Ab initio modelling: recovering 3D shapes from 1D data

Clement Blanchet, EMBL Hamburg

EMBL

How can one obtain a 3d model from the 1D SAXS data?

- SAS curve contain information about distances.

- How to use this information to build model based on the SAXS data.

Outline

- Computing form factor from geometrical shape
- Bead modelling:
  - Principle
  - Target function and minimization
  - Bead modeling
  - Dummy residue modeling
- Words of caution

Form factor of simple geometrical shape

EMBL

Computing form factor from simple geometrical shape

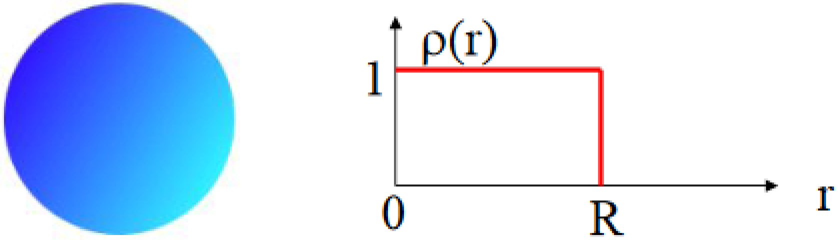
$$I(\boldsymbol{s}) = A(\boldsymbol{s}) \cdot A^*(\boldsymbol{s})$$

$$A(\boldsymbol{s}) = \int_{V_r} \rho(\boldsymbol{r}).\, e^{-i\boldsymbol{r}_j\boldsymbol{s}}\ dV_r$$

For simple geometrical shapes, the form factor can be computed from the electron density

Example: form factor of a solid sphere

From factor of a solid sphere

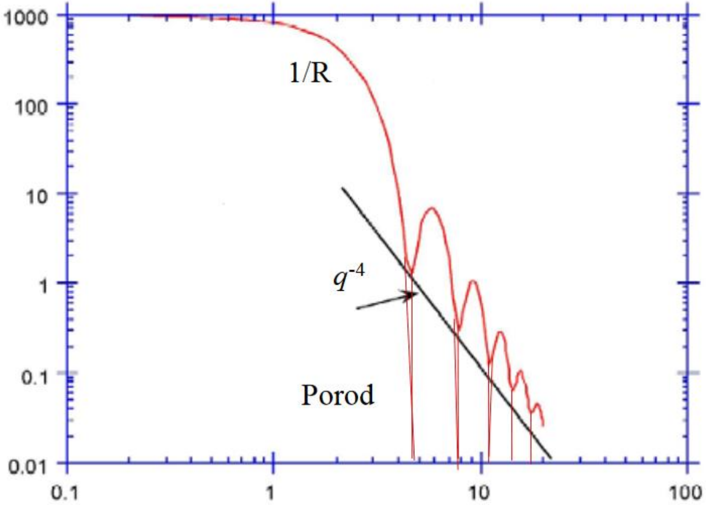


$$A(q) = 4\pi\int_0^\infty \rho(r)\frac{\sin(qr)}{qr}r^2dr = 4\pi\int_0^R \frac{\sin(qr)}{qr}r^2dr$$

$$= \frac{4\pi}{q}\int_0^R \sin(qr)\, rdr =$$

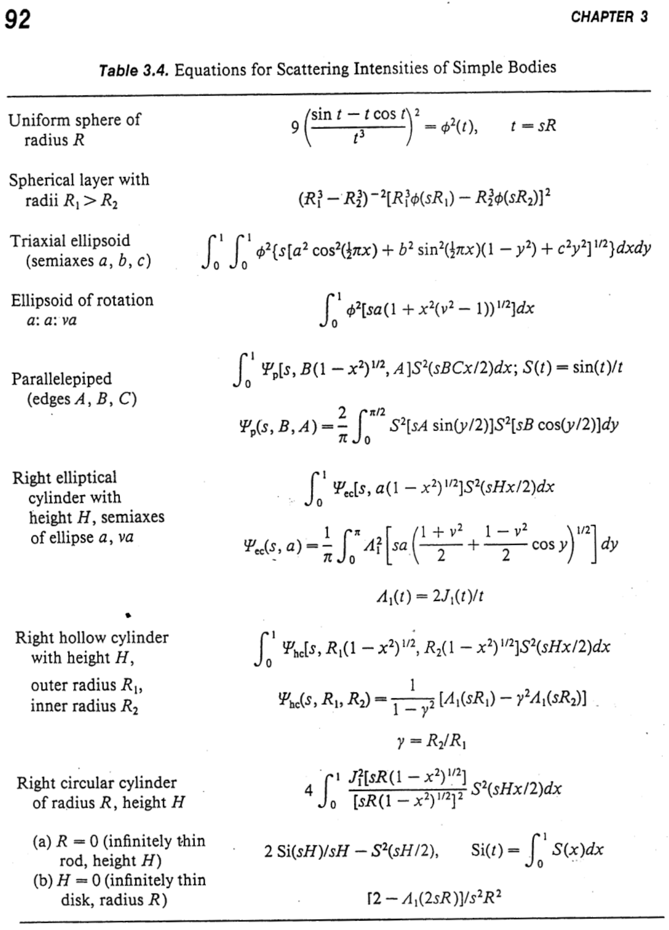
Form factor of sphere

$P(q) = A(q)^2/V^2$



Jan Skov Pedersen

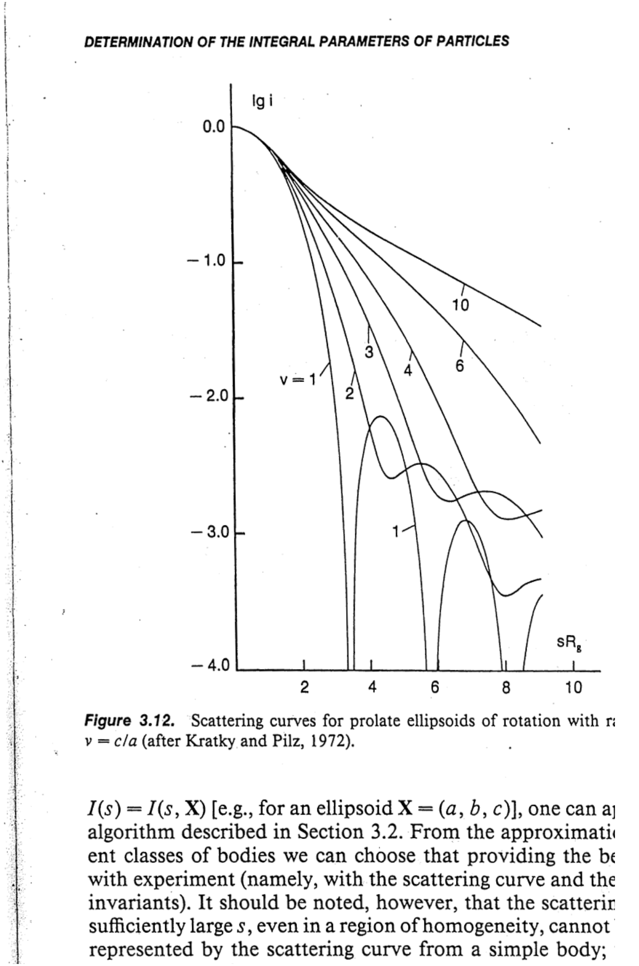
The sphere case is trivial,
It quickly become complicated



92 CHAPTER 3

Table 3.4. Equations for Scattering Intensities of Simple Bodies

Uniform sphere of radius R	$9\left(\frac{\sin t - t\cos t}{t^3}\right)^2 = \phi^2(t), \qquad t = sR$
Spherical layer with radii $R_1 > R_2$	$(R_1^3 - R_2^3)^{-2}[R_1^3\phi(sR_1) - R_2^3\phi(sR_2)]^2$
Triaxial ellipsoid (semiaxes a, b, c)	$\int_0^1\int_0^1 \phi^2\{s[a^2\cos^2(\frac{1}{2}\pi x) + b^2\sin^2(\frac{1}{2}\pi x)(1-y^2) + c^2y^2]^{1/2}\}dxdy$
Ellipsoid of rotation $a: a: va$	$\int_0^1 \phi^2[sa(1 + x^2(v^2-1))^{1/2}]dx$
Parallelepiped (edges A, B, C)	$\int_0^1 \Psi_p[s, B(1-x^2)^{1/2}, A]S^2(sBCx/2)dx; \; S(t) = \sin(t)/t$ $\Psi_p(s, B, A) = \frac{2}{\pi}\int_0^{\pi/2} S^2[sA\sin(y/2)]S^2[sB\cos(y/2)]dy$
Right elliptical cylinder with height H, semiaxes of ellipse a, va	$\int_0^1 \Psi_{ec}[s, a(1-x^2)^{1/2}]S^2(sHx/2)dx$ $\Psi_{ec}(s, a) = \frac{1}{\pi}\int_0^{\pi} \Lambda_1^2\left[sa\left(\frac{1+v^2}{2} + \frac{1-v^2}{2}\cos y\right)^{1/2}\right]dy$ $\Lambda_1(t) = 2J_1(t)/t$
Right hollow cylinder with height H, outer radius R_1, inner radius R_2	$\int_0^1 \Psi_{hc}[s, R_1(1-x^2)^{1/2}, R_2(1-x^2)^{1/2}]S^2(sHx/2)dx$ $\Psi_{hc}(s, R_1, R_2) = \frac{1}{1-\gamma^2}[\Lambda_1(sR_1) - \gamma^2\Lambda_1(sR_2)]$ $\gamma = R_2/R_1$
Right circular cylinder of radius R, height H	$4\int_0^1 \frac{J_1^2[sR(1-x^2)^{1/2}]}{[sR(1-x^2)^{1/2}]^2} S^2(sHx/2)dx$
(a) $R = 0$ (infinitely thin rod, height H)	$2\,\mathrm{Si}(sH)/sH - S^2(sH/2), \qquad \mathrm{Si}(t) = \int_0^1 S(x)dx$
(b) $H = 0$ (infinitely thin disk, radius R)	$[2 - \Lambda_1(2sR)]/s^2R^2$



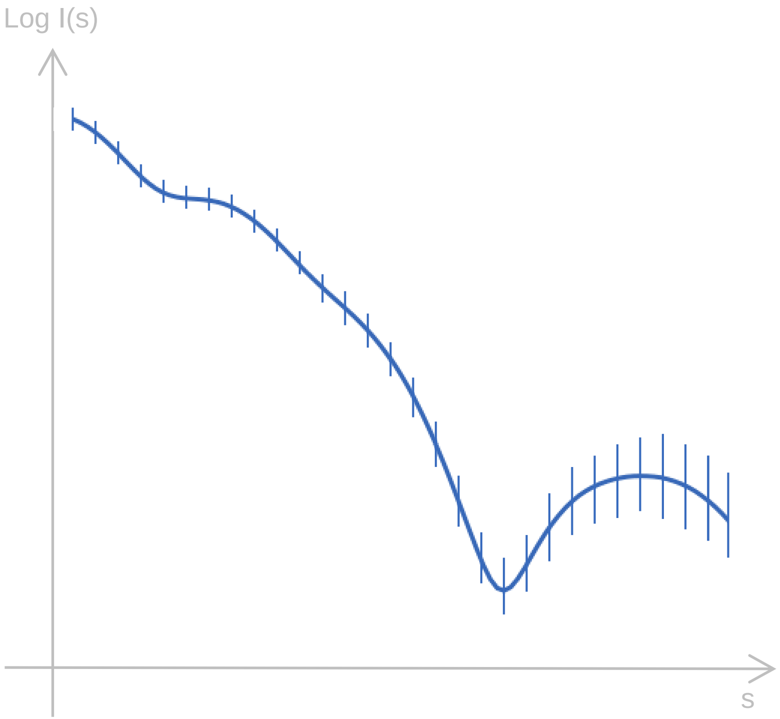
DETERMINATION OF THE INTEGRAL PARAMETERS OF PARTICLES

Figure 3.12. Scattering curves for prolate ellipsoids of rotation with ra
$v = c/a$ (after Kratky and Pilz, 1972).

$I(s) = I(s, \mathbf{X})$ [e.g., for an ellipsoid $\mathbf{X} = (a, b, c)$], one can a
algorithm described in Section 3.2. From the approximati
ent classes of bodies we can choose that providing the be
with experiment (namely, with the scattering curve and the
invariants). It should be noted, however, that the scatterin
sufficiently large s, even in a region of homogeneity, cannot
represented by the scattering curve from a simple body;

Feigin, L.A. and Svergun, D.I. Structure Analysis by Small-Angle X-Ray and Neutron Scattering. Plenum Press 1987

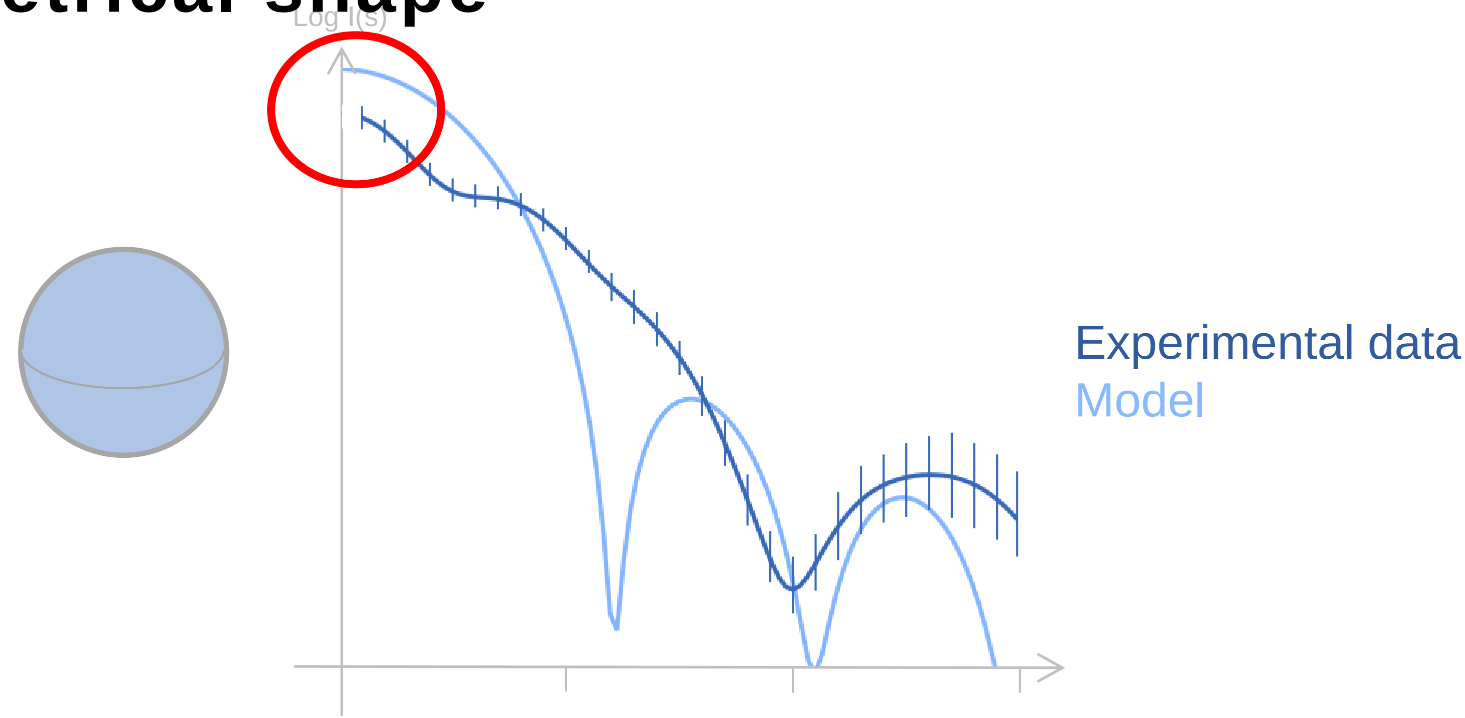
SAXS curve



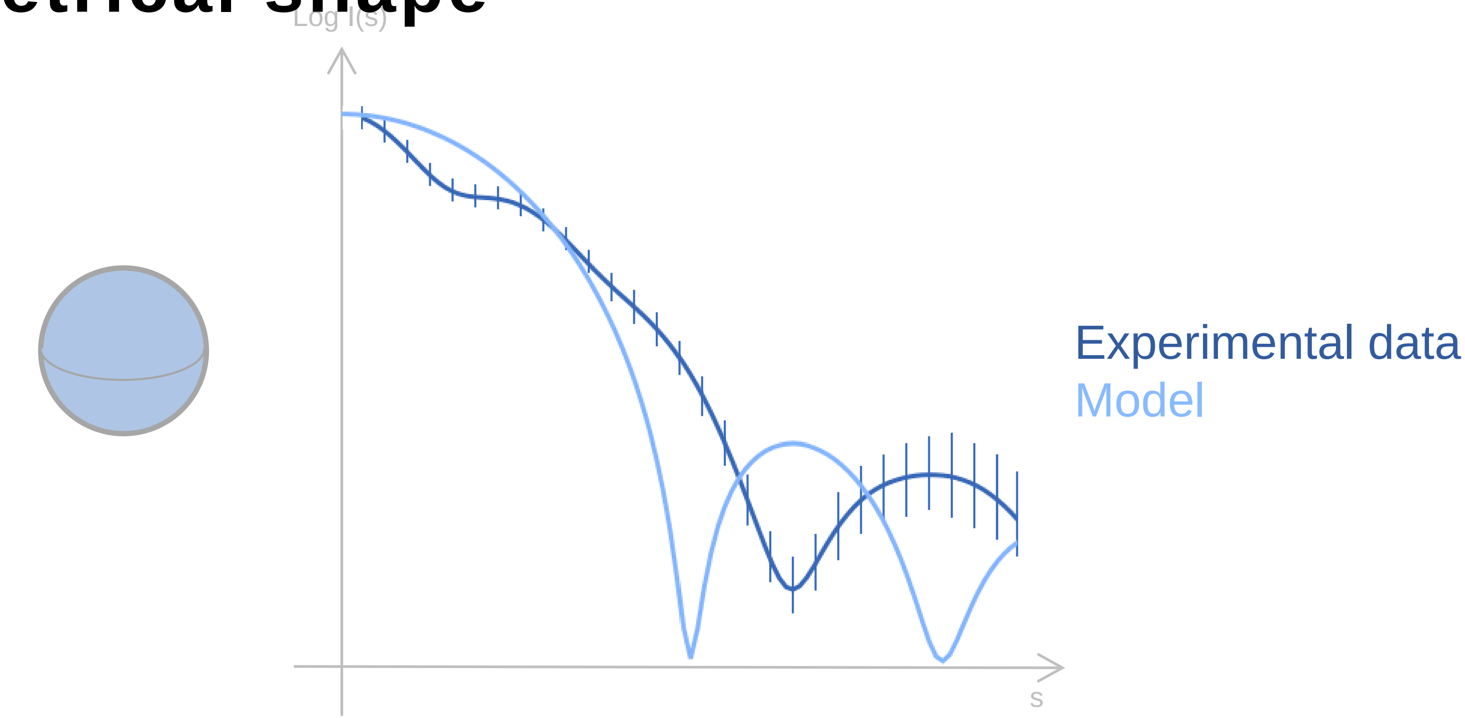
Experimental data

EMBL

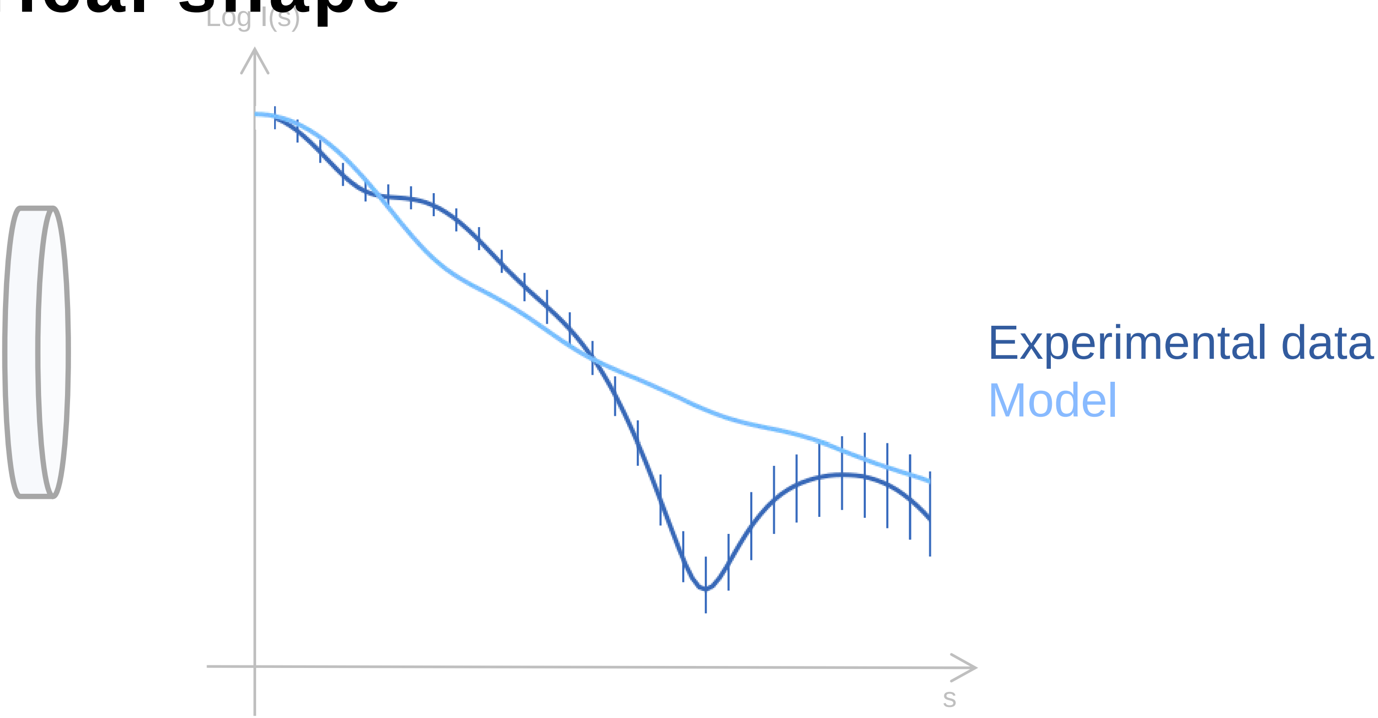
Form factor computed from simple geometrical shape



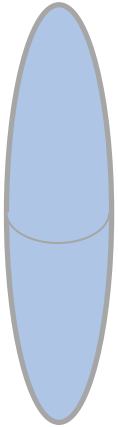
Form factor computed from simple geometrical shape

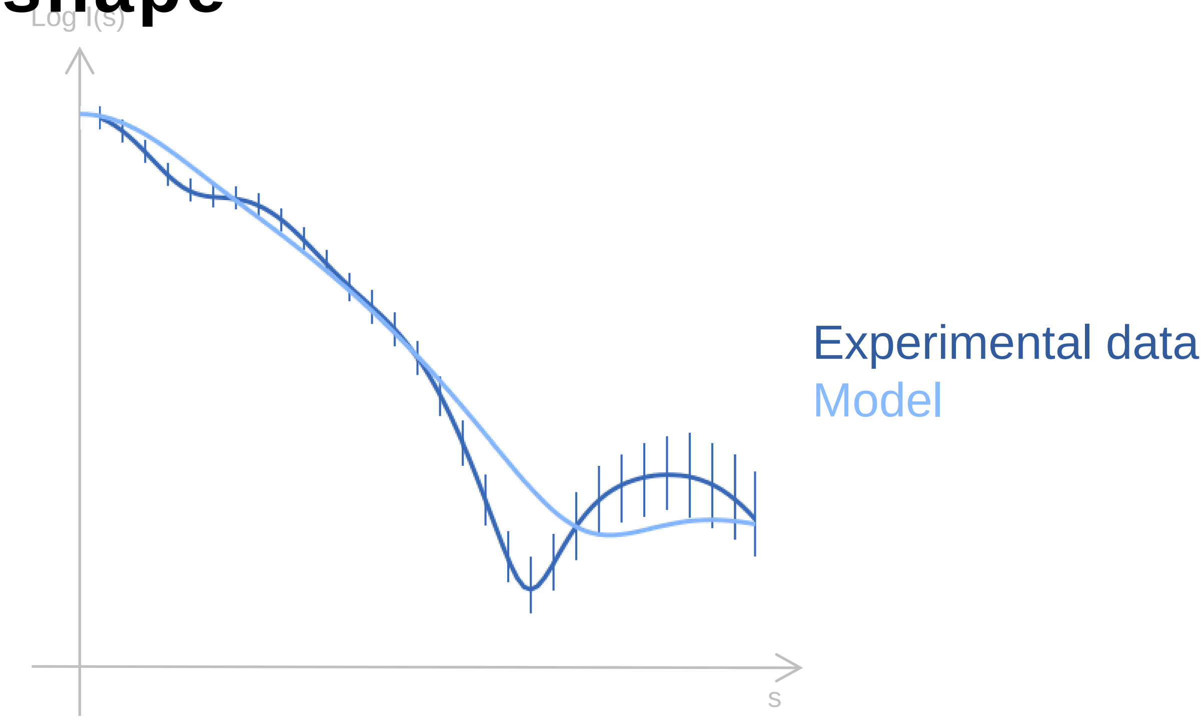


Form factor computed from simple geometrical shape

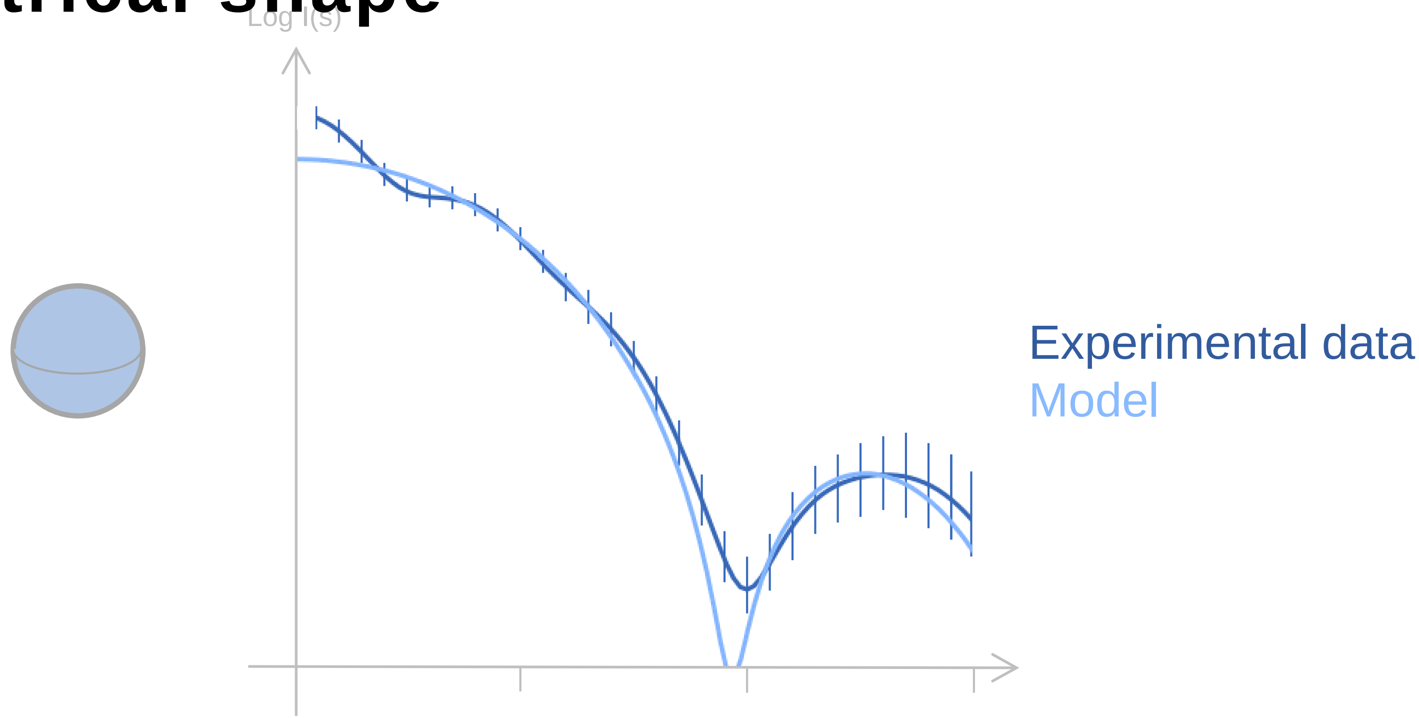


Form factor computed from simple geometrical shape

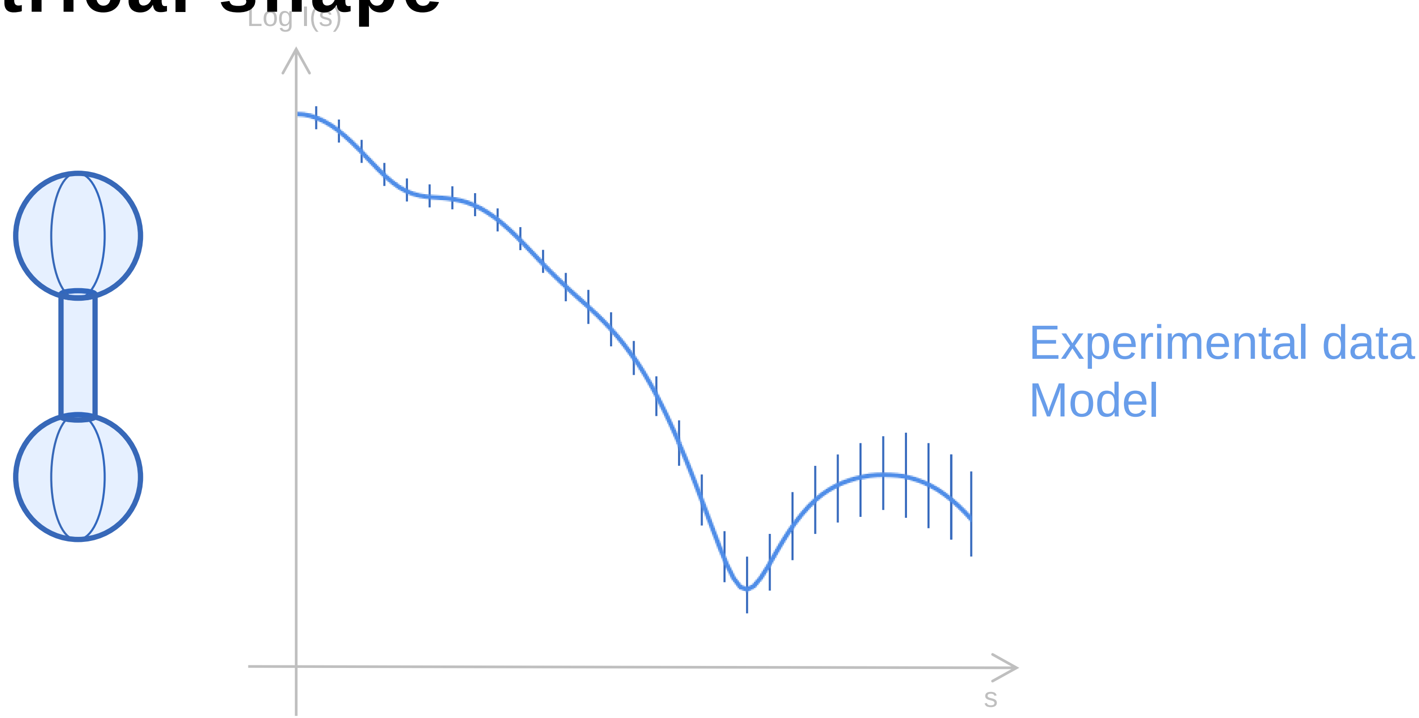




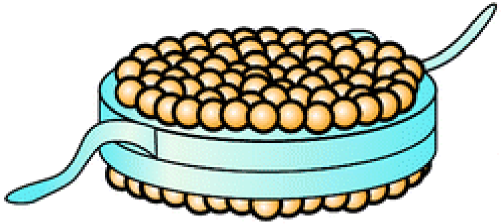
Form factor computed from simple geometrical shape

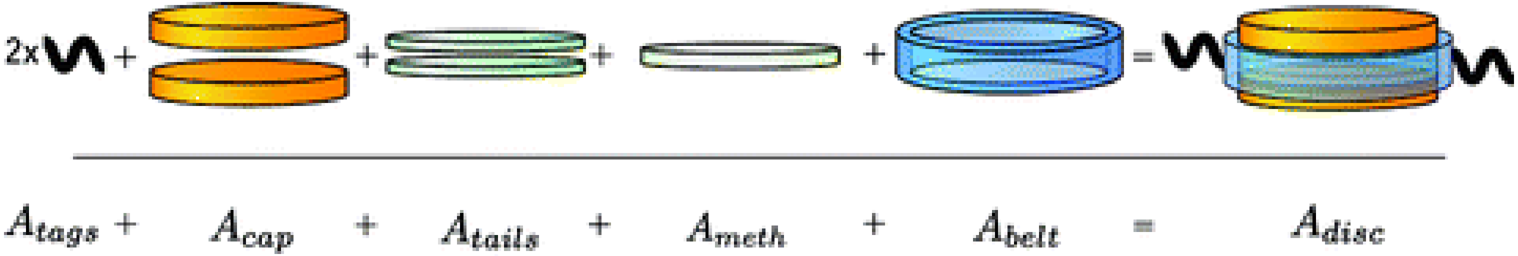


Form factor computed from simple geometrical shape



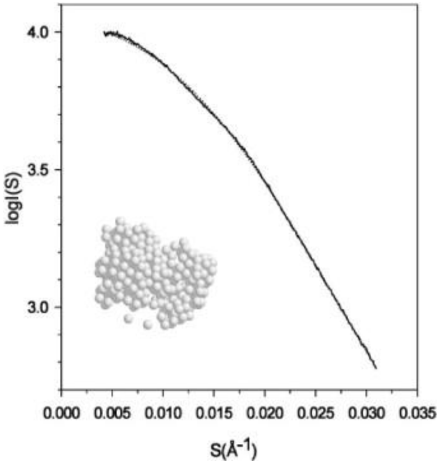
Example: modelling nano-disc





$A_{tags} + A_{cap} + A_{tails} + A_{meth} + A_{belt} = A_{disc}$

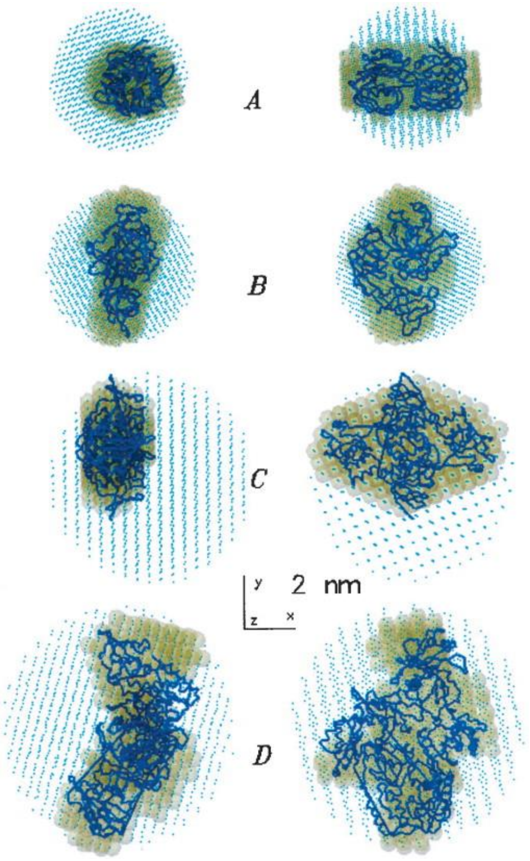
Skar-Gislinge *et al. J. Am. Chem. Soc.*

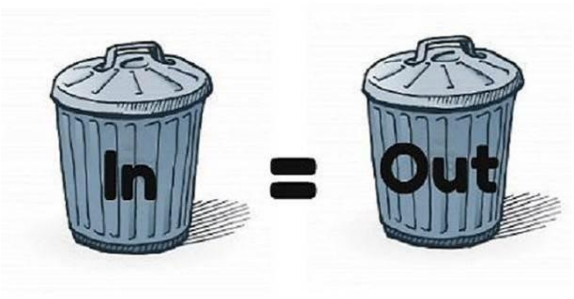


Chacón, P. *et al.*(1998)
*Biophys. J.*74, 2760-2775.
→ minimization using genetic algorithm

Bead models

Svergun, D.I. (1999)
*Biophys. J.*76, 2879-2886
→ minimisation using simulated annealing

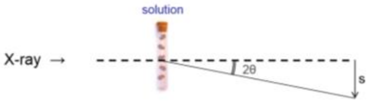




- Ab initio modelling (contrary to many other modelling approach) will always give a nice looking model that fit the data , even if the data are completely wrong.

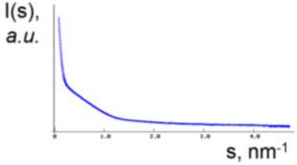
- Make sure that the SAXS curves used for ab initio correspond to the form factor of the solutes you are trying to measure.

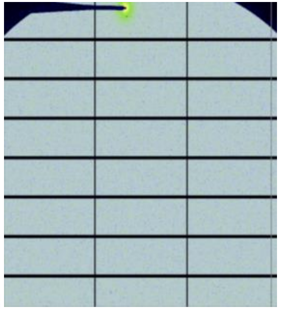
Data reduction



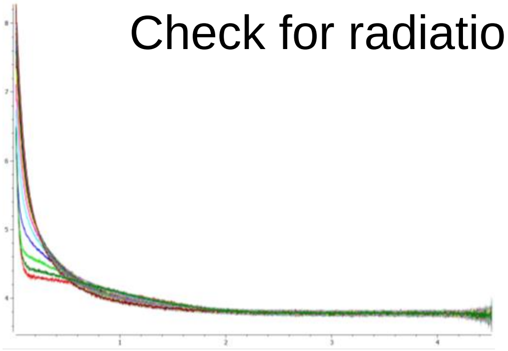
$|s| = 4\pi \sin\theta/\lambda$

2θ – scattering angle
λ – wavelength
s – scattering vector
I(s) – intensity

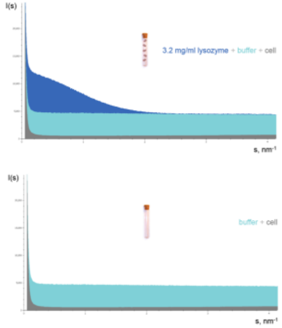


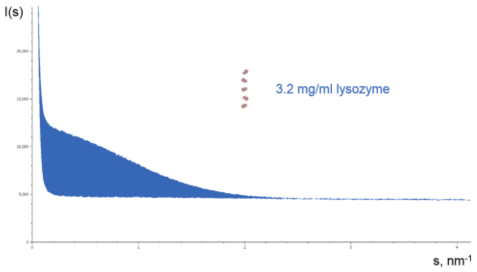


Check for radiation damage



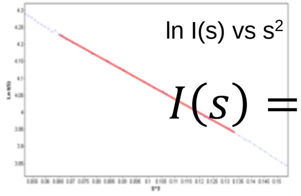
Buffer subtraction





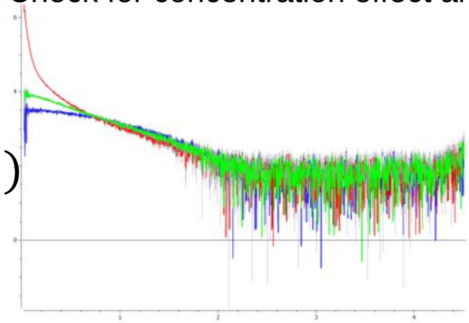
Check overall parameters, before ab initio modelling

Radius of gyration (Guinier)



$$I(s) = I_0 \exp(-\frac{1}{3}s^2{R_g}^2)$$

Check for concentration effect and aggregation



Estimation of molecular weight by forward scatteri

$$MW = \frac{I(0)}{c} \cdot \frac{(c_{st} \cdot MW_{st})}{I(0)_{st}}$$

Porod Volume

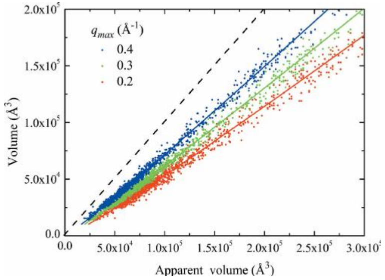
$$V_{Porod} = 2 \cdot \pi^2 \frac{I(0)}{Q}$$

Volume of correlation

$$V_C = \frac{I(0)}{\int q \cdot I(q) dq}$$

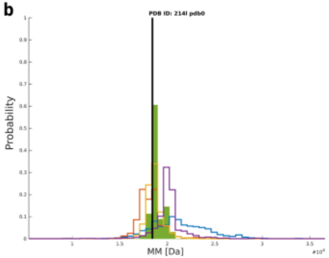
Rambo RP, Tainer JA. *Nature*. 2013

SAXS mow



Fischer, H. et al. J. Appl. Cryst. 2010

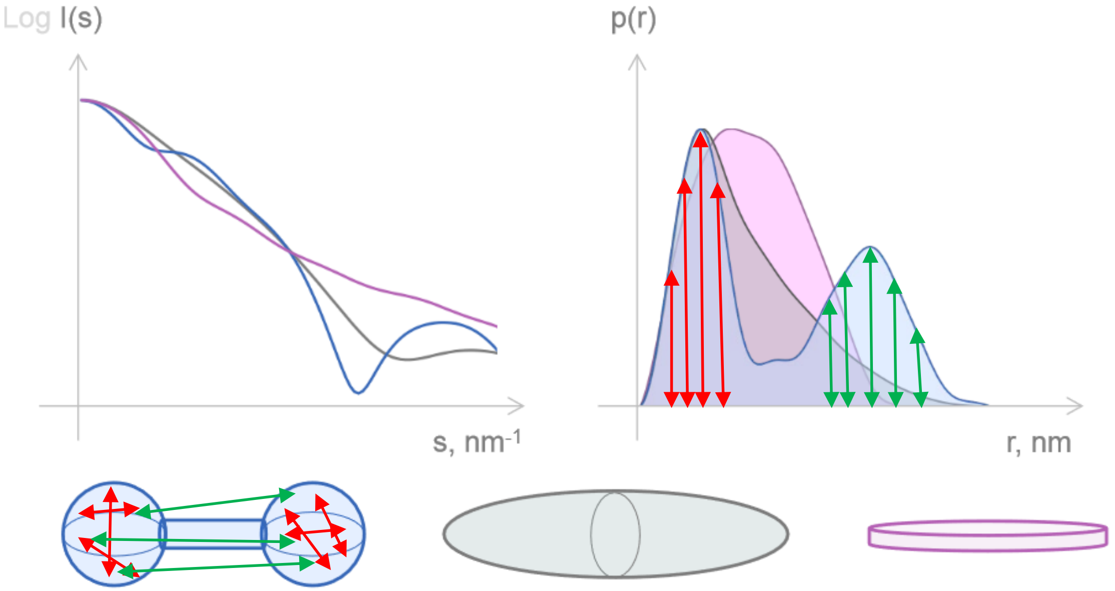
Bayesian MW



Hajizadeh NR, et al. *Sci. Rep.* 8:7204 (2018)

EMBL

Distance distribution function

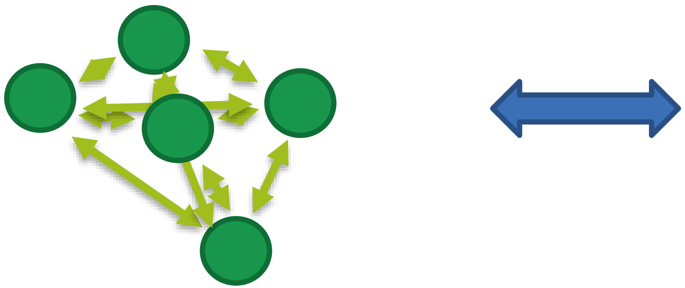


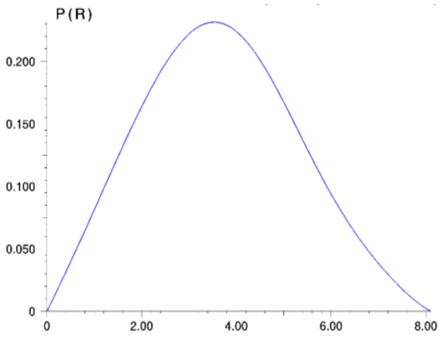
$$p(r) = \rho^2 \gamma_0(r) V r^2$$

Where $\gamma_0(r)$ is the probability of finding a point within the particle at a distance r from a given points.

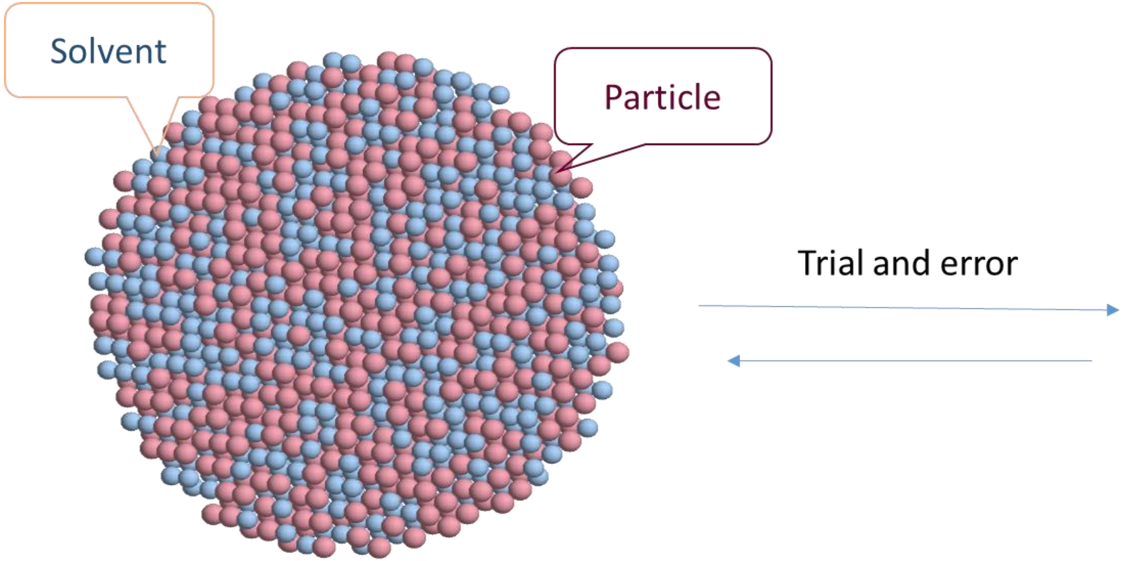
Ab initio bead modelling: Basic idea

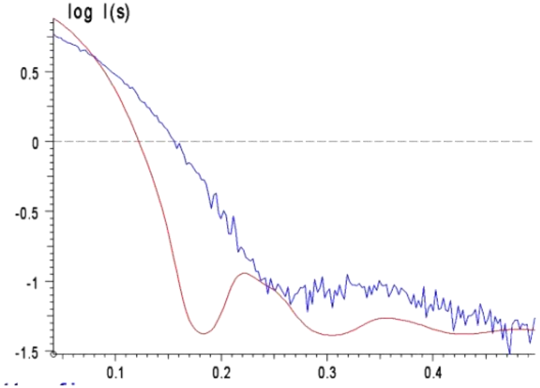
- Find an ensemble of beads with the inter-bead distances are consistent with the p(r)





Beads on a grid





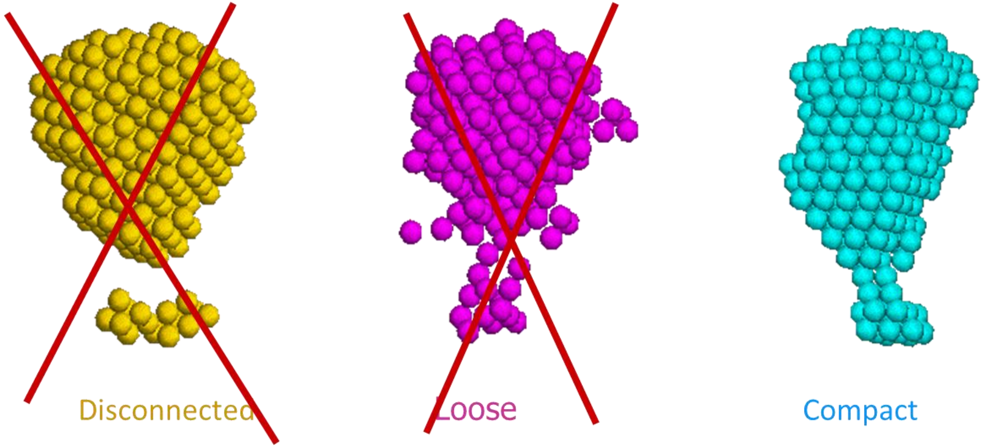
Computation of the theoretical SAXS curve from the bead ensemble and fit to the experimental SAXS data.

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

EMBL

Penalty terms

- Bead configuration should not only fit the data but also provide a compact model. This can be reinforce by the use of penalty terms.
- The looseness penalty term is computed from the bead configuration and is small when the bead ensemble has a compact configuration

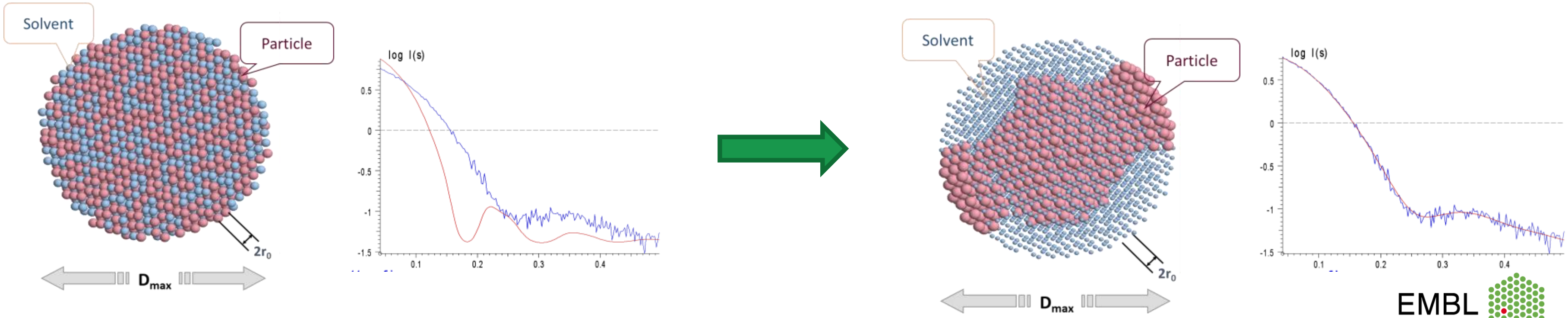


Finding good bead ensemble

Find the bead ensemble that minimized the target function:

$$f(X) = \chi^2 + \alpha \cdot P(X)$$

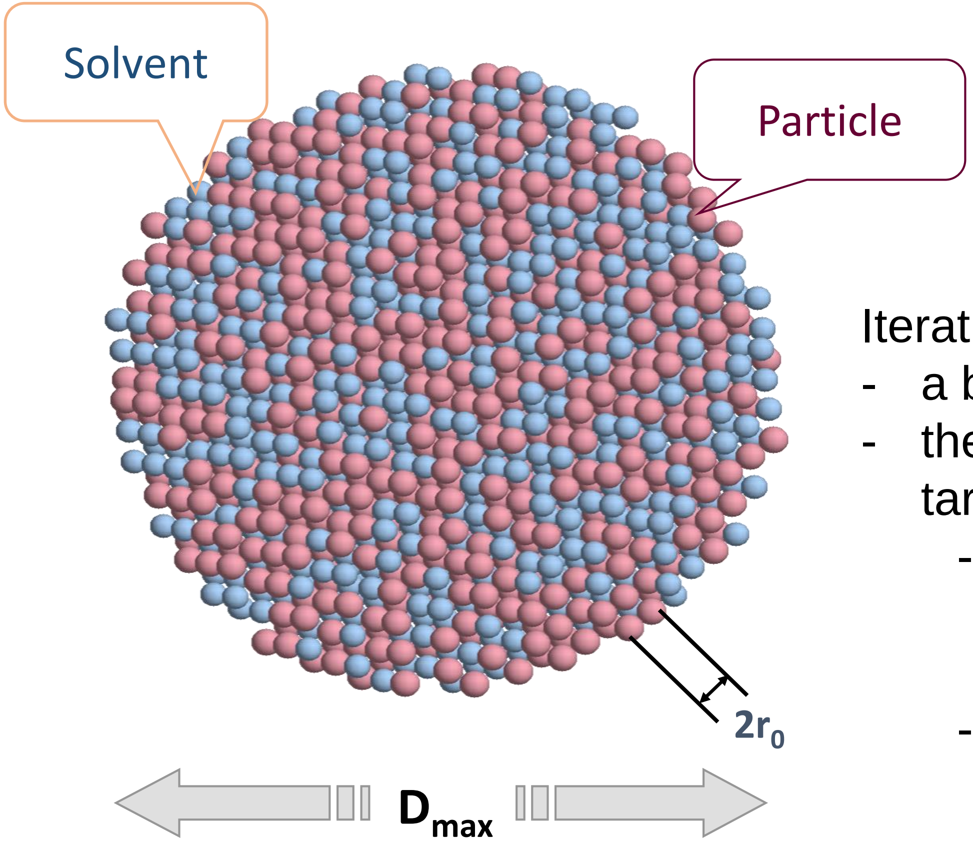
$$\chi^2 = \frac{1}{N-1} \sum_j \left[\frac{I_{exp}(s_j) - cI(s_j)}{\sigma(s_j)} \right]^2$$



EMBL

Minimization of the target function

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



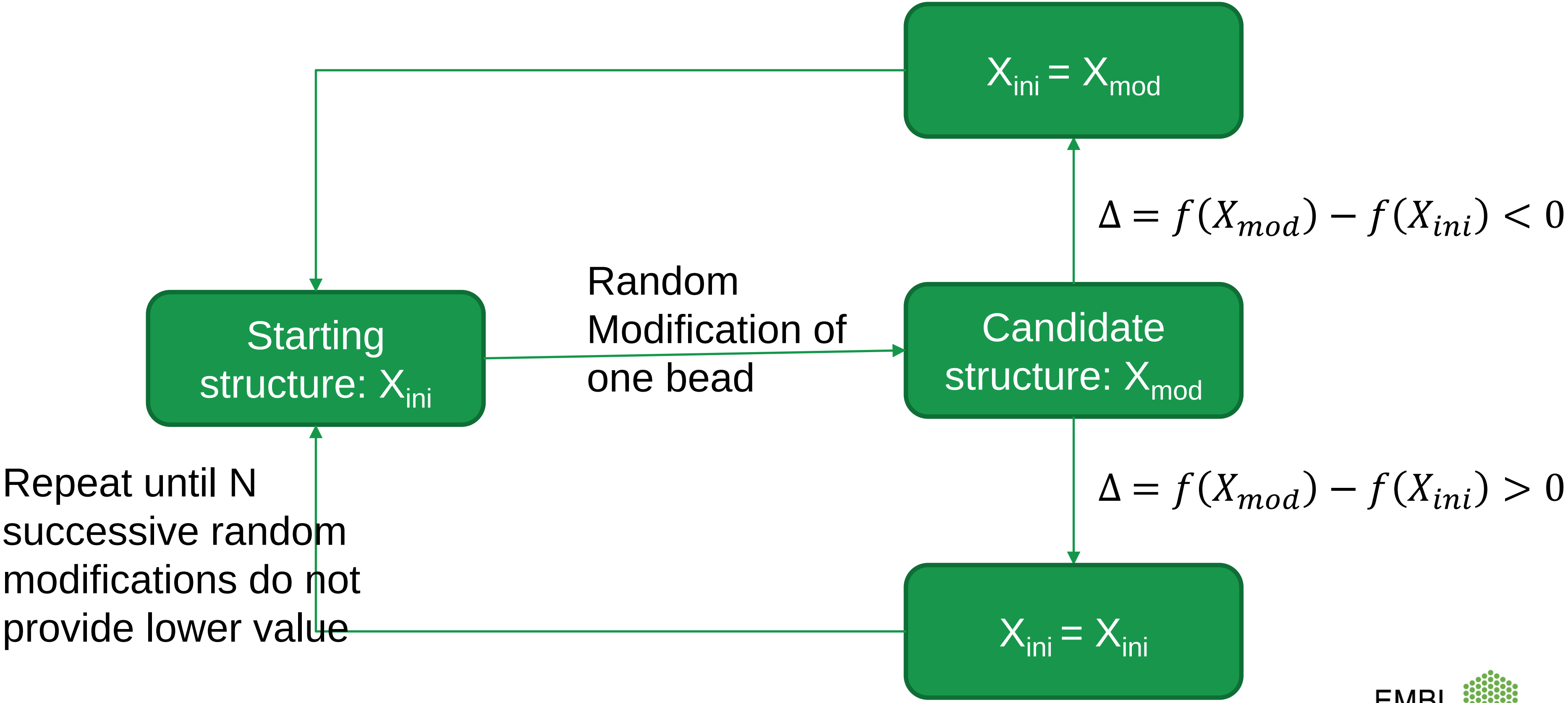
$$f(X) = \chi^2 + \alpha \cdot P(X)$$

Iterative approach:

- a bead can be changed
- the effect of this change is evaluated: is the target function smaller after this change?
  - If yes, the changed structure is the new starting configuration for the next iteration.
  - If not, the unchanged structure is used.

EMBL

Pure Monte Carlo



Starting structure: X_{ini}

Random Modification of one bead

Candidate structure: X_{mod}

$\Delta = f(X_{mod}) - f(X_{ini}) < 0$

$X_{ini} = X_{mod}$

$\Delta = f(X_{mod}) - f(X_{ini}) > 0$

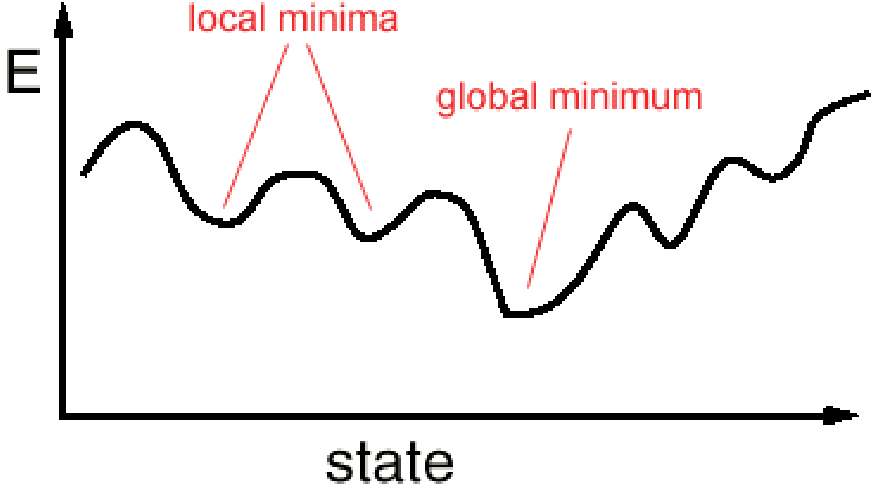
$X_{ini} = X_{ini}$

Repeat until N successive random modifications do not provide lower value

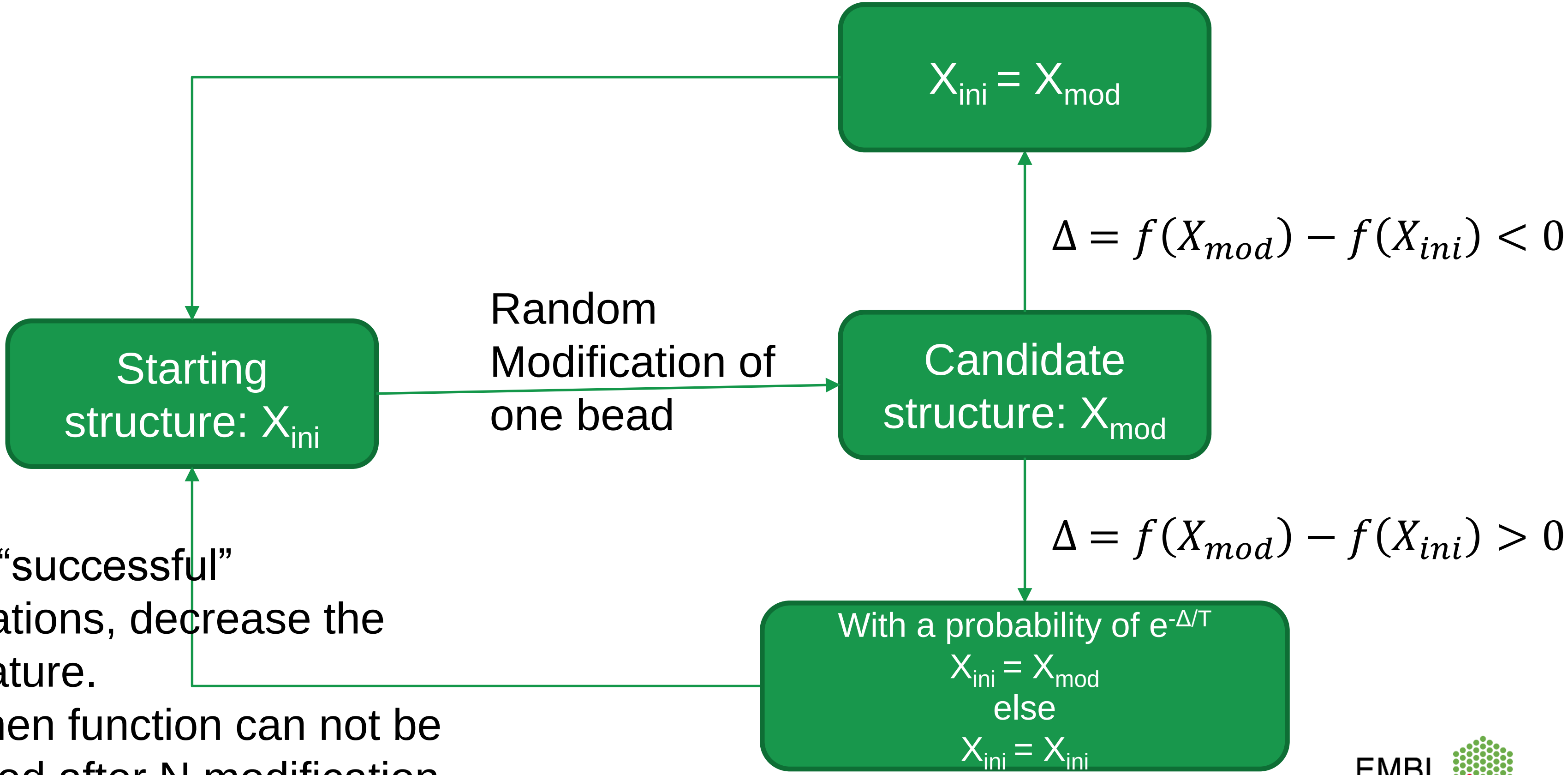
Local Minima vs global minimum

Local search can be trapped in a local minimum.
Pure Monte-Carlo search always goes to the closest local minimum (nature: rapid quenching and vitreous ice formation)

To get out of local minima, global search must be able to (sometimes) go to a worse point.
Slower annealing allows to search for a global minimum (nature: normal, e.g. slow freezing of water and ice formation)



Simulated annealing



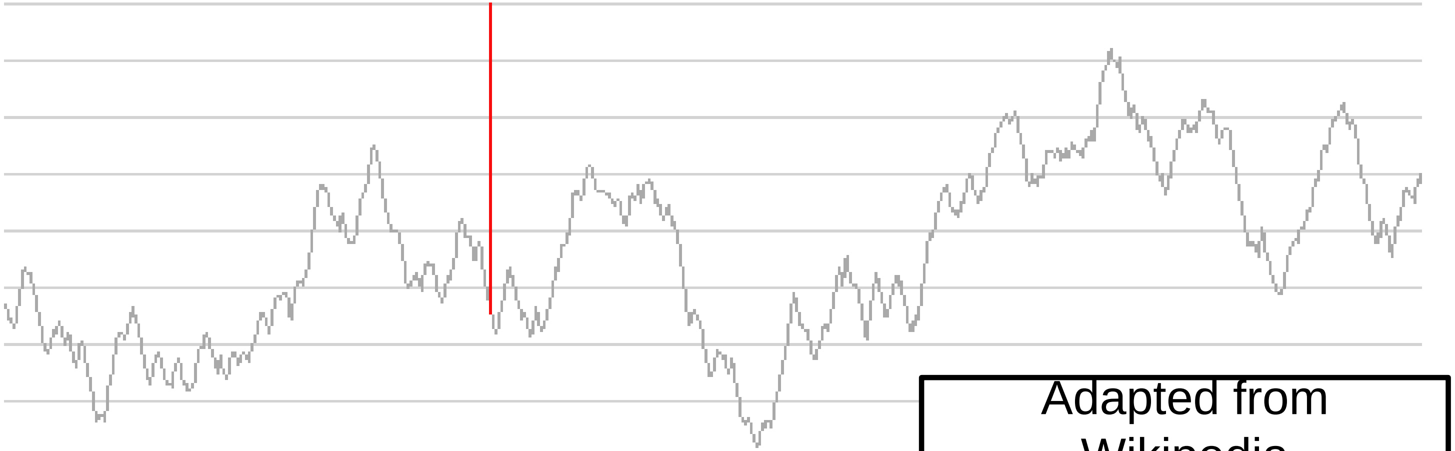
$X_{ini} = X_{mod}$
Starting structure: X_{ini}
Random Modification of one bead
Candidate structure: X_{mod}
With a probability of $e^{-\Delta/T}$
$X_{ini} = X_{mod}$
else
$X_{ini} = X_{ini}$

$$\Delta = f(X_{mod}) - f(X_{ini}) < 0$$

$$\Delta = f(X_{mod}) - f(X_{ini}) > 0$$

Repeat.
after M “successful” modifications, decrease the temperature.
Stop when function can not be minimized after N modification.

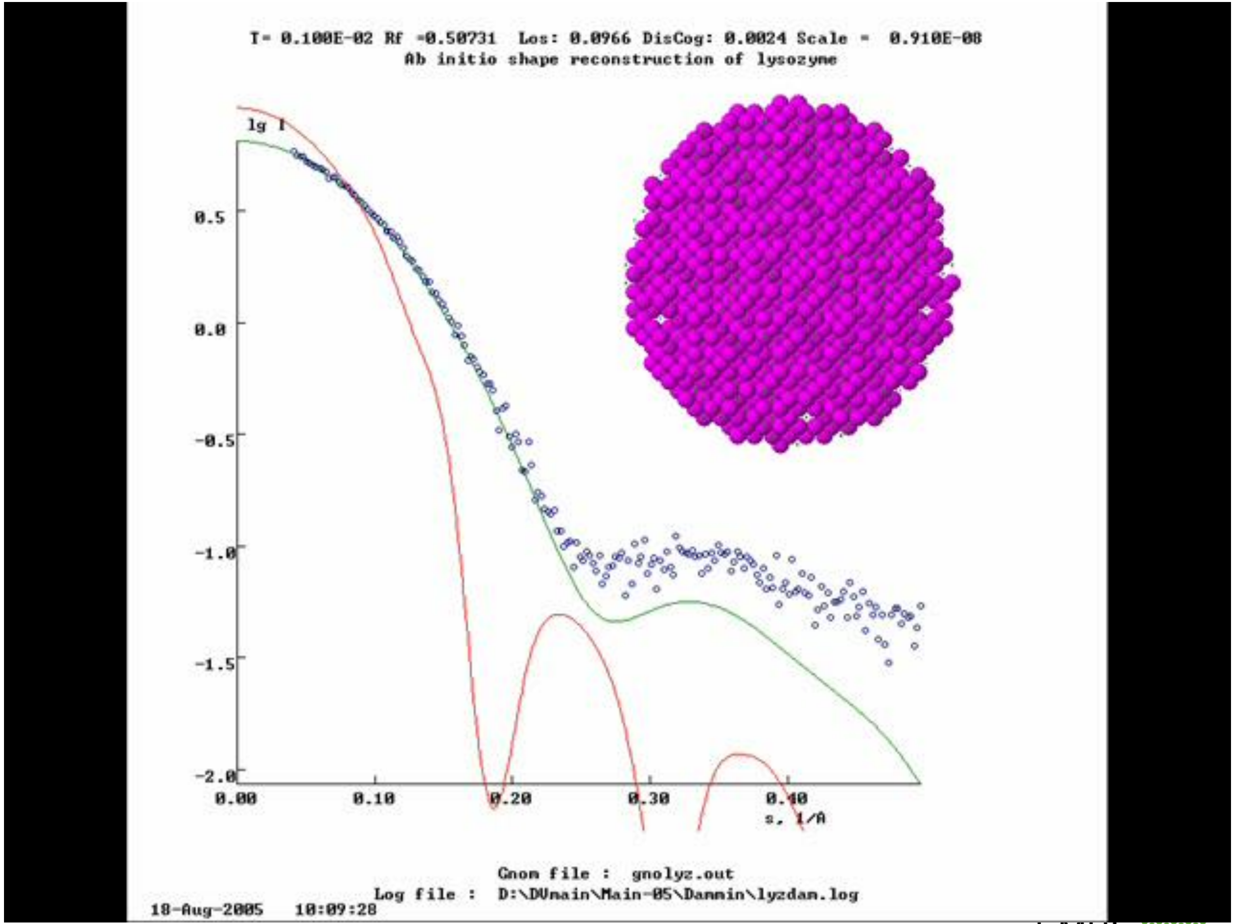
Simulated annealing



Adapted from Wikipedia

Ab initio program

DAMMIN

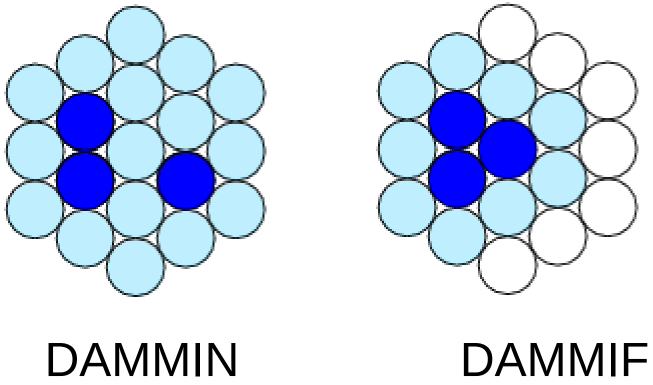


```
T= 0.100E-02 Rf =0.50731  Los: 0.0966 DisCog: 0.0024 Scale =  0.910E-08
               Ab initio shape reconstruction of lysozyme
```

```
                              Gnom file :  gnolyz.out
                    Log file :  D:\DVmain\Main-05\Dammin\lyzdam.log
18-Aug-2005    10:09:28
```

DAMMIF

- Reimplementation of DAMMIN written in object oriented code
- About 25 to 40 times faster (about 1-2 min for fast run on a PC)
- Make use of multiple CPU
- Use adaptive search volume

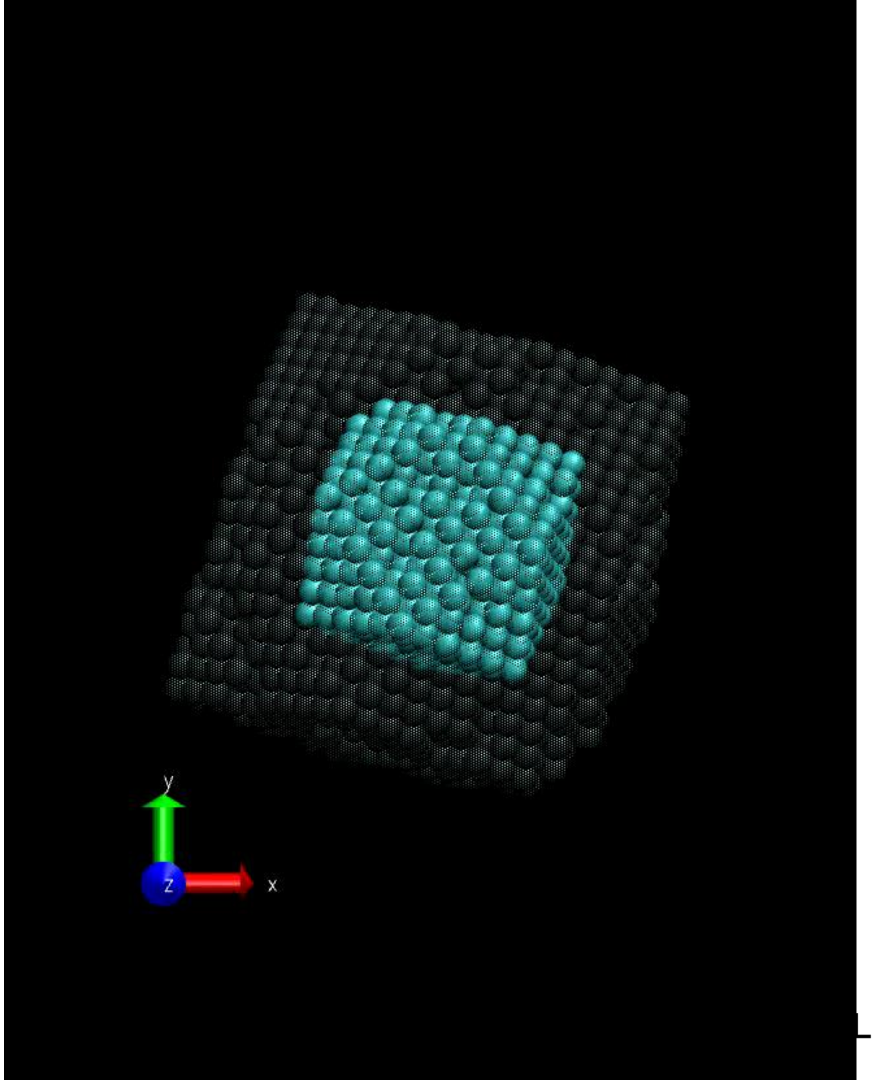


At the current iteration:

- dark blue particle, might become solvent
- light blue solvent, might become particle
- white solvent, won’t change

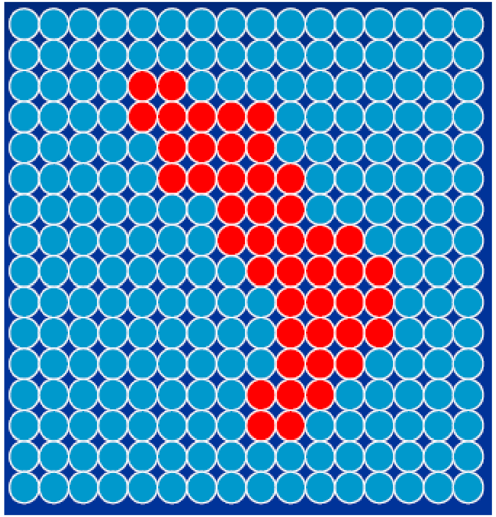
Franke, D. & Svergun, D. I. (2009) *J. Appl. Cryst.* **42**, 342–346

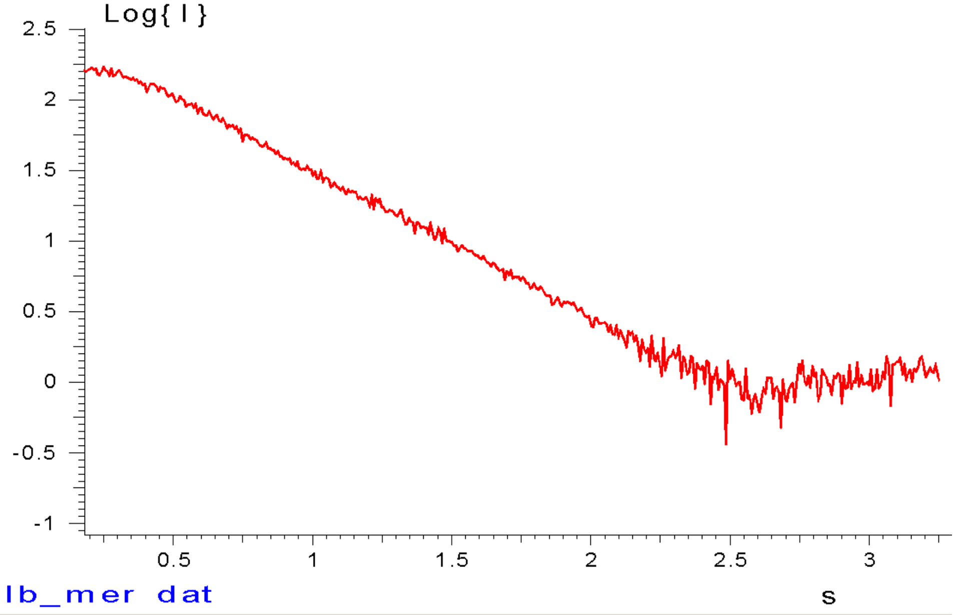
DAMMIF in action



Shape analysis for multi-component systems: principle

MONSA



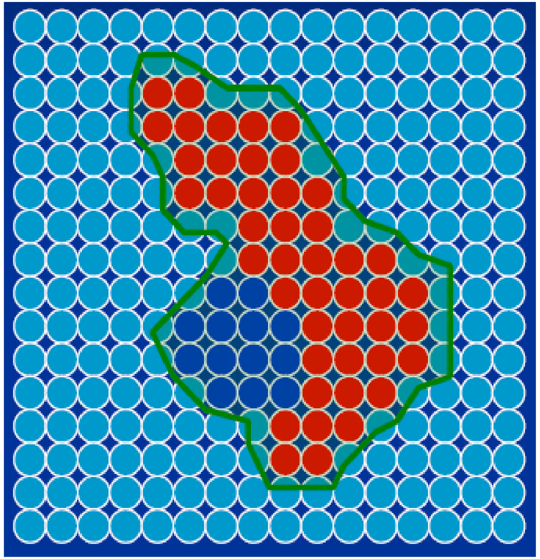


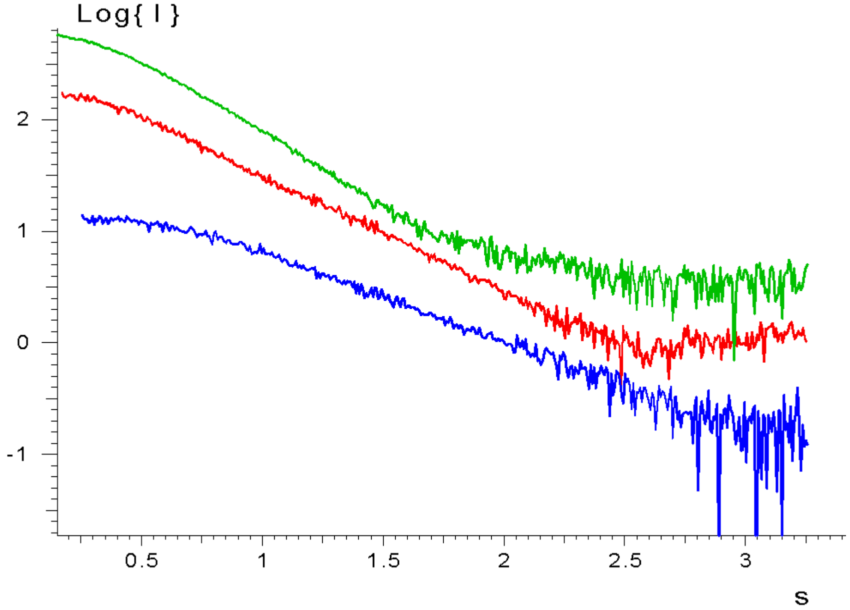
One component, one scattering pattern:
“normal” shape determination

Chacón, P. et al. (1998) Biophys. J. 74, 2760-2775

Shape analysis for multi-component systems: principle

MONSA





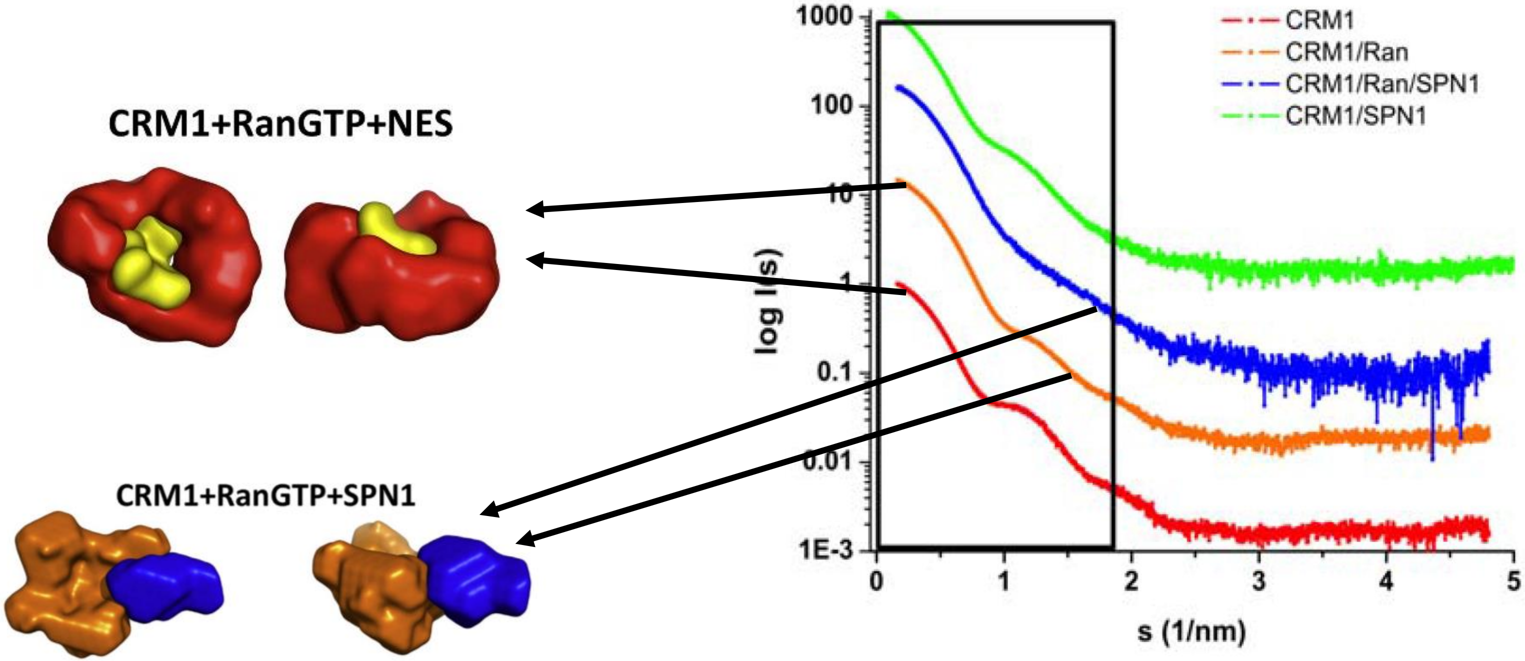
Many components, many scattering patterns: shape and internal structure

MONSA

Svergun, D.I. (1999) Biophys. J. **76**, 2879-2886

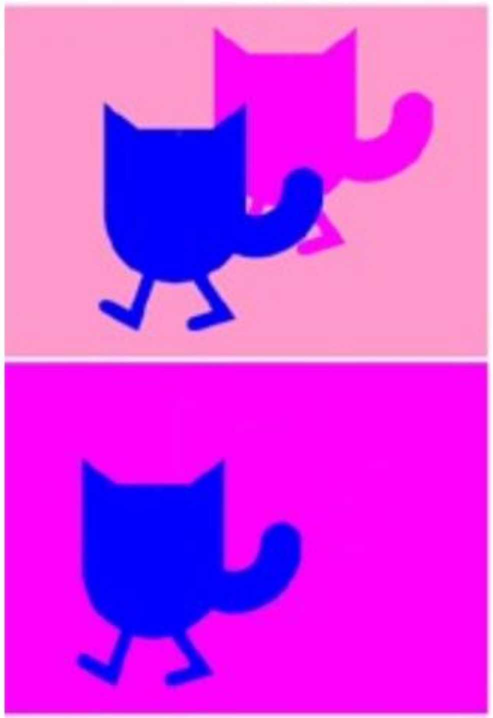
Svergun, D.I. & Nierhaus, K.H. (2000) J.

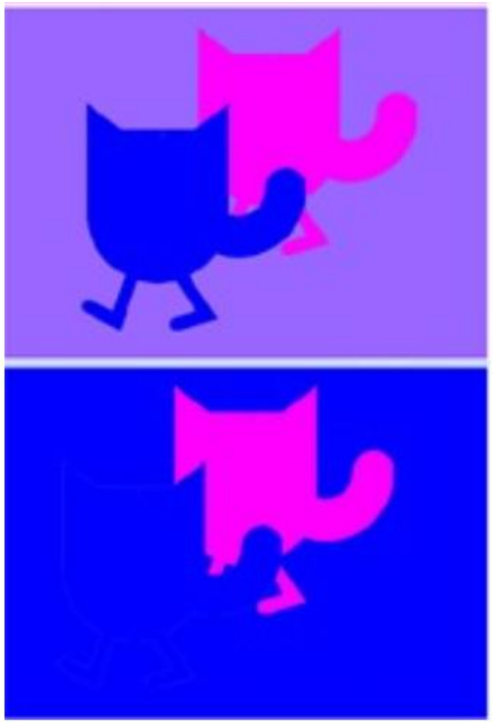
Example multi-component system



Dölker, Blanchet et al., Structure 2013

This approach is very useful for contrast matched data.

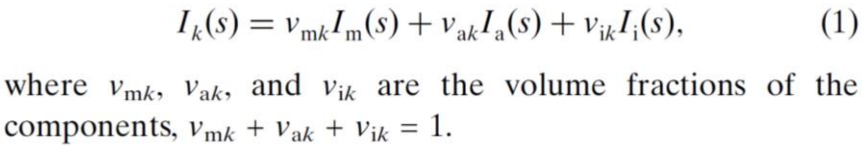




Konarev, P. V. & Svergun, D. I. (2018). IUC

DAMMIX

- Dummy atom modelling on mixture with known volume fraction

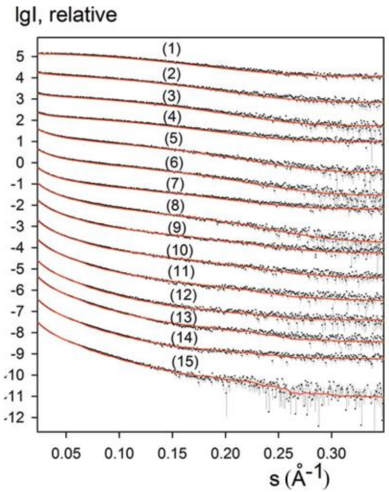


$$I_k(s) = v_{\mathrm{m}k} I_{\mathrm{m}}(s) + v_{\mathrm{a}k} I_{\mathrm{a}}(s) + v_{\mathrm{i}k} I_{\mathrm{i}}(s), \qquad (1)$$

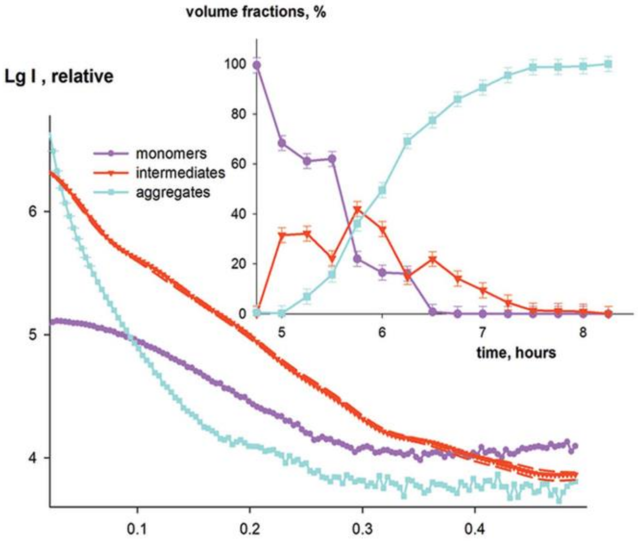
where $v_{\mathrm{m}k}$, $v_{\mathrm{a}k}$, and $v_{\mathrm{i}k}$ are the volume fractions of the components, $v_{\mathrm{m}k} + v_{\mathrm{a}k} + v_{\mathrm{i}k} = 1$.

$$F(X) = \chi^2(X) + P(X),$$

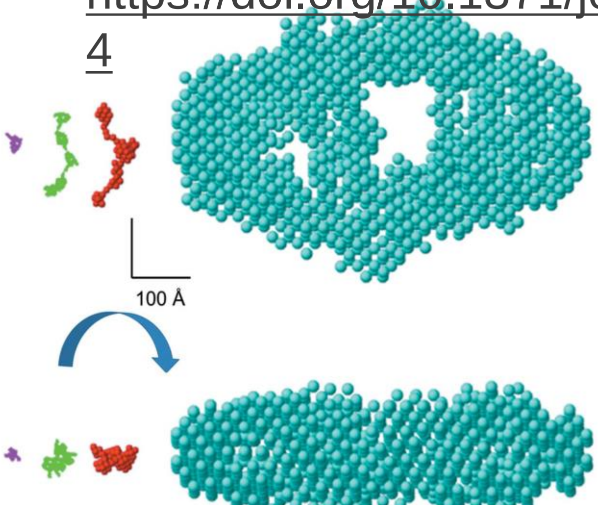
$$F(X) = \sum_{k=1}^{K} \chi_k^2(X) + \sum_j W_j \times P_j(X).$$



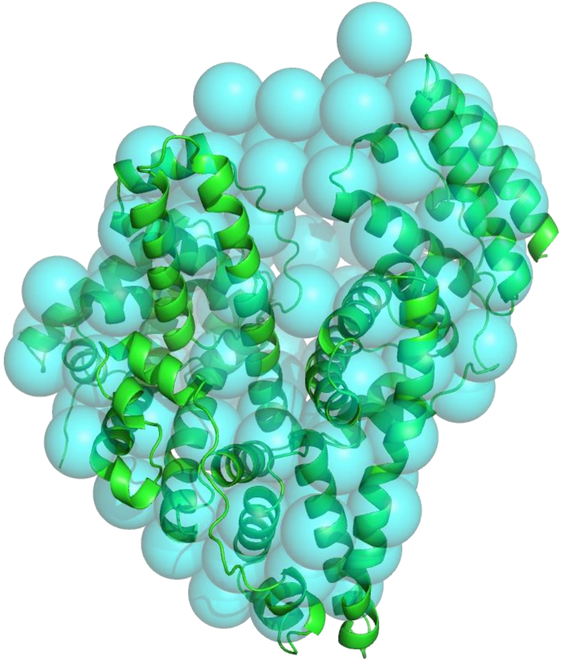
Konarev, P. V. & Svergun, D. I. (2018). IUC

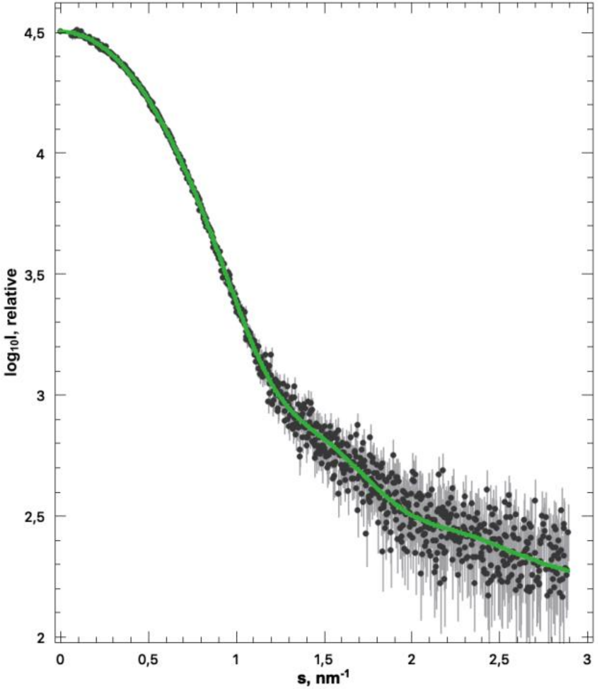


Vestergaard B, Groenning M, Roessle M, Kastrup JS, de Weert Mv, et al. (2007) A Helical Structural Nucleus Is the Primary Elongating Unit of Insulin Amyloid Fibrils . PLOS Biology 5(5): e134. https://doi.org/10.1371/journal.pbio.0050134



DATMIF





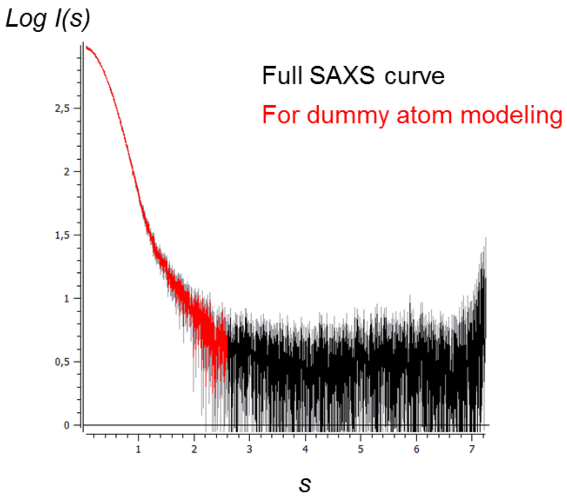
- Fit experimental data directly
- Debeye formula
- Only penalty: minimize surface area
- Runtime: minutes

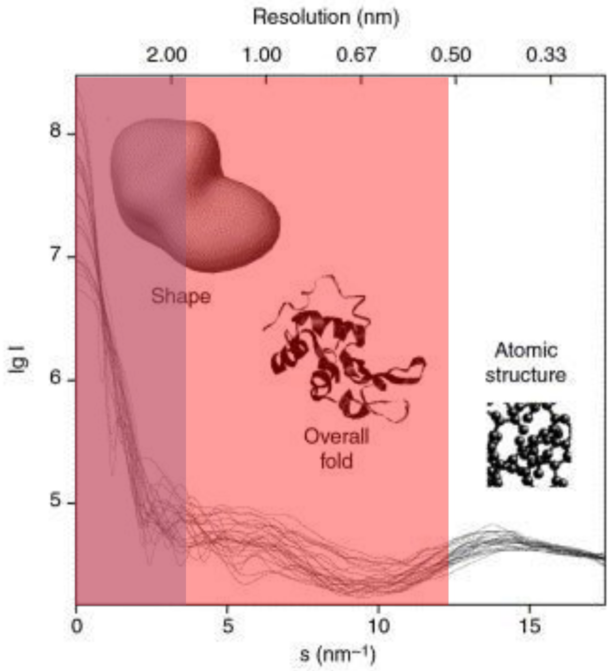
- Example:
  - BSA monomer from SEC-SAXS
  - 967 experimental data points
  - Model superposition to monomer of 4F5S
  - Fit: red. χ^2=0.9
  - MW: ~70 kDa
  - D_{max}: ~9.0 nm

EMBL

Resolution limit of dummy atom model

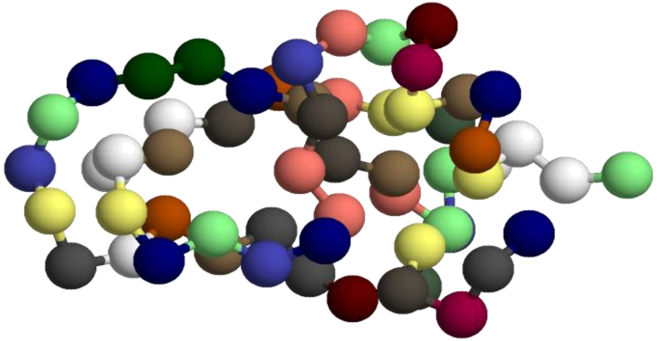
- For dummy atom models, the electron density within the protein is considered as homogeneous





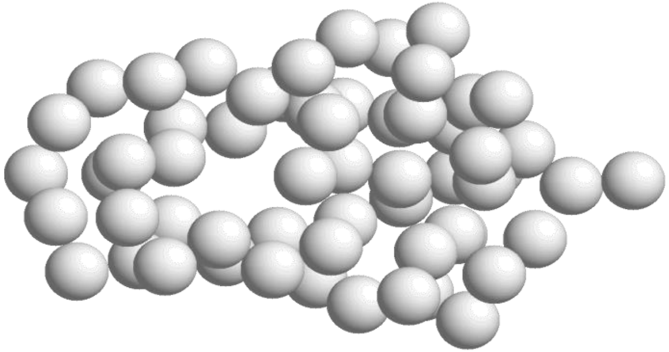
Dummy residue models

- 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)



Dummy residue models

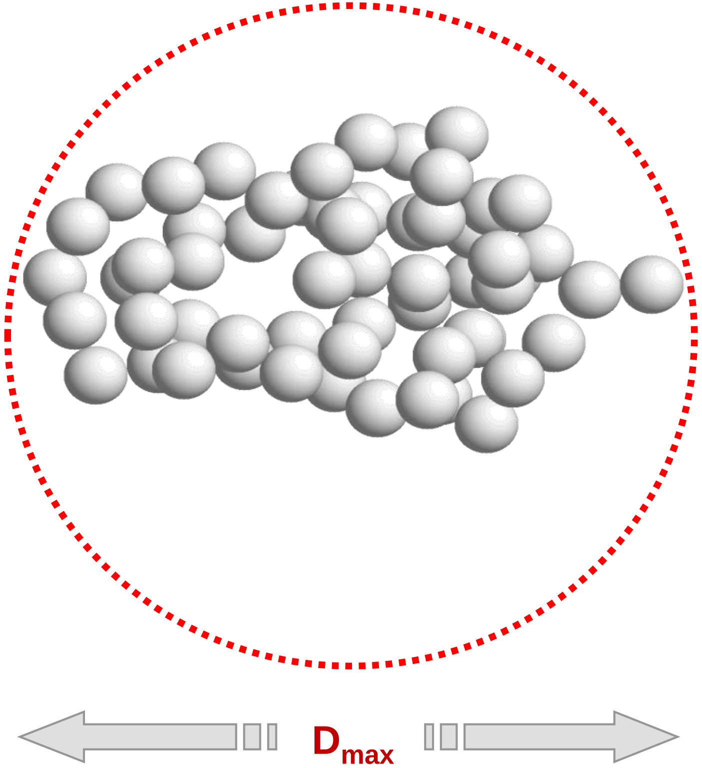
- Proteins typically consist of folded polypeptide chains composed of amino acid residues
- At a resolution of 0.5 nm each amino acid can be represented as one entity (dummy residue)
- In GASBOR a protein is represented by an ensemble of *K* dummy residues that are
  - Identical
  - Have no ordinal number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



Dummy residue models

GASBOR

- GASBOR finds coordinates of *K* dummy residues within its search volume (red)

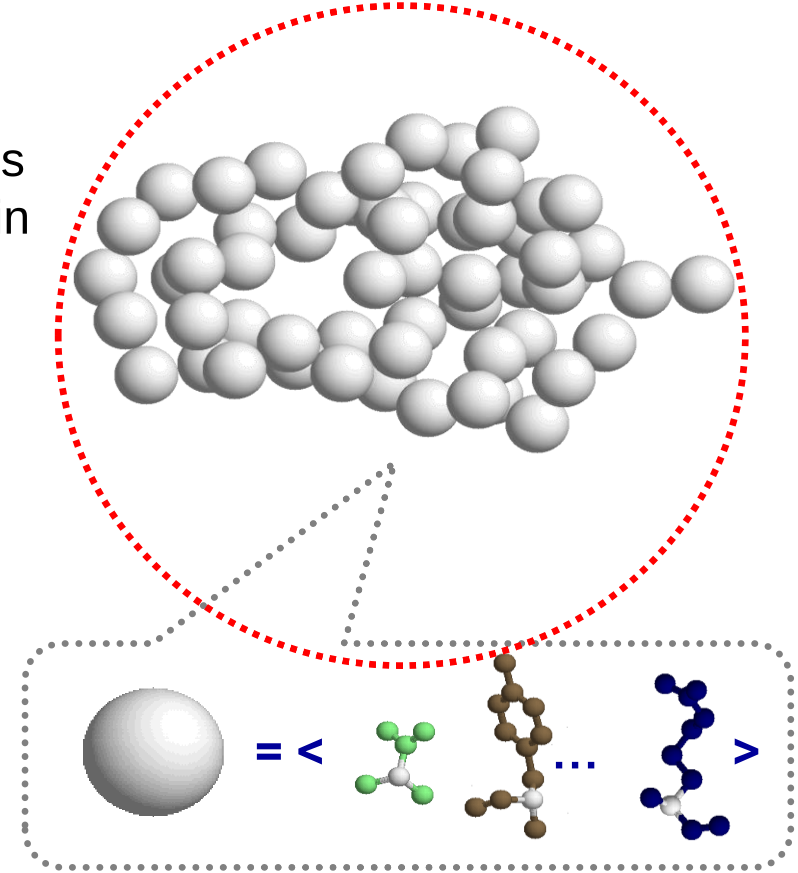


EMBL

Dummy residue models

GASBOR

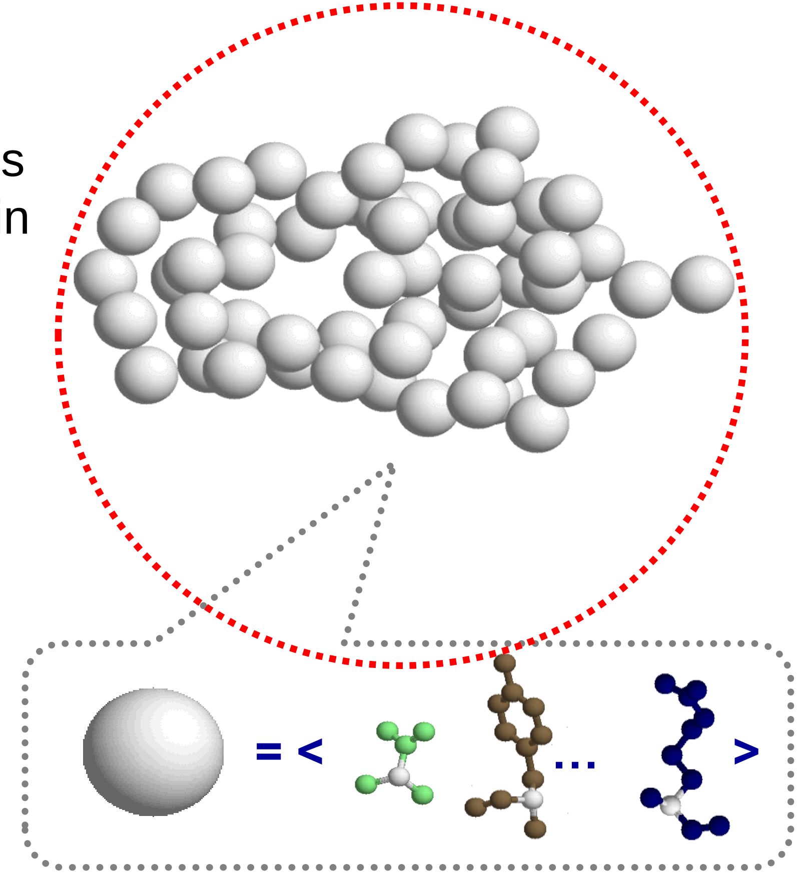
- GASBOR finds coordinates of *K* dummy residues within its search volume (red)
- Requires polypeptide chain-compatible arrangement of dummy residues



Dummy residue models

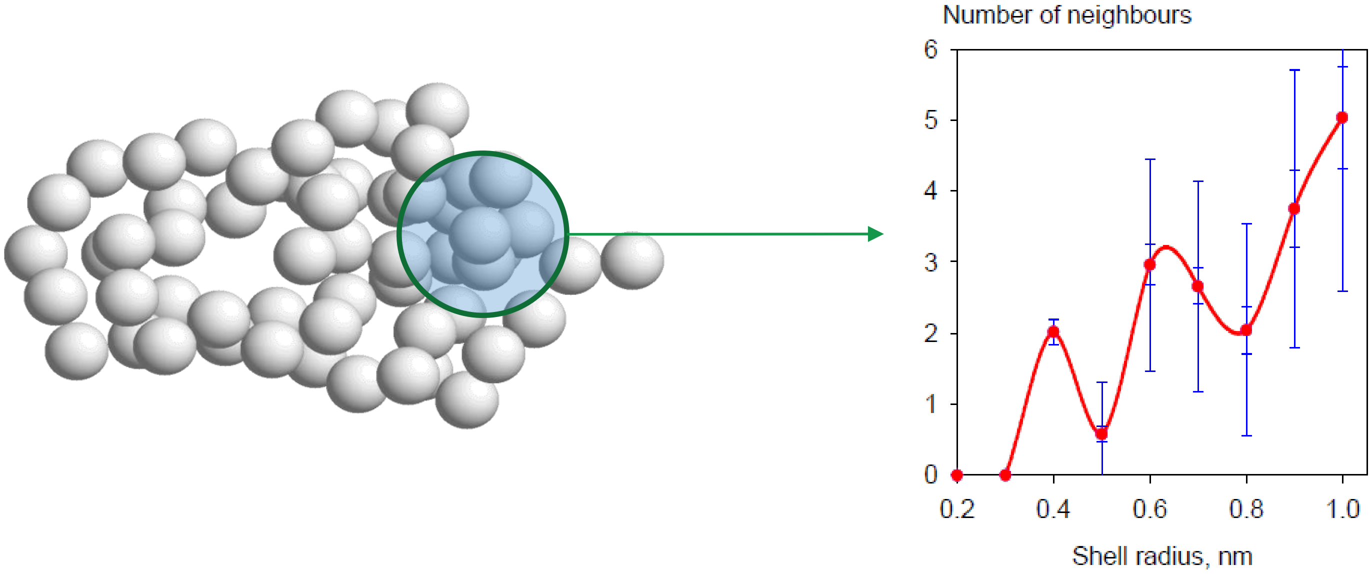
GASBOR

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



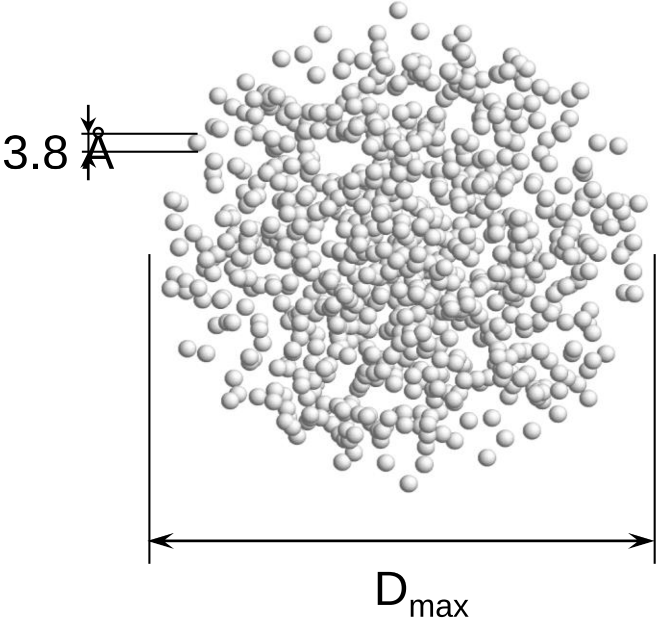
Distribution of neighbours

- Excluded volume effects and local interactions lead to a characteristic distribution of nearest neighbors around a given residue in a polypeptide chain



EMBL

Ab initio reconstruction: dummy residue modelling

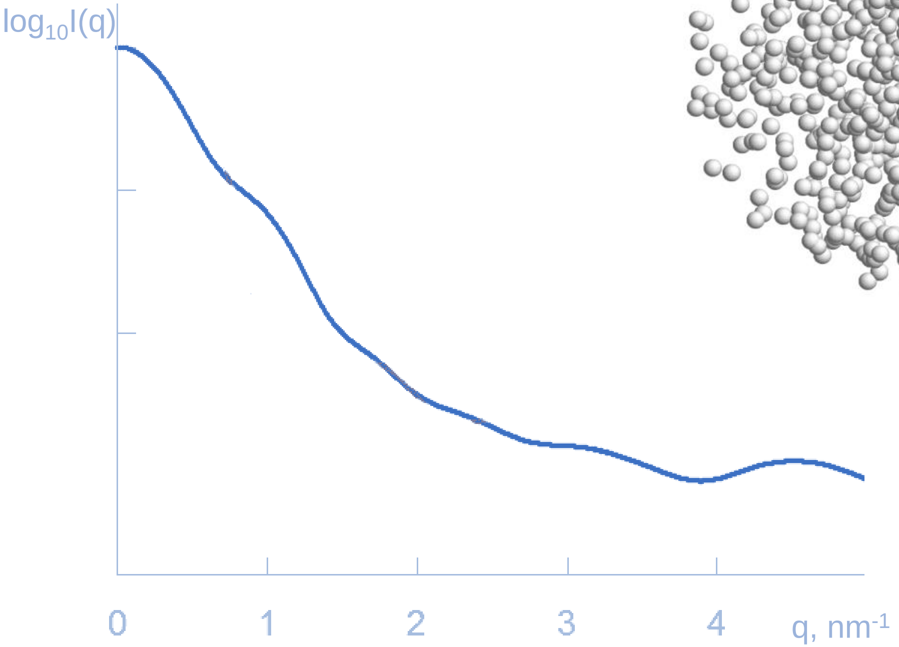


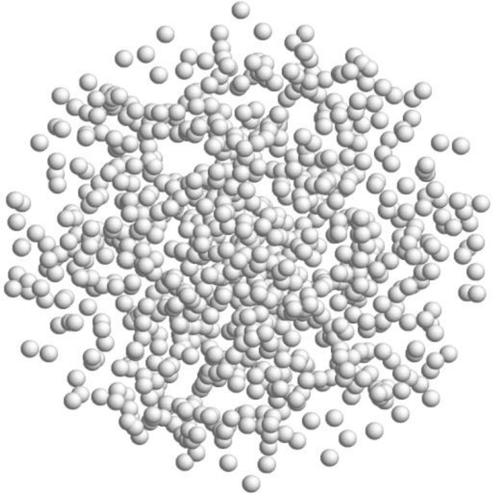
GASBOR

Svergun, D.I., Petoukhov, M.V, Koch, M.H.J. (2001)

Biophys J 80, 2946–2953.

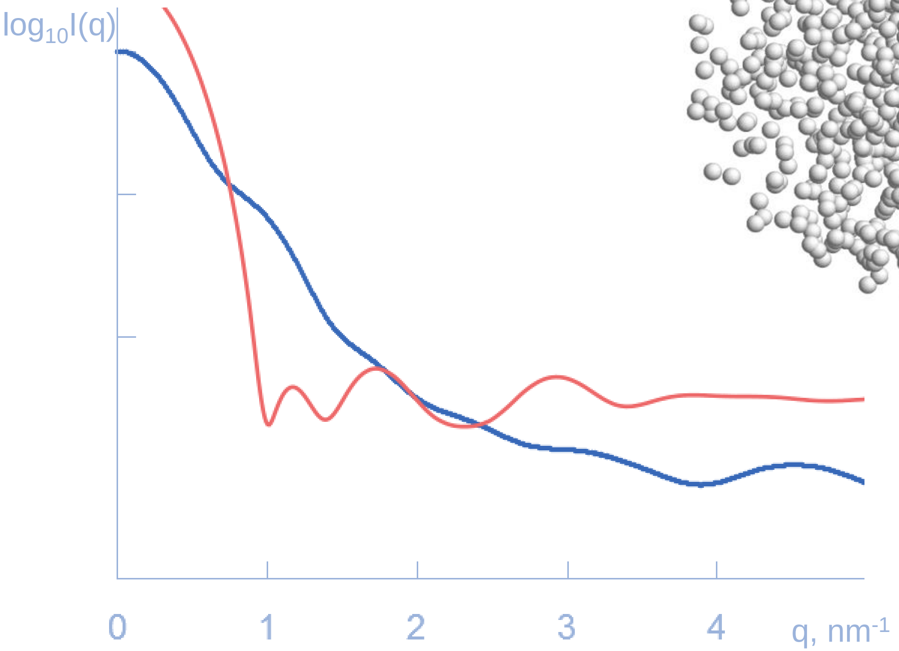
Ab initio reconstruction: dummy residue modelling

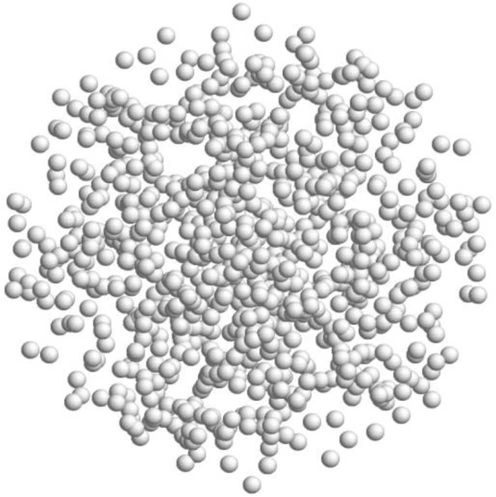




EMBL

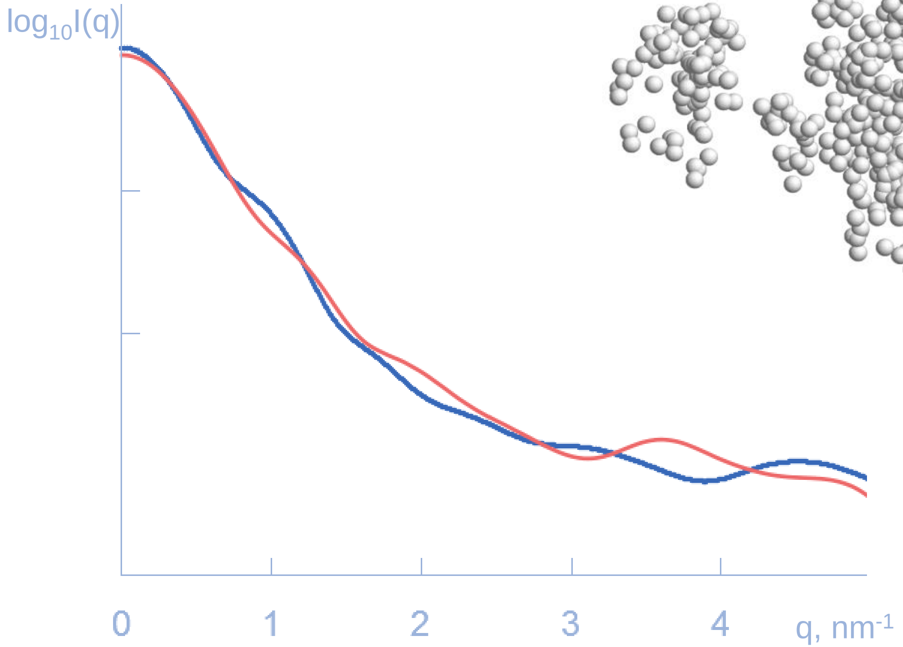
Ab initio reconstruction: dummy residue modelling

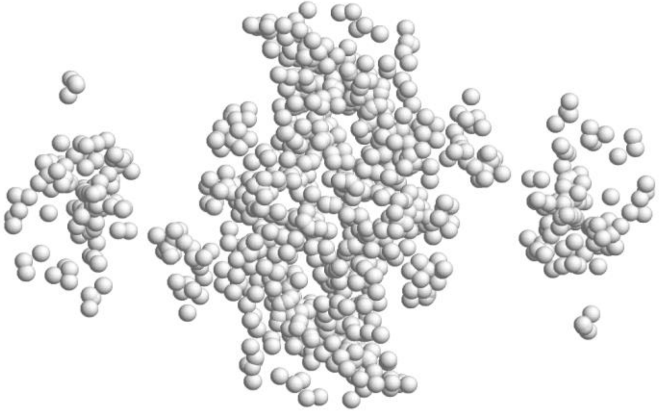




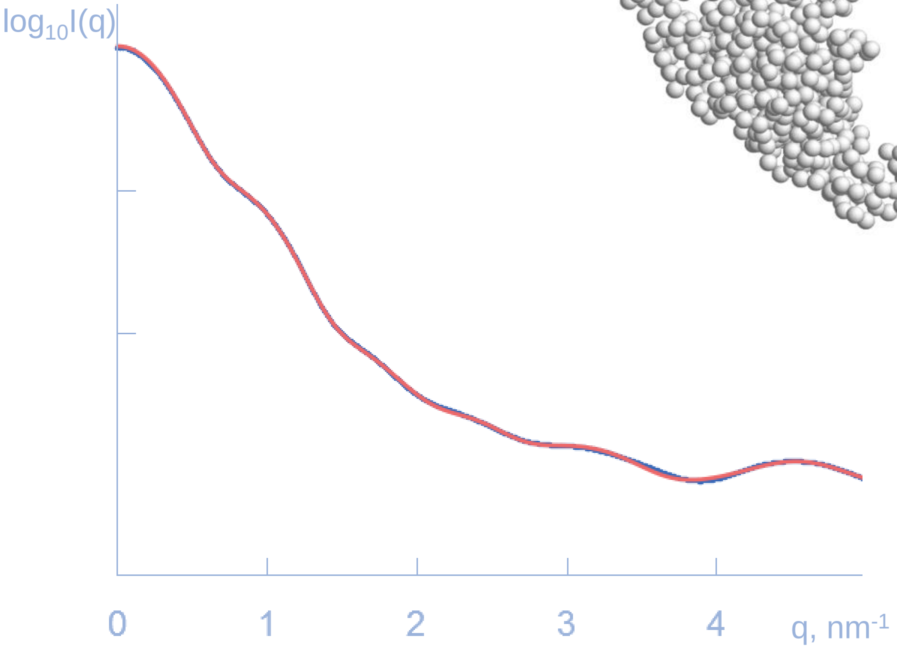
EMBL

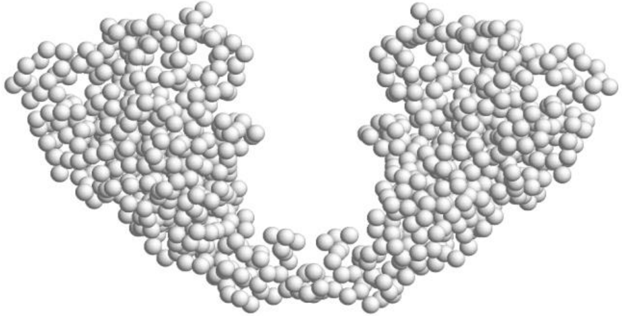
Ab initio reconstruction: dummy residue modelling





Ab initio reconstruction: dummy residue modelling





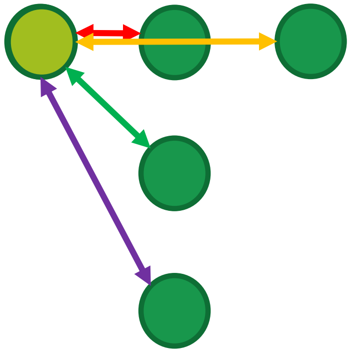
GASBOR

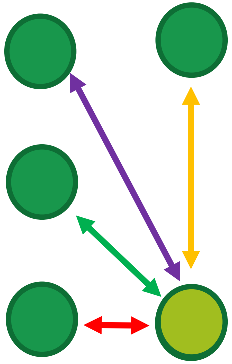
Svergun, D.I., Petoukhov, M.V, Koch, M.H.J. (2001)
Biophys J 80, 2946–2953

- Use dummy residues with average density (**fixed radius of 1.9 Å**)
- Number(dummy residues) = Number(AA) = *K* (**fixed number**)
- Distances to neighbor “residues” like for proteins
- Fixed search space
- Scattering is computed using Debye formula
- Use **higher angles** (up to 12 nm^{-1})
- Only for proteins **smaller than 660 kDa**

Words of caution

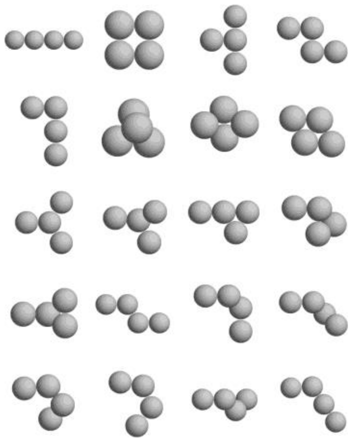
Ambiguity in SAXS: C-T

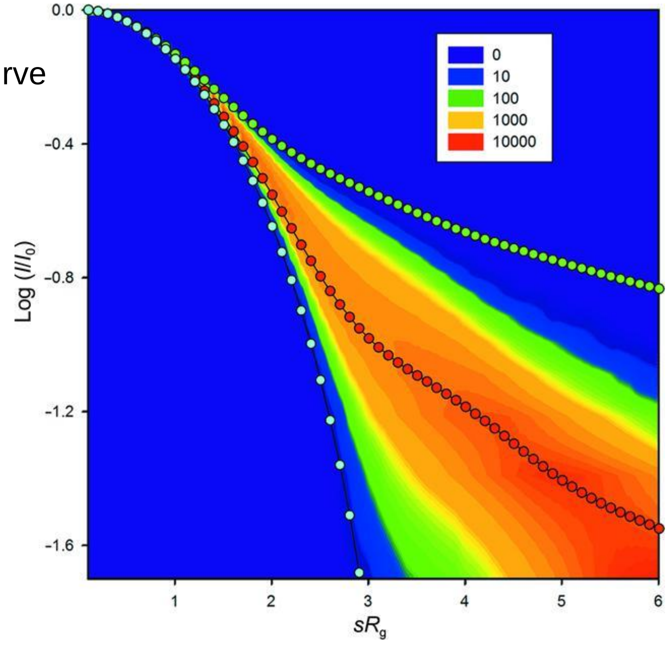




Measure of the ambiguity

Quantitative measure of the ambiguity from the SAXS curve





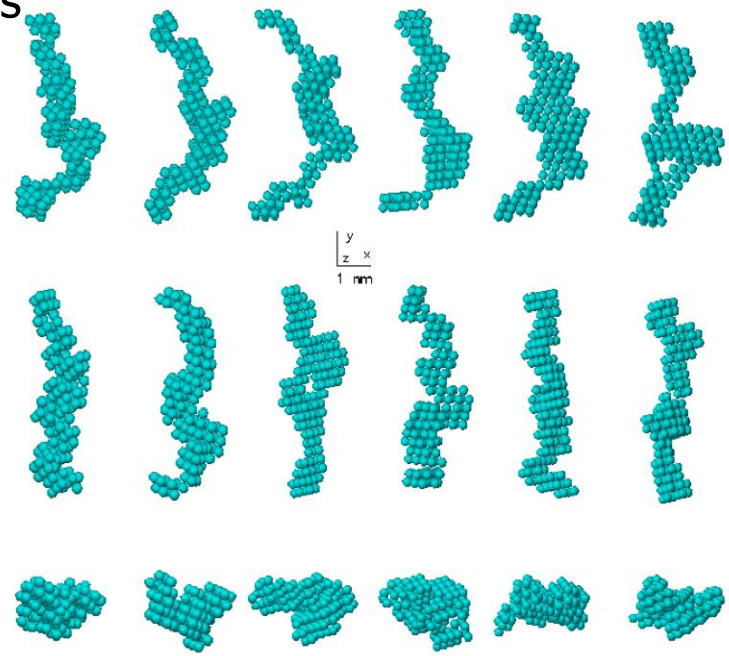
Ambimeter

- 14000 shape topologies generated (up to seven beads closely packed on hexagonal grid).
- Scattering curves computed and rescaled to keep only shape topology information.
- Scattering map computed from these curves.
- By plotting the experimental SAXS curves on the map, ambiguity intrinsic to the curve can be estimated .

Petoukhov, M. V., & Svergun, D. I. (2015). *Acta. Cryst. D.*

Ab initio model validity

Shape determination of 5S RNA: six DAMMIN models yielding identical fits

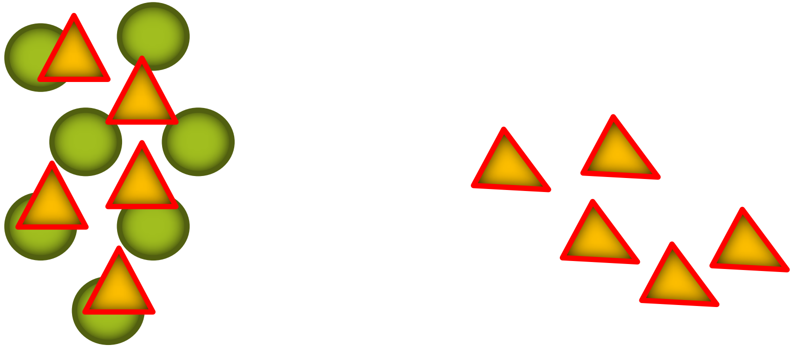


Funari et al. (2000) *J. Biol. Chem.* 275, 31283-31288

Ab initio model validity

SUPCOMB

- Superimpose models by minimizing the Normalized Spatial Discrepancy (NSD)
- Steps
  - Principle axes alignment
  - Gradient minimization
  - Local grid search



SUPALM

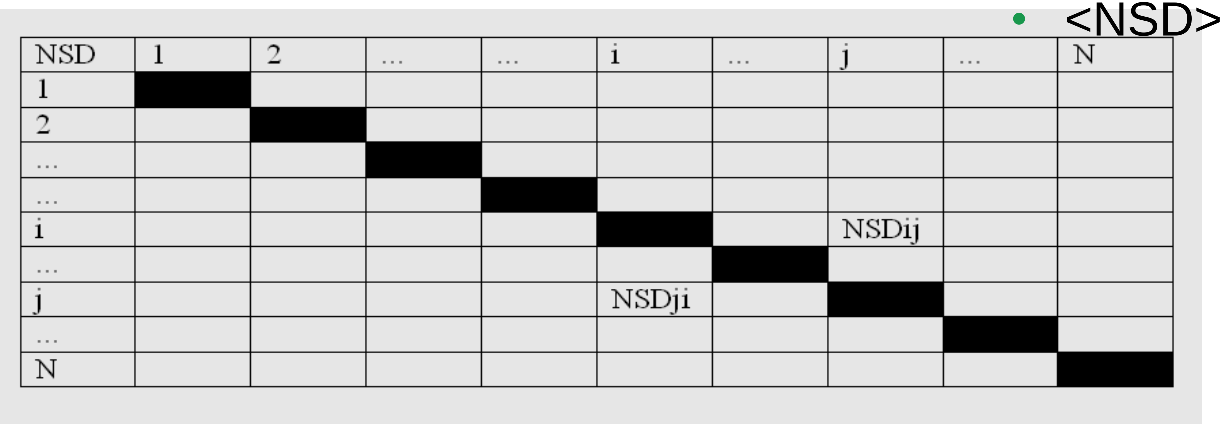
- Aligns models in Fourier space using spherical harmonics representation
- For MDa size particles – about 10 times faster than SUPCOMB

Reduce ambiguity of ab-initio model

- To reduce ambiguity, several models are built, averaged and compared

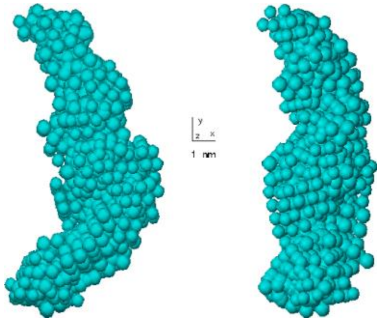
- NSD_i = $<NSD_{ij}>_j$
- MIN(NSD_i) => typical (most proba
- <NSD> + 2 σ (NSD) => threshold

DAMAVER

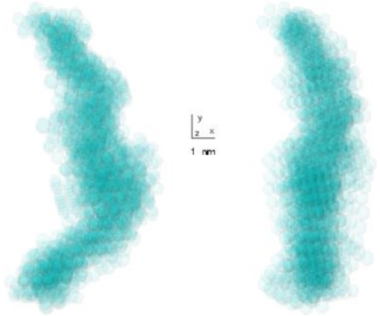


NSD	1	2	…	…	i	…	j	…	N
1	■								
2		■							
…			■						
…				■					
i					■		NSDij		
…						■			
j					NSDji		■		
…								■	
N									■

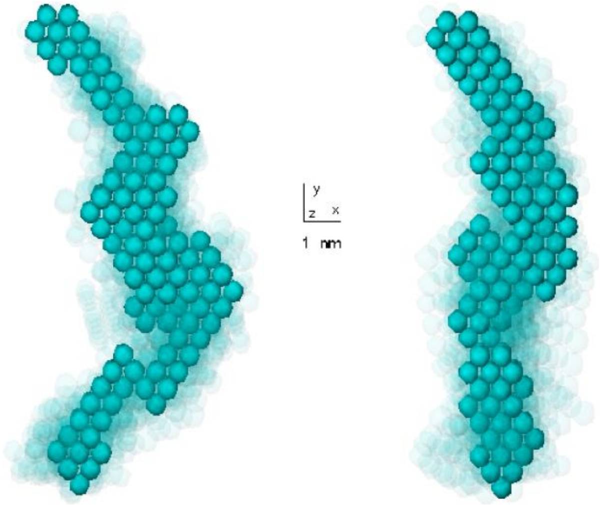
Model validity



5S RNA – Solution spread region



5S RNA – Most Populated Volume

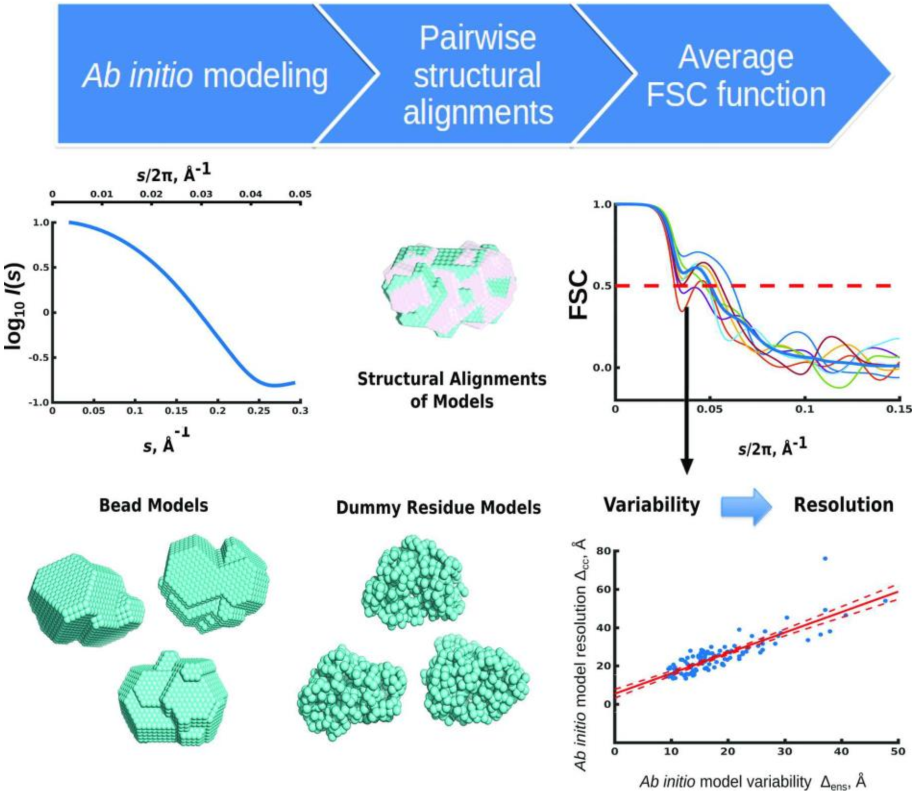


5S RNA – Final Solution within the Spread Region

Funari et al. (2000) *J. Biol. Chem.* 275, 31283-31288

Resolution of ab initio models

SASRES

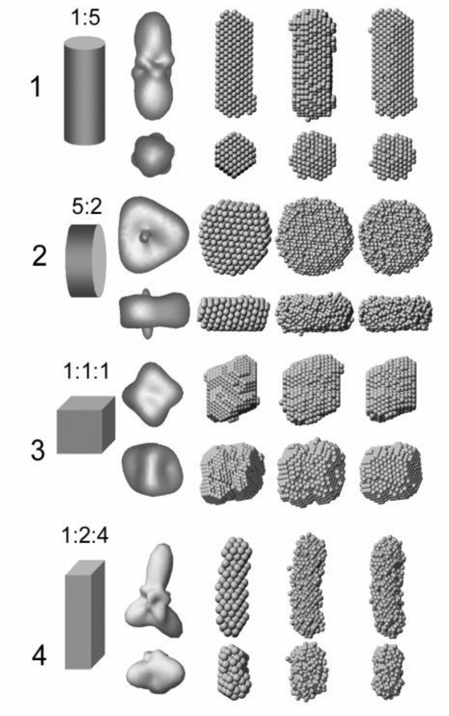


- “Measure ambiguity to estimate resolution”
- Resolution estimated from a set of (10-20) bead model.
- Model compared and aligned.
- Measure of the variability gives an estimation of the resolution

Tuukkanen et al., IUCr J. (2016)

Can all shapes be reconstructed by ab initio modelling?

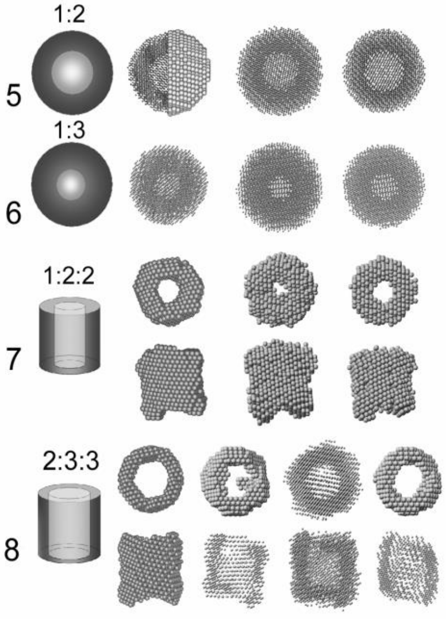
Volkov, V. V. & Svergun, D. I. (2003). J. Appl. Cryst. 36, 860-864.



solid bodies with moderate anisometry
(elongated particles up to 1:5 and flattened up to 5:2) can be reliably reconstructed from the SAXS data.
Mean value NSD : 0.4-0.7

Globular solid particles

Hollow globular models can also be well reconstructed

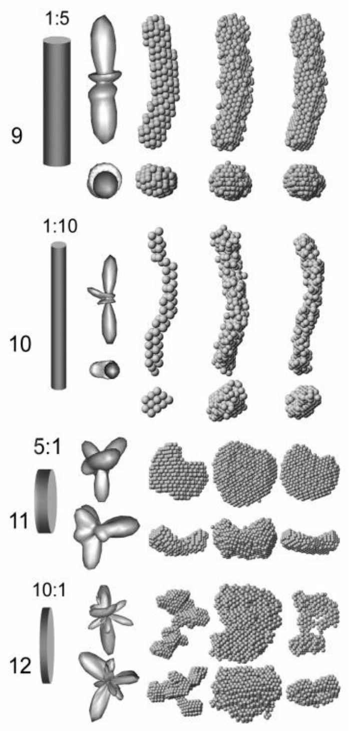


Hollow globular particles

EMBL

Can all shapes be reconstructed by ab initio modelling?

Volkov, V. V. & Svergun, D. I. (2003). J. Appl. Cryst. 36, 860-864.



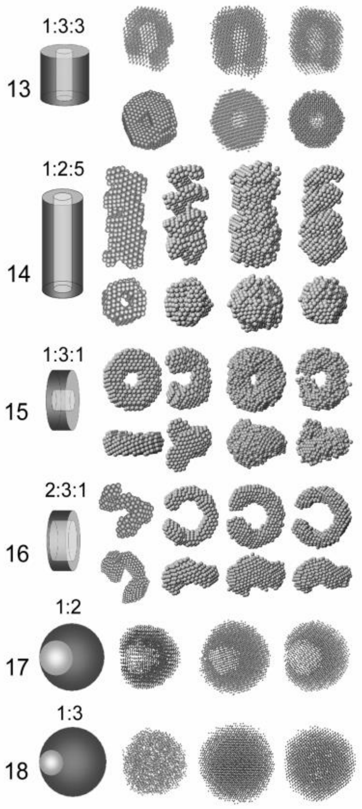
Shape reconstructions of anisometric particles are less stable and reliable.

Elongated hollow body: the channels may appear closed from one or both sides

For hollow flattened the resulting shapes may show a helical turn instead of a hollow disk, even after the averaging.

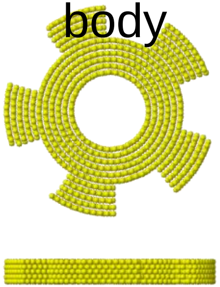
Acentric voids in hollow spheres are only reconstructed if r/R is about 0.5

anisotropic solid particles



Hollow anisotropic and acentric

Use of symmetry.



Original body

Conclusion

- Ab initio methods are powerful tools to build model from SAXS data.

- Ab initio methods always provide good looking models that fit the data (even if they shouldn’t → Beware of what data you put in)

- Different kind of models can be built (dummy atom model, dummy residue model, multiphase)

- The models built are of low resolution and have some ambiguity but methods now exist to estimate this ambiguity and resolution

- Further reduce ambiguity → add information

Questions?