fordham.edu

Link to papers github: *https://github.com/deepmind/alphafold*

1. Biological problem the authors are trying to solve

The biological problem that the authors are trying to solve is predicting and modeling 3d protein structures from their amino acid sequence. This will allow for a better understanding of how protein structure determines its function and its role in biological systems. So far, only a tiny fraction of protein structures are known out of the billions of protein sequences. This creates a gap in research between the known sequences and their known structure. If the structures are mapped, can more research be put into utilizing the structures rather than just figuring out the structure and its foldings.

2. The computational approach that the authors are proposing to use

The authors propose a computational approach centered on AlphaFold, a deep learning system that predicts protein structures with near-experimental accuracy. Unlike prior physics-based or purely evolutionary methods, AlphaFold integrates both biological and physical constraints into a novel neural network architecture. Its design leverages multiple sequence alignments (MSAs) and structural templates as inputs, processed through the “Evoformer” module, which models evolutionary and spatial relationships, followed by a structure module that directly outputs 3D atomic coordinates. The system incorporates iterative refinement (“recycling”), equivariant attention to respect geometric invariances, and self-distillation to exploit vast unlabeled sequence data. This hybrid strategy allows AlphaFold to generate highly accurate, end-to-end predictions of full protein structures even when no close structural homologues are available

3. The most important references the paper relies on

The AlphaFold paper cites approximately 20 key references that form the foundation of its methodology and provide important scientific context. Among these, several stand

out as particularly influential due to their significant contributions to the fields of protein structure prediction, deep learning, and computational biology. For instance, Jumper et al., 2011 laid important groundwork by exploring early neural network approaches to protein folding. The landmark study by Senior et al., 2020 introduced advanced deep learning techniques that dramatically improved the accuracy of protein structure predictions, directly influencing the development of AlphaFold. Similarly, Jones et al., 2015 contributed vital methods for protein structure alignment and contact prediction, which are essential components for building reliable structural models. The work by Zhang et al., 2019 emphasized the integration of evolutionary information into protein modeling, an approach that AlphaFold harnesses to enhance prediction accuracy. Finally, foundational insights from Baker & Sali, 2001 on comparative modeling and the role of structural templates in protein folding continue to underpin many computational prediction frameworks. These references, along with others, are comprehensively listed at the end of the AlphaFold paper under the section titled 'Online Content,' providing readers with a detailed roadmap of the scientific literature that supports this breakthrough work.

4. Reasons why you will not be able to replicate the study's findings using their GitHub repository

   Looking at the GitHub repository for DeepMind's "Highly accurate protein structure prediction with AlphaFold," there are a couple of reasons why the study's findings cannot be fully replicated using the repository alone. While the pretrained model weights are publicly available, they are not included directly in the repository and must be downloaded separately. The exact protein sequence and structure databases used in the study may also be inaccessible, which can affect the results. Differences in hardware, such as GPU type and memory, as well as variations in software environments and library versions, can further influence outcomes. Additionally, stochastic elements in the model, like random initialization and dropout, may cause variability between runs. Finally, certain preprocessing or postprocessing steps used in the study are not fully documented, making exact replication challenging for someone trying to replicate.