Bayesian inference of conformational ensembles from limited structural data

Wojtek Potrzebowski

Data Management and Software Centre, ESS

Biochemistry and Structural Biology, LU

www.europeanspallationsource.se

7 April, 2019

European Spallation Source

60% complete

13 funding members, Sweden and Demark serving as host countries

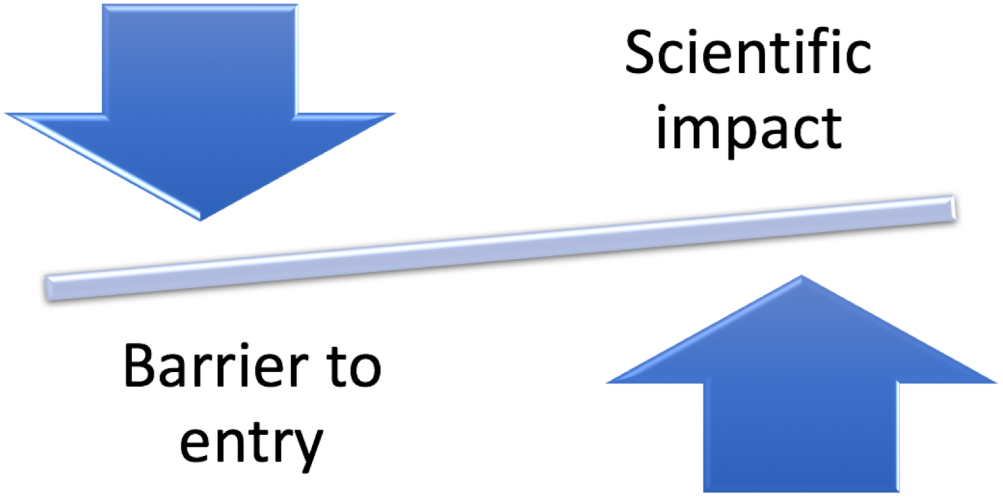
Pulsed neutron source

More and 100 times brighter neutrons than currently available

Exponentially more data to collect and analyse

Data Management and Software Centre (DMSC) balances scientific impact with entry barrier

Minimise the time it takes to understand experimental data



Maximise the scientific impact and success of ESS

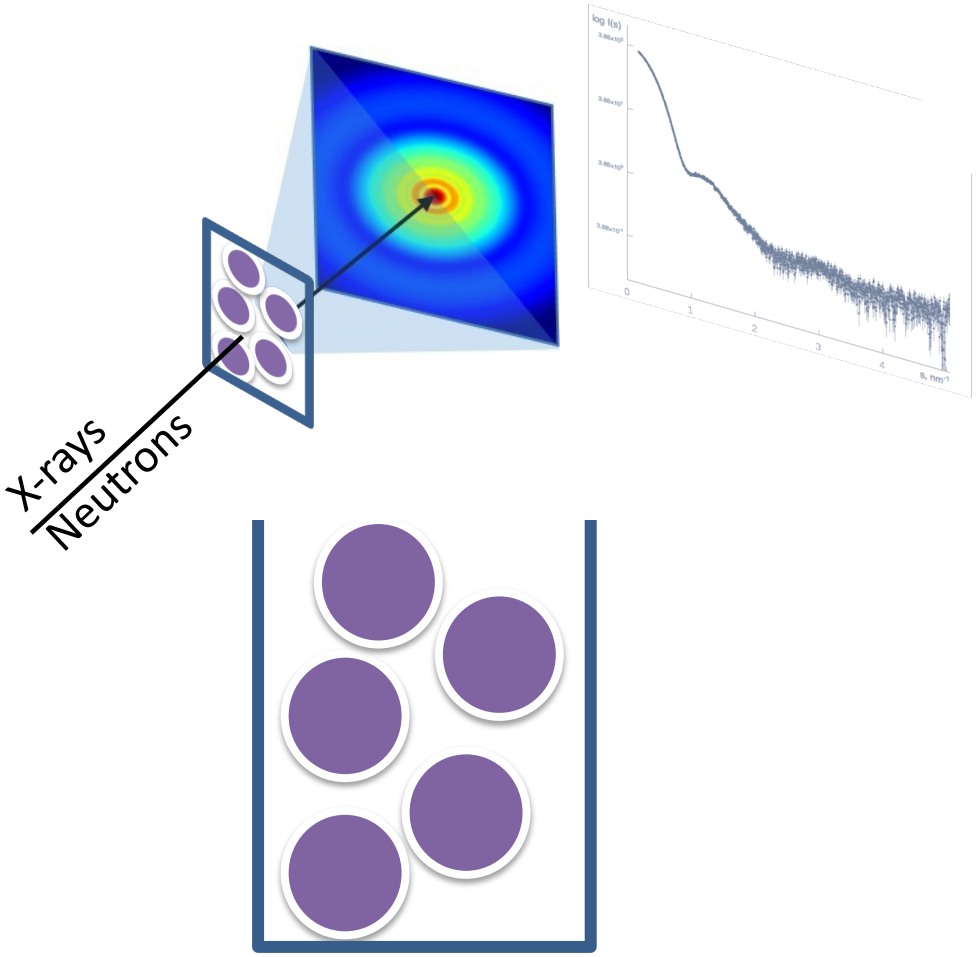
We follow the best software development practice!

ESS/DMSC engages in the collaborations

Ingemar André, Lund University

Jill Trewhella, Univ. of Sydney

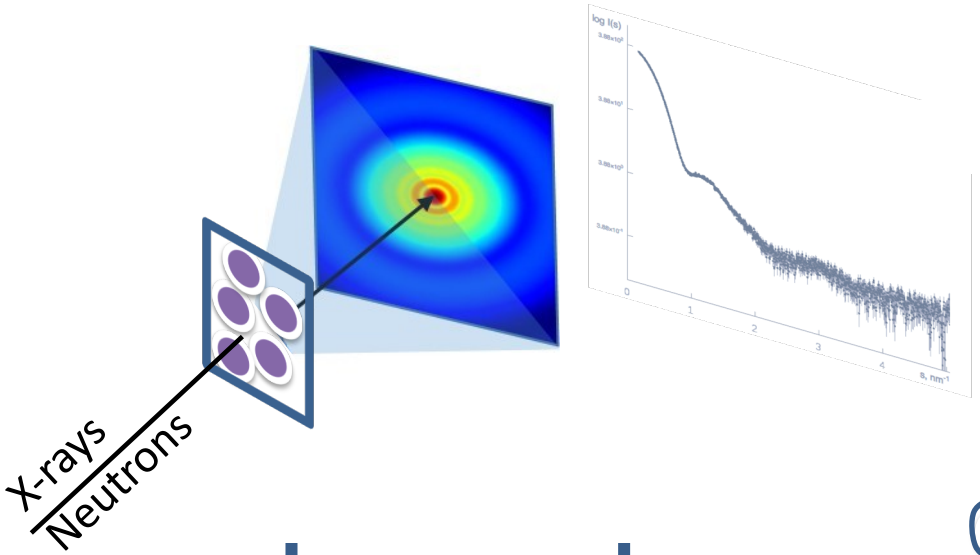
Monodispersity is desired in SAS experiment

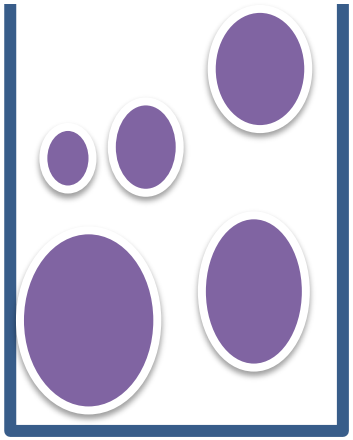


Monodisperse sample

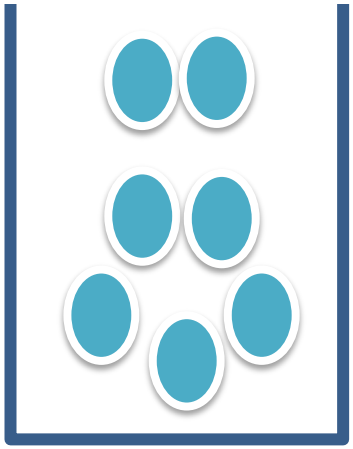
Adapted from Di Cola et al. Pharmaceutics, 2017

Polydispersity complicates structural determination

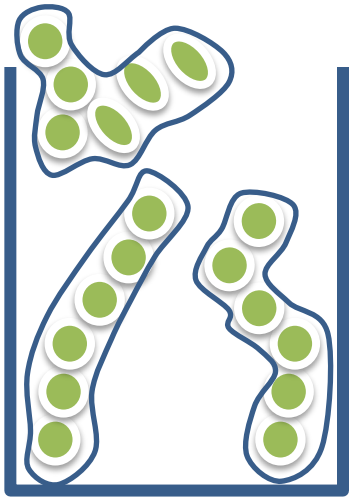




Shape polydispersity

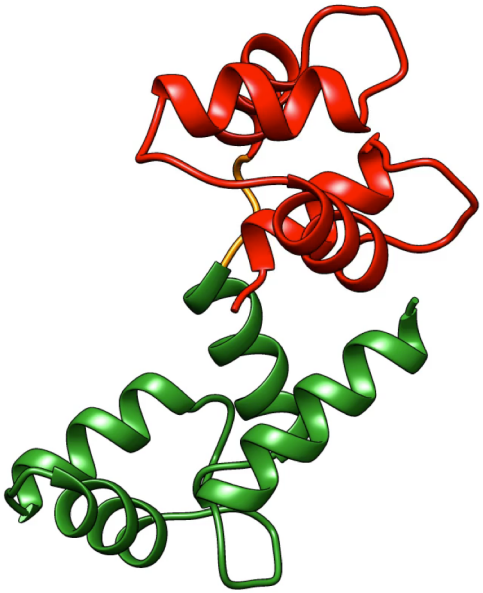


Size polydispersity



Conformational polydispersity

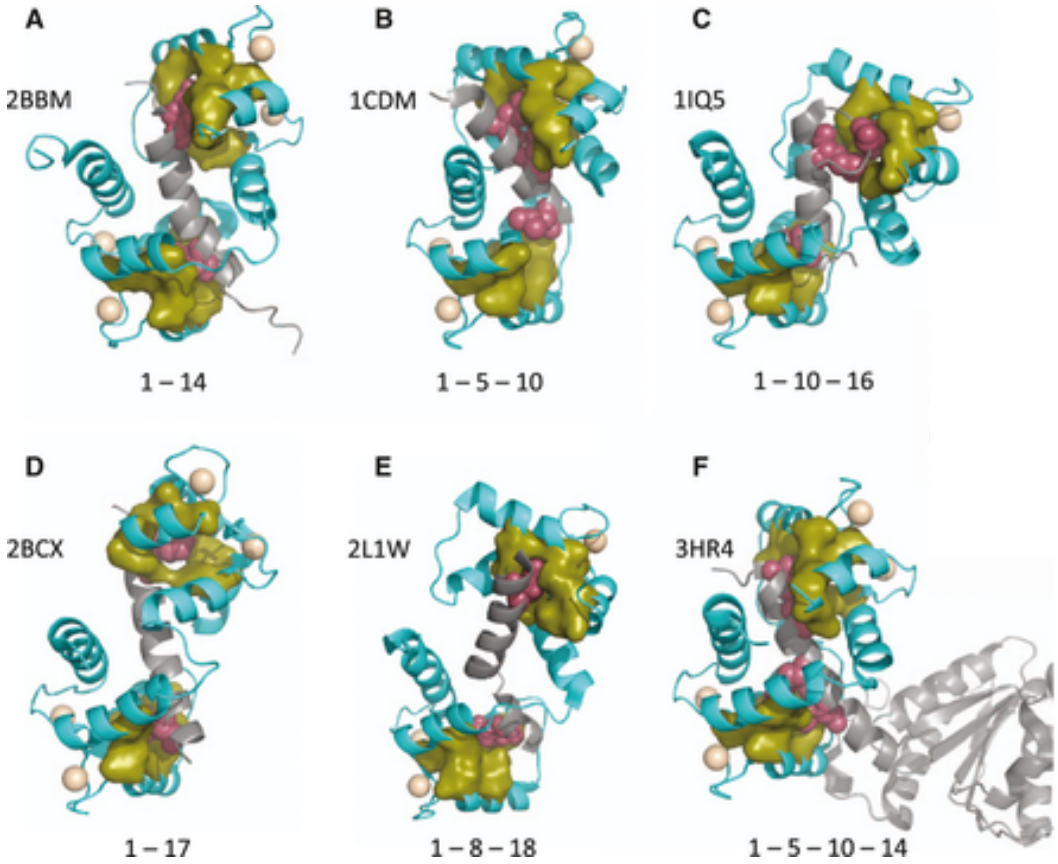
Proteins with flexible linkers are attractive targets for SAS experiments



Calmodulin (CaM)

Cardiac myosin binding protein C (*Δ*mC2)

Domain-domain flexibility plays a central role in biology

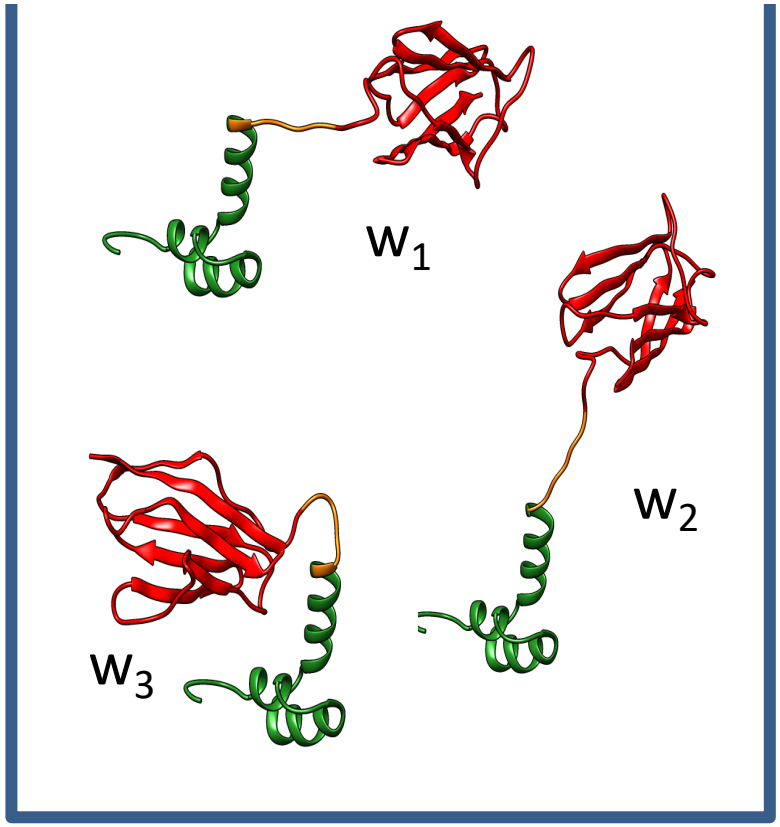


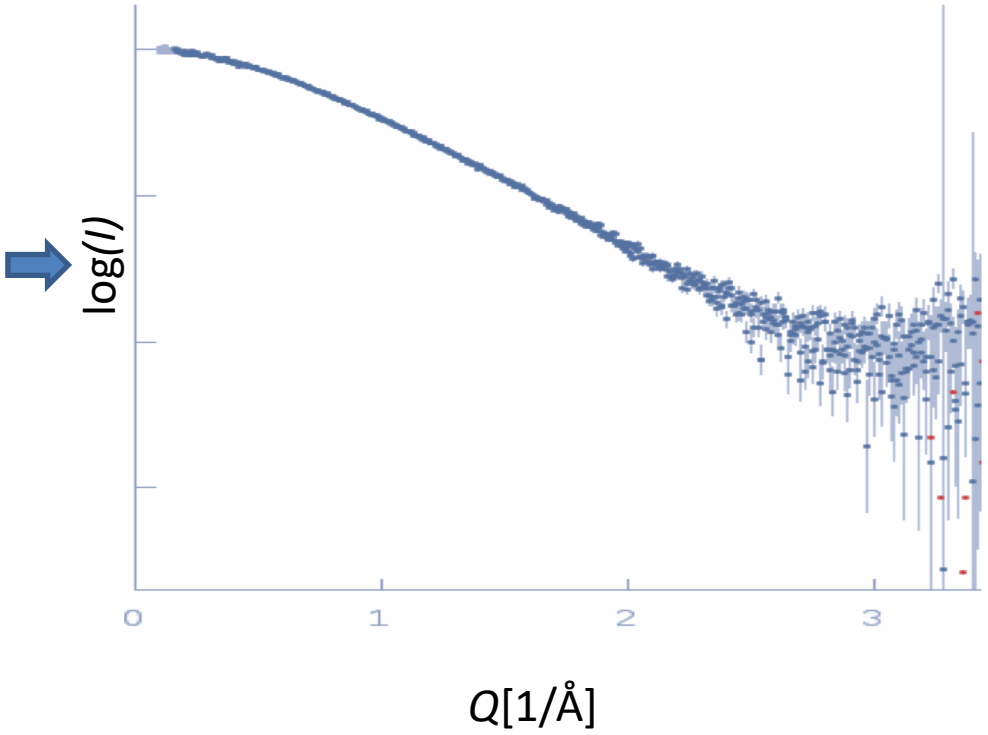
Structural diversity of calmodulin binding to its target sites, Volume: 280, Issue: 21, Pages: 5551-5565, First published: 20 April 2013, DOI: (10.1111/febs.12296)

Challenges for analyzing poldydisperse data

- Scattering from the mixture of components
- Low information content
- Overfitting to experimental noise
- Optimization problem – thousands of parameters

Multiple conformations contribute to scattering curve

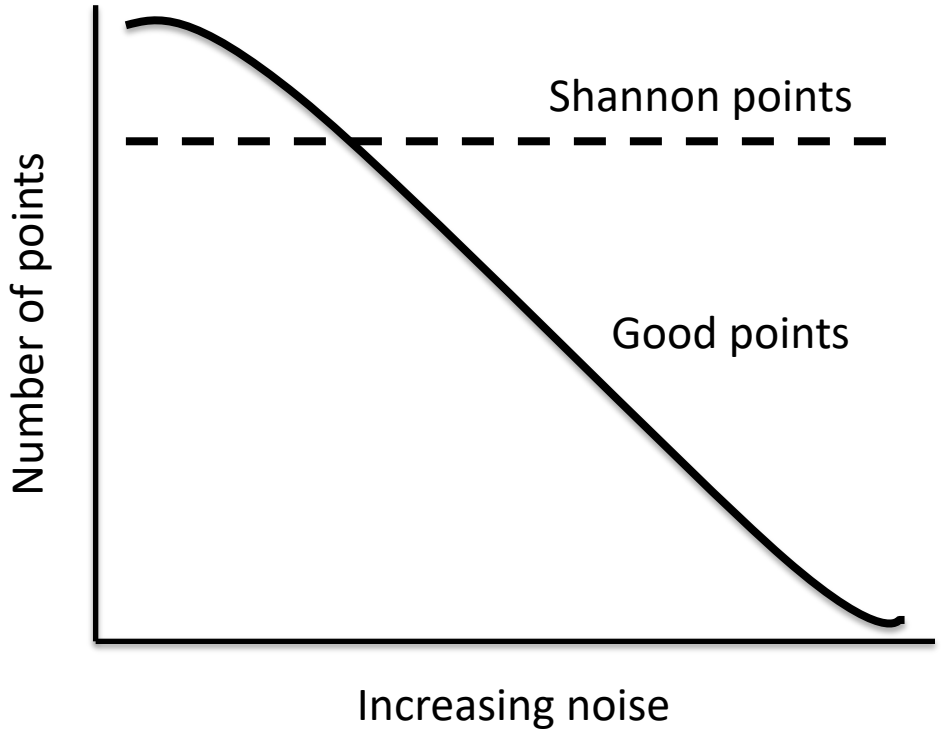




$w_1 + w_2 + w_3 = 1$

SAXS data contains limited structural information

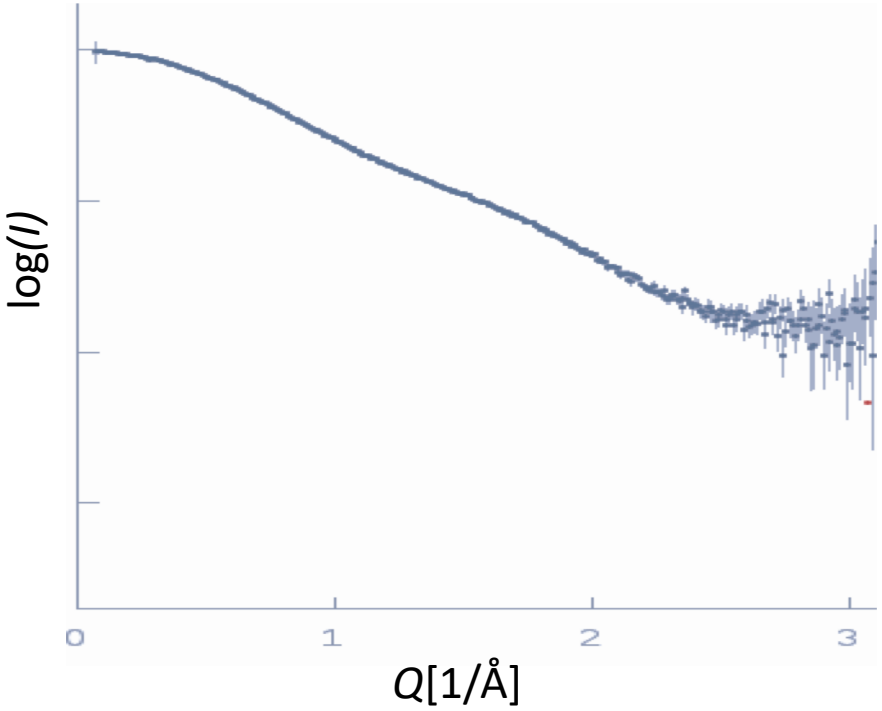
The number of independent data points estimated from Nyquist-Shannon sampling theorem or maximum entropy regularization (Vestergaard and Hansen (2006))



The number of independent measurement rarely exceeds **5-15**.

SAXS data contains limited structural information

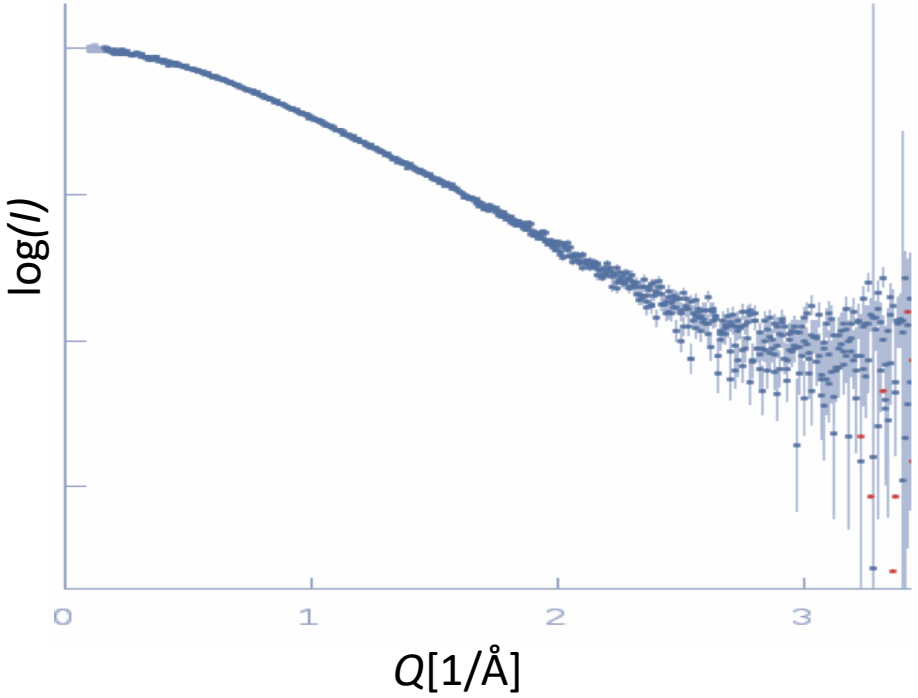
Calmodulin (SEC-SAXS)



$N_s = 6.4$

$N_g = 5.1$

Cardiac myosin-binding protein C (SAXS)

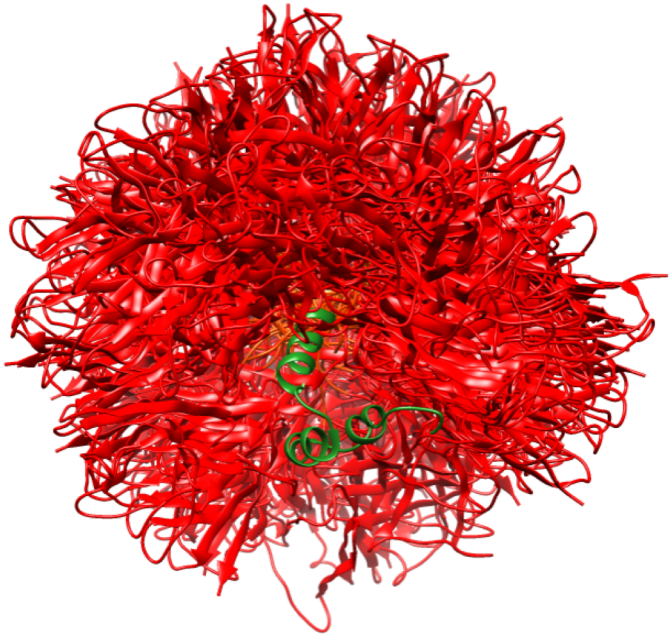


$N_s = 8.7$

$N_g = 6.1$

Many parameters describe conformational ensembles

- Thousands of parameters may be required to define ensemble
- Ensembles can be represented in discrete or continuous space



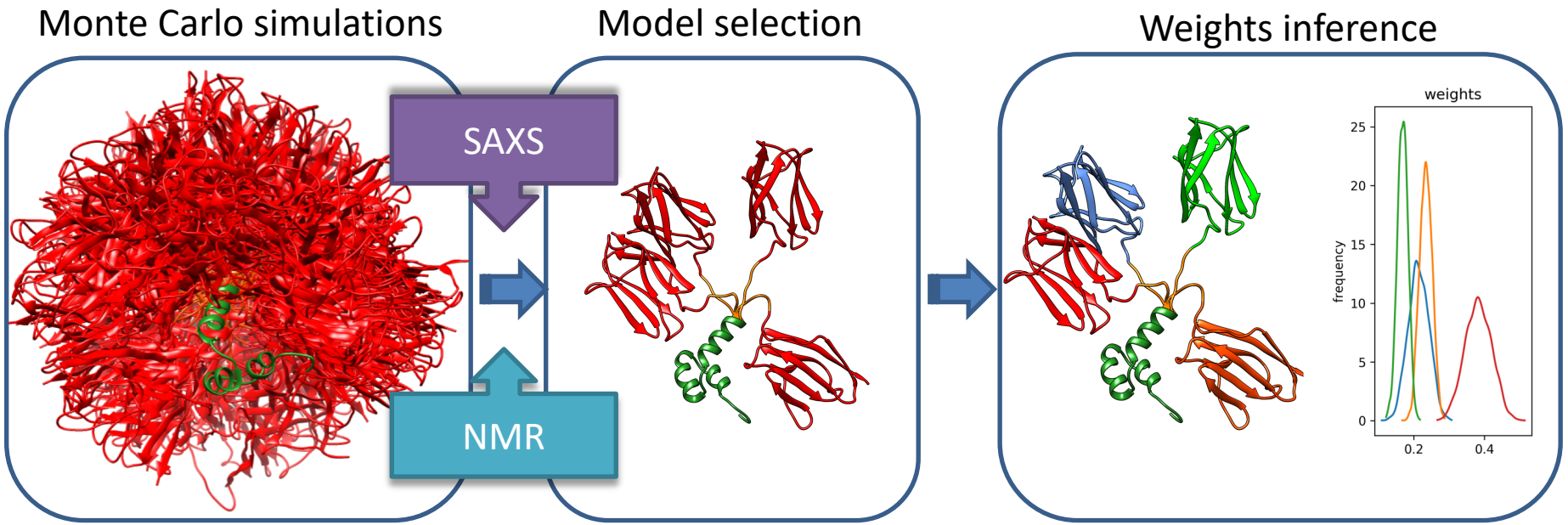
- χ^2 based optimization leads to overfitting

What do we want from our inference method?

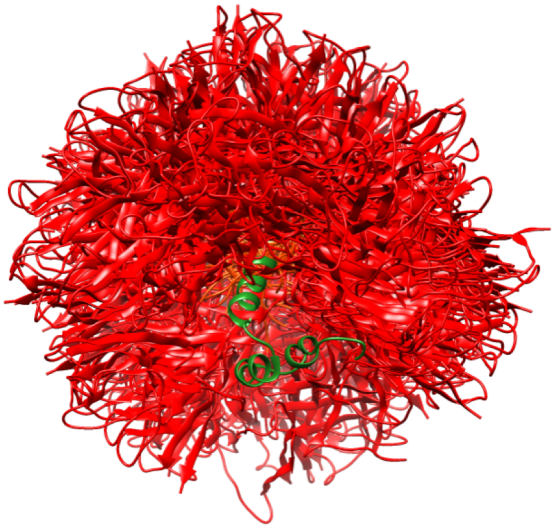
- We want to extract as much information from the experimental data – but not more than that
- Avoid overfitting to experimental noise
- Quantify the uncertainty in the inference
- Combine multiple sources of data: simulation and experimental data sources

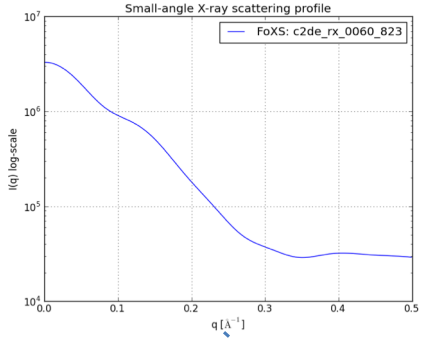
-> Bayesian Inference to the rescue!

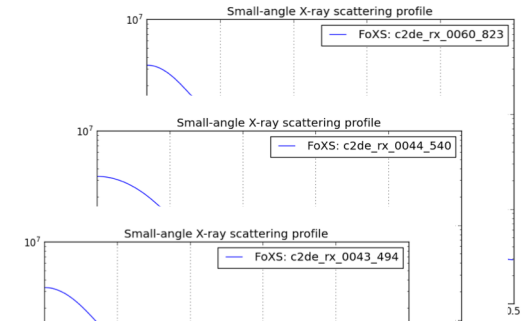
Method to infer conformational ensembles



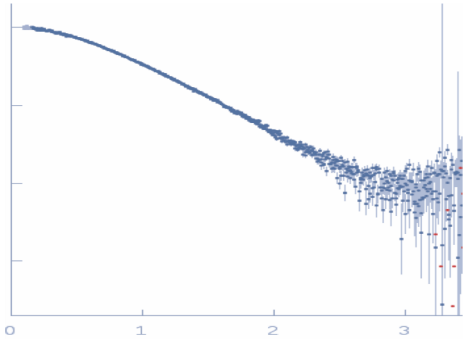
Scattering patterns are generated after Rosetta modeling (once)







Crysol, FOXS



Experimental

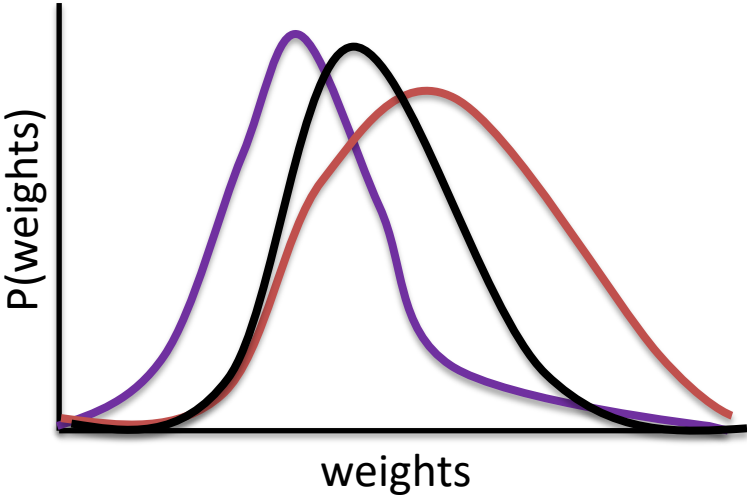
Bayesian statistics can be applied to infer weights

Bayes rule:

Thomas Bayes

$$P(w|Data) = \frac{P(Data|w) * P(w)}{P(Data)}$$

$$Posterior \propto likelihood * prior$$



Posterior $\left(P(w|Data)\right)$: Probability of w after observing data

Prior $(P(w))$: Information before observation of data

Likelihood $(P(Data|w))$: Probability of observing data given w

$P(Data)$: Normalization constant also refered to Model evidence

Approximate and complete inference complements each other

Approximate Bayesian inference – fast, less accurate:

Normal distribution

Dirichlet distribution

Simulated annealing → $P(w, Ensemble|Data) \propto P(Data|w, Ensemble) * P(w|Ensemble)$

Assume Dirichlet distribution

Complete Bayesian inference - slow:

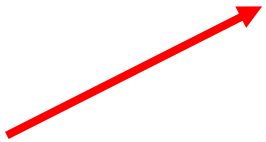
Normal distribution

Dirichlet distribution

Hamiltonian Monte Carlo → $P(w|Data) \propto P(Data|w) * P(w)$

Combining multiple data sources increases structural information

$$P(w|SAXS, NMR, conformer\ energy) \propto P(SAXS|w) * P(NMR|w) * P(w(conformer\ energy))$$



SAXS data – Rigid body orientation of domains

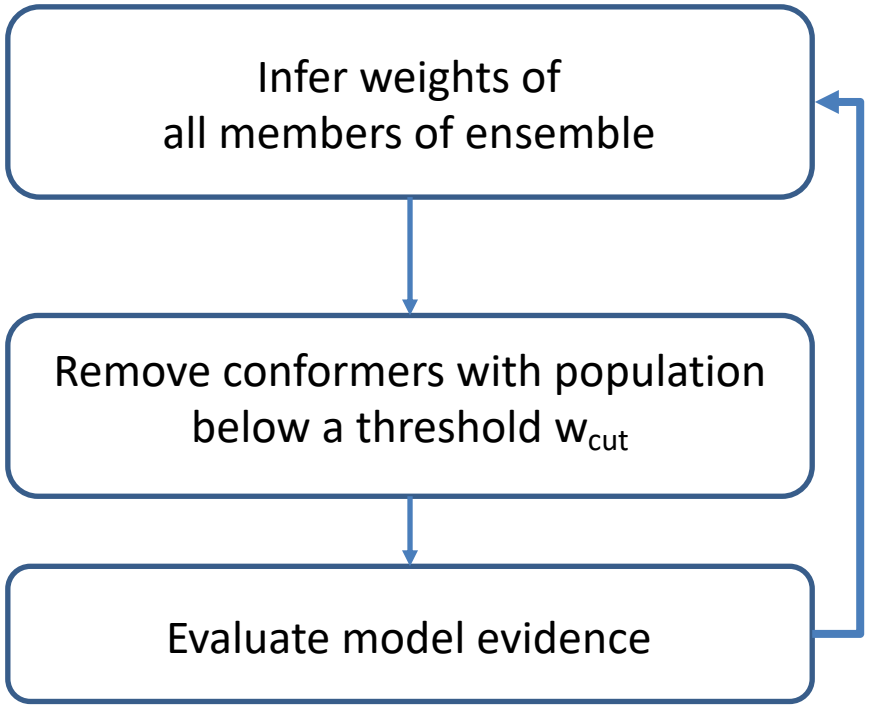
NMR chemical shift data – Local contacts

Energy of conformer from energy function – stereochemistry and intramolecular interactions

Approximate Bayesian inference balances complexity with fit to the data

Fisher CK, Ullman O, Stultz C. (2012) Pacific symposia on biocomputing.

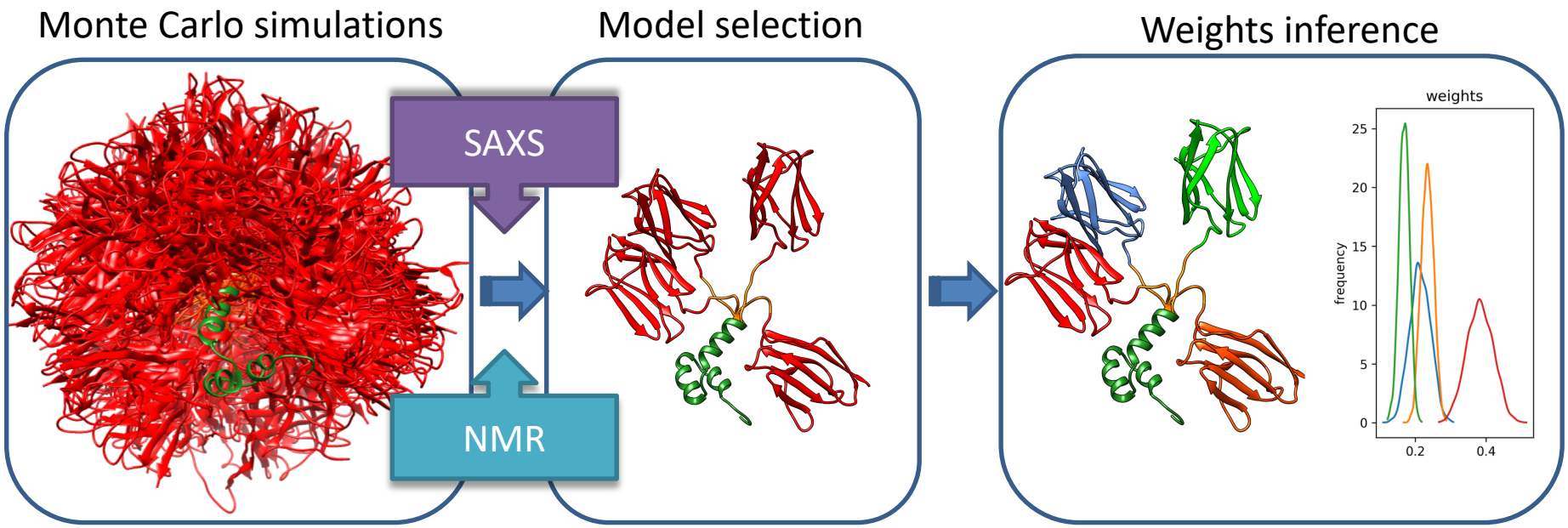
For a given set of conformers



$$P(w, Ensemble|Data) = \frac{P(Data|w, Ensemble) * P(w|Ensemble)}{P(Data|Ensemble)}$$

Model evidence $\longrightarrow P(Data|Ensemble) \propto P(Ensemble|Data)$

Method to infer conformational ensembles



Rosetta structure prediction

10 000 decoys
(1000 low energy)

Approximate Bayesian

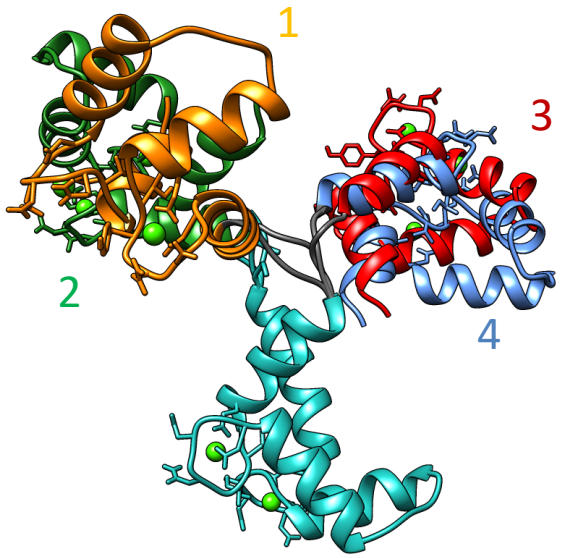
3-5 conformers

Complete Bayesian

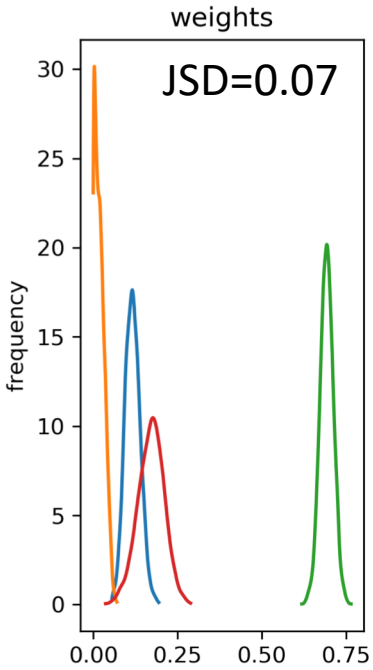
3-5 conformers

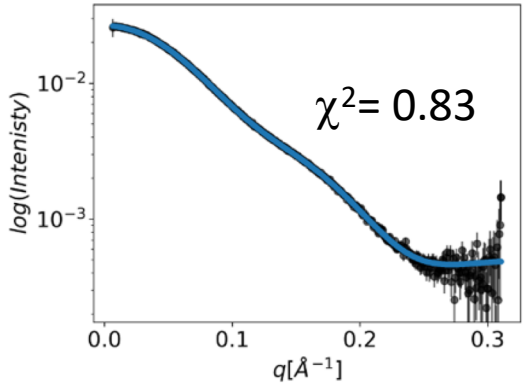
Calmodulin ensemble inferred from SAXS data

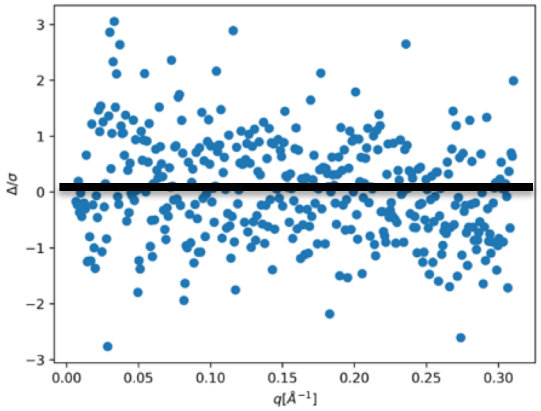
posterior ∝ likelihood(SAXS) · prior(weights)



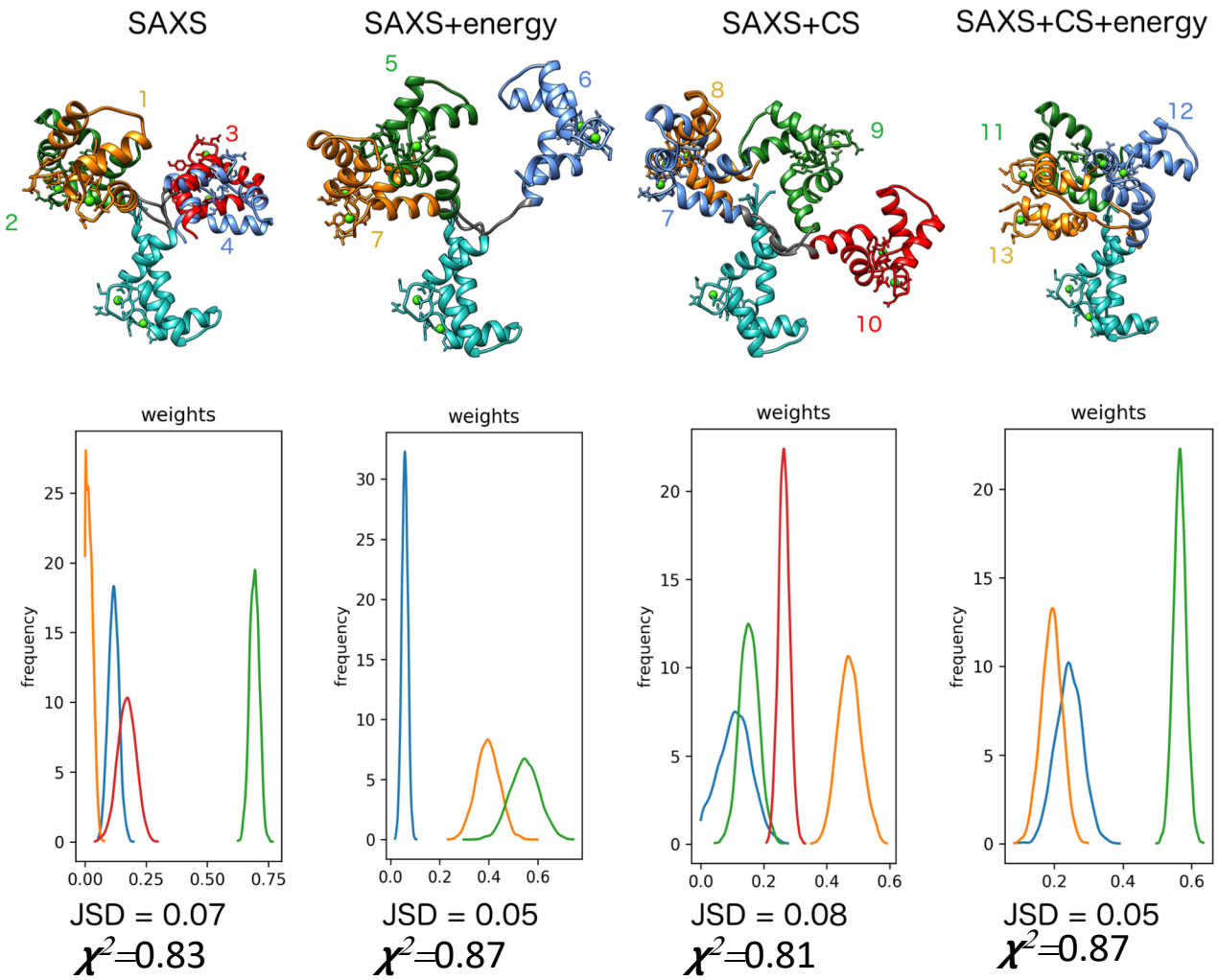
Conformer	Population weight
1	0.02 ± 0.01
2	0.69 ± 0.02
3	0.17 ± 0.04
4	0.12 ± 0.02





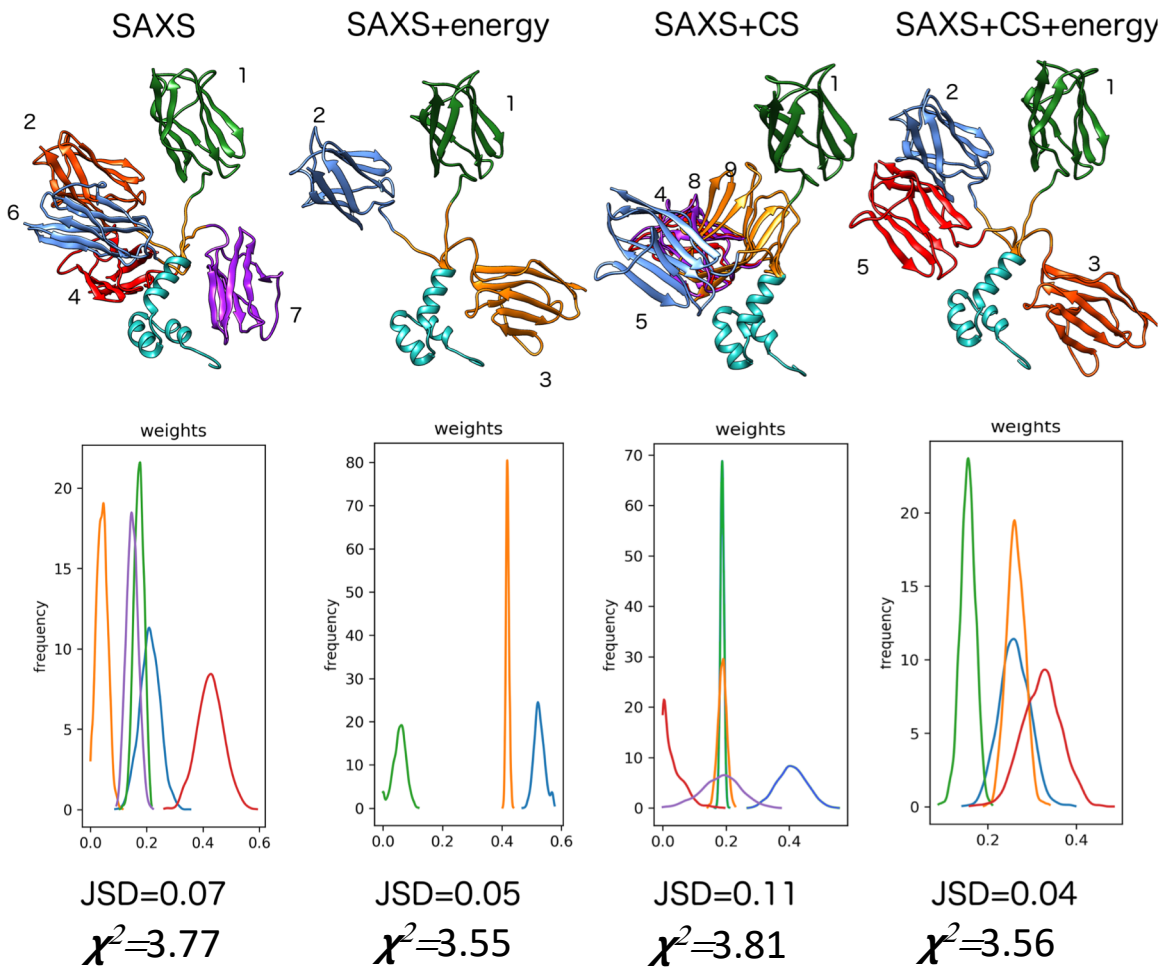


Combination of different data sources explains SAXS data equally well



energy = energy from molecular simulations
CS = NMR chemical shifts
JSD = Jensen-Shannon weights divergence

Combination of different data sources improves definition of ΔmC2 ensemble



energy = energy from molecular simulations
CS = NMR chemical shifts
JSD = Jensen-Shannon weights divergence

The drop in χ^2 is the result in the improved quality of ensemble

Conclusions I

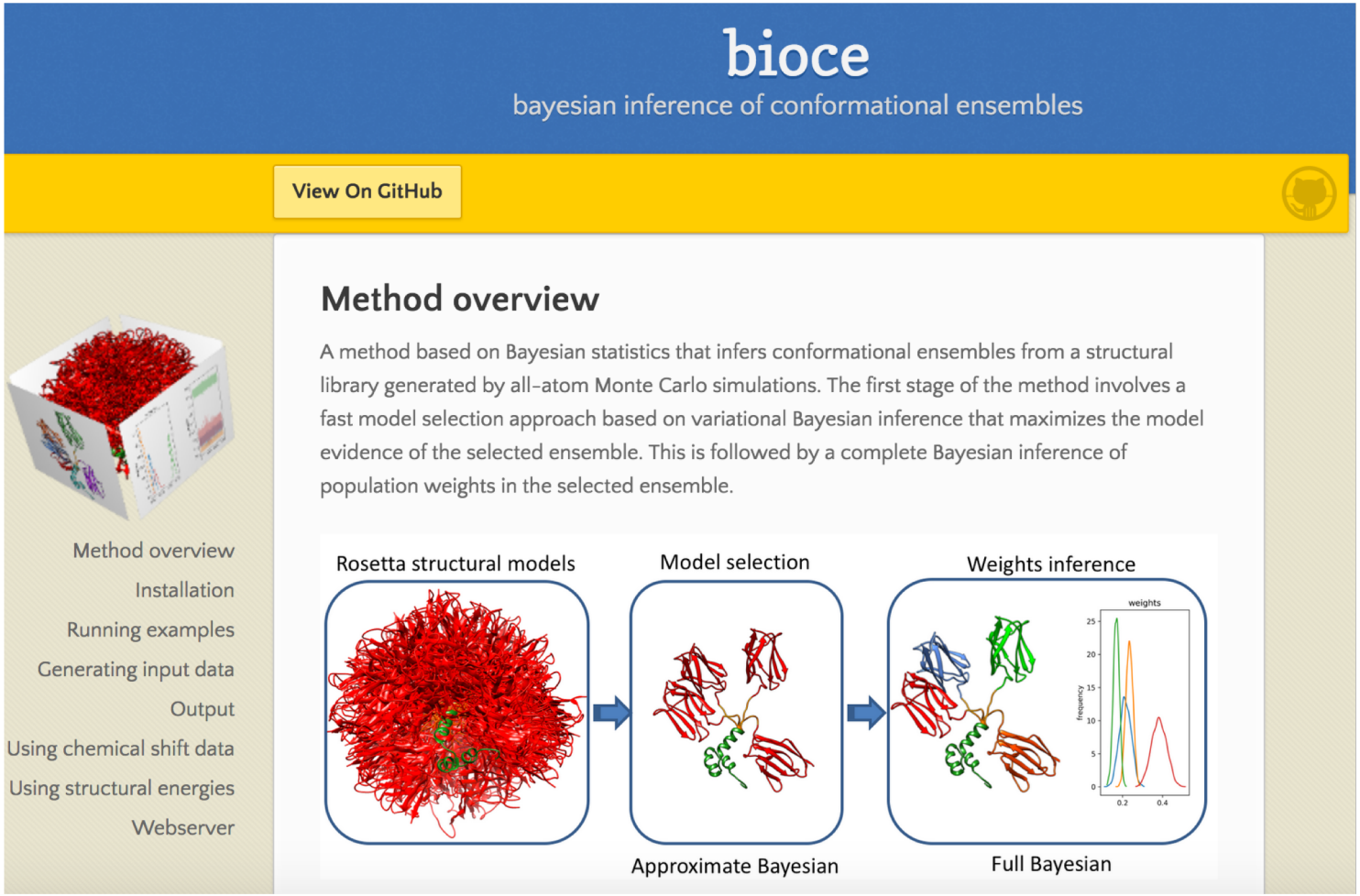
Bayesian Inference:

- can be used to optimally balance model complexity and fit to experimental data
- gives information about the uncertainty of inferred ensembles and parameters
- allows for combining information from different data sources in one model

Our method combines the speed of alternative methods with the benefits of Bayesian inference

Method is available for download

https://andre-lab.github.io/bioce/



Exploring protein association pathways with time resolved SANS/SAXS

Ingemar André
(LU)

Ryan Olivier
(LU)

Najet Mahmoudi
(ISIS)

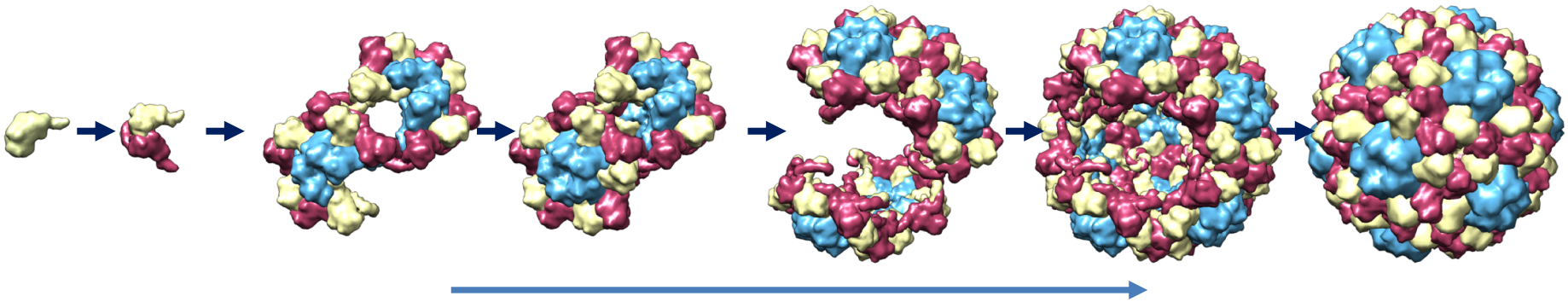
Thomas Holm Rod
(ESS)

Martin Nors Pedersen
(KU)

Lise Arleth
(KU)

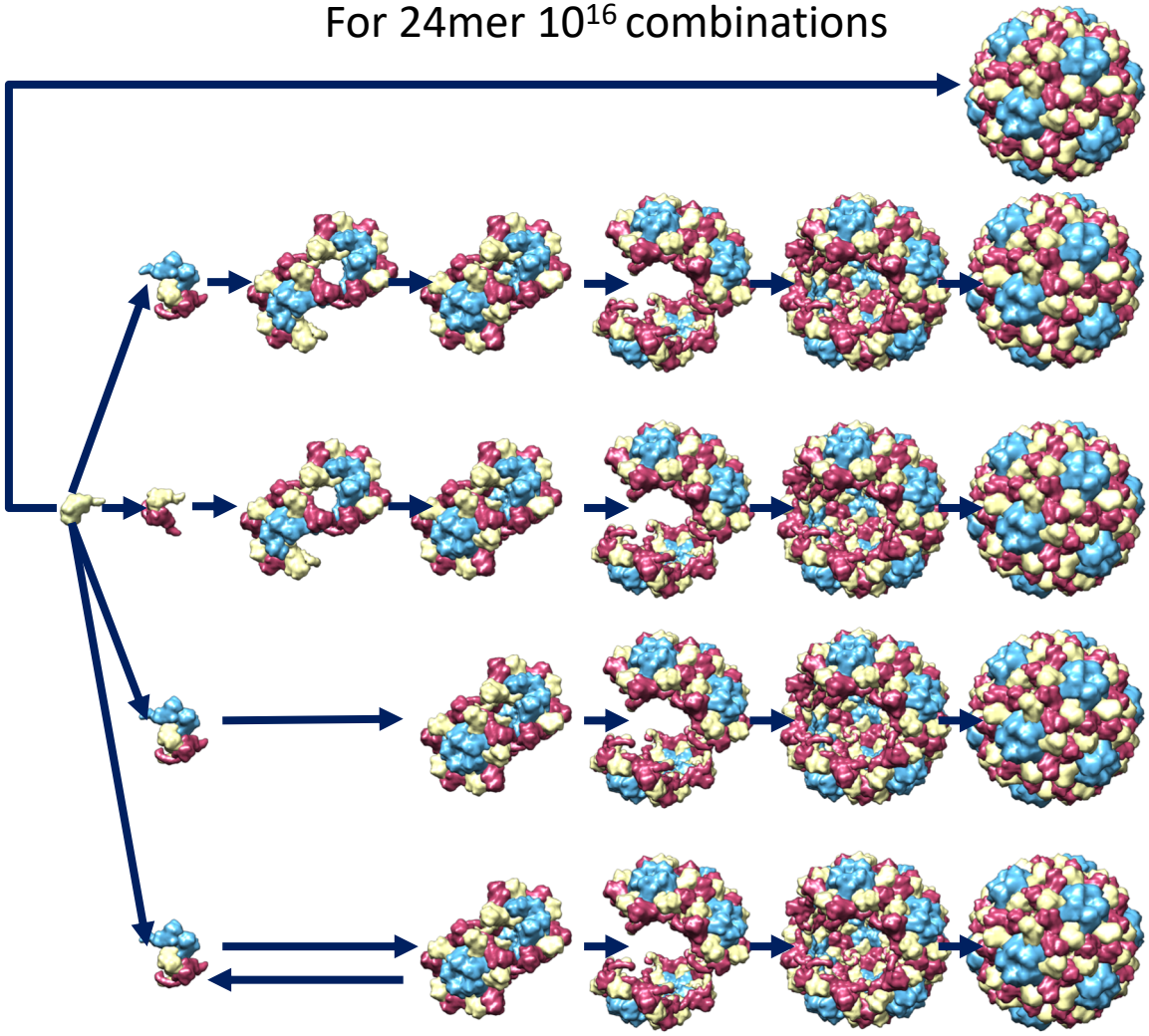
Why to study protein association pathways?

- Self-assembly central to many processes in different domains
- Key to understanding life processes at molecular level
- For virus capsids:
  - Start and end points are often known
  - But not what happens in between
  - E.g. critical in designing antiviral therapies

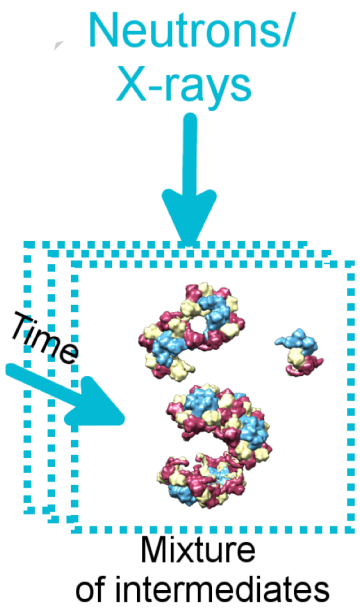


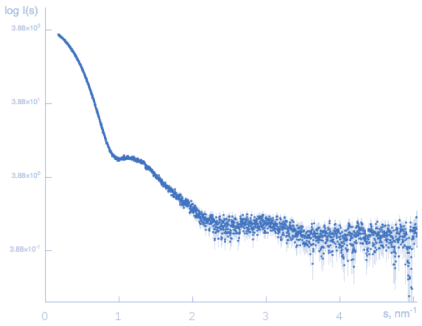
Cowpea chlorotic mottle virus association pathway?

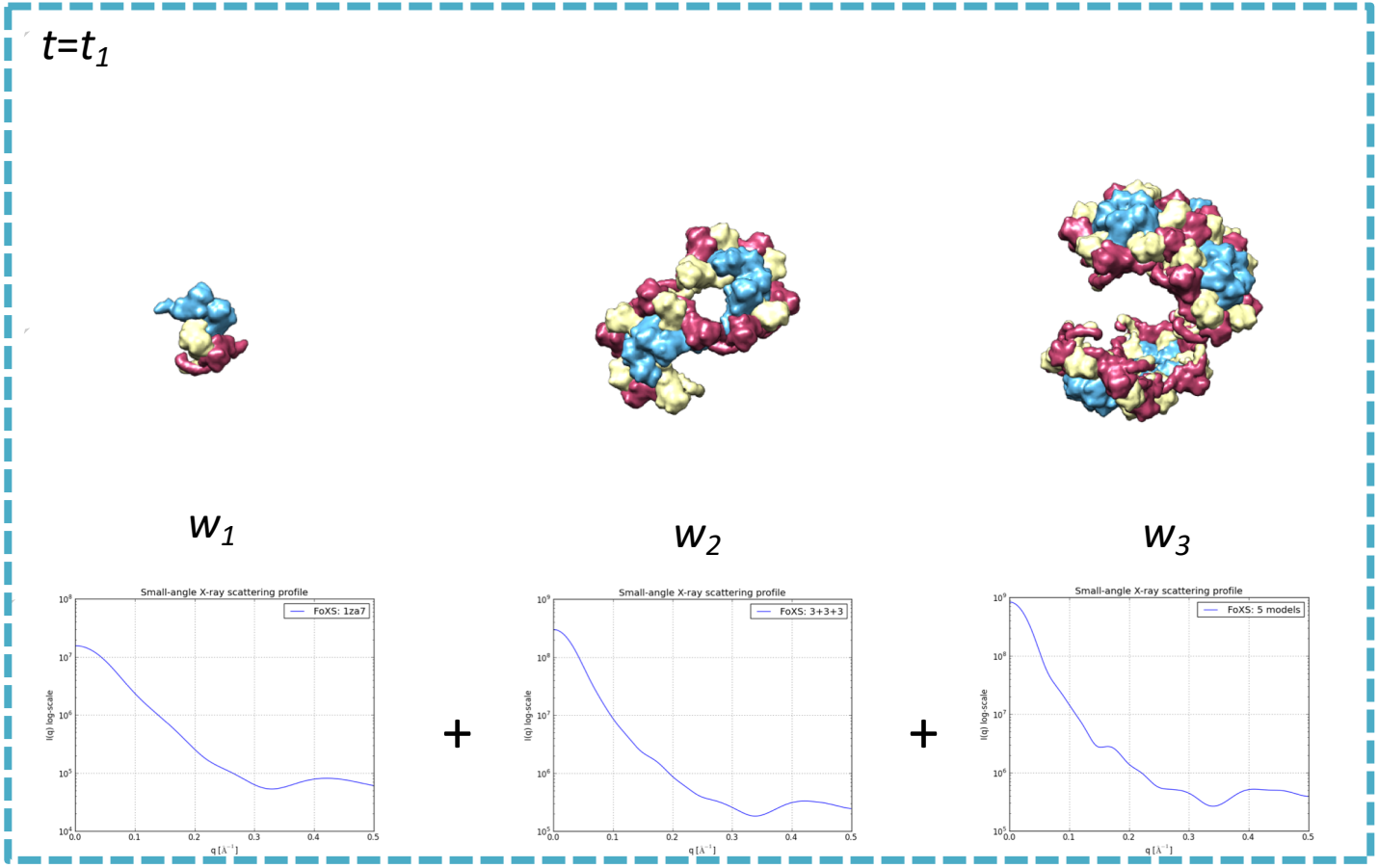
Humongous number of possible pathways



Method for analyzing SAS data from the mixture of components

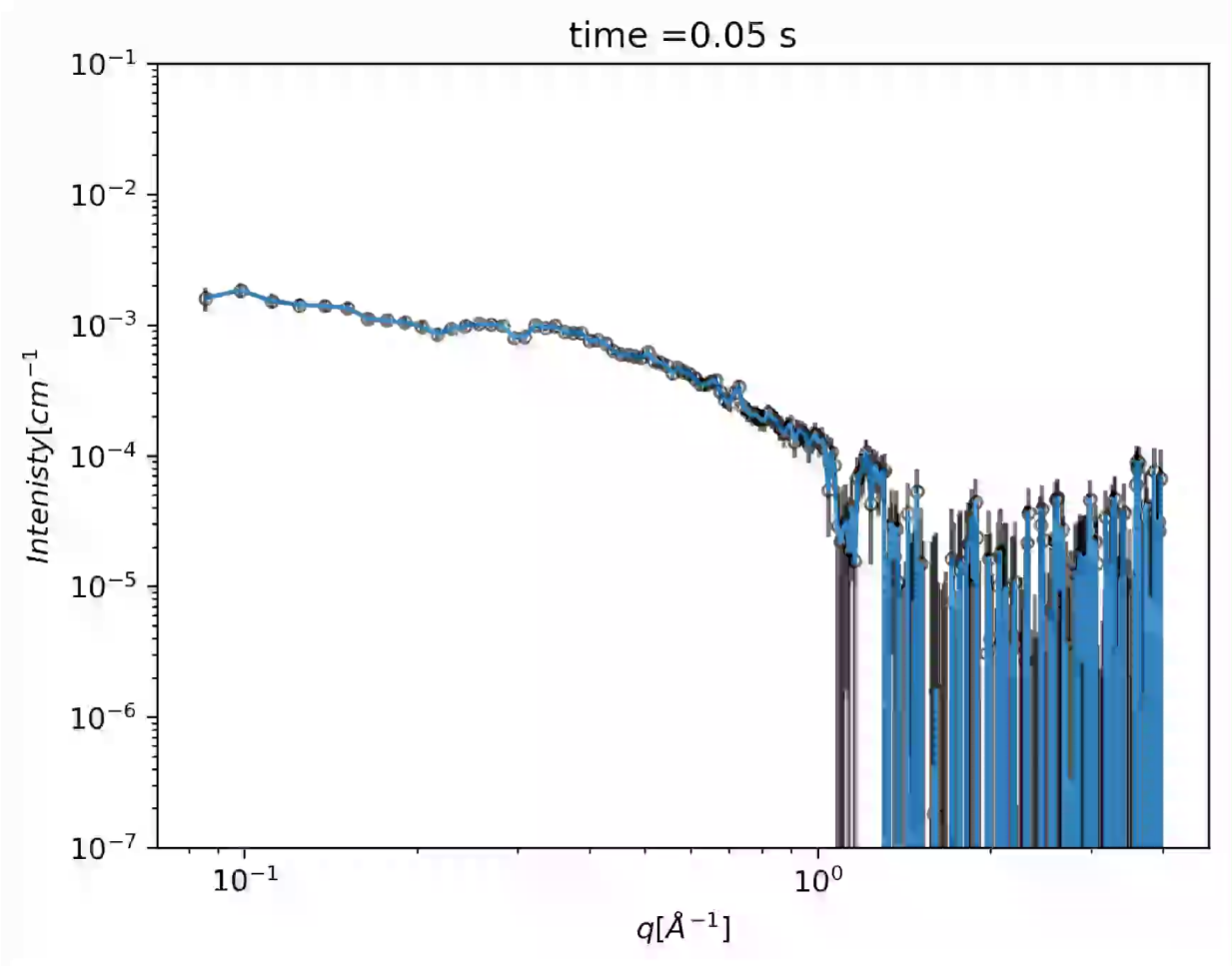






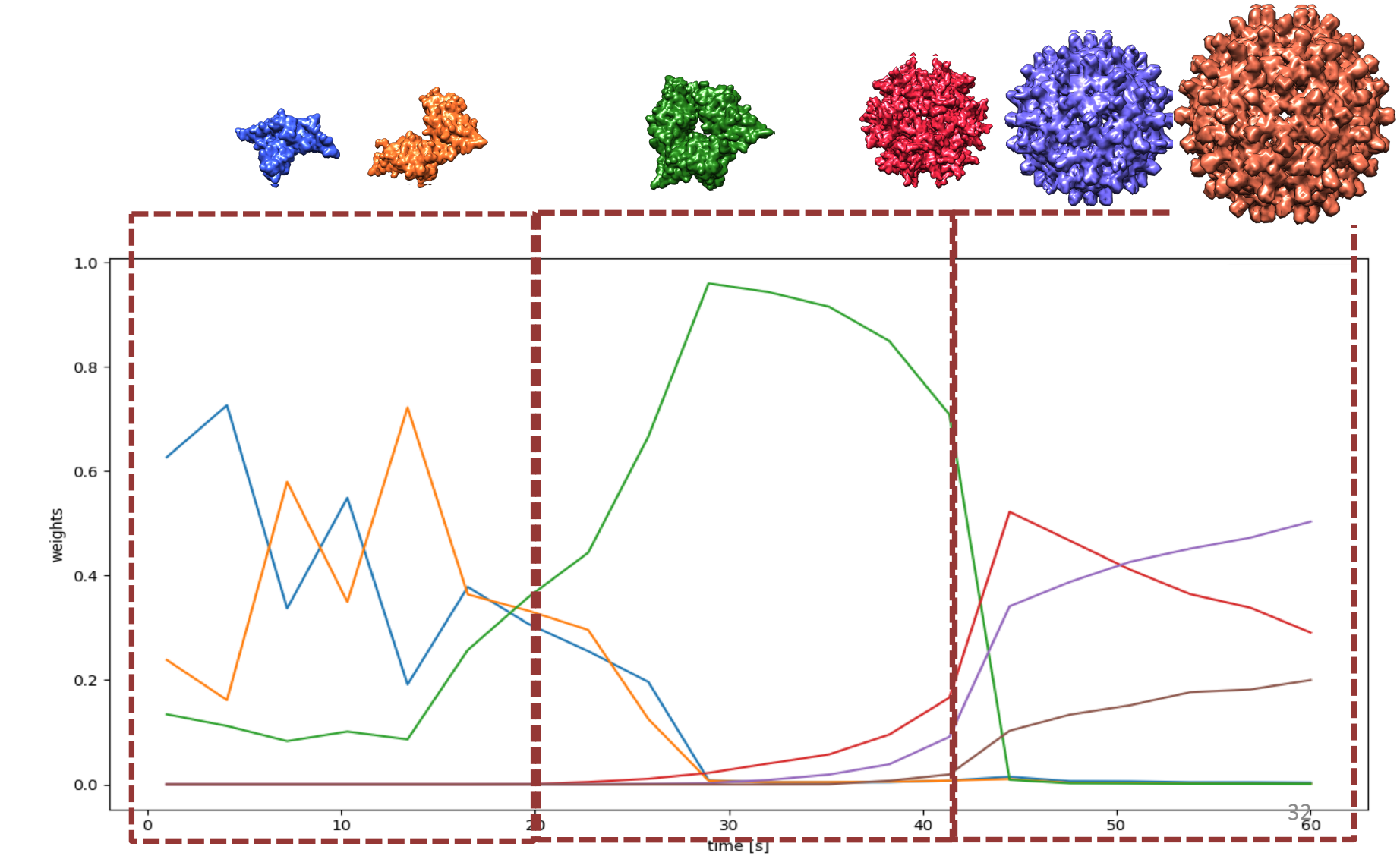
Kinetic model $\alpha D \underset{k_1^-}{\overset{k_1^+}{\rightleftharpoons}} I$ $\beta I + \alpha D \underset{k_2^-}{\overset{k_2^+}{\rightleftharpoons}} C$

Time-resolved SAXS captures virus capsid assembly



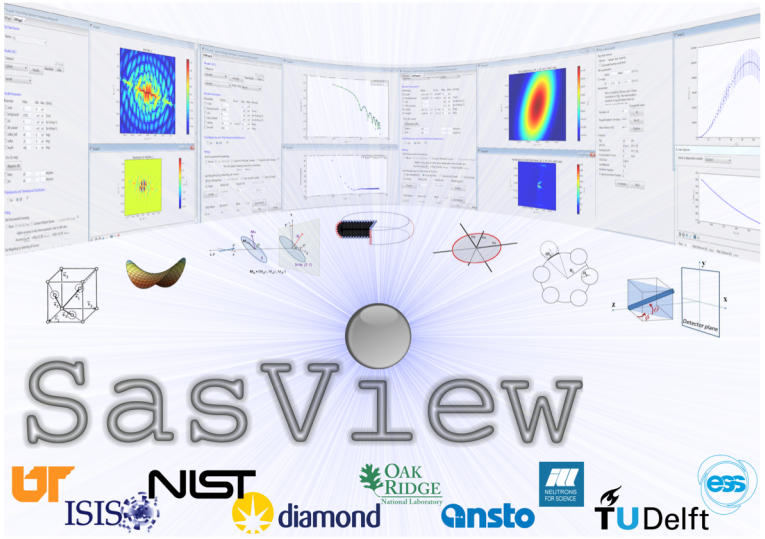
Data collected at ESRF, ID02, Nov. 2018

Intermediates can be inferred from time-resolved SAXS data



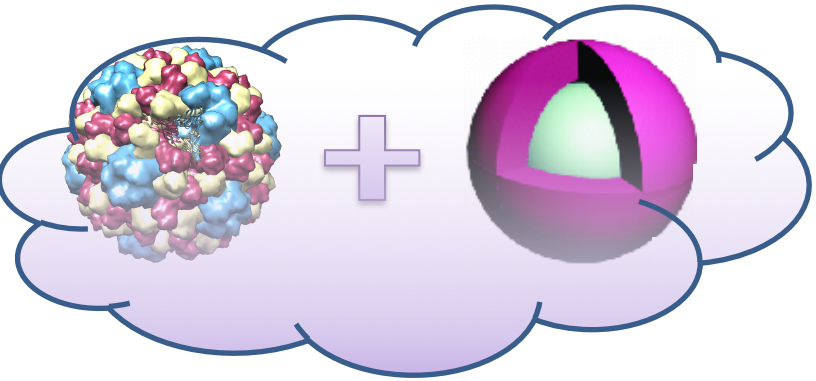
SasView analyzes and models SAS data

- Community driven project
- 30 developers from X-ray and neutron facilities
- Engaged user community
- Regular code camps and releases
- **We are always on the lookout for new collaborations**

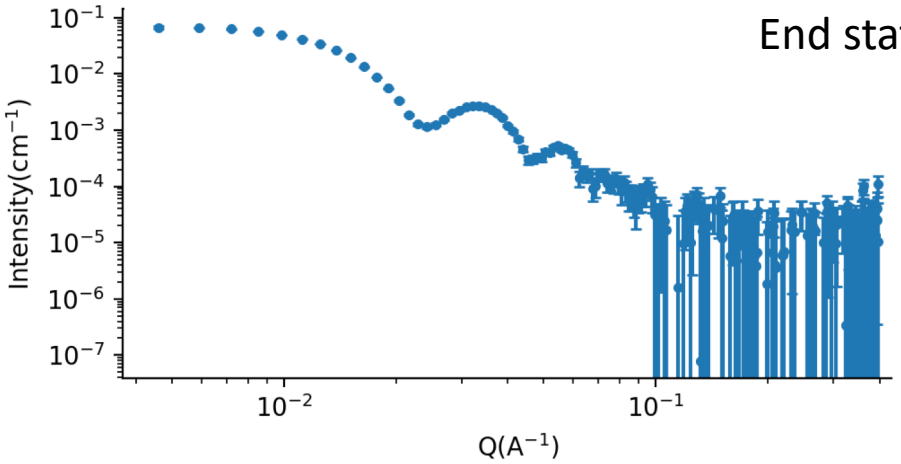


- Fitting geometrical models to data
- P(r) inversion
- Model-free calculations
- Correlation functions
- Numerous utility tools

Bayesian inference combines atomistic and coarse-grained models



End state



Parameter	Value
Weight T3 capsid	0.4
Weight T4 capsid	0.15
Weight core shell sphere	0.45
Core radius [Å]	90
Shell radius [Å]	30

Conclusions II

- Bayesian inference methods can be adapted to analyze time-resolved SAXS data
- Models of various accuracy can be combined into a single statistical framework

Thank you!

Lund University:
Ingemar André
Ryan Oliver

ESS:
Thomas Holm Rod

Univ. of Sydney:
Jill Trewhella

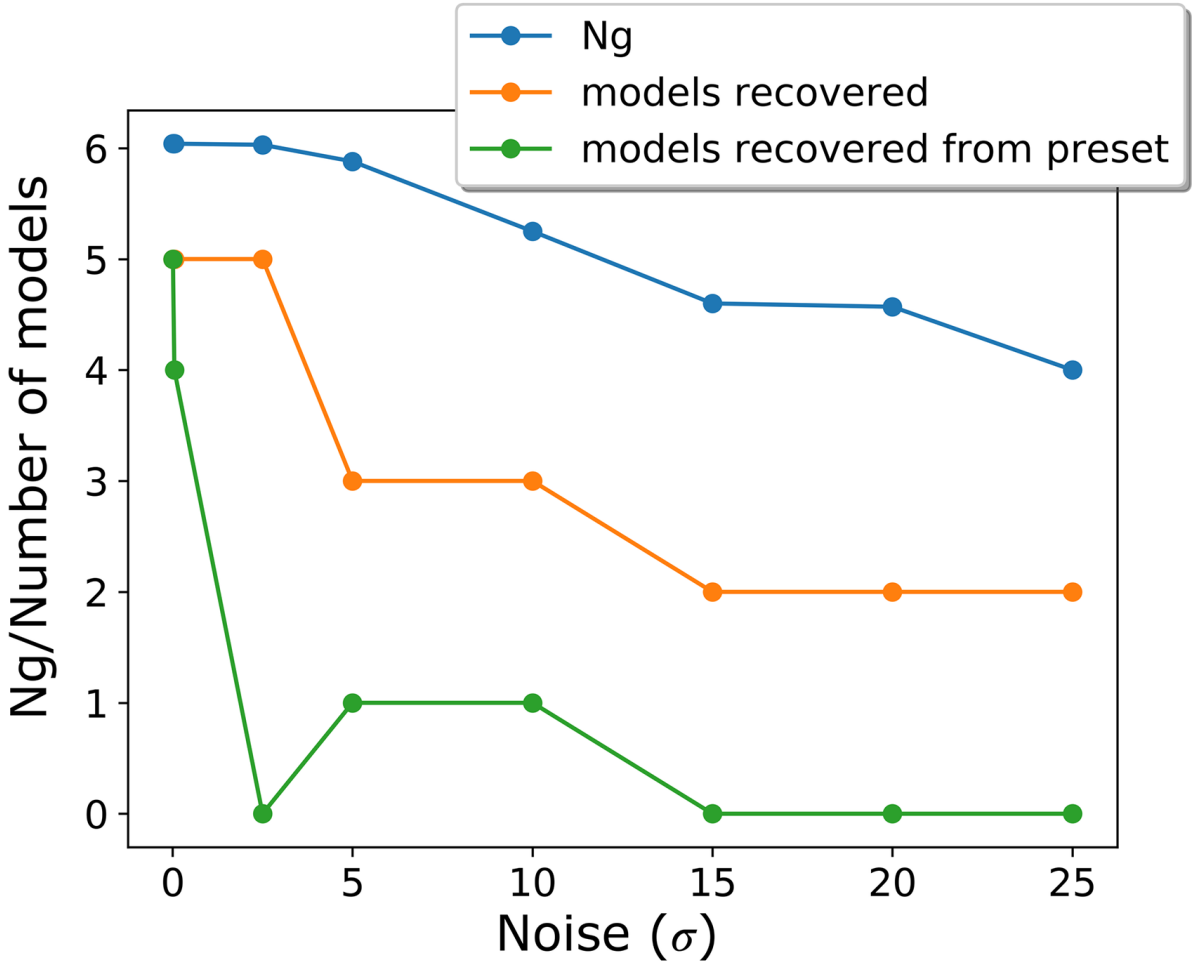
ISIS neutron and muon source:
Najet Mahmoudi

Univ. of Copenhagen:
Martin Nors Pedersen
Lise Arleth

SasView collaboration

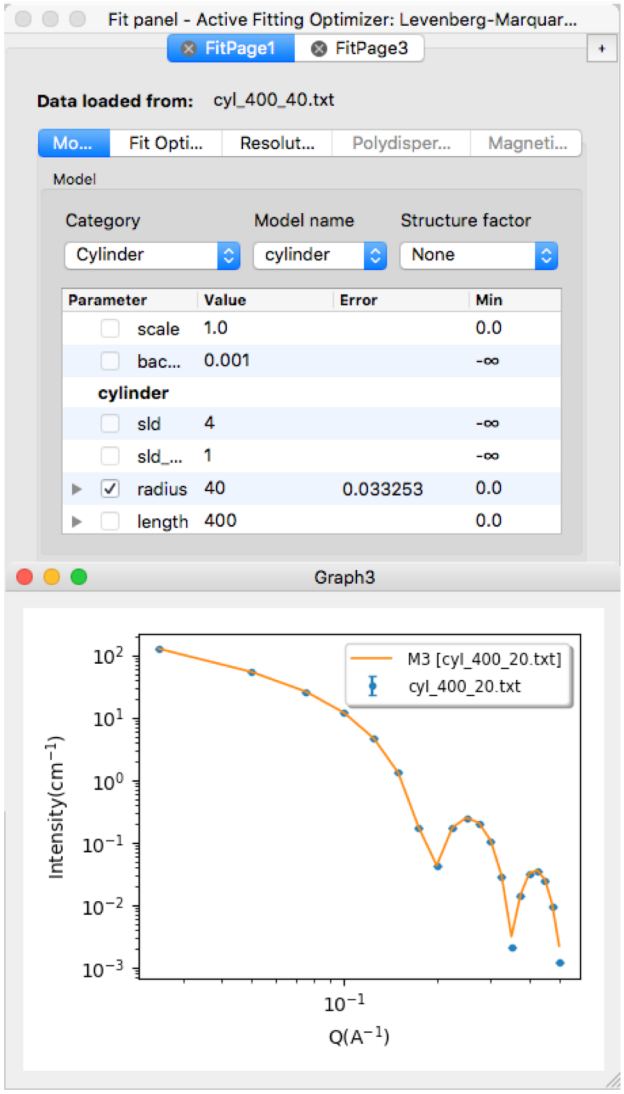
Funding:

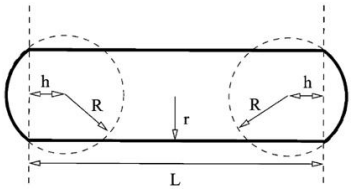
Approxiamte bayesian recovers ensembles up to high level of noise

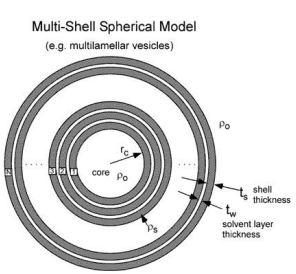


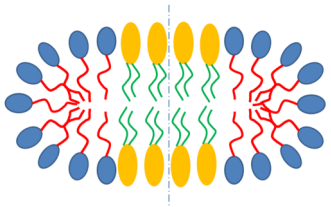
Simulated ensemble with 5 members from a library of 100 conformers

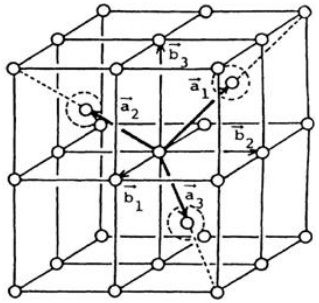
SasView has large collection of models

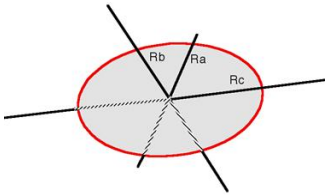


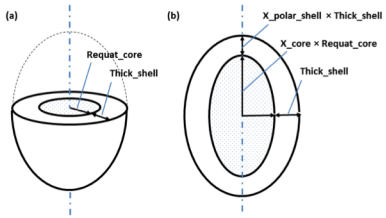


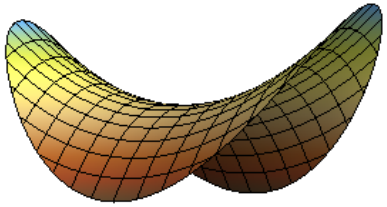


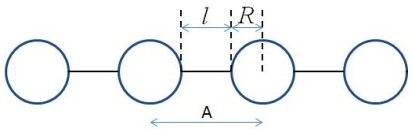


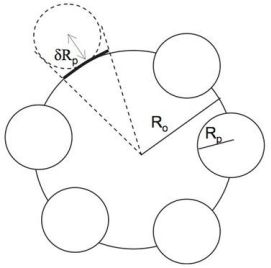




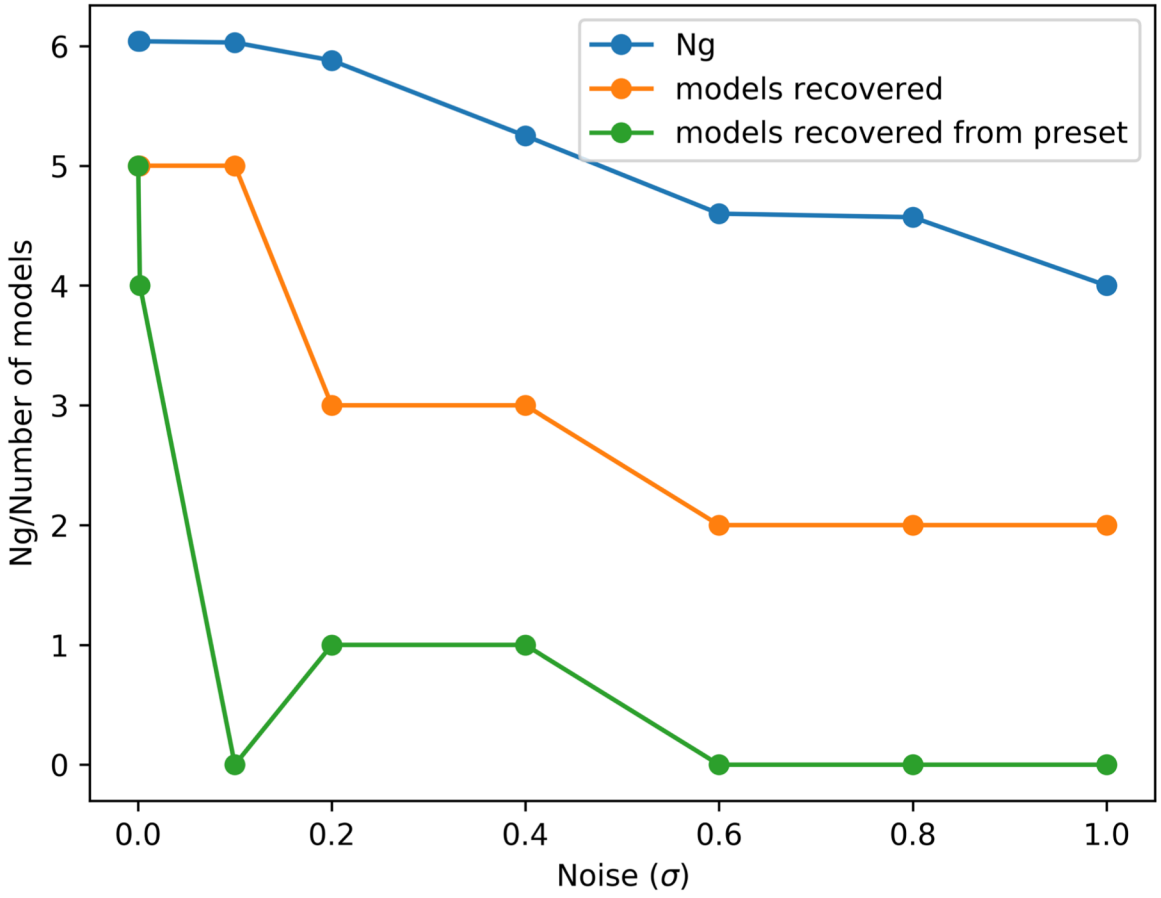








Variational Bayesian Inference on noisy data



Simulated ensemble with 5 members from a library of 100 conformers

Information content in SAS

From Nyqvist-Shannon sampling theorem

- The scattering curve $I(q)$ of object with maximum diameter d is fully represented by a set of points

$$q_n = n\pi/\mathrm{d}, \text{ where } n = (1,\ldots,\infty).$$

- The number of independent data points (Shannon channels)

$$N_s = d(q_{\max}-q_{\min})/\pi$$

Bayesian ensemble modeling: complete inference

A ***model*** (**M**) is a set a set of structures S. For a given model:

$$P(\boldsymbol{w}|\boldsymbol{D}) = \frac{P(\boldsymbol{D}|\boldsymbol{w}) * P(\boldsymbol{w})}{\int P(\boldsymbol{D}|\boldsymbol{w}) * P(\boldsymbol{w}) d\boldsymbol{w}}$$

The **prior** of **w** can be selected as a flat prior or based on Boltzmann distribution

$$w_i = \frac{e^{-U(S_i)/k_B T}}{\sum_{j=1}^{n} e^{-U(S_i)/k_B T}}$$

The **likelihood** of **w** can be modeled as a Gaussian distribution for each data point, d_i

$$P(d_i|\boldsymbol{w}, \lambda) = \frac{1}{\sqrt{2\pi {\varepsilon_{SAXS,i}}^2}} \exp\left(-\frac{(d_i - \lambda \sum w_i S_i)^2}{2\varepsilon_{SAXS,i}^2}\right)$$

$$P(\boldsymbol{D}|\boldsymbol{w}, \lambda) = \prod_{i=1}^{N} P(d_i|w, \lambda)$$

Bayesian ensemble modeling

Point estimate of w's:

$$w_j^B \equiv \langle w_j \rangle_{posterior} = \int w_j P(\boldsymbol{w}|D) d\boldsymbol{w}$$

Uncertainty is qualified with a posterior divergence $\sigma_{\boldsymbol{w}}$ for a vector **w**:

$$\sigma_{\boldsymbol{w}} = \left\langle \boldsymbol{\Omega^2}(\boldsymbol{w^B}, \boldsymbol{w}) \right\rangle_{\boldsymbol{posterior}}$$

$0 \leq \sigma_{\boldsymbol{w}} \leq \mathbf{1}$

Where $\boldsymbol{\Omega^2}(\boldsymbol{w^B}, \boldsymbol{w})$ is the Jensen-Shannon divergence between vector $\boldsymbol{w^B}$ and $\boldsymbol{w}$

$\sigma_{\boldsymbol{w}}$=0 if there is no uncertainty in population weights

Approximate inference with variational Bayes

$$P(\boldsymbol{w}|\boldsymbol{D},\boldsymbol{S}) = \frac{P(\boldsymbol{D}|\boldsymbol{w},\boldsymbol{S}) * P(\boldsymbol{w}|\boldsymbol{S})}{P(\boldsymbol{D}|\boldsymbol{S})}$$

$\mathrm{G}(w|\boldsymbol{\alpha},\boldsymbol{S})$ is selected to make analytical derivation possible

$$\mathrm{G}(w|\boldsymbol{\alpha},\boldsymbol{S}) = \frac{\Gamma(\alpha_0)}{\sum_{i=1}^{l}\Gamma(i)}\prod_{i=1}^{l} w_i^{\alpha_i-1}$$

This is the Dirichlet distribution where $\sum_{i=1}^{l} w_i = 1$

For the prior we select an non-informative prior (Jeffrey's) Dirichlet distribution

$$\mathrm{P}(w|\boldsymbol{S}) = \frac{\Gamma(n/2)}{\sum_{i=1}^{l} n\Gamma(1/2)}\prod_{i=1}^{l} w_i^{-1/2}$$

Or informative prior based on Boltzmann distribution

Overfitting

More complex models will always fit the data better but risk fitting to noise

Model fitting typically involves optimization of likelihoods (Maximum likelihood)

$$\chi^2 = \frac{\sum_{i=1}^{N}(I(q_i)_{data} - I(q_i)_{model})^2}{\sigma_i^2}$$

Optimization based solely on χ^2 can lead to overfitting

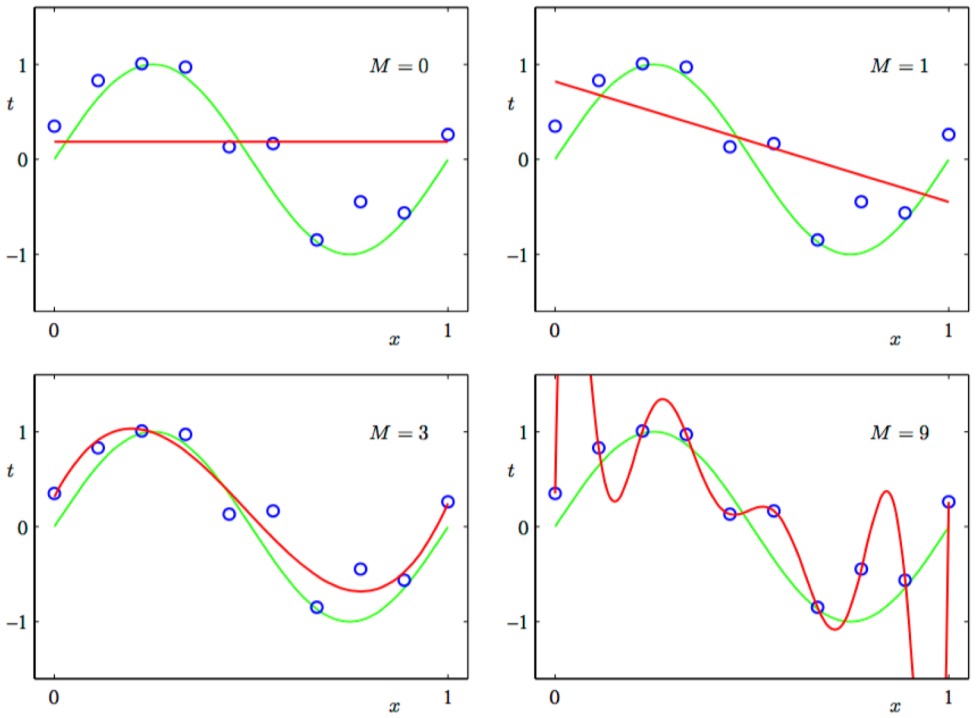
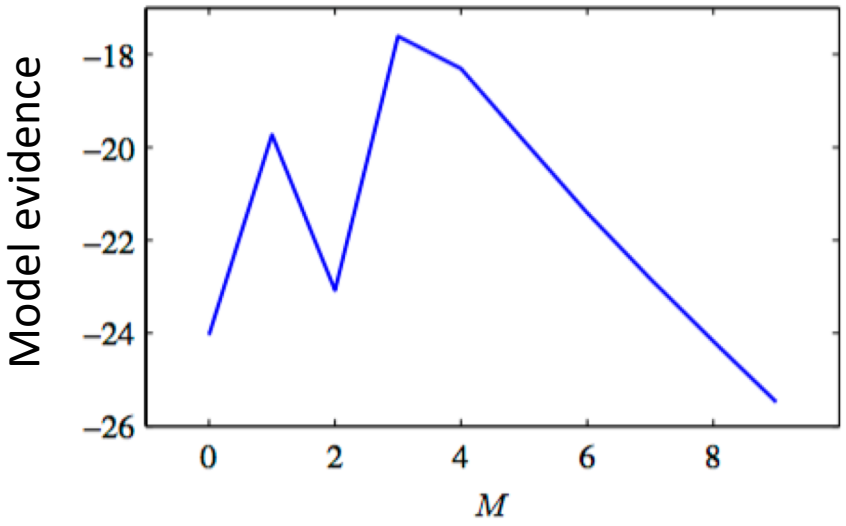


Figure 1.4 Plots of polynomials having various orders M, shown as red curves, fitted to the data set shown in Figure 1.2.

From "Pattern recognition and machine learning" Bishop (2006)

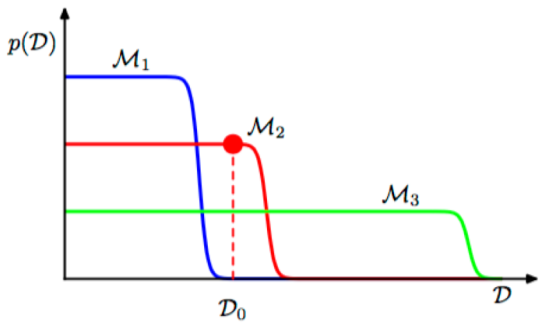
Overfitting

Bayes to the rescue!



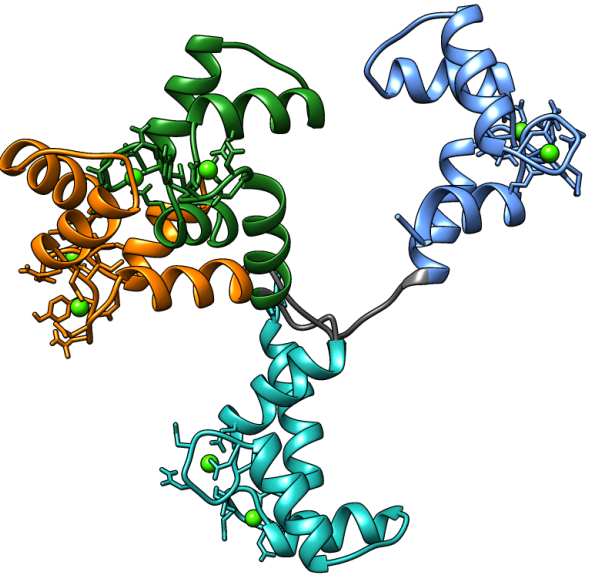
Embodies the Occam's razor principle: simpler models are better

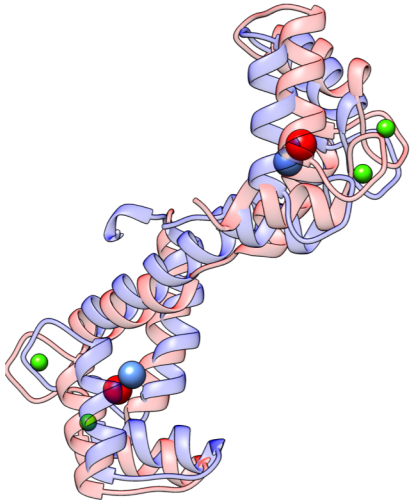
$$P(D|M_i) = \int P(D|\theta, M_i) * \mathrm{P}(\theta|M_i) d\theta$$



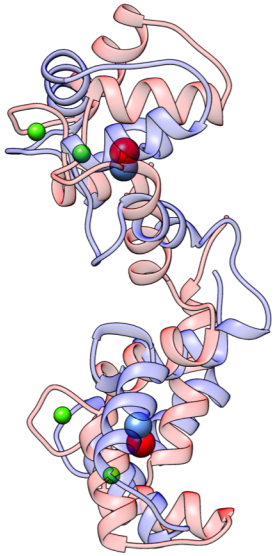
Figures from "Pattern recognition and machine learning" Bishop (2006)

Calmodulin ensemble compared with PDB structures

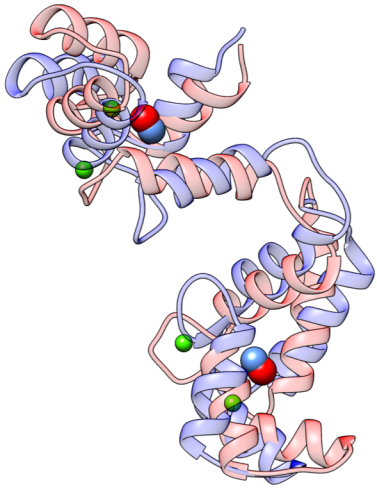




RMSD=4.6 A



RMSD = 4.8 A



RMSD = 7.9 A