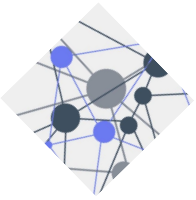


# Cas12a Variant Structural Predictions With AlphaFold2

Jason Laird, Bioinformatics Scientist <sup>1</sup>

1. Research Technology, TTS, Tufts University





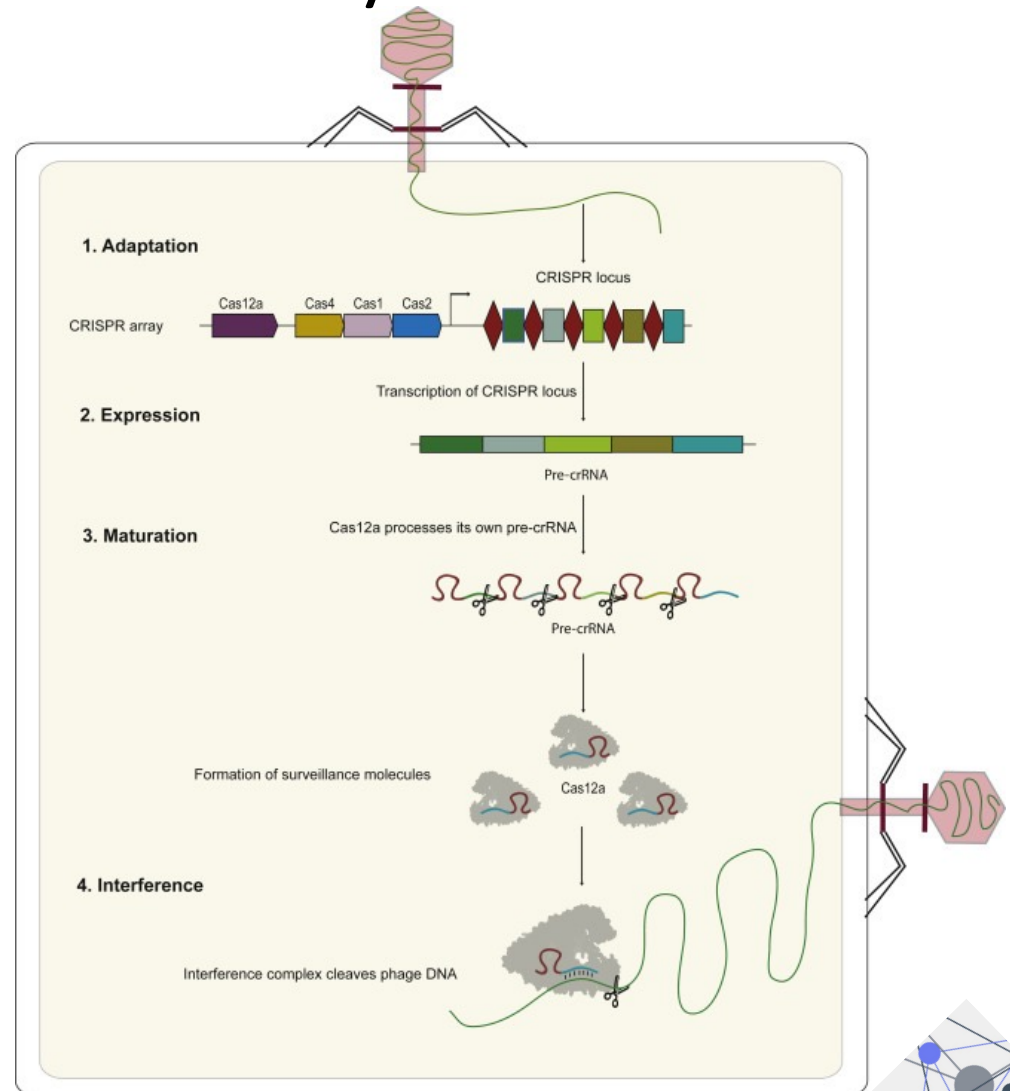
# Overview

- CRISPR-Cas Systems
- CRISPR-Cas9 & CRISPR-Cas12a
- Cas12a Variant Structural Predictions With AlphaFold2
- Visualization with PyMOL



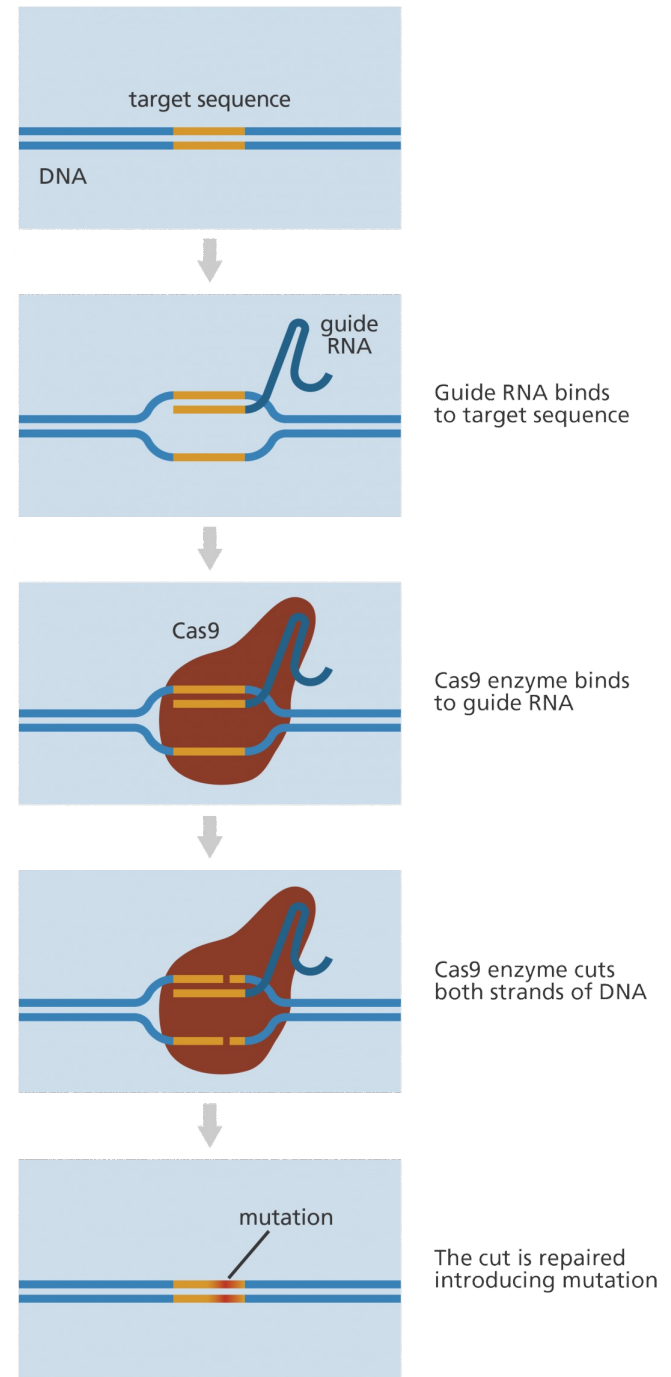
# CRISPR-Cas Immunity

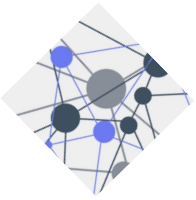
- Used by Prokaryotes as a defense against mobile genetic elements and can be broken into 3 main stages
- **Adaptation:** identification and incorporation of the protospacer into the CRISPR array
- **Expression/Maturation:** CRISPR array is transcribed into pre-CRISPR RNA
- **Interference:** CRISPR RNA complexes with effector protein to cut a target DNA sequence



# CRISPR-Cas Genome Editing




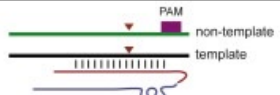
- We can leverage this system by swapping the guide RNA that binds to the target sequence
- Cas proteins will then cleave this sequence - Opening the possibility of introducing sequences of interest into the genome



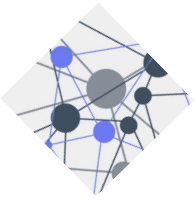


# Cas12a versus Cas9

- CRISPR-Cas9 enzymes have been widely characterized over the past decade
- Cas12a differs from Cas9 in several ways:
  - only needs a single CRISPR RNA molecule rather than two
  - Has a single nuclease site
  - Produces a staggered double strand break

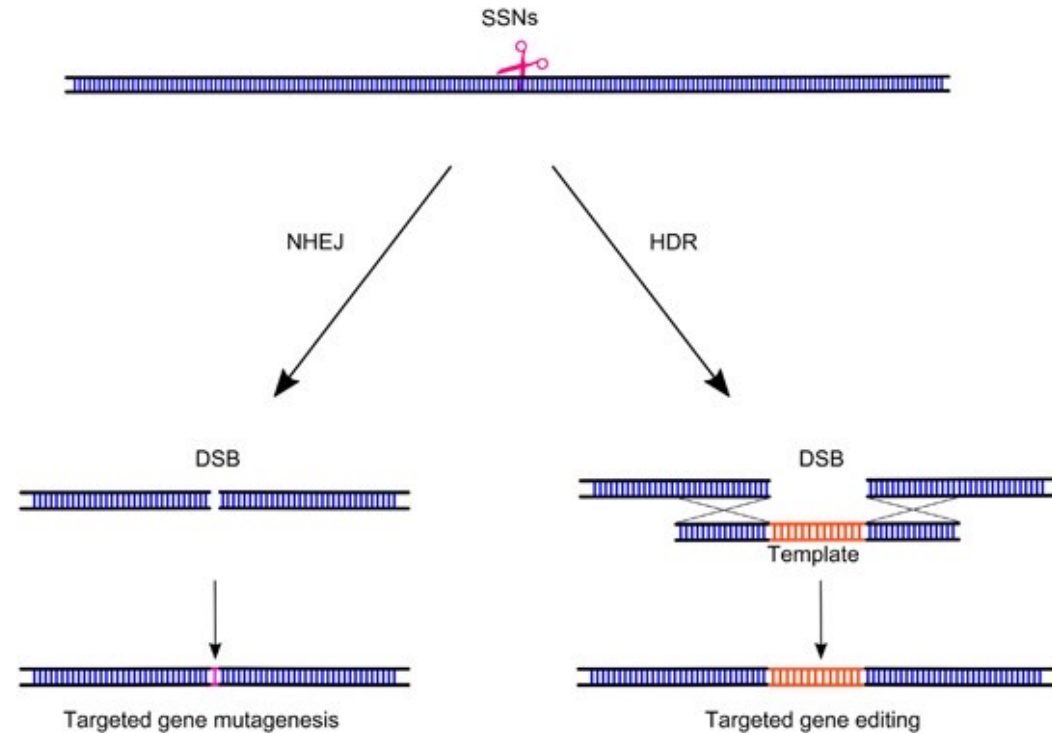
	Cas12a	Cas9
Size of protein	~1300 amino acids	~1000-1600 amino acids
RNA	 crRNA Single RNA molecule	 crRNA tracrRNA Two RNA molecules
Nuclease sites	Single nuclease site RuvC-Nuc	2 nuclease domains HNH and RuvC
Type of cut	 PAM non-template template Staggered ends	 PAM non-template template Blunt ends
PAM requirements	Recognises 5' T-rich PAM sequences of 3-4 nt	Recognises 3' G-rich PAM sequences of 3-5 nt
precrRNA processing	possesses intrinsic RNase activity to process precr-RNA	requires host RNase III and tracrRNA

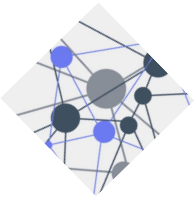




# The Type of Cut Matters

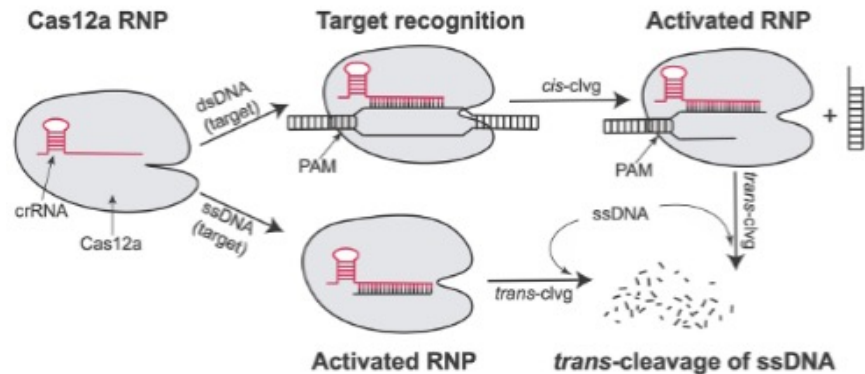
- CAS12a produces a staggered double strand break – which can be mended by Homology directed repair (HDR) **instead** of non-homologous end joining (NHEJ)
- HDR utilizes template DNA to repair the break site
- NHEJ will directly join the two broken ends. Tends to be more error prone

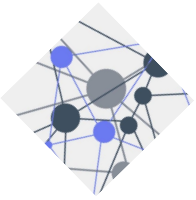




# Cas12a: Disadvantages

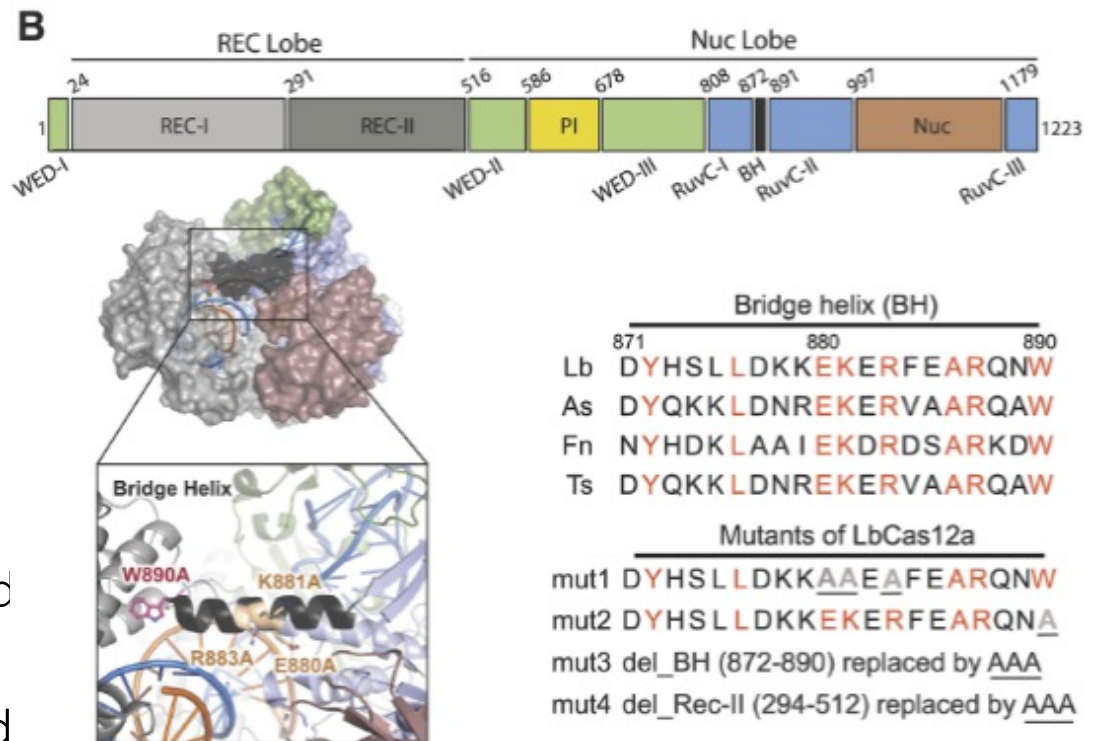
- While CAS12a encourages less error prone HDR of double stranded DNA breaks (**cis-activity**), it will also indiscriminately degrade single stranded DNA (**trans-activity**). Potentially introducing off target effects
- So, to use Cas12a as a genome editing tool, it would be beneficial to reduce this **trans-activity**





# Today's Study

- Today we will be looking at a study by Ma et al. 2022, where they engineer Cas12a variants with reduced trans-activity while maintaining cis-activity
- They start by screening multiple mutants and identify mutant 2 as having diminished trans-activity
- Variants were then introduced in mutant 2 to create a variant with diminished trans-activity, and maintained cis-activity

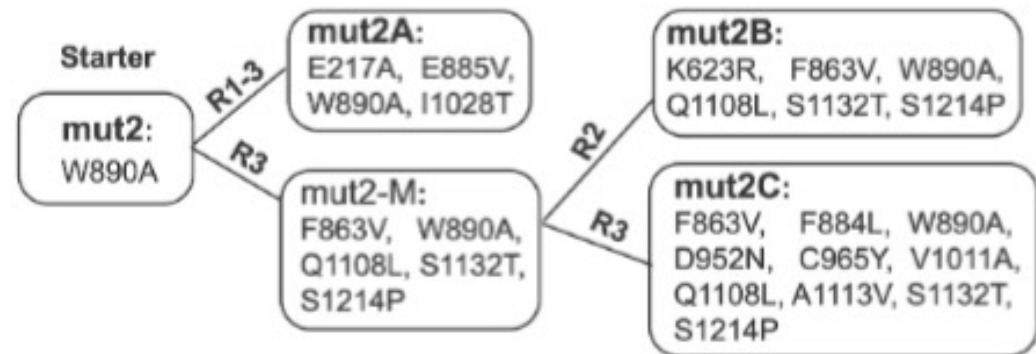


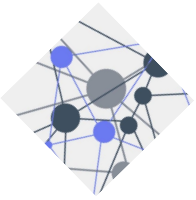




# Variant Structure Prediction With AlphaFold2

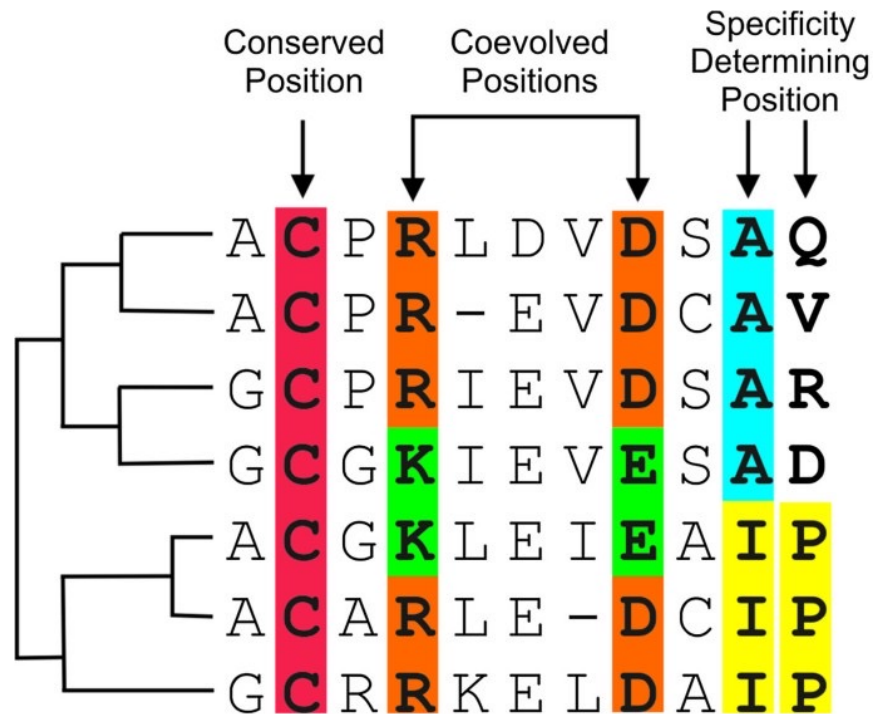
- Three variants were ultimately refined: mut2B-W, mut2C-W, and mut2C-WF
- To investigate these variants, their structure was predicted with **AlphaFold2**
- **AlphaFold2** is a protein structure prediction software that takes an input protein sequence and outputs a 3D structure that can be analyzed

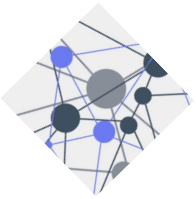




# How Can You Get Structure From A Sequence?

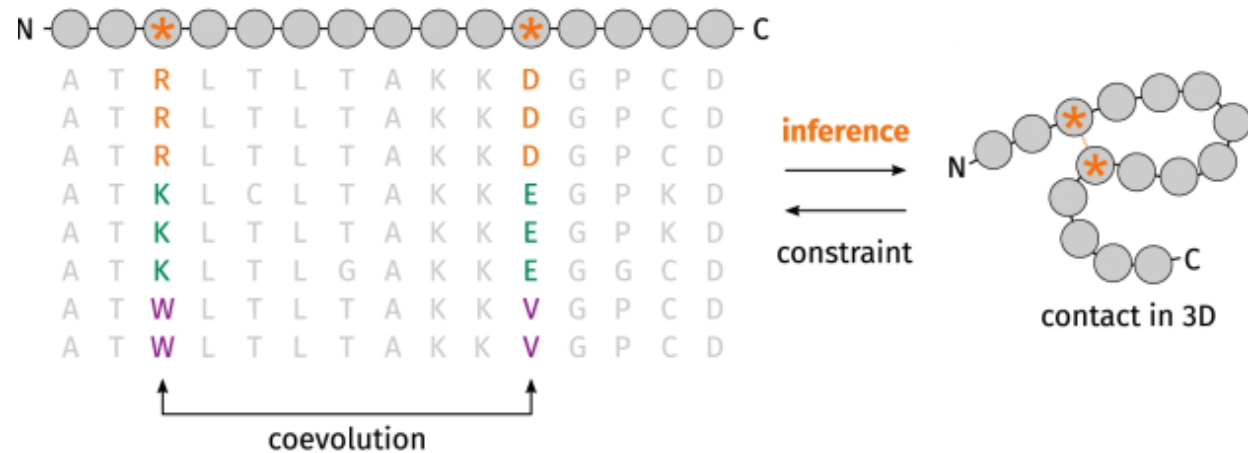
- Protein structural information can be gained by understanding multiple sequence alignments (MSA)
- When we align similar protein sequences we identify:
  - **Conserved positions:** where the letter does not change
  - **Coevolved positions:** where the letter will change with another letter
  - **Specificity determining positions:** where the letter is consistently different





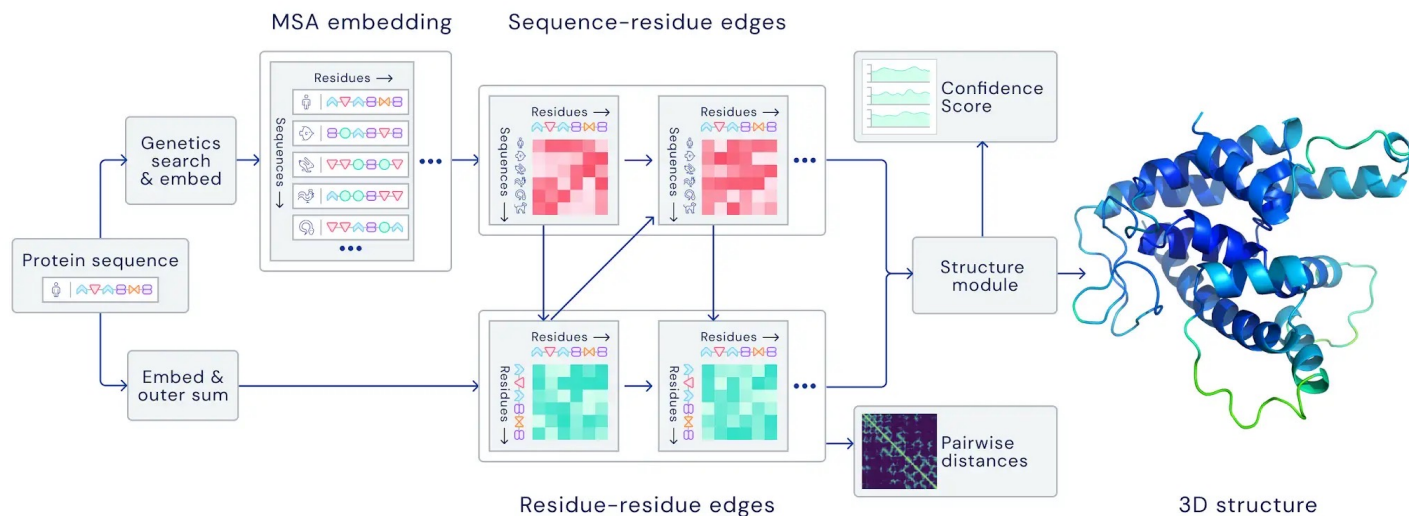
# Residue Coevolution

- With an MSA we can identify residues that coevolve, or change together
- We can then reason that residues that change together must be close together in 3D space



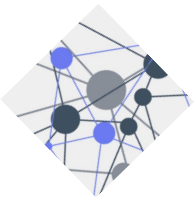


# AlphaFold2 Algorithm



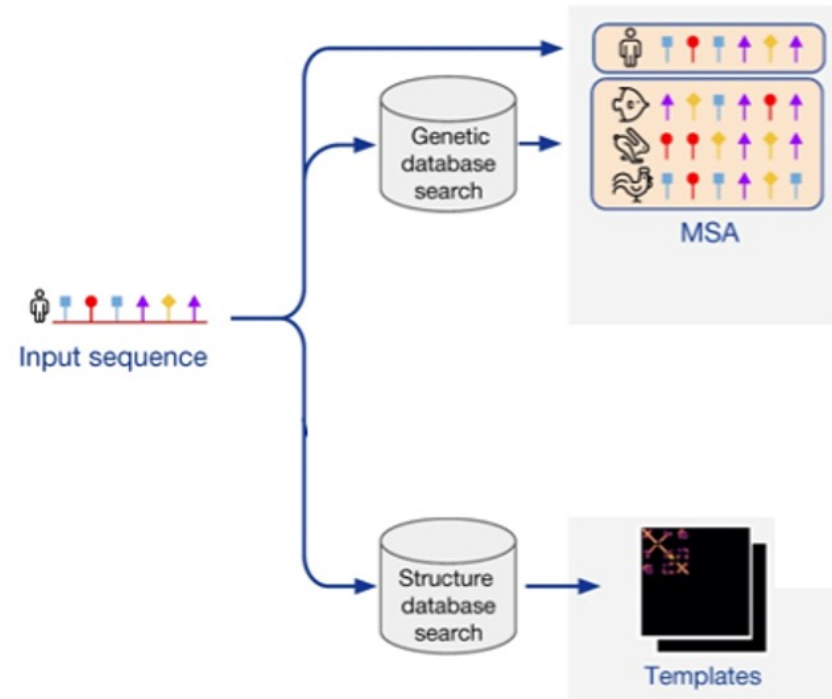
- starts with a user's query protein sequence (Fasta File)
- finding similar sequences to that query
- aligns these sequences to create an MSA
- uses available structure data based on query sequence to create initial distances between residues
- uses a neural network to iteratively update the distances between residues by using information from the sequence alignment
- passes this to another neural network to determine how these residues are oriented in 3D space (PDB File)

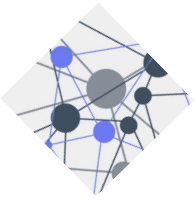




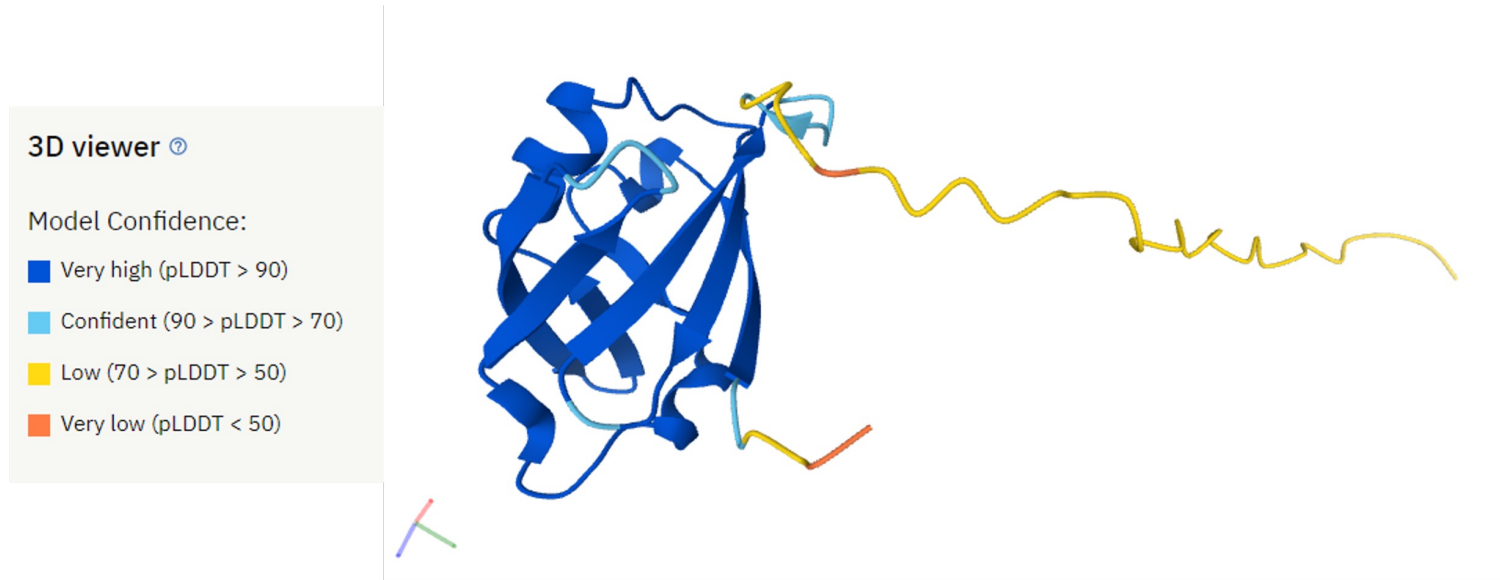
# AlphaFold2 Sequence/Structure Information

- Where does AlphaFold2 get Sequence/Structure Information?
- **Sequence Information:** Gathers this from UniRef, a collection of protein sequences from the [UniProt Knowledgebase](https://www.uniprot.org/help/uniref) - stored as FASTA Files
- **Structure Information:** Gathers this from the [Protein Data Bank \(PDB\)](https://www.rcsb.org/) which contains 3D coordinates for residues in a protein - stored as PDB Files



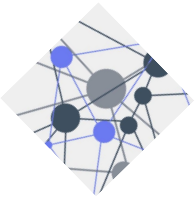


# AlphaFold2 Accuracy - Predicted Local Distance Difference Test



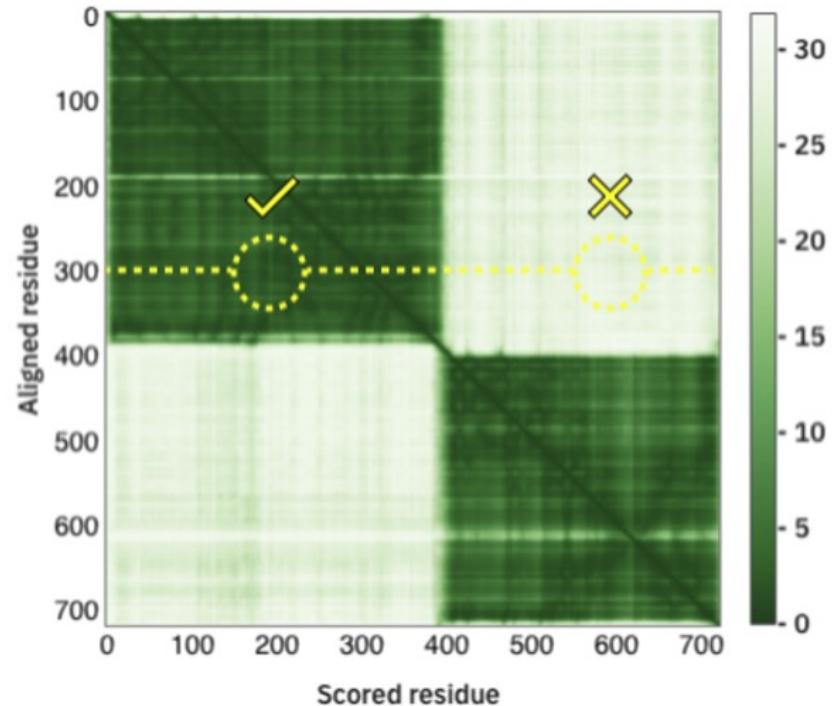
- The Predicted Local Distance Difference Test (pLDDT) is a per-residue confidence metric ranging from 0-100 (100 being the highest confidence)
- Regions below 50 could indicate disordered regions





# AlphaFold2 Accuracy - Predicted Alignment Error

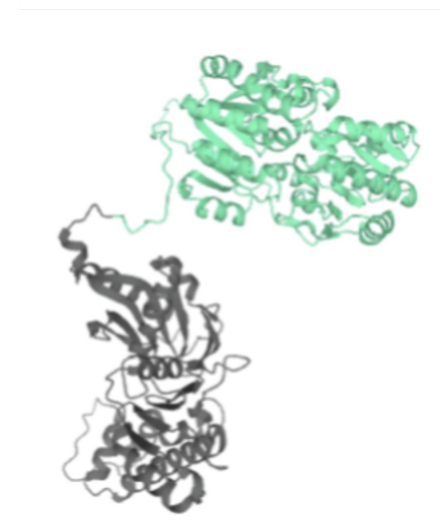
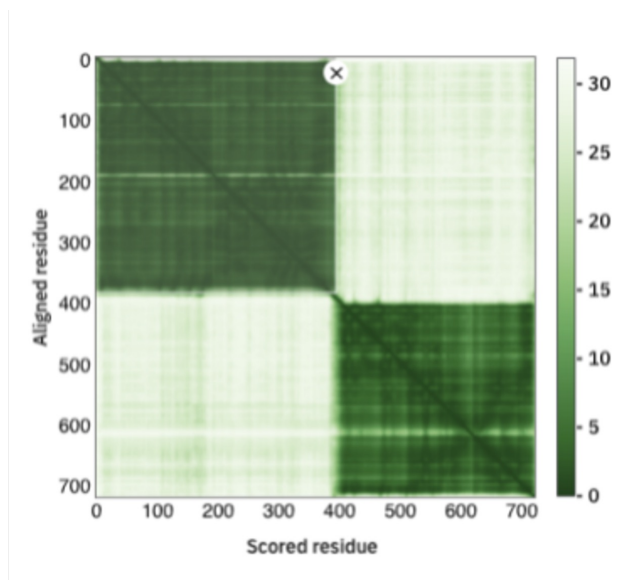
- The Predicted Alignment Error (PAE) gives us an expected distance error based on each residue.
- If we are more confident that the distance between two residues is accurate, then the PAE is lower (darker green). If we are less confident that the distance between two residues is accurate, the PAE is higher (lighter green)





# AlphaFold2 Accuracy – PAE Example

- In the Example to the right, we have a protein dimer
- Here we see two dark green patches which indicate two regions of low PAE
- These regions indicate that we are more confident in the distance between residues **within a subunit** and less confident when comparing residues **between subunits**

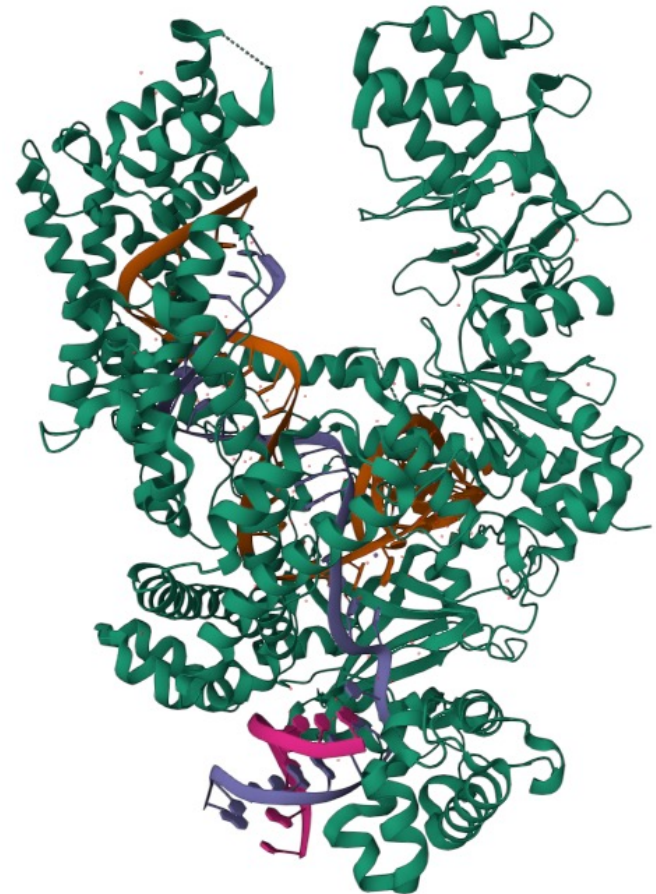






# AlphaFold2 Limitations

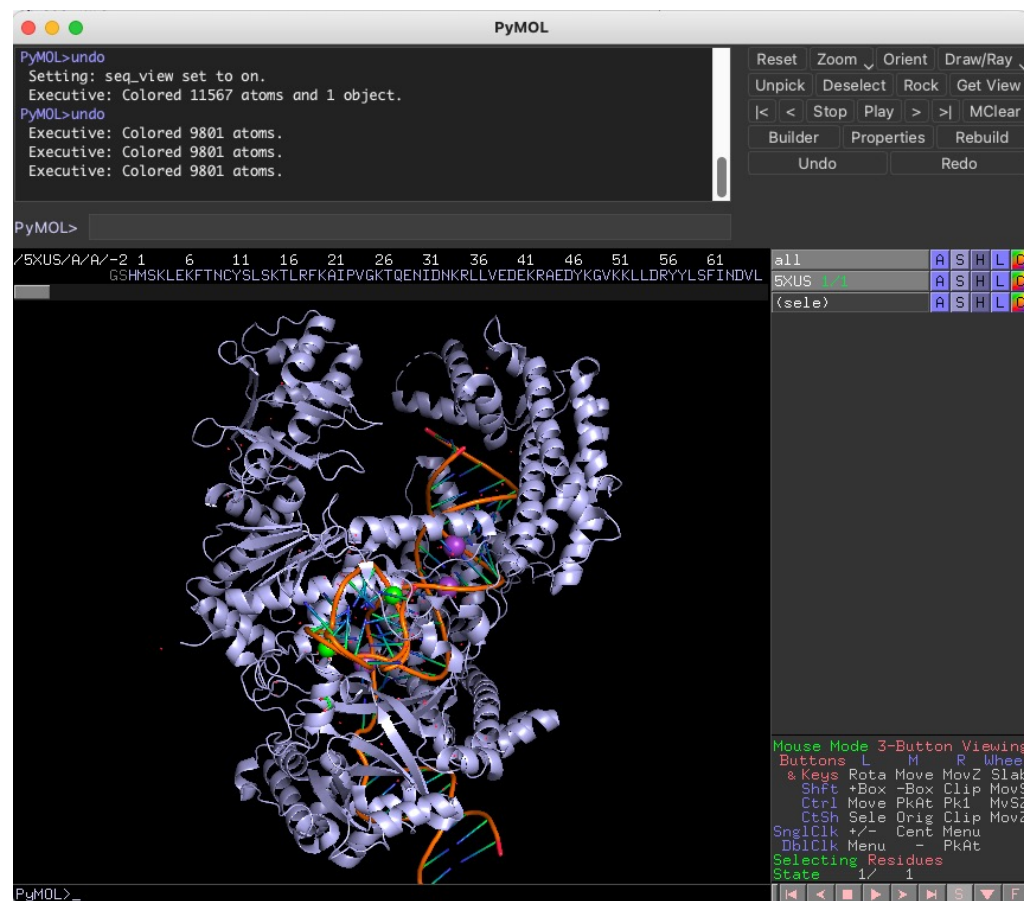
- While AlphaFold2 has been an amazing leap forward in structural biology it isn't perfect
- PDB structures are usually created from experiments where the context of that structure is specific to the study question. (bound to ions, chemically modified, etc.)
- Protein interactions/multimers might not be captured in the [PDB](#) database. Given this, AlphaFold2's multimeric prediction might not be reflective of the true interaction structure.
- Proteins can also contain disordered regions (i.e. loops), which are difficult to crystallize and as such AlphaFold's prediction of these disordered regions is bound to be poor.

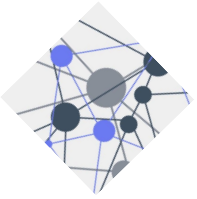




# Protein Structure Visualization

- We can visualize the output of AlphaFold2 (a PDB File) with a visualization software called PyMOL
- Here we can visually investigate residues on a protein

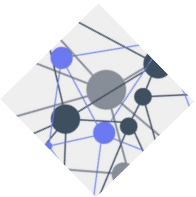




# Lab Exercise!

- [Click here to go to the webpage for the web exercise!](#)





# References

1. <https://www.sciencedirect.com/science/article/pii/S2319417019305050>
2. <https://www.yourgenome.org/facts/what-is-crispr-cas9/>
3. <https://www.nature.com/articles/emm2016111>
4. <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC9825149/>
5. <https://bmcbioinformatics.biomedcentral.com/articles/10.1186/1471-2105-13-235>
6. <https://www.blopiq.com/blog/2021/07/alphafold-2-is-here-whats-behind-the-structure-prediction-miracle/>
7. <https://www.deepmind.com/blog/alphafold-a-solution-to-a-50-year-old-grand-challenge-in-biology>
8. <https://www.nature.com/articles/s41586-021-03819-2>
9. <https://www.uniprot.org/help/uniref>
10. <https://www.rcsb.org/>
11. <https://alphafold.ebi.ac.uk/faq>
12. <https://alphafold.com/entry/Q9FX77>
13. <https://www.rcsb.org/3d-view/5XUS/1>
14. <https://pymol.org/2/>

