SwedNess

SANS data analysis II: ATSAS and other methods

for biological SAS (but only)

Wojtek Potrzebowski

EUROPEAN
SPALLATION
SOURCE

LUNDS
UNIVERSITET

What SasView does not cover

- Ab initio modeling
- Rigid body modeling
- Shape and Conformational Polydispersity
- Efficient intensity calculation from PDB file

ATSAS

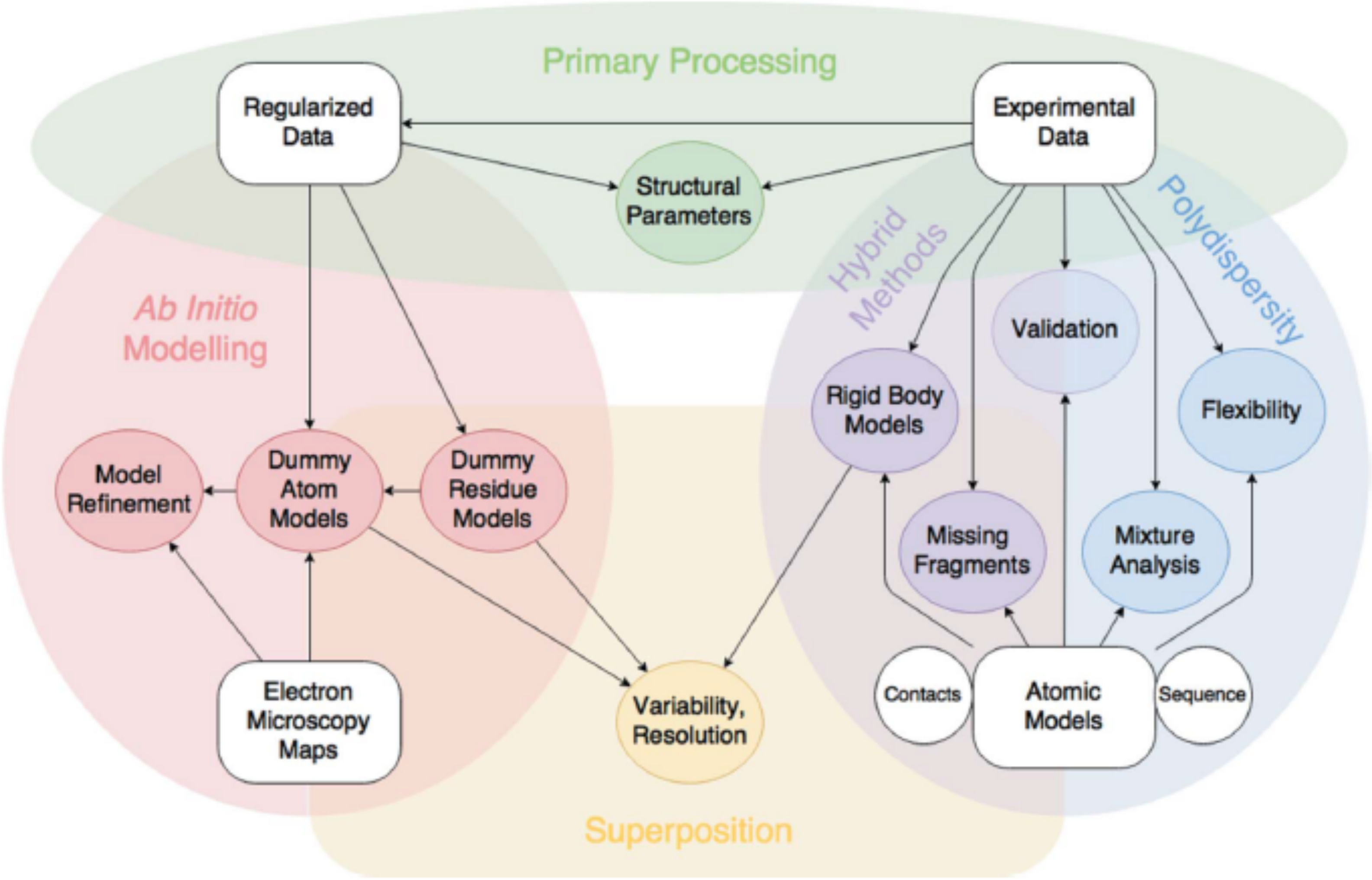
- Molecular Dynamics — Sassie
- 3D particle electron densities from SAS data — DENSS
- And much more smallangle.org/content/software

ATSAS software package 3.0

- Over 90 programs
- Operating systems:
  - Windows 8 and 10,
  - macOS 10.12 Sierra, 10.13 High Sierra and 10.14 Mojave,
  - Red Hat/CentOS 7 and 8,
  - Ubuntu 16 and 18,
  - Debian 9 and 10.
- Free for academic users: https://www.embl-hamburg.de/biosaxs/download.html

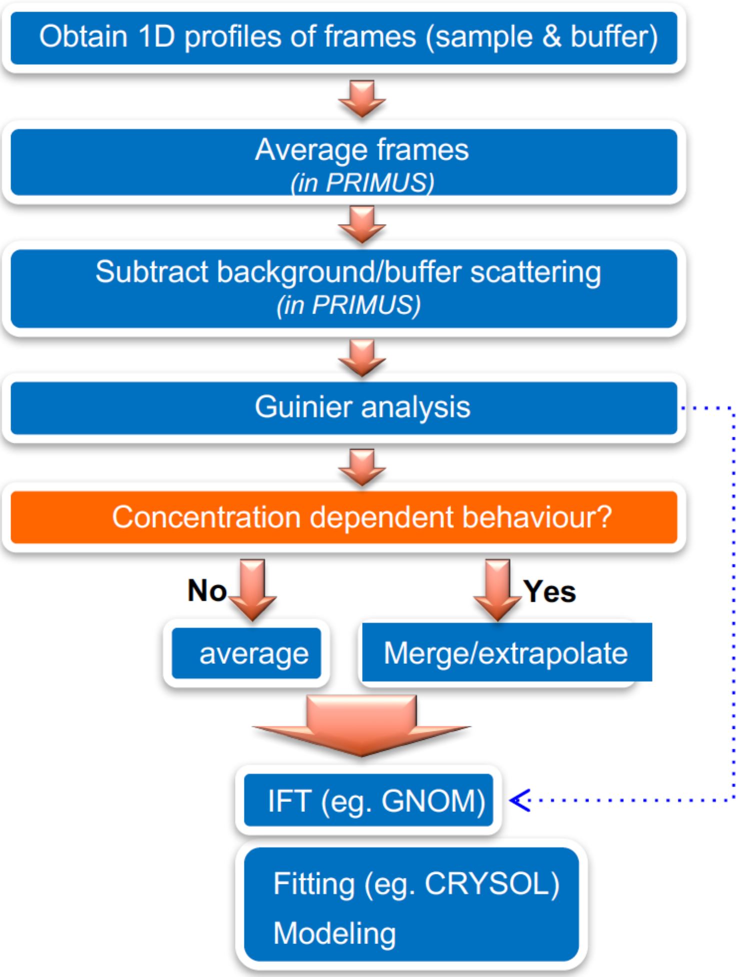
K. Manalastas-Cantos, P.V. Konarev, N.R. Hajizadeh, A.G. Kikhney, M.V. Petoukhov, D.S. Molodenskiy, A. Panjkovich, H.D.T. Mertens, A. Gruzinov, C. Borges, C.M. Jeffries, D.I. Svergun and D. Franke (2020)
ATSAS 3.0: Expanded functionality and new tools for small-angle scattering data analysis
J. Appl. Cryst., submitted

Atsas software overview



Franke et al. J.Appl.Cryst. (2017). 50, 1212–1225

Primary processing workflow



Obtain 1D profiles of frames (sample & buffer)

Average frames
(in PRIMUS)

Subtract background/buffer scattering
(in PRIMUS)

Guinier analysis

Concentration dependent behaviour?

average

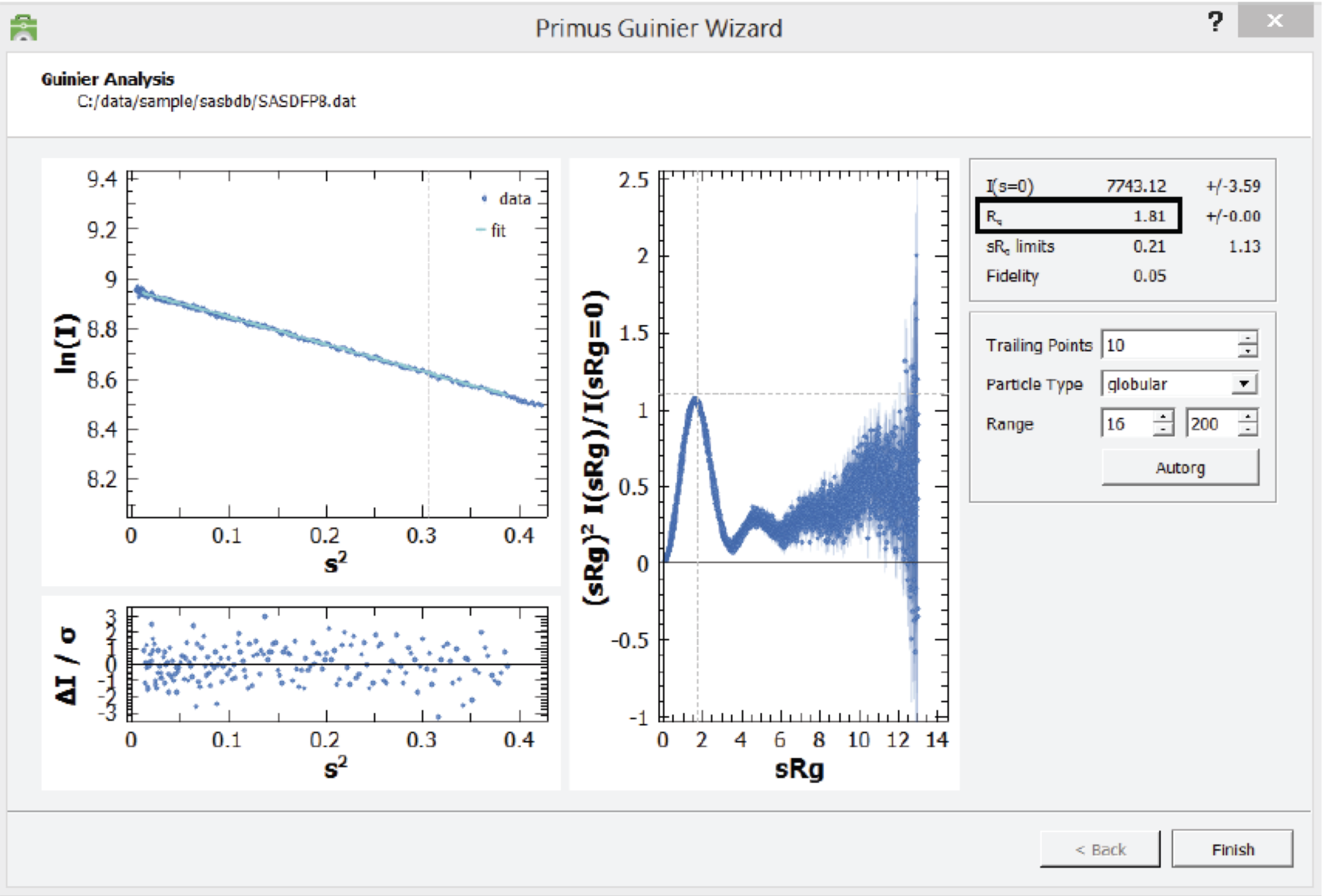
Merge/extrapolate

IFT (eg. GNOM)

Fitting (eg. CRYSOL)
Modeling

Adapted from Haydyn Mertens (EMBL-Hamburg), EMBO 2017

Guinier approximation in PRIMUS



Primus Guinier Wizard

Guinier Analysis

C:/data/sample/sasbdb/SASDFP8.dat

I(s=0)	7743.12	+/-3.59
R_g	1.81	+/-0.00
sR_g limits	0.21	1.13
Fidelity	0.05	

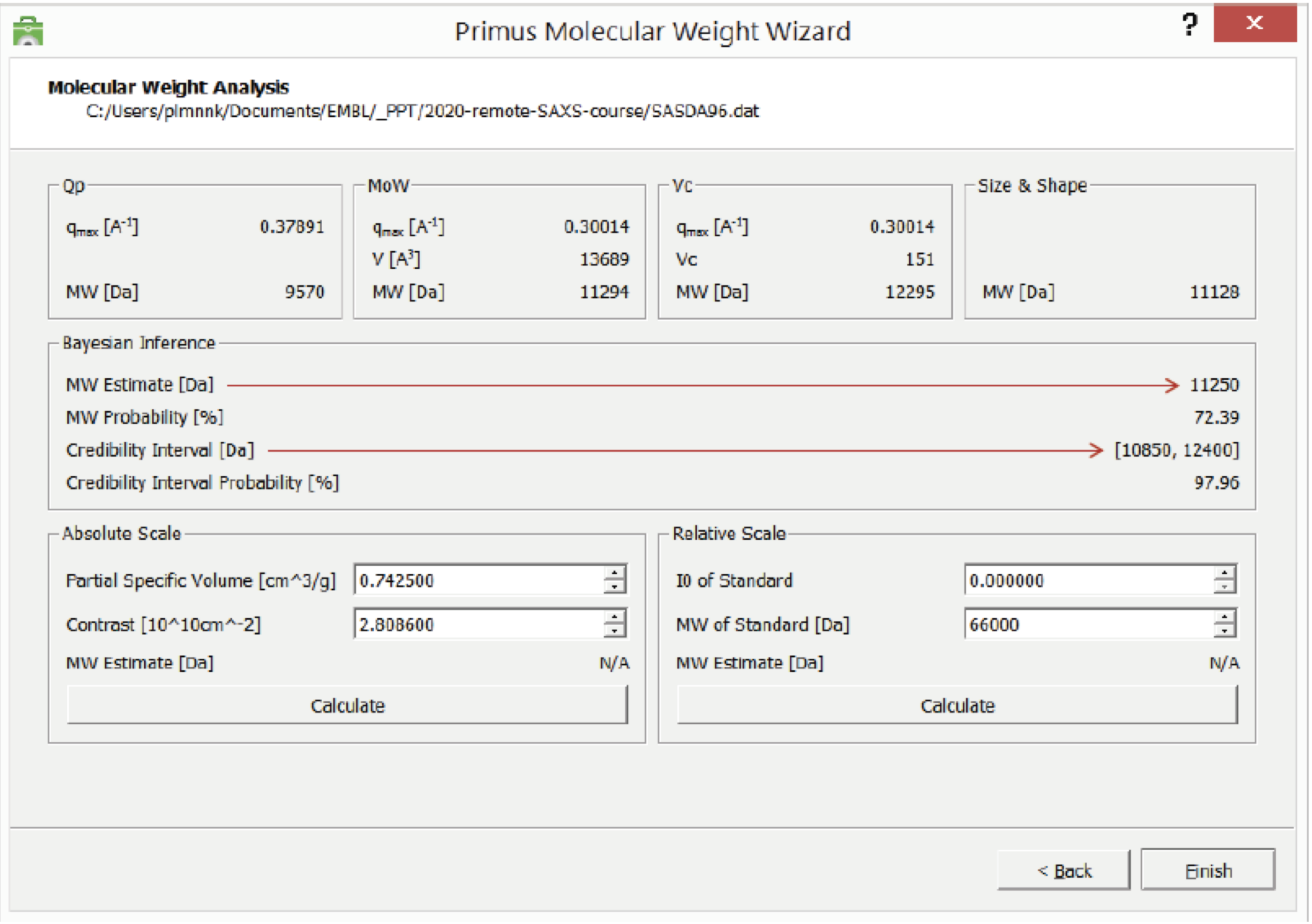
Trailing Points: 10

Particle Type: globular

Range: 16 200

Adapted from Al Kikhney

Molecular Weight in PRIMUS



Primus Molecular Weight Wizard

Molecular Weight Analysis

C:/Users/plmnnk/Documents/EMBL/_PPT/2020-remote-SAXS-course/SASDA96.dat

Qp	
q_{max} [A^{-1}]	0.37891
MW [Da]	9570

MoW	
q_{max} [A^{-1}]	0.30014
V [A^3]	13689
MW [Da]	11294

Vc	
q_{max} [A^{-1}]	0.30014
Vc	151
MW [Da]	12295

Size & Shape	
MW [Da]	11128

Bayesian Inference	
MW Estimate [Da]	11250
MW Probability [%]	72.39
Credibility Interval [Da]	[10850, 12400]
Credibility Interval Probability [%]	97.96

Absolute Scale	
Partial Specific Volume [cm^3/g]	0.742500
Contrast [10^10cm^-2]	2.808600
MW Estimate [Da]	N/A

Relative Scale	
I0 of Standard	0.000000
MW of Standard [Da]	66000
MW Estimate [Da]	N/A

P(r) inversion

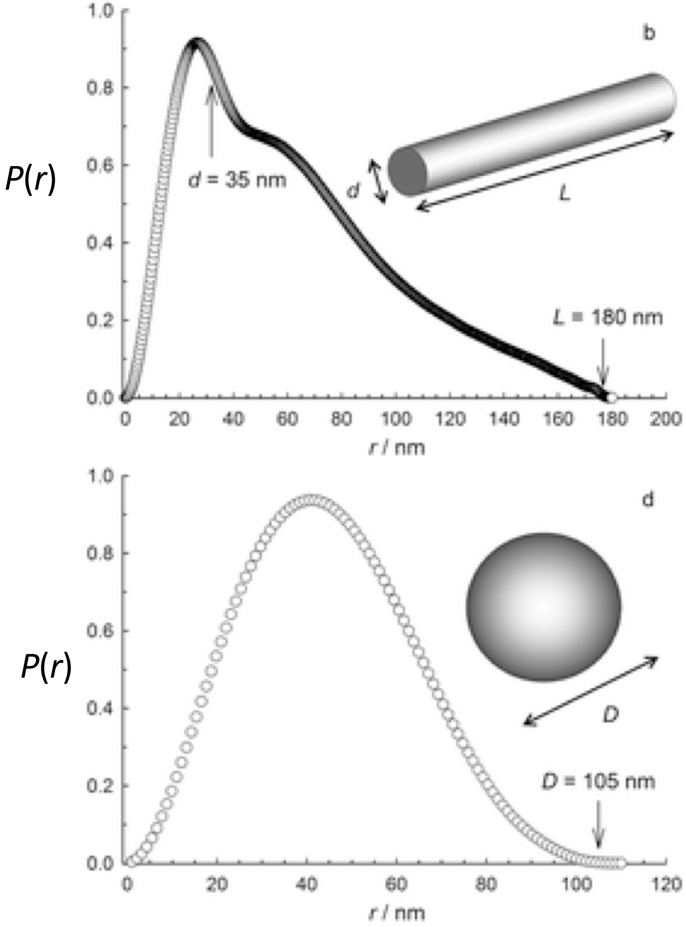
- Calculates a real-space pair-distance distribution function

$$I(q) = 4\pi \int_0^{D\max} p(r) \frac{\sin(qr)}{qr} .dr$$

FT $\Uparrow\Downarrow$ FT^{-1}

$$p(r) = \frac{r^2}{2\pi^2} \int_0^{\infty} q^2 I(q) \frac{\sin(qr)}{qr} .dq$$

- Calculated by Indirect Fourier Transform (Fourier transform of noisy data).
- Popular methods: Glatter, Moore
- Maximum *d*, noise level, regularization constants have to be chosen



F. Grrohn, Soft Matter, 2010,6, 4296-4302

P(r) calculation (SasView)

- Using IFT method (Moore, 1980)
- *P(r)* is set to be equal to an expansion of base functions of the type

$$P(r) = \sum_{n=0}^{N} c_n \Phi_n$$

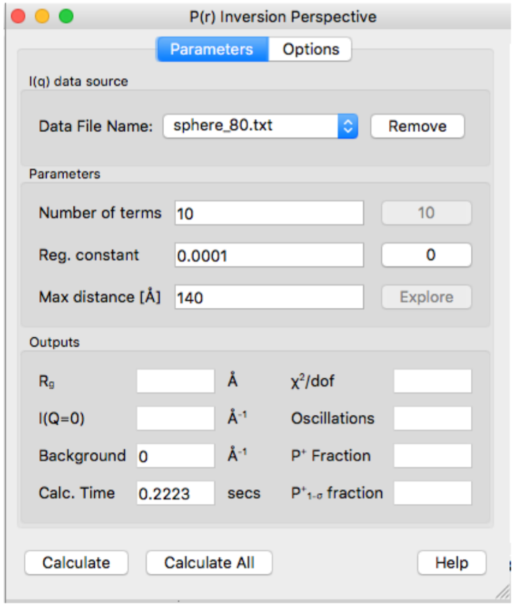
$$\Phi_{n(r)} = 2rsin(\frac{\pi n r}{D_{max}})$$

- The coefficient c_n of each base function in the expansion is found by performing a least square fit

$$\chi^2 = \frac{\sum_i (I_{meas}(Q_i) - I_{th}(Q_i))^2}{error^2} + Reg\_term$$

P(r) calculation SasView

- *Number of terms*: the number of base functions in the P(r) expansion.
- *Regularization constant*: a multiplicative constant to set the size of the regularization term.
- *Maximum distance*: the maximum distance between any two points in the system.



I(q) data source

Data File Name: sphere_80.txt

Parameters

Number of terms: 10 (10)

Reg. constant: 0.0001 (0)

Max distance [Å]: 140

Outputs

R_g: Å

χ^2/dof:

I(Q=0): $Å^{-1}$

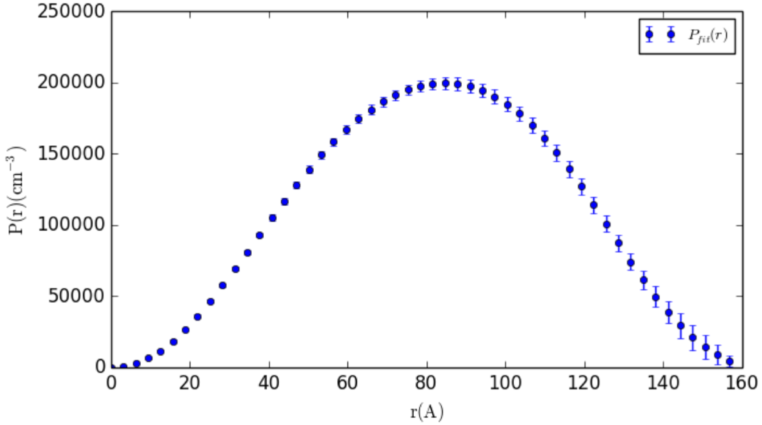
Oscillations:

Background: 0 $Å^{-1}$

P^+ Fraction:

Calc. Time: 0.2223 secs

$P^+_{1-\sigma}$ fraction:



P(r) calculation (ATSAS)

Indirect Fourier Transformation (IFT) of SAXS data

- Solution is Indirect Fourier Transformation (IFT), (Glatter, 1977)
- Fit a function to the SAXS data and transform → *p(r)*
- Regularisation parameter (*α*) helps balance between the fit and the **FT**.

$$p(r) = \sum_{k=1}^{K} c_k \phi_k(s_i)$$

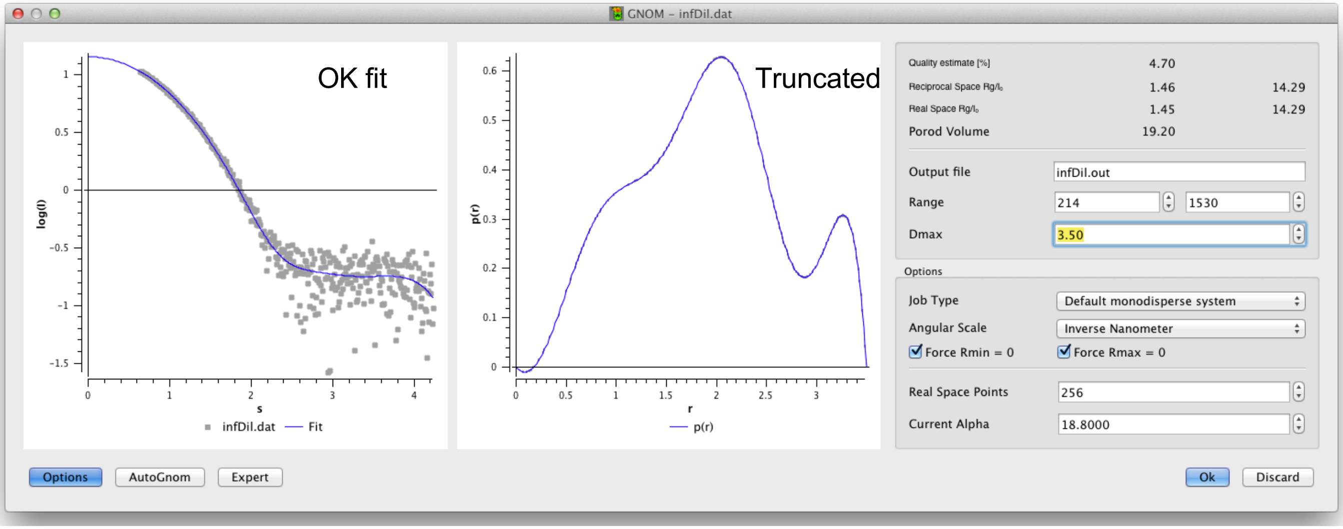
$$\begin{tikzcd}
\Phi = \chi^2 + \alpha P(p) \arrow[dr] \arrow[dd] & \\
& P(p) = \int_0^{D_{\max}} \left[p'\right]^2 dr \\
\chi^2 = \frac{1}{N-1} \sum_{j=1}^{N} \left[\frac{I_{exp}(s_j) - cI_{calc}(s_j)}{\sigma(s_j)} \right]^2 &
\end{tikzcd}$$

Adapted from Haydyn Mertens (EMBL-Hamburg), EMBO 2017

P(r) calculation with GNOM

So, what is a good *p(r)*? How do I know a good solution?

- D_{max} estimate (3.5 nm) poor solution – ***too small***



OK fit

Truncated

Quality estimate [%] 4.70

Reciprocal Space Rg/I₀ 1.46 14.29

Real Space Rg/I₀ 1.45 14.29

Porod Volume 19.20

Output file: infDil.out

Range: 214 1530

Dmax: 3.50

Options

Job Type: Default monodisperse system

Angular Scale: Inverse Nanometer

☑ Force Rmin = 0 ☑ Force Rmax = 0

Real Space Points: 256

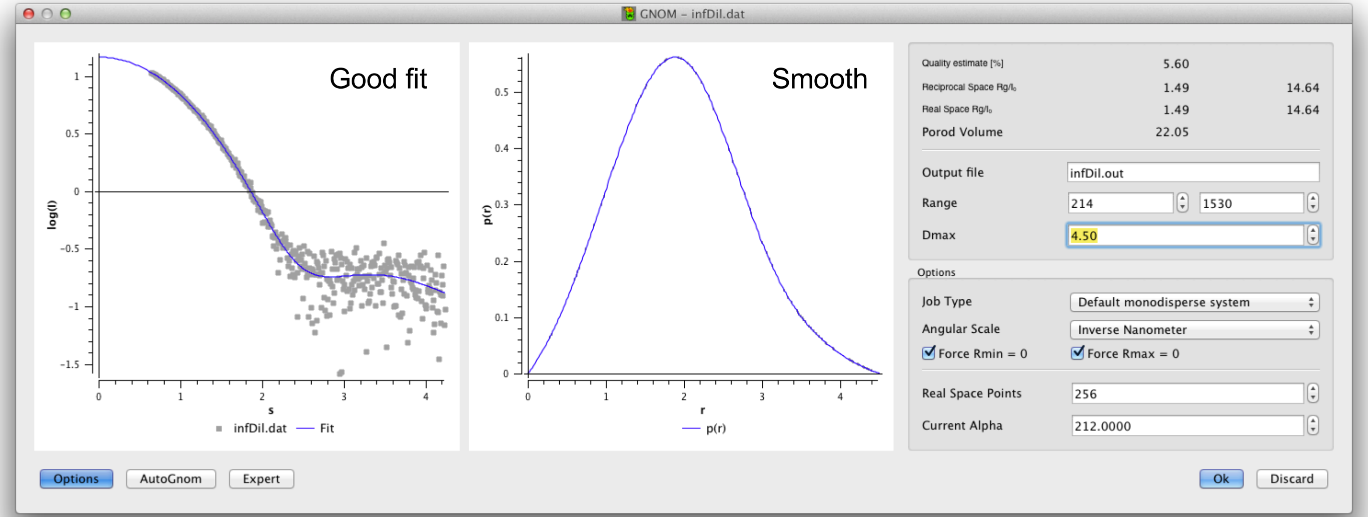
Current Alpha: 18.8000

Adapted from Haydyn Mertens (EMBL-Hamburg), EMBO 2017

P(r) calculation with GNOM

So, what is a good *p(r)*? How do I know a good solution?

- D_{max} estimate (4.5 nm) - ***good solution***



GNOM – infDil.dat

Good fit

Smooth

Quality estimate [%] 5.60

Reciprocal Space Rg/I₀ 1.49 14.64

Real Space Rg/I₀ 1.49 14.64

Porod Volume 22.05

Output file: infDil.out

Range: 214 1530

Dmax: 4.50

Options

Job Type: Default monodisperse system

Angular Scale: Inverse Nanometer

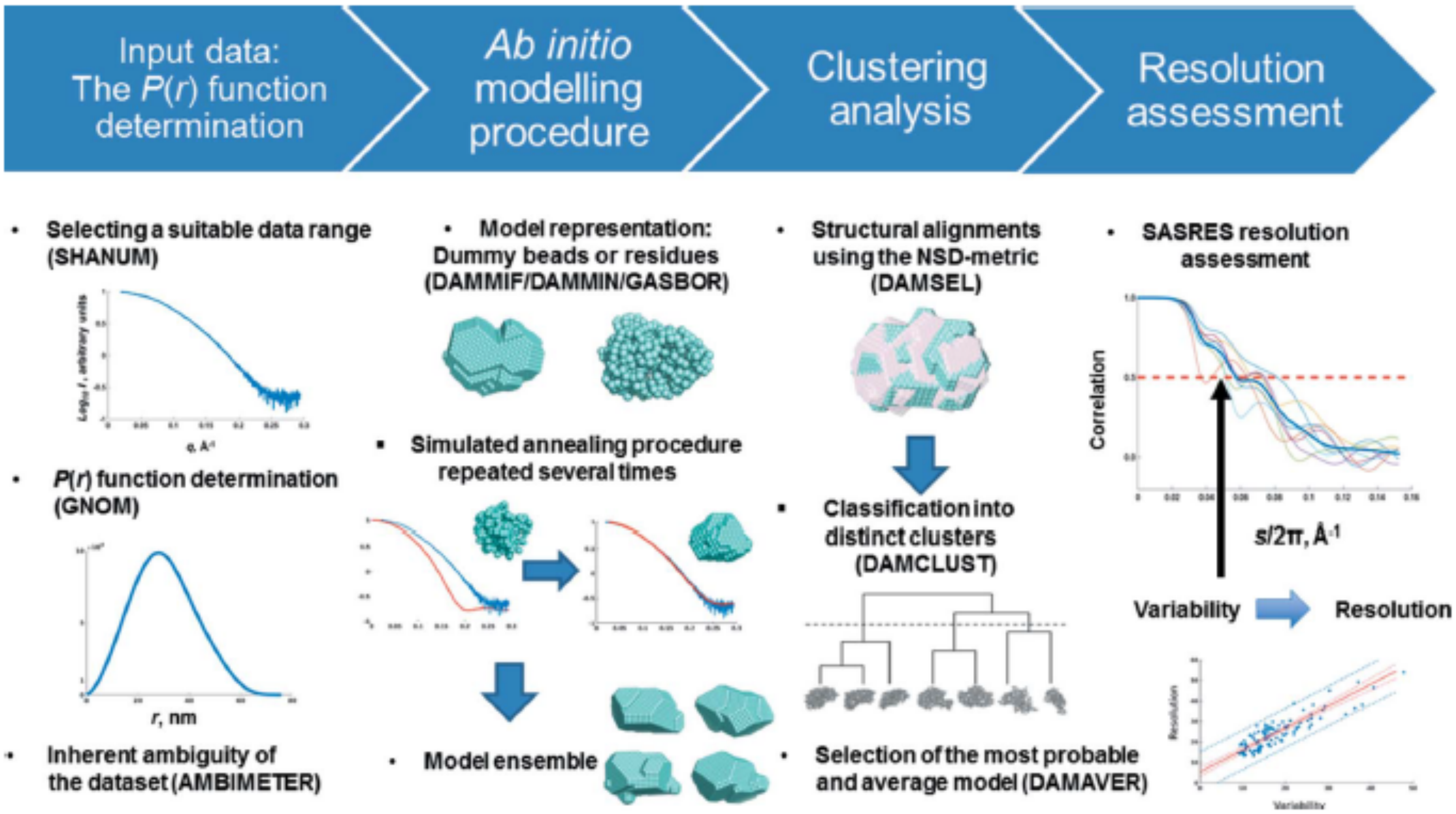
☑ Force Rmin = 0 ☑ Force Rmax = 0

Real Space Points: 256

Current Alpha: 212.0000

Adapted from Haydyn Mertens (EMBL-Hamburg), EMBO 2017

Ab initio modeling overview



Input data: The *P*(*r*) function determination

Ab initio modelling procedure

Clustering analysis

Resolution assessment

- Selecting a suitable data range (SHANUM)
- *P*(*r*) function determination (GNOM)
- Inherent ambiguity of the dataset (AMBIMETER)

- Model representation: Dummy beads or residues (DAMMIF/DAMMIN/GASBOR)
- Simulated annealing procedure repeated several times
- Model ensemble

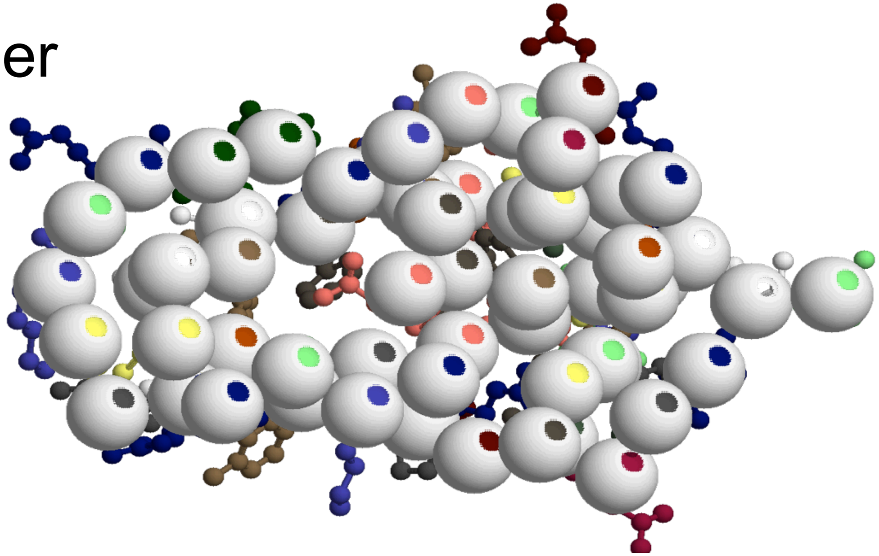
- Structural alignments using the NSD-metric (DAMSEL)
- Classification into distinct clusters (DAMCLUST)
- Selection of the most probable and average model (DAMAVER)

- SASRES resolution assessment

Variability → Resolution

Dummy residues

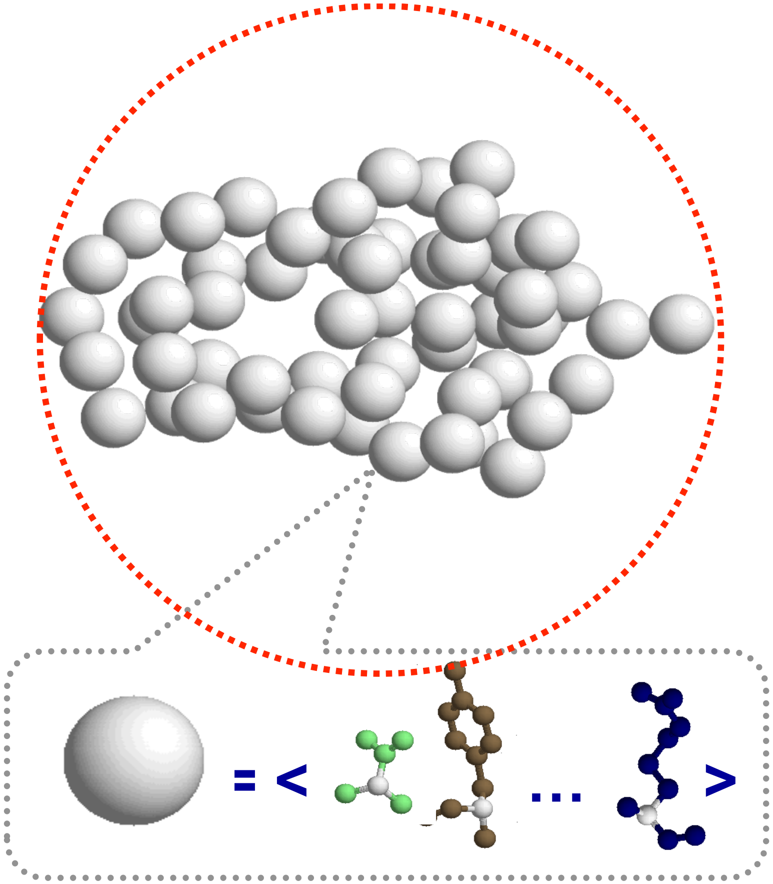
- Proteins typically consist of folded polypeptide chains composed of amino acid residues
- At a resolution of 0.5 nm each amino acid can be represented as one entity (dummy residue)
- In GASBOR a protein is represented by an ensemble of *K* dummy residues that are
  - Identical
  - Have no ordinal number
  - For simplicity are centered at the Cα positions



Adapted from Daniel Franke (EMBL-Hamburg), EMBO 2017

Dummy residues

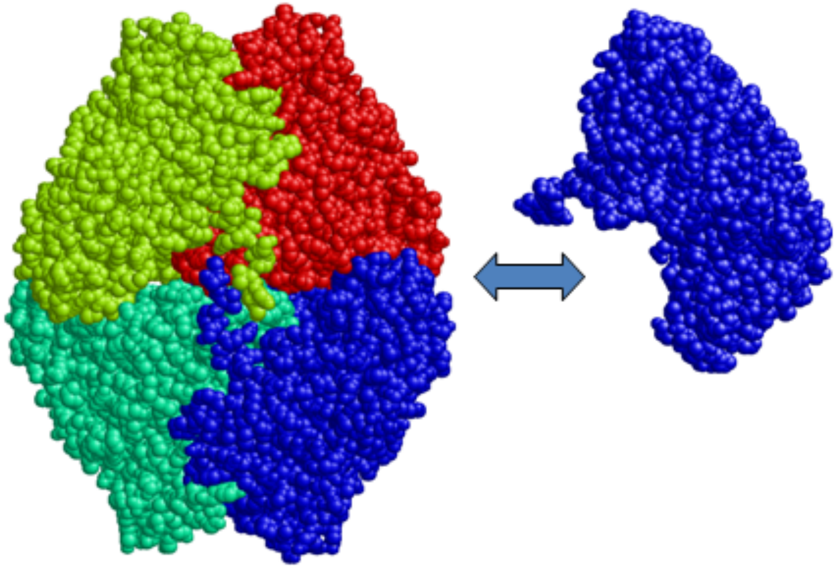
- GASBOR finds coordinates of *K* dummy residues within its search volume (red)
- Scattering is computed using the Debye (1915) formula
- Requires polypeptide chain-compatible arrangement of dummy residues



Adapted from Daniel Franke (EMBL-Hamburg), EMBO 2017

Dummy residues for mixture models

- GASBORMX extension to equilibrium mixtures
- Reconstructs the monomer and a symmetric multimer together
- Interconnectivity is required for the monomer and the multimer



Adapted from Daniel Franke (EMBL-Hamburg), EMBO 2017

Single Phase Dummy Atom Models

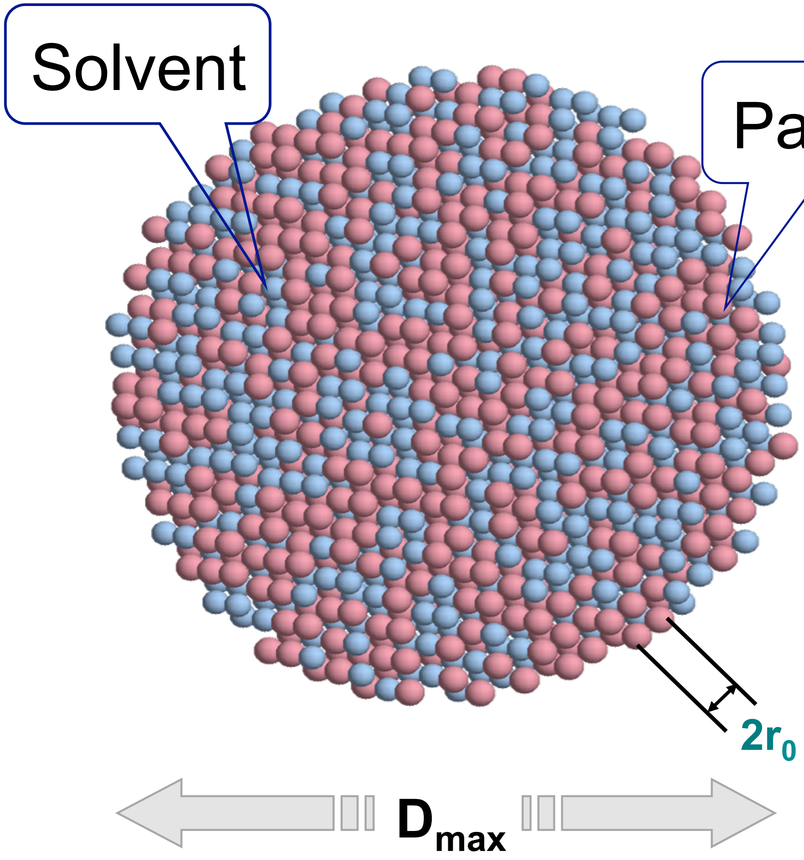
Dummy atoms:

- Act as a placeholder for, but does not resemble, a real atom
- Occupy a known position in space
- Have a known scattering pattern
- May either contribute to solvent or particle
- Are also known as beads

Adapted from Daniel Franke (EMBL-Hamburg), EMBO 2017

Single Phase Dummy Atom Models

A volume is filled by densely packed beads of radius $r_0 << D_{max}$



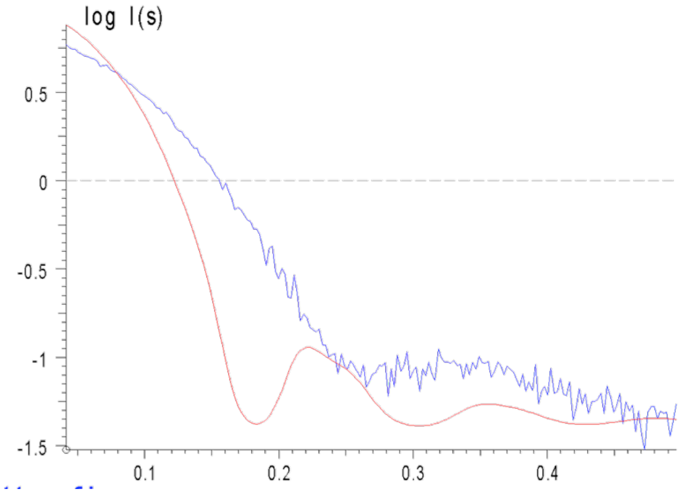
Solvent

Particle

$2r_0$

D_{max}

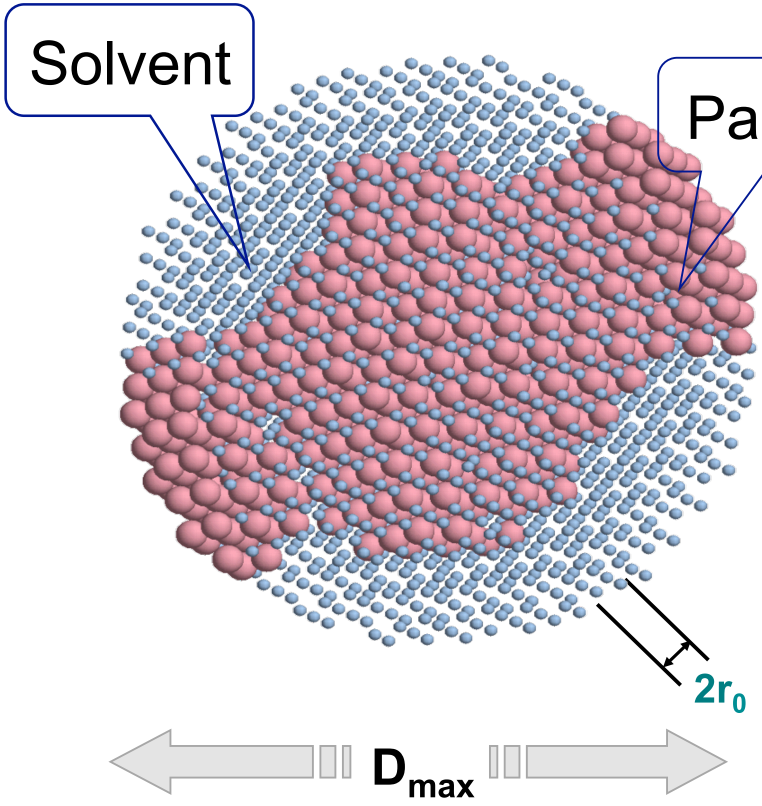
Parametrization:
a binary vector,
0 if solvent, 1 if particle



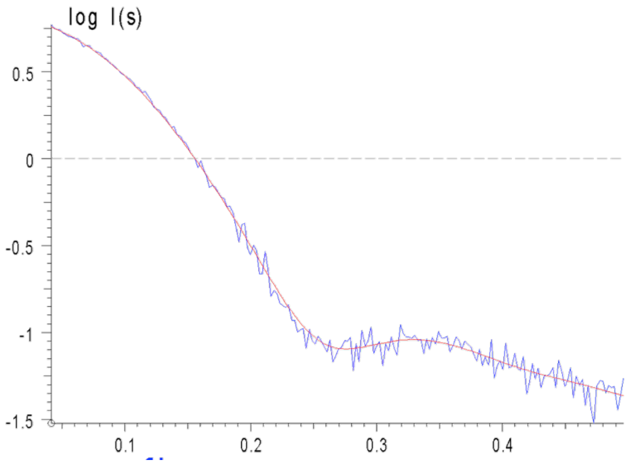
Adapted from Daniel Franke (EMBL-Hamburg), EMBO 2017

Single Phase Dummy Atom Models

A volume is filled by densely packed beads of radius $r_0 << D_{max}$



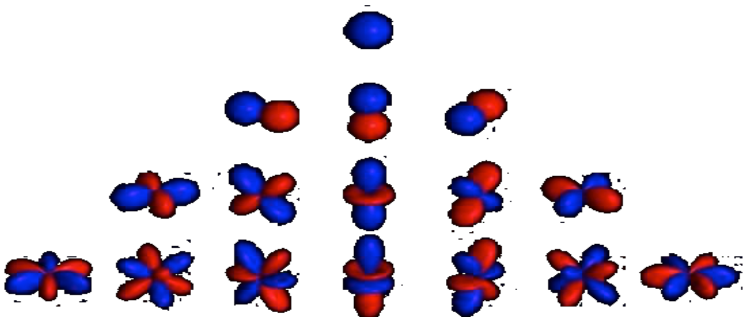
Parametrization:
a binary vector,
0 if solvent, 1 if particle



Adapted from Daniel Franke (EMBL-Hamburg), EMBO 2017

Single Phase Dummy Atom Models

- Scattering intensity is computed using spherical harmonics

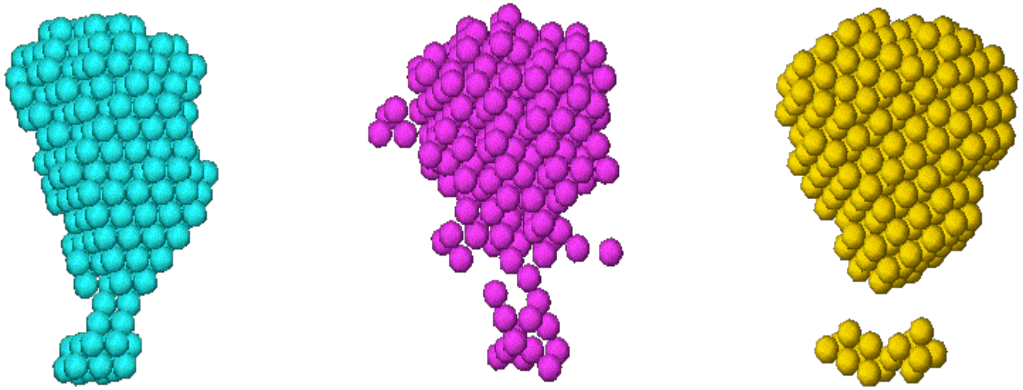


- Penalty terms ensure compactness and connectivity

compact

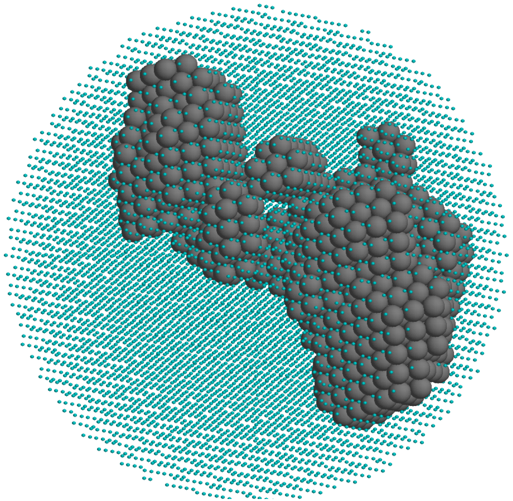
loose

disconnected

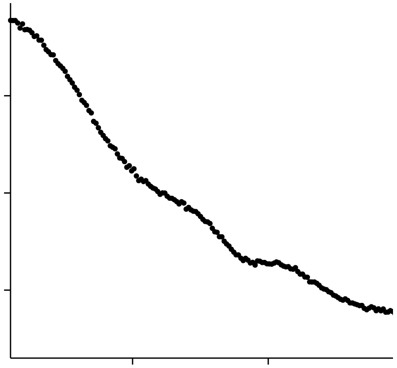


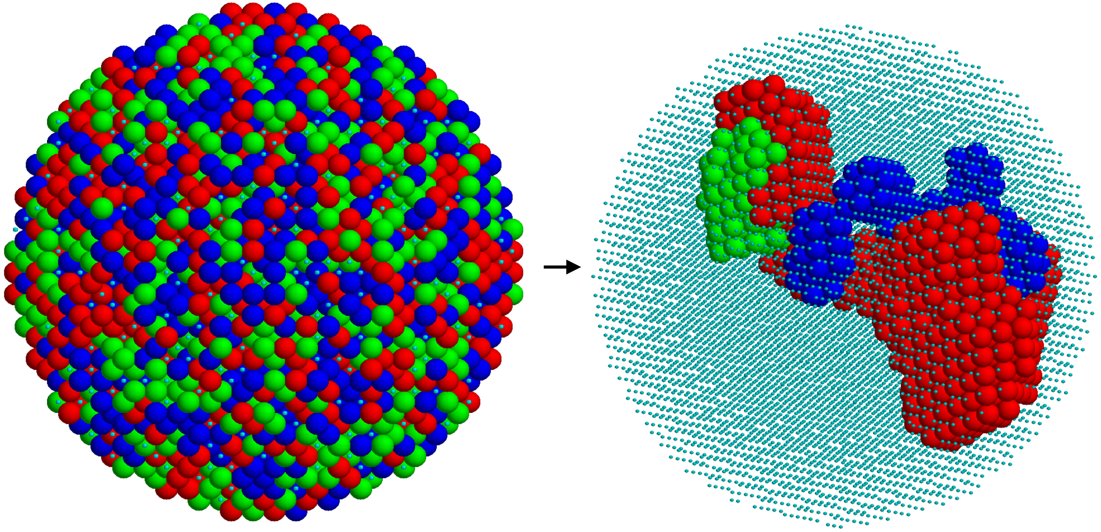
Adapted from Daniel Franke (EMBL-Hamburg), EMBO 2017

Multi Phase Dummy Atom Models



Single phase shape determination





- One can differentiate between distinct parts of the particle
- Several curves are required
- Assuming the same arrangement of the parts in different samples

Adapted from Daniel Franke (EMBL-Hamburg), EMBO 2017

Dummy Atom Models

	DAMMIN	DAMMIF	MONSA
Objects	any	any	any
Max # of phases	1	1	4
Angular range	lower part	lower part	lower part
Resolution	low	low	low
Search volume	fixed	growing	fixed
Constrains	Symmetry, Interconnectivity, Compactness	Symmetry, Interconnectivity, Compactness	Symmetry, Interconnectivity, Compactness
Performace	slow	fast	very slow
Limitations		DAMMIN has better symmetry support	

Warning: results are not atomic models, just a filled volume!

Adapted from Daniel Franke (EMBL-Hamburg), EMBO 2017

Applicability to SAS data

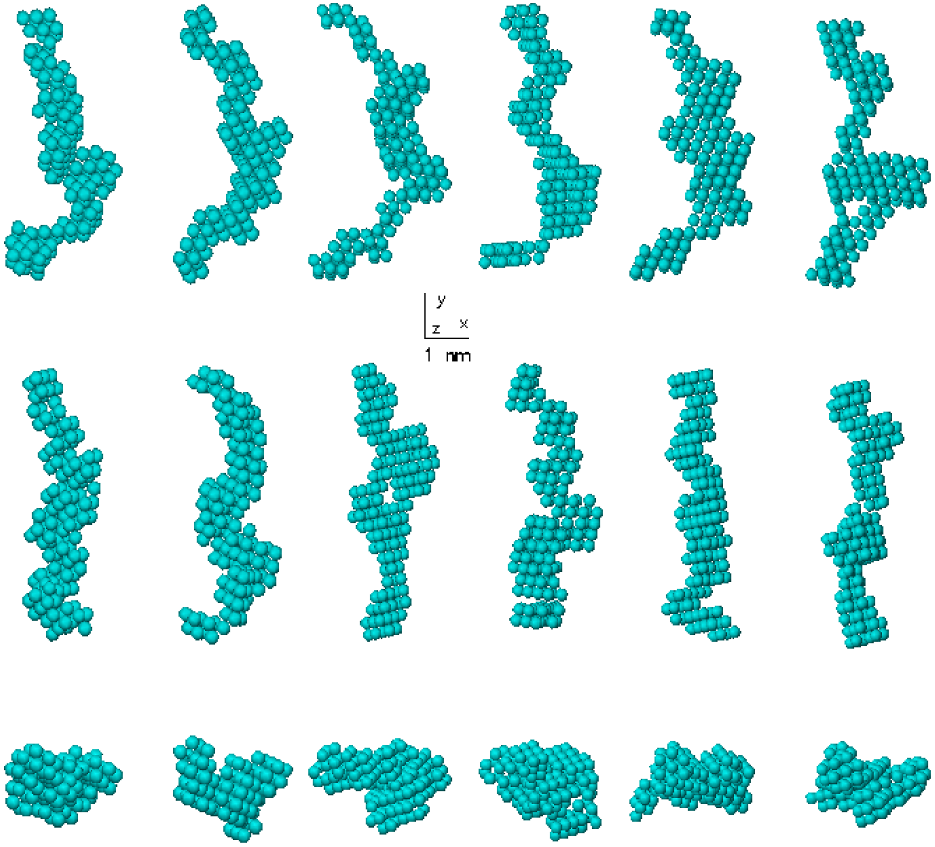
Program	SAXS	SANS
GASBOR/GASBORMX	✓	✗ **
DAMMIN/DAMMIF	✓	✓ *
MONSA	✓	✓

* May be used if contrast is high and the particle is homogeneous
** Dummy residue form factors are available for X-rays only

Adapted from Daniel Franke (EMBL-Hamburg), EMBO 2017

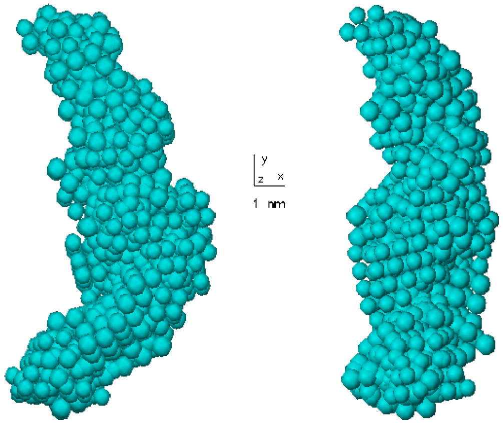
Model post processing

- 5S RNA models
- A variety of DAMMIN models explains data equally well

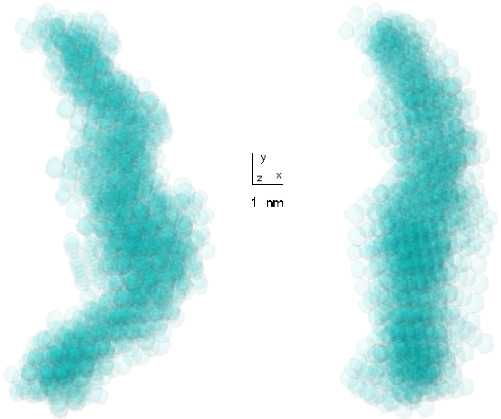


Adapted from Daniel Franke (EMBL-Hamburg), EMBO 2017

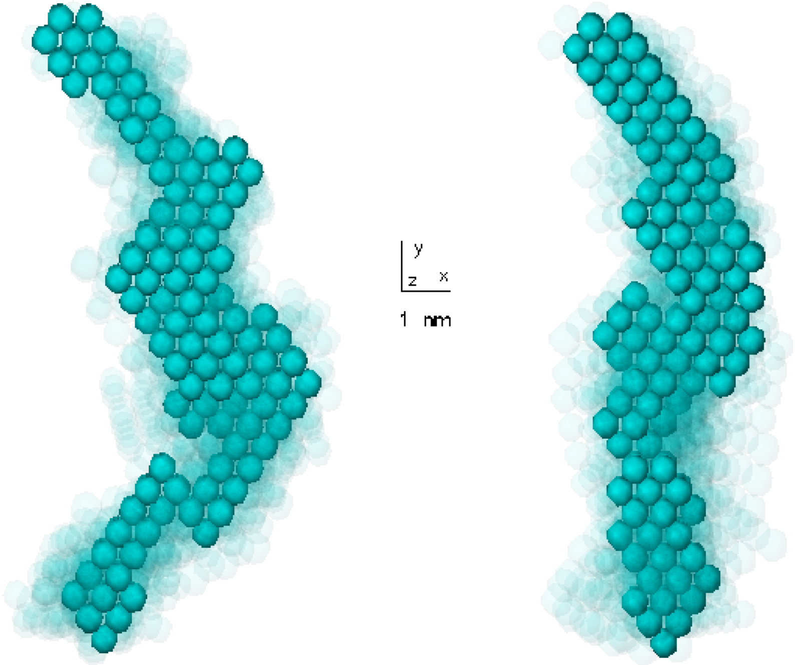
Model post processing



5S RNA – Solution spread region



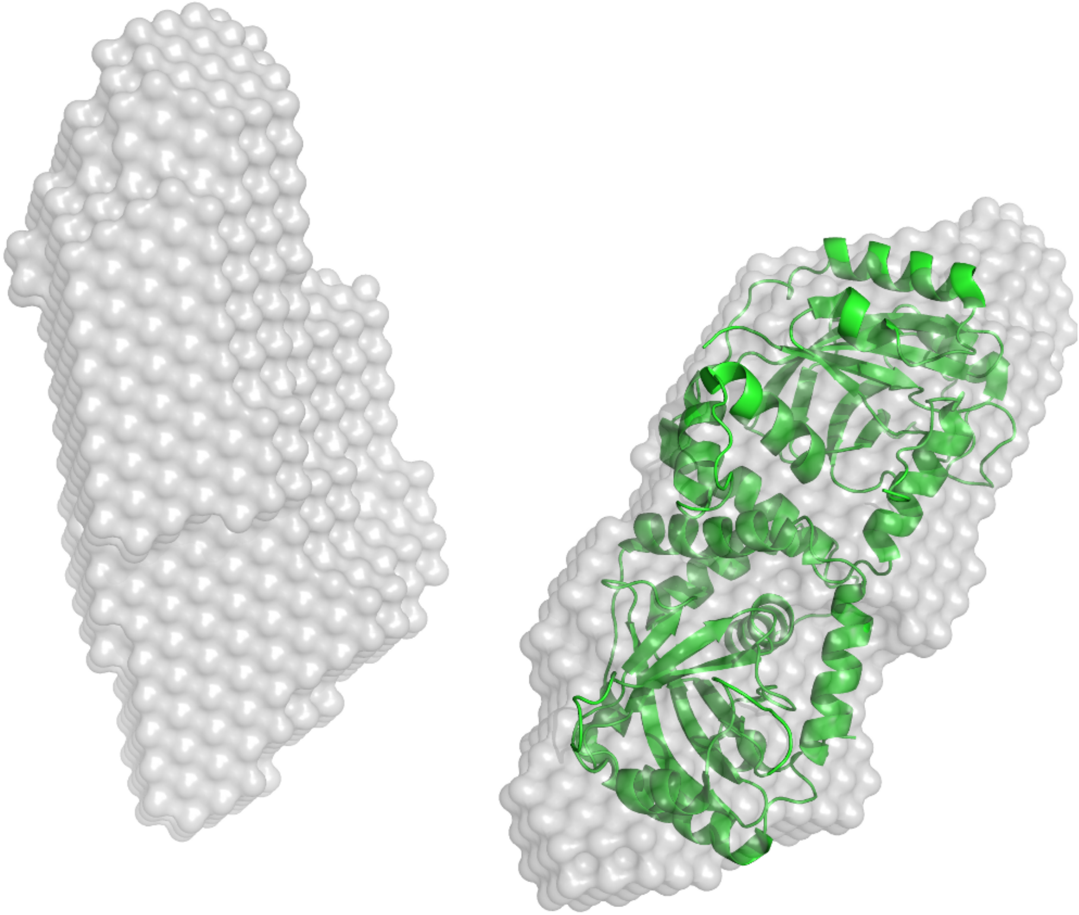
5S RNA – Most Populated Volume



5S RNA – Final Solution within the Spread Region

Adapted from Daniel Franke (EMBL-Hamburg), EMBO 2017

Superposition with hi-res model



SUPALM

Konarev PV, Svergun DI. *IUCrJ.* 2015 Apr 21;2(Pt 3):352-60

Model Validity

- Validate your input data
- Check for
  - Aggregation
  - Noise at higher angles
- Keep in mind: it is easy to model noise

→ Garbage in, garbage out

Adapted from Daniel Franke (EMBL-Hamburg), EMBO 2017

Fitting high-resolution structures to SAS data

Scattering from macromolecule in solution

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

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

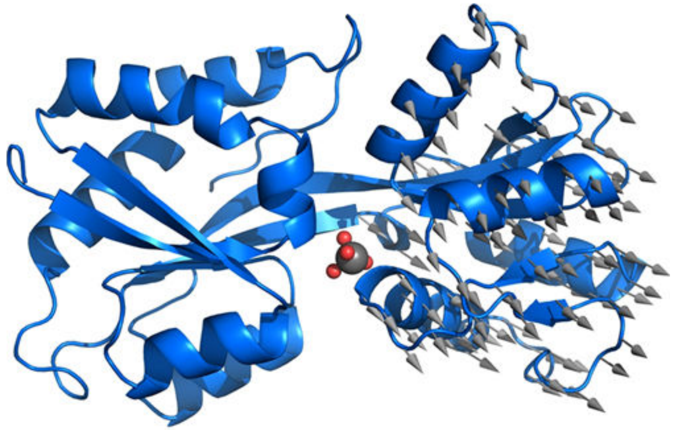
Programs:

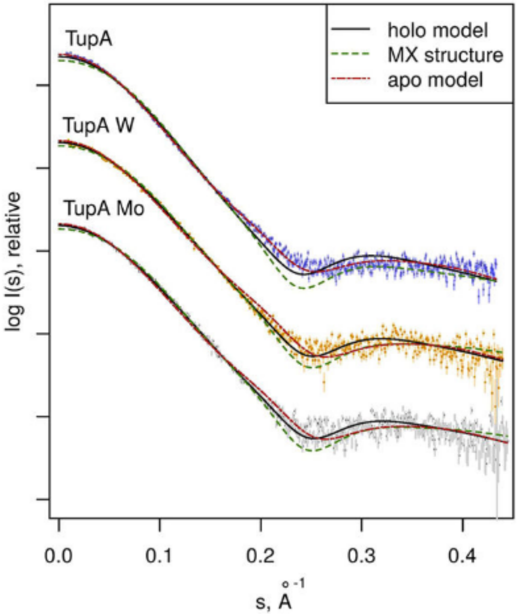
- CRYSOL (X-rays): Svergun *et al.* (1995) *J. Appl. Cryst.* **28**, 768
- CRYSON (neutrons): Svergun *et al.* (1998) *P.N.A.S USA* **95**, 2267

Adapted from Alejandro Panjkovich (EMBL-Hamburg), EMBO 2017

CRYSOL on PDB structure

How does the atomic model fit the solution scattering profile?





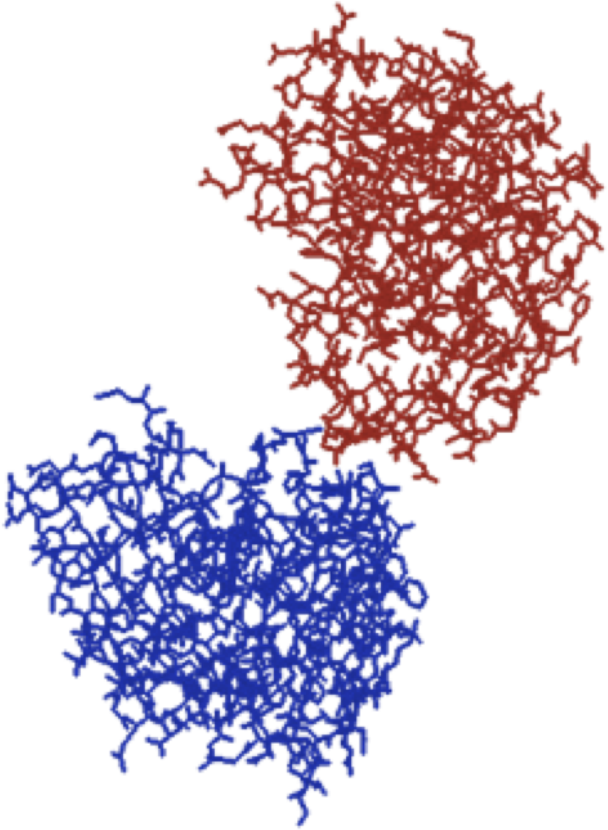
$$\chi^2 = \frac{1}{N}\sum_{i=1}^{N_p}\left(\frac{I_e(s_i) - cI(s_i)}{\sigma(s_i)}\right)^2 \qquad (7)$$

Adapted from Alejandro Panjkovich (EMBL-Hamburg), EMBO 2017

Rigid and flexible modeling

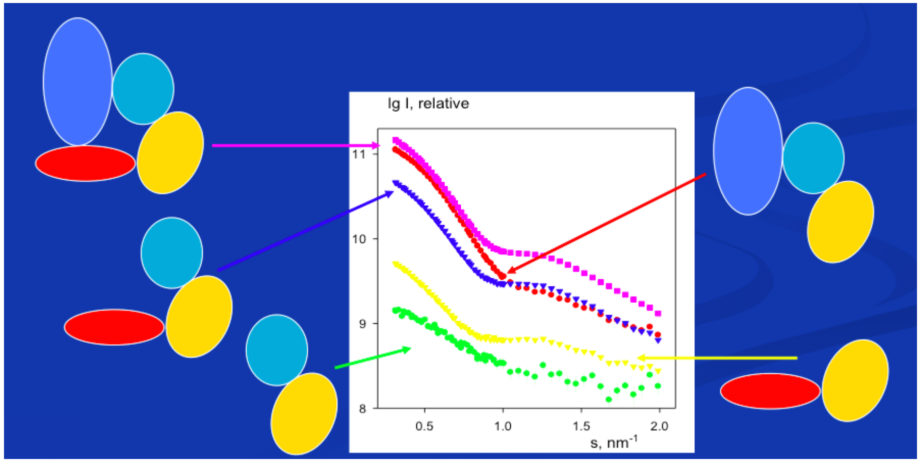
Rigid body fitting

- The structures of two subunits in reference positions are known.
- Arbitrary complex can be constructed by moving and rotating the second subunit.
- This operation depends on three Euler rotation angles and three Cartesian shifts.



Adapted from Alejandro Panjkovich (EMBL-Hamburg), EMBO 2017

Rigid body modeling with SASREF

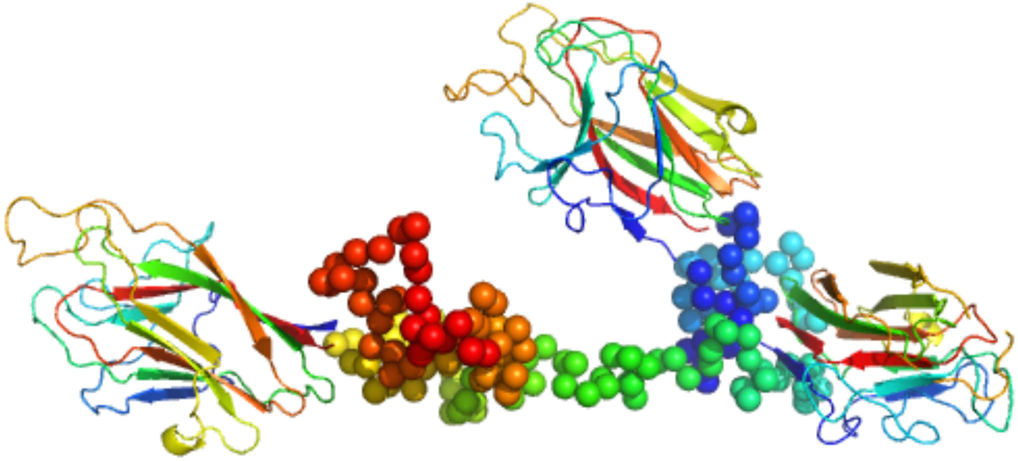


- Fits (multiple X-ray and neutron) scattering curve(s) from partial constructs or contrast variation using simulated annealing
- Requires models of subunits, builds interconnected models without steric clashes.
- Uses constrains: symmetry, distance, relative orientation if applicable.

Petoukhov & Svergun (2005). *Biophys J.* **89**, 1237; Petoukhov & Svergun (2006). *Biophys J.* **35**, 567

Adapted from Alejandro Panjkovich (EMBL-Hamburg), EMBO 2017

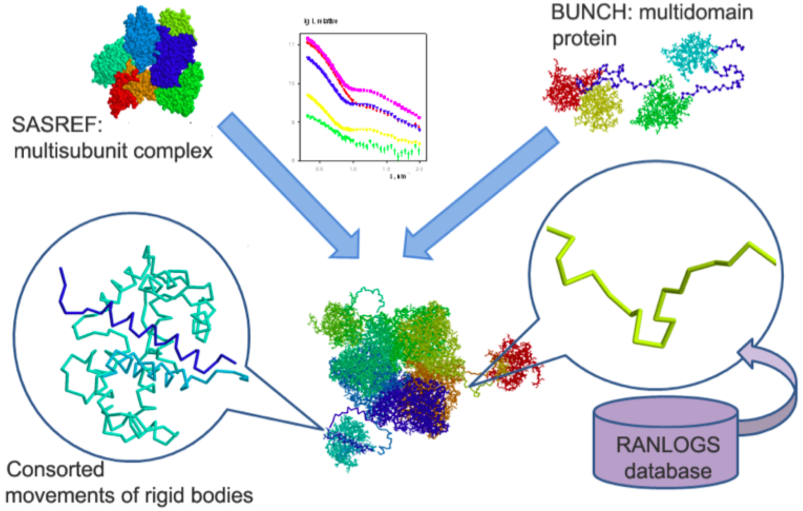
Addition of missing fragments with BUNCH



- BUNCH combines rigid body and *ab-initio* modelling to find the positions and orientations of rigid domains and probable conformations of flexible linkers represented as dummy residues chains
- Multiple experimental scattering data sets from partial constructs (e.g. deletion mutants) can be fitted simultaneously with the data of the full-length protein.
- accounts for symmetry, allows one to fix some domains and to restrain the model by contacts between specific residues

Adapted from Alejandro Panjkovich (EMBL-Hamburg), EMBO 2017

Addition of missing fragments with CORAL



- A combination of SASREF and BUNCH to account for missing loops in multi-subunit biological macromolecules.
- Loops are modeled based on known high-resolution structures.

Adapted from Alejandro Panjkovich (EMBL-Hamburg), EMBO 2017

Summary

- Atsas is a powerful toolbox to analyze SAS data from biological macromolecules
- Dummy models gives an idea of an overall shape of molecule
- Hi-res structures can be compared with SAS data either or used as a building blocks in rigid or flexible modeling
- Potential flaw: you will always get answer but not always correct.

Questions?