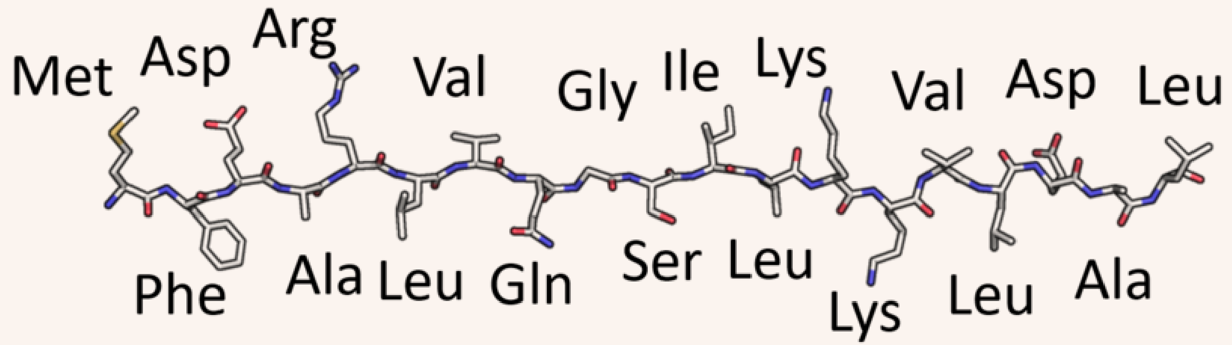
USING WRITHE TO PRODUCE REALISTIC PROTEIN STRUCTURE PREDICTIONS FROM BIOSAXS DATA

Arron Bale, Supervised by Chris Prior

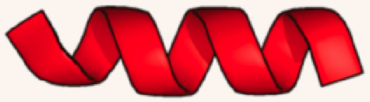
WHAT IS A PROTEIN?

PRIMARY STRUCTURE

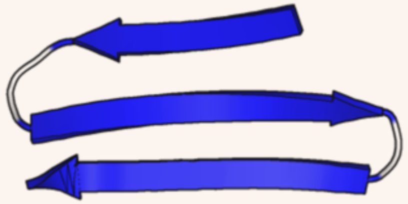


SECONDARY STRUCTURE

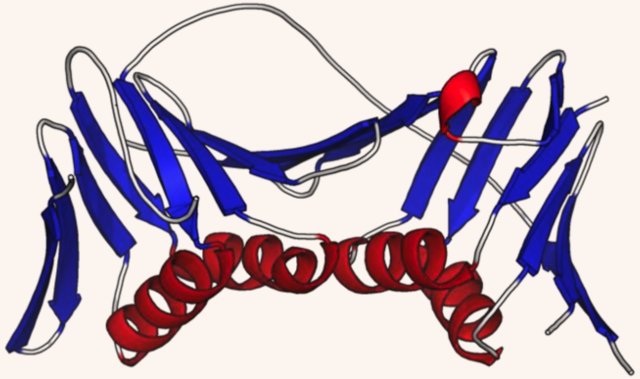
ALPHA HELICES



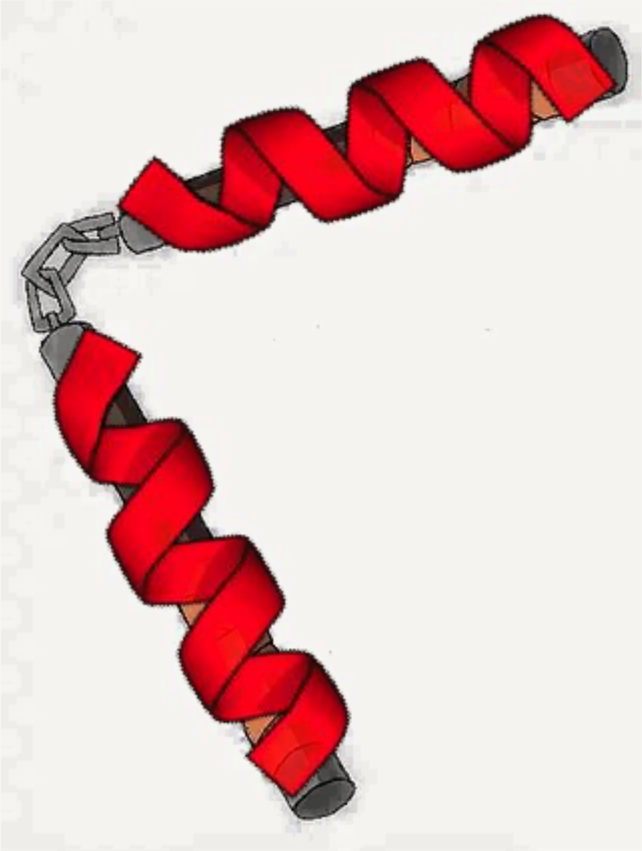
BETA-SHEETS



TERTIARY STRUCTURE



HOW I LIKE TO VISUALIZE THINGS

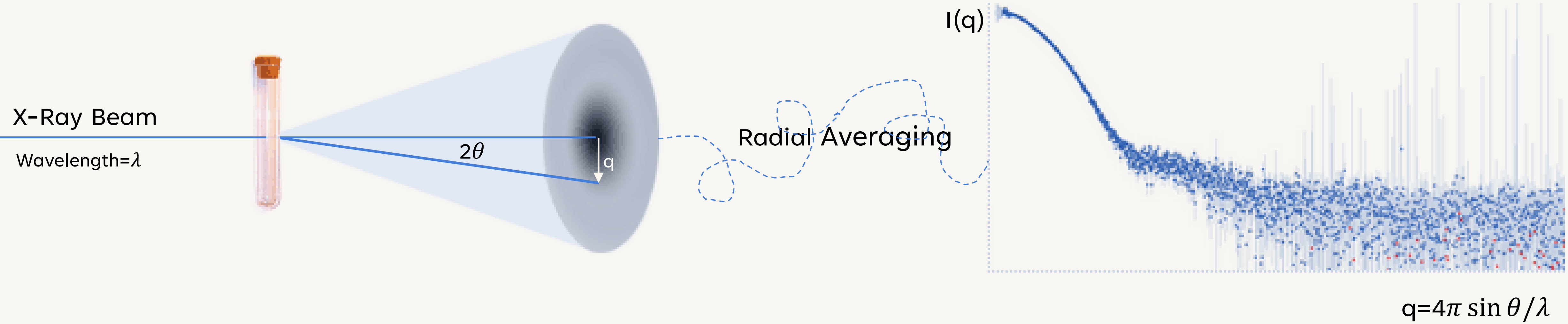


linker sections are key in understanding tertiary structure, as they take a much wider range of geometries.

When making predictions, we can in some sense ignore secondary structure (more on this later).

Vague open question: Inverse stick number? i.e. given n sticks, how many “different” configurations could I make?

SMALL ANGLE X-RAY SCATTERING



X-Ray Beam

Wavelength=λ

2θ

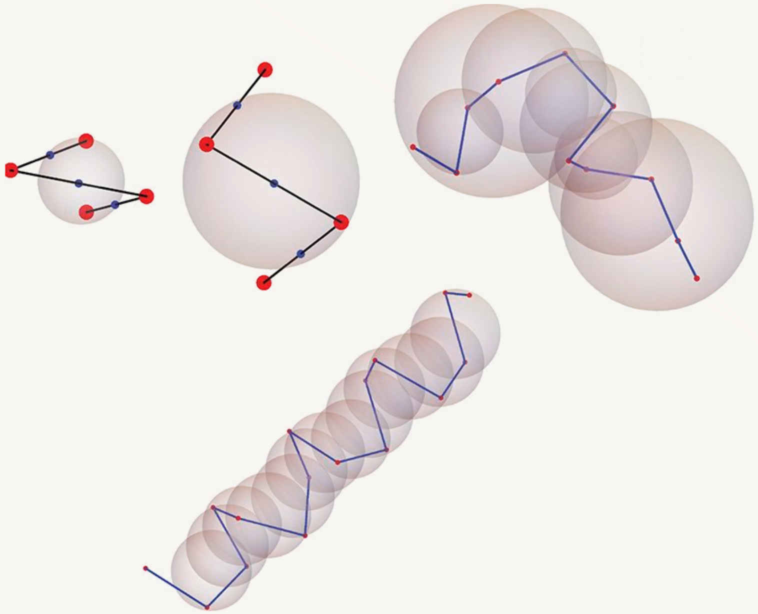
q

Radial Averaging

I(q)

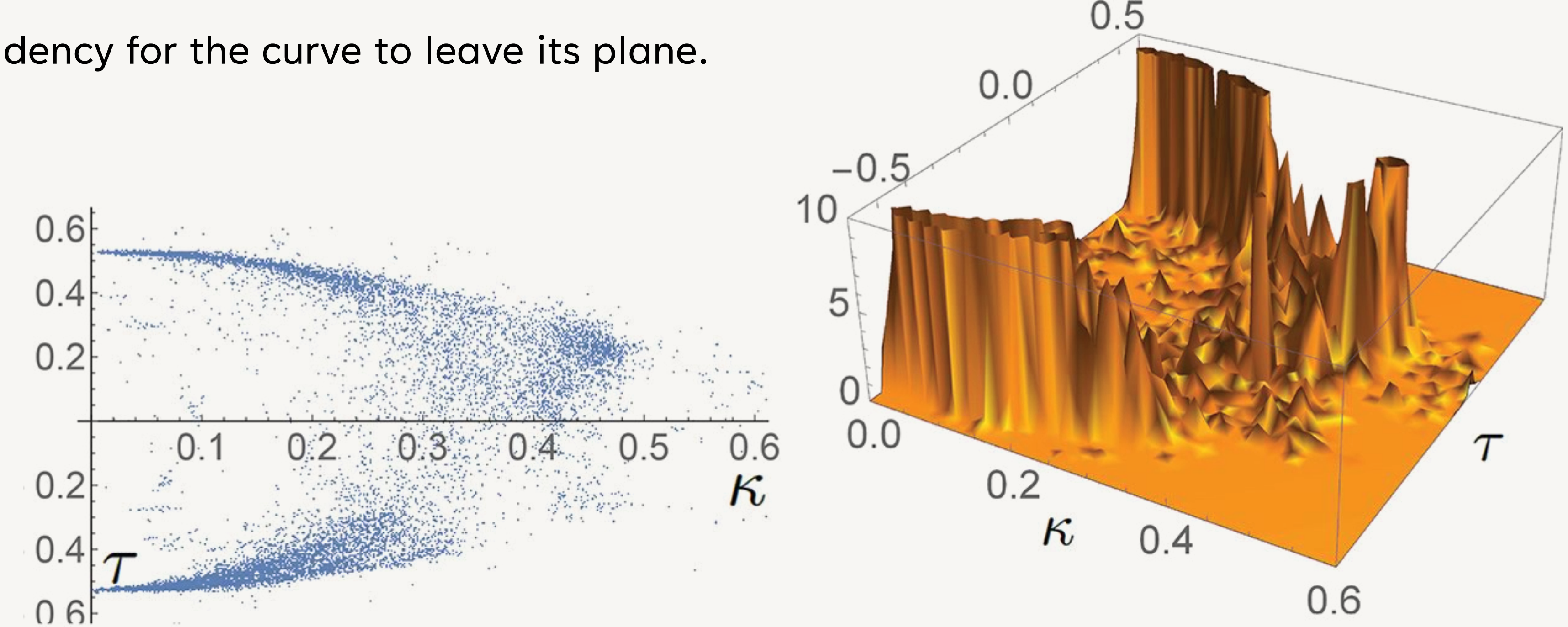
q=$4\pi \sin\theta/\lambda$

We can produce **locally** realistic protein curves using just two parameters, curvature and torsion.



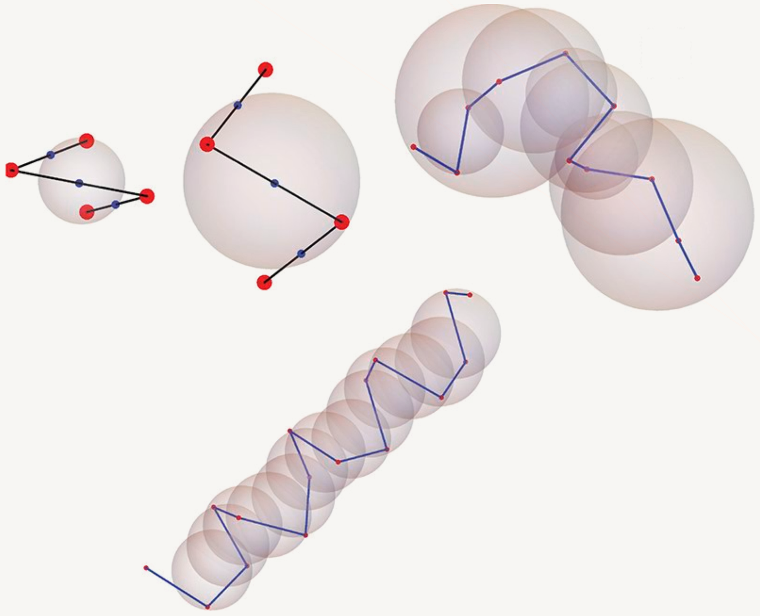
Curvature: inverse of the radius of the sphere defined by midpoints of three edges.

Torsion: tendency for the curve to leave its plane.



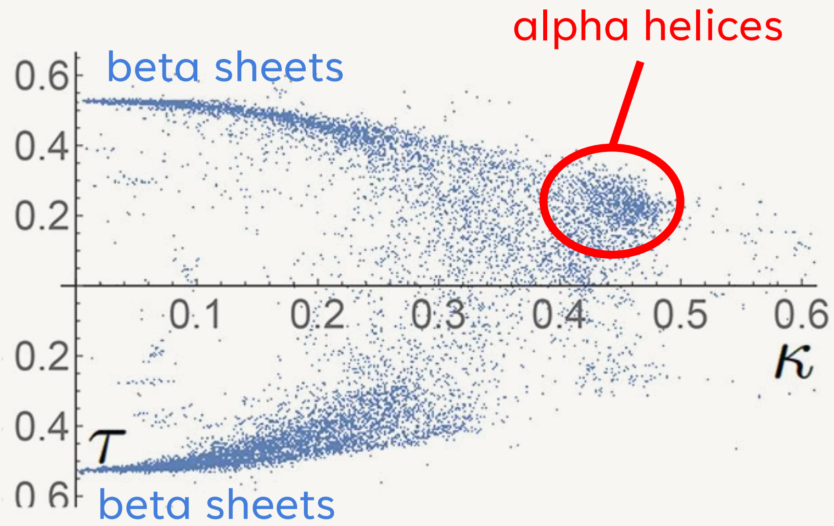
(a) (b)

We can produce **locally** realistic protein curves using just two parameters, curvature and torsion.

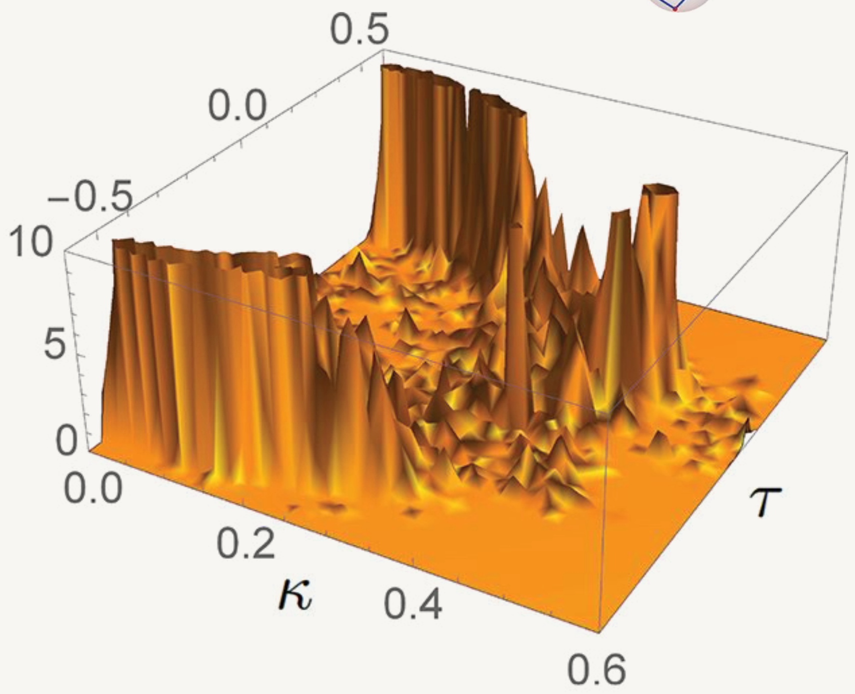


Curvature: inverse of the radius of the sphere defined by midpoints of three edges.

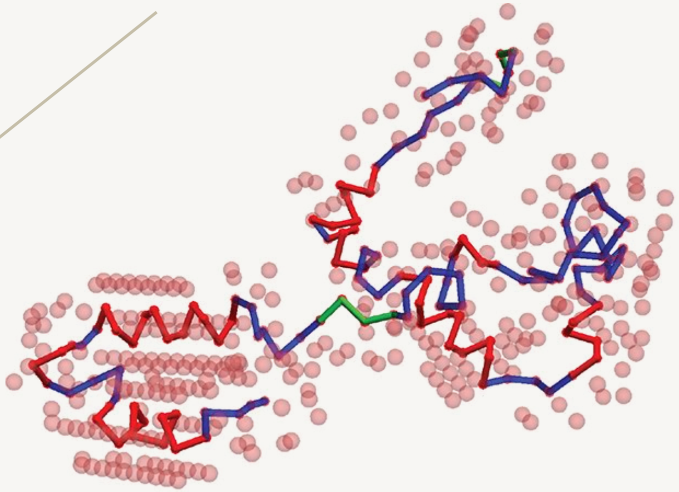
Torsion: tendency for the curve to leave its plane.



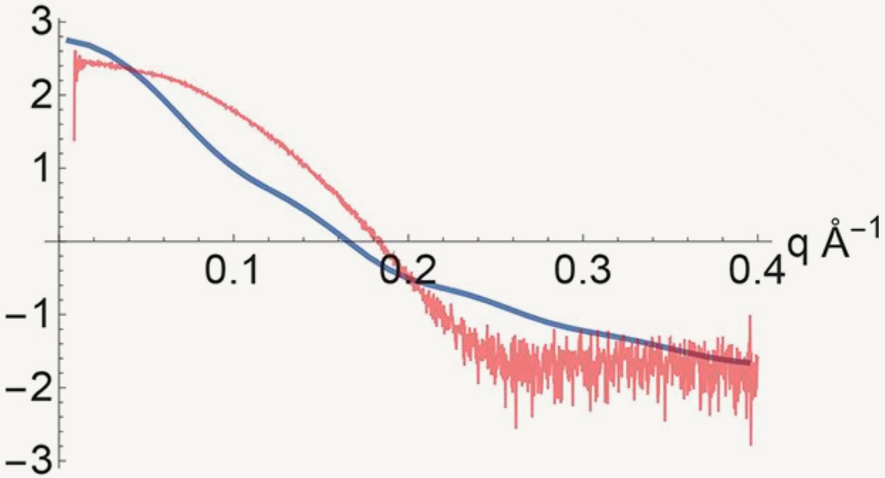
(a)



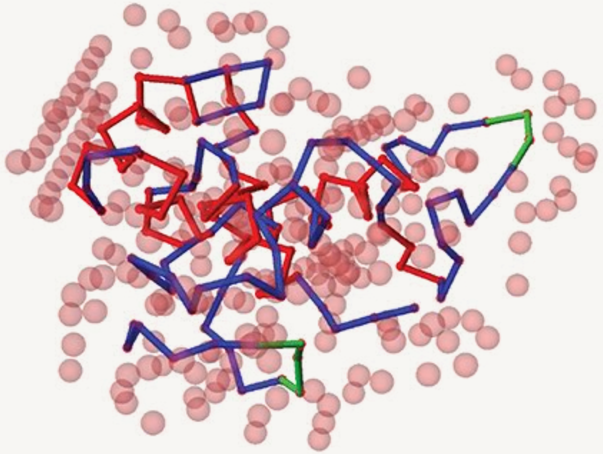
(b)



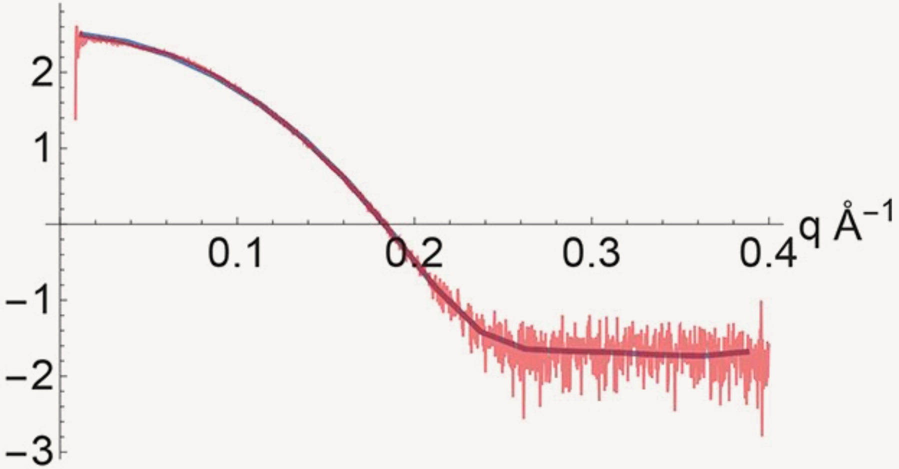
A starting configuration. Alpha helices and beta strands are shown in red and green respectively.



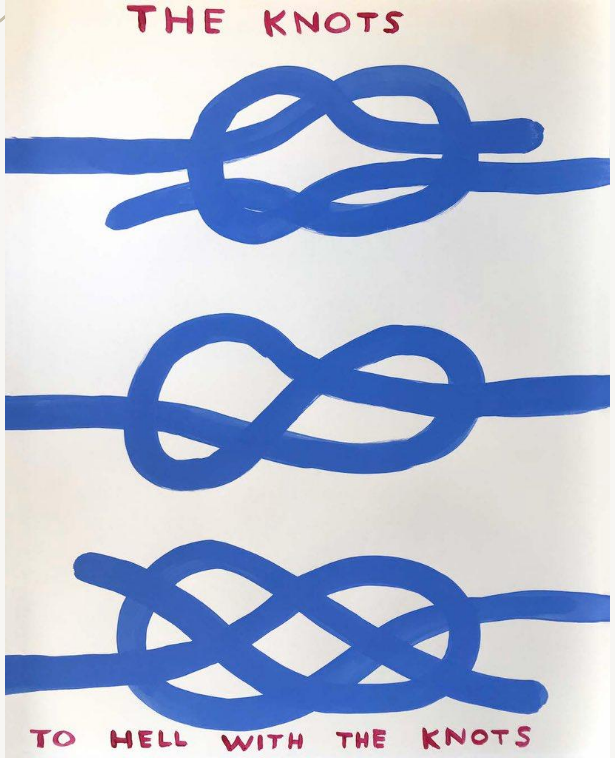
In red: The experimental scattering data. In blue: The scattering profile of the starting configuration (not good...)



A good fit to the scattering data.



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



THE PROBLEM

Want realistically entangled predictions

Fitting the SAXS data isn't enough.

Knotted proteins are rare

Mansfield, M. Are there knots in proteins?. *Nat Struct Mol Biol* **1,** 213–214 (1994).

But proteins are open ended..

Probabilistic closures required, changing the geometry.

We'll use the writhe, is this rare?

$$|\,\mathrm{Wr}(K)| < \frac{1}{4}\left(\frac{L}{R}\right)^{\frac{4}{3}}$$

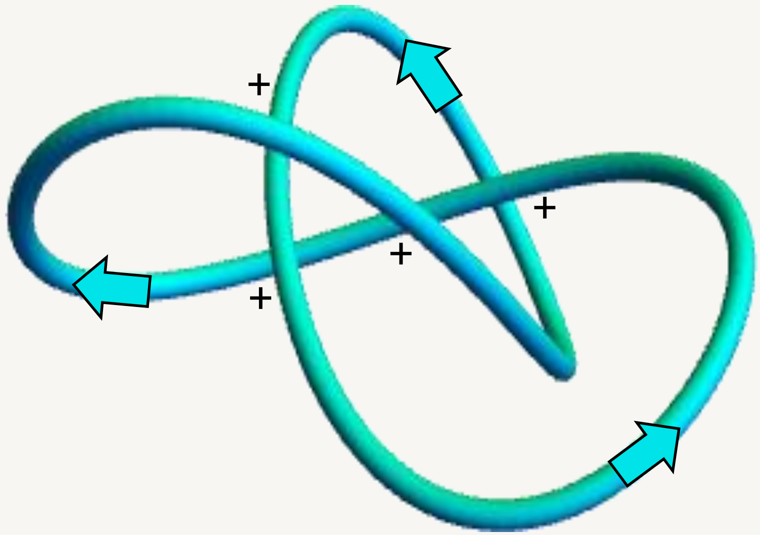
Cantarella, et al. "Upper bounds for the writhing of knots and the helicity of vector fields." *AMS IP Studies in Advanced Mathematics* 24 (2001): 1-22

~spherical averaging

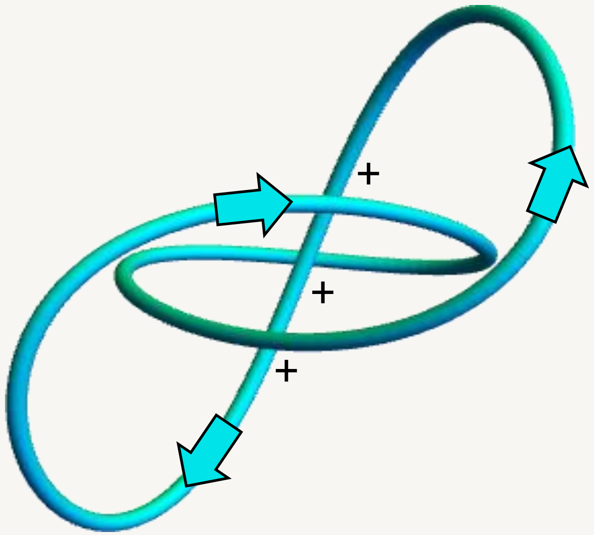
~sign of the crossing

~“strength” of the crossing

$$Wr = \frac{1}{4\pi}\oint_x \oint_x \mathbf{T}_{\mathbf{x}}(s) \times \mathbf{T}_{\mathbf{x}}(t) \cdot \frac{\mathbf{x}(s) - \mathbf{x}(t)}{\|\mathbf{x}(s) - \mathbf{x}(t)\|^3} \, \mathrm{d}s \; \mathrm{d}t$$



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

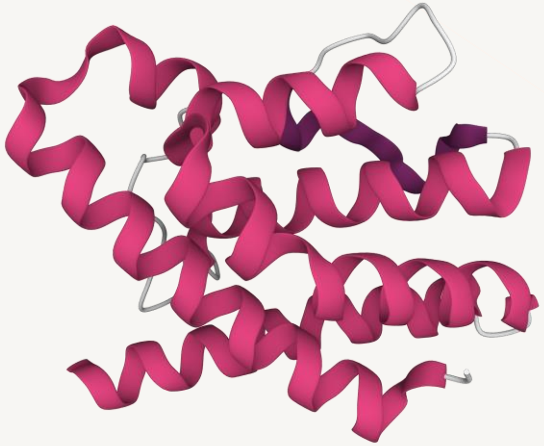


From this angle, we see +3 crossings on the trefoil knot.

SMOOTHING

- Given the secondary structure fingerprint of a protein.
- Take the start point ($c\alpha$ 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.

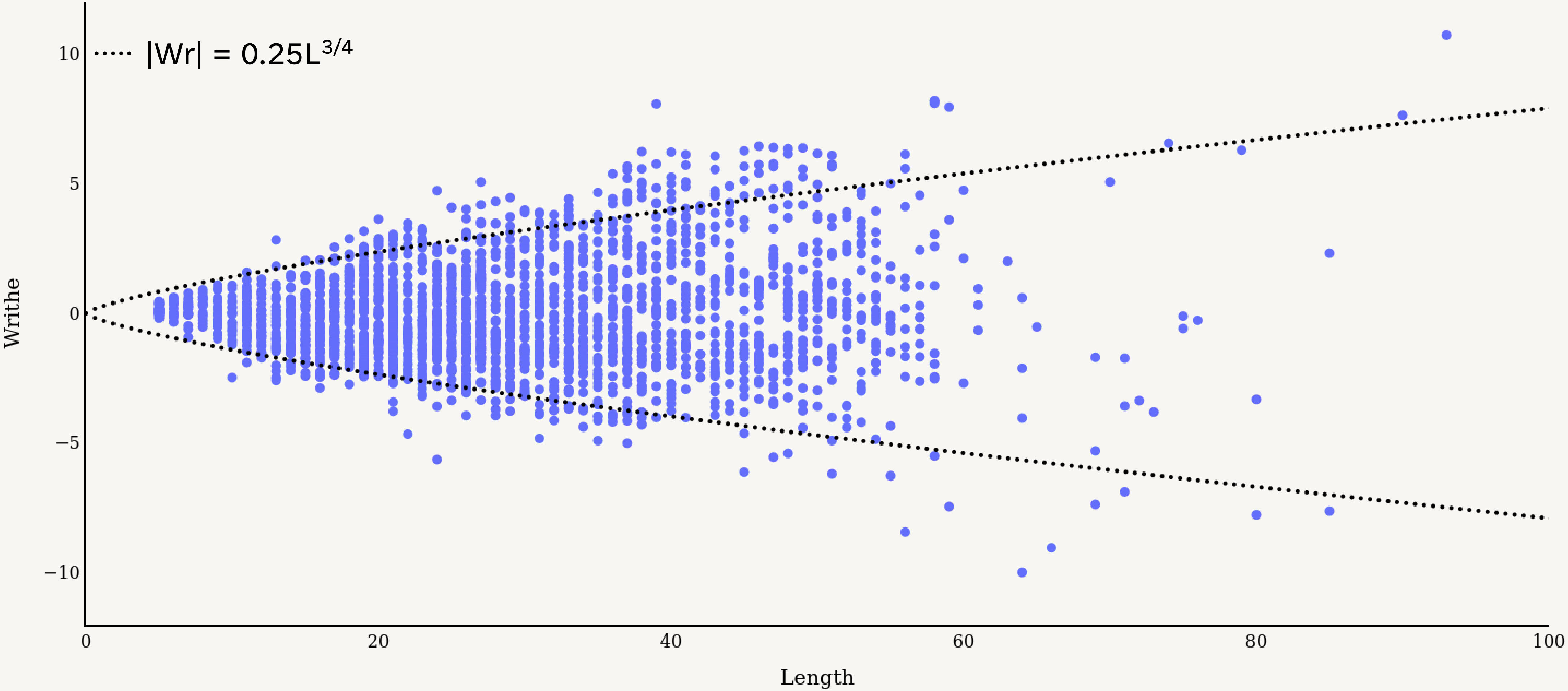


Cartoon representation of 2OEB. In pink are the alpha-helices and in white the flexible linkers.

Our smoothed representation of 2OEB. No real global entanglement.

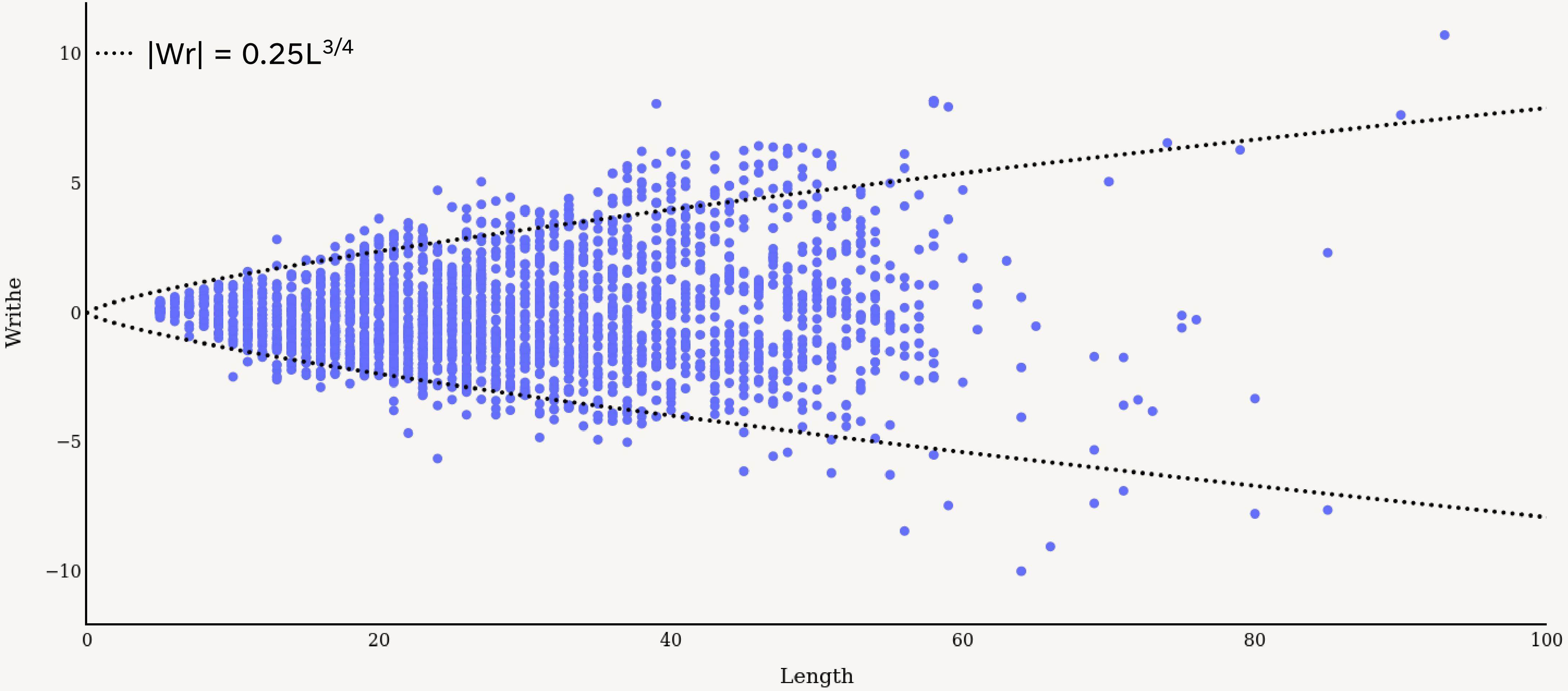
Recall Cantarella: $|\mathrm{Wr}(K)| < \frac{1}{4}\left(\frac{L}{R}\right)^{\frac{4}{3}}$

An empirical bound on the writhe for a representative sample of >2000 proteins from the PDB

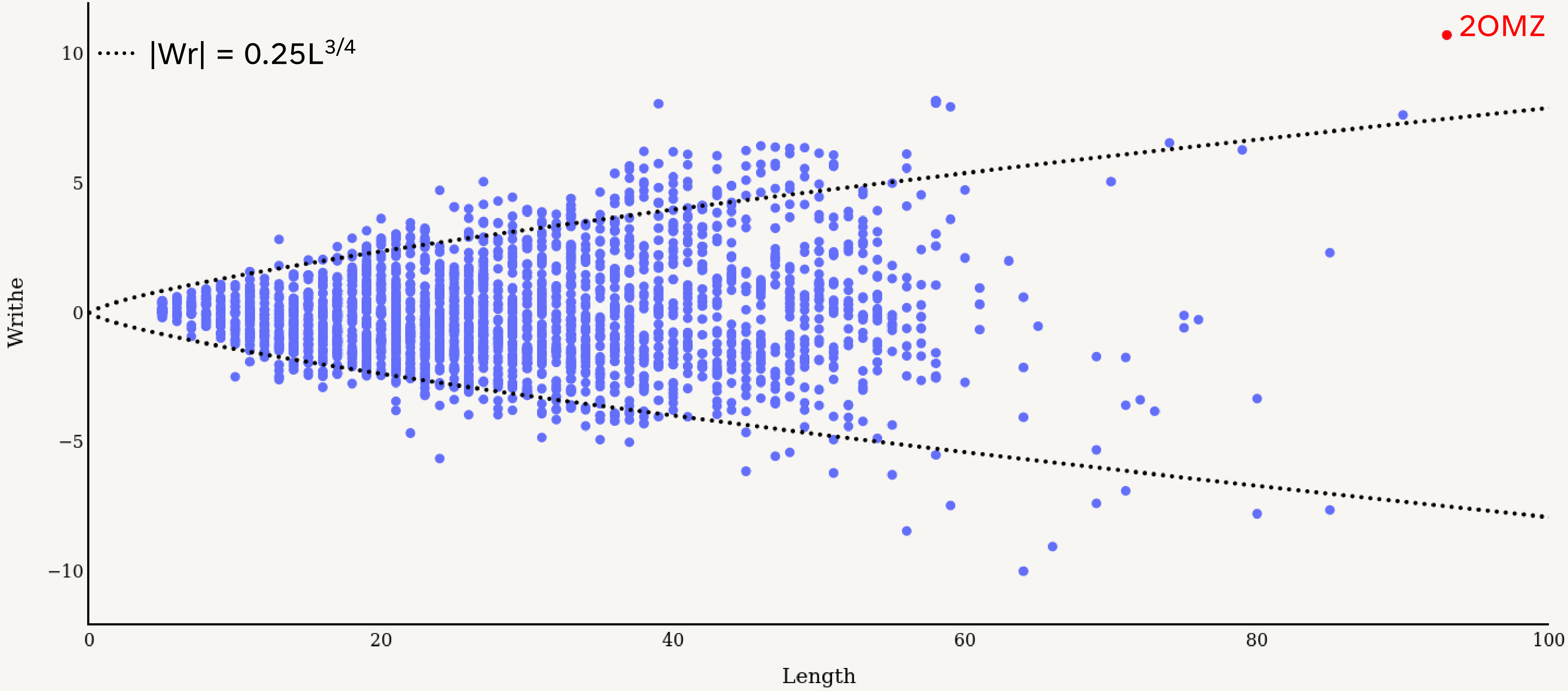


Recall Cantarella: $|\mathrm{Wr}(K)| < \frac{1}{4}\left(\frac{L}{R}\right)^{\frac{4}{3}}$

An empirical bound on the writhe for a representative sample of >2000 proteins from the PDB

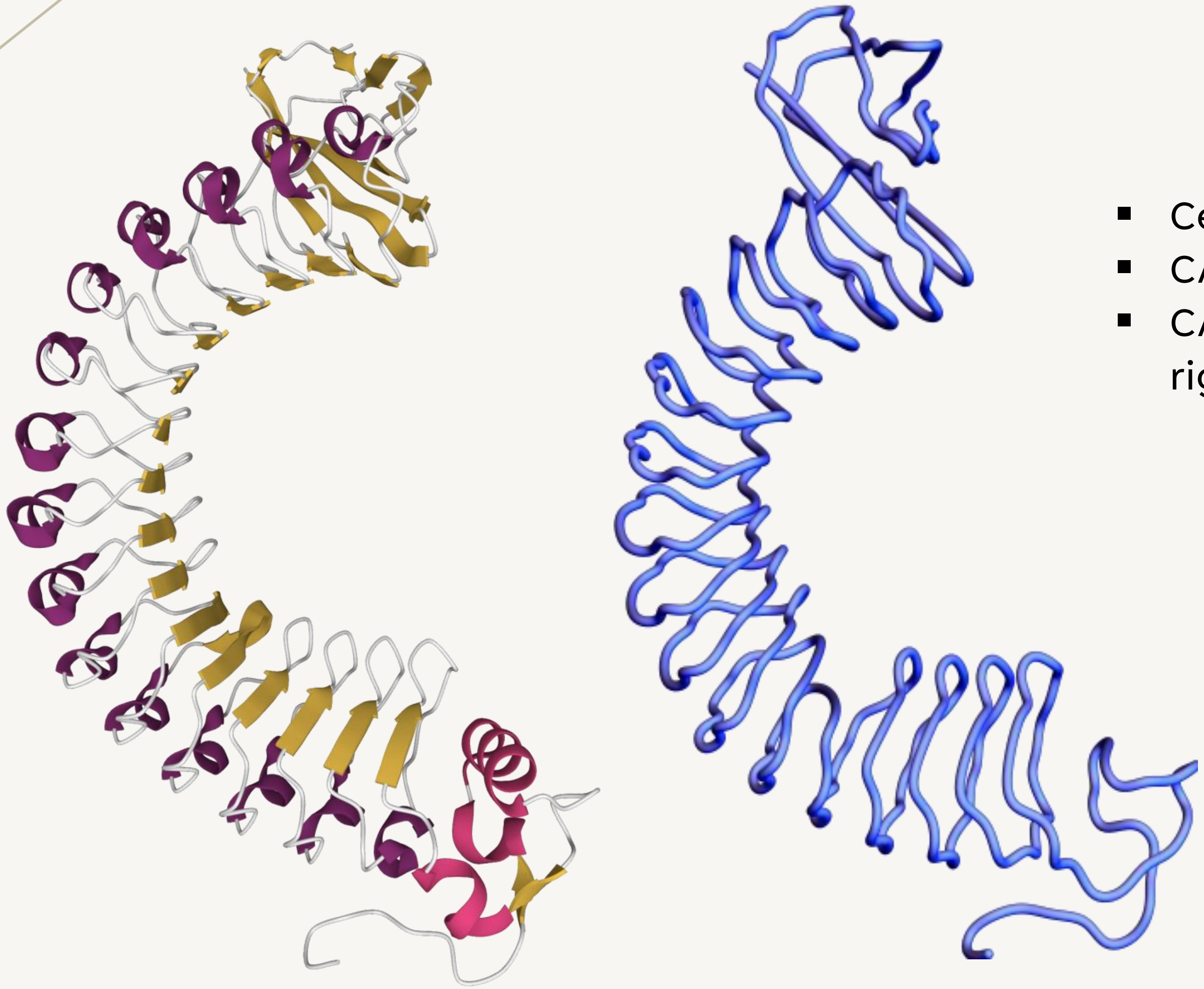


An empirical bound on the writhe for a representative sample of >2000 proteins from the PDB



2OMZ

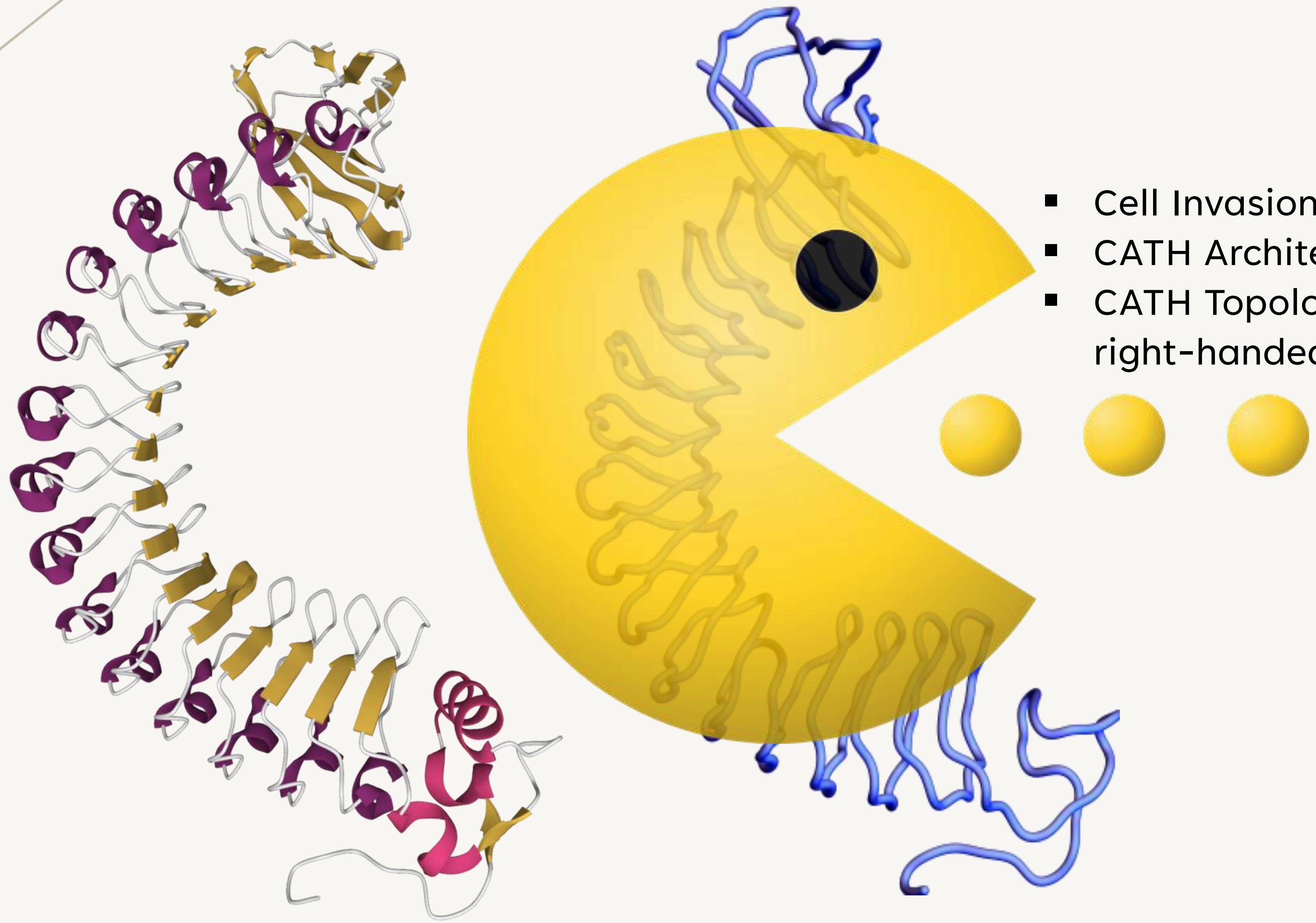
Wollert, T., Heinz, D.W., Schubert, W.D.
(2007) Proc Natl Acad Sci U S A 104: 13960-13965



- Cell Invasion/Cell Adhesion protein
- CATH Architecture: alpha-beta horseshoe
- CATH Topology: Leucine-rich repeat (LRR), right-handed beta-alpha superhelix

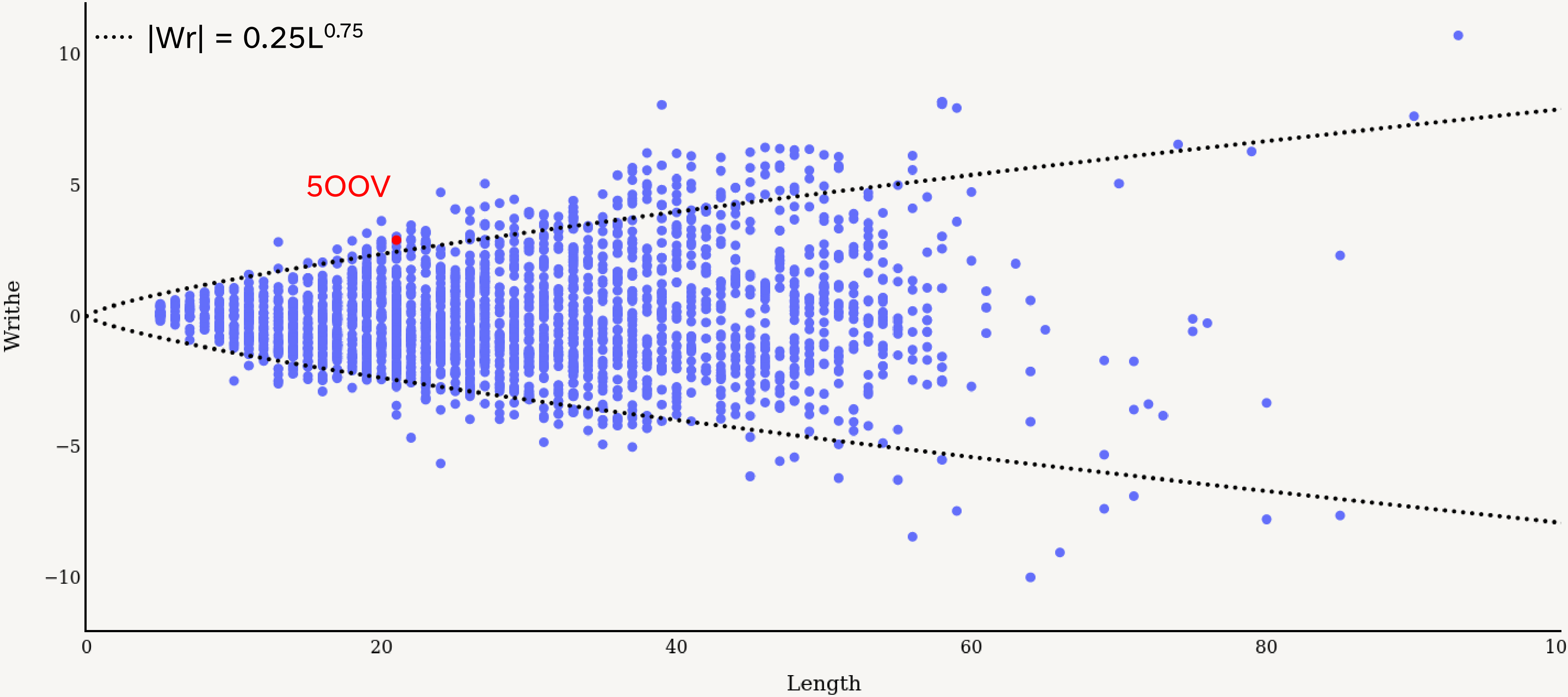
2OMZ

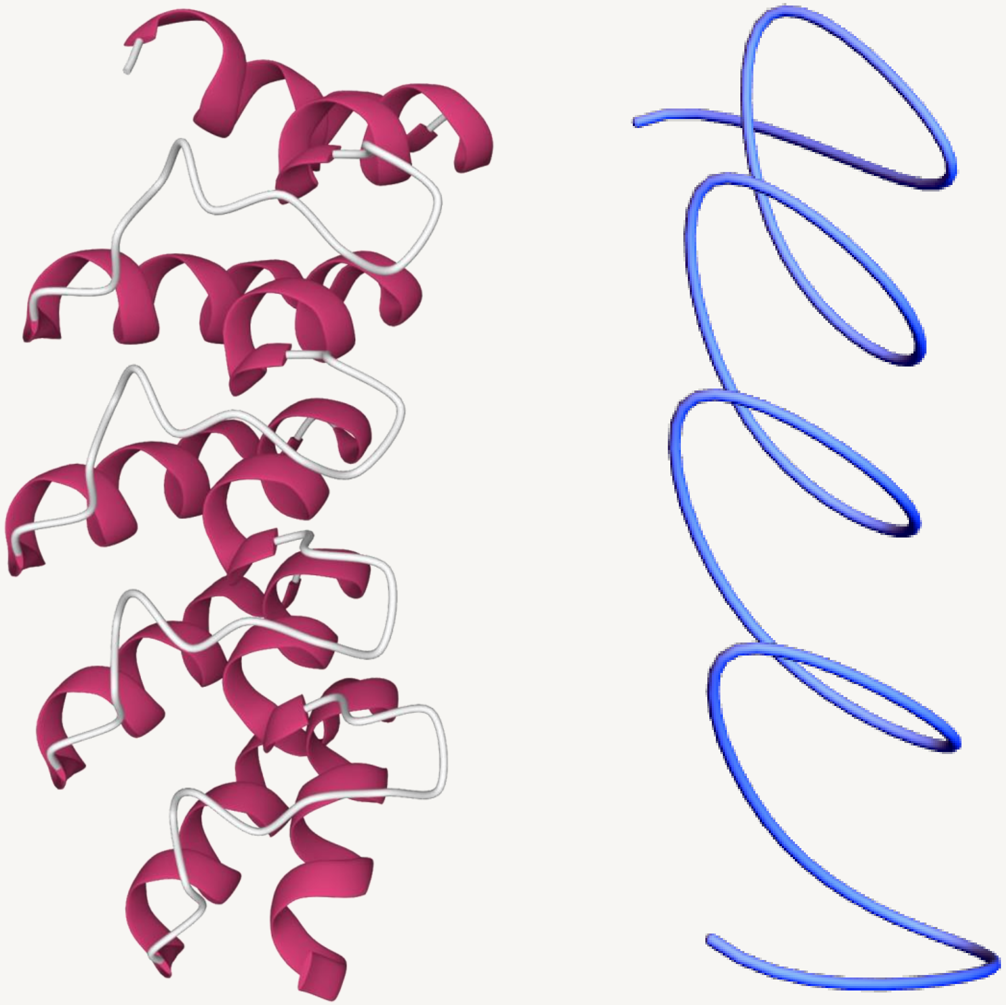
Wollert, T., Heinz, D.W., Schubert, W.D.
(2007) Proc Natl Acad Sci U S A 104: 13960-13965



- Cell Invasion/Cell Adhesion protein
- CATH Architecture: alpha-beta horseshoe
- CATH Topology: Leucine-rich repeat (LRR), right-handed beta-alpha superhelix

An empirical bound on the writhe for a representative sample of >2000 proteins from the PDB





5OOV

Designed Ankyrin Repeat Protein (DARPin) ETVD-1 in complex with Lysozyme
Fischer, G.
To be published.

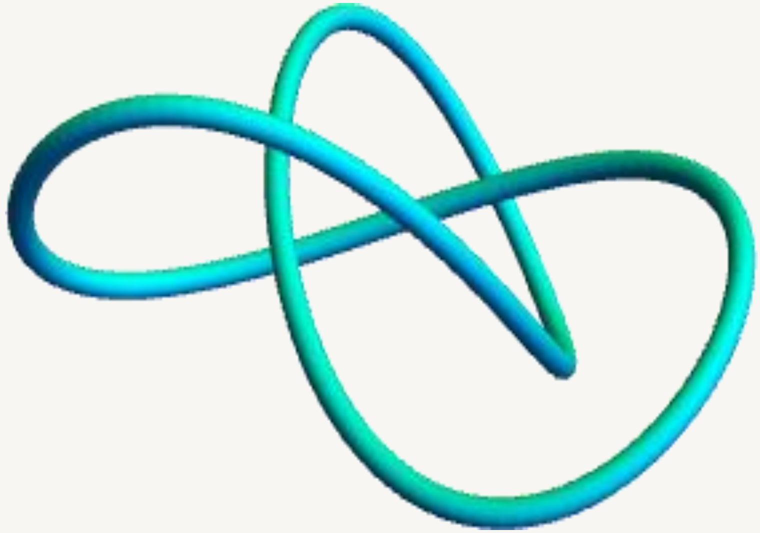
- De Novo Synthetic Protein
- CATH Arcitecture: alpha horseshoe
- CATH Topology: Ankyrin repeat

~spherical averaging

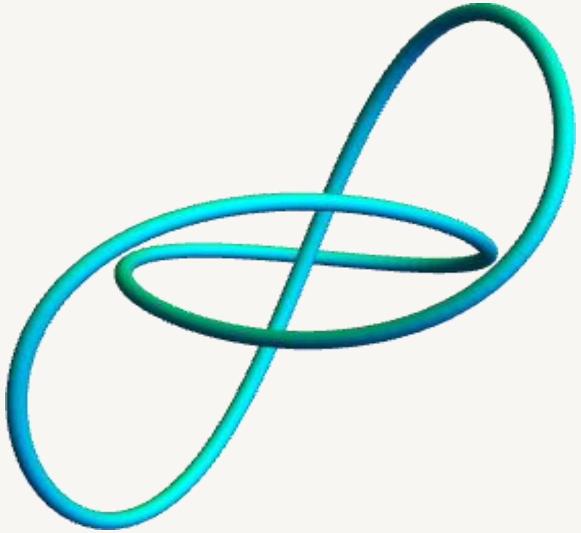
ignore sign!

~“strength” of the crossing

$$Wr = \frac{1}{4\pi} \oint_x \oint_x \left| \mathbf{T}_{\mathbf{x}}(s) \times \mathbf{T}_{\mathbf{x}}(t) \right| \cdot \frac{\mathbf{x}(s) - \mathbf{x}(t)}{\left\| \mathbf{x}(s) - \mathbf{x}(t) \right\|^3} \, \mathrm{d}s \, \mathrm{d}t$$

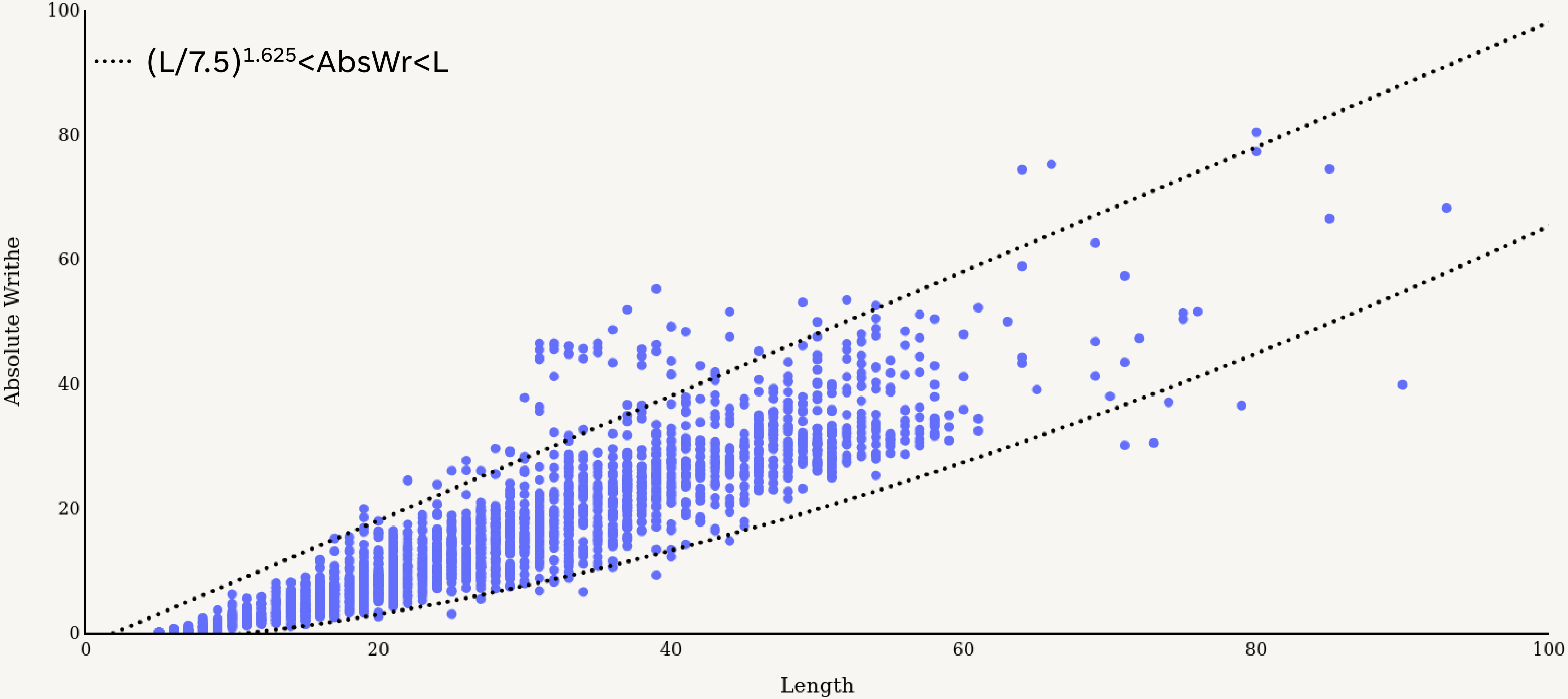


From this angle, we see 4 crossings on the trefoil knot.

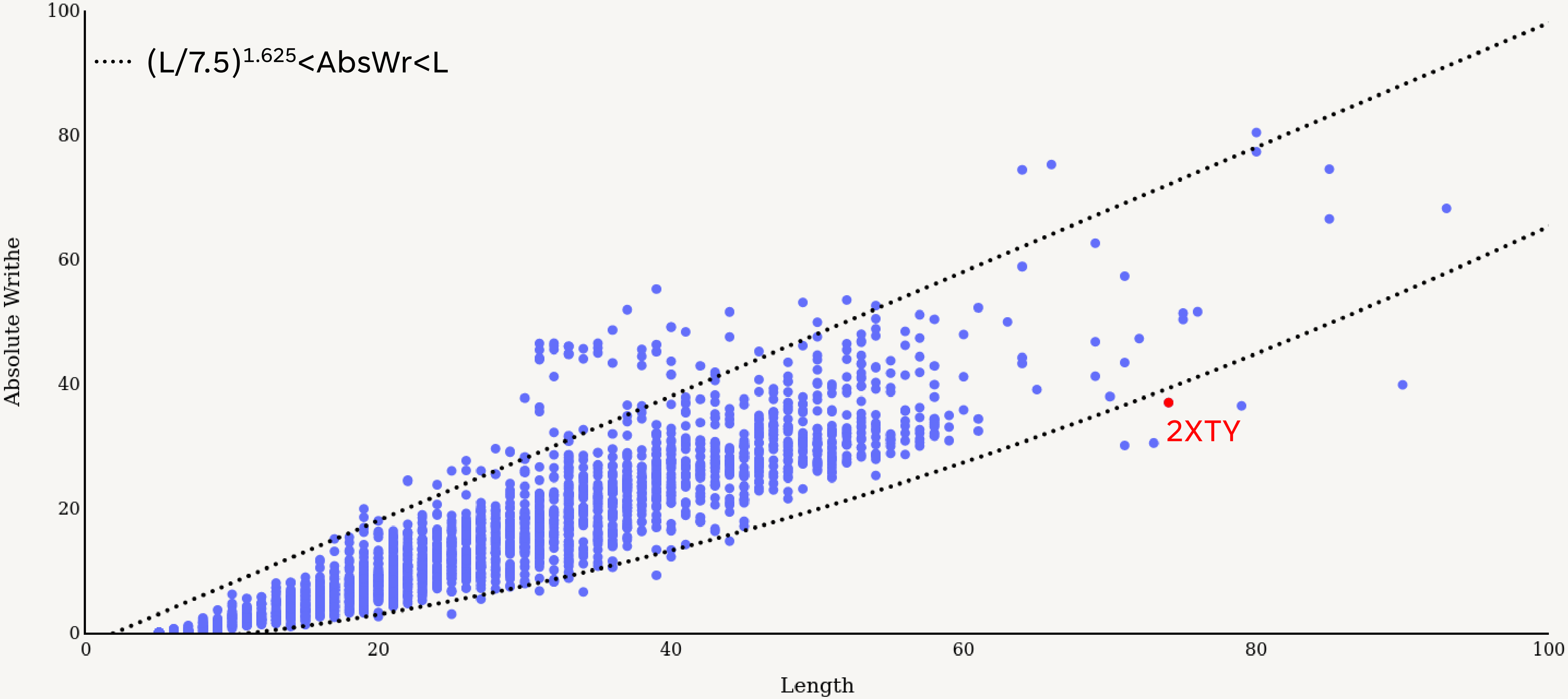


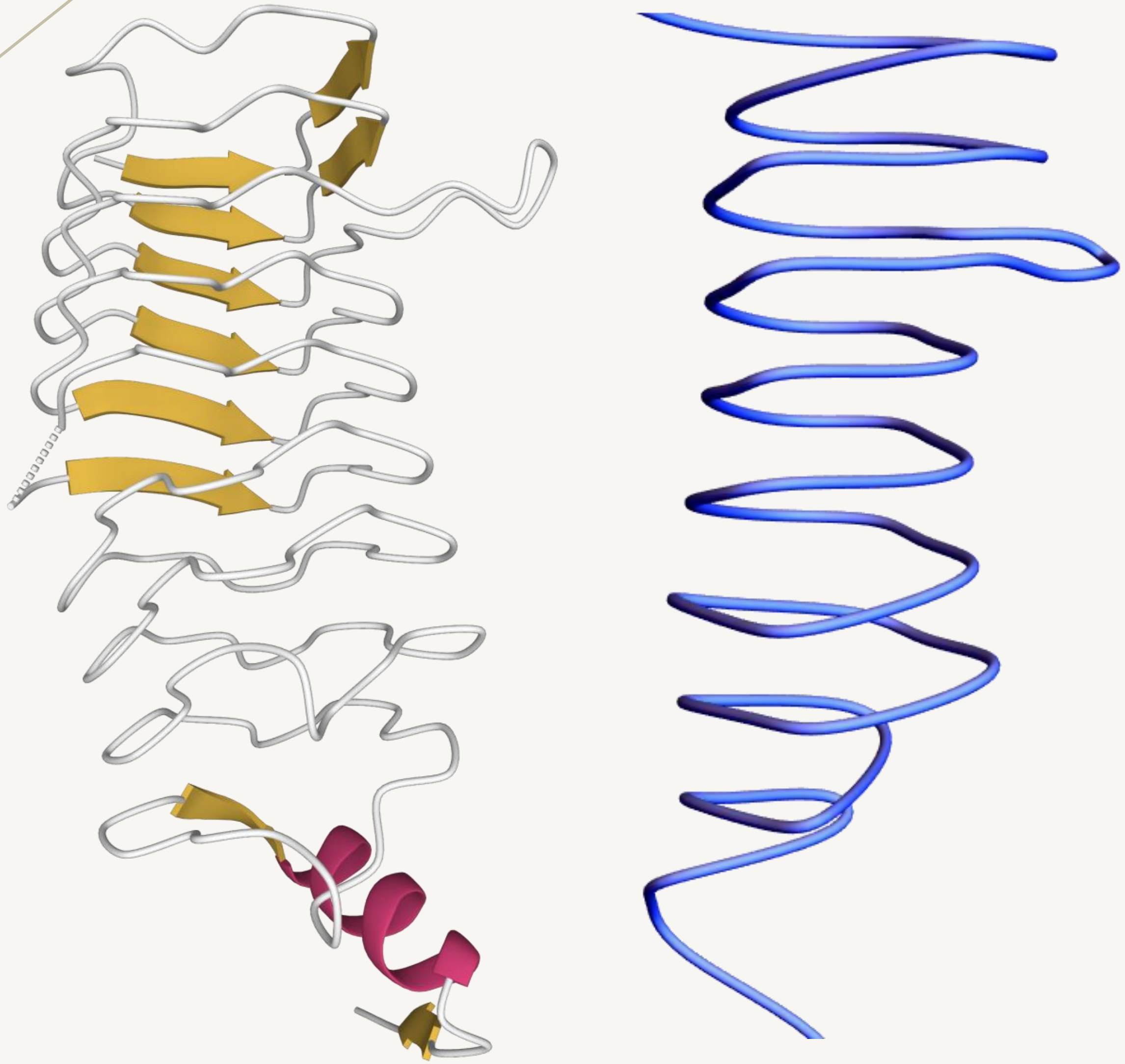
From this angle, we see 3 crossings on the trefoil knot.

An empirical bound on the absolute writhe for a representative sample of >2000 proteins from the PDB



An empirical bound on the absolute writhe for a representative sample of >2000 proteins from the PDB



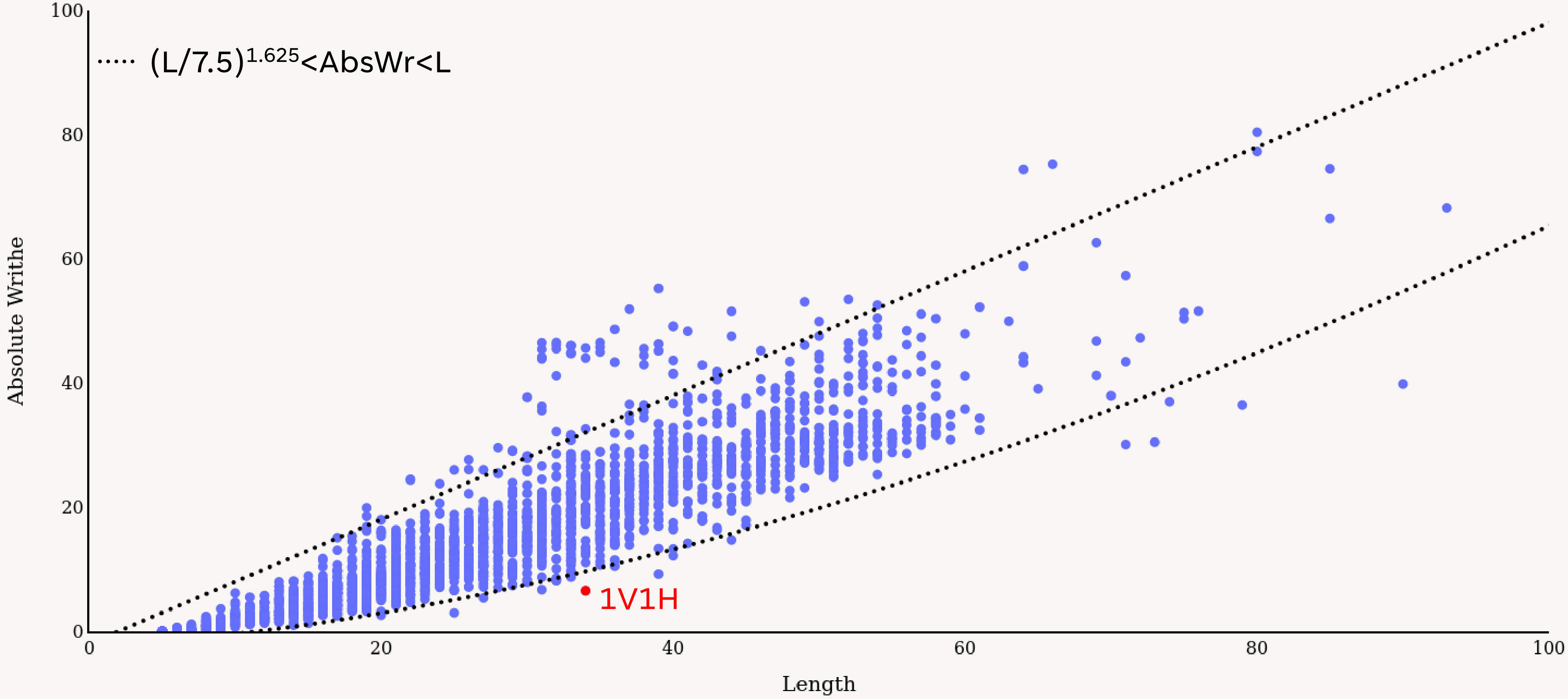


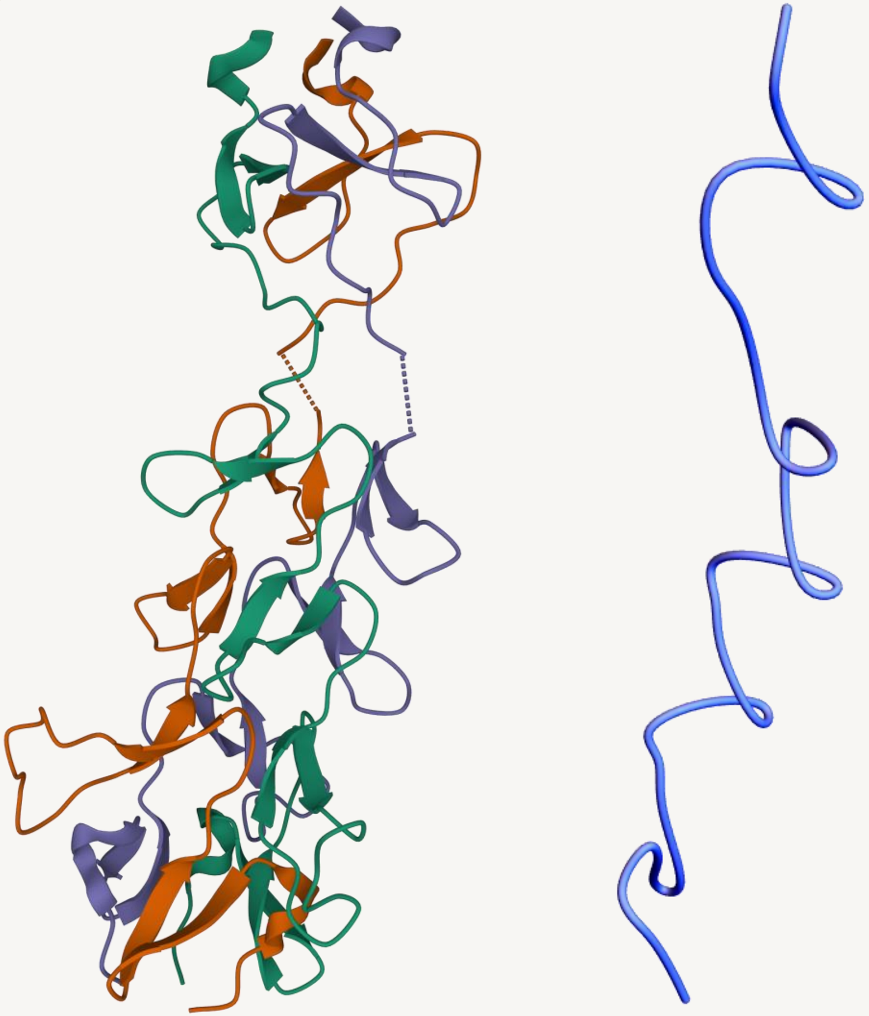
2XTY

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 prediction not always reliable
- With correct length, this falls well inside the lower bound

An empirical bound on the absolute writhe for a representative sample of >2000 proteins from the PDB



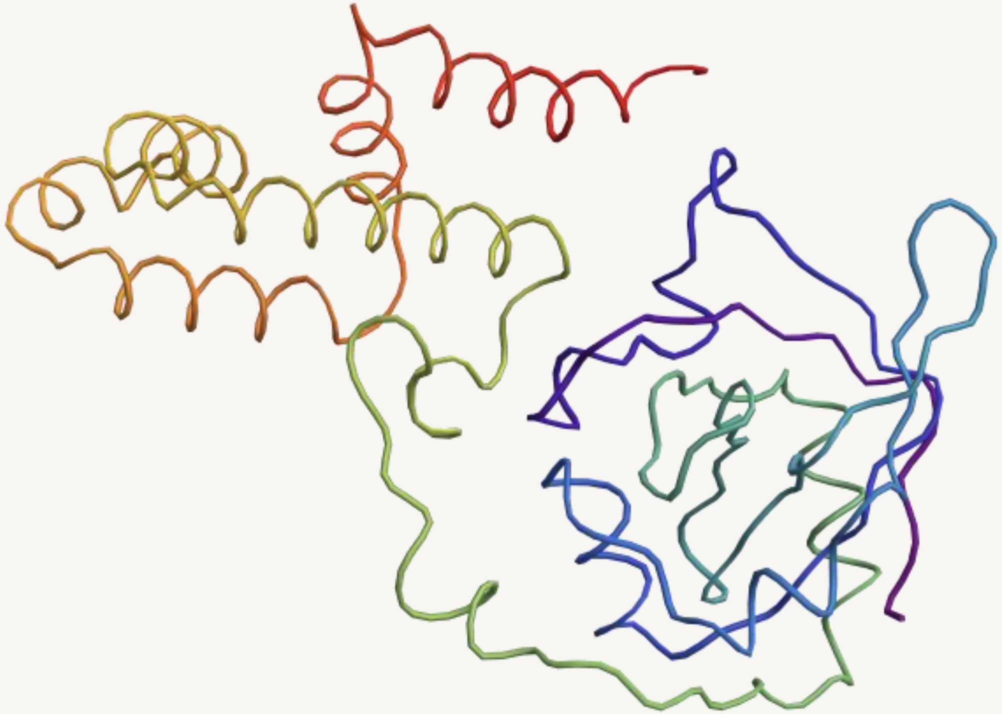


1V1H

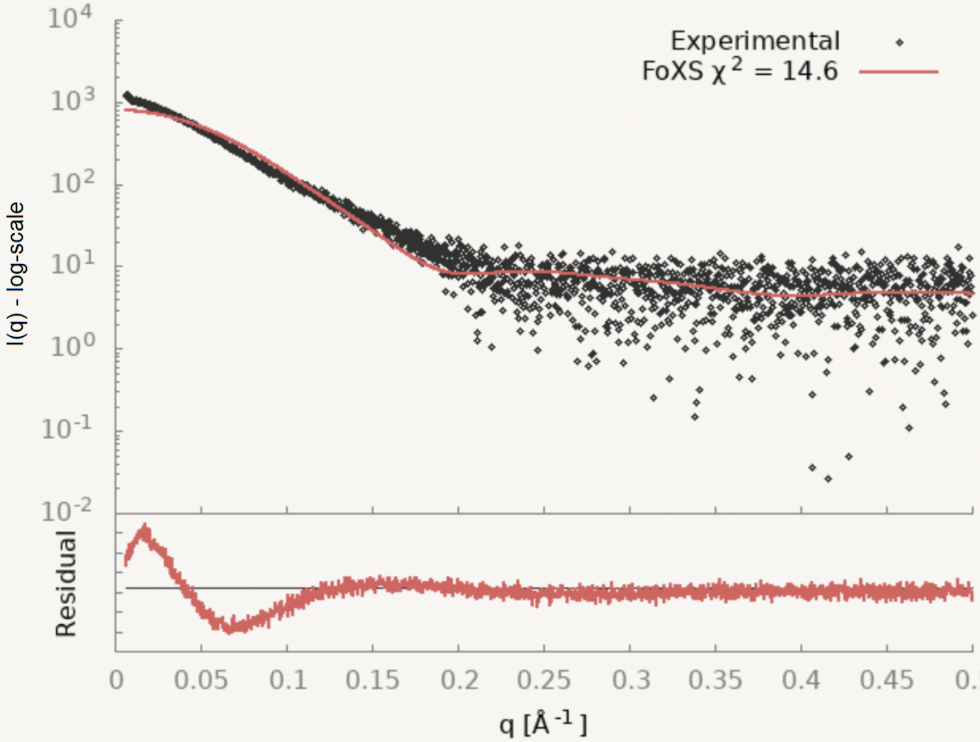
Papanikolopoulou, K., Teixeira, S., Belrhali, H., Forsyth, V.T., Mitraki, A., van Raaij, M.J.
(2004) J Mol Biol 342: 219

- We only consider monomers in our study, or single chains of a multimer
- The overall topology of the protein is interesting, roughly a triple helix.
- The single chain however not so much.

MICA



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

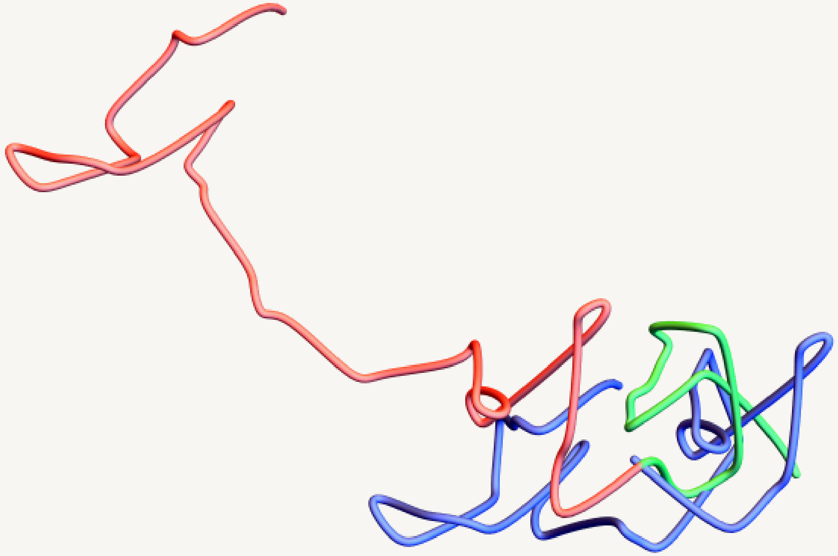


In red, the FoXS[2] scattering profile for the crystallographic structure. This is a poor fit, especially at low q, indicating the global structure is not right.

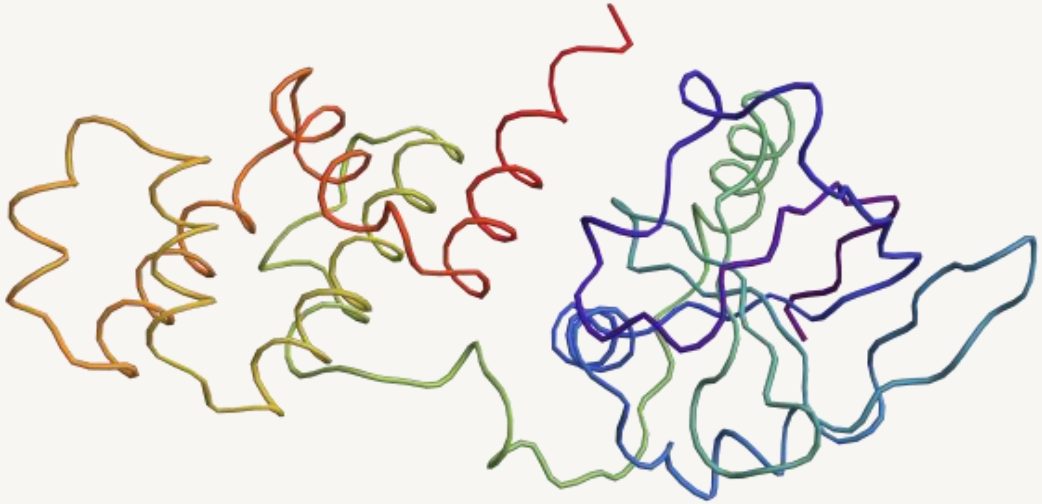
---

[1] Jumper, J et al. Highly accurate protein structure prediction with AlphaFold. Nature (2021).

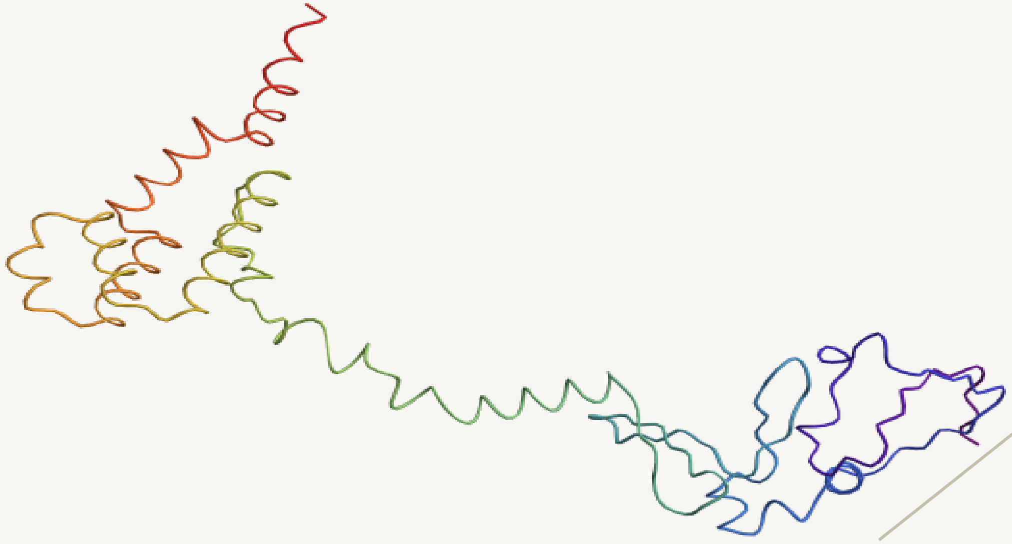
[2] Schneidman-Duhovny D, Hammel M, Tainer JA, and Sali A. Accurate SAXS profile computation and its assessment by contrast variation experiments. Biophysical Journal 2013. 105 (4), 962-974



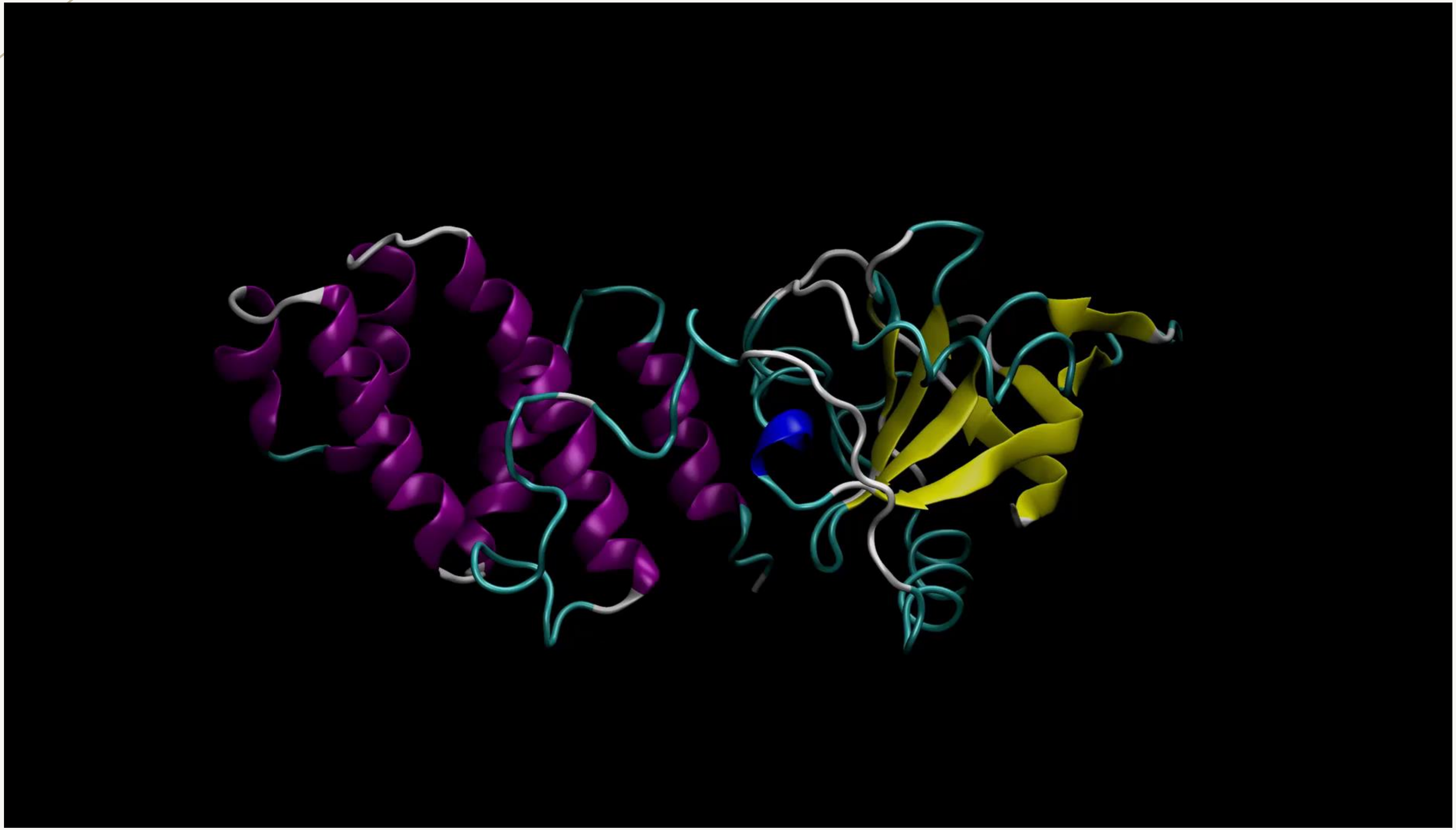
Two good fits to the scattering data.
Where they agree is shown in green. The conformations on the left and right are shown in blue and red respectively



A good fit to the scattering data that falls well above the lower bound on absolute writhe



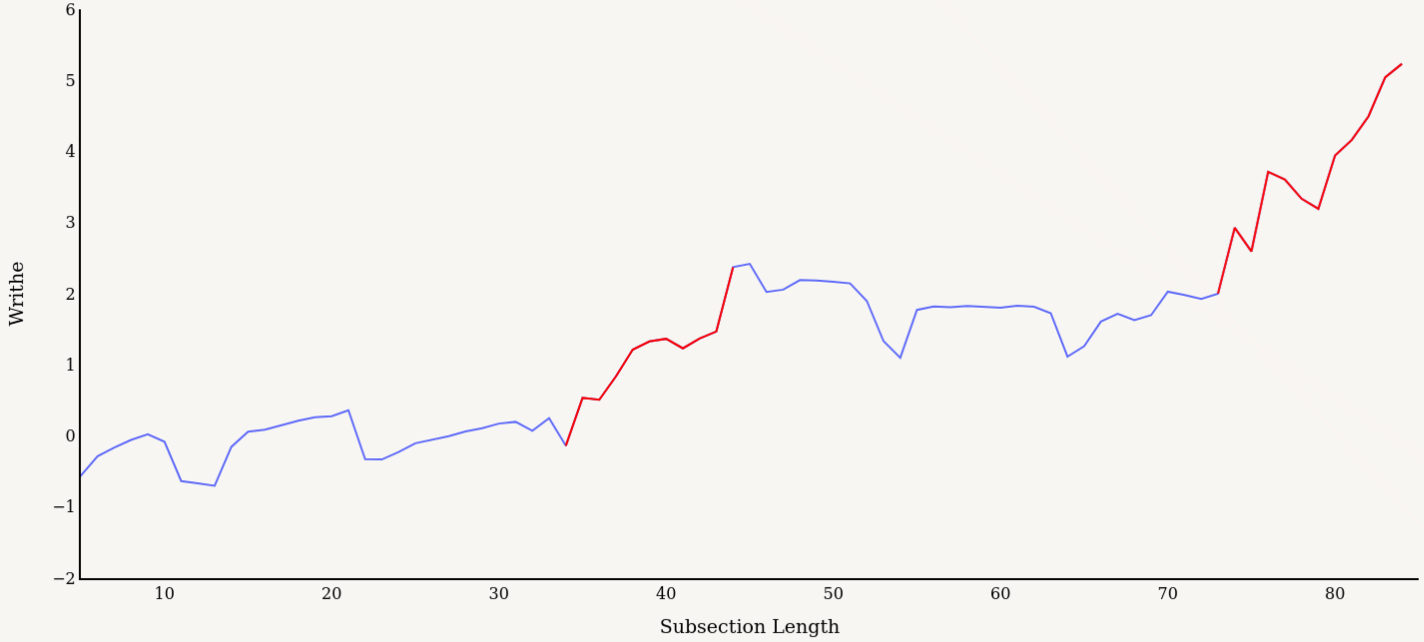
Another good fit to the scattering data, however this one falls well below the lower bound on absolute writhe

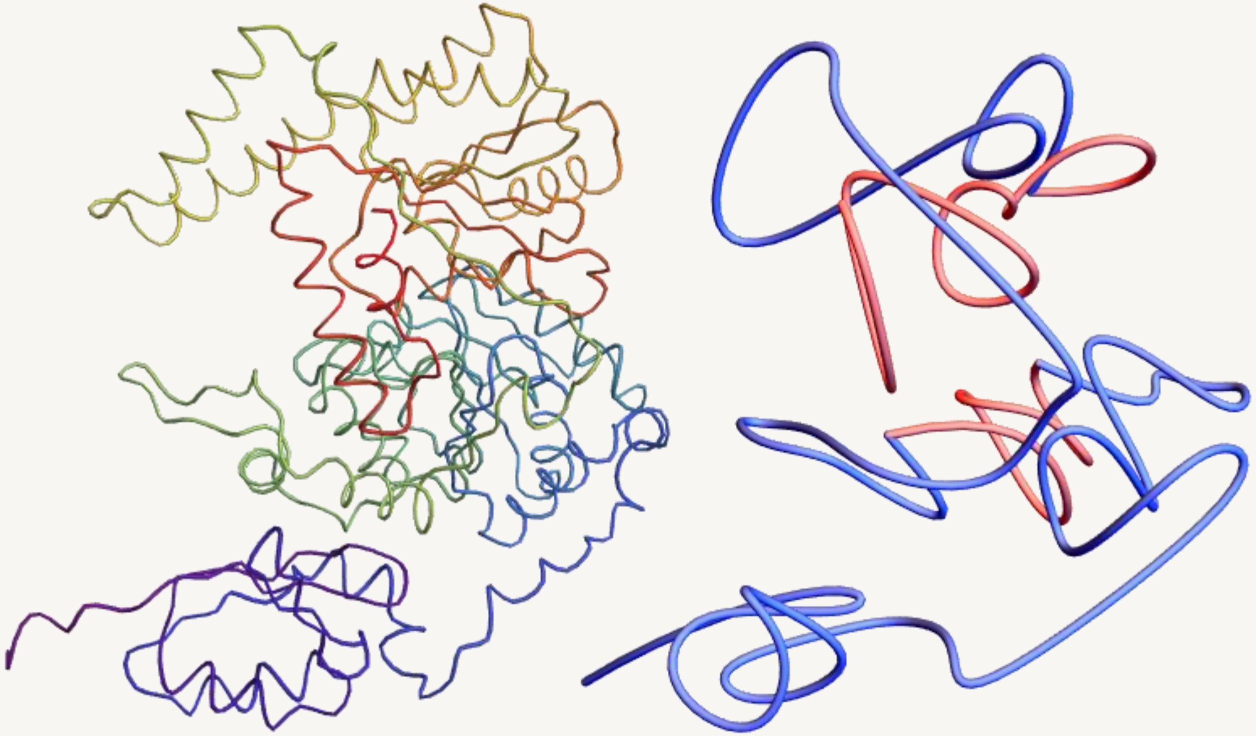


MD simulation and videos provided by Josh McKeown

Human SMARCAL1

- Consider the curve given by the first n points of the (smoothed) protein.
- Plot the writhe of this subsection of the protein.
- Repeat for all n to see how the writhe varies as the protein "grows".





Two subsections of maximal increase in writhe.
Both exhibit the secondary structure pattern seen in 2omz earlier.

CONCLUSIONS

The writhe of proteins is bounded

Proteins that attain their writhe bound provide interesting space for functional study.

The absolute writhe of proteins is bounded.

This allows us to make realistic protein predictions for molecular dynamics simulations.

Subsections of proteins can have maximal writhe

Following the same pattern as the globally writhe bound beating proteins.

THANK YOU FOR LISTENING

Any questions?