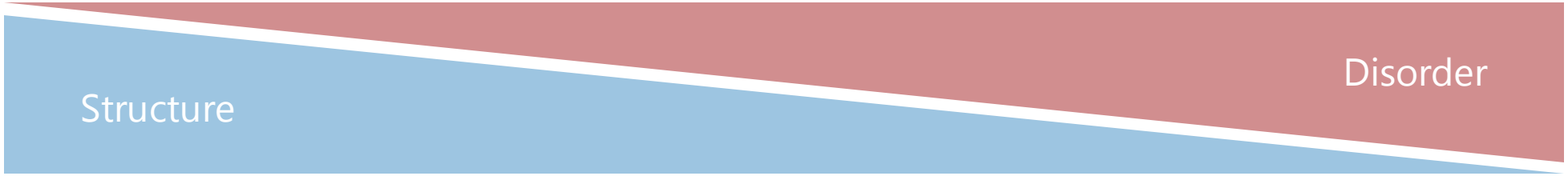
Data-driven modeling of highly dynamic proteins and reweighting of conformational ensembles against SAXS data

Giulio Tesei

UNIVERSITY OF COPENHAGEN

Intrinsically disordered proteome

Around **1/3** of protein sequences encoded by the **human genome** are **disordered** and adopt **ensembles** of structures

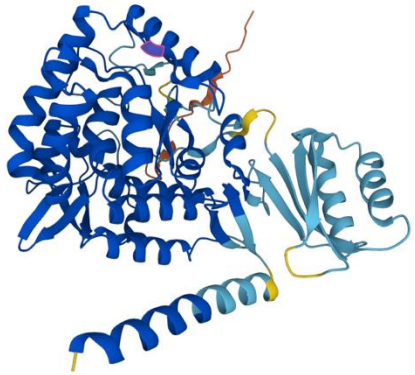


Many proteins are fully folded and globular

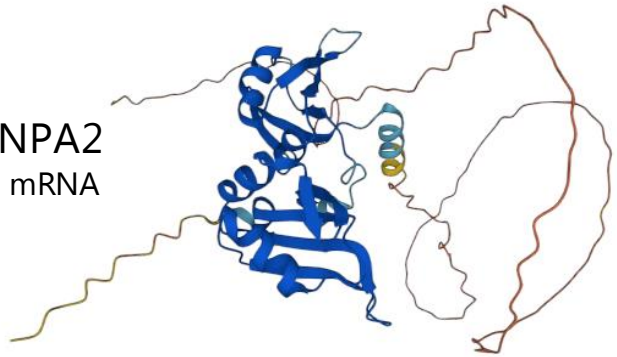
Folded domain are often connected to disordered regions

5 % of human proteins lack any folded domains

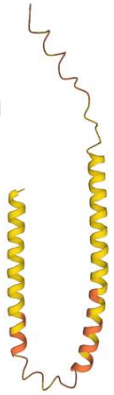
Phenylalanine hydroxylase
enzyme



hnRNPA2
binds mRNA

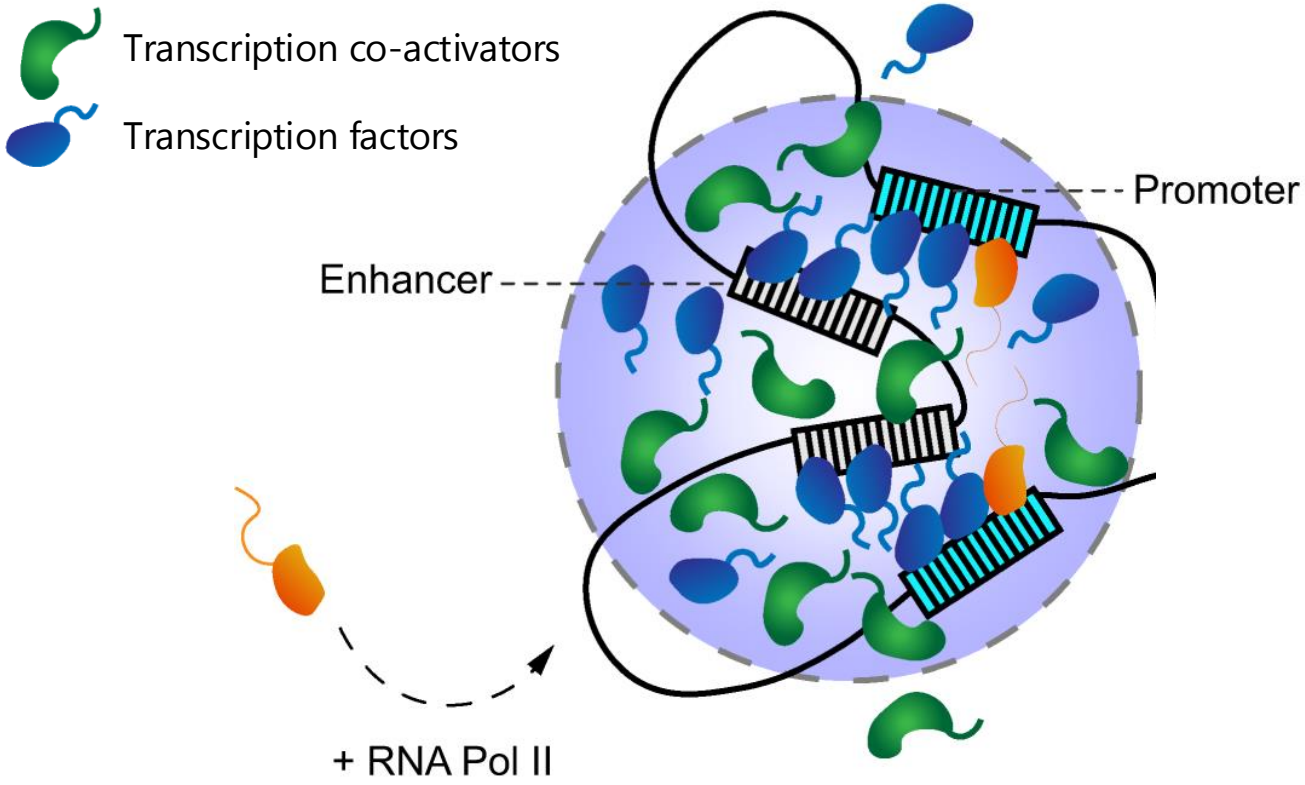


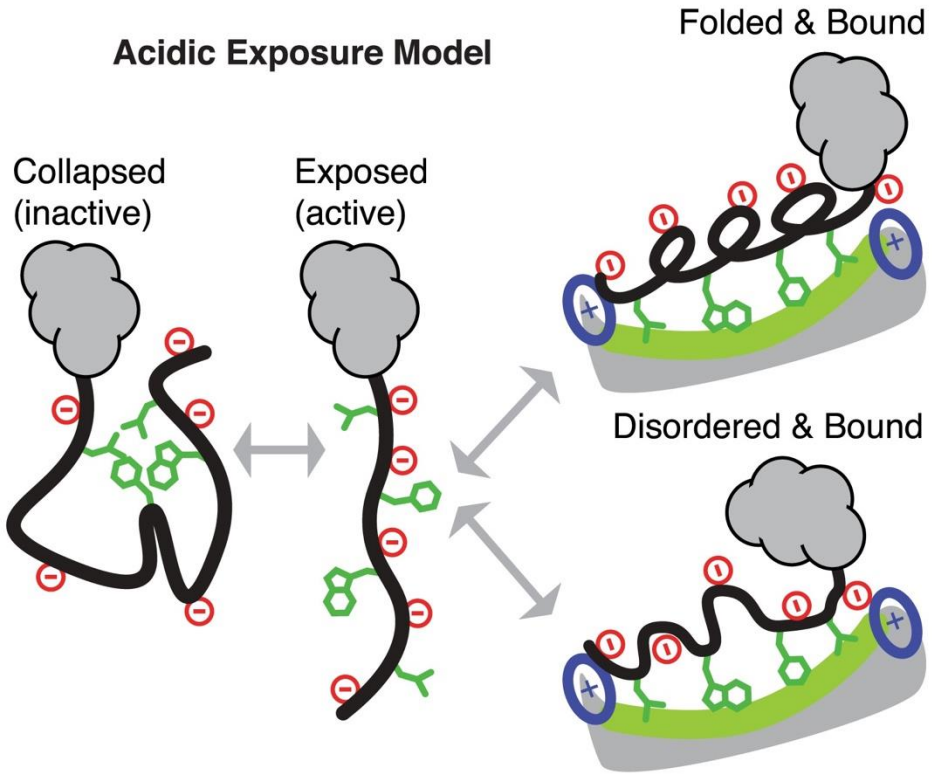
α-Synuclein (neuronal protein)



van der Lee et al. *Chem. Rev.* 2014;114(13):6589–6631

Complex sequence-ensemble-function relationship





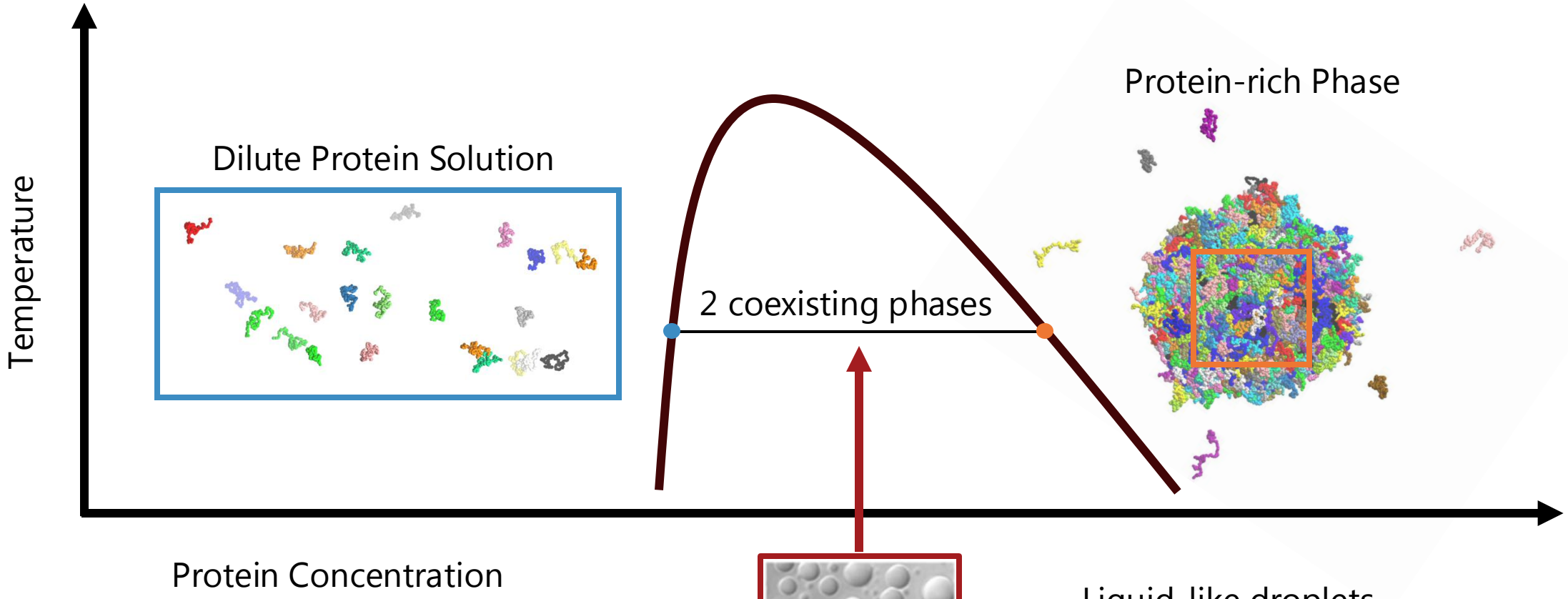
IDRs of transcription factors have clusters of aromatic and leucine residues embedded in a region rich in negatively charged residues

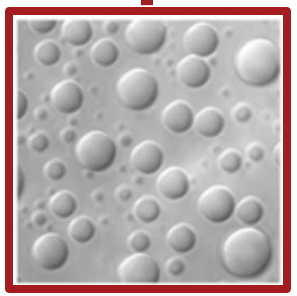
Patterning and balance between acidic and aromatic residues is crucial for binding to coactivators

Laflamme and Mekhail *Commun Biol.* 2020; 3(1):773 (CC BY 4.0)

Kotha and Staller *Genetics*. 2023; 225(2):iyad131 (CC BY 4.0)

Many IDRs drive formation of condensates



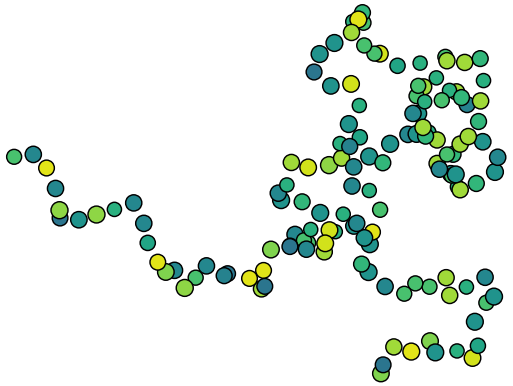


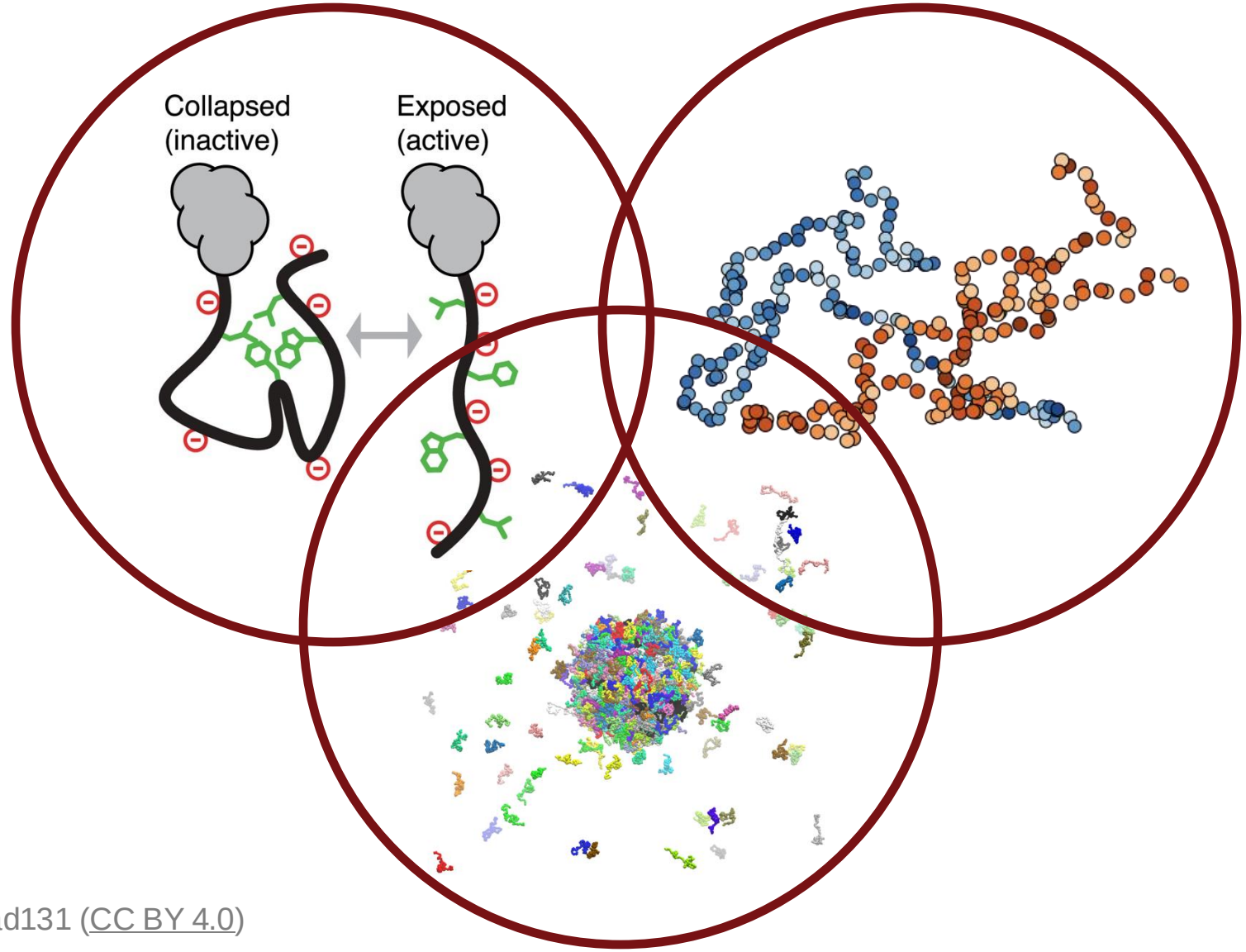
Liquid-like droplets formed through a reversible phase separation process

https://animationlab.utah.edu/phase-separation

Modelling disordered proteins and condensates

MDVFMKGLSKAKEGVVAAAEKTKQGVAE
AAGKTKEGVLYVGSKTKEGVVHGVATVA
EKTKEQVTNVGGAVVTGVTAVAQKTVEG
AGSIAAATGFVKKDQLGKNEEGAPQEGI
LEDMPVDPDNEAYEMPSEEGYQDYEPEA

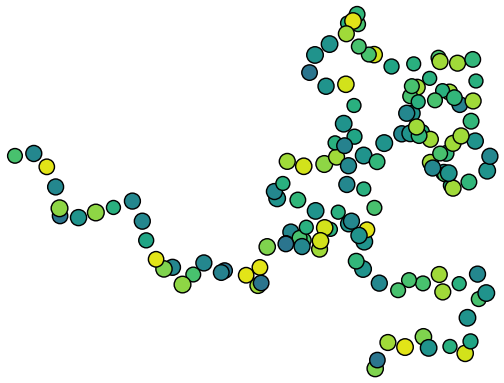


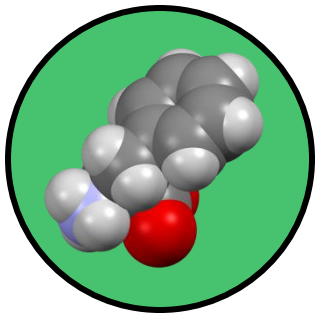


Kotha and Staller *Genetics*. 2023; 225(2):iyad131 (CC BY 4.0)

The CALVADOS model

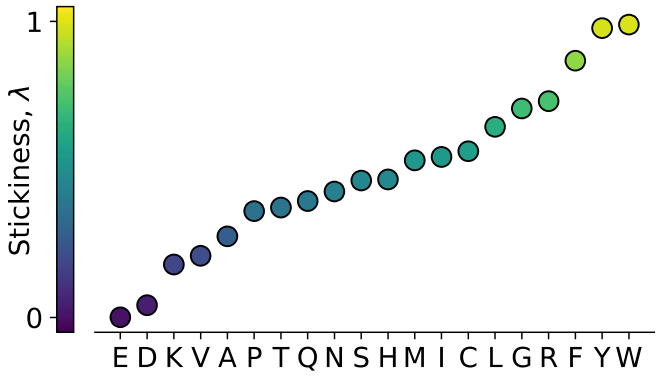
- Residues represented as single beads
- Solvent as dielectric continuum with given salt concentration
- Salt-screened electrostatic interactions between charged residues

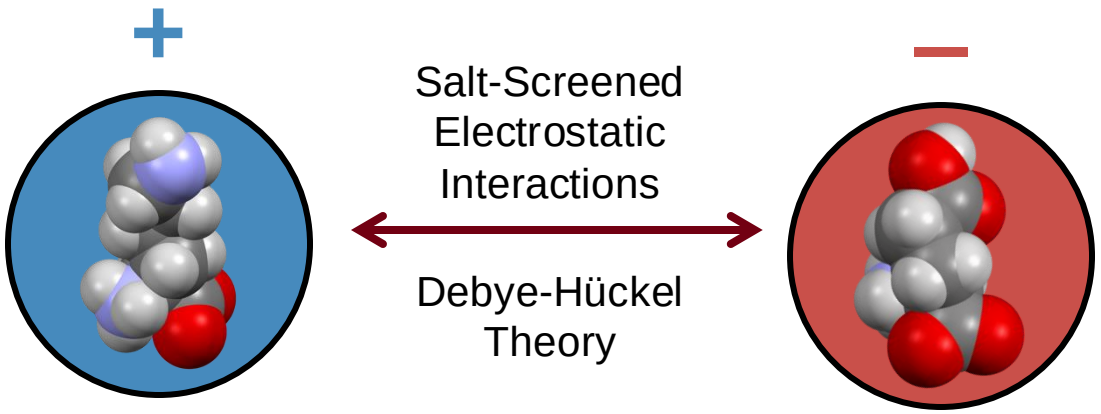




- diameter, σ
- charge
- stickiness, λ: hydrophobicity, H-bonding, cation-π, ...

- Stickiness parameters quantify strength of nonionic interactions



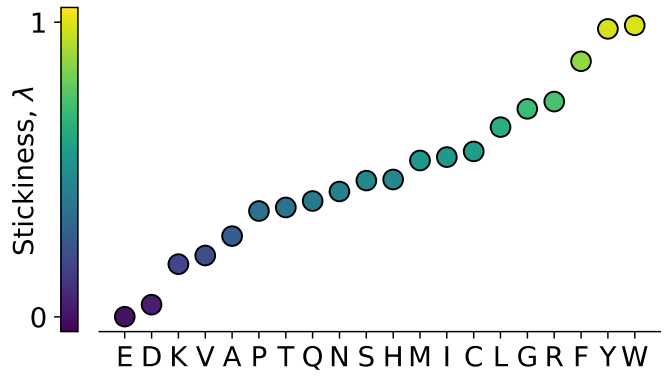


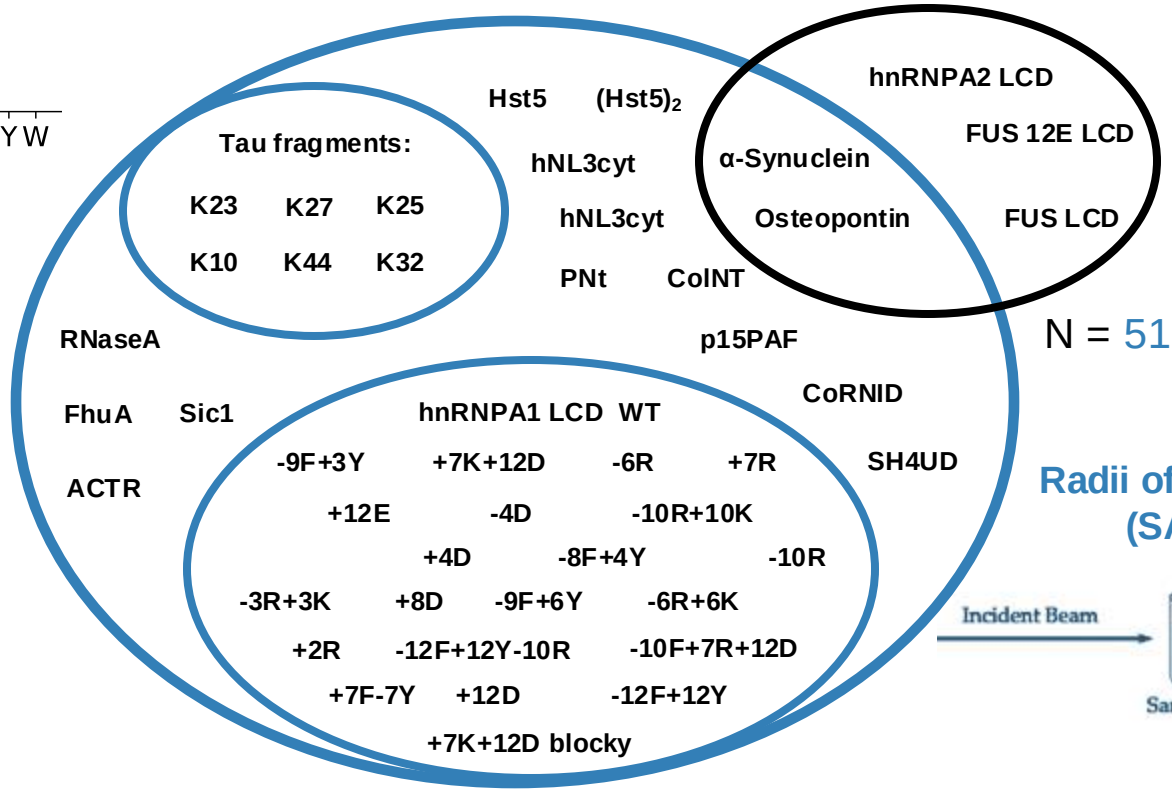
jupyter CO github.com/KULL-Centre/CALVADOS GitHub

Image courtesy of Benjah-bmm27. Public domain via Wikimedia Commons
Tesei et al. *Proc Natl Acad Sci U S A.* 2021;118(44):e2111696118

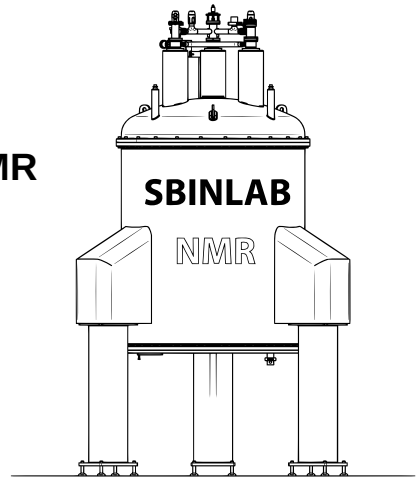
Tesei and Lindorff-Larsen *Open Res Eur.* 2023

Optimization based on experimental data



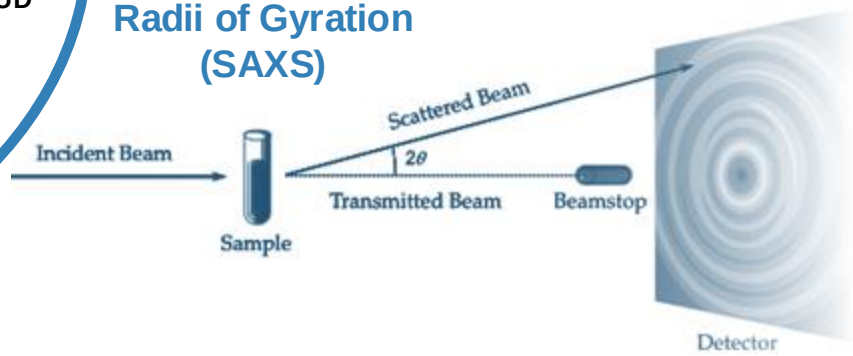


PRE NMR Data



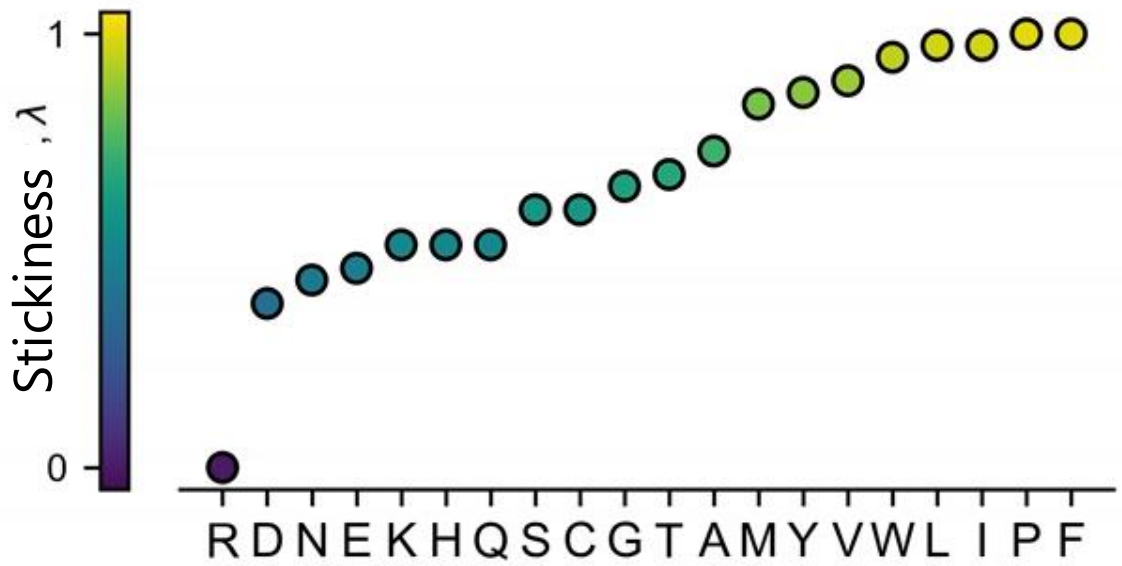
N = 51 + 5

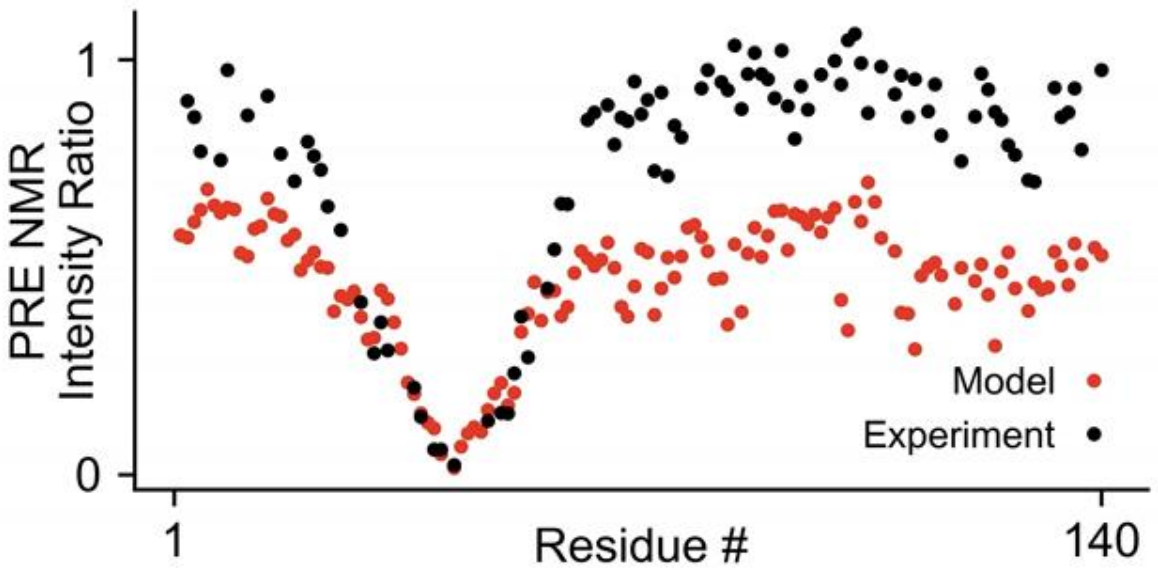
Radii of Gyration (SAXS)



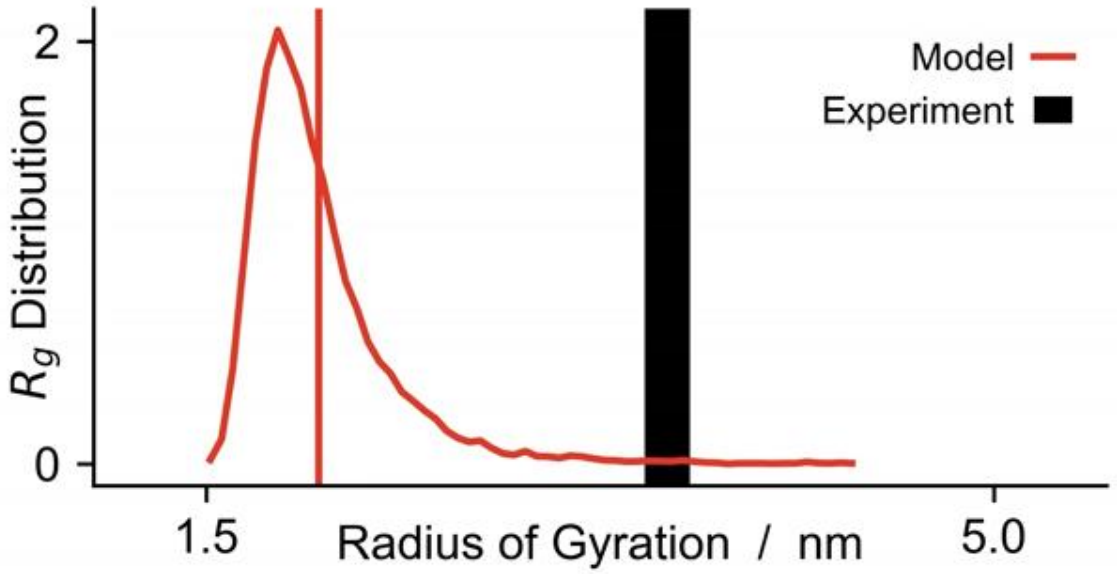
We thank all who measured and made available the experimental data used in this study: Skepö, Kjærgaard, Mittag, Petridis, Bernado, Lakey, Silman, Sosnick, Kragelund, Sattler, Svergun, Kriwacki, Pappu, Sukenik, Konrat, Fawzi, and more

Regy et al. *Protein Sci.* 2021;30(7):1371-1379
Dannenhoffer-Lafage and Best *J Phys Chem B*. 2021;125(16):4046-4056 Latham and Zhang *J Chem Theory Comput.* 2020;16(1):773-781



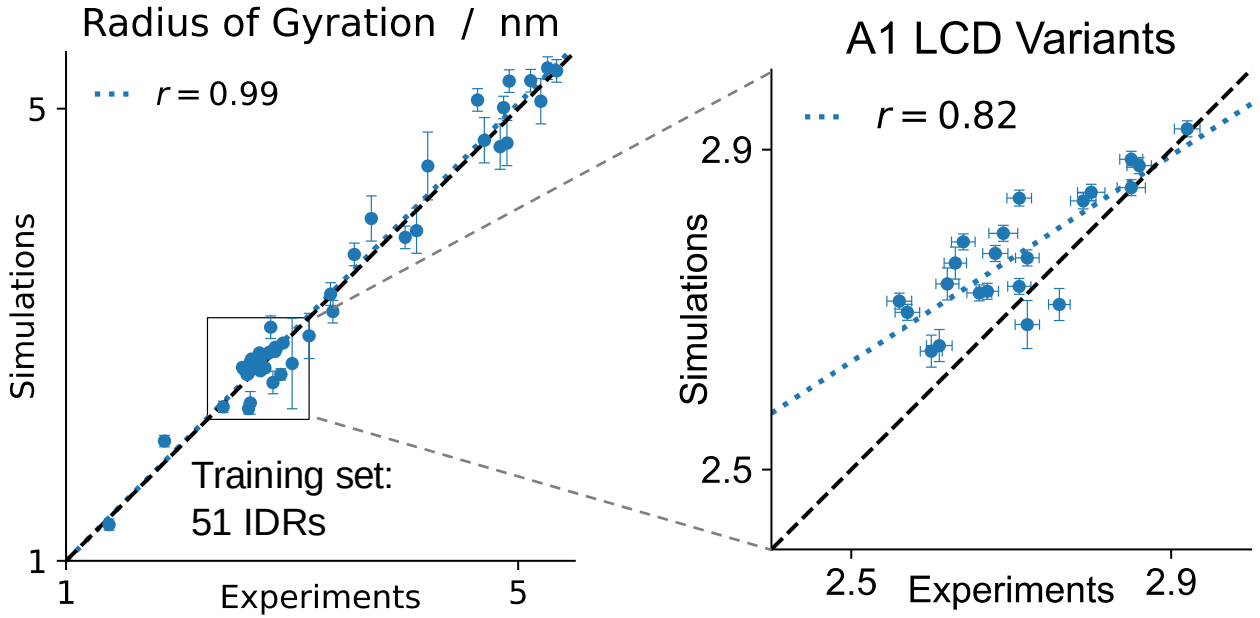


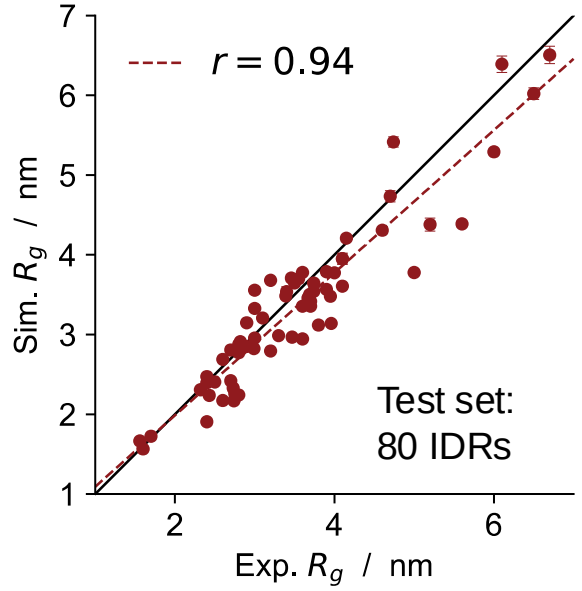
1st Reweighting Cycle

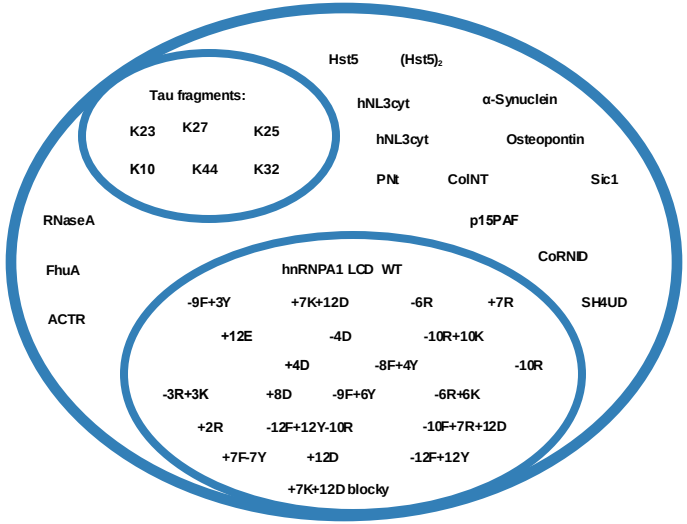


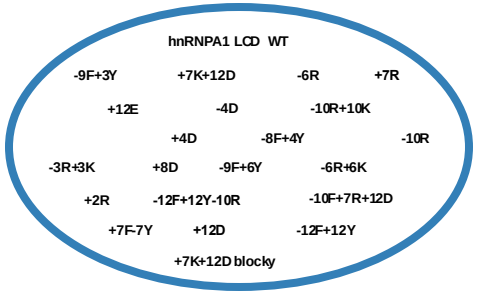
YouTube video: https://youtu.be/r-eFzoBiQZ4

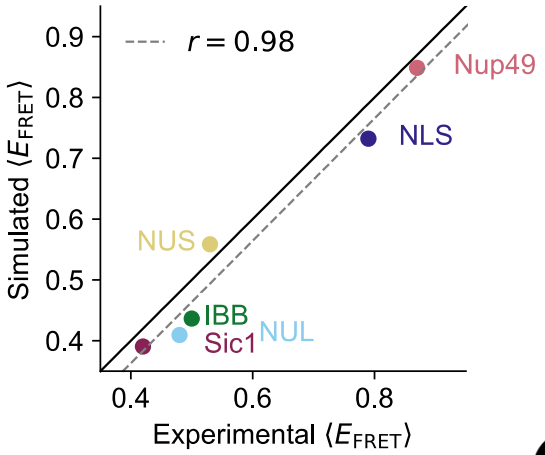
Testing chain compaction







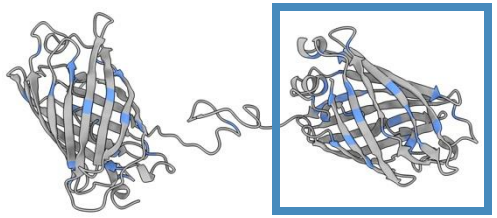




Bremer, Farag, Borcherds et al.
Nat Chem. 2022;14(2):196-207

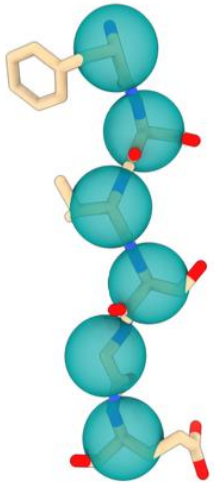
github.com/KULL-Centre/FRETpredict GitHub

CALVADOS 3: multi-domain proteins

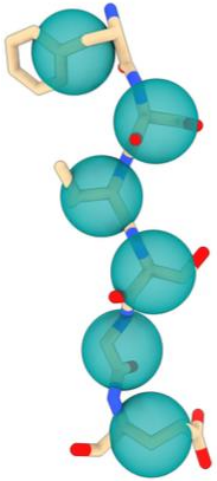


Rigid structure from PDB or AF2

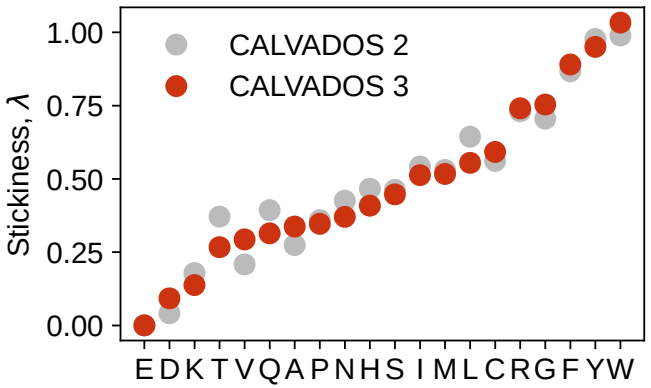
Cα mapping for IDRs

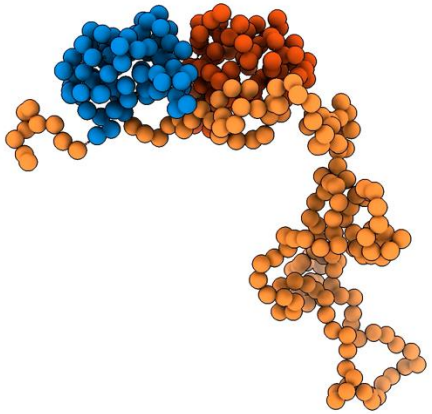


COM mapping for folded domains



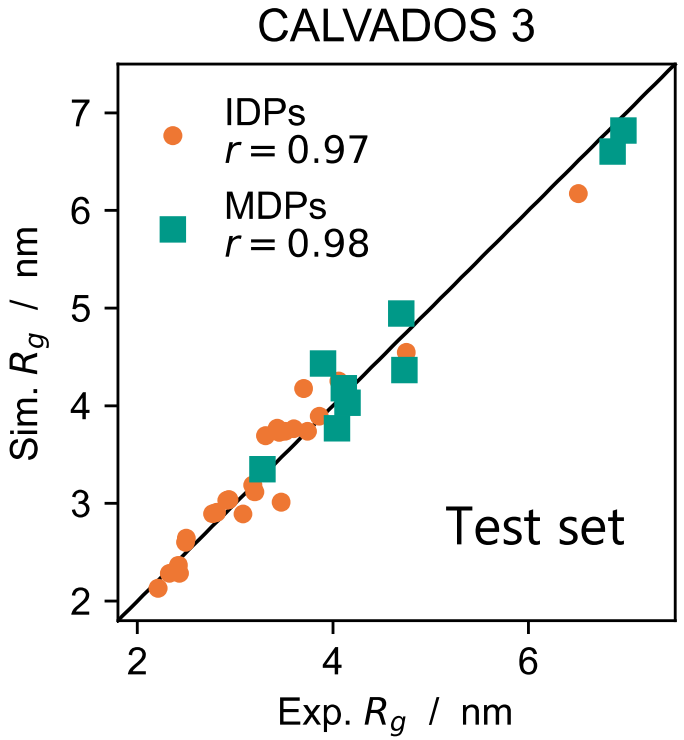
Optimization of 56 IDPs & 14 MDPs





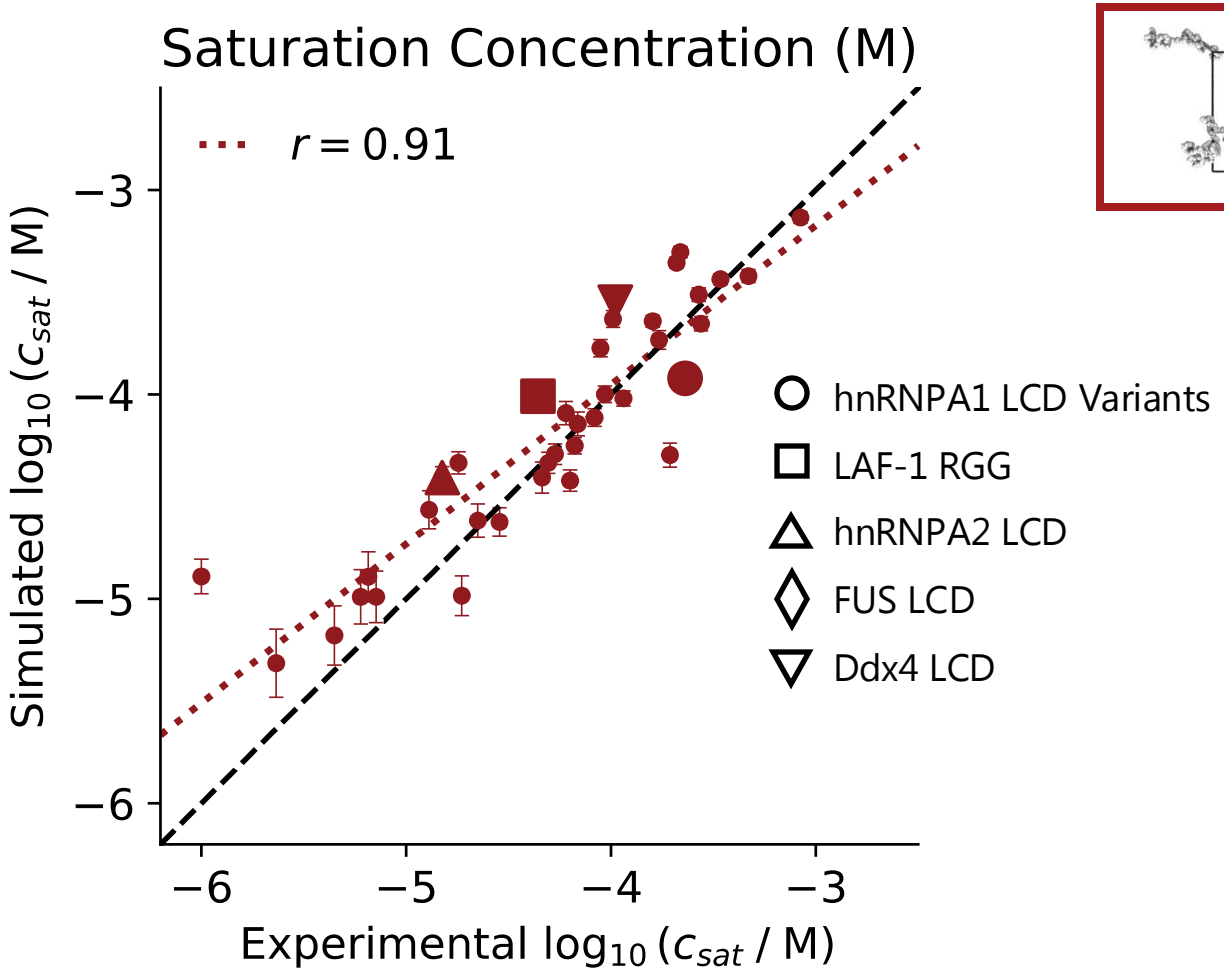
Fan Cao

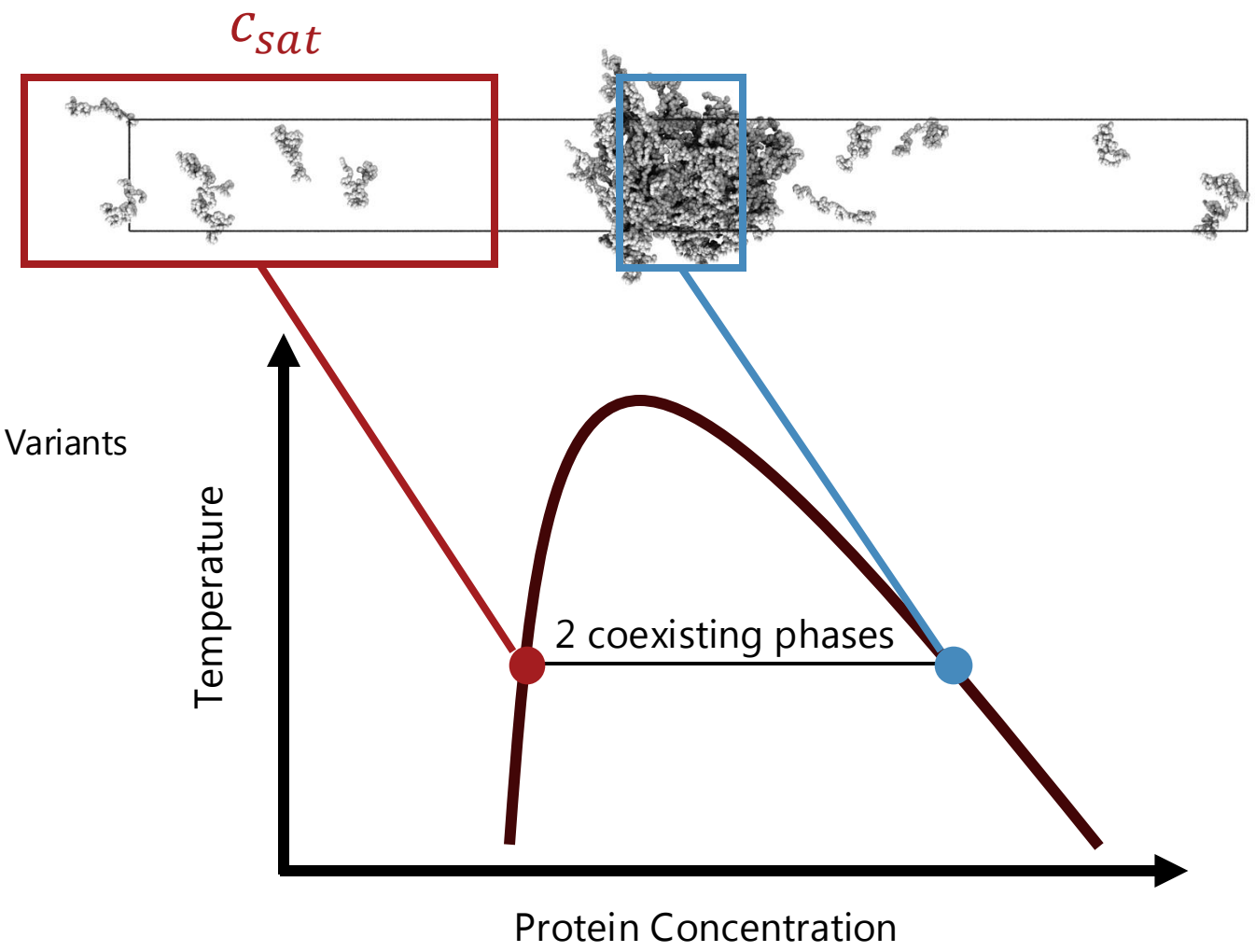
Sören von Bülow



Cao et al. A coarse-grained model for disordered and multi-domain proteins *Protein Sci.* 2024;33(11):e5172

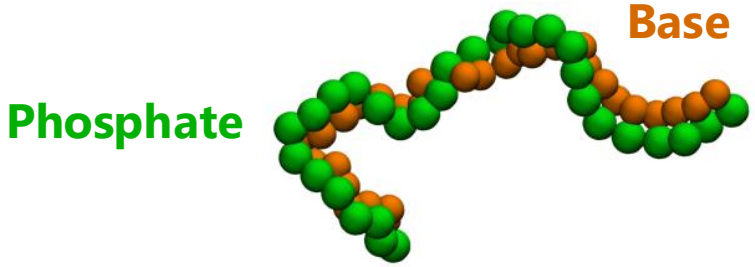
Simulation of biomolecular condensates: Protein components

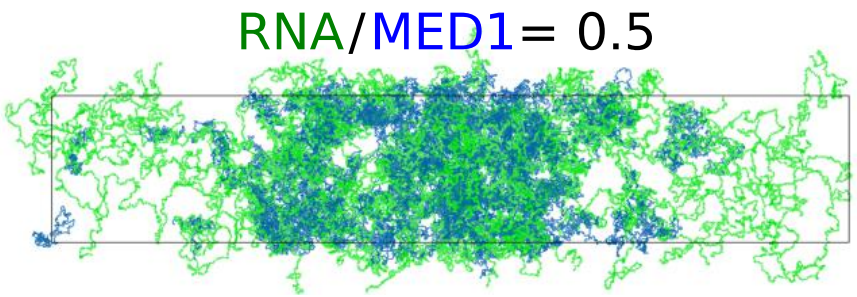




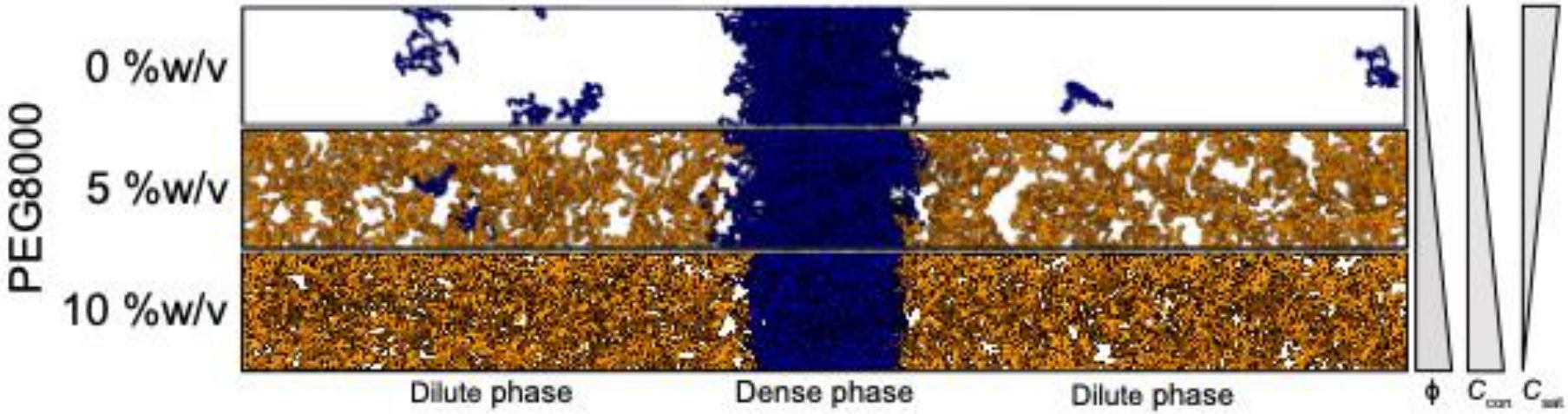
Tesei et al. *Proc Natl Acad Sci U S A.* 2021;118(44):e2111696118
Tesei and Lindorff-Larsen *Open Res Europe* 2022; 2:94

Simulation of biomolecular condensates: RNA and PEG





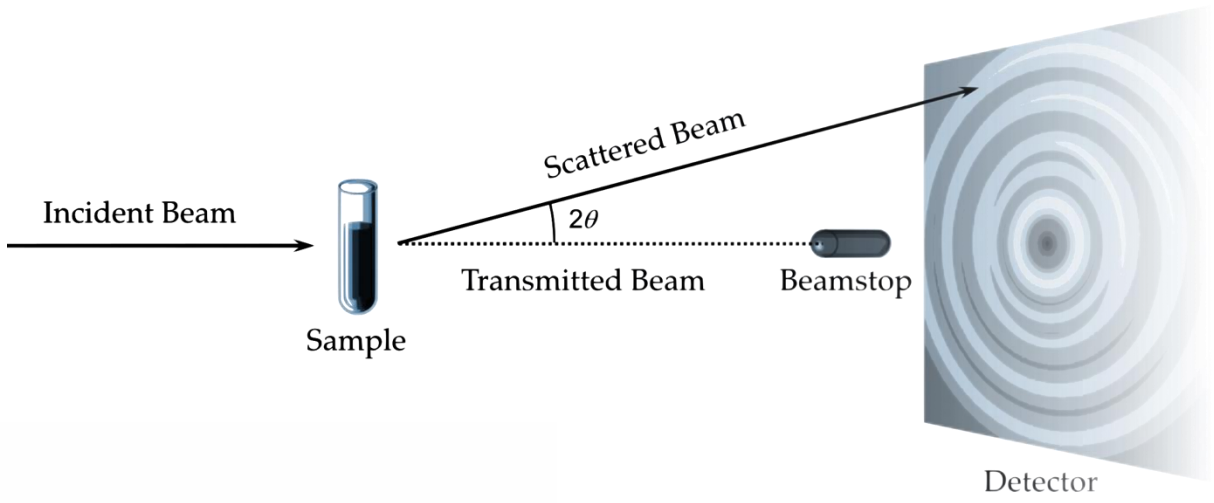
Ikki Yasuda

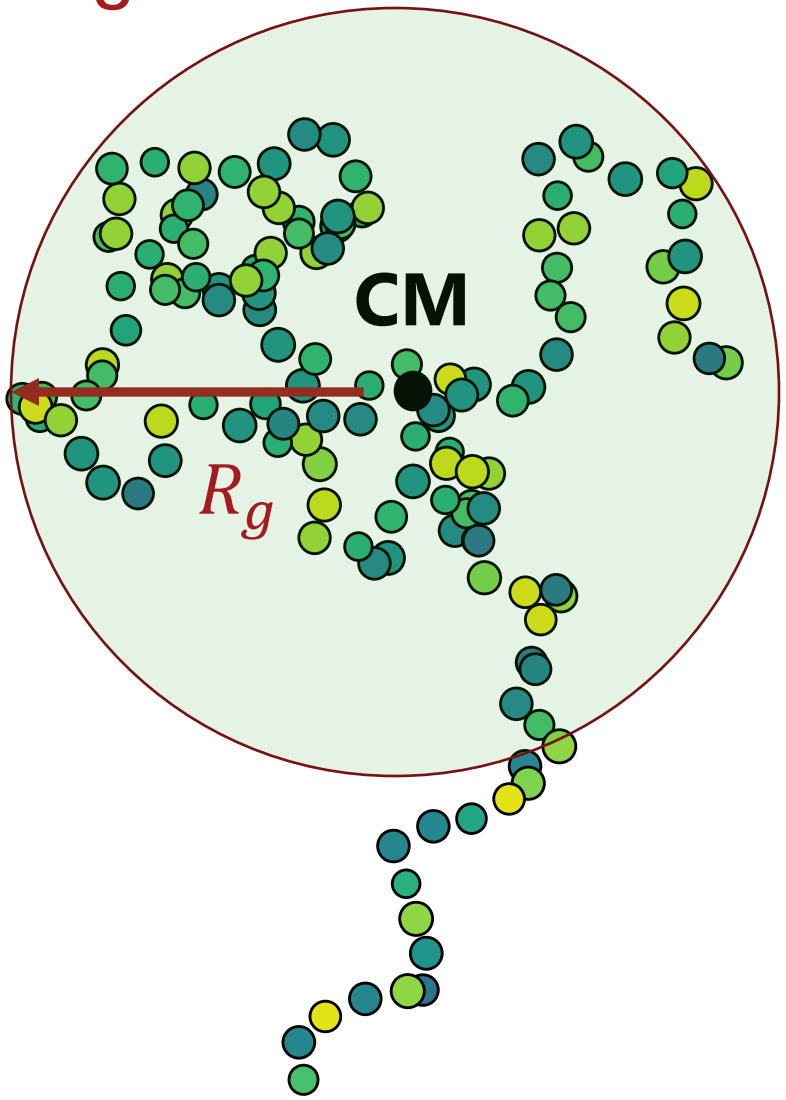


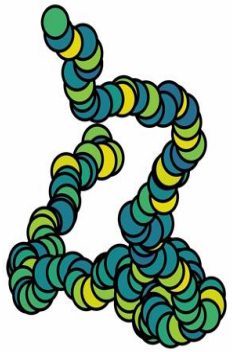
Arriën Rauh

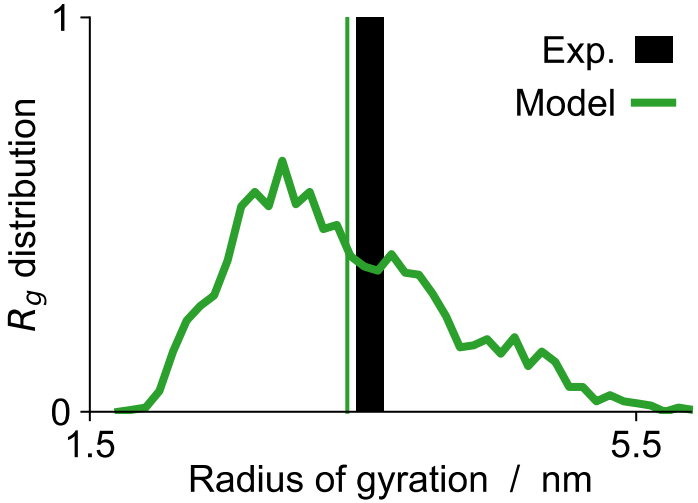
Yasuda et al. Coarse-Grained Model of Disordered RNA for Simulations of Biomolecular Condensates *J Chem Theory Comput. 2025*
Rauh et al. A coarse-grained model for disordered proteins under crowded conditions *bioRxiv.* 2025 DOI: 10.1101/2025.03.01.640997

Experiments often report on ensemble averages

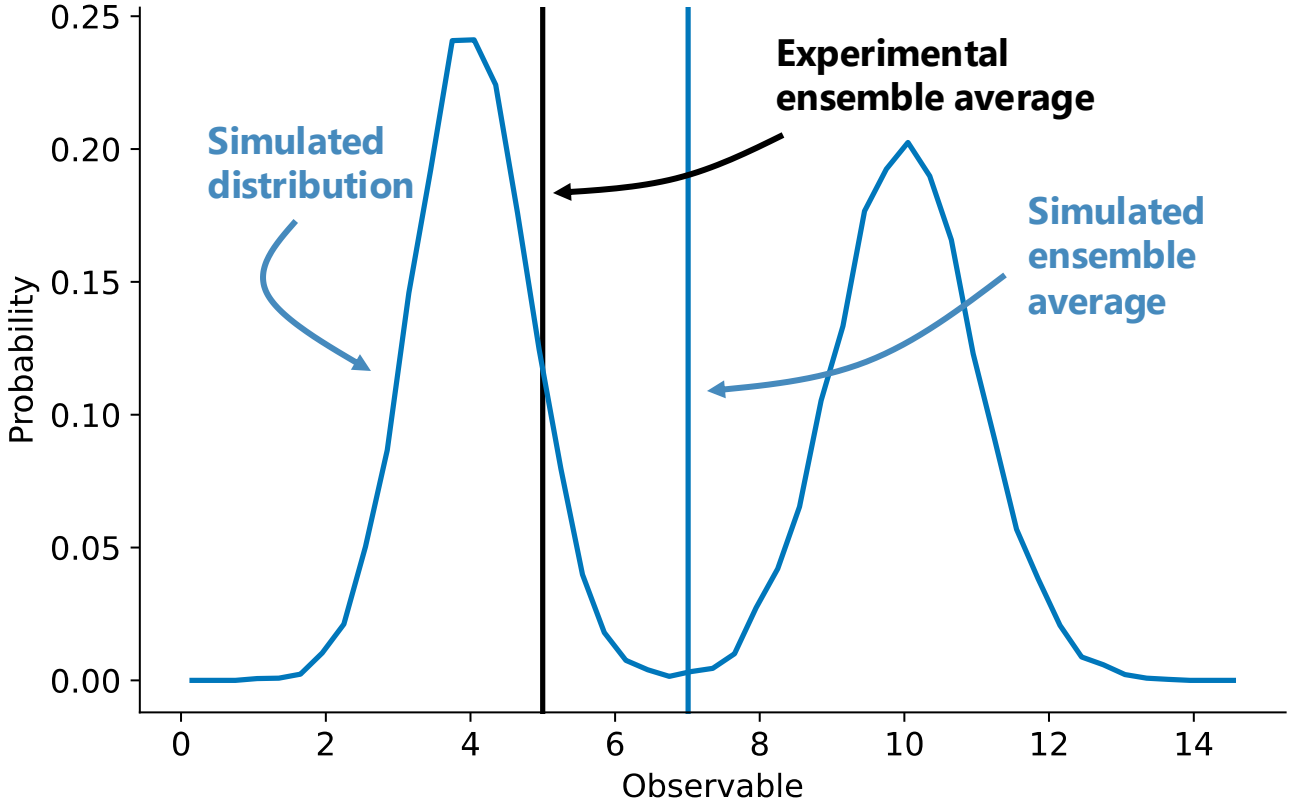


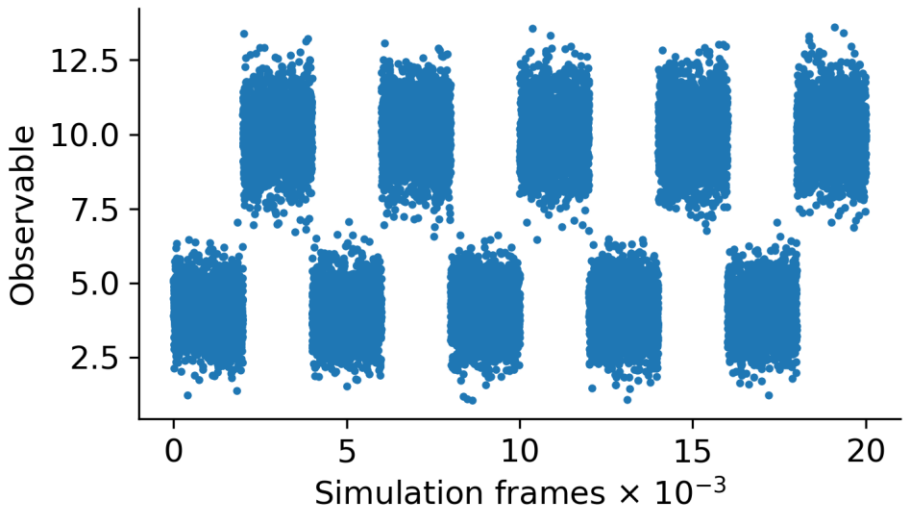


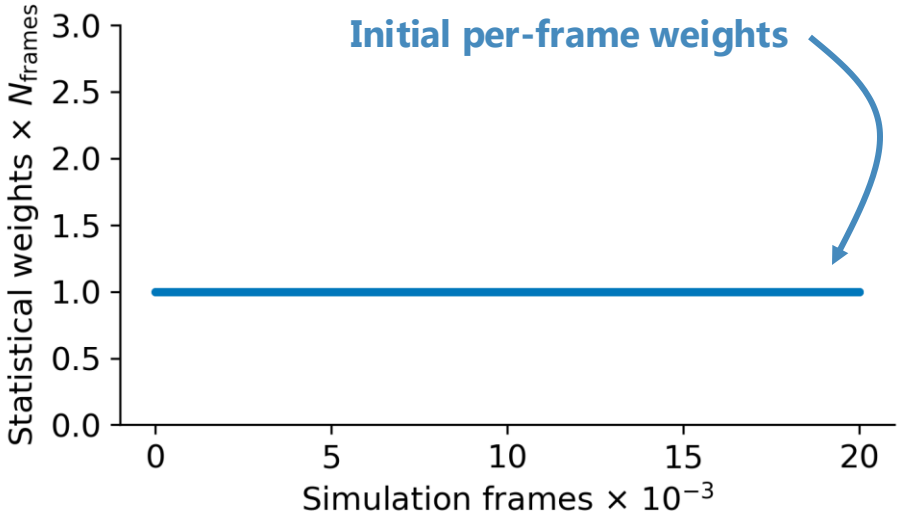




The underlying distribution is often unknown



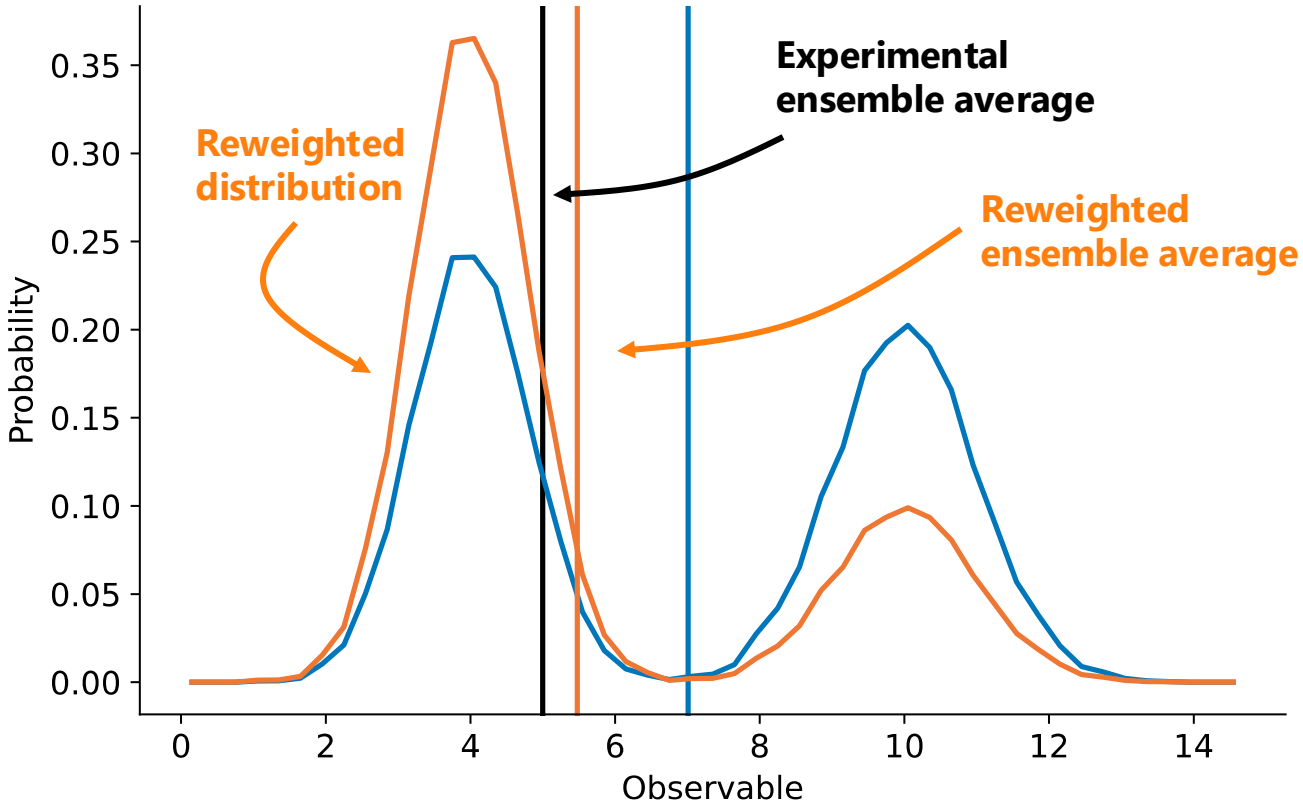


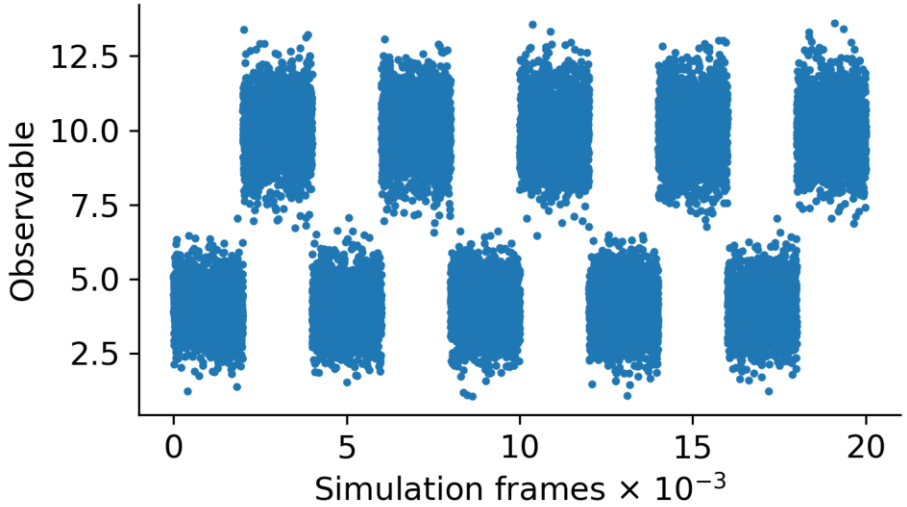


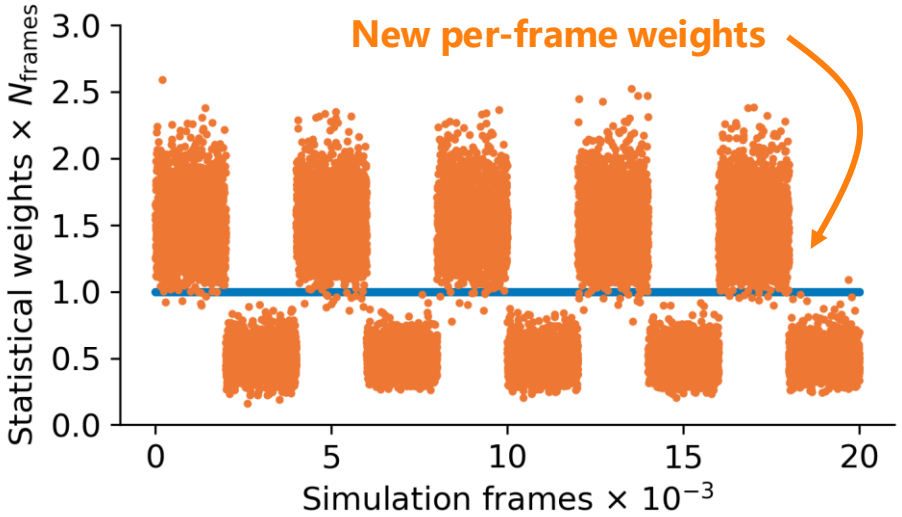
Boomsma et al *PLOS Comp Biol.* 2014
Bottaro, Bussi, Kennedy, Turner, Lindorff-Larsen *Sci Adv.* 2018
Bottaro, Bengtsen and Lindorff-Larsen *Struct Bioinf.* 2020
Orioli, Larsen, Botaro and Lindorff-Larsen *Prog Mol Bio Trans Sci.* 2020

Also related work by Hummer, Bussi, Vendruscolo, Chodera and many others

The underlying distribution is often unknown







Boomsma et al *PLOS Comp Biol.* 2014
Bottaro, Bussi, Kennedy, Turner, Lindorff-Larsen *Sci Adv.* 2018
Bottaro, Bengtsen and Lindorff-Larsen *Struct Bioinf.* 2020
Orioli, Larsen, Botaro and Lindorff-Larsen *Prog Mol Bio Trans Sci.* 2020

Also related work by Hummer, Bussi, Vendruscolo, Chodera and many others

BME: Bayesian/maximum entropy reweighting

$$G = H - TS$$

$$\Gamma(w, \theta) = \chi^2(w) - \theta S_{rel}(w)$$

$\chi^2(w)$ Quantifies deviation between experiment and simulation

$S_{rel}(w)$ Quantifies deviation from force field ("prior")

θ A temperature-like parameter that determines the balance between the "enthalpy" and "entropy", and determines how precisely we fit the data

Bottaro, Bengtsen and Lindorff-Larsen *Methods Mol. Biol.* 2020 2112:219–240

https://github.com/KULL-Centre/BME GitHub

Balancing prior distribution and experimental data

- Loss function to minimize

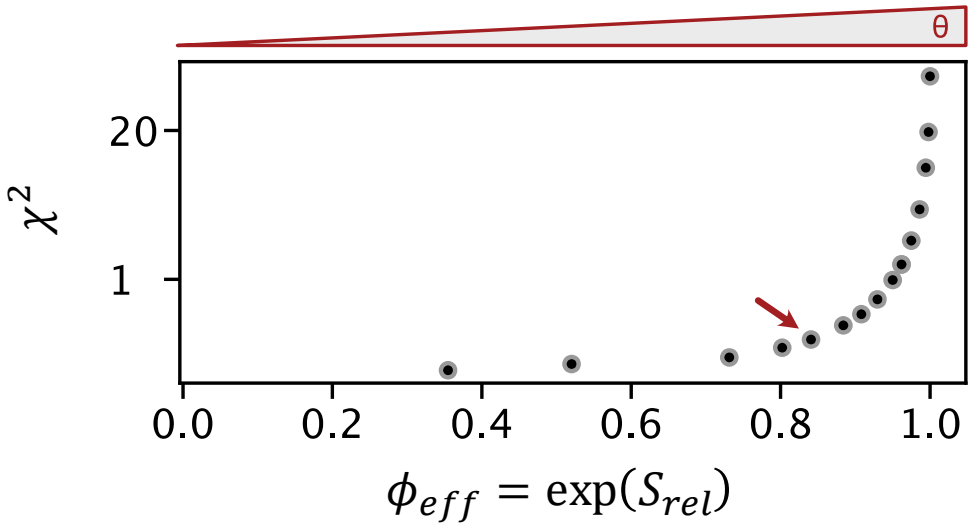
$$\mathscr{L}(w_1 \dots w_n) = \frac{m}{2}\chi^2(w_1 \dots w_n) - \theta\, S_{rel}(w_1 \dots w_n)$$

- Deviation from experimental SAXS curve

$$\chi^2(w_1 \dots w_n) = \frac{1}{m}\sum_i^m \frac{\left(\sum_j^n w_j I_j^{CALC}(q_i) - I^{EXP}(q_i)\right)^2}{\sigma(q_i)^2}$$

- Deviation from initial ensemble from simulation

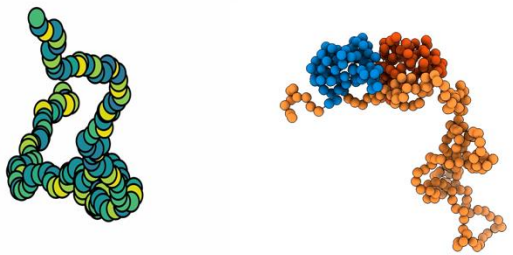
$$-S_{rel}(w_1 \dots w_n) = \sum_j^n w_j \ln\left(\frac{w_j}{w_j^0}\right) = \sum_j^n w_j \ln\left(w_j \times n\right)$$



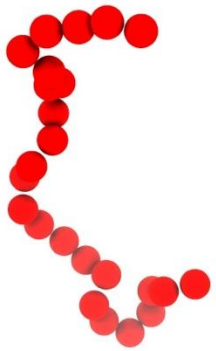
Note: We do not expect to be able to recover the true landscape with finite and noisy data. Just generally a better one than the one from the force field.

Workflow of the hands-on session

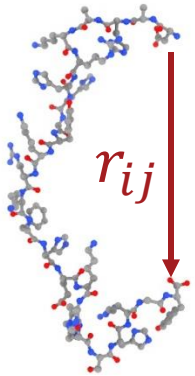
CALVADOS simulation



Convert coarse-grained to all-atom

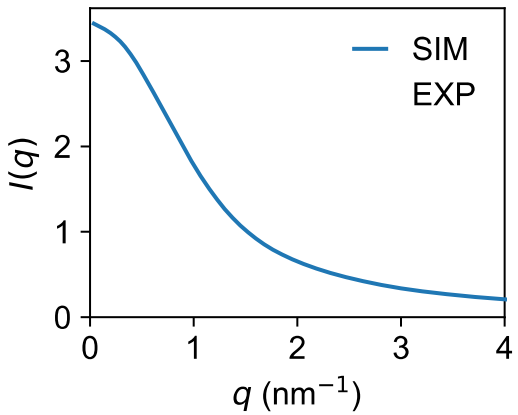


cg2all

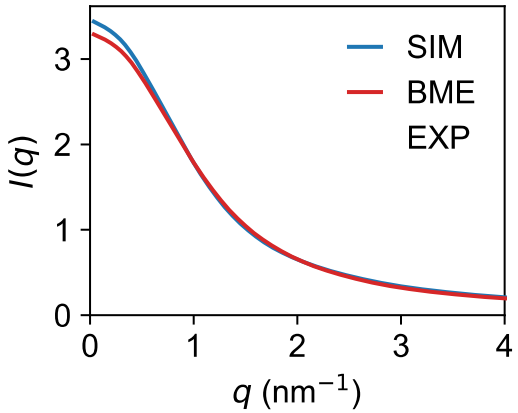


$$I(q) = \sum_i^N \sum_j^N f_i(q) f_j(q) \frac{\sin(q r_{ij})}{q r_{ij}}$$

Calculate SAXS curve using Pepsi-SAXS



BME reweighting



Heo and Feig *Structure* 2024;32(1):97–111.e6
Pesce and Lindorff-Larsen *Biophys J.* 2021;120(22):5124–5135

Lab exercise – CALVADOS simulations + BME reweighting

https://github.com/KULL-Centre/ColabCALVADOS

Open in Colab

Francesco Pesce

Fan Cao

name: Ubq4

sequence: MQIFVKTLTGKTITLEVEPSDTIENVKAKIQDKEGIPPDQQRLIFAGKQLEDGRTLSDYNIQKESTLHLVLRLRGGMQIFVKTLTGKTITLEVEPSDTIENVKAKIQDKEGIPPDQQRLIFAGKQLEDGRTLSDYNIQKESTLHLVLRLRGGMQIFVKTLTGKTITLEVEPSD

temperature: 293

ionic_strength: 0.33

pH: 8.0

**Units: temperature [K], ionic strength [M]*

Define domain boundary -- ignore this if `isIDP` is toggled on:

Residue ranges are delimited by `-`, e.g. residues 1 to 200 are `1-200`;

Separate domains are delimited by `,`, e.g. `1-200,201-400`;

Discontinuous segments are delimited by `_`, e.g. `1-200,201-300_350-400`.

domain_boundary: 1-72,77-148,153-224,229-300

Lab exercise – CALVADOS simulations + BME reweighting

- NLS.fasta
- ProTa.dat
- ProTa.fasta
- RS.dat
- RS.fasta
- Sic1.dat
- Sic1.fasta
- THB_C2.dat
- THB_C2.fast
- THB_C2.pdb
- TIA1.dat
- TIA1.fasta
- TIA1.pdb
- Tau.dat
- Tau.fasta
- Ubq2.dat
- Ubq2.fasta
- Ubq2.pdb
- Ubq3.dat

Go to file t

4ca1390 · 4 months ago History

Download raw file

Raw

Choose files Ubq4.pdb

- **Ubq4.pdb**(n/a) - 388363 bytes, last modified: 30/11/2024 - 100% done

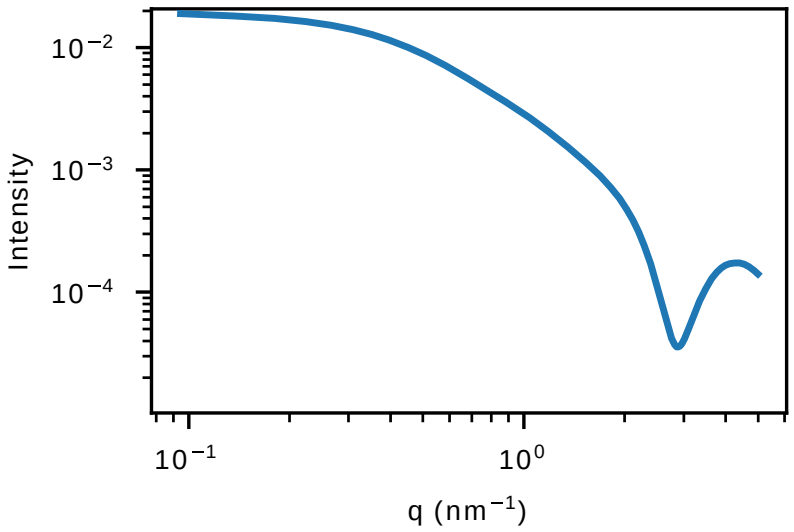
```
Saving Ubq4.pdb to Ubq4.pdb
Ubq4/Ubq4.pdb successfully uploaded.
```

Lab exercise – CALVADOS simulations + BME reweighting

6. Run MD simulation

3m Show code

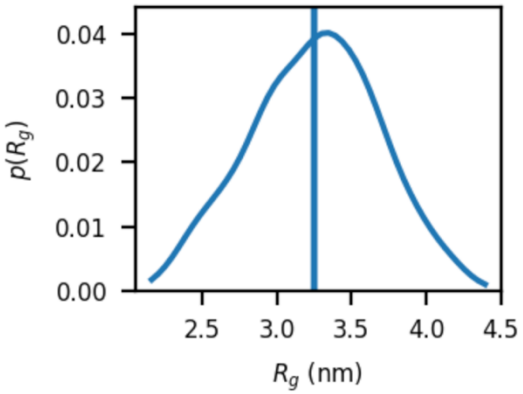
```
pH: 8.0
Ionic strength: 0.33 M
Temperature: 293.0 K
Starting from pdb structure Ubq4/Ubq4.pdb
rc 2.0 nm
Using GPU
Total frames: 1010; the first 10 frames will be discarded
100.00% [1010/1010 03:02<00:00]
Ubq4 total simulation time: 0.0h 3.0min 2.78s
```

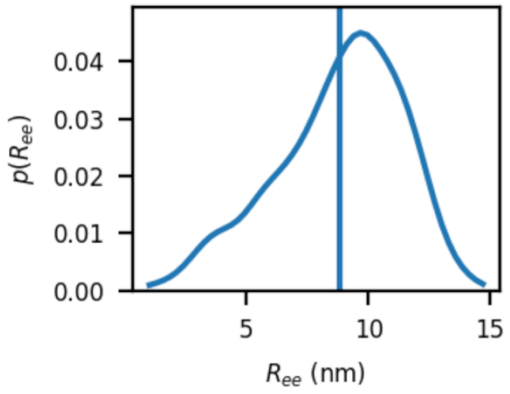


8. Visualize trajectory

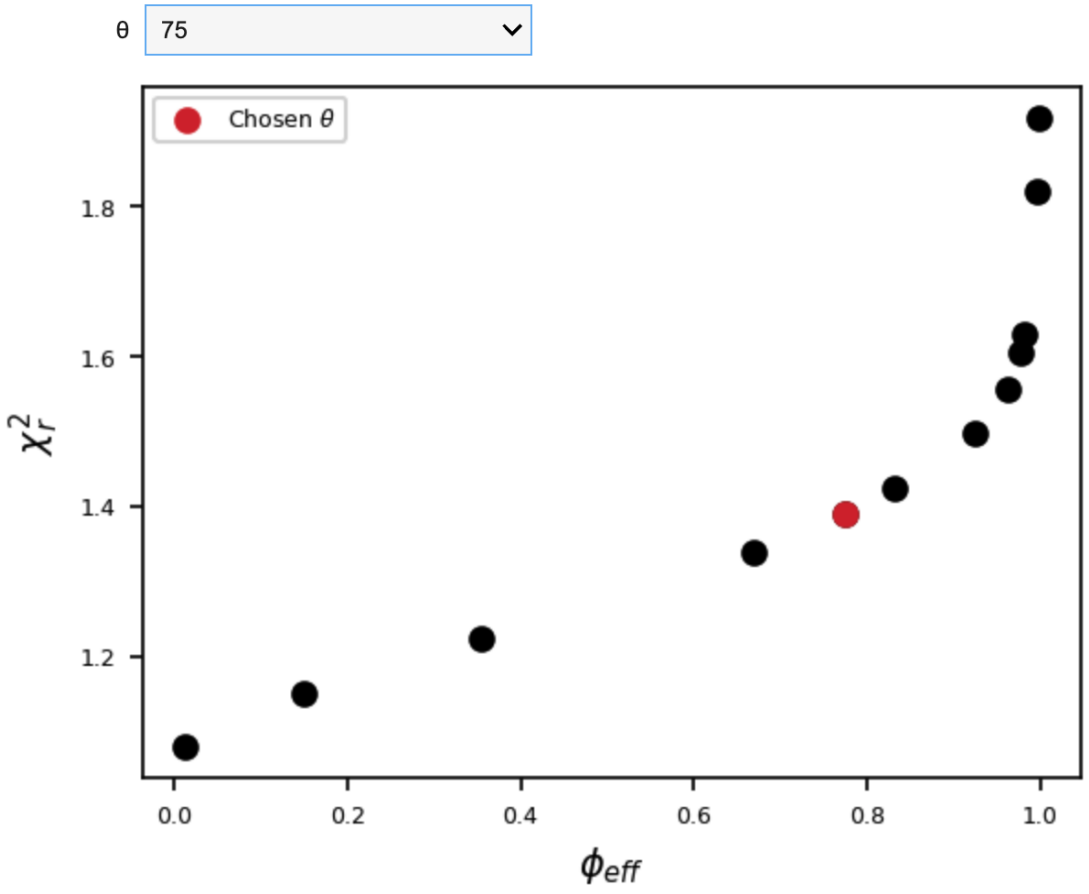
28s Show code

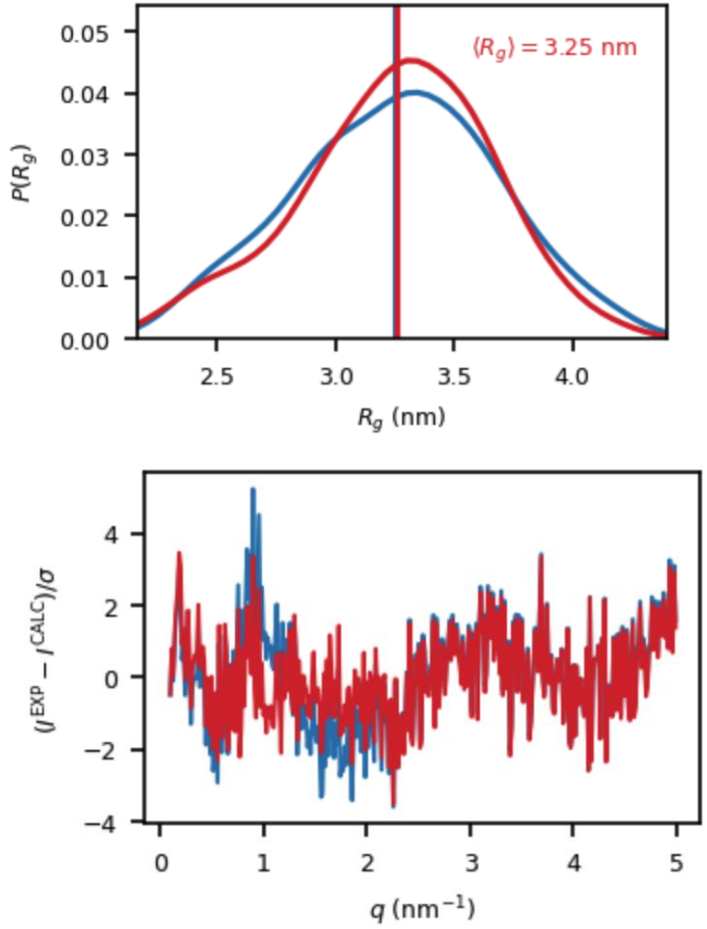
9s Show code





Lab exercise – CALVADOS simulations + BME reweighting





› 11. Download results

Acknowledgements

Kresten Lindorff-Larsen
UCPH

Ramon Crehuet
IQAC Barcelona

Thea K. Schulze
UCPH

Ikki Yasuda
Keio University

Sören von Bülow
UCPH

Fan Cao
UCPH

Francesco Pesce
Acellera Therapeutics

Arriën Rauh
UCPH

Computational resources

- Biocomputing Core Facility (UCPH)
- Centre for Scientific Computing Aarhus
- Danish e-Infrastructure Cooperation

DeiC

MARIE CURIE ACTIONS

BRAINSTRUC

LUNDBECKFONDEN

novo nordisk fonden

UNIVERSITY OF COPENHAGEN