AlphaFold Structure Determination

**Shallee T. Page** & Rachel Sterne-Marr

# Objectives

* Explain how proteins fold based on primary structure (amino acid sequence).
* Use the AlphaFold to predict tertiary structure of:
* A precalculated protein
* A protein *ab initio*
* Assess the reliability of an *ab initio* prediction
* Analyze protein predictions for sequence-structure relationships

#### Prerequisites

* Define primary, secondary, tertiary and quaternary structure of proteins.
* Explain the role of intermolecular forces in protein folding.
* Briefly describe how the structure of a protein is determined by X-ray crystallography

#### Resources & Tools

* RCSB Protein Data Base: <https://www.rcsb.org/>
* Jumper *et al.* 2021. Highly Accurate Protein Structure Prediction with AlphaFold. Nature. 596: 583-589. <https://www.nature.com/articles/s41586-021-03819-2>
* AlphaFold, a Practical Guide, EMBL: <https://www.ebi.ac.uk/training/online/courses/alphafold/accessing-and-predicting-protein-structures-with-alphafold/choosing-how-to-access-alphafold/>
* AlphaFold Protein Structure Database: <https://alphafold.ebi.ac.uk/>
* Gene Record Finder: [thegep.org/finder](https://thegep.org/finder)
* AlphaFold2 for *ab initio* prediction: <https://colab.research.google.com/github/sokrypton/ColabFold/blob/main/AlphaFold2.ipynb>

Table of Contents

[Objectives 1](#_Toc172645975)

[Introduction 2](#_Toc172645976)

[Learning Outcomes 2](#_Toc172645977)

[Part 1: Deciding what program to use 3](#_Toc172645978)

[Part 2: AlphaFold *ab initio* 4](#_Toc172645979)

[Investigation: Finding protein structures 6](#_Toc172645980)

[I. Investigation 1: Experimentally Determined Structures 7](#_Toc172645981)

[II. Investigation 2: PreComputed Structures 8](#_Toc172645982)

[III. Investigation 3: AlphaFold Prediction ab initio 9](#_Toc172645983)

# Introduction

In this module, you will examine how to model protein structure based on amino acid sequence. In recent years, computer software has gotten much better at predicting how primary structure dictates secondary and tertiary structure and AlphaFold is one of the best algorithms for doing so. We will explore the program, its strengths and weaknesses and how to use it to look at proteins.

# Learning Outcomes

In this modules, students will:

…explain the role of the AlphaFold program

…decide whether AlphaFold is needed

…find three-dimensional protein structures in the appropriate database

…use AlphaFold to predict a protein structure

# Part 1: Deciding what program to use

Before we begin the analysis, we will want to determine if a structure is already in the database that matches your protein of interest.

1. If an X-Ray crystallography (XRC) structure exists, that is the gold standard for 3-dimensional protein structure. For mobile or intrinsically disordered proteins, or transmembrane proteins, there may be nuclear magnetic resonance (NMR) structures available. You can search for these experimentally determined structures at the [RCSB website.](https://www.rcsb.org/) (Figure 1)

A screen shot of a computer

Description automatically generated

Figure 1-Screenshot of the search bar of RCSB.

1. If the structure has not been determined by NMR or XRC, then AlphaFold may be useful for structure prediction. For some proteins or protein families, AlphaFold has already been used to predict a structure and you can look up that pre-computed structure in AlphaFold. This saves time and computational power. These can also be searched at the [RCSB website](https://www.rcsb.org/) (Figure 2). You can also look for these by entering amino acid sequences in the [AlphaFold website](https://alphafold.ebi.ac.uk/) (Figure 3). You can choose to do either.

A blue arrow pointing to a blue box

Description automatically generated

Figure 2-Screenshot of the RCSB website, showing the link for precomputed structures

A screenshot of a computer

Description automatically generated

Figure 3-Screenshot of AlphaFold Structure Database

1. If there are no structures in the AlphaFold database, you can run the program to determine the structure from scratch (*ab initio*) using the amino acid sequence. ([Part 2](#_Part_2:_Alpha), below)

# Part 2: AlphaFold *ab initio*

For proteins sequences that don’t match to something in the AlphaFold database, we can use AlphaFold to do a prediction in the cloud. The collaborative folding site is in the cloud and allows us to use significant computing resources without downloading or running the program on our computers.

That ColabFold site has a sample sequence preloaded (Fig. 4).

A screenshot of a computer

Description automatically generated

Figure -ColabFold Website

To use that test sequence, you will just assign a new jobname, and leave the defaults in place (“Investigation” below). By clicking the white-on-gray arrows for each section of code, we will run each section. This type of page with code to run a set of mini-programs is called a “notebook”.

The sequence coverage plot (Fig. 5) shows us if there are similar sequences in the database that AlphaFold can compare to. The x-axis is the number of the amino acid in the primary sequence query (almost 60 amino acids in the input). Each sequence in the database that matches is a horizontal line and they are color coded (0-1 or 0-100%) where blue means that the sequence is almost identical and the red means lower of sequence identity. Those horizontal lines for the sequence matches are sorted in order of increasing sequence identity from bottom to top. The coverage is the length of each line that aligns with the query sequence.

A screen shot of a graph

Description automatically generatedSo, in this search, the program found over 8000 sequences in the database that match the query that we inputted. Most of those sequences are orange lines, indicating around 20 % sequence identity. The really good matches at the top are blue (almost 1.0). Some of the lines only cover a portion of the

Figure -Sequence Coverage Plot

sequence; those regions that are absent or don’t match are colored white. You can see most of the horizontal lines show about 80% coverage of the sequence. The black line shows that the coverage is greatest for all the 8000+ sequences in the middle; the N-terminus coverage is lower and the C-terminus coverage is even lower.

Below the sequence coverage plot is our structure prediction (Fig. 6).

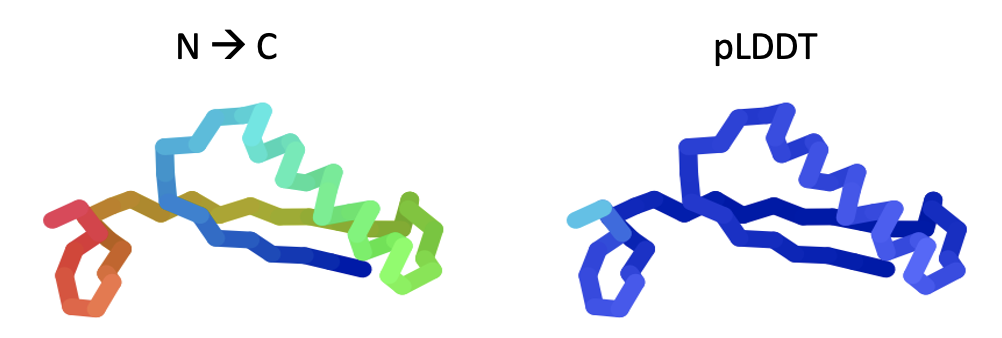


Figure -Sample Structure Prediction.

The lefthand picture is coloring the structure in a rainbow from the N-terminus (blue) to the C-terminus (red). The righthand image shows the predicted local distance difference test (pLDDT) score, a measure of how confident AlphaFold is for each section (closer to 100% in blue is good, less than 40% in red is much less confident. The blue above indicates high confidence). Low confidence might mean a flexible region, an intrinsically disordered region, or just not enough similar structures to compare to. There are alternative possible structures below.

# Investigation: Finding protein structures

In this walkthrough, we will look up an experimentally determined structure, two pre-computed structures and then we will use AlphaFold to examine the structure of a protein that is unknown.

At the ribosome, mRNA is translated into a chain of amino acids. The sequence of the amino acids is known as the protein *primary structure*. As those peptide chains are produced, each amino acid is exposed to a variety of intermolecular forces in three dimensions that cause the protein to fold up.

The dipole-dipole interactions from the water exclude the hydrophobic amino acid side chains, encouraging them to collapse into a hydrophobic core dominated by Van der Walls interactions. Intramolecular dipole-dipole interactions, ion-dipole interactions, and ion-ion interactions (“salt bridges”) also form within the protein as it folds up into *secondary structures*. As it folds up, the intermolecular interactions between water and the amino acids help to stabilize the structure. The shape and domain structure of the folded protein is known as the *tertiary structure.*

There are several methods for determining the protein structure. Two of the most powerful are X-Ray Crystallography (XRC) and Nuclear Magnetic Resonance (NMR). In the method of XRC, proteins are allowed to form regular crystal forms and then subjected to X-rays. When the X-rays hit the electron clouds of the protein atoms, they are diffracted in various directions and hit an X-ray detector. By tracing the trajectory of the X-ray path backward, it can be determined where the protein electron clouds were, and thus the shape of the protein. At high resolution, it is clear what amino acids are present, their position and orientation. NMR is done by subjecting the protein in solution to a magnetic field and then to radiofrequency electromagnetic energy. This disrupts the spin of the nuclei such that it can be shielded by the nearby electron clouds. In this way, it can be calculated which amino acids are nearby each other in the protein. NMR structures are often lower resolution than XRC and are limited to small proteins. NMR, however, is useful for proteins that don’t readily crystallize, and NMR can measure molecular motions of the proteins. An NMR study results in several predicted structures for a protein.

Because there are a huge number of potential interactions between each of the amino acids, and between the amino acids and water, there is an astronomical number of possible ways that an amino acid chain can fold into a protein. For a long time, it was thought that we would never have enough computational power to predict protein structure from the primary structure. But, computers are getting better and better at making these predictions.

AlphaFold uses artificial intelligence to predict protein structure from the primary sequence. It doesn’t do it completely *de novo*. It looks at multiple sequence alignments with known structures, and it maps two-dimensional interactions from three dimensionals structural information. With proteins with potential similar blocks and domains, it creates a model. It calculates how well the model fits, and then recalculates the structure prediction dozens of times, finally generating a three-dimensional model. At this point in time, AlphaFold and RoseTTAFold are the most powerful tools for predicting protein structure.

## I. Investigation 1: Experimentally Determined Structures

Let’s look for the structure of the *Drosophila melanogaster (Dmel)* glycolytic protein enolase.

When you go to PDB (<https://www.rcsb.org/> ), and type “melanogaster enolase”, a dropdown appears suggesting the crystal structure of the enolase (Fig. 7). For now, do not click the toggle switch that says “include CSM”.

A screenshot of a computer

Description automatically generated

Figure -PDB Search for Dmel enolase

1. Clicking on that protein takes you to the record of the enolase (Fig. 8). Notice that it shows the method (X-ray crystallography) and the resolution (2.015 Angstroms, where the smaller number is better resolution). The block letters at the top (“5WRO”) is a stable accession number; any future search using that accession code will bring you to that exact record in the database.

A close-up of a computer screen

Description automatically generated

Figure -Dmel enolase PDB record

1. You can click on the box beneath the thumbnail structure to view and move the structure.

**Q1.** Make a sketch showing the diffraction in X-Ray Crystallography.

**Q2.** Is the 2.015 Angstroms considered high resolution?

## II. Investigation 2: PreComputed Structures

1. Let’s look for the structure of Neprilysin-2. Return to the PDB database.

In the advanced search box, (Fig. 9), type “computed melanogaster Neprilysin-2”. (include the inside quotation marks). Make sure the button is toggled **on** that says “Include Computed Structure Models (CSM)”. Then click “Search”. The best hit will be a compound related to neprilysin-2 (neprilysin-2-like). The result will provide information about the structure and how it was determined (Fig. 10).

A screenshot of a computer

Description automatically generated

Figure -Neprilysin computational structure search

A screenshot of a computer

Description automatically generated

Figure -Result of Neprilysin-2 search.

**Q3.** Describe the secondary and tertiary structure predicted.

**Q4.** Is the program confidence high or low? What regions seem to be of lower confidence?

## III. Investigation 3: AlphaFold Prediction *ab initio*

Let’s look for a structure of a protein of unknown function. This is a protein from a bacteriophage named Chickadee whose sequence indicates it might bind to DNA but that is all we know.

1. Navigate to the ColabFold site. <https://colab.research.google.com/github/sokrypton/ColabFold/blob/main/AlphaFold2.ipynb#scrollTo=AzIKiDiCaHAn>
2. Copy the sequence of the unknown protein: MAEKLHDLRSFAEVAGVKYTTMRRYHATATKRRAEAAADPEVKLPAWLIPPPDDRVGQSPVWRDRTVRKWIESRPRAATHATT
3. Paste the sequence into the “query sequence” search box for AlphaFold (red arrow in Fig. 11).

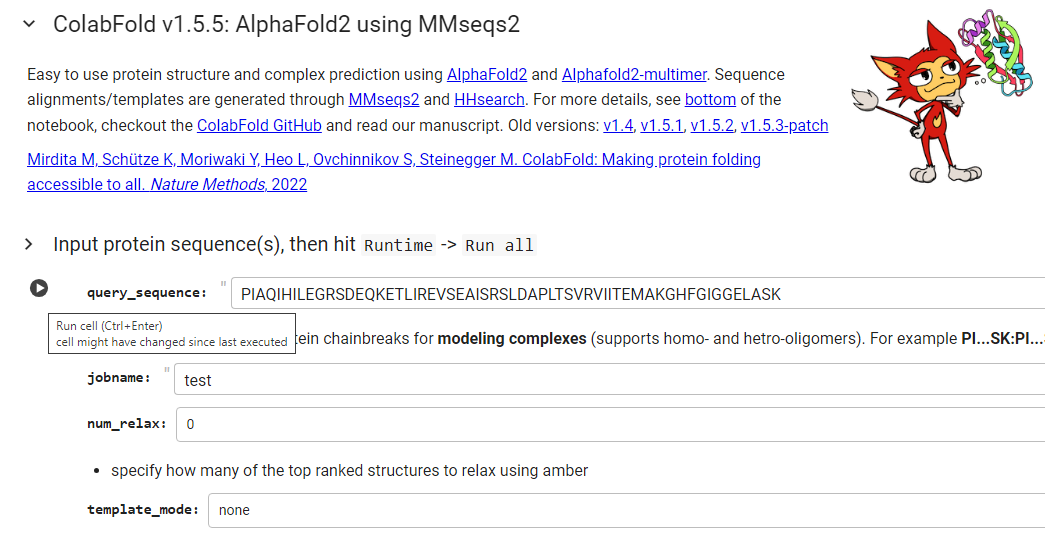


Figure 11-Screenshot of Query Sequence Data Box

1. Click on the white-on-gray arrow to the left of query sequence and wait a few seconds until a wee green arrow appears. (A warning will probably come up that the code was not written by Google; if so, click “run anyway”.)
2. In the next box down, “jobname”, it says “test”. Change that to give it your own name, like “Chickadee gp87”.
3. Scroll down to the next box of code called “Install dependencies”. The white-on-gray arrow should appear as you mouse over. The first arrow says, “show code”. Click on that and wait for the green arrow. The next one down says, “MSA Options”----click on that (Fig. 12, red arrow).

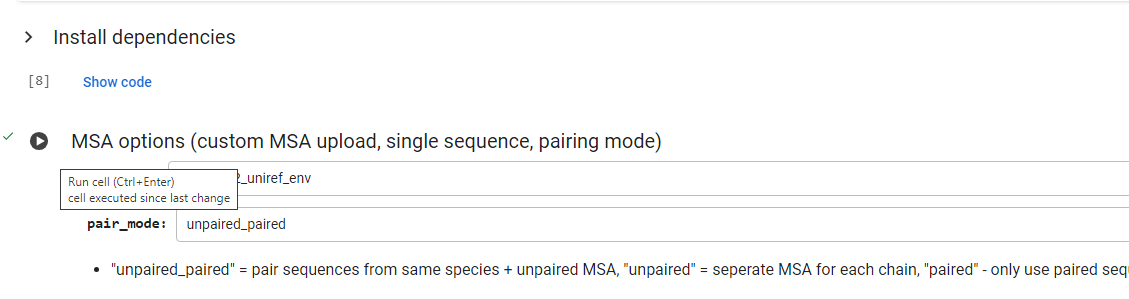


Figure -Screenshot of MSA Options

1. Similarly, click on the arrow that says, “Advanced settings”. We are not going to change any of those setting (use the default settings that are already there). MSA options and Advanced Settings code chunks should take almost no time to run.
2. Scroll down to “Run Prediction”. This is the big launch. You can do a countdown 3-2-1 if you want. Make sure the blue “display images” checkbox is checked (Fig 13, green arrow). Click the white-on-gray arrow (Fig 13, red arrow). This one may take a minute to run, depending on the time of day.

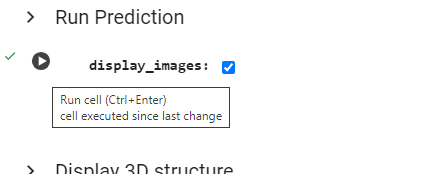


Figure 13-Screenshot of Run Prediction Code Block. Check green arrow first then click at the red arrow.

1. **Alternative:** Note that you can skip steps 7-11 and just enter the protein sequence and go to the menu under the drop-down menu “Runtime” and choose “run all” (Fig. 14). When the server is getting a lot of jobs, this may not work as well.

A screenshot of a computer

Description automatically generated

Figure 14-Screenshot of Run All command.

1. Look at the first graph “sequence coverage”.

**Q5.** Sketch the sequence coverage graph shown or include a screenshot.

**Q6.** Explain what the graph tells you about the structure prediction.

Scroll down and look at the structure prediction images.

**Q7.** Describe the secondary and tertiary structure of the prediction.

**Q8.** What does the pLDDT structure tell you? How does it relate to the sequence coverage graph?

**Q9.** As you scroll down, below the predicted structure are some alternate structures. Are the alternate structures significantly different? If so, explain what regions differ and why you think that might be.

1. Scroll down to the “Display 3-D Structure”. Click on the white-on-gray arrow (Fig 15, red arrow).

Click on the structure and drag to rotate it in three dimensions.

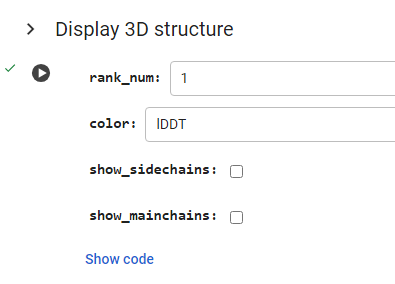


Figure 15-Screenshot of 3D Structure codeblock.

**Q10.** Examine each of the regions. Are they low or high confidence? Explain what is notable about each.

1. Go back to the box above “Display 3D Structure” Click the small box that says, “show\_sidechains”. Rerun the code and examine the new 3D structure. If you have a pad or mouse touchpad, you can put both fingers on and stretch them out to zoom in.

**Q11. Optional** **-ask your instructor -**Examine the amino acids in the central low confidence region. Identify the amino acids and speculate on what each one is interacting with.

**Q12. Optional** **-ask your instructor -**Examine the amino acids in the central low confidence region. Identify an amino acid that is inconsistent with secondary structure in that region.

**Further Explorations**

If your instructor directs you to do so, scroll down to “Package and Download Results” and save the results of the AlphaFold run. The zip file contains:

* PDB formatted structures sorted by avg. pLDDT and complexes are sorted by pTMscore. (unrelaxed and relaxed if use\_amber is enabled).
* Plots of the model quality.
* Plots of the MSA coverage.
* Parameter log file.
* A3M formatted input MSA.
* A predicted\_aligned\_error\_v1.json using AlphaFold-DB's format and a scores.json for each model which contains an array (list of lists) for PAE, a list with the average pLDDT and the pTMscore.
* BibTeX file with citations for all used tools and databases.

You can import cif format into PyMol:. You can also pull the entire zip file into PyMol: Open PyMOL and click Plugin > Plugin Manager. Then, click the Install new plugin tab, select Choose file, and select the .zip. Choose a directory to save it, then unzip the files.