Mapping protein interactions with AI

Ingelise Holland-Kaye[1], Siu-Shing Wong[1], Jordan Raff[1]

[1]Sir William Dunn School of Pathology, University of Oxford Contact: ingelise.holland-kaye@path.ox.ac.uk

Biotechnology and Biological Sciences Research Council

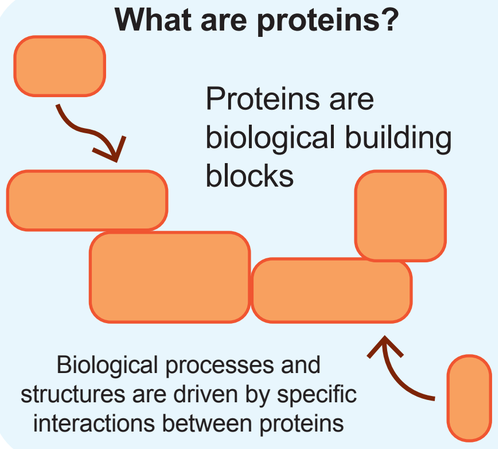
1. Summary

- Understanding **which proteins interact** and **how** is key to understanding biological structures and systems
- AlphaFold is an AI tool that predicts protein interaction structures
- To adapt AlphaFold for **large scale interaction screening**, new methods are needed to reduce numbers of false positive interactions
- Here we present methods to improve screening in AlphaFold, by fragmenting input sequences to improve accuracy of output, and analysing predictions using a combination of new and existing metrics

2. Proteins

What are proteins?

Proteins are biological building blocks

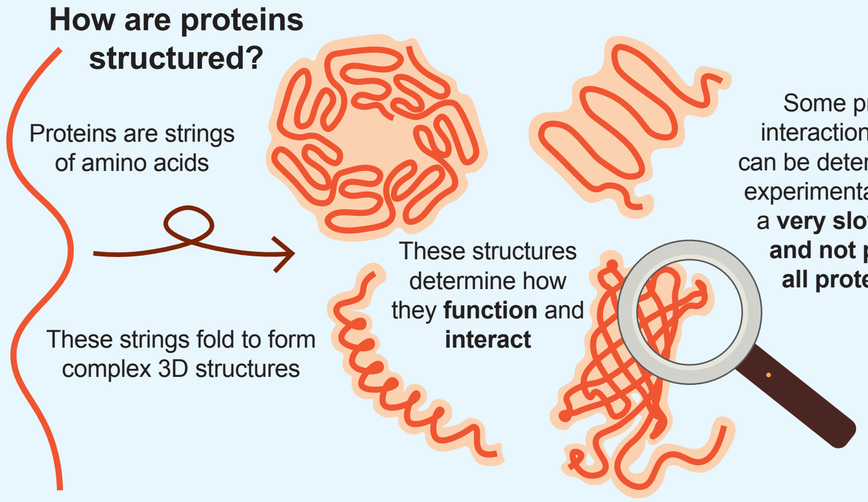


Biological processes and structures are driven by specific interactions between proteins

How are proteins structured?

Proteins are strings of amino acids

These strings fold to form complex 3D structures

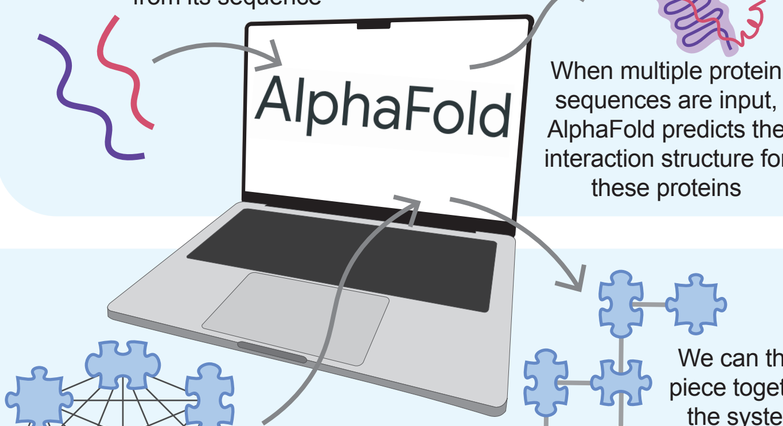


Some protein and interaction structures can be determined experimentally, but this is a **very slow process and not possible for all proteins**

3. Screening for protein interactions with AlphaFold

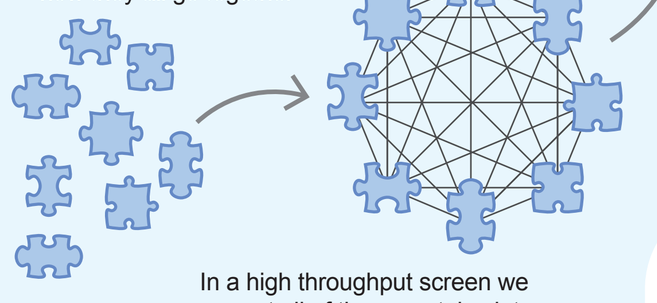
What is AlphaFold?

AlphaFold is an AI based program that predicts the 3D structure of a protein from its sequence



When multiple protein sequences are input, AlphaFold predicts the interaction structure for these proteins

For many systems and structures, we know all the proteins involved, but not how they all go together



In a high throughput screen we can put all of these proteins into AlphaFold, in every possible combination of pairs

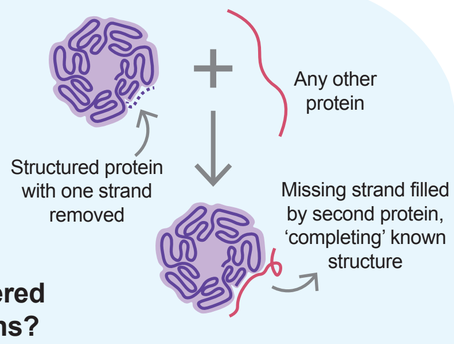
We can then piece together the system based on which pairs AlphaFold thinks interact

An issue with this is that **AlphaFold assumes that input proteins will interact, and predicts how** - so interaction structures are still predicted for non interacting proteins

4. Optimising screen input

AlphaFold makes **more accurate** predictions, and is **more confident** in its predictions, when **smaller proteins** are input

Cutting large proteins into sections before running them through AlphaFold can improve prediction accuracy

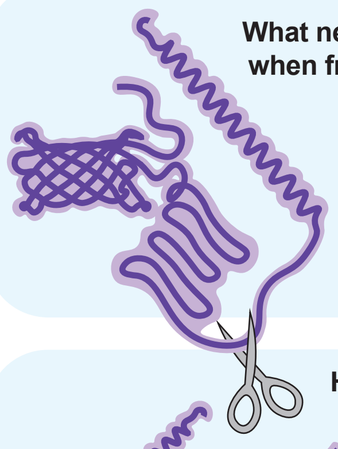


What needs to be considered when fragmenting proteins?

Proteins have distinct structural features, which often carry out separate interactions

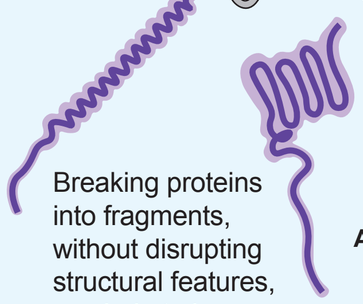
When these features are broken, AlphaFold often makes impossible predictions to try to 'complete' them

This means **fragmentation can be disadvantageous if it breaks these structured regions**



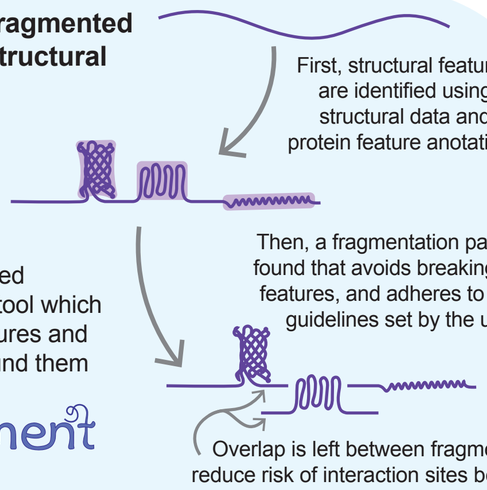
How can proteins be fragmented without disrupting structural features?

Breaking proteins into fragments, without disrupting structural features, can help to improve screening accuracy



We have developed **AlphaFragment**, a new tool which identifies structural features and fragments proteins around them

AlphaFragment



First, structural features are identified using structural data and protein feature anotations

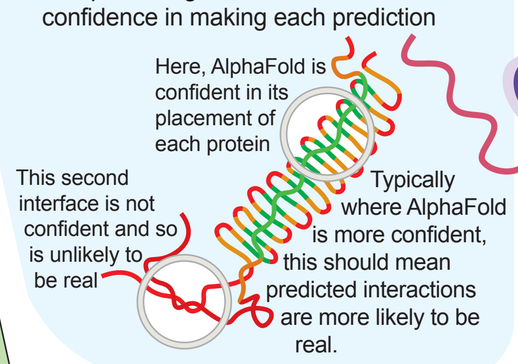
Then, a fragmentation pattern is found that avoids breaking these features, and adheres to length guidelines set by the user

Overlap is left between fragments to reduce risk of interaction sites being split

5. Analysing screen output

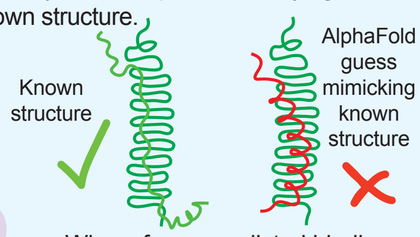
AlphaFold confidence

AlphaFold gives measures of its confidence in making each prediction



Protein promiscuity

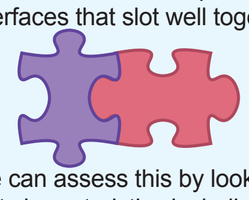
AlphaFold is trained on known protein and interaction structures. When certain structures are overrepresented in this data, proteins may have many incorrect predictions trying to mimic a known structure.



Where fewer predicted binding partners are seen for a protein, this may make these predictions more likely to be true

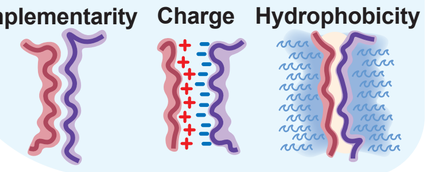
Interface fit

Proteins that interact are expected to have interfaces that slot well together



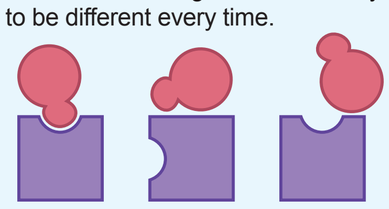
we can assess this by looking at characteristics including:

Complementarity **Charge** **Hydrophobicity**

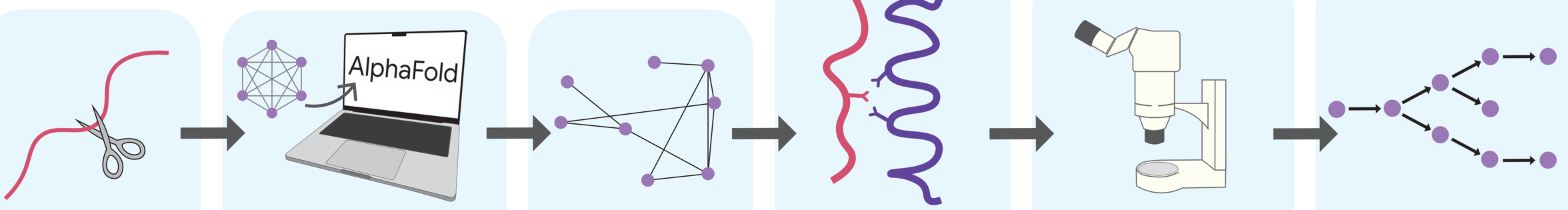


Prediction repeatability

Real interfaces are more likely to be predicted the same across multiple AlphaFold runs, whereas random guesses are likely to be different every time.



6. Protein interaction screening workflow



Fragment protein sequences, keeping structural features intact

Use AlphaFold to predict interactions between each fragment of the proteins of interest

Narrow down likely interactions using new metrics

Use prediction structures to identify the structural basis of interactions

Test predictions in lab - see whether blocking predicted interaction affects cell function

Use confirmed predictions to put together the protein interaction network

Jumper, J. et al. "Highly accurate protein structure prediction with AlphaFold." Nature, 596, pages 583–589 (2021). DOI: 10.1038/s41586-021-03819-2