SAS data: Modelling.

Cy Jeffries, EMBL Hamburg

EMBL
40 YEARS | 1974–2014

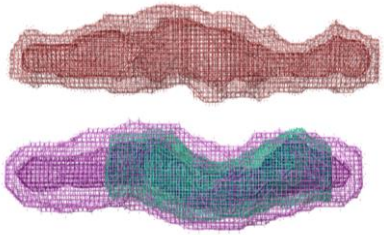
An obsession with models

Complexes and assemblies

S-layer proteins

Fagan *et al* Mol. Microbiol (2009)

α-synuclein oligomers



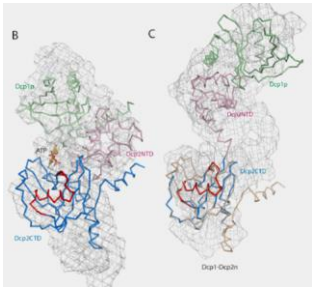
Giehm *et al*
PNAS USA (2011)

insulin fibres

Vestergaard *et al PLoS Biology* 2007

Domain and quaternary structure

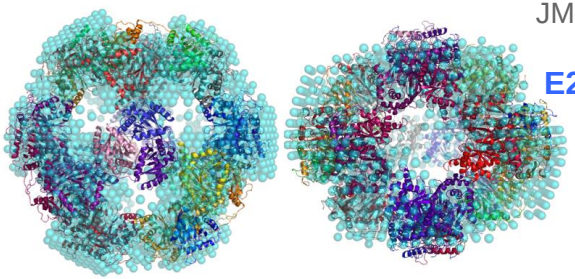
Dcp1/Dcp2 complex



She *et al*, Mol Cell (2008)

Toxin B

Albesa-Jové *et al*
JMB (2010)

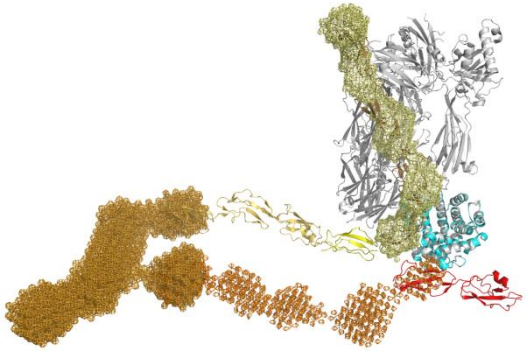


E2 multienzyme complex

Marrott *et al*
FEBS J. 2012

Flexible/transient systems

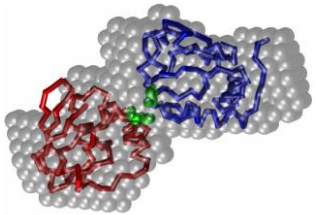
Complement factor H



Morgan *et al*
Nature Struct. Mol. Biol. (2011)

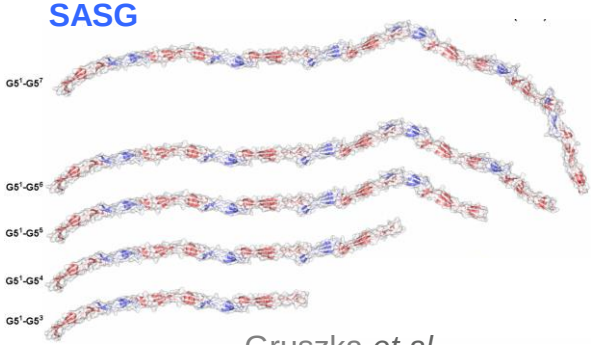
Structures and structural transitions

Cytochrome/adrenodoxin

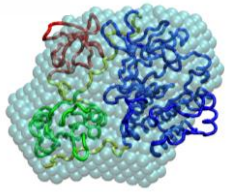


Xu *et al*
JACS (2008)

SASG



Gruszka *et al*
Nature Comm. (2015)



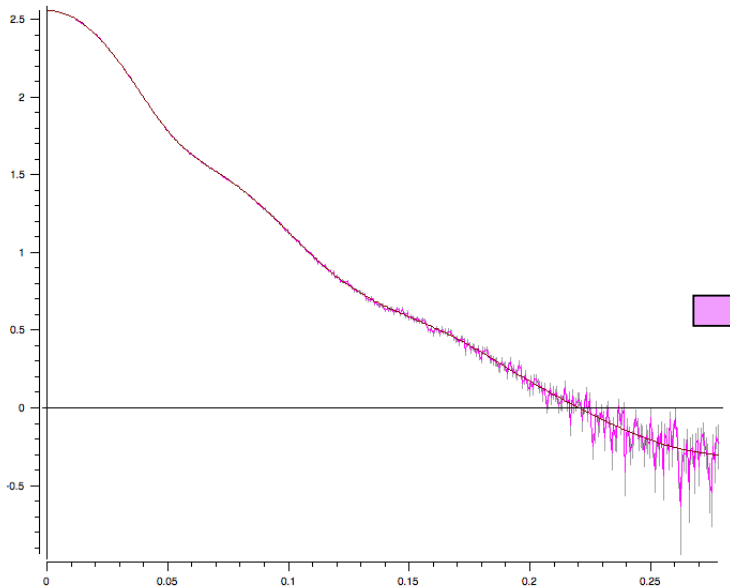
Bernado *et al*
JMB (2008)

Src kinase

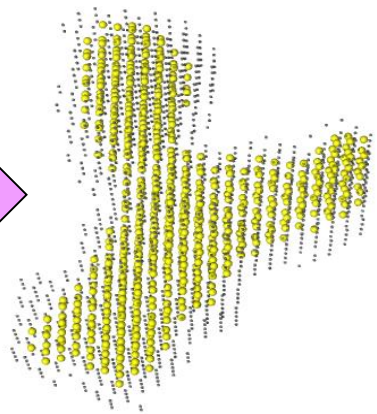
EMBL

Modelling is a multi-dimensional problem

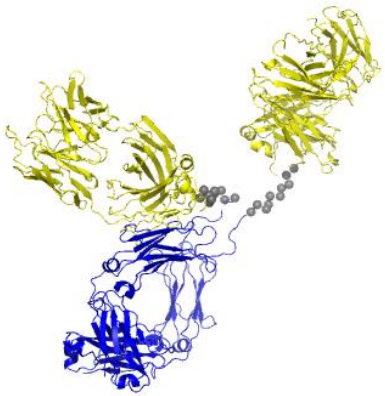
1-D Scattering profile in reciprocal space



3-D model in real space



Ab initio model



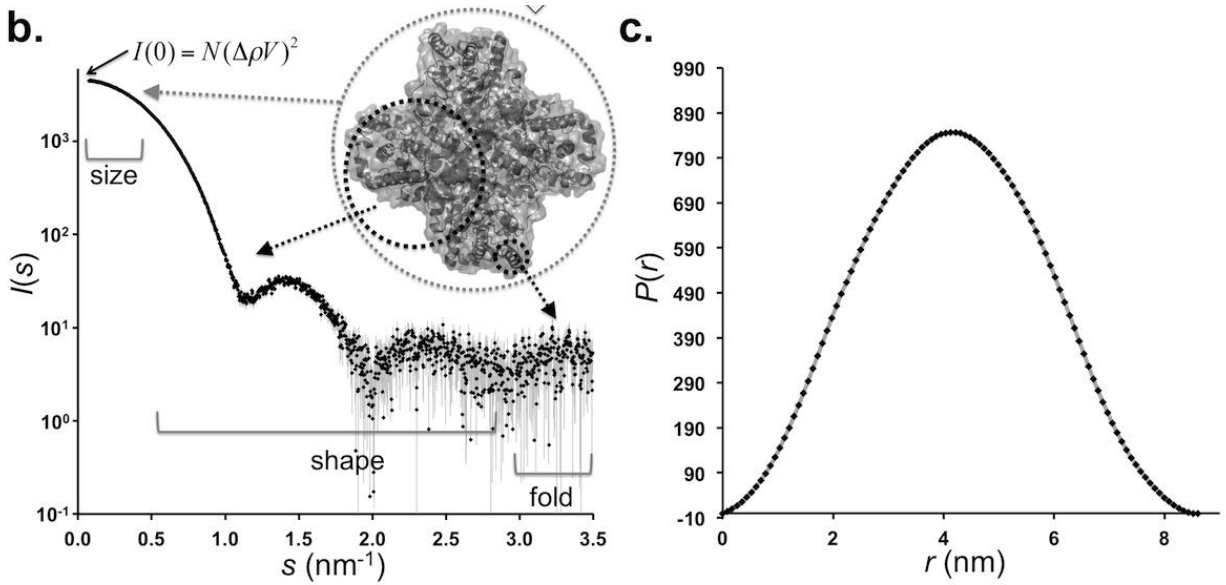
Atomistic model

Loss of orientation information due to spherical averaging of the scattering amplitudes from the time and ensemble isotropic tumbling of particles in solution.

x,y,z spatial coordinates, i.e., orientation information is restored.

Does anyone see the inherent problem?

- An issue of ‘resolution’ – you cannot make conclusions about resolution an atomic level of detail.
- An issue of ambiguity. Different kinds of structures can produce identical shape scattering.



$$I(s) = N(\Delta\rho V)^2 P(s) \quad \Rightarrow \quad I(s) = 4\pi \int_0^{D_{max}} P(r) \frac{\sin(sr)}{sr} dr$$

reciprocal-space form factor scattering → *real-space probable distance distribution*

Does the model fit the data?

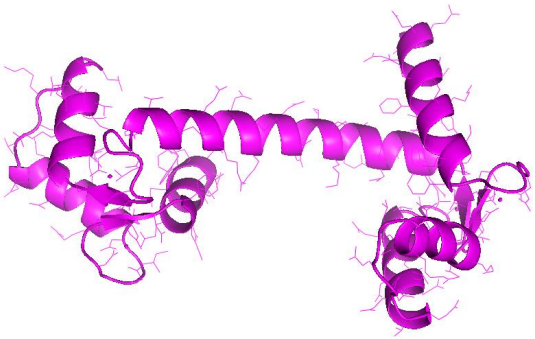
Are you *really sure* the model fits the SAS data?
(– this is a trap question.)

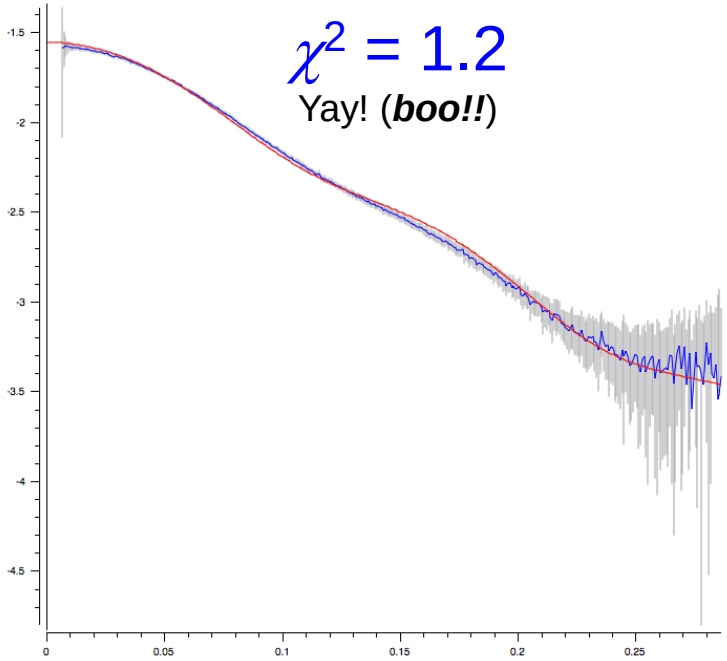
Yes! The model fits!
…but…maybe there is an alternative (oops).

…you (really want to) ***BELIEVE*** the model fits – even though it does not…

?

SAS is about *describing* what is going on in solution (in terms of structural biology) within the limitations of the data and the limitations of the models in context of the experiments and the question!





Modelling – main points

- Understand the data (understand the sample and the instrument).
- Model the data without modelling it (??)
- *Ab initio* bead or dummy residue modelling – (Refer to C. Blanchet lecture.) Advantage = few assumptions.
- Rigid-body modelling. Automatically introduces bias – but this is not a bad thing!

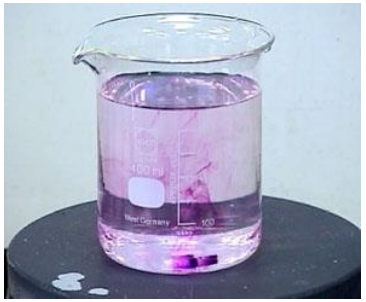
Modelling a three-dimensional (3D) shape of a protein derived from a 1D scattering pattern representing a rotationally and time-averaged sample is not trivial.

Lets quickly recap what scattering data is…

Solution small-angle scattering (SAS)

X-ray crystallography:

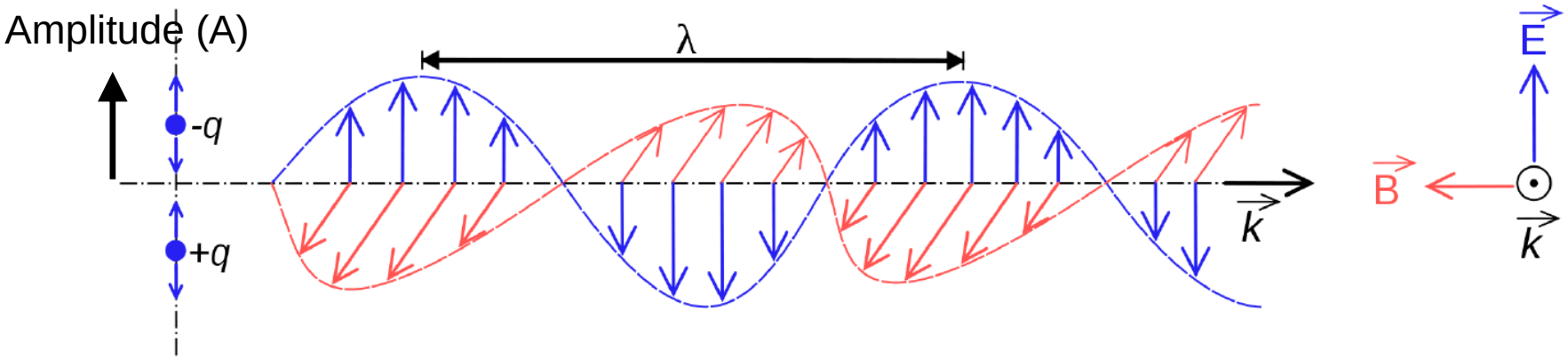
SAS:



- Amplification of intensities (*I*) via SCATTERING through a Bragg-lattice (i.e., a diffraction grid). ***Strong signal.***
- Ordered distances between unit-cells. ***High spatial resolution***.
- ***Directional information*** – convert to x,y,z in real space.

- No amplification of intensities through a grid, just scattering from ***ALL*** electrons (for SAXS) or ***ALL*** atomic nuclei (for SANS) in the illuminated volume. In solution = water! ***Very weak signal.***
- ***Loss*** of spatial resolution + directional information – isotropic tumbling in solution through time: Intensities = the sum of the time and rotationally averaged scattering from *each* particle (e.g., protein) in solution.

What is SAXS?



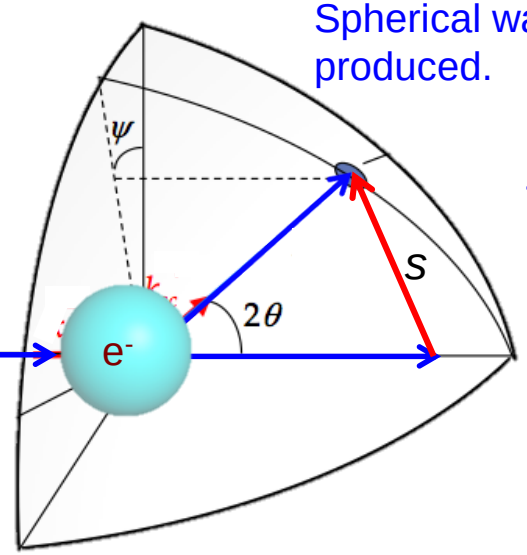
- At the X-ray energies used for SAXS (4-20 keV) X-rays primarily interact with **electrons***.*
- *Three main outcomes when illuminating a sample with X-rays*
- 1) Nothing happens – straight through (= Transmission)
- 2) Get absorbed and re-emitted at a different wavelength, e.g., fluorescence.
- 3) ***Scatter***. Elastic scattering = NO CHANGE IN ENERGY.

  Inelastic scattering = CHANGE in energy.

- PROBABILITY: The probability of a ***SINGLE*** electron to scatter an X-ray through a solid angle in a given time = (differential) **cross section**.
- Cross section can be conceptualized as a 'probability circle' with a radius. The radius relates to the X-ray **scattering length** of the electron. It is essentially a measure of the potential of the electron-X-ray interaction.

Spherical wave front is produced.

Incident electromagnetic field, i.e., an X-ray of wavelength λ.



The modulus of the scattering vector *s*, or momentum transfer.

$$s = \frac{4\pi \sin\theta}{\lambda}$$

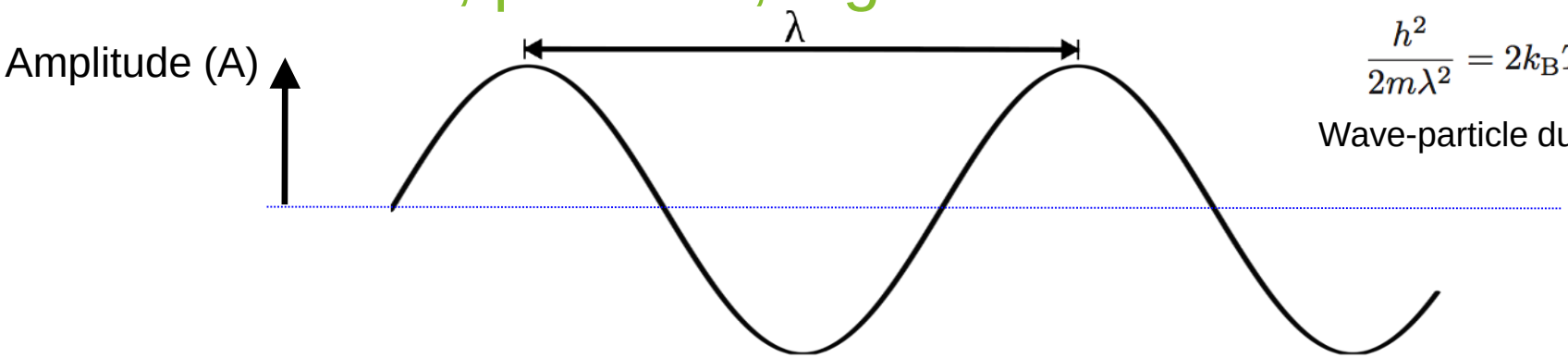
'Harmonic oscillation' of the electron in the field (di-pole oscillation) = momentum transfer to the photon *without* a loss of energy= elastic scattering

!! *s* can be defined in a number of ways!!

$Q = q = h = \mu = k = s$

Sometimes; $S = 2\sin\theta/\lambda = 2\pi s$

SANS - waves, particles, nightmare!



$$\frac{h^2}{2m\lambda^2} = 2k_{\mathrm{B}}T,$$

Wave-particle duality

- For SANS - neutrons primarily interact with **atomic nuclei**.
- *Three main outcomes when illuminating a sample with neutrons*
- 1) Nothing happens – straight through (= Transmission)
- 2) Get absorbed.
- 3) ***Scatter***. Elastic scattering = NO CHANGE IN ENERGY.

  Inelastic scattering = CHANGE in energy.

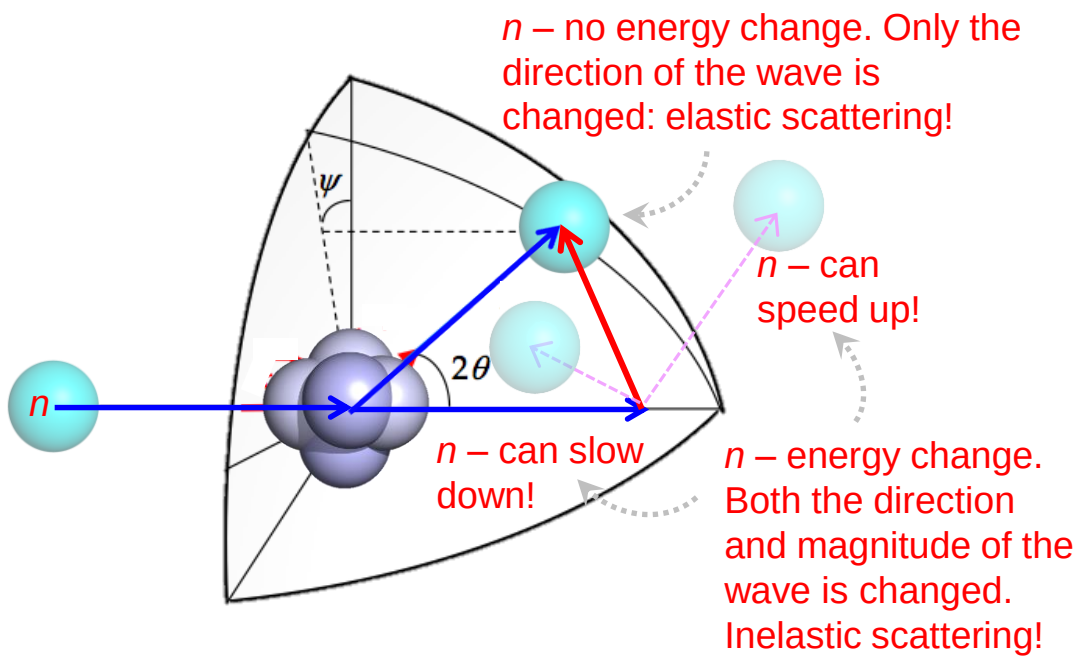
  Coherent scattering.

  Incoherent scattering.

  Combinations thereon!

  ‘Magnetic’ scattering.

- PROBABILITY: The probability of a ***SINGLE*** nucleus to scatter a neutron through a solid angle in a given time = (differential) **cross section**.
- Cross section can be conceptualized as a ‘probability circle’ with a radius. The radius relates to the neutron **scattering length** of the specific nucleus. It is essentially a measure of the potential of the nucleus-neutron interaction.



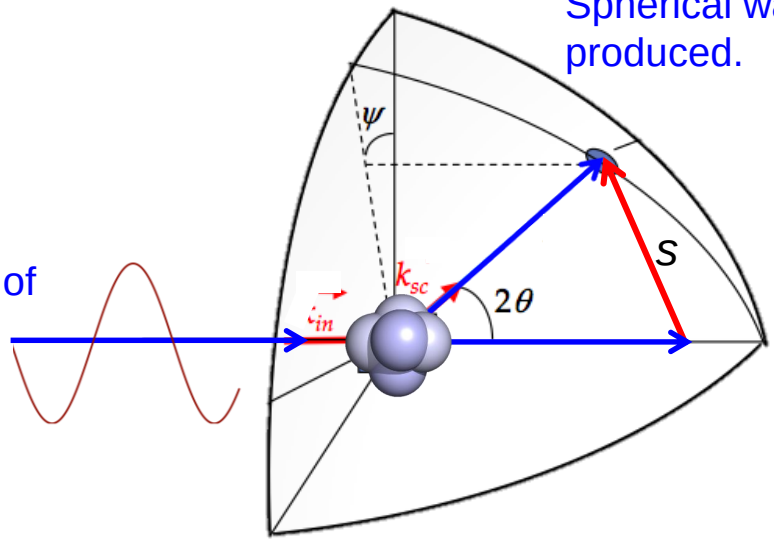
Point source – *s*-wave scattering (i.e., spherical)

For SANS we are after the (coherent) *elastic* scattering component.

Spherical wave front is produced.

The modulus of the scattering vector *s*, or momentum transfer.

Incident neutron of wavelength λ.



$$s = \frac{4\pi \sin\theta}{\lambda}$$

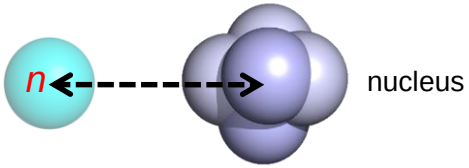
Potential interaction between the compound neutron and nucleus in the nuclear resonance field = momentum transfer to the neutron *without* a loss of energy= elastic scattering

!! *s* can be defined in a number of ways!!

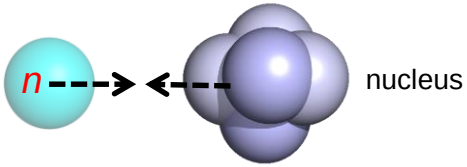
$Q = q = h = \mu = k = s$

Sometimes; $S = 2\sin\theta/\lambda = 2\pi s$

Repulsive energy potential



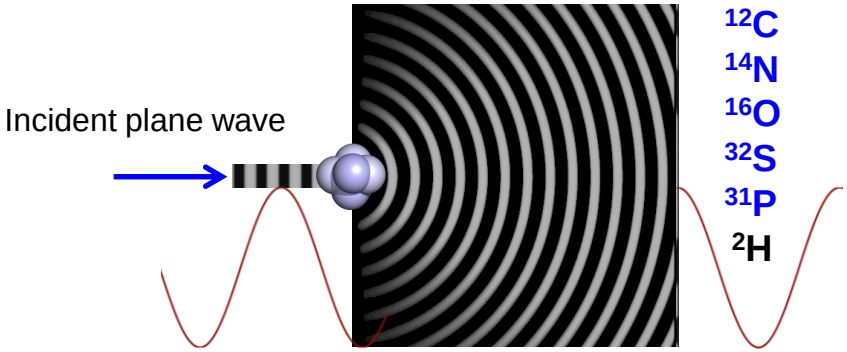
Attractive energy potential



In 2-dimensions: elastic scattering

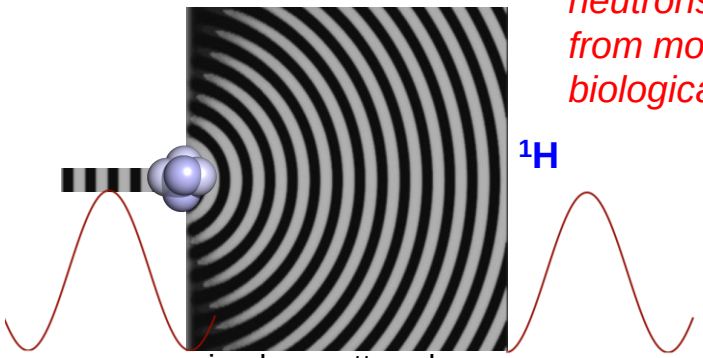
Neutrons scattered from ^{1}H are 180° out of phase with neutrons scattered from most other biological isotopes!

Repulsive energy potential



circular scattered wave
180° phase shift

Attractive energy potential



circular scattered wave
0° phase shift

However, a neutron scattering length consists of two terms:

$$b_n = b_c + b_i$$

Where the b_c represents the coherent neutron scattering length, and b_i the represents the incoherent scatteing length.

For SANS it is the ELASTIC COHERENT scattering that can be used to determine the atom-pair separation within a macromolecule.

The incoherent term can be expressed as:

$$b_i = 2B(\boldsymbol{IS})$$

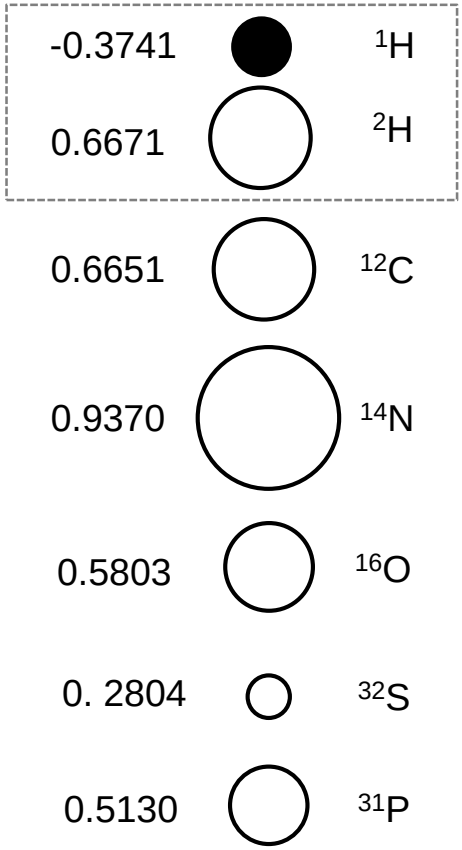
Where B is the spin scattering length of an isotope (*I* and *S* are spin operators of the neutron and nucleus).

Incoherent spin-interactions can cause a problem for SANS experiments. If the spins of the atoms comprising the sample and the incoming neutrons are not ordered, neutrons will scatter incoherently. In other words, scattering-pair distance correlations encoded within the scattered wave amplitudes no longer exist.

Incoherent scattering – that in basic terms ‘ripples across’ the entire coherent wave front – produces significant background noise in a SANS experiment.

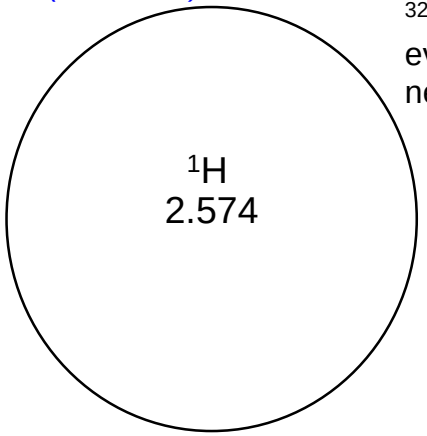
Neutron scattering lengths, biological elements.

$b_{(coherent)}$ values (10^{-12} cm)

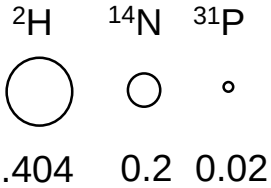


The *b* for ^{1}H is negative: Attractive interaction potential.

$B_{(incoherent)}$ (10^{-12} cm)

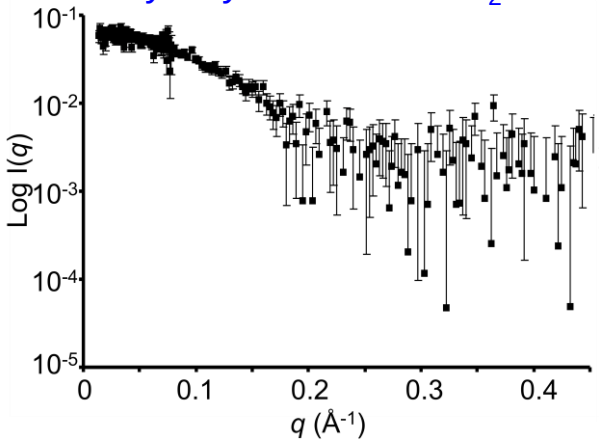


The spin state of ^{12}C, ^{16}O and ^{32}S disallow incoherent scattering events from the ½ spin state of neutrons.

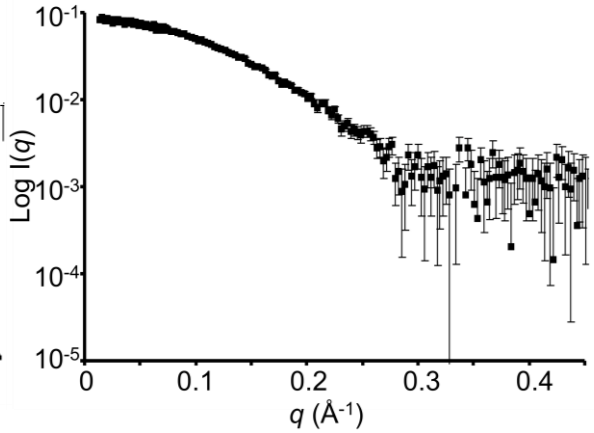


As it happens, the incoherent scattering length of ^{1}H is enormous – one of the longest incoherent scattering lengths!

Lysozyme in 100% $^{1}H_2O$



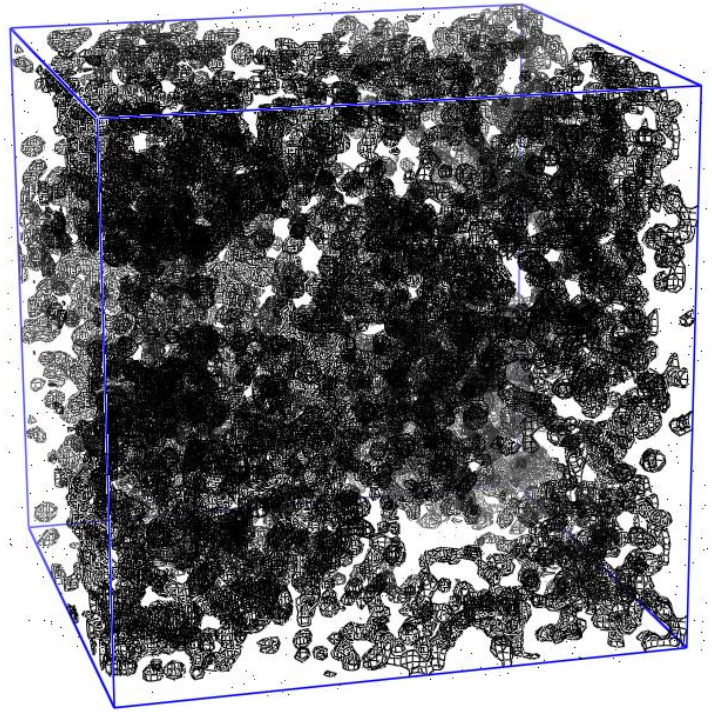
Lysozyme in 100% $^{2}H_2O$



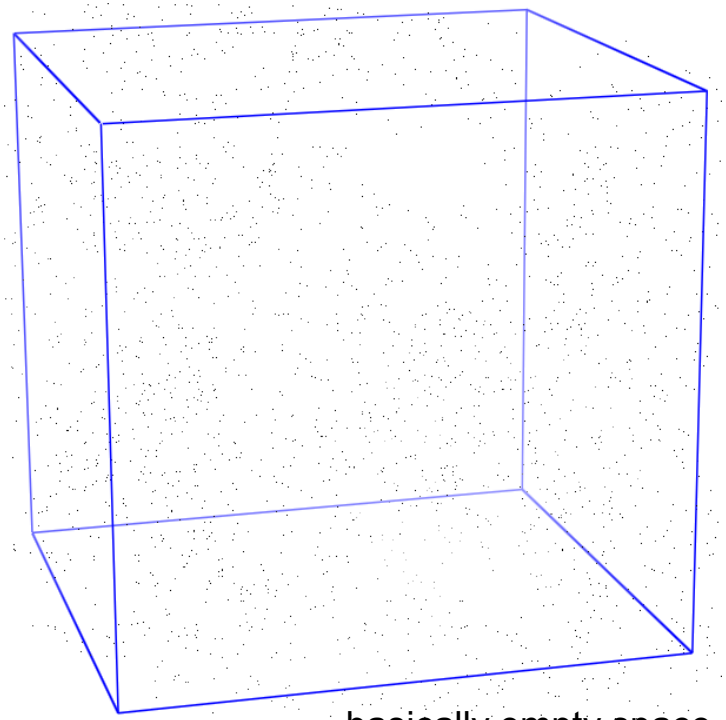
Incoherent neutron scattering in $^{1}H_2O$ is quite evident!

What X-rays and neutrons ‘feel’

X-rays are scattered by electrons.

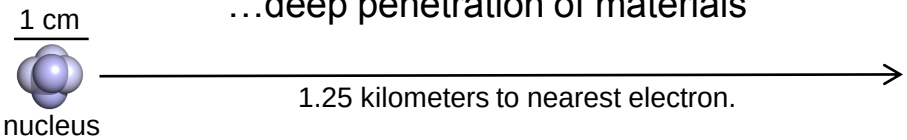


Neutrons are primarily scattered by atomic nuclei.



…basically empty space

…deep penetration of materials



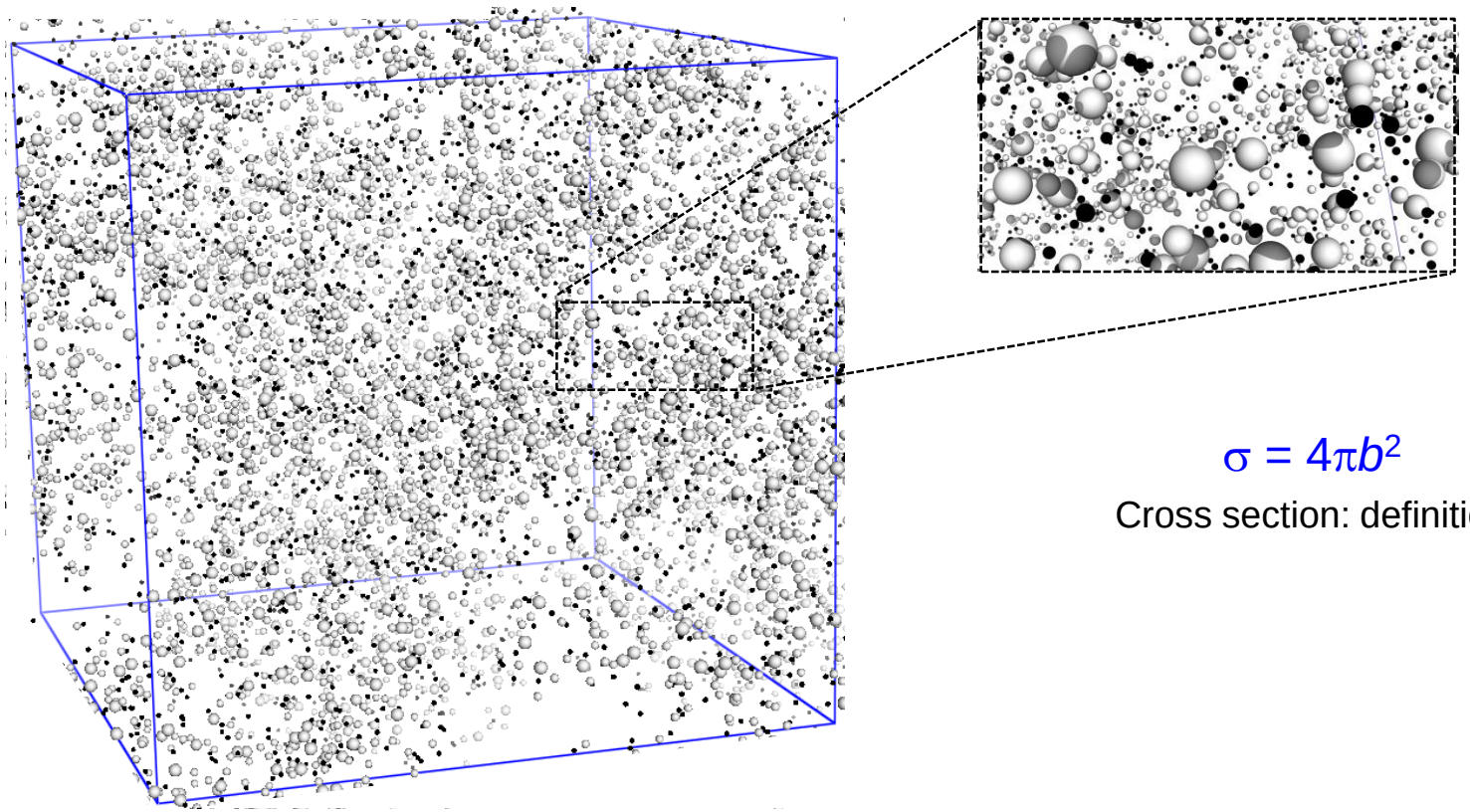
…unlike X-rays, low-to-no radiation damage!

Remember, neutrons have a magnetic moment. Neutrons undergo magnetic scattering from unpaired electrons in ordered magnetic lattices. Very useful in material science!

…but in terms of probability, for both SAXS and SANS

The probability of interaction, for SAXS or SANS, is represented as a conceptual ‘circle’ – or cross section.

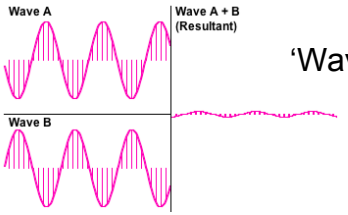
The radius of the circle has a length and relates to what is termed the scattering length, *b*.



$$\sigma = 4\pi b^2$$

Cross section: definition

Elastic coherent small-angle scattering =
'Sum the waves game.'

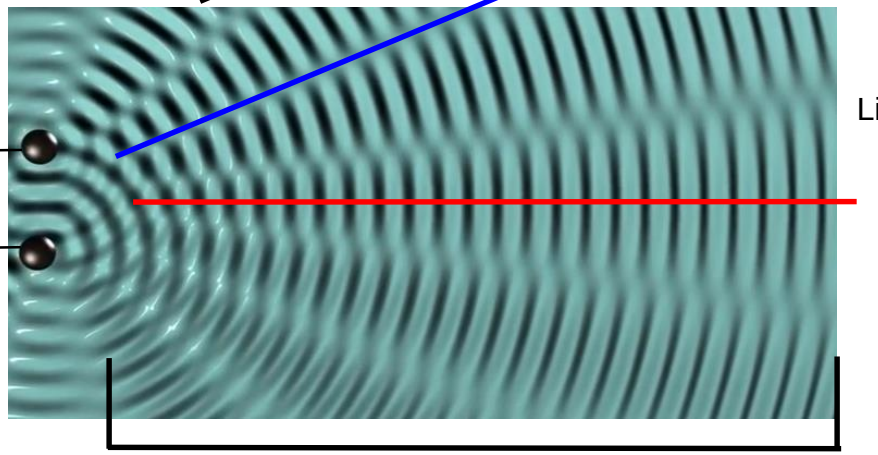


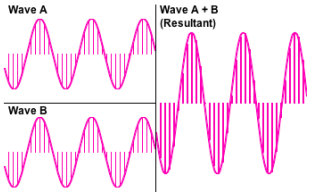
'Waves cancel'

Line of destructive interference

Line of constructive interference

Preserved distance correlation between two scattering centres.





'Waves add'

Coherent wave front

elastic scattering – no wavelength change

Of course, macromolecules have many, many atom pair distance correlations within extent of their volume boundary.

The coherent wave front is derived from the sum of the scattered waves from all of these correlations – time and rotationally averaged – as a function of angle.

For macromolecules in solution…

If the distances, *r*, between the atoms of a marcomolecule are preserved then the amplitudes of the *coherent* wave front through *s* are proportionate to the *sum* of the atomic scattering factors (i.e., probability to scatter) weighted by the distribution of the distances between scattering pairs.

The amplitudes

$$A(s) = \sum_{i=1}^{N} b_i e^{i\vec{s}\bullet\vec{r}}$$

Spherical wave bit

'Scattering factor': relates to the atomic cross section, i.e., scattering length, or probability of an atom to scatter for every atom in the sample.

The issue? We cannot access the amplitudes experimentally. We measure the *intensity* of the scattered radiation.

$$I(s) = A(s)A(s)^*$$

Amplitudes squared (actually the amplitudes multiplied by the complex conjugate of the amplitudes).

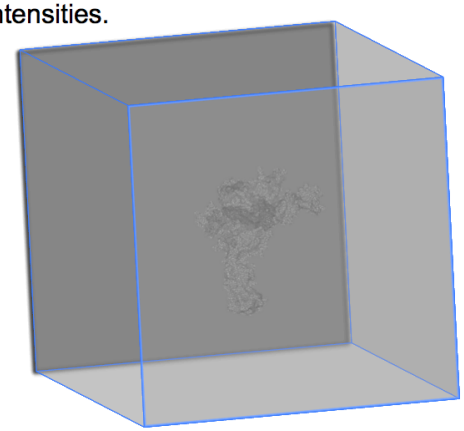
Contrast

I(*s*) in the small-angle region depends, and indeed only arises, if there is a difference between the average scattering length density of the solvent and the average scattering length density of the particles of interest. This difference is known as ***contrast*** and is represented as

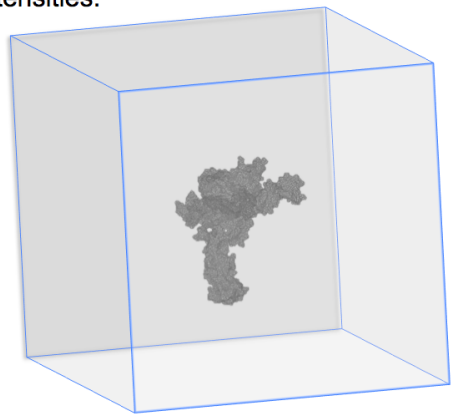
$$\Delta\rho = \bar{\rho} - \bar{\rho}_{\mathrm{S}},$$

where $\bar{\rho}$ and $\bar{\rho}_{\mathrm{S}}$ are the mean scattering length densities of the particle and the solvent, respectively.

Low contrast = weaker coherent scattering intensities.

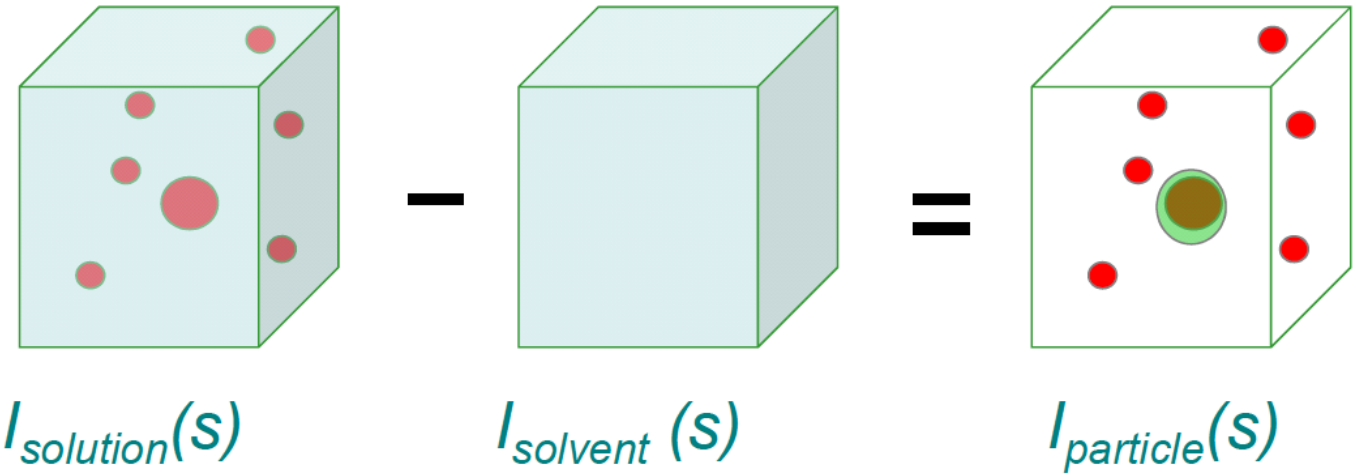


High contrast = stronger coherent scattering intensities.



$$I(s) \propto \Delta\rho^2$$

Solution bioSAS is a subtractive technique



To obtain scattering from the particles, solvent scattering must be subtracted to yield the effective *EXCESS* (time-preserved) scattering length density distribution $\Delta\rho = \langle\rho(r) - \rho_s\rangle$, where ρ_s is the average scattering density of the solvent.

How do I calculate the contrast?

http://smb-research.smb.usyd.edu.au/NCVWeb/

MULCh: Modules for the analysis of small-angle neutron contrast variation data from bio-molecular assemblies.

Whitten, A. E., S. Cai, and J. Trewhella (2008) MULCh: ModULes for the analysis of small-angle neutron contrast variation from biomolecular assemblies. *J. Appl. Crystallogr*. 41:222–226.

Jeffries et al., (2016) Nature Protocols 11:2122-2153

A: Define the solvent

ModULes For The Analysis Of Small-angle Neutron Contrast Variation Data From Bio-molecular Assemblies

Contrast: Module For Estimating The Contrast Of Bio-molecular Assemblies

Upload an existing input file: Upload Contrast File **(Upload txt input if avaialble)**

Project Title: VH Ab lysozyme **i) Input title of project**

Number of contrast points: 0

f_{D_2O}(0 - 1)	I(0)	σ(I(0))	Protein conc.

Number dissolved species in the solvent: 3 **ii) # molecules in solvent = 3**

Substance Type	Formula	Conc. (mol/L)	Volume (Å³)
P D R M	NaCl	0.1	0.0
P D R M	C4H11NO3	0.05	0.0
P D R M	C6H7NaO6	0.001	0.0

iii) For small molecules: input atomic formula and concentrations

P = protein; D = DNA; R = RNA; M = molecule

B: Define macromolecules

Number of components in subunit 1: 1 **i) # components in subunit 1 = 1**

Deuteration level(0 - 1): 0.0 Fraction of acidic protons accessible by the solvent: 0.95 **ii) Choose level of deuteration.**

Substance Type	Formula	$N_{molecules}$	Volume (Å³)
P D R M	KVFGRCELAAAMKRHGLDNYRGYSLGNWVCAAKFESNFNTQATNRNTDGSTDYGILQINSRWWCNDGRTPGSRNLCNIPCSALLSSDITASVNCAKKIVSDGNGMNAWVAWRNRCKGTDVQAWIRGCRL	1	0.0

iii) Amino acid sequence

Number of components in subunit 2: 2 **iv) # components in subunit 2 = 2**

Deuteration level (0 - 1): 0.6 Fraction of acidic protons accessible by the solvent: 0.95 **v) Choose level of deuteration**

Substance Type;	Formula	$N_{molecules}$	Volume (Å³)
P D R M	DVQLQASGGGSVQAGGSLRLSCAASGYTIGPYCMGWFRQAPGKEREGVAINMGGGITYYADSVKGRFTISQDNAKNTVYLLMNSLEPEDTAIYYCAADSTIYASYYECGHGLSTGGYGYDSWGQGTOVT	1	0.0
P D R M	Ca	2	0.0

vi) Amino acid sequence

vii) Bound calcium; 2 per subunit.

Submit Cancel

C: ρ and $\Delta\rho$ output

Individual component and solvent scattering length density — Individual component and whole complex contrasts

Tabulated scattering length densities and contrasts

	ρ (10^{10}cm^{-2})			$\Delta\rho$ (10^{10}cm^{-2})		
	1	2	Solvent	1	2	Total
X-RAY	12.515	12.580	9.454	3.061	3.127	3.095
NEUTRON						
0.0	1.957	4.579	-0.545	2.502	5.123	3.869
0.1	2.112	4.712	0.146	1.967	4.566	3.322
0.2	2.268	4.844	0.836	1.432	4.008	2.775
0.3	2.423	4.977	1.526	0.897	3.451	2.228
0.4	2.579	5.110	2.217	0.362	2.893	1.682
0.5	2.734	5.242	2.907	-0.173	2.335	1.135
0.6	2.890	5.375	3.597	-0.707	1.778	0.588
0.7	3.045	5.508	4.288	-1.242	1.220	0.042
0.8	3.201	5.641	4.978	-1.777	0.663	-0.505
0.9	3.356	5.773	5.668	-2.312	0.105	-1.052
1.0	3.512	5.906	6.359	-2.847	-0.453	-1.598
Calculated match-point (f_{D_2O})				0.468	0.919	0.708

i) X-ray contrast for SAXS

Fraction ^2H_2O in solvent, (SANS).

ii) Total neutron contrast for SANS

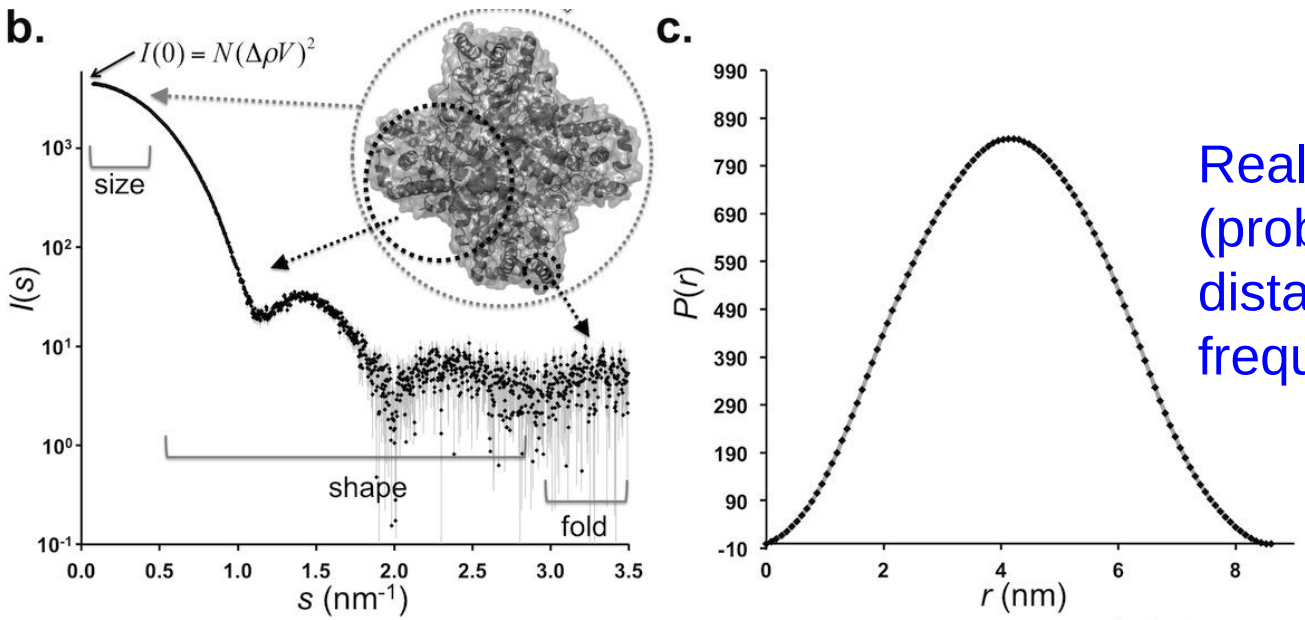
Individual component SANS matchpoints: v/v ^2H_2O.

Whole complex SANS matchpoint: v/v ^2H_2O.

The scattering intensity $I(s)$ – and thus the associated form factor in reciprocal space – relates to an atom-pair distance distribution function of the particle $p(r)$ in real space by a Fourier transform:

$$I(s) = 4\pi \int_0^{D_{max}} P(r) \frac{\sin(sr)}{sr} dr \qquad p(r) = \frac{r^2}{2\pi} \int_0^{\infty} s^2 I(s) \frac{\sin(sr)}{sr} ds$$

Reciprocal space intensity



Real space (probable) distance frequencies

Combine the:

1) Atomic scattering factors with;

2) The excess scattering length density distribution within the ***volume*** of a macromolecule and…

you obtain a very simple set of relationships for the measured intensities (*reciprocal space* scattering):

$$I(s) \propto P(s)$$

Form factor of the entire macromolecule – relates to real space distance distribution between all scattering pairs inside the volume boundary.

$$I(s) \propto \Delta\rho^2$$

The difference in scattering length density SQUARED

$$I(s) \propto V^2$$

The particle volume SQUARED

The scattering intensity

Is the SUM of all macromolecules averaged over all orientations.

The structure factor or 'between particle' contributions

$$I(s) = \sum_{i}^{n} [(\Delta\rho_i V_i)^2 P_i(s)] S(s)$$

Weighted by the contrast and volume SQUARED of all macromolecules

The form factor of all macromolecules within the sample

For a PURE, MONODISPERSE and IDEAL sample.

The ***concentration***.

$$I(s) = N(\Delta\rho V)^2 P(s)$$

If all particles are identical, and do not interact, the *I*(*s*) profile (after background solvent scattering has been subtracted) will represent the time and rotationally averaged scattering from a ***SINGLE PARTICLE***.

How do I know I have an ideal system?

The stability of molecular mass, MM, and volume (V) estimates through a concentration series.

The MM, the MM, the MM, the MM, the MM, the MM.

(+/-10 %)

Think about this – there is no point generating a single model to describe a 100 kDa protein if the experimental MW of the protein from SAS is 125 kDa (probably a mixture).

$I(0)$

At zero angle (s = 0) the magnitude of $I(s)$ will primarily depend on the number of scattering centres within the bound squared-volume of a macromolecule – independent of the shape – weighted by the concentration and contrast squared:

$$I(0) \approx N(\Delta\rho V)^2$$

From this parameter, it is possible to obtain the ***molecular weight*** of, for example, a protein.

$$MW = \frac{I(0)N_A}{c(\Delta\rho\upsilon)^2}$$

Absolute scaling – requires Avigadros number, N_A (and partial specific volume υ).

Data scaled to a standard protein with a KNOWN concentration and molecular weight

$$MW_{protein} = \frac{I(0)_{protein} \bullet c_{standard} \bullet MW_{standard}}{c_{protein} \bullet I(0)_{standard}}$$

An assumption that a target has a similar scattering length density and partial specific volume as the secondary standard! If NOT you have to correct the above relationship!

For SANS with contrast variation

The forward scattering intensity at zero angle, *I*(0), is basically the total scattering derived from all distance correlations within the volume of the particle (assuming no interparticle interactions weighted by the contrast)

For a single component
(One scattering length density)

$$I(q_{\text{total}}) \propto \Delta\rho^2$$

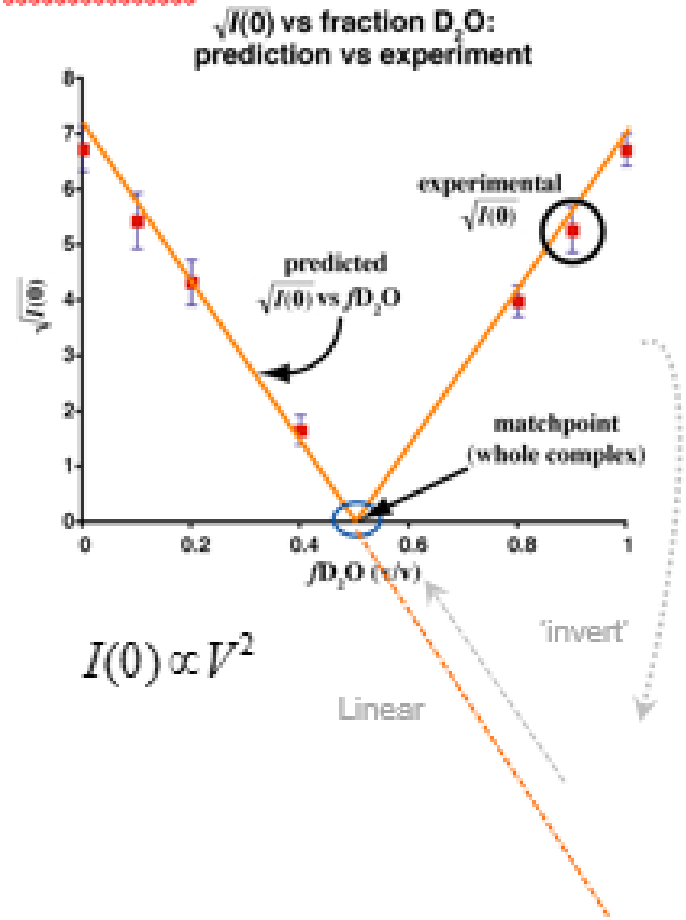
For a complex
(Two scattering length densities)

$$I(q_{\text{total}}) \propto \Delta\rho_1^2 I(q)_1 + \Delta\rho_1\Delta\rho_2 I(q)_{12} + \Delta\rho_2^2 I(q)_2$$

Component 1 contribution

'between' component contributions (cross-term)

Component 2 contribution



$$I(0) \propto V^2$$

A deviation from linearity for a two-component system through the contrast series is an indication that there is something wrong with the sample:

Incorrect % v/v ^2H_2O.
Aggregation.
The complex falls apart in ^2H_2O.

ATSAS tools for volume and volume-based MM estimation for SAXS.

ATSAS tool: *datporod*

ATSAS tool: *datmow*

ATSAS tool: *datvc*

At the command prompt (.cmd, terminal, etc) type:

datporod filename.out

↓

Porod volume estimate.

For proteins, convert to MM by dividing by 1.6

datmow filename.out

↓

MM estimate of proteins using the method of Fischer et al. *SAXMOW*

datvc filename.out

↓

MM estimate of proteins using the method of Rambo and Tainer.

Bayesian Inference

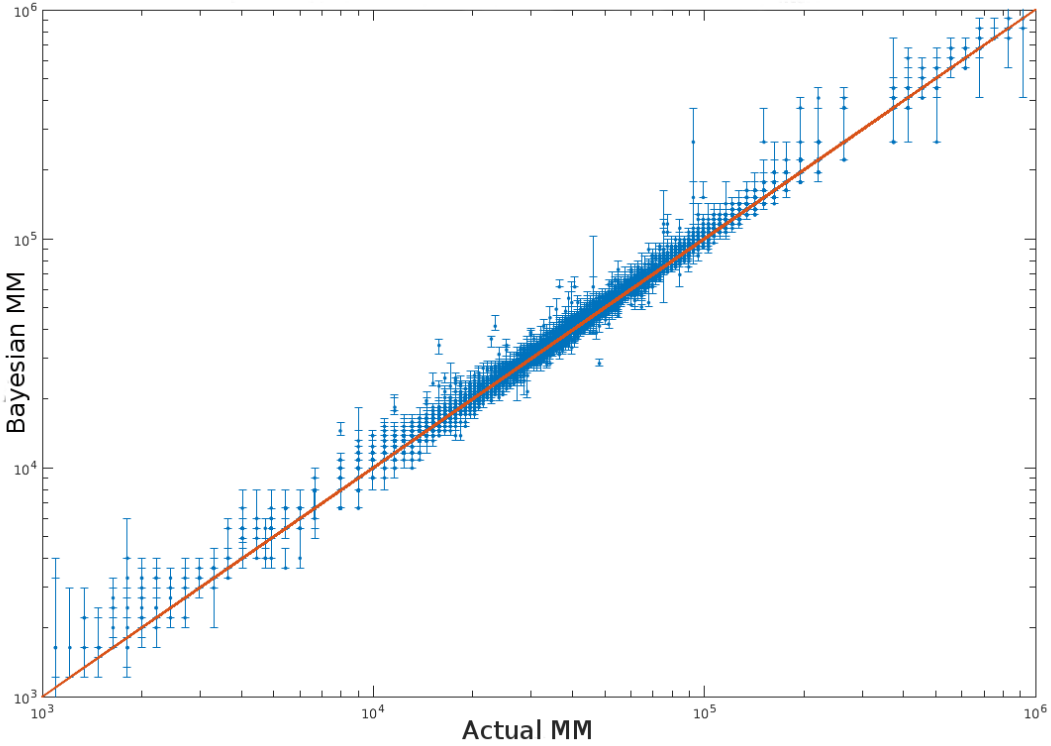
- Combines the different integral based approaches for concentration independent MM estimation.
- Provides the most likely MM within a confidence interval.

ATSAS tool: *datmw*

At the command prompt (.cmd, terminal, etc) type:

```
>datmw  filename.out
```

- Assess BOTH concentration dependent and independent methods of MM determination!



Model the data: Setup.

Modelling SAS data – before you leap into danger

- Understand the data – get the unit right, nm or Å, etc.
- Extract structural parameters and additional information *BEFORE* you begin modelling: if there is one thing you can trust it is the structural parameters from SAS data!

Part 1 of your validation toolbox →

- Radius of gyration (R_g) maximum particle dimension (D_{max}), volume (V).
- Molecular mass estimates (*MM*).
- Probable frequency of distances (r) within single particles ($p(r)$ vs r), i.e., *global* shape and structural information.
- *Scaling parameters* – compact, flexible, flat, rod, hollow.
- *Useful data range!*
- *The AMBIGIUTY of the data!*
- Size distributions and volume fractions.

Modelling SAS data – before you leap into danger

- Obtain as much information as possible about your system.

Part 2 of your validation toolbox →

- For example, obtain the ***EXACT*** amino acid sequence of the protein actually used for the SAS experiment. ***ALL atoms scatter***, so you have to take into account ***ALL of the mass*** in your modelling!
- Obtain the ***CORRECT PDB*** files (atomic coordinate files). ***as ALL atoms scatter***, so you have to take into account ***ALL of the mass*** in your modelling!
- If required, calculate the ***CONTRAST*** of your system; (on occasion, for SAXS, convert to electron density difference.)
- Obtain restraints derived from complementary methods – in particular ***CONTACT*** information (e.g., from NMR, cross-linking mass-spectrometry, FRET.)
- Know the ***STOICHIOMETRY*** and from this, the estimated ***SYMMETRY***. Obtain the MM estimate from SAS or other methods, e.g., MALLS.

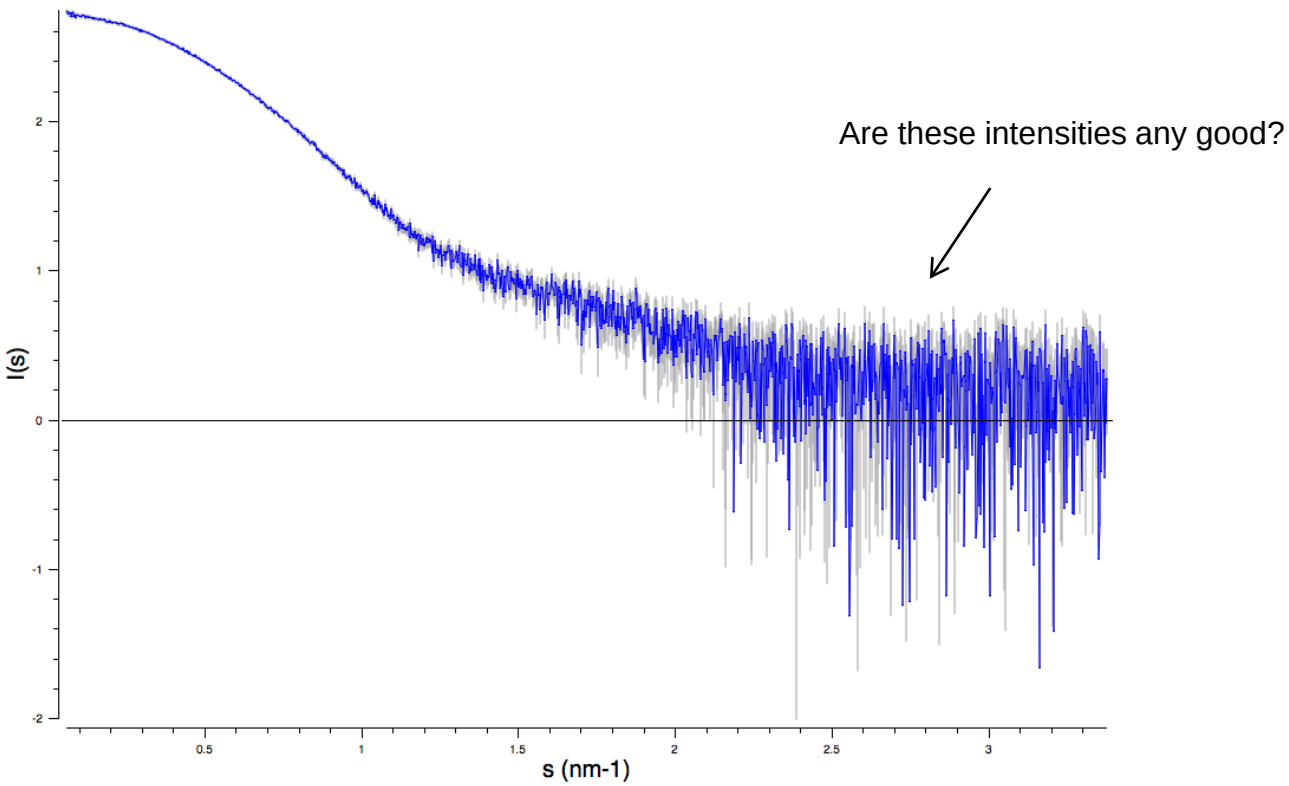
Model the data *without* modelling the data!

What the?

Four programs:

- SHANUM – define the useful data range.
- DATCLASS – machine-learning methods for the rapid geometric classification of SAXS data (from proteins).
- DARA – kd-tree searching of the PDB for similar scattering profiles.
- AMBIMETER – assess the ambiguity of the scattering data.

What is signal and what is noise? First assess the information content.



Going back to perfection:

If we can calculate the ***EXACT*** scattering amplitudes

Therfore...

We can calculate the ***EXACT*** scattering intensities.

$$I(s) = \left\langle \left| A(\mathbf{s}) \right|^2 \right\rangle_\Omega$$

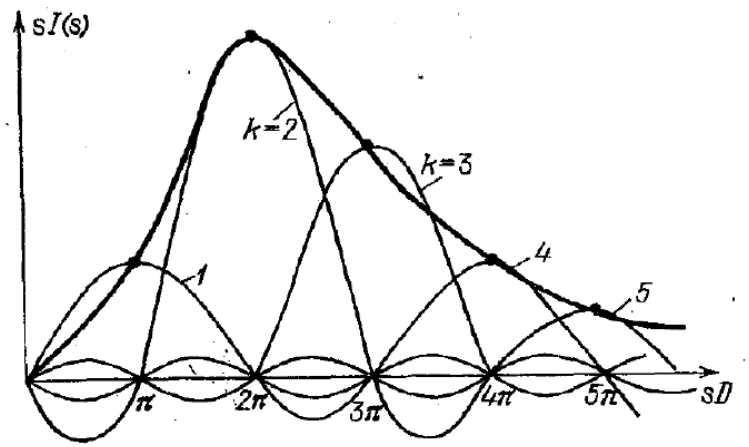
Question: So do we require all of the EXACT amplitudes across a continuous angular range or can we sample the intensities at a given, or defined interval of *s* - to describe the scattering profile?

I suppose you could say, is it necessary to have every single continuous point across all of *x* to reconstruct the function $y^3 = x$?

Shannon channels, information content and some wizardry (that only about three people have ever understood…Shannon and Moore being two of them.)

Shannon sampling theorem: the scattering intensity from a particle with the maximum size *D* is defined by its values on a grid $s_k = k\pi/D$ (Shannon channels):

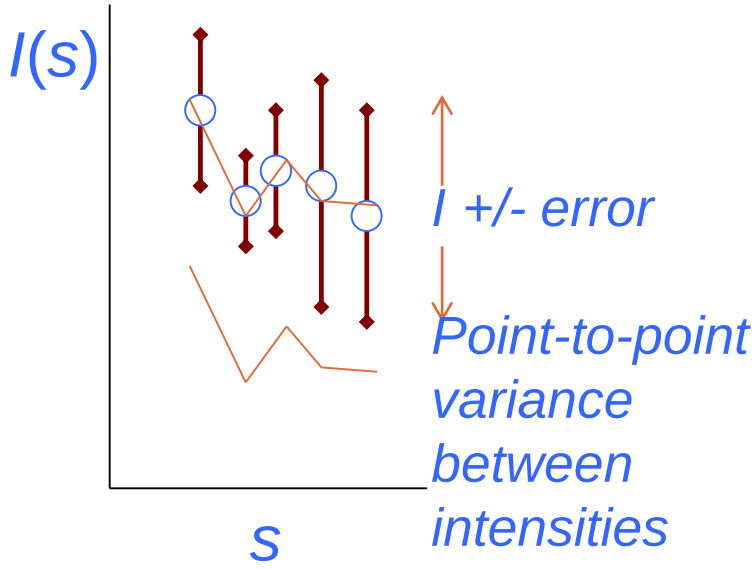
$$sI(s) = \sum_{k=1}^{\infty} s_k a_k \left[\frac{\sin D(s - s_k)}{D(s - s_k)} - \frac{\sin D(s + s_k)}{D(s + s_k)} \right]$$

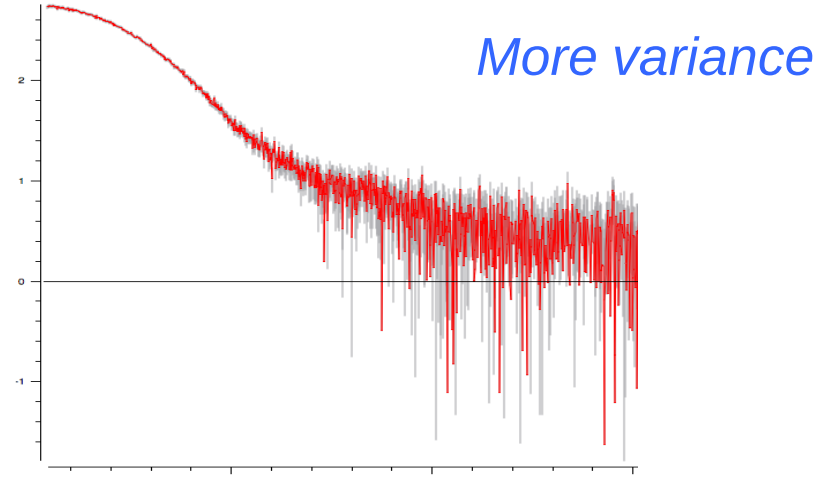


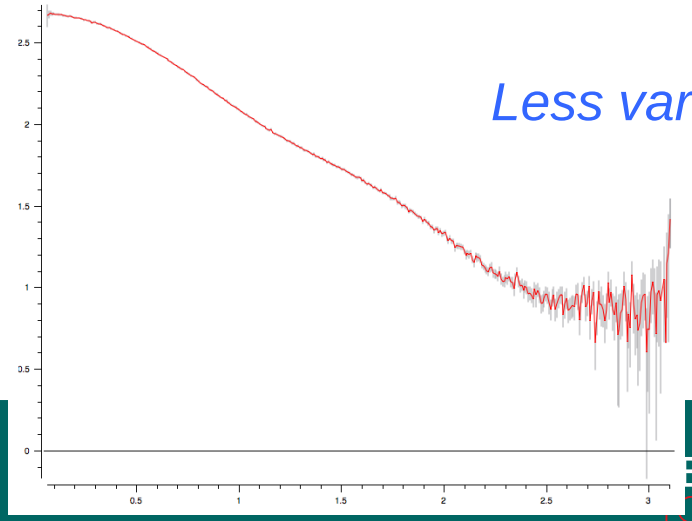
Shannon sampling was utilized by many authors (e.g. Moore, 1980). An estimate of the number of channels in the experimental data range ($N_s = s_{max}D/\pi$) is often used to assess the information content in the measured data taking into account the over-sampling and variance.

The issue is...

Although we can calculate the exact intensities and calculate exact D_{max} we can never ***MEASURE*** the exact intensities or the exact D_{max}.
So in reality you CANNOT describe a scattering pattern using only a limited number of points on a kπ/*D* grid, even though in theory, you can…







So what is the useful data range?

The first Shannon channel is defined as:

$$s_k = k\pi/D_{max}$$

Where k = 1

So, to ‘grab’ the first channel, i.e., to encompass information regarding the *LONGEST VECTORS* you have to measure to a s_{min} of π/D.

But what about s_{max}?

You do not want to model data that is effectively ‘useless’ random noise at higher angles!

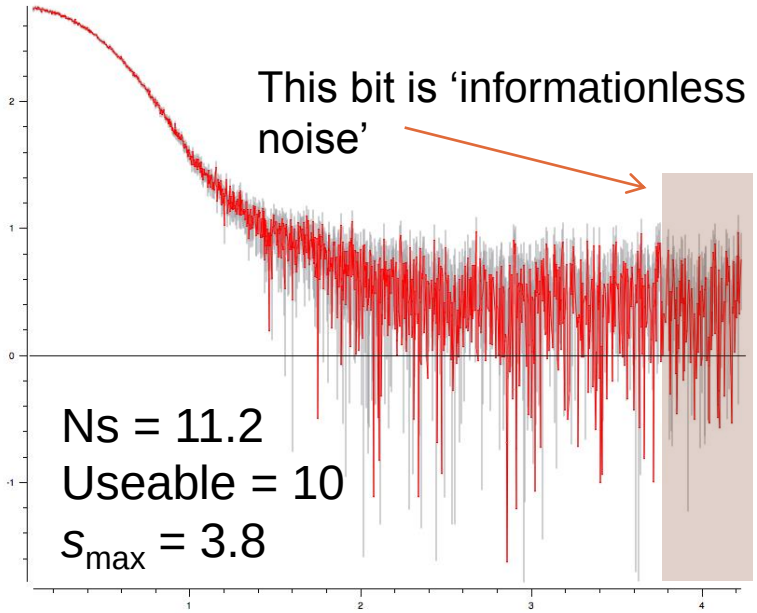
Shanum

Shanum will also estimate D_{max} (or you can enter it yourself)

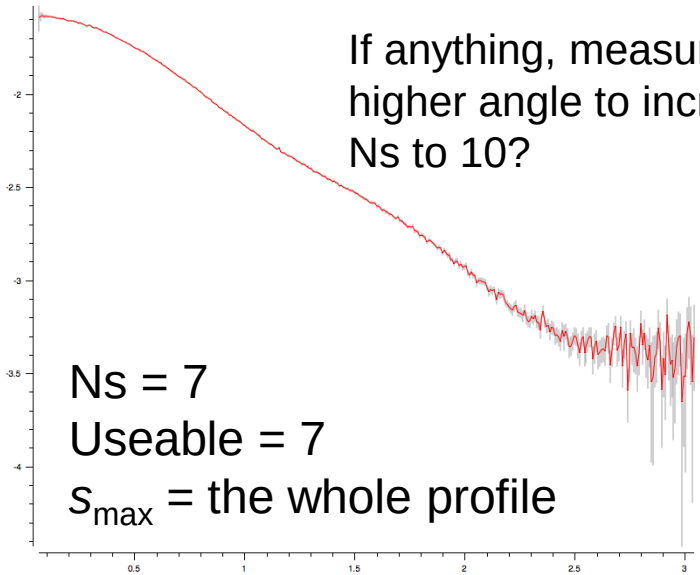
ATSAS tool: *shanum*

At the command prompt (.cmd, terminal, etc) type:

shanum filename.dat



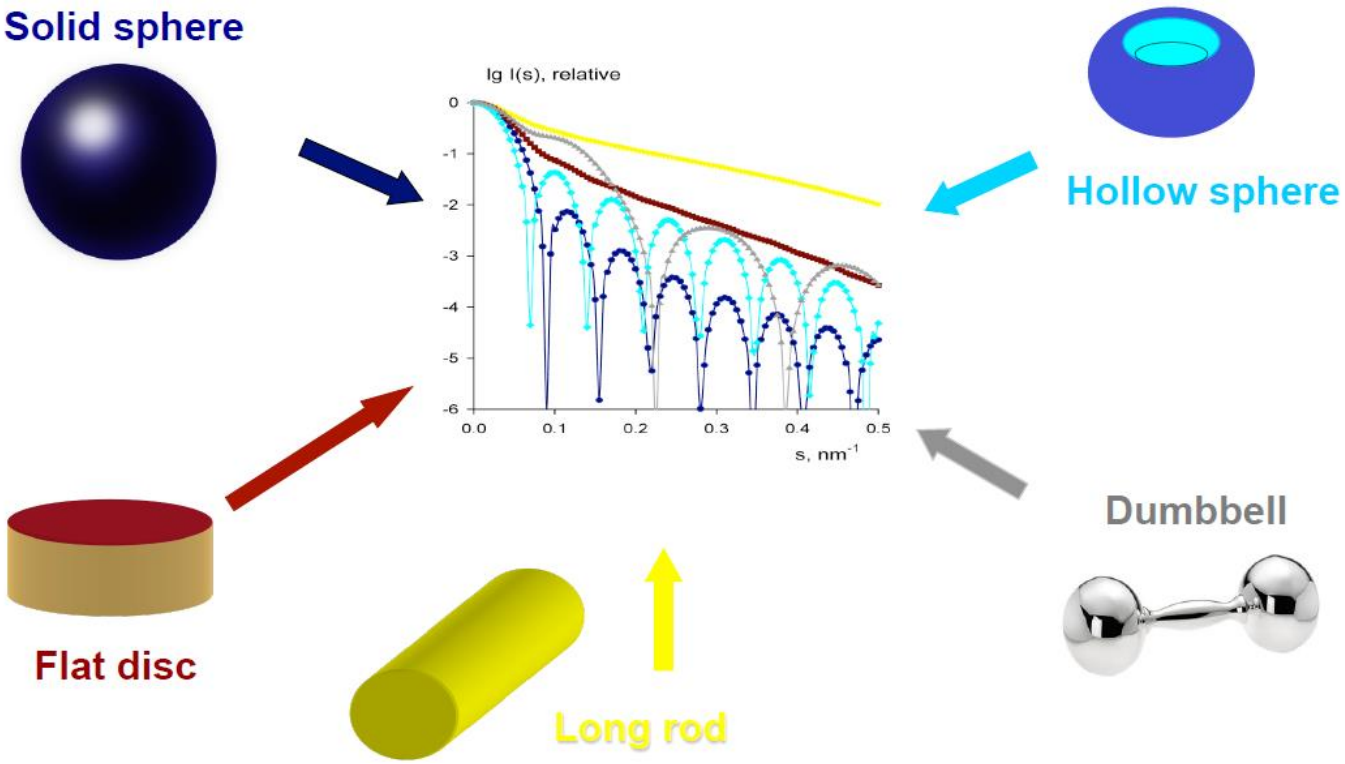
shanum BSA.dat 8.2



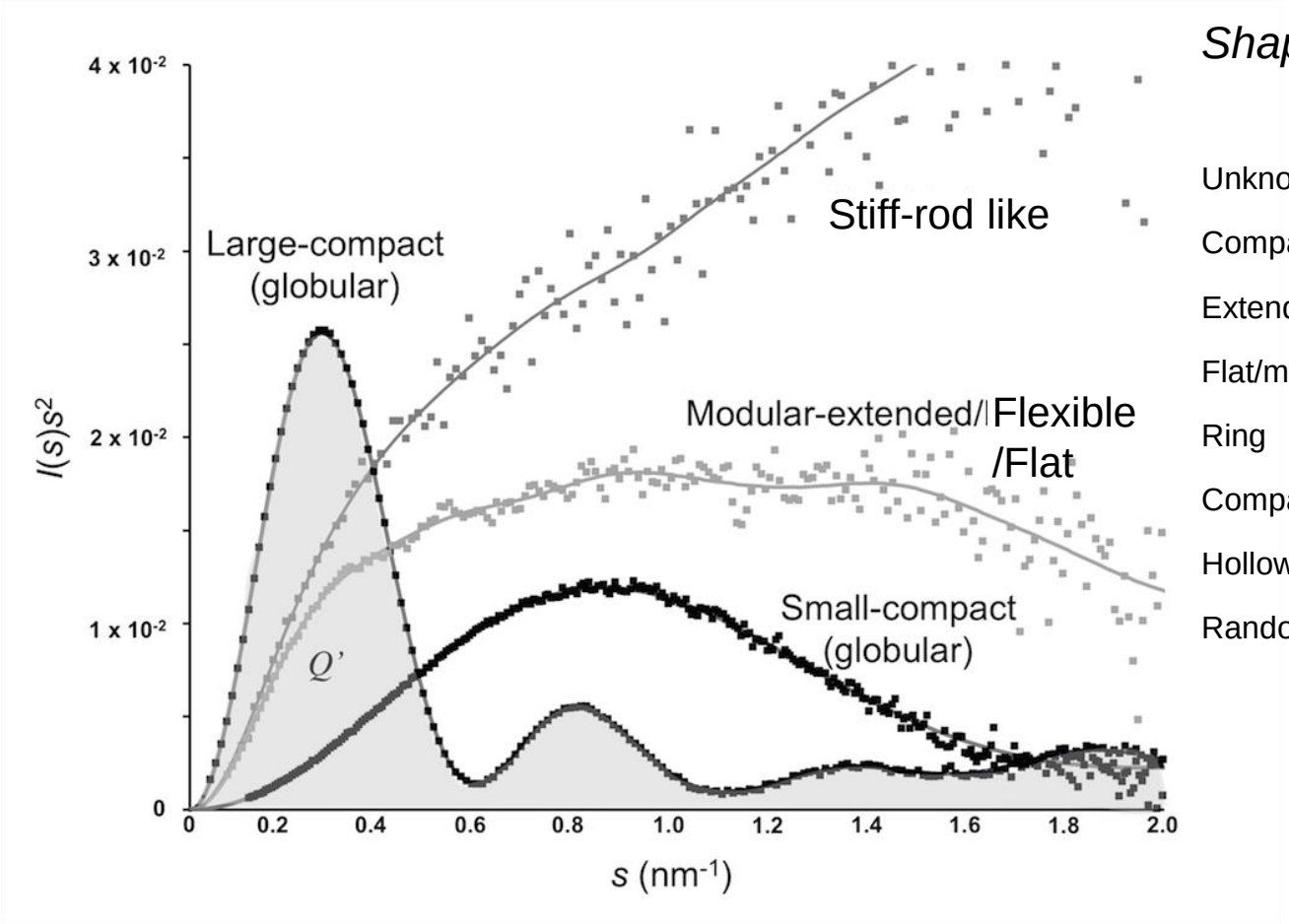
shanum CAM.dat 7.2

DATCLASS

- Classification of a protein shape using machine learning methods based on the scattering profiles calculated from a continuum of 488 000 geometric objects



Shape classification



Shape categories

Unknown

Compact

Extended

Flat/modular

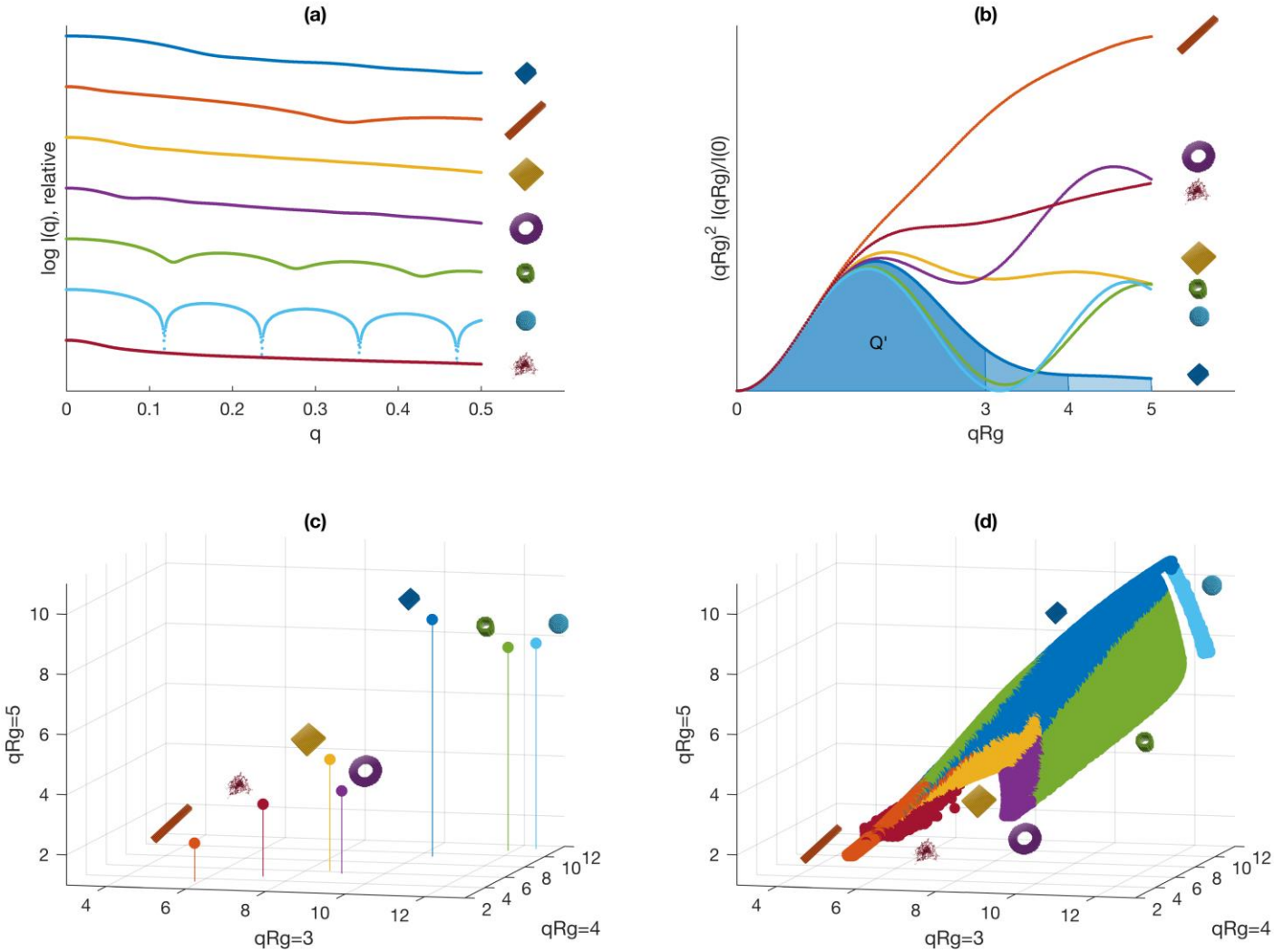
Ring

Compact-hollow

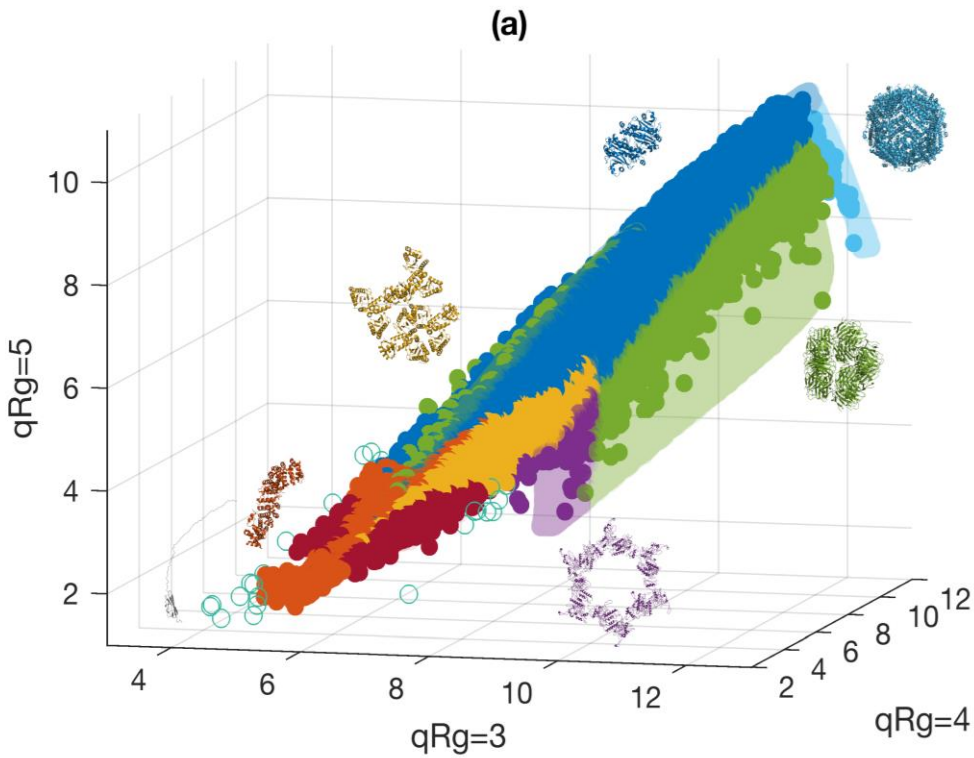
Hollow-sphere

Random-chain

Dimensionless Kratky plot at different qR_g



99.98% of the PDB maps into the classifier space.

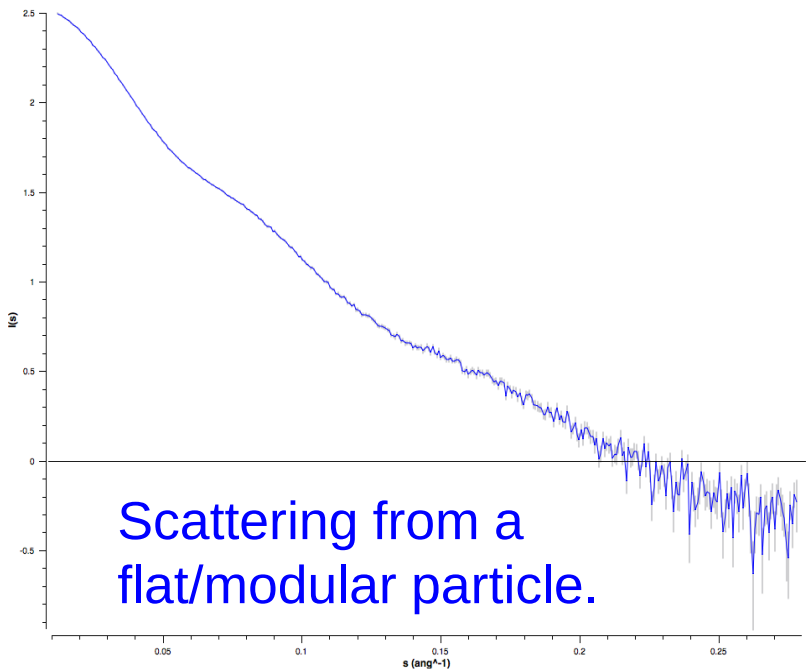


Class Label	PDB	
Unknown	25	0. 02 %
Compact	122.913	74.05 %
Extended	5.382	3.24 %
Flat	9.734	5.86 %
Ring	154	0.09 %
Compact hollow	26.909	16.21 %
Hollow sphere	125	0.08 %
Random Chain	740	0.45 %
Total	165.982	100.00 %

ATSAS tool: *datclass*

Running datclass

- .dat or GNOM.out files.



Scattering from a flat/modular particle.

At the command prompt (.cmd, terminal, etc) type:

For GNOM.out files:

>datclass filename.out

For data files:

>datclass filename.dat –rg=XXX –i0=YYY

Where the rg and i0 are calculated from Guinier.

The output from datclass will be the shape classification plus the MM estimate plus the D_{max} (in angstroms) Check this against the D_{max} from $p(r)$ vs r.

MM does NOT work for random chains!

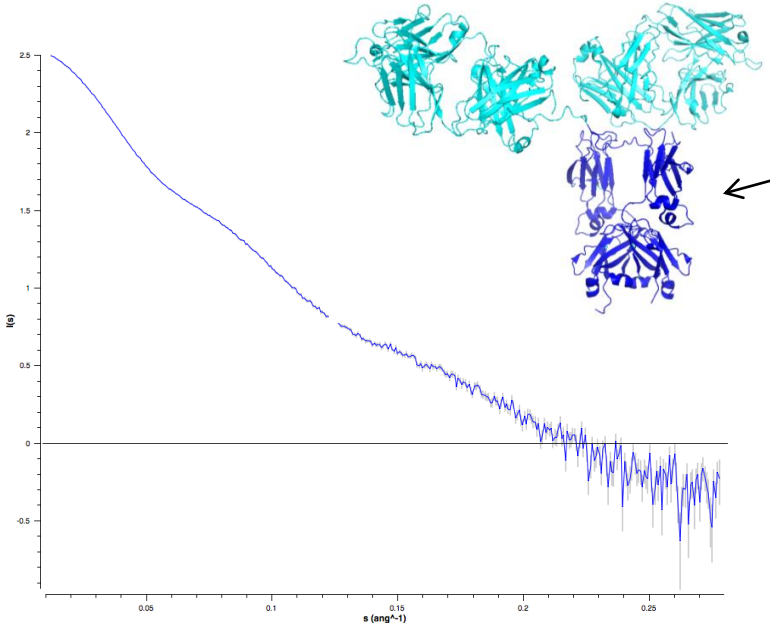
DARA.

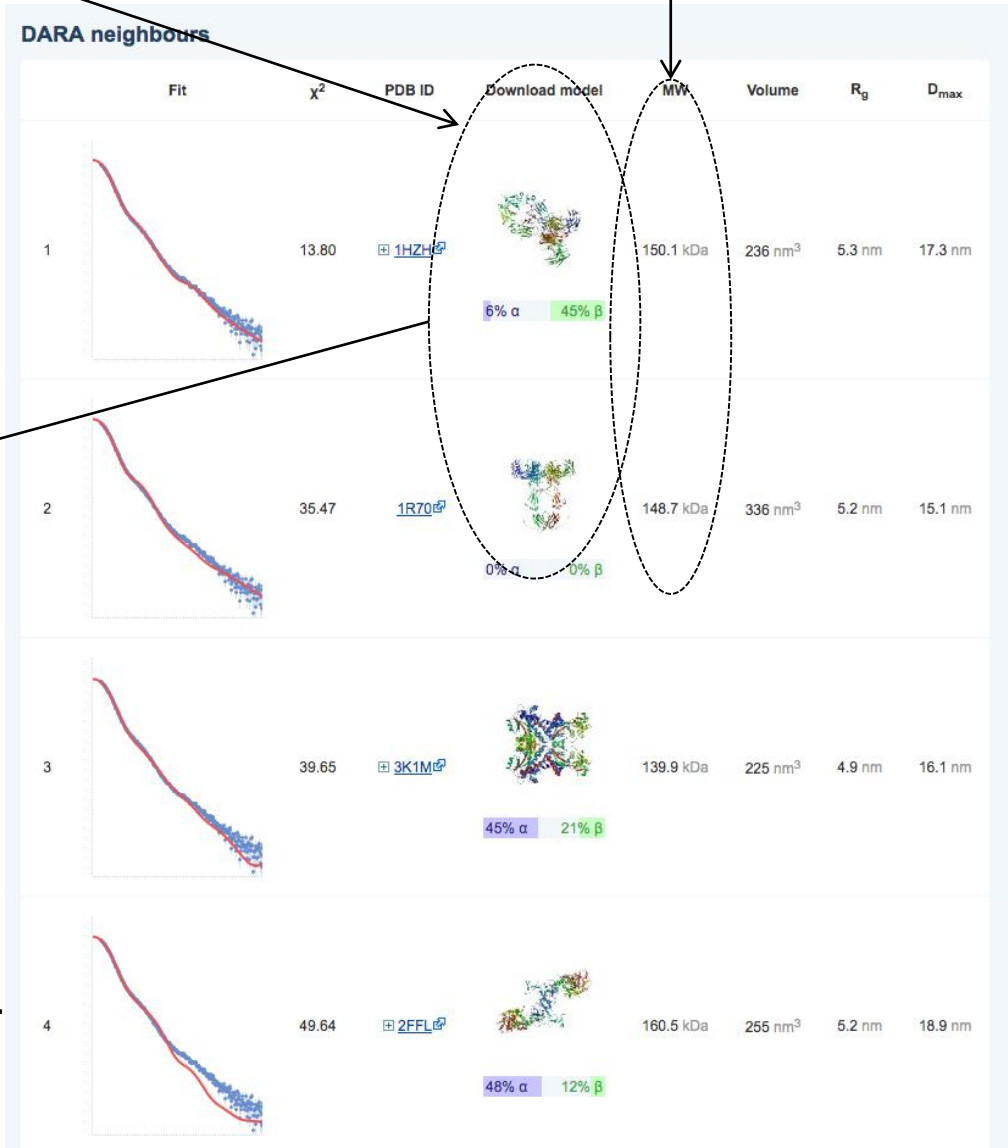
IgG or IgA like scattering

MW estimates!

SAXS data or GNOM out file

https://dara.embl-hamburg.de/





DARA neighbours

	Fit	χ^2	PDB ID	Download model	MW	Volume	R_g	D_{max}
1		13.80	1HZH	6% α 45% β	150.1 kDa	236 nm^3	5.3 nm	17.3 nm
2		35.47	1R70	0% α 0% β	148.7 kDa	336 nm^3	5.2 nm	15.1 nm
3		39.65	3K1M	45% α 21% β	139.9 kDa	225 nm^3	4.9 nm	16.1 nm
4		49.64	2FFL	48% α 12% β	160.5 kDa	255 nm^3	5.2 nm	18.9 nm

Combine DARA output with secondary structure prediction (predicted all β-strand).
E.g., YAPSIN:
http://www.ibi.vu.nl/programs/yaspinwww/
E.g., ProteinPredict

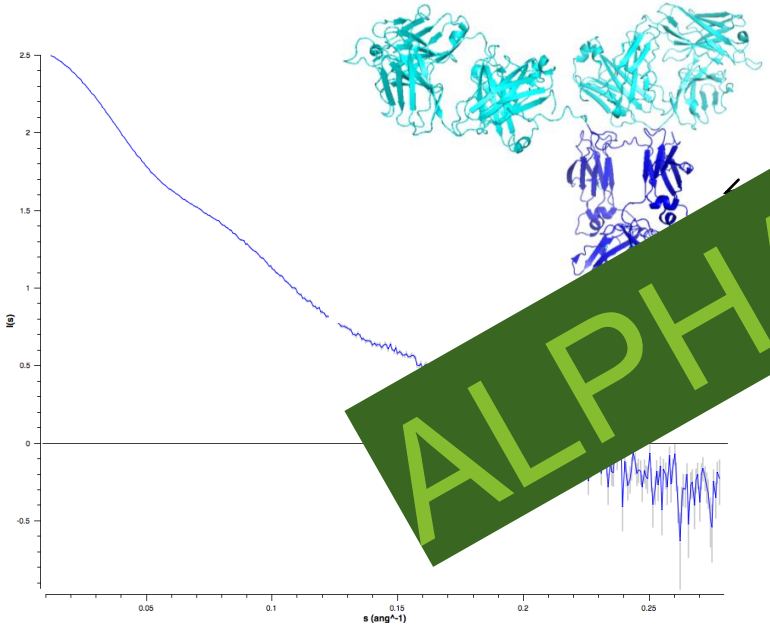
DARA.

IgG or IgA like scattering

MW estimates!

SAXS data or GNOM out file

https://dara.embl-hamburg.de/



ALPHAFOLD.

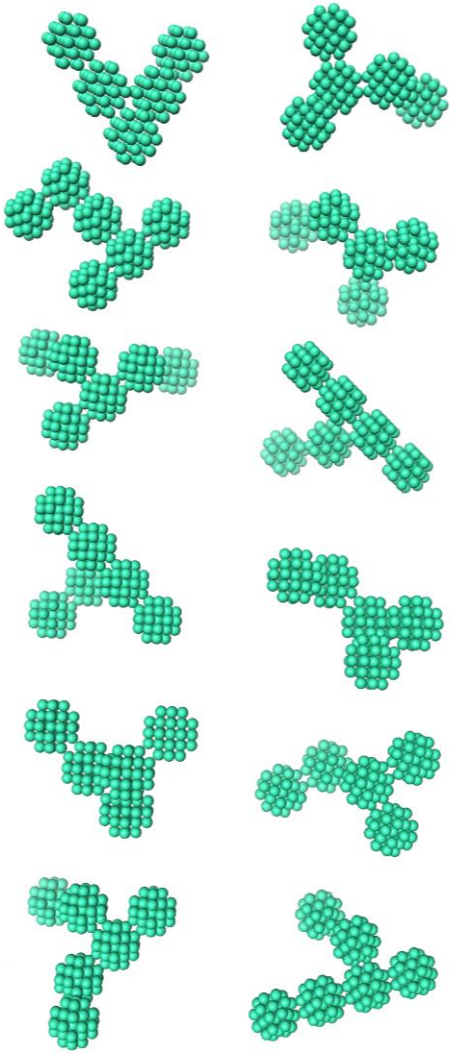
DARA neighbours

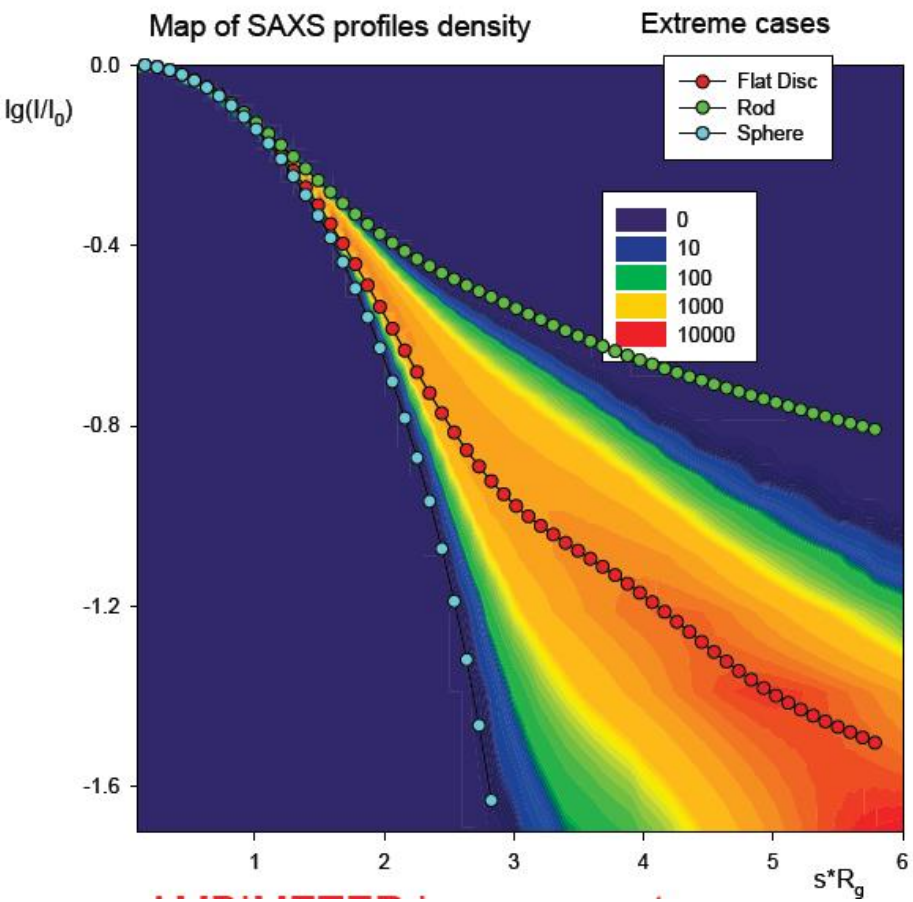
	Fit	χ^2	PDB ID	Download model	MW	Volume	R_g	D_{max}
1				45% β	150.1 kDa	236 nm³	5.3 nm	17.3 nm
		35.47	1R70	0% α 0% β	148.7 kDa	336 nm³	5.2 nm	15.1 nm
3		39.65	3K1M	45% α 21% β	139.9 kDa	225 nm³	4.9 nm	16.1 nm
4		49.64	2FFL	48% α 12% β	160.5 kDa	255 nm³	5.2 nm	18.9 nm

Combine DARA output with secondary structure prediction (predicted all β-strand).
E.g., YAPSIN:
http://www.ibi.vu.nl/programs/yaspinwww/
E.g., ProteinPredict

Ambiguity: Ambimeter

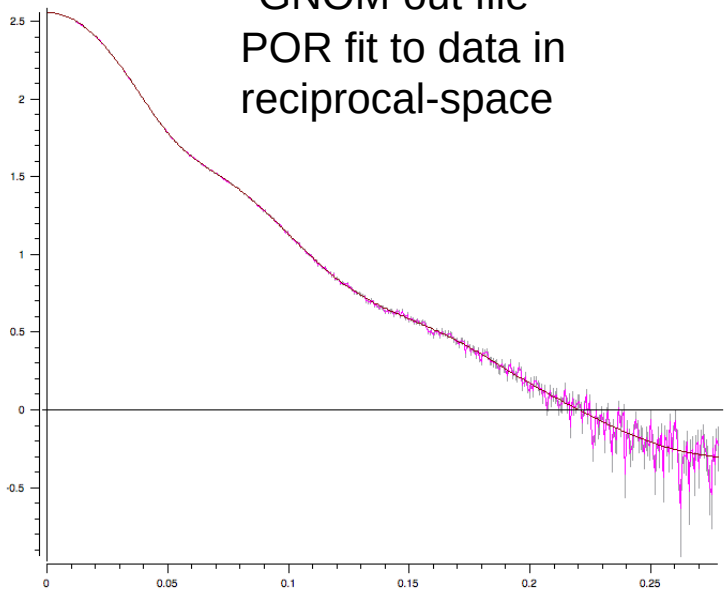
Based on a set of (several thousand) shape topologies with pre calculated scattering profiles.



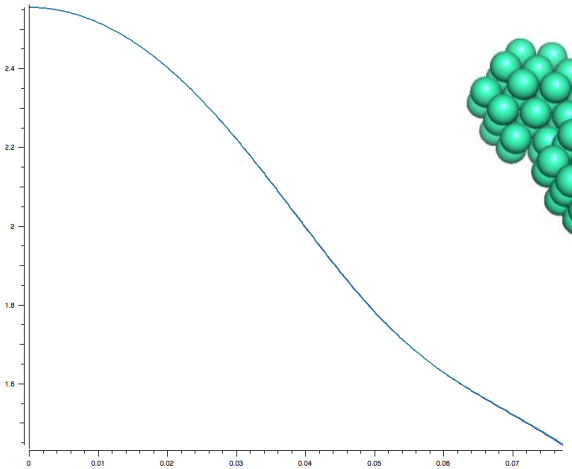


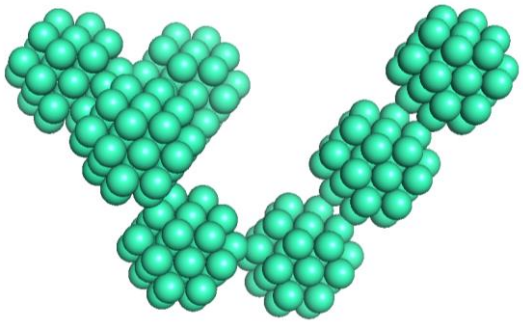
Ambimeter input.

GNOM out file
POR fit to data in
reciprocal-space



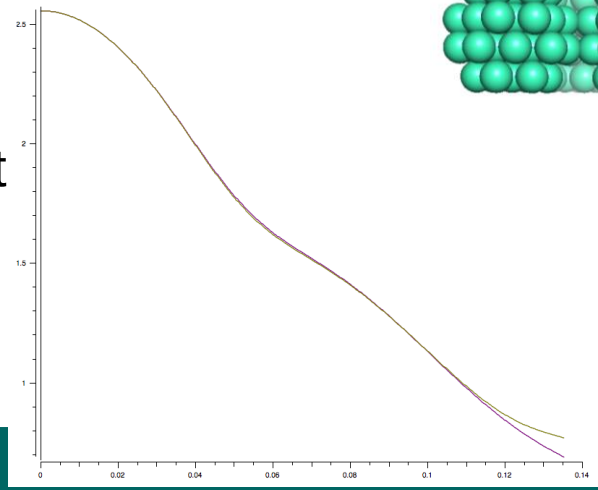
Shape-topology fit to sR_g = 4

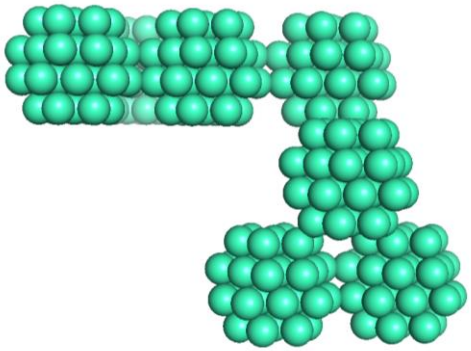




…845 shape skeletons fit the SAXS data!
...ambiguity score = 2.9 (very high!)

Shape-topology fit to sR_g = 7



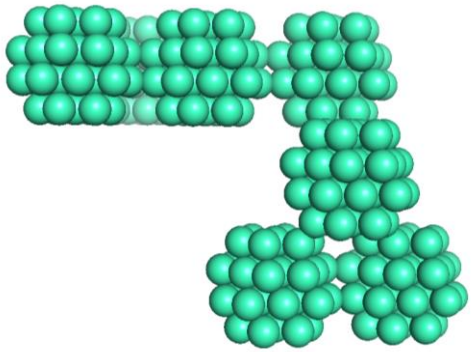


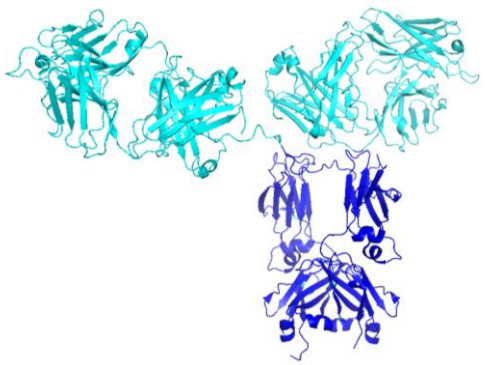
ATSAS tool: *ambimeter*

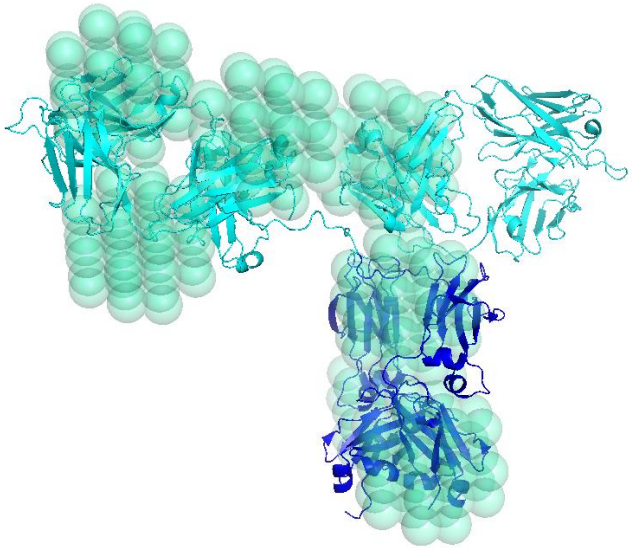
>ambimeter filename.out
>ambimeter –f=best –r=7 filename.out

Including higher-angle information…3 shape skeletons almost-fit the SAXS data!
...ambiguity score = 0.5 (between 0 to 1.5ish are 'potentially unique')

DARA and ambimeter...







To align two structures

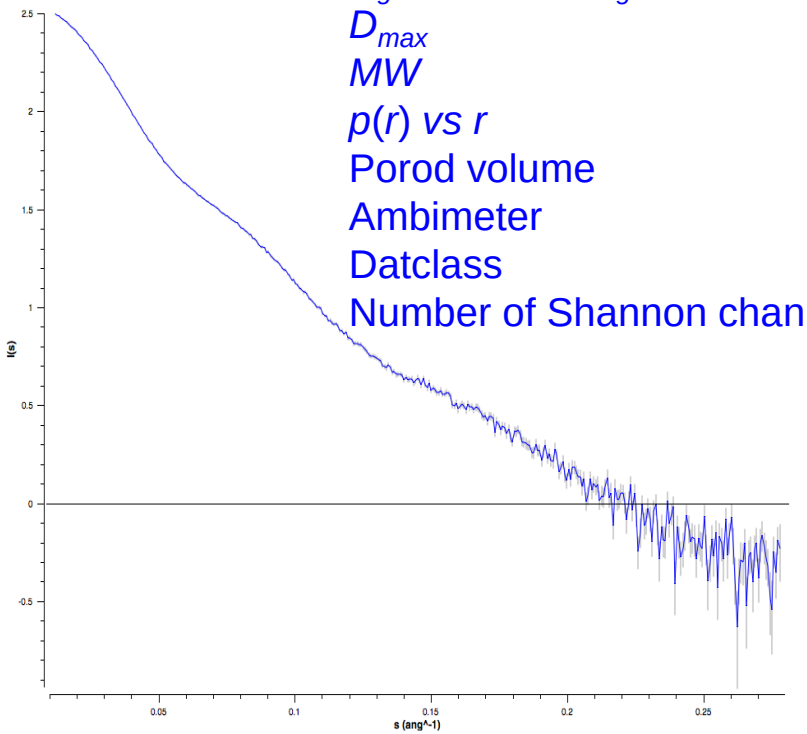
ATSAS tool: *supcomb*

At the command prompt (.cmd, terminal, etc) type:

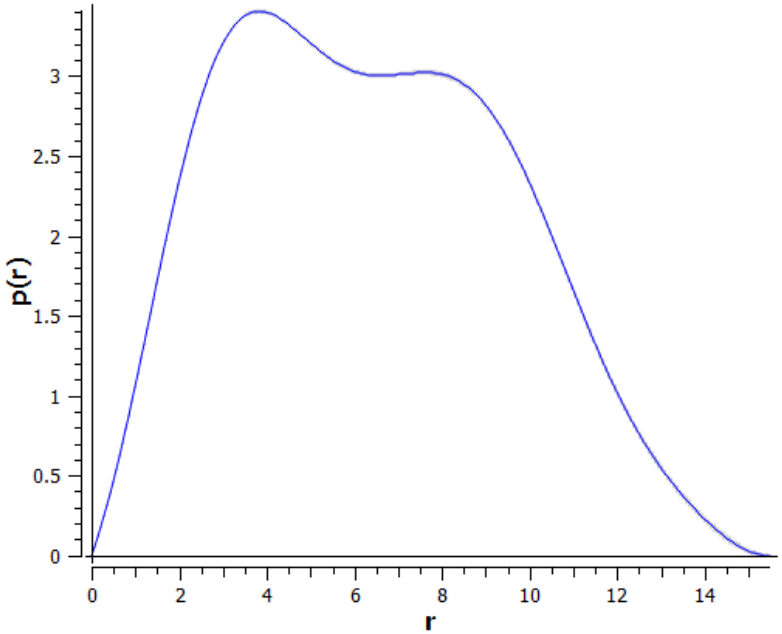
supcomb file1.pdb file2.pdb

Remember that enantiomorphs (mirror images) generate the same scattering profiles as each other! *DISABLE* when aligning atomistic structures! (supcomb file1.pdb file2.pdb –e=no)

You get all of this... *without* making 3D models

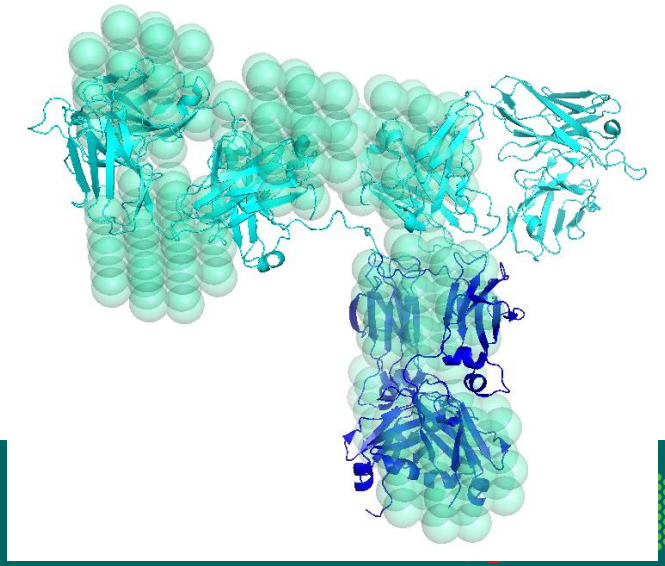


R_g (for rods, R^c_g; *flat thickness* R^t_g)
D_{max}
MW
p(*r*) *vs r*
Porod volume
Ambimeter
Datclass
Number of Shannon channels



Scattering from a flat/modular particle.

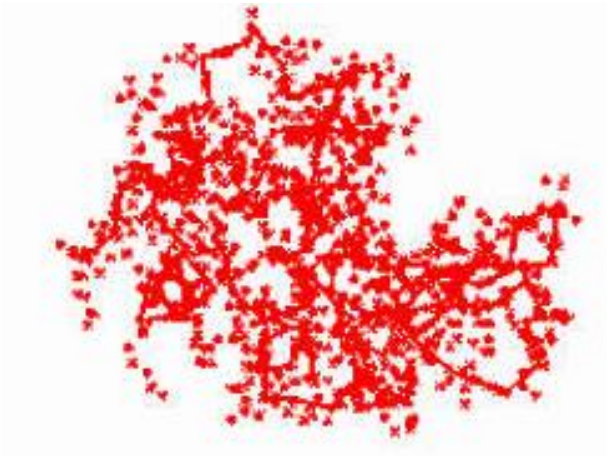
Might be ambiguous.



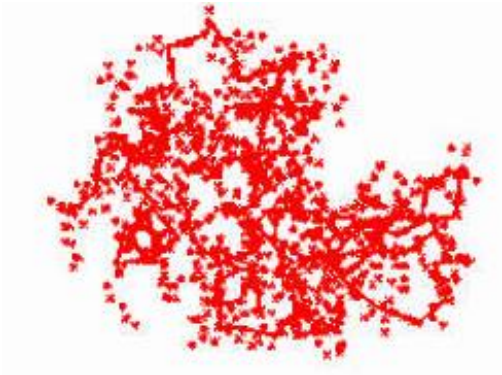
Just remember to always check the structural parameters!

- R_g and $I(0)$ from Guinier and $p(r)$.
- Molecular mass estimates.
- Identify concentration independent interparticle interactions: coulombic-repulsive or aggregation.

Lets do some high-resolution model FITTING!



Intuitively…

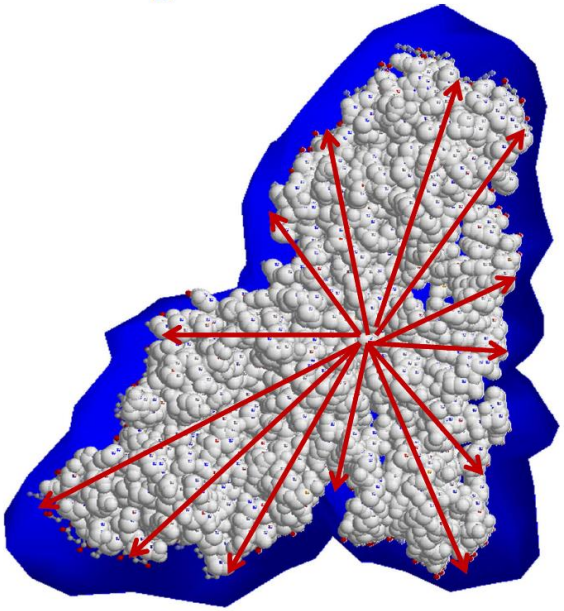


- *All Atom* .
- Use the Debye formula to calculate the modelled SAXS intensities

$$I(s) = \mathrm{f}^2(s) \sum_{i=1}^{M} \sum_{j=1}^{M} X_i X_j \frac{\sin(sr_{ij})}{sr_{ij}}$$

- Where rij = ri-ij, the distance between atoms i and j and f(s) the atomic form factor.
- Of course, it is never that easy.

In terms of fitting high-resolution structures to SAS data



Atomic scattering

- Excluded volume

+ Shell scattering

Convert the atomic coordinates of a model into a convenient mathematical expression for fitting or modelling.

Calculate the envelope function from the centre of the macromolecule from a common/coincident grid origin.

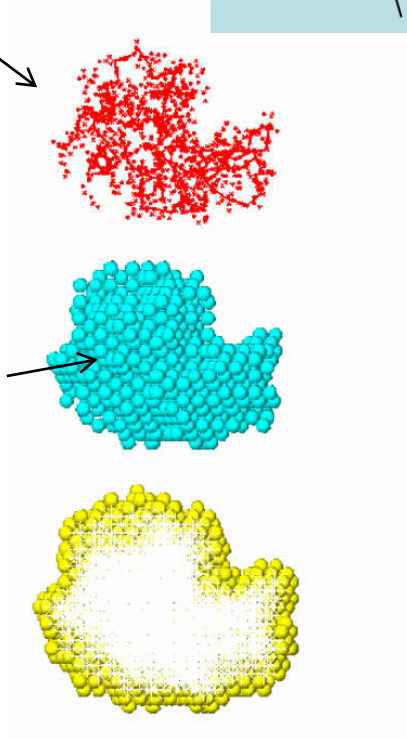
Take into account the atomic scattering, the excluded volume and hydration shell scattering.

SLD of the solvent

$$I(s) = \left\langle \left| A(\mathbf{s}) \right|^2 \right\rangle_\Omega = \left\langle \left| A_a(\mathbf{s}) - \rho_s A_s(\mathbf{s}) + \delta\rho_b A_b(\mathbf{s}) \right|^2 \right\rangle_\Omega$$

Electrons (nuclei) are 'points'

…atoms have volume and the macromolecule takes up space is in a solution



- $A_a(\mathbf{s})$: atomic scattering in vacuum
- $A_s(\mathbf{s})$: scattering from the excluded volume
- $A_b(\mathbf{s})$: scattering from the hydration shell

***CRYSOL* (X-rays):** Svergun et al. (1995). *J. Appl. Cryst.* **28**, 768

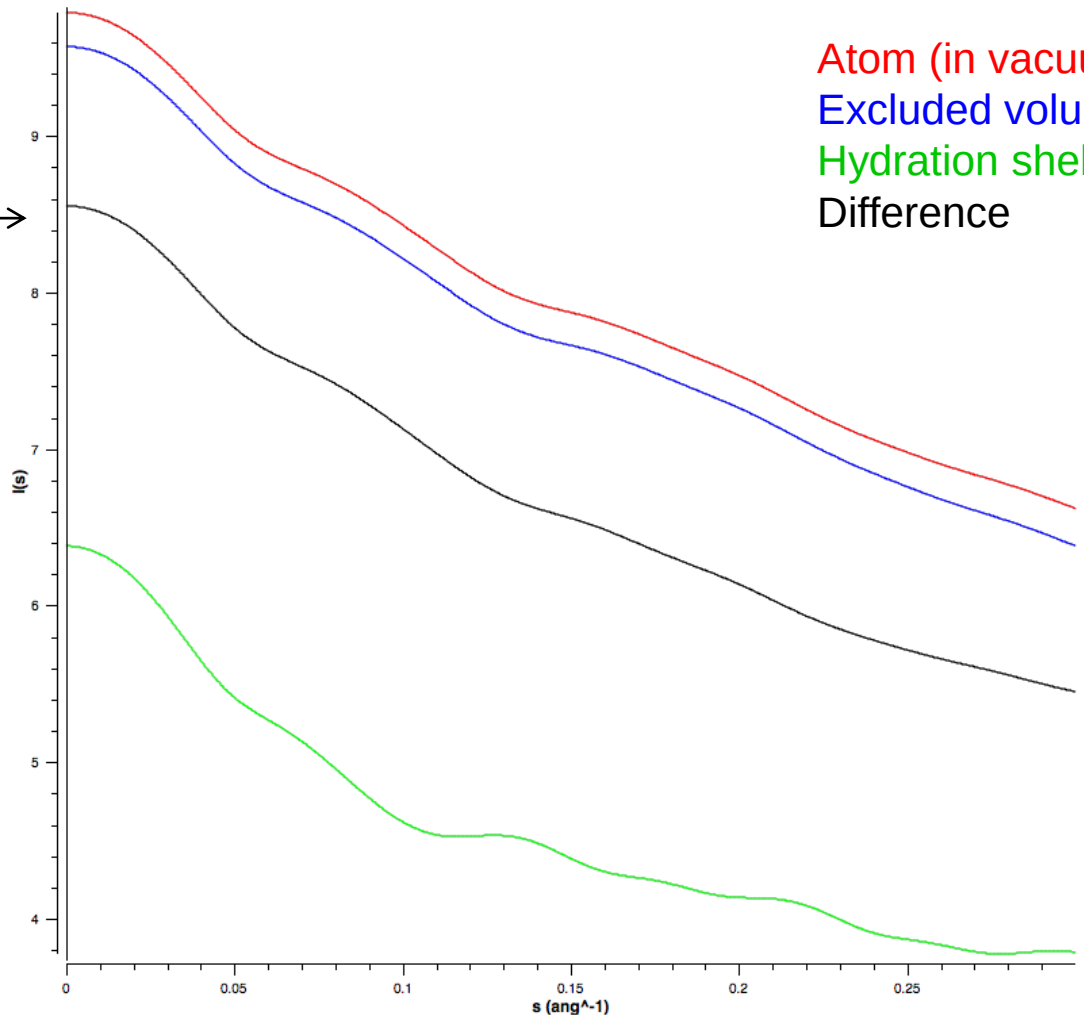
***CRYSON* (neutrons):** Svergun et al. (1998) *P.N.A.S. USA*, **95**, 2267

CRYSOL and CRYSON

$$I(s) = \left\langle \left| A(\mathbf{s}) \right|^2 \right\rangle_\Omega = \left\langle \left| A_a(\mathbf{s}) - \rho_s A_s(\mathbf{s}) + \delta\rho_b A_b(\mathbf{s}) \right|^2 \right\rangle_\Omega$$

- Either fit the experimental data by varying the density of the hydration layer δ*ρ* (affects the third term) and the total excluded volume (affects the second term).
- Or predict the scattering from the atomic structure using default parameters (theoretical excluded volume and bound solvent density of 1.1g/cm3).
- Provide output files (scattering amplitudes) for rigid body refinement routines.
- Compute particle envelope function F(ω)

Calculated contrast-weighted scattering from atomic coordinates



Atom (in vacuum)
Excluded volume (shape)
Hydration shell (envelope)
Difference

Why the hydration layer is important.

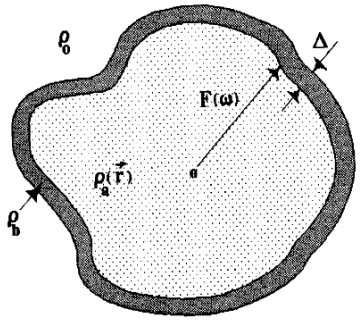


Fig. 1. Schematic representation of a macromolecule in solution. For explanations see text.

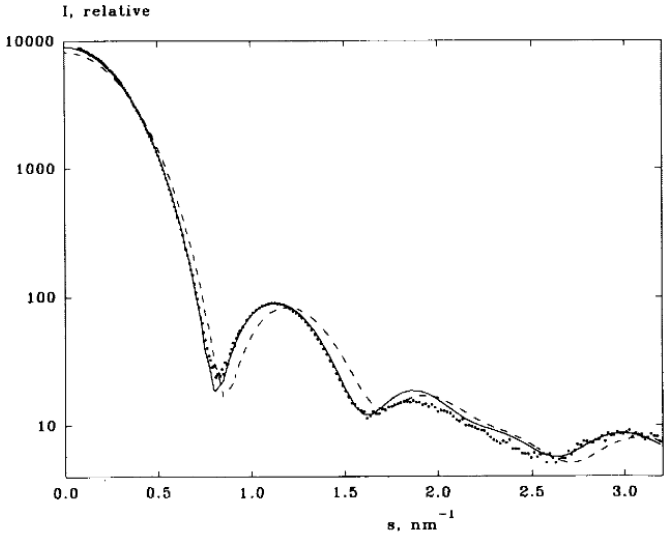
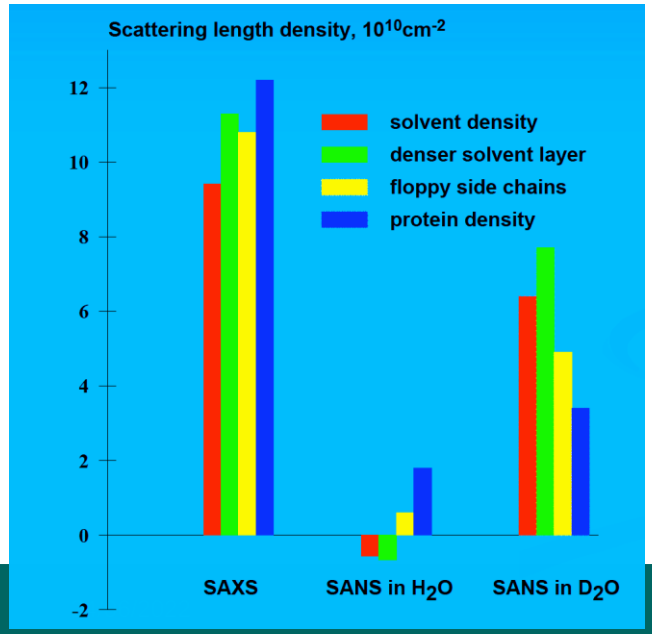
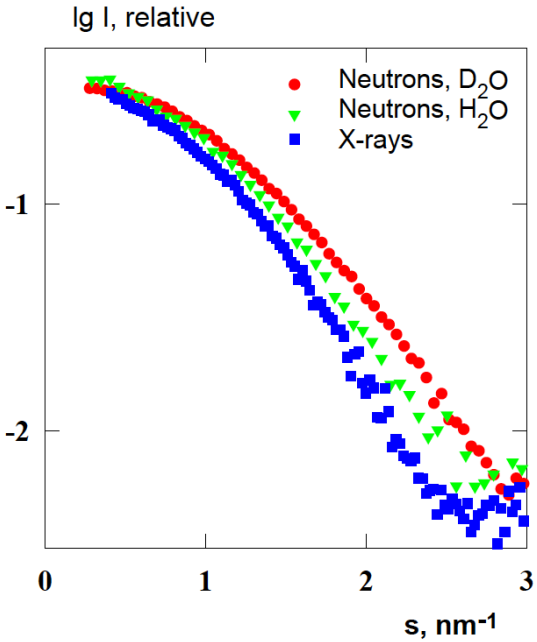


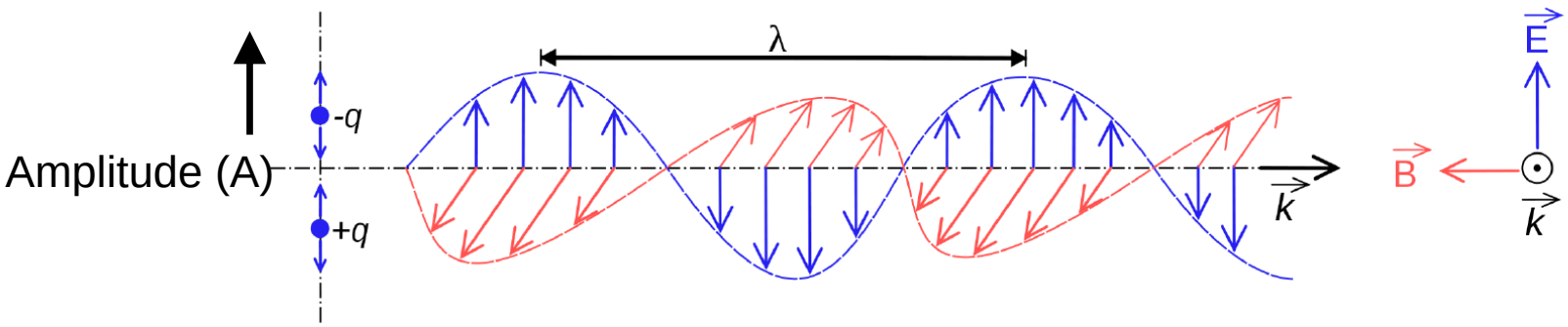
Fig. 5. Experimental solution scattering from ATCase and the fits with and without solvation shell. Notation is as in Fig. 3.

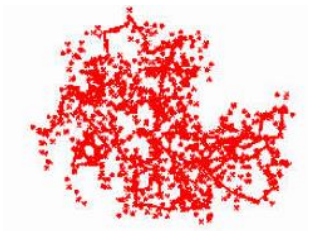




Lysozyme: appears larger for X-rays and smaller for neutrons in D_2O

How are the scattering amplitudes calculated? …





$$A(s) = \sum_{i=1}^{N} b_i \, e^{i\vec{s}\bullet\vec{r}}$$

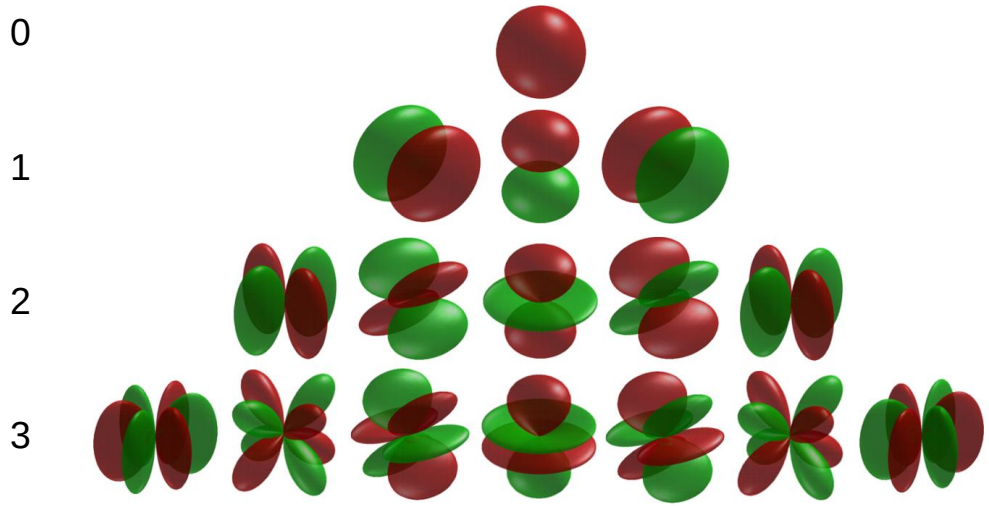
Spherical wave bit

'Scattering factor': relates to the atomic cross section, i.e., scattering length, or probability of an atom to scatter for every atom in the sample.

The 'spherical wave bit' can be mathematically expressed in terms of a summed set of independent ***spherical harmonics*** (as a multipole expansion):

$$\exp(i\mathbf{sr}) = 4\pi \sum_{l=0}^{\infty} \sum_{m=-l}^{l} i^l j_l(sr) Y^*_{lm}(\omega) Y_{lm}(\Omega)$$

What does this mean?



Essentially given a set of atomic coordinates in 3-dimensions (i.e., x, y, z coordinates), and knowing the identity of each atom at that coordinate (i.e., the atomic form factor), as well as the atomic volumes and scattering length densities, we can calculate the scattering amplitudes from the entire structure. As a result we can calculate the scattering intensities (i.e., the square of the scattering amplitudes.)

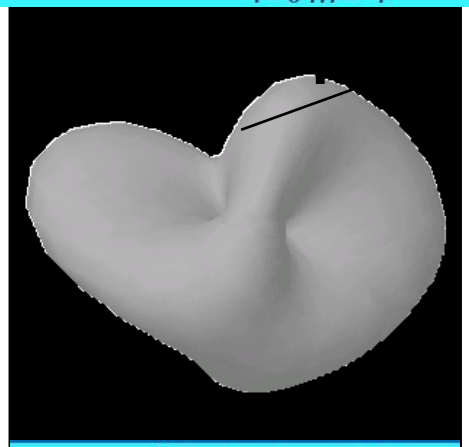
$$I(s) = 2\pi^2 \sum_{l=0}^{\infty} \sum_{m=-l}^{l} \left| A_{lm}(s) \right|^2$$

In 1970, Stuhrmann showed that the information content of a SAXS profile can be conveniently described in terms of a sum of spherical harmonic functions.

The envelope function is expressed as a sum of spherical harmonics

$F(\omega)$

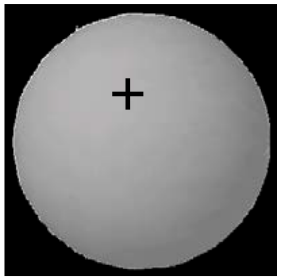
$$F(\omega) \cong F_L(\omega) = \sum_{l=0}^{L} \sum_{m=-l}^{l} f_{lm} \cdot Y_{lm}(\omega)$$



$$\rho(\mathbf{r}) = \begin{cases} 1, & 0 \le r \le F(\omega) \\ 0, & r > F(\omega) \end{cases}$$

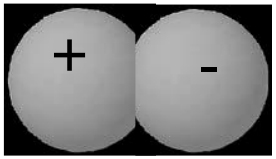
=

$A_{00}(s)$



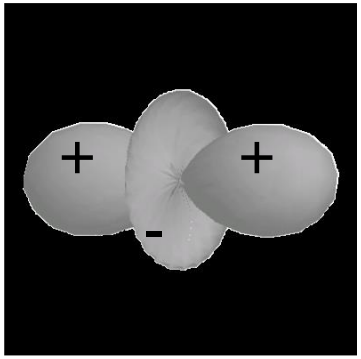
+

$A_{11}(s)$



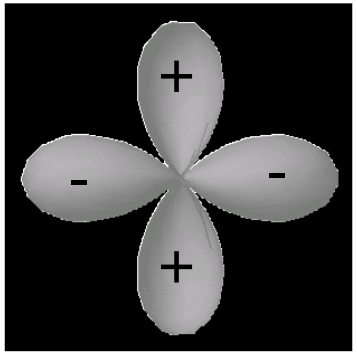
+

+



$A_{20}(s)$

+



$A_{22}(s)$

+ …

Remember $I(s)$ = <$I(s)$> = the Fourier transform of $\rho(r)$ squared i.e., <(F$\rho(r)$)^2>

How many spherical harmonics to use?

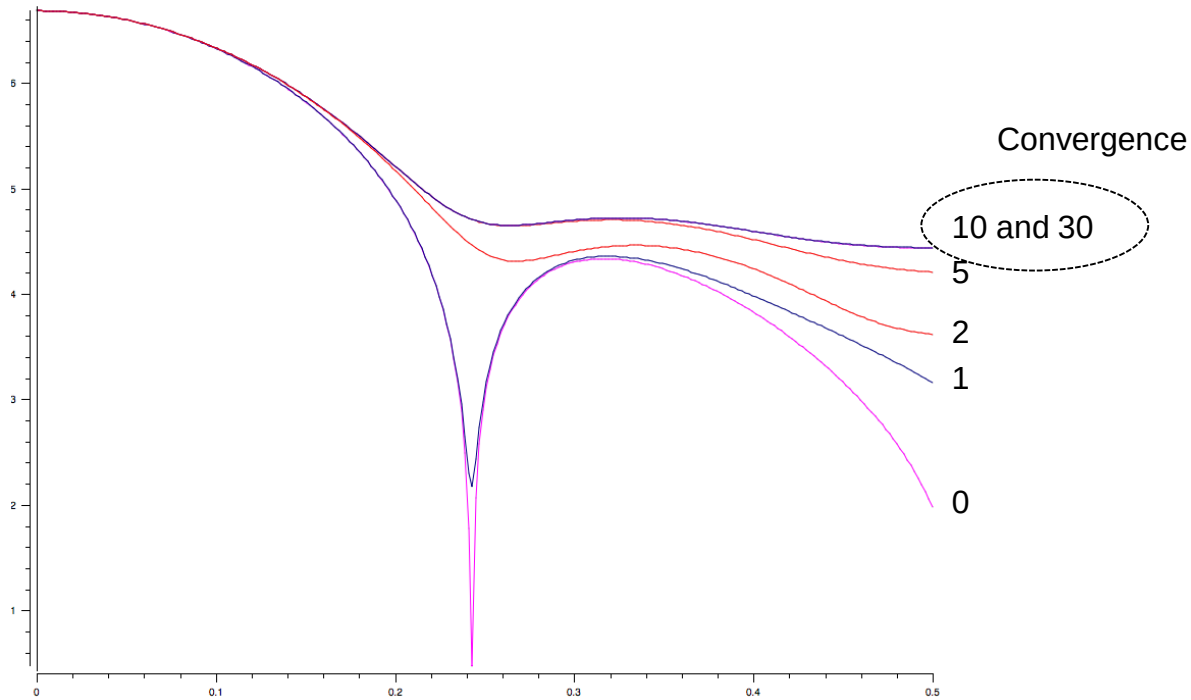
If you use the first harmonic only, i.e., zeroth-order, then the calculated intensities from the model will be a sphere. This is okay only if you want to describe the overall SIZE of the object, i.e., at the very lowest of angles in the Guinier region of the scattering profile. The zeroth-order harmonic dominates the very lowest angles of a calculated scattering profile!

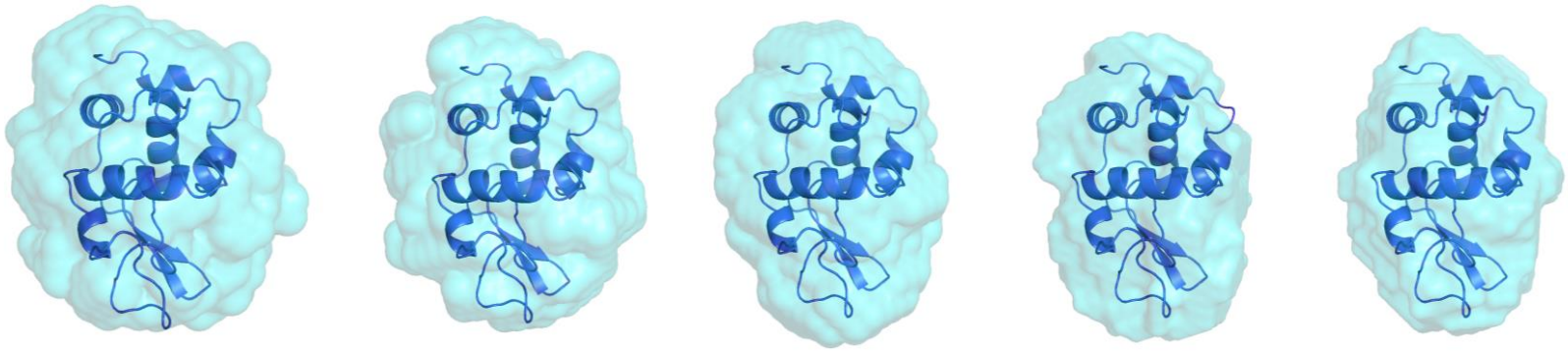
If you use two harmonics, you will introduce an additional 'shape feature' into the calculated scattering intensities across ***s***…but the resulting shape will probably still look like a sphere..with a couple of very low humps.

If you continue to increase the number of harmonics, you introduce atdditional shape features across ***s***. However, the more harmonics you introduce the less impact on the overall calculated scattering is observed at the low angles (i.e., in the SAXS regeime).

Typically 15-30 harmonics are used to describe size and the shape of the object. However, this depends on the <u>CLASSIFICATION</u> of an object. Clearly, if the object is an extended rod, you probably need additional spherical harmonics terms.

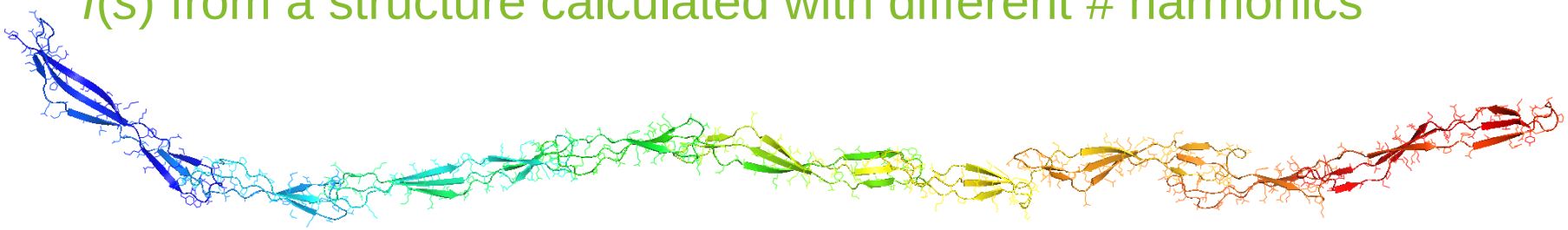
I(*s*) from a structure calculated with different # harmonics



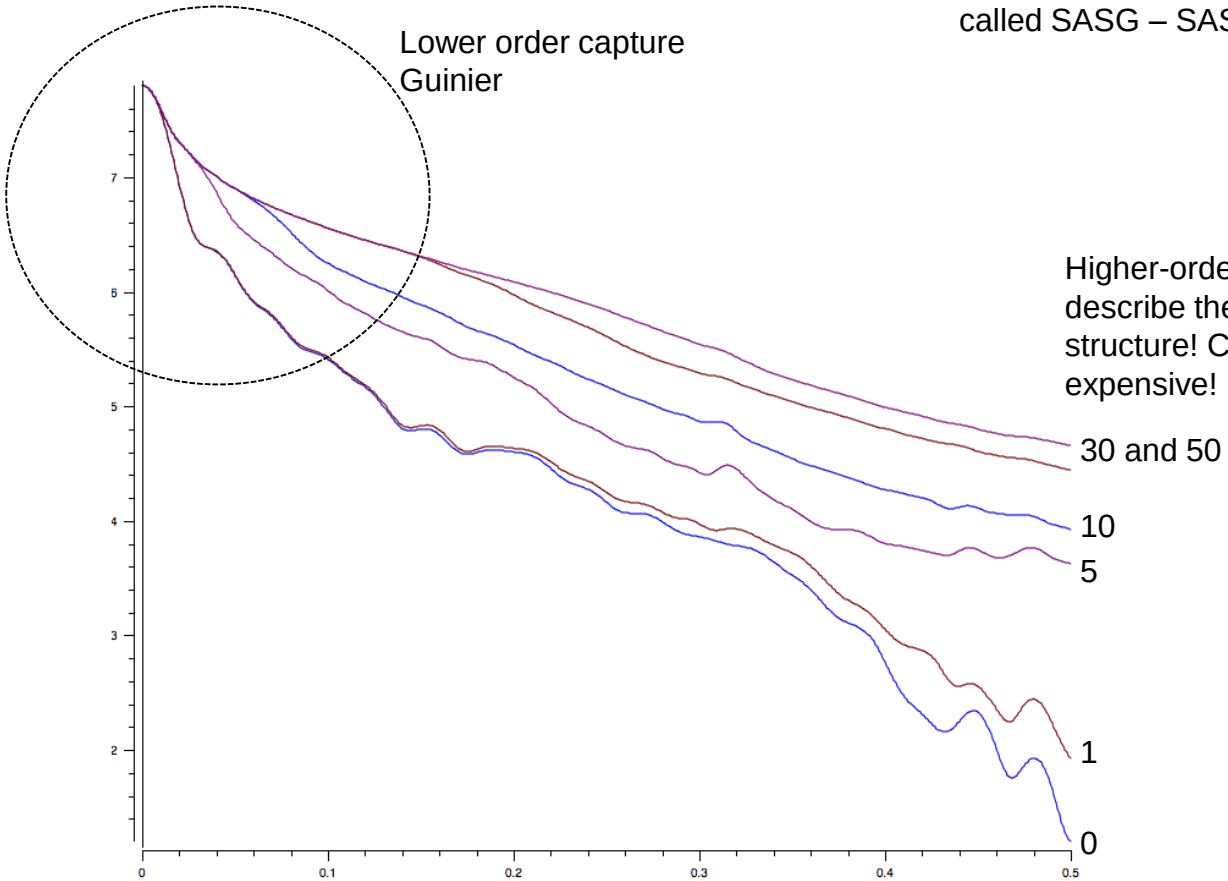


0 → 30

I(*s*) from a structure calculated with different # harmonics



…yes this protein is real. It is called SASG – SASBDB search it.



Higher-orders required to describe the anisotropic structure! Computationally expensive!

Centre your atomic models – always.

- THE MODEL SCATTEING AMPLITUDES (and therefore the resulting intensities) MUST BE CALCULATED FROM THE ORIGIN, i.e., the models must be centred, otherwise you loose low-order harmonic contributions.

ATSAS tool: *alpraxin*

Go to the folder where you have an atomic model.

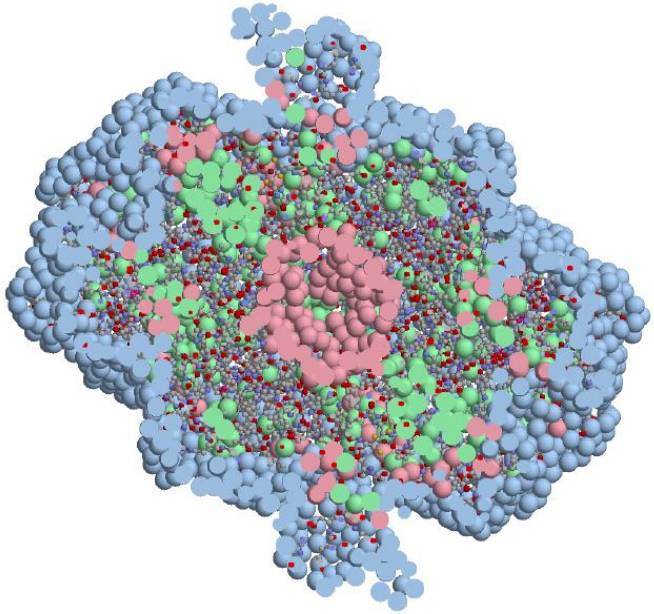
At the command prompt (.cmd, terminal, etc) type:

```
>alpraxin xxx.pdb
```

CRYSOL 3

Hydration shell representation as envelope function (CRYSOL) or dummy solvent beads (CRYSOL 3)

Fitting with hydration tuning or with default parameters



Outer shell

Internal shell

Small cavities (excluded volume adjustment)

Especially important for ring-shaped, hollow sphere, very small (less than 10 kDa) or very extended particles. Otherwise CRYSOL is fine.

Dealing with the hydration layer

Approach	Modeling of the hydration layer	Representation of the molecule	References
CRYSOL	Implicit layer using an envelope function	All-atom	Svergun et al. *J. Appl. Cryst*. (1995)
AXES	Explicit water molecules using equilibrated water boxes	All-atom	Grishaev *et al*. *JACS* (2010)
FoXS	Implicit layer based on surface accessibility	All-atom or coarse-grained	Schneidman-Duhovny *et al*. *NAR* (2010)
HyPred	Explicit water molecules based on MD simulations	All-atom	Virtanen *et al*. *Biophys. J*. (2011)
AquaSAXS	Solvent-density map using the dipolar PB-Langevin approach	All-atom	Poitevin *et al*. *NAR* (2011)

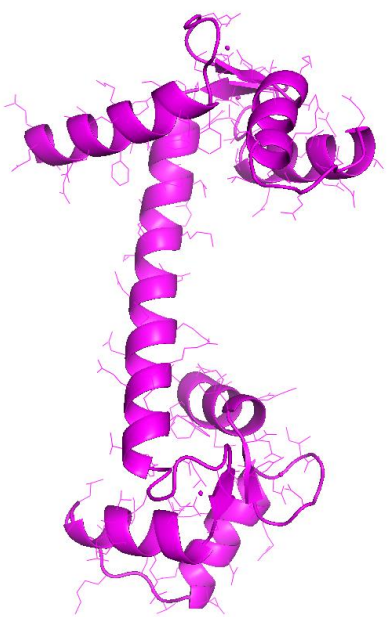
WAXIS – molecular dynamics for the hydration layer!

…knowing what models do NOT fit the data can be as valuable as knowing what models do fit the data.

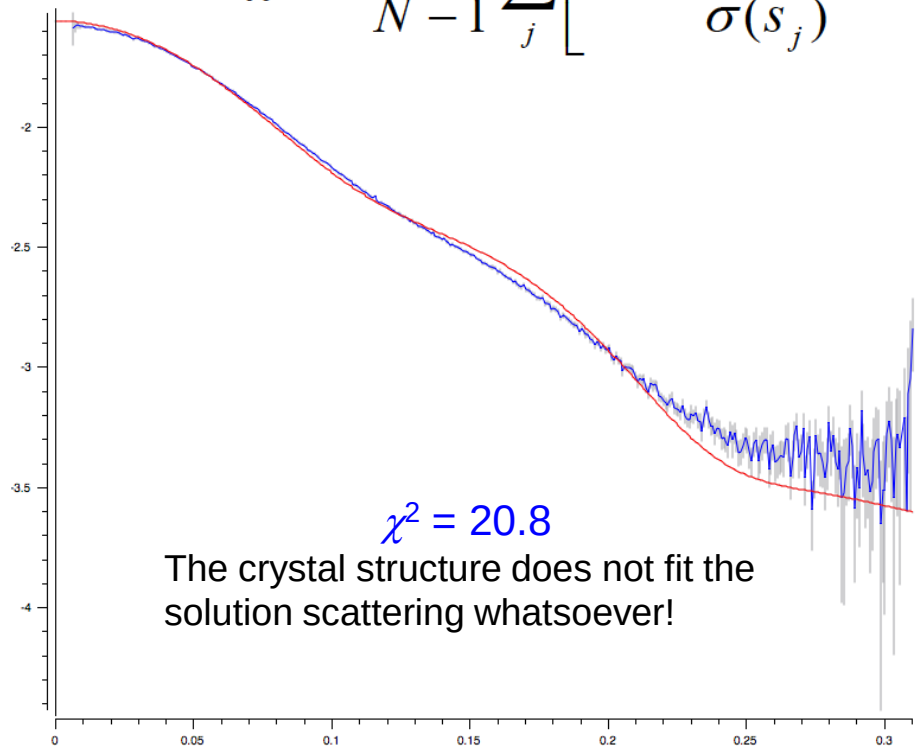
CRYSOL fit to the SAXS data. The goodness of fit is described by the χ^2 discrepancy.

$$\chi^2 = \frac{1}{N-1}\sum_j \left[\frac{I_{\exp}(s_j) - cI(s_j)}{\sigma(s_j)}\right]^2$$

Calmodulin: X-ray crystal structure



PDB: 3CLN



χ^2 = 20.8

The crystal structure does not fit the solution scattering whatsoever!

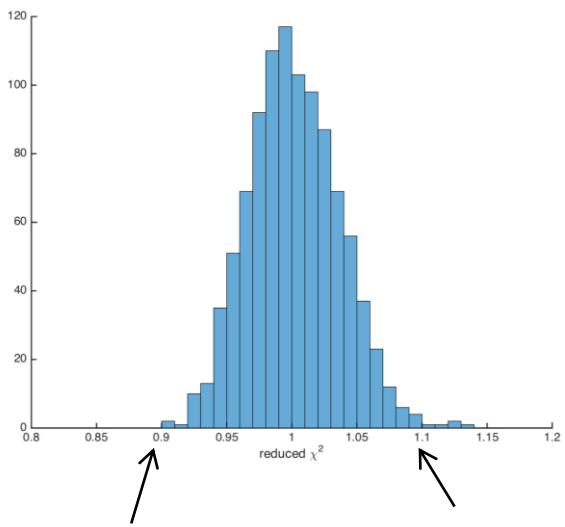
A note on χ^2.

$$\chi^2 = \frac{1}{N-1}\sum_j \left[\frac{I_{\exp}(s_j) - cI(s_j)}{\sigma(s_j)}\right]^2$$

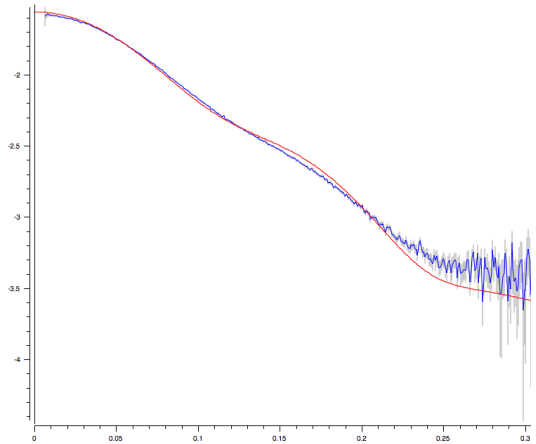
You need to correctly specify the errors on the scattering intensities, otherwise the test is, by default, absolutely INVALID.

If the errors are correctly specified and no significant (systematic) deviations are present between the experimental and modeled intensities, the value should lie in the range of approximately 0.9-1.1 depending on the number of points.

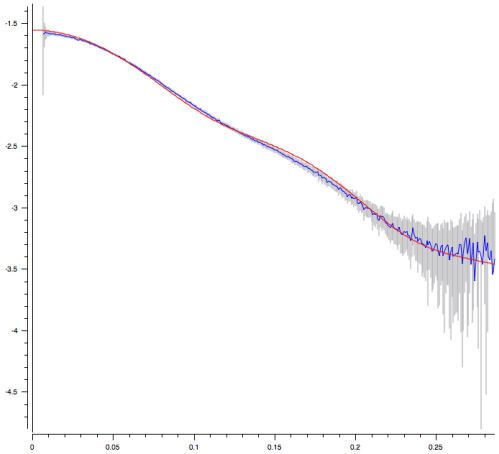
Same intensities, same model, but different error estimates



χ^2 = 20.8
With correct errors



χ^2 = 1.2
With incorrect errors



Correlation Map

$$J = \begin{pmatrix} \vdots \\ I(q_k) \\ \vdots \end{pmatrix}, \quad \Sigma = \begin{pmatrix} \ddots & & & \\ & \sigma(I(q_k))^2 & \cdots & \sigma(I(q_k), I(q_l)) \\ & \vdots & \ddots & \vdots \\ & \sigma(I(q_l), I(q_k)) & \cdots & \sigma(I(q_l))^2 \\ & & & & \ddots \end{pmatrix}$$

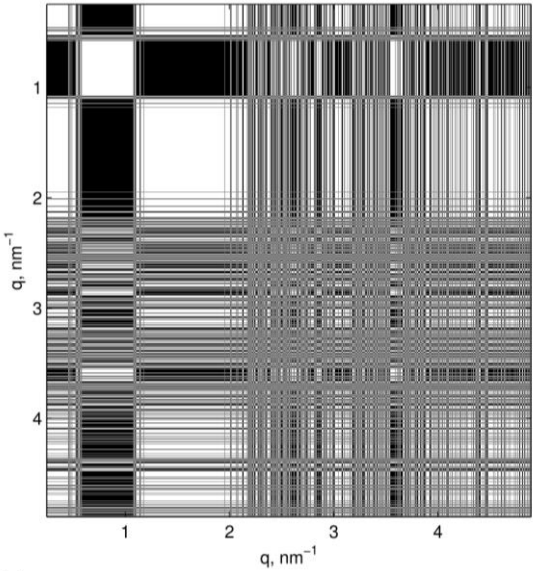
$$\sigma(I_{\exp}(q_k))^2 = \frac{1}{m-1}\sum_{i=1}^{m}(I_{\exp}(q_k)_i - \bar{I}_{\exp}(q_k))^2$$

On-diagonal variance.

$$\sigma(I_{\exp}(q_k), I_{\exp}(q_l)) =$$

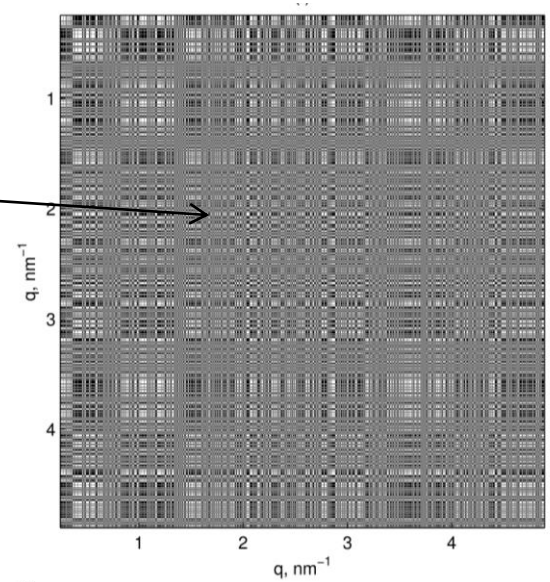
$$\frac{1}{m-1}\sum_{i=1}^{m}(I_{\exp}(q_k)_i - \bar{I}_{\exp}(q_k))(I_{\exp}(q_l)_i - \bar{I}_{\exp}(q_l))$$

Off-diagonal co-variance between all point-to-point q_k and q_l.



P < 0.01

View as a +/- 1 'map': random small patches = low probability of systematic differences (i.e., the pairwise comparison fits)!



P > 0.01

Remember this?

The scattering intensity

Is the SUM of all macromolecules averaged over all orientations.

The structure factor or 'between particle' contributions

$$I(s) = \sum_{i}^{n} [(\Delta\rho_i V_i)^2 P_i(s)] S(s)$$

Weighted by the contrast and volume SQUARED of all macromolecules

The form factor of all macromolecules within the sample

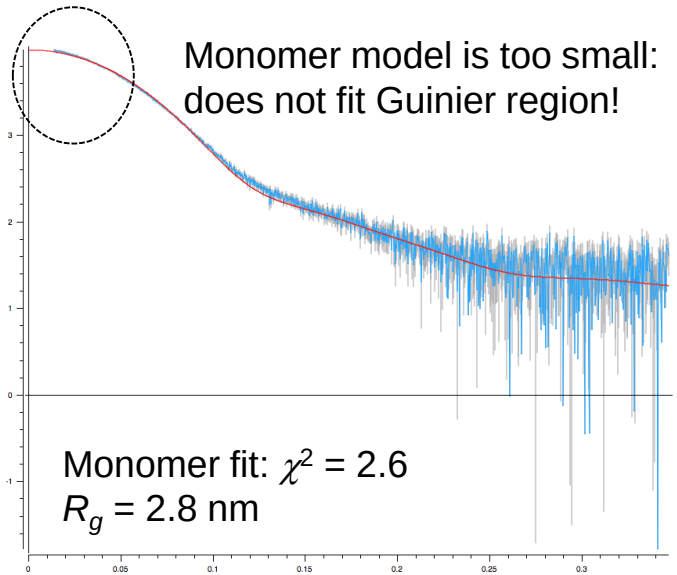
Scattering from Mixtures.

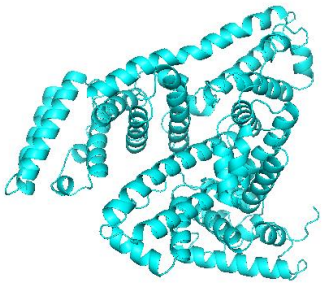
- Possible to obtain the volume fraction contribution to the total scattering profile of individual components of mixtures.

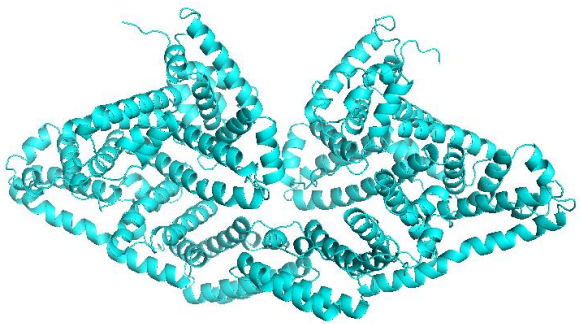
$$I(s)=\sum_{k} v_k I_k(s)$$

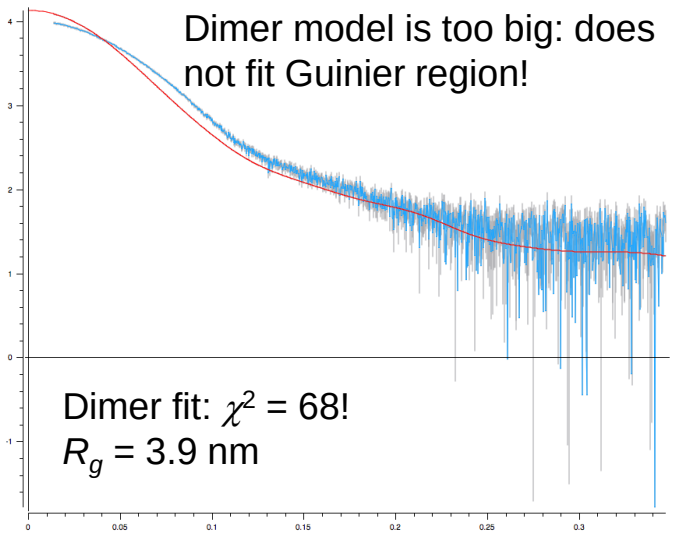
- Equilibrium analysis through a concentration series.

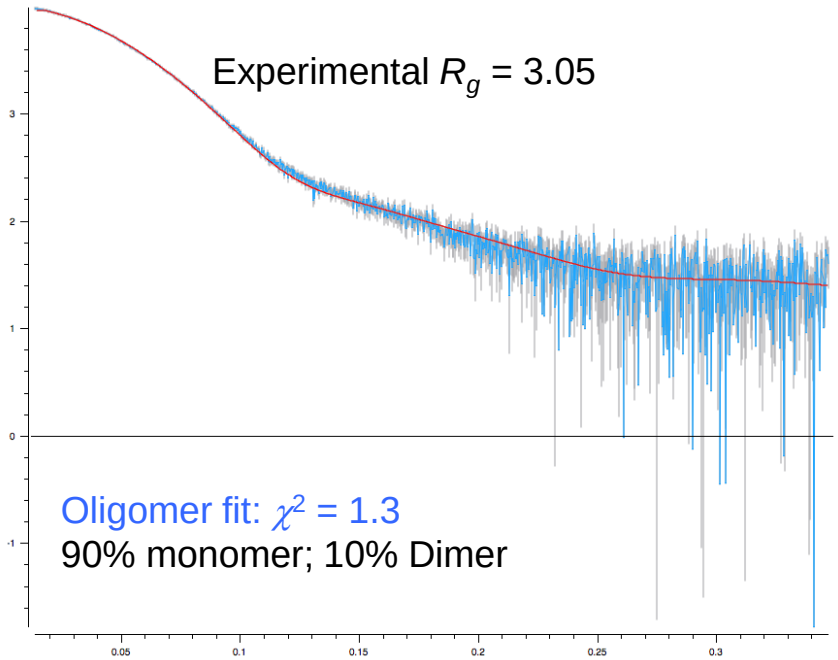
Dealing with mixtures: Use OLIGOMER











Structure does not fit? Try some Rigid body modelling…

Modelling 3D-structure**s** that fit SAS data is perhaps the fundemental ‘art’ of small-angle scattering.

The major considerations to keep in mind when modelling SAS data are:

There is often more than one model that fits the data equally well.

SAS data is inherently noisy.

SAS data is inherently ambiguous.

Lets do the easy bit first: get the right sequence and the right PDB file(s).

- You should know the amino acid sequence of the protein (or polynucleotide) used for the SAS experiment. The expected protein sequence can be obtained from the gene sequence (yes I am a biologist!)
- You should know what rigid-body (or bodies) you want to use for the modelling, i.e., the atomic coordinate PDB files (.pdb format).

- Extract the amino acid sequence from the PDB file.
- Align the PDB amino acid sequence with the amino acid sequence of the ***EXACT*** protein used for the SAS experiment.
- Deal with missing side-chains in the atomic coordinate file (account for ***ALL OF THE MASS***).

Amino acid sequence of protein used for SAS

HMHHHHHHHTRGSNNEEAICSLCDKKIRDRFVS
KVNGRCYHSSCLRCSTCKDELGATCFLREDSM
YCRAHFYKKFGTKCSSCNEGIVPDHVVRKASN
HVYHVECFQCFICKRSLETGEEFYLIADDARLV
CKDDYEQARDGGSGGHMGSGGGIGPLMVQP
ATPHIDNTLGGPIDIQHF

↓

Align the sequences using Clustal Omega

http://www.ebi.ac.uk/Tools/msa/clustalo/

↑

GSNNEEAICSLCDKKIRDRFVSKVNGRCYHSS
CLRCSTCKDELGATCFLREDSMYCRAHFYKKF
GTKCSSCNEGIVPDHVVRKASNHVYHVECFQC
FICKRSLETGEEFYLIADDARLVCKDDYEQARD
GGSGGHMGSGGGIGPLMVQPATPHIDNTLGG
PIDIQHF

Amino acid sequence of protein from PDB file

← This will save the sequence in the text file called 'filename.txt'

↓

pdb2seq filename.pdb > filename.txt

↓

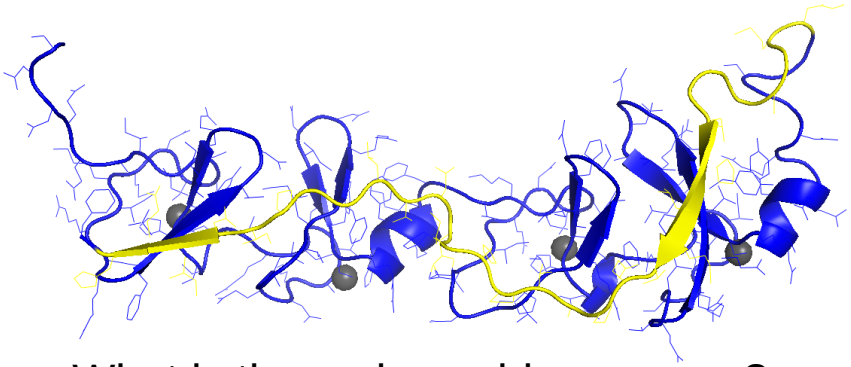
ATSAS tool: *pdb2seq*

At the command prompt (.cmd, terminal, etc) type:

What is the amino acid sequence?

↓

Atomistic model from PDB file (filename.pdb)



www.ebi.ac.uk/Tools/msa/clustalo/

EMBL-EBI | Services | Research | Training | Industry | About us | EMBL-EBI Hinxton

Clustal Omega

Input form | Web services | Help & Documentation | Feedback | Share

Tools > Multiple Sequence Alignment > Clustal Omega

Multiple Sequence Alignment

Clustal Omega is a new multiple sequence alignment program that uses seeded guide trees and HMM profile-profile techniques to generate alignments between **three or more** sequences. For the alignment of two sequences please instead use our pairwise sequence alignment tools.

Important note: This tool can align up to 4000 sequences or a maximum file size of 4 MB.

STEP 1 - Enter your input sequences

Enter or paste a set of

PROTEIN

sequences in any supported format:

```
>SAS_Protein
HMHHHHHHTRGSNNEEAICSLCDKKIRDRFVSKVNGRCYHSSCLRCSTCKDELGATCFLREDSMYCRAHFYKKFGTKCSSCNEGIVPDHVVRKASN
HVYHVECFQCFICKRSLETGEEFYLIADDARLVCKDDYEQARDGGSGGHMGSGGGIGPLMVQPATPHIDNTLGGPIDIQHF

>PDB_sequence
GSNNEEAICSLCDKKIRDRFVSKVNGRCYHSSCLRCSTCKDELGATCFLREDSMYCRAHFYKKFGTKCSSCNEGIVPDHVVRKASNHVYHVECFQCF
ICKRSLETGEEFYLIADDARLVCKDDYEQARDGGSGGHMGSGGGIGPLMVQPATPHIDNTLGGPIDIQHF
```

Or, upload a file: Browse... No file selected.

EMBL-EBI | Services | Research | Training | Industry | About us

Clustal Omega

Input form | Web services | Help & Documentation

Tools > Multiple Sequence Alignment > Clustal Omega

Results for job clustalo-I20170905-075837-0782-22232

Alignments | Result Summary | Phylogenetic Tree | Submission Details

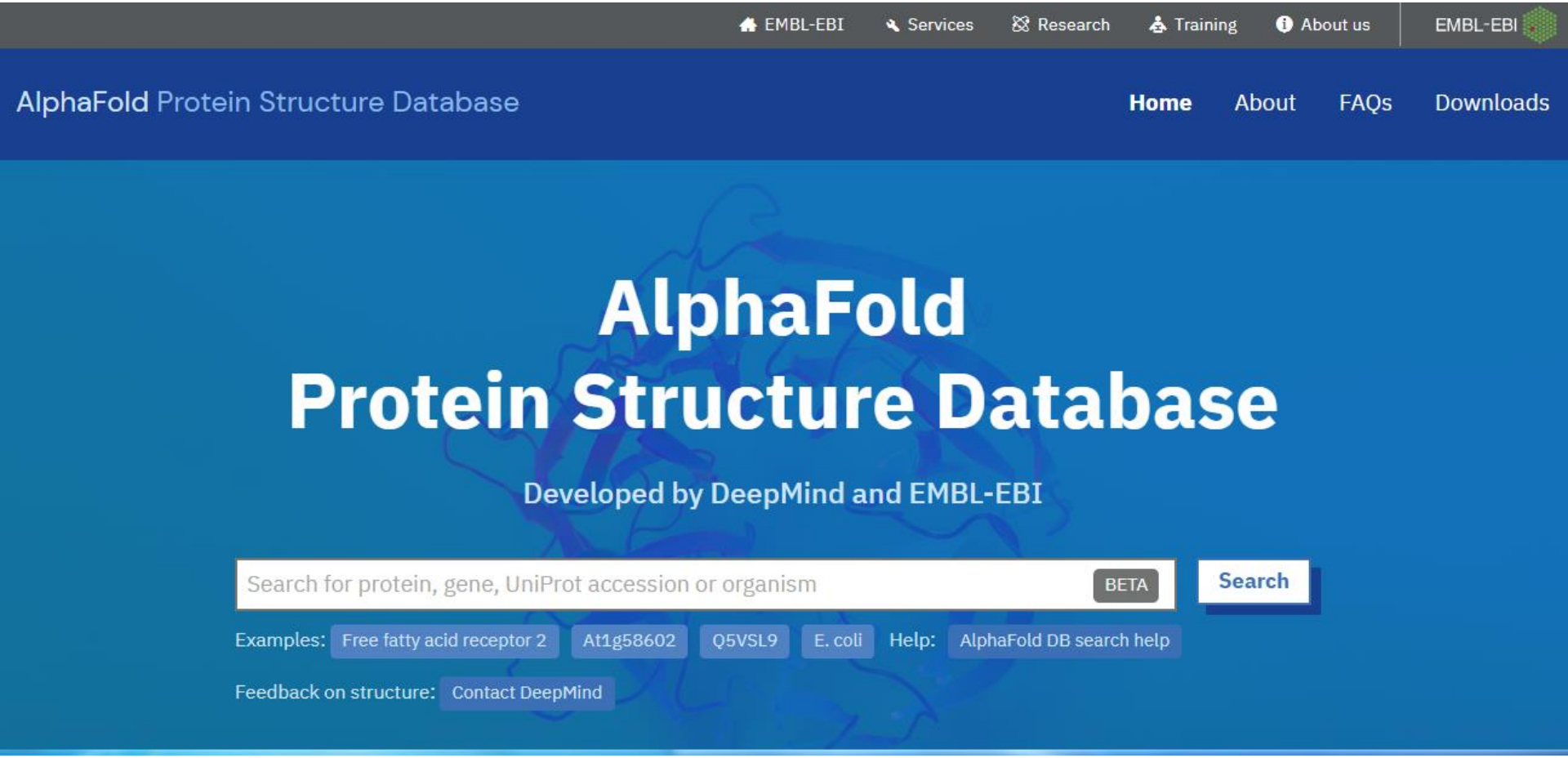
Download Alignment File | Show Colors | Send to Simple_Phylogeny

```
CLUSTAL O(1.2.4) multiple sequence alignment


SAS_Protein       HMHHHHHHTRGSNNEEAICSLCDKKIRDRFVSKVNGRCYHSSCLRCSTCKDELGATCFLR
PDB_sequence      ----------GSNNEEAICSLCDKKIRDRFVSKVNGRCYHSSCLRCSTCKDELGATCFLR
                            **********************************************

SAS_Protein       EDSMYCRAHFYKKFGTKCSSCNEGIVPDHVVRKASNHVYHVECFQCFICKRSLETGEEFY
PDB_sequence      EDSMYCRAHFYKKFGTKCSSCNEGIVPDHVVRKASNHVYHVECFQCFICKRSLETGEEFY
                  ************************************************************

SAS_Protein       LIADDARLVCKDDYEQARDGGSGGHMGSGGGIGPLMVQPATPHIDNTLGGPIDIQHF
PDB_sequence      LIADDARLVCKDDYEQARDGGSGGHMGSGGGIGPLMVQPATPHIDNTLGGPIDIQHF
                  ***********************************************************
```

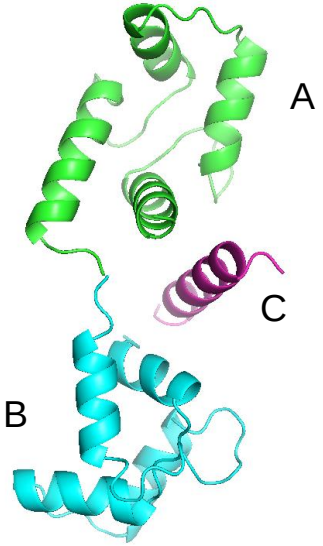
PLEASE NOTE: Showing colors on large alignments is slow.

Oops! Part of the sequence missing in the PDB file! This missing fragment will have to be built. Do not worry…ATSAS rigid-body modelling programs can deal with this!

ALPHAFOLD - https://alphafold.ebi.ac.uk/



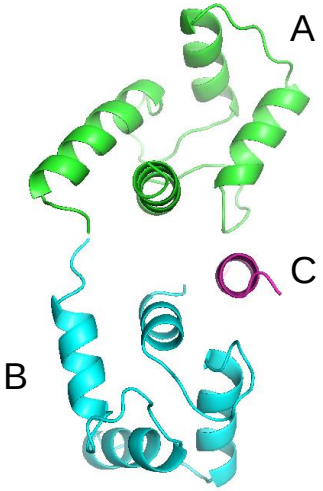
Rigid body modelling basics

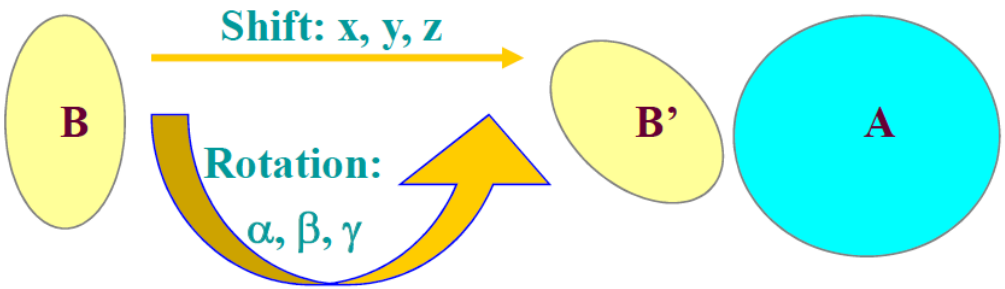


The structures of two (or more) subunits in reference positions are known.

Arbitrary complex can be constructed by moving and rotating the subunits.

This operation depends on three Euler rotation angles and three Cartesian shifts.





The partial amplitudes of a rotated and displaced subunit are expressed *via* the initial amplitudes, three Euler rotation angles and three Cartesian shifts):

$$A^{(i)}_{lm}(s) = A^{(i)}_{lm}(s)\ \{A_0{}^{(i)}_{lm}(s),\ \alpha^{(i)},\ \beta^{(i)},\ \gamma^{(i)},\ x^{(i)},\ y^{(i)},\ z^{(i)}\}.$$

$$I(s) = 2\pi^2 \sum_{l=0}^{L} \sum_{m=-l}^{l} \left| \sum_{n} A^{n}{}_{lm}(s) \right|^2$$

For symmetric particles, there are fewer parameters and the calculations are faster

Svergun, D.I. (1991). *J. Appl. Cryst.* **24**, 485-492

The target function:

$$E(\{X\}) = \chi^2[(I(s), I_{\exp}(s)] + \sum_i \alpha_i P_i$$

is minimized…basically χ^2 plus penalties!

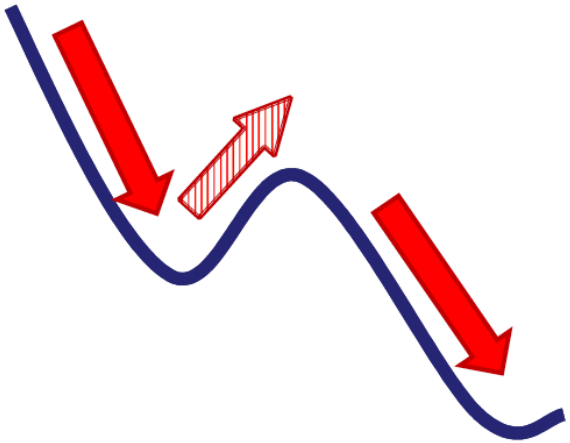
Penalties describe model-based restraints and/or introduce the available additional information from other methods: MX, NMR, EM, Alphafold etc).

A brute force (grid) search is applied if the number of free parameters is small.

Otherwise a Monte-Carlo based technique (e.g. simulated annealing) is employed to perform the minimization of E({X}).

Simulated annealing

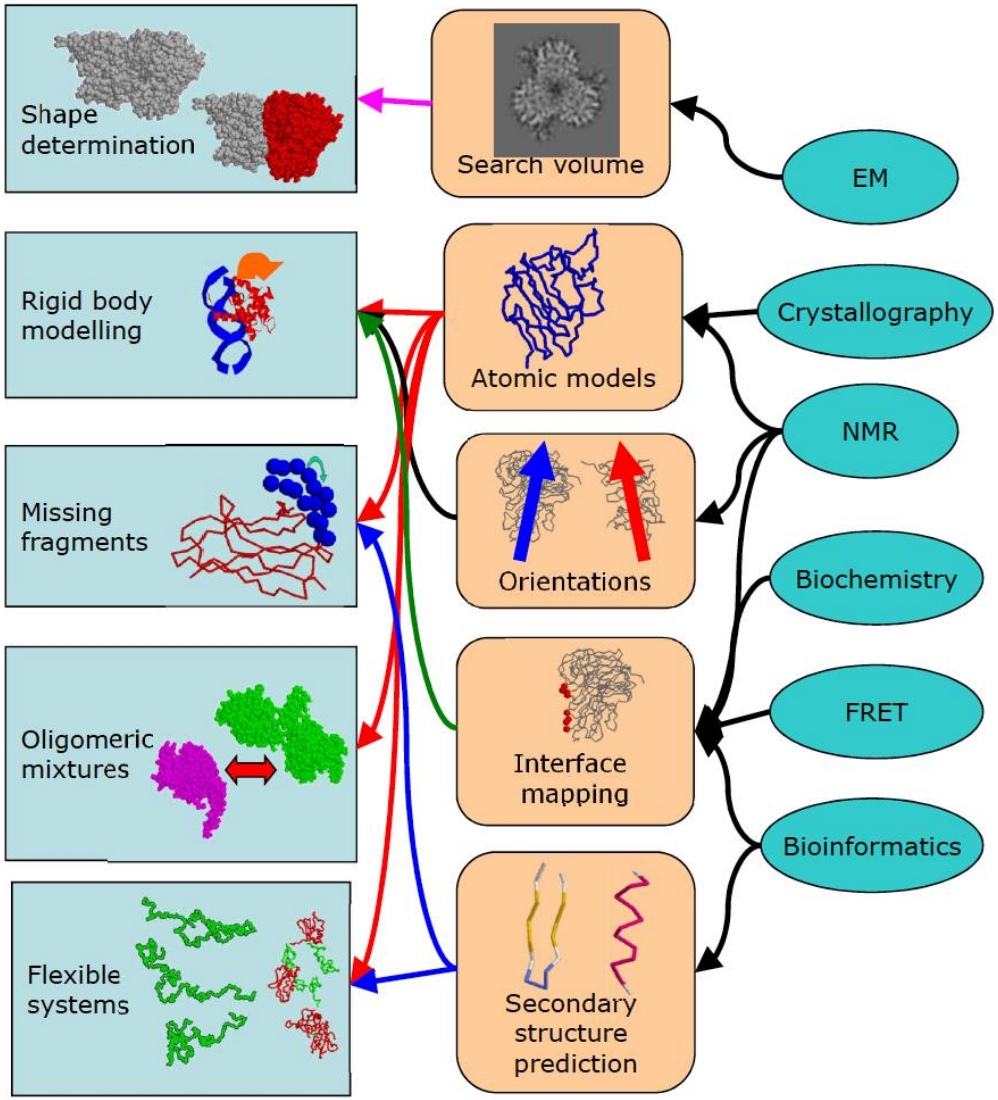
- **Main idea:** Minimization of the target function *E(**X**)* by random modifications of the system always moving to configurations that decrease *E(**X**)* but to also occasionally move to configurations that increase the scoring function.



A note on χ^2.

$$\chi^2 = \frac{1}{N-1}\sum_j \left[\frac{I_{\exp}(s_j) - cI(s_j)}{\sigma(s_j)}\right]^2$$

$$E(\{X\}) = \chi^2[(I(s), I_{\exp}(s)] + \sum_i \alpha_i P_i$$



+

Default ‘sensible’ modelling restraints like:

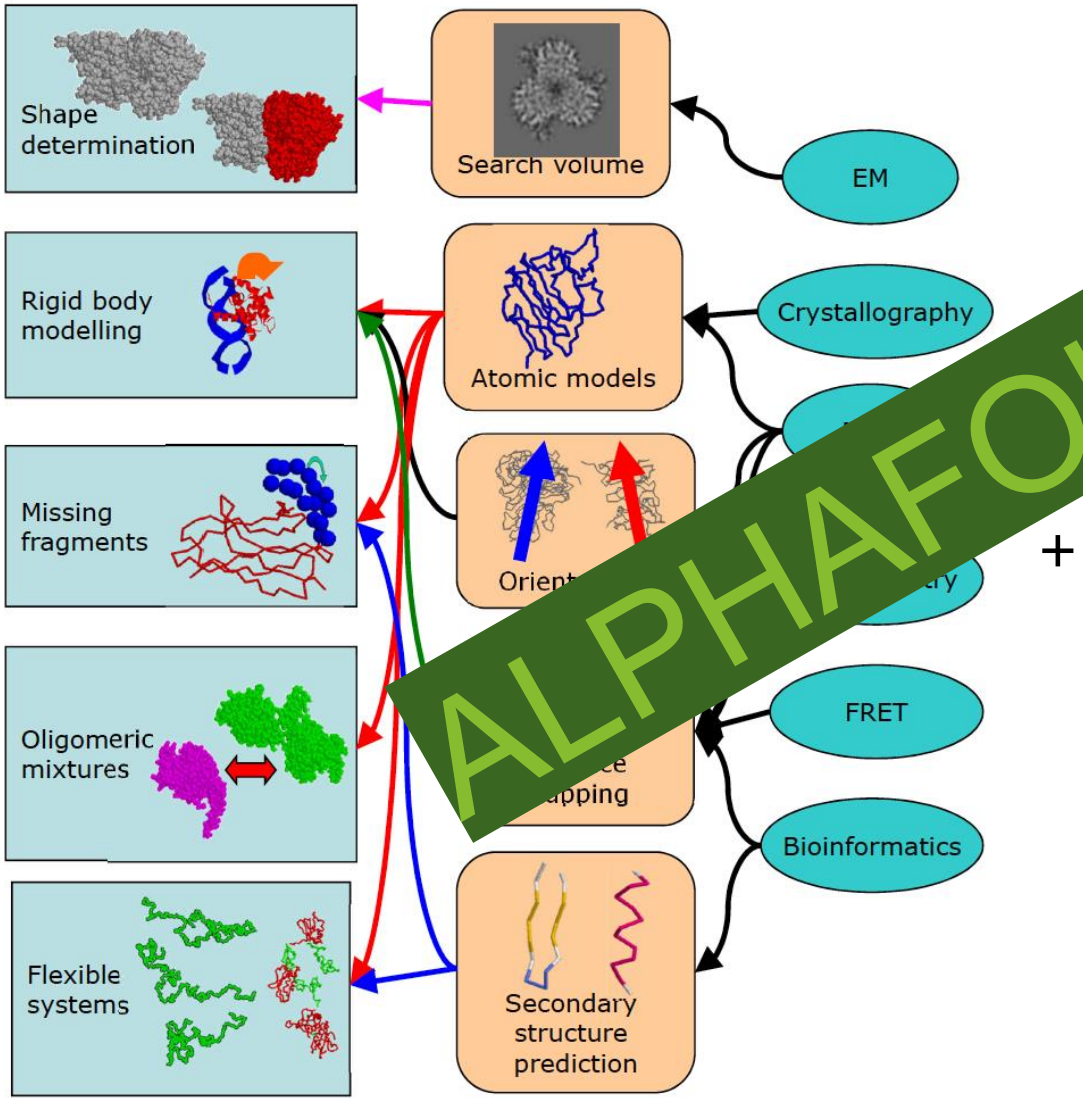
Minimise clashes.

Maintain contacts.

Don’t shift too far from the origin!

For dummy residues, make dihedral angles and Ramachandran geometry sensible.

Do not inter-penetrate subunits (interconnectivity).



ALPHAFOLD.

+

Default ‘sensible’ modelling restraints like:

N ashes.

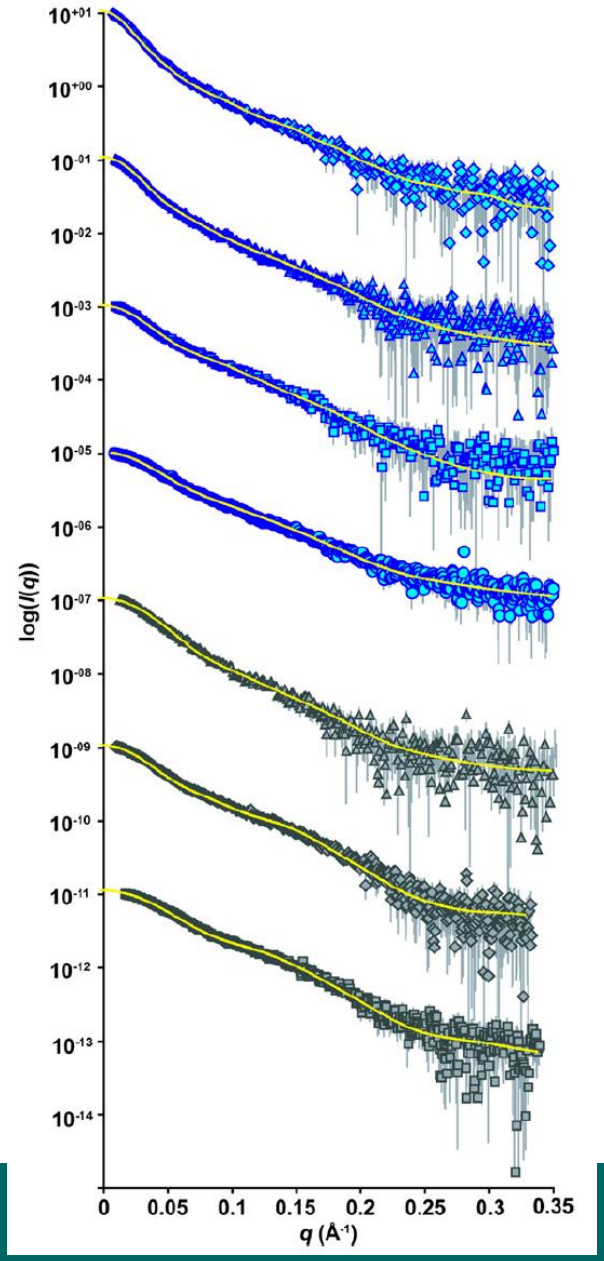
tain contacts.

Don’t shift too far from the origin!

For dummy residues, make dihedral angles and Ramachandran geometry sensible.

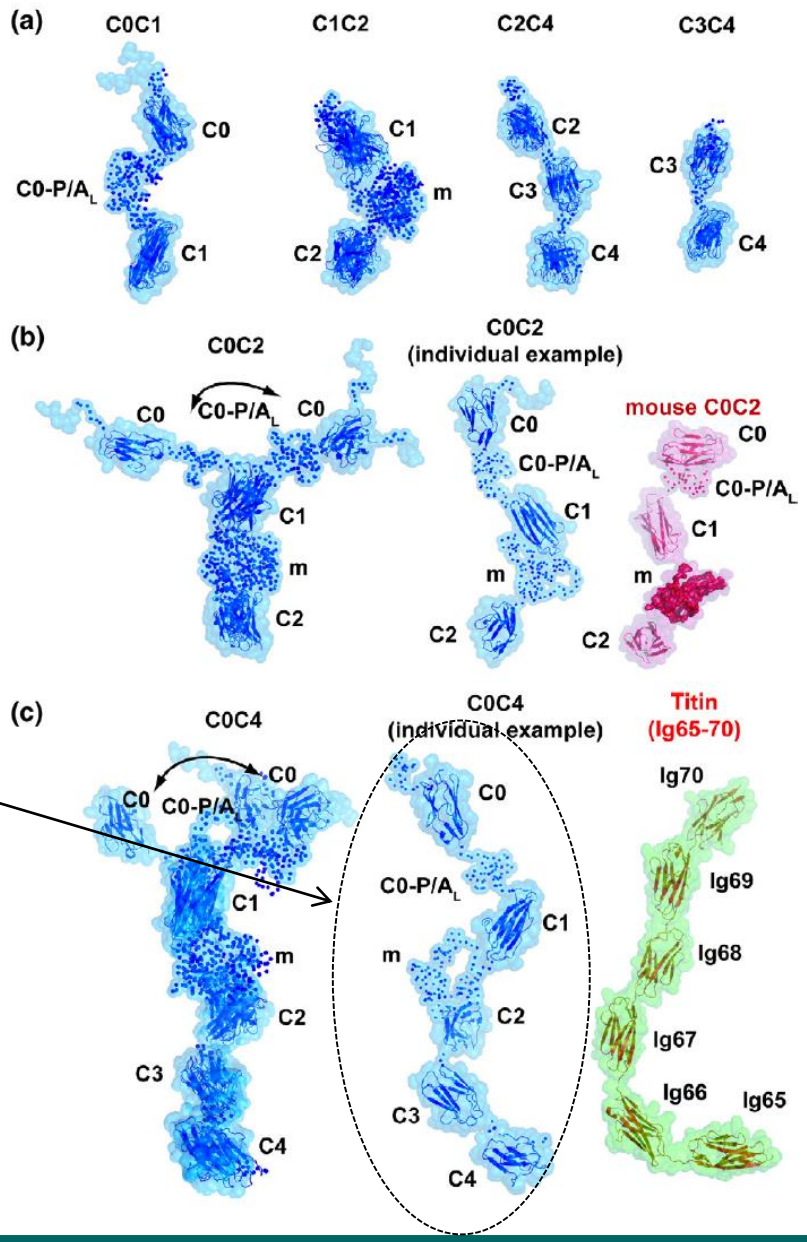
Do not inter-penetrate subunits (interconnectivity).

…more information = less ambiguity!



Parallel SAXS modelling of domain and domain constructs (truncation mutants)

Final Target: fits but importantly DOES NOT describe *in toto* what is going on



SANS with Contrast variation

Contrast variation means to collect SANS data from samples and buffers across several % v/v ^2H_2O concentrations in the solvent.

0%, 20% 40%, 80% 90% 100%

A series of linear equations can be used to *extrapolate* the component scattering functions.

For a single component
(One scattering length density)

$$I(q_{\text{total}}) \propto \Delta\rho^2$$

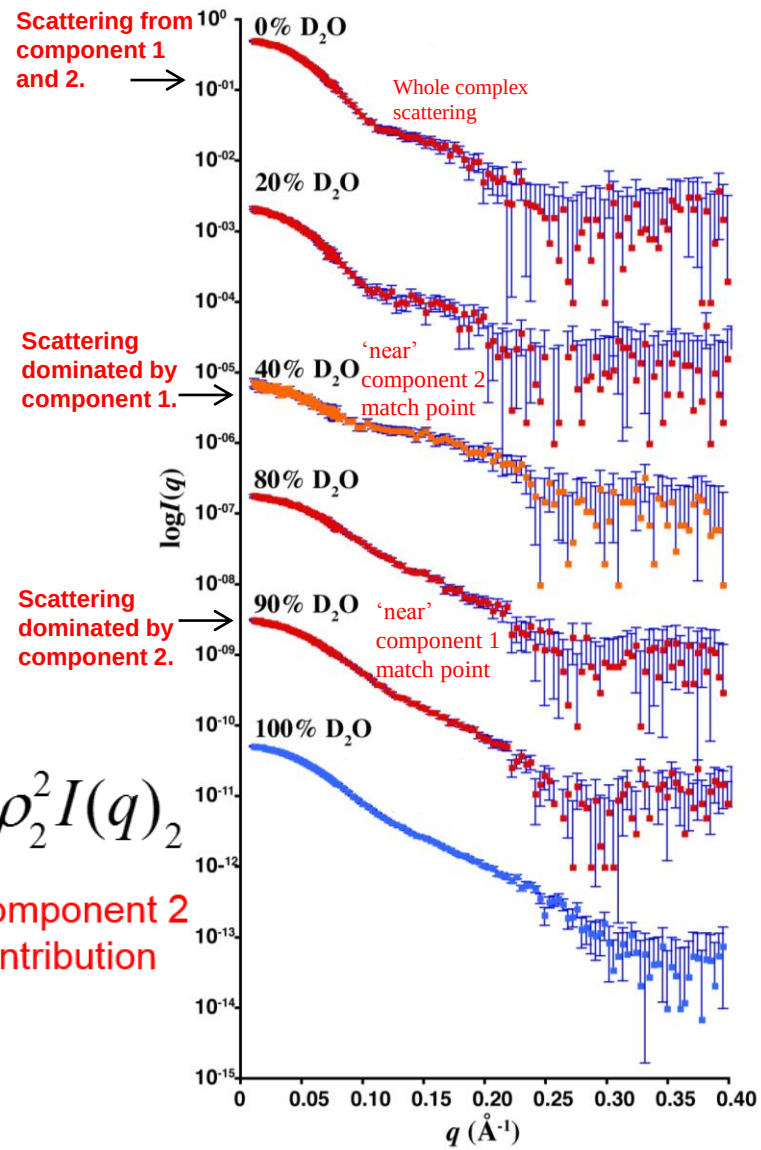
For a complex
(Two scattering length densities)

$$I(q_{\text{total}}) \propto \Delta\rho_1^2 I(q)_1 + \Delta\rho_1\Delta\rho_2 I(q)_{12} + \Delta\rho_2^2 I(q)_2$$

Component 1 contribution

'between' component contributions (cross-term)

Component 2 contribution

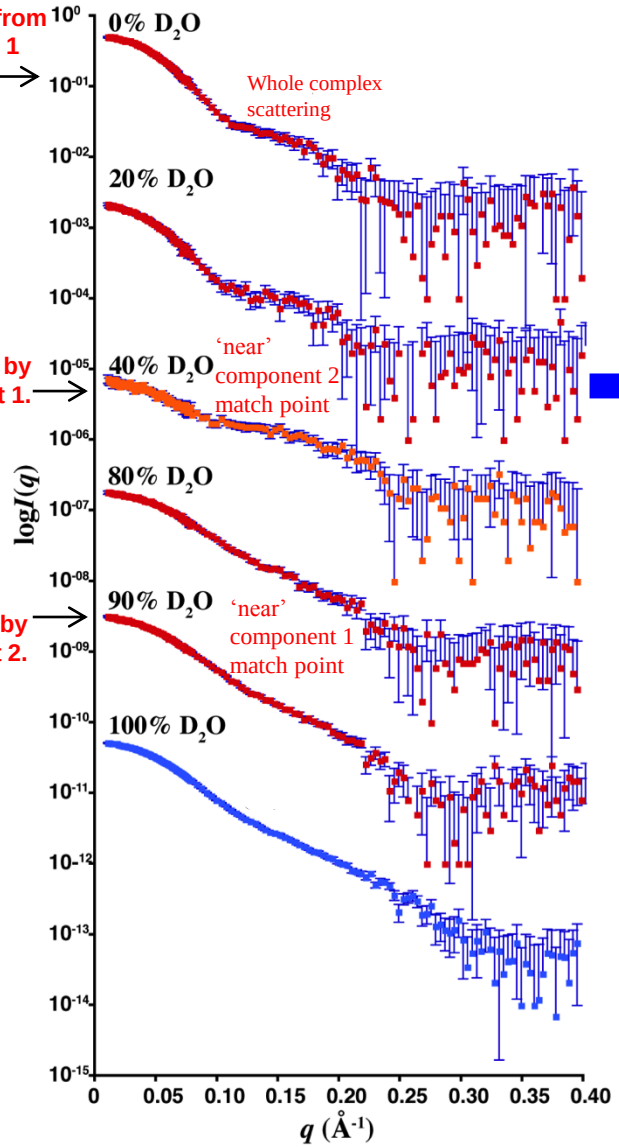


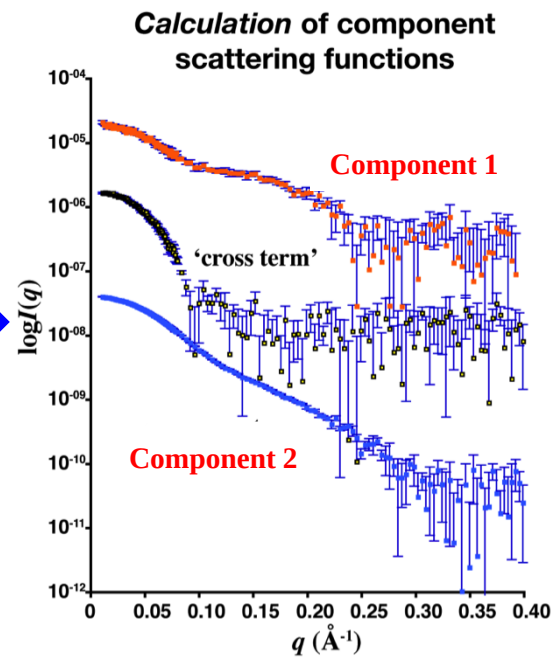
SANS with Contrast variation

Scattering from component 1 and 2.

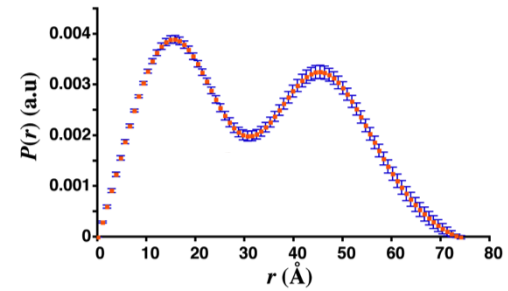
Scattering dominated by component 1.

Scattering dominated by component 2.

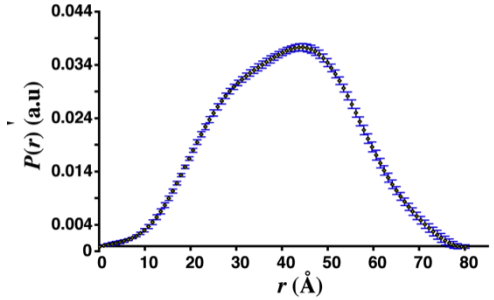




Component 1 P(*r*) vs *r*



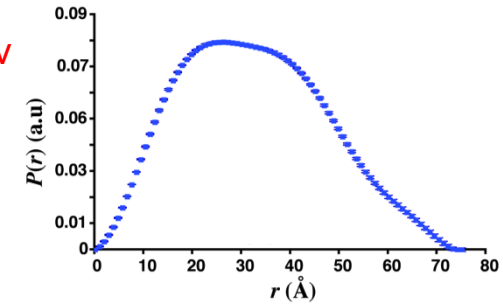
Cross term P(*r*) vs *r*



P(*r*) vs *r* can also be calculated for each contrast point, including the 'whole complex' scattering in 0% v/v ^2H_2O!

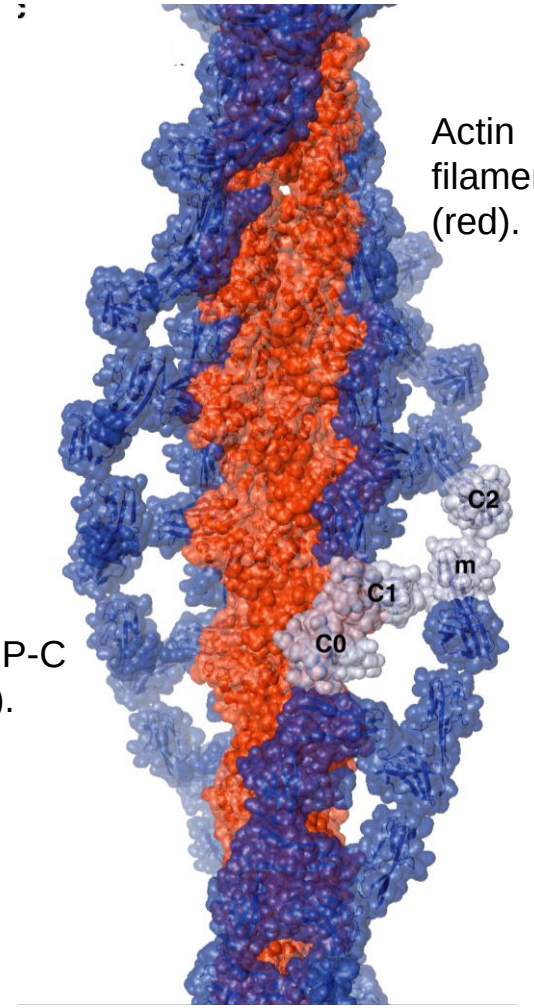
Add in SAXS data for modelling!!

Component 2 P(*r*) vs *r*



....some models

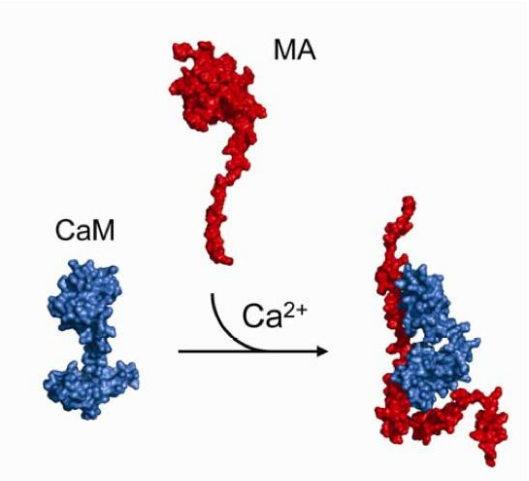
Macromolecular assembly: myosin binding protein C and F-actin



Actin filament (red).

cMyBP-C (blue).

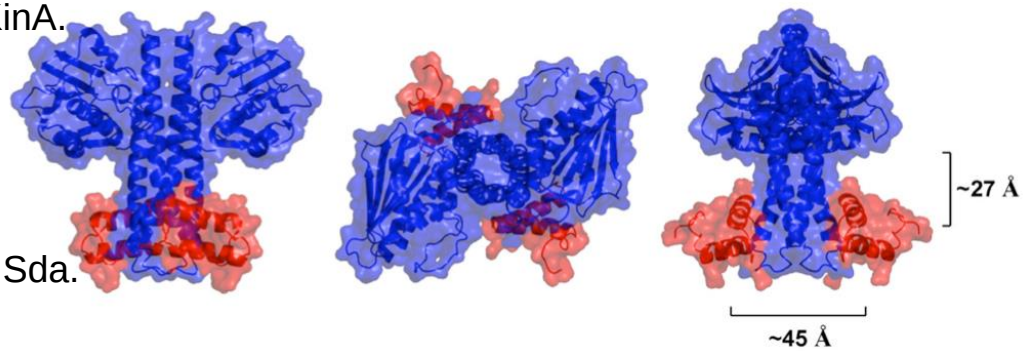
Unfolding of HIV1 matrix protein on binding calmodulin.



KinA/Sda bacterial sporulation control complex.

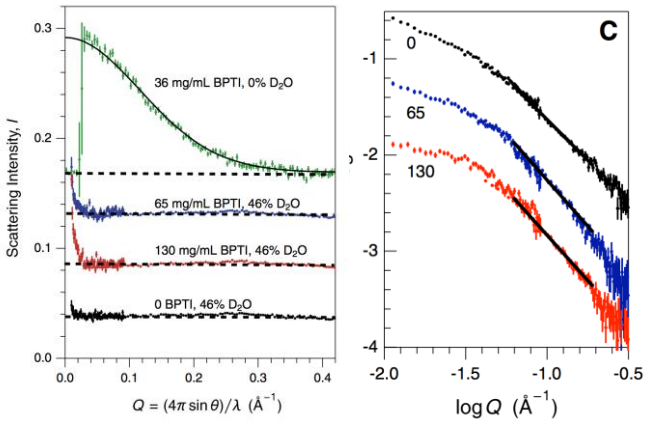
KinA.

Sda.

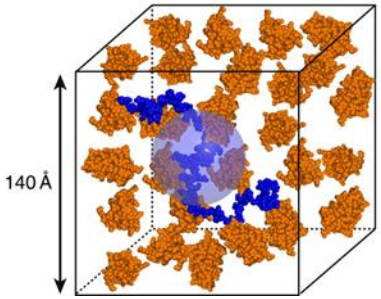


More exotic examples:

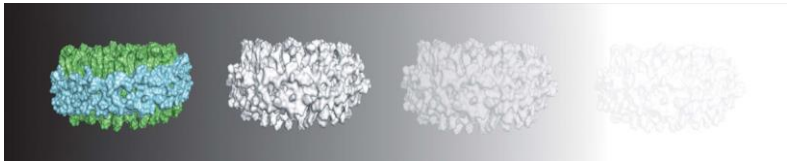
Macromolecular crowding of an intrinsically disordered deuterated protein in high-concentration non-deuterated protein solutions.

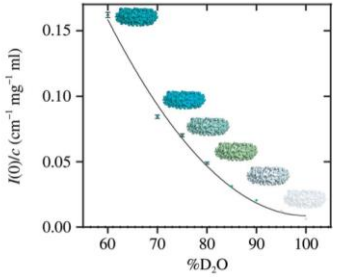


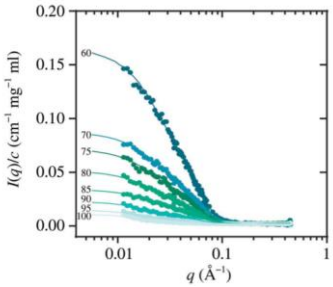
1) N-protein (IDP) experiences minimal compaction as a result of crowding by BBTI.
2) Less aggregated in more crowded environments

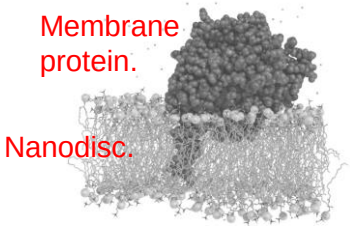


'Stealth nanodiscs'. Selma Maric; Lise Arleth





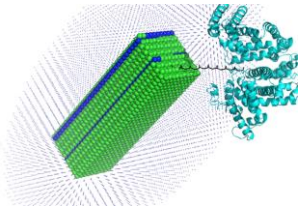




Ultimate goal is to use nanodiscs as a support for membrane proteins!

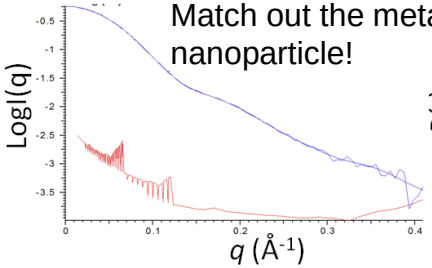
Acta Cryst. (2014). D70, 317–328.

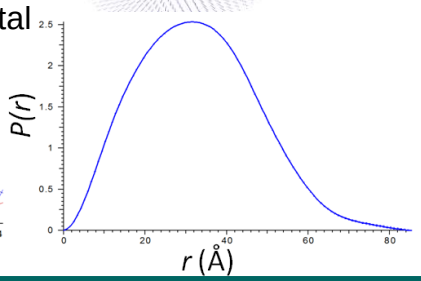
Nanomagnetite particles



Added to human serum albumin (HSA)

Match out the metal nanoparticle!





SASREF (for SAXS), SASREFcv (for SAXS and SANS)

Each subunit is treated as an individual rigid body. Protein, DNA, RNA, etc.

Assumes the atomistic models are ***COMPLETE i.e., no missing fragments or mass!***

Options to perform ***MIXTURE*** modelling (e.g., monomer-dimer; SASREFmx) or ***CONTRAST VARIATION*** (SAXS and SANS; SASREFcv).

Start from arbitrary initial orientations of the subunits – at the grid origin.

Simulated annealing is employed.

Search of interconnected spatial arrangement of the subunits without clashes.

Random movement/rotation at one SA step.

Fitting the scattering data by minimizing the target function.

Additional restraints may be applied.

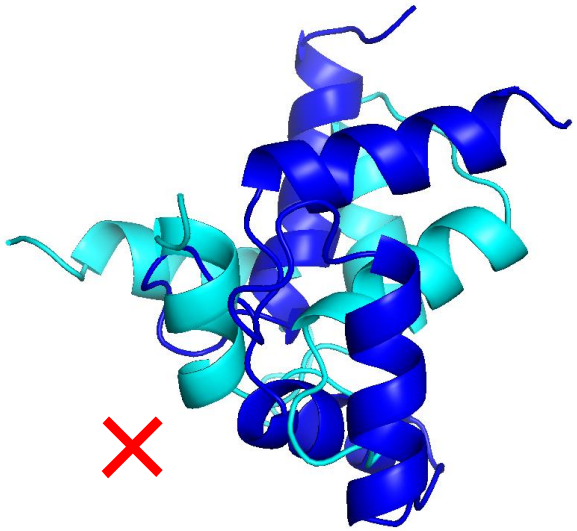
Petoukhov, M. V., and Svergun, D. I. (2006). ***Eur Biophys J.***, 35, 567-576

SASREFcv for SANS with contrast variation

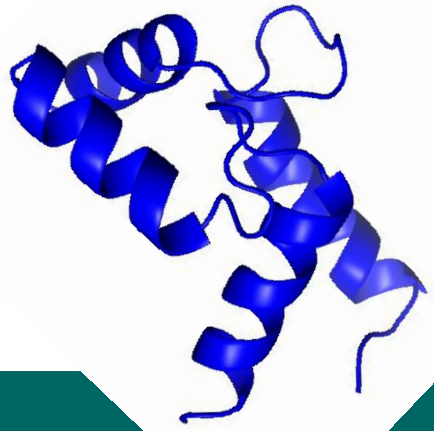
- Get a handle on the experimentally determined contrasts.
- Get a handle on the average non-exchangeable ^{1}H per unit volume of each component.
- Estimate the %-exchangeable ^{1}H (typically around 90-95%)

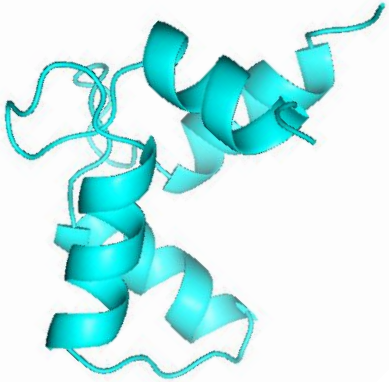
SASREF restraints

Subunit clashes or disconnected models are penalised!



Inter penetrating subunits are penalised.

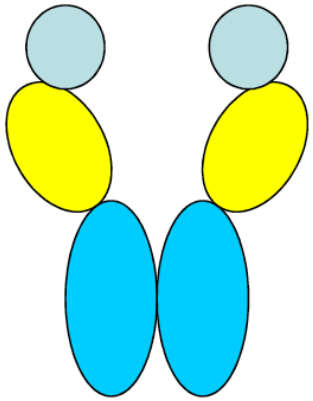




Disconnected models are penalised.

…a little bit of caution: you will bias the search-space with symmetry employed. Maybe do P1 first, then Px: compare and contrast…etc.

Symmetry constraint

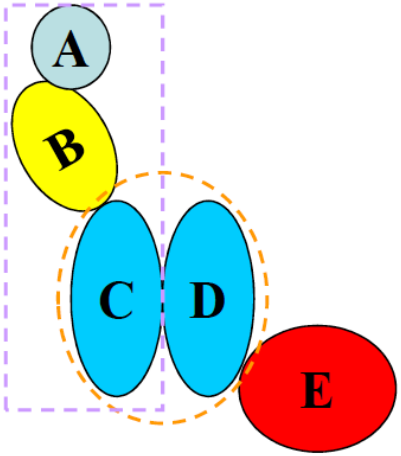


Groups Pn / Pn2 (n=1..6), P23, P432 and icosahedral symmetry can be taken into account.

- fewer spatial parameters to describe the model

Fixation of subset

Some subunits can be fixed at the initial positions and orientations to keep their mutual arrangement



SASREF and SASREFcv inputs

For SAXS:

SAXS data.
Rigid body starting models – centred to an origin.
Scattering amplitude files of each rigid-body model (partial scattering amplitudes of the subunits). Calculated using **CRYSOL**.
Contacts file (optional).

For SANS:

SANS data.
Rigid body starting models – centred to an origin.
Scattering amplitude files of each rigid-body model (partial scattering amplitudes of the complex). Calculated using **CRYSON**. Takes into component deuteration and % D_2O in the solvent.
Contacts file (optional).
The neutron contrasts.

Missing Stuff…

BUNCH

For SAXS only!

Single residue polypeptide chain only, i.e., ‘protein domains’!

With or without symmetry.

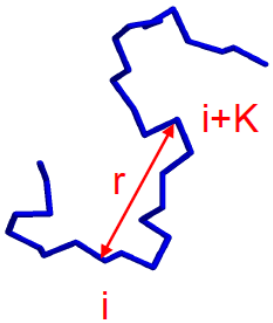
Models missing linkers and mass as a set of dummy residues.

A two step procedure.

pre_bunch
bunch

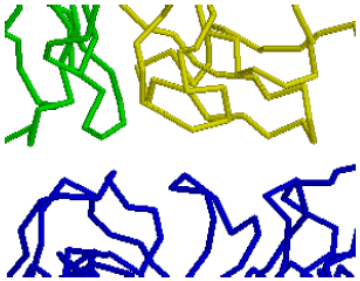
Requires the domain PDB files and the EXACT protein sequence (along with the SAXS data and scattering amplitudes calculated by CRYSOL.)

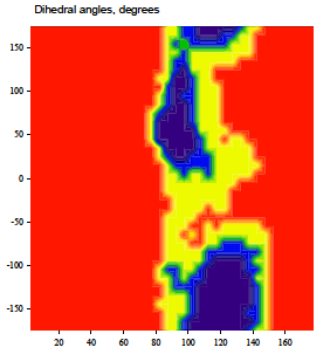
BUNCH, will optimize the position dummy amino acids during modelling.



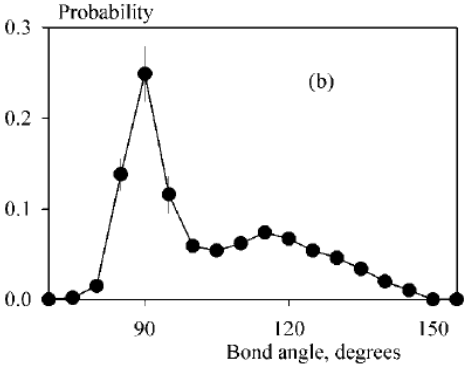
Neighbors distribution along the sequence

Absence of steric clashes





Bond angles & dihedrals distribution



Loop compactness may also be required $Rg_{id} = 3\sqrt[3]{n_l}$

BUNCH – in words, will…

- Search of the optimal positions and orientations of rigid domains and probable conformations of DR linkers, those fit the SAXS data.
- Proper bond and dihedral angles in the DR chains are required together with the absence of overlaps.
- The scattering pattern is calculated from partial amplitudes of domains and form-factors of DR comprising the loops using spherical harmonics.

$$I(s) = 2\pi^2 \sum_{l=0}^{\infty} \sum_{m=-l}^{l} \left| \sum_k A^{(k)}{}_{lm}(s) + \sum_i D^{(i)}{}_{lm}(s) \right|^2$$

- Multiple scattering curves fitting from deletion mutants

Petoukhov M.V., Svergun, D.I. (2005). *Biophys. J.* **89**, 1237-1250

CORAL

SASREF – is good for modelling whole/complete complexes against SAXS data.

BUNCH – is good form modelling single polypeptide chains with missing fragments against SAXS data

CORAL combines both concepts into one!

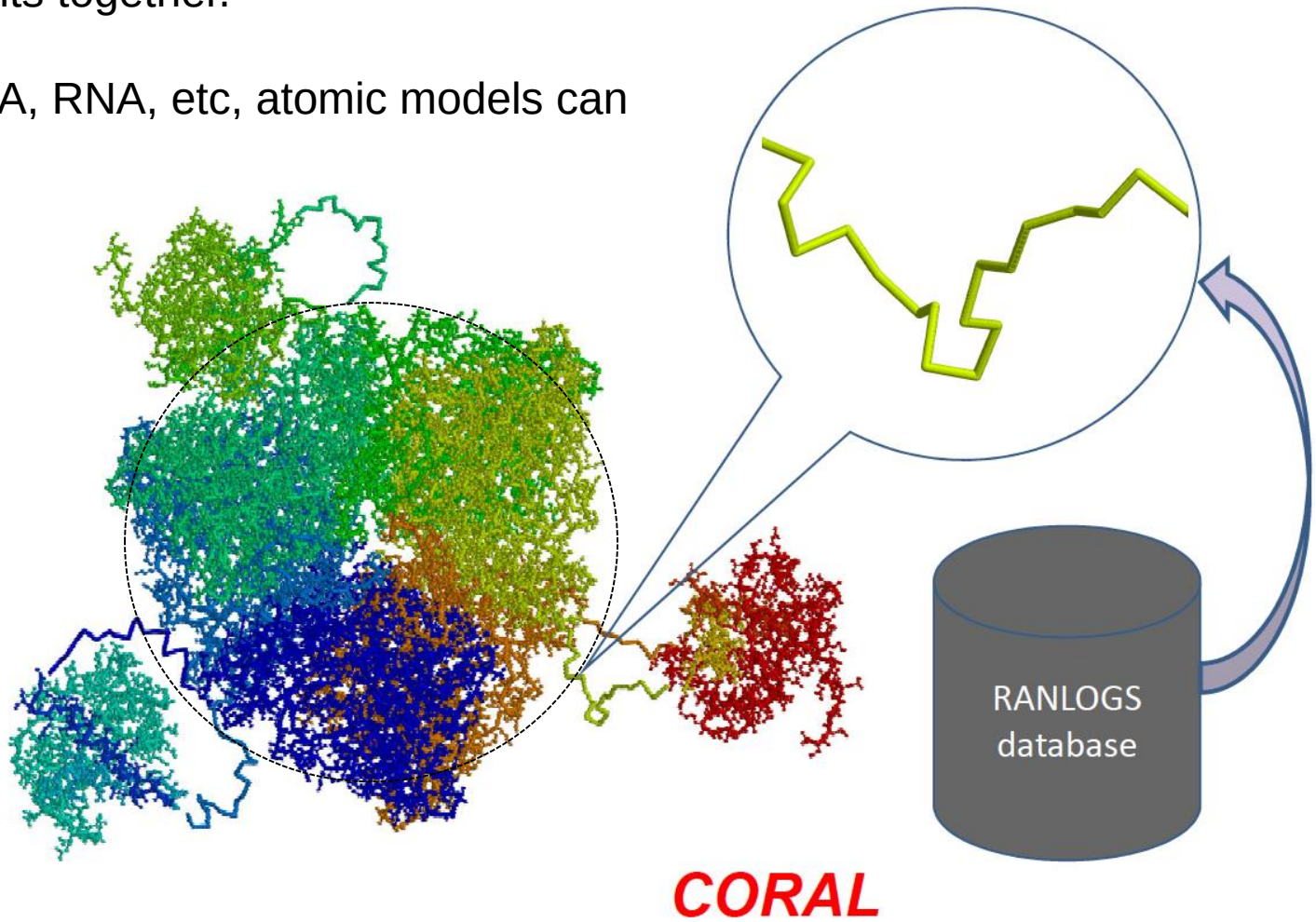
CORAL is also a great deal faster than BUNCH (CORAL can be used to model single polypeptide chains as well, and it is much faster!).

...***SAXS only!!***

CORAL has many options:

Known subunit interfaces can be preserved by grouping subunits together.

Protein and DNA, RNA, etc, atomic models can be used.



Always check the final model fits using Crysol, Cryson (for neutrons), or…for SAXS

Your favorite programs!

Approach	Modeling of the hydration layer	Representation of the molecule	References
CRYSOL	Implicit layer using an envelope function	All-atom	Svergun et al. *J. Appl. Cryst*. (1995)
AXES	Explicit water molecules using equilibrated water boxes	All-atom	Grishaev *et al*. *JACS* (2010)
FoXS	Implicit layer based on surface accessibility	All-atom or coarse-grained	Schneidman-Duhovny *et al*. *NAR* (2010)
HyPred	Explicit water molecules based on MD simulations	All-atom	Virtanen *et al*. *Biophys. J*. (2011)
AquaSAXS	Solvent-density map using the dipolar PB-Langevin approach	All-atom	Poitevin *et al*. *NAR* (2011)

WAXIS

ALWAYS REMEMBER AMBIGUITY

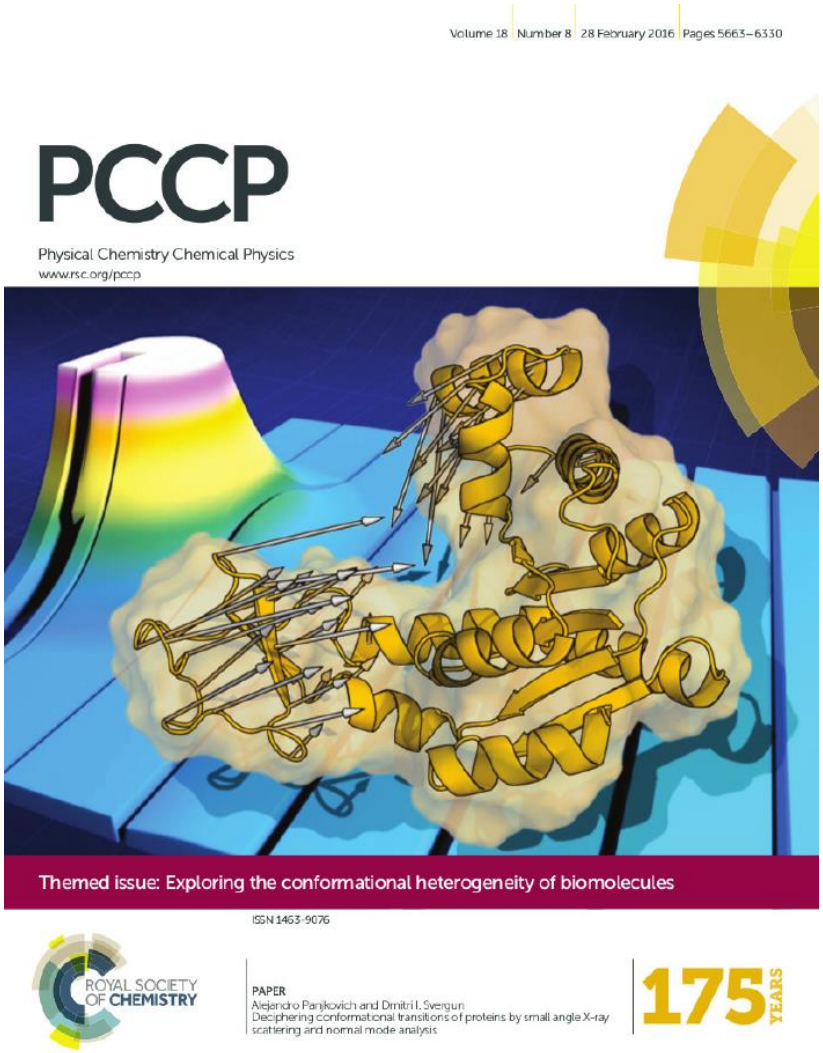
You ***must*** run your selected rigid body modelling routines at least 10 times and check for the spatial consistency of the models (spatial alignment using supcomb).

Double-check the fits with CRYSOL or CRYSON! Use Correlation Map to assess fits if you are unsure about your errors!

…also apply common sense.

But my structure almost fits…
Can I just wiggle it a bit?

SREFLEX



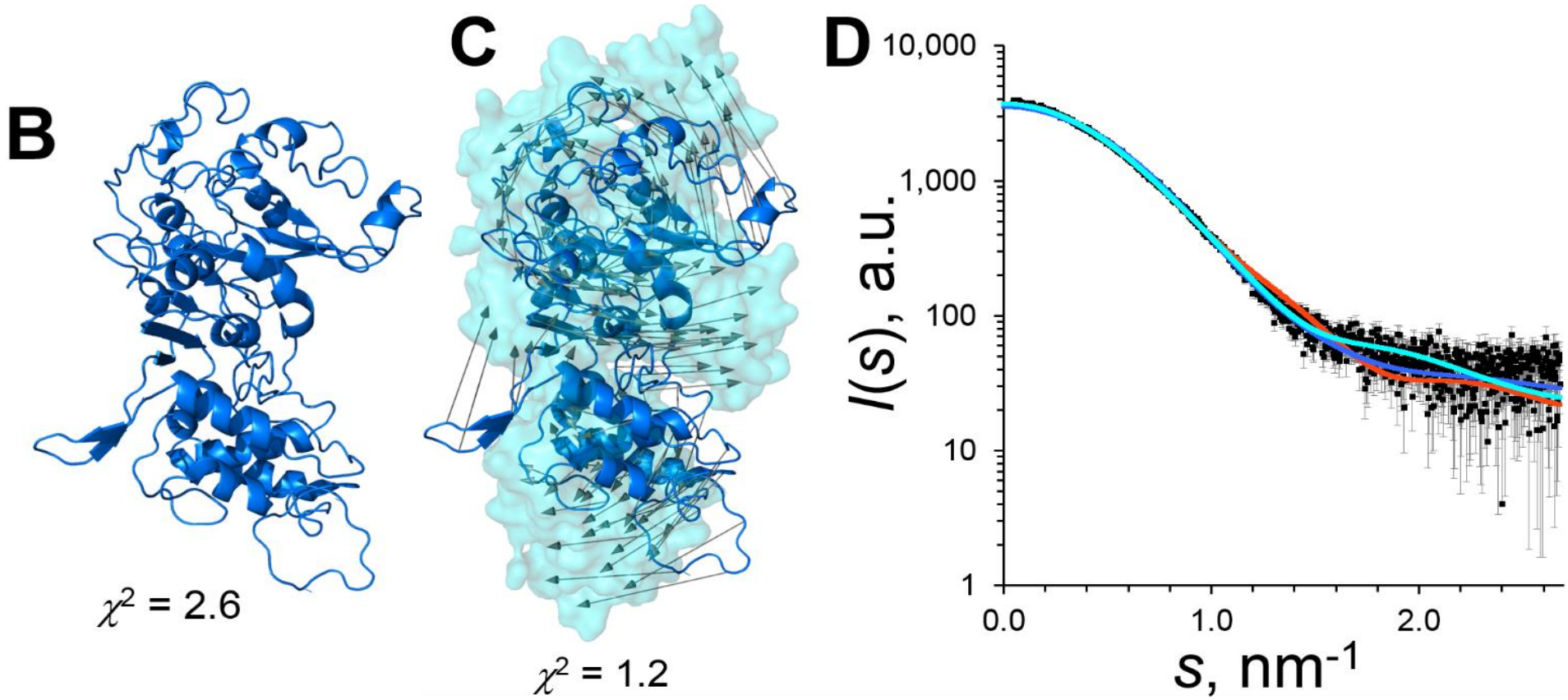
Deciphering conformational transitions of proteins by small angle X-ray scattering and normal mode analysis

A. Panjkovich, D.I. Svergun (2016) *Phys Chem Chem Phys.* 18, 5707-19

Used for refinement of models: small structural adjustments.

Great for assessing whether slight conformational movements are required to fit SAXS data (e.g., from crystal structures).

Limited to single unmodified polypeptide chains.

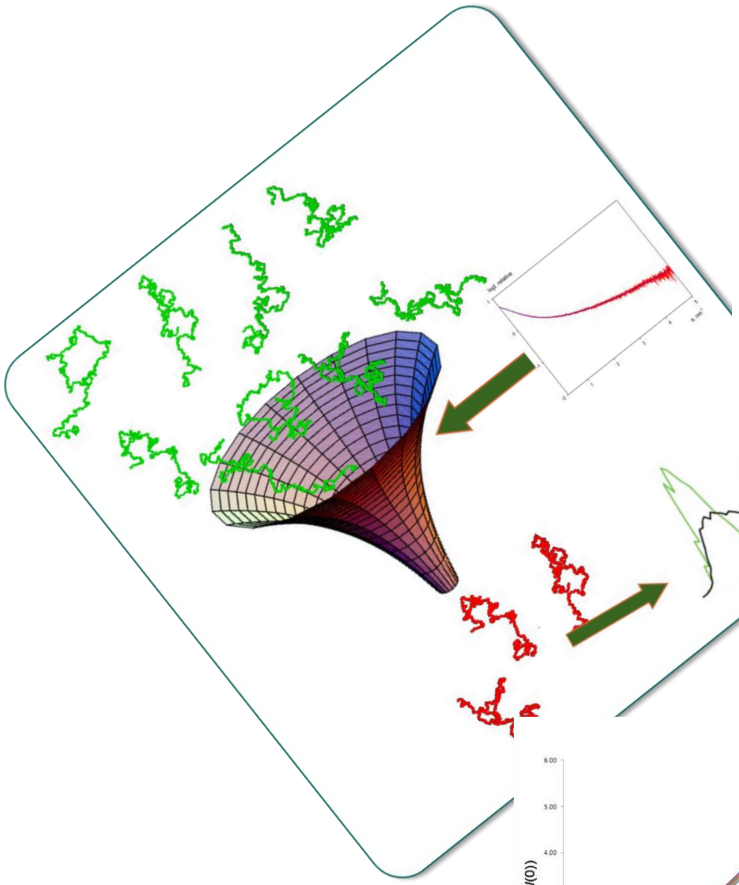


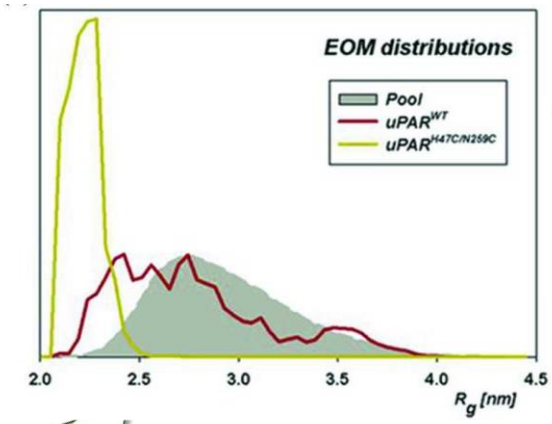
ATSAS online version applied additional CONCORD refinement

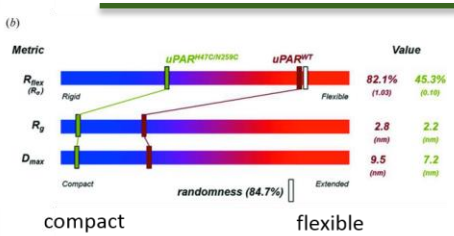
OMG, my structure is moving all over the shop!

Flexible proteins

…then wait for Pau Bernadó's lecture!

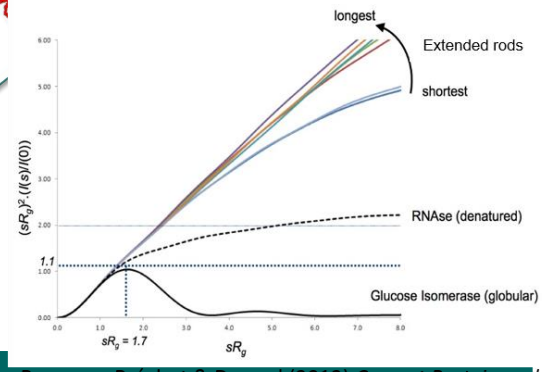


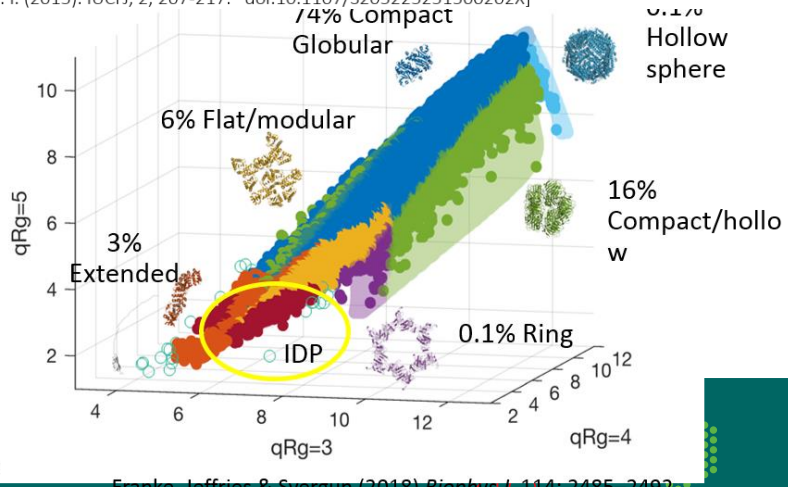




compact flexible

Characterization of the flexibility of uPARWT and the mutated uPARH47C-N259C using EOM 2.0. (a) Size distributions (Rg) of uPARWT and uPARH47C-N259C, providing only a qualitative assessment through direct comparison of the distributions of the selected ensembles and the pool. (b) The metrics Rflex and Rσ enable characterization of the flexibility quantitatively, with Rflex = ~82% and Rflex = ~45%, for uPARWT and uPARH47C-N259C, respectively, reflecting a significant change in compactness of the particle upon mutation (with a threshold of randomness of ~85% calculated from the pool). [Tria, G., Mertens, H. D. T., Kachala, M. and Svergun, D. I. (2015). IUCrJ, 2, 207-217. doi:10.1107/S205225251500202X]





Receveur-Bréchot & Durand (2012) *Current Protein and Peptide Science*, 13, 55-75.

Durand D, Vivès C, Cannella D, Pérez J, Pebay-Peyroula E, Vachette P, Fieschi F. (2010) *J Struct Biol.* 169: 45-53.

Franke, Jeffries & Svergun (2018) *Biophys J.* 114, 2485–2492.

ATSAS online

Web interface for **SASREF** --- Page 1 of 3

Main parameters of the project	
Total number of **curves**	6
Total number of **subunits**	3
Overall symmetry	P1

SUBMIT

Web interface for **SASREF** --- Page 2 of 3
no. of **subunits: 3** - no. of **curves: 6** - overall symmetry: **P1**

Curve	File		D2O fraction	Symmetry	Angular units $4\pi\sin(\theta)/\lambda$	Fraction to fit	Setting	Weight	Use a constant?
1	C:\project1\xcomp.dat	Browse...	-1.00	P1	Å-1	1.00	0	1.00	No
2	C:\project1\x-12.dat	Browse...	-1.00	P1	Å-1	1.00	0	1.00	No
3	C:\project1\x-23.dat	Browse...	-1.00	P1	Å-1	1.00	0	1.00	No
4	C:\project1\nc_0.dat	Browse...	0	P1	Å-1	1.00	1	1.00	Yes
5	C:\project1\nc_p50_0.d	Browse...	0	P1	Å-1	1.00	1	1.00	Yes
6	C:\project1\nc_p50_10(	Browse...	1.00	P1	Å-1	1.00	2	1.00	Yes

Subunit	File		Shift?	Fix?	Symmetry
1	C:\project1\sub1.pdb	Browse...	Yes	No	P1
2	C:\project1\sub2.pdb	Browse...	Yes	Yes	P1
3	C:\project1\sub3.pdb	Browse...	Yes	No	P1

Optional files:

C:\project1\smear.res Browse... Smearing parameters (*.res)
C:\project1\contacts.cn(Browse... Contacts conditions (*.cnd)

SUBMIT

Web interface for **SASREF** --- Page 3 of 3

Perdeuterations of the subunits in each construct
(specify "-1.0" if the subunit does not appear in the construct)

	sub1.pdb	sub2.pdb	sub3.pdb
xcomp.dat	0.00	0.00	0.00
x-12.dat	0.00	0.00	-1.0
x-23.dat	-1.0	0.00	0.00
nc_0.dat	0.00	0.00	0.00
nc_p50_0.dat	0.00	0.5	0.00
nc_p50_100.dat	0.00	0.5	0.00

SUBMIT

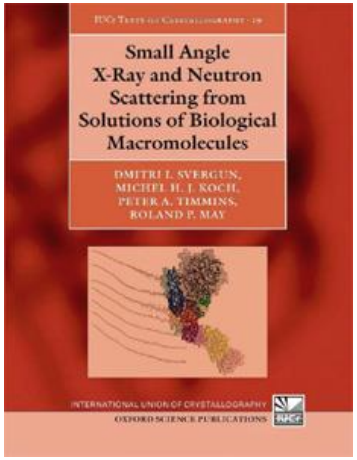
http://www.embl-hamburg.de/biosaxs/atsas-online

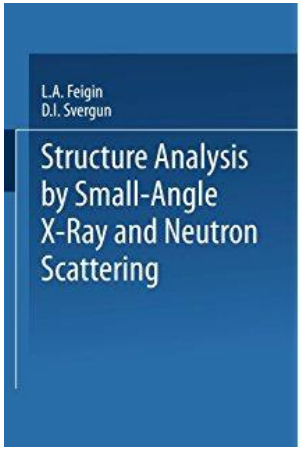
https://www.embl-hamburg.de/biosaxs/publications.html

- Franke, D., Petoukhov, M.V., Konarev, P.V., Panjkovich, A., Tuukkanen, A., Mertens, H.D.T., Kikhney, A.G., Hajizadeh, N.R., Franklin, J.M., Jeffries, C.M. and Svergun, D.I. (2017) ATSAS 2.8: a comprehensive data analysis suite for small-angle scattering from macromolecular solutions *J. Appl. Cryst.* 50 © IUCr DOI

Latest ATSAS paper

- Panjkovich, A. and Svergun, D.I. (2016) Deciphering conformational transitions of proteins by small angle X-ray scattering and normal mode analysis *Phys. Chem. Chem. Phys.* 18, 5707-5719 DOI
- Kikhney, A.G., Panjkovich, A., Sokolova, A.V. and Svergun, D.I (2016) DARA: a web server for rapid search of structural neighbours using solution small angle X-ray scattering data *Bioinformatics* 32(4), 616-618 DOI
- Kikhney, A.G. and Svergun, D.I (2015) A practical guide to small angle X-ray scattering (SAXS) of flexible and intrinsically disordered proteins *FEBS Lett.* 589(19A), 2570-2577 DOI
- Tria, G., Mertens, H. D. T., Kachala, M. and Svergun, D. I. (2015) Advanced ensemble modelling of flexible macromolecules using X-ray solution scattering *IUCrJ* 2, 207-217 DOI
- Konarev, P.V. and Svergun, D.I (2015) *A posteriori* determination of the useful data range for small-angle scattering experiments on dilute monodisperse systems *IUCrJ* 2, 352-360 © IUCr DOI
- Petoukhov, M.V. and Svergun, D.I. (2015) Ambiguity assessment of small-angle scattering curves from monodisperse systems *Acta Cryst.* D71, 1051-1058 © IUCr DOI
- Svergun D.I., Barberato C., and Koch M.H.J. (1995) CRYSOL - a program to evalate X-ray solution scattering of biological macromolecules from atomic coordinates *J. Appl. Cryst.* 28, 768-773
- Petoukhov, M.V. and Svergun, D.I. (2005) Global rigid body modelling of macromolecular complexes against small-angle scattering data *Biophys. J.* 89, 1237-1250 © Biophysical Journal





Other approaches/programs I

- J. Bardhan, S. Park and L. Makowski (2009) SoftWAXS: a computational tool for modeling wide-angle X-ray solution scattering from biomolecules *J. Appl. Cryst.* **42**, 932-943 *A program to compute WAXS*
- Schneidman-Duhovny D, Hammel M, Sali A. (2010) FoXS: a web server for rapid computation and fitting of SAXS profiles. Nucleic Acids Res. **38** Suppl:W540-4. *Debye-like computations, Fox Web server*
- Grishaev A, Guo L, Irving T, Bax A. (2010) Improved Fitting of Solution X-ray Scattering Data to Macromolecular Structures and Structural Ensembles by Explicit Water Modeling. *J Am Chem Soc.* **132,** 15484-6. *Generate bulk and bound waters with MD, do fit the data to the model*
- Poitevin F, Orland H, Doniach S, Koehl P, Delarue M (2011). AquaSAXS: a web server for computation and fitting of SAXS profiles with non-uniformally hydrated atomic models. Nucleic Acids. Res. 39, W184-W189 *Generate waters around proteins using MD (AquaSol program)*
- Virtanen JJ, Makowski L, Sosnick TR, Freed KF. (2011) Modeling the hydration layer around proteins: applications to small- and wide-angle x-ray scattering. *Biophys J.* **101**, 2061-9. *Use a "HyPred solvation" model to generate the shell, geared towards WAXS.*
- Chen P, Hub JS (2014) Validating solution ensembles from molecular dynamics simulations by wide-angle X-ray scattering data. Biophys. J., 107, 435-447. *Use MD simulations to generate excluded/bound waters, WAXSIS Web server.*

Acknowledgements

- Maxim Petoukhov
- Petr Konarev
- Daniel Franke
- Alejandro Panjkovich
- Nelly Hajizadeh
- Dmitri Svergun
- And of course, all members of the Svergun Group

Helmholtz-Zentrum Geesthacht
Zentrum für Material- und Küstenforschung

BioStructX

BioSAXS

DESY

Bundesministerium für Bildung und Forschung

BIOSCAT

The research has received funding from the European Community's Seventh Framework Programme (FP7/2007-2013) under BioStruct-X (grant N° 283570) and BMBF research grant BIOSCAT (Contract no: 05K12YE1).

EMBL 40 YEARS | 1974–2014