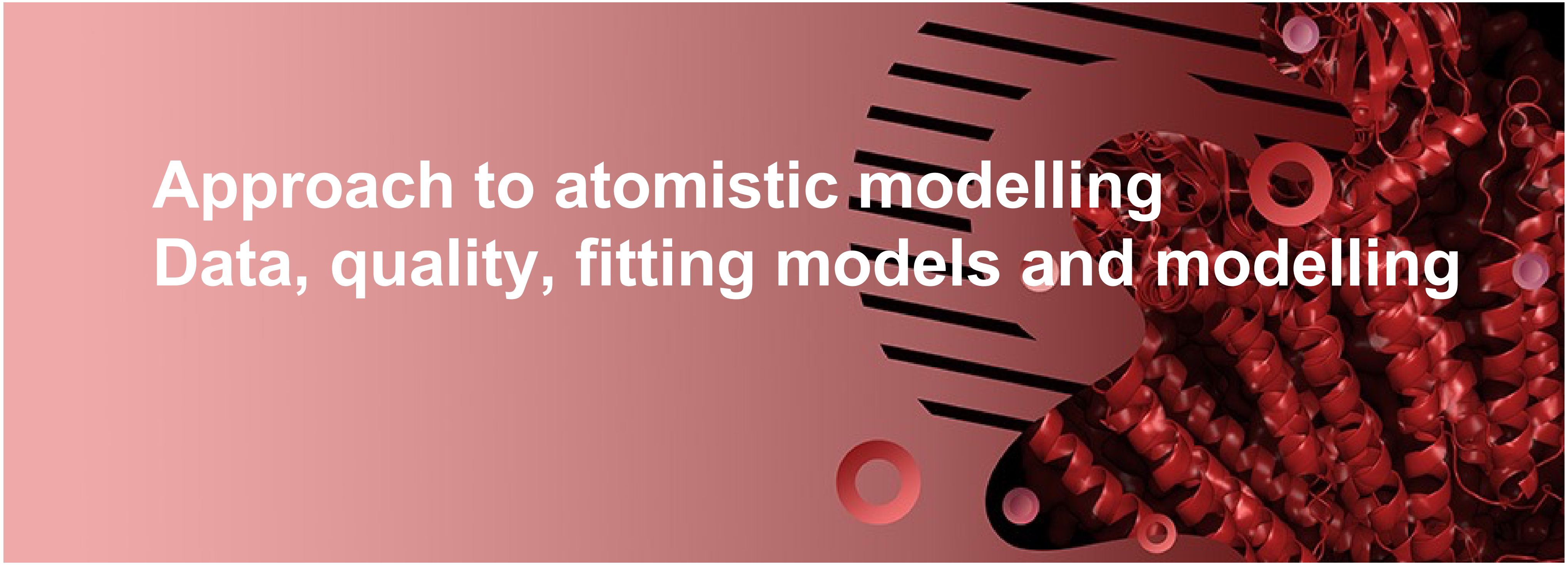
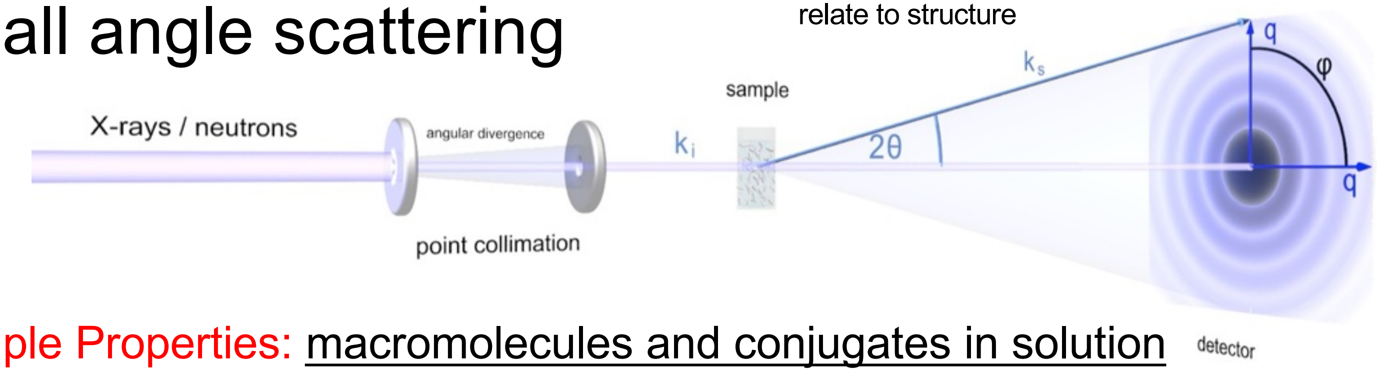


Approach to atomistic modelling
Data, quality, fitting models and modelling

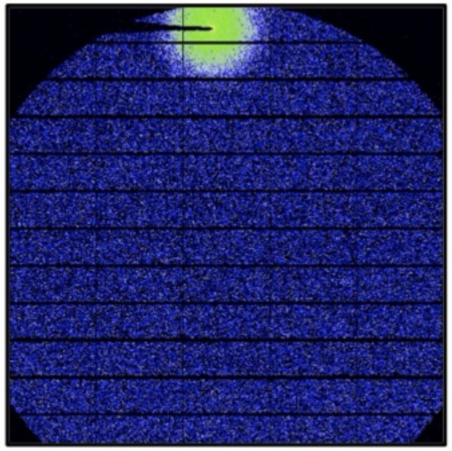
Cy Jeffries
EMBL Hamburg

Small angle scattering

Scattering intensities, *I*, as a function of angle (*s*, or *q*) – relate to structure



2D detector image



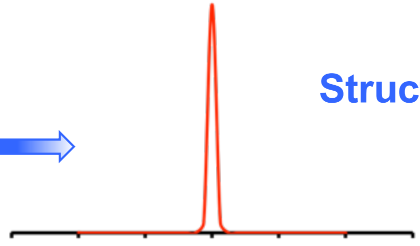
Sample Properties: macromolecules and conjugates in solution

always illuminate a population

Sample

Monodisperse

$$I(s)=\sum_k v_k I_k(s)$$



Narrow population state

Structurally homogeneous

Single-state model

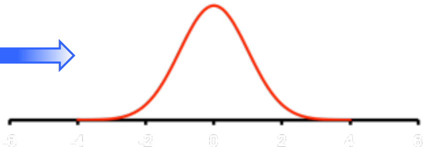
Polydisperse

Different conformations

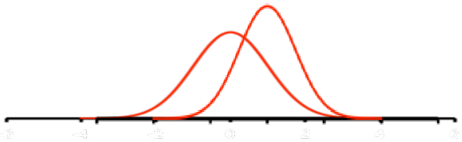
Different oligomeric states

Structurally heterogeneous

Multi-state models

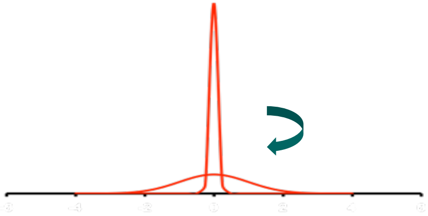


population state(s) distribution



multiple population states:mixtures

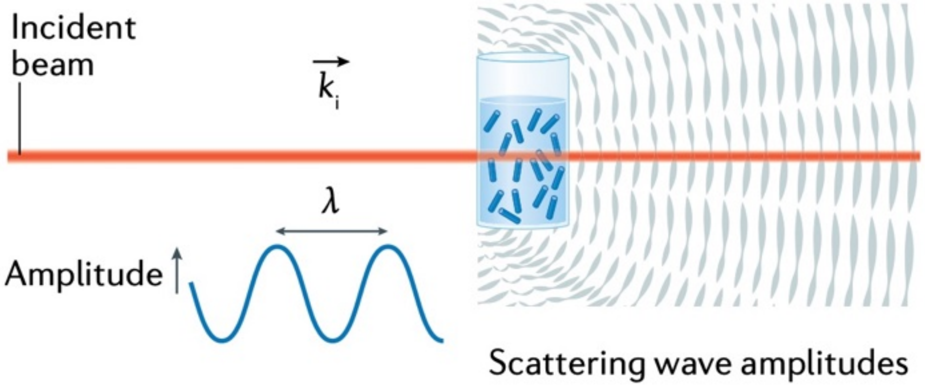
Temperature
Time
Chemical environment



population state transition

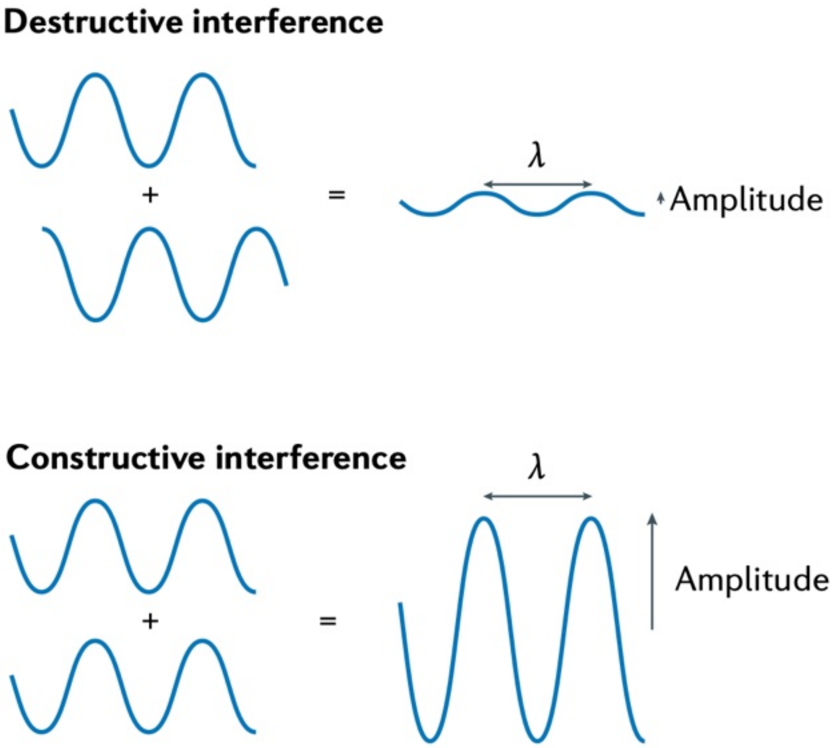
Small angle scattering

Wave properties: Scattering amplitudes

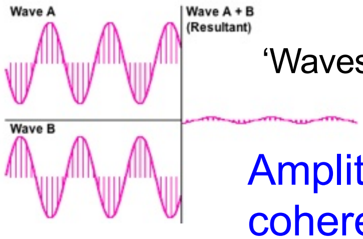


When the incident beam is **scattered elastically** (no change in λ) and if preserved distance correlations exist between the scattering centres, a **coherent wavefront** develops that emanates from the sample where both constructive and destructive interference occurs in the wave amplitudes.

The magnitude of the **coherent scattering amplitudes** as a function of angle relates to spatial correlations between scattering centres.



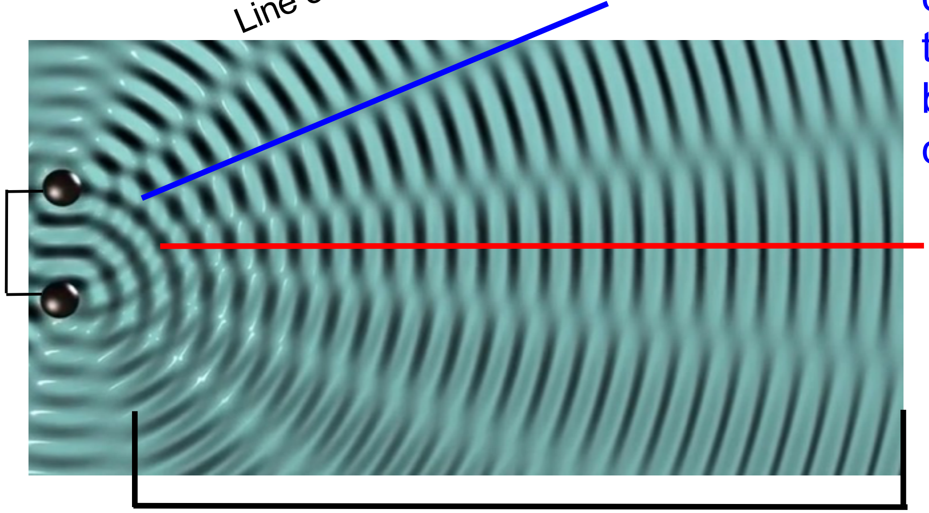
Elastic coherent small-angle scattering = ‘Sum the waves game.’



‘Waves cancel’ (amplitudes cancel)

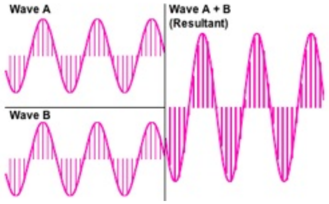
Line of destructive interference

Amplitude pattern across the coherent wave front relates to the correlated distance between the two scattering centres



Preserved distance correlation between two scattering centres.

Line of constructive interference



‘Waves add’ (amplitudes 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 as a function of angle.

More formally:

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.

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

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

Q = q = h = μ = k = s

Sometimes; S = 2sinθ/λ = 2πs

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

Spherical wave bit

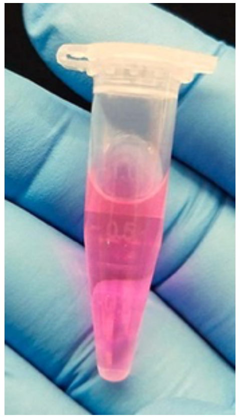
‘Scattering factor’: relates to the atomic cross section, i.e., scattering length, or probability of an atom to scatter for every atom in the sample.

The issue?

We cannot access the amplitudes experimentally. We measure the *intensity* of the scattered radiation, *I*(*s*).

For solution-based SAXS, the sample particles are tumbling in solution!

Sample



$$I(s) = \left\langle A(s)A(s)^{*} \right\rangle \xrightarrow{\hspace{2cm}}$$

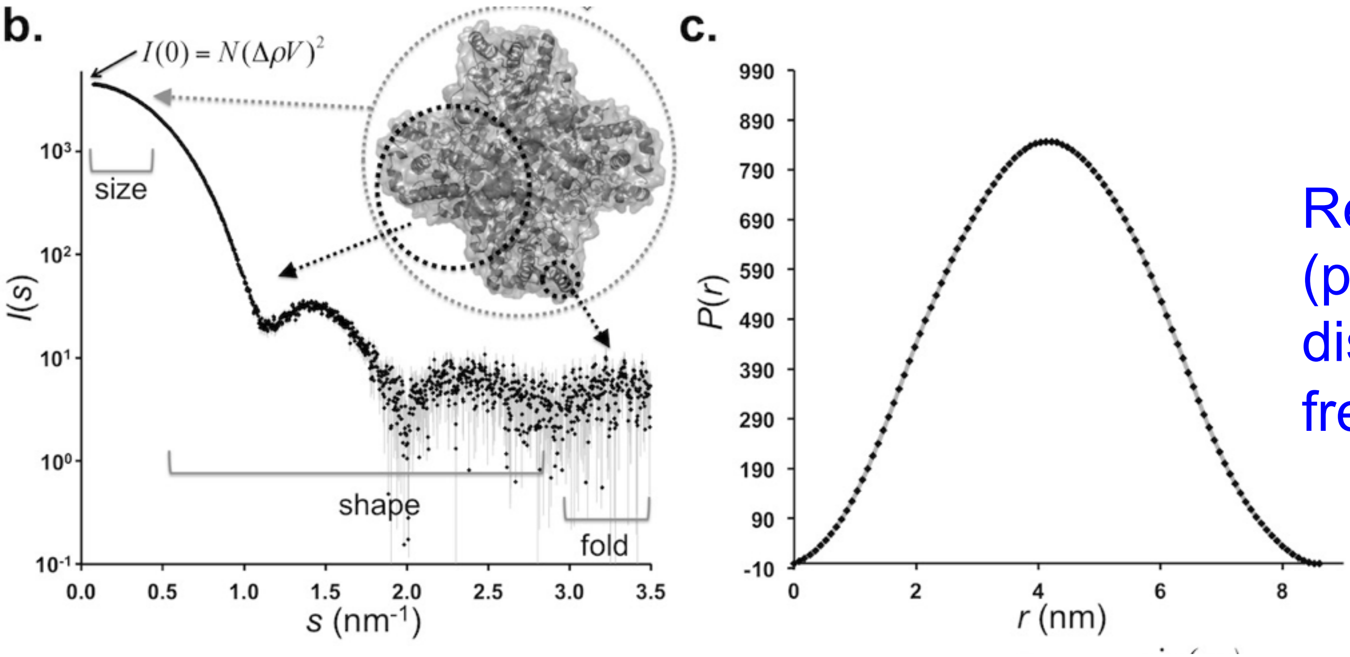
All orientations considered, i.e., isotropic scattering

I(*s*) fundamentally boils down to the form factor of the particles, *P*(*s*), their volume and scattering power!

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:

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

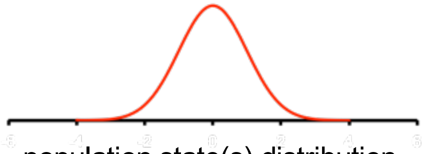
Reciprocal space intensity



Real space (probable) distance frequencies

How does the sample properties combined with the measurement impact our approach to modelling data?

Populations of macromolecules in the sample

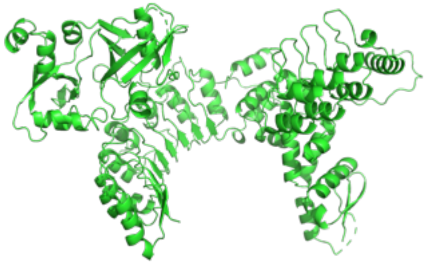


population state(s) distribution

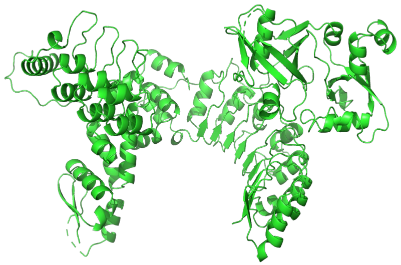
Sum of spherically averaged scattering amplitudes

$$I(s) = \left\langle A(s)A(s)^{*} \right\rangle$$

Distance distributions are encoded in the scattering intensities – not x,y,z atomic coordinates



L-amino acids



D-amino acids

The Complication

For biological macromolecules in solution, ***we forgot the solution!***

It is obvious that macromolecules of a sample will scatter. The amplitudes arising from preserved distance correlations will ***sum*** to produce coherent scattering intensities at low angle.

The lower the angle (lower *s*), the *longer* the correlated distances, *d*:

$$s = 2\pi/d$$

However, the solution, i.e., the solvent of the sample, also scatters! As the solvent (hopefully) does not have any time-preserved long-range distance correlations, its scattering contributions add as a ‘flat incoherent background’ in the SAS regime.

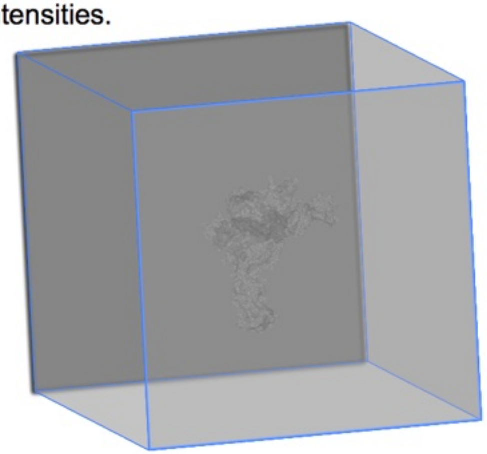
But you need contrast as well: ‘excess scattering power’

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

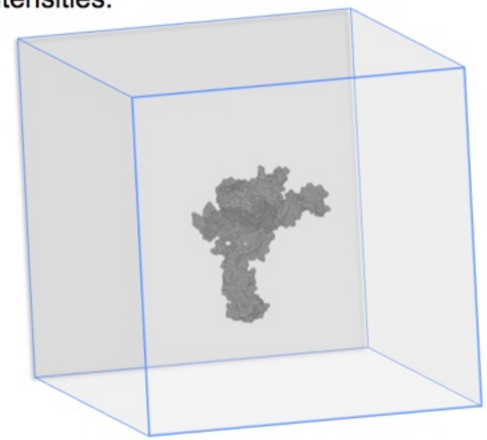
$$\Delta\rho = \overline{\rho} - \rho_{\mathrm{s}},$$

where $\overline{\rho}$ and ρ_{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.



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

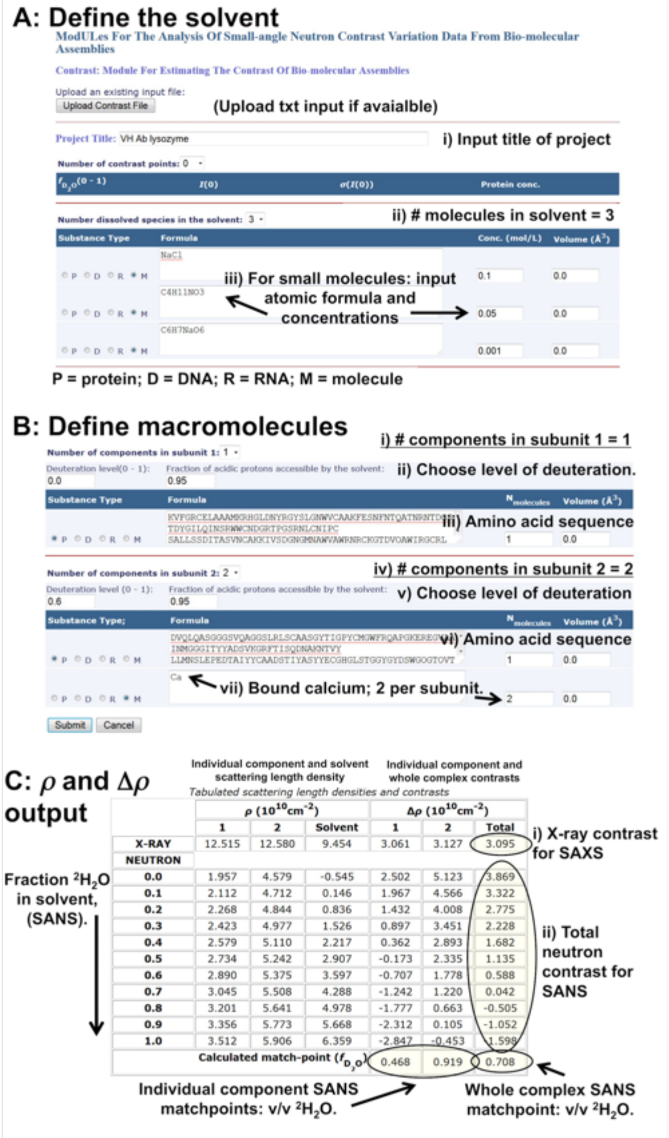
How do I calculate the Contrast

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

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

For X-rays: Convert the SLD, ρ (10^{10} cm^{-2}) to electron density by dividing by the Thomson electron radius: 2.8179 x 10^{-13} cm. The answer is in e/cm^3, so divide again by 10^{24} to get e/Å^3…or more quickly:

$$\frac{\rho}{28.179} \quad \text{e/Å}^3$$



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 txt input if avaialble)

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

Number of contrast points: 0

f_{D_2O}(0 - 1)	$I(0)$	$\sigma(I(0))$	Protein conc.

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

Substance Type	Formula	Conc. (mol/L)	Volume (Å^3)
○ 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

i) # components in subunit 1 = 1

Number of 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$_{\text{molecules}}$	Volume (Å^3)
● P ○ D ○ R ○ M	KVFGRCELAAAMKRHGLDNYRGYSLGNWVCAAKFESNFNTQATNRNTDG TDYGILQINSRWWCNDGRTPGSRNLCNIPC SALLSSDITASVNCAKKIVSDGNGMNAWVAWRNRCKGTDVQAWIRGCRL	1	0.0

iii) Amino acid sequence

iv) # components in subunit 2 = 2

Number of 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$_{\text{molecules}}$	Volume (Å^3)
● P ○ D ○ R ○ M	DVQLQASGGGSVQAGGSLRLSCAASGYTIGPYCMGWFRQAPGKEREG INMGGGITYYADSVKGRFTISQDNAKNTVY LLMNSLEPEDTAIYYCAADSTIYASYYECGHGLSTGGYGYDSWGQGTQVT	1	0.0
○ P ○ D ○ R ● M	Ca	2	0.0

vi) Amino acid sequence

vii) Bound calcium; 2 per subunit.

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.

After background subtraction…

I(*s*) will represent the time and rotationally averaged squared scattering amplitudes from the particle population expressed as the summed contribution from each individual particle, *i*, in the sample.

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 scattering has been subtracted) will represent the time and rotationally averaged squared scattering amplitudes, i.e., the scattering intensity, from a ***SINGLE PARTICLE***.

How do I ***maybe*** know I have an ideal system?

The molecular mass estimates through a concentration series.

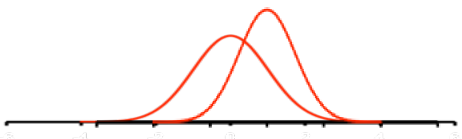
The MM, the MM, the MM, the MM, the MM, the MM.

(+/-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).

$$I(s)=\sum_i^n[(\Delta\rho_i V_i)^2 P_i(s)]$$

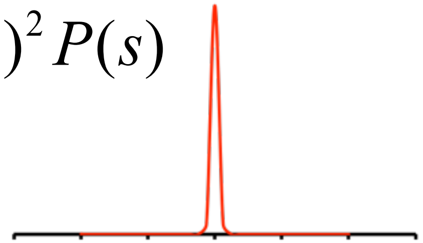
If *i* are not identical, model as a mixture



multiple population states: mixtures

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

If *i* are all identical, model as a single particle

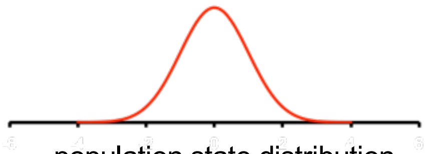


Narrow population state

The case of intrinsically disordered proteins or modular proteins connected by flexible linkers.

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

Model as a structural ensemble!



population state distribution

MM is correct!

BUT

Still not ideal, i.e., cannot be modelled as a non-interacting single particle because the protein is structurally heterogeneous!

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***.

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

$$\mathrm{MW}_{sample} = \frac{I(0)_{sample}\mathrm{N}_{\mathrm{A}}}{c_{sample}(\Delta\rho\upsilon_{\mathrm{sample}})^2}$$

Absolute scaling - requires partial specific volume and contrast.

$$\mathrm{MW}_{sample} = \frac{I(0)_{sample}}{I(0)_{standard}} \times \frac{c_{standard}\Delta\rho^2_{standard}}{c_{sample}\Delta\rho^2_{sample}} \times MW_{standard}$$

An assumption that a target has a similar scattering length density and partial specific volume as the secondary standard!

Porod volumes and Kratky plot

The determination of MW from *I*(0) requires an accurate assessment of the concentration of a protein in solution that in and of itself can be difficult to determine!

An alternative concentration-independent estimate of MW is based on the volume of a protein in solution. Porod showed that for uniform particles with sharp boundaries the excluded volume *Vp* can be calculated as:

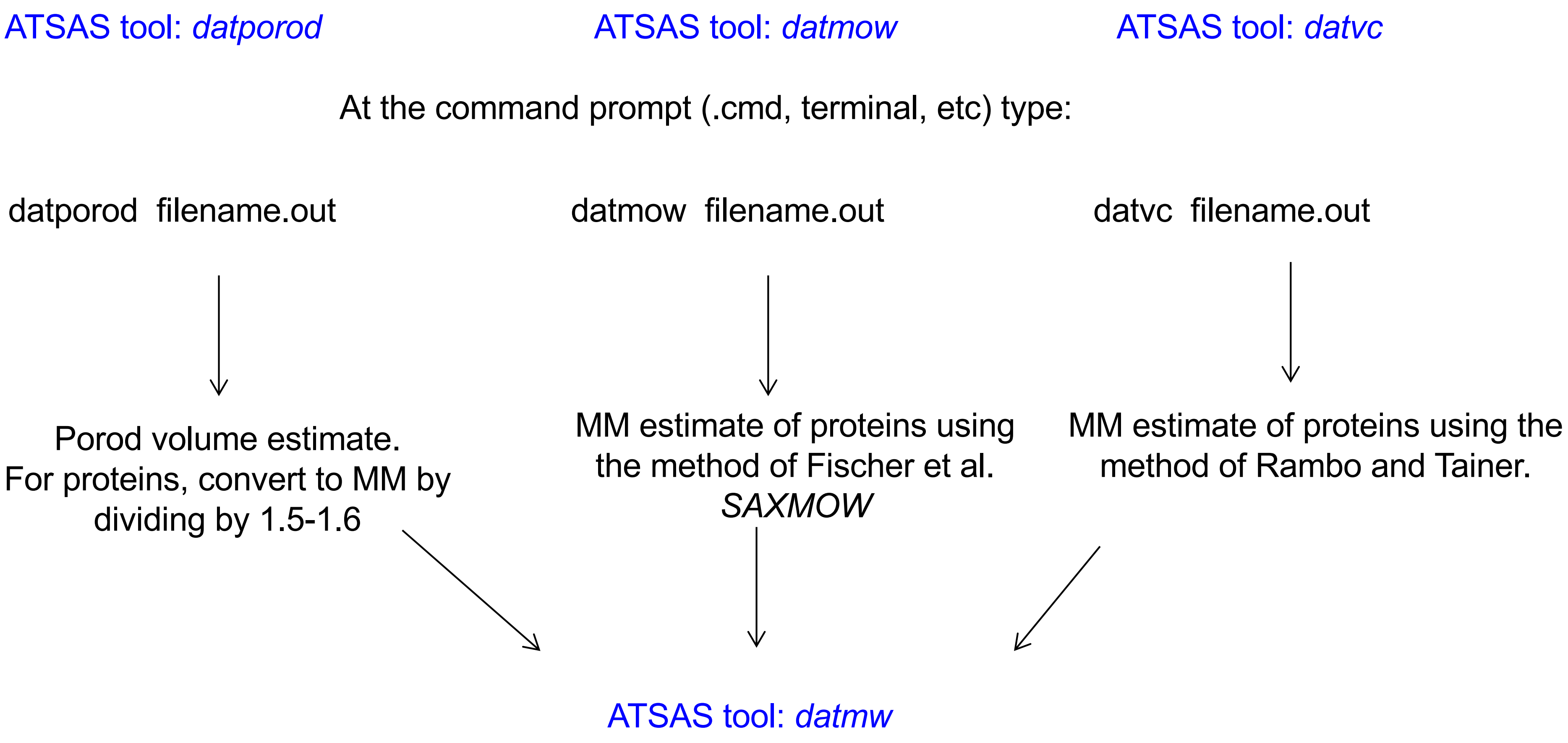
$$V_p = \frac{2\pi^2 I(0)}{Q}$$

where *Q* is the Porod invariant or the area under a plot of *I*(s)s^2 vs *s* calculated to s = ∞, or Kratky plot.

The *Vp* of a protein in nm^3 is typically 1.5–1.6 times the MW in kilodaltons (kDa).

However, caution must be applied when dealing with highly anisotropic or highly flexible/disordered proteins. In the case of flexible, or rod-like proteins, the decay in scattering intensities at high angle deviates sufficiently from Porod’s law that the estimation of *Q* will incur errors in the volume estimation!

Useful ATSAS tools



ATSAS tool: *datporod*

ATSAS tool: *datmow*

ATSAS tool: *datvc*

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

```
datporod  filename.out
```

```
datmow  filename.out
```

```
datvc  filename.out
```

Porod volume estimate.
For proteins, convert to MM by dividing by 1.5-1.6

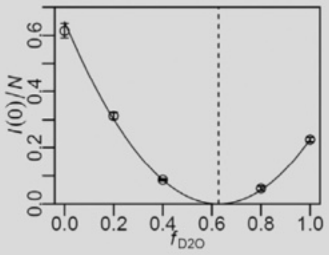
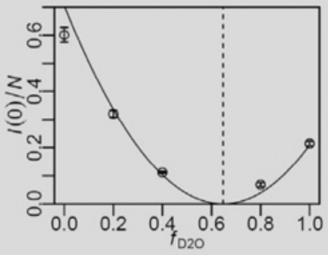
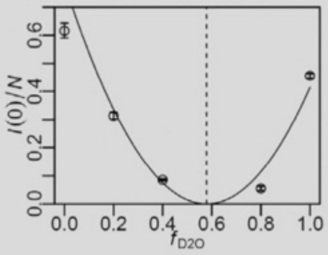
MM estimate of proteins using the method of Fischer et al. *SAXMOW*

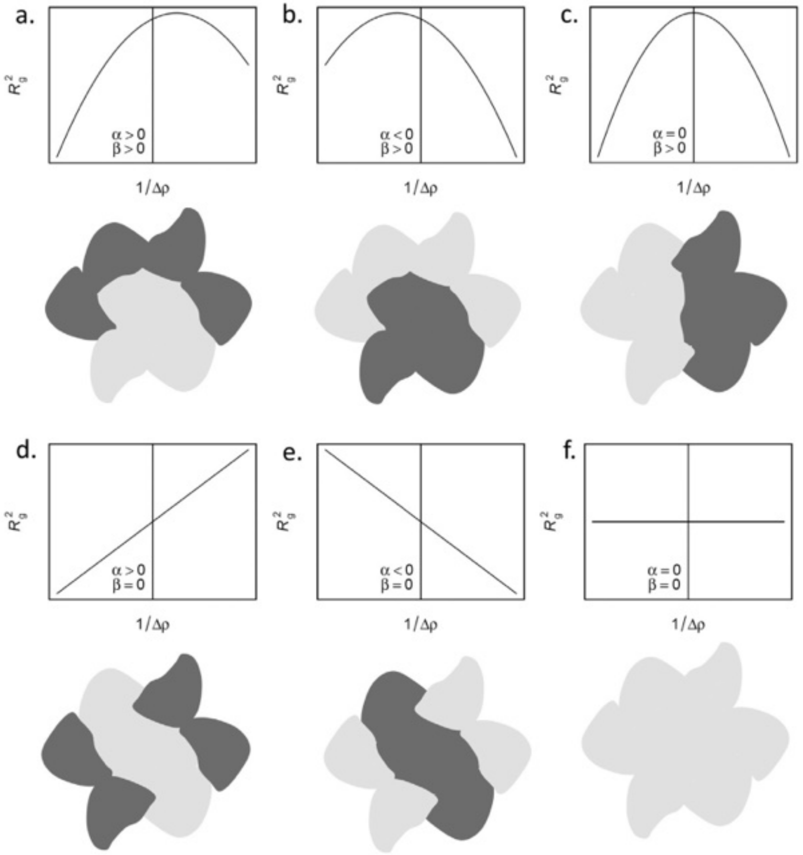
MM estimate of proteins using the method of Rambo and Tainer.

ATSAS tool: *datmw*

Why is all this stuff important for SANS?

$$R^2_{g,obs} = R^2_C + \frac{\alpha}{\Delta\bar{\rho}} - \frac{\beta}{\Delta\bar{\rho}^2},$$

a. Homogeneous Solution	b. Dissociating Complex	c. Oligomerisation/ Aggregation
		
The dependency of $I(0)$ with f_{D2O} should be represented accurately by a parabolic curve that intersects the f_{D2O} axis at a single point, corresponding to the match-point of the complex.	The dependency of $I(0)$ with f_{D2O} should be reasonably well represented by a parabolic curve that intersects the f_{D2O} axis at a single point. The parabola will typically miss data points at the extremities of the plot, but the deviation of the estimated match-point from the real value is usually small. Of note, $I(0)$ values from a dissociating complex can be accurately represented by general quadratic function, but that function will not intersect the f_{D2O} axis.	The dependency of $I(0)$ with f_{D2O} will not be reasonably well represented by a parabolic curve that intersects the f_{D2O} axis at a single point. The parabola will typically miss many data points, and the deviation of the estimated match-point from the real value is usually significant.



EMBL

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.

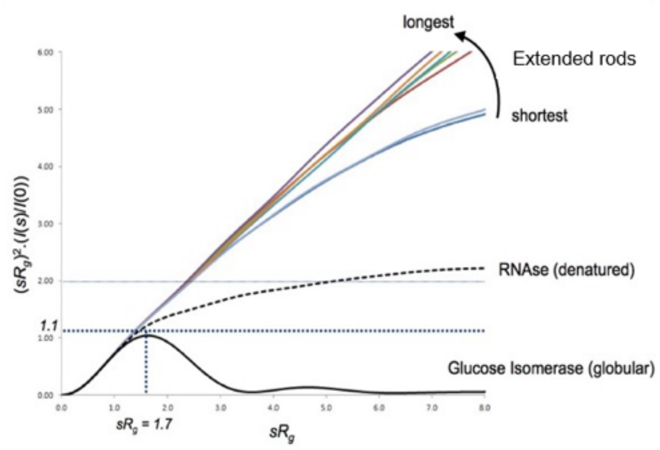
Part 2 of your validation toolbox →

- For example, obtain the ***EXACT*** sequence of the protein(s), RNA, DNA, glycans, etc actually used for the SAS measurement. ***ALL atoms scatter***, so you have to take into account ***ALL of the mass*** in your modelling!
- Obtain the ***CORRECT PDB (or cif)*** files (i.e., atomic coordinate files). ***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 – especially important for neutrons; (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, bioinformatics, Alphafold.)
- Know the ***STOICHIOMETRY*** and from this, the estimated ***SYMMETRY***. Obtain the MM estimate from SAS or other methods, e.g., MALLS, AUC, mass-photometry, etc.

Get to *know your data* before you model!

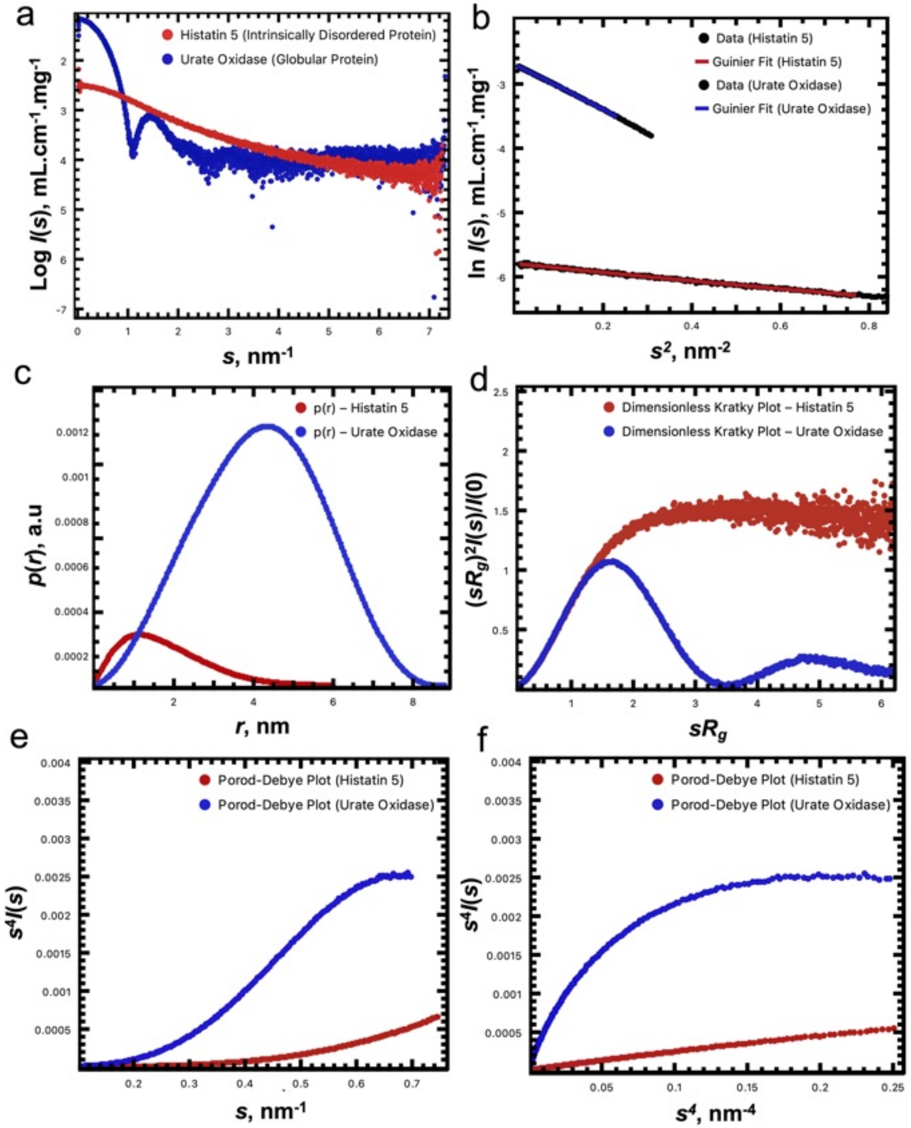
Fundamental plots

1) Guinier
2) P(*r*)
3) Dimensionless Kratky
4) Porod-Debye



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

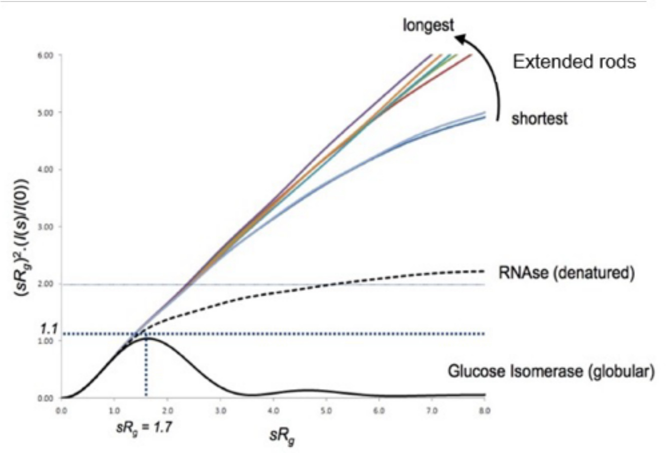
Durand D, Vivès C, Cannella D, Pérez J, Pebay-Peyroula E, Vachette P, Fieschi F. (2010) *J Struct Biol.* 169: 45-53.



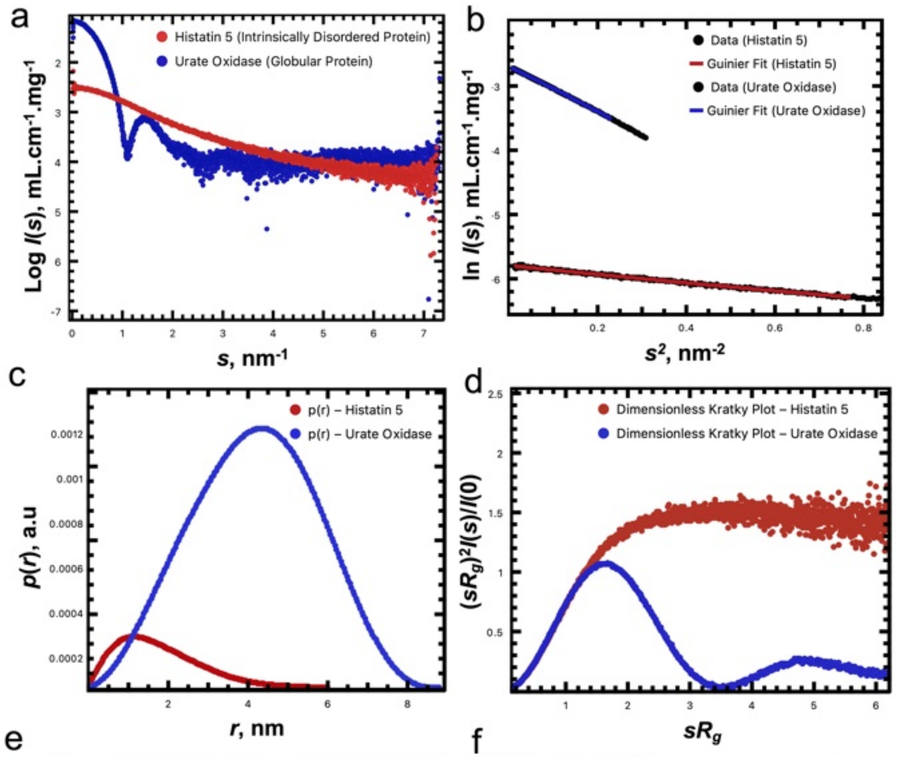
Get to *know your data* before you model!

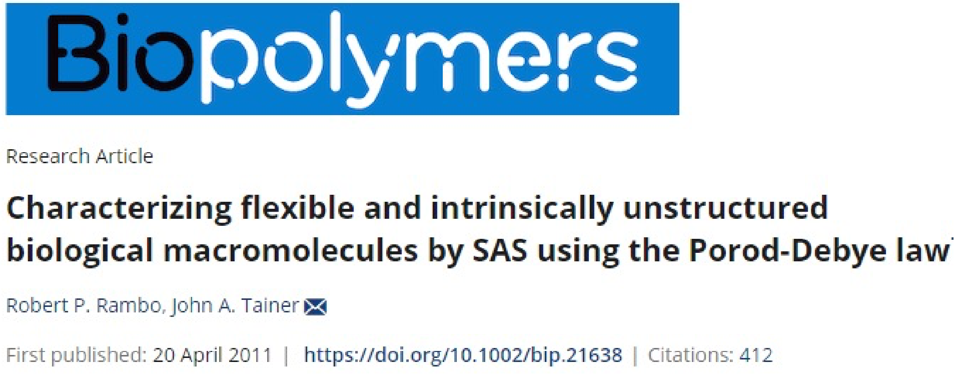
Fundamental plots

1) Guinier
2) P(r)
3) Dimensionless Kratky
4) Porod-Debye



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.




Biopolymers

Research Article

Characterizing flexible and intrinsically unstructured biological macromolecules by SAS using the Porod-Debye law

Robert P. Rambo, John A. Tainer

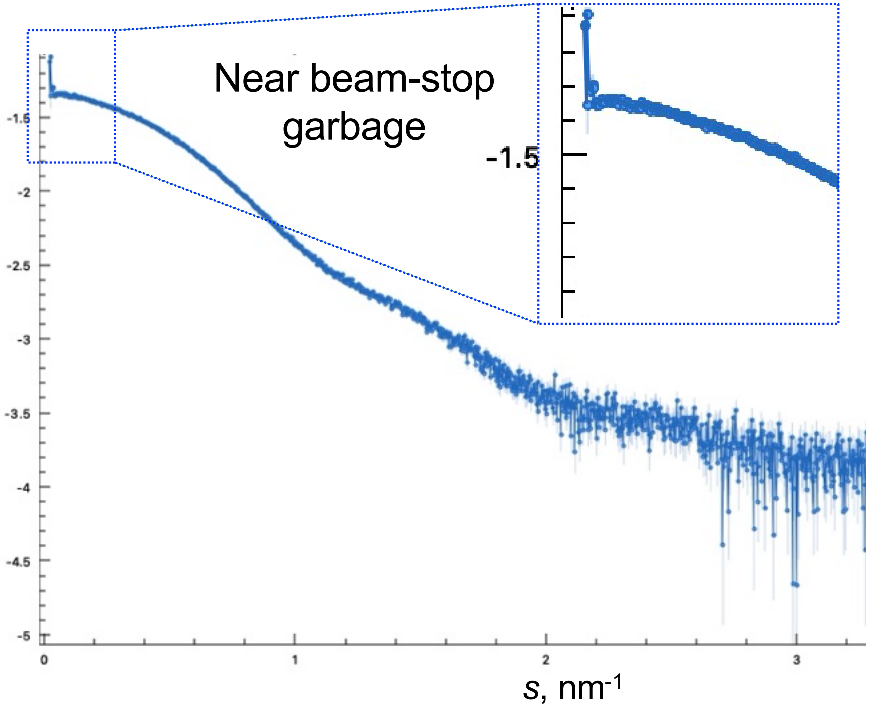
First published: 20 April 2011 | https://doi.org/10.1002/bip.21638 | Citations: 412

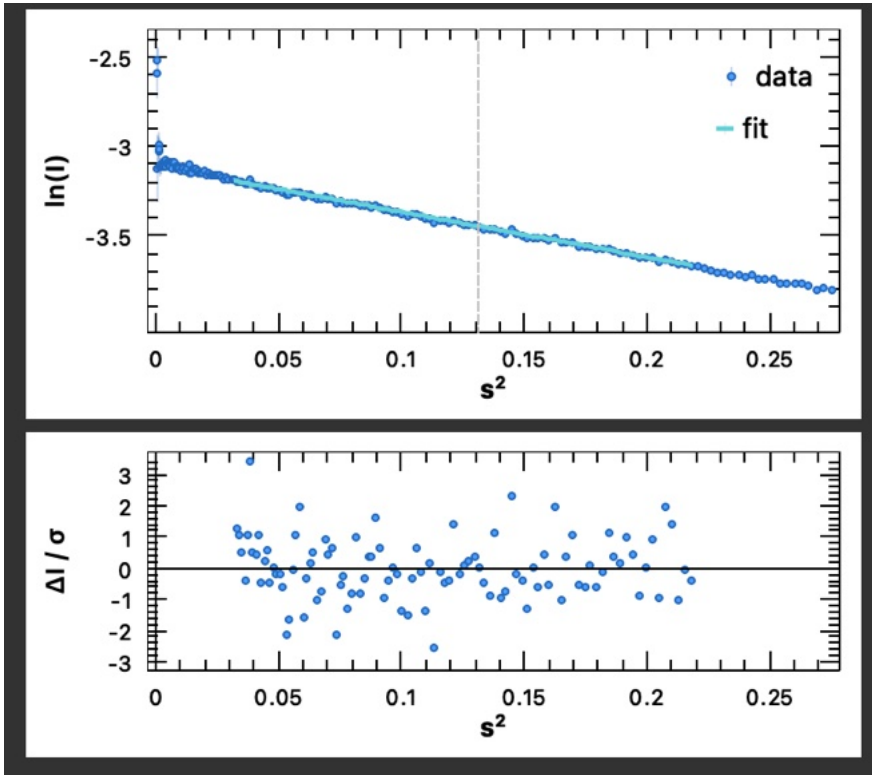
Get to *know your data* before you model!

2) Five programs:

- AutoRg – for first assessments of s_{min}
- SHANUM – define the useful s_{max}.
- DATCLASS – machine-learning method for the rapid geometric classification of SAXS data (from proteins).
- DARA – kd-tree searching of the PDB + Alphafold DB for similar scattering profiles.
- AMBIMETER – assess the ambiguity of the scattering data.

AutoRg and s_{min}





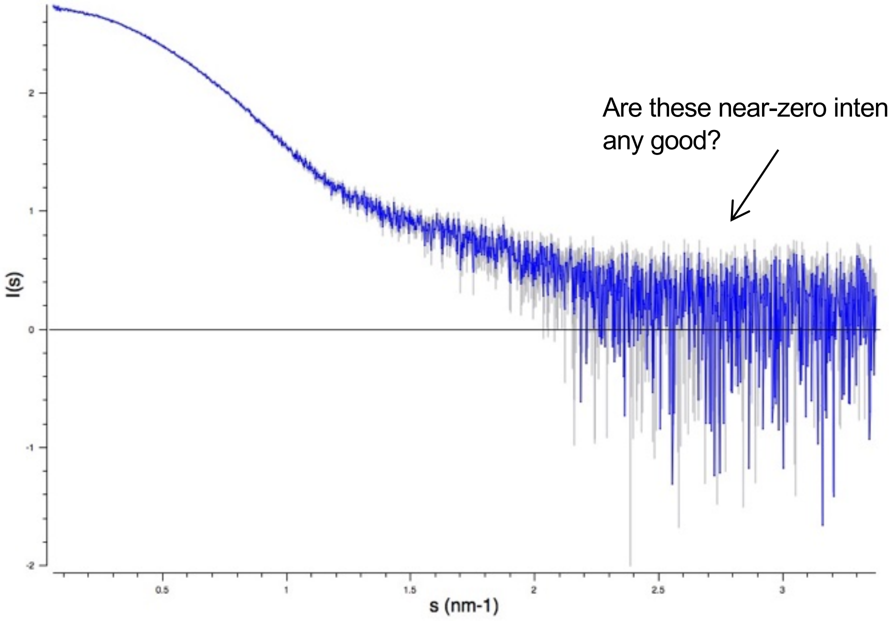
After cleaning up! **Is there sufficient data at the very lowest *s* to encompass the particle size?**

At minimum, s_{min} should = π/D_{max}
Better rule of thumb, $s_{\mathrm{min}} = 1/D_{\mathrm{max}}$

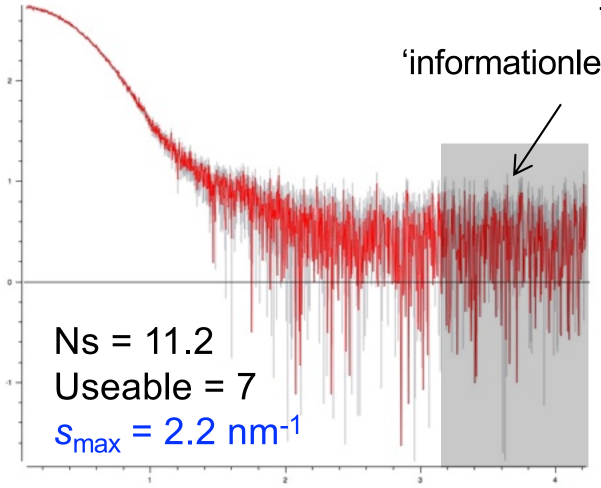
EMBL

Shanum and s_{max}

Shanum takes into account the statistical variance in the data to assess the useable s_{max}.

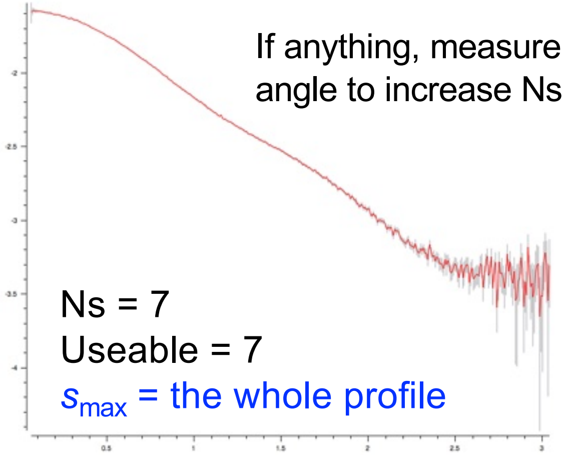


Are these near-zero intensities any good?



This bit is 'informationless noise'

Ns = 11.2
Useable = 7
s_{max} = 2.2 nm^{-1}



If anything, measure to a higher angle to increase Ns to 10?

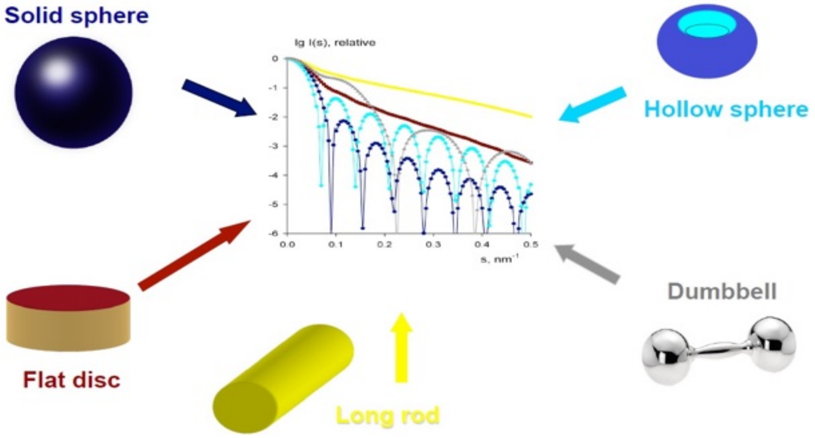
Ns = 7
Useable = 7
s_{max} = the whole profile

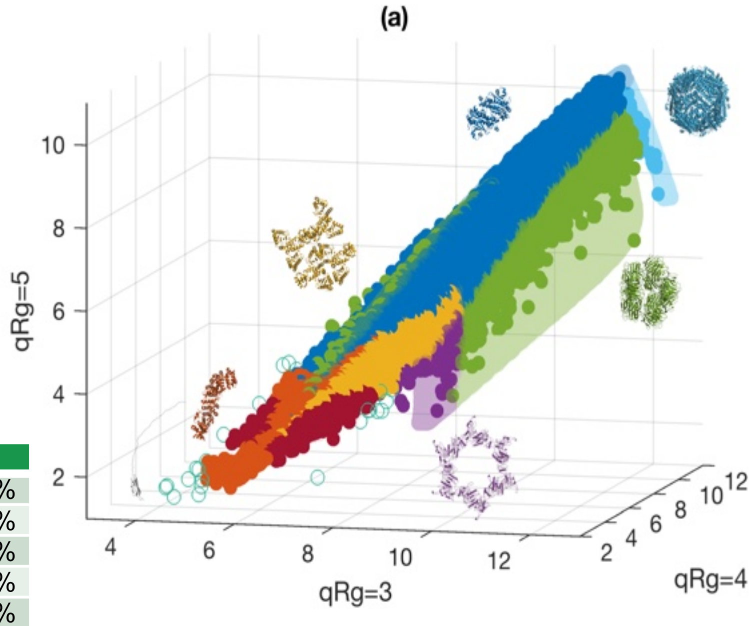
Shanum will also estimate D_{max} for you (or you can enter it yourself)

Datclass

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

99.98% of the PDB maps into the classifier space.





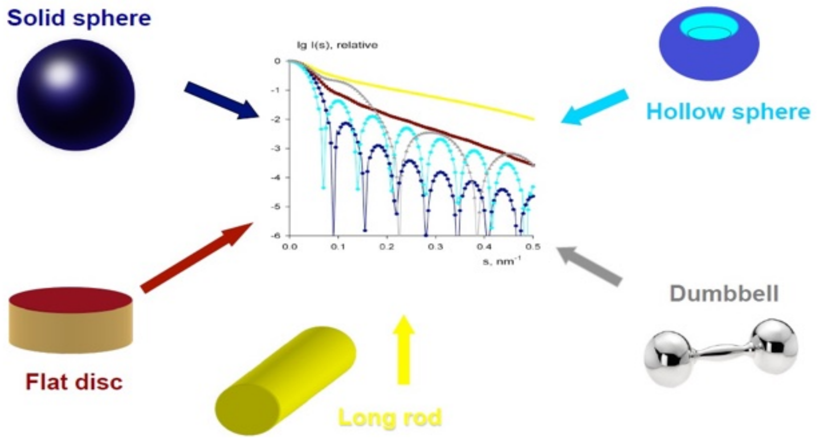
Class Label	PDB	
Unknown	25	0. 02 %
Compact	122.913	74.05 %
Extended	5.382	3.24 %
Flat	9.734	5.86 %
Ring	154	0.09 %
Compact hollow	26.909	16.21 %
Hollow sphere	125	0.08 %
Random Chain	740	0.45 %
Total	165.982	100.00 %

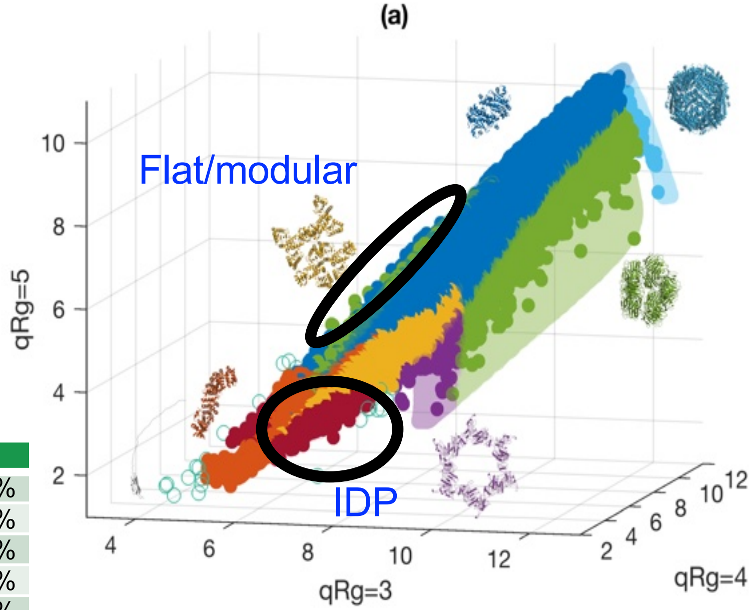
EMBL

Datclass

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

99.98% of the PDB maps into the classifier space.





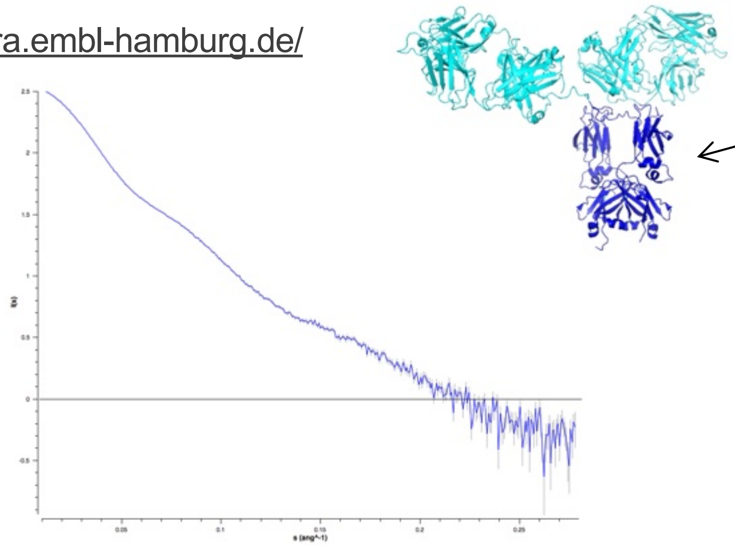
Class Label	PDB	
Unknown	25	0. 02 %
Compact	122.913	74.05 %
Extended	5.382	3.24 %
Flat	9.734	5.86 %
Ring	154	0.09 %
Compact hollow	26.909	16.21 %
Hollow sphere	125	0.08 %
Random Chain	740	0.45 %
Total	165.982	100.00 %

For IDP and Flat/modular = an ensemble approach might be considered!

DARA

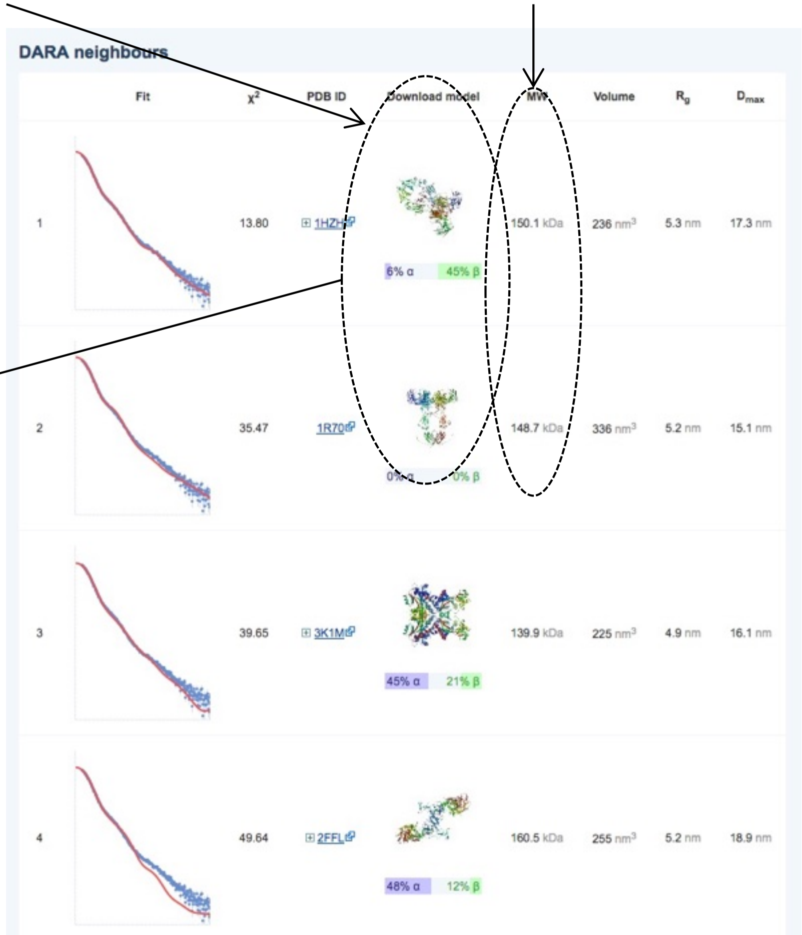
Kd-tree nearest neighbour search of a .dat file or GNOM.out file against calculated SAXS profiles – PDB and Alphafold.

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



IgG or IgA like scattering

MW estimates!



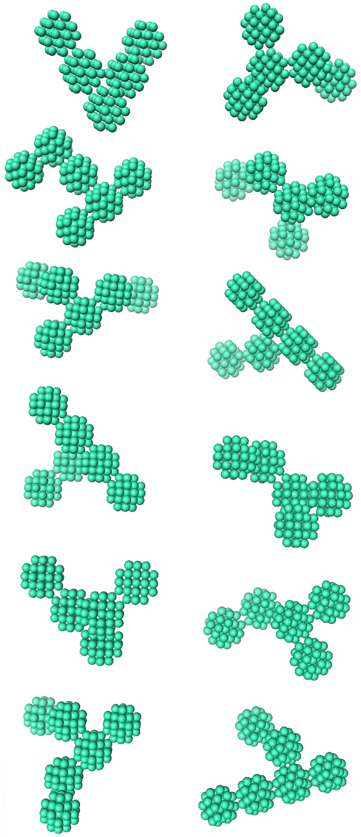
DARA neighbours

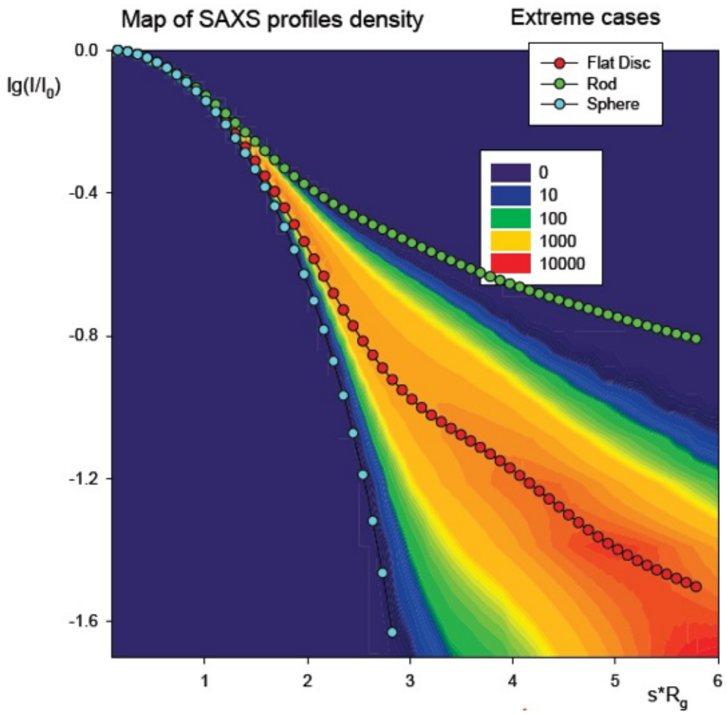
	Fit	χ^2	PDB ID	Download model	MW	Volume	R_g	D_{max}
1		13.80	⊞ 1HZH	6% α 45% β	150.1 kDa	236 nm^3	5.3 nm	17.3 nm
2		35.47	1R70	0% α 0% β	148.7 kDa	336 nm^3	5.2 nm	15.1 nm
3		39.65	⊞ 3K1M	45% α 21% β	139.9 kDa	225 nm^3	4.9 nm	16.1 nm
4		49.64	⊞ 2FFL	48% α 12% β	160.5 kDa	255 nm^3	5.2 nm	18.9 nm

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

Ambimeter

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





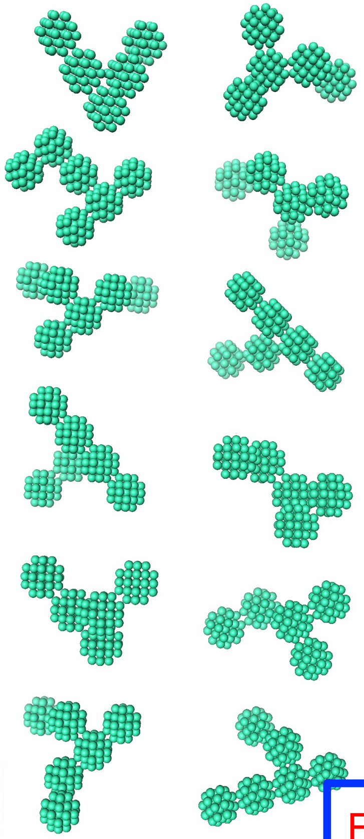
- Provides a sense of how ambiguous a dataset is with respect to fitting models.
- An ambimeter score of 0 to 1.5(ish) are ‘potentially unique’ shapes.
- An ambimeter score of 2.9, for example, is very highly ambiguous.

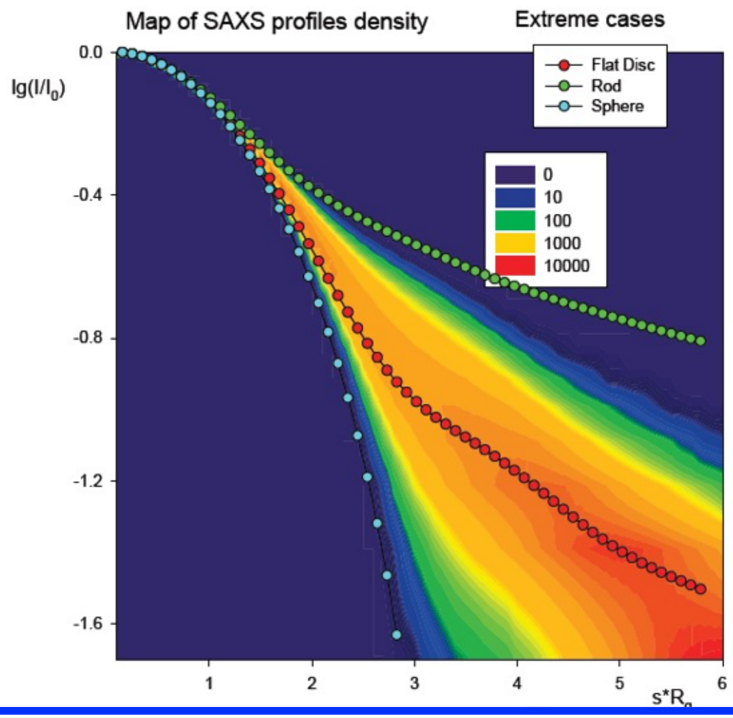
What to do?

- Always run modelling routines several times!
- Use information from other techniques.
- Perform parallel modelling against several SAS datasets.

Ambimeter

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





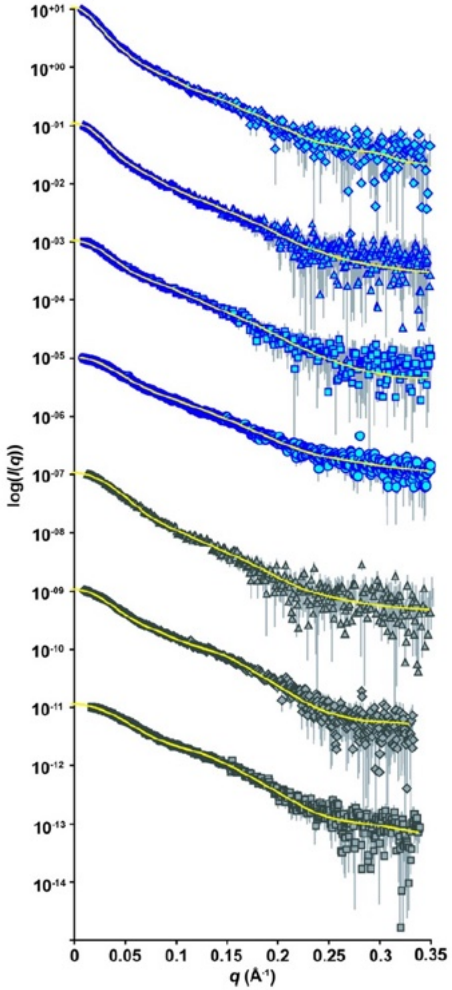
Flat things are highly ambiguous. Classification as 'flat/modular' – are these modular with flexible linkers or just highly ambiguous?

- Provides a sense of how ambiguous a dataset is with respect to fitting models.
- An ambimeter score of 0 to 1.5(ish) are 'potentially unique' shapes.
- An ambimeter score of 2.9, for example, is very highly ambiguous.

What to do?

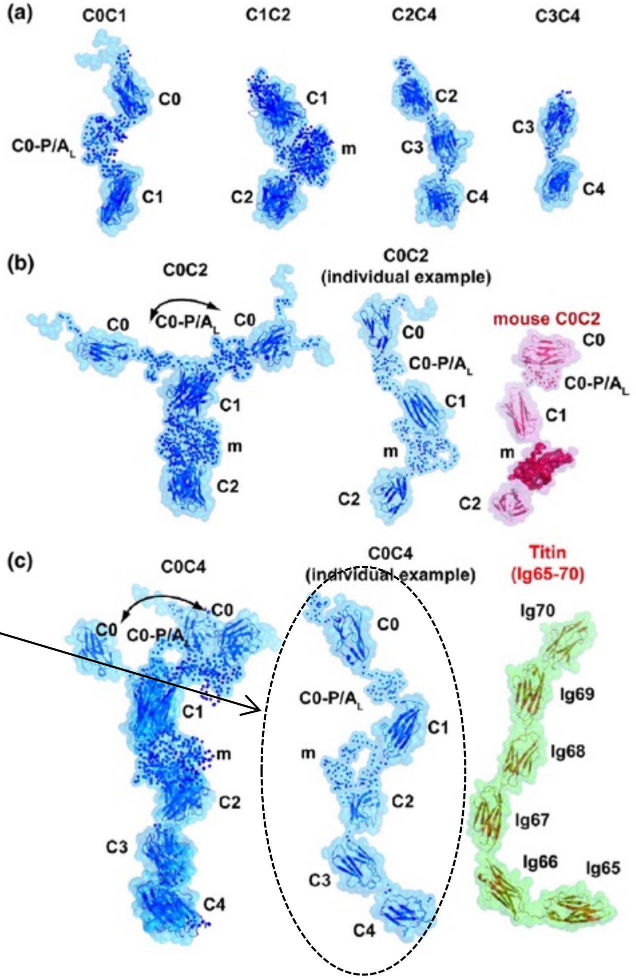
- Always run modelling routines several times!
- Use information from other techniques.
- Perform parallel modelling against several SAS datasets.

More information = less ambiguity!



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

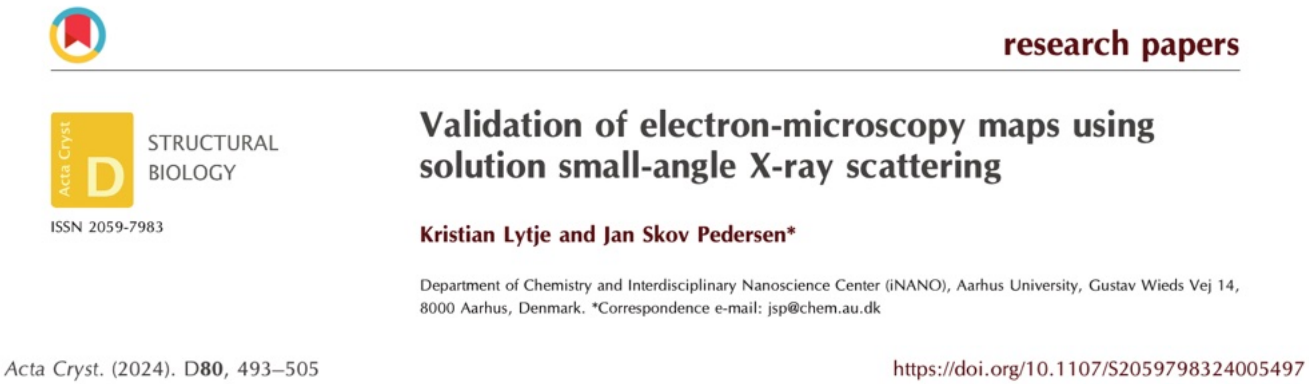
Final Target: fits but importantly DOES NOT describe *in toto* what is going on



EMBL

More information = less ambiguity! From other methods.

Xay-ray crystallography
NMR
FRET
Electron microscopy
Mass-spec (HDX, cross linking)
Predictive methods (Alphafold)



research papers

Acta Cryst D STRUCTURAL BIOLOGY

ISSN 2059-7983

Validation of electron-microscopy maps using solution small-angle X-ray scattering

Kristian Lytje and Jan Skov Pedersen*

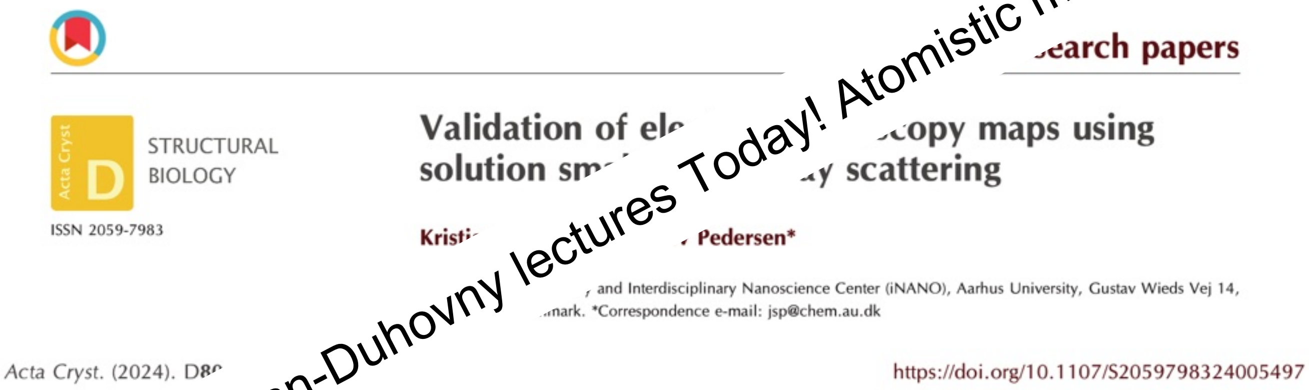
Department of Chemistry and Interdisciplinary Nanoscience Center (iNANO), Aarhus University, Gustav Wieds Vej 14, 8000 Aarhus, Denmark. *Correspondence e-mail: jsp@chem.au.dk

Acta Cryst. (2024). D**80**, 493–505

https://doi.org/10.1107/S2059798324005497

More information = less ambiguity! From other methods.

Xay-ray crystallography
NMR
FRET
Electron microscopy
Mass-spec (HDX, cross linking)
Predictive methods (Alphafold)



…earch papers

Acta Cryst D
STRUCTURAL
BIOLOGY

ISSN 2059-7983

Validation of ele… …copy maps using solution sm… …ay scattering

Kristi… …Pedersen*

…, and Interdisciplinary Nanoscience Center (iNANO), Aarhus University, Gustav Wieds Vej 14, …mark. *Correspondence e-mail: jsp@chem.au.dk

Acta Cryst. (2024). D8…

https://doi.org/10.1107/S2059798324005497

Wait for Sergei Grudinin and Dina Schneidman-Duhovny lectures Today! Atomistic modelling

Summary: Know your data.

- R_g and $I(0)$ from Guinier and $p(r)$ – check for consistency through a concentration series. Identify concentration independent interparticle interactions: coulombic-repulsive or aggregation. Deal with it.
- *Prepare your data for modelling:* s_{min} and s_{max} from AUTORG and SHANUM. Make sure s_{min} is at least = π/D_{max}, or better yet, $1/D_{\text{max}}$!
- Molecular mass estimates – very important for guiding the modelling approach!
- Ambiguity.

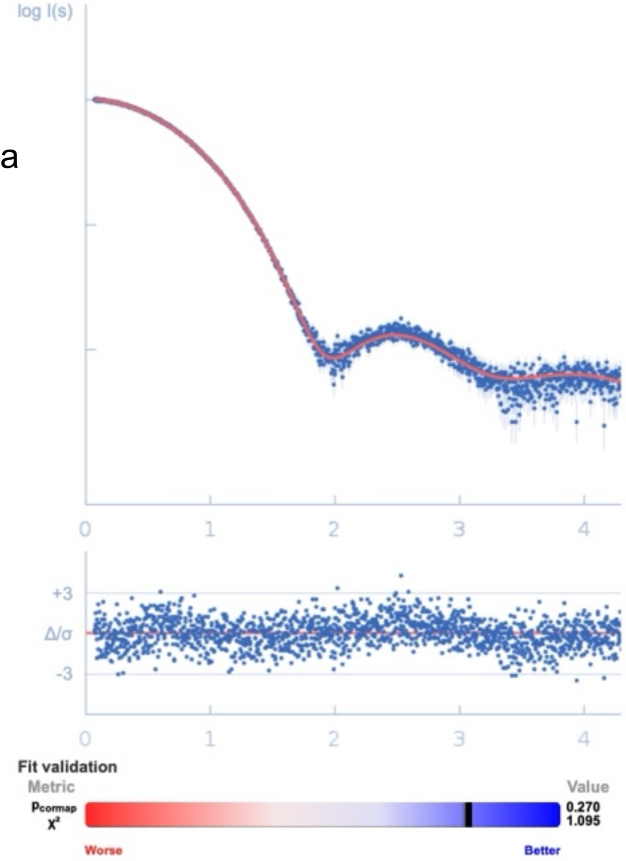
Any questions so far?

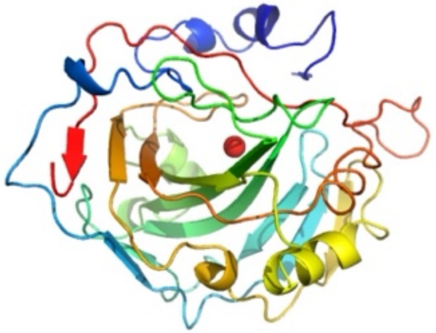
EMBL

Lets do some atomistic model fitting!

https://www.sasbdb.org/data/SASDFP8/

Assessing SAS-data model fit



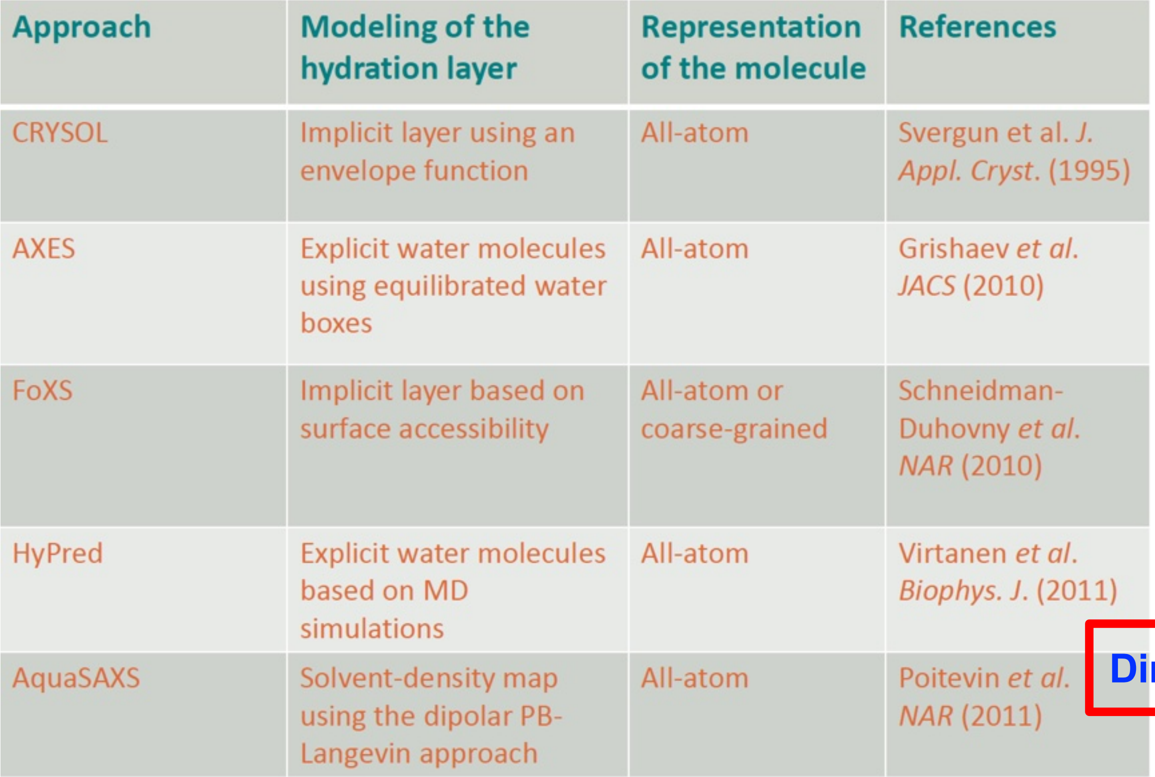


Atomistic all-atom model – e.g., from the PDB, Alphafold, etc

Error-normalized residual plot– looking systematic deviations between the calculated model scattering intensities and the experimental data. CORRECT ERRORS.

CorMap *P* and χ^2 evaluations

Assessing SAS-data model fits – Methods abound!



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)

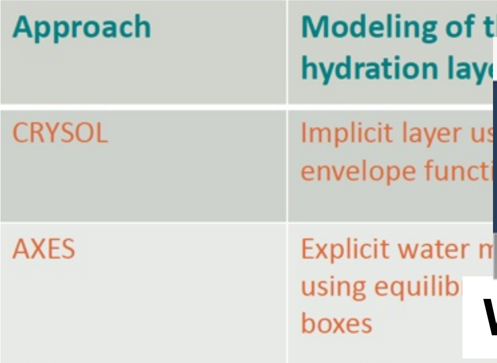
CRYSON – for SANS

FoXS – Debye formula

$$I(q) = \sum_i \sum_j w_i w_j \frac{\sin qr_{ij}}{qr_{ij}}$$

Dina Schneidman-Duhovny…is here!

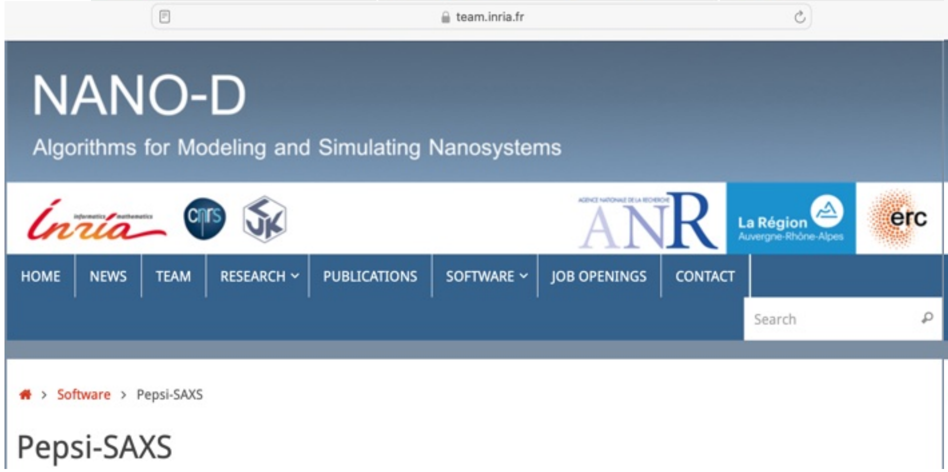
Assessing SAS-data model fits – Methods abound!

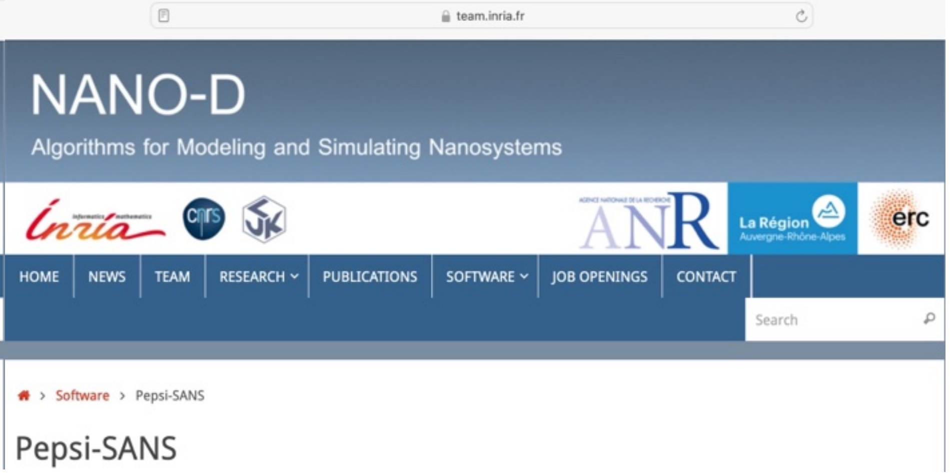


Approach	Modeling of the hydration lay	Representation	References
CRYSOL	Implicit layer us envelope functi		
AXES	Explicit water m using equilib boxes		

WAXS in Solvent (WAXSiS) computes small- and wide-angle X-ray scattering curves based on explicit-solvent all-atom molecular dynamics simulations.

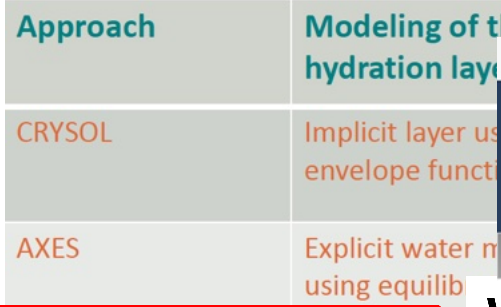
WAXIS – molecular dynamics for the hydration layer!





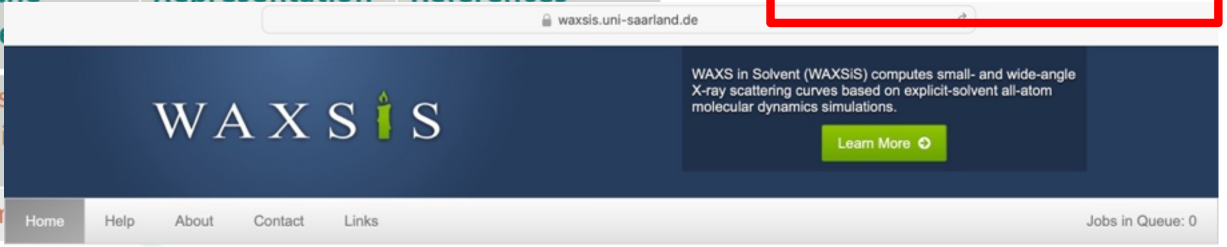
PEPSI-SAXS and PEPSI-SANS

Assessing SAS-data model fits – Methods abound!



Approach	Modeling of the hydration lay	Representation	References
CRYSOL	Implicit layer us envelope functi		
AXES	Explicit water m using equilib boxes		

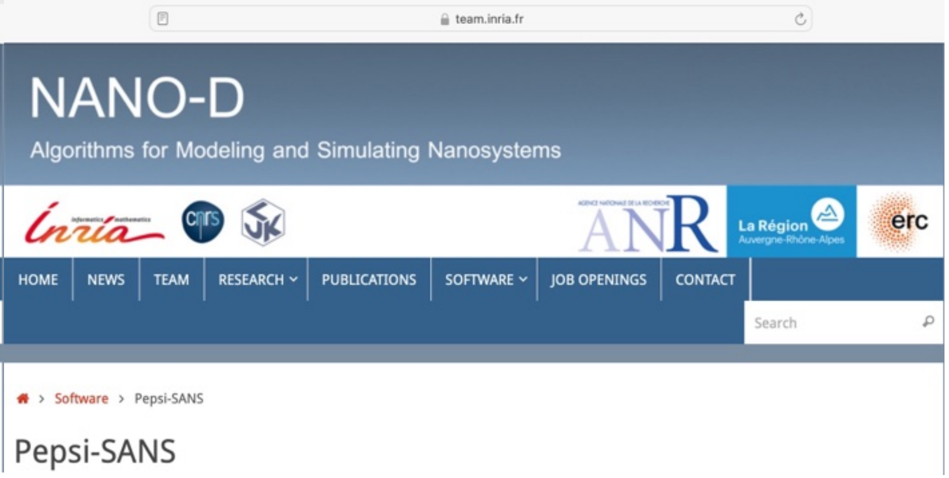
Jochen Hub…is here!



WAXS in Solvent (WAXSiS) computes small- and wide-angle X-ray scattering curves based on explicit-solvent all-atom molecular dynamics simulations.

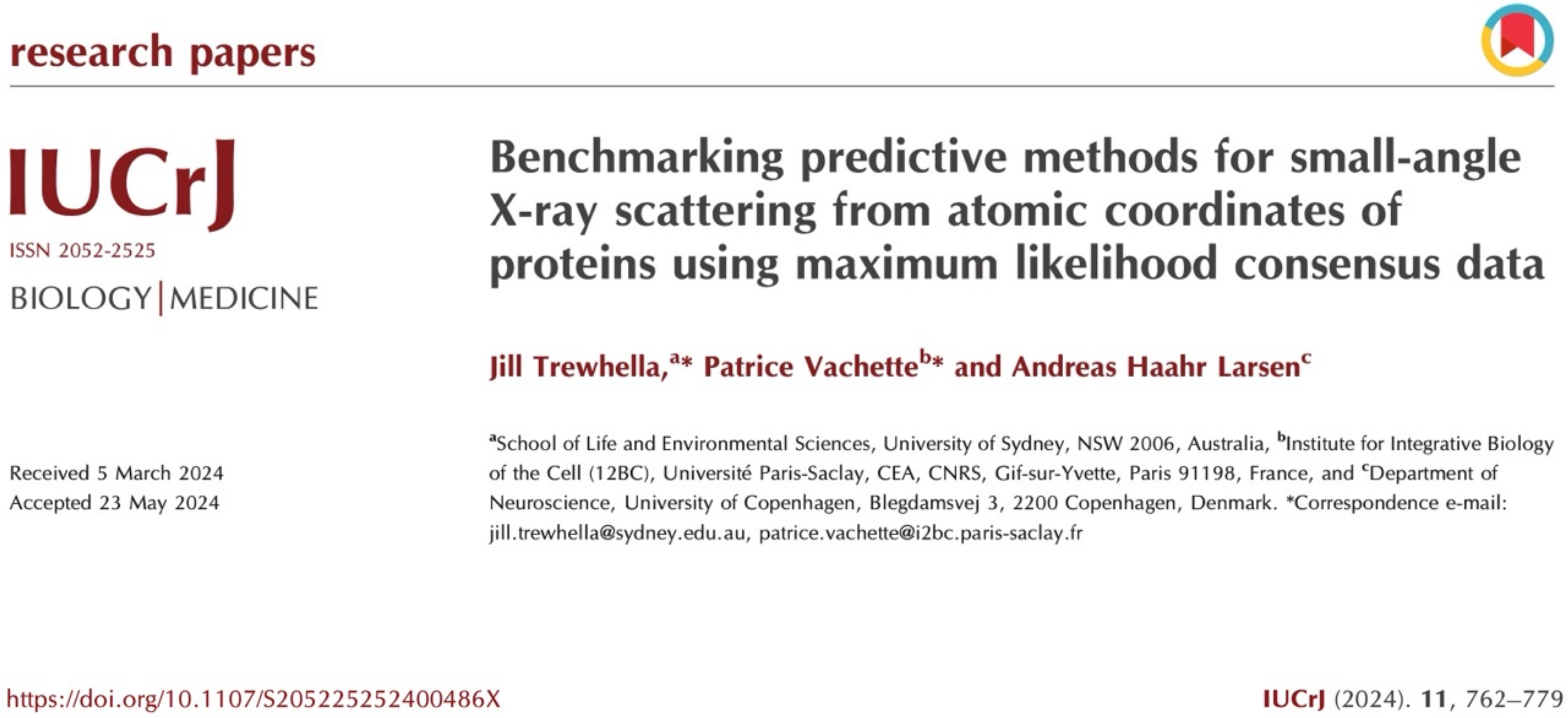
WAXIS – molecular dynamics for the hydration layer!

Sergei Grudinin…is here!



PEPSI-SAXS and PEPSI-SANS

Assessing SAS-data model fits – Methods abound!



research papers

IUCrJ

ISSN 2052-2525

BIOLOGY | MEDICINE

Benchmarking predictive methods for small-angle X-ray scattering from atomic coordinates of proteins using maximum likelihood consensus data

Jill Trewhella,[a]* Patrice Vachette[b]* and Andreas Haahr Larsen[c]

[a]School of Life and Environmental Sciences, University of Sydney, NSW 2006, Australia, [b]Institute for Integrative Biology of the Cell (12BC), Université Paris-Saclay, CEA, CNRS, Gif-sur-Yvette, Paris 91198, France, and [c]Department of Neuroscience, University of Copenhagen, Blegdamsvej 3, 2200 Copenhagen, Denmark. *Correspondence e-mail: jill.trewhella@sydney.edu.au, patrice.vachette@i2bc.paris-saclay.fr

Received 5 March 2024
Accepted 23 May 2024

https://doi.org/10.1107/S205225252400486X

IUCrJ (2024). **11**, 762–779

Workhorse in ATSAS: CRYSOL (for SAXS)

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.

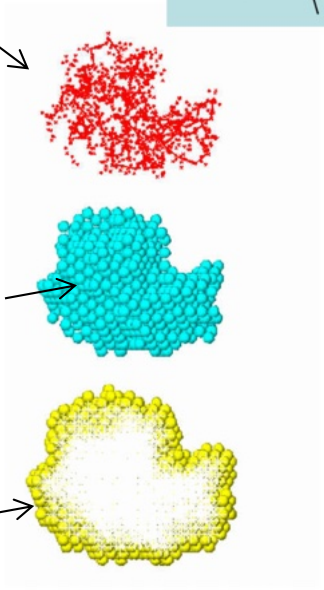
SLD of the solvent

$$I(\mathbf{s})=\left\langle\left|A(\mathbf{s})\right|^{2}\right\rangle_{\Omega}=\left\langle\left|A_{a}(\mathbf{s})-\rho_{s}A_{s}(\mathbf{s})+\delta\rho_{b}A_{b}(\mathbf{s})\right|^{2}\right\rangle_{\Omega}$$

Electrons (nuclei) are ‘points’

Excluded volume with the SLD of the solvent.

SLD of the hydration layer is slightly different to bulk.



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

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

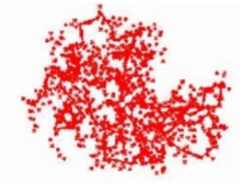
***CRYSON* (neutrons):** Svergun et al. (1998) *P.N.A.S. USA*, **95**, 2267

EMBL

Workhorse in ATSAS: CRYSOL

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

- 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/cm^3).
- Provide output files (scattering amplitudes) for rigid body refinement routines.
- Compute particle envelope function F(ω)



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

Spherical wave bit

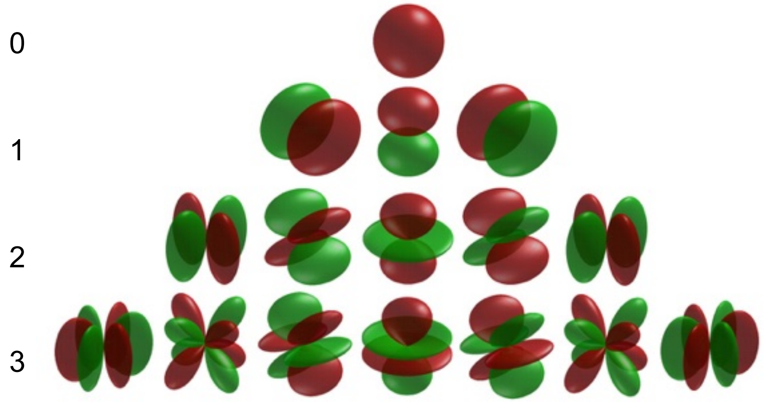
‘Scattering factor’: relates to the atomic cross section, i.e., scattering length, or probability of an atom to scatter for every atom in the sample.

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

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

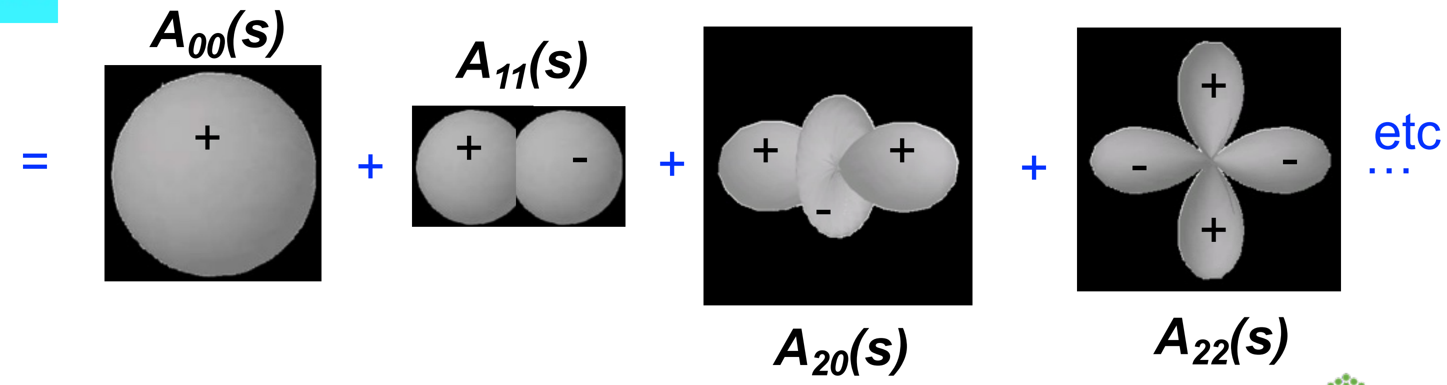
In 1970, Stuhrmann showed that the information content of a SAXS profile can be conveniently described in terms of a sum of spherical harmonic functions.

Spherical Harmonics



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.)

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



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

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

How many spherical harmonics to use in CRYSOL?

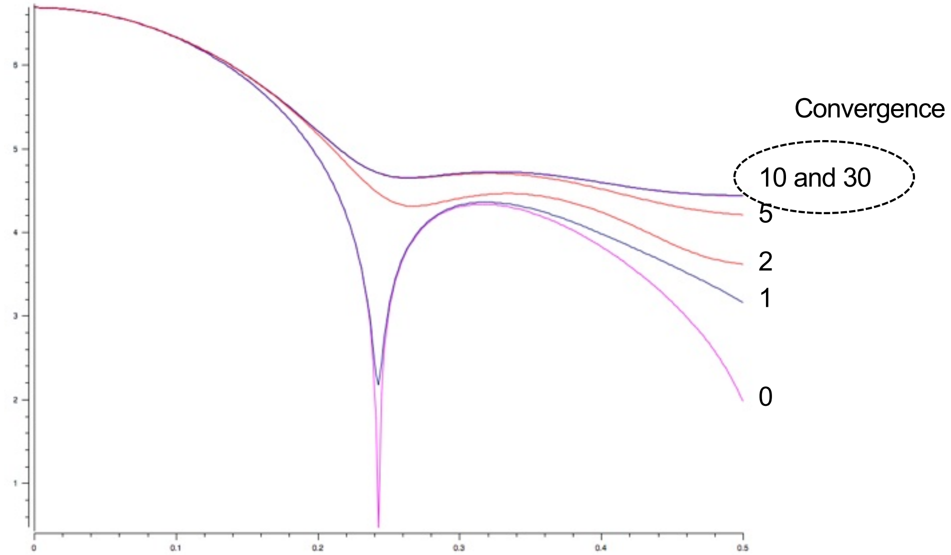
If you use the first harmonic only, i.e., zeroth-order, then the calculated intensities from the model will be a sphere. This is okay only if you want to describe the overall SIZE of the object, i.e., at the very lowest of angles in the Guinier region of the scattering profile. The zeroth-order harmonic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

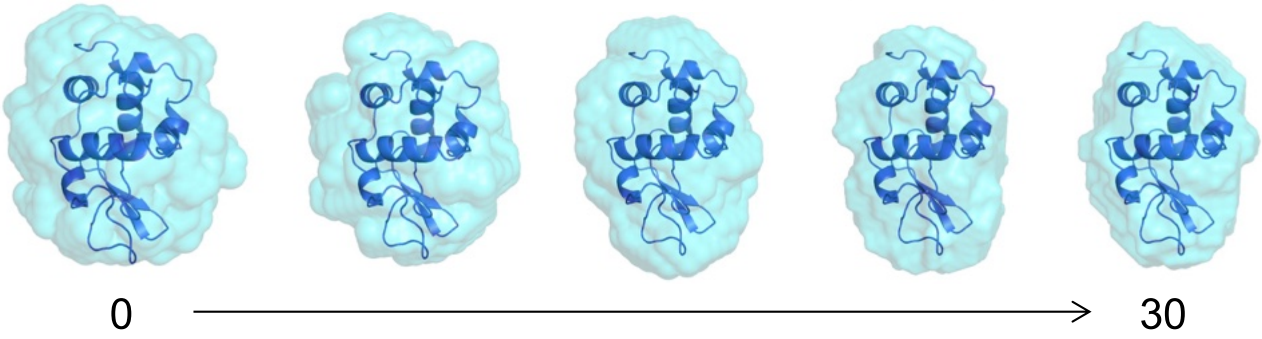
If you use two harmonics, you will introduce an additional ‘shape feature’ into the calculated scattering intensities across ***s***…but the resulting shape will probably still look like a sphere..with a couple of very low humps.

If you continue to increase the number of harmonics, you introduce additional 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).

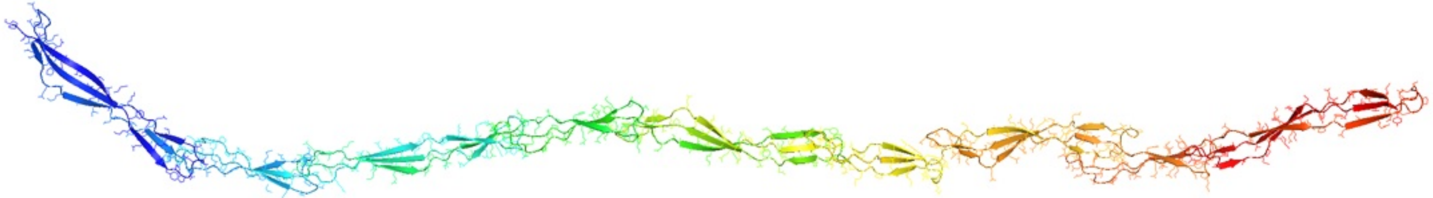
Typically 15-30 harmonics are used to describe size and the shape of the object. However, this depends on the CLASSIFICATION of an object. Clearly, if the object is an extended rod, you probably need additional spherical harmonics terms.

I(*s*) from a globular structure using different numbers of harmonics

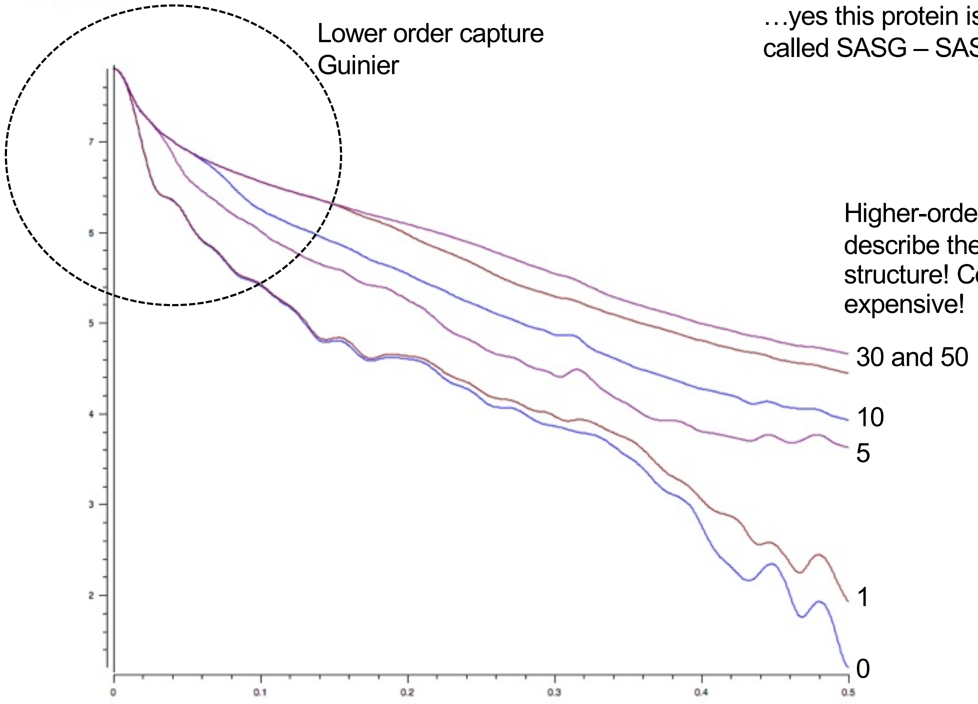




I(*s*) from an extended structure using different numbers of harmonics



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



Lower order capture Guinier

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

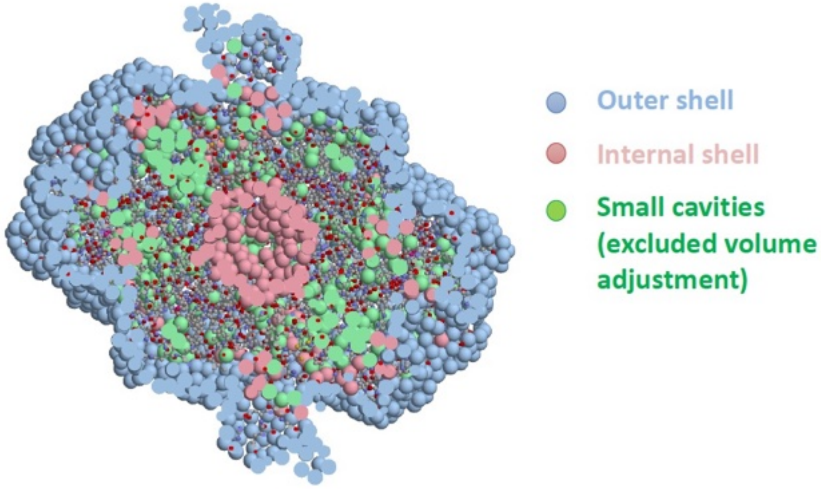
Centre your atomic models!

- 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*

For macromolecules with cavities and holes – explicit hydration water using CRYSOL3

- Hydration shell representation as envelope function (CRYSOL – implicit solvent layer) or dummy solvent beads, i.e., explicit solvent layer (CRYSOL 3).
- Explicit solvent modelling is important for internal cavities!



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

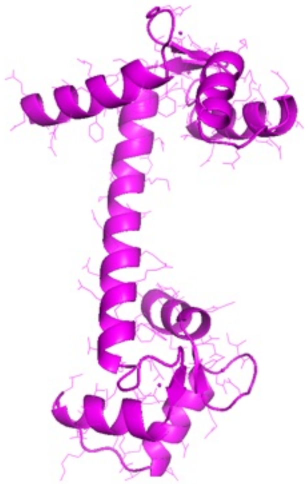
Assessing data-model fits – χ^2

…knowing what model does NOT fit the data can be as valuable as knowing what model(s) do fit the data!

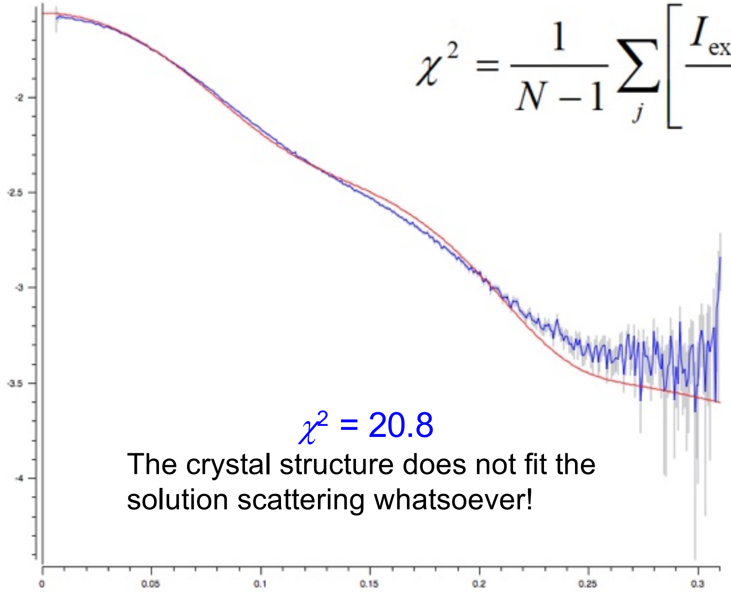
CRYSOL fit to the SAXS data. The goodness of fit is described by the reduced χ^2 discrepancy.

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

Calmodulin: X-ray crystal structure



PDB: 3CLN



χ^2 = 20.8

The crystal structure does not fit the solution scattering whatsoever!

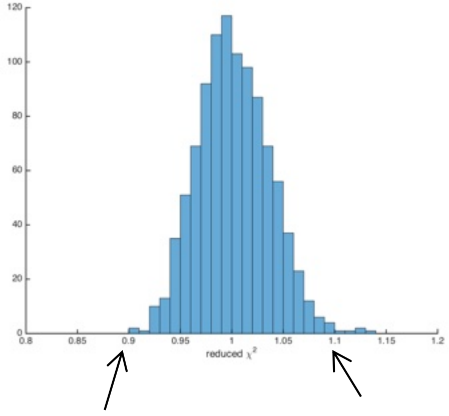
A note on χ^2

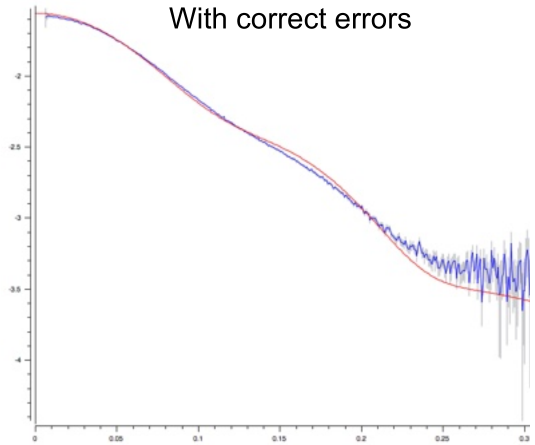
$$\chi^2 = \frac{1}{N-1}\sum_j \left[\frac{I_{\exp}(s_j) - cI(s_j)}{\boxed{\sigma(s_j)}}\right]^2$$

The errors on the scattering intensities need to be correctly specified, otherwise the test is, by default, INVALID. Errors follow Poisson counting statistics that limit to a gaussian distribution after many repetitions (for photon counting detectors).

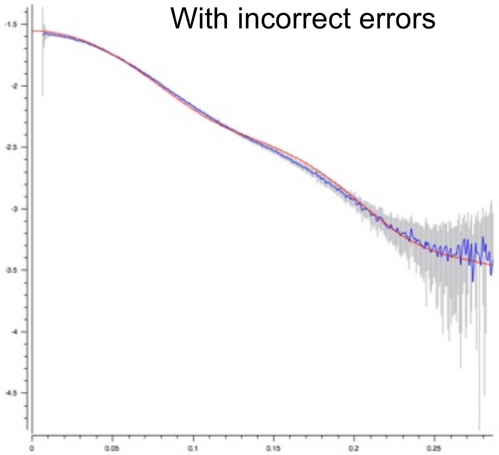
If the errors are correctly specified and no significant (systematic) deviations are present between the experimental and modeled intensities, the value should lie in the range of approximately 0.9-1.1 depending on the number of points in the dataset (0.9-1.1 is typical for over-sampled SAXS data on modern detectors).

Same intensities, same model, but different error estimates





χ^2 = 20.8
With correct errors



χ^2 = 1.2
With incorrect errors

MBL

Correlation Map: CorMap *P*

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

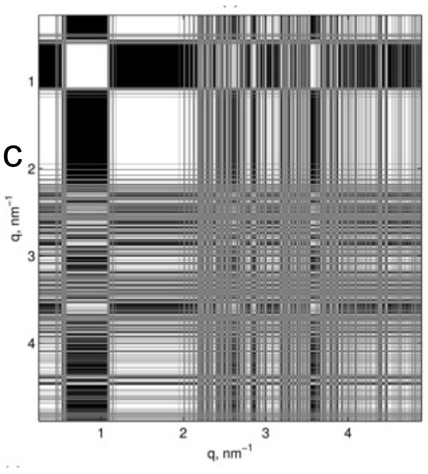
$$\sigma(I_{\exp}(q_k))^2 = \frac{1}{m-1}\sum_{i=1}^{m}(I_{\exp}(q_k)_i - \bar{I}_{\exp}(q_k))^2$$

On-diagonal variance.

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

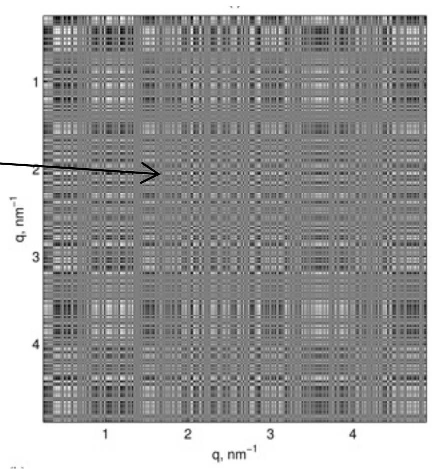
Off-diagonal co-variance between all point-to-point q_k and q_l.

Statistically significant, systematic differences between the modelled and experimental intensities



$P < 0.01$

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



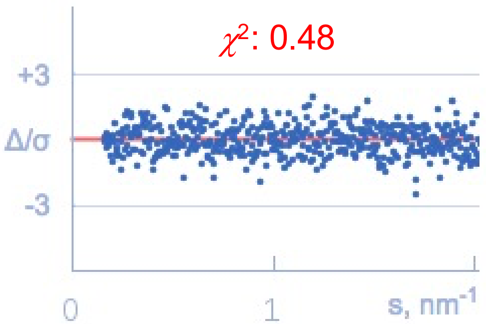
$P > 0.01$

EMBL

Error normalized residual plots

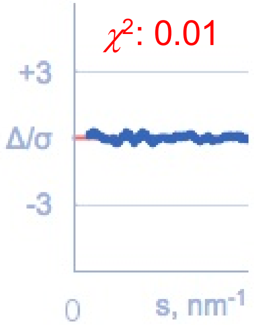
- Model fits to the data are also evaluated using normalized residual plots to help assess systematic model-fit deviations from the data in addition to over or under-estimation of the errors.

$$resudual = \frac{(I(s)experiment - cI(s)model)}{\sigma(I(s))}$$



Over-estimated errors

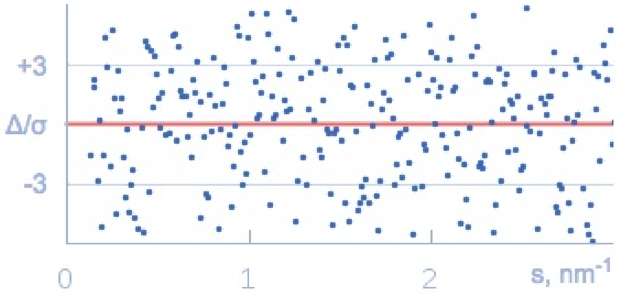
CorMap P > 0.01



Severely over-estimated errors

CorMap P < 0.01

χ^2: 12.5

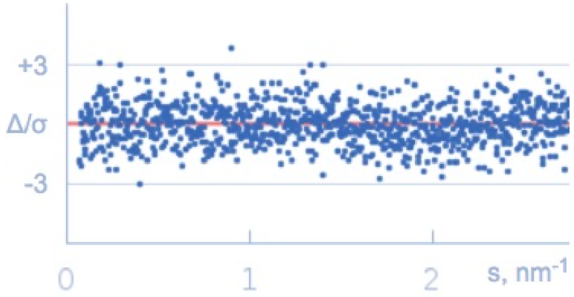


Severely under-estimated errors

CorMap P > 0.01

χ^2: 1.0

CorMap P > 0.01



✔ *Correctly specified errors*

EMBL

Lets do some atomistic model building!

SREFLEX
OLIGOMER
SASREF
BUNCH
CORAL
Ensemble Optimization Method (EOM)

Modelling 3D-structures that fit SAXS data is perhaps the fundamental ‘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 (or .cif) file(s).

- You should know the amino acid sequence of the protein (or polynucleotide, any other macromolecule, etc.) used for the SAS experiment. You should also know if the macromolecule binds metals, ligands, lipids, detergents, is glycosylated, etc.

- For proteins, use UNIPROT as a a fundamental resource to obtain the correct canonical sequence: www.uniport.org

- You should know what rigid-body (or bodies) you want to use for the modelling, i.e., the atomic coordinate PDB or .cif files.

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

EMBL

Amino acid sequence of protein used for SAS

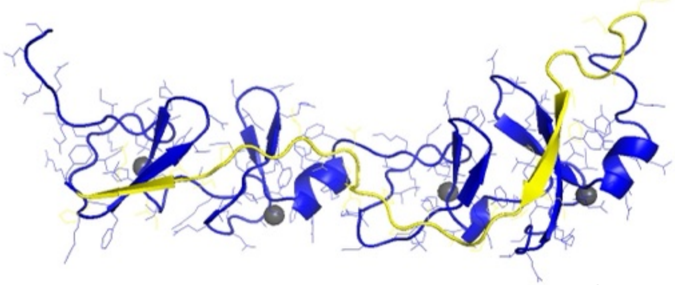
```
HMHHHHHHTRGSNNEEAICSLCDKKIRDRFVS
KVNGRCYHSSCLRCSTCKDELGATCFLREDS
MYCRAHFYKKFGTKCSSCNEGIVPDHVVRKA
SNHVYHVECFQCFICKRSLETGEEFYLIADDA
RLVCKDDYEQARDGGSGGHMGSGGGIGPLM
VQPATPHIDNTLGGPIDIQHF
```

↓

Align the sequences using Clustal Omega

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

Atomistic model from PDB file (filename.pdb)



What is the amino acid sequence?

↓

ATSAS tool: *pdb2seq*

↓

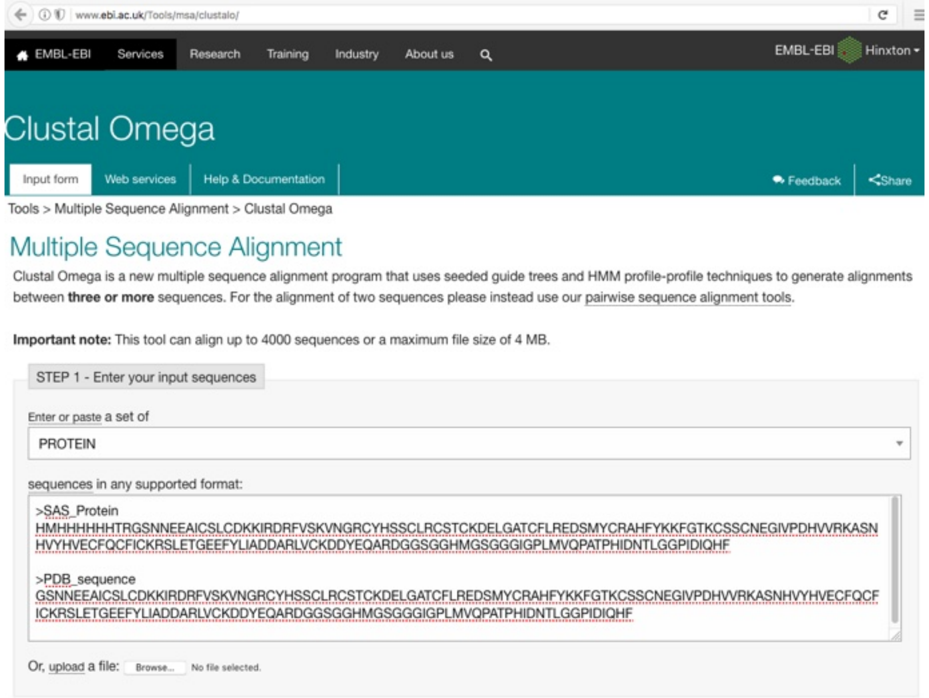
This will save the sequence in the text file called ‘filename.txt’

←

↑

```
GSNNEEAICSLCDKKIRDRFVSKVNGRCYHSS
CLRCSTCKDELGATCFLREDSMYCRAHFYKK
FGTKCSSCNEGIVPDHVVRKASNHVYHVECF
QCFICKRSLETGEEFYLIADDARLVCKDDYEQ
ARDGGSGGHMGSGGGIGPLMVQPATPHIDNT
LGG
PIDIQHF
```

Amino acid sequence of protein from PDB or .cif file



Clustal Omega

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

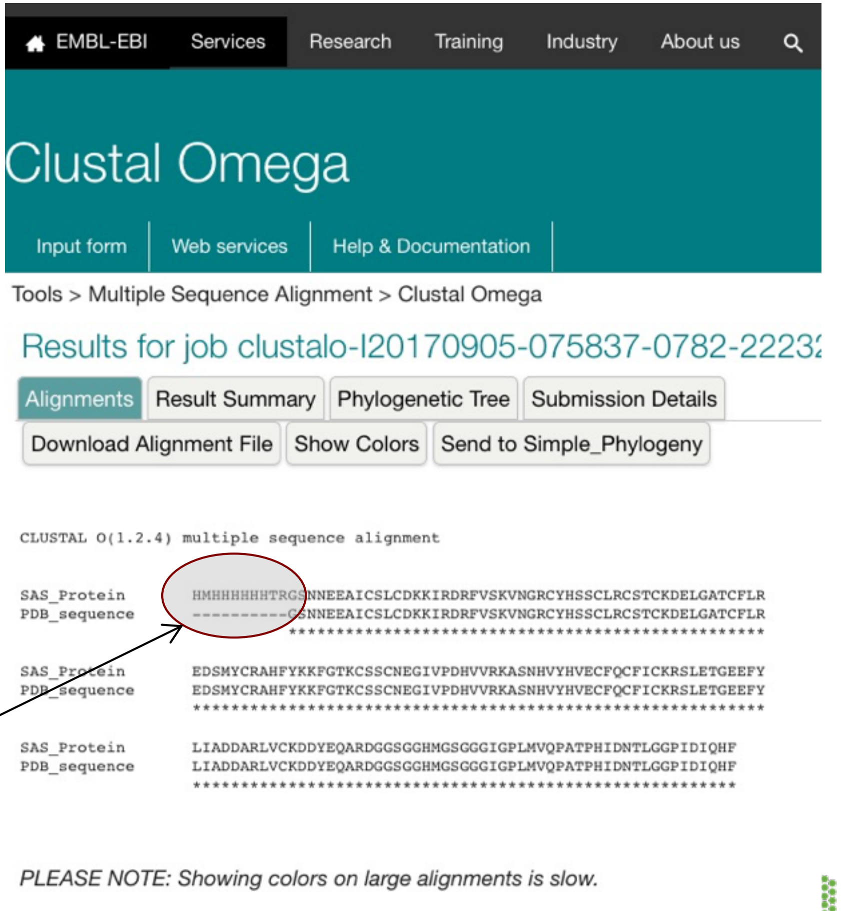
Enter or paste a set of

PROTEIN

sequences in any supported format:

```
>SAS_Protein
HMHHHHHHTRGSNNEEAICSLCDKKIRDRFVSKVNGRCYHSSCLRCSTCKDELGATCFLREDSMYCRAHFYKKFGTKCSSCNEGIVPDHVVRKASN
HVYHVECFQCFICKRSLETGEEFYLIADDARLVCKDDYEQARDGGSGGHMGSGGGIGPLMVQPATPHIDNTLGGPIDIQHF

>PDB_sequence
GSNNEEAICSLCDKKIRDRFVSKVNGRCYHSSCLRCSTCKDELGATCFLREDSMYCRAHFYKKFGTKCSSCNEGIVPDHVVRKASNHVYHVECFQCF
ICKRSLETGEEFYLIADDARLVCKDDYEQARDGGSGGHMGSGGGIGPLMVQPATPHIDNTLGGPIDIQHF
```

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



Clustal Omega

Tools > Multiple Sequence Alignment > Clustal Omega

Results for job clustalo-I20170905-075837-0782-2223

```
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!

Amino acid sequence of protein used for SAS

HMHHHHHHTRGSNNEEAICSLCDKKIRDRFVS
KVNGRCYHSSCLRCSTCKDELGATCFLREDS
MYCRAHFYKKFGTKCSSCNEGIVPDHVVRKA
SNHVYHVECFQCFICKRSLETGEEFYLIADDA
RLVCKDDYEQARDGGSGGHMGSGGGIGPLM
VQPATPHIDNTLGGPIDIQHF

↓

Align the sequences using Clustal Omega

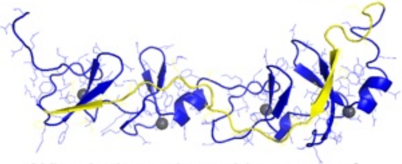
http://www.ebi.ac.uk/Tools/msa/clustalo/

↑

GSNNEEAICSLCDKKIRDRFVSKVNGRCYHSS
CLRCSTCKDELGATCFLREDSMYCRAHFYKK
FGTKCSSCNEGIVPDHVVRKASNHVYHVECF
QCFICKRSLETGEEFYLIADDARLVCKDDYEQ
ARDGGSGGHMGSGGGIGPLMVQPATPHIDNT
LGG
PIDIQHF

Amino acid sequence of protein from PDB or .cif file

Atomistic model from PDB file (filename.pdb)



What is the amino acid sequence?

↓

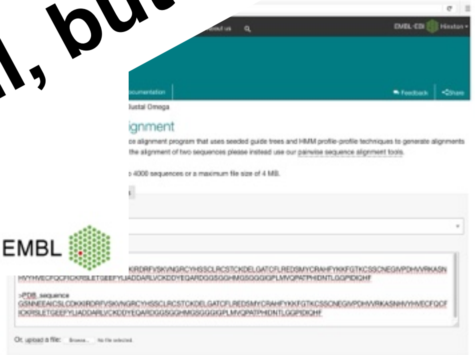
ATSAS tool: *pdb2seq*

↓

This will save the sequence in th 'filename.txt'

←

Good to know and is useful, but this is old school!

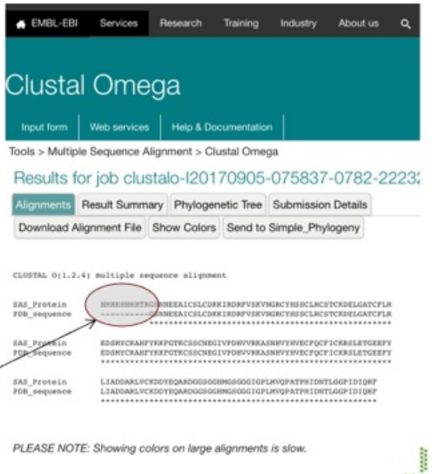


ignment

ce alignment program that uses seeded guide trees and HMM profile-profile techniques to generate alignments
the alignment of two sequences please instead use our pairwise sequence alignment tools.

4000 sequences or a maximum file size of 4 MB.

```
KIRDRFVSKVNGRCYHSSCLRCSTCKDELGATCFLREDSMYCRAHFYKKFGTKCSSCNEGIVPDHVVRKASN
HVYHVECFQCFICKRSLETGEEFYLIADDARLVCKDDYEQARDGGSGGHMGSGGGIGPLMVQPATPHIDNTLGGPIDIQHF

>PDB_sequence
GSNNEEAICSLCDKKIRDRFVSKVNGRCYHSSCLRCSTCKDELGATCFLREDSMYCRAHFYKKFGTKCSSCNEGIVPDHVVRKASNHVYHVECFQCF
ICKRSLETGEEFYLIADDARLVCKDDYEQARDGGSGGHMGSGGGIGPLMVQPATPHIDNTLGGPIDIQHF
```



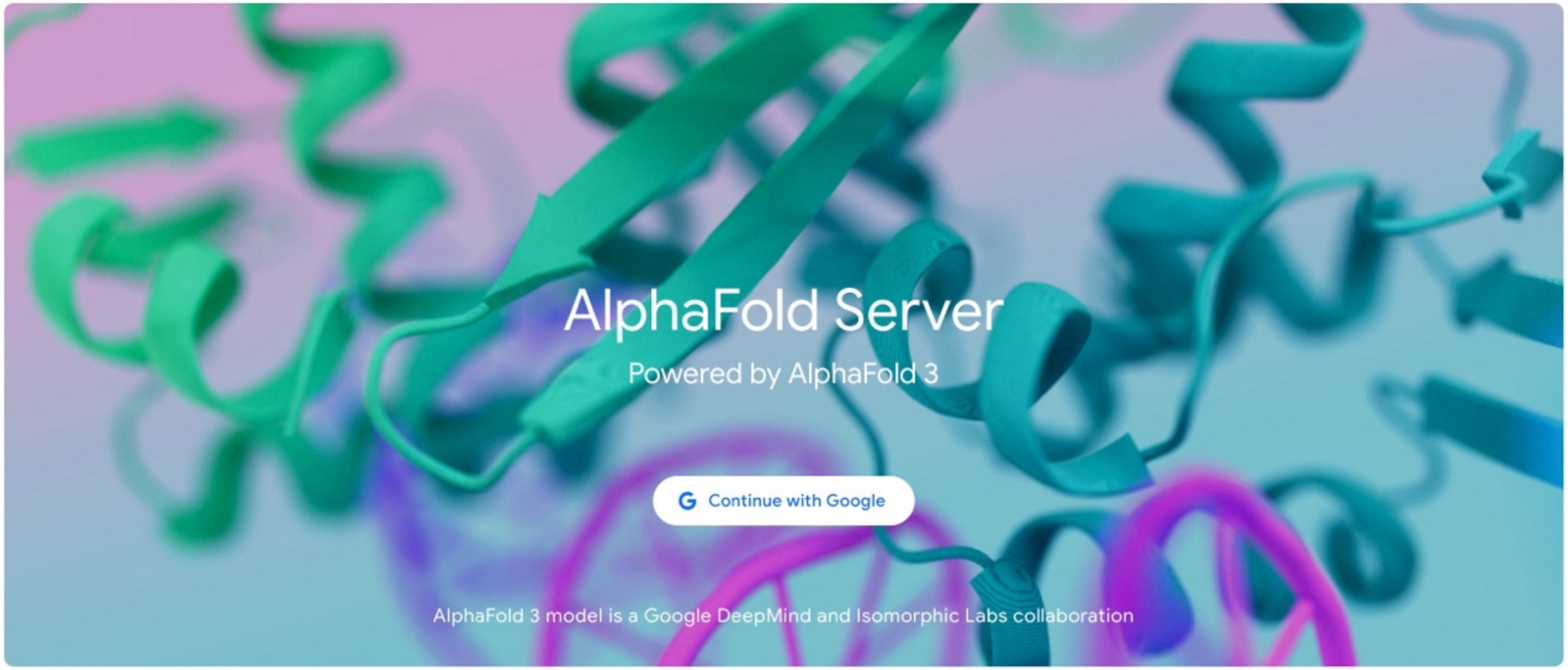
```
CLUSTAL O(1.2.4) multiple sequence alignment


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!

For proteins RNA, DNA, etc, just use AlphaFold3

https://golgi.sandbox.google.com/



AlphaFold Server

Powered by AlphaFold 3

AlphaFold 3 model is a Google DeepMind and Isomorphic Labs collaboration

But my structure *almost* fits the data, can I just wiggle it a bit? - SREFLEX

Employs normal modes pattern of motion on domain-partitioned structures.

Automated or manual domain partitioning possible.

Works with proteins!

Volume 18 Number 8 28 February 2016 Pages 5663–6330

PCCP

Physical Chemistry Chemical Physics
www.rsc.org/pccp

Themed issue: Exploring the conformational heterogeneity of biomolecules

ISSN 1463-9076

ROYAL SOCIETY OF CHEMISTRY

PAPER
Alejandro Panjkovich and Dmitri I. Svergun
Deciphering conformational transitions of proteins by small angle X-ray scattering and normal mode analysis

175 YEARS

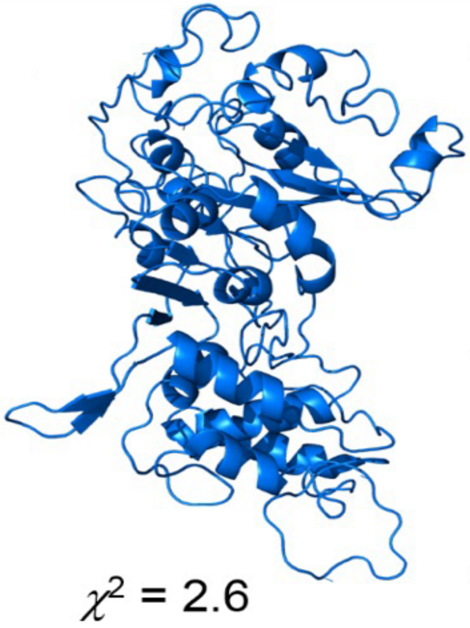
Deciphering conformational transitions of proteins by small angle X-ray scattering and normal mode analysis

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

Used for spatial refinement of models using small structural adjustments.

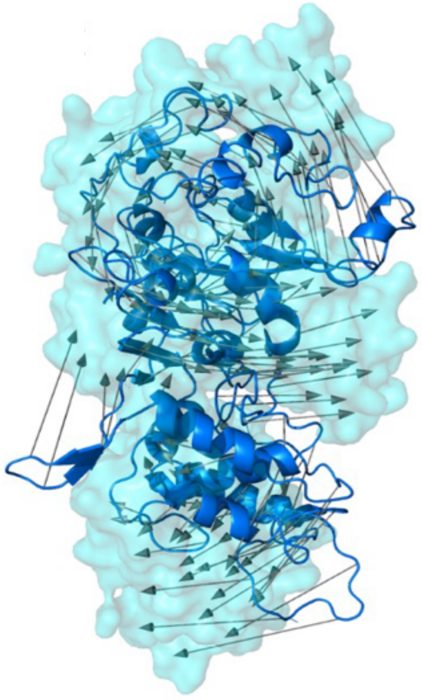
Great for assessing whether slight conformational movements are required to fit SAXS data (e.g., from crystal or AF-predicted structures).

Start structure

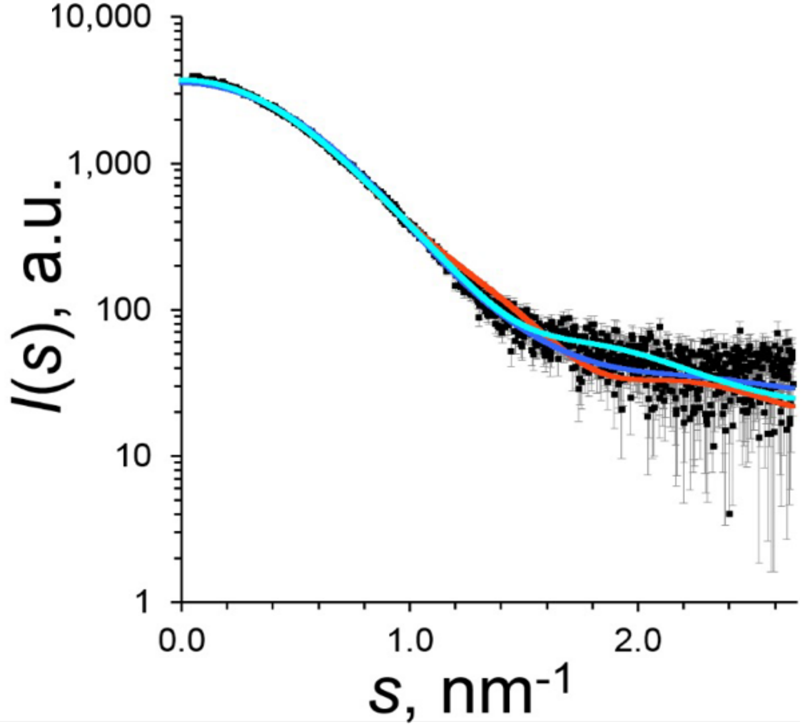


$\chi^2 = 2.6$

After normal mode

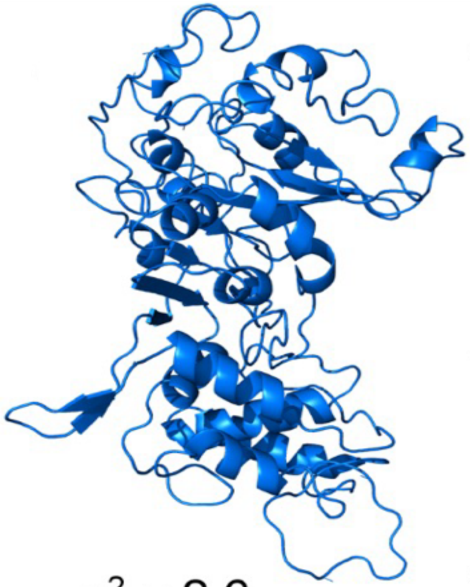


$\chi^2 = 1.2$



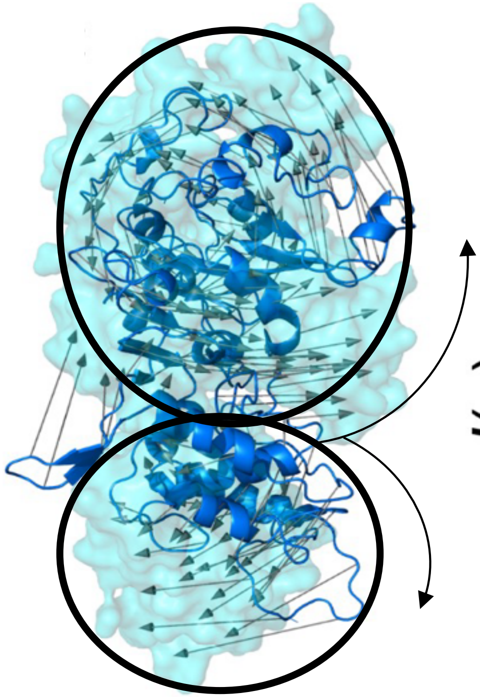
ATSAS online version applied additional CONCORD refinement

Start structure

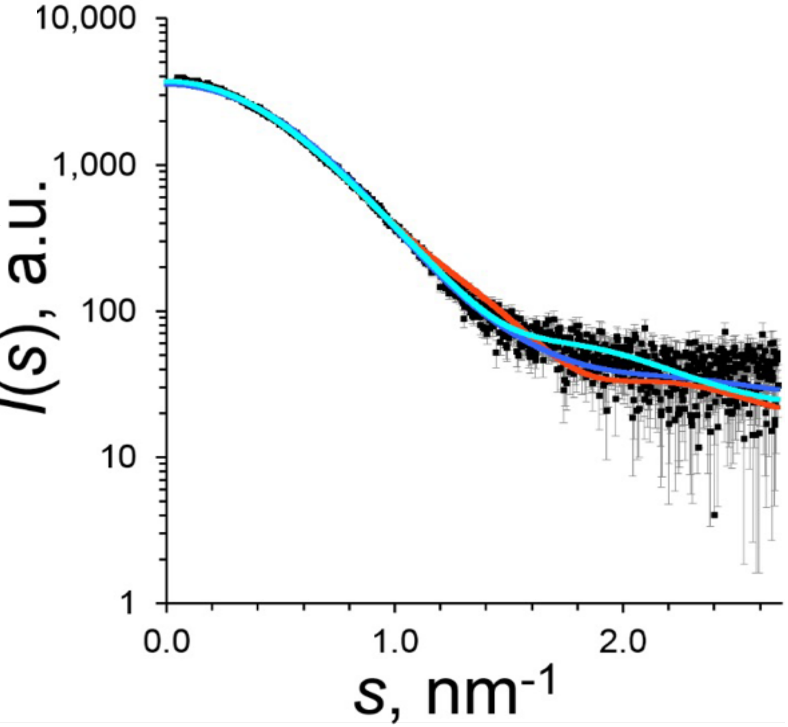


$\chi^2 = 2.6$

After normal mode

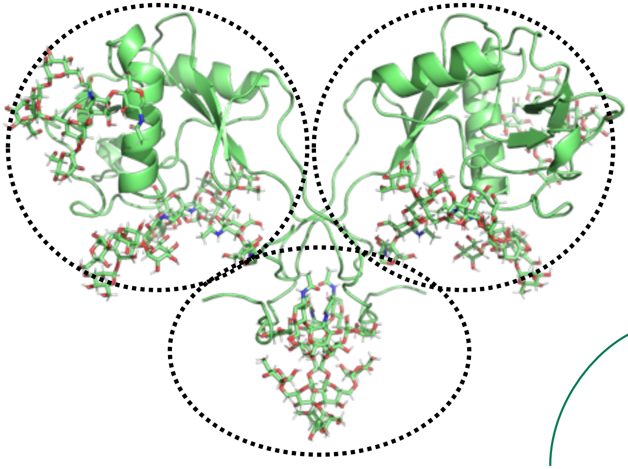


$\chi^2 = 1.2$



ATSAS online version applied additional CONCORD refinement

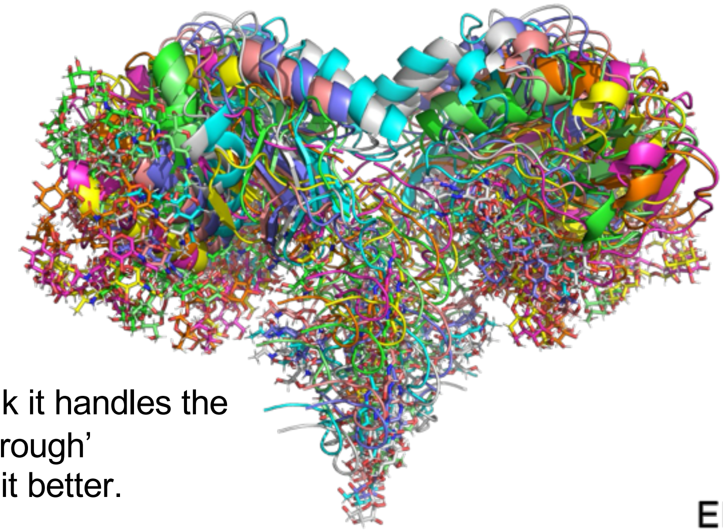
Combine SREFLEX with Multi-FoXS



Initial model: Manually define the rigid bodies in SREFLEX

Ask SREFLEX to output the normal mode models. Generate a normal mode pool. Do some basic scoring.

Get Multi-FoXS to fit the resulting NMA ensemble



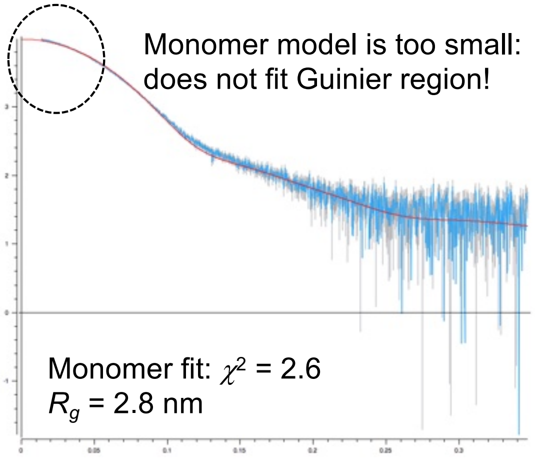
Why FoXS? Well…I think it handles the somewhat complicated ‘rough’ glycosylated surface a bit better.

Scattering from mixtures

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

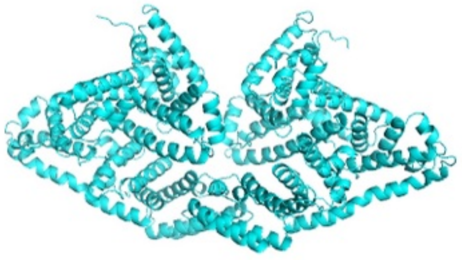
$$I(s)=\sum_{k} v_k I_k(s)$$

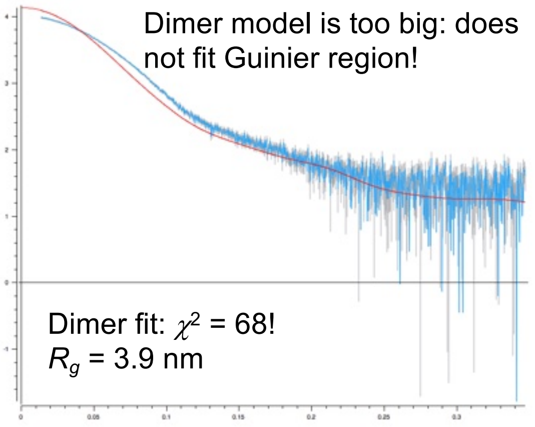
ATSAS program: OLIGOMER



Monomer model is too small: does not fit Guinier region!

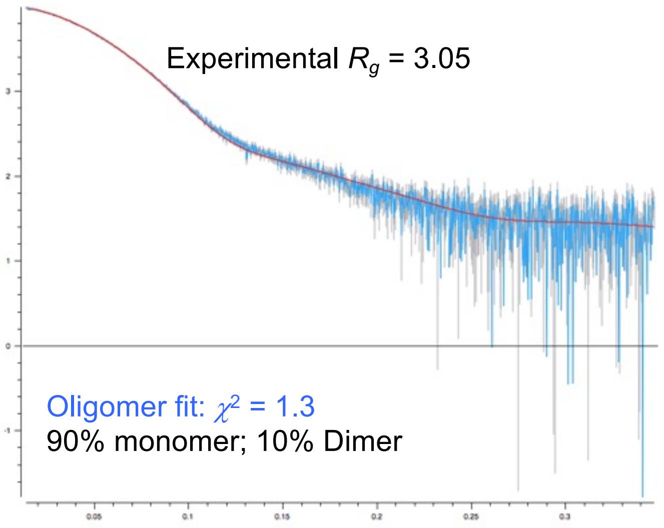
Monomer fit: χ^2 = 2.6
R_g = 2.8 nm





Dimer model is too big: does not fit Guinier region!

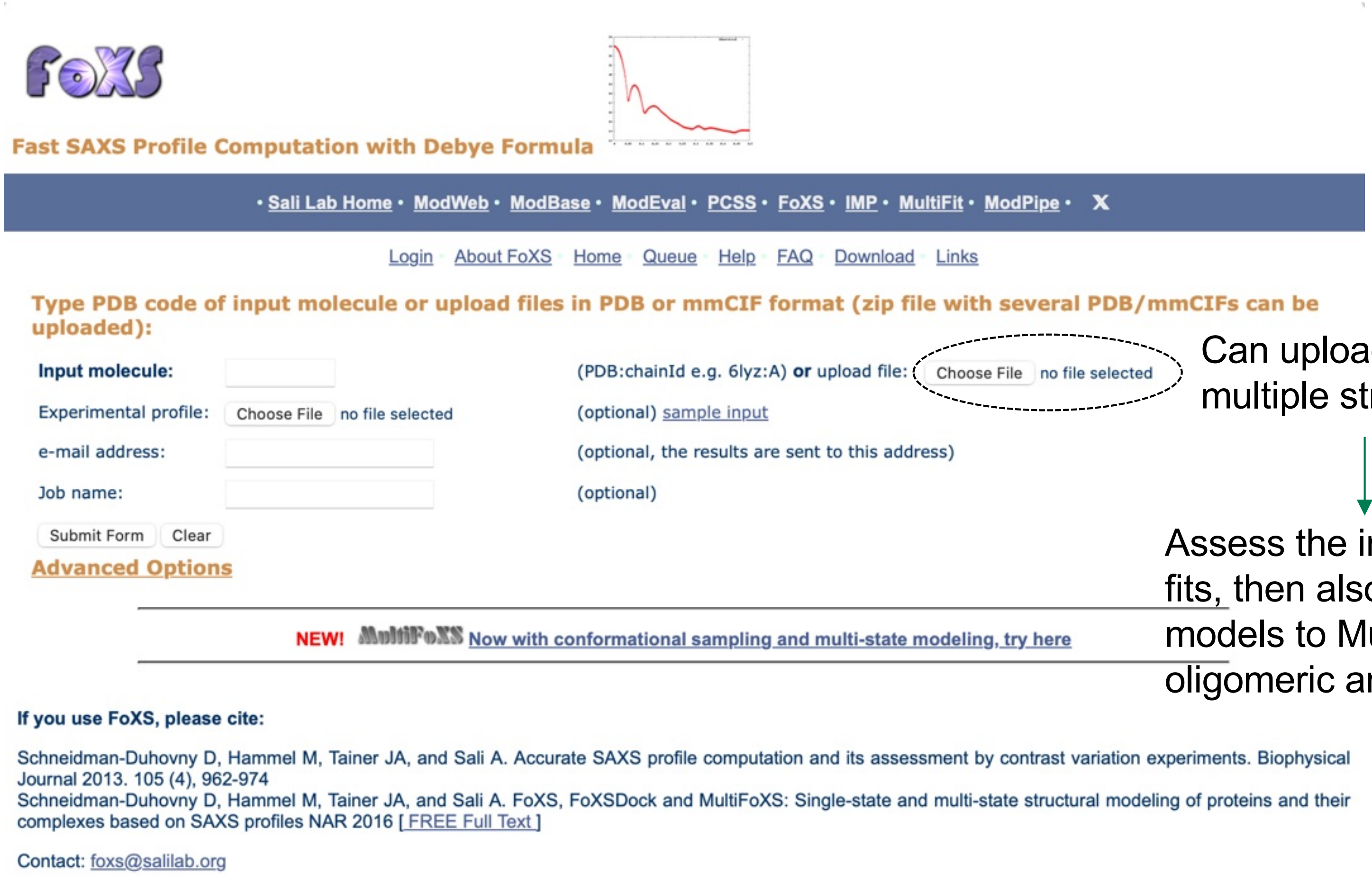
Dimer fit: χ^2 = 68!
R_g = 3.9 nm



Experimental R_g = 3.05

Oligomer fit: χ^2 = 1.3
90% monomer; 10% Dimer

foXS combined with Multi-foXS!



FoXS
Fast SAXS Profile Computation with Debye Formula
• Sali Lab Home • ModWeb • ModBase • ModEval • PCSS • FoXS • IMP • MultiFit • ModPipe • X
Login · About FoXS · Home · Queue · Help · FAQ · Download · Links

Type PDB code of input molecule or upload files in PDB or mmCIF format (zip file with several PDB/mmCIFs can be uploaded):

Input molecule:		(PDB:chainId e.g. 6lyz:A) **or** upload file: Choose File no file selected
Experimental profile:	Choose File no file selected	(optional) sample input
e-mail address:		(optional, the results are sent to this address)
Job name:		(optional)

Submit Form
Clear
Advanced Options

NEW! MultiFoXS Now with conformational sampling and multi-state modeling, try here

If you use FoXS, please cite:

Schneidman-Duhovny D, Hammel M, Tainer JA, and Sali A. Accurate SAXS profile computation and its assessment by contrast variation experiments. Biophysical Journal 2013. 105 (4), 962-974
Schneidman-Duhovny D, Hammel M, Tainer JA, and Sali A. FoXS, FoXSDock and MultiFoXS: Single-state and multi-state structural modeling of proteins and their complexes based on SAXS profiles NAR 2016 [FREE Full Text]

Contact: foxs@salilab.org

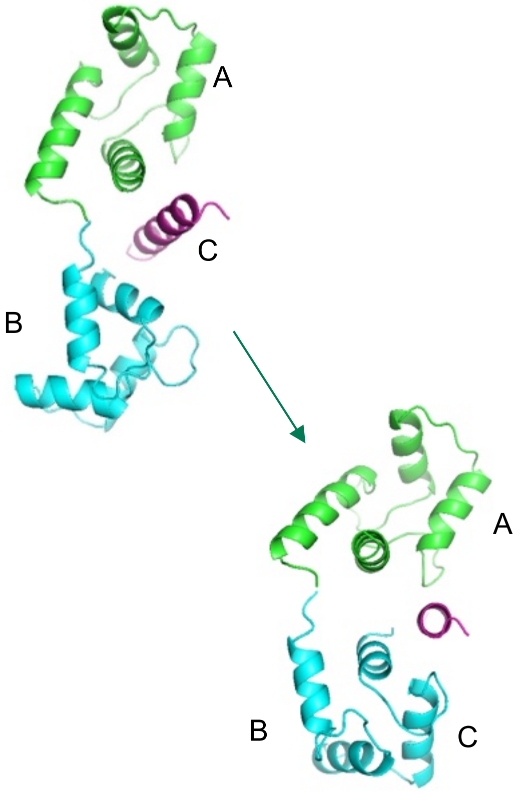
Can upload a zip file with multiple structures

Assess the individual model fits, then also pass the models to Multi-foXS for oligomeric analysis

EMBL

Structure still does not fit – try some rigid body modelling

- 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.

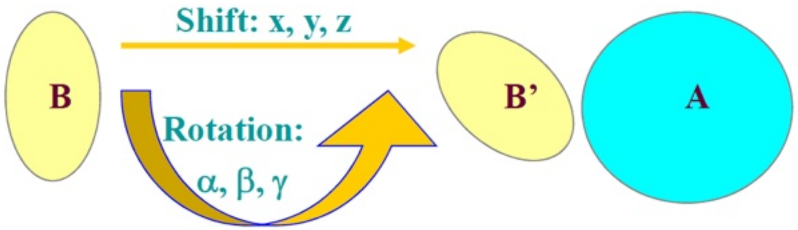


Structure does not fit – try some rigid body modelling

- The structures o
are known.

- Arbitrary comple
the subunits.

- This operation d
Cartesian shifts.



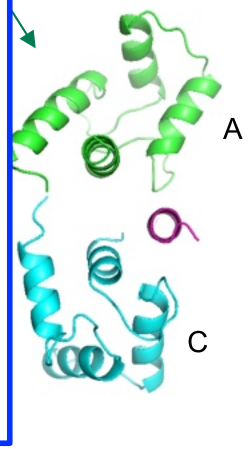
The partial amplitudes of a rotated and displaced subunit are expressed *via* the initial amplitudes, three Euler rotation angles and three Cartesian shifts):

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

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

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

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



EMBL

The target function:

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

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.

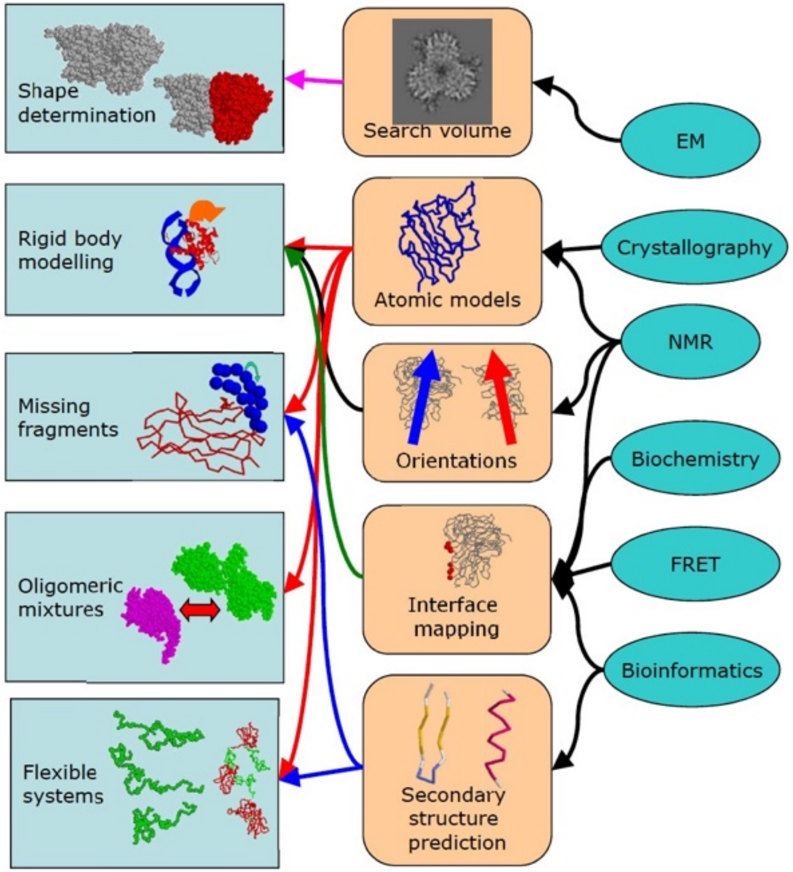
Otherwise a Monte-Carlo based technique (e.g. simulated annealing) is employed to perform the minimization of E({X}).

A note on χ^2

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

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

Incorporate information from EM, crystallography, NMR, biochemistry (e.g., cross linking, Mass-spec), FRET and bioinformatics…and of course for proteins…Alphafold!



+

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).

EMBL

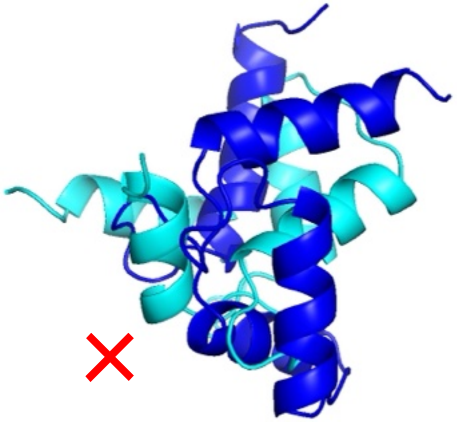
SASREF (for SAXS), SASREFcv (for SAXS and SANS)

- 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

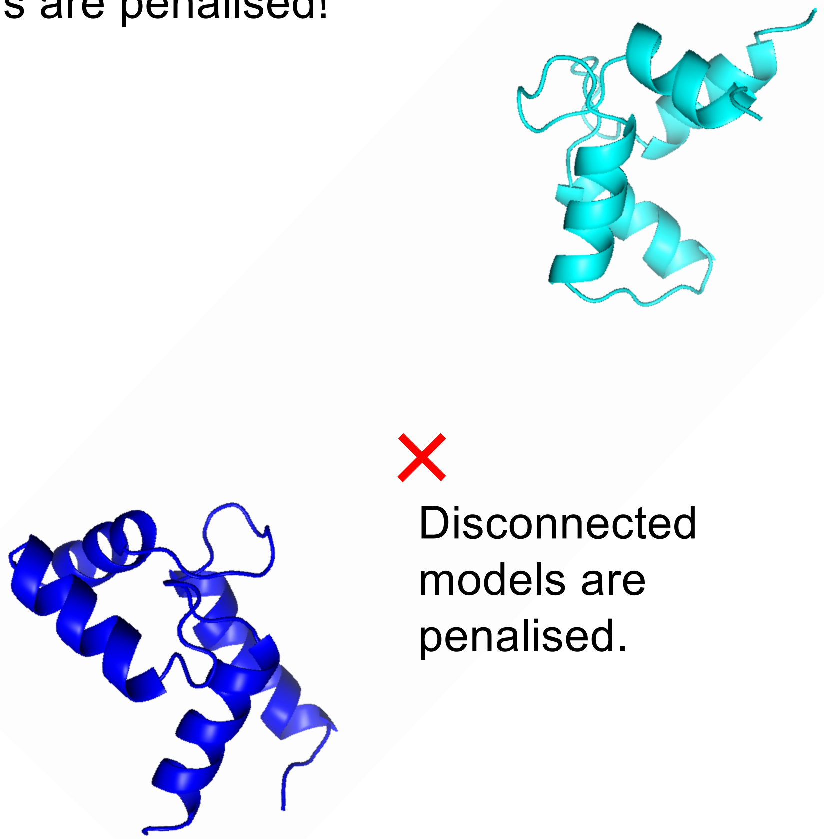
SASREF restraints

Subunit clashes or disconnected models are penalised!



✗

Inter penetrating subunits are penalised.



✗

Disconnected models are penalised.

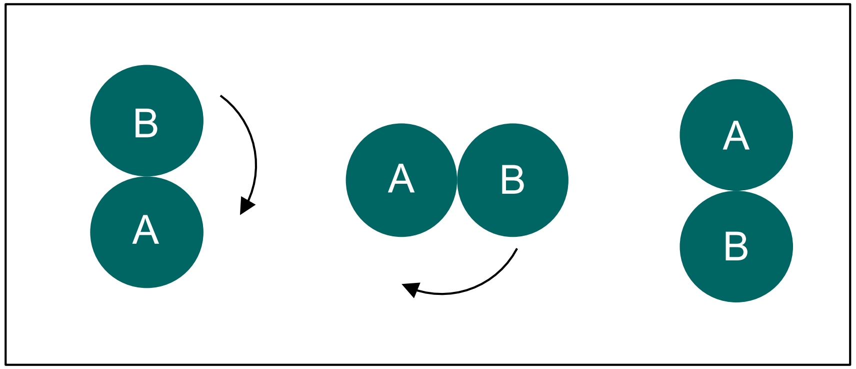
SASREF inputs

For SAXS:

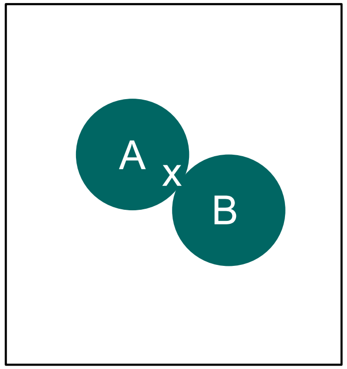
- Rigid body starting models – centred to an origin. Protein, DNA, RNA, etc.
- Scattering amplitude files of each rigid-body model calculated using **CRYSOL**.
- Contacts file (optional).
- Symmetry information.

Alphafold3 of course be used if no contact information is available.

Contact information can be exceptionally useful!



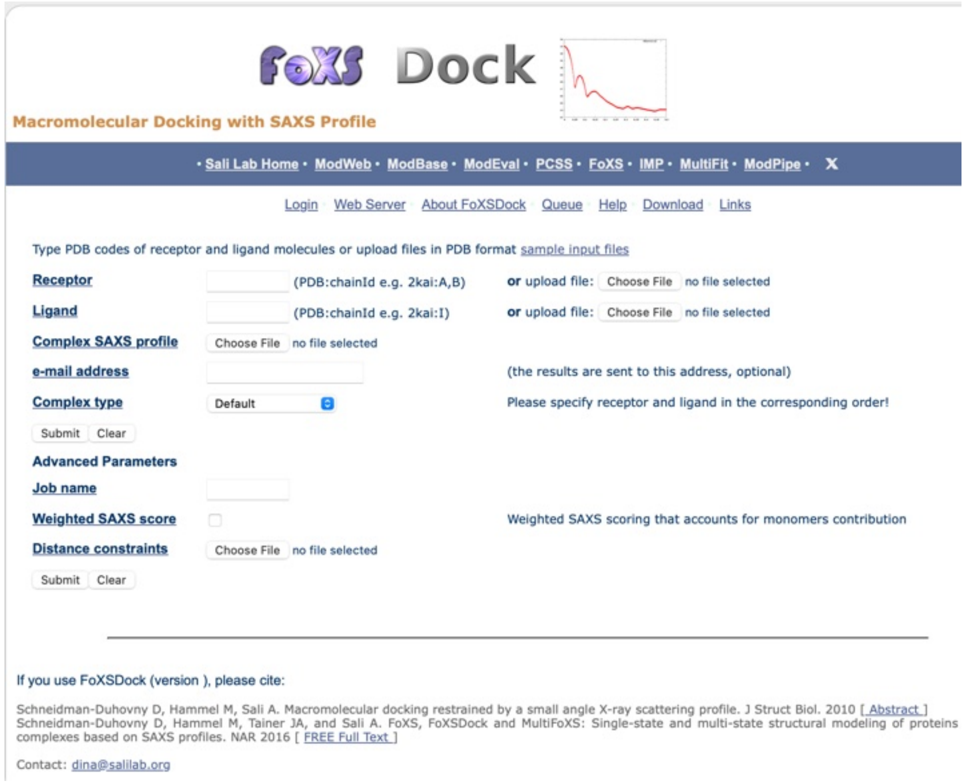
No contact information



Single contact

EMBL

Other docking methods



FoXS Dock

Macromolecular Docking with SAXS Profile

• Sali Lab Home • ModWeb • ModBase • ModEval • PCSS • FoXS • IMP • MultiFit • ModPipe • X
Login · Web Server · About FoXSDock · Queue · Help · Download · Links

Type PDB codes of receptor and ligand molecules or upload files in PDB format sample input files

Receptor	(PDB:chainId e.g. 2kai:A,B)	or upload file: Choose File no file selected
Ligand	(PDB:chainId e.g. 2kai:I)	or upload file: Choose File no file selected
Complex SAXS profile	Choose File no file selected	
e-mail address		(the results are sent to this address, optional)
Complex type	Default	Please specify receptor and ligand in the corresponding order!

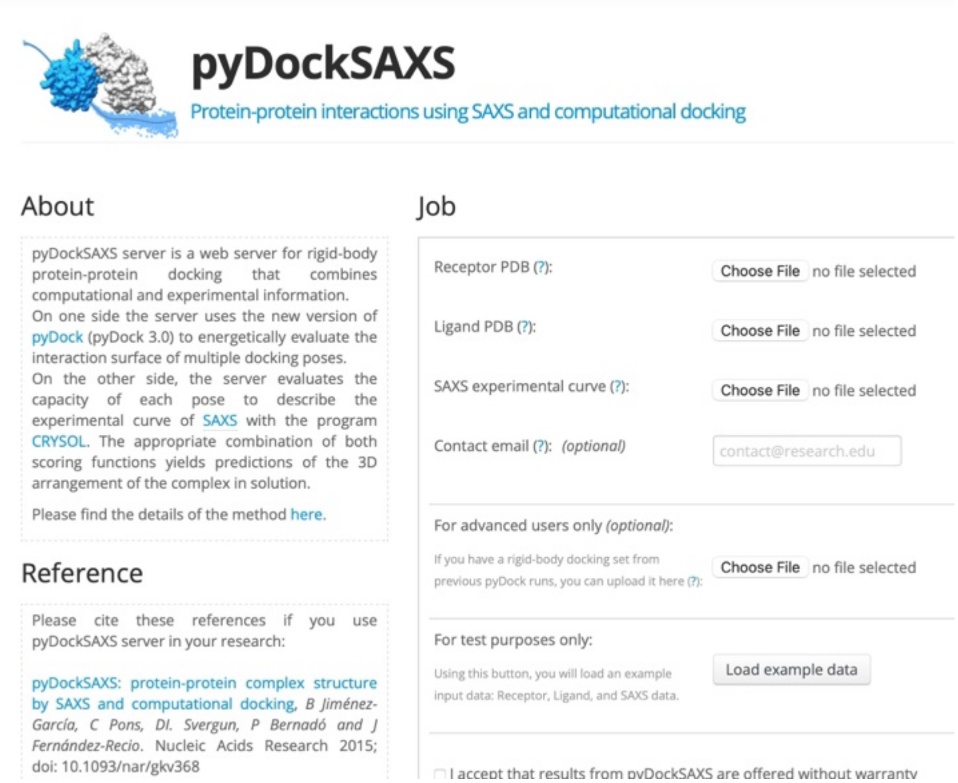
Advanced Parameters

Job name		
Weighted SAXS score	□	Weighted SAXS scoring that accounts for monomers contribution
Distance constraints	Choose File no file selected	

If you use FoXSDock (version), please cite:

Schneidman-Duhovny D, Hammel M, Sali A. Macromolecular docking restrained by a small angle X-ray scattering profile. J Struct Biol. 2010 [Abstract]
Schneidman-Duhovny D, Hammel M, Tainer JA, and Sali A. FoXS, FoXSDock and MultiFoXS: Single-state and multi-state structural modeling of proteins complexes based on SAXS profiles. NAR 2016 [FREE Full Text]

Contact: dina@salilab.org



pyDockSAXS

Protein-protein interactions using SAXS and computational docking

About

pyDockSAXS server is a web server for rigid-body protein-protein docking that combines computational and experimental information. On one side the server uses the new version of pyDock (pyDock 3.0) to energetically evaluate the interaction surface of multiple docking poses. On the other side, the server evaluates the capacity of each pose to describe the experimental curve of SAXS with the program CRYSOL. The appropriate combination of both scoring functions yields predictions of the 3D arrangement of the complex in solution.

Please find the details of the method here.

Reference

Please cite these references if you use pyDockSAXS server in your research:

pyDockSAXS: protein-protein complex structure by SAXS and computational docking, *B Jiménez-García, C Pons, DI. Svergun, P Bernadó and J Fernández-Recio*. Nucleic Acids Research 2015; doi: 10.1093/nar/gkv368

Job

Receptor PDB (?):	Choose File no file selected
Ligand PDB (?):	Choose File no file selected
SAXS experimental curve (?):	Choose File no file selected
Contact email (?): *(optional)*	contact@research.edu

For advanced users only *(optional)*:

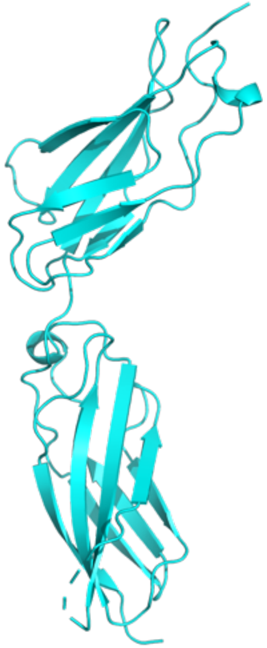
If you have a rigid-body docking set from previous pyDock runs, you can upload it here (?): Choose File no file selected

For test purposes only:

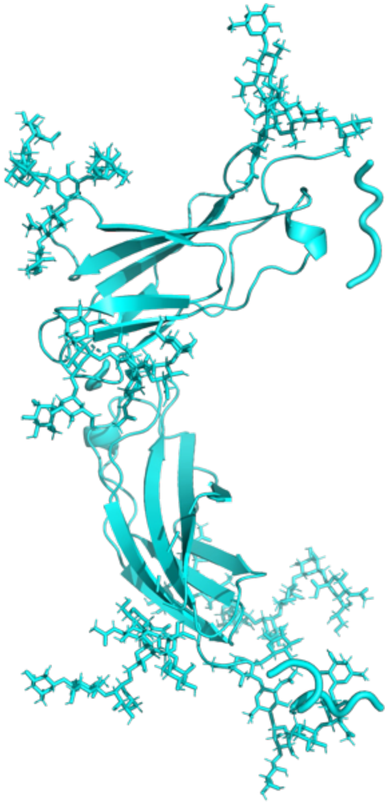
Using this button, you will load an example input data: Receptor, Ligand, and SAXS data. Load example data

□ I accept that results from pyDockSAXS are offered without warranty

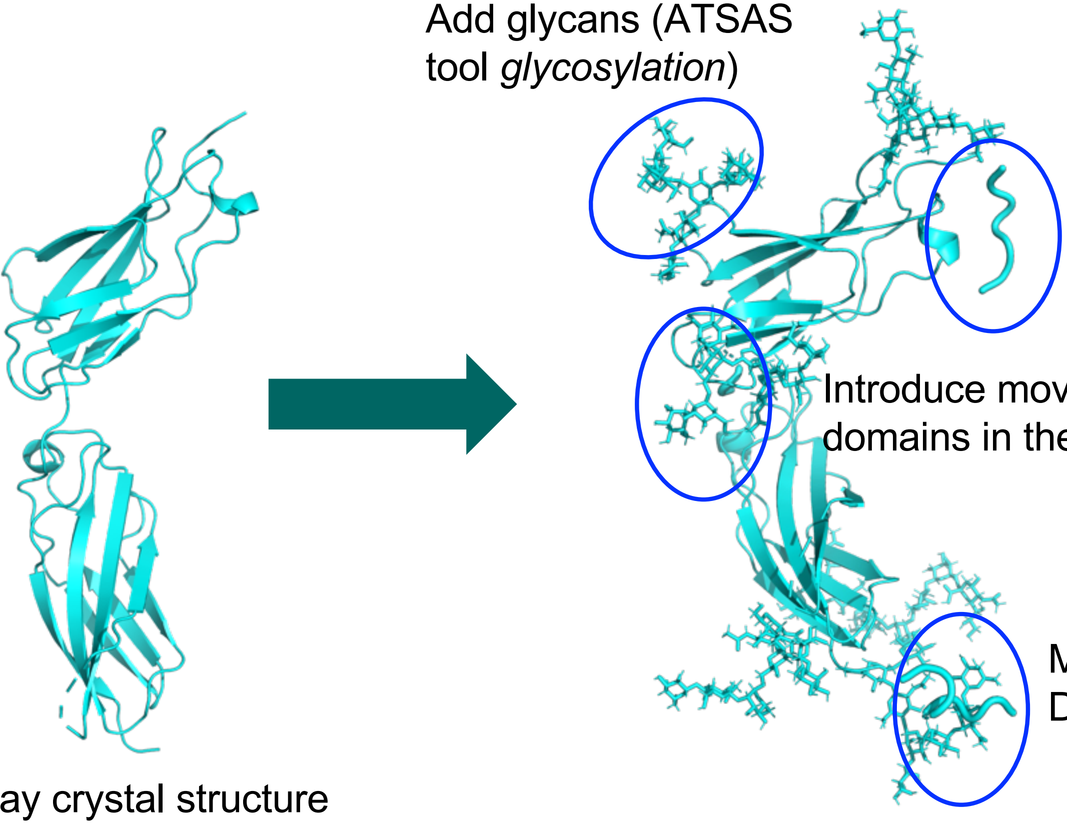
More complicated examples – Dealing with missing stuff, linkers, etc.



X-ray crystal structure



Missing stuff, linkers, etc



X-ray crystal structure

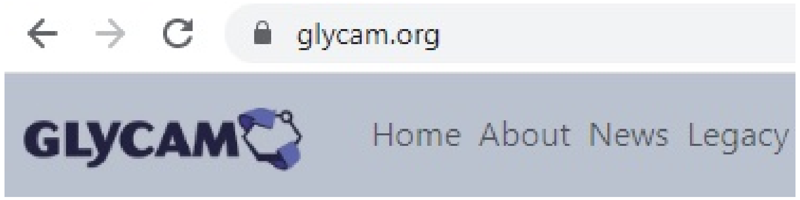
Add glycans (ATSAS tool *glycosylation*)

Missing N-terminus – Dummy residue addition

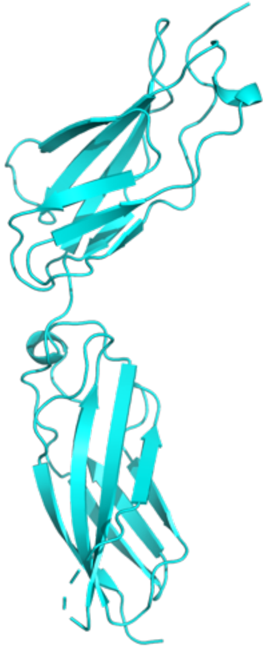
Introduce movement between domains in the connecting linker

Missing C-terminus – Dummy residue addition

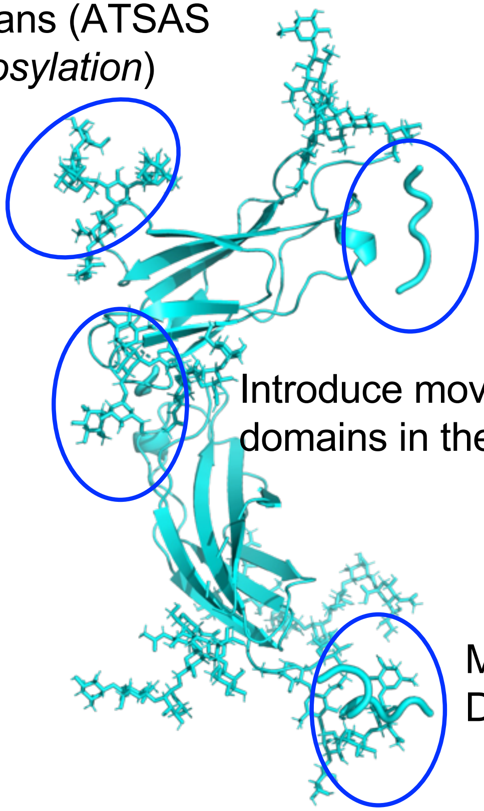
Missing stuff, linkers, etc



Add glycans (ATSAS tool *glycosylation*)



X-ray crystal structure



Missing N-terminus – Dummy residue addition

Introduce movement between domains in the connecting linker

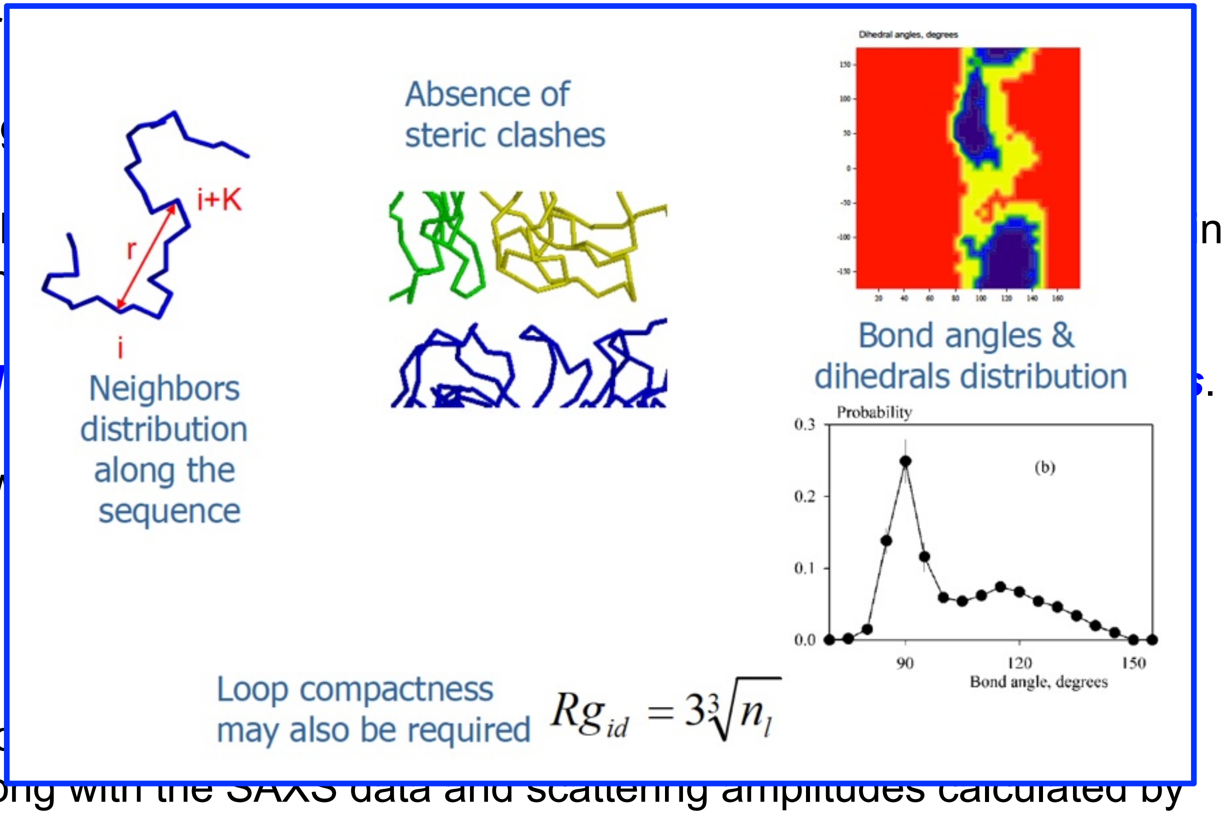
Missing C-terminus – Dummy residue addition

BUNCH – will optimize domain and dummy-amino acid positions

- For SAXS only!
- Single residue **polypeptide chain only, i.e., ‘protein domains’!**
- With or without symmetry – **multiple curves allowed**, e.g., domain truncation mutants.
- ***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 domain and dummy-amino acid positions

- For
- Sin
- Witl n trun
- ***M*** ***s***.
- A tw
- Rec (along with the SAXS data and scattering amplitudes calculated by CRYSC


i+K
r
i
Neighbors distribution along the sequence
Absence of steric clashes
Dihedral angles, degrees
150
100
50
0
-50
-100
-150
20
40
60
80
100
120
140
160
Bond angles & dihedrals distribution
Probability
0.3
0.2
0.1
0.0
(b)
90
120
150
Bond angle, degrees
Loop compactness may also be required
$Rg_{id} = 3\sqrt[3]{n_l}$

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

ATSAS - CORAL

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

- BUNCH – is good form modelling single polypeptide chains with missing fragments, or linkers connecting modules/domains against SAXS data

  - ***CORAL*** combines both concepts into one!

- CORAL – Protein, DNA, RNA, glycosylated systems and complexes...all are possible!

- Known subunit interfaces can be preserved by grouping subunits together.

- 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!).

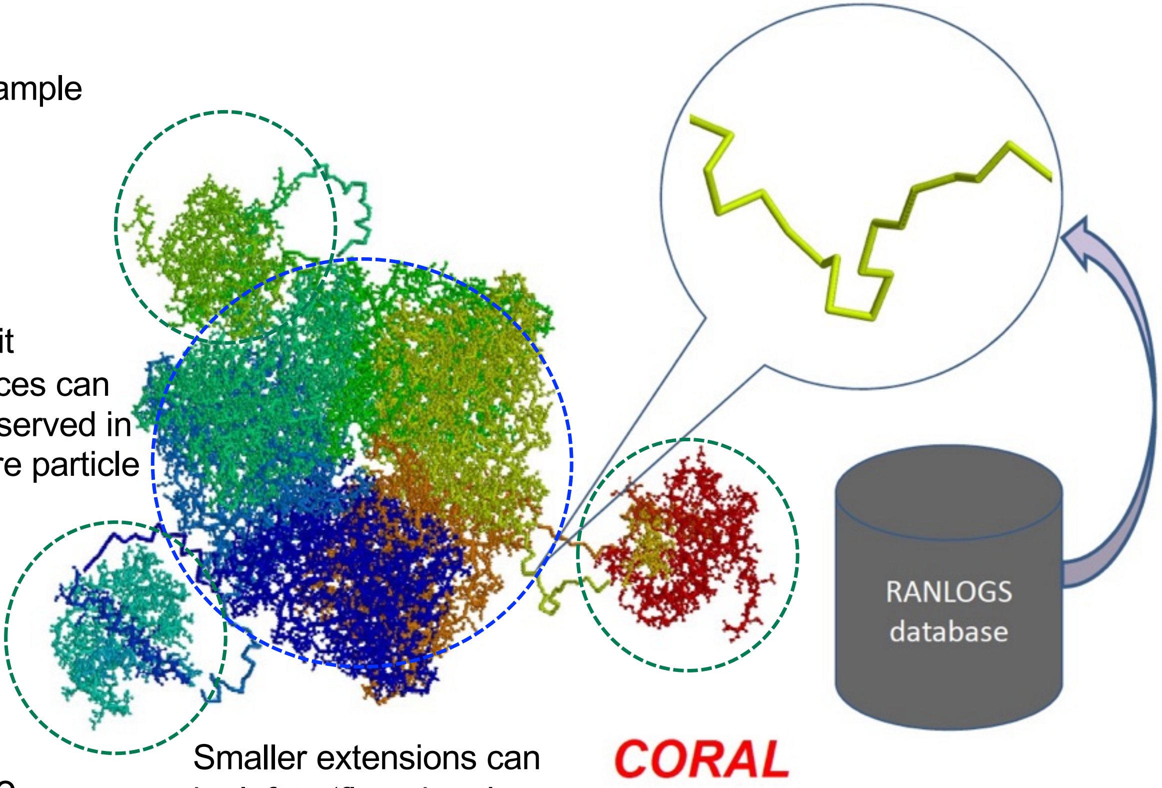
- ...***SAXS only!!***

EMBL

- CORAL requires the SAXS data, domain/subunit atomic coordinate files along with the scattering amplitudes calculated by CRYSOL. A contact file is also possible!
- CORAL requires and additional .con file telling the program where to generate the linkers for each subunit:

```
NTER 6
KD_monomer1_1coral.pdb
LINK 10
KD_monomer2_1coral.pdb
CTER 10
NTER 4
KD_monomer2_coral.pdb
CTER 10
DNA.pdb
```

For example



Subunit interfaces can be preserved in the core particle

Smaller extensions can be left to 'flop about' without obeying symmetry

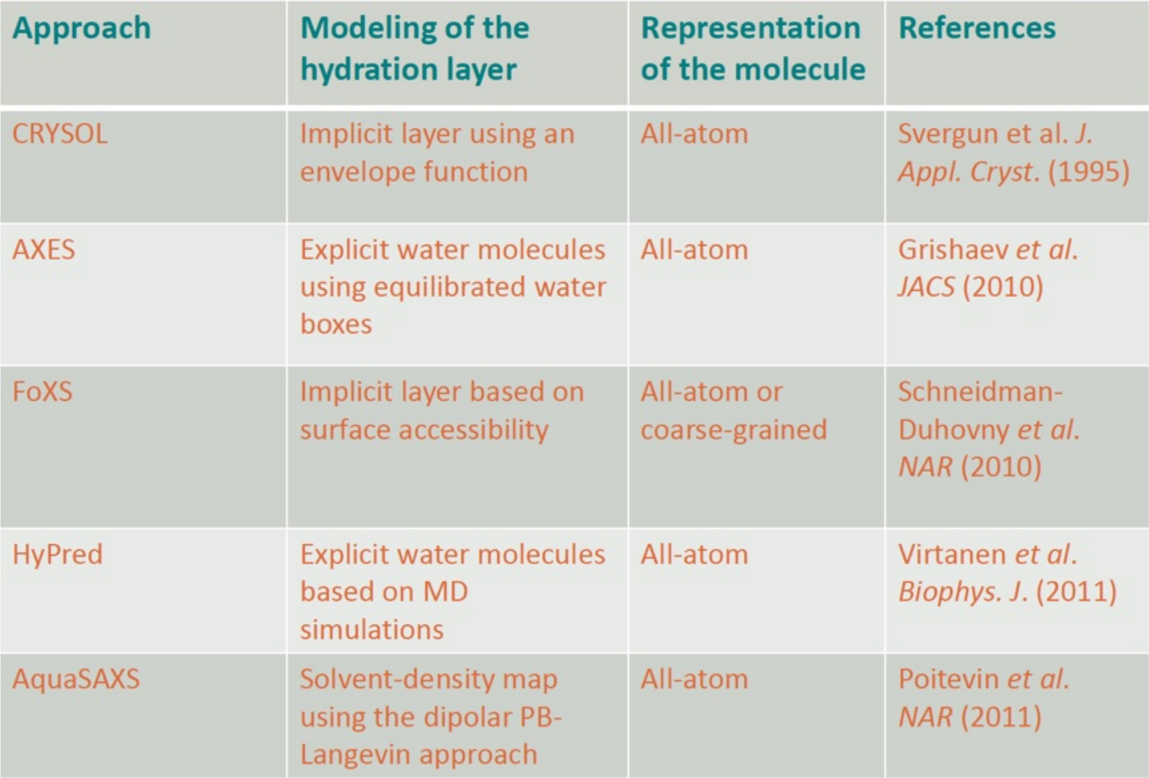
- At some point the program will ask:

```
Pair of domains to group
```

Where you can specify to preserve the spatial orientation between subunits. In the above, e.g., 3,4 and also 4,3 to preserve KD_monomer2_coral.pdb with DNA.pdb

EMBL

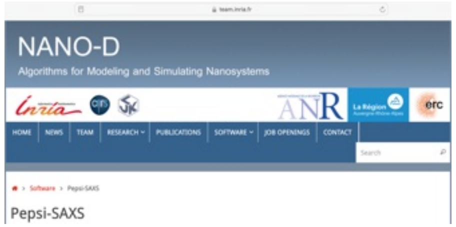
Always check the final model fits using CRYSOL

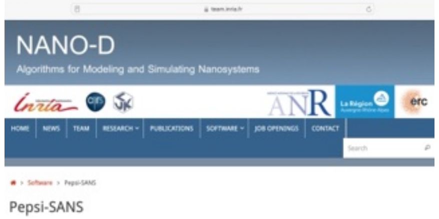


Approach	Modeling of the hydration layer	Representation of the molecule	References
CRYSOL	Implicit layer using an envelope function	All-atom	Svergun et al. *J. Appl. Cryst.* (1995)
AXES	Explicit water molecules using equilibrated water boxes	All-atom	Grishaev *et al. JACS* (2010)
FoXS	Implicit layer 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)

WAXS in Solvent (WAXSiS) computes small- and wide-angle X-ray scattering curves based on explicit-solvent all-atom molecular dynamics simulations.

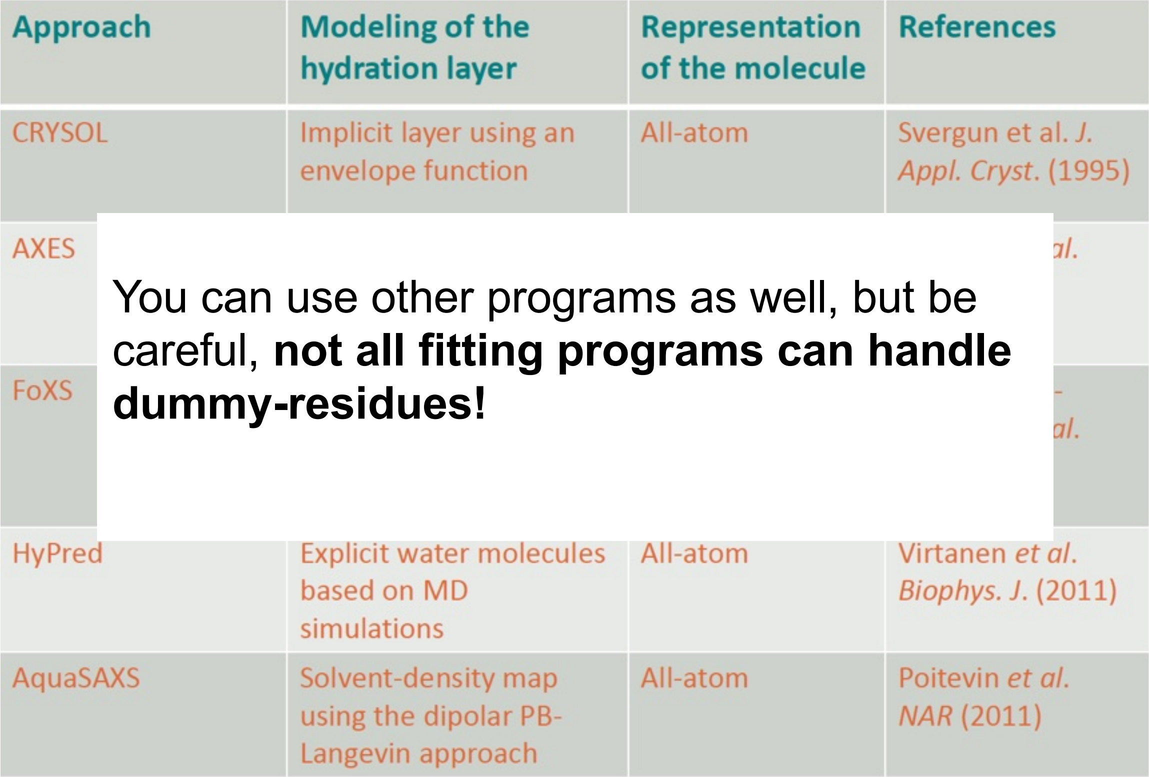
WAXIS





PEPSI-SAXS and PEPSI-SANS

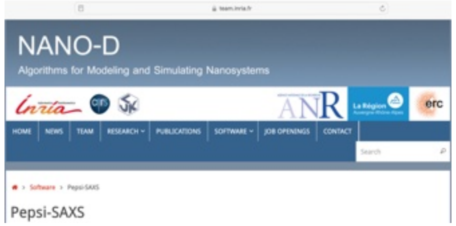
Always check the final model fits using CRYSOL

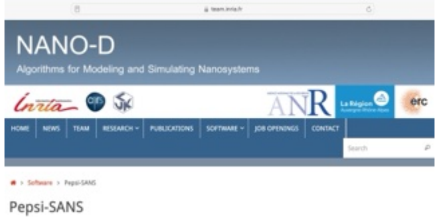


Approach	Modeling of the hydration layer	Representation of the molecule	References
CRYSOL	Implicit layer using an envelope function	All-atom	Svergun et al. *J. Appl. Cryst.* (1995)
AXES			*al.*
FoXS			- *al.*
HyPred	Explicit water molecules based on MD simulations	All-atom	Virtanen *et al. Biophys. J.*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

You can use other programs as well, but be careful, **not all fitting programs can handle dummy-residues!**

WAXIS





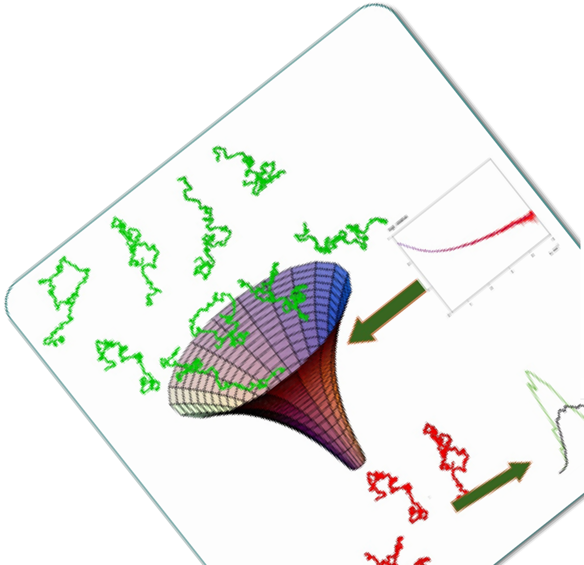
PEPSI-SAXS and PEPSI-SANS

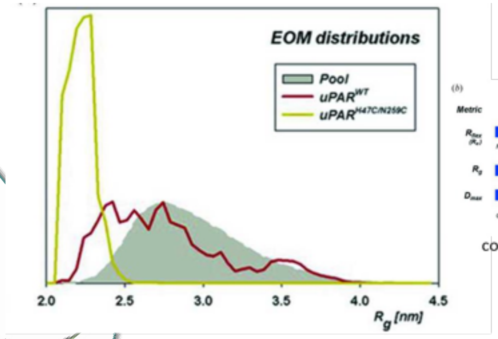
If there is one message, I want to get across today – always consider ambiguity!

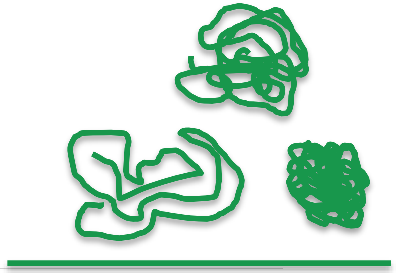
- You ***must*** run your selected rigid body modelling routines at least 10 times and check for the spatial consistency of the models (spatial alignment using supcomb).
- *At the end of a BUNCH, SASREF or CORAL run check the fits with CRYSOL!*
- *Use Correlation Map to assess fits if you are unsure about your experimental errors!*
- *Error normalized residual plots are a great tool to visually assess systematic differences between modelled and experimental scattering intensities.*
- *…also apply common sense.*
- I usually do 20 modelling runs, check the individual model fits with CRYSOL (using 30 harmonics, minimum), then order the CRYSOL fits in terms of χ^2 and CorMap *P*, then spatially align all models that fit the data to assess consistency.

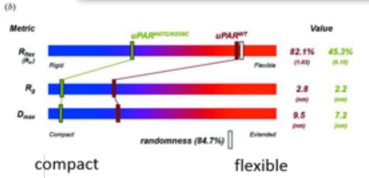
My structure is moving all over the shop!

Ensemble optimization method (EOM)

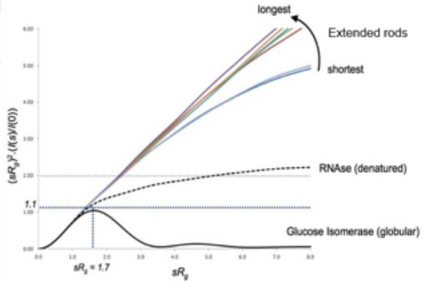






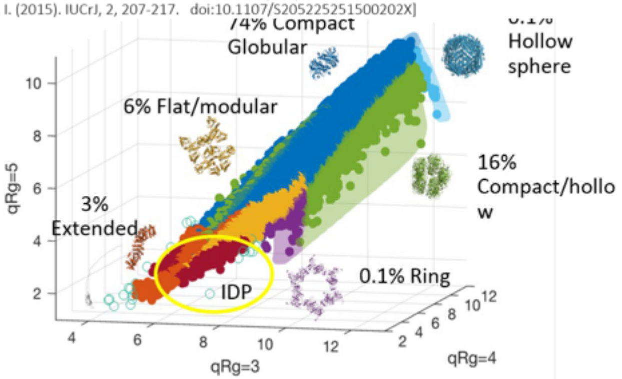


Characterization of the flexibility of uPARWT and the mutated uPARH47C-N259C using EOM 2.0. (a) Size distributions (Rg) of uPARWT and uPARH47C-N259C, providing only a qualitative assessment through direct comparison of the distributions of the selected ensembles and the pool. (b) The metrics Rflex and 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.



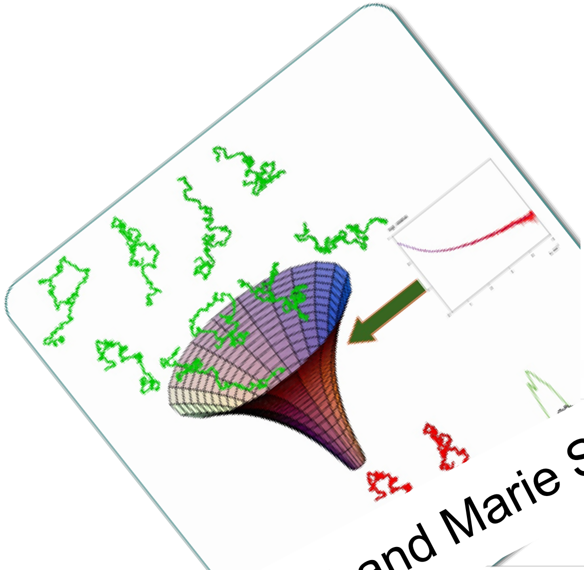
Franke, Jeffries & Svergun (2018) *Biophys J.* 114: 2485–2492

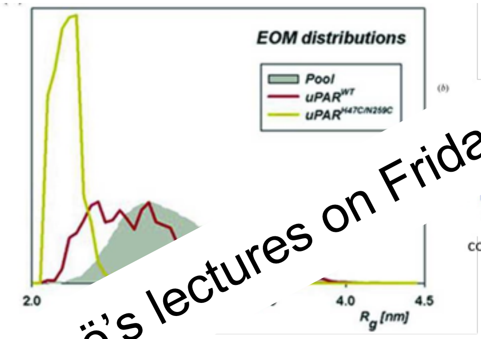
EMBL

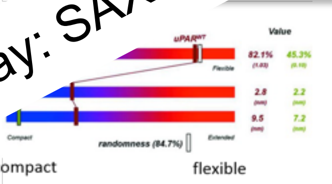
My structure is moving all over the shop!

Ensemble optimization method (EOM)

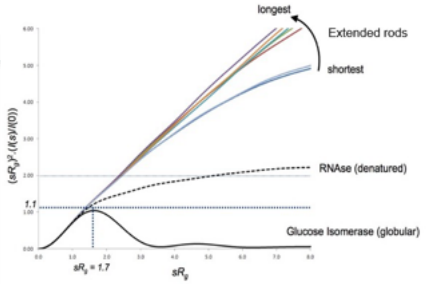
Wait for Pau Bernadó and Marie Skepö's lectures on Friday: SAXS and flexibility



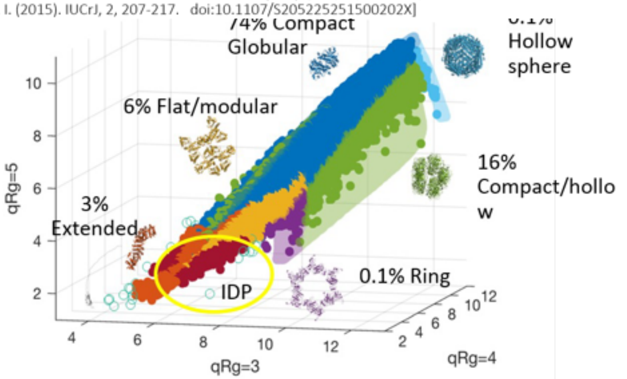




…of the flexibility of uPARWT and the mutated uPARH47C-N259C using EOM 2.0. (a) Size …ns (Rg) of uPARWT and uPARH47C-N259C, providing only a qualitative assessment through …comparison of the distributions of the selected ensembles and the pool. (b) The metrics Rflex and Ro enable characterization of 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]



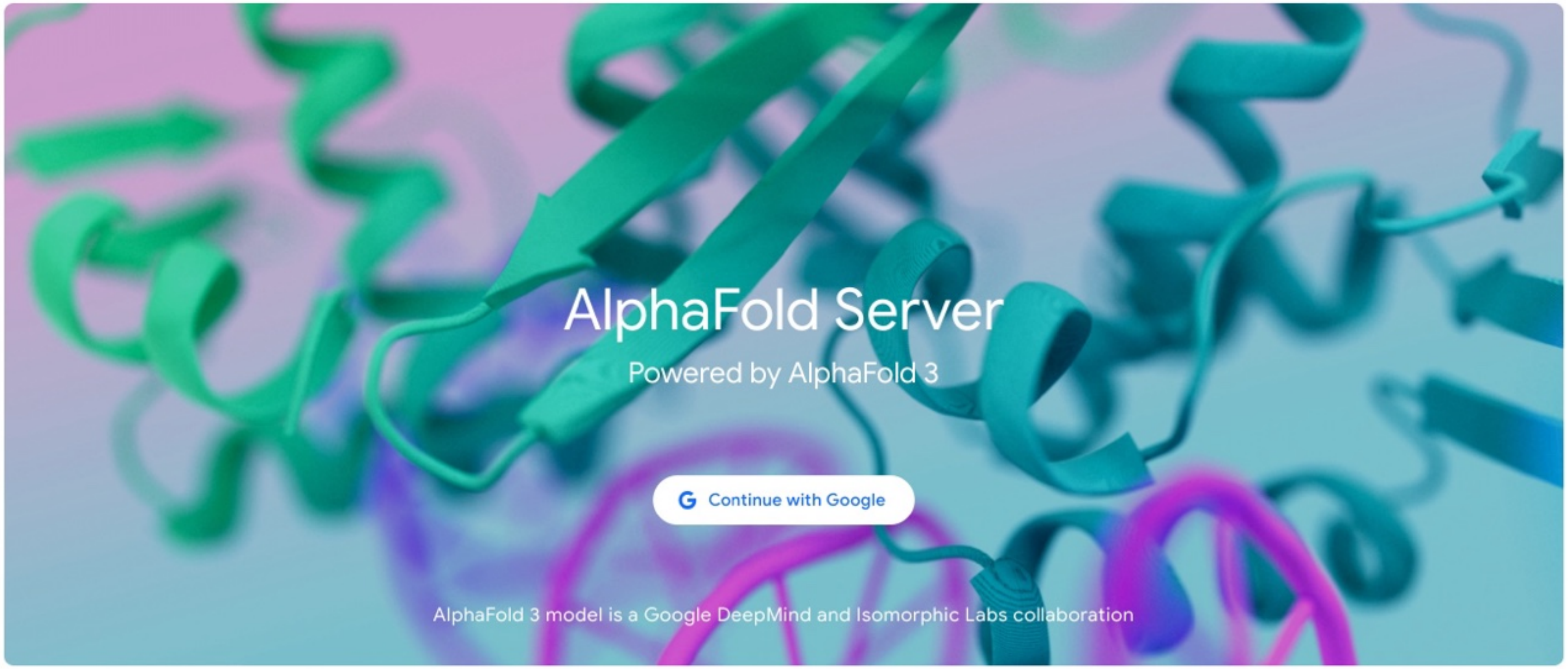
Receveur-Bréchot & Durand (2012) *Current Protein and Peptide Science*, 13, 55-75.
Durand D, Vivès C, Cannella D, Pérez J, Pebay-Peyroula E, Vachette P, Fieschi F. (2010) *J Struct Biol.* 169: 45-53.



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

EMBL

https://golgi.sandbox.google.com/



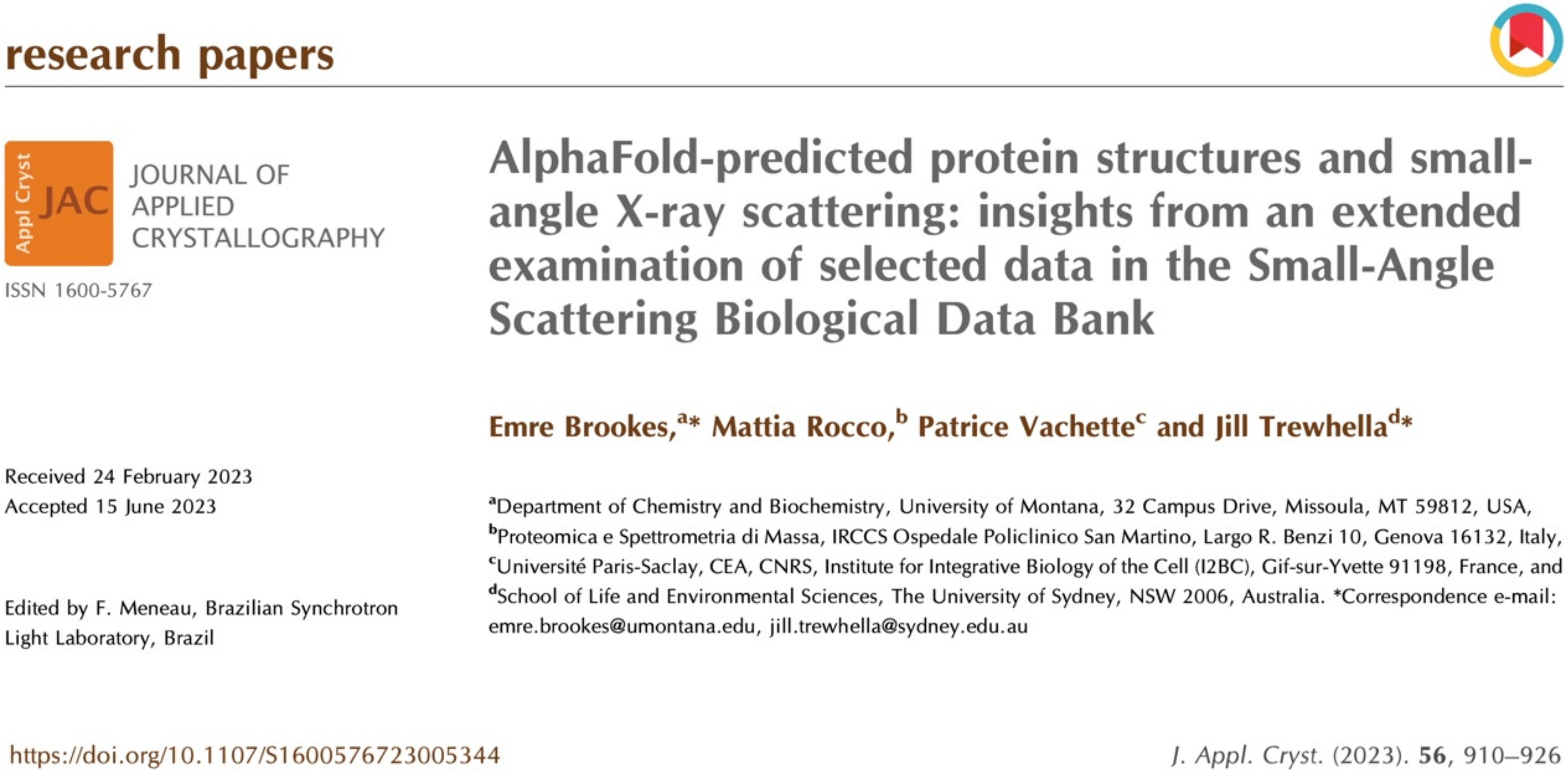
AlphaFold Server

Powered by AlphaFold 3

AlphaFold 3 model is a Google DeepMind and Isomorphic Labs collaboration

EMBL

Combine with Alphafold!



research papers

JOURNAL OF APPLIED CRYSTALLOGRAPHY

Appl Cryst
JAC

ISSN 1600-5767

AlphaFold-predicted protein structures and small-angle X-ray scattering: insights from an extended examination of selected data in the Small-Angle Scattering Biological Data Bank

Emre Brookes,[a]* Mattia Rocco,[b] Patrice Vachette[c] and Jill Trewhella[d]*

[a]Department of Chemistry and Biochemistry, University of Montana, 32 Campus Drive, Missoula, MT 59812, USA, [b]Proteomica e Spettrometria di Massa, IRCCS Ospedale Policlinico San Martino, Largo R. Benzi 10, Genova 16132, Italy, [c]Université Paris-Saclay, CEA, CNRS, Institute for Integrative Biology of the Cell (I2BC), Gif-sur-Yvette 91198, France, and [d]School of Life and Environmental Sciences, The University of Sydney, NSW 2006, Australia. *Correspondence e-mail: emre.brookes@umontana.edu, jill.trewhella@sydney.edu.au

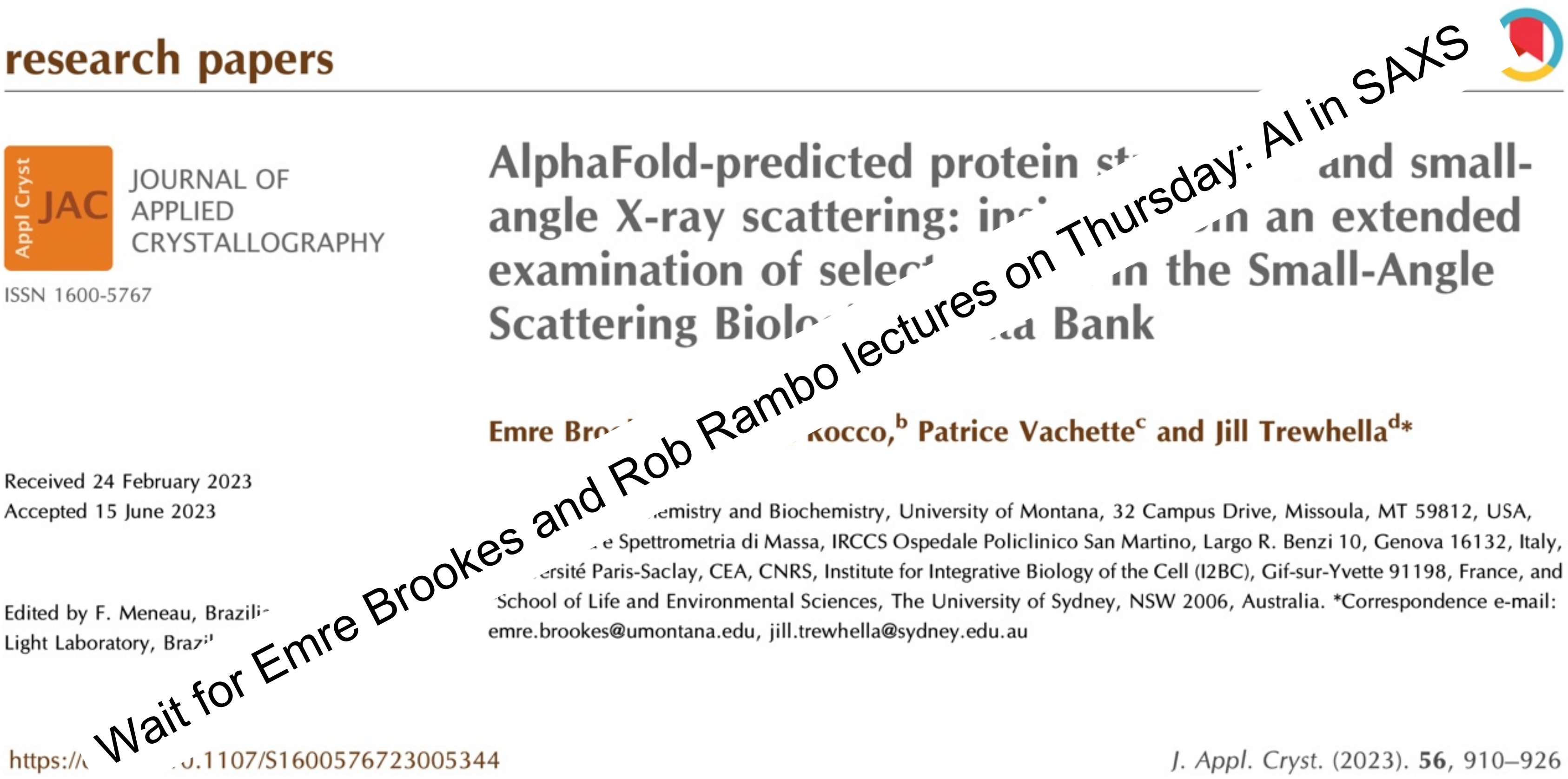
Received 24 February 2023
Accepted 15 June 2023

Edited by F. Meneau, Brazilian Synchrotron Light Laboratory, Brazil

https://doi.org/10.1107/S1600576723005344

J. Appl. Cryst. (2023). **56**, 910–926

Combine with Alphafold!



research papers

JOURNAL OF APPLIED CRYSTALLOGRAPHY
ISSN 1600-5767

AlphaFold-predicted protein st and small-
angle X-ray scattering: in in an extended
examination of selec n the Small-Angle
Scattering Biolo a Bank

Emre Bro Rocco,[b] Patrice Vachette[c] and Jill Trewhella[d]*

emistry and Biochemistry, University of Montana, 32 Campus Drive, Missoula, MT 59812, USA,
e Spettrometria di Massa, IRCCS Ospedale Policlinico San Martino, Largo R. Benzi 10, Genova 16132, Italy,
ersité Paris-Saclay, CEA, CNRS, Institute for Integrative Biology of the Cell (I2BC), Gif-sur-Yvette 91198, France, and
School of Life and Environmental Sciences, The University of Sydney, NSW 2006, Australia. *Correspondence e-mail:
emre.brookes@umontana.edu, jill.trewhella@sydney.edu.au

Received 24 February 2023
Accepted 15 June 2023

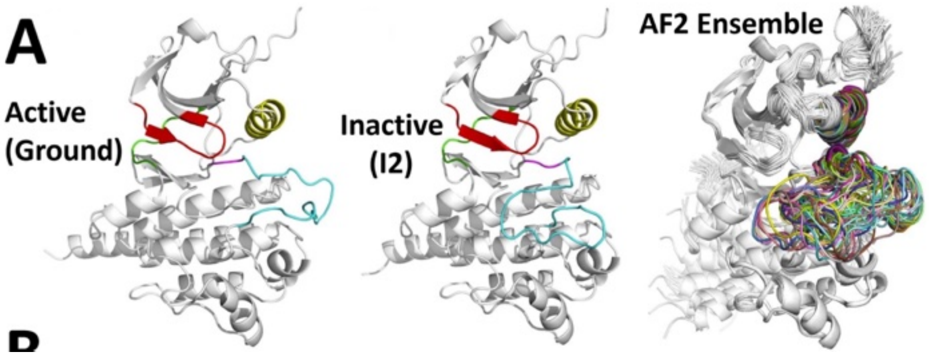
Edited by F. Meneau, Brazili
Light Laboratory, Braz

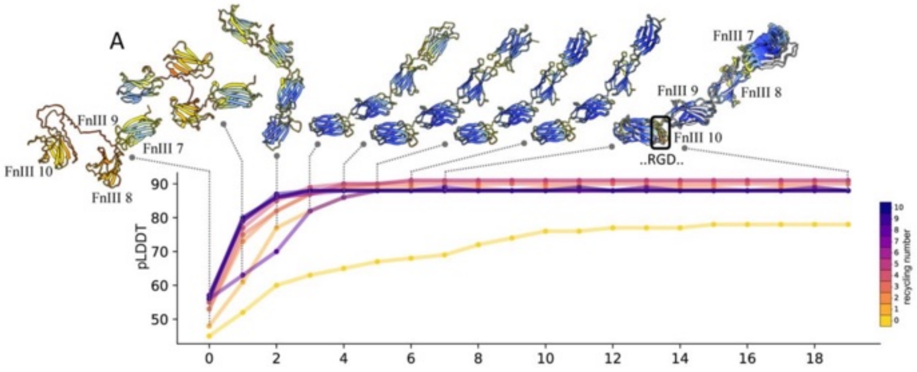
https:// .1107/S1600576723005344

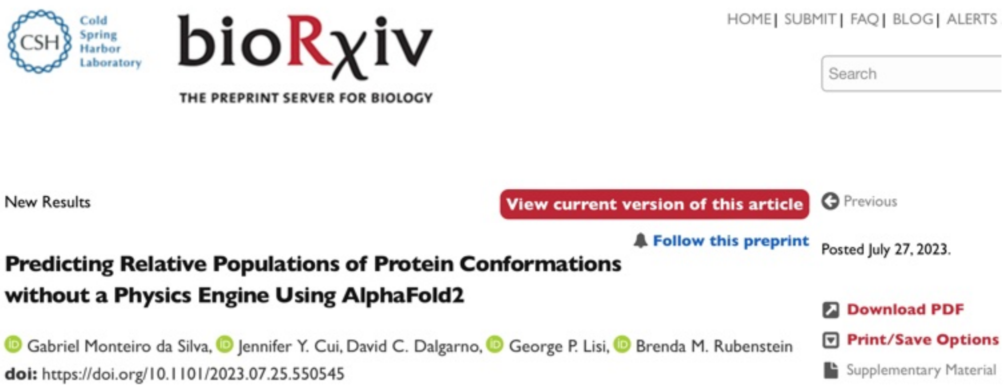
J. Appl. Cryst. (2023). 56, 910–926

Wait for Emre Brookes and Rob Rambo lectures on Thursday: AI in SAXS

Is it only a matter of time before AlphaFold can build ensemble models by itself? …yes







Cold Spring Harbor Laboratory

bioRxiv

THE PREPRINT SERVER FOR BIOLOGY

HOME | SUBMIT | FAQ | BLOG | ALERTS

Search

New Results

View current version of this article

Follow this preprint

Previous

Posted July 27, 2023.

Predicting Relative Populations of Protein Conformations without a Physics Engine Using AlphaFold2

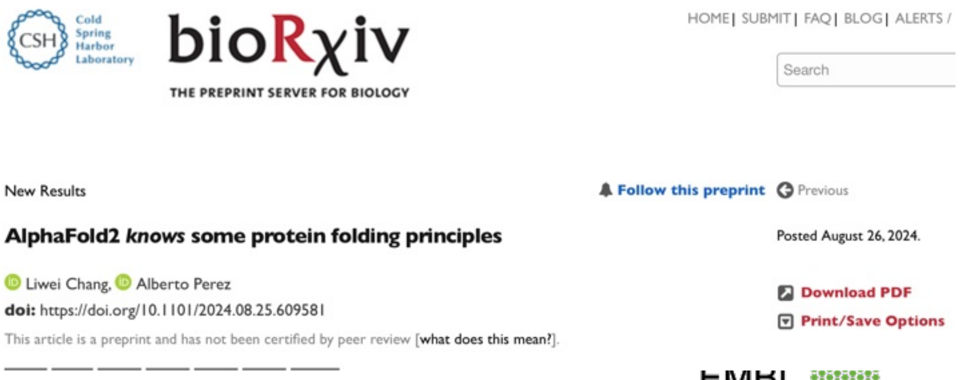
Gabriel Monteiro da Silva, Jennifer Y. Cui, David C. Dalgarno, George P. Lisi, Brenda M. Rubenstein

doi: https://doi.org/10.1101/2023.07.25.550545

Download PDF

Print/Save Options

Supplementary Material



Cold Spring Harbor Laboratory

bioRxiv

THE PREPRINT SERVER FOR BIOLOGY

HOME | SUBMIT | FAQ | BLOG | ALERTS /

Search

New Results

Follow this preprint

Previous

AlphaFold2 *knows* some protein folding principles

Posted August 26, 2024.

Liwei Chang, Alberto Perez

doi: https://doi.org/10.1101/2024.08.25.609581

This article is a preprint and has not been certified by peer review [what does this mean?].

Download PDF

Print/Save Options

Thank you and goodbye!

SAXS Team@EMBL

Dmytro Soloviov
Melissa Gräwert
Clement Blanchet
Aleksi Sutinen

Everyone involved in ATSAS over the years!

EMBL

DFG Deutsche Forschungsgemeinschaft

elixir

European Commission

SEVENTH FRAMEWORK PROGRAMME

Bundesministerium für Bildung und Forschung

iNEXT DISCOVERY

FACH HOCHSCHULE LÜBECK University of Applied Sciences