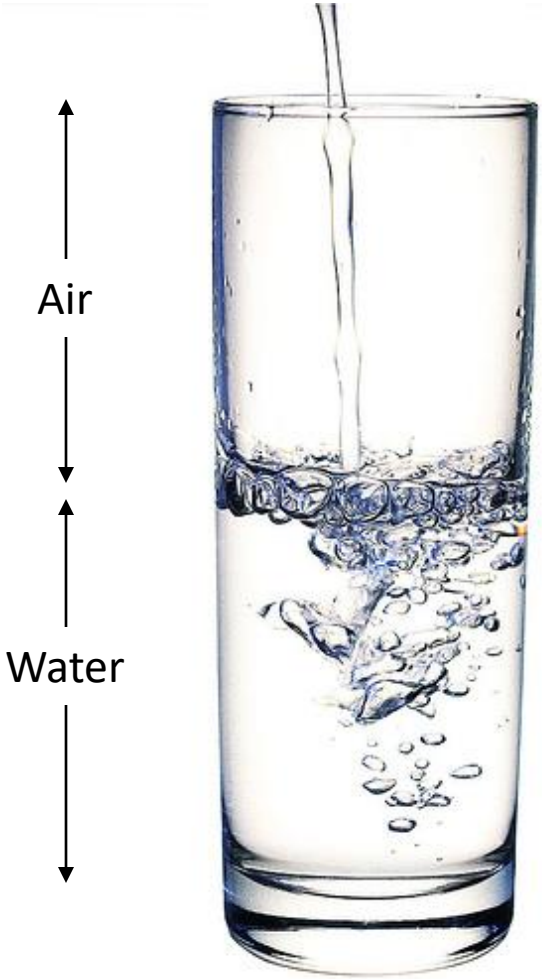
Small Angle X-ray Scattering as a Complementary Tool in the Structural Biology Laboratory: A case study with tRNA synthetase

Edward H. Snell

Hauptman Woodward Medical Research Institute, Buffalo, NY 14203, USA.

Why SAXS?

Pessimists, Optimists, and Crystallographers



Consider a glass of water

Pessimist
(the glass is half empty)

Optimist
(the glass is half full)

Crystallographer
(the glass is completely full)

Fantasy

Acta Crystallographica Section F
Structural Biology and Crystallization Communications
ISSN 1744-3091

On the need for an international effort to capture, share and use crystallization screening data

Janet Newman,[a]* Evan E. Bolton,[b] Jochen Müller-Dieckmann,[c] Vincent J. Fazio,[a] Travis Gallagher,[d] David Lovell,[e] Joseph R. Luft,[f,g] Thomas S. Peat,[a] David Ratcliffe,[e] Roger A. Sayle,[h] Edward H. Snell,[f,g] Kerry Taylor,[e] Pascal Vallotton,[i] Sameer Velanker[j] and Frank von Delft[k]

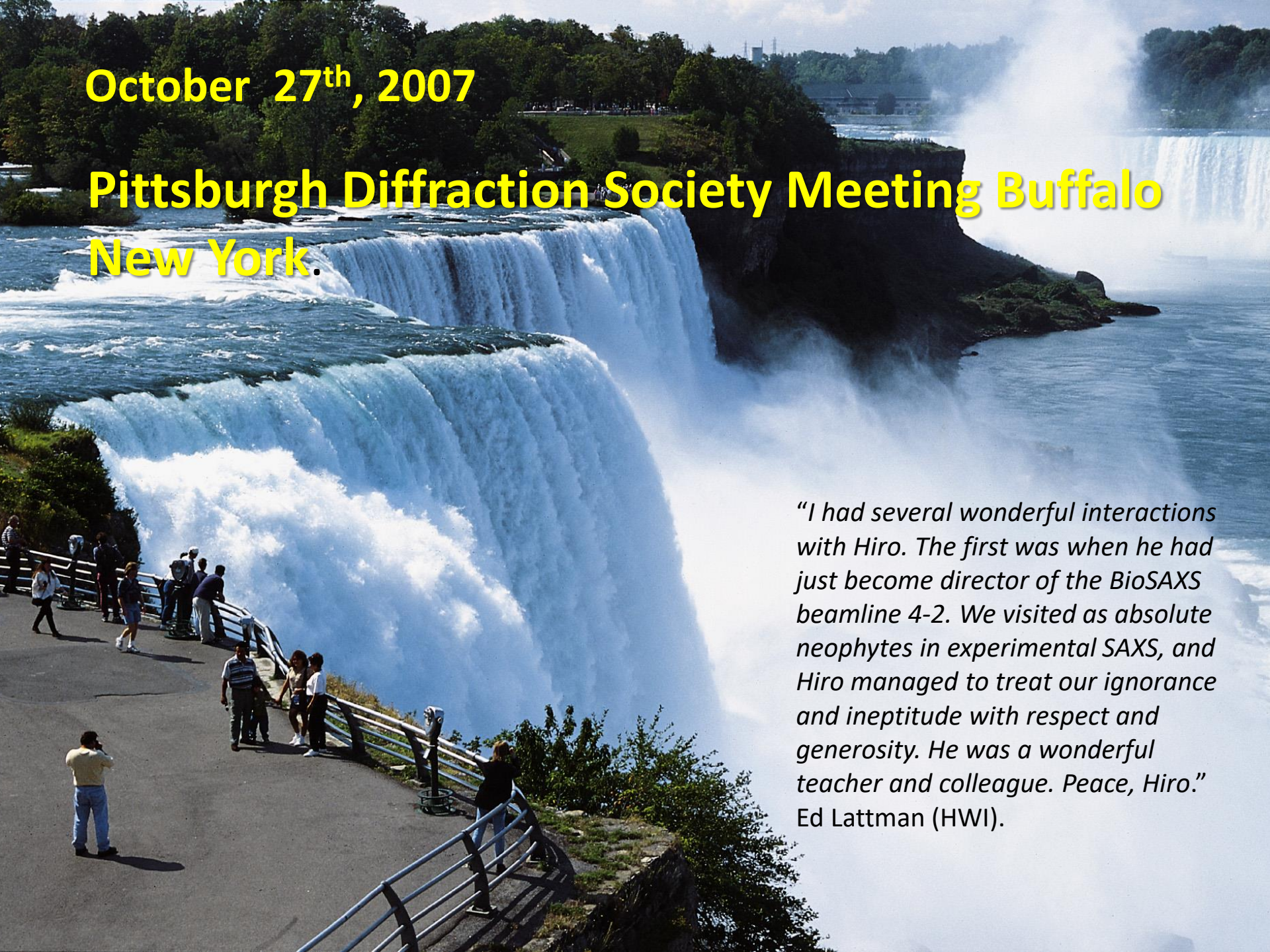
[a]Materials Science and Engineering, CSIRO, 343 Royal Parade, Parkville, VIC 3052, Australia, [b]NCBI, NLM, NIH, Department of Health and Human Services, 8600 Rockville Pike, Bethesda, MD 20894, USA, [c]EMBL Hamburg Outstation c/o DESY, Notkestrasse 85, D-22603 Hamburg, Germany, [d]National Institute for Standards and

When crystallization screening is conducted many outcomes are observed but typically the only trial recorded in the literature is the condition that yielded the crystal(s) used for subsequent diffraction studies. The initial hit that was optimized and the results of all the other trials are lost. These missing results contain information that would be useful for an improved general understanding of crystallization. This paper provides a report of a crystallization data exchange (XDX) workshop organized by several international large-scale crystallization screening laboratories to discuss how this information may be captured and utilized. A group that administers a significant fraction of the world's crystallization screening results was convened, together with chemical and structural data informaticians and computational scientists who specialize in creating and analysing large disparate data sets. The development of a crystallization ontology for the crystallization community was proposed. This paper (by the attendees of the workshop) provides the thoughts and rationale leading to this conclusion. This is brought to the attention of the wider audience of crystallographers so that they are aware of these early efforts and can contribute to the process going forward.

Only approximately 11% of the proteins we target for crystallography yield a crystallographic structure.

At least 99.8% of crystallization experiments produce an outcome other than crystallization.

There exists a large quantity of soluble purified protein that remains structurally uncharacterized.



October 27th, 2007

Pittsburgh Diffraction Society Meeting Buffalo New York.

"I had several wonderful interactions with Hiro. The first was when he had just become director of the BioSAXS beamline 4-2. We visited as absolute neophytes in experimental SAXS, and Hiro managed to treat our ignorance and ineptitude with respect and generosity. He was a wonderful teacher and colleague. Peace, Hiro."
Ed Lattman (HWI).

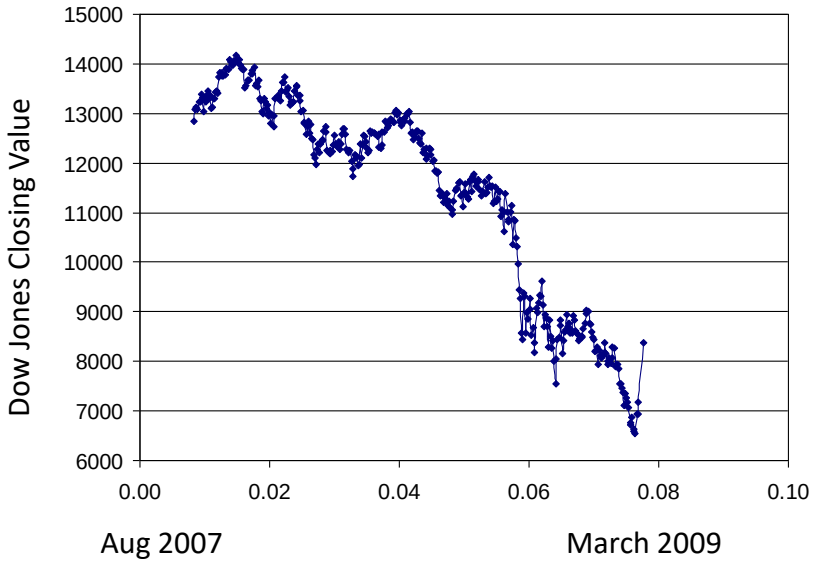
High throughput SAXS

- At our high-throughput crystallization facility we have run ~12,500 different proteins.
- Crystals result in about 50% of cases.
- Where we track results (PSI samples) about 50% of samples that give crystals go on to a PDB deposition.

Frustration

- All our samples are in solution.
- So … since meeting Hiro in 2007 we have been developing high-throughput strategies to take the remaining dregs of crystallization samples from NESG (~60 microL) and gathering SAXS data.
- To date, SAXS data from 800 different proteins (3 concentrations each)

Can we use X-ray solution scattering?

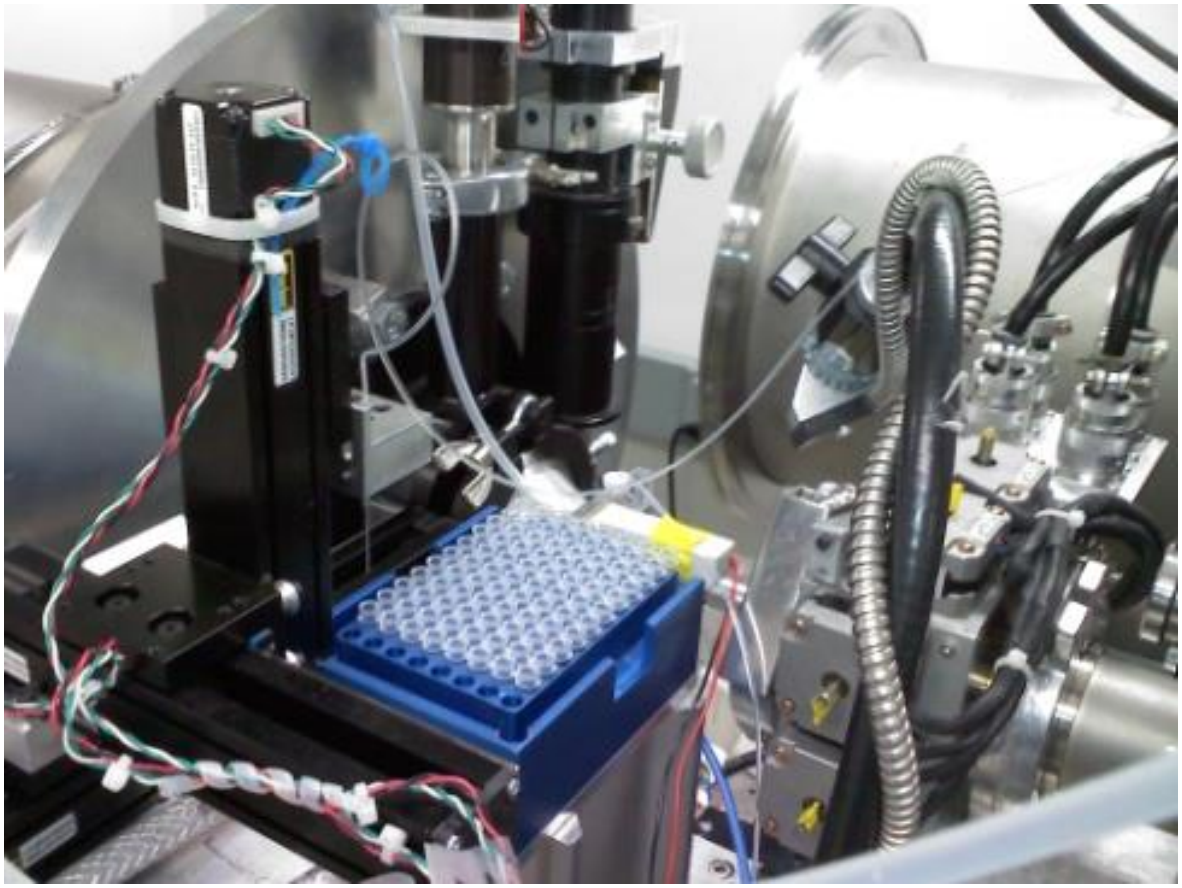


The scattering data from SAXS provides a 1D Fourier transform of the envelope of the particle.

It's possible to fit multiple envelopes to the data.

You will always get an envelope despite the data!

High-throughput SAXS



High throughput protocol

Up to 12 different PCR strips.

3-7 different concentrations per sample.

For high-throughput studies, 2 samples per strip, 24 samples in total

Start with buffer then lowest concentration first. End with buffer

8 exposures, 1-2s each dependent on sample molecular weight, buffer and concentration.

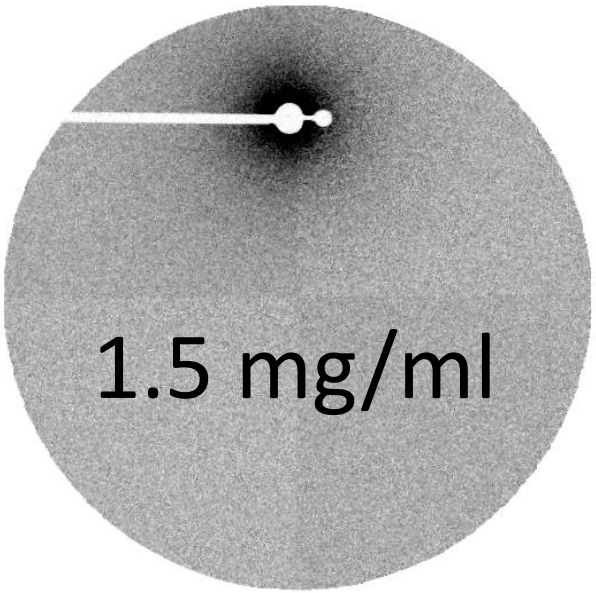
Oscillate sample to minimize radiation damage

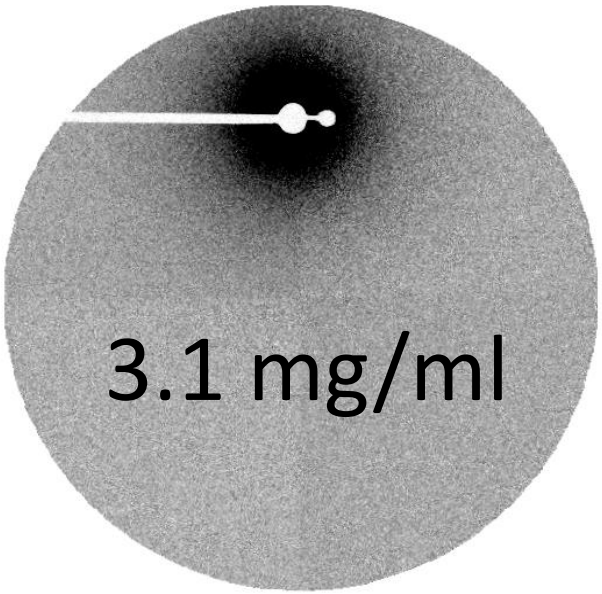
Repeat the buffer.

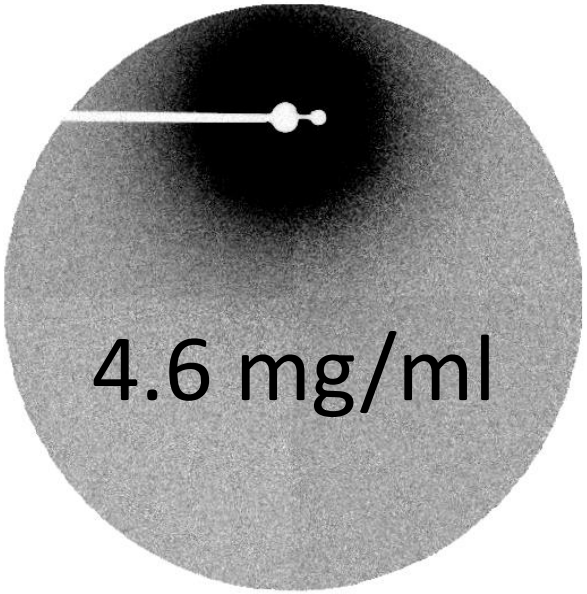
Load next sample

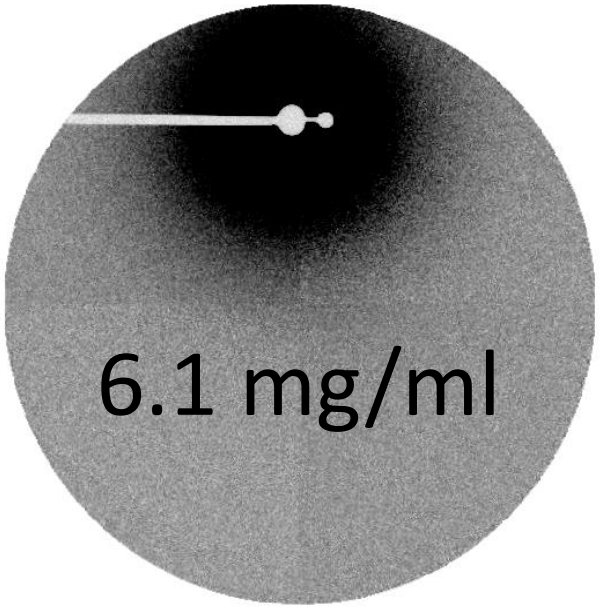
Time per concentration series – approximately 10 to 15 minutes. In high-throughput mode 24 samples in 3 to 4 hours.

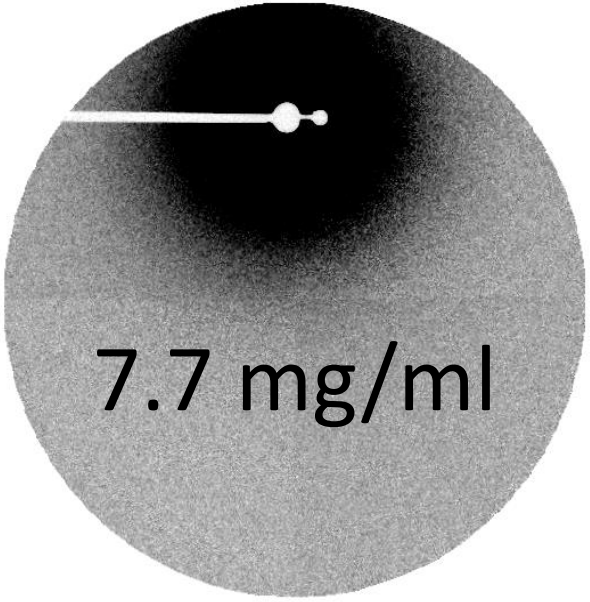
Enables two important things – eat and sleep!


1.5 mg/ml

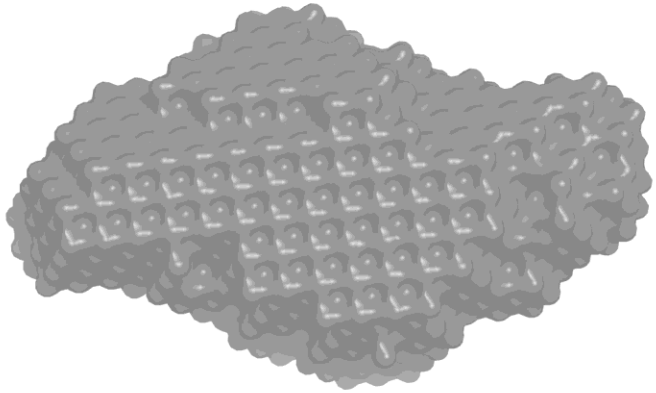

3.1 mg/ml


4.6 mg/ml

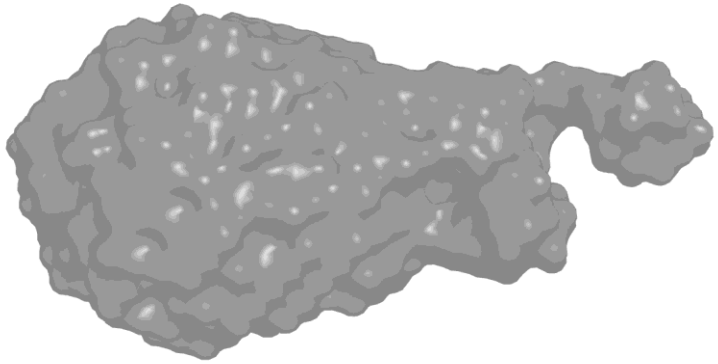

6.1 mg/ml


7.7 mg/ml

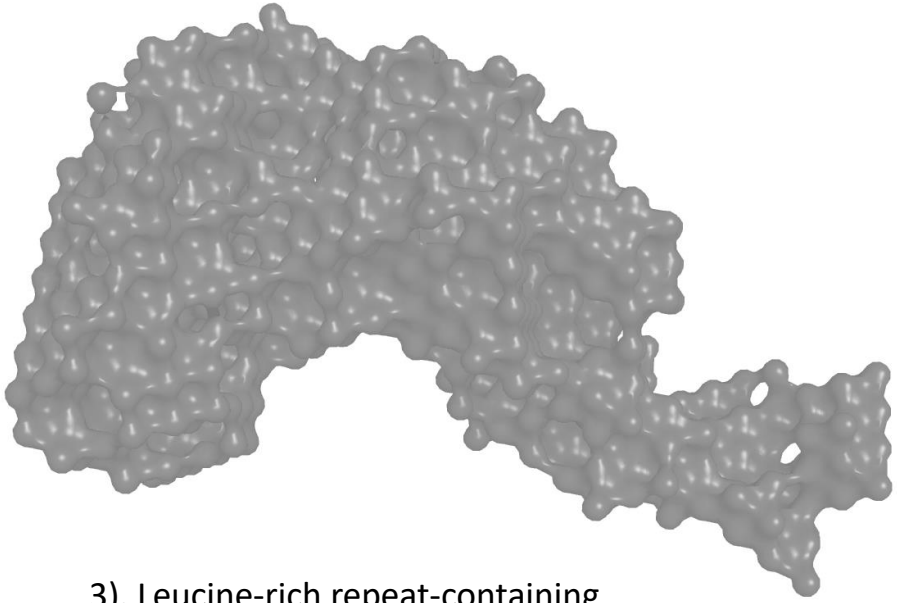
Ab intio envelopes



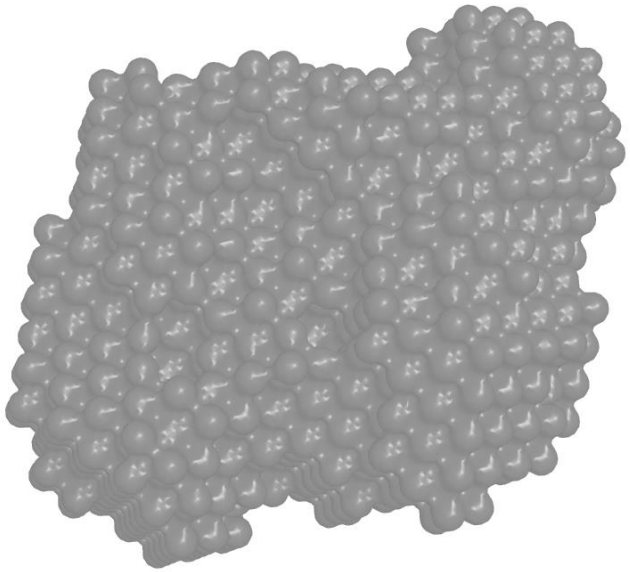
1). alr0221 protein from Nostoc (18.6 kDa)



2). C-terminal domain of a chitobiase (17.9 kDa)

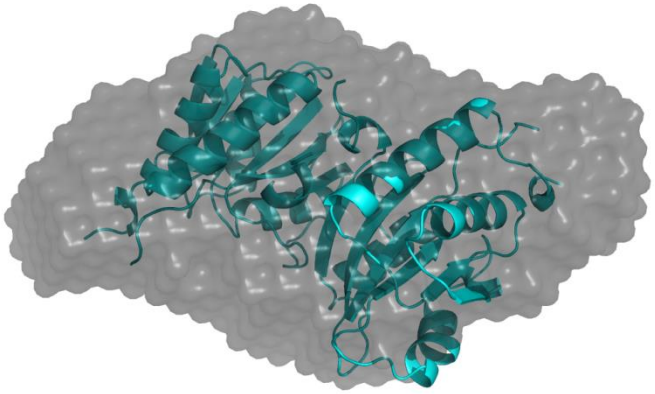


3). Leucine-rich repeat-containing protein LegL7 (39 kDa)

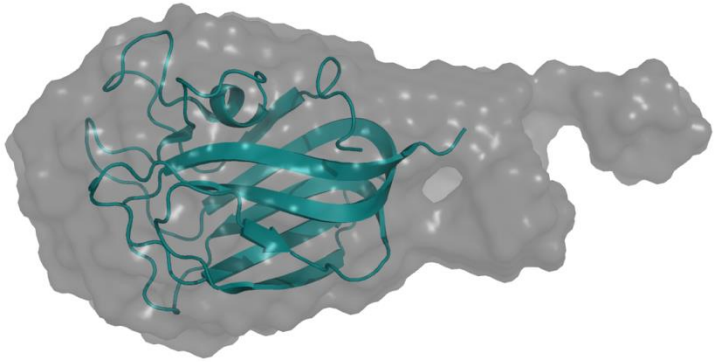


4). *E. Coli.* Cystine desulfurase activator complex (170 kDa)

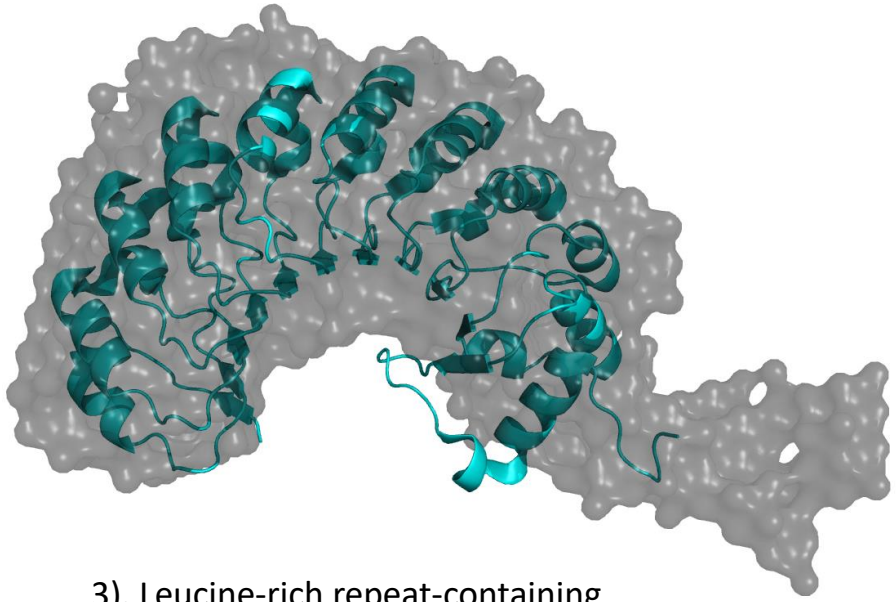
Overlaid with subsequent X-ray structures



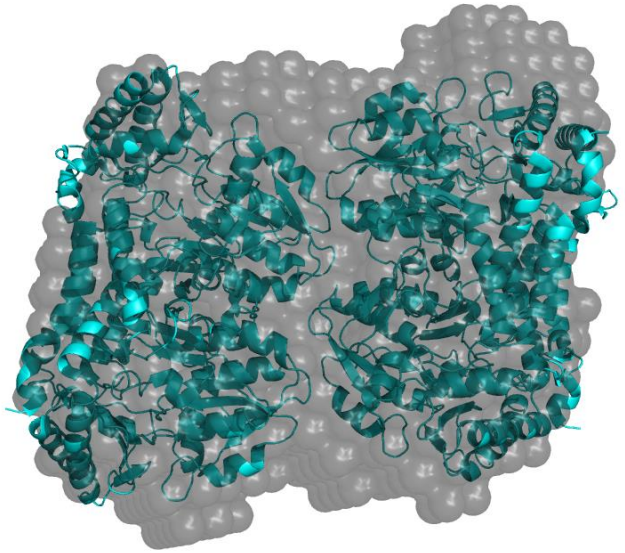
1). alr0221 protein from Nostoc (18.6 kDa)



2). C-terminal domain of a chitobiase (17.9 kDa)

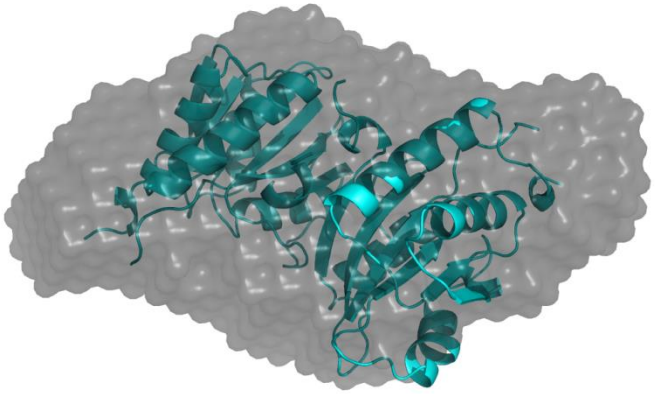


3). Leucine-rich repeat-containing protein LegL7 (39 kDa)

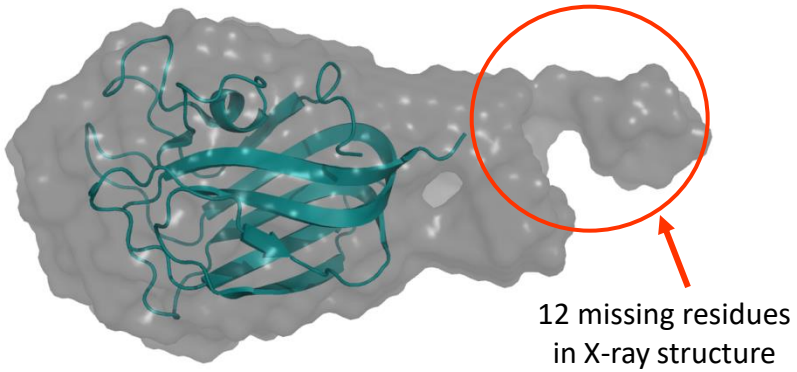


4). *E. Coli.* Cystine desulfurase activator complex (170 kDa)

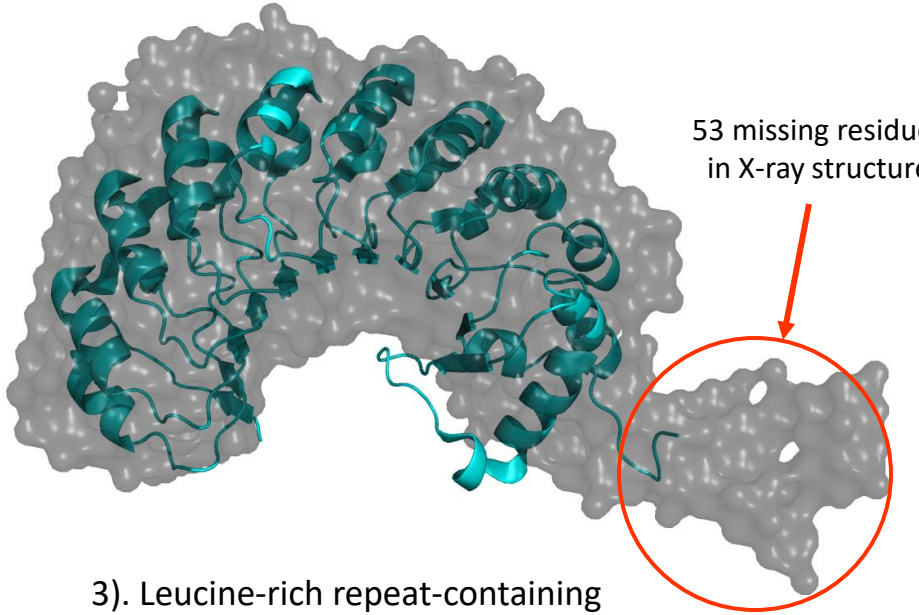
And data on what was missing …



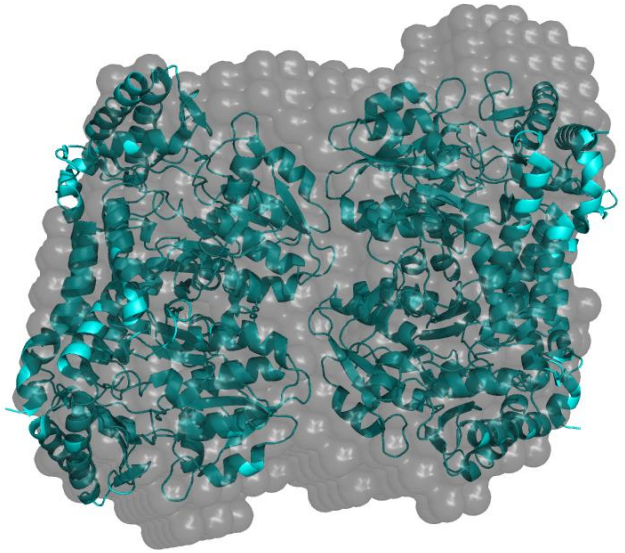
1). alr0221 protein from Nostoc (18.6 kDa)



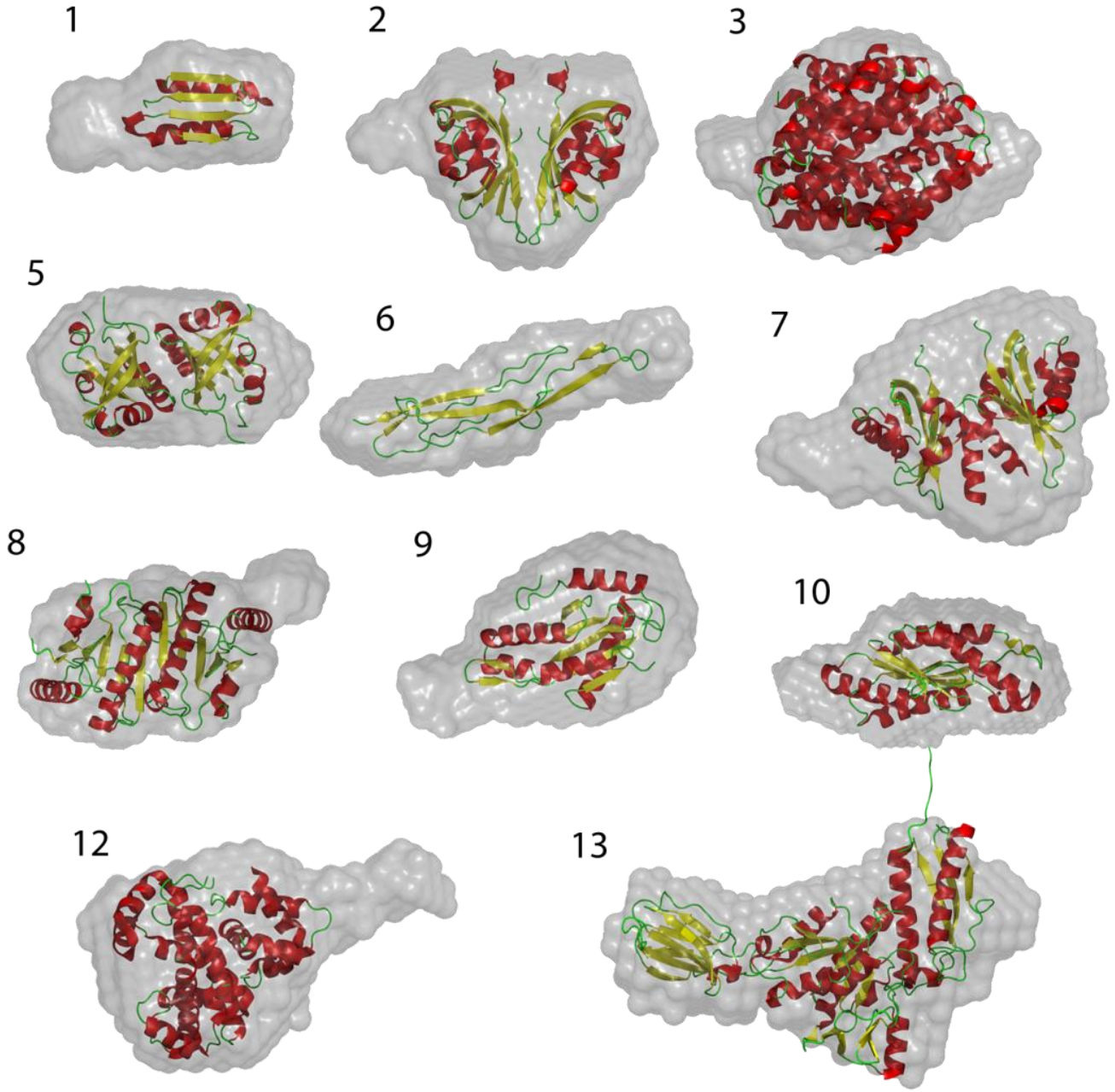
2). C-terminal domain of a chitobiase (17.9 kDa)



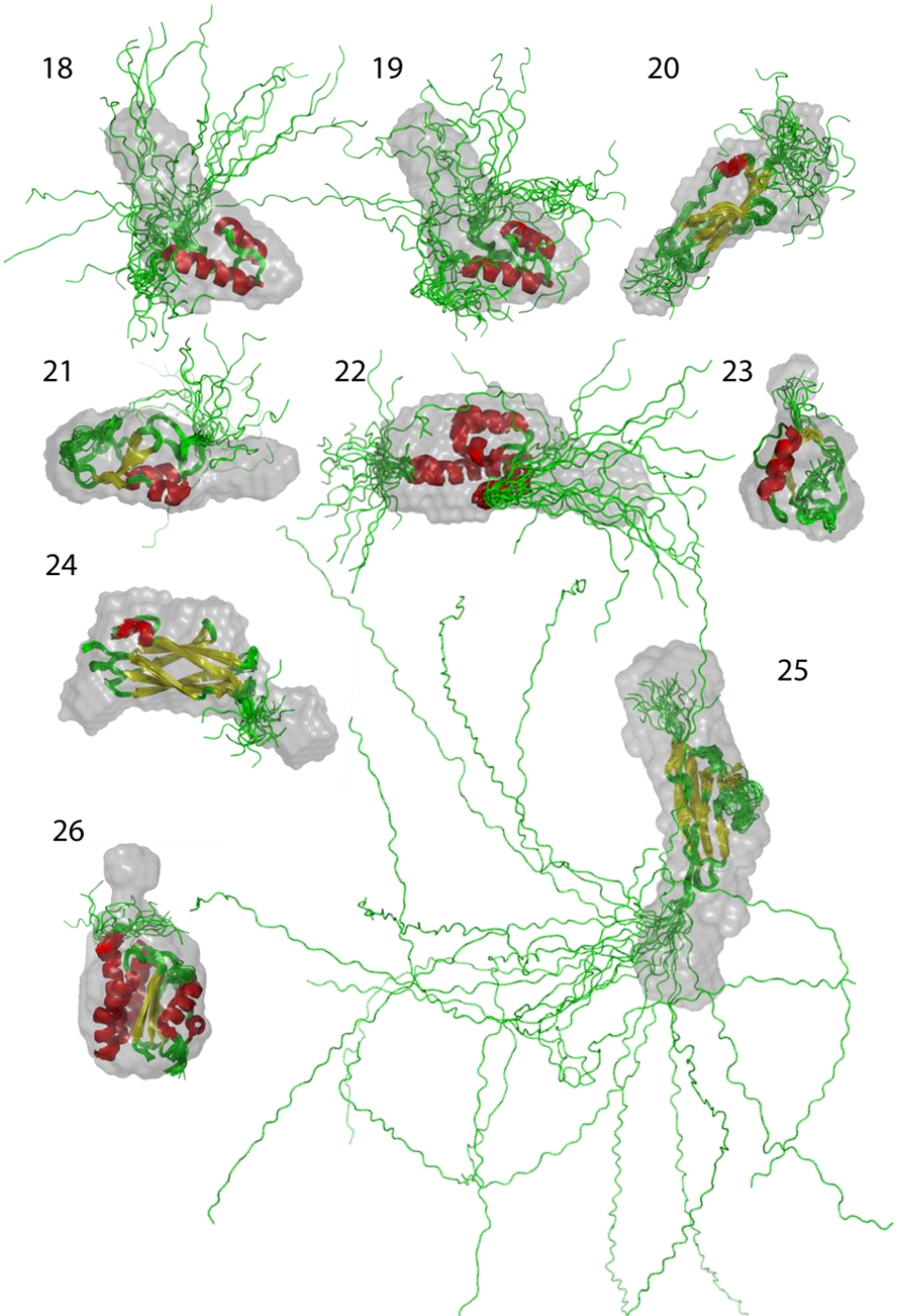
3). Leucine-rich repeat-containing protein LegL7 (39 kDa)



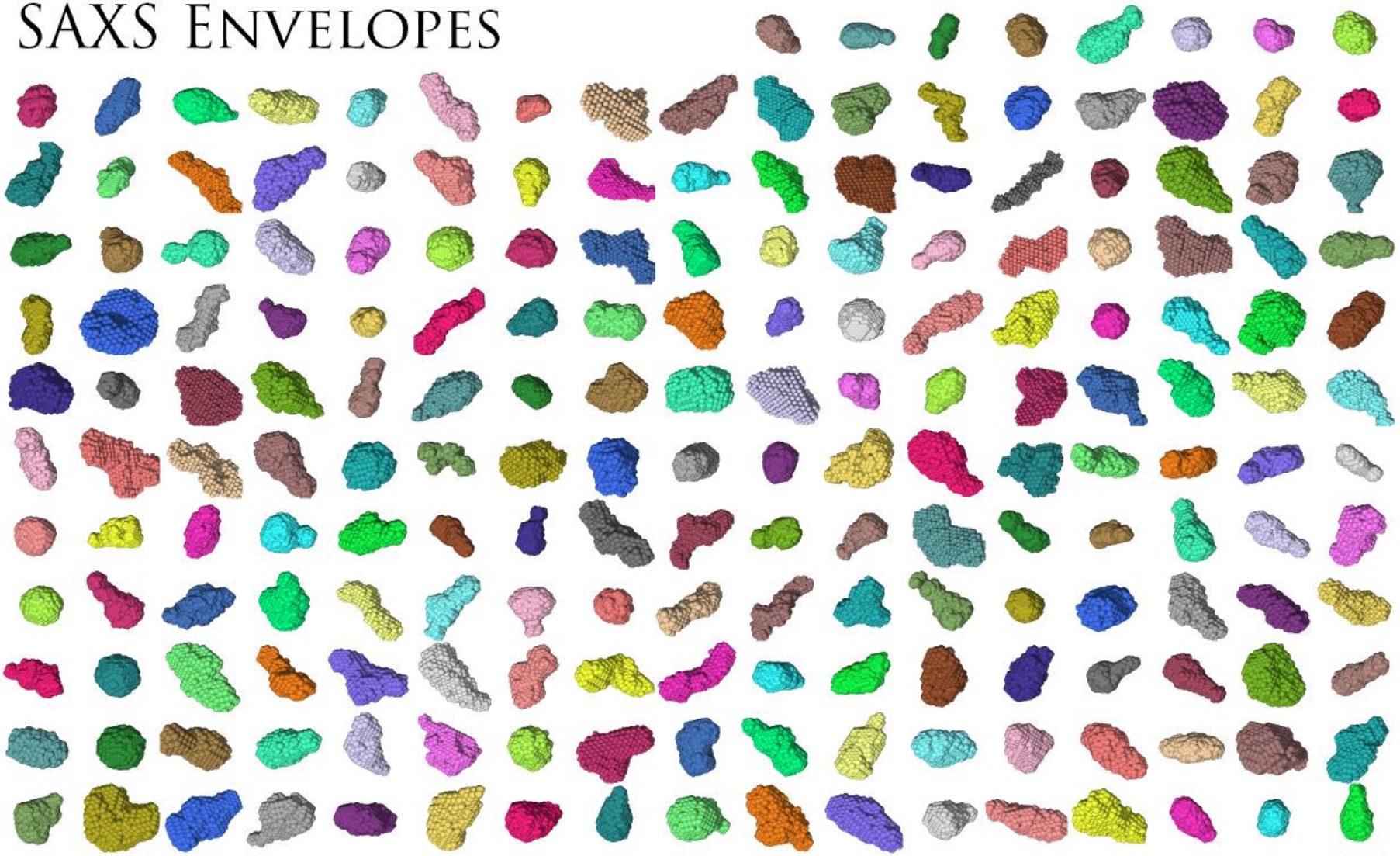
4). *E. Coli.* Cystine desulfurase activator complex (170 kDa)



Comparing X-ray structures



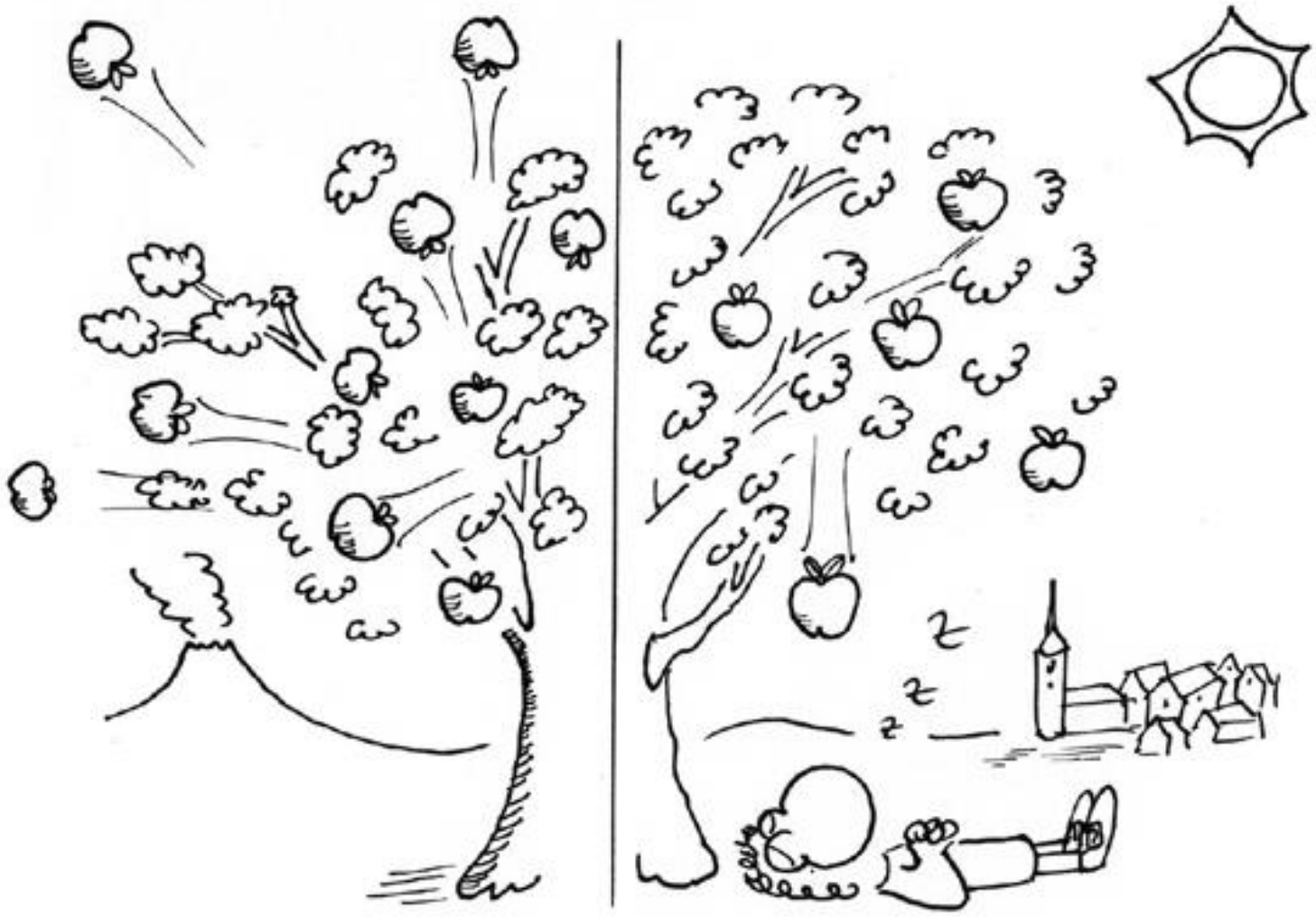
Comparing NMR structures



SAXS : the T-shirt (Tom Grant LLC)

A Biological Puzzle

PHYSICS FOR BIOLOGISTS

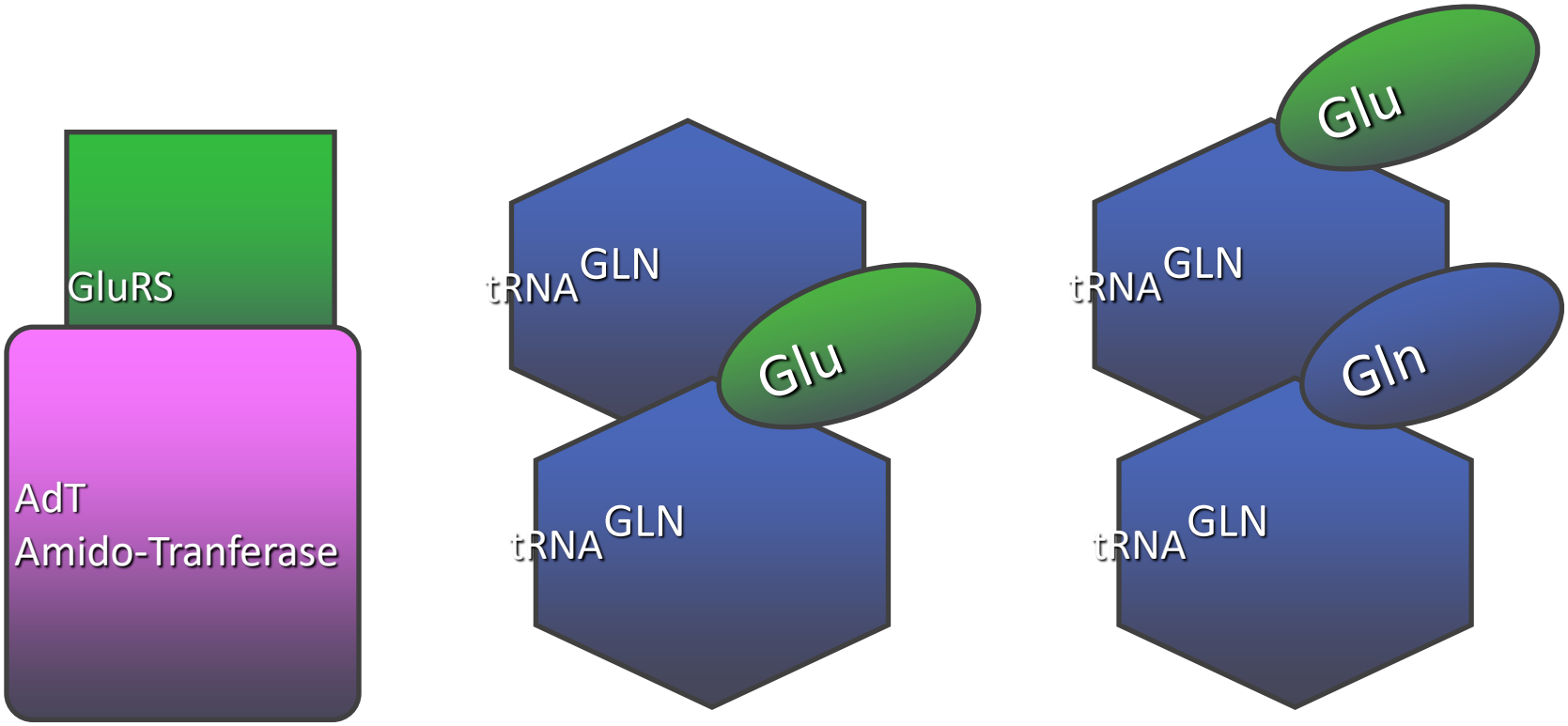


A long time ago the apple trees used to shoot the apples in all directions. Only those that did it downward got reproduced. Then, after millions years of natural selection and evolution, gravity was finally discovered.

tRNA Synthetases

Two routes of gln-tRNA^{GLN} Formation

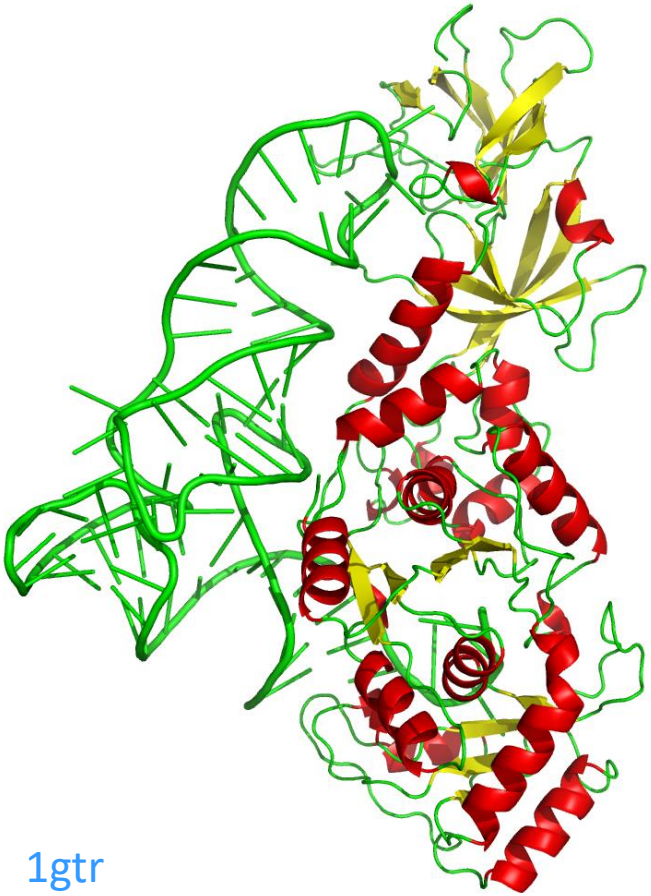
Two routes of gln-tRNA^{GLN} Formation



tRNA synthetase of Eukaryotes and Prokaryotes

Appended Domains

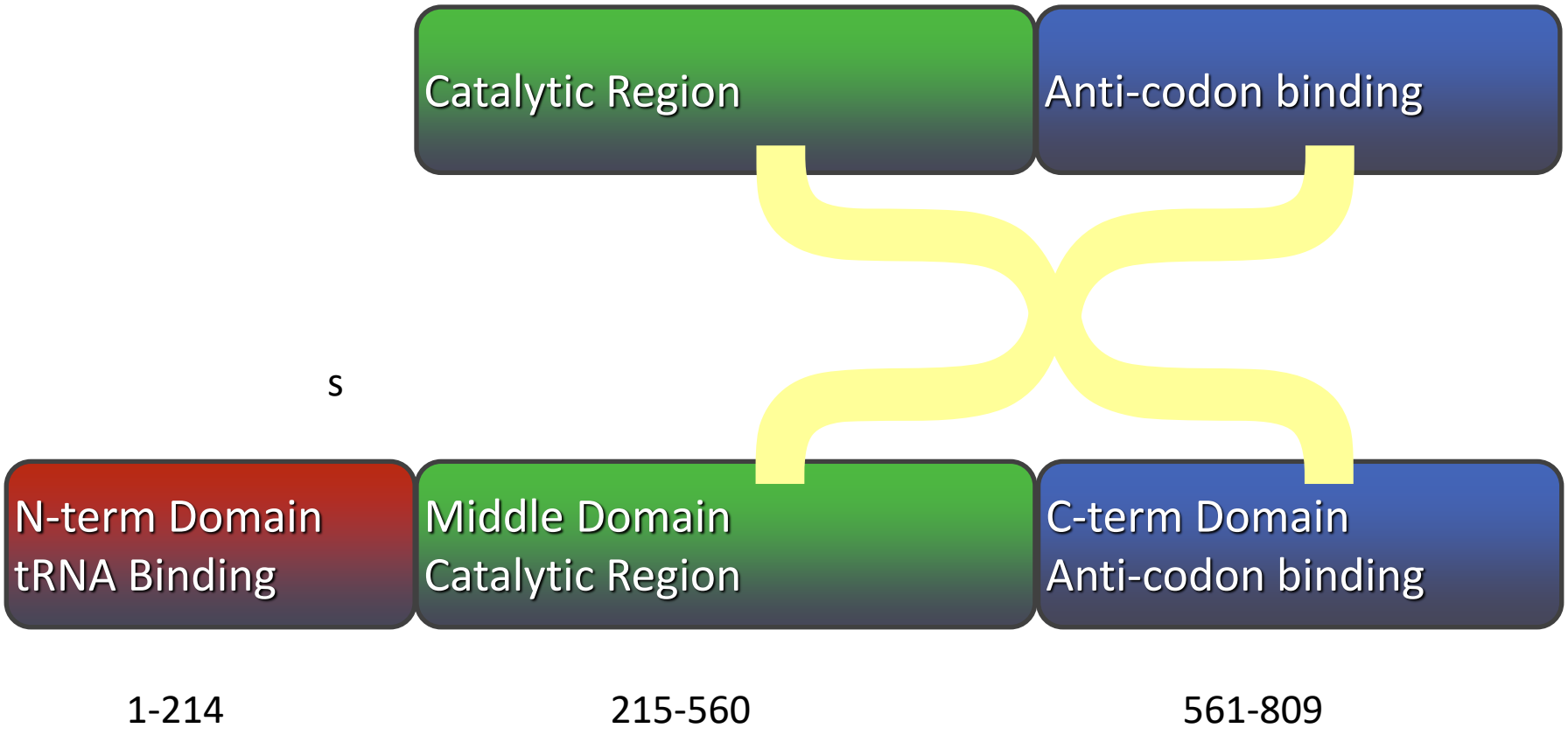
Structural model of *E. coli* glutaminyl-tRNA synthetase



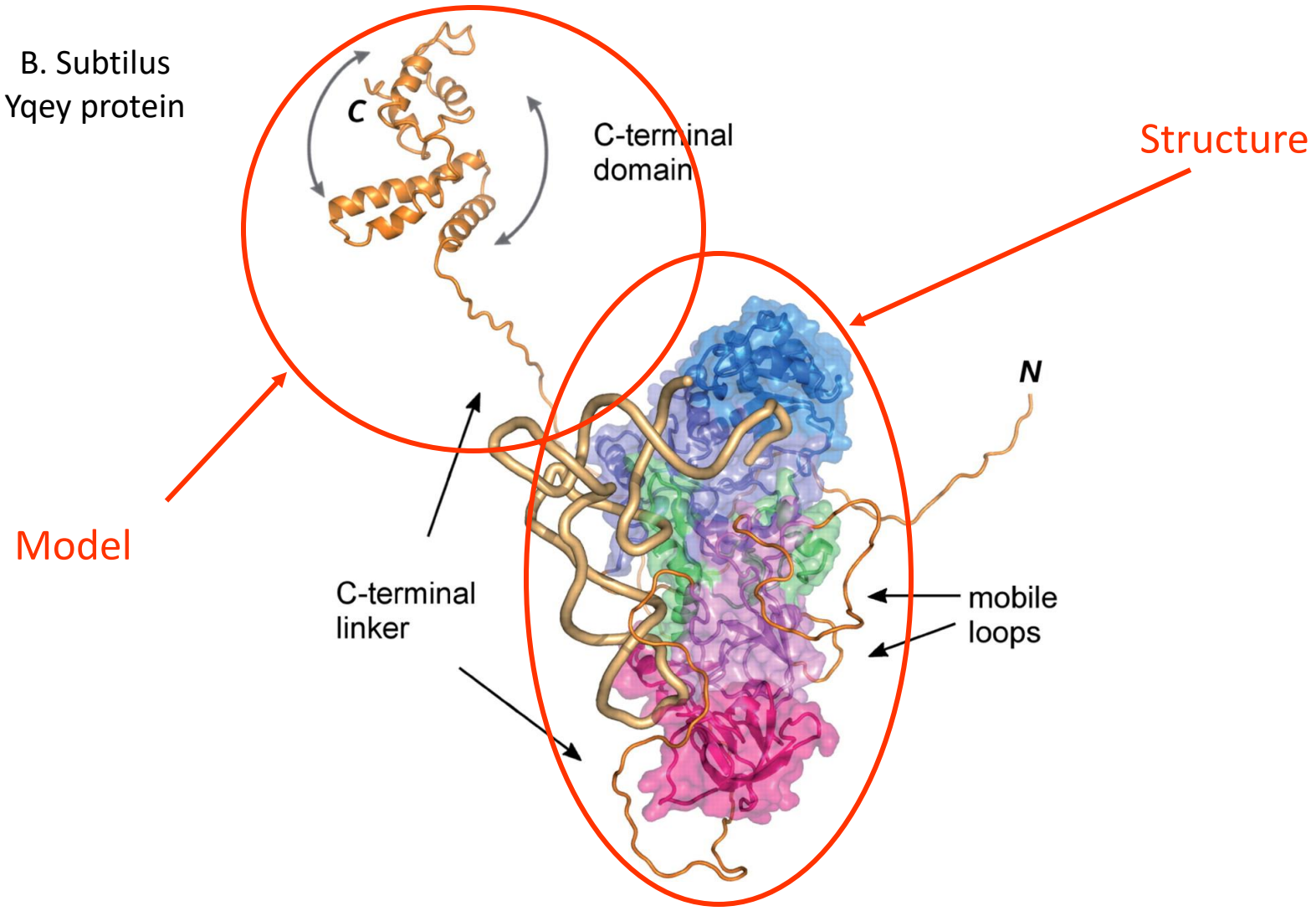
1gtr

Structures only known from *E.coli* and *D. radiodurans*

Glutamine tRNA Synthetase



Model of *D. radiodurans* GlnRStRNA^{Gln} complex



Target

- Our target today is Glutaminyl tRNA synthetase (Gln4) from yeast *Saccharomyces cerevisiae*
- Yeast *Saccharomyces cerevisiae is* a well-established model system for understanding fundamental cellular processes of higher eukaryotic organisms.

- Many eukaryotic tRNA synthetases like Gln4 differ from their prokaryotic homologs by the attachment of an additional domain appended to their N or C-terminus, but it is unknown how these domains contribute to tRNA synthetase function, and why they are not found in prokaryotes

- The 228 amino acid N-terminal domain of Gln4 is among the best studied of these domains, but is structurally uncharacterized.

- The N-terminal domain appears to have non specific RNA binding.

- The role of a nonspecific RNA binding domain in the function of a highly specific RNA binding enzyme is baffling, but clearly crucial given its prevalence among tRNA

Crystallization/Data collection

- Gln4 Screened against 1536 different biochemical conditions, ~1000 forming an incomplete factorial of chemical space and ~500 representing commercially available screens.
- Crystal leads seen, several were chosen based on ease of cryoprotection of the native hit.
- Crystals were optimized with a Drop Volume Ratio versus Temperature (DVR/T) technique.
- Cryoprotected and 'drop' shipped to SSRL by FedEx.

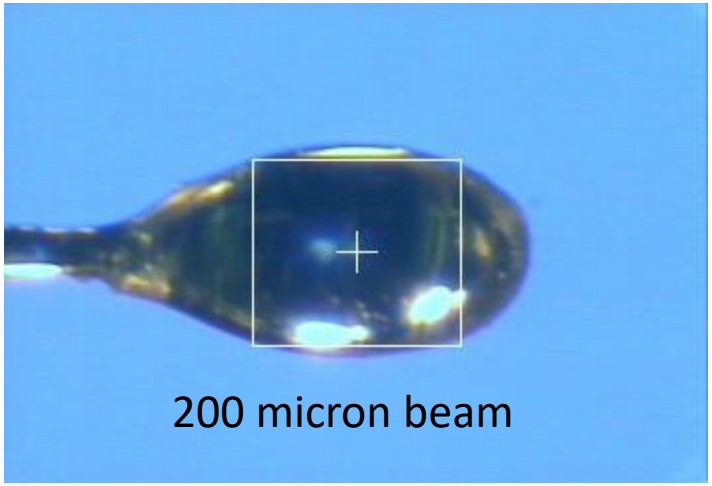
- Only 2 structures for related glutaminyl tRNA synthetases are available (~40% sequence homology), we had 228 extra residues (almost 40% more residues) therefore we expected problems in molecular replacement and didn't have a SeMet example.

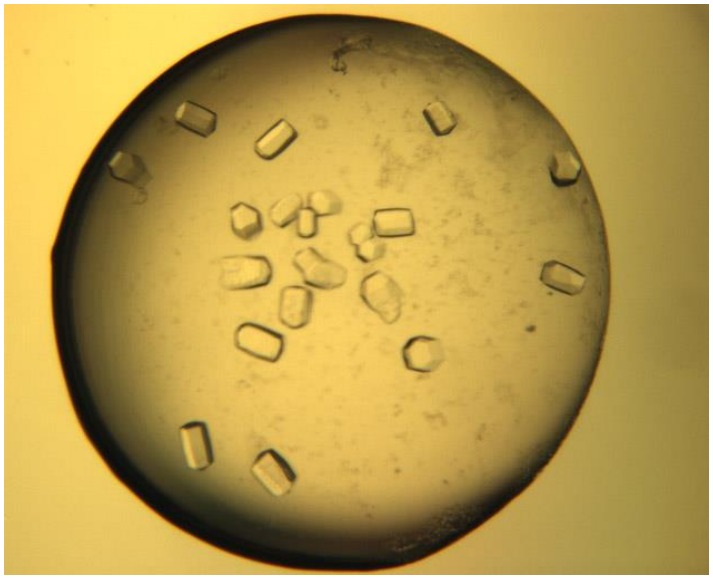
- EXAFS data indicate Zinc present in the *E. coli.* Case (not seen in the X-ray structure). The zinc acts to stabilize the structure in a pseudo zinc finger motif.

- We collected data remotely with an excitation scan to determine if Zinc was present.

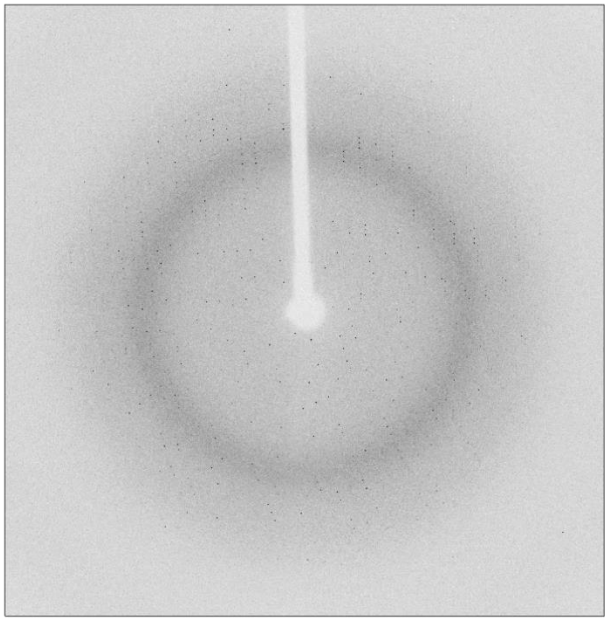
- It was!

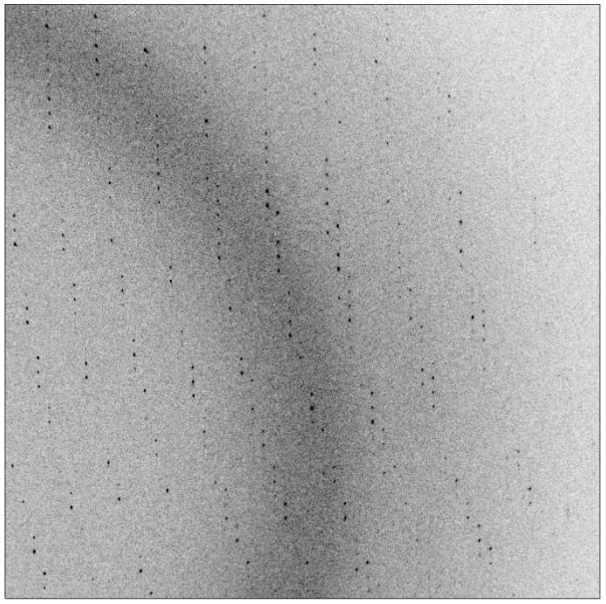
Crystallography





80% PEG 400 in the
crystallization cocktail

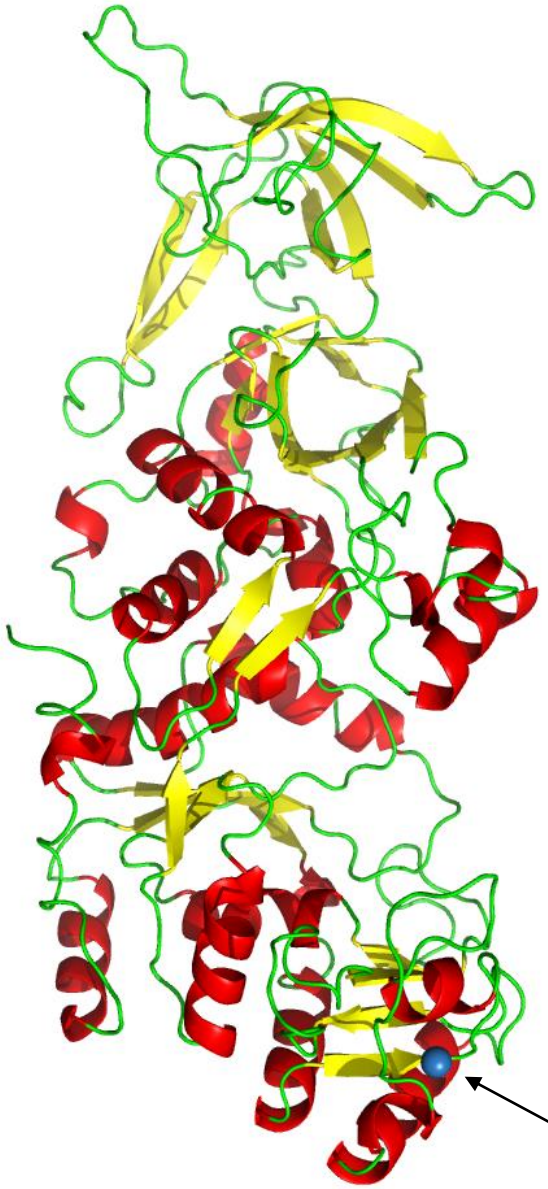




Data collection/Processing

- We used beamline 11-1 at SSRL with a Mar 325 CCD detector, 340 mm crystal to detector distance.
- We collected 200˚ of data, 0.4˚ per frame, 500 images, 3.7s per frame, wavelength 1.169 Å (as close as we could get to Zinc on the beamline used) (deliberately high redundancy for the anomalous signal).
- We indexed in P3121, a=b=176.75 Å, c=72.22 Å, α=β=90, gamma=120˚

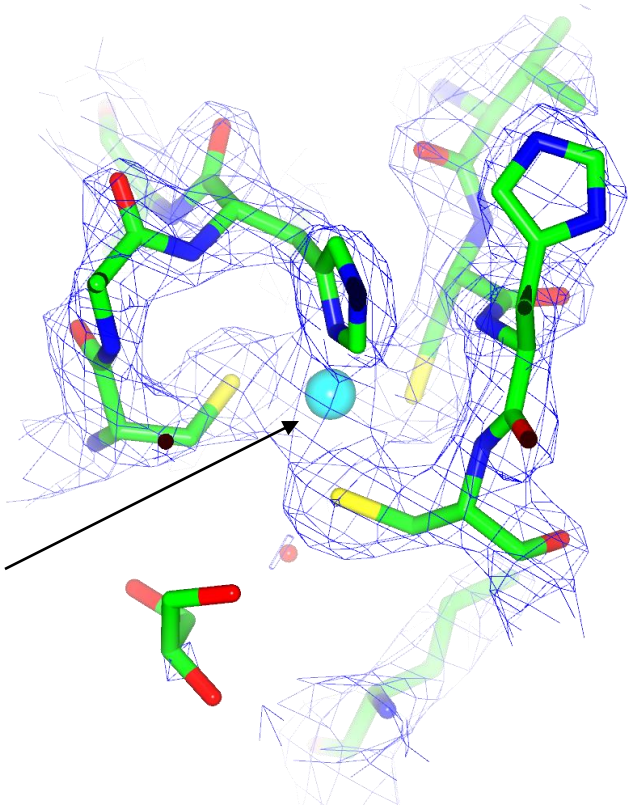
	Overall	Inner Shell	Outer Shell
Low resolution limit (Å)	40.00	40.00	2.64
High resolution limit (Å)	2.5	7.91	2.5
R_{merge}	0.104	0.036	0.743
R_{pim}	0.032	0.011	0.273
	3.2%	1.1%	27.3%
Total number of observations	508484	17694	51511
Total number unique	44752	1523	6332
Mean((I)/sd(I))	24.6	86.6	2.2
Completeness (%)	99.7	99.9	97.9
Multiplicity	11.4	11.6	8.1

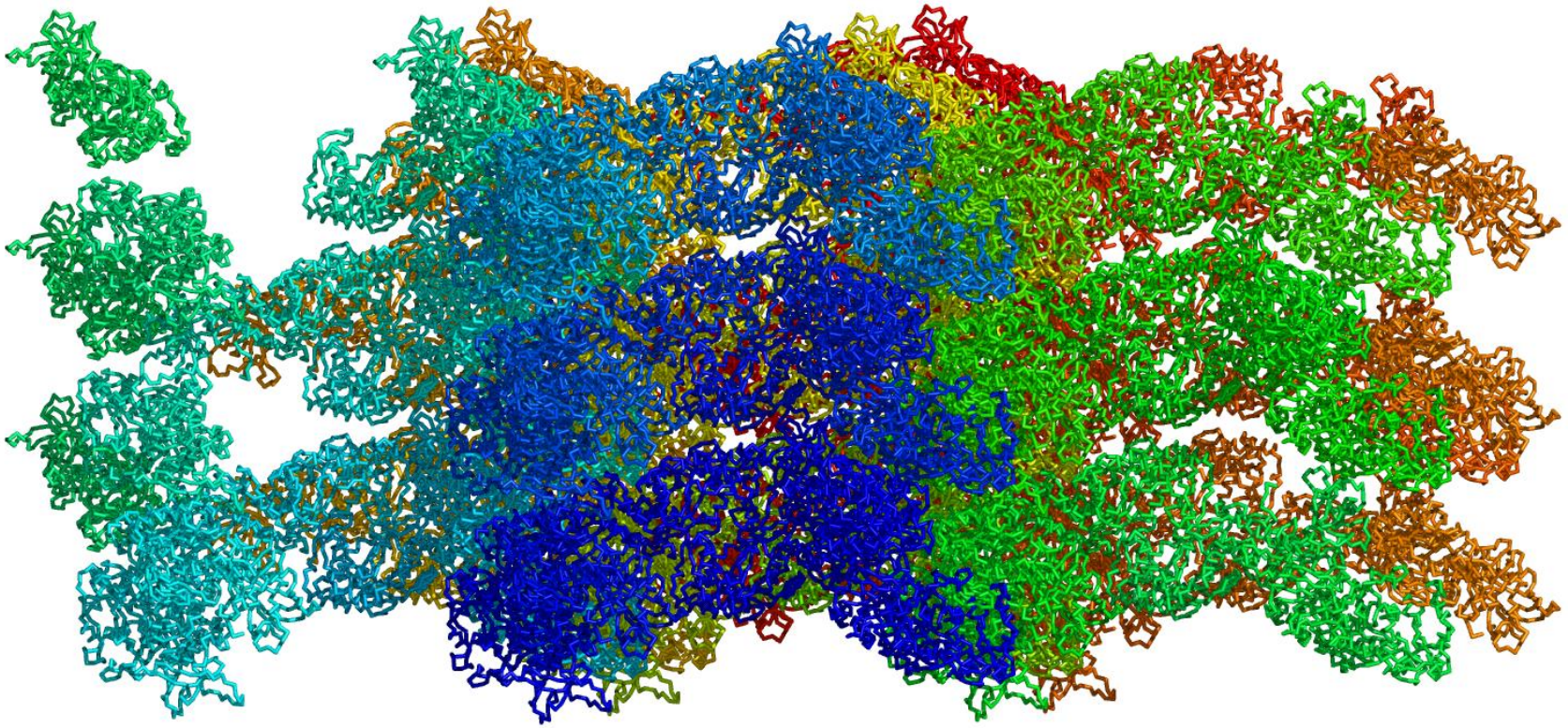


Zinc

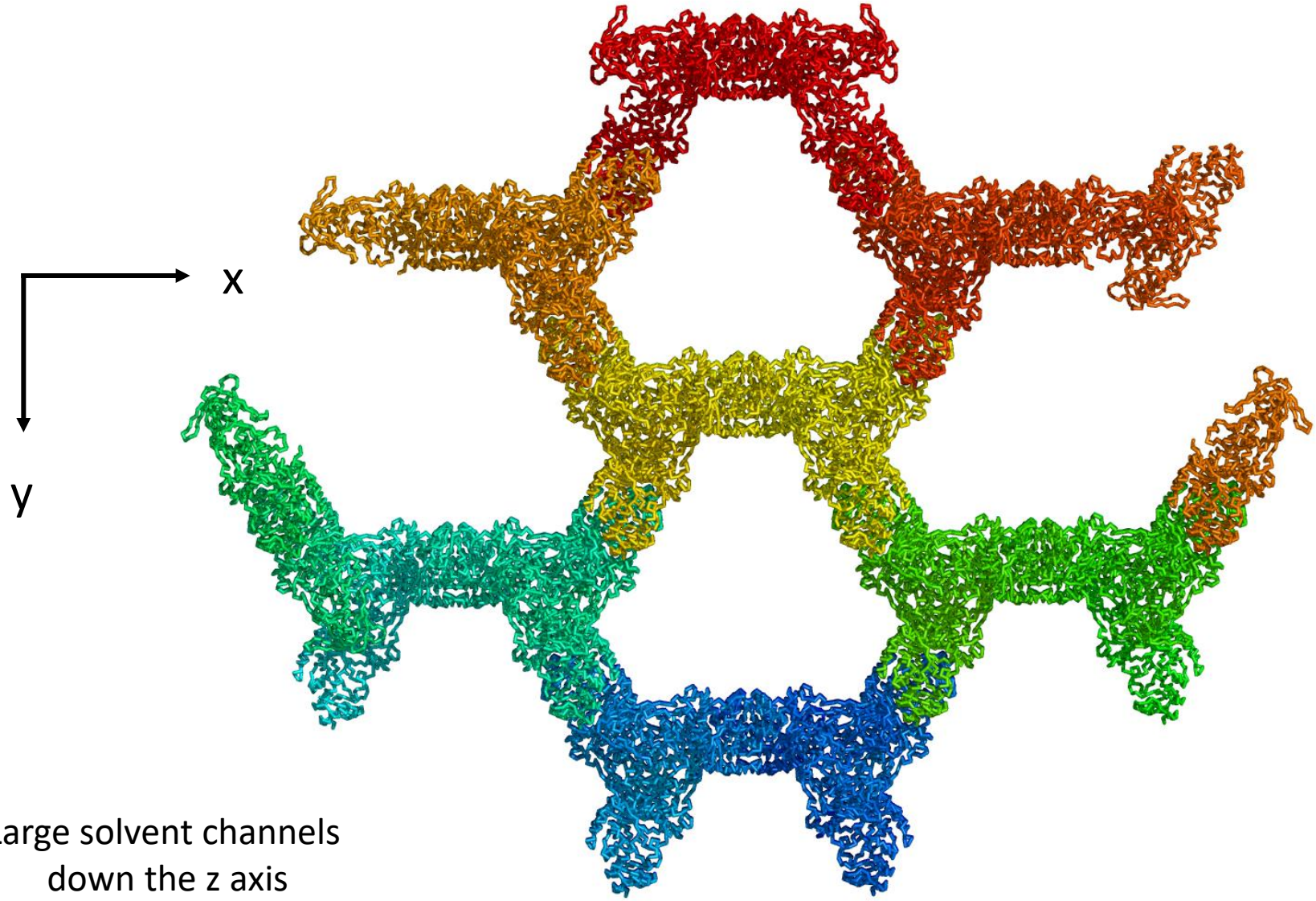
Structure solved (with help of the zinc signal) and refined with Phenix.

Overall R and R_{free} are 14.2 and 19.8% respectively.

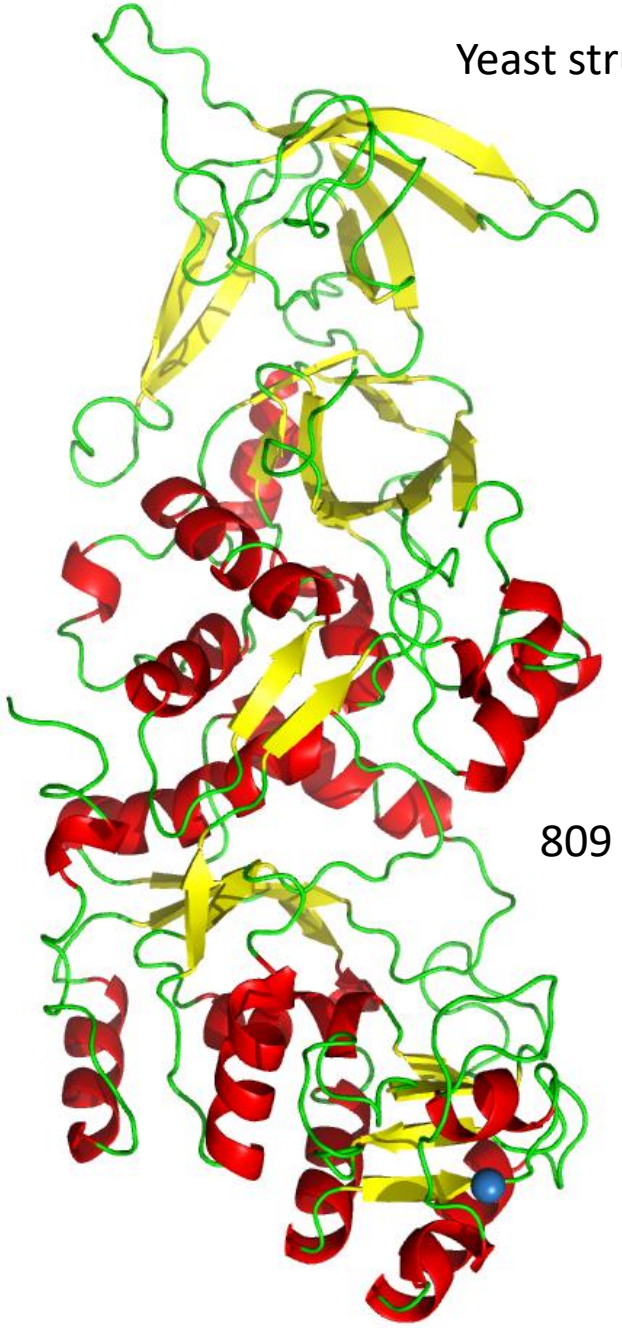




Tight packing in z and y



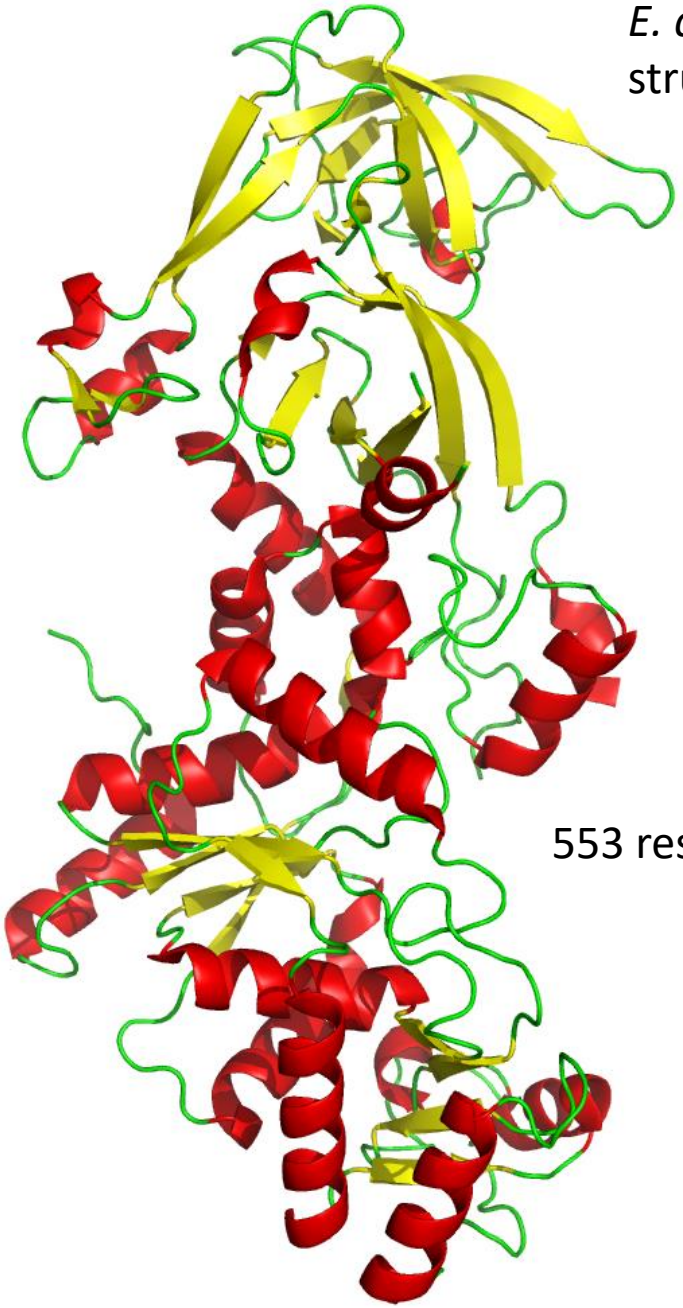
Large solvent channels
down the z axis



Yeast structure

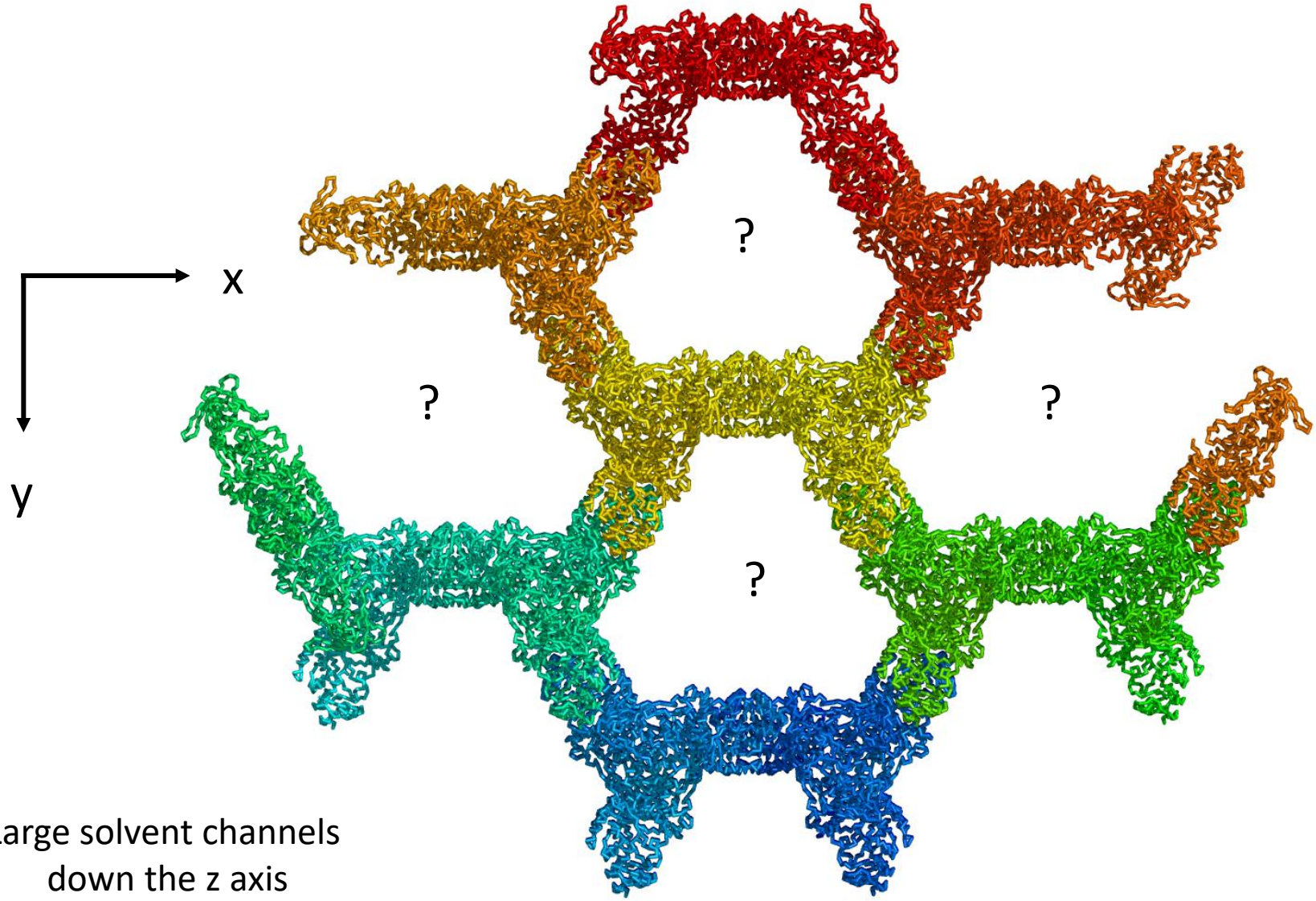
?

809 residues

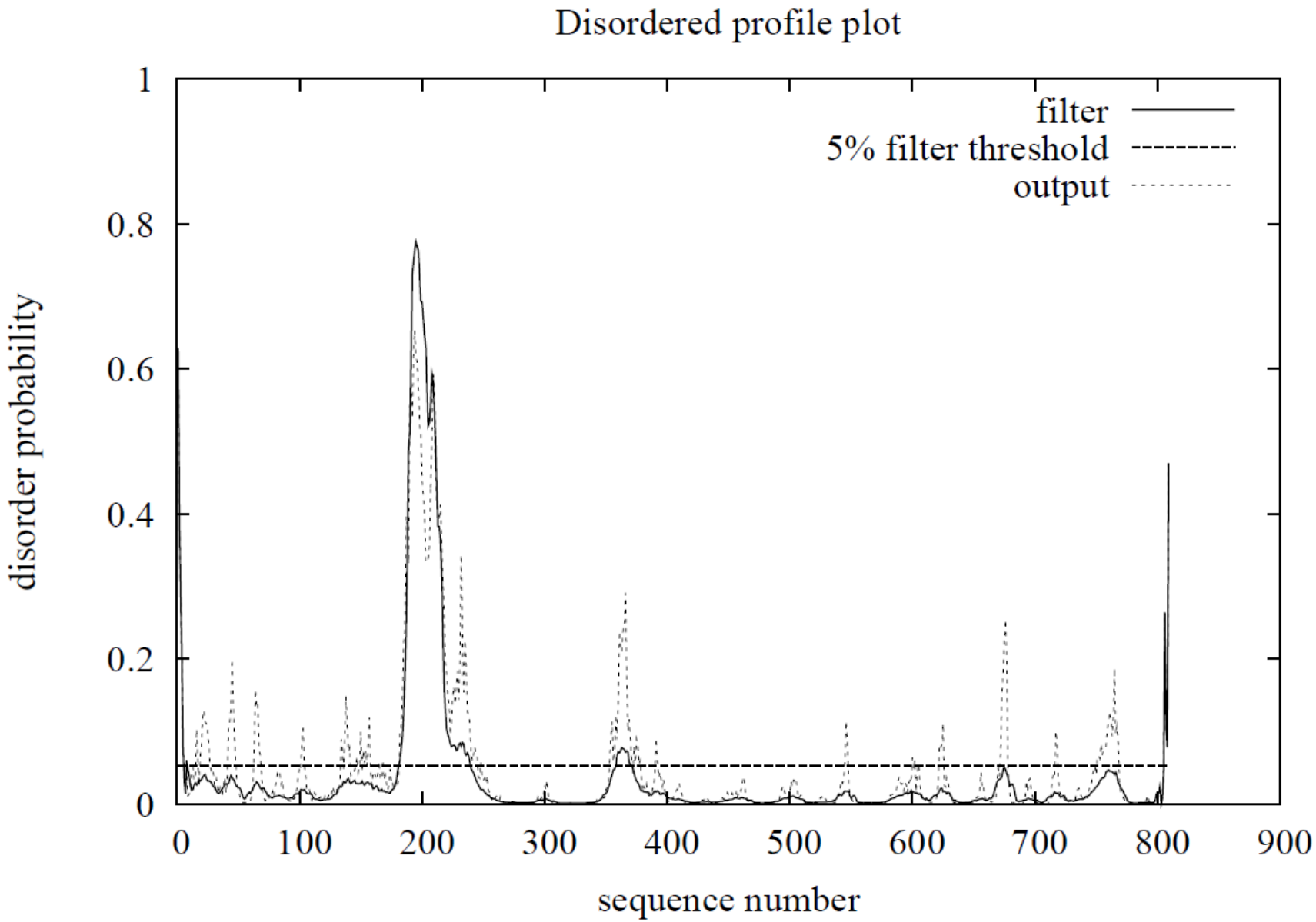


E. coli. structure

553 residues

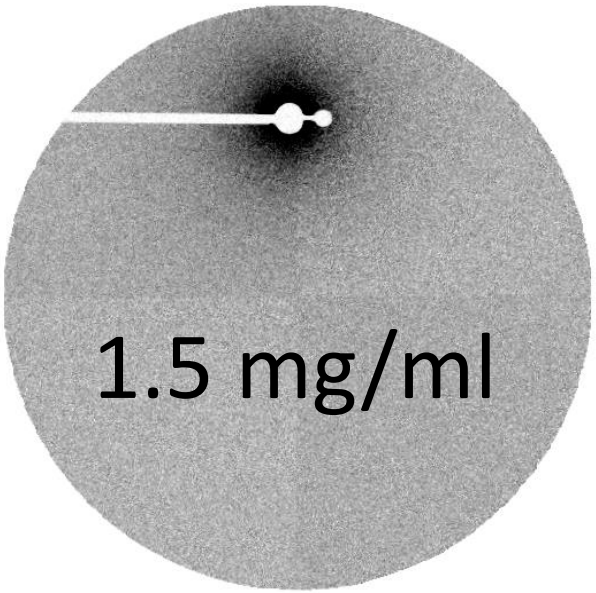


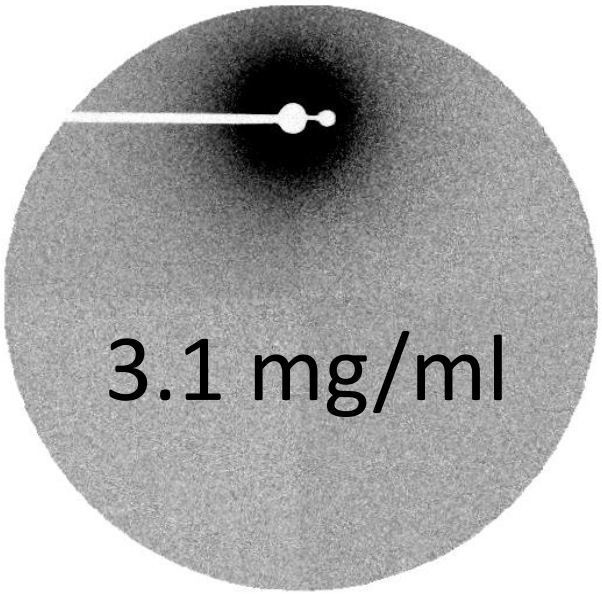
Large solvent channels
down the z axis

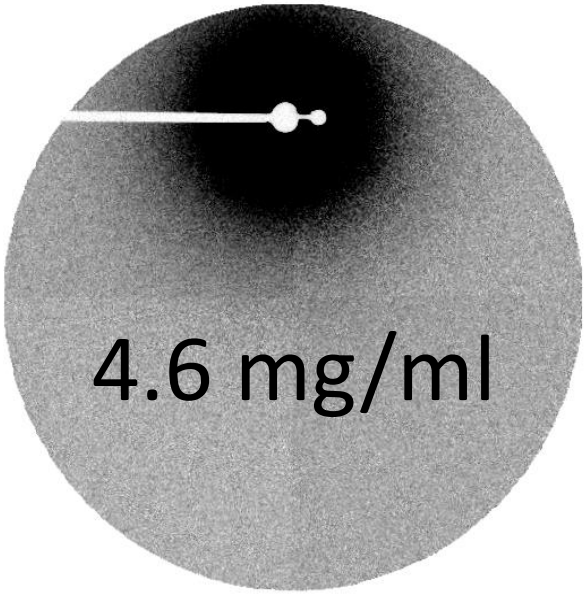

Disordered profile plot
filter
5% filter threshold
output
disorder probability
sequence number
0
0.2
0.4
0.6
0.8
1
0
100
200
300
400
500
600
700
800
900

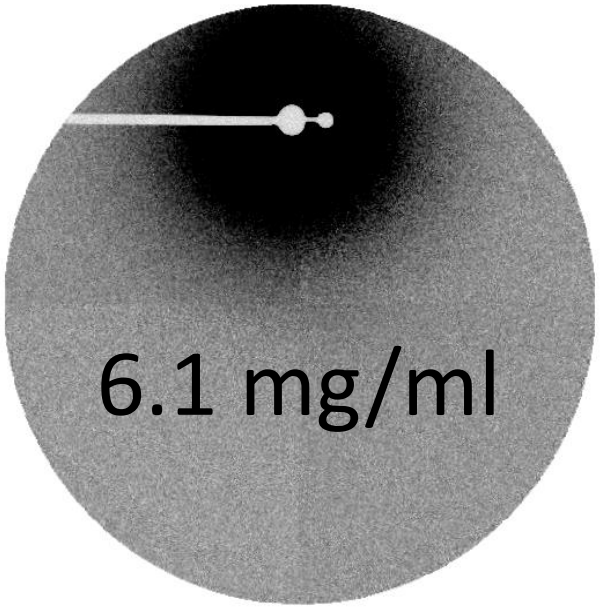
Missing residues

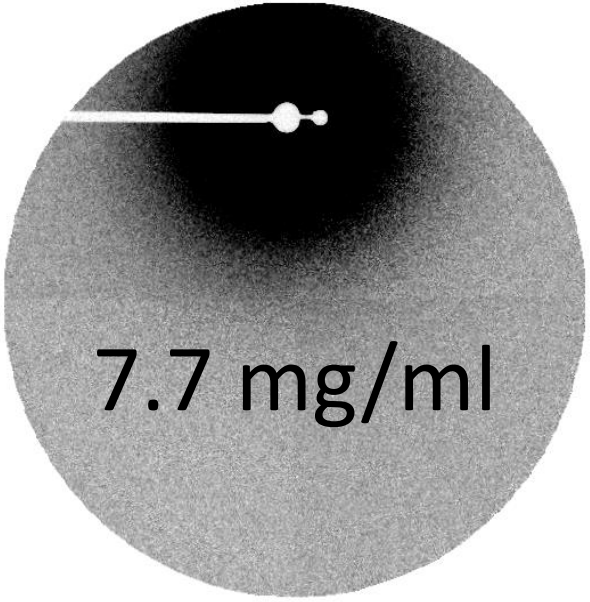
- There were 216 missing residues from the structure, 95% of the N-terminal domain.
- Where they in the mix to start with?.
- SDS PAGE gel on the remaining crystals indicated that the full length protein was present.
- For a more concrete answer the protein was re-expressed with a His tag attached to the N-terminal domain.
  - It was purified with a nickel affinity column.
  - It was crystallized and the structure solved, again with missing residues.
  - A western blot on the dissolved crystals confirmed the presence of the N-terminal domain His tag.
  - No protein degradation had taken place during crystallization.
- For the re-expressed protein the full N-terminal domain was present in the protein but not seen in the crystallographic structure.

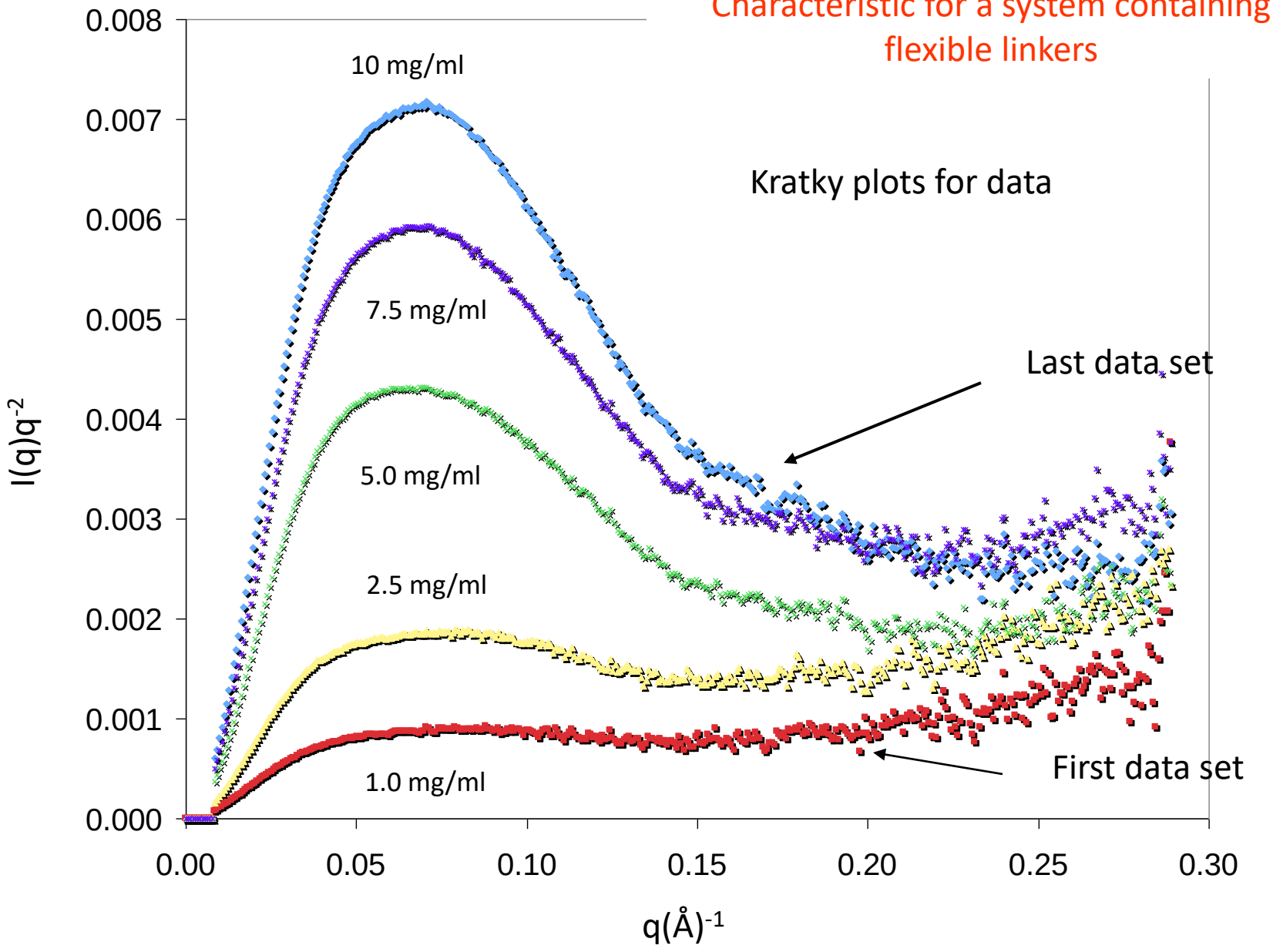

1.5 mg/ml

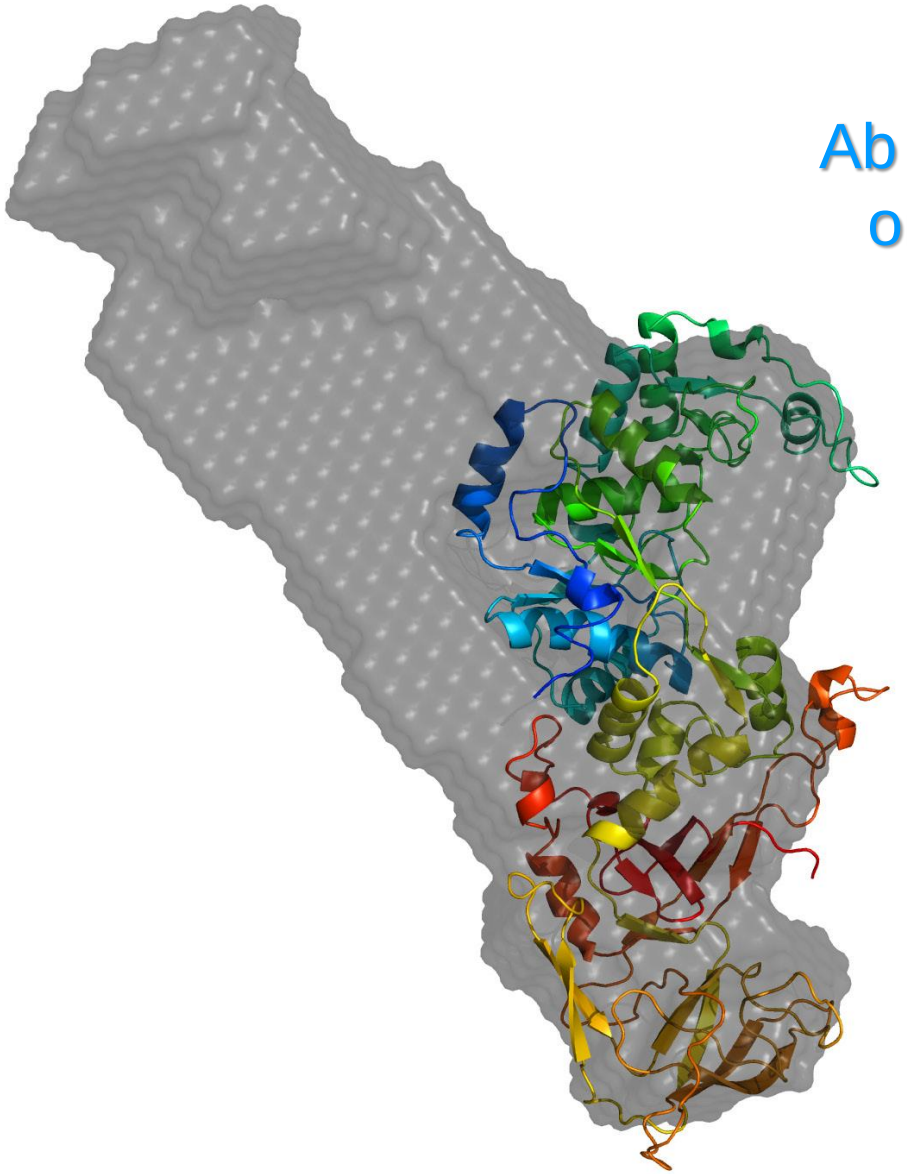

3.1 mg/ml


4.6 mg/ml


6.1 mg/ml

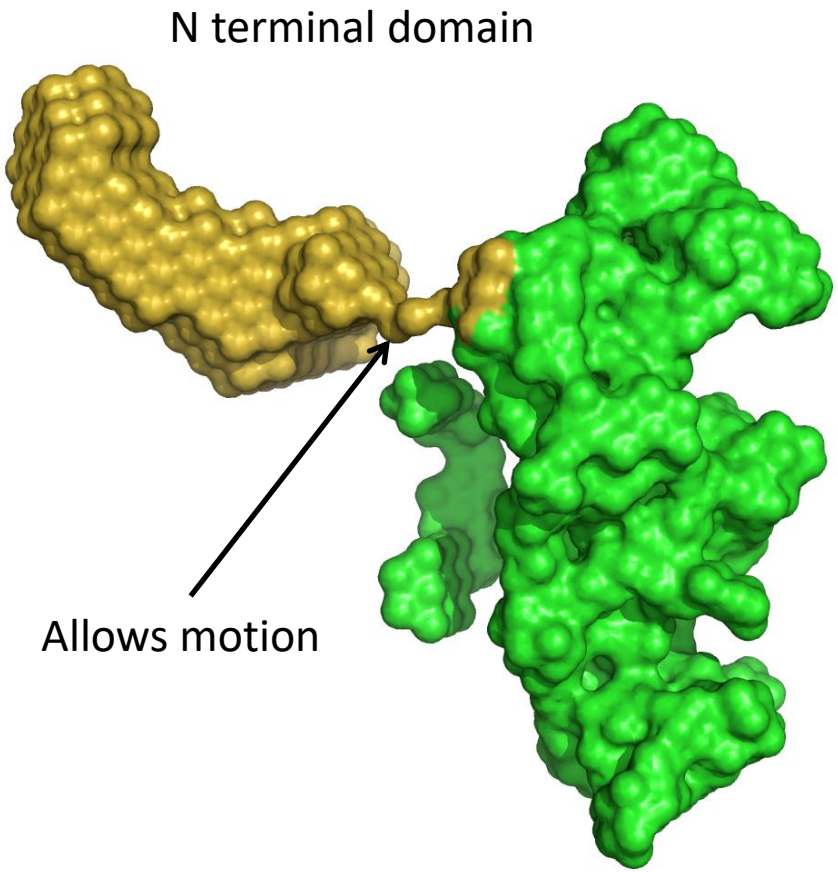

7.7 mg/ml


Characteristic for a system containing flexible linkers
Kratky plots for data
10 mg/ml
7.5 mg/ml
5.0 mg/ml
2.5 mg/ml
1.0 mg/ml
Last data set
First data set
I(q)q-2
q(Å)-1
0.000
0.001
0.002
0.003
0.004
0.005
0.006
0.007
0.008
0.00
0.05
0.10
0.15
0.20
0.25
0.30

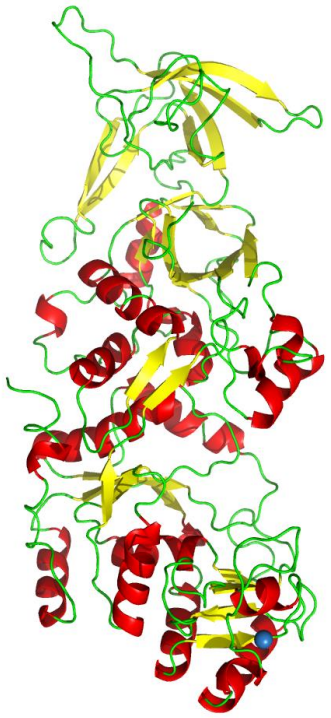


Ab initio structure overlaid on the crystallographic structure

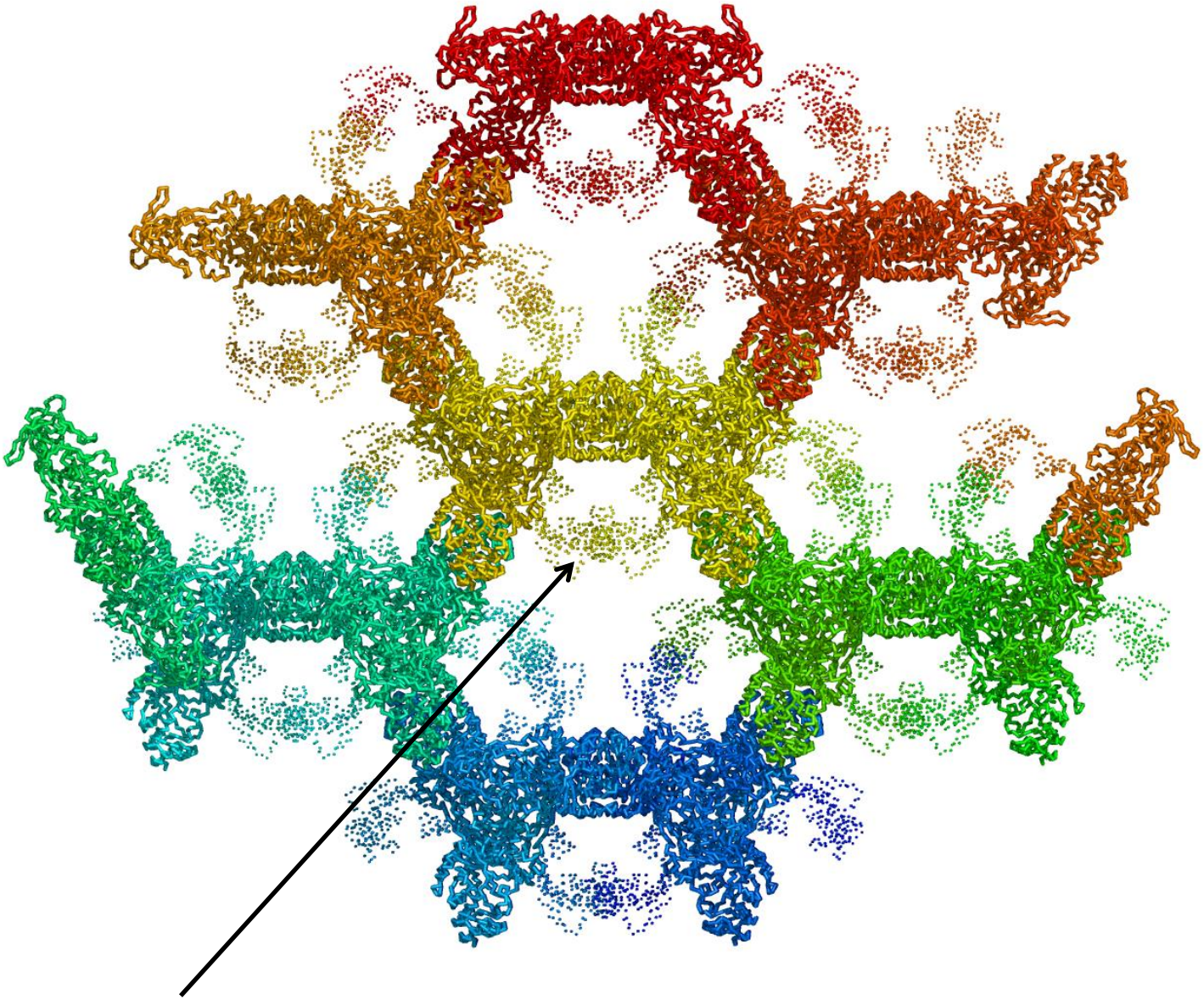
Envelope reconstruction using the crystallographic structure



C terminal domain



The crystal structure (which shows only the C-domain)

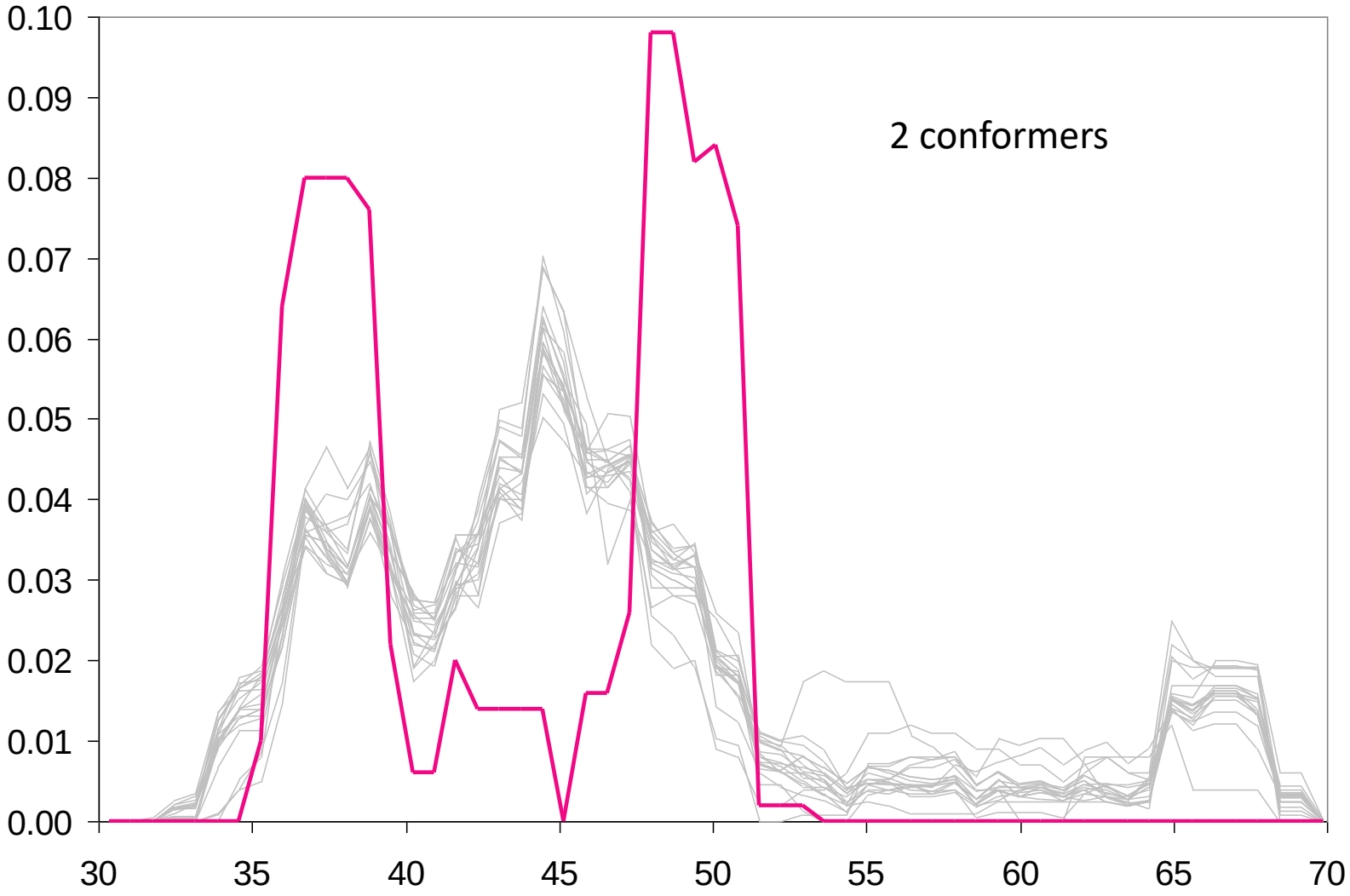


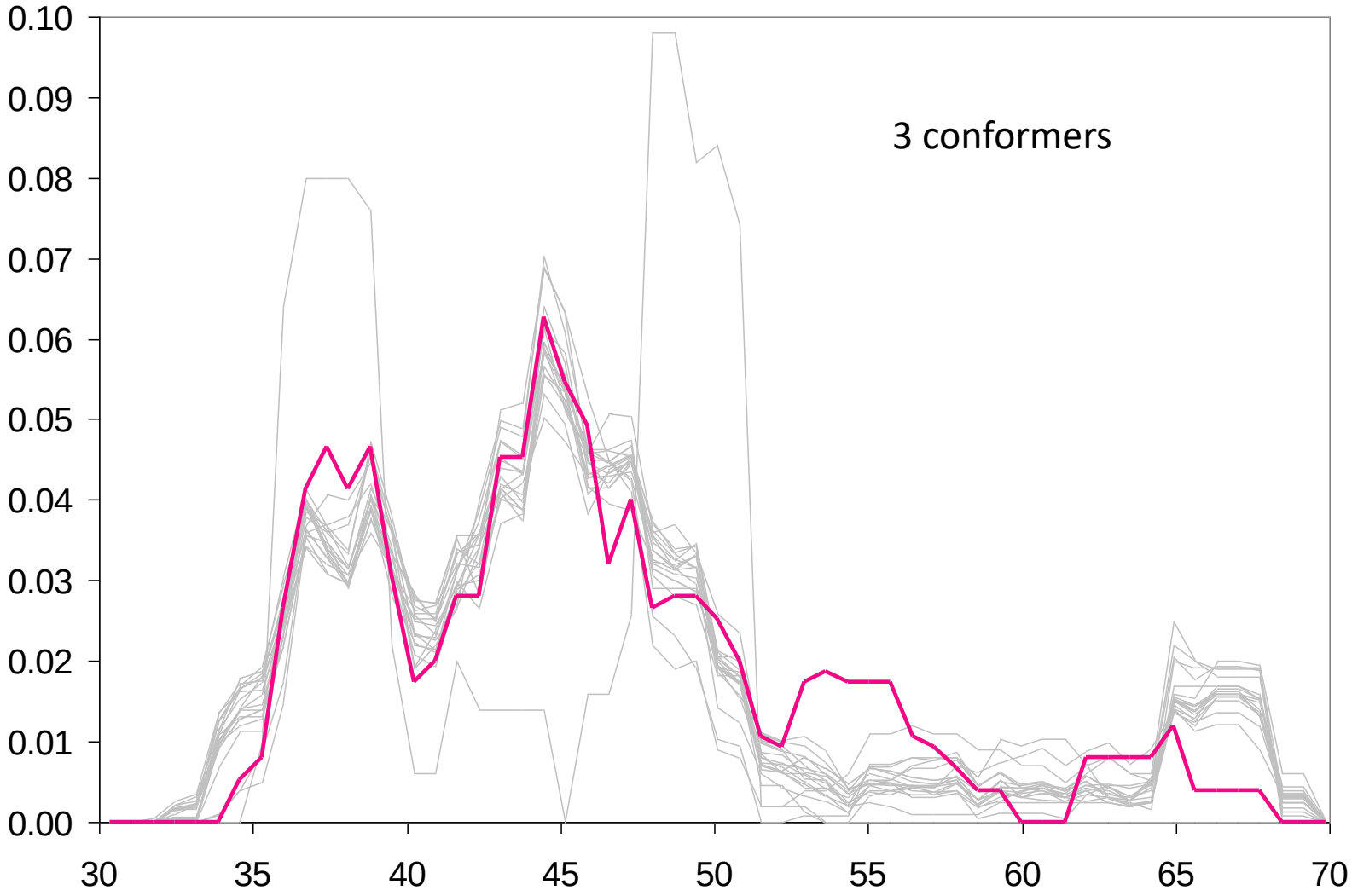
The N-terminal ‘arm’ is completely compatible with the crystal structure

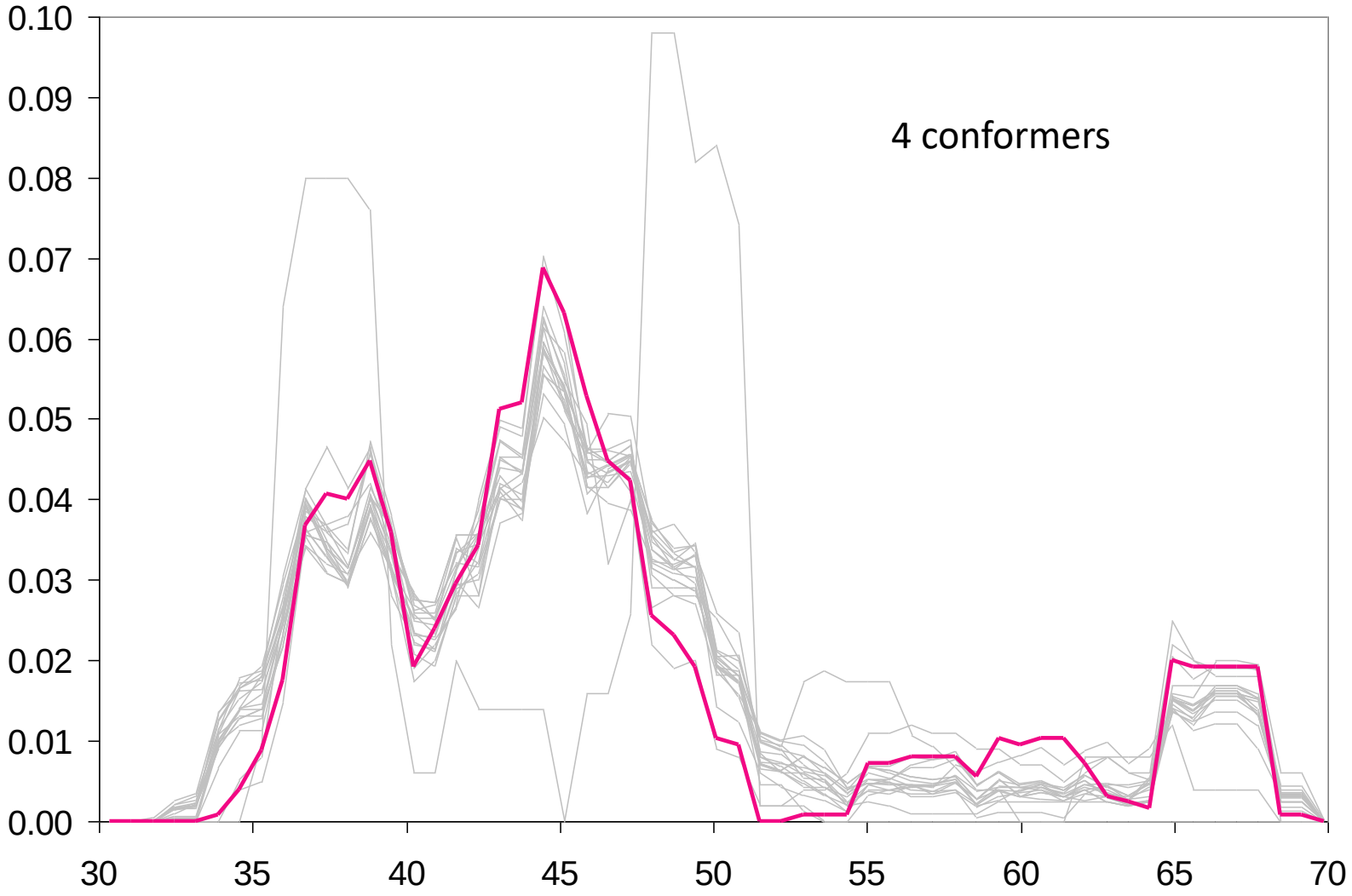
Ensemble optimization

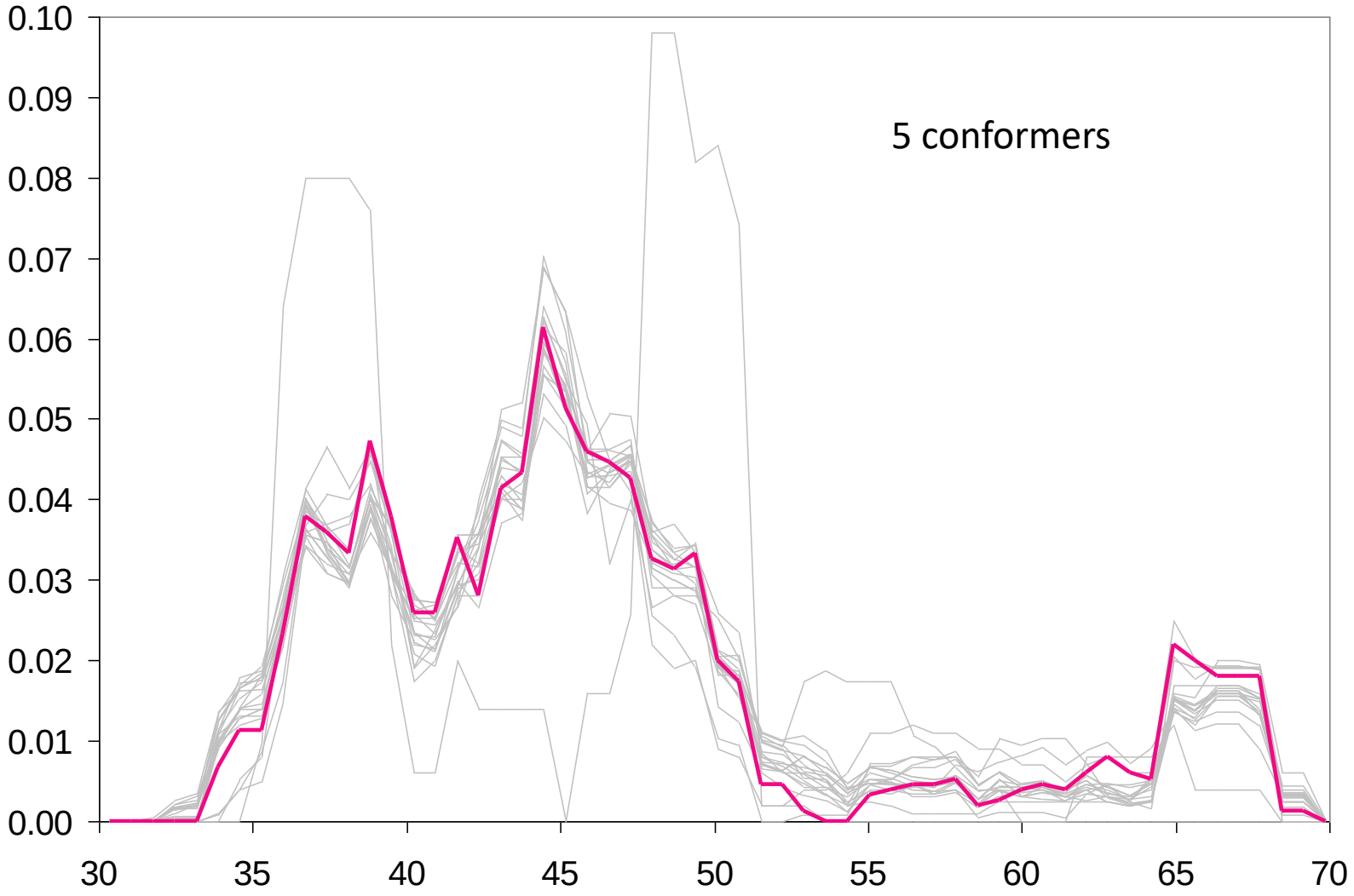
- The Ensemble Optimization Method (EOM) was used to assess the flexibility of the Gln4 N-terminal domain.
- RanCh (**Ran**dom **Ch**ain Generator) generated 10,000 conformers of the N-terminal sequence of Gln4 covering all possible configuration space.
- Sets of these conformers were binned to create ensembles.
- GAJOE (**G**enetic **A**lgorithm **J**udging **O**ptimization of **E**nsembles) optimized the ensembles by comparing the average scattering profile of their conformers to the experimental data.
- Plotting the *Rg* distribution for successive runs, each using an increasing number of conformers per ensemble, allows us to identify the optimal number of conformers that most accurately characterizes the system.
- Analysis of chi (an error indicator) shows an systematic decrease, converging at eight conformers in each ensemble.

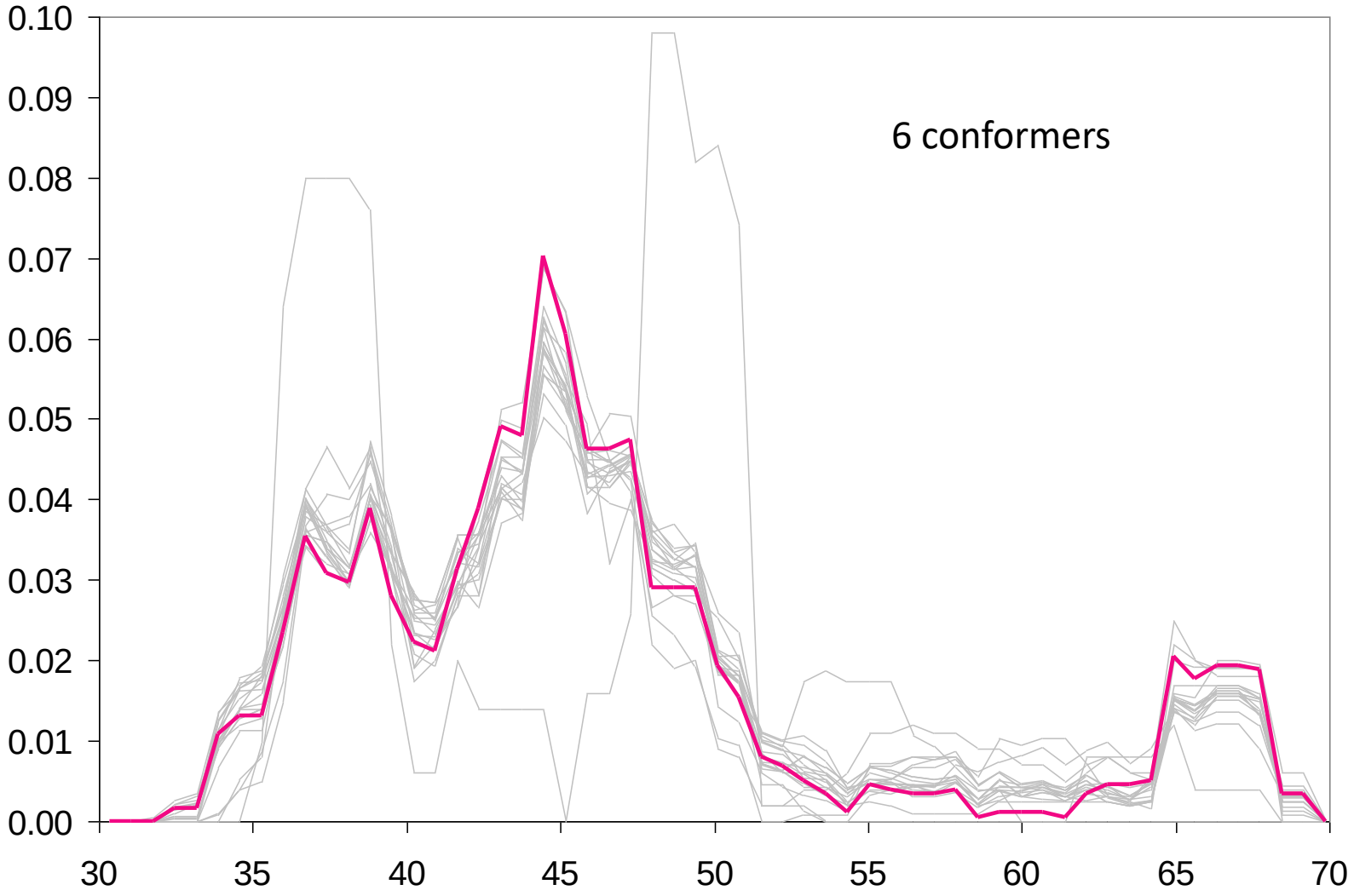
The convergence of the population distribution on distinct populations indicates that dynamic motion or different species are present - when this is not the case the distribution is monomodal (confirmed by similar analyses on static systems).

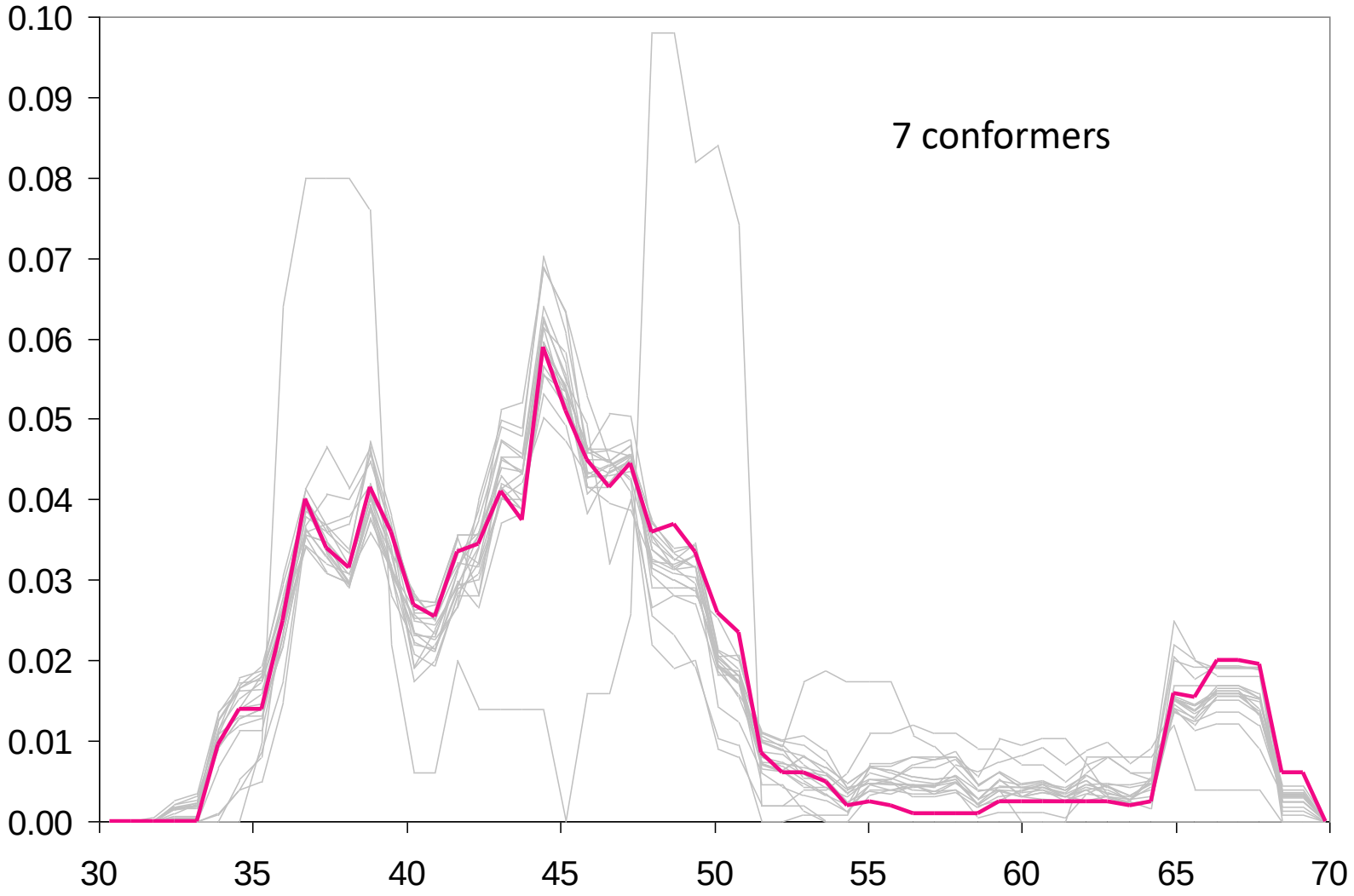

0.10
0.09
0.08
0.07
0.06
0.05
0.04
0.03
0.02
0.01
0.00
30
35
40
45
50
55
60
65
70
2 conformers

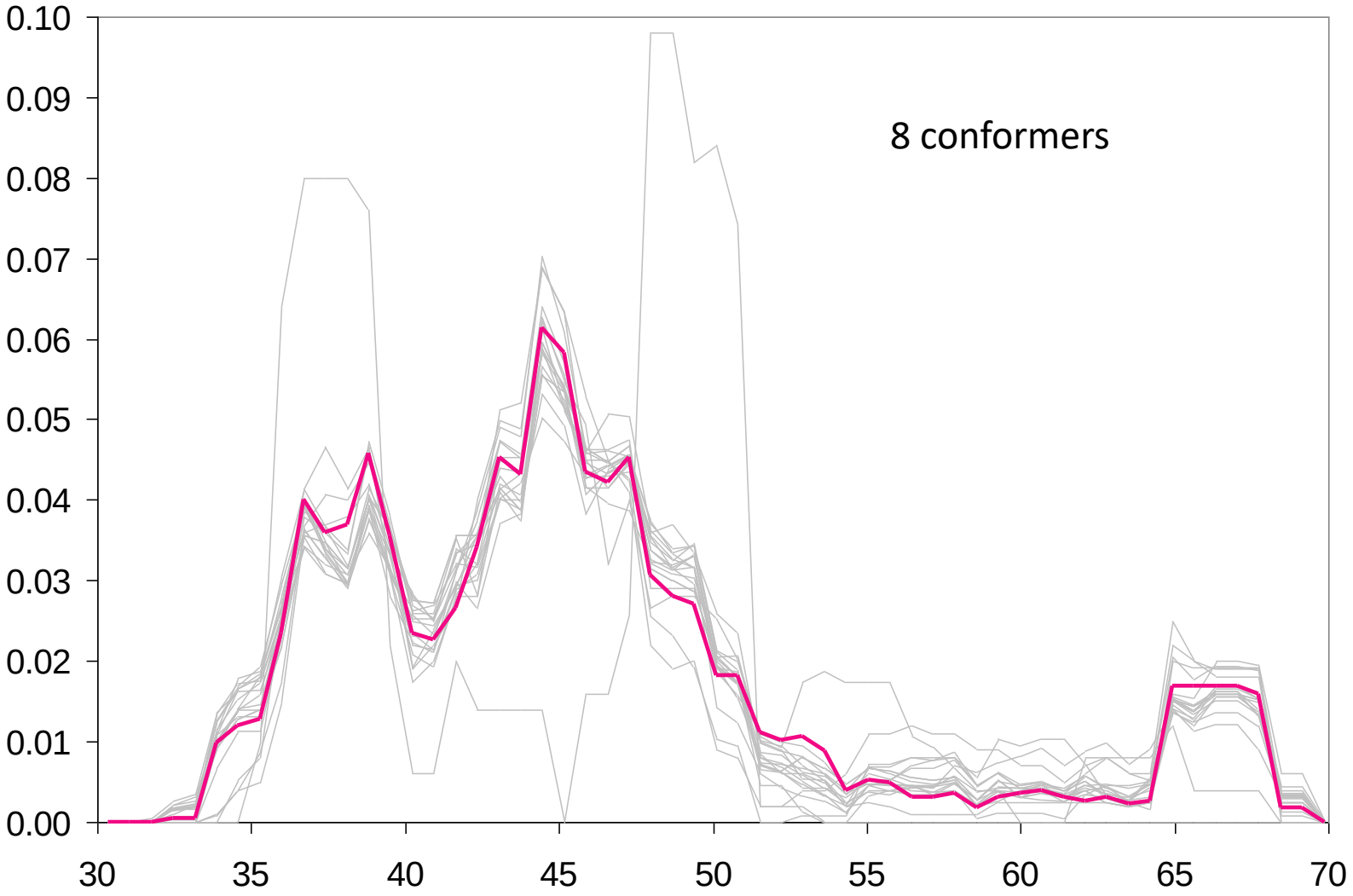

3 conformers
0.10
0.09
0.08
0.07
0.06
0.05
0.04
0.03
0.02
0.01
0.00
30
35
40
45
50
55
60
65
70

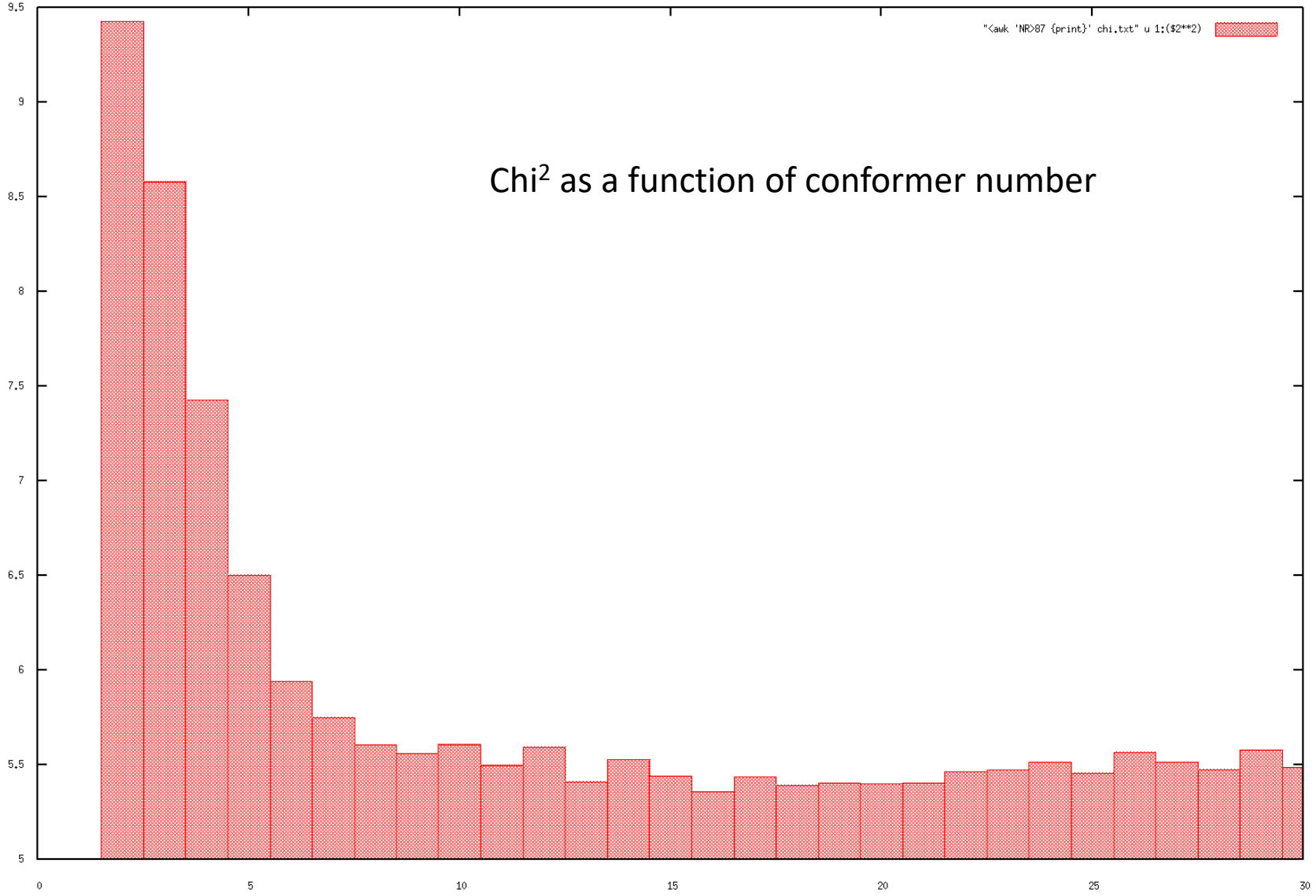

4 conformers
0.10
0.09
0.08
0.07
0.06
0.05
0.04
0.03
0.02
0.01
0.00
30
35
40
45
50
55
60
65
70


5 conformers
0.10
0.09
0.08
0.07
0.06
0.05
0.04
0.03
0.02
0.01
0.00
30
35
40
45
50
55
60
65
70


6 conformers
0.10
0.09
0.08
0.07
0.06
0.05
0.04
0.03
0.02
0.01
0.00
30
35
40
45
50
55
60
65
70

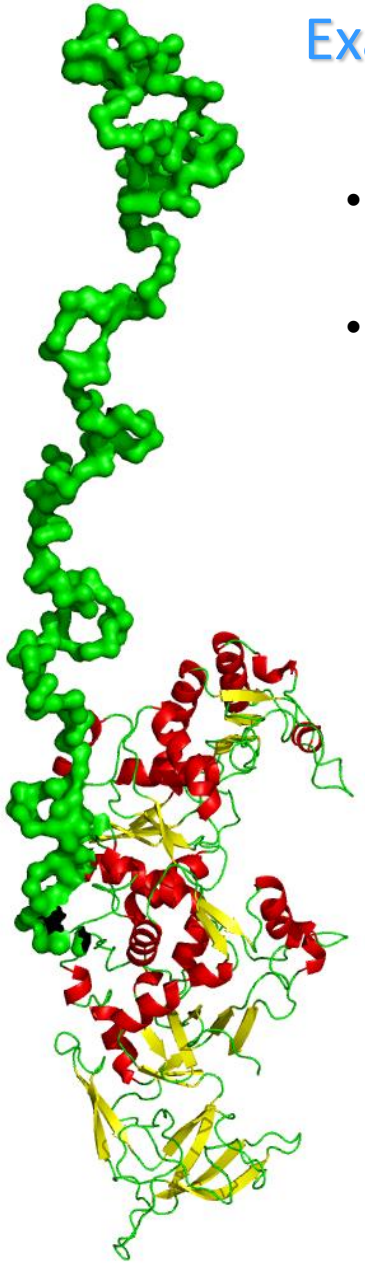

0.10
0.09
0.08
0.07
0.06
0.05
0.04
0.03
0.02
0.01
0.00
30
35
40
45
50
55
60
65
70
7 conformers

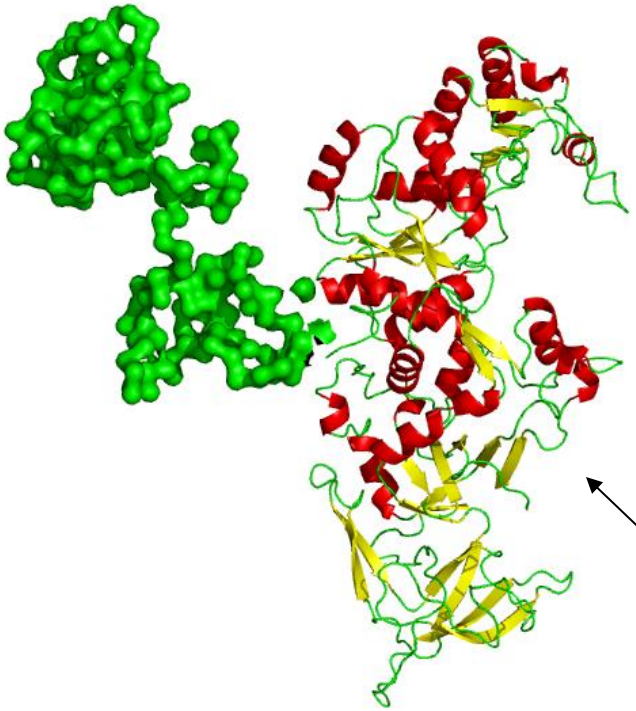

8 conformers
0.10
0.09
0.08
0.07
0.06
0.05
0.04
0.03
0.02
0.01
0.00
30
35
40
45
50
55
60
65
70

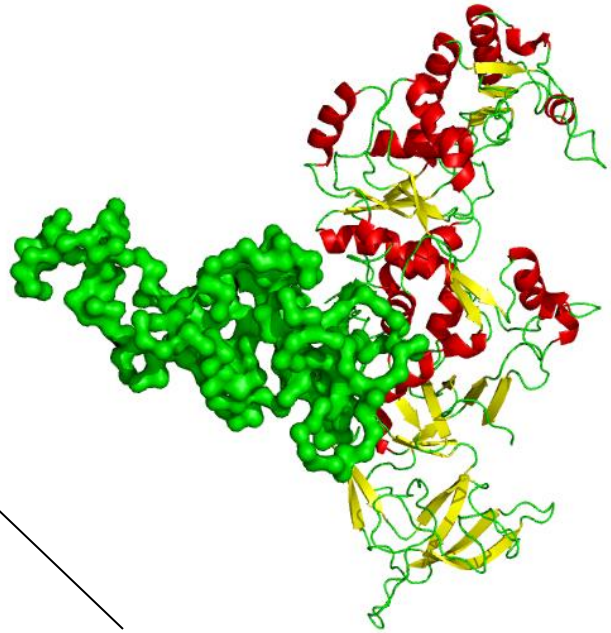

"<awk 'NR>87 {print}' chi.txt" u 1:($2**2)
Chi2 as a function of conformer number
9.5
9
8.5
8
7.5
7
6.5
6
5.5
5
0
5
10
15
20
25
30

Example Conformers from the Ensemble optimization

- Ensemble optimization told us that the SAXS data could be best explained with a minimum of 8 different conformers.
- The single *ab initio* model produced by traditional techniques represents the average conformation in solution.







Crystallographic structure used

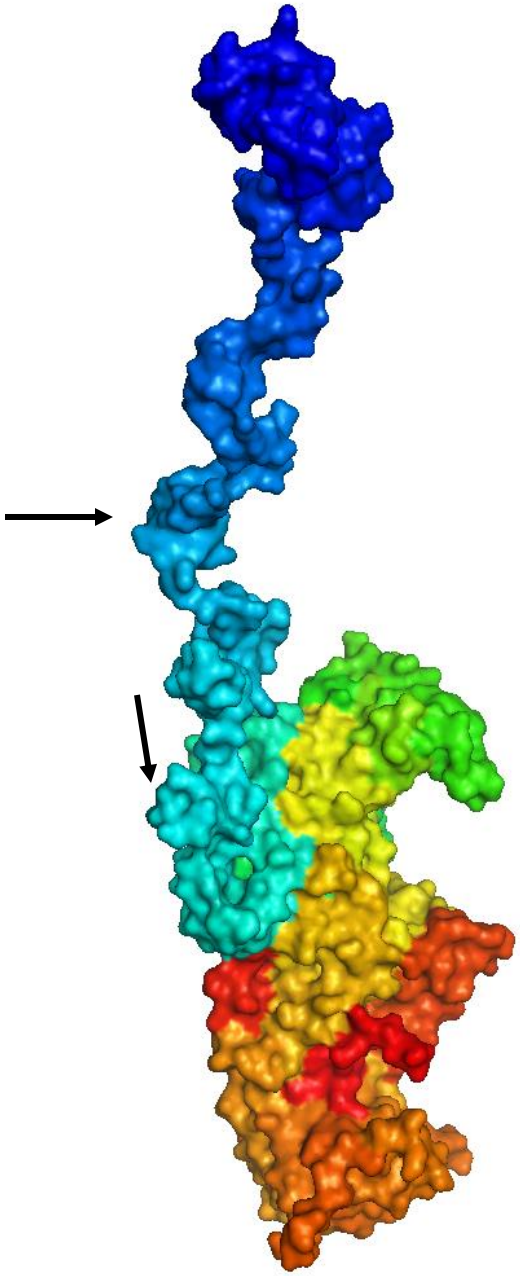
Outline

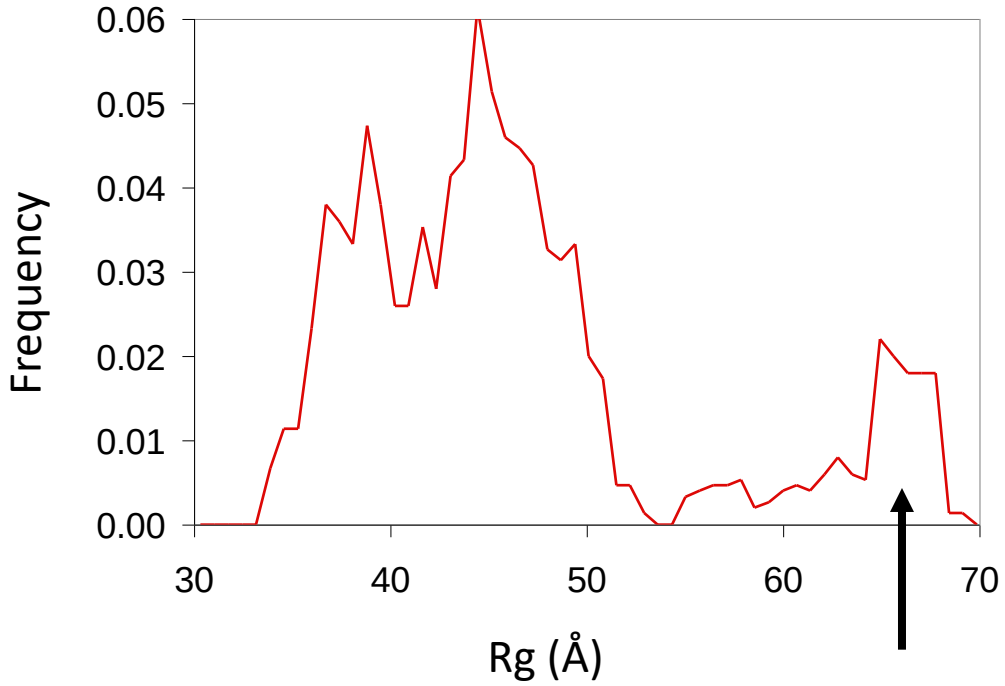
- Biological dynamics
- Crystallography
- Small Angle X-ray Scattering
- A biological case
- A structure, a puzzle, a question?
- A solution
- **Computational analysis**
- An answer, more questions, more solutions
- Where are we going next?

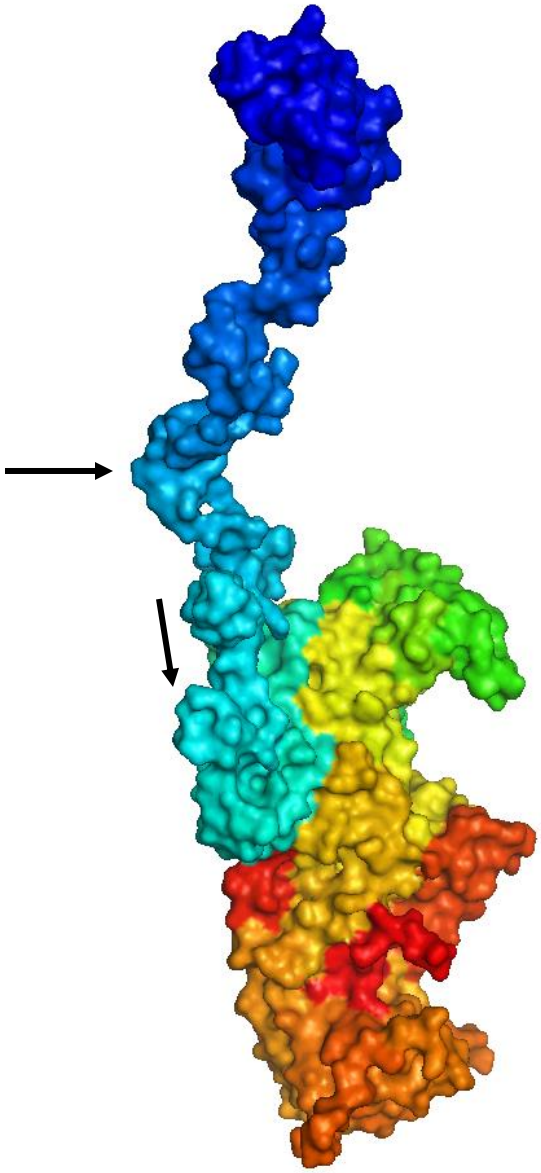
Computationally model the motion

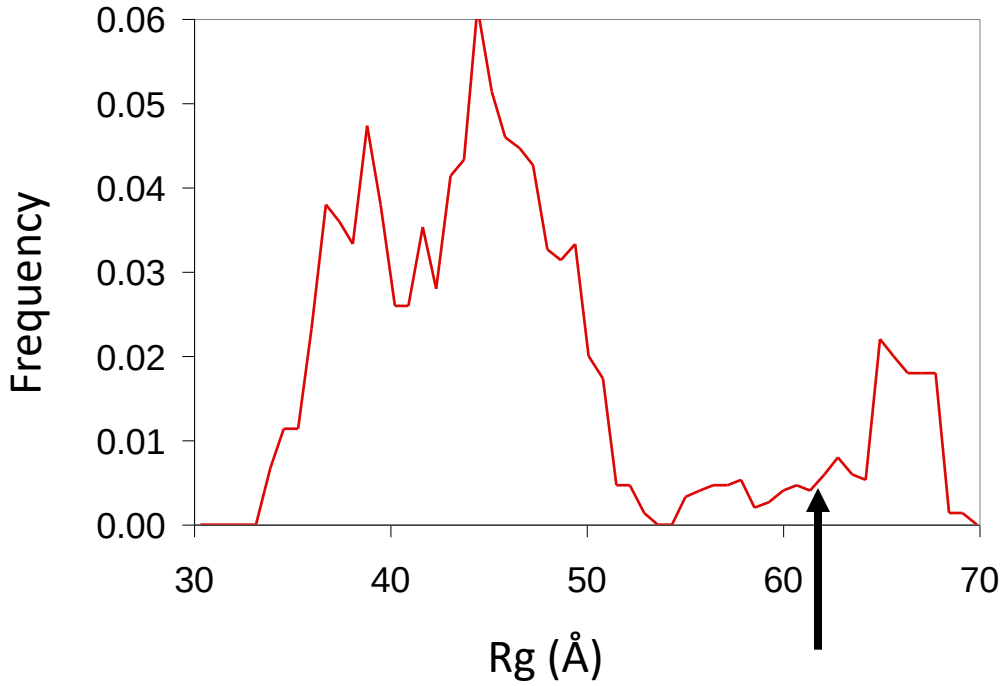
- The motion is too large for us to perform full molecular dynamics simulations with the computing capacity currently available to us (a ~200 processer cluster in-house and shared time on a neighboring 2000+ processer cluster).
- We took the most compact form and the most extended form and using an energy minimization procedure with Morph Server calculated a pathway between the two forms.

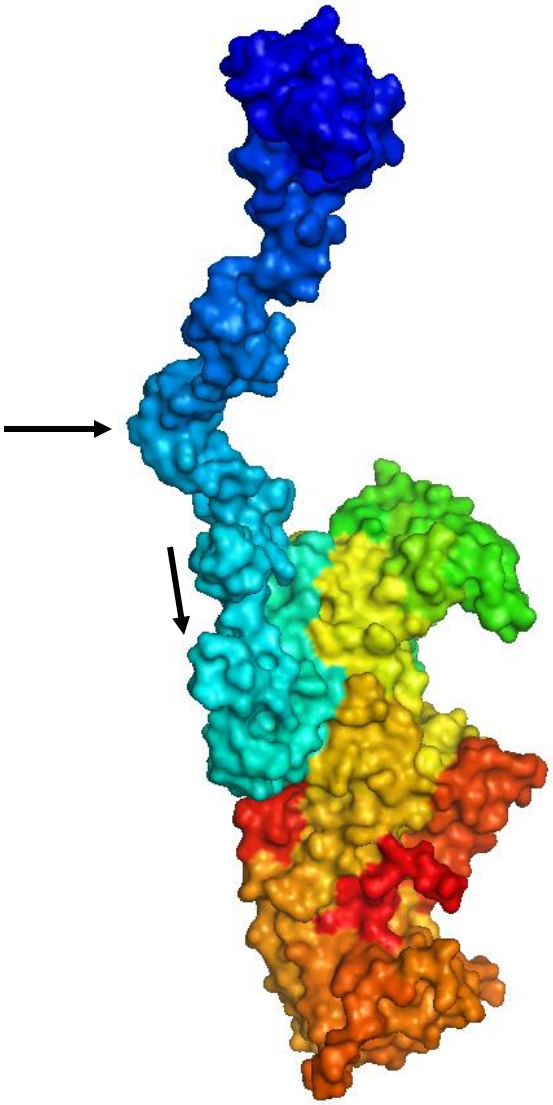
- This is a preliminary analysis. A future approach will be to run molecular dynamics simulations on each conformer to evaluate the pathway between nearest neighbors. This appears to be computationally feasible.

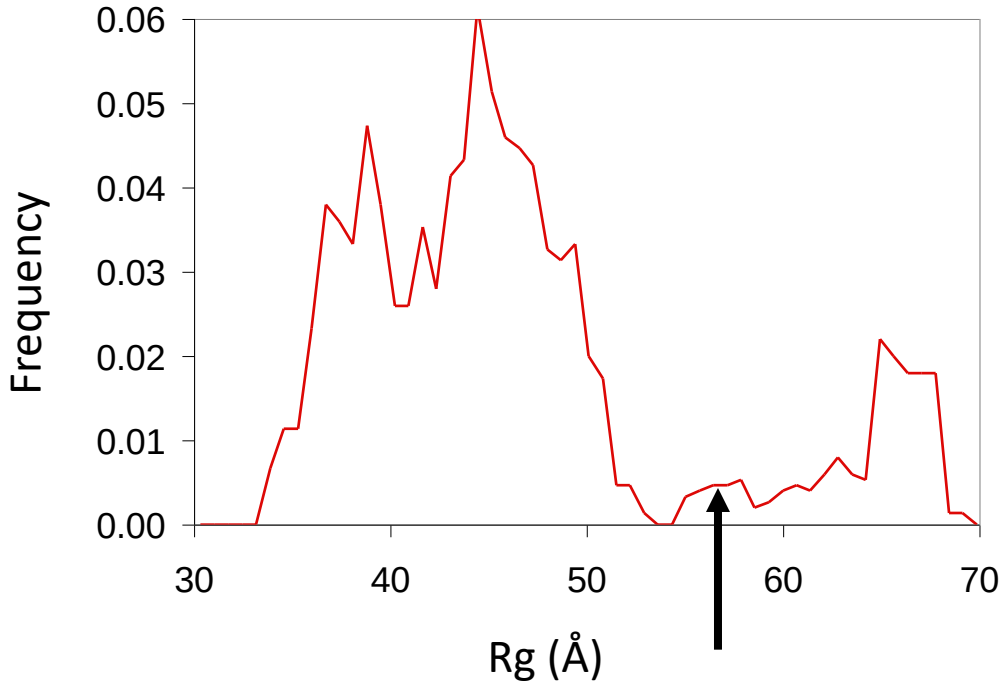


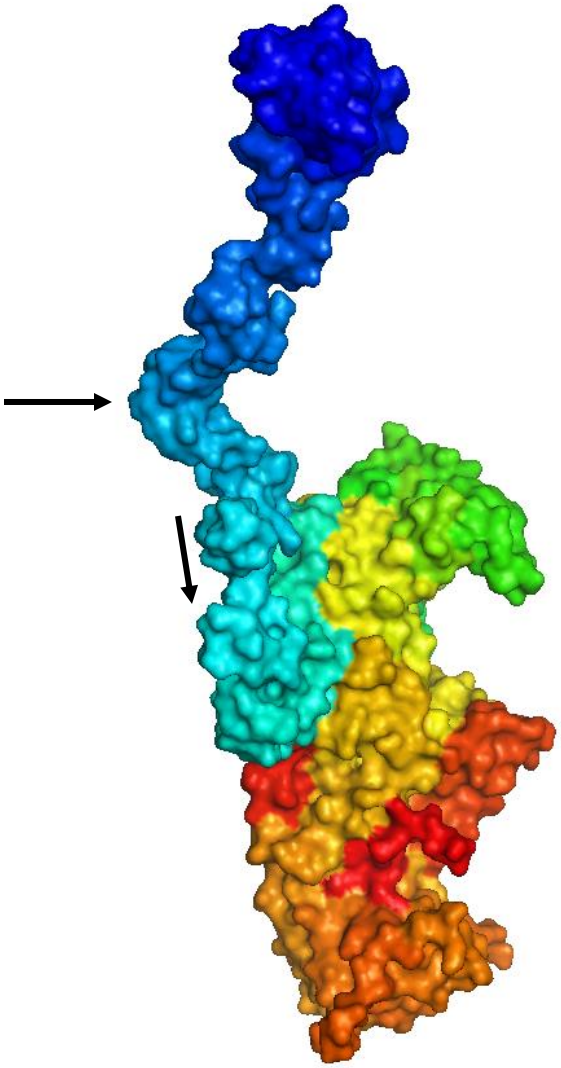

0.06
0.05
0.04
0.03
0.02
0.01
0.00
Frequency
30
40
50
60
70
Rg (Å)

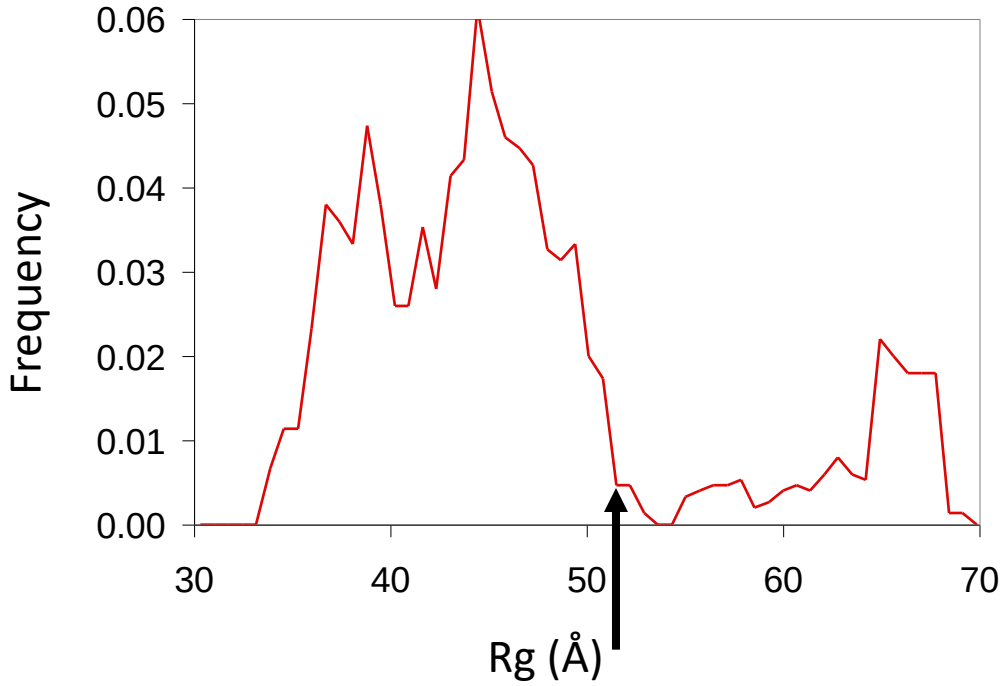


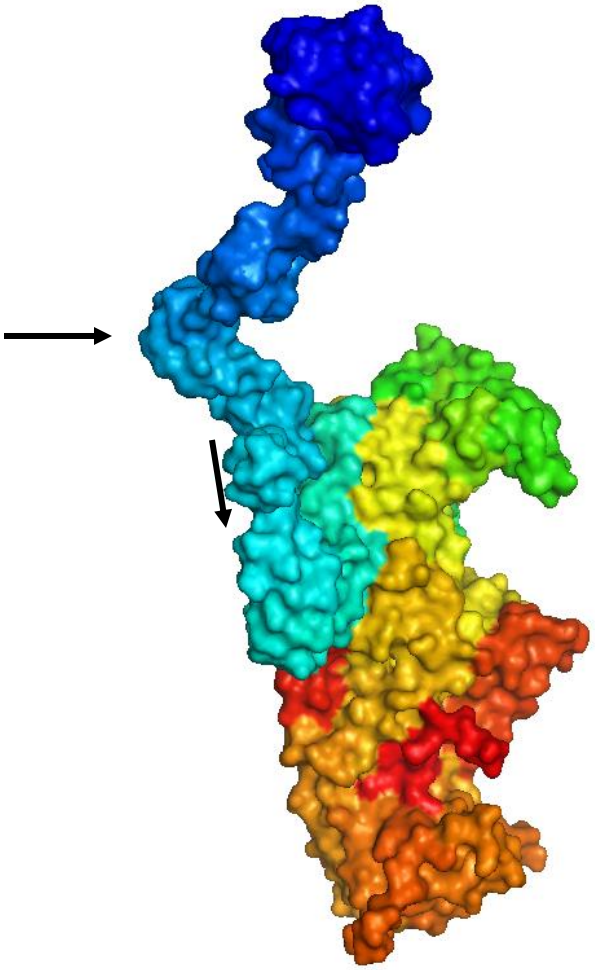

0.06
0.05
0.04
0.03
0.02
0.01
0.00
Frequency
30
40
50
60
70
Rg (Å)

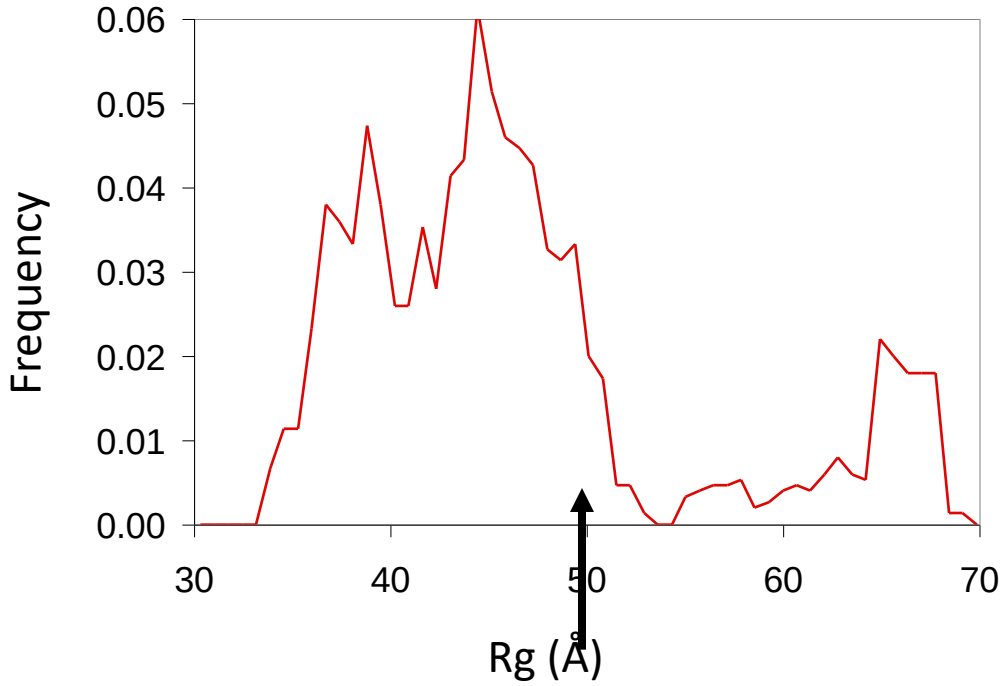


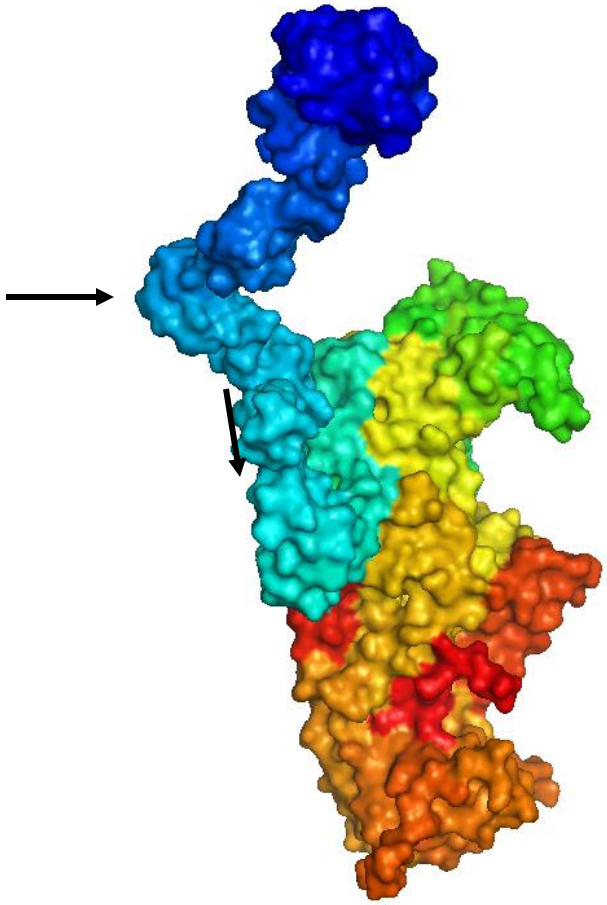

0.06
0.05
0.04
0.03
0.02
0.01
0.00
Frequency
30
40
50
60
70
Rg (Å)

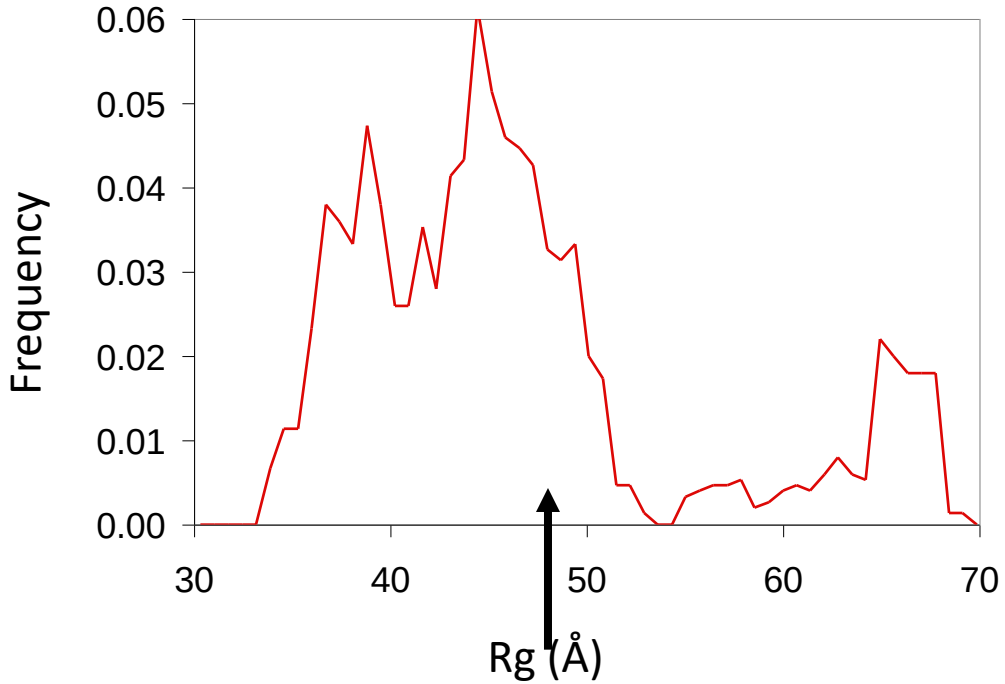


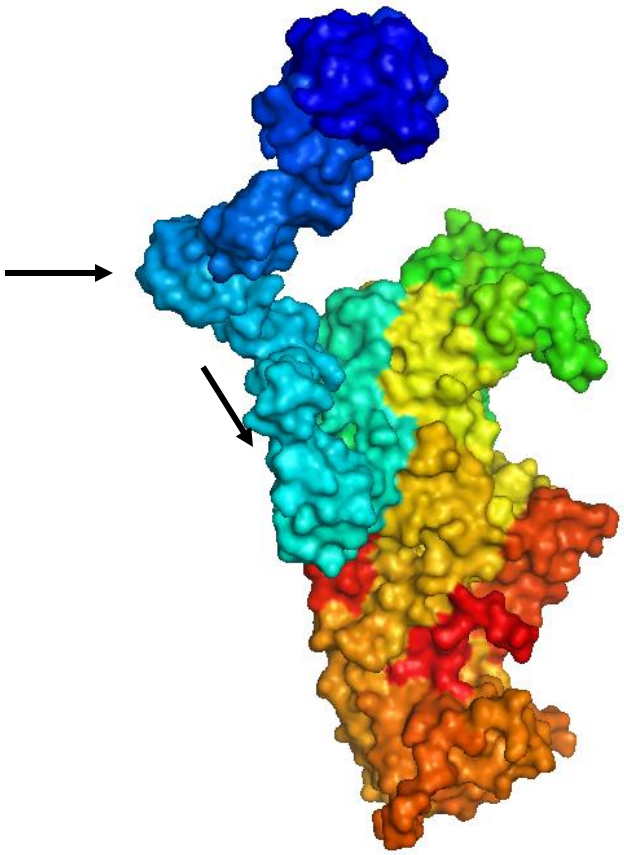

0.06
0.05
0.04
0.03
0.02
0.01
0.00
Frequency
30
40
50
60
70
Rg (Å)

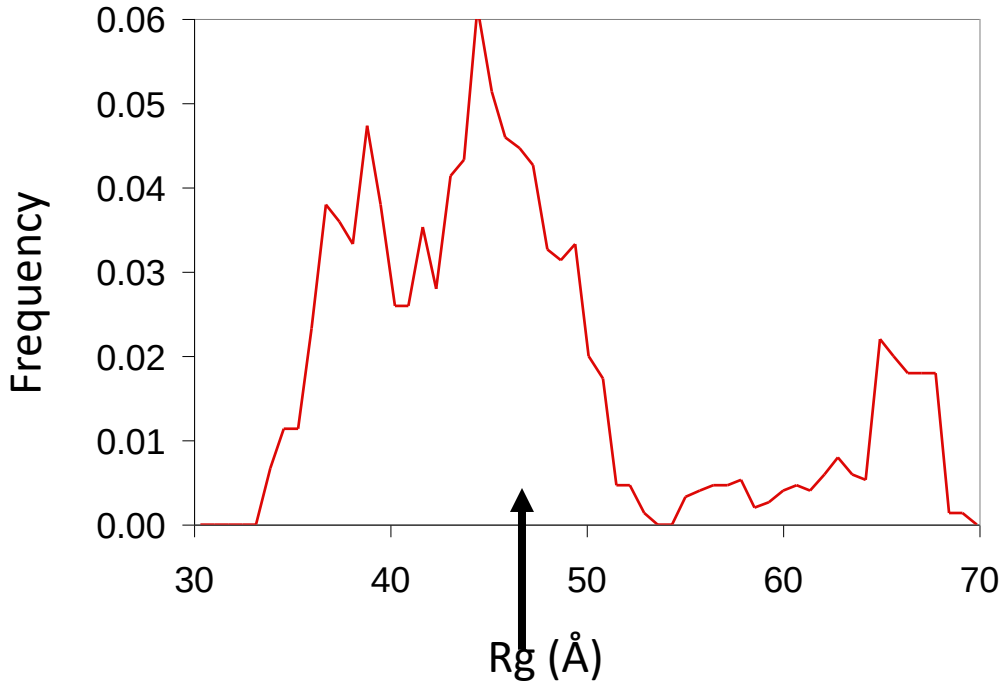


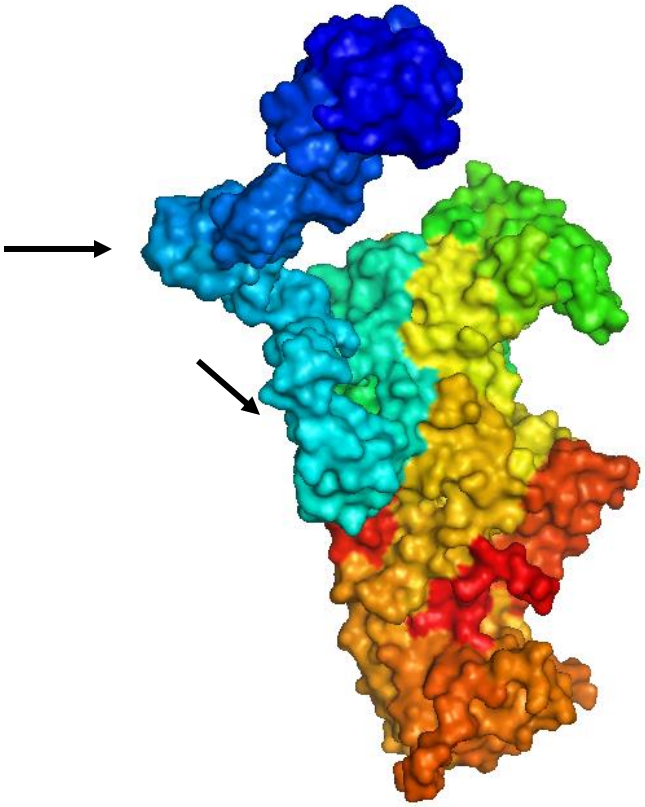

0.06
0.05
0.04
0.03
0.02
0.01
0.00
Frequency
30
40
50
60
70
Rg (Å)

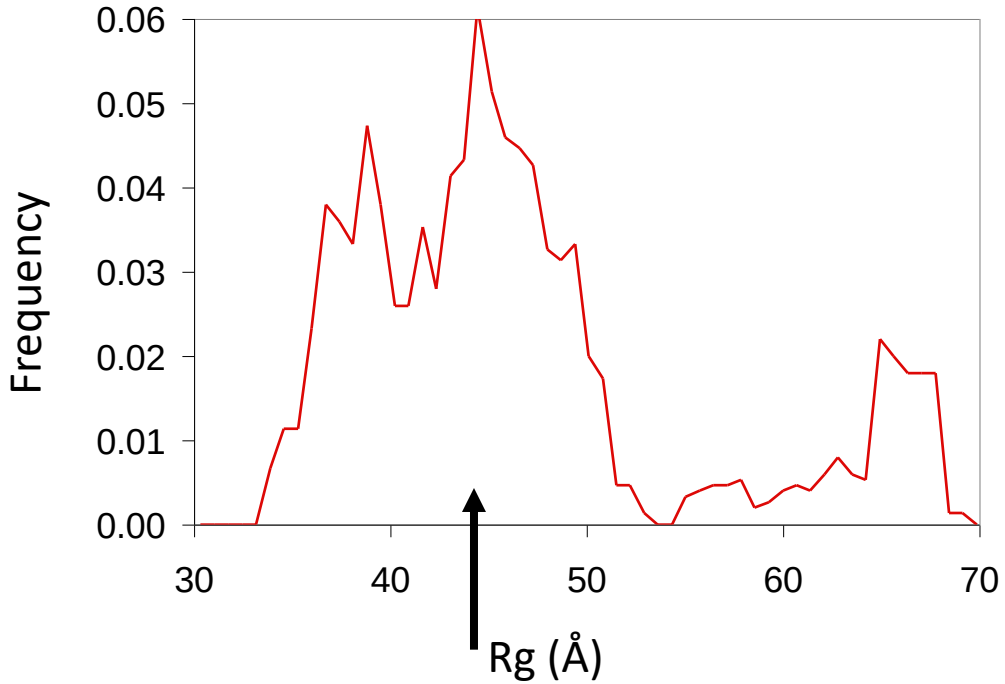


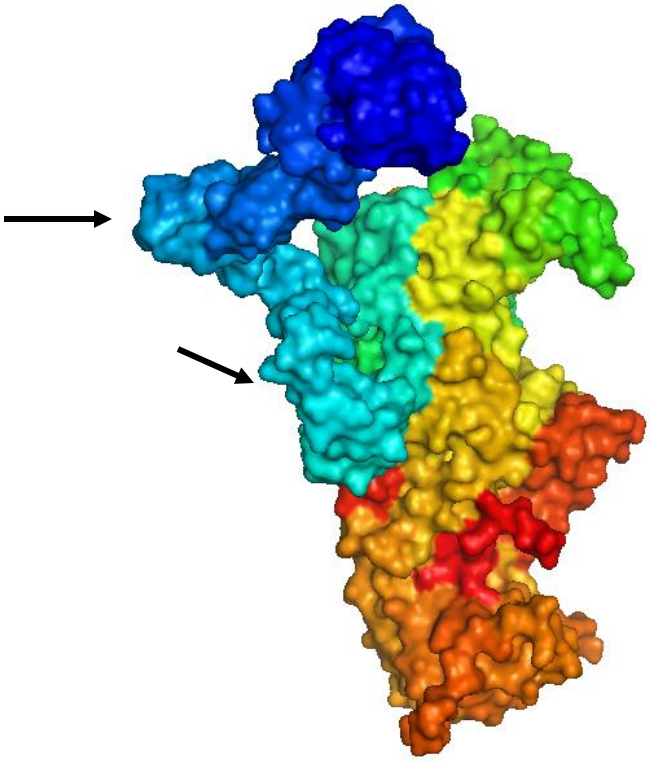

0.06
0.05
0.04
0.03
0.02
0.01
0.00
Frequency
30
40
50
60
70
Rg (Å)

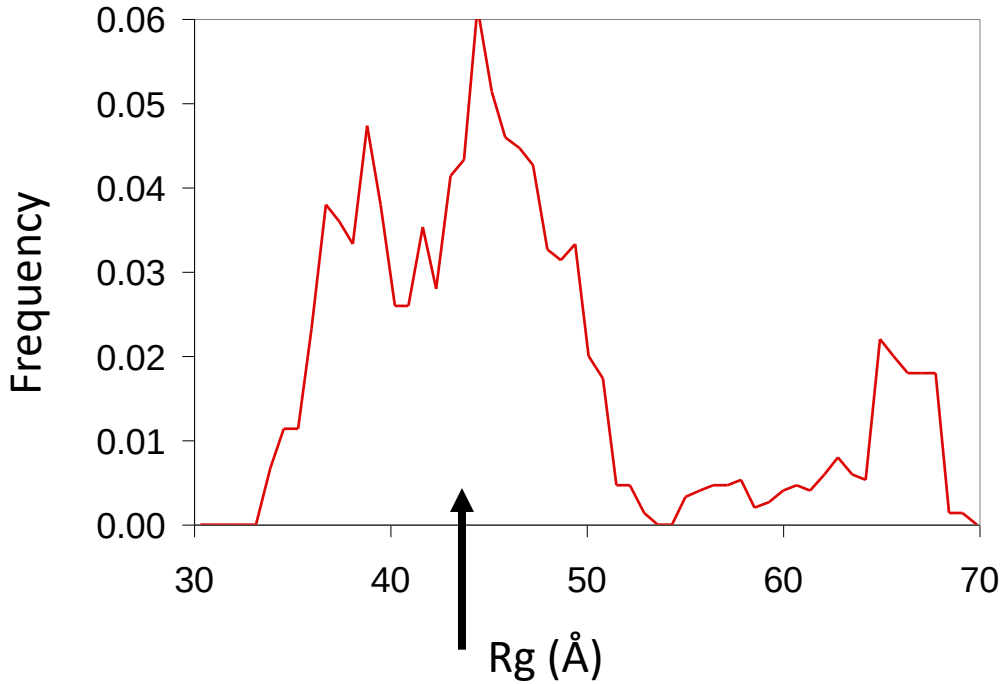


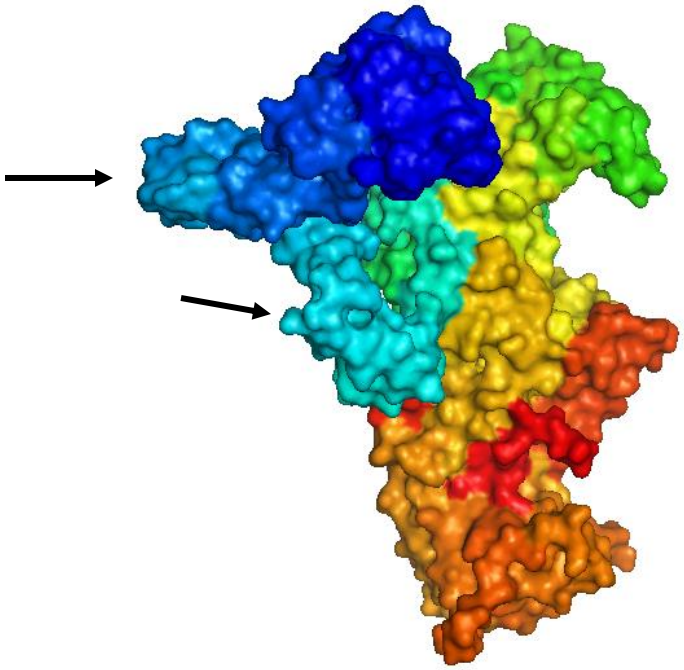

0.06
0.05
0.04
0.03
0.02
0.01
0.00
Frequency
30
40
50
60
70
Rg (Å)

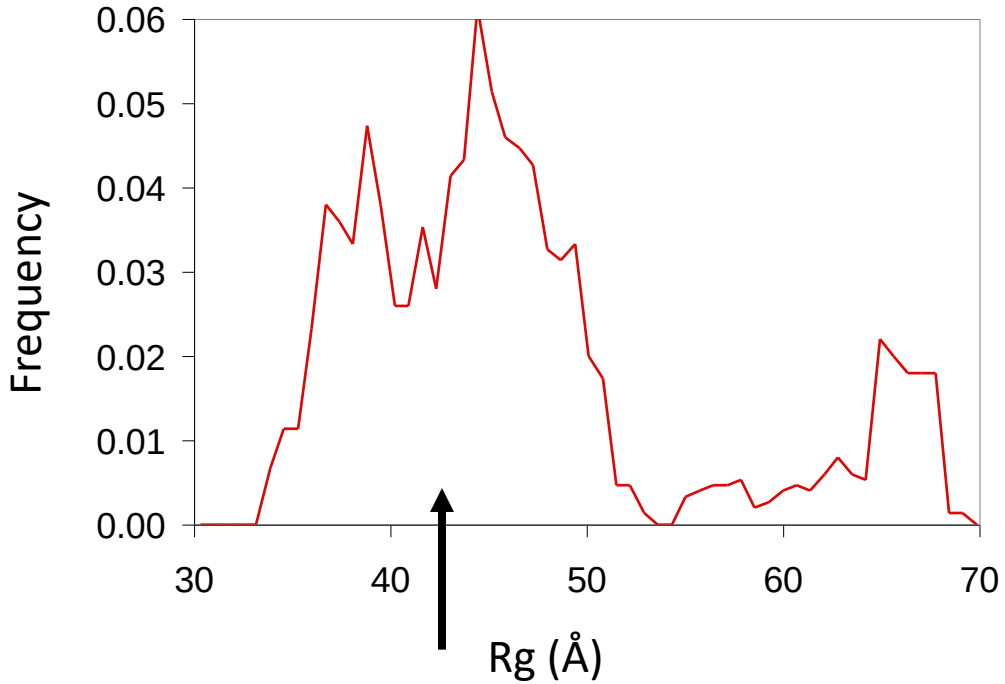


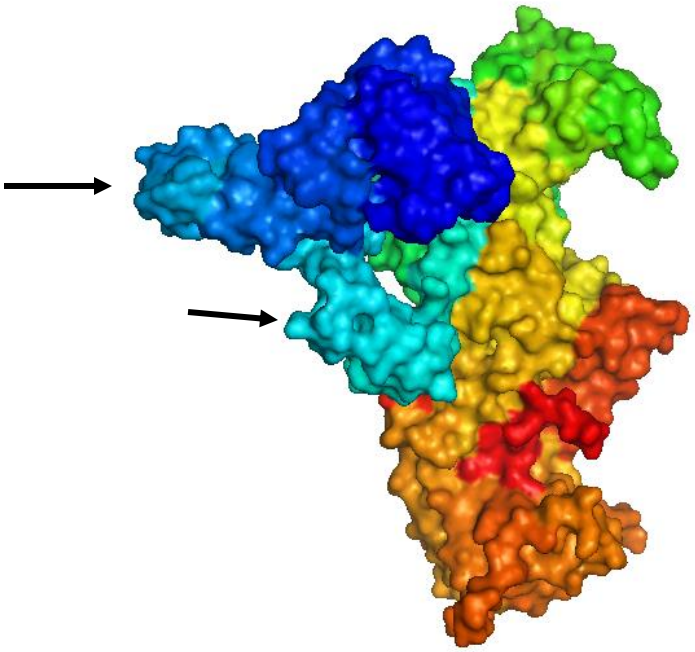

0.06
0.05
0.04
0.03
0.02
0.01
0.00
Frequency
30
40
50
60
70
Rg (Å)

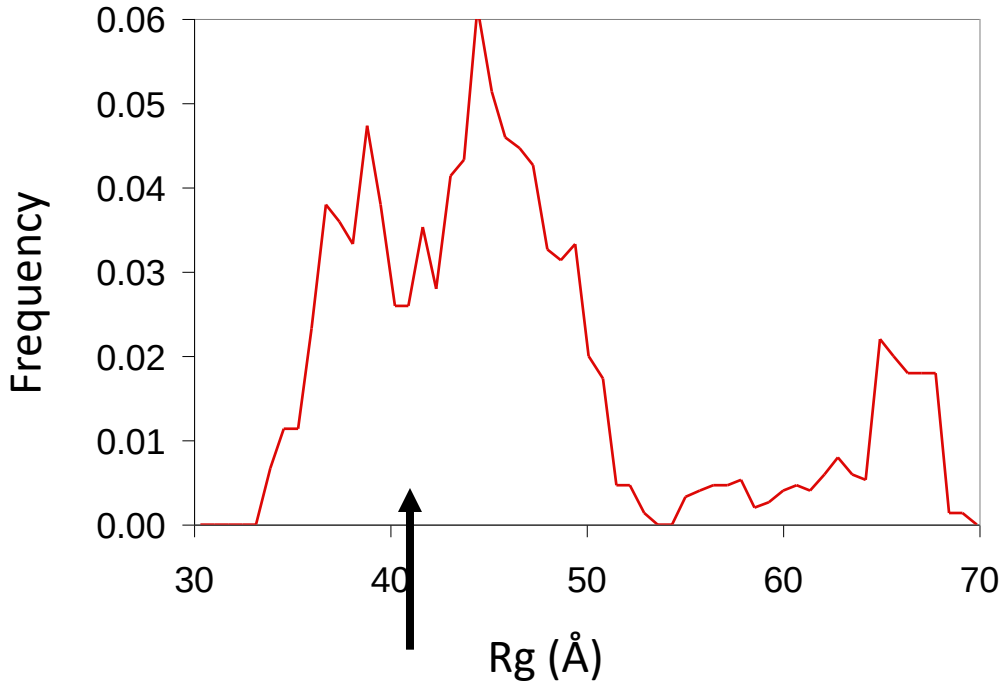


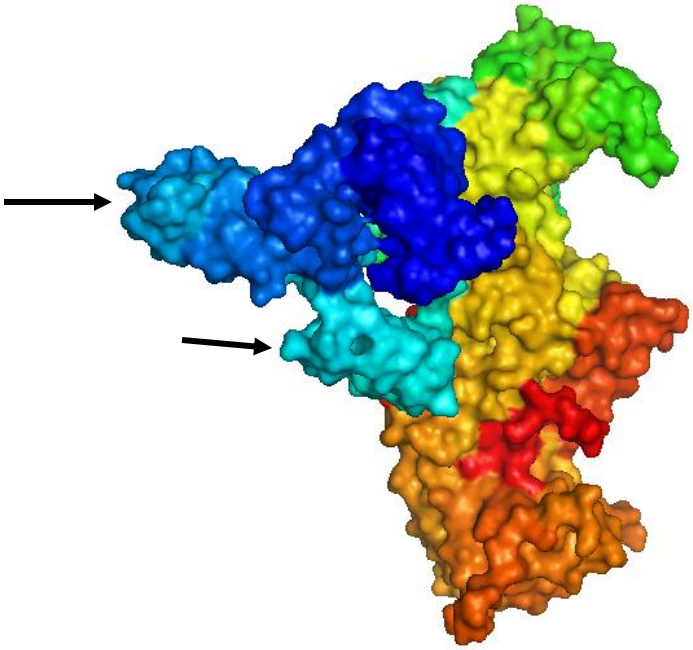

0.06
0.05
0.04
0.03
0.02
0.01
0.00
Frequency
30
40
50
60
70
Rg (Å)

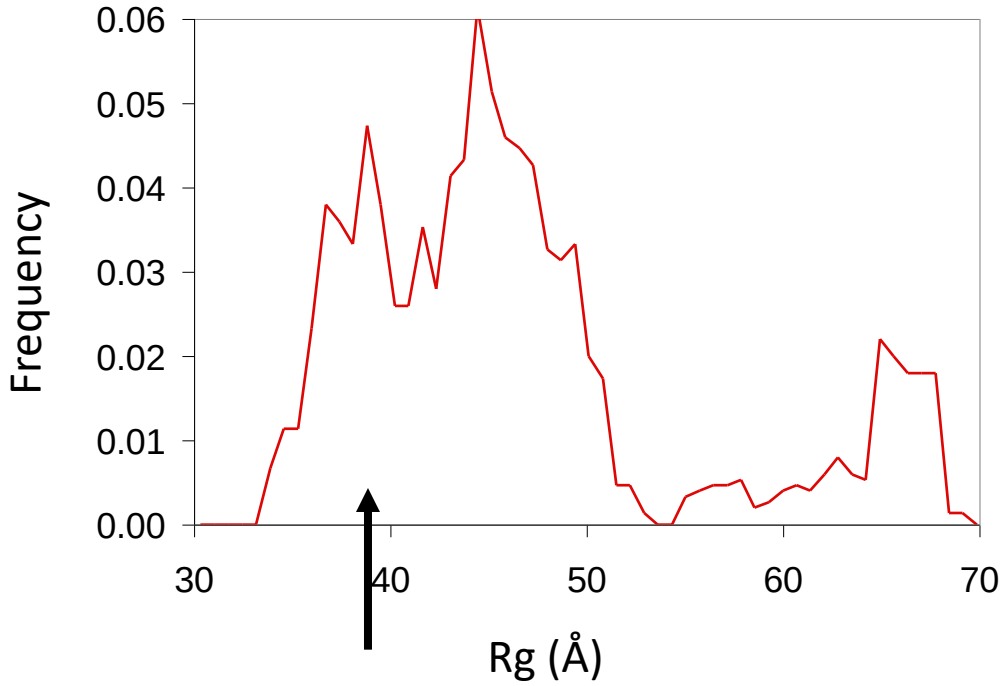


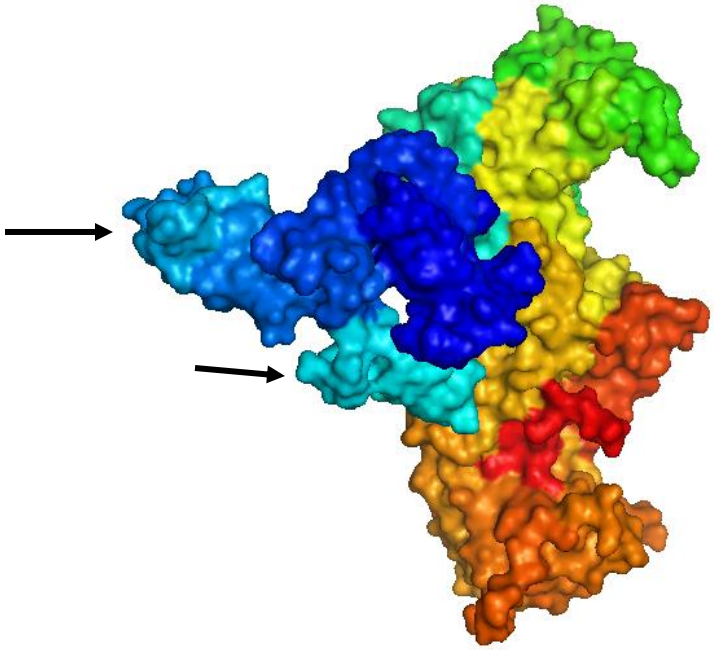

0.06
0.05
0.04
0.03
0.02
0.01
0.00
Frequency
30
40
50
60
70
Rg (Å)

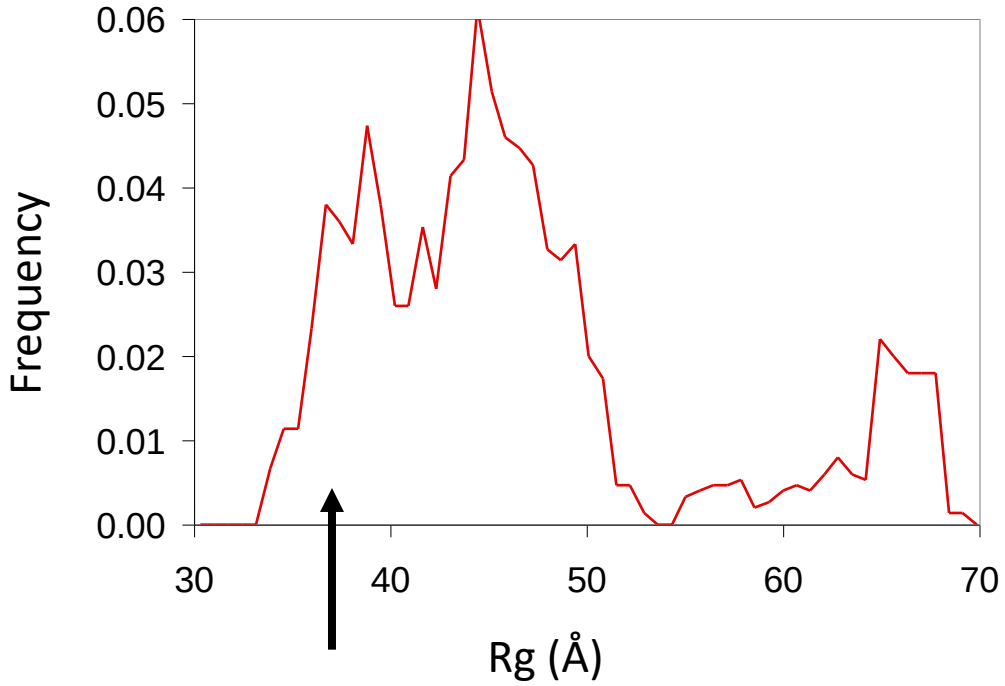


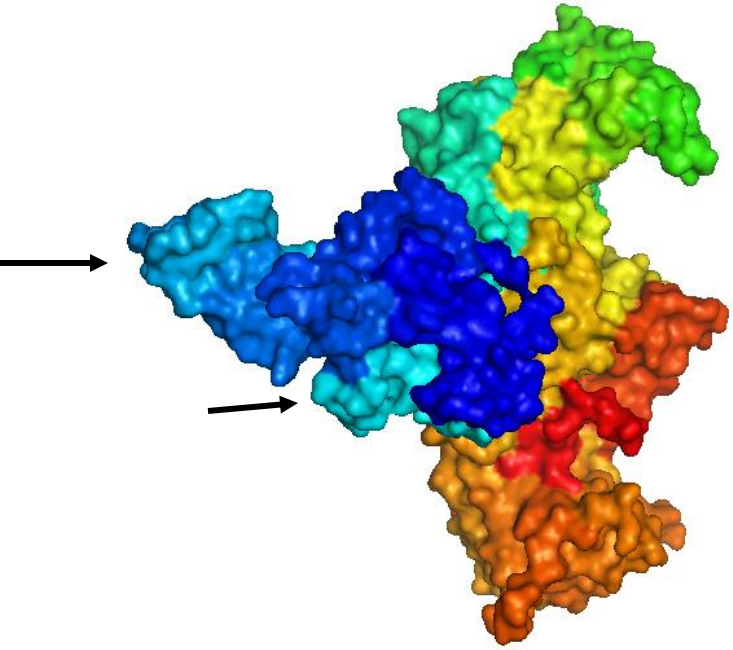

0.06
0.05
0.04
0.03
0.02
0.01
0.00
Frequency
30
40
50
60
70
Rg (Å)

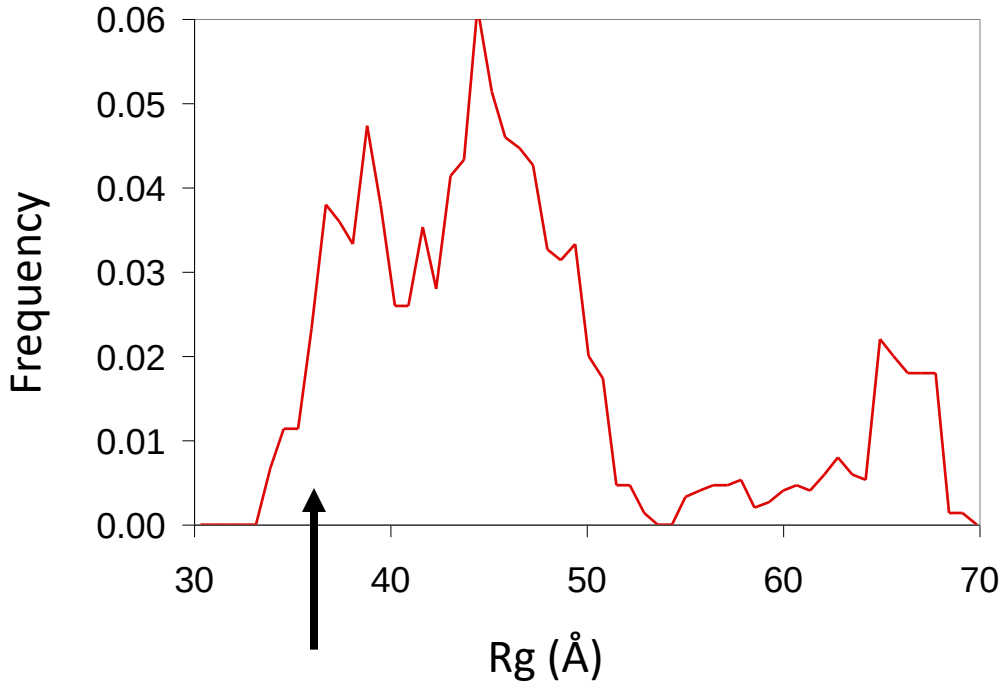


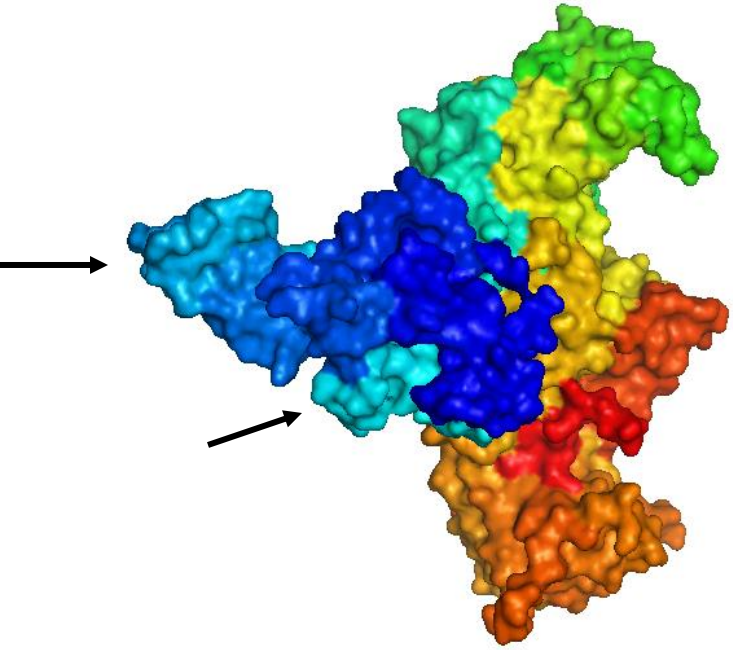

0.06
0.05
0.04
0.03
0.02
0.01
0.00
Frequency
30
40
50
60
70
Rg (Å)

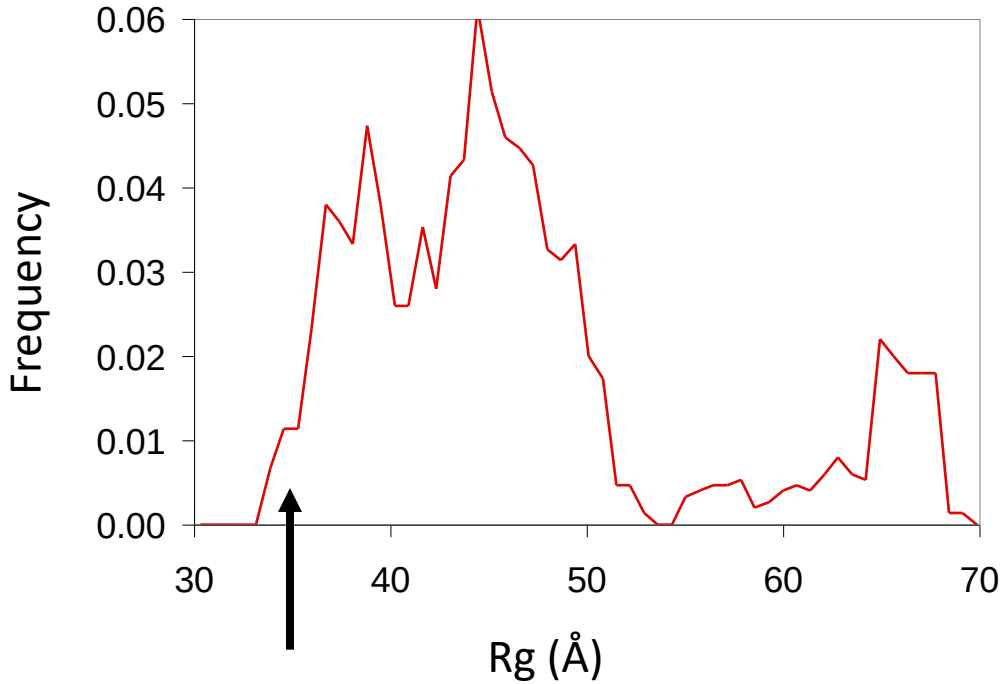



0.06
0.05
0.04
0.03
0.02
0.01
0.00
Frequency
30
40
50
60
70
Rg (Å)

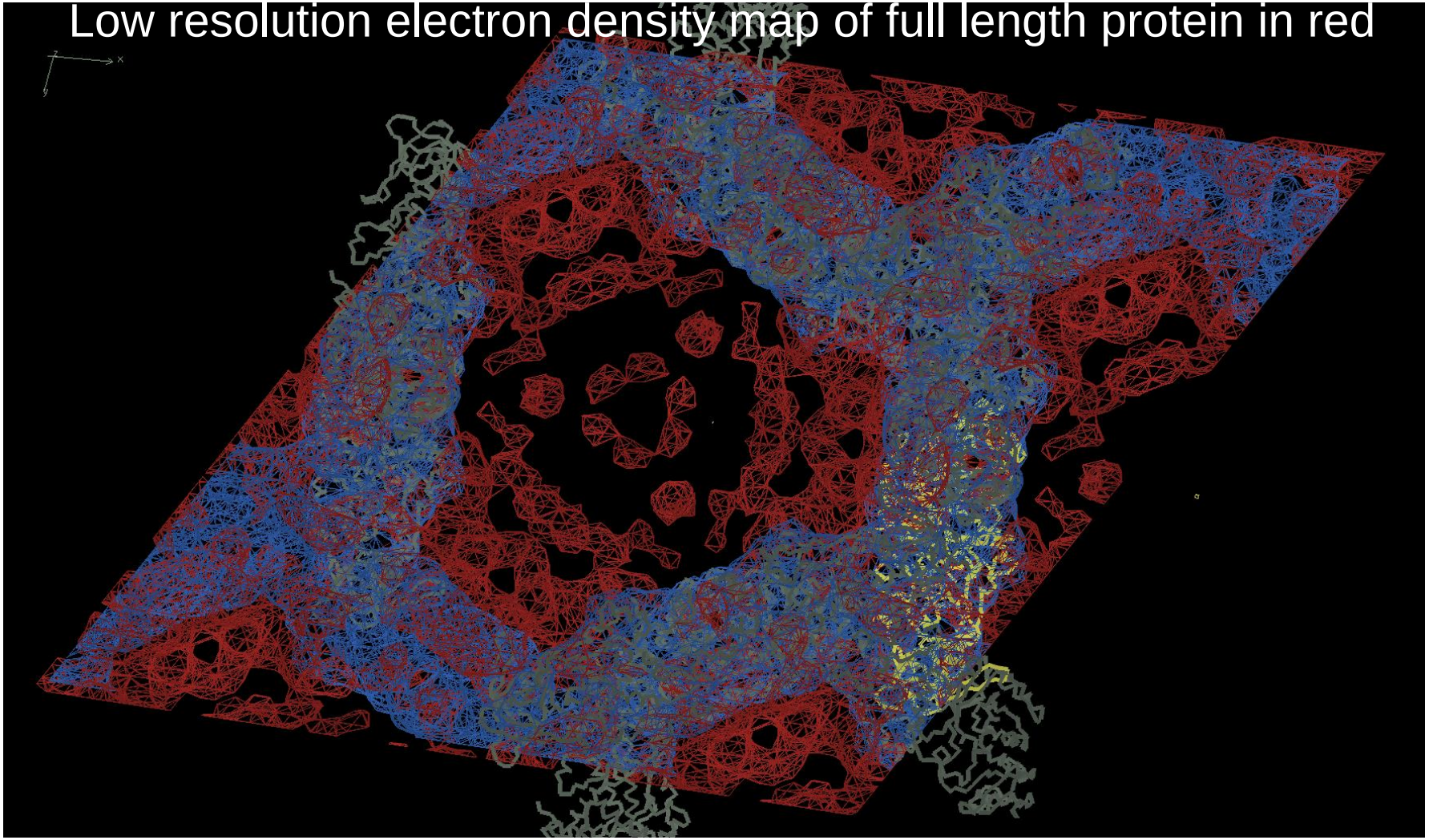



0.06
0.05
0.04
0.03
0.02
0.01
0.00
Frequency
30
40
50
60
70
Rg (Å)

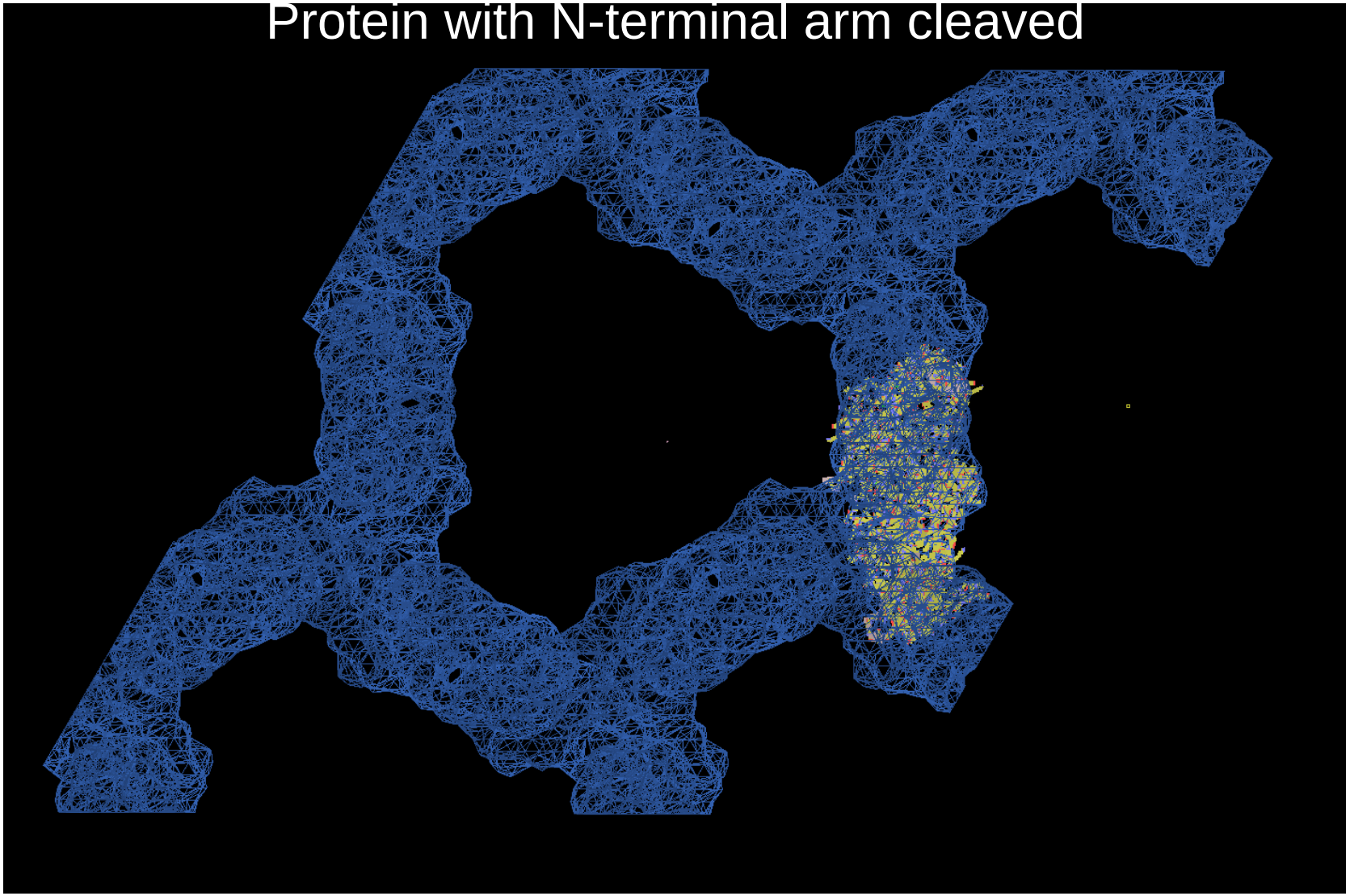



0.06
0.05
0.04
0.03
0.02
0.01
0.00
Frequency
30
40
50
60
70
Rg (Å)

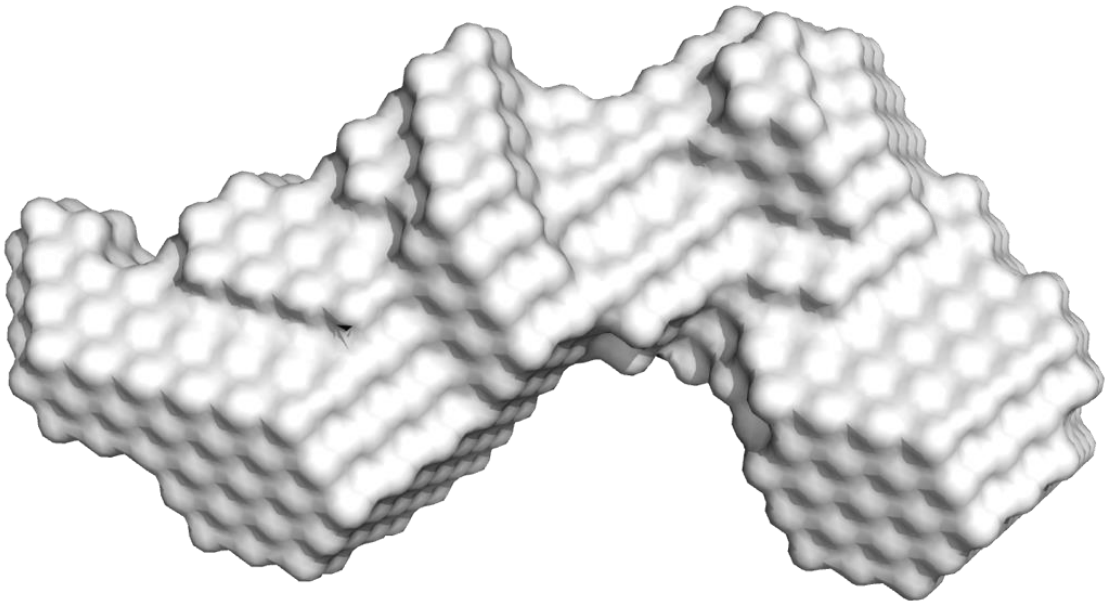
Low resolution electron density map of full length protein in red



Protein with N-terminal arm cleaved

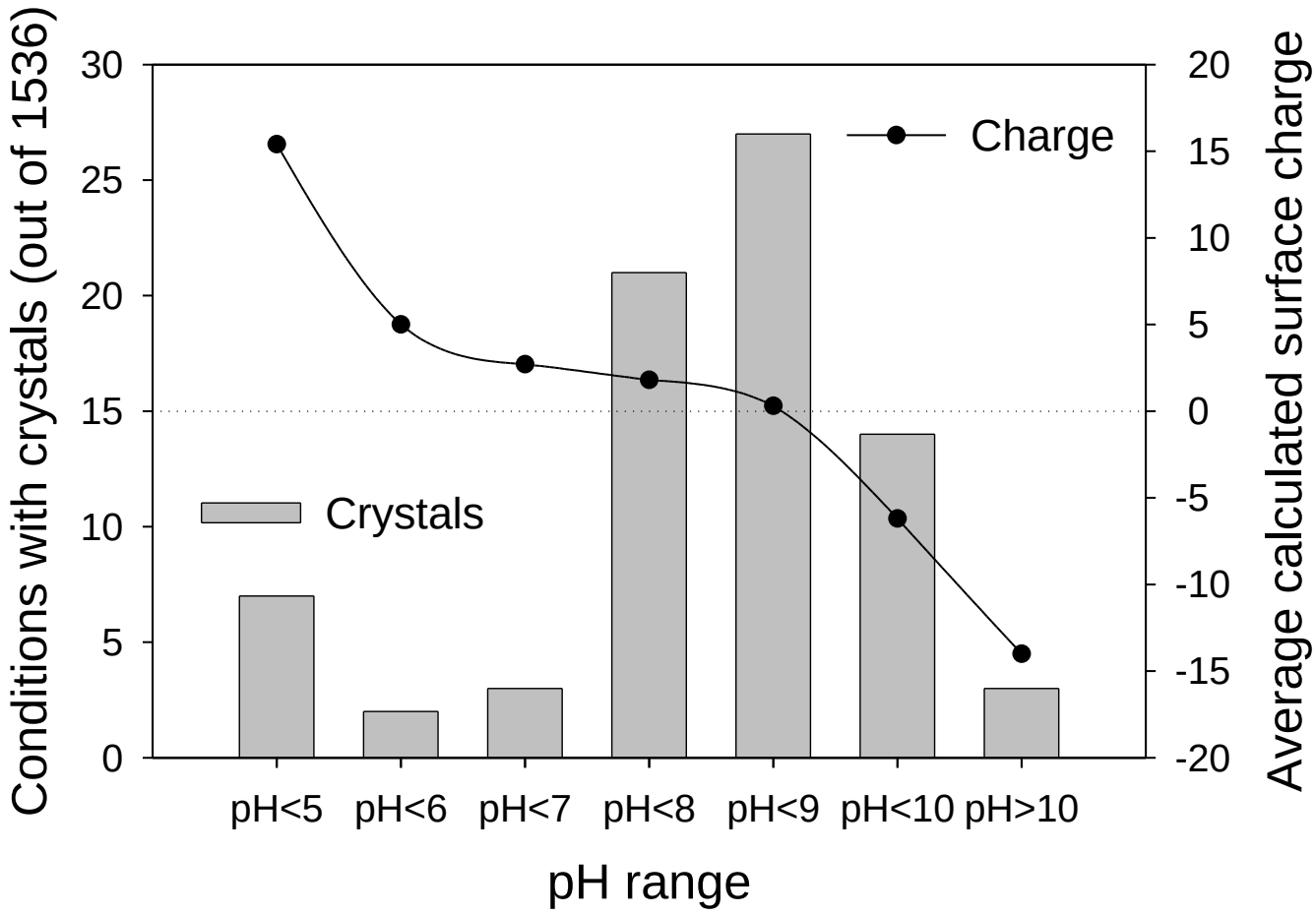


Envelope reconstruction of the N-terminal domain

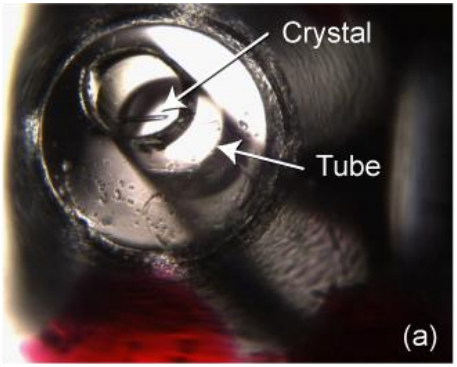


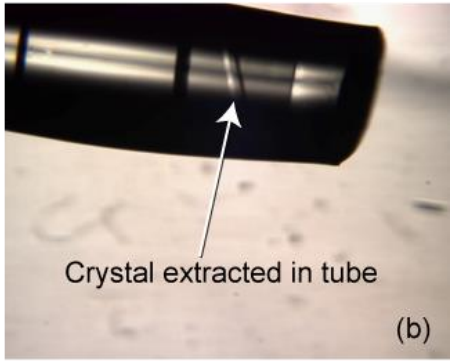
Back to our crystallography

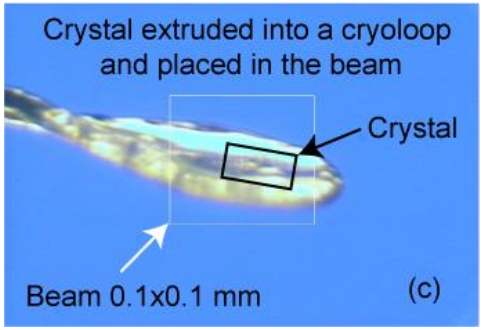
Crystallization trials of the N-terminal domain

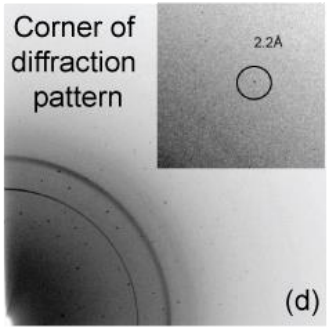


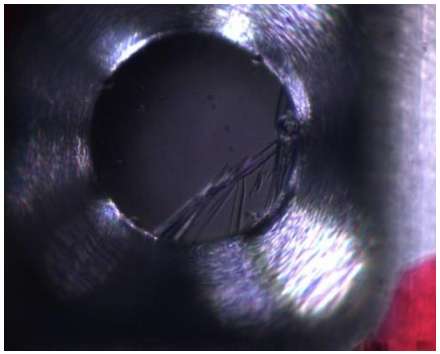
Does it diffract? Screening before the synchrotron

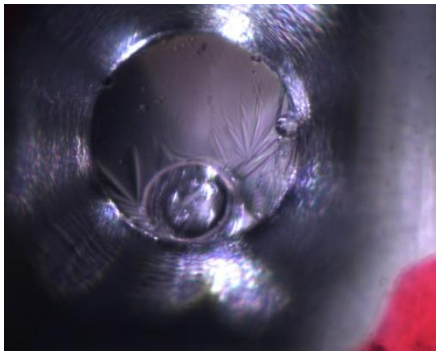


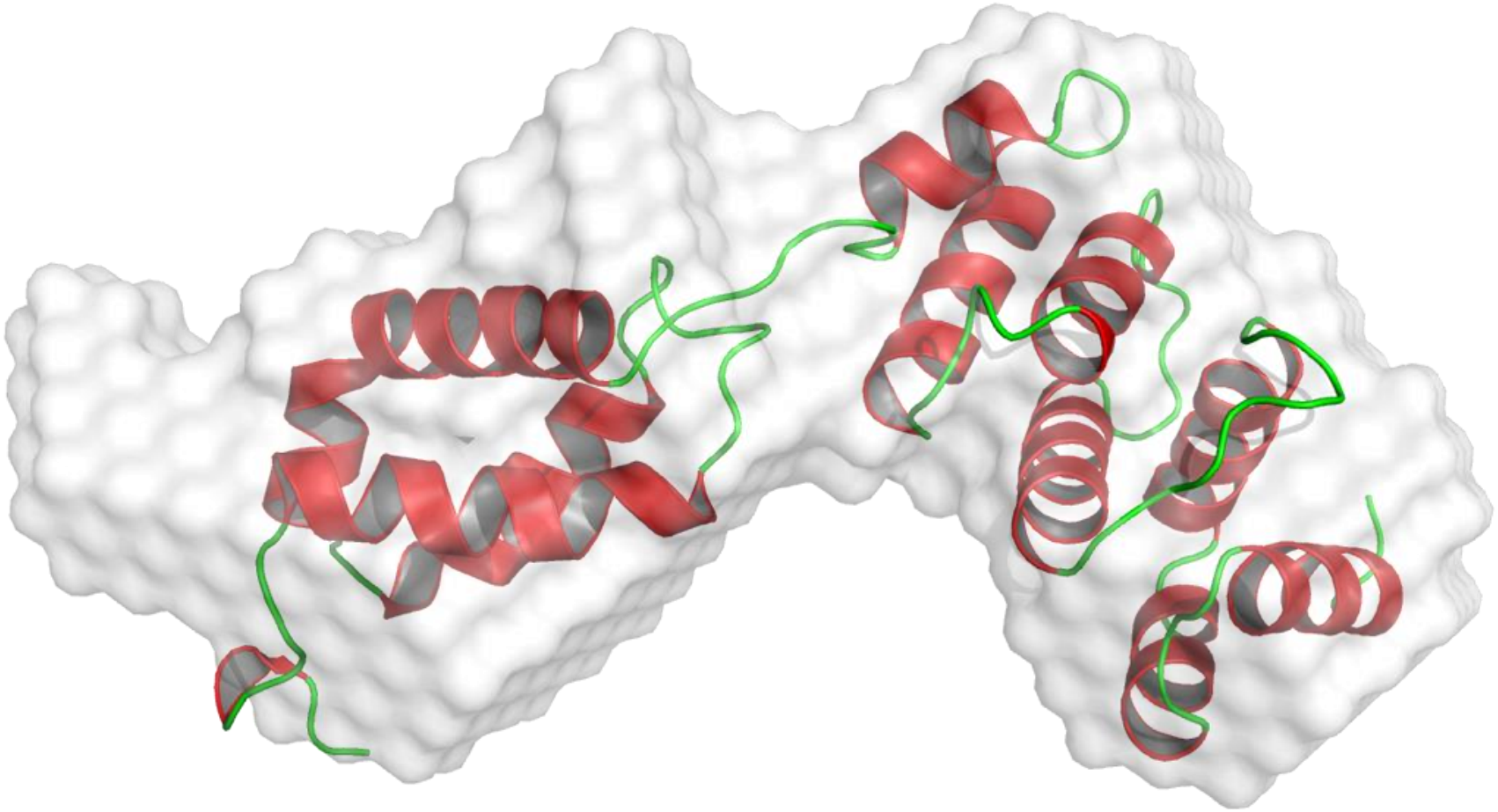


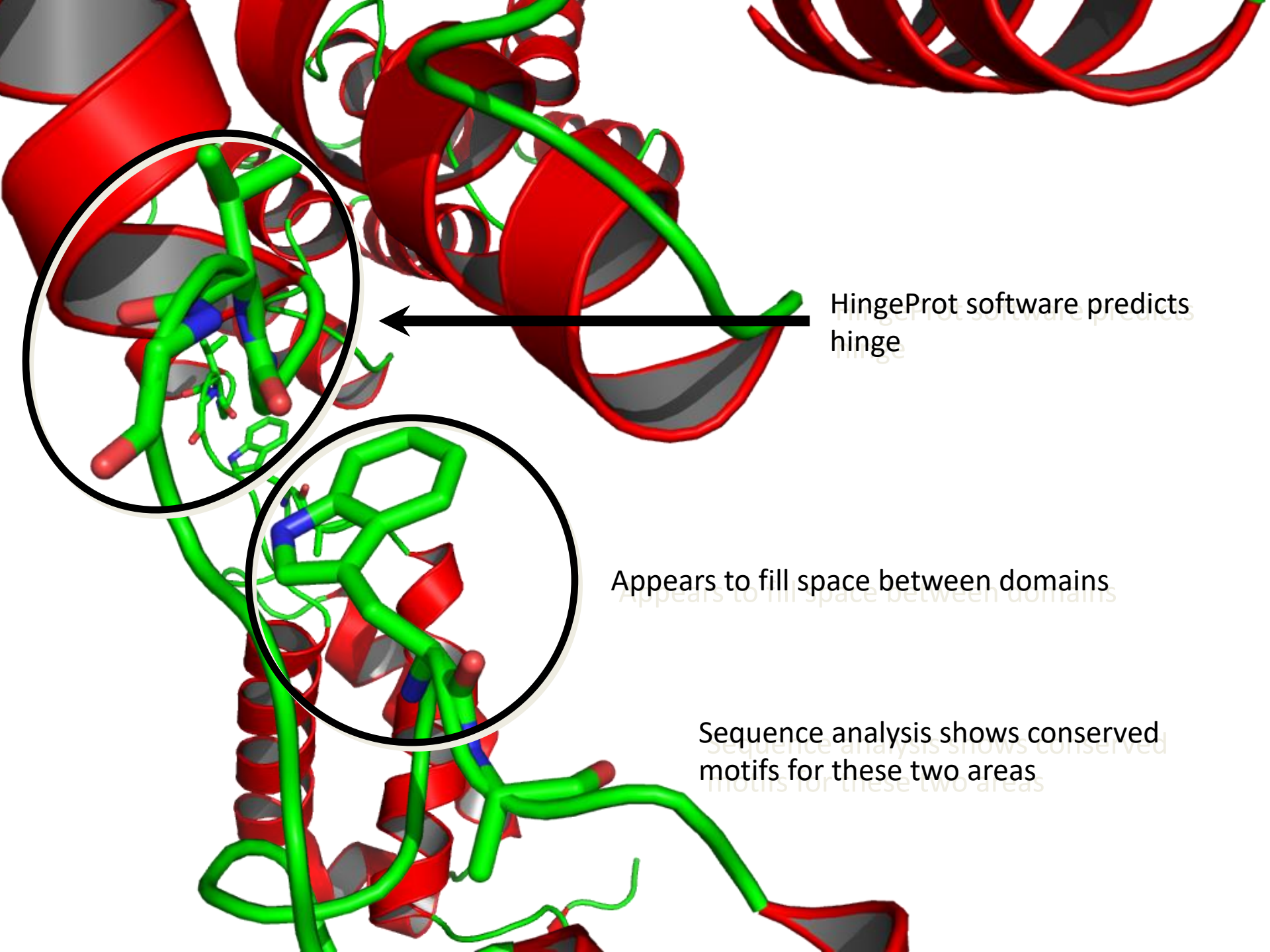


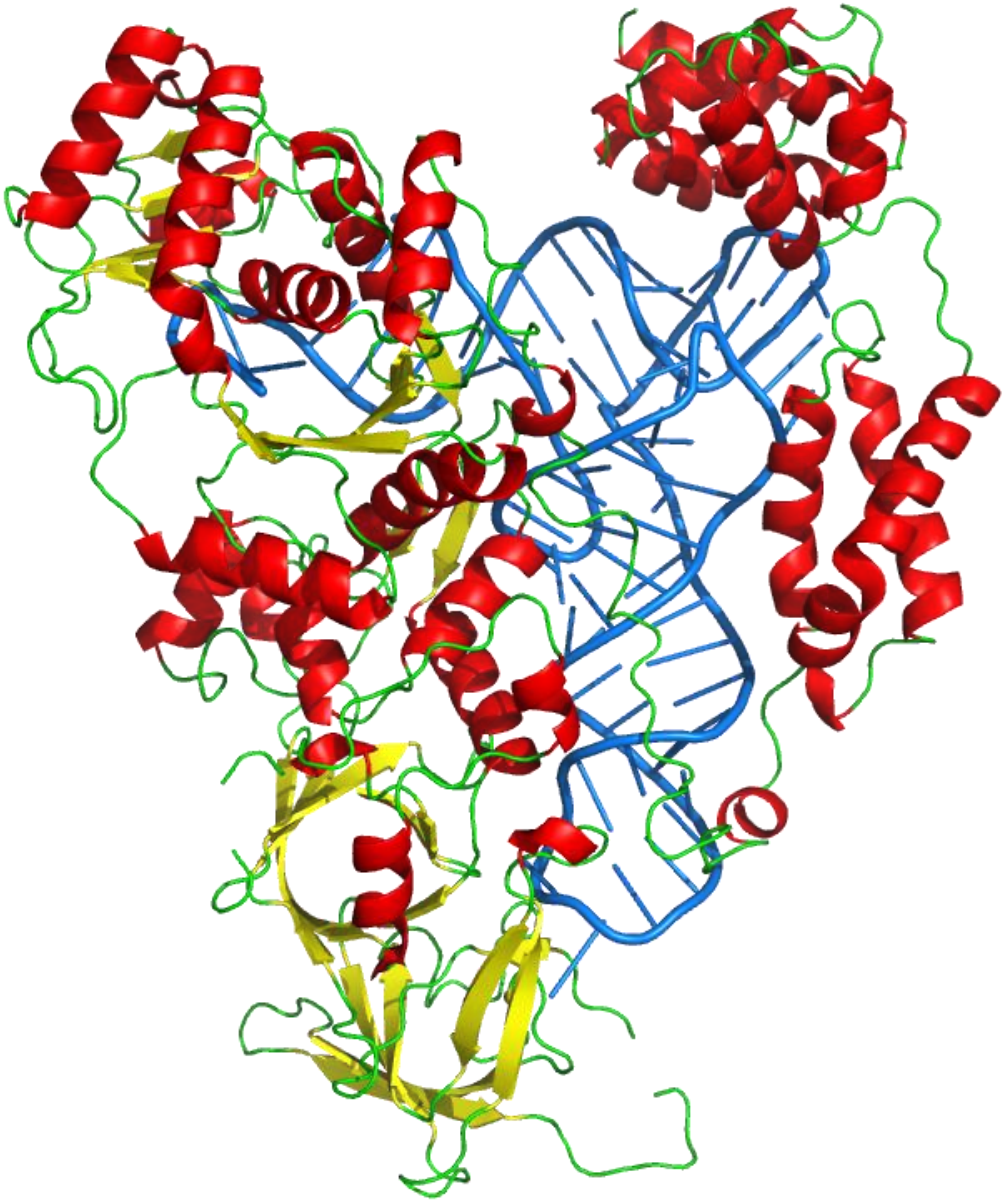









HingeProt software predicts hinge
Appears to fill space between domains
Sequence analysis shows conserved motifs for these two areas

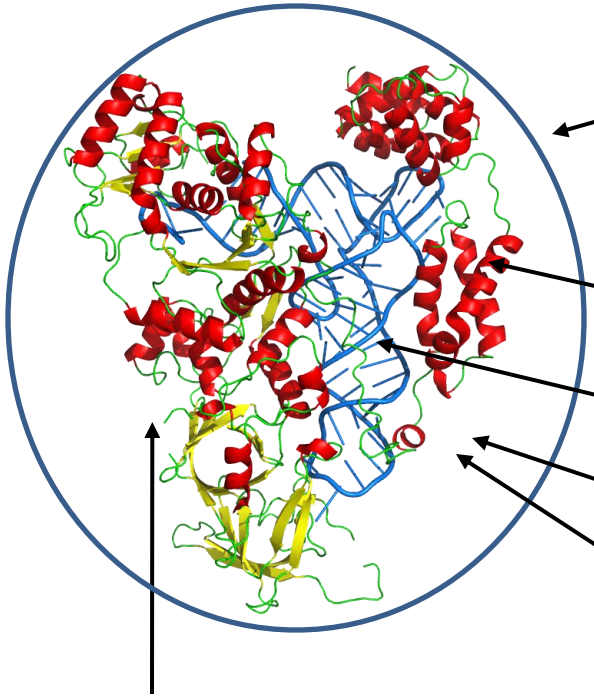


Structural Homologs

A blast search did not reveal structural homologs – having the structure of the N-terminal arm was critical.

Summary

Eukaryotic Gln tRNA synthetase



SAXS data indicating a larger but well folded system in solution

A Sherlock analysis indicated a preferential pH

The truncated terminal was crystallized

It was extracted directly from the screening plate and X-rayed to give the structure.

tRNA was docked in

SAXS aided by sequence analysis identified a flexible region

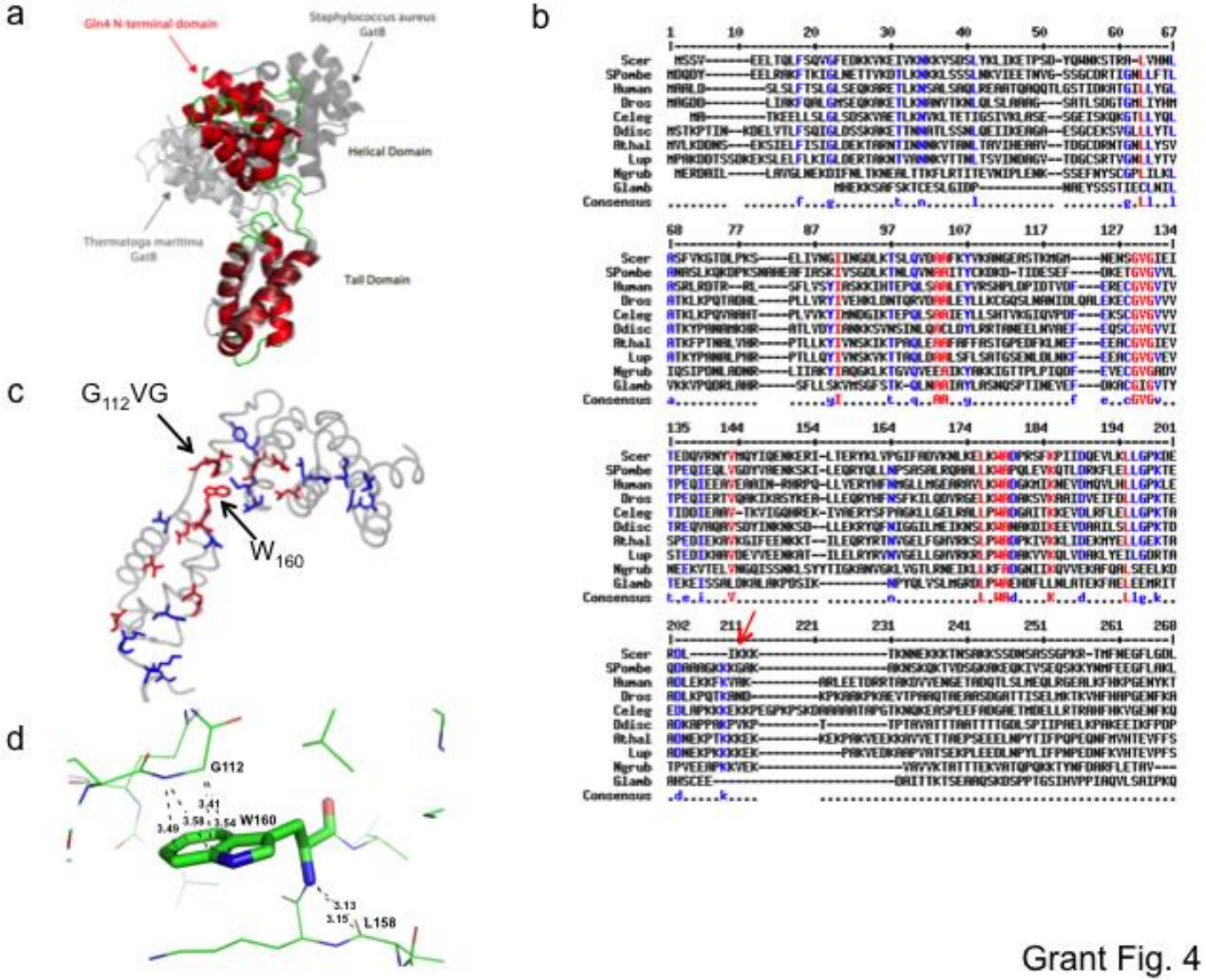
Homology modeling (FREAD) gave the flexible region

Crystallized the C-terminal in the standard screen, conditions chosen that were already known to be good cryo-conditions.

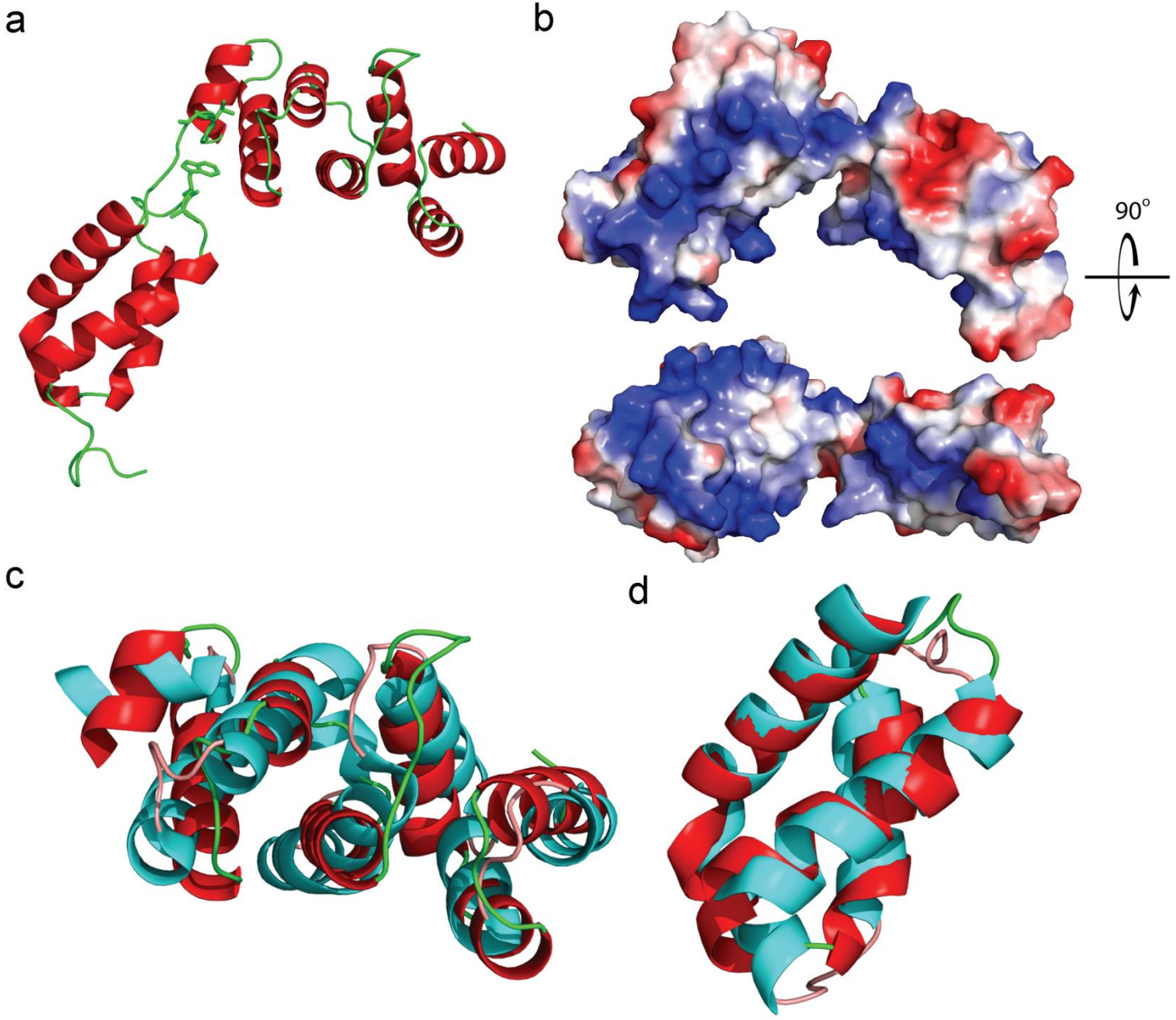
A combination of crystallography, SAXS, homology modeling and computational modeling was used to give the complete structure and tested by biochemical analysis.

Outline

- Biological dynamics
- Crystallography
- Small Angle X-ray Scattering
- A biological case
- A structure, a puzzle, a question?
- A solution
- Computational analysis
- **An answer, more questions, more solutions**
- Where are we going next?


a
Gln4 N-terminal domain
Staphylococcus aureus GatB
Helical Domain
Thermatoga maritima GatB
Tail Domain
b
c
$G_{112}VG$
W_{160}
d
G112
3.41
3.58
3.54
3.49
W160
3.13
3.15
L158

Grant Fig. 4



Grant Fig. 3