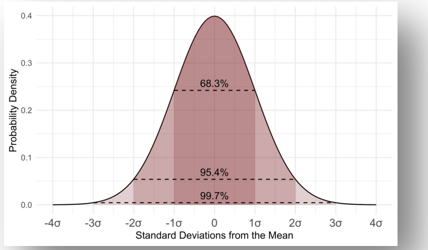
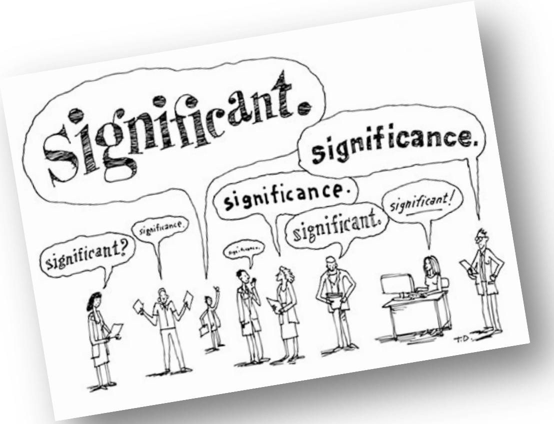
Statistical methods in SAXS and SANS

FASEM PhD School (March 2024)

Andreas Haahr Larsen
Department of Neuroscience,
University of Copenhagen





You will learn how statistics can:

- Help you to assess if a model is a good fit to data
- Help you to adjust your model

Part I:
use statistics to assess if a fit is good

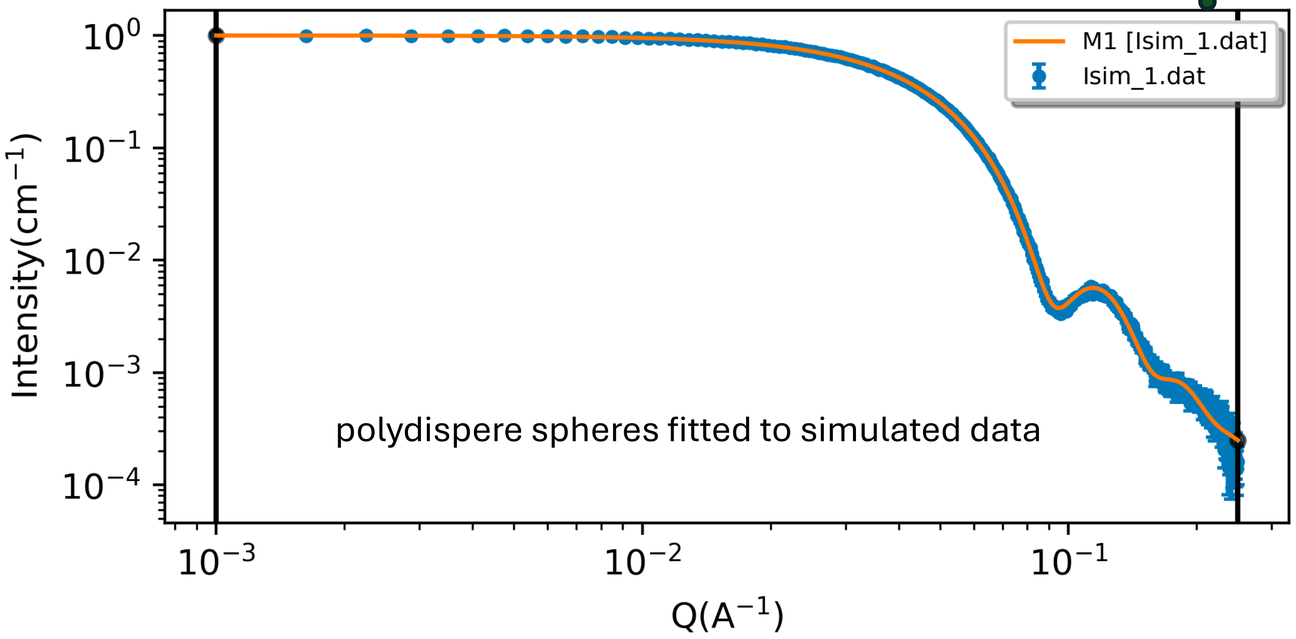
Checklist: Goodness of fit

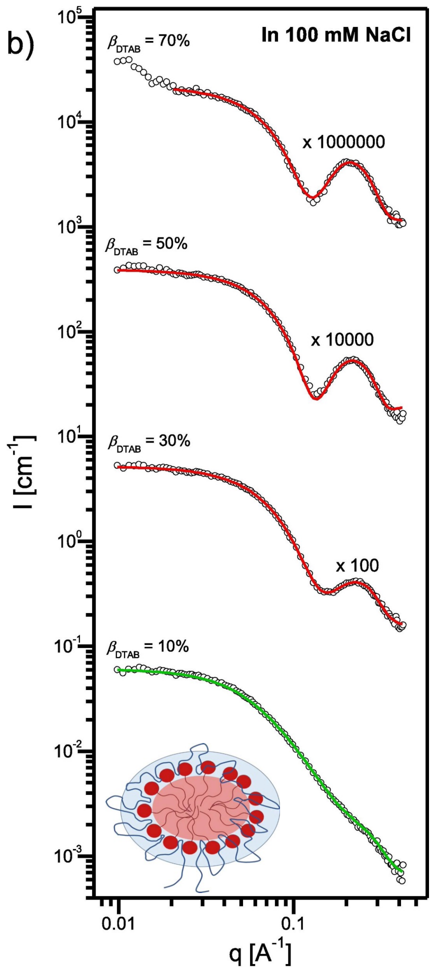
1) Look at the fit and chi-square ☐

2) Residuals (and runs statistics) ☐

3) Are the parameters reasonable? ☐

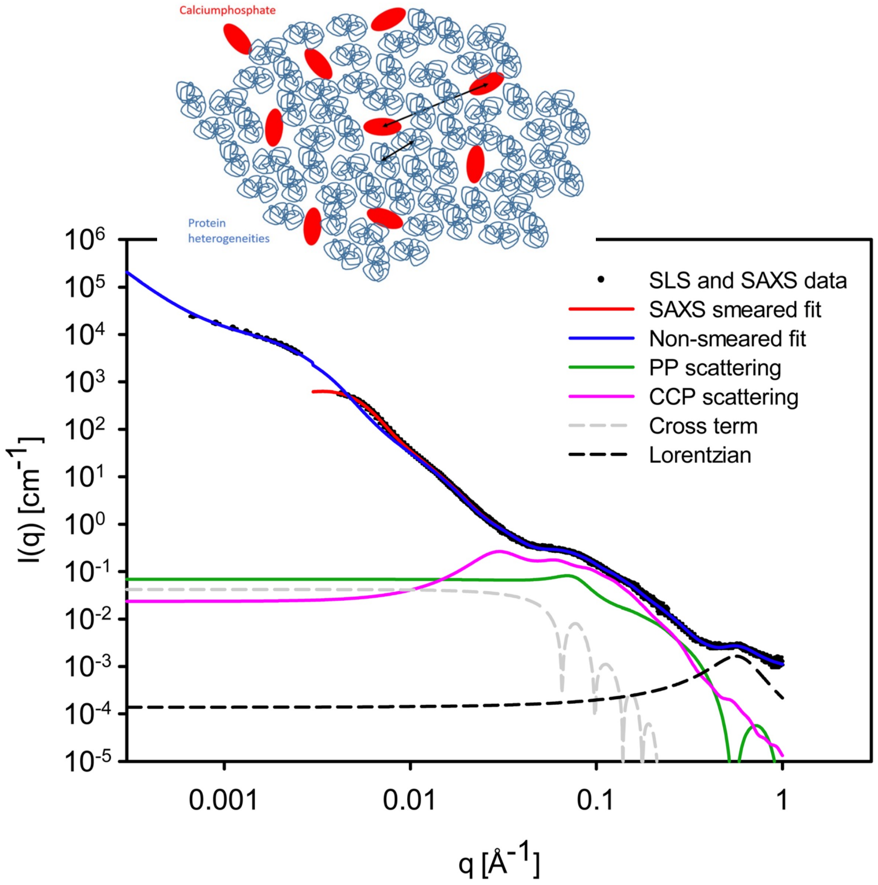
is it a good fit??

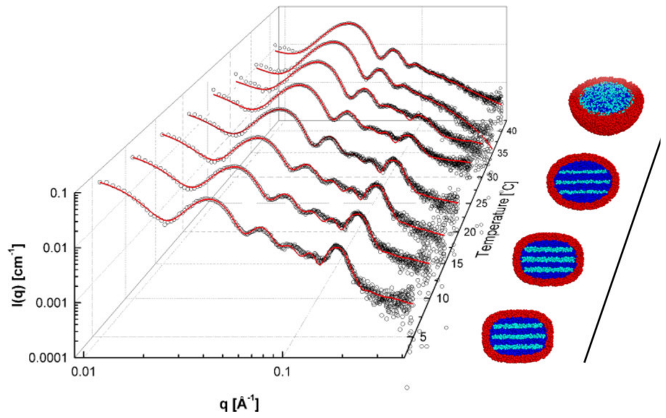


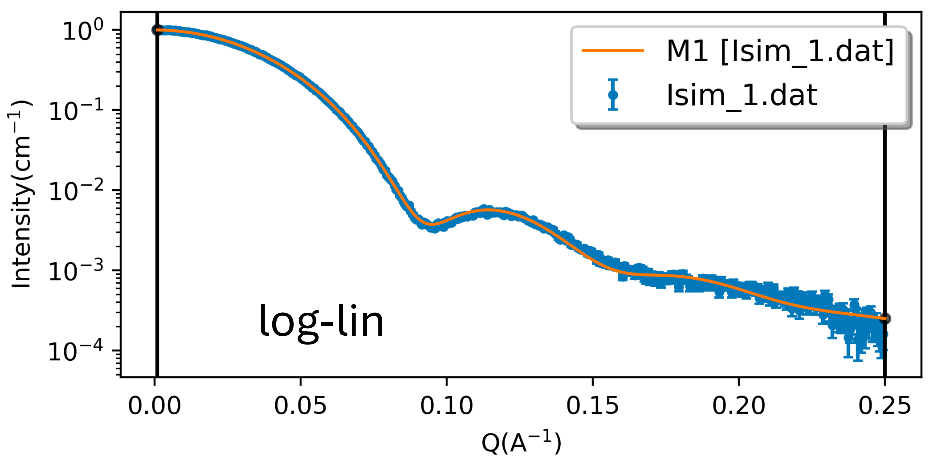


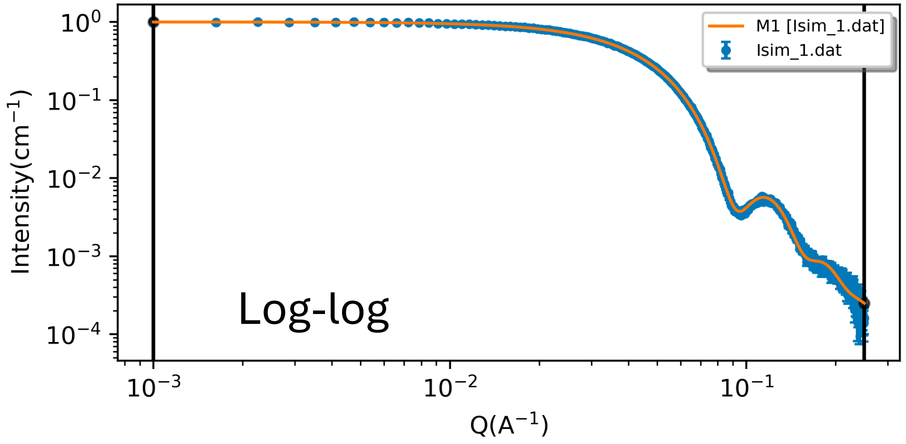
Visual inspection

- log(q) and lin(q)
- Good fit in the whole *q*-range, or part of it
- Does the model holds for series of contrasts/concentrations/temperatures/…









A simple model fitting reasonably can be more informative than a complex model fitting perfectly

“chi-square”:

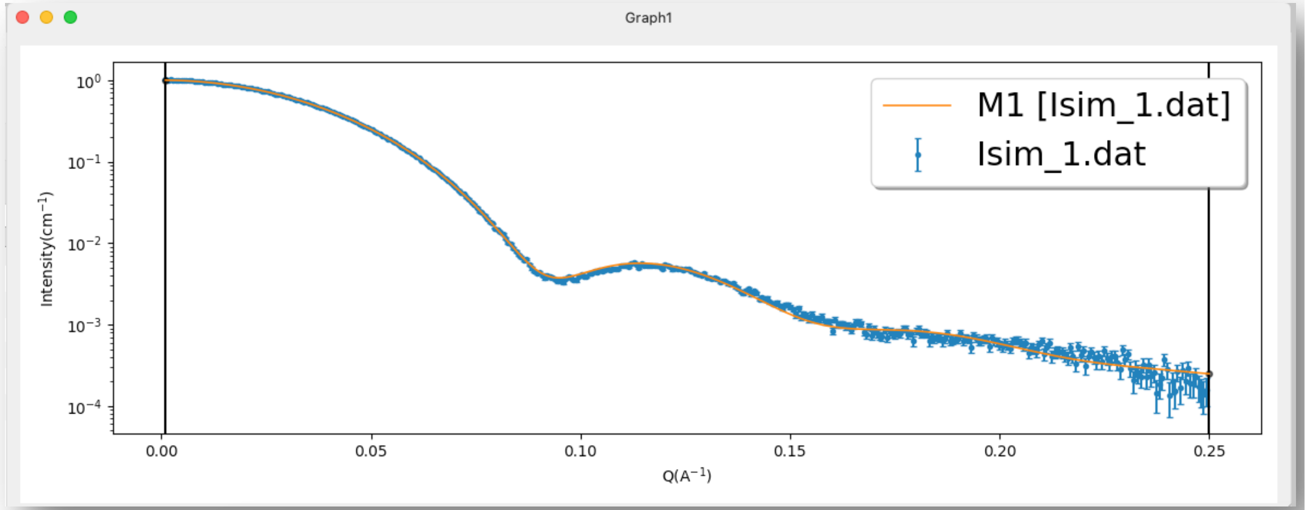
$$\chi^2 = \sum_{i=1}^{N} \left(\frac{I_i^{\mathrm{exp}} - I^{\mathrm{mod}}(q_i)}{\sigma_i} \right)^2$$

“reduced chi-square”:

$$\chi_r^2 = \frac{\chi^2}{\mathrm{Expected}\ \chi^2} = \frac{\chi^2}{N - K}$$

Notation alert:
As the *reduced* chi-square is always reported, “reduced” is often omitted.

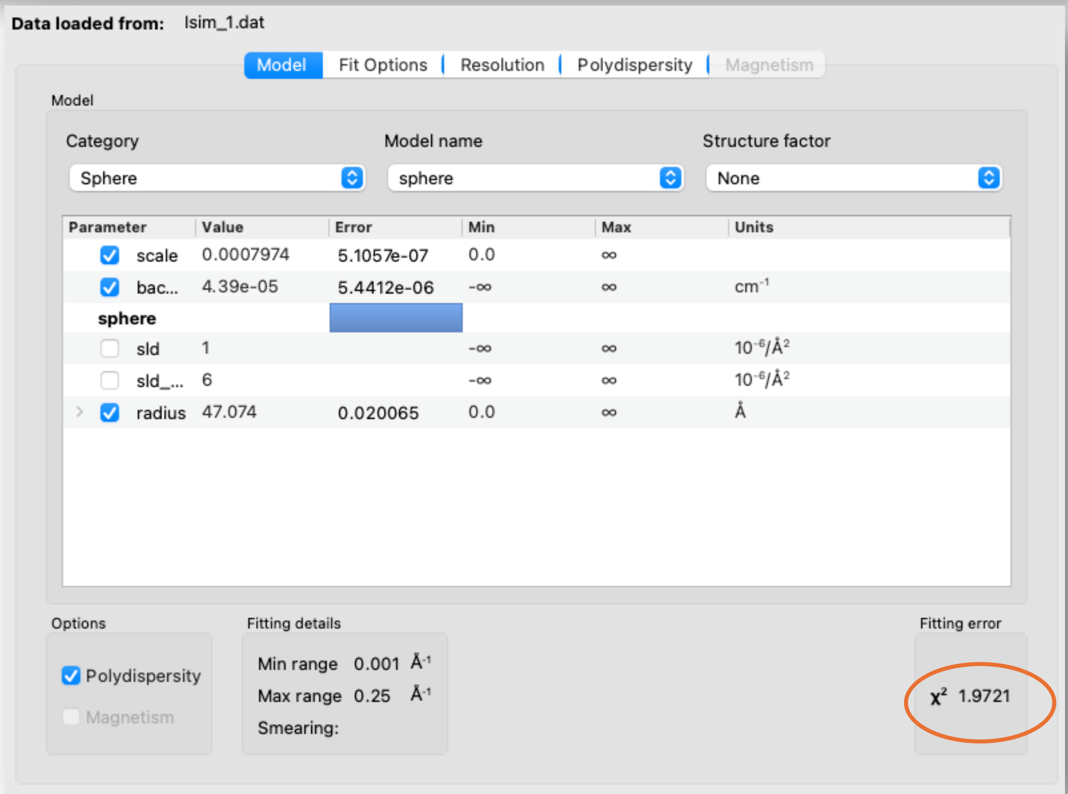
So, when “chi-squared” is used, it (usually) refers to the “reduced chi-square”

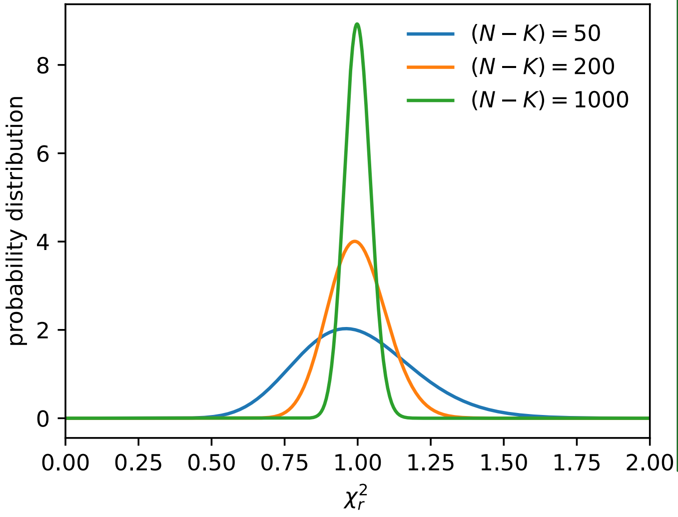


K: number of *independent* model parameters
N- K: the degrees of freedom

rules of thumb for the reduced chi-square:

$\chi_r^2 \sim 1$:	perfect fit
$\chi_r^2 > 1.5$:	model could be improved (or underestimated errors)
$\chi_r^2 < 0.7$:	overfitting (or overestimated errors)





p-value: What is the **probability** of getting this χ_r^2, assuming the model is true?

significance: if p > α
α: **significance criteria**, e.g. 5% or 1%

p-values are often extremely low in SAXS/SANS analysis due to:

- Approximate models
- Systematic errors

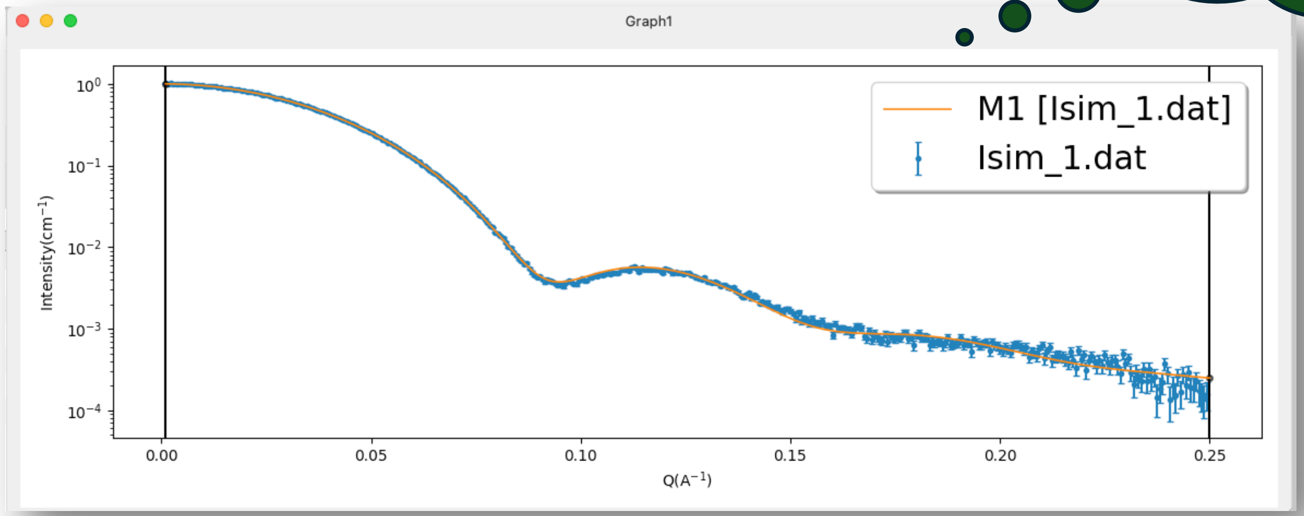
Checklist: Goodness of fit

1) Look at the fit and chi-square (☑)

1) Residuals (and runs statistics) ☐

2) Are the parameters reasonable? ☐

is it a good fit??

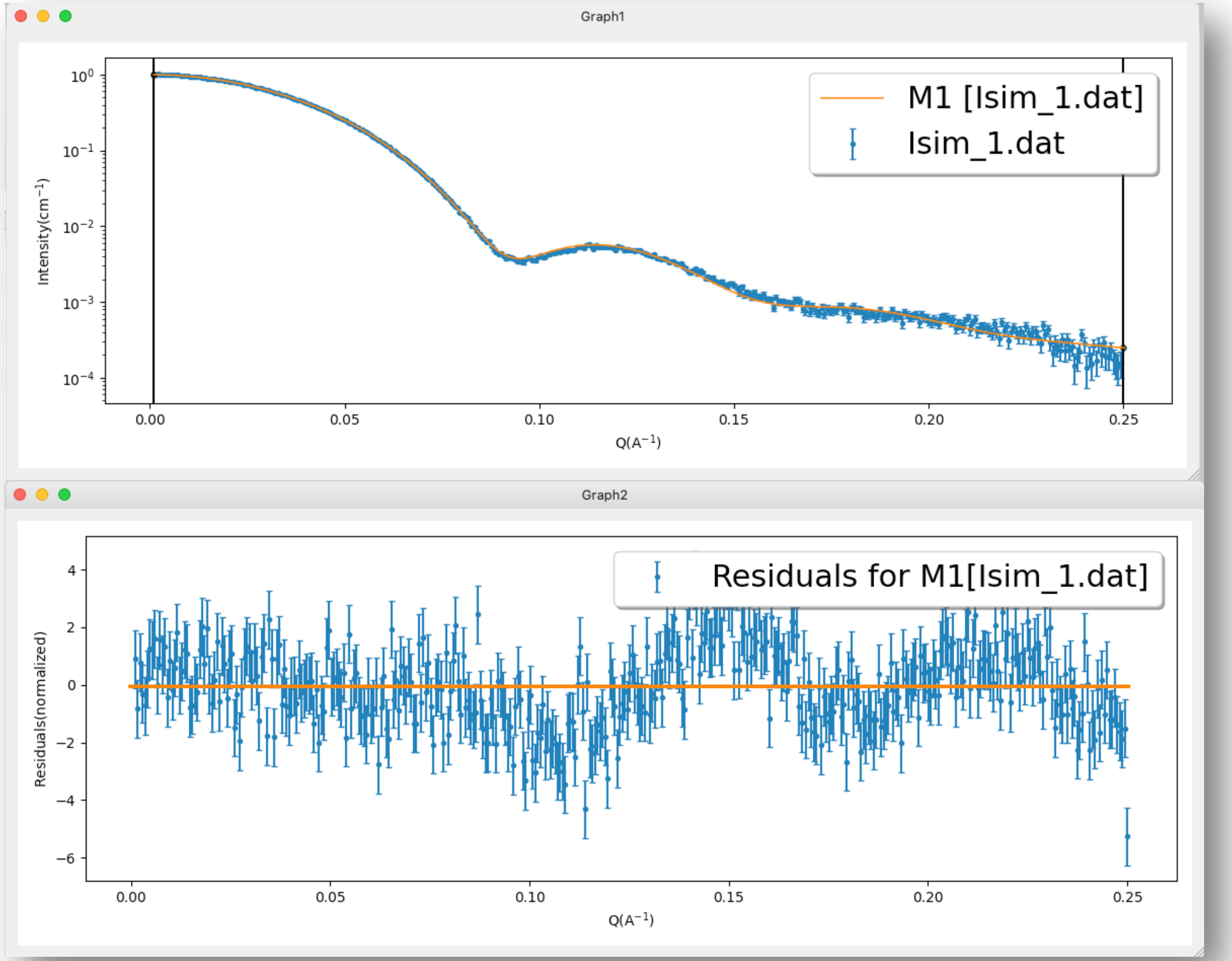


Normalized residuals:

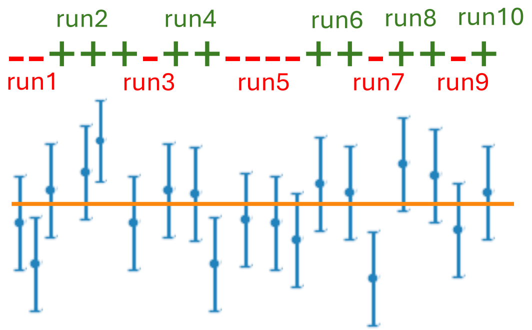
$$\left(\frac{\Delta I}{\sigma}\right)_i = \frac{I_i^{\text{exp}} - I^{\text{mod}}(q_i)}{\sigma_i}$$

Residuals and chi-square

$$\chi^2 = \sum_{i=1}^{N} \left(\frac{I_i^{\text{exp}} - I^{\text{mod}}(q_i)}{\sigma_i}\right)^2 = \sum_{i=1}^{N} \left(\frac{\Delta I}{\sigma}\right)_i^2$$



Runs statistics



In this example:

$N = 19,$

$N_+ = 10,$

$N_- = 9,$

number of runs: $R^M = 10$

longest run: $R^L = 4$

- Independent on error estimates
- Reduced *R* values should be around one

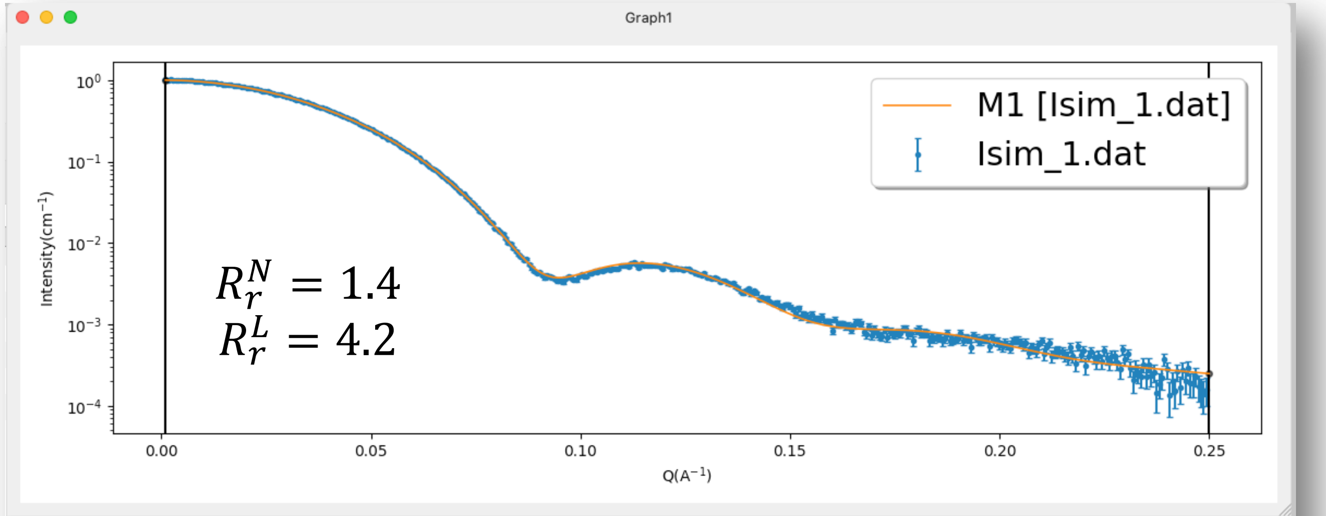
$$R_r^L = \frac{R^L}{\text{Expected } R^L} = \frac{R^L}{\log_2(N - K) - 1} \sim 1$$

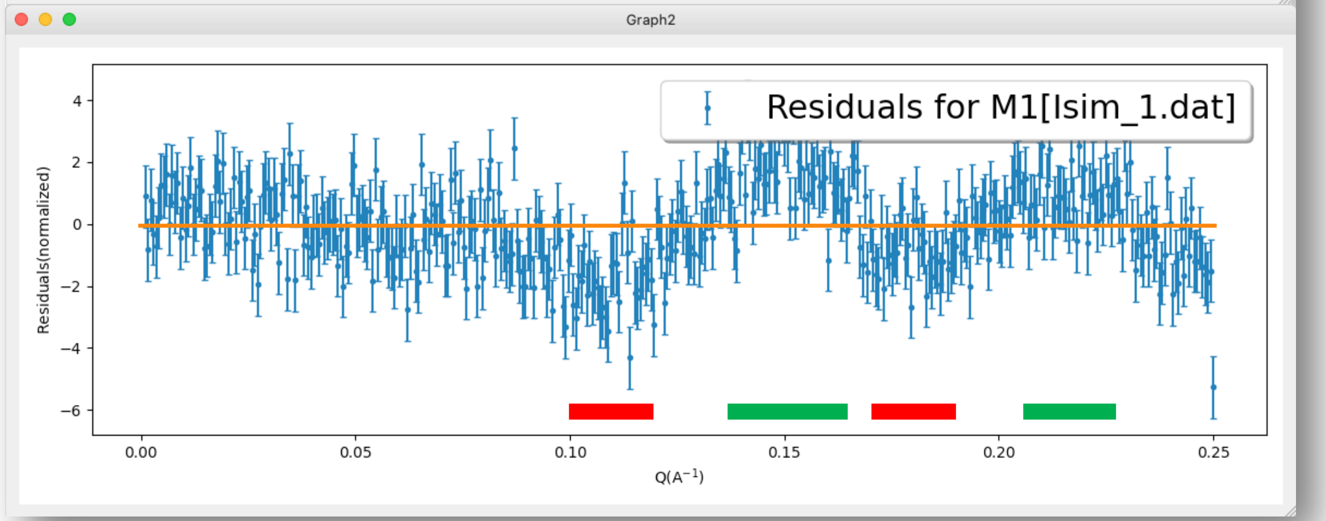
$$R_r^N = \left(\frac{R^N}{\text{Expected } R^N}\right)^{-1} = \left(\frac{R^N}{1 + 2\, N_+ N_- / (N - K)}\right)^{-1} \sim 1$$

p-values can be calculated but are often not very useful for model comparison

Disclamer: work in progress…

The use of runs statistics in SAXS/SANS is (yet) not standard, and runs statistics are not reported in most programs or papers.





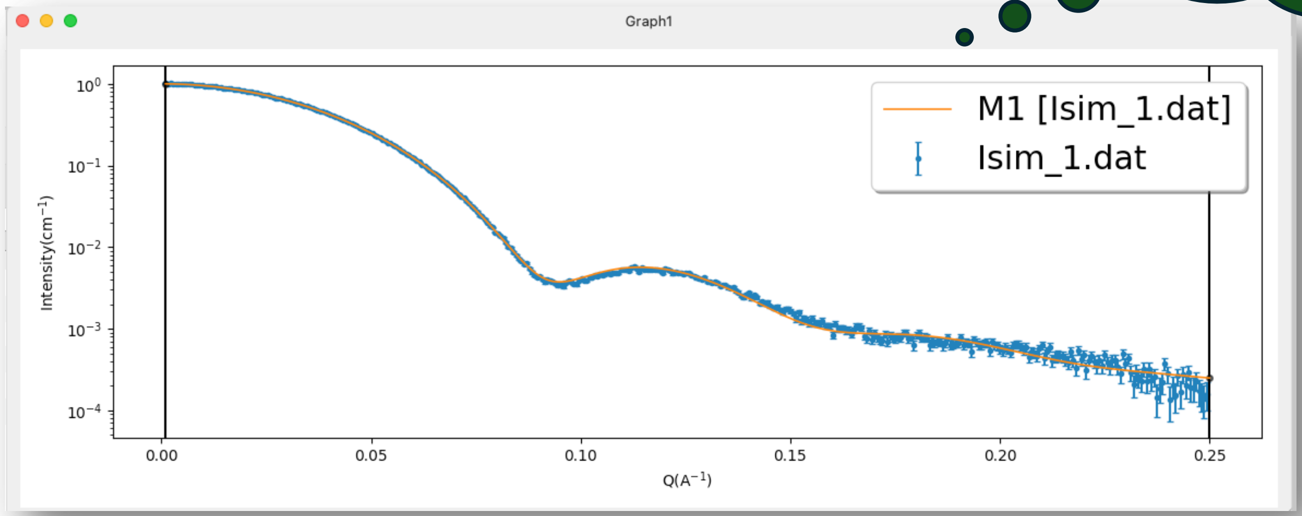
Checklist: Goodness of fit

1) Look at the fit and chi-square ☑

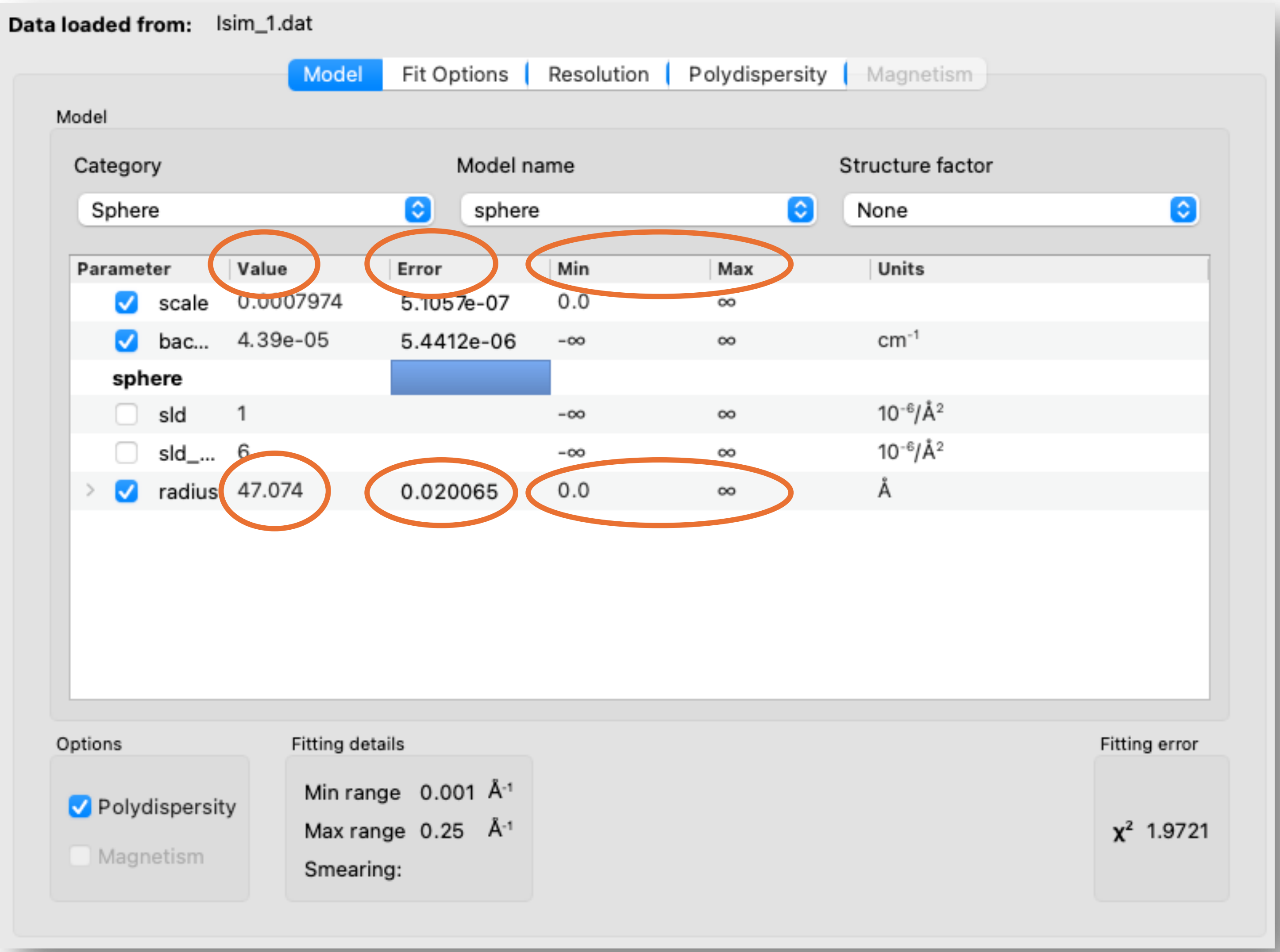
1) Residuals (and runs statistics) ☑

2) Are the parameters reasonable? ☐

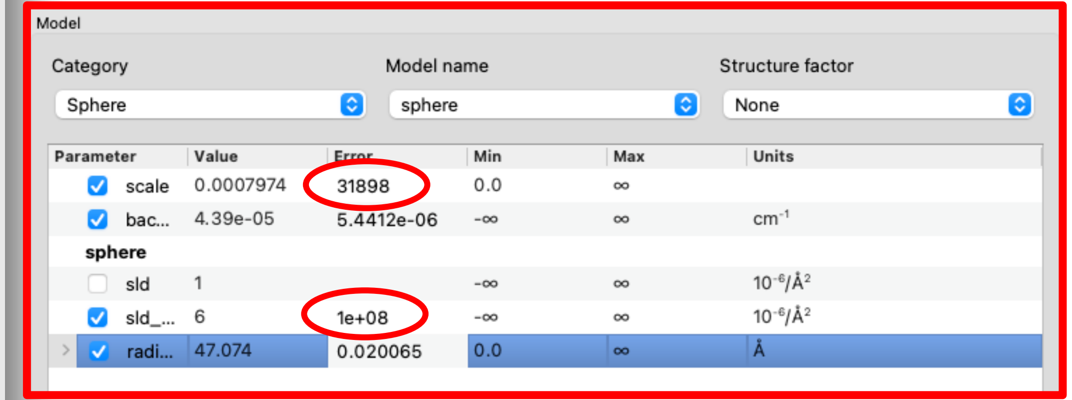
is it a good fit??



Are the parameters reasonable?



- Mean value close to expectation?
- Within expected range (min/max)?
- Be cautious if refined value equals min or max
- If error is large:
  → correlation between parameters



$$I(q) \propto scale \cdot (\Delta SLD)^2$$

100% correlated

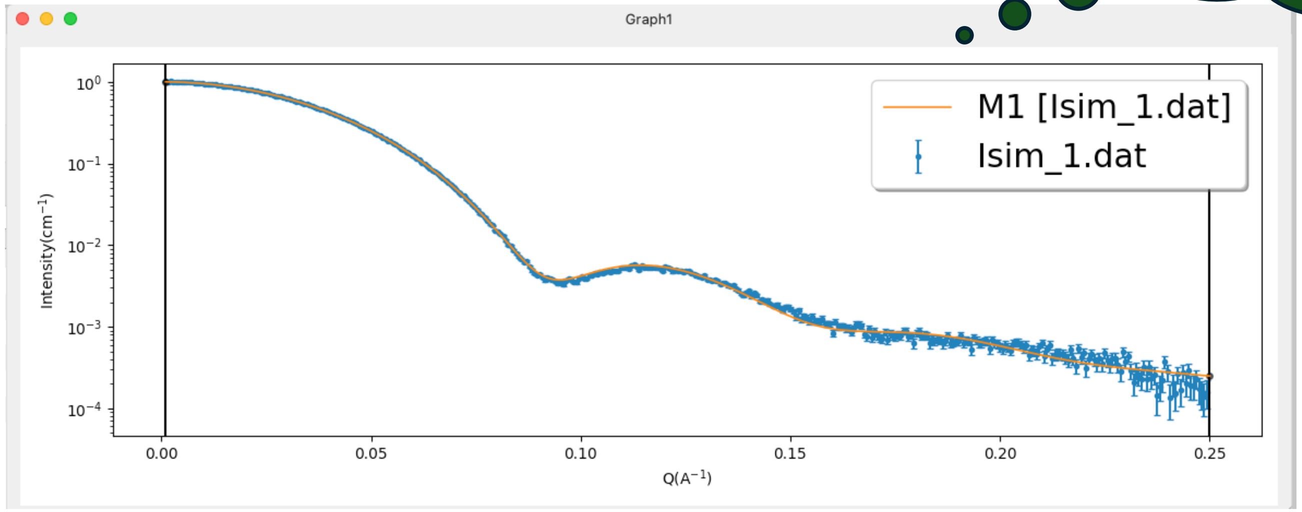
Checklist: Goodness of fit

1) Look at the fit and chi-square (✓)

1) Residuals (and runs statistics) (✓)

2) Are the parameters reasonable? ✓

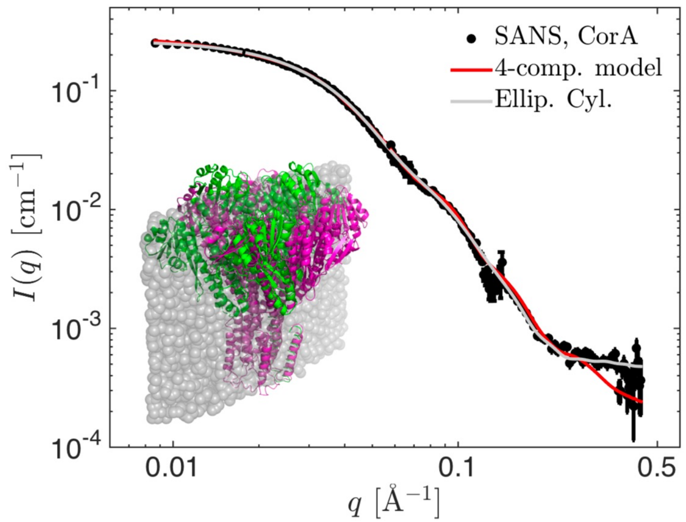
Questions for Part I?



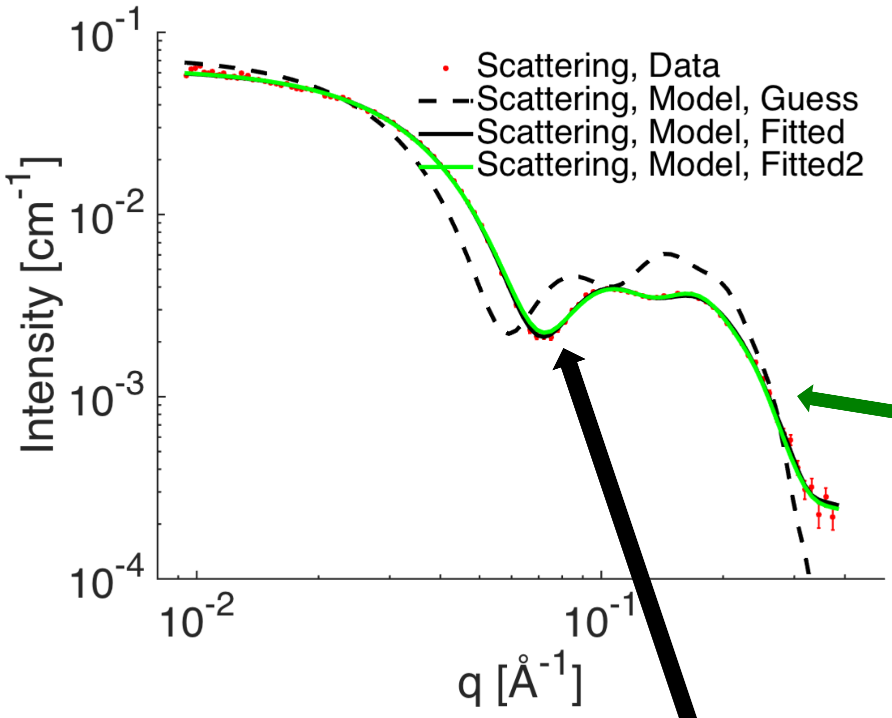
Part II:
use statistics to find the best model

Or, how to build our knowledge into the model

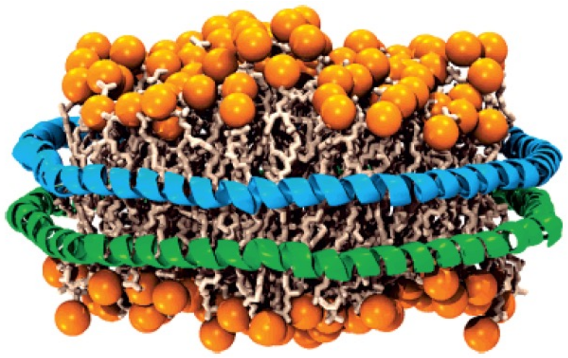
Motivation: structure determination with SAXS and SANS is an **ill-posed problem**

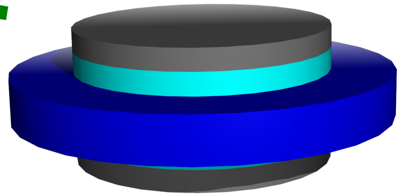


What is a "simple model"?
"not surprising" may be a better term



a nanodisc

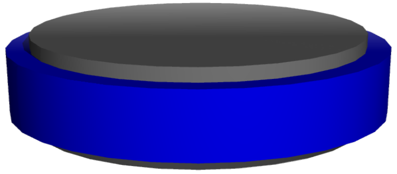




nanodisc
model 2

Therefore: we must constrain the solutions to realistic models

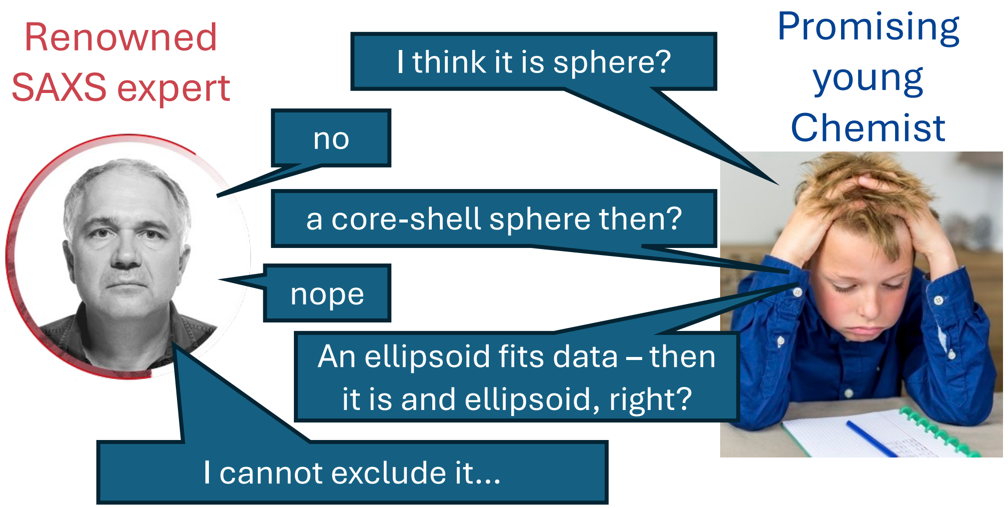
- parameter limits
- choice of model



nanodisc
Model 1

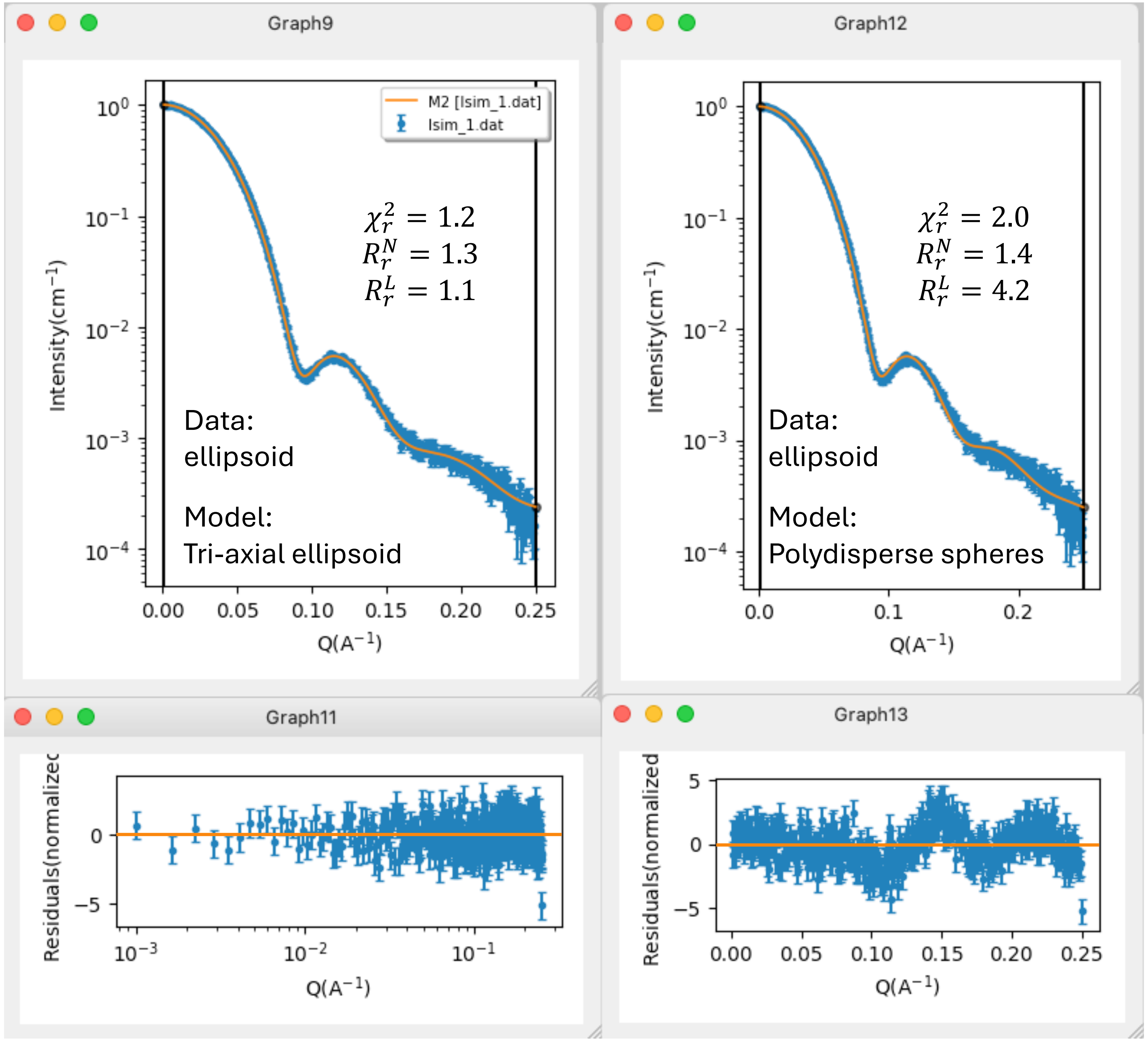
Motivation: structure determination with SAXS and SANS is an **ill-posed problem**

Take-home message:
One SAXS/SANS dataset does not correspond to a single, unique structure
However, SAXS is good at saying NO



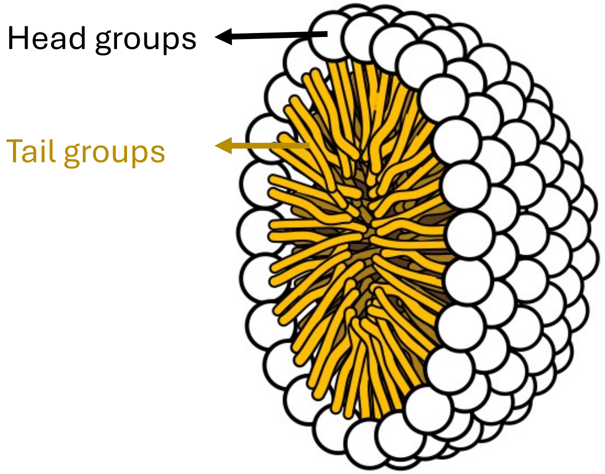
Therefore: we must constrain the solutions to realistic models

- parameter limits
- choice of model

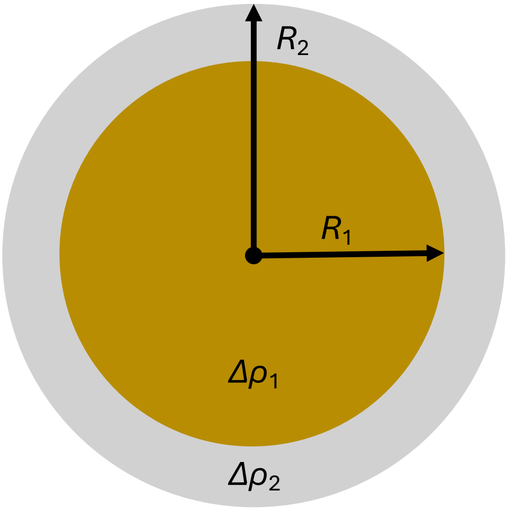


Inclusion of **molecular restraints**
(by reparameterization)

Detergent Micelle



Core-shell Model



Example

scale, R_1, R_2, $\Delta\rho_1$, $\Delta\rho_2$ → **c, N, V_{head}, V_{tail}**

- *c* → *Micelle number density (**concentration**)*
- *N* → ***Number** of detergents per micelle*
- V_{head} → ***Volume** of one **head** group*
- V_{tail} → ***Volume** of one **tail** group*

$$\frac{4\pi}{3} R_1^3 = N V_{tail}$$

$$\frac{4\pi}{3} (R_2^3 - R_1^3) = N V_{head}$$

$$\Delta\rho_1 V_{tail} = b_{tail}$$

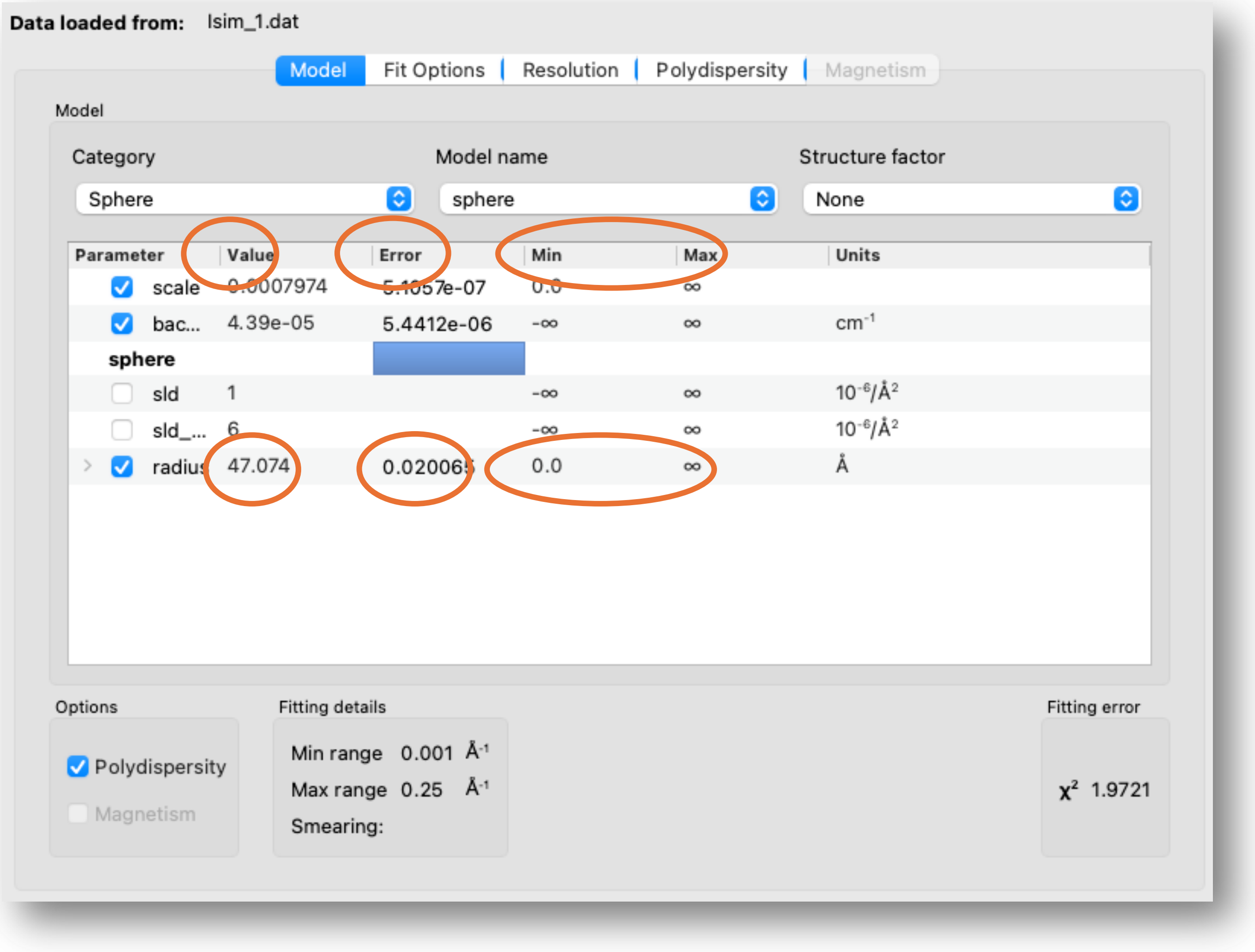
$$\Delta\rho_2 V_{head} = b_{head}$$

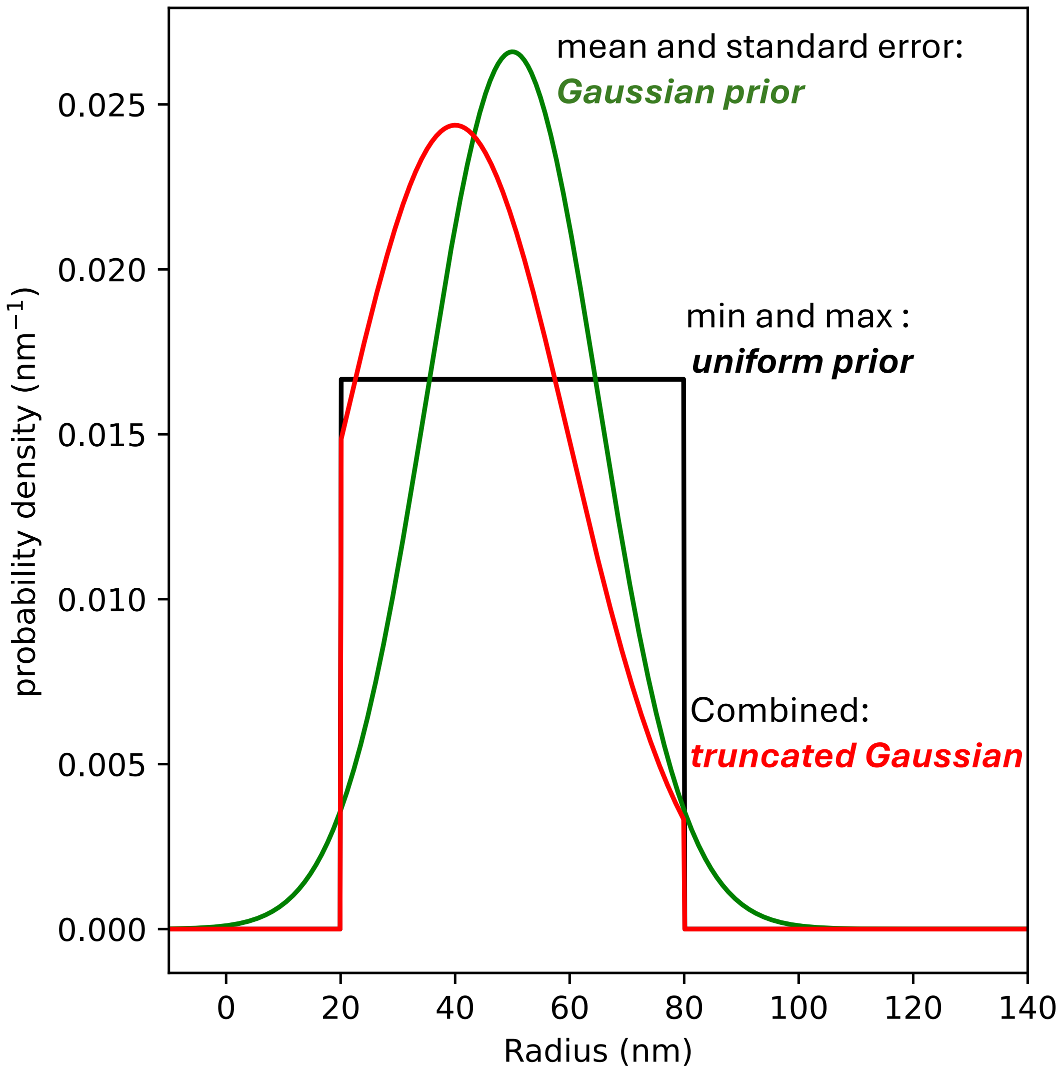
Scattering lengths, b_{head} and b_{tail}, can be calculated from the chemical composition

What do we achieve?

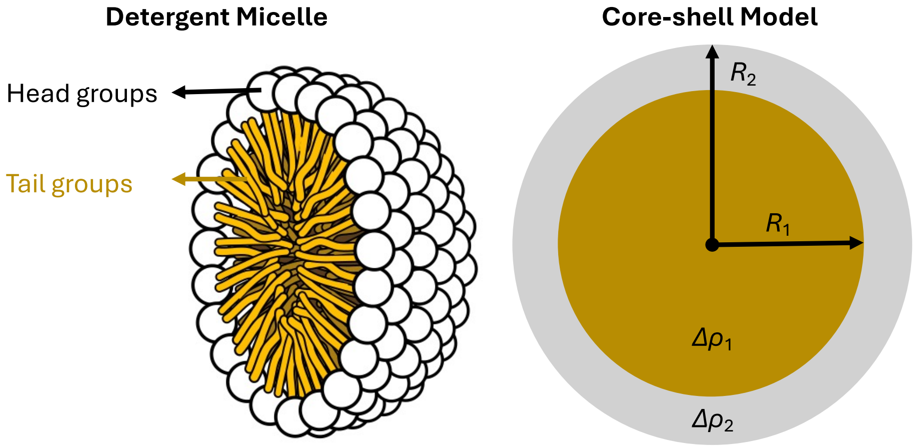
- Reduce from 5 to 4 parameters
- Can use prior knowledge:
  - measurements of V_{head} and V_{tail}
  - concentration estimate

Quantify our prior knowledge as probability distributions: “*priors*”





Inclusion of the priors in the model



Bayes Theorem for probabilities:

$$\underbrace{P(c, N, V_{head}, V_{tail} | \text{data})}_{\text{Posterior (total probability of solution)}} \propto \underbrace{P(c)\, P(N)\, P(V_{head})\, P(V_{tail})}_{\text{Prior (consistency with prior knowledge)}}\ \underbrace{P(\text{data} | c, N, V_{head}, V_{tail})}_{\text{Likelihood (how well does the model fit the data)}}$$

Posterior
(total probability of solution)

Prior
(consistency with prior knowledge)

Likelihood
(how well does the model fit the data)

Priors ensure :

Given alternative models that fit the data well, the most realistic is selected

Goal: find the parameters (c, N, V_{head}, V_{tail}) that maximize the **posterior** probabilities

Leads to **regularized** problem, where $\Gamma = -2\log(\text{Posterior})$ is minimized:

$$\Gamma = \chi^2 + \alpha S$$

likelihood (χ^2)
prior (αS)

a balance between likelihood and prior

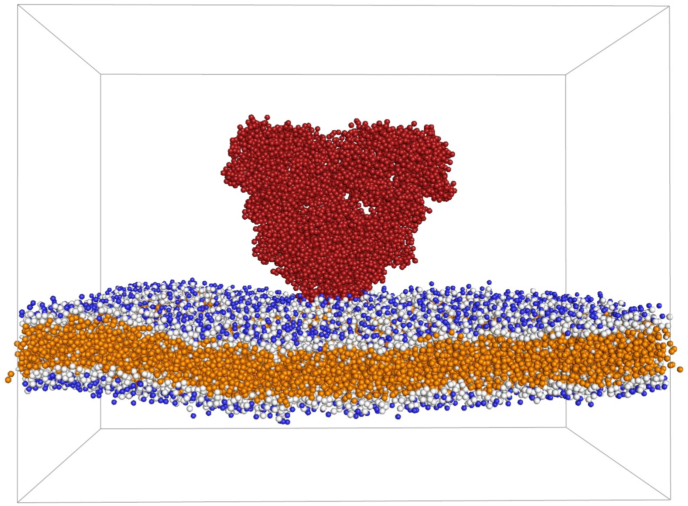
α: adjusts the balance between prior and likelihood

Analogue:
free energy (G), enthalpy (H) and entropy (S)

G = H + T S

T: temperature

Using a simulations as model



Example: structural ensemble of the AMPA receptor

prior weights: $w_{prior,1} = 1.0$ $w_{prior,2} = 1.0$ … $w_{prior,3} = 1.0$

"reweighting"

posterior weights: $w_{prior,1} = 0.8$ $w_{prior,2} = 0.9$ … $w_{prior,3} = 1.3$

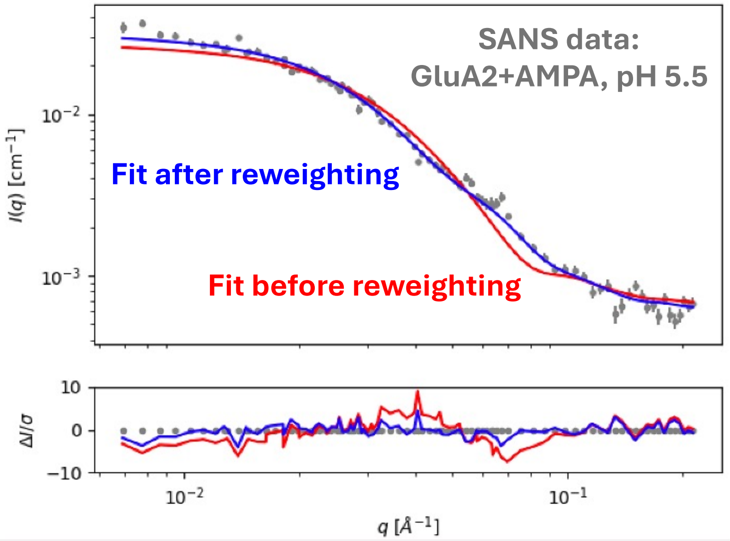
$$S = \sum_{j=1}^{N_W} w_j \left(\frac{w_j}{w_{prior,j}} \right)$$

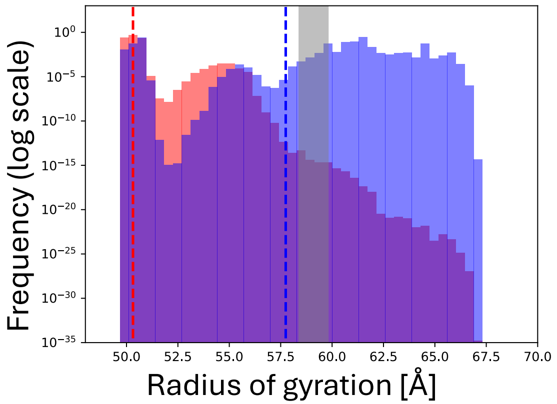
$$\Gamma = \chi^2 + \alpha S$$

likelihood

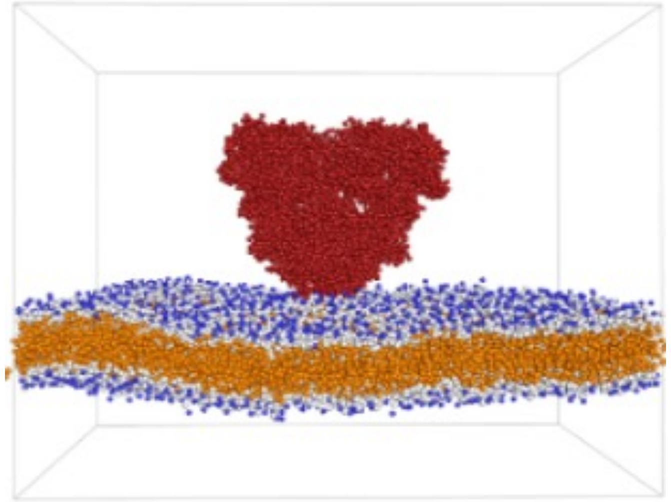
prior

a balance between likelihood and prior





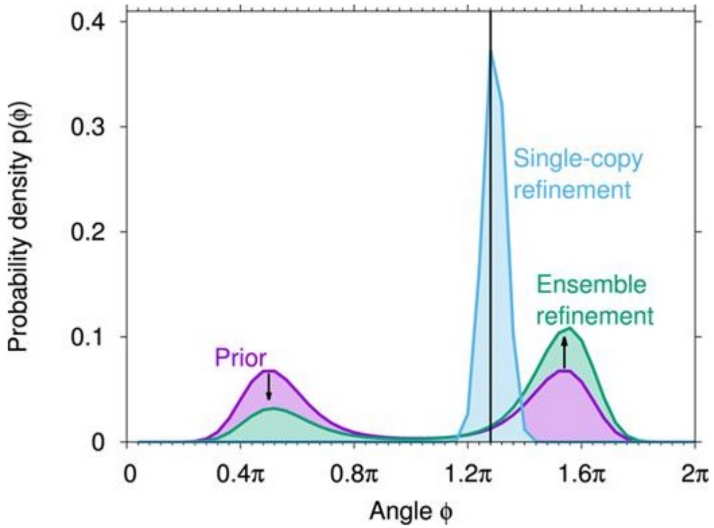
Using a simulations as model



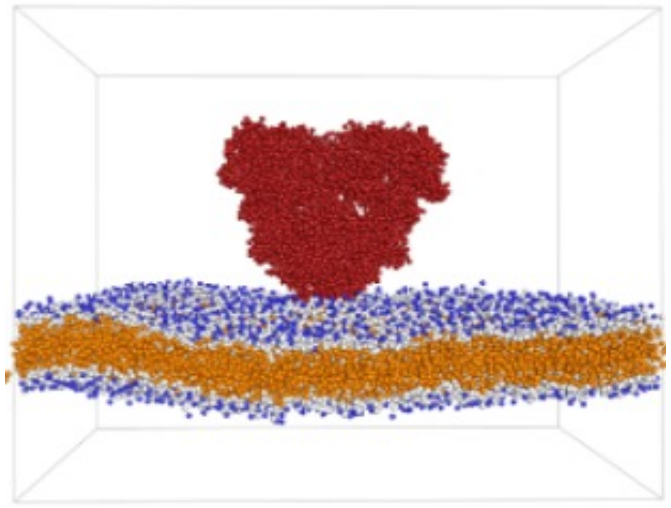
Add a potential to the simulation:

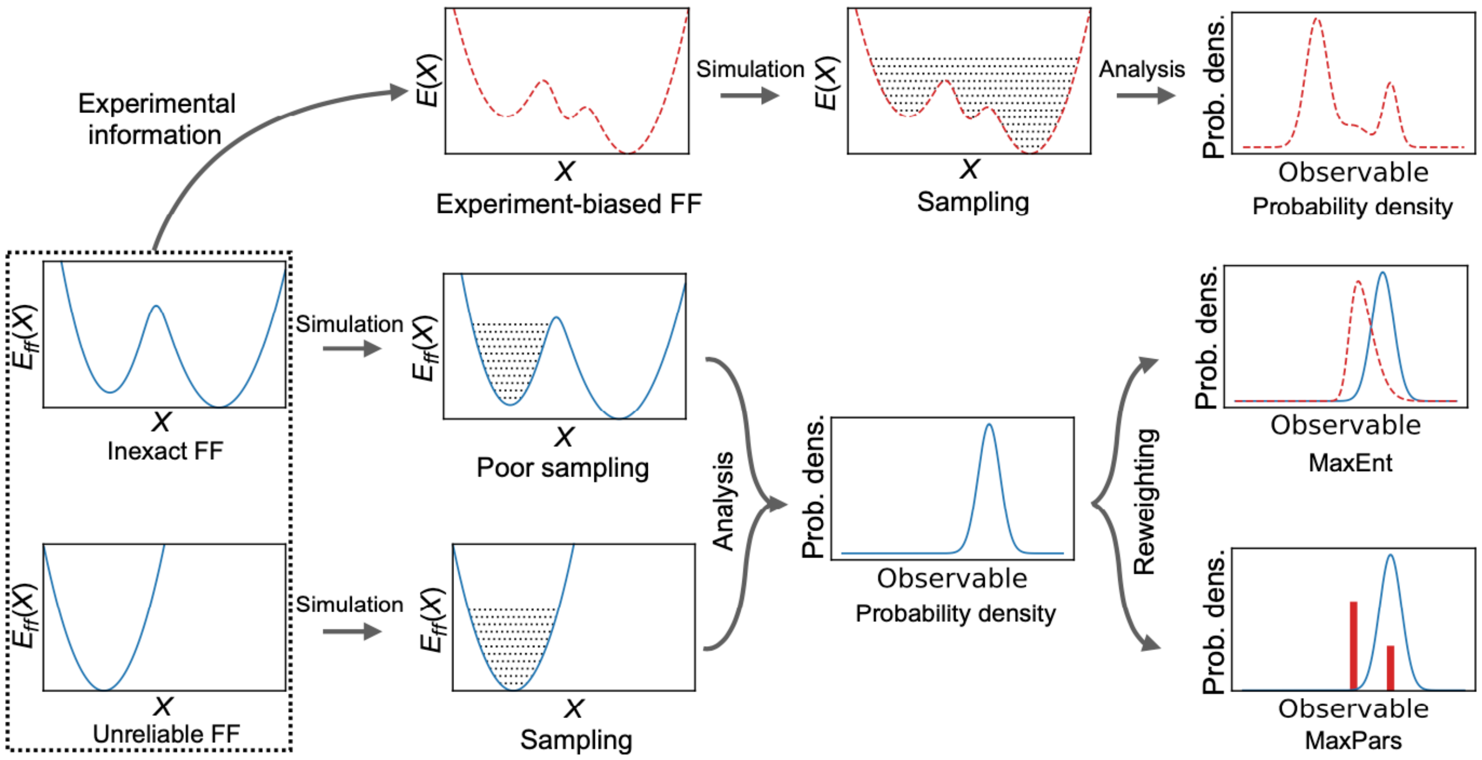
Important:
Each structures should not fit the data, only the average

"sample-and-select" methods not applicable for ensembles



Using a simulations as model





Add a potential to the simulation:

$$E_{\text{hybrid}} = E_{\text{forcefield}}\,(\boldsymbol{R}) + E_{\text{exp}}(\boldsymbol{R}, \boldsymbol{w}, \text{data})$$

Alternative:
Directly use the data to change the force field
"Bayesian update" of force field parameters (*θ*):

P(*θ*|data) ∝ P(*θ*) P(data| *θ*)

Thank you for your attention!
I hope you got a *significantly* better understanding of how to use statistics in SAXS and SANS

A few links for further reading (very incomplete list):

Chi-square tests in SAXS/SANS:

Introduction: https://doi.org/10.1016/S0001-8686(97)00312-6

Error assessment: https://doi.org/10.1107/S1600576721006877

Runs test in SAXS:

Various runs tests: 10.26434/chemrxiv-2021-mdt29-v3

Specific on longest runs test: 10.1038/nmeth.3358

Bayesian model refinement:

Analytical model: https://doi.org/10.1107/S1600576718008956

Multiple datasets: https://arxiv.org/abs/2311.06408

Combine with simulations: https://doi.org/10.1371/journal.pcbi.1005800

Ensembles: https://doi.org/10.1063/1.4937786

10.1371/journal.pcbi.1006641

MaxEntropy reweighting: https://pubmed.ncbi.nlm.nih.gov/32006288/

Reviews on combining simulations and experiments:

https://doi.org/10.1016/bs.pmbts.2019.12.006

https://www.science.org/doi/10.1126/science.aat4010

DOI: 10.1039/c5cp04077a