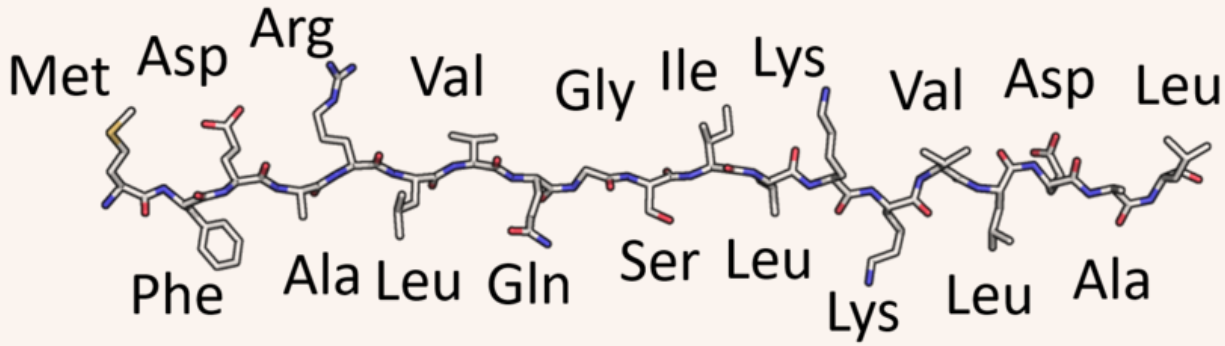
USING WRITHE TO PRODUCE REALISTIC PROTEIN STRUCTURE PREDICTIONS FROM BIOSAXS DATA

Arron Bale, Supervised by Chris Prior

START EASY: WHAT IS A PROTEIN?

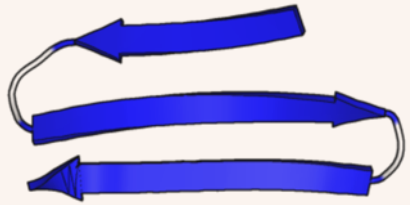
PRIMARY STRUCTURE



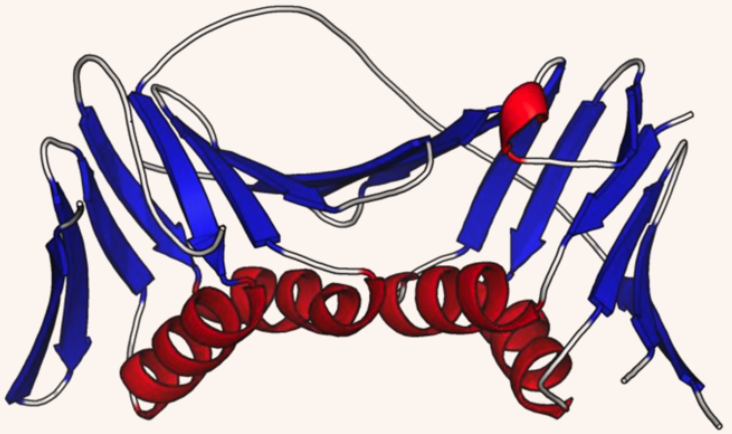
ALPHA HELICES

SECONDARY STRUCTURE

BETA-SHEETS



TERTIARY STRUCTURE



SAXS Data Collection

diamond

Initial Predictions

carbonara

Molecular Dynamics

SAXS Data Collection

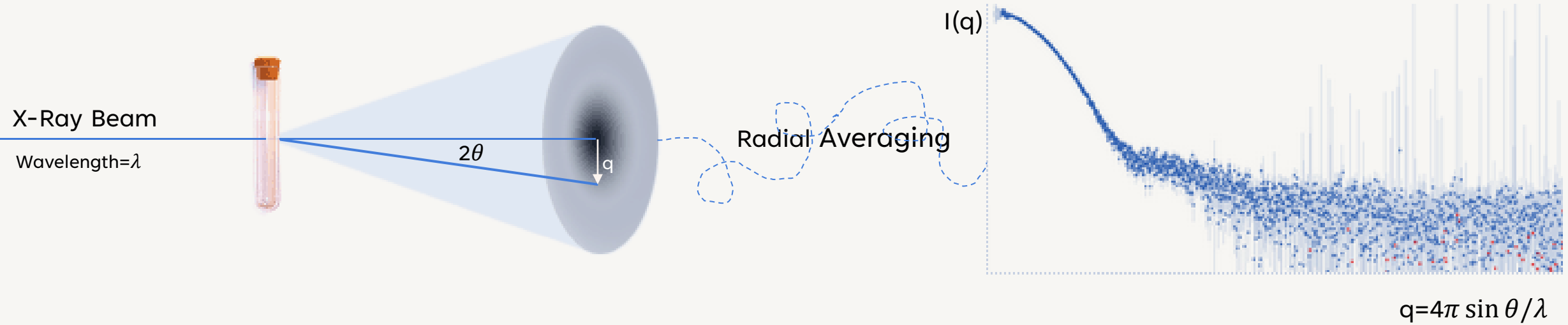
diamond

Initial Predictions

carbonara

Molecular Dynamics

SMALL ANGLE X-RAY SCATTERING



$q=4\pi \sin\theta/\lambda$

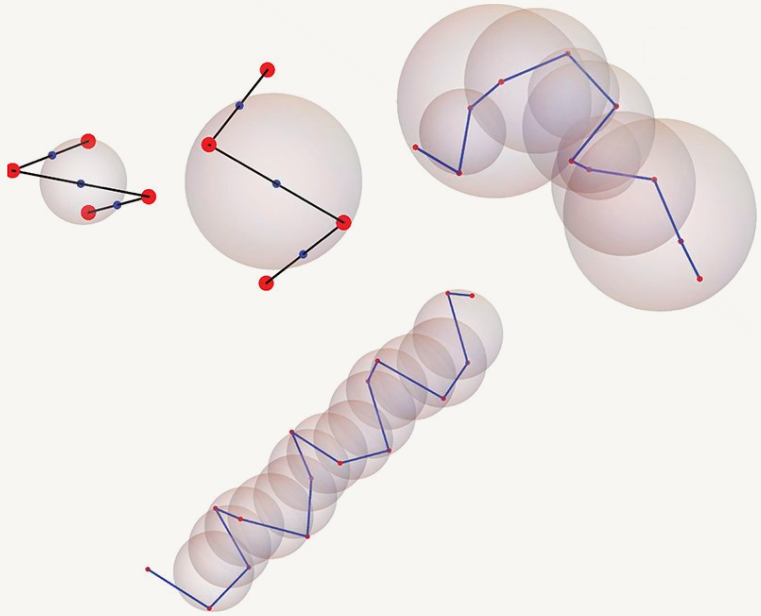
SAXS Data Collection

Initial Predictions

carbonara

Molecular Dynamics

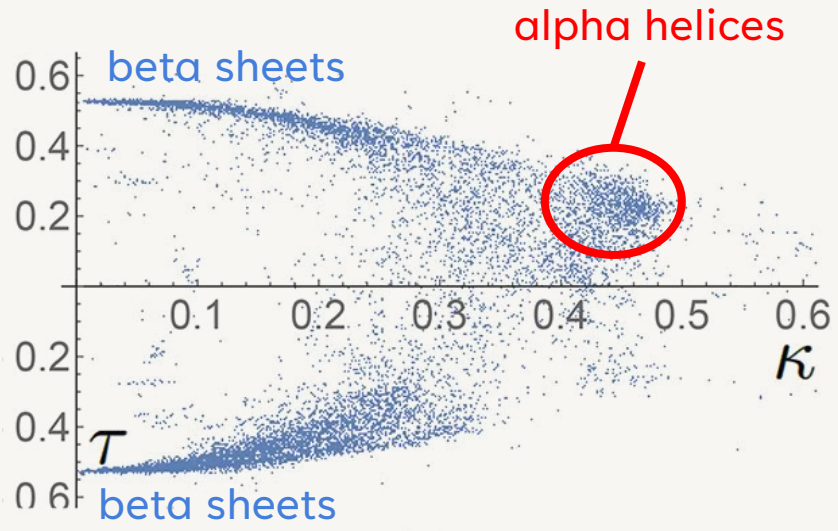
Tell me the secondary structure,
I'll build a backbone curve



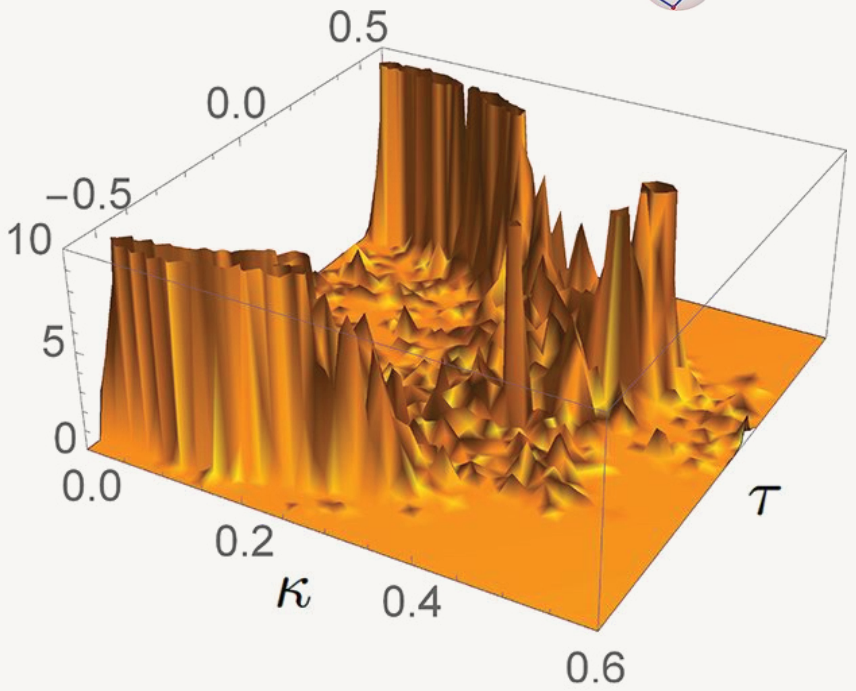
We can produce **locally** realistic protein curves using just two parameters:

Curvature (κ) ~ how tightly wound is my curve.

Torsion (τ) ~ how helical is my curve.



(a)



(b)

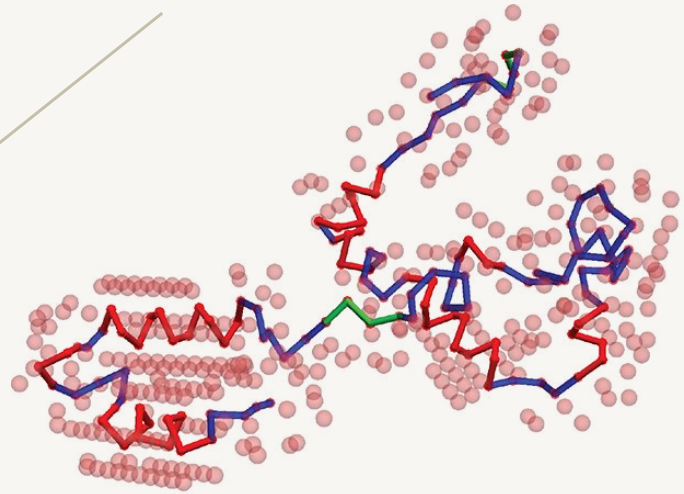
SAXS Data Collection

diamond

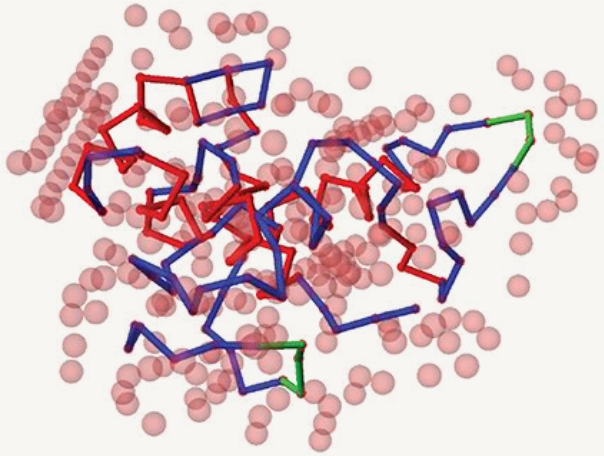
Initial Predictions

carbonara

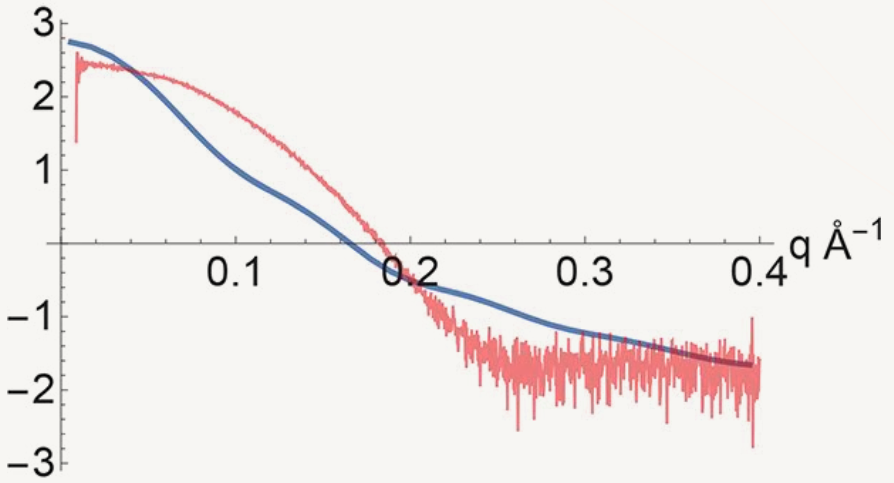
Molecular Dynamics



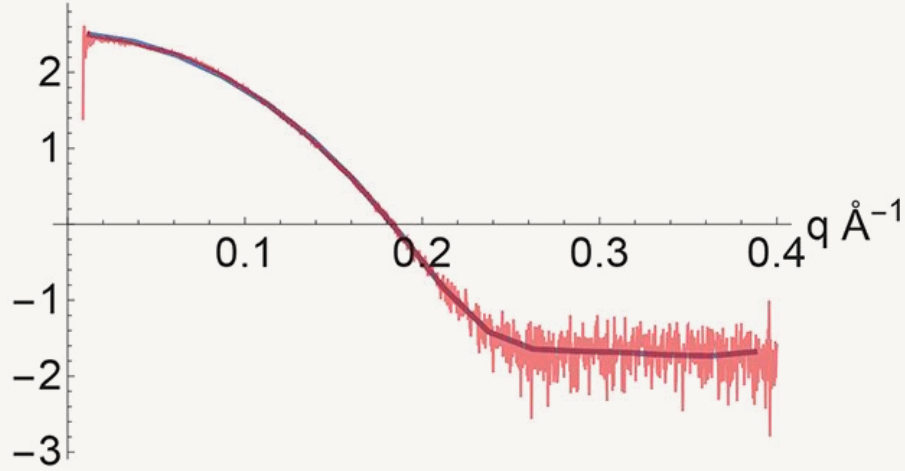
A starting configuration. Alpha helices and beta strands are shown in red and green respectively.



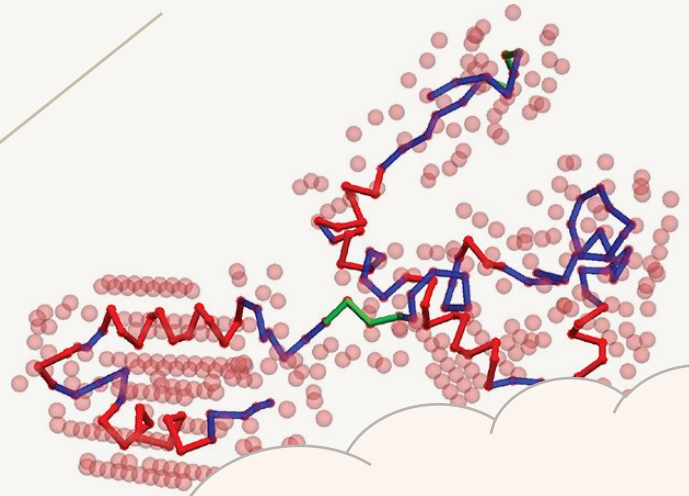
A good fit to the scattering data.



In red: The experimental scattering data. In blue: The scattering profile of the starting configuration (not good...)

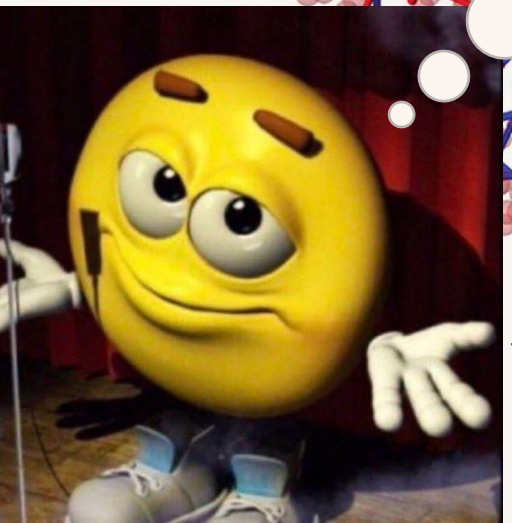


In red: The experimental scattering data. In blue: The scattering profile of the final configuration (good!)

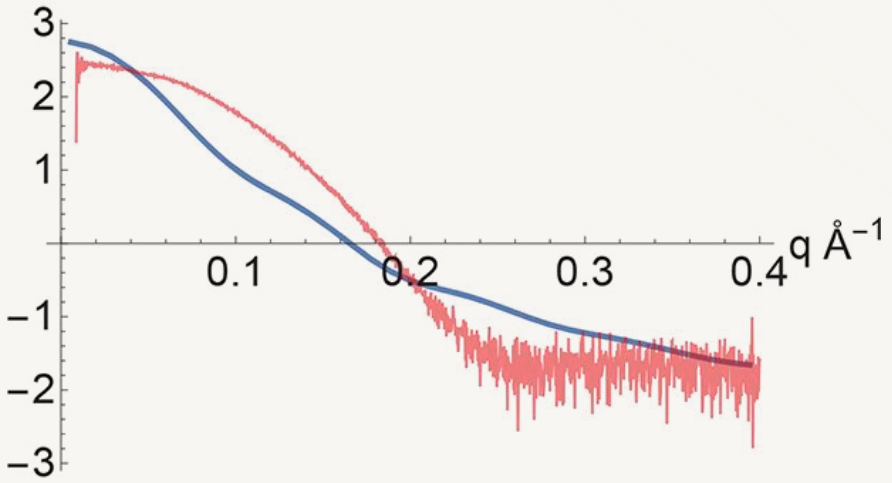


A starting
beta str
respec

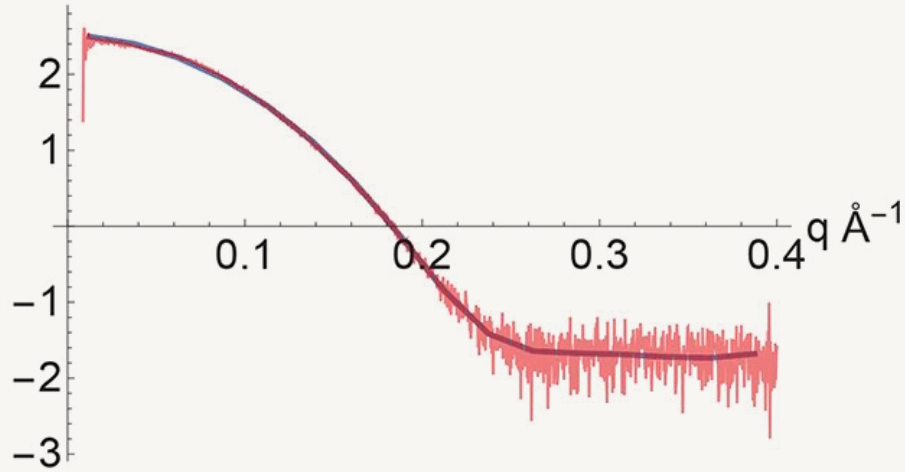
But... AlphaFold's solved this?



the scattering data.

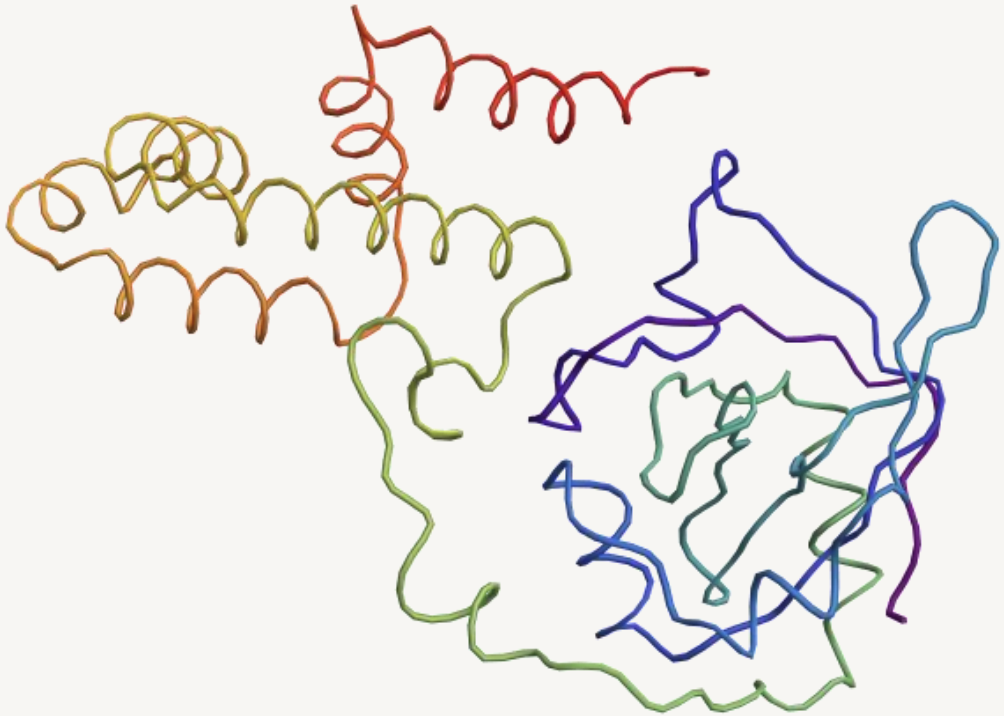


In red: The experimental scattering data. In blue: The scattering profile of the starting configuration (not good...)

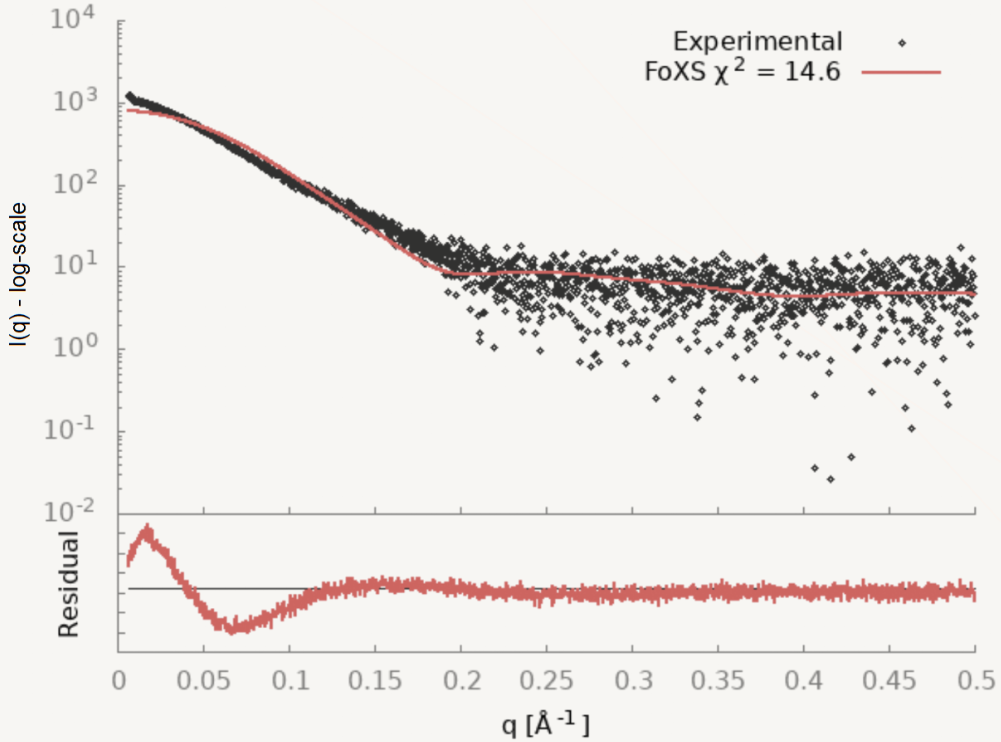


In red: The experimental scattering data. In blue: The scattering profile of the final configuration (good!)

NOT THIS TIME ALPHAFOLD!

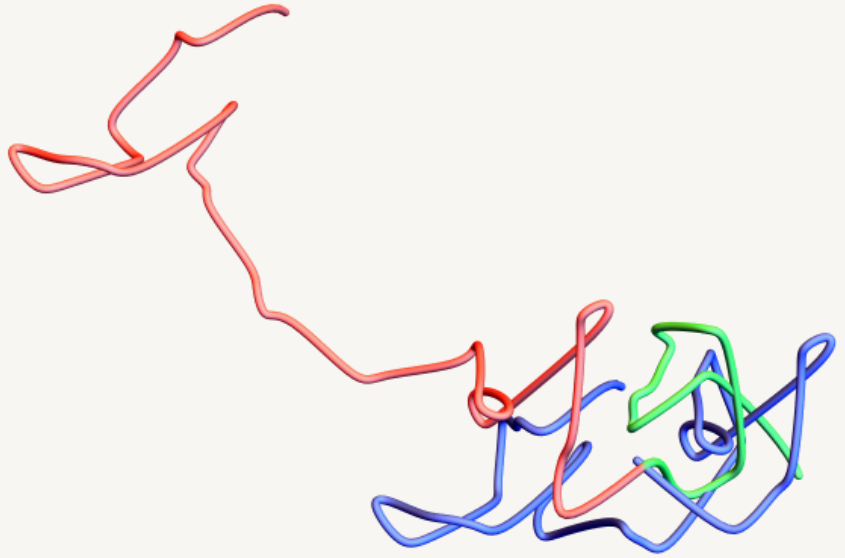


The crystallographic (=AlphaFold[1] predicted) structure.

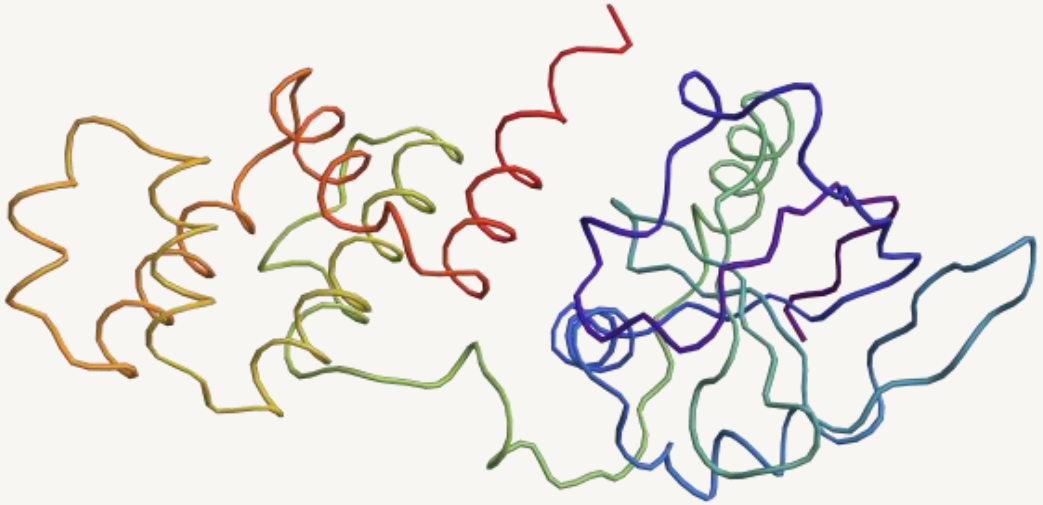


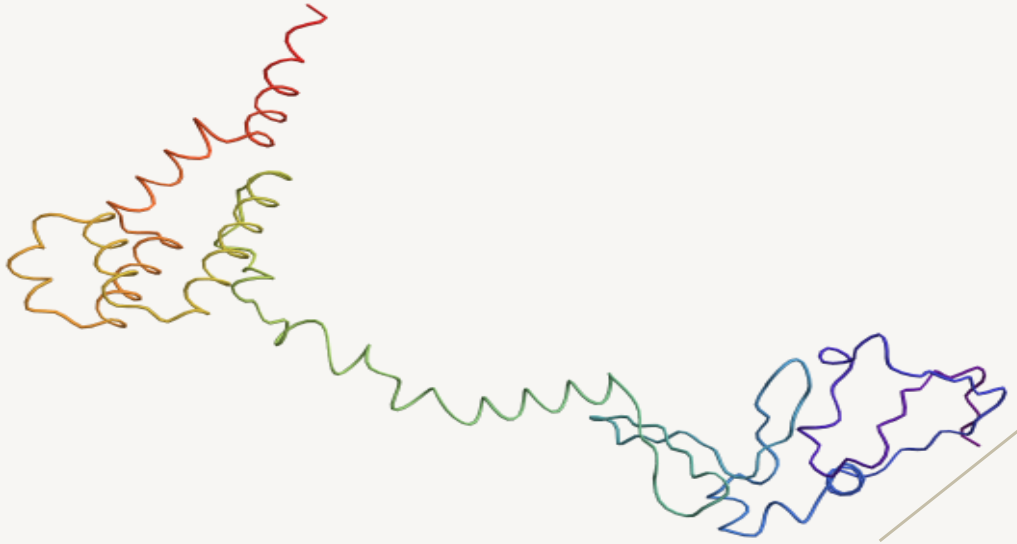
In red, the FoXS[2] scattering profile for the crystallographic structure. This is a poor fit, especially at low q, indicating the global structure is not right.

[1] Jumper, J et al. Highly accurate protein structure prediction with AlphaFold. Nature (2021).
[2] Schneidman-Duhovny D, Hammel M, Tainer JA, and Sali A. Accurate SAXS profile computation and its assessment by contrast variation experiments. Biophysical Journal 2013. 105 (4), 962-974

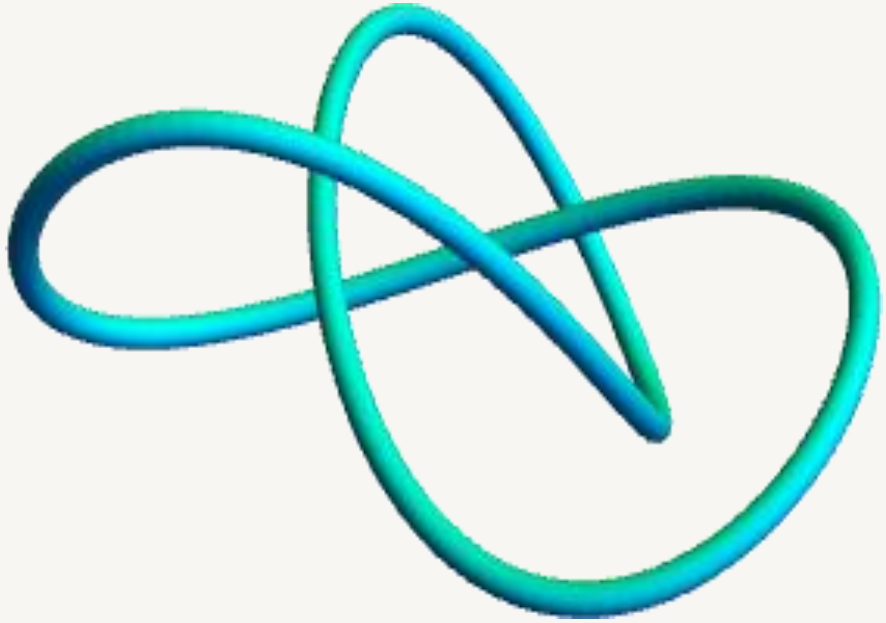


Two good fits to the scattering data.
Where they agree is shown in green. The conformations on the left and right are shown in blue and red respectively

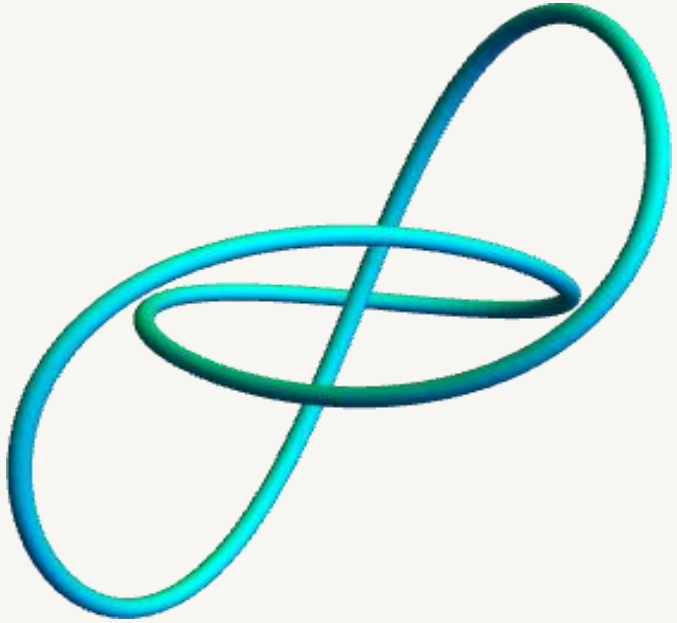




Average Crossing Number (ACN)



From this angle, we see 4 crossings on the trefoil knot.

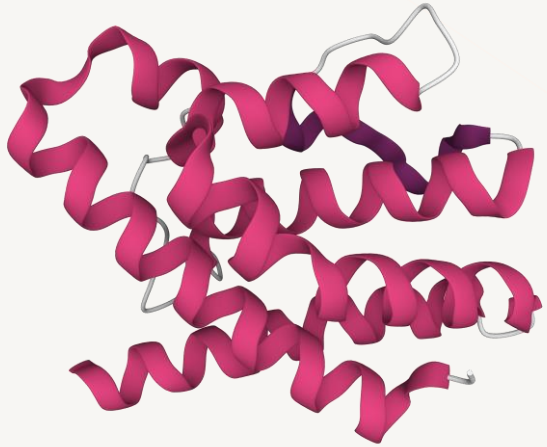


From this angle, we see 3 crossings on the trefoil knot.

SMOOTHING

- Given the secondary structure fingerprint of a protein.
- Take the start point (cα molecule) of each SSE.
- Construct the discrete curve given by this subset of points (and the end points of the protein).

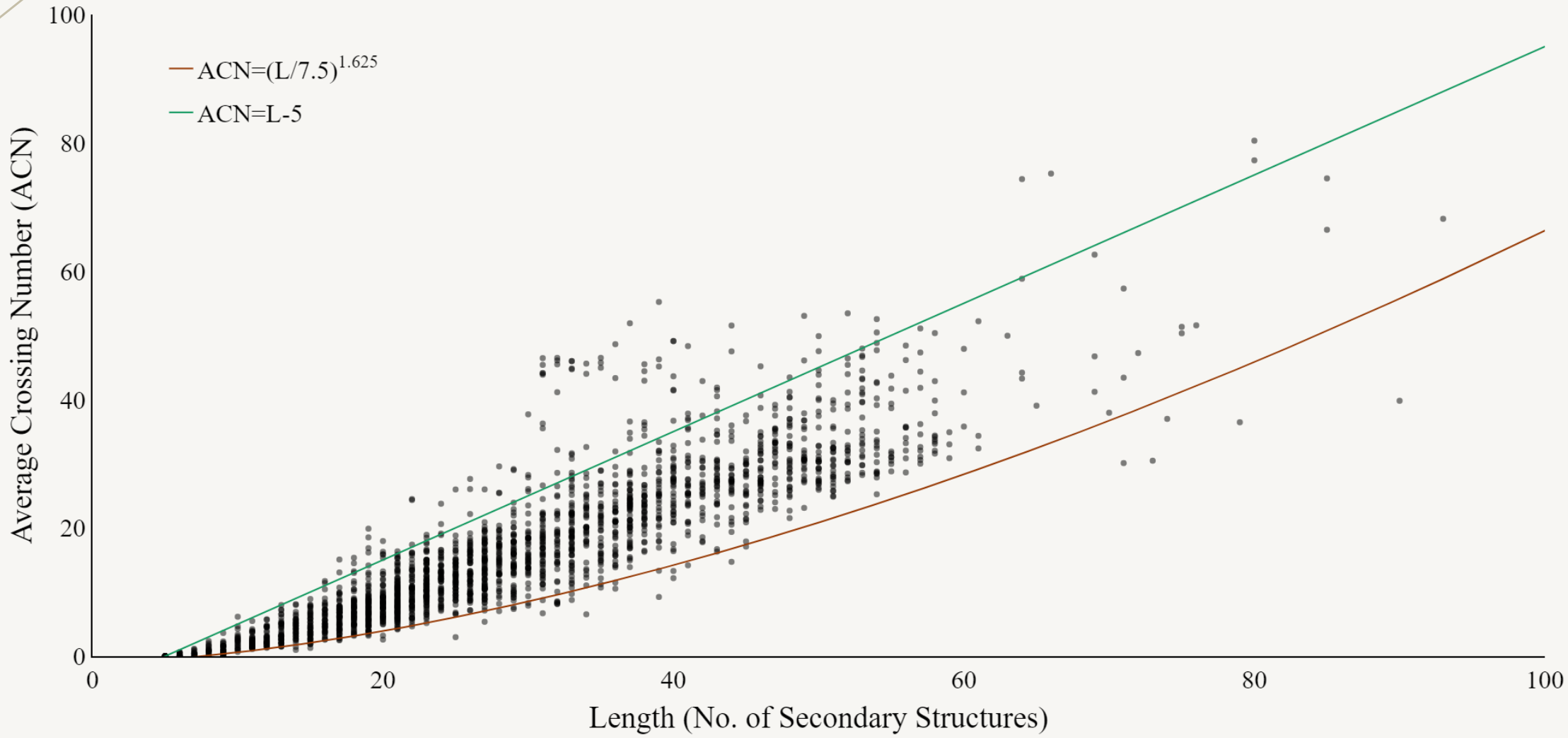
By smoothing in this way, we get a much clearer picture of the global entanglement of the protein, and work on a resolution closer to that of most SAXS data.



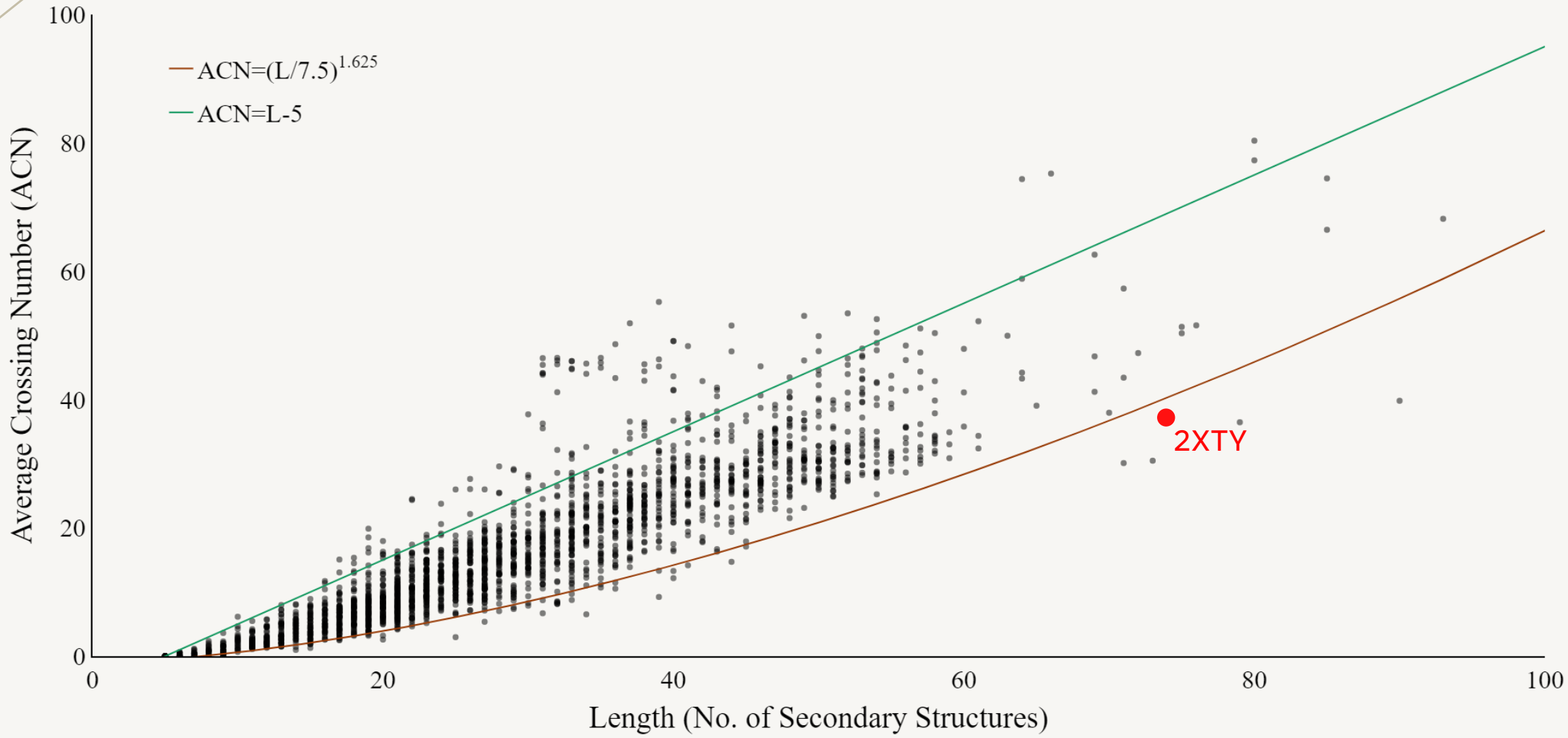
Cartoon representation of PDB: 2OEB. In pink are the alpha-helices and in white the flexible linkers.

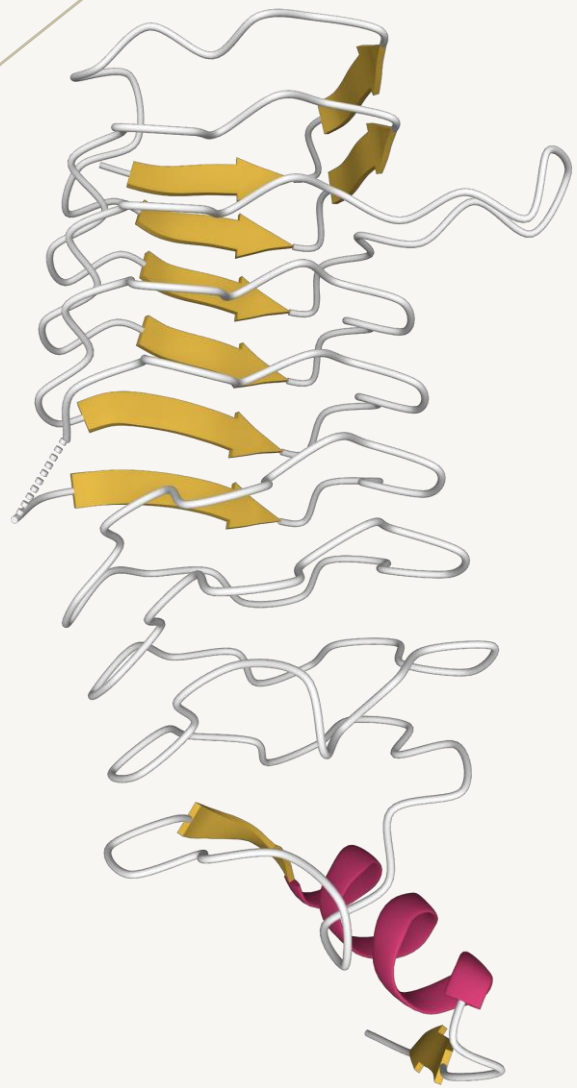
Our smoothed representation of 2OEB. Note the lack of real global entanglement.

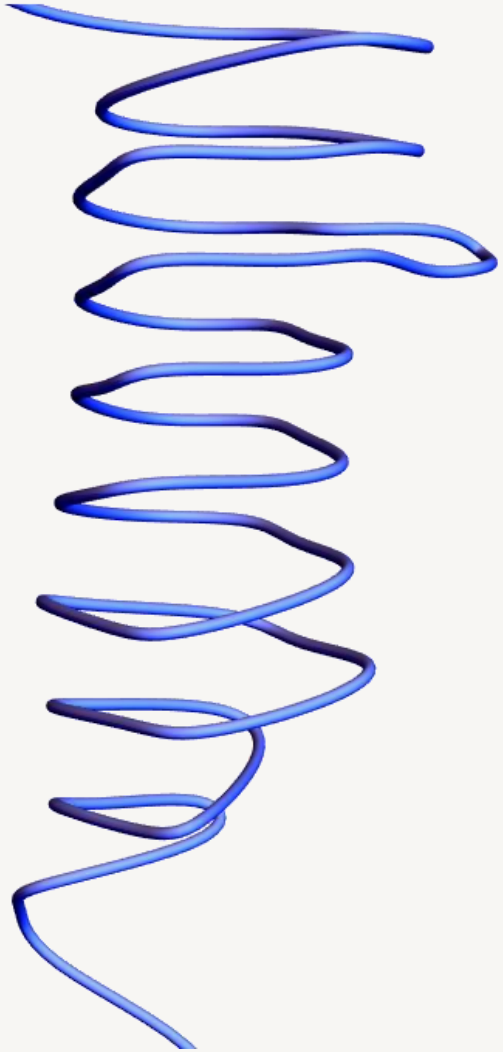
An empirical bound on the average crossing number for a representative sample of >2000 proteins from the PDB



An empirical bound on the average crossing number for a representative sample of >2000 proteins from the PDB





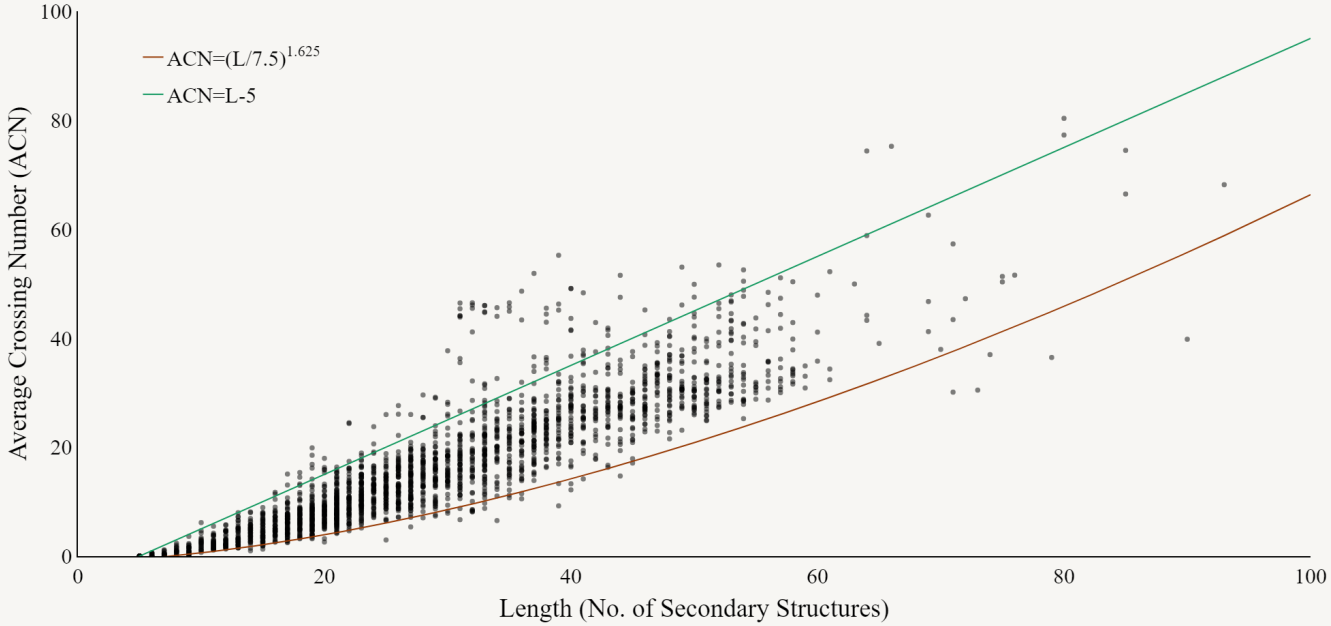


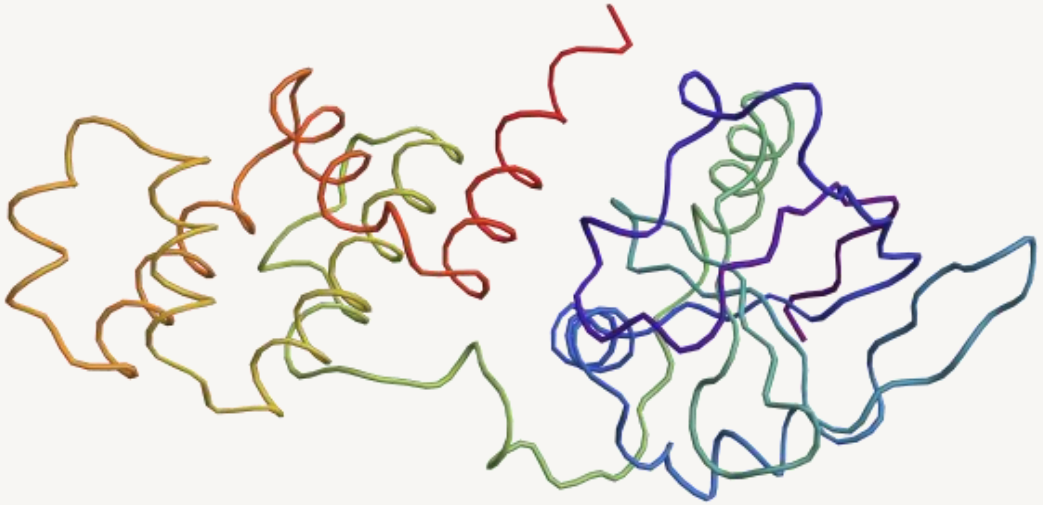
2XTY

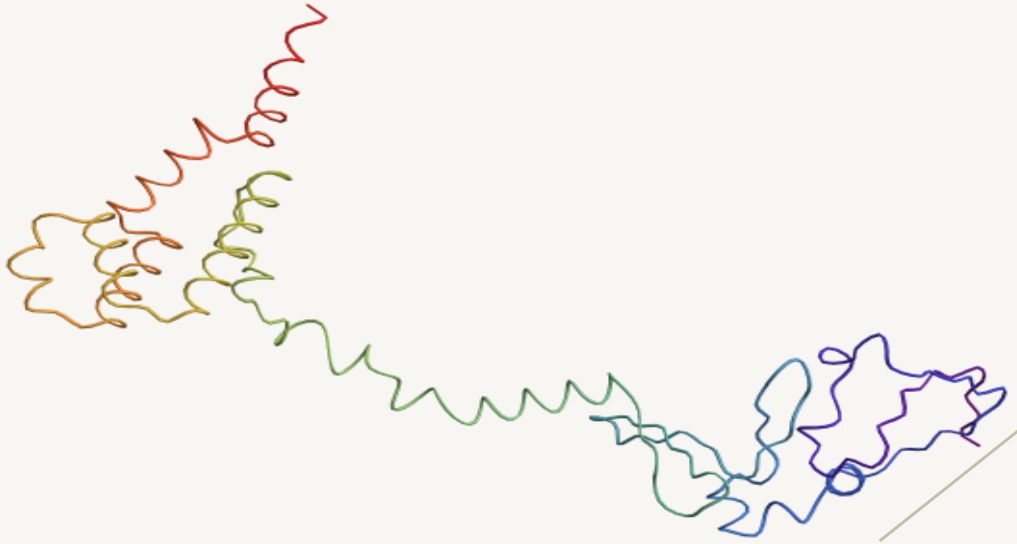
Vetting, M.W., Hegde, S.S., Wang, M., Jacoby, G.A., Hooper, D.C., Blanchard, J.S.
(2011) J Biol Chem 286: 25265

- Secondary structure assignment not always reliable
- Here, DSSP finds many single amino acid long β-strands, opposite those seen in yellow in the left figure.
- With the correct length, this falls well inside the lower bound

- With initial predictions from e.g. AlphaFold, not an issue, the secondary structure assignment is clean.

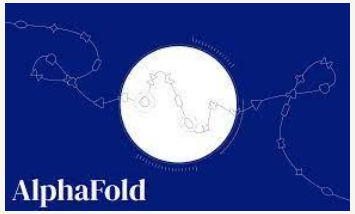

ACN=(L/7.5)^1.625
ACN=L-5
Average Crossing Number (ACN)
Length (No. of Secondary Structures)
0
20
40
60
80
100





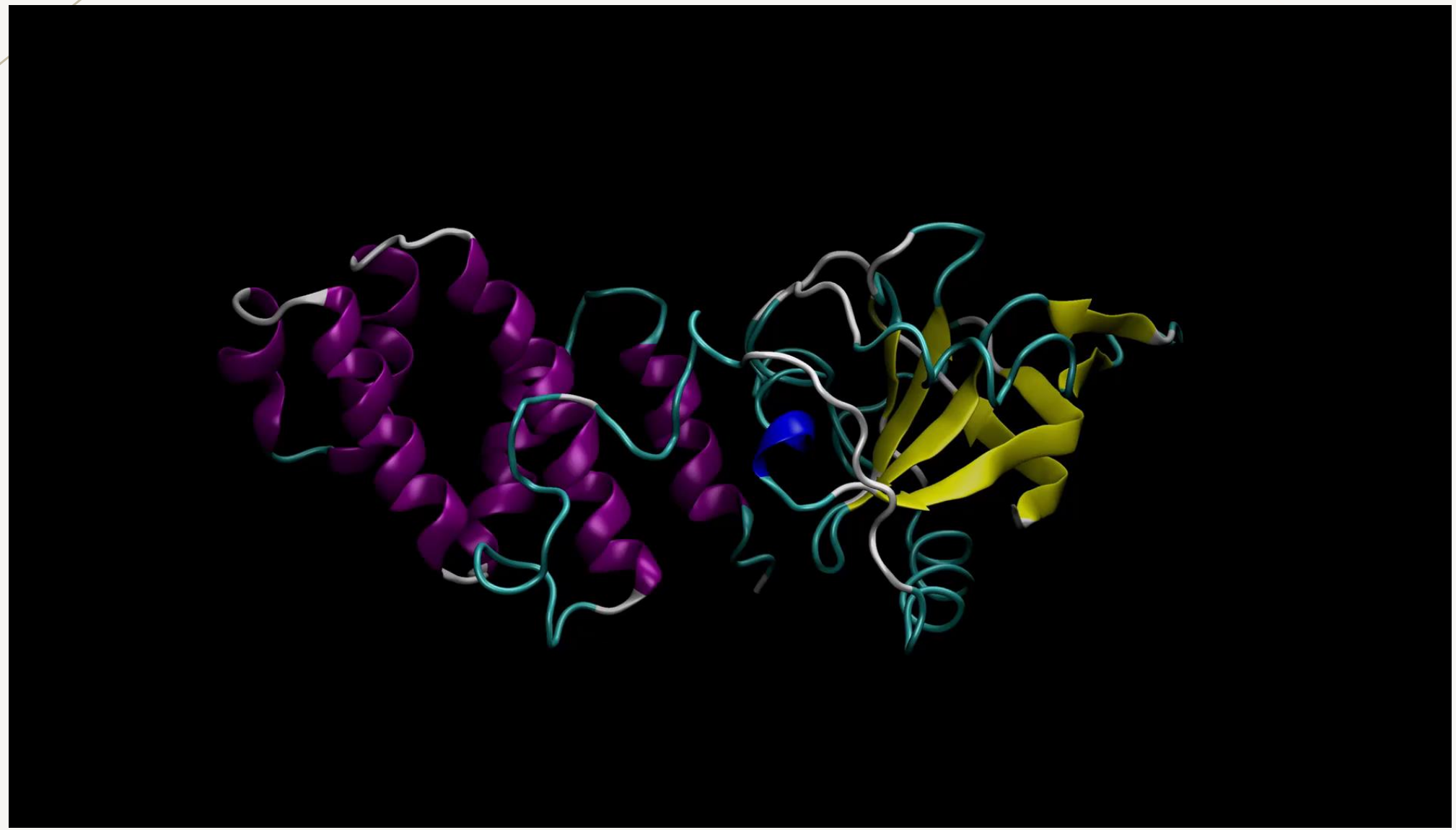
SAXS Data Collection

Initial Predictions



carbonara

Molecular Dynamics



MD simulation videos provided by Josh McKeown

CONCLUSIONS

carbonara is able to produce novel structures which fit solution scattering data.

The Average Crossing Number (ACN) of proteins is bounded from above and below.

The ACN ensures we make realistic predictions, which behave well in molecular dynamics simulations.

THANK YOU FOR LISTENING

Any questions?