MoSMed
Molecular Sciences for Medicine

# Using writhe to produce realistic protein structure predictions from BioSAXS data

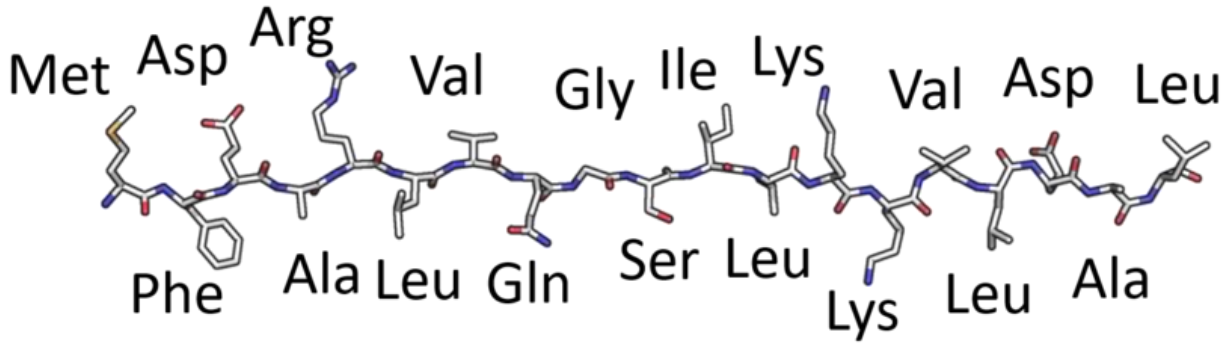
Arron Bale, supervised by Chris Prior and Rob Rambo

diamond
Durham University
Newcastle University
UKRI Engineering and Physical Sciences Research Council

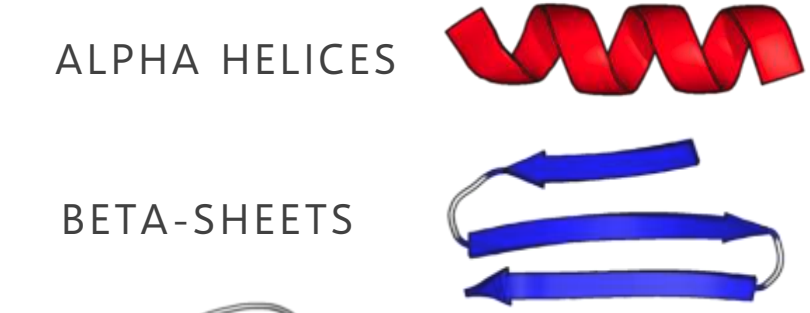
# Start easy: What is a protein?

MoSMed
Molecular Sciences for Medicine

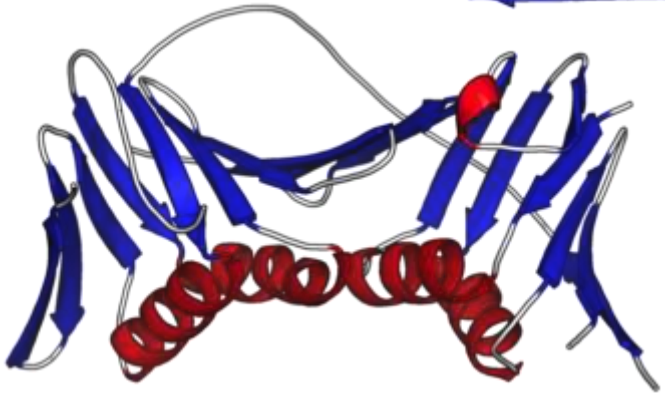
Primary Structure




Secondary Structure

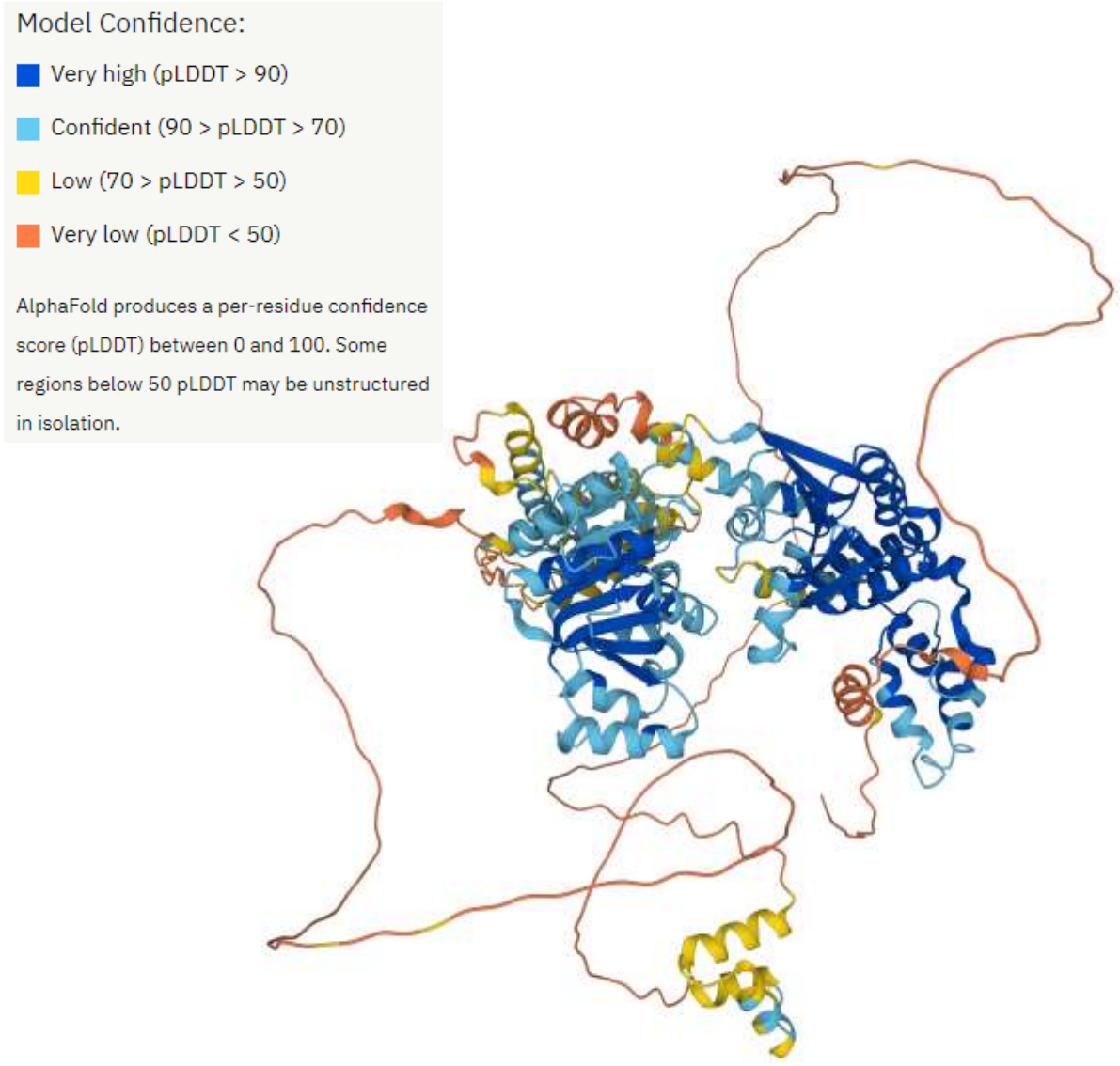



Tertiary Structure



diamond
Durham University
Newcastle University
UKRI Engineering and Physical Sciences Research Council

# But.. But.. AlphaFold?




Low accuracy predictions expected for:

- regions without homologous sequences in the PDB
- regions with multiple homologs with widely different structures

For example, Intrinsically Disordered Proteins (IDRs) would typically have low confidence predictions.

Most importantly, proteins are flexible. Any prediction should be thought of as a representative of conformational space

# The Menu

MoSMed
Molecular Sciences for Medicine

SAXS Data Collection

diamond

Initial Predictions

Carbonara

Molecular Dynamics

diamond

Durham University

Newcastle University

UKRI Engineering and Physical Sciences Research Council

# The Menu

MoSMed
Molecular Sciences for Medicine

SAXS Data Collection

diamond

Initial Predictions

Carbonara

Molecular Dynamics

diamond

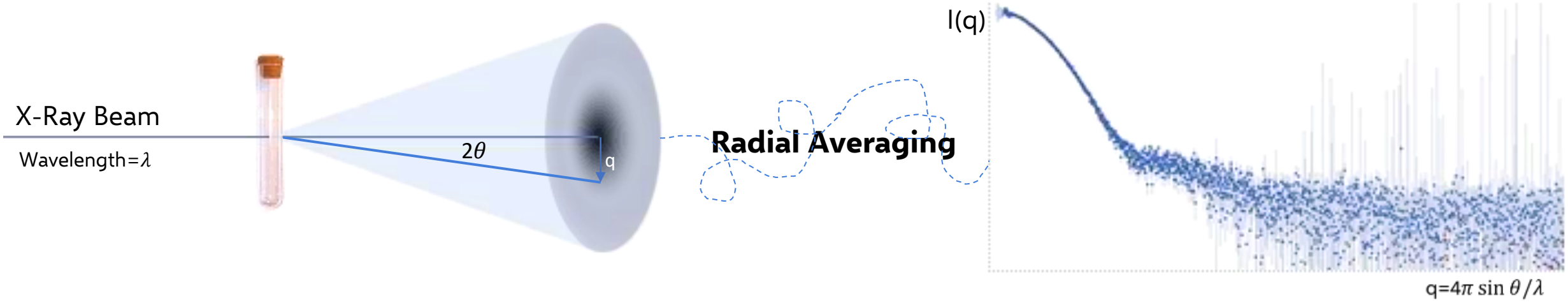
Durham University

Newcastle University

UKRI Engineering and Physical Sciences Research Council

# SAXS Data

- Can be used for proteins which fail to crystallise, or that are too large for NMR.
- Ideal technique to study proteins in solution, often indicating that obtained crystal structures do not match the shapes adopted in solution.
- The cost of this method is the difficulty in interpreting BioSAXS data, with the random motion of molecules in solution leading to isotropy of the scattering pattern.



# The Menu

MoSMed
Molecular Sciences for Medicine

SAXS Data Collection

diamond

Initial Predictions

Carbonara

Molecular Dynamics

diamond

Durham University

Newcastle University

UKRI Engineering and Physical Sciences Research Council

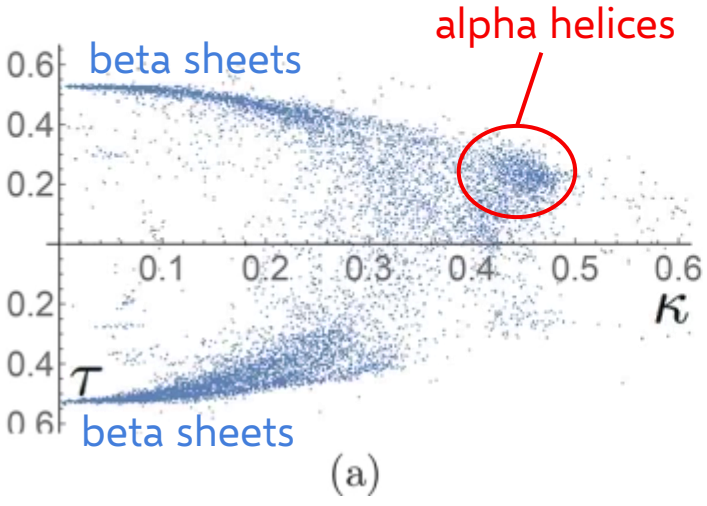
# What is a protein to me?

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

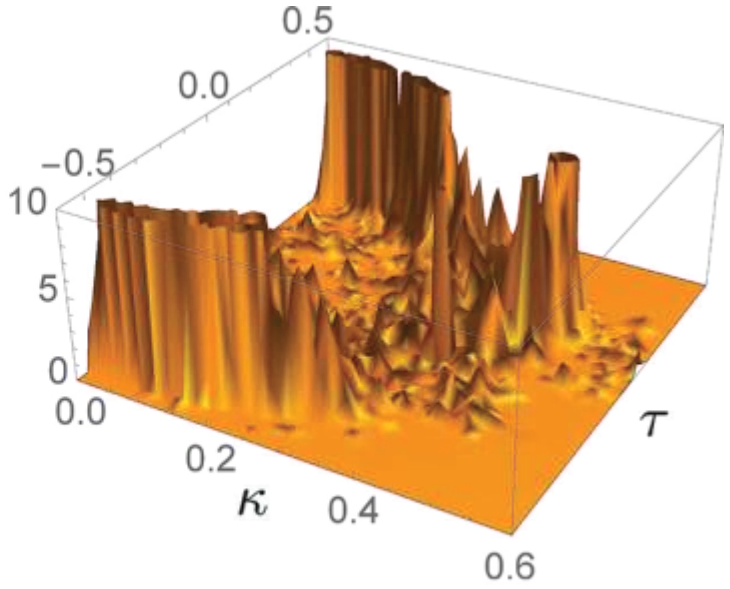
**Curvature** ($\kappa$) ~ how tightly wound is my curve.

**Torsion** ($\tau$) ~ how helical is my curve.

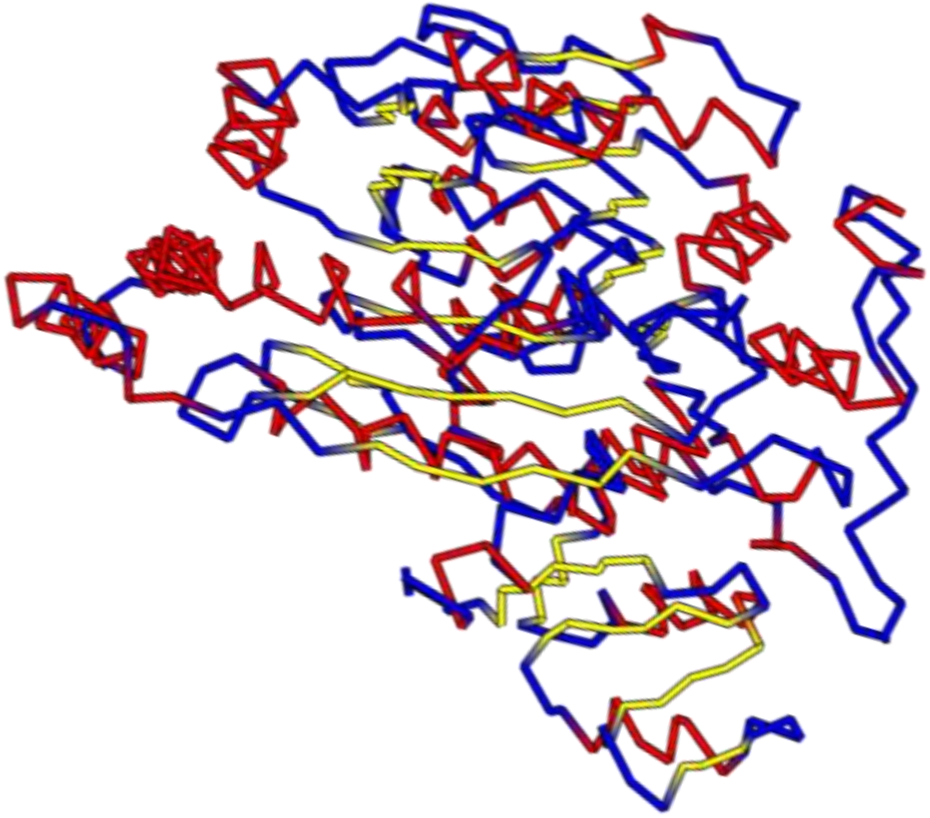
*Think Ramachandran angles*




(a)




(b)



diamond Durham University Newcastle University UKRI Engineering and Physical Sciences Research Council

# The Menu

MoSMed
Molecular Sciences for Medicine

SAXS Data Collection

diamond

Initial Predictions

Carbonara

Molecular Dynamics

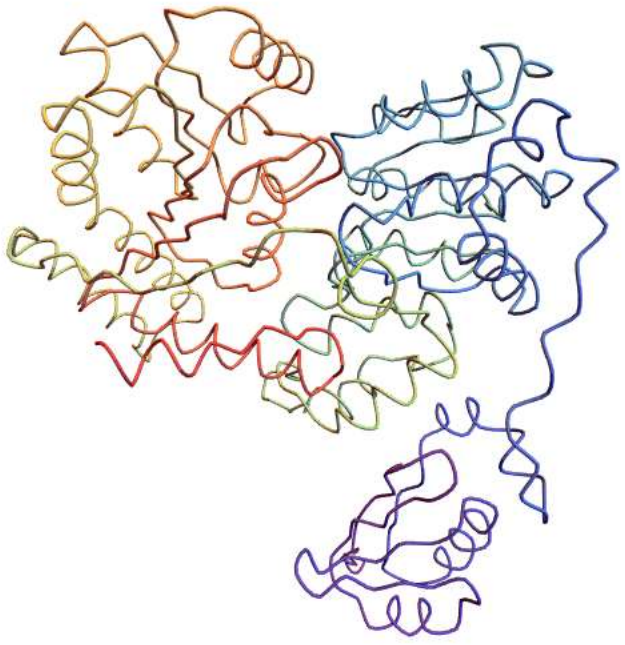
diamond

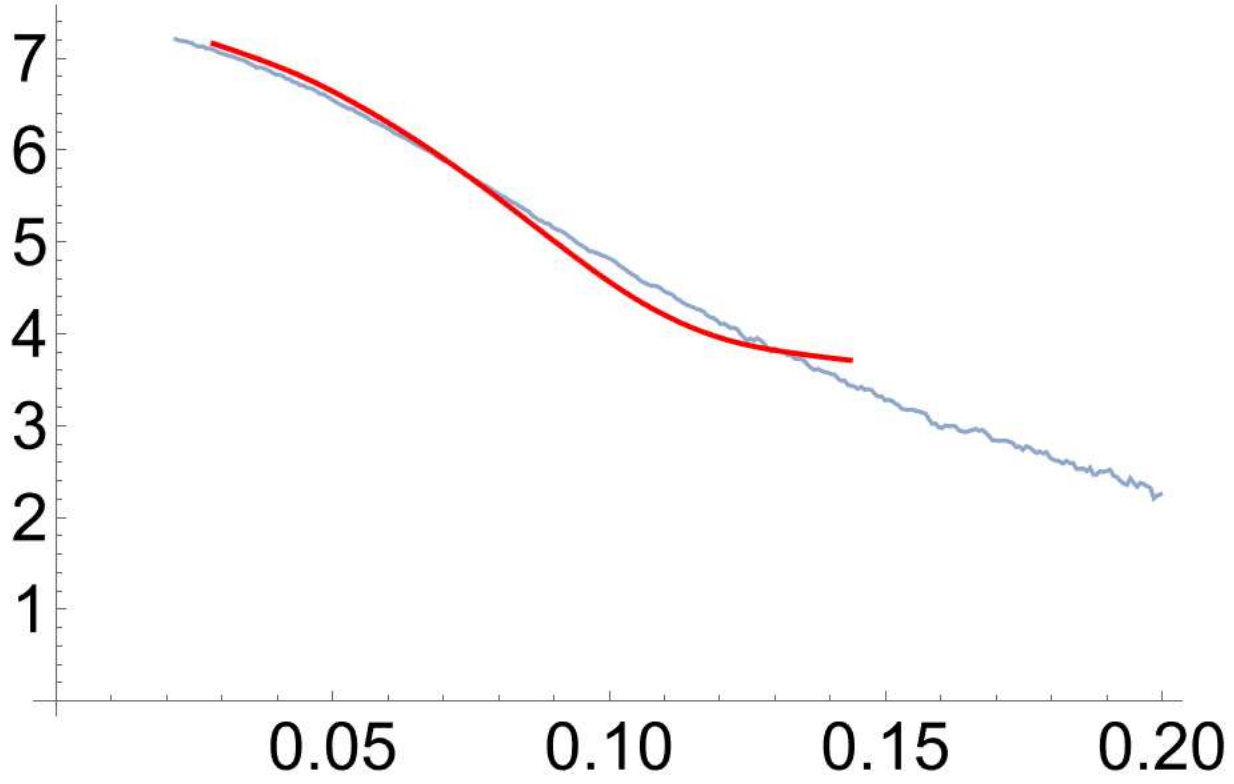
Durham University

Newcastle University

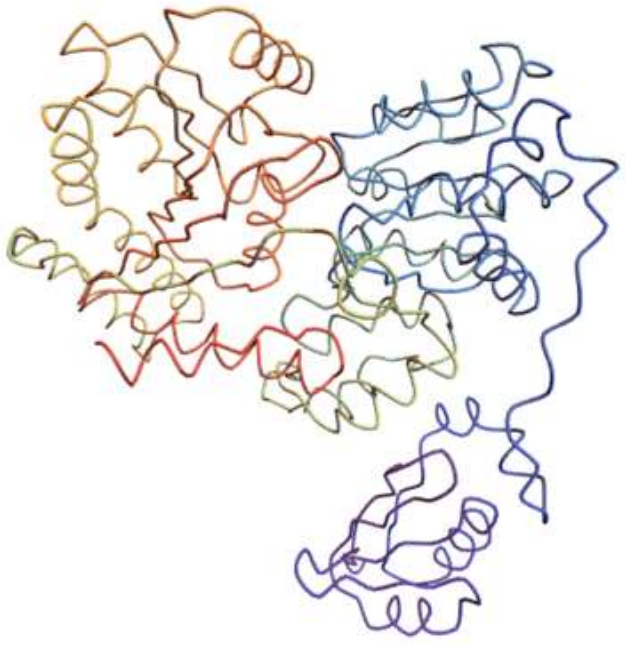
UKRI Engineering and Physical Sciences Research Council

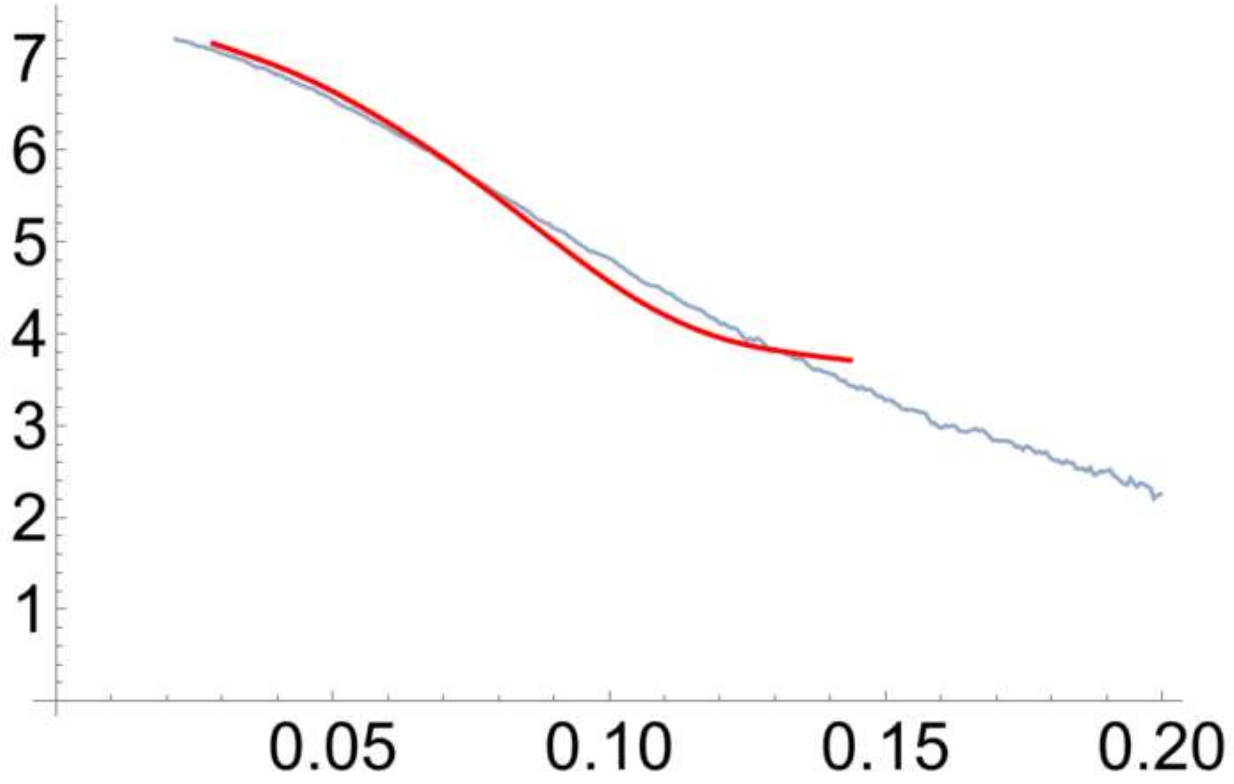
# Initial Fit



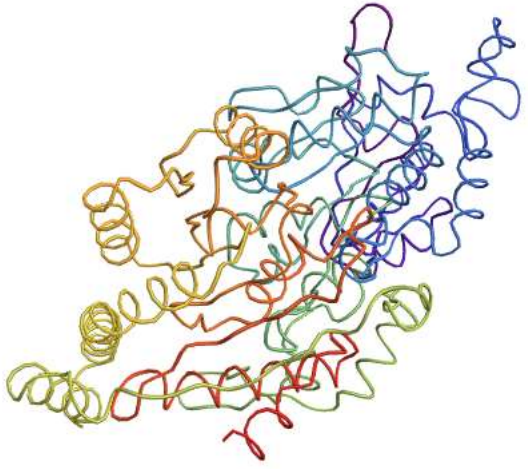


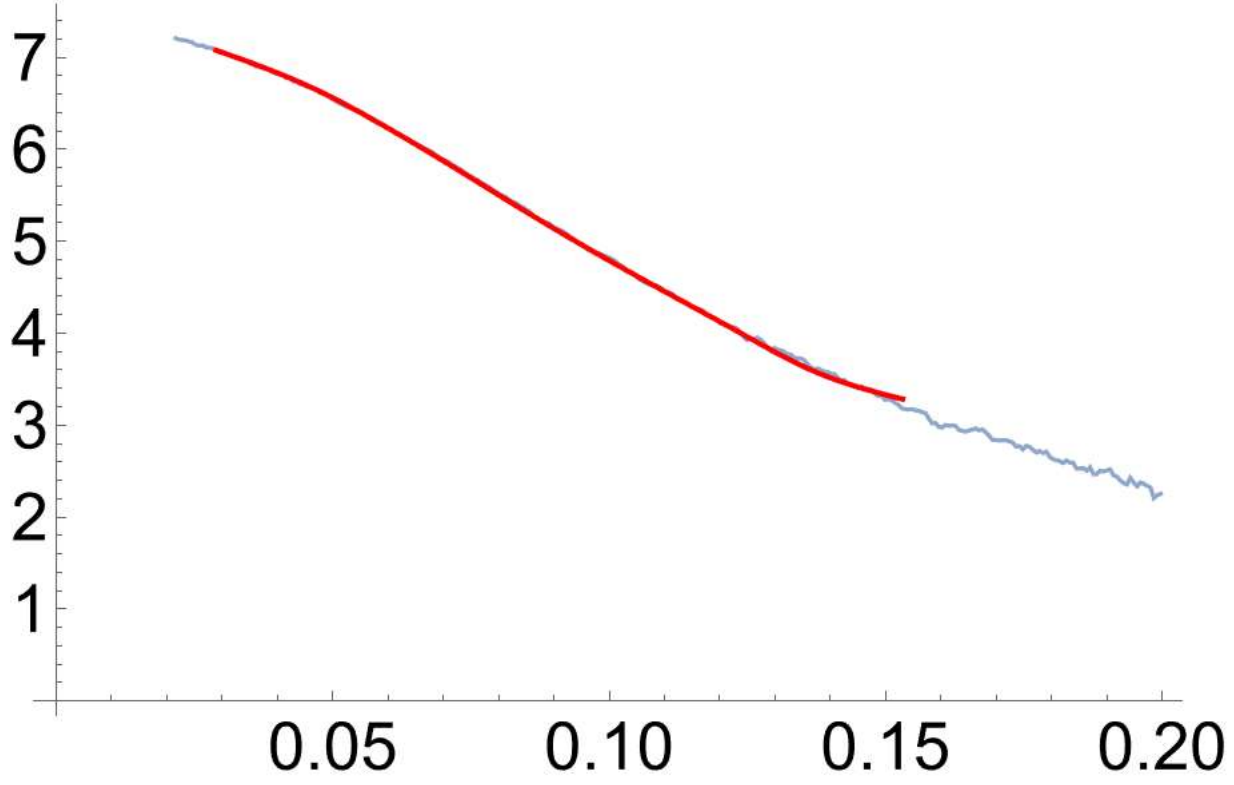
# In action



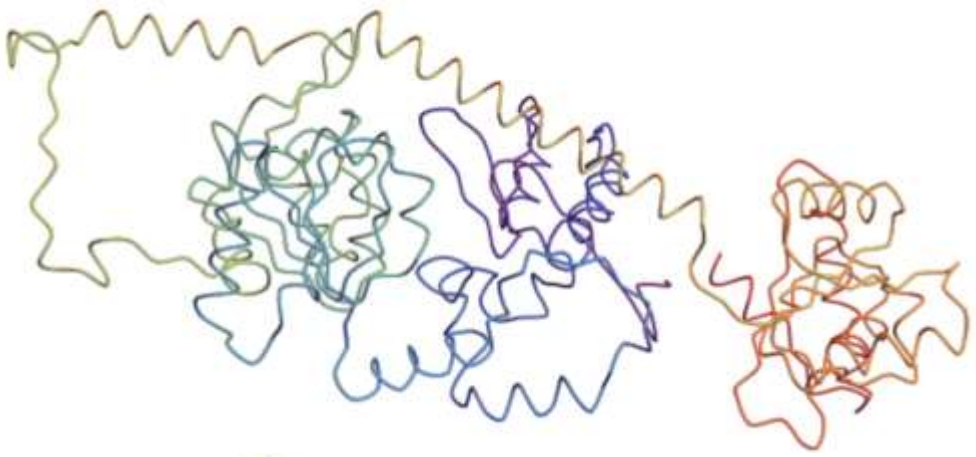


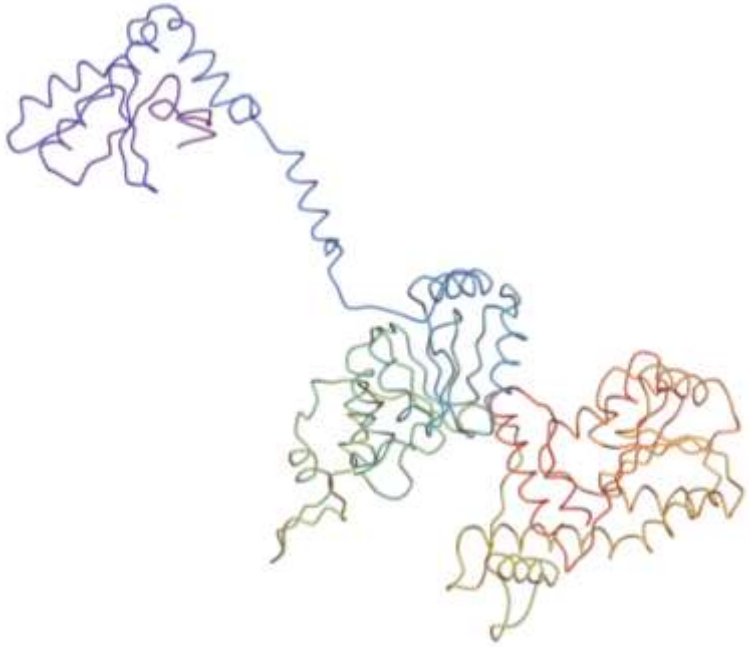
# Final fit

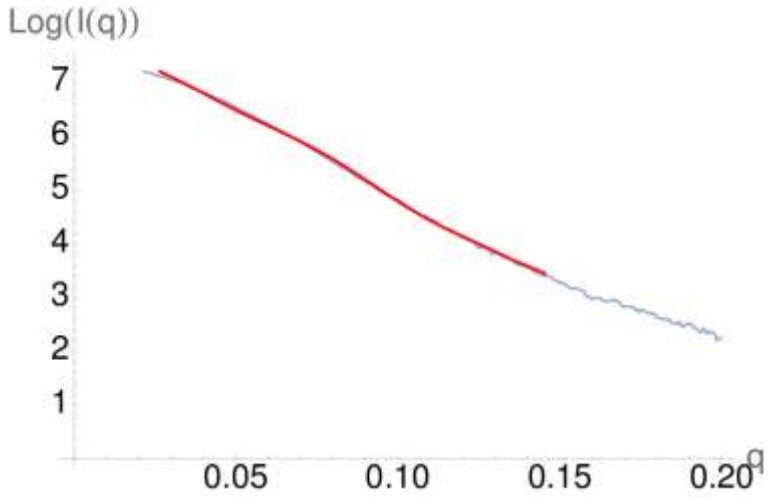


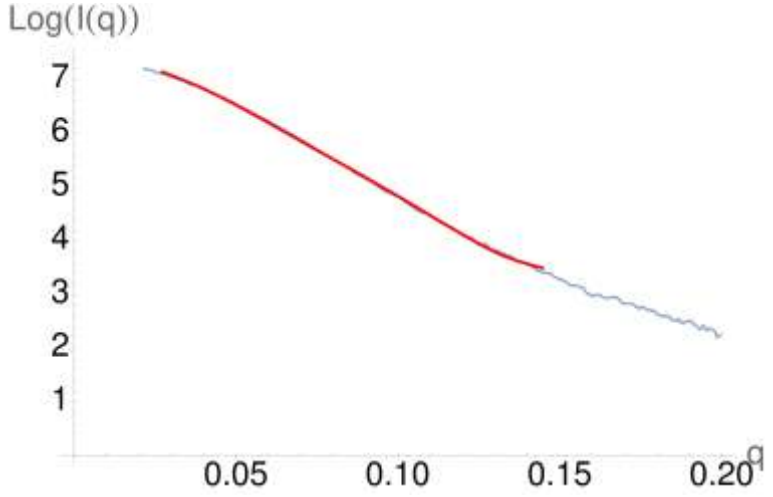


# So we're done?





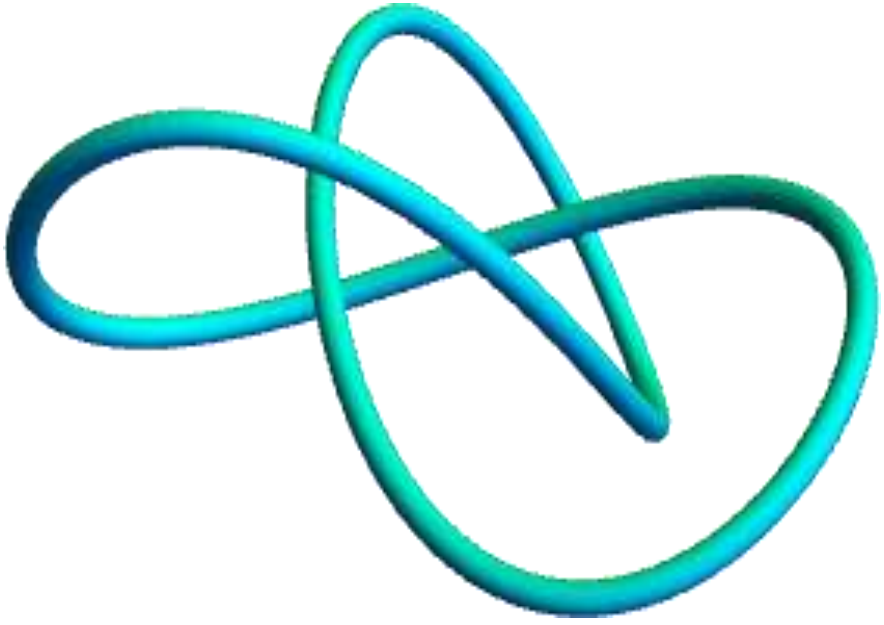


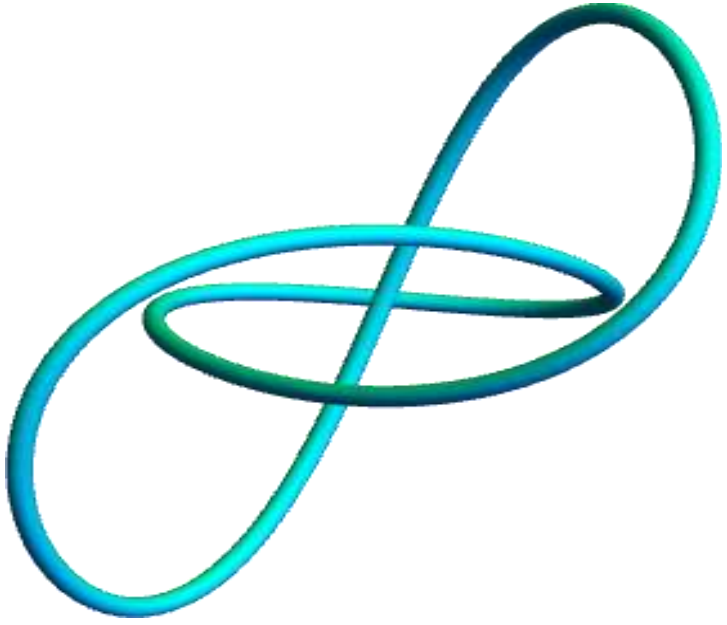

We need a simple metric that can tell us when the protein is realistically folded

# Average Crossing Number (ACN)

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



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

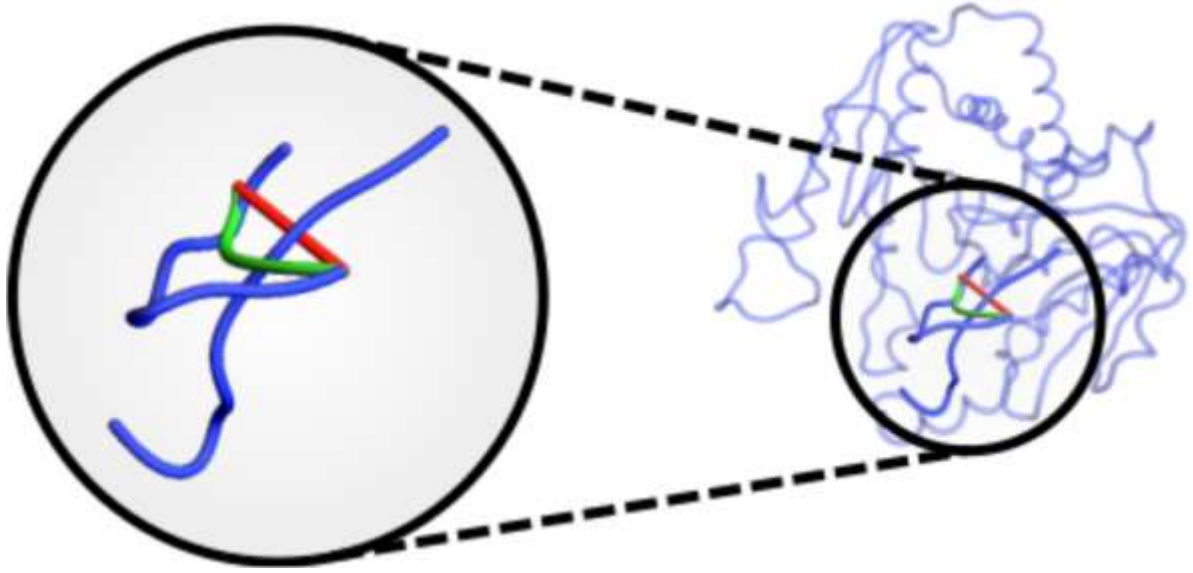


The ACN is thus an average of the crossings over all projections of a curve.

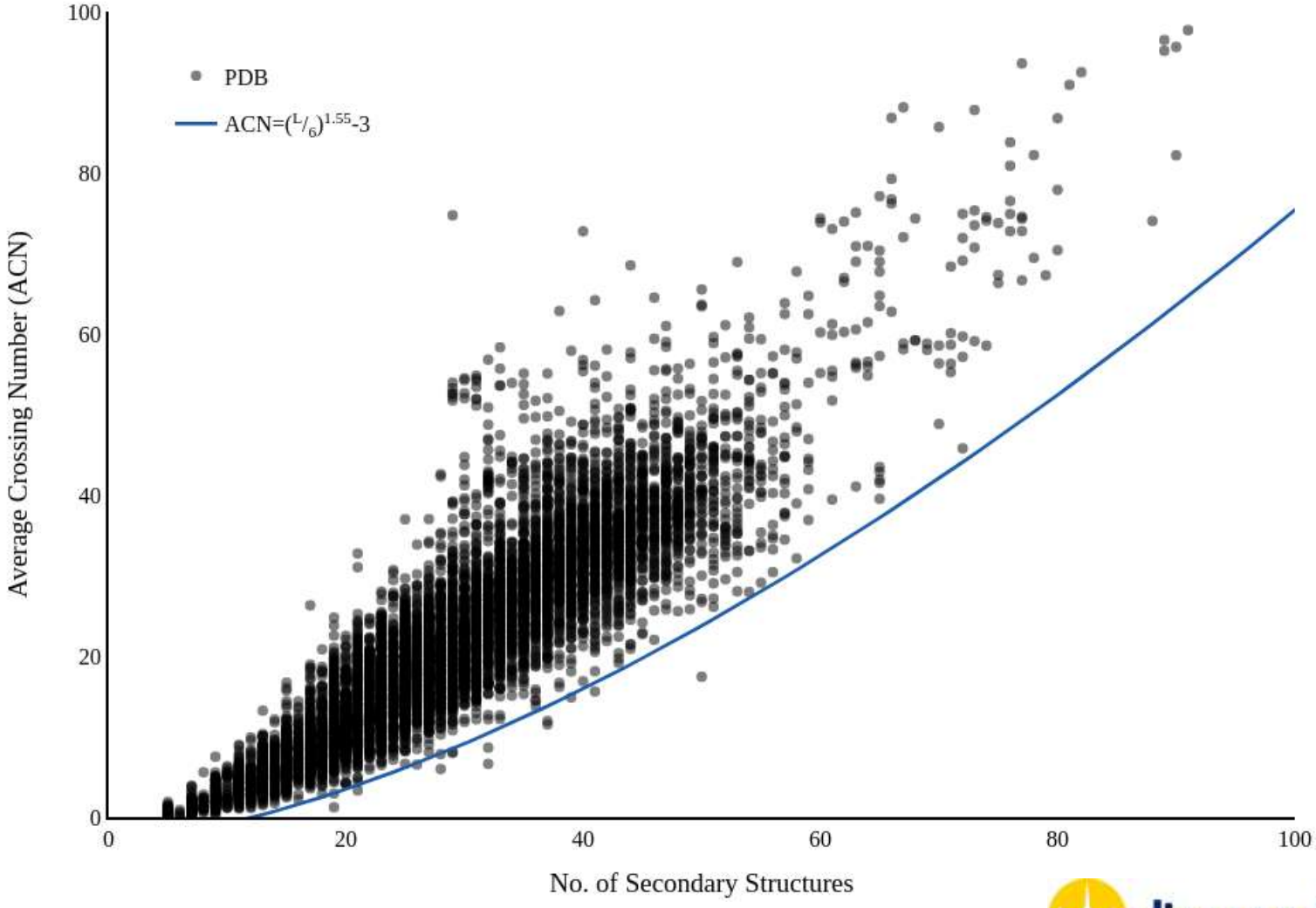
*SWRITHE: a python package to identify surprising protein tertiary structure relationships and aid new structure search.* In preparation

# Smoothing the backbone

- Secondary structure introduces lots of noise to the ACN.
- We need to smooth out the backbone, whilst preserving the underlying tertiary fold
- To do this, replace each alpha helix and beta strand with a straight edge.
- For linkers, we have developed an algorithm to reduce it to a minimal representation which preserves essential crossings.

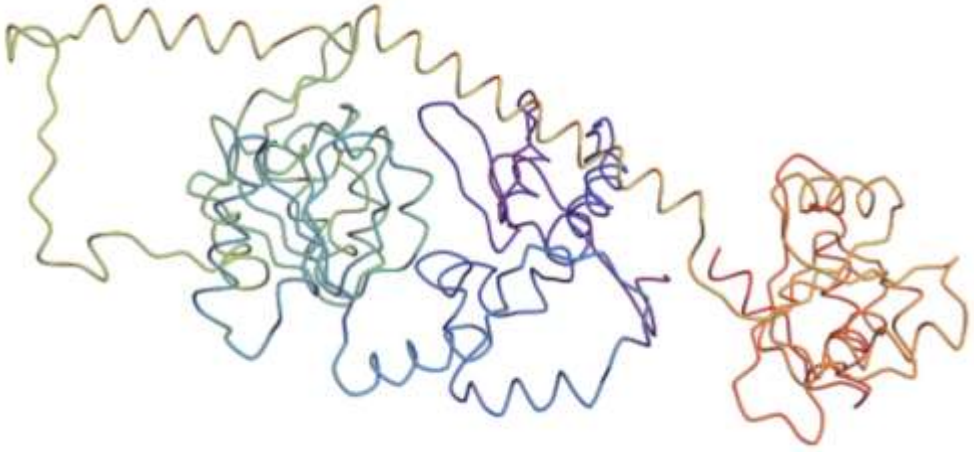


# Bounding the ACN for proteins

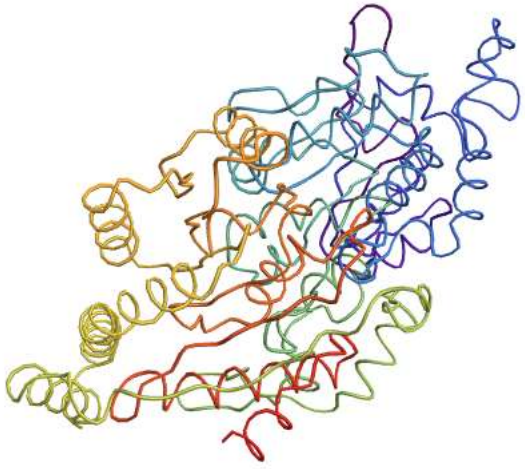



- 99.86% of the data lies above the blue bounding curve.
- Those that do lie below are the relatively trivially folded single unit of a multimer structure.
- This lower bound gives us a way to quantify how realistically folded our predictions are.

# Putting this all together



ACN = 48.3



ACN = 59.2

By including a penalty for structures whose ACN falls below the blue bounding curve, we can restrict the search space to globally realistic folds. The ACN for a protein of this length should be at least 56.

In doing so, we speed up the fitting process, and ensure that structures we pass into MD simulations have a fighting chance of survival

# The Menu

MoSMed
Molecular Sciences for Medicine

SAXS Data Collection

diamond

Initial Predictions

Carbonara

Molecular Dynamics

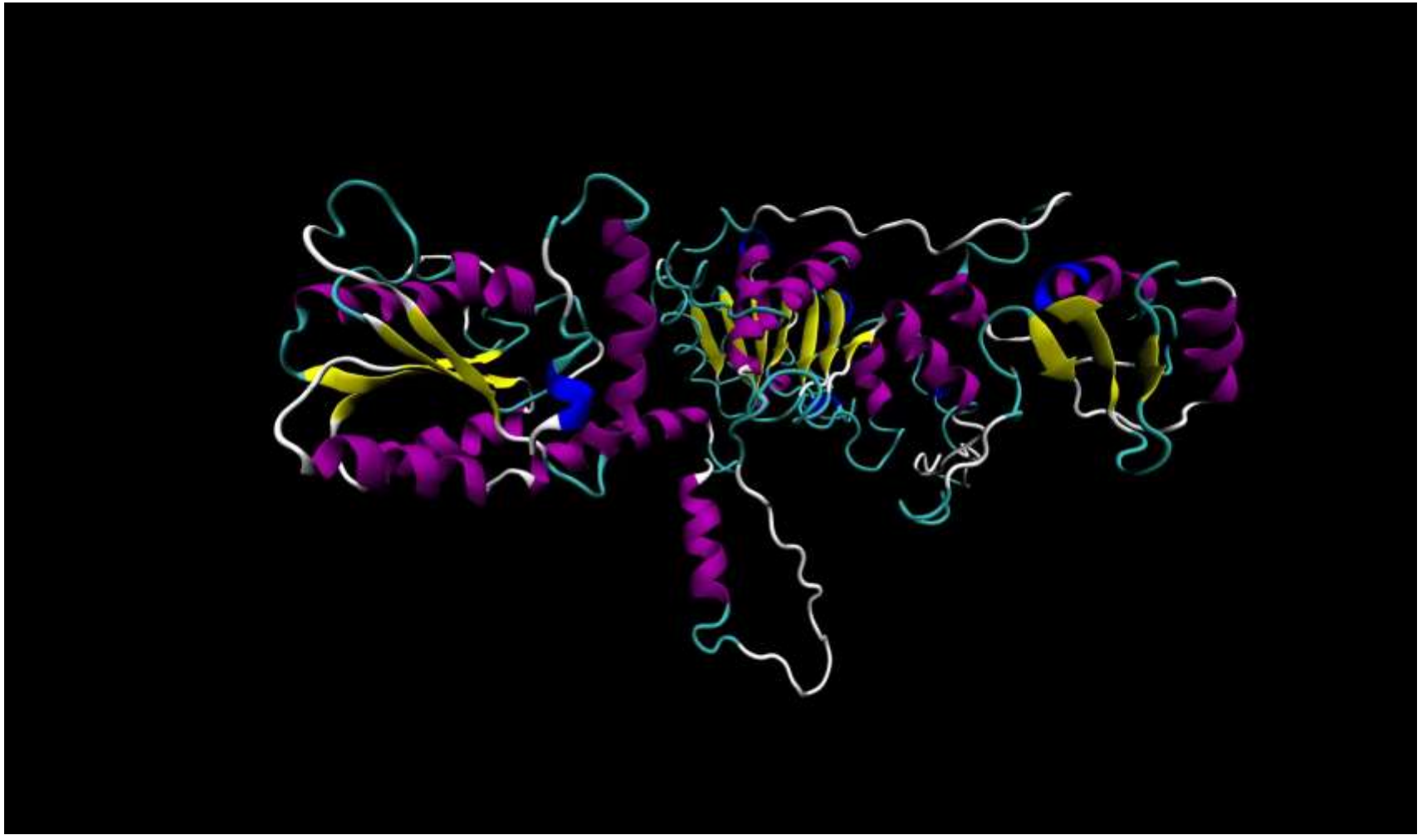
diamond

Durham University

Newcastle University

UKRI Engineering and Physical Sciences Research Council

# MD Simulations



MoSMed
Molecular Sciences for Medicine

diamond
Durham University
Newcastle University
UKRI Engineering and Physical Sciences Research Council

# Conclusions and Future Work

- Carbonara is able to produce novel protein structural predictions which fit solution scattering data.
- The ACN of proteins is bounded from above and below, providing a guideline for the entanglement of tertiary structures.
- By including the ACN as penalty in the fitting procedure, we can produce realistic structural predictions which behave well in MD simulations

- Finalise production of Carbonara for a full release, including a loop modelling package.
- Using variants of the ACN, develop a structural database for cross referencing predictions.

# Acknowledgements


- My supervisors, Chris Prior, Ehmke Pohl, and Rob Rambo
- Collaborators at Diamond, Nathan Cowieson and Tim Snow
- The MoSMed CDT
- My funders EPSRC and Diamond

# Conclusions and Future Work

- Carbonara is able to produce novel protein structural predictions which fit solution scattering data.
- The ACN of proteins is bounded from above and below, providing a guideline for the entanglement of tertiary structures.
- By including the ACN as penalty in the fitting procedure, we can produce realistic structural predictions which behave well in MD simulations

- Finalise production of Carbonara for a full release, including a loop modelling package.
- Using variants of the ACN, develop a structural database for cross referencing predictions.