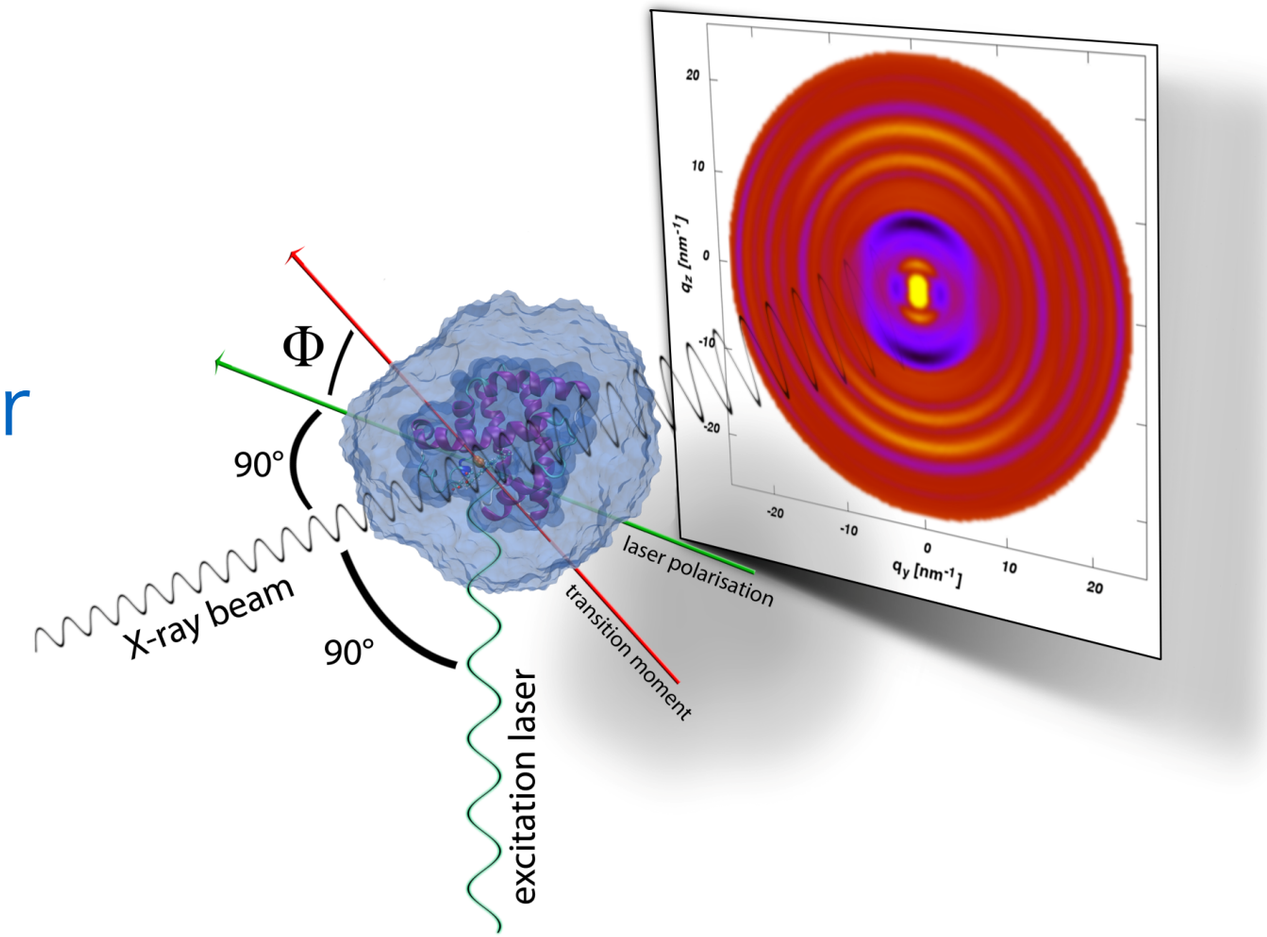
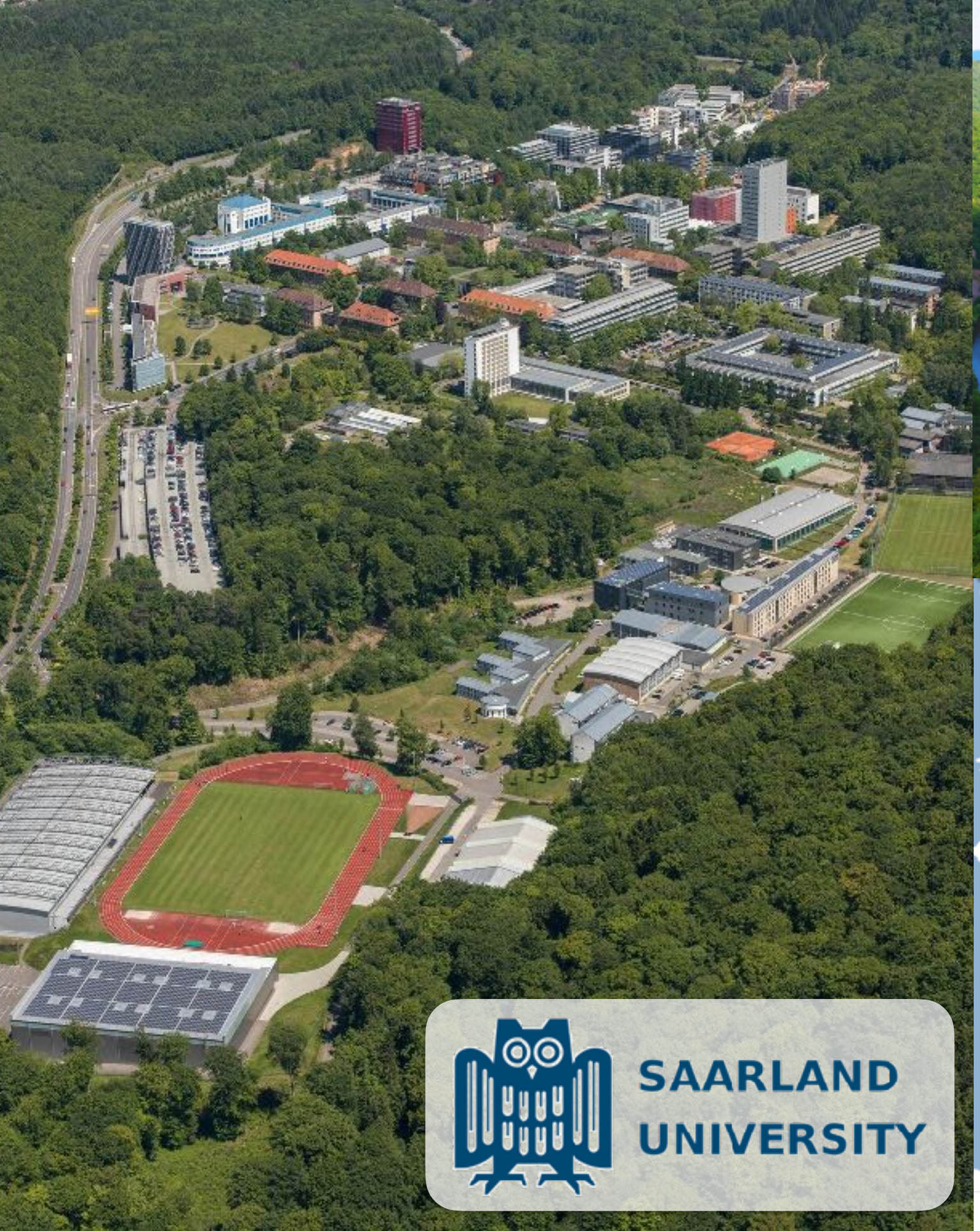
Interpretation of SAXS/ SANS data using Molecular Dynamics Simulations



Jochen Hub, Saarland University

EMBO Practical Course, Grenoble


SAARLAND
UNIVERSITY


Saarland
Germany
Saarland
France

Detecting elephant structures and dynamics

What we want

- High resolution
- Native conditions

Detecting elephant structures and dynamics

What we want

- High resolution
- Native conditions

What we get from
Crystallography / cryo-EM

- Atomic resolution
- No dynamics

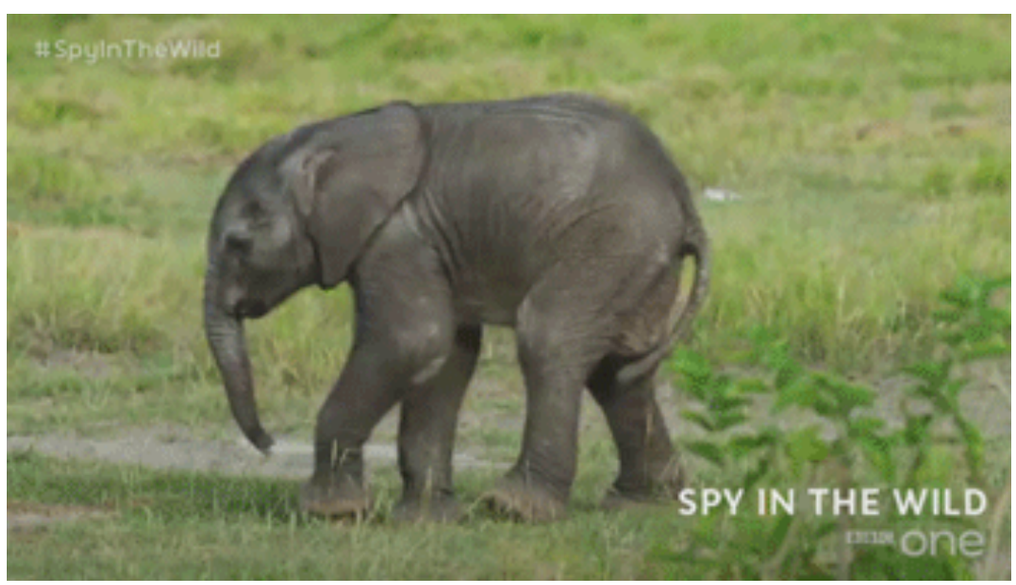
Detecting elephant structures and dynamics

What we want

- High resolution
- Native conditions

What we get from
SAXS

- Poor resolution
- Native conditions



?

Interpretation of SAS data of biomolecules

Few independent data points
$N_{\text{Shannon}} \approx D_{\max} q_{\max} / \pi$
(low information content)

→

Structural interpretation
(many degrees of freedom)

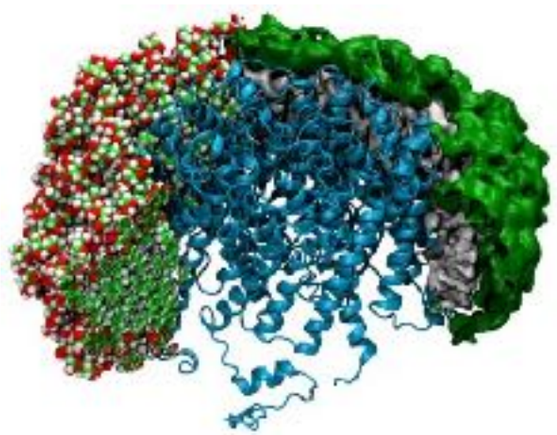
Three ingredients needed

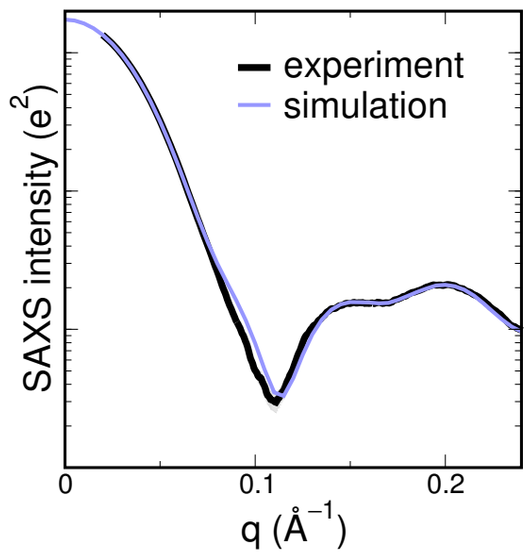
Physical model/ physical information	**Forward model/ *I*(*q*) prediction**	**Sampling algorithm**
Rigid body + volume exclusion	Implicit-solvent (Crysol, FoXS, PepsiSAXS)	Simulated annealing/ "best fit to the data"
…up to…	…up to…	…up to…
All atom/explicit solvent force field	Explicit-solvent (WAXSiS, trjSAXS, …)	Bayesian methods, Maximum Entropy ensemble refinement
$E_{\text{force field}}(\mathbf{R})$	$I_{\text{SAS}}(\mathbf{R})$	**Our work**

Hub, *Curr Opin Struct Biol*, 2018

Validating structures and dynamics against SAXS

Validate/disprove structural models





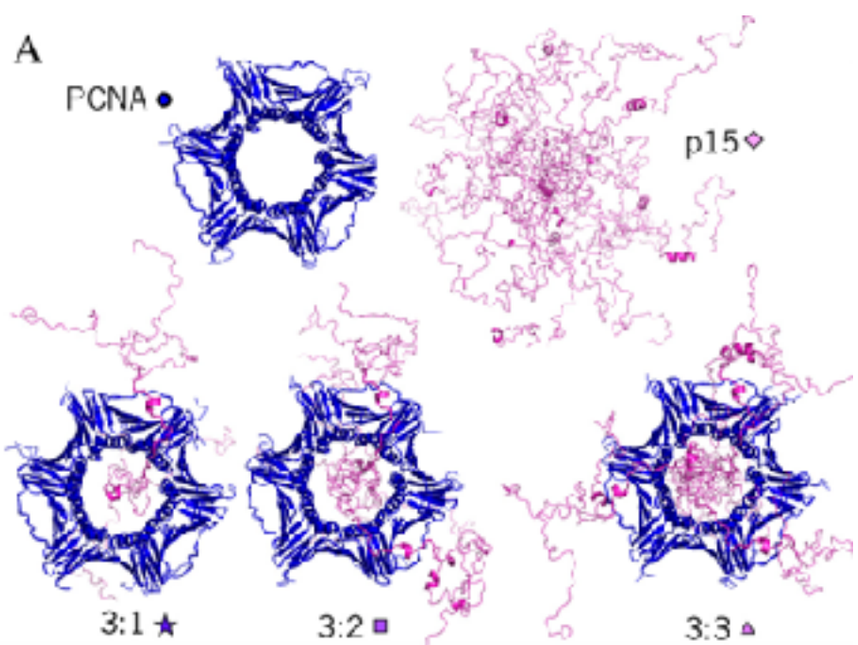
Chen and Hub, Biophys J, 2014
Chen and Hub, *J Phys Chem Lett* (2015)

WAXSiS

Web server for prediction of SAXS data

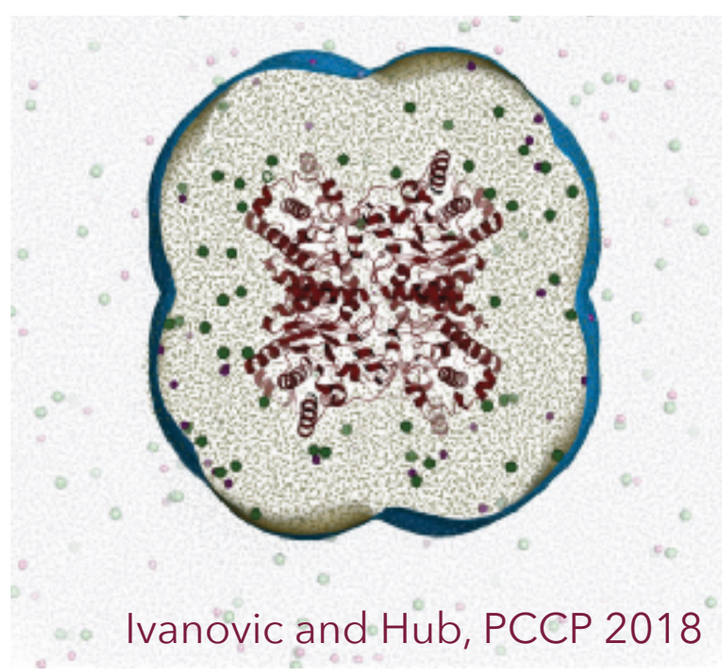
http://waxsis.uni-saarland.de

Knight and Hub, Nucleic Acids Res (2015)



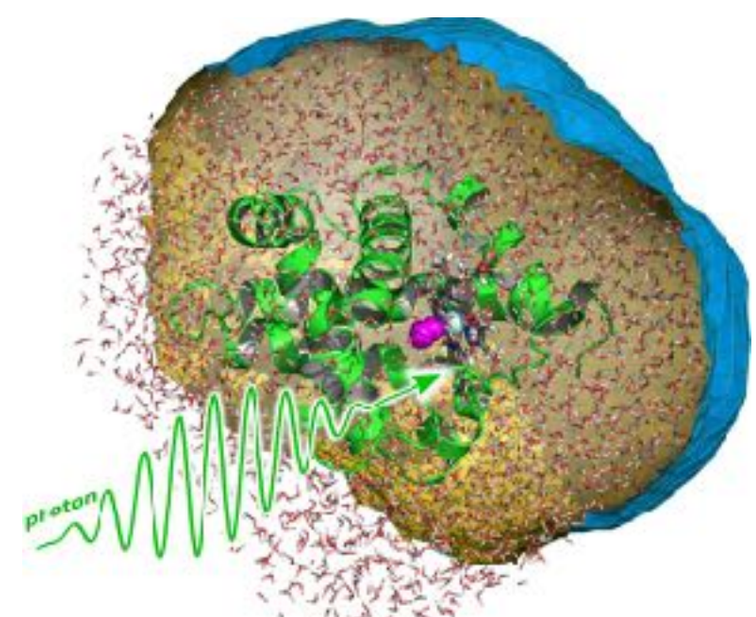
Cordeiro et al., *Nucleic Acids Res* (2017)

Ion cloud effects on SAXS



Ivanovic and Hub, PCCP 2018

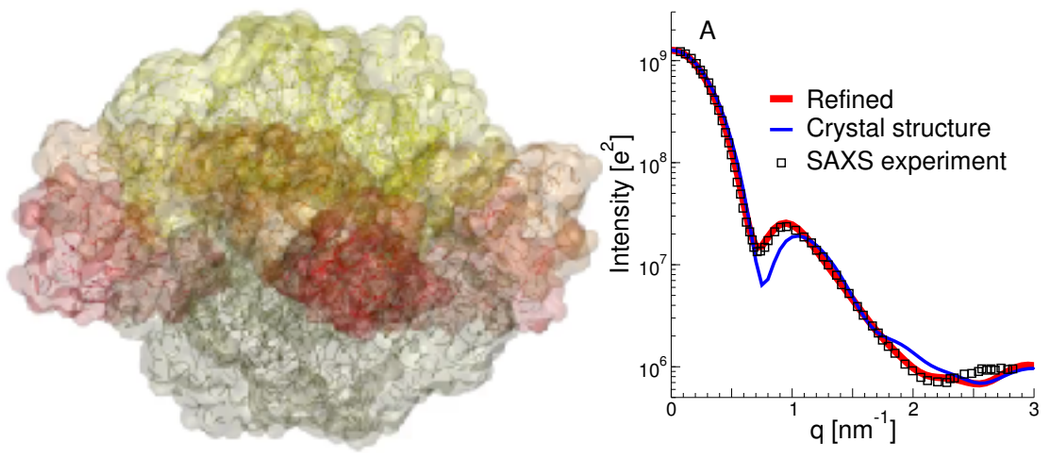
Interpretation of time-resolved SAXS/WAXS data



Brinkmann and Hub., *PNAS* (2016)
Brinkmann and Hub., *J Chem Phys (2015)*

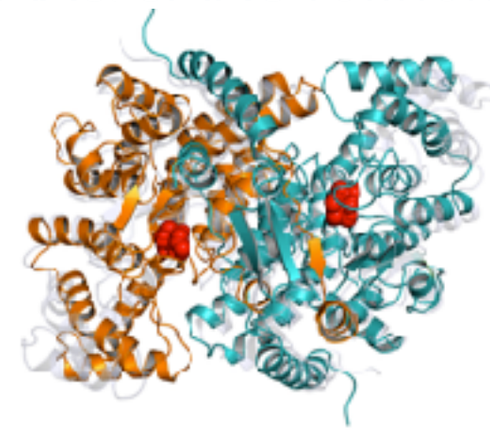
Structure refinement against SAXS (and SANS)

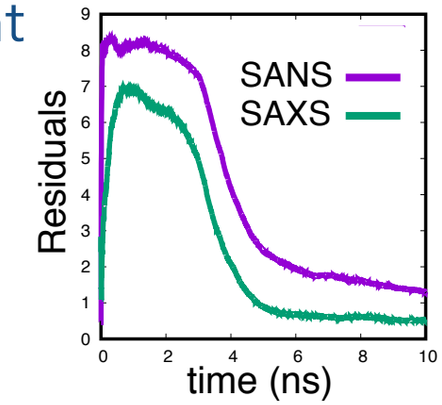
Structure refinement again SAXS data



Chen and Hub, *Biophys J* (2015)

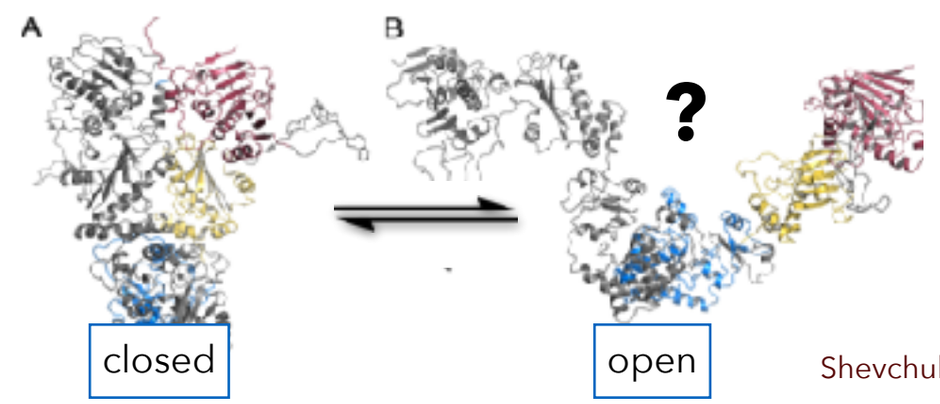
SAXS/SANS refinement

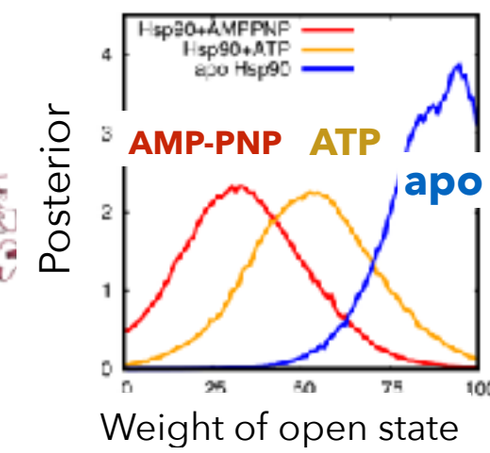




Chen et al., JCTC (2019)

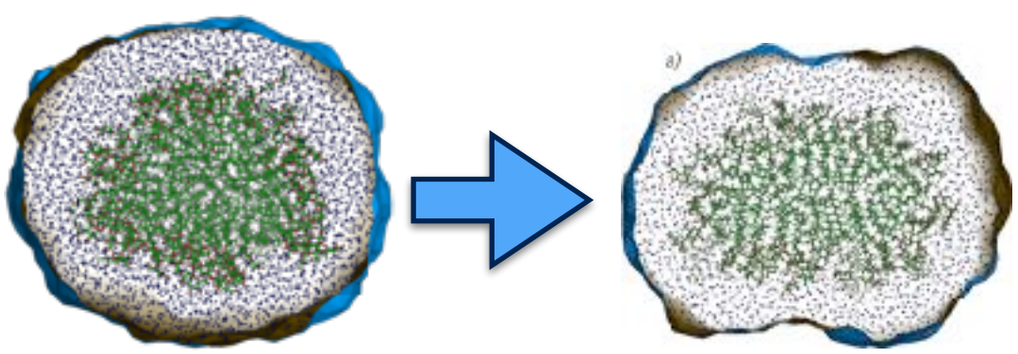
Bayesian refinement

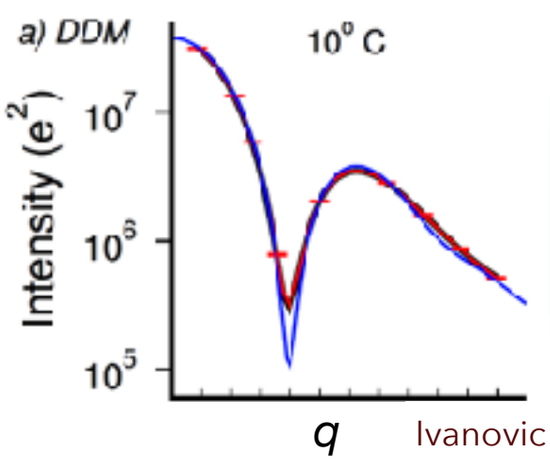




Shevchuk & Hub, PLoS Comp Biol (2017)

Soft-matter systems

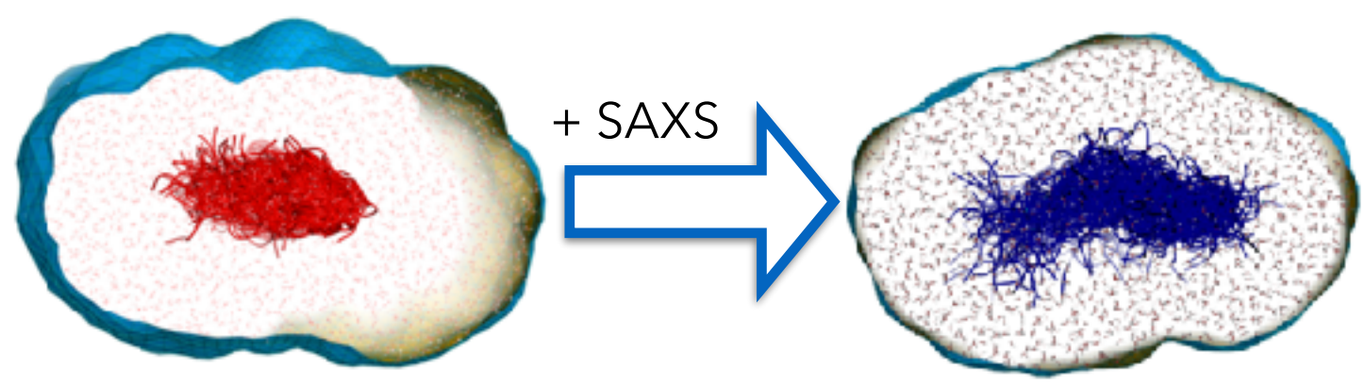




Experiment
Free simulation
Refined

Ivanovic et al., Angew Chem Int Ed, 2018

Maximum-entropy ensemble refinement



Hermann & Hub, JCTC (2019)

Information content of SAXS data

Information content of SAXS data

Number of independent data points:

(by Shannon sampling theorem)

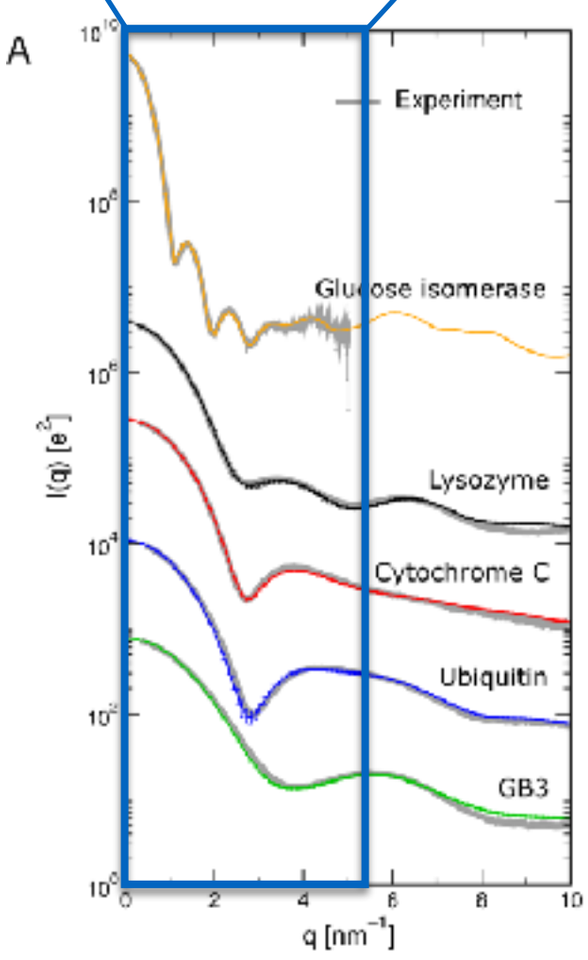
$$N_{\text{Shannon}} = (q_{\max} - q_{\min}) D_{\text{solute}} / \pi$$

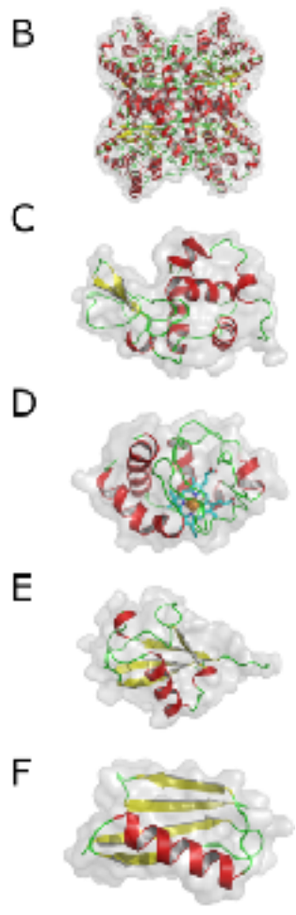
Moore, *J Appl Cryst*, 1980

Somewhat under debate, but a good guess

Assuming q_{max} = 0.5Å

q_{max} = 0.5Å^{-1}





Protein	D_{solute}(Å)	$N_{Shannon}$
Glucose isomerase	94	15
Lysozyme	43	6.8
Cytochrome C	38	6
Ubiquitin	43	6.8
GB3 domain	34	5.3

Information content of SAXS data

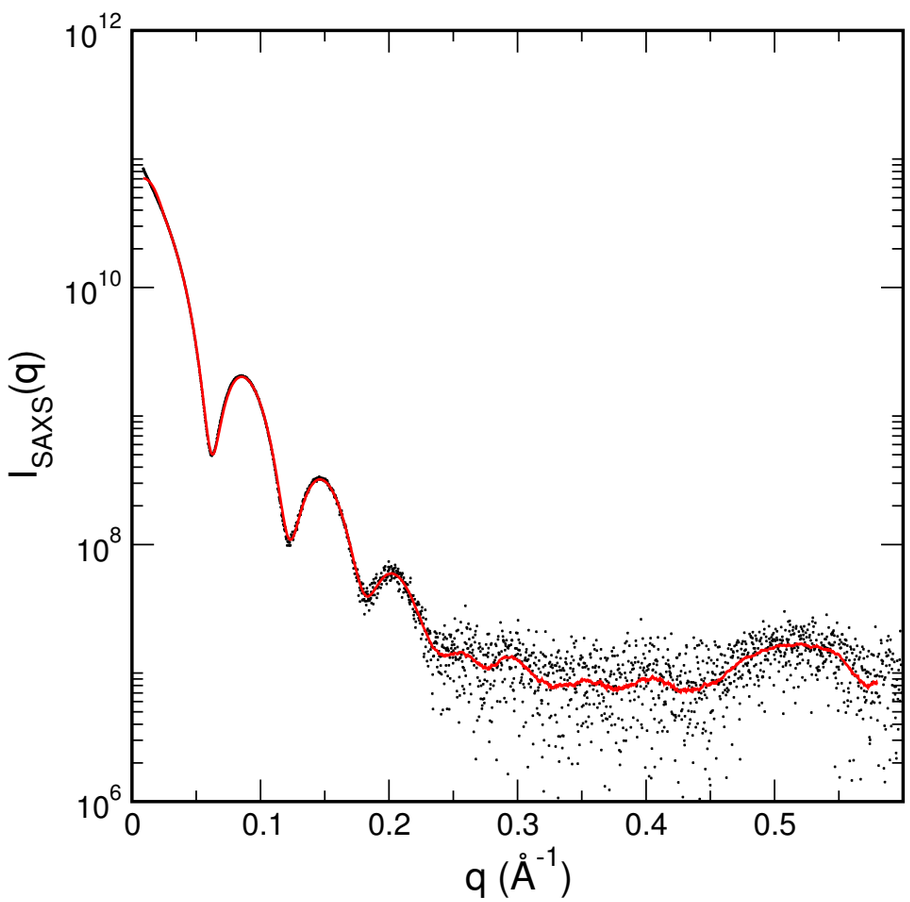
Number of independent data points:

(by Shannon sampling theorem)

$$N_{\text{Shannon}} = (q_{\max} - q_{\min}) D_{\text{solute}} / \pi$$

Moore, *J Appl Cryst*, 1980

Somewhat under debate, but a good guess



Example: Apoferritin from SASBDB

- SAXS curve highly oversampled
- Points contain independent estimates for the underlying SAXS curve, but not independent structural information!

Overfitting

The Party Pooper's topic… :-)

John von Neumann:

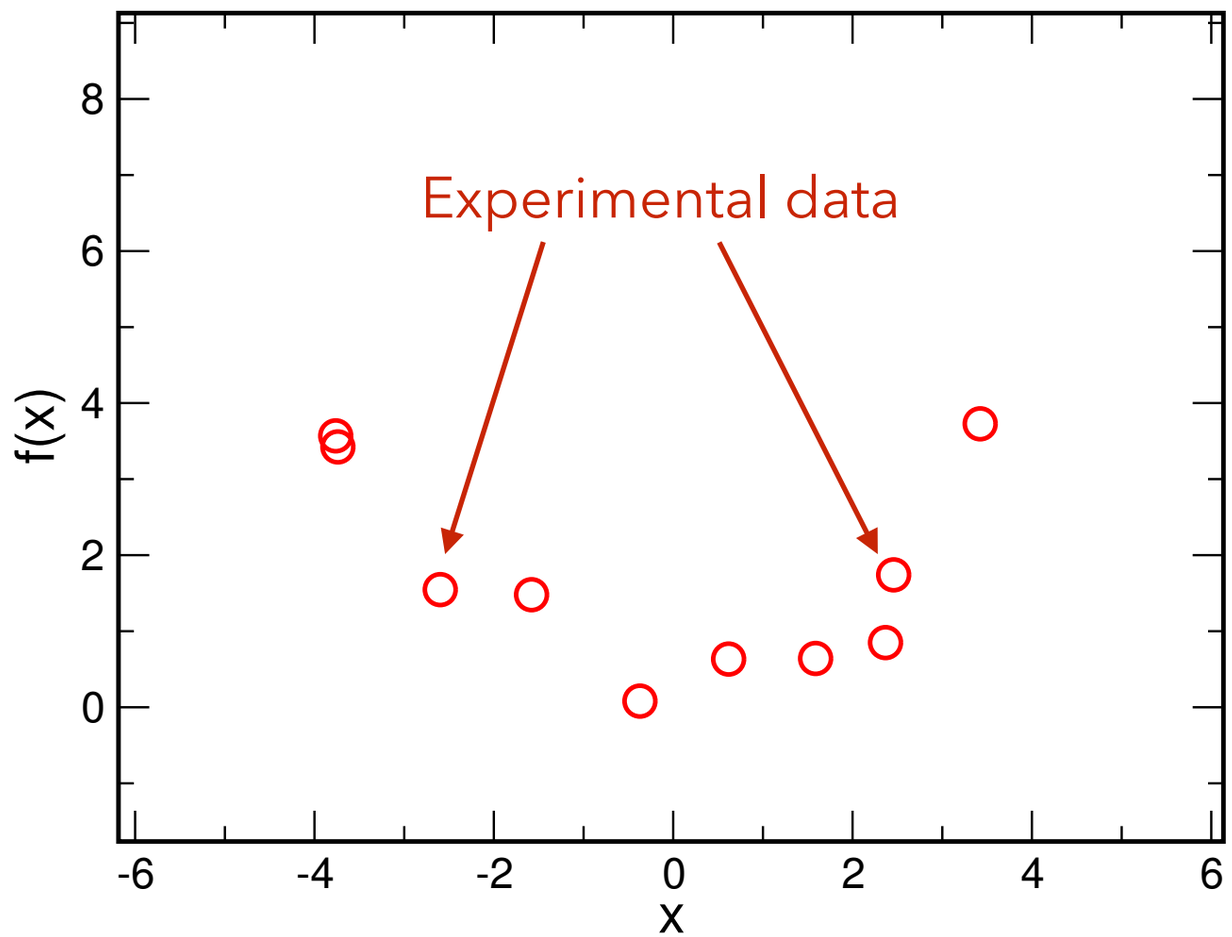
"With four parameters I can fit an elephant, and with five I can make him wiggle his trunk."

Or what he meant: Don't be impressed if you can make a complex model (with many parameters) fit some data.

Polynomial curve fitting

Example: Polynomial curve fitting

The Party Pooper's topic... :-)

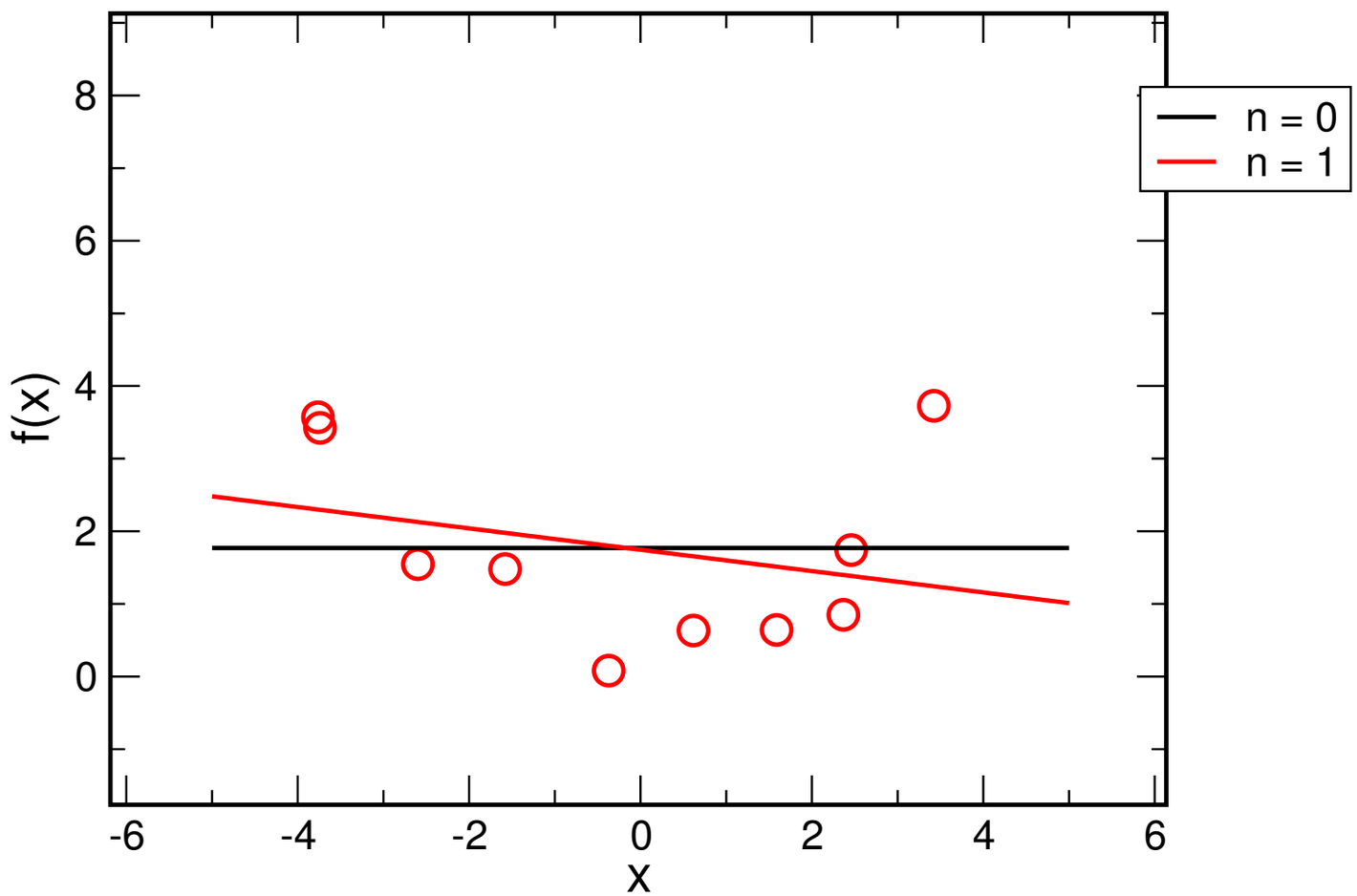


$$f_{\text{fit}}(x) = c_n x^n + c_{n-1} x^{n-1} + \ldots + c_1 x + c_0$$

Adopted from: Bishop, Maschine Learning and Pattern Recognition.

Polynomial curve fitting

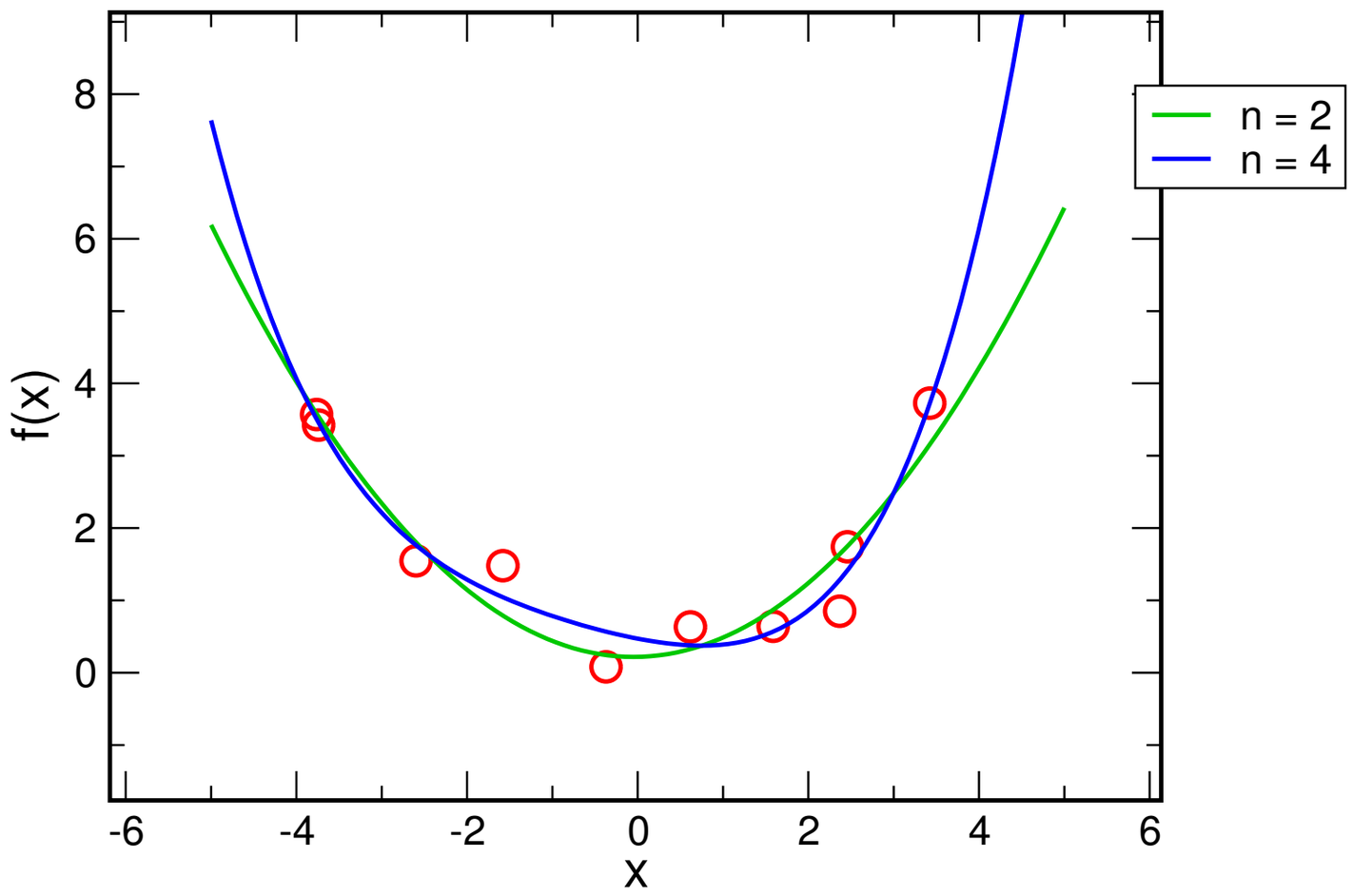
Example: Polynomial curve fitting



$$f_{\text{fit}}(x) = c_n x^n + c_{n-1} x^{n-1} + \ldots + c_1 x + c_0$$

Polynomial curve fitting

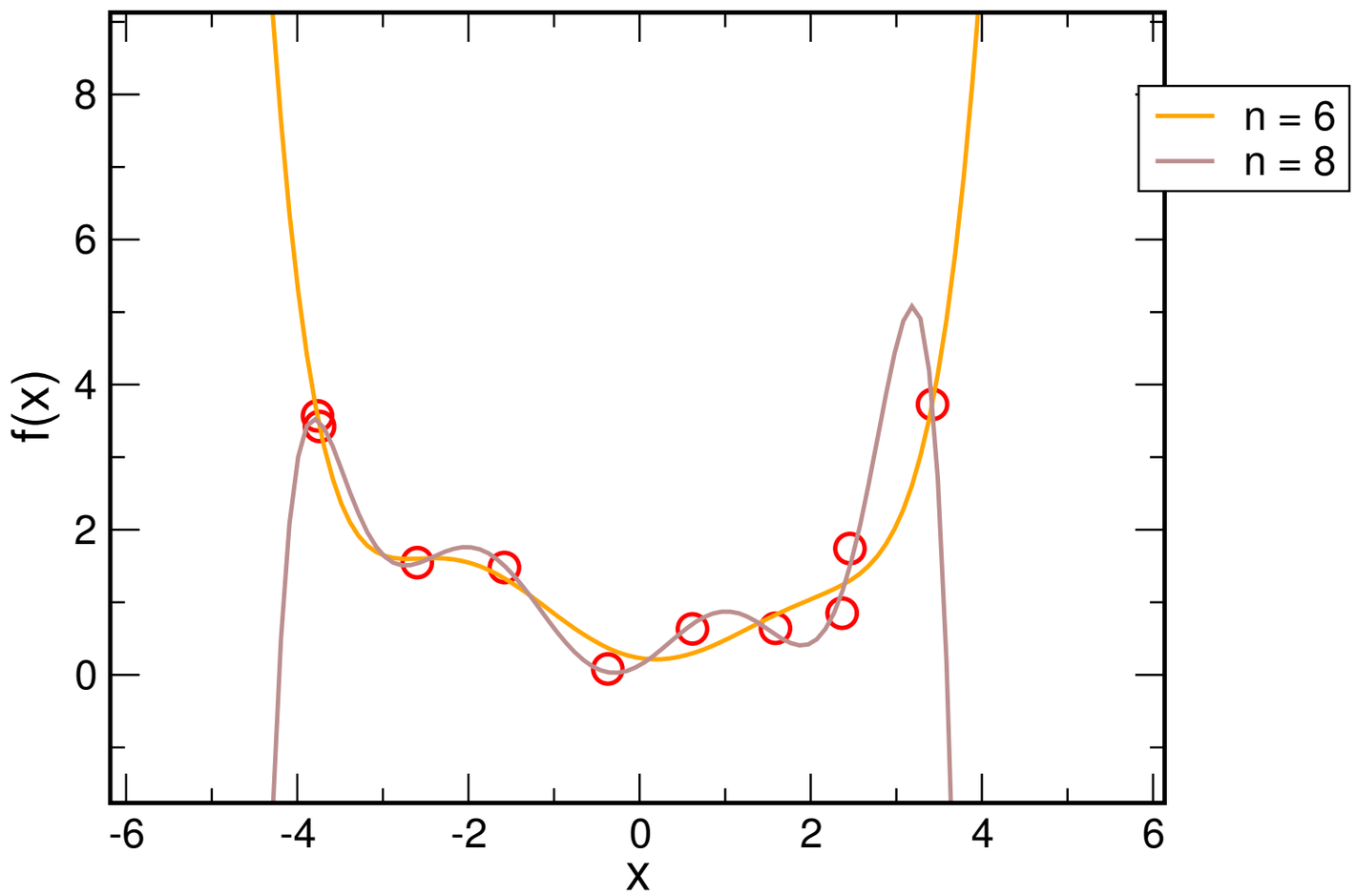
Example: Polynomial curve fitting



$$f_{\text{fit}}(x) = c_n x^n + c_{n-1} x^{n-1} + \ldots + c_1 x + c_0$$

Polynomial curve fitting

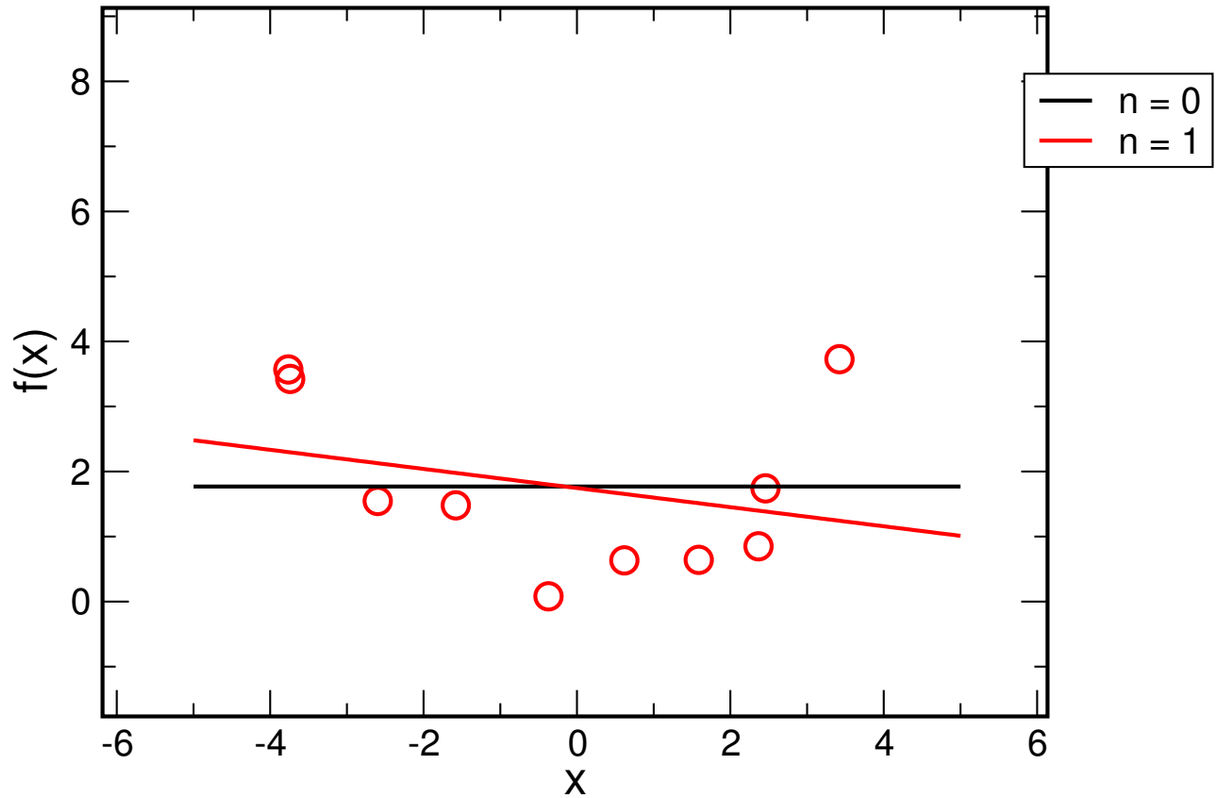
Example: Polynomial curve fitting

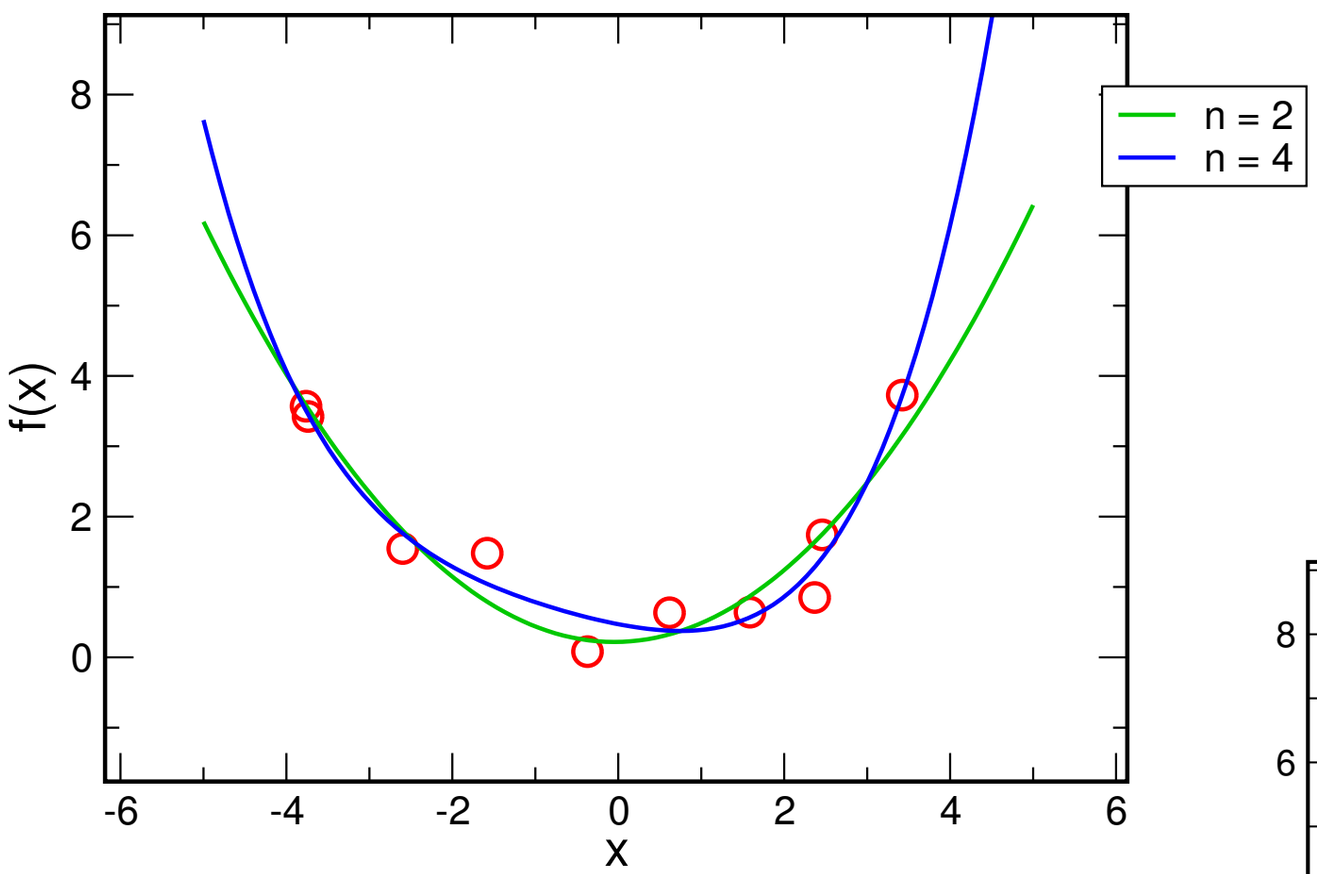


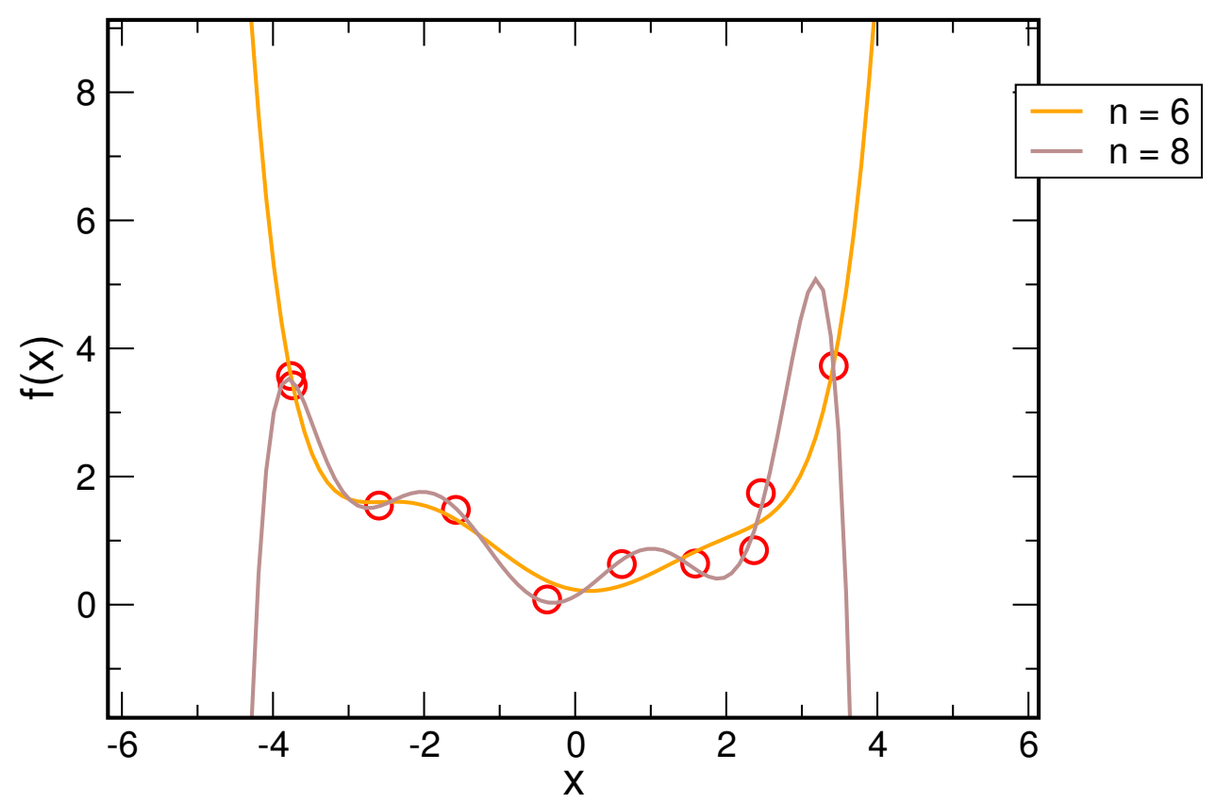
$$f_{\text{fit}}(x) = c_n x^n + c_{n-1} x^{n-1} + \ldots + c_1 x + c_0$$

Which is the correct "model"?

Example: Polynomial curve fitting

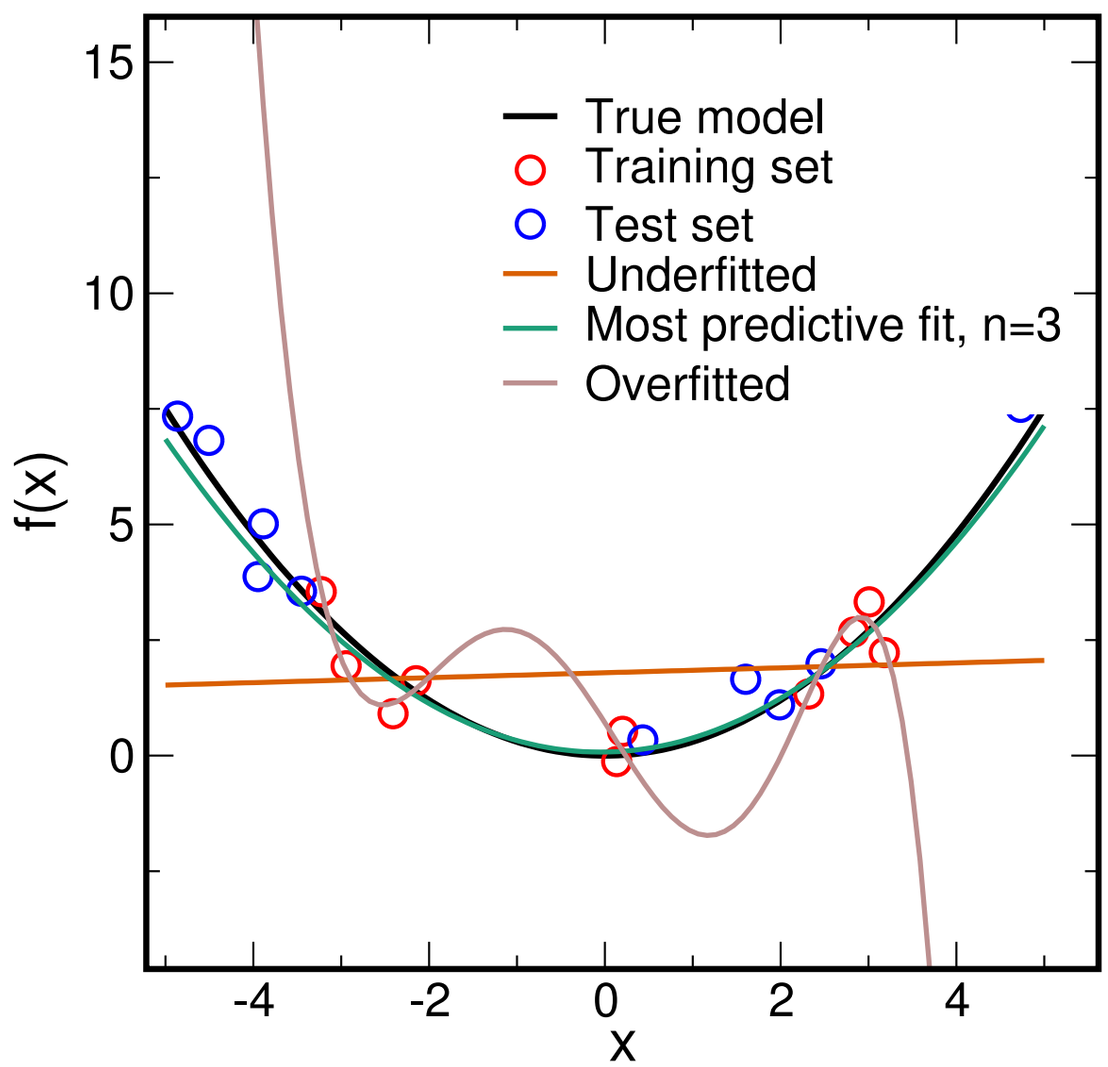




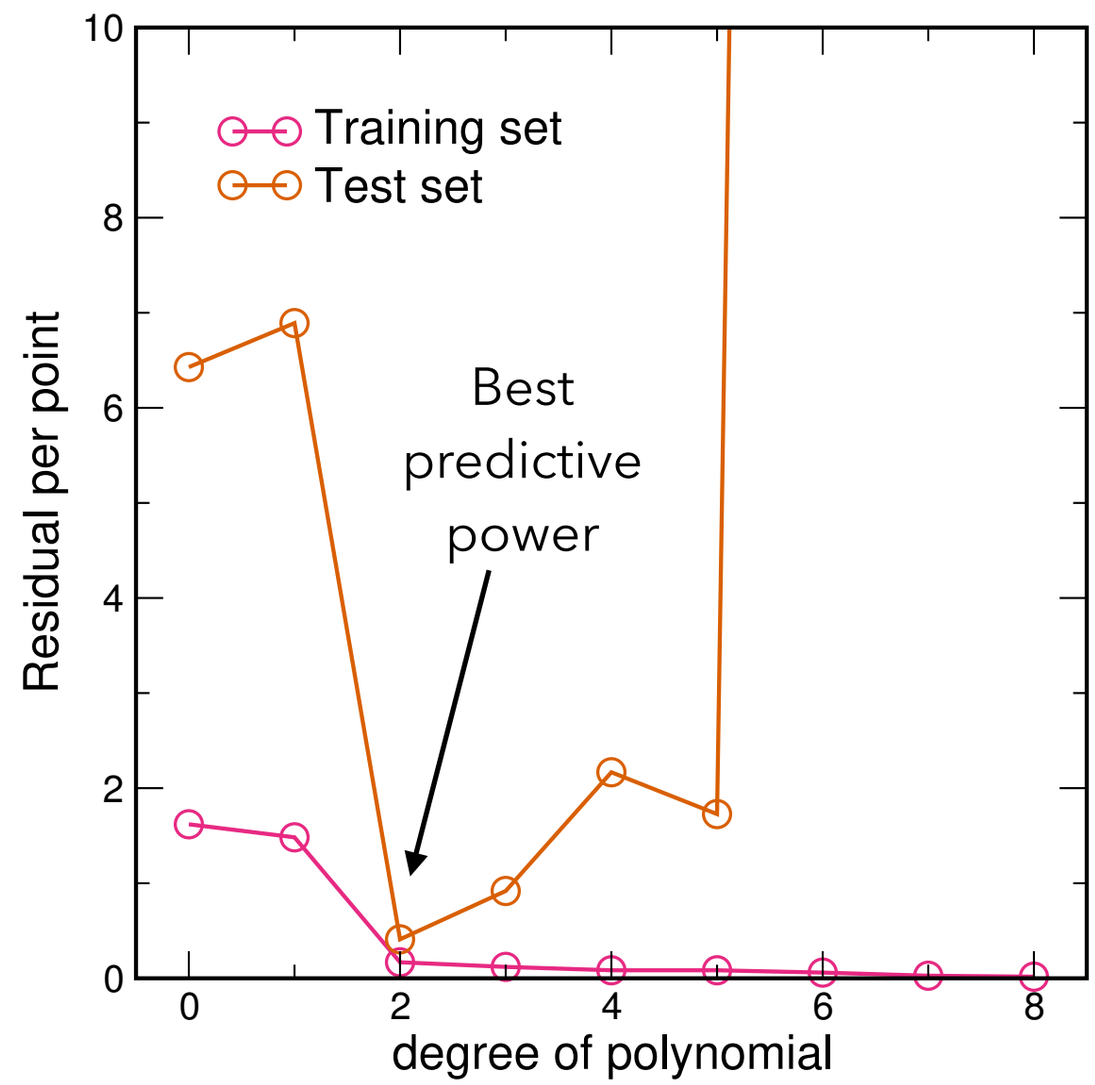


Avoiding overfitting: Training and test set

Example: Polynomial curve fitting



Avoid overfitting with a test set



Overfitting

John von Neumann:

"With four parameters I can fit an elephant, and with five I can make him wiggle his trunk."

Or what he meant: Don't be impressed if you can make a complex model (with many parameters) fit some data.

Key message:

- Having a "good fit" does not guarantee that you have learned anything about the underlying structure / physics
- There may be many other models that explain / fit the data
- The adjusted parameter (here: polynomial coefficients) may not reflect the physically correct values but merely minimize the residuals.

Overfitting

Overfitting is a major problem if:

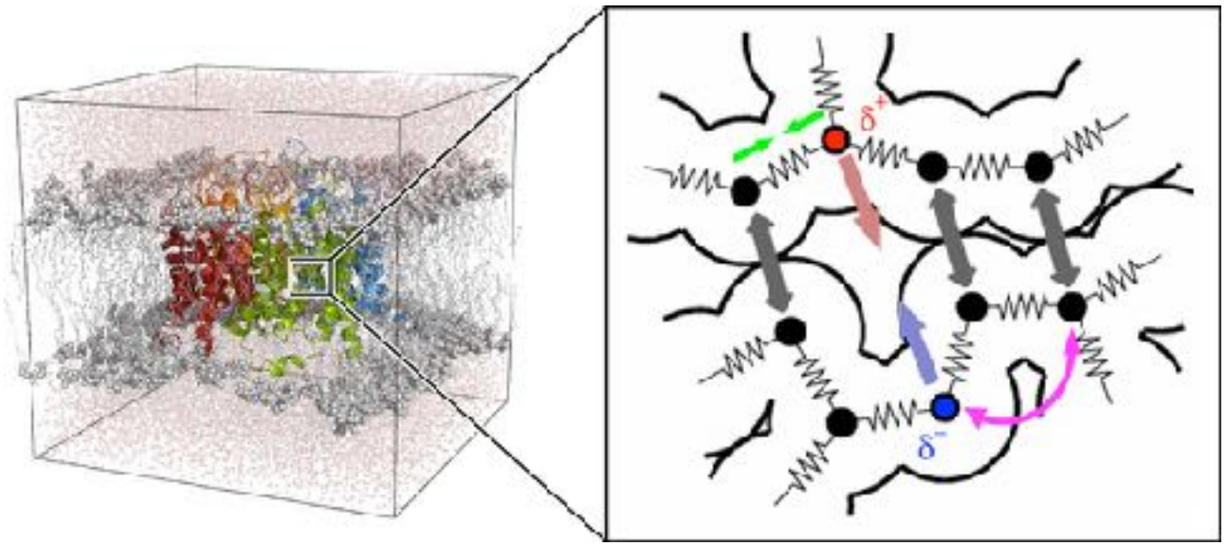
Like in a protein

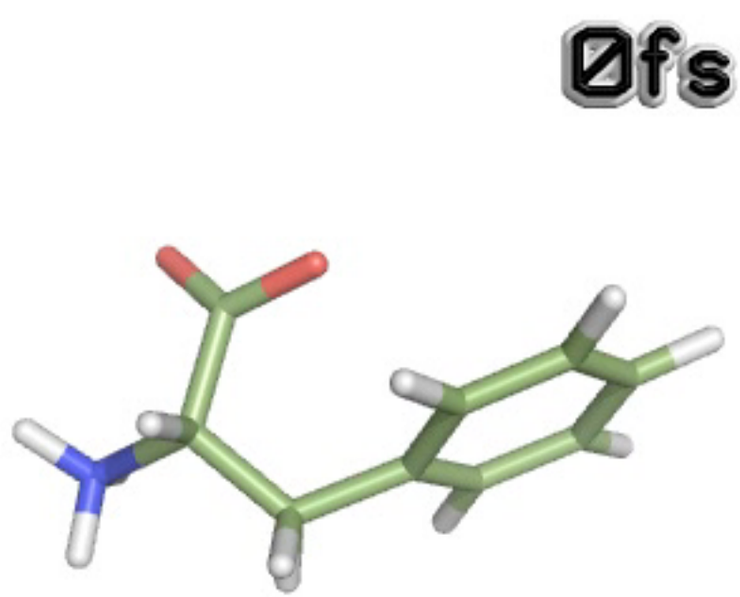
- Number of degrees of freedom of model exceeds the number of independent data points
- No test set available Like in SAXS
- If data in test and training sets are correlated
- (Noisy data → fitting noise instead of underlying model)

Avoiding overfitting:

- Validate fitted model against a independent (!) test set
- Add additional information, e.g.
  1) a reasonable maximum polynomial order
  2) knowledge on the polynomial coefficients
- Use Bayesian inference

Molecular Dynamics (MD) Simulations





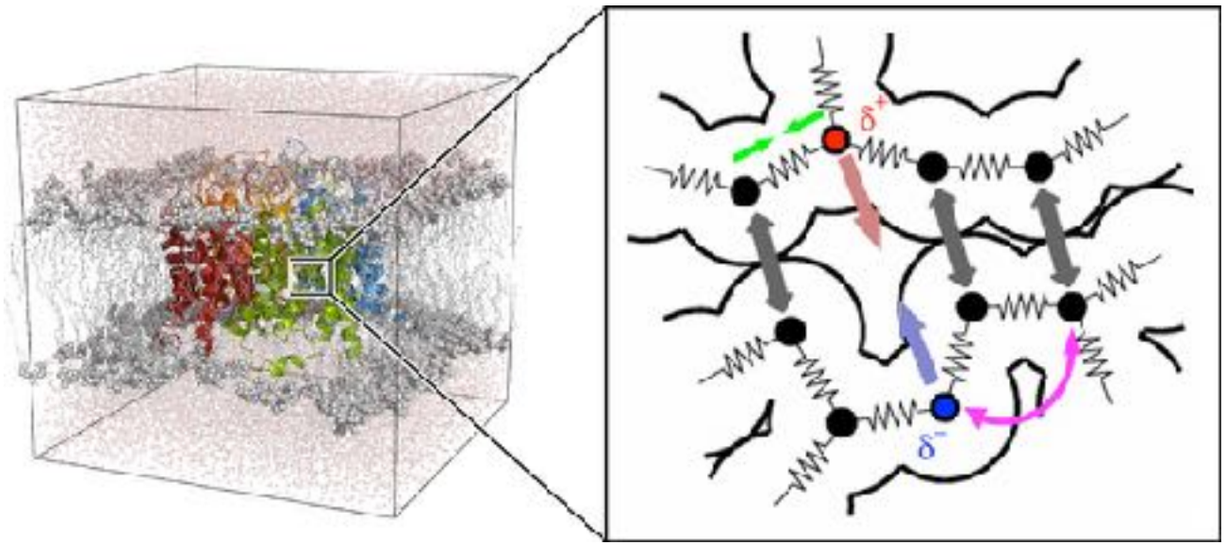
"Force field"

$$
\begin{aligned}
V(\mathbf{x}) &= V_{\text{bonded}}(\mathbf{x}) + V_{\text{non-bonded}}(\mathbf{x}) \\
&= \sum_{\text{bonds } i} k_i^{(\text{b})}(r_i - r_{0,i})^2/2 \\
&\quad + \sum_{\text{angles } i} k_i^{(\theta)}(\theta_i - \theta_{0,i})^2/2 \\
&\quad + \sum_{\text{dihedrals } i} V_i^{(\phi)}(\phi_i) \\
&\quad + \sum_{\text{impropers } i} k_i^{(\xi)}(\xi_i - \xi_{0,i})^2/2 \\
&\quad + \sum_{\text{atoms } i,j} \frac{q_i q_i}{4\pi\epsilon_0 r_{ij}} \\
&\quad + \sum_{\text{atoms } i,j} \frac{C_{ij}^{(12)}}{r_{ij}^{12}} - \frac{C_{ij}^{(6)}}{r_{ij}^{6}}
\end{aligned}
$$

Newton's equation of motion

$$\mathbf{F} = m\mathbf{a}$$

Molecular Dynamics (MD) Simulations



"Force field"

$$
\begin{aligned}
V(\mathbf{x}) &= V_{\text{bonded}}(\mathbf{x}) + V_{\text{non-bonded}}(\mathbf{x}) \\
&= \sum_{\text{bonds } i} k_i^{(b)}(r_i - r_{0,i})^2/2 \\
&\quad + \sum_{\text{angles } i} k_i^{(\theta)}(\theta_i - \theta_{0,i})^2/2 \\
&\quad + \sum_{\text{dihedrals } i} V_i^{(\phi)}(\phi_i) \\
&\quad + \sum_{\text{impropers } i} k_i^{(\xi)}(\xi_i - \xi_{0,i})^2/2 \\
&\quad + \sum_{\text{atoms } i,j} \frac{q_i q_j}{4\pi\epsilon_0 r_{ij}} \\
&\quad + \sum_{\text{atoms } i,j} \frac{C_{ij}^{(12)}}{r_{ij}^{12}} - \frac{C_{ij}^{(6)}}{r_{ij}^{6}}
\end{aligned}
$$

First step: Fit to quantum-chemical calculations, e.g.

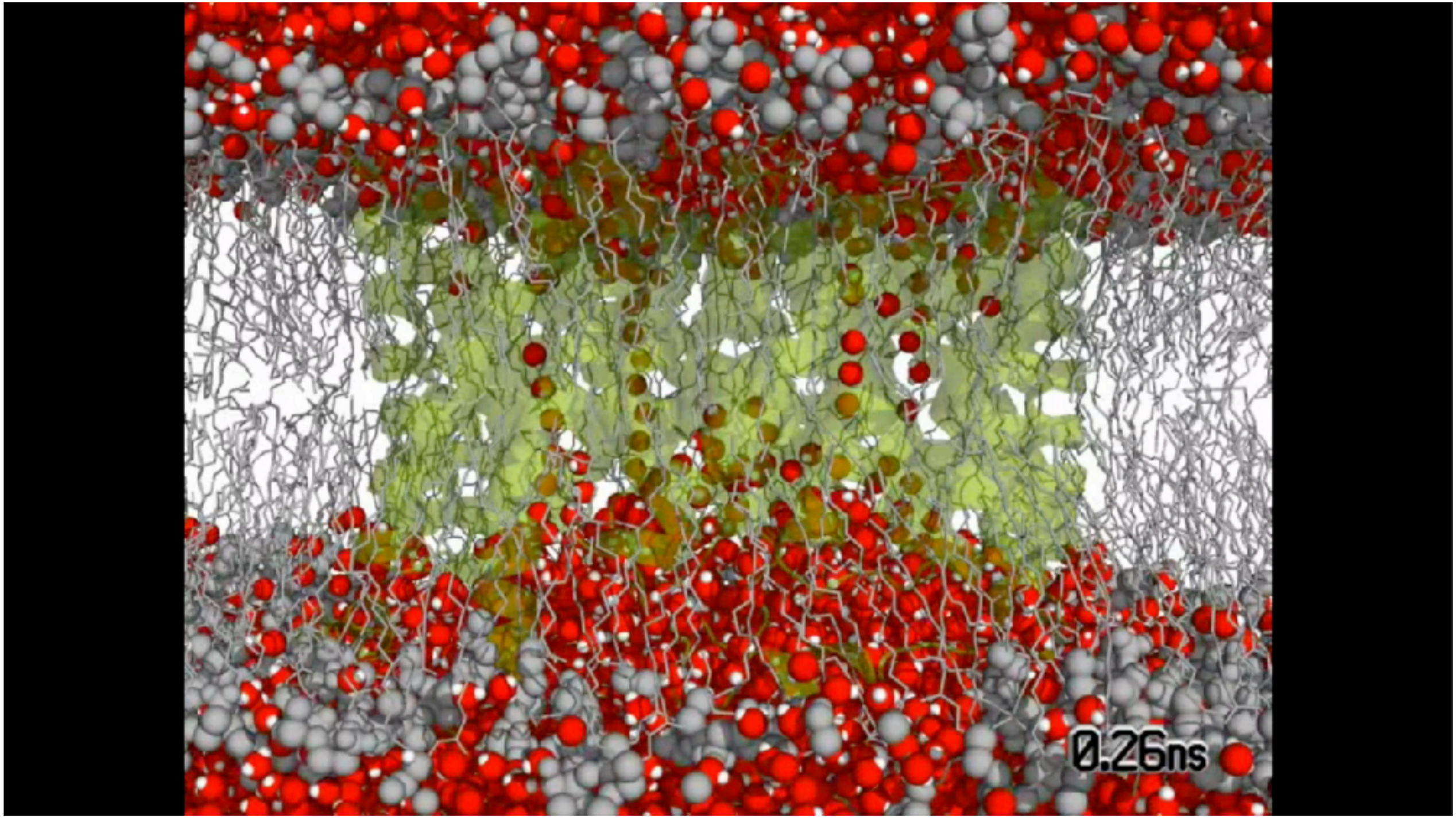
- Post-Hartree-Fock methods
- Density functional theory

Second step: Refine against experimental data

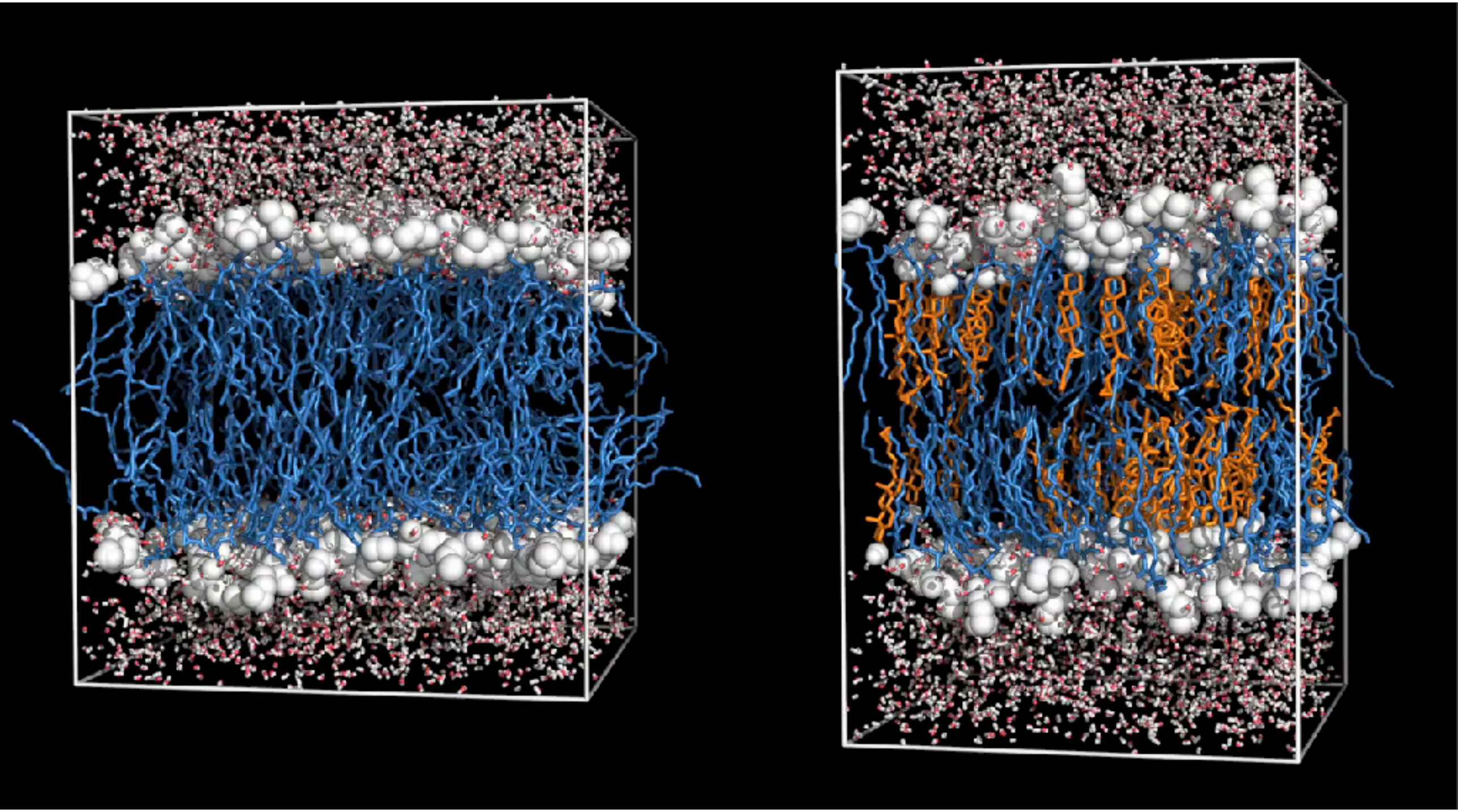
- Hydration free energies
- Partition coefficients
- Densities
- NMR data
- CD spectra
- …

Force field contains physicochemical information

Water permeation through Aquaporin



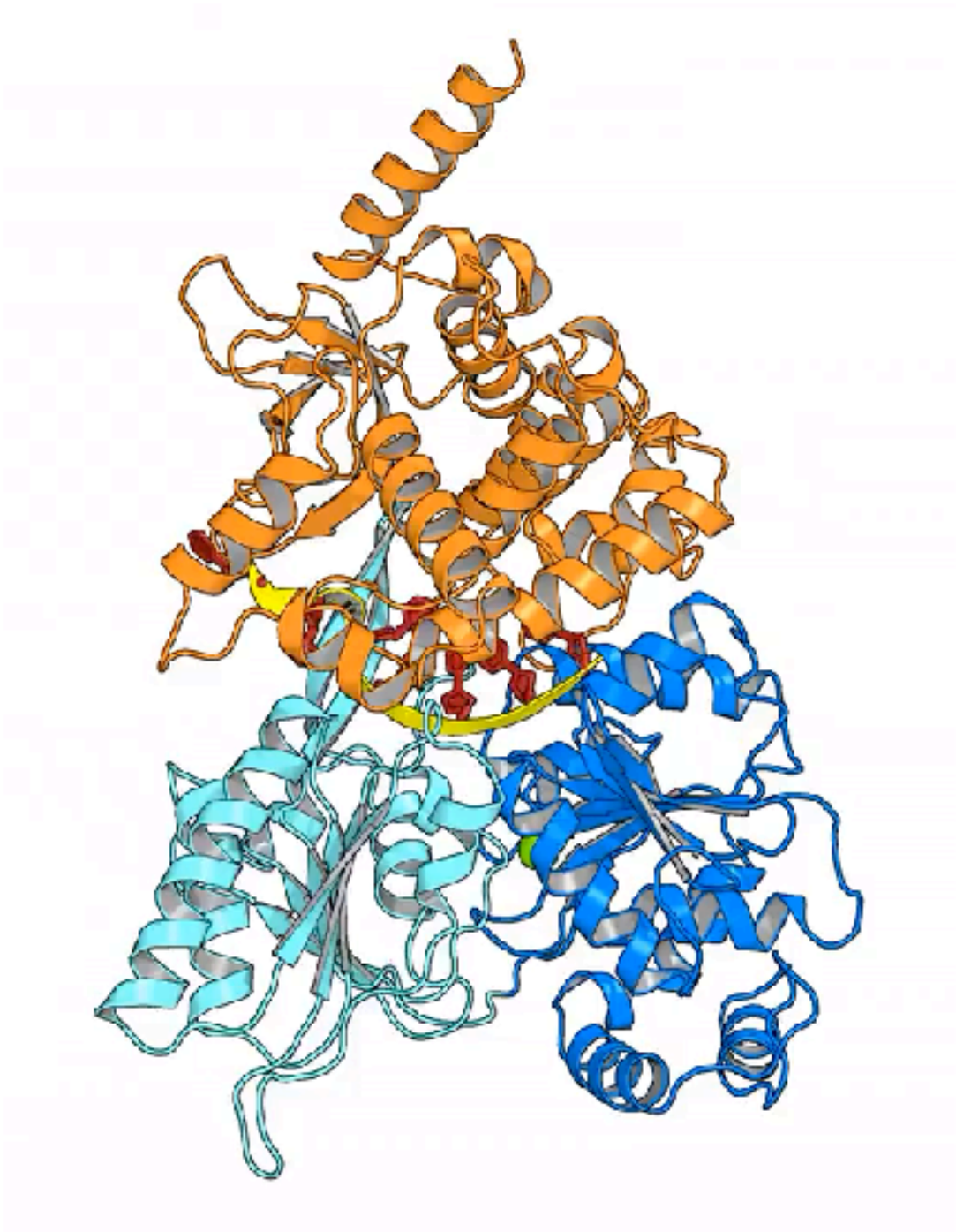
Lipid membrane simulation



DOPC

DOPC + 40% Cholesterol

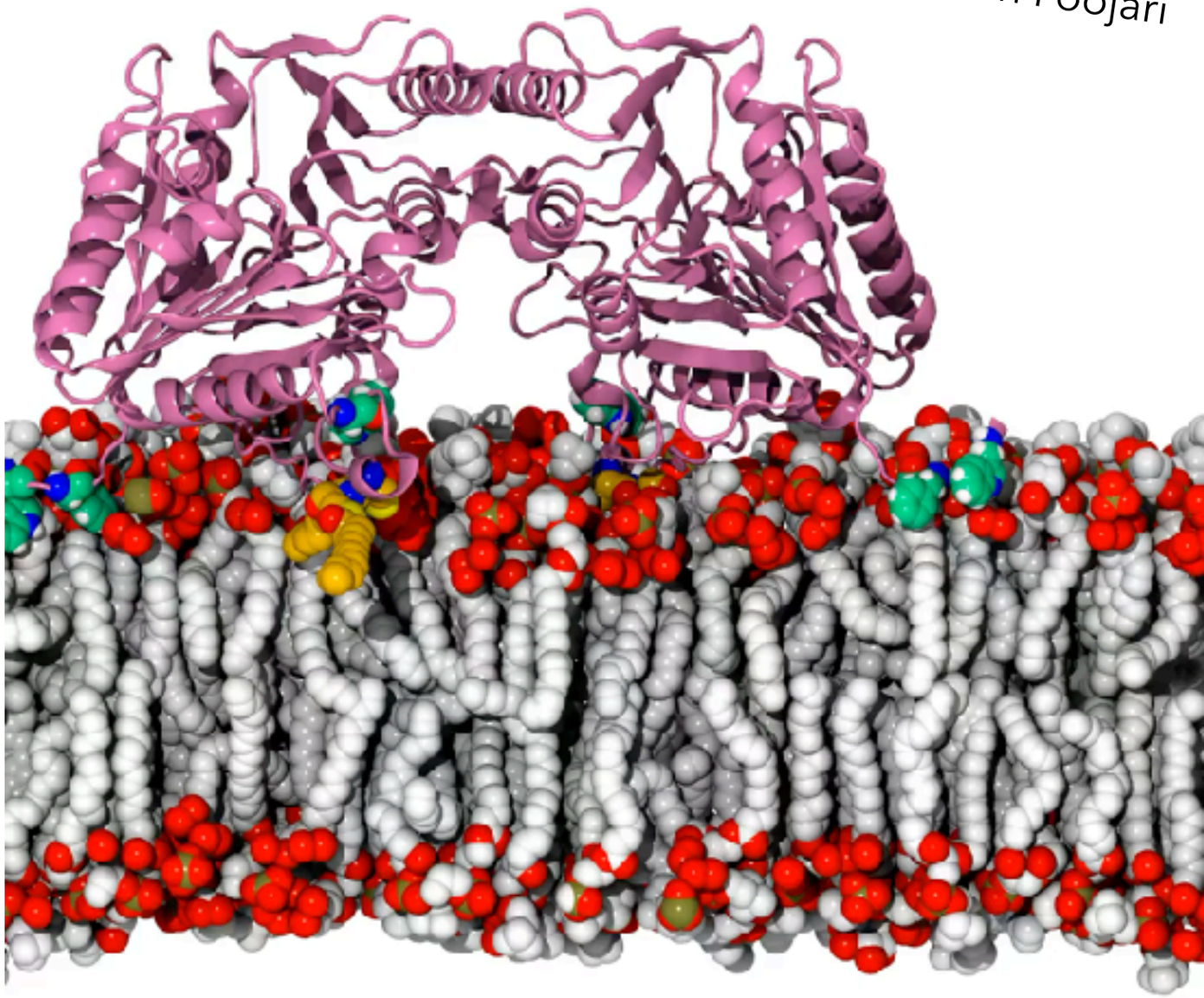
More MD simulations



Helicase, a molecular motor

Becker & Hub, *Commun Biol* (2023)

Movie by Chetan Poojari

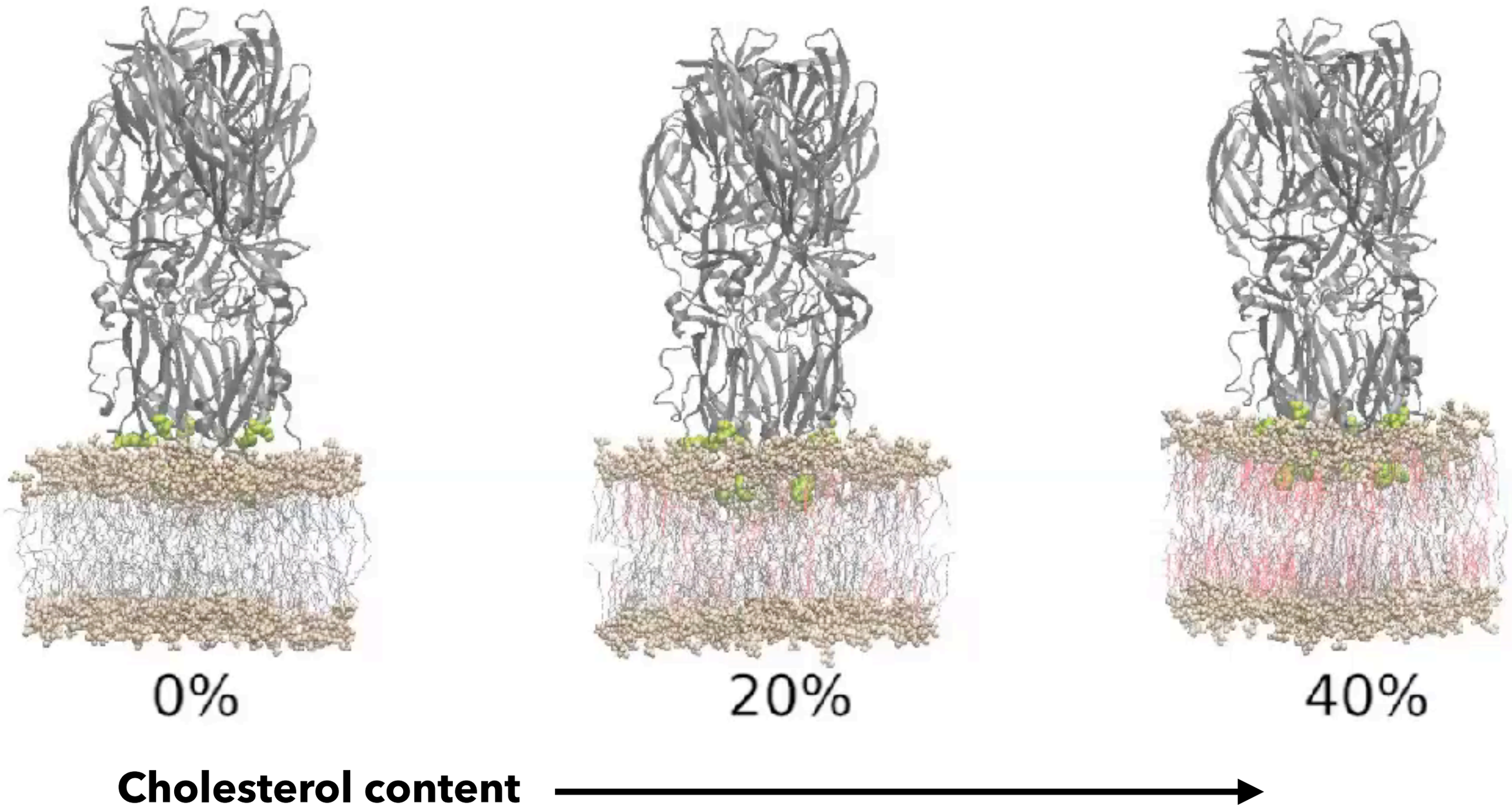


Membrane-anchored protein (unpublished)

Vernuccio, Martinez Leon, Poojari, …, Hub, Guardado-Calvo, *submitted*

Viral fusion protein binding to host membrane

Cooperation with
Félix Rey (Institut Pasteur)



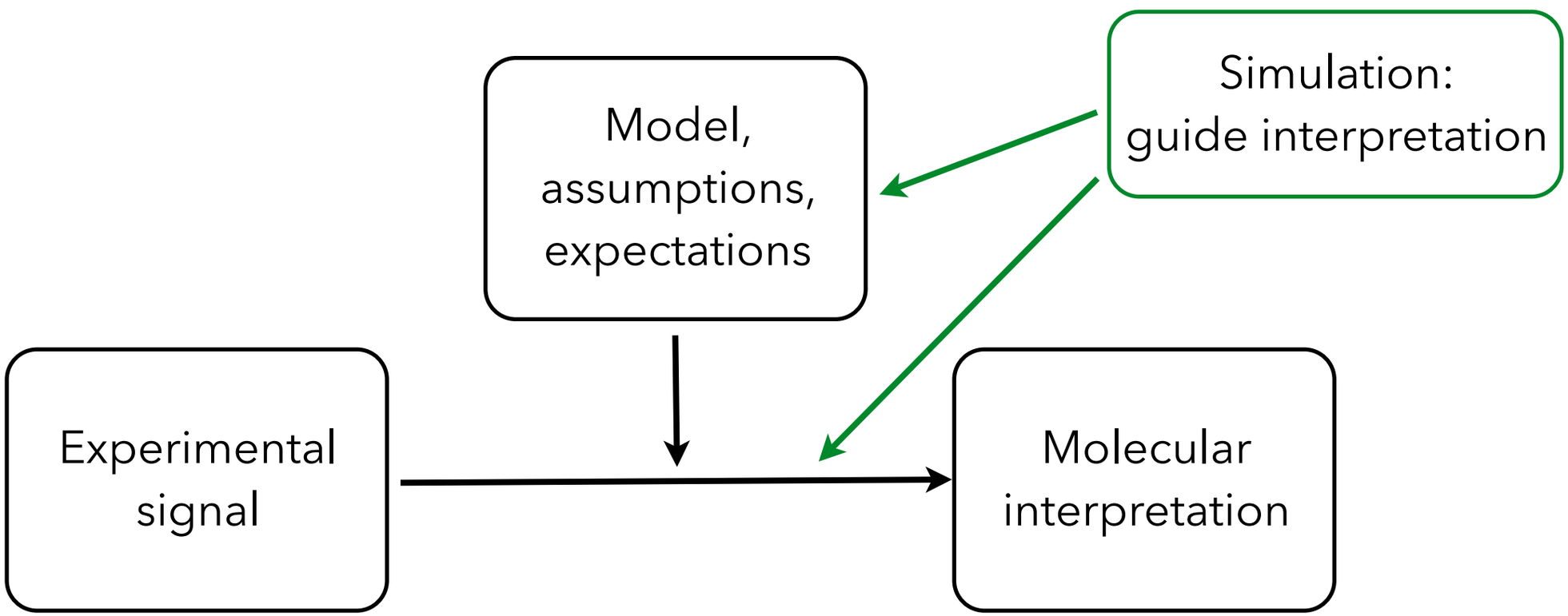
Guardado-Calvo, *et al.*, *Science* 2017

Why Molecular Dynamics Simulations?

Why simulations of biomolecular process?

Atomic details at
full time resolution

(Free) energies and forces available
➙ driving forces for
biomolecular processes



Combining SAXS/WAXS with MD simulations

Guide simulations

SAXS / WAXS

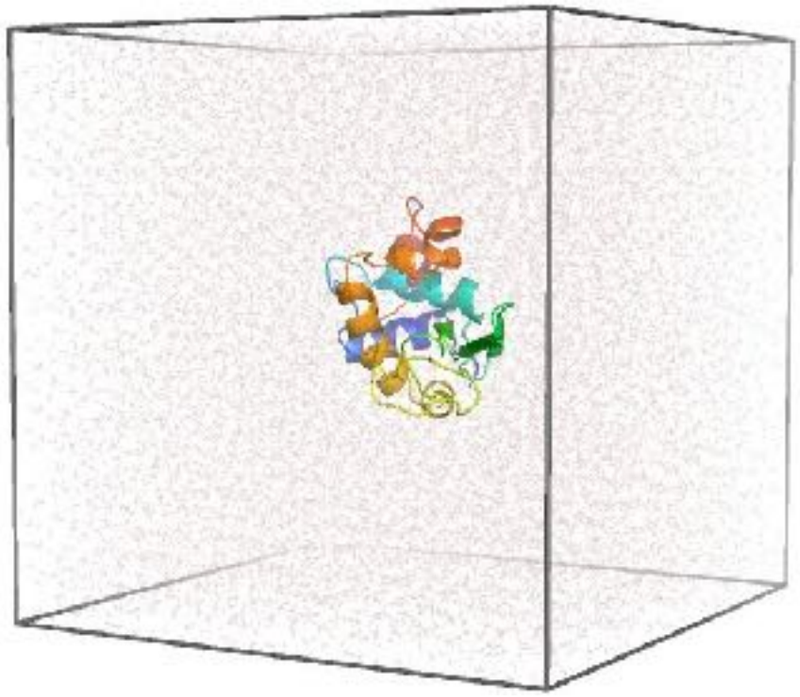
Explicit-solvent MD simulations

Add information

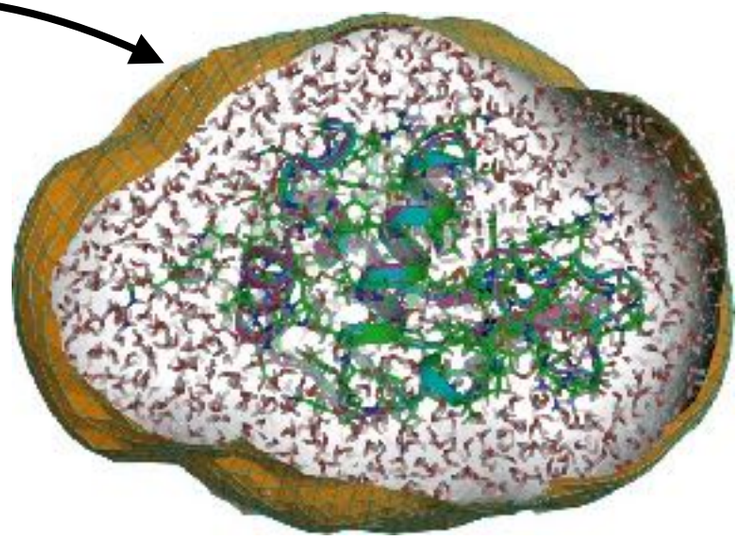
- It's experimental data!
- Hard to interpret
- Low information content
- Scattering contributions from the solvent

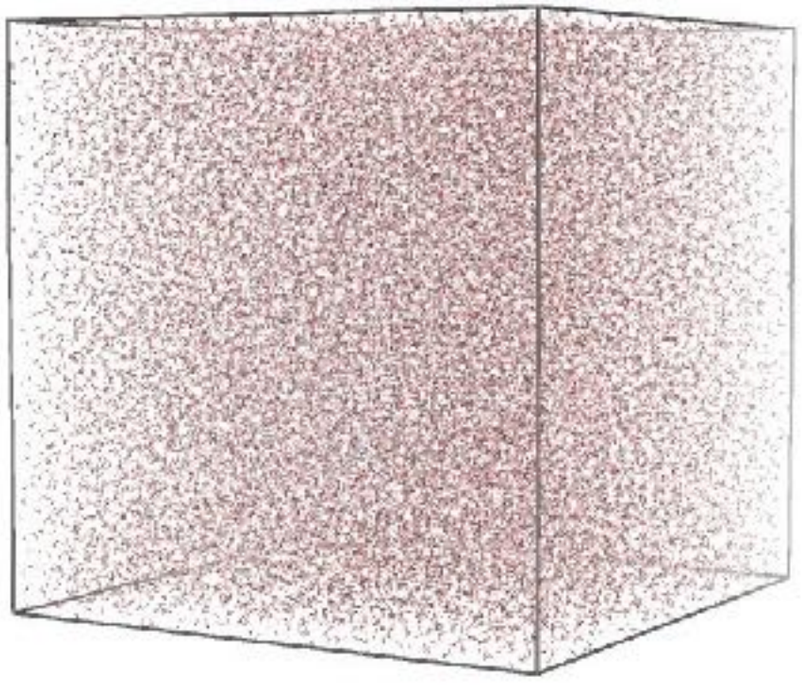
- Quite accurate physical model
- Atomic details, full time resolution
- Thermodynamic driving forces
- Often too slow for large-scale transition (limited to microseconds)
- Force field inaccuracies

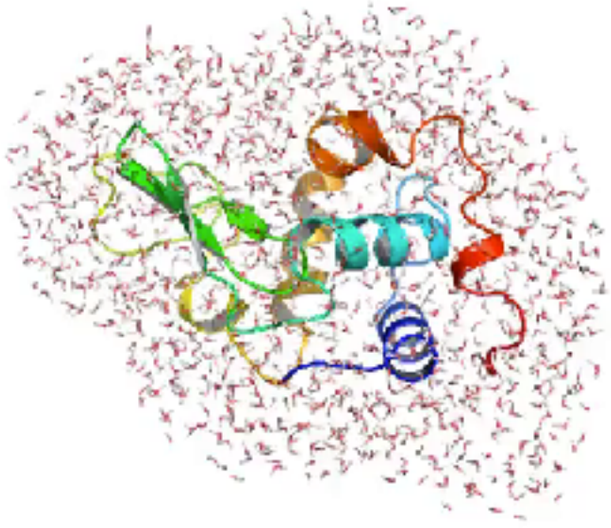
SAXS/WAXS patterns from atomistic MD



Envelope



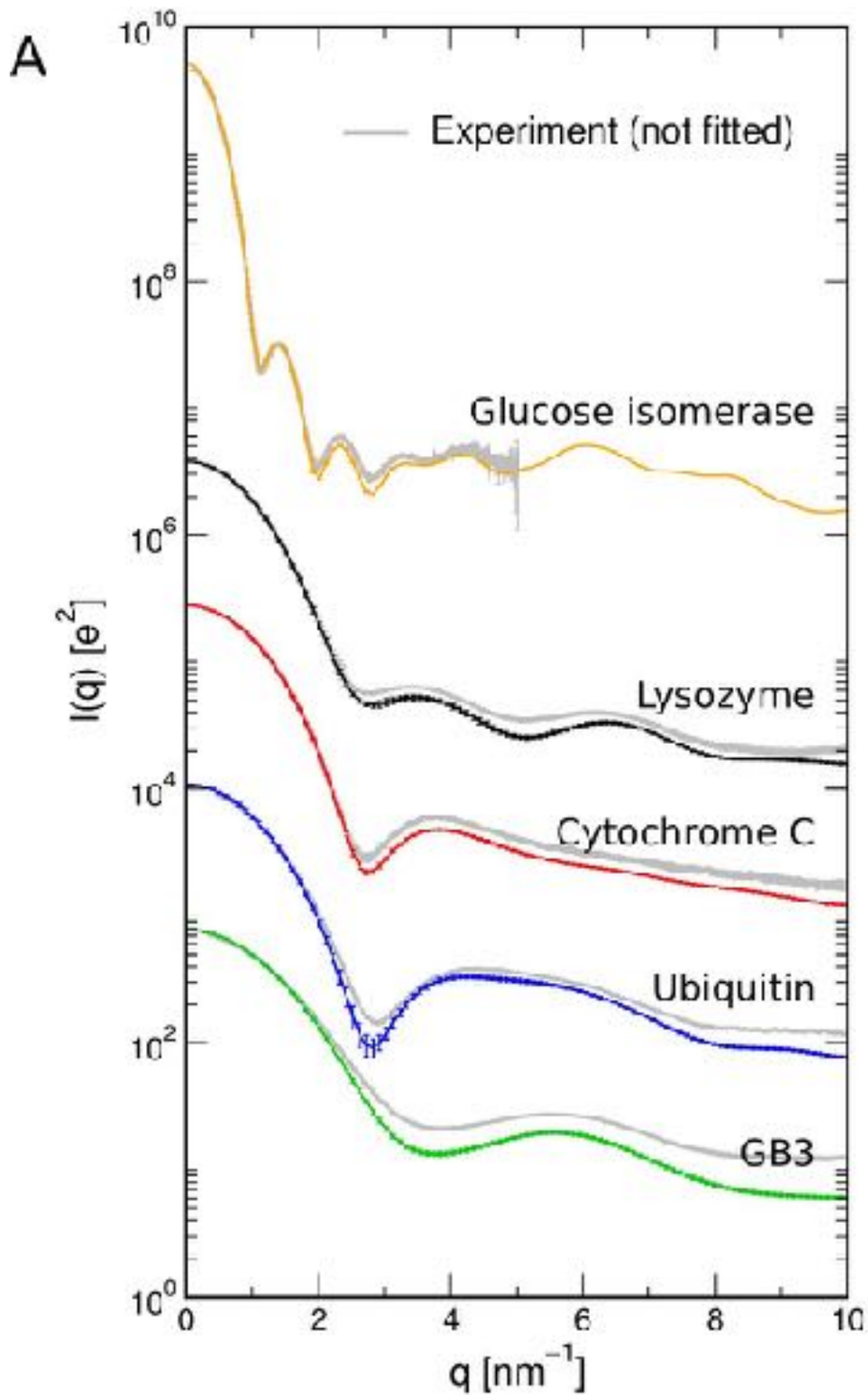


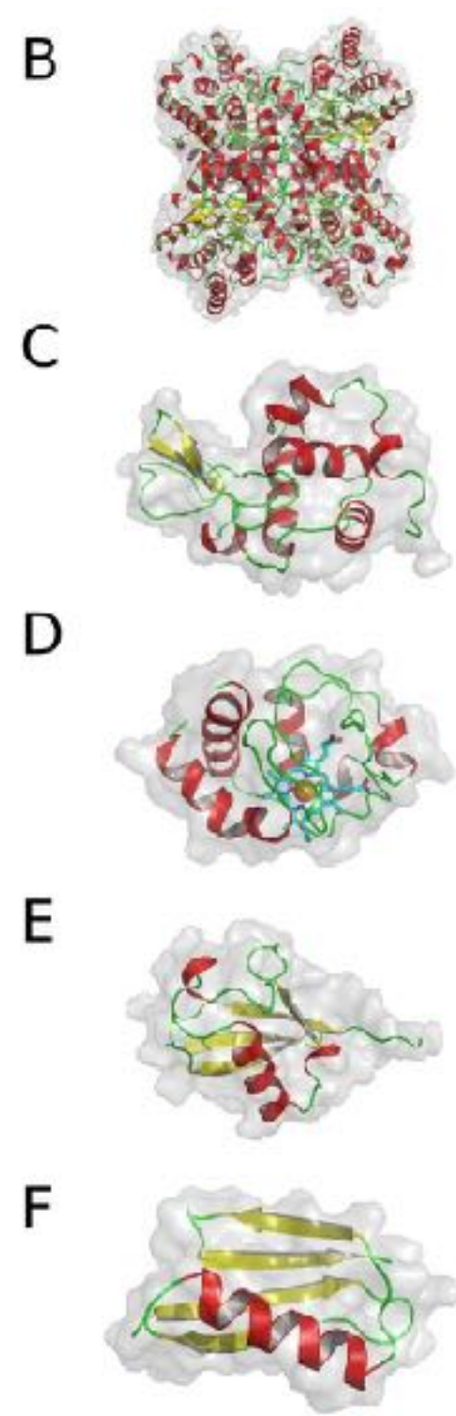


- Accurate hydration layer & excluded solvent
- No solvent-related fitting required

Chen and Hub, *Biophys. J. 107:435-447* (2014)
Park, Bardhan, Roux, Makowski, *J Chem Phys* (2009)

SAXS/WAXS patterns from atomistic MD

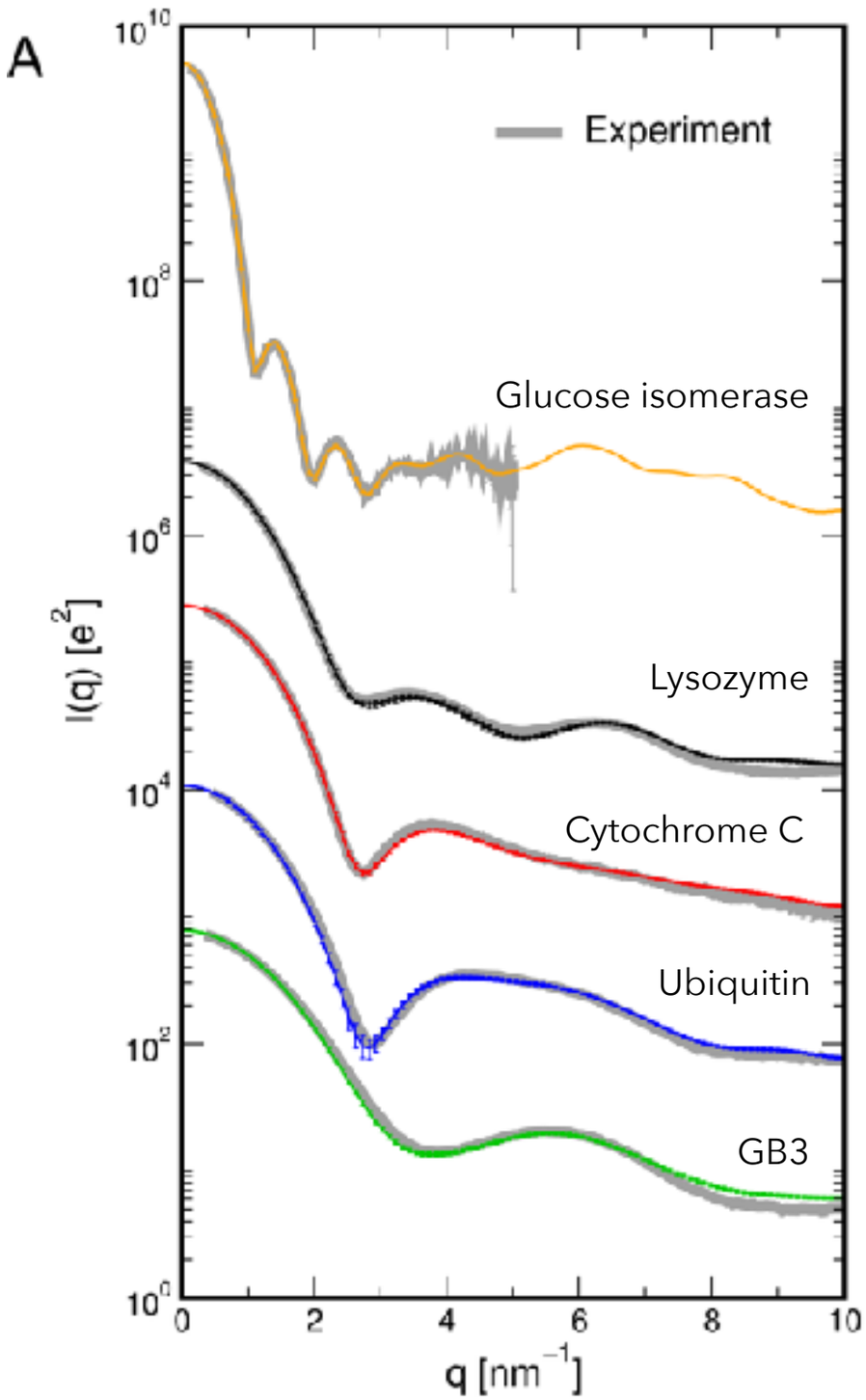


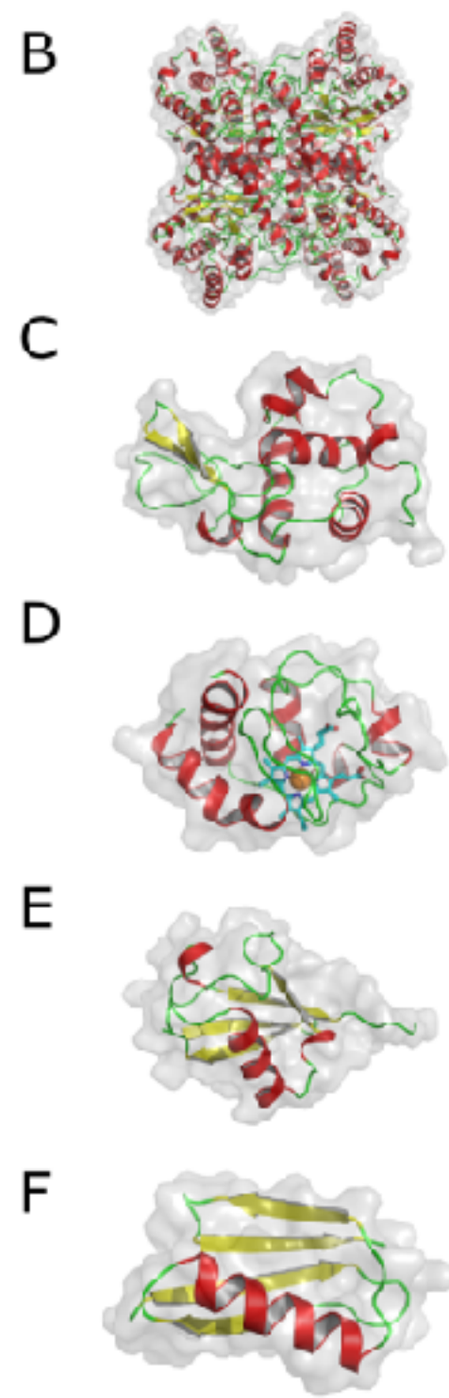


Uncertainty in buffer subtraction?!

Experimental data:
Grishaev *et al.*, *JACS* 2010
Lachlan Chasey, University of Queensland

Chen and Hub, *Biophys. J. 107:435-447* (2014)



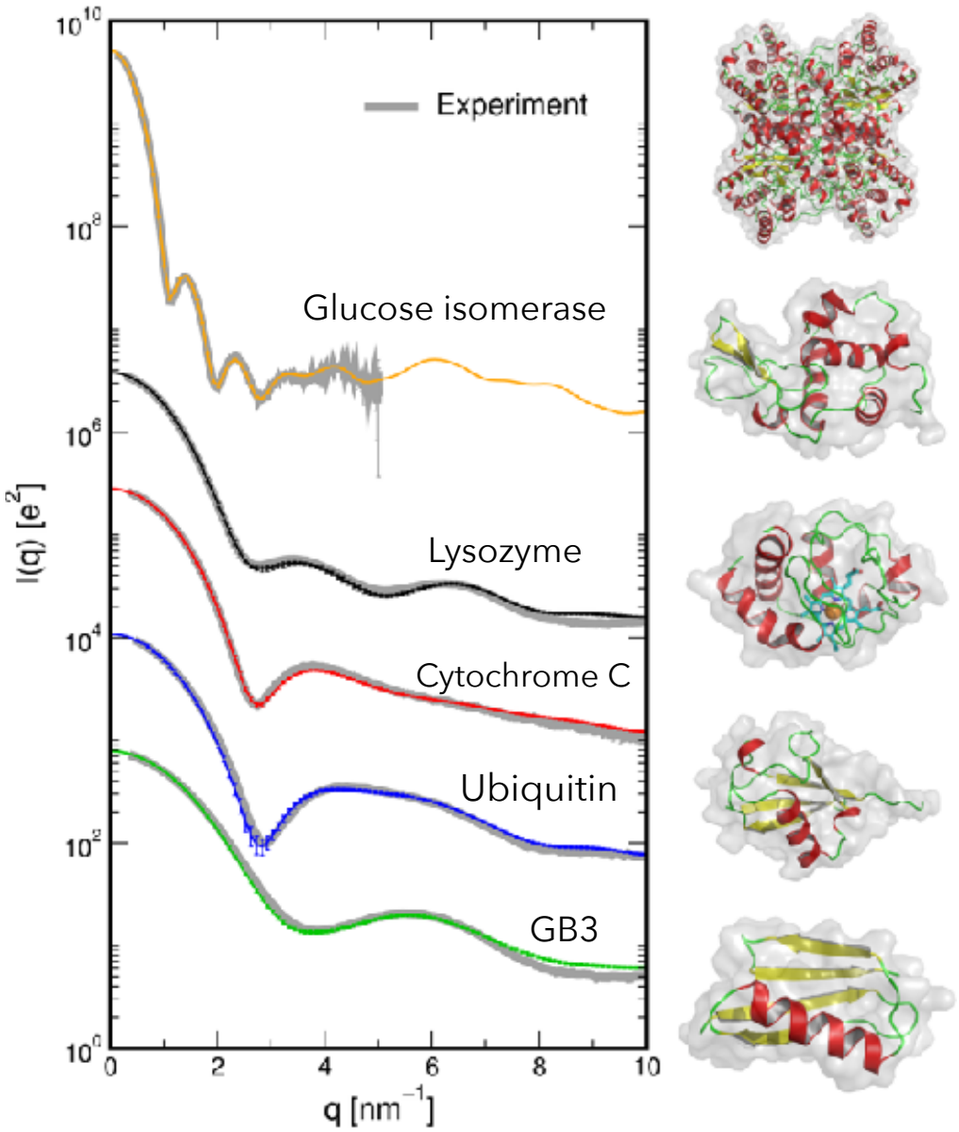


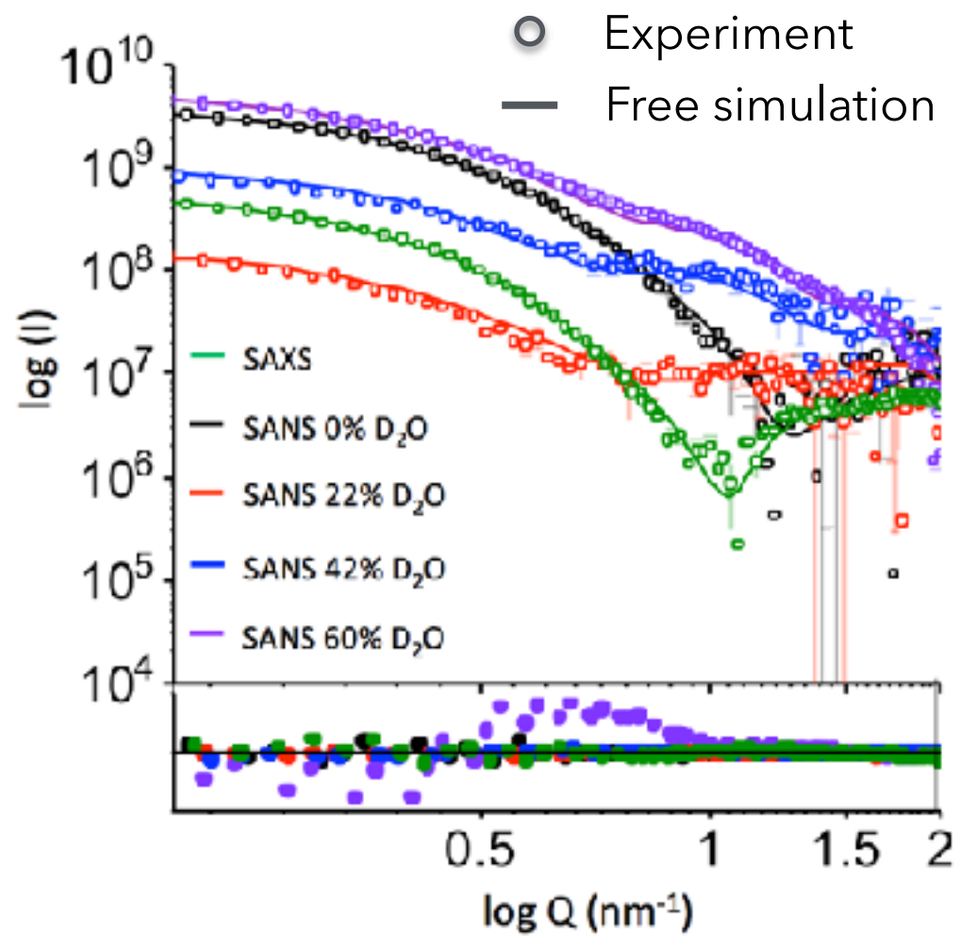
$$I_{\mathrm{fit}}(q) = f I_{\mathrm{exp}}(q) + c$$

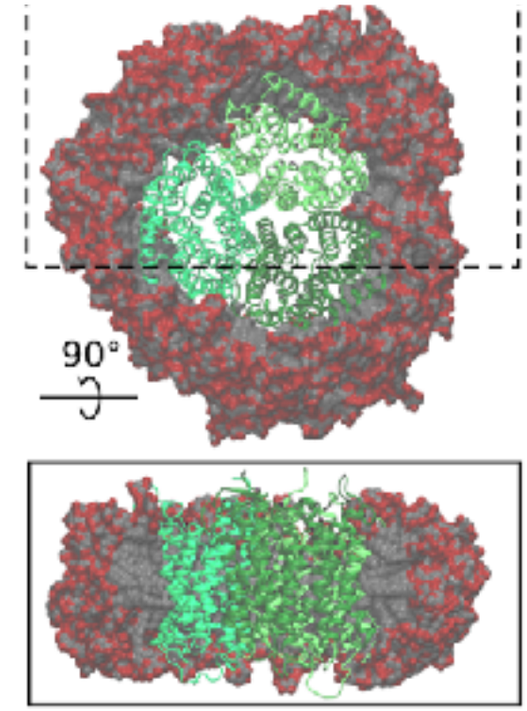
Experimental data:
Grishaev *et al.*, *JACS* 2010
Lachlan Chasey, University of Queensland

Chen and Hub, *Biophys. J. 107:435-447* (2014)

I(*q*) predictions using explicit solvent







Collaboration with:
Arnaud Javalle (Glasgow)
Frank Gabel (Grenoble)
Ulrich Zacharias (Dundee)

Experimental data:
Grishaev *et al.*, *JACS* 2010
Lachlan Chasey, University of Queensland

$$I_{\text{fit}}(q) = f I_{\text{exp}} + c$$

Absorb buffer mismatch / incoherent scattering

No fitting of:
- Hydration layer
- Excluded solvent

Chen and Hub, *Biophys J*, 2014
Dias Mirandela *et al.*, *J Phys Chem Lett*, 2018

WAXSiS

WAXS in Solvent (WAXSiS) computes small- and wide-angle X-ray scattering curves based on explicit-solvent all-atom molecular dynamics simulations.

Learn More

Home | Help | About | Contact | Links | Jobs in Queue: 0

Jobs can be submitted by entering a PDB ID, uploading a PDB file (max 20 MB), or uploading trajectory files. PDB files may have 300 to 40000 heavy atoms.

PDB ID | PDB File | Trajectory

Please select one of the above options.

Email Address (Optional)

Confirm Email Address

Job will be submitted using default options.

Review Options

Submit Job

Please visit at:

http://waxsis.uni-saarland.de

Offline use of WAXSiS ?

- Talk to us
- Buy Yasara, € 500 / 3 years

Computational Molecular Biophysics Group,
Institute for Microbiology and Genetics,
Georg-August University, Göttingen

YASARA GWDG Emmy Noether-Programm

Knight and Hub, *Nucleic Acids Res.* 43, W225-W230 (2015)

WAXS in Solvent - WAXSiS webserver

Basic Options

Ligands
- Keep ligands, try to remove crystallization agents
- Keep both ligands and crystallization agents
- Remove all

q Scattering
Specify the maximum *q* scattering vector (Å^{-1})
1.00

Buffer Subtraction
Select the buffer subtraction method:
- Buffer scattering reduced by solute volume
- Total buffer scattering subtracted

Experimental Curve - Optional
Fit an experimental SAXS / WAXS curve to the calculated curve.
Upload File No experimental curve
Units: Å^{-1} nm^{-1}
Scattering Convention: *q* *s*

Advanced Options

Output *q* Units
Select the output *q* units:
Å^{-1} nm^{-1}

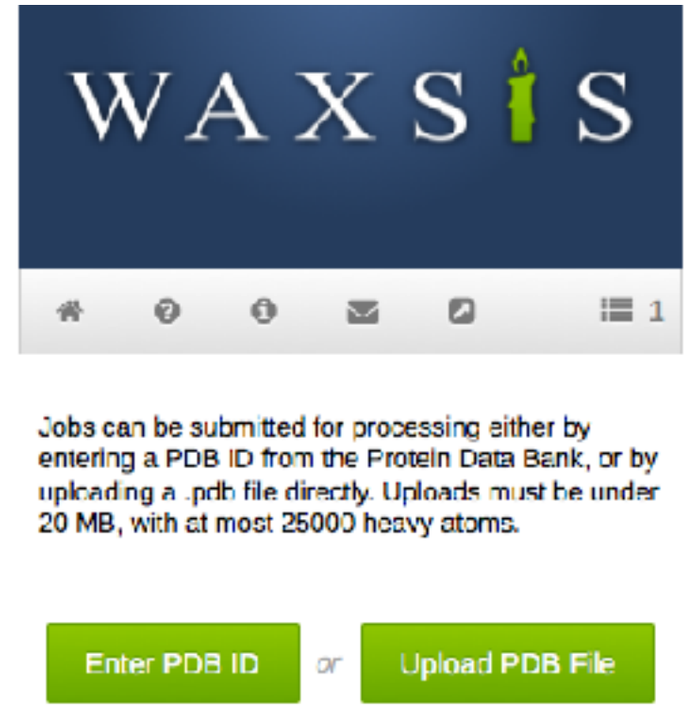
Solvent Density
Specify the solvent density (e/nm^{-3})
334

Selenomethionine
Replace selenomethionine with methionine?
- Yes
- No

Envelope Distance
Specify the envelope distance (Å)
7

Convergence
- Quick
- Normal
- Thorough

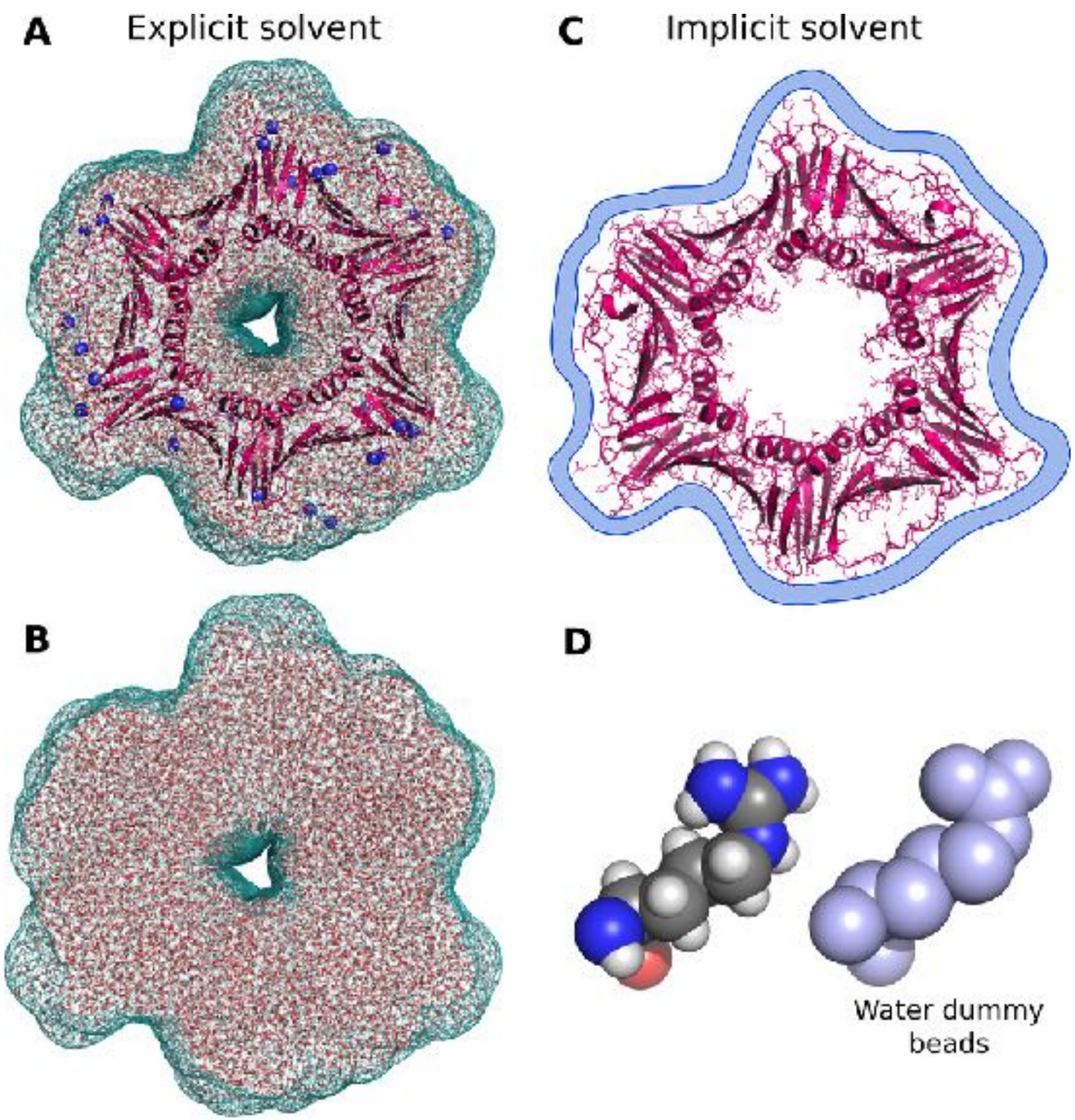
Smartphone / tablet



Feedback appreciated !!!

Knight and Hub, *Nucleic Acids Res.* 43, W225-W230 (2015)

Why explicit solvent?



Explicit solvent:

- Atomic representation of …
  - hydration layer and
  - excluded solvent
- Reproduces increase of R_g
- Works for inhomogenous biomolecules

Implicit-solvent methods

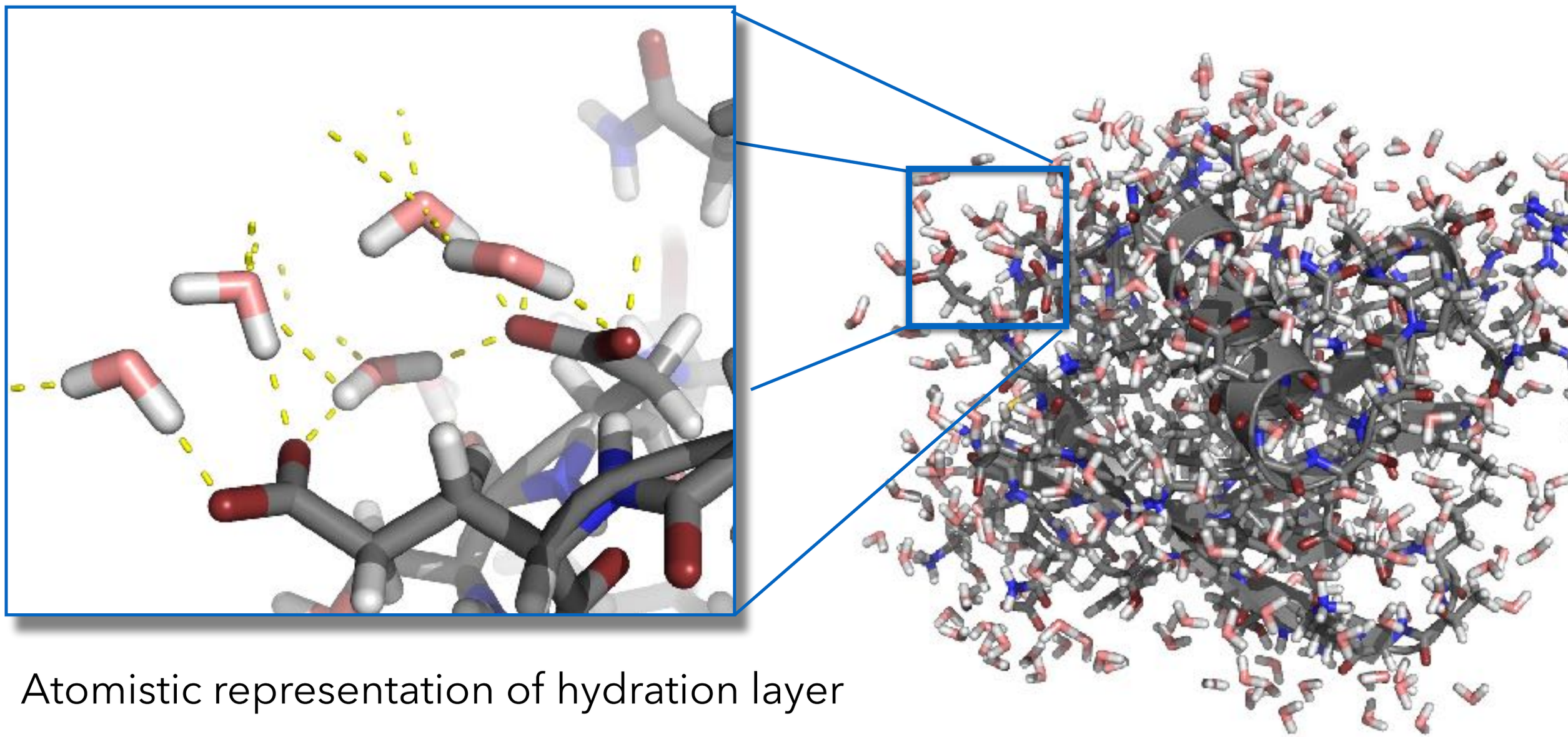
(CRYSOL, FoXS, SASTBX, Pepsi-SAXS,…)

- Continuum model of hydration layer
- Water dummy beads for excluded solvent, which volumes to use?
- Hydration layer and excluded volume are fitted
- Accuracy for inhomogenous solutes?

Chatzimagas & Hub, ArXiv (2022)

Water packing on the protein surface

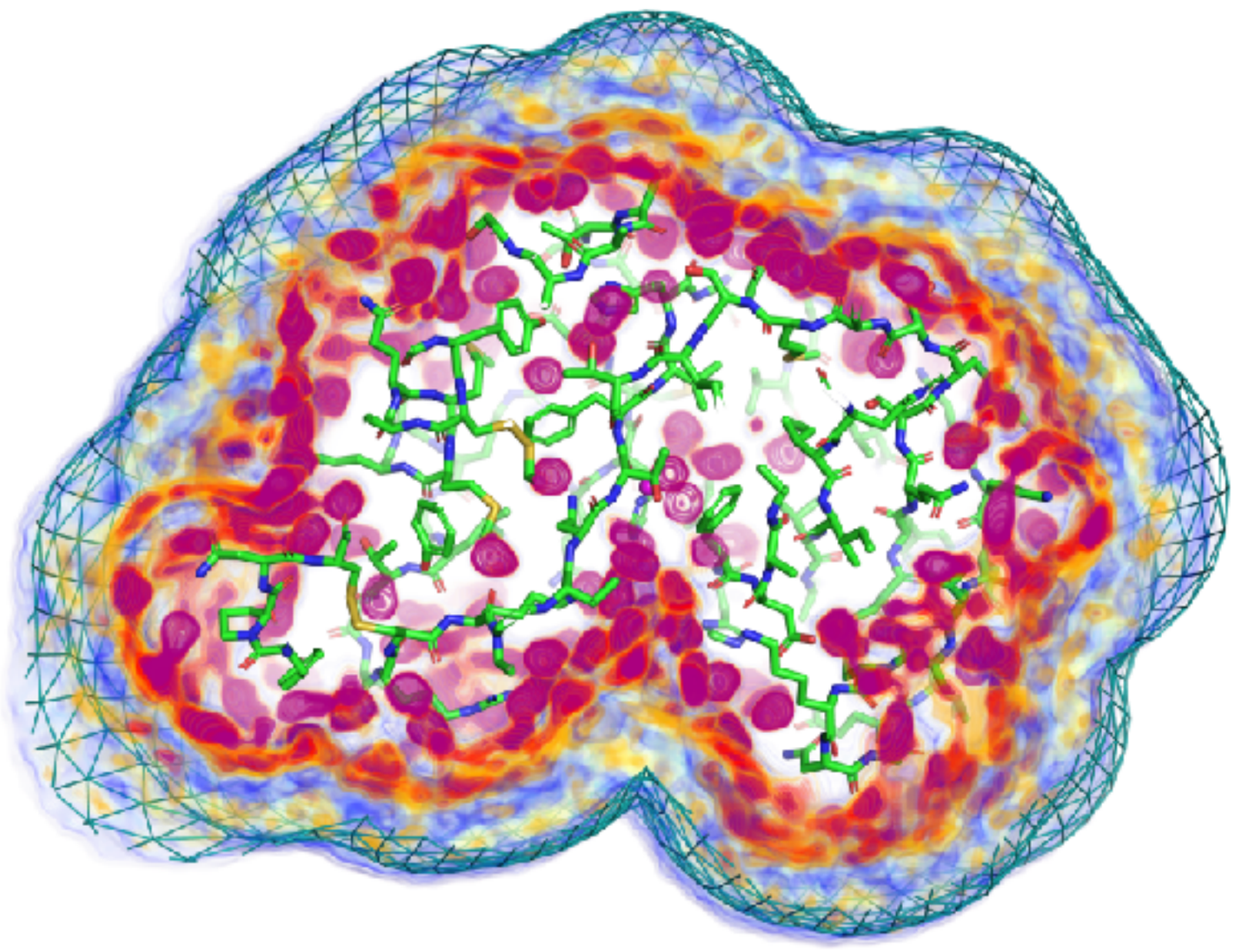
Hydration layer:
- typically more dense than bulk water
- Increases the apparent radius of gyration R_g



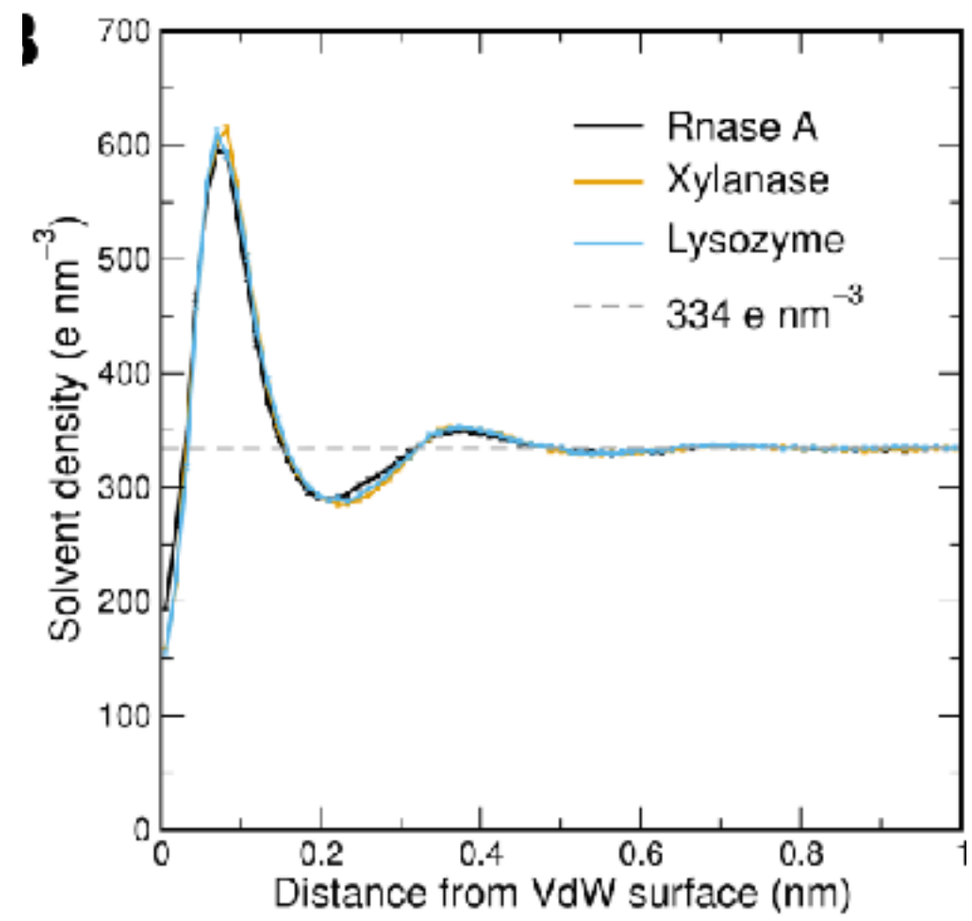
Atomistic representation of hydration layer

Why explicit solvent?

Hydration layer:
- typically more dense than bulk water
- Increases the apparent radius of gyration R_g



3D density water (at fixed protein atoms)



Chatzimagas & Hub, ArXiv (2022)

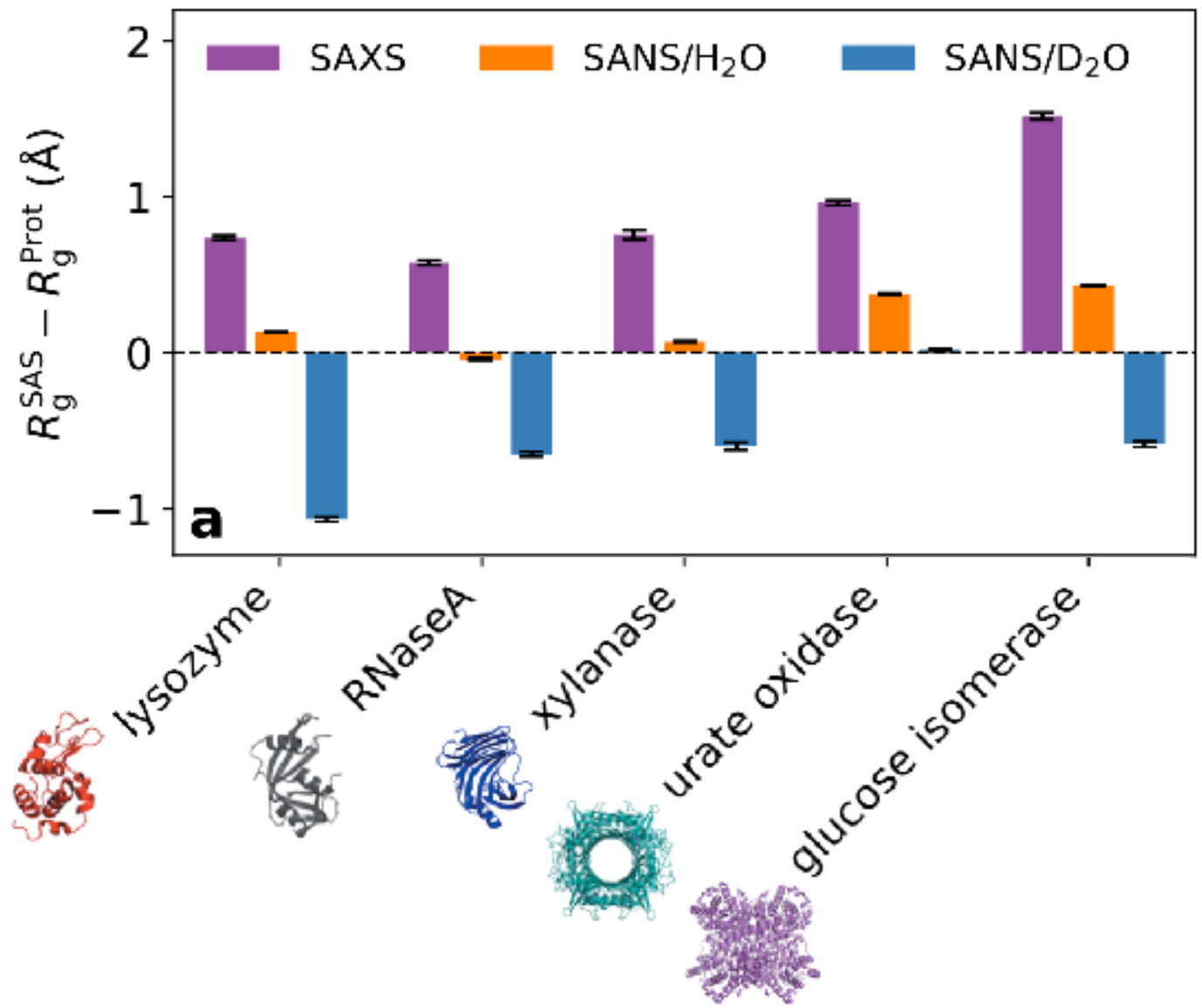
R_g increase due to hydration layer

$$\Delta R_g = R_g^{\mathrm{SAS}} - R_g^{\mathrm{Prot}}$$

- SAXS $\Delta R_g > 0$

  SANS^{D2O} $\Delta R_g < 0$

- ΔR_g depends on protein



Trewella, Vachette, et al., Acta Crystal D (2022)

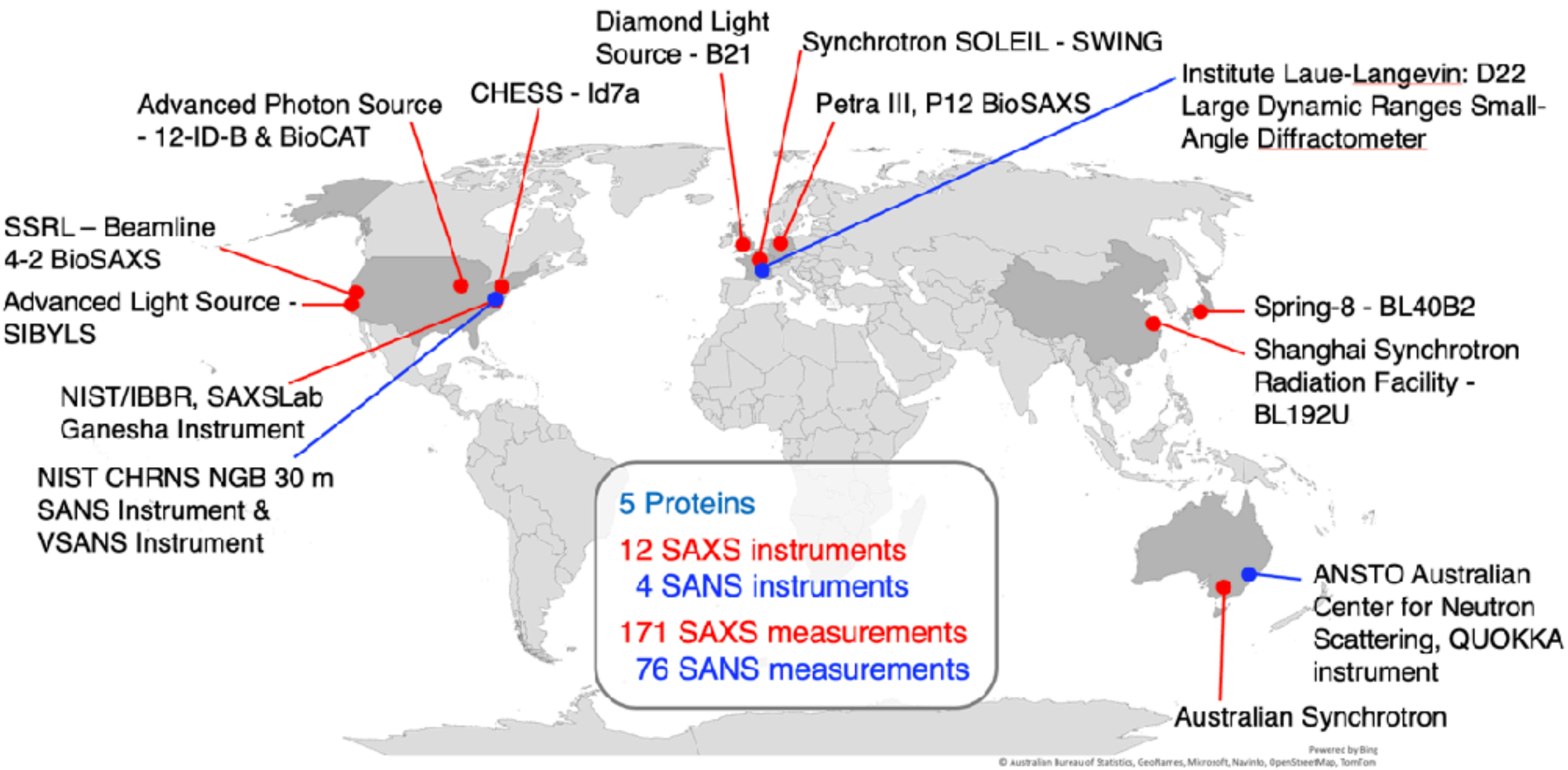
Linse and Hub, *Commun Chem* (2023)

A round-robin approach provides a detailed assessment of biomolecular small-angle scattering data reproducibility and yields consensus curves for benchmarking

Jill Trewhella,[a*] Patrice Vachette,[b*] Jan Bierma,[c] Clement Blanchet,[d] Emre Brookes,[e] Srinivas Chakravarthy,[f] Leonie Chatzimagas,[g] Thomas E. Cleveland IV,[h,i] Nathan Cowieson,[j] Ben Crosset,[k] Anthony P. Duff,[l] Daniel Franke,[d] Frank Gabel,[m] Richard E. Gillilan,[n] Melissa Graewert,[d] Alexander Grishaev,[o,p] J. Mitchell Guss,[q] Michal Hammel,[r] Jesse Hopkins,[s] Qingqui Huang,[t] Jochen S. Hub,[u] Greg L. Hura,[v] Thomas C. Irving,[w] Cy Michael Jeffries,[d] Cheol Jeong,[x] Nigel Kirby,[y] Susan Krueger,[z] Anne Martel,[a] Tsutomu Matsui,[b] Na Li,[c] Javier Pérez,[d] Lionel Porcar,[e] Thierry Prangé,[f] Ivan Rajkovic,[g] Mattia Rocco,[h] Daniel J. Rosenberg,[i] Timothy M. Ryan,[j] Soenke Seifert,[k] Hiroshi Sekiguchi,[l] Dmitri Svergun,[m] Susana Teixeira,[n] Aurelien Thureau,[o] Thomas M. Weiss,[p] Andrew E. Whitten,[q] Kathleen Wood[r] and Xiaobing Zuo[s]

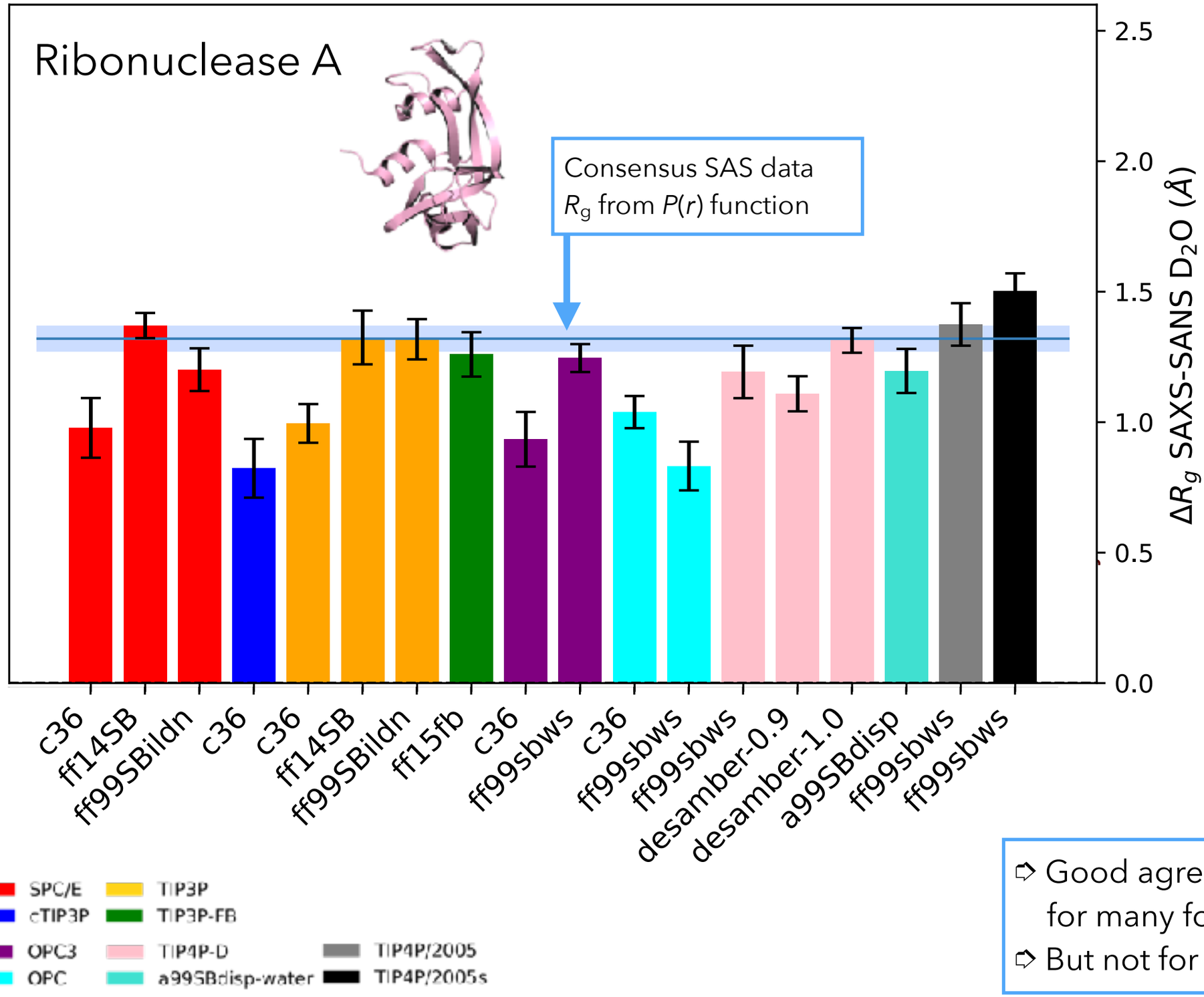
Worldwide community SAS benchmark

developed by Jill Trewhella and Patrice Vachette



Trewhella, Vachette, …, Chatzimagas, …, Hub, …, J Appl Cryst (2022)

Hydration shell validated against consensus SAXS/SANS data

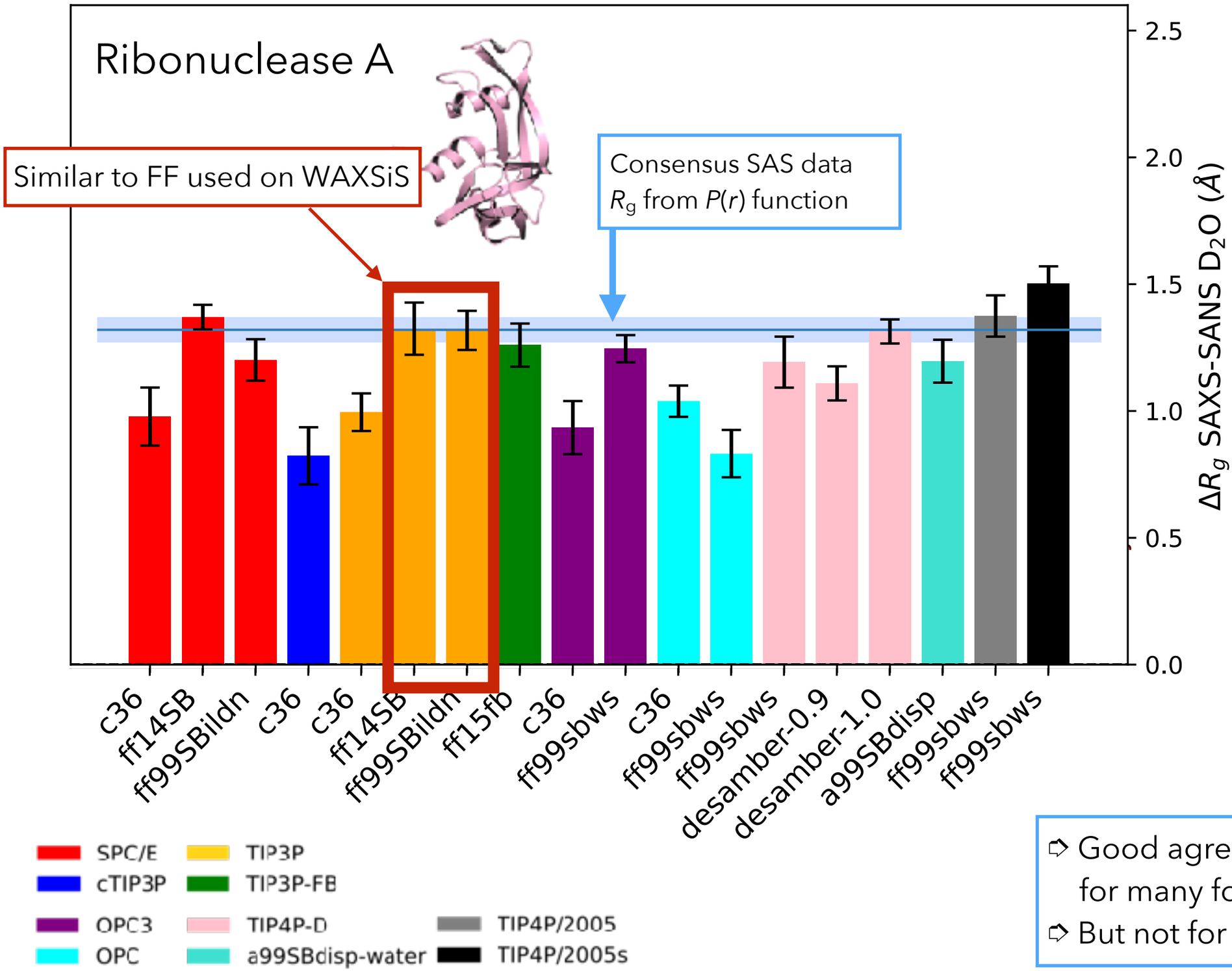


- ➪ Good agreement with experiment, for many force fields
- ➪ But not for all…

Trewella, Vachette, et al., Acta Crystal D (2022)

Linse and Hub, *Commun Chem* (2023)

Hydration shell validated against consensus SAXS/SANS data



- ➪ Good agreement with experiment, for many force fields
- ➪ But not for all…

Trewella, Vachette, et al., Acta Crystal D (2022)

Linse and Hub, *Commun Chem* (2023)

Where the volumes come from…

J. Appl. Cryst. (1978). **11**, 693–694

An Improved Method for Calculating the Contribution of Solvent to the X-ray Diffraction Pattern of Biological Molecules

By R. D. B. Fraser, T. P. MacRae and E. Suzuki

Division of Protein Chemistry, CSIRO, Parkville (Melbourne), Victoria 3052, Australia

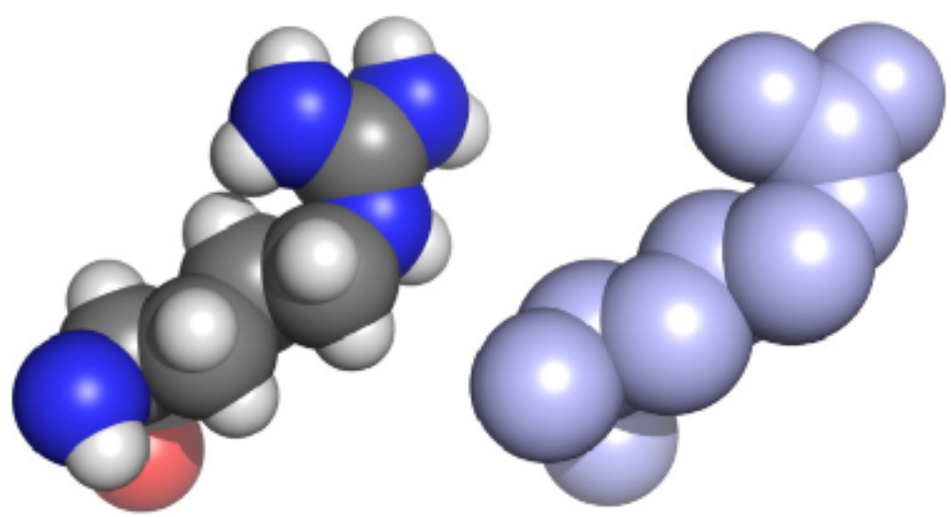
Fraser *et al.*, J Appl Cryst (1978)

	Calculated volume			Observed volume
	(*a*) (Å^3)	(*b*) (Å^3)	(*c*) (Å^3)	(*d*) (Å^3)
H	—	7·24	7·24	5·15
C	17·16	20·58	20·58	16·44
N	10·31	14·14	15·60	2·49
O	11·49	11·49	14·71	9·13

References: (*a*) Okuyama *et al.* (1976); (*b*) Arnott & Hukins (1973); (*c*) Bondi (1964); (*d*) Traube (1899), Zamyatnin (1972).

Highly cited in the SAXS field

Needed by implicit-solvent methods



Water dummy beads

Where the volumes come from…

J. Appl. Cryst. (1978). **11**, 693–694

An Improved Method for Calculating the Contribution of Solvent to the X-ray Diffraction Pattern of Biological Molecules

By R. D. B. Fraser, T. P. MacRae and E. Suzuki

Division of Protein Chemistry, CSIRO, Parkville (Melbourne), Victoria 3052, Australia

Fraser *et al.*,J Appl Cryst (1978)

	Calculated volume			Observed volume
	(*a*) (Å^3)	(*b*) (Å^3)	(*c*) (Å^3)	(*d*) (Å^3)
H	—	7·24	7·24	5·15
C	17·16	20·58	20·58	16·44
N	10·31	14·14	15·60	2·49
O	11·49	11·49	14·71	9·13

References: (*a*) Okuyama *et al.* (1976); (*b*) Arnott & Hukins (1973); (*c*) Bondi (1964); (*d*) Traube (1899), Zamyatnin (1972).

Zamyatnin (1972).

PROTEIN VOLUME IN SOLUTION

A. A. Zamyatnin

Institute of Biophysics of the USSR Academy of Sciences, Pushchino-on-Oka, Moscow Region, USSR

Zamyatnin, Prog. Phys. Mol. Biol. (1972)

Table 4. The Values of the Molar Volumes of the Main Atomic Groups Composing Proteins (Cohn and Edsall, 1943a)

Atomic group	—NH_2	—CH_2	—COOH	—CONH	—OH
The volume of group (ml/mole)	7.7	16.3[a]	18.9	20.0	5.4

Where the volumes come from…

J. Appl. Cryst. (1978). **11**, 693–694

An Improved Method for Calculating the Contribution of Solvent to the X-ray Diffraction Pattern of Biological Molecules

By R. D. B. Fraser, T. P. MacRae and E. Suzuki

Division of Protein Chemistry, CSIRO, Parkville (Melbourne), Victoria 3052, Australia

Fraser *et al.*,J Appl Cryst (1978)

	Calculated volume			Observed volume
	(*a*) (Å^3)	(*b*) (Å^3)	(*c*) (Å^3)	(*d*) (Å^3)
H	—	7·24	7·24	5·15
C	17·16	20·58	20·58	16·44
N	10·31	14·14	15·60	2
O	11·49	11·49	14·71	9

References: (*a*) Okuyama *et al.* (1976); (*b*) Arnott & Hukins (1973); (*c*) Bondi (1964); (*d*) Traube (1899), Zamyatnin (1972).

Traube (1899).

Traube, J. (1899). Quoted by Partington (1951).

Where the volumes come from…

J. Appl. Cryst. (1978). **11**, 693–694

An Improved Method for Calculating the Contribution of Solvent to the X-ray Diffraction Pattern of Biological Molecules

By R. D. B. Fraser, T. P. MacRae and E. Suzuki

Division of Protein Chemistry, CSIRO, Parkville (Melbourne), Victoria 3052, Australia

Fraser *et al.*,J Appl Cryst (1978)

	Calculated volume			Observed volume
	(*a*) (Å^3)	(*b*) (Å^3)	(*c*) (Å^3)	(*d*) (Å^3)
H	—	7·24	7·24	5·15
C	17·16	20·58	20·58	16·44
N	10·31	14·14	15·60	2[illegible]
O	11·49	11·49	14·71	9[illegible]

References: (*a*) Okuyama *et al.* (1976); (*b*) Arnott & Hukins (1973); (*c*) Bondi (1964); (*d*) Traube (1899), Zamyatnin (1972).

Traube (1899).

Traube, J. (1899). Quoted by Partington (1951).

537. J. Traube: Ueber das Molekularvolumen.

[9. Abhandlung.]

(Eingegangen am 29. October.)

Traube, Berichte der deutschen chemischen Gesellschaft, 2722-2728 (**1895**)

Die folgende Tabelle enthält die von mir gefundenen Volumconstanten:

		ccm
Molekularcontraction in Wasser	bei 15°	13.5
Molekulare Dilatationsconstante	bei 15°	12.4
Kohlenstoff		9.9
Dreiwerthiger Stickstoff (Amine, Imide, Ringe)		1.5
Fünfwerthiger Stickstoff (Ammonium, Ringammonium)		ca. 10.7
Stickstoff in Nitroverbindungen		ca. 8.5—10.7

Where the volumes come from…

J. Appl. Cryst. (1978). **11**, 693–694

An Improved Method for Calculating the Contribution of Solvent to the X-ray Diffraction Pattern of Biological Molecules

By R. D. B. Fraser, T. P. MacRae and E. Suzuki

Division of Protein Chemistry, CSIRO, Parkville (Melbourne), Victoria 3052, Australia

Fraser *et al.*,J Appl Cryst (1978)

	Calculated volume (*a*) (Å^3)	(*b*) (Å^3)	(*c*) (Å^3)	Observed volume (*d*) (Å^3)
H	—	7·24	7·24	5·15
C	17·16	20·58	20·58	16·44
N	10·31	14·14	15·60	2·49
O	11·49	11·49	14·71	9·13

References: (*a*) Okuyama *et al.* (1976); (*b*) Arnott & Hukins (1973); (*c*) Bondi (1964); (*d*) Traube (1899), Zamyatnin (1972).

16·44
2·49

537. J. Traube: Ueber das Molekularvolumen.

[9. Abhandlung.]

(Eingegangen am 29. October.)

Traube, Berichte der deutschen chemischen Gesellschaft, 2722-2728 (**1895**)

Die folgende Tabelle enthält die von mir gefundenen Volumconstanten:

	ccm
Molekularcontraction in Wasser } bei 15°	13.5
Molekulare Dilatationsconstante } bei 15°	12.4
Kohlenstoff	9.9
Dreiwerthiger Stickstoff (Amine, Imide, Ringe)	1.5
Fünfwerthiger Stickstoff (Ammonium, Ringammonium)	ca. 10.7
Stickstoff in Nitroverbindungen	ca. 8.5—10.7

Carbon / Kohlenstoff

9.9 ccm/mol = 16.44 Å^3

3-valued nitrogen

1.5 ccm/mol = 2.49 Å^3

Where the volumes come from…

from Voronoi tesselation of high-res crystal structure cores

Atomic group	Pontius *et al.* [Å^3]	Fraser *et al.*, 1978 [Å^3]	Traube, 1895 [ccm/mol]
H		5.15	3.1
C		16.44	9.9
N	8.8 (0.8)	2.49	1.5 (trivalent)
			10.7 (pentavalent)
			8.5–10.7 (in nitro compound)
O	22.3 (0.4)	9.13	5.5 (carbonyl oxygen)
			2.3 or 0.4 (hydroxy oxygen)
OH	23.9 (0.9)	14.28	
NH	14.1 (0.3)	7.64	
CH	11.8 (0.6)	21.59	
CH_2	20.9 (1.8)	26.74	
CH_3	33.9 (1.2)	31.89	

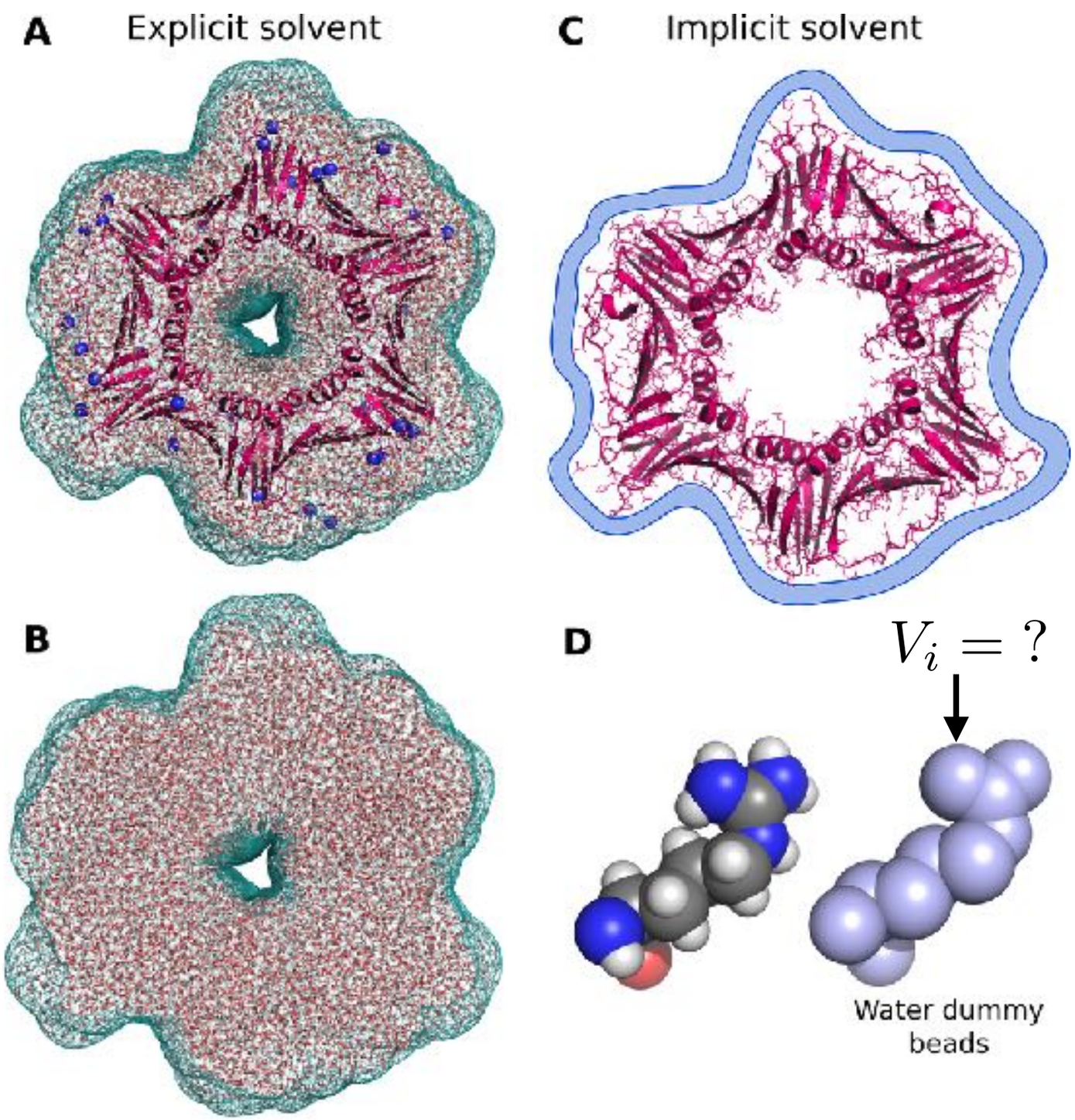
Cancellation of errors for fixed C-to-N ratio?

	Pointius	Fraser/Traube	Experiment
Hexadecane volume (Å^3)	360.4	438.1	486.6

Pontius et al., JMB (1996)

Chatzimagas & Hub, *Methods Enzymol* (2022)

Why explicit solvent?



Explicit solvent:

- Atomic representation of …
  - hydration layer and
  - excluded solvent
- Reproduces increase of R_g
- Works for inhomogeneous biomolecules

Implicit-solvent methods

(CRYSOL, FoXS, SASTBX, Pepsi-SAXS,…)

- Continuum model of hydration layer
- Water dummy beads for excluded solvent, which volumes to use?
- Hydration layer and excluded volume are fitted
- Accuracy for inhomogenous solutes?

Chatzimagas & Hub, *Methods Enzymol* (2022)

(Incomplete list of) SAXS prediction methods

ID	Name/authors	Year	$\delta\rho_{\mathrm{fit}}/f_{\mathrm{red}}$	Resol.	Fluct.	Avail.	Refs.
Implicit solvent methods:							
1	CRYSOL	1995	yes/yes	atom.	-	D/W	[25]
2	ORNL-SAS	2007	yes/yes	atom.	-	D	[64]
3	SoftWAXS	2009	yes/-	atom.	-	-	[65]
4	Fast-SAXS-pro	2009	yes/yes	CG	yes	D/W	[30, 36]
5	FoXS	2010	yes/yes	atom.	-	D/W	[66, 29]
6	PHAISTOS	2010	yes/yes	CG	-	D	[67]
7	AquaSAXS/AquaSol	2011	yes/yes	atom.	-	W	[27]
8	SASbtx/Zernike	2012	yes/-	atom.	-	W	[68]
9	Nguyen et al./RISM	2014	-/yes	atom.	-	D	[69]
10	BCL::SAXS	2015	yes/yes	atom.	-	D	[70]
11	Pepsi-SAXS	2017	yes/yes	atom.	-	D	[71]
Explicit solvent methods:							
12	SASSIM/Sassena	2002	-/yes	atom.	yes	D	[72]
13	MD-SAXS	2009	-/-	atom.	yes	-	[73, 74]
14	AXES	2010	yes/-	atom.	-	W	[26]
15	Park et al.	2009	-/-	atom.	-	-	[75]
16	Köfinger & Hummer	2013	-/-	atom.	yes	D	[76]
17	WAXSiS	2014	-/-	atom.	yes	W	[38, 77]

W = Webserver
D = Download

Hub, *Curr Opin Struct Biol*, 2018

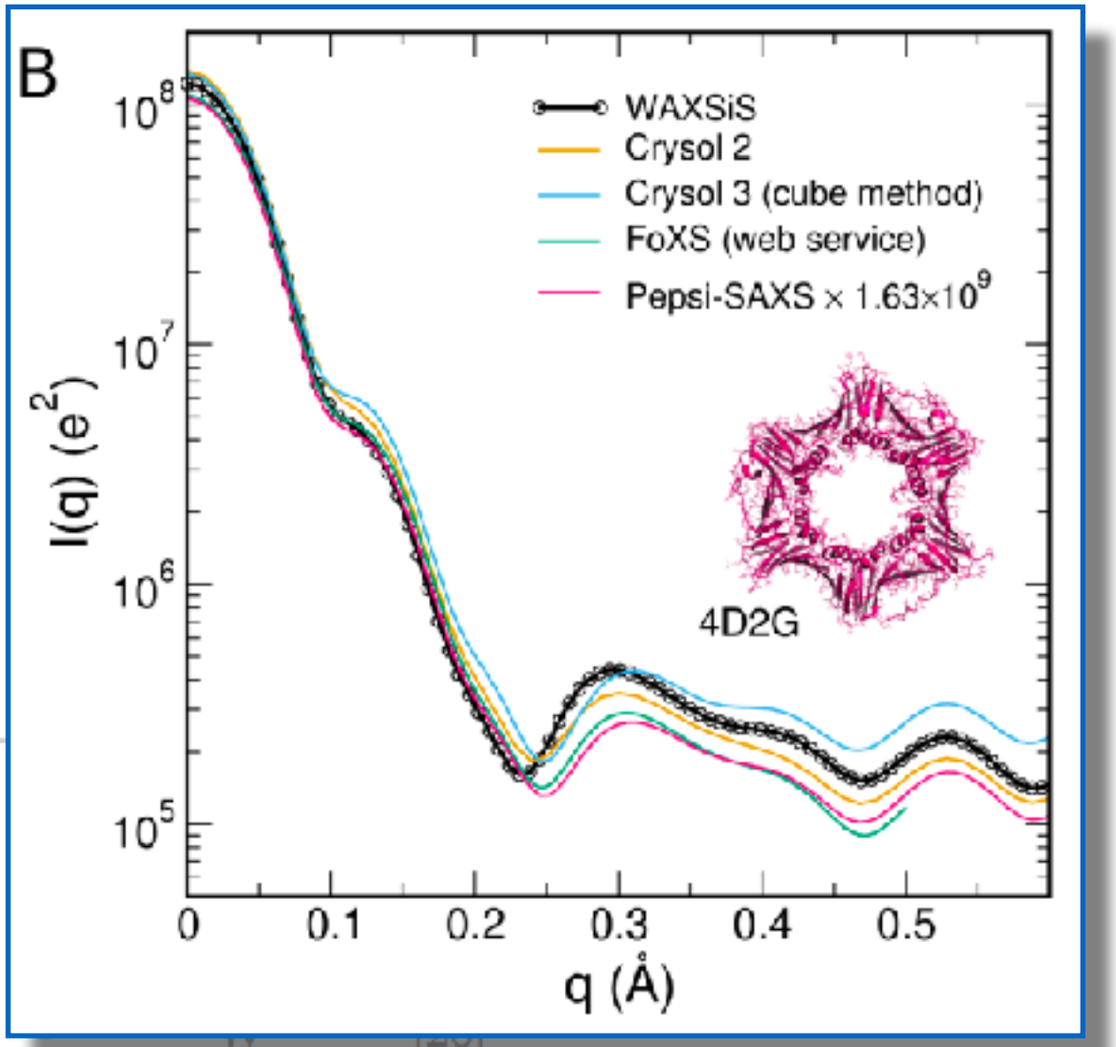
SAXS prediction methods

ID	Name/authors	Year	$\delta\rho_{\mathrm{fit}}/f_{\mathrm{red}}$	Resol.	Fluct.	Avail.	Refs.
Implicit solvent methods:							
1	CRYSOL	1995	yes/yes	atom.	-	D/W	[25]
					-	D	[64]
15	Park et al.	2009	-/-	atom.	-	-	[75]
16	Köfinger & Hummer	2013	-/-	atom.	yes	D	[76]
17	WAXSiS	2014	-/-	atom.	yes	W	[38, 77]

W = Webserver
D = Download

Bottom line:

1) Methods yield different results.
2) Accurate predictions of SAXS curves from structural models is a matter of ongoing research.
3) Strengths and weaknesses of available methods?
4) Generally accepted SEC-SAXS benchmark suites required.



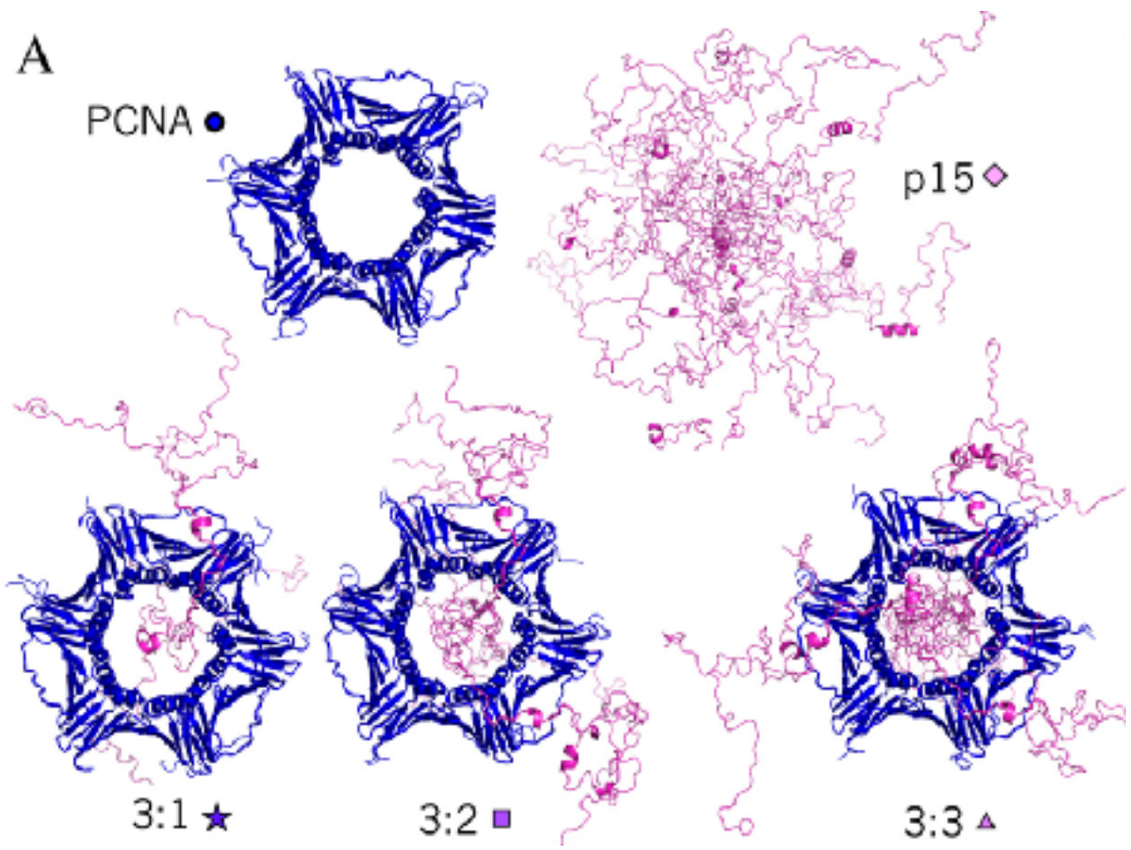
Hub, Curr Opin Struct Biol, *in revision*

WAXSiS application to PCNA

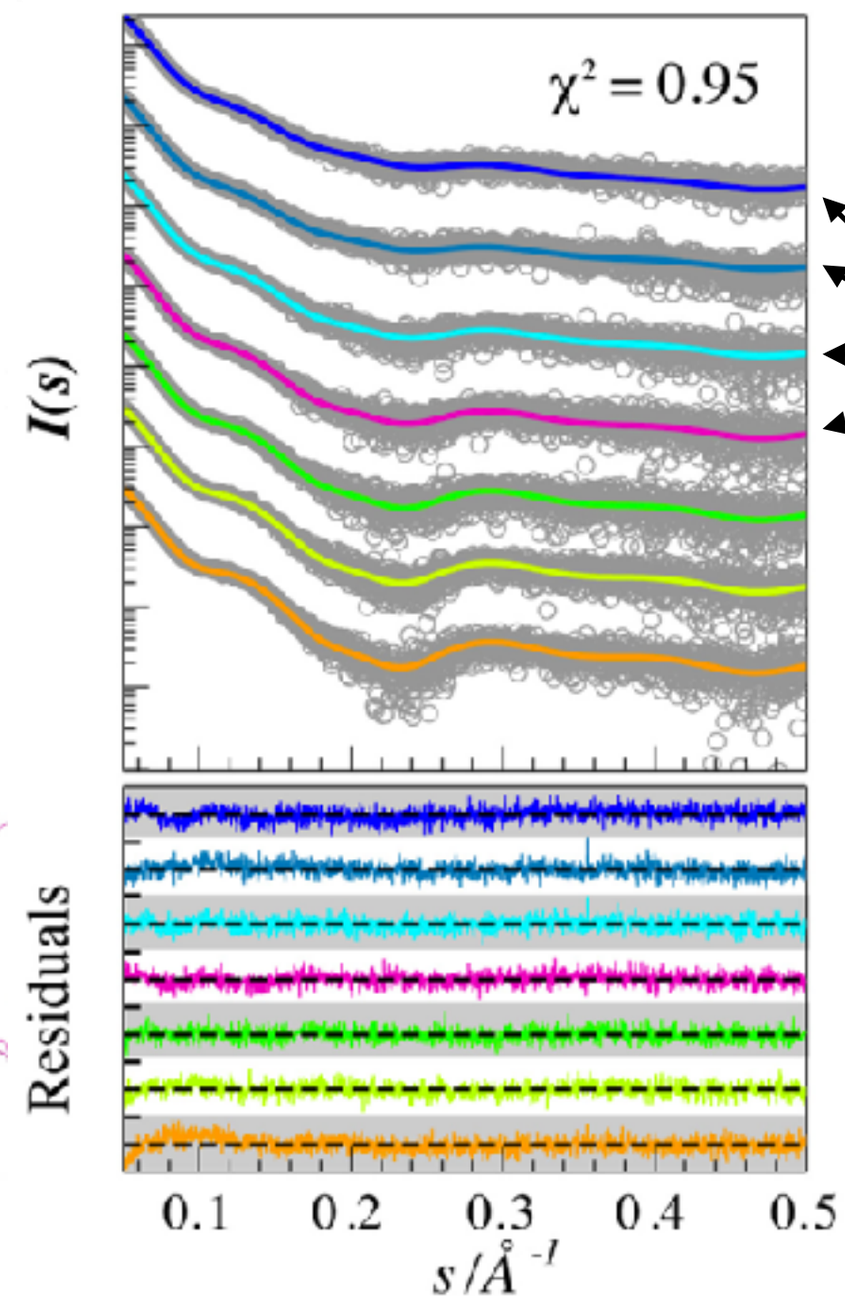
Collaboration with
Pau Bernado

PCNA / p15 complexes:

- Transient
- Disordered
- Multivalent



Only 1 fitting parameter: $K_d \approx 30$ μM

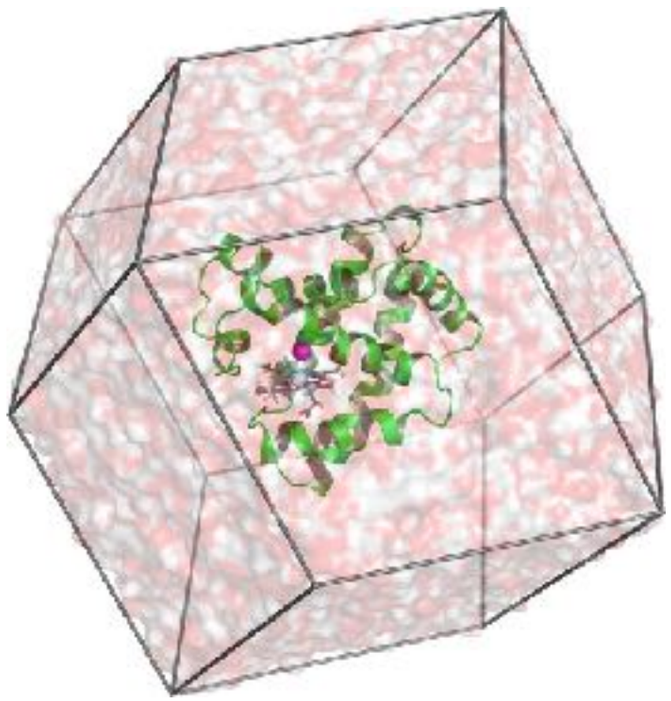


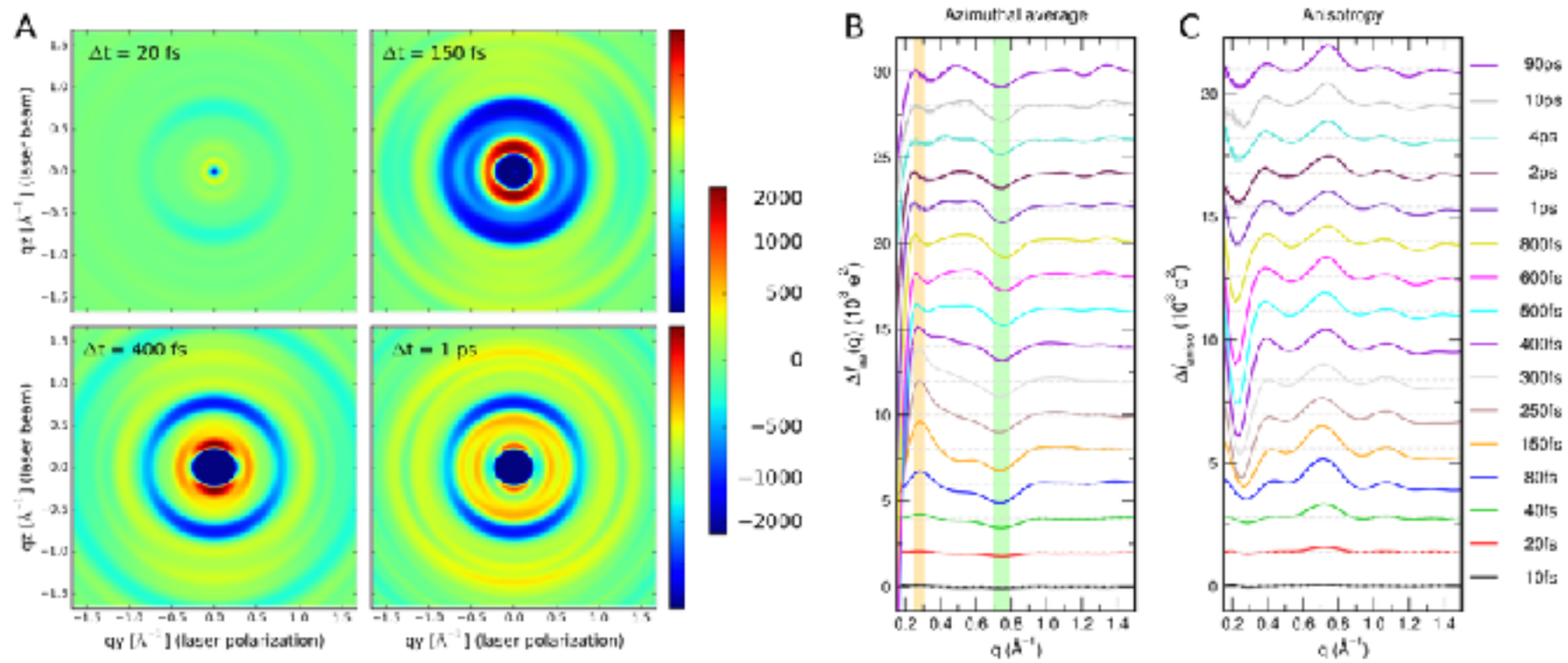
p15 titration points

Cordeiro et al., *Nucleic Acids Res* (2016)

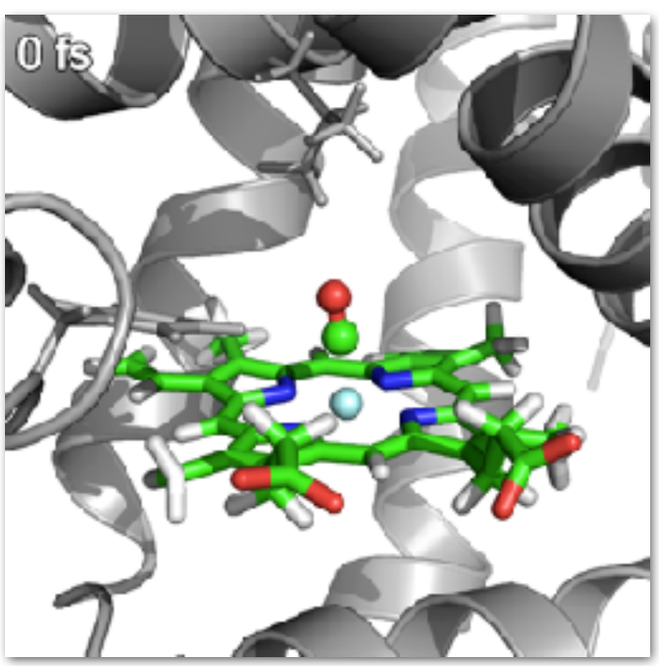
Glimpse on: time-resolved SAXS

Myoglobin simulation system



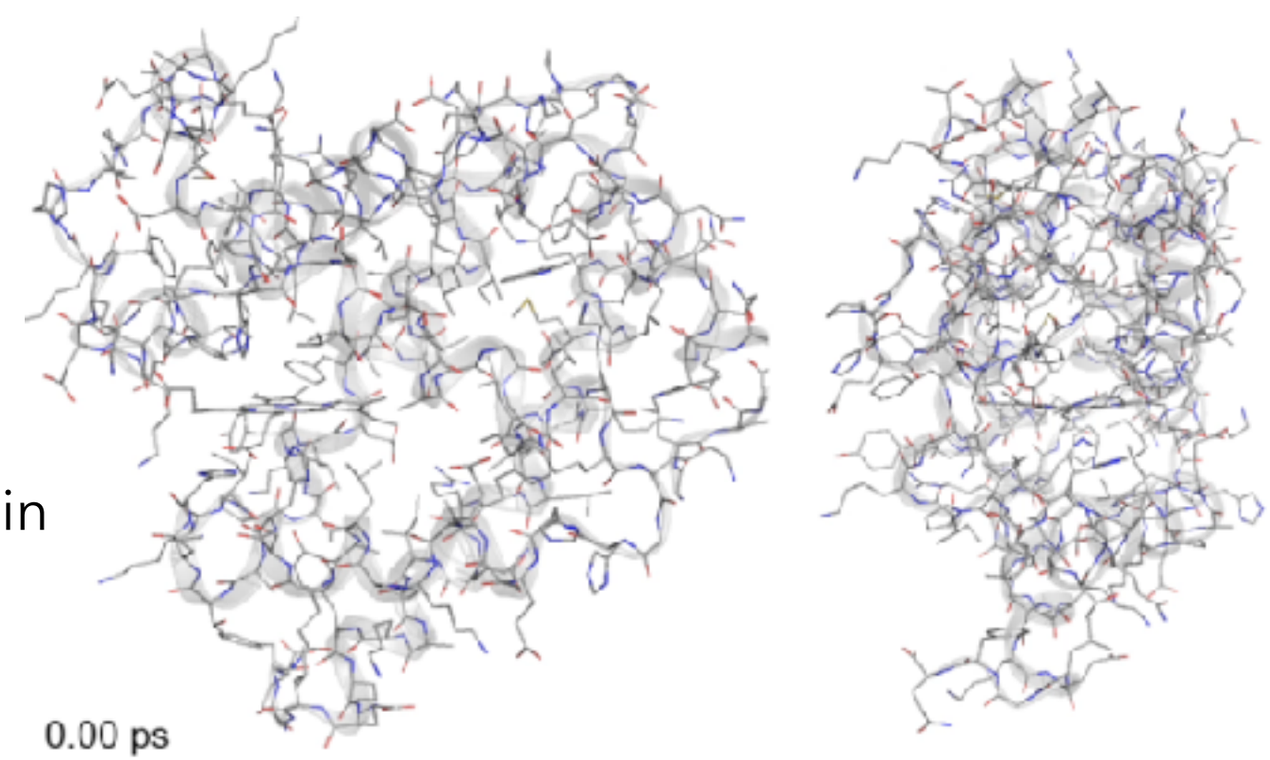


Computed time-resolved SAXS/WAXS patterns



CO dissociation in myoglobin

Protein quake in myoglobin



Brinkmann and Hub, *PNAS* (2016)

Large Aap fibrils

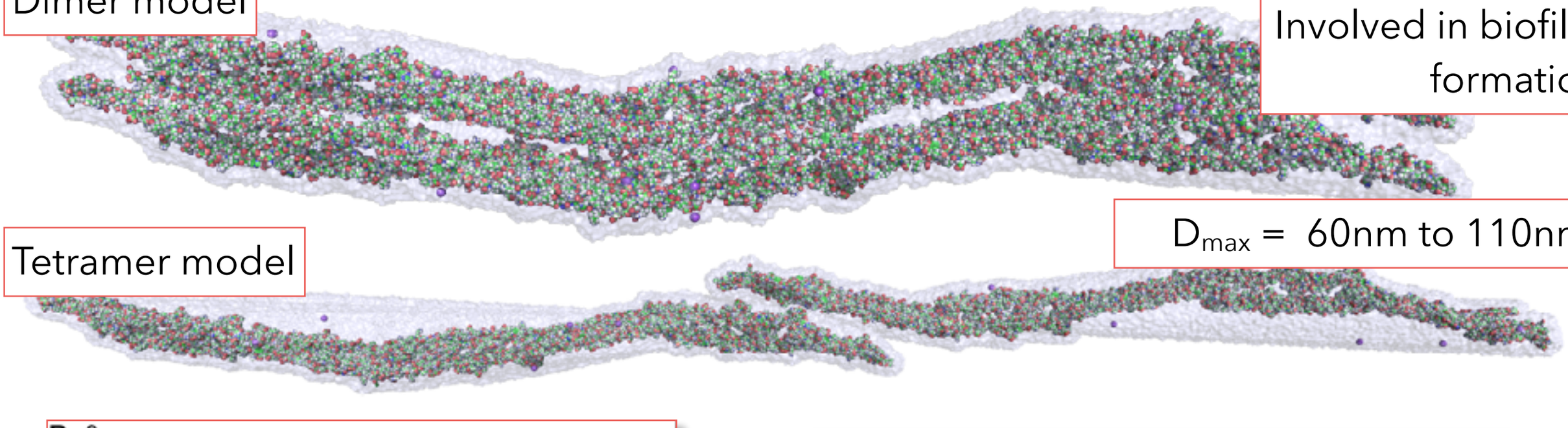
Collaboration with Herr lab (Cincinnati)

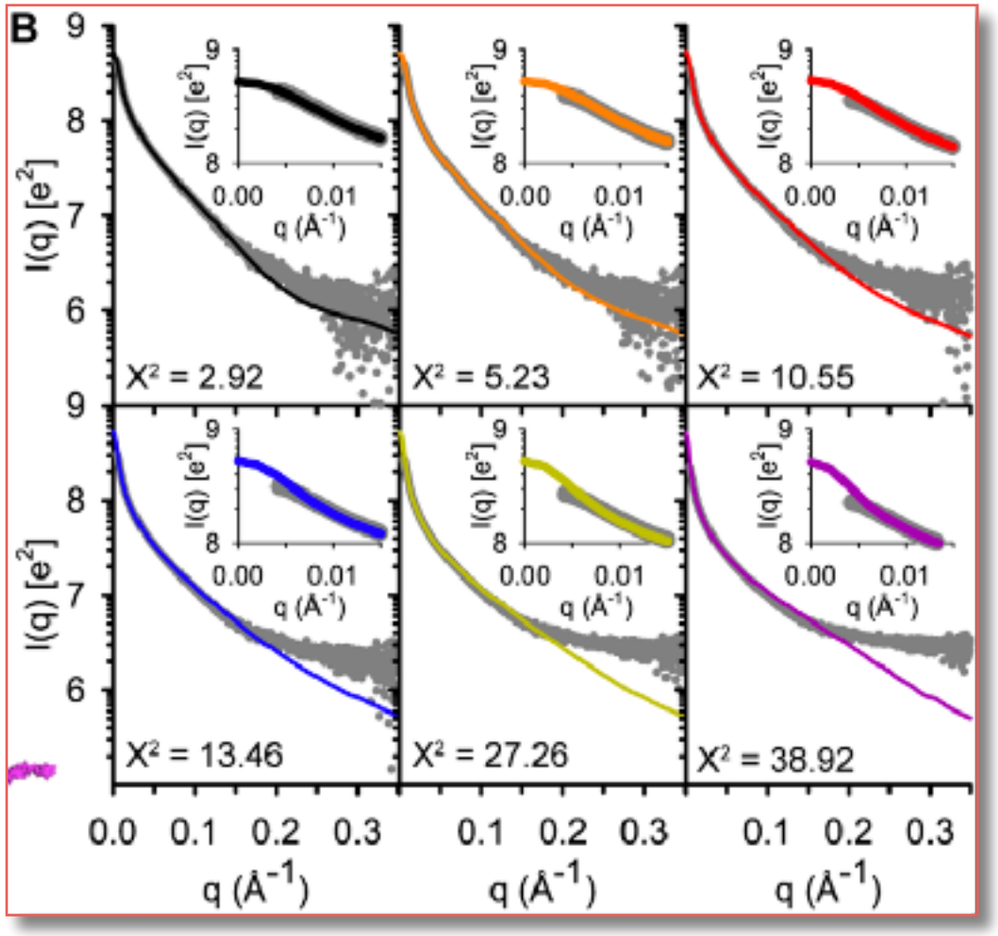
Dimer model

Involved in biofilm formation

Tetramer model

D_{max} = 60nm to 110nm





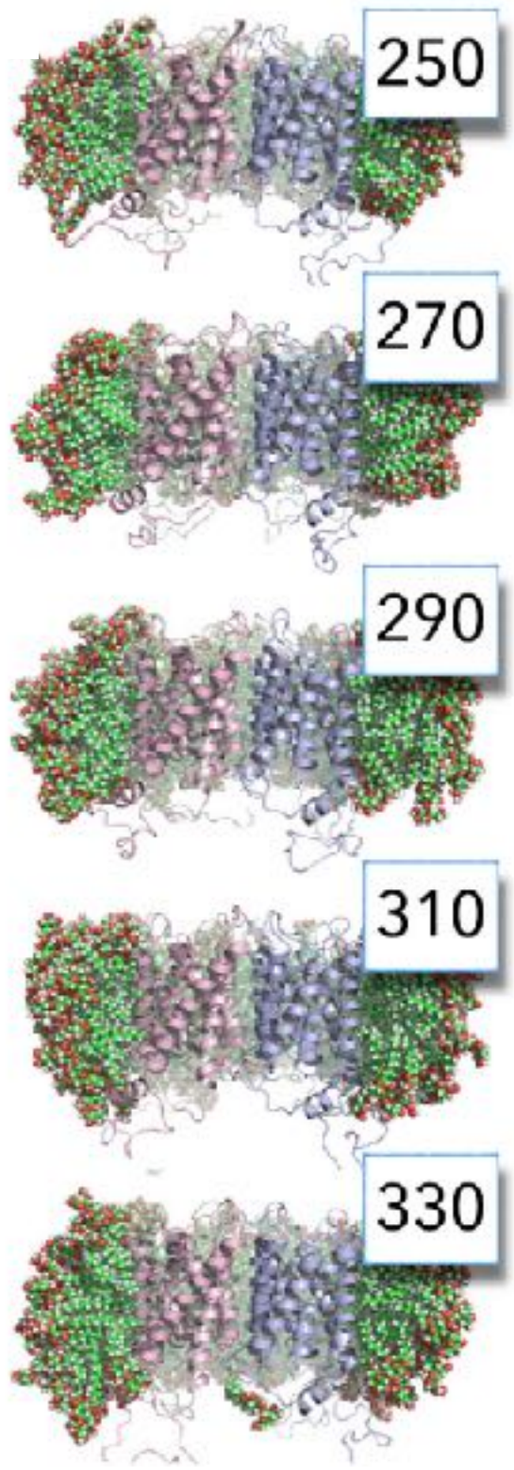
	WAXSiS		
Model	**Chi²**	**R_g (Å)**	**R_g (anhydrous)**
Exp data	-	167*	-
Dimer_0	2.92	162.7	170.5
Dimer_1	5.23	172.6	183.8
Dimer_2	10.55	200.9	209.1

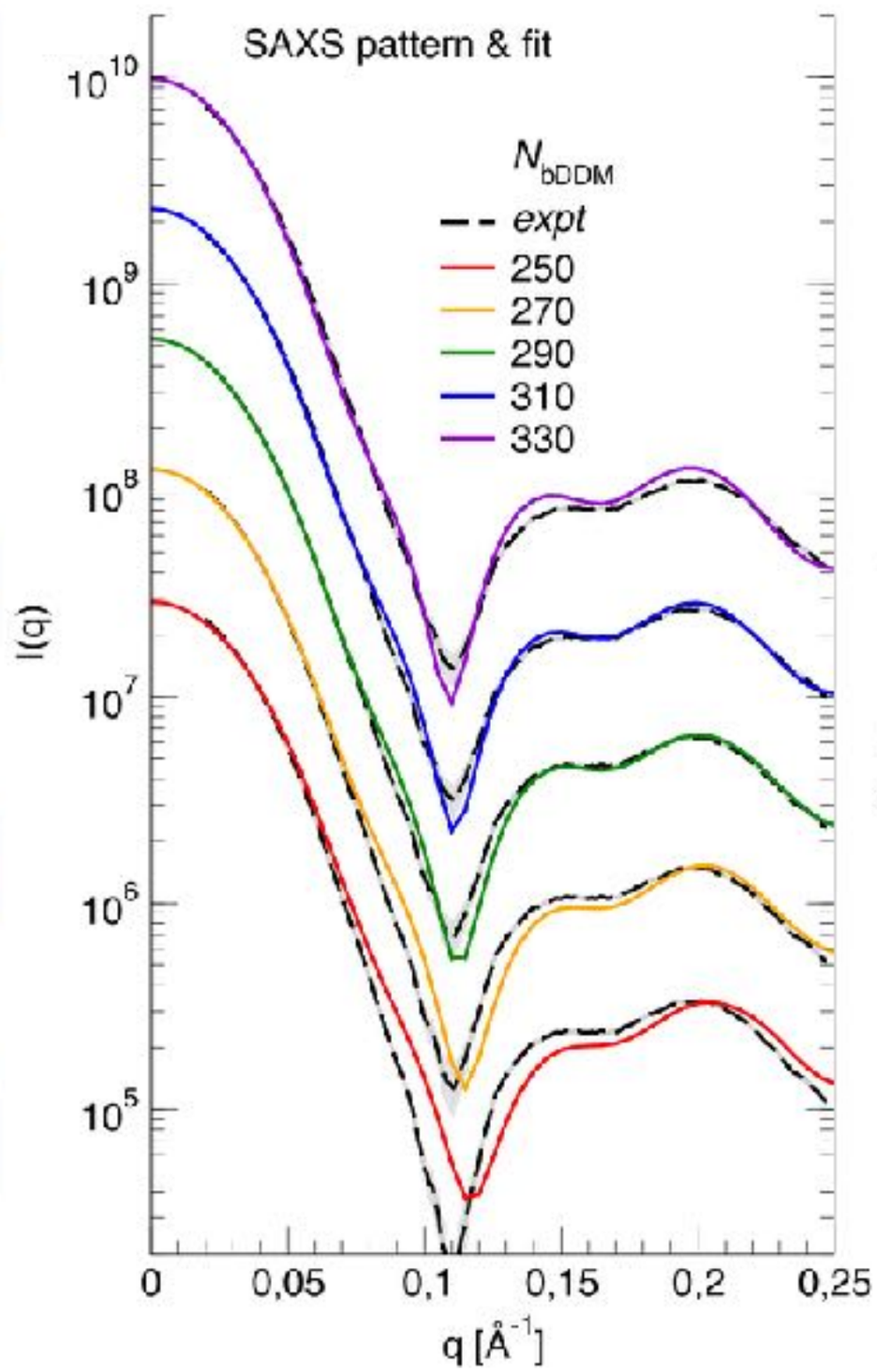
Negative effect of hydration layer on! R_g

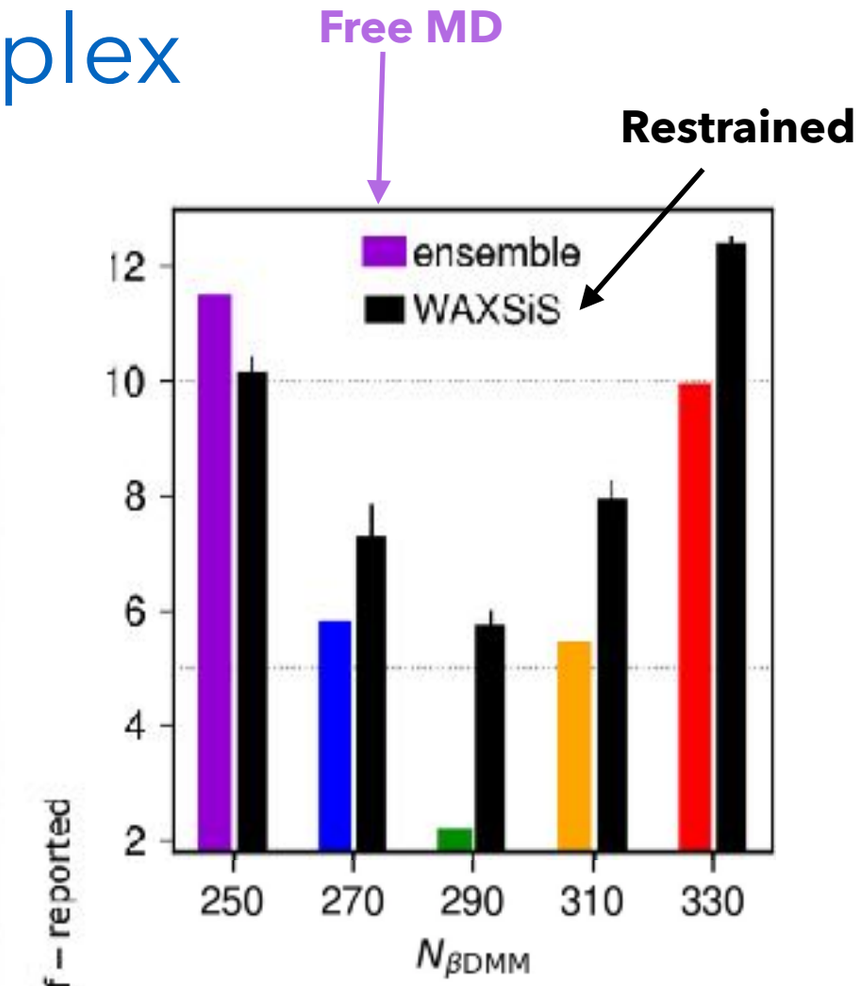
Yarawsky, Hopkins, Chatzimagas, Hub, Herr, J Mol Biol (2022)

Information added by explicit solvent?

Aquaporin protein-detergent complex



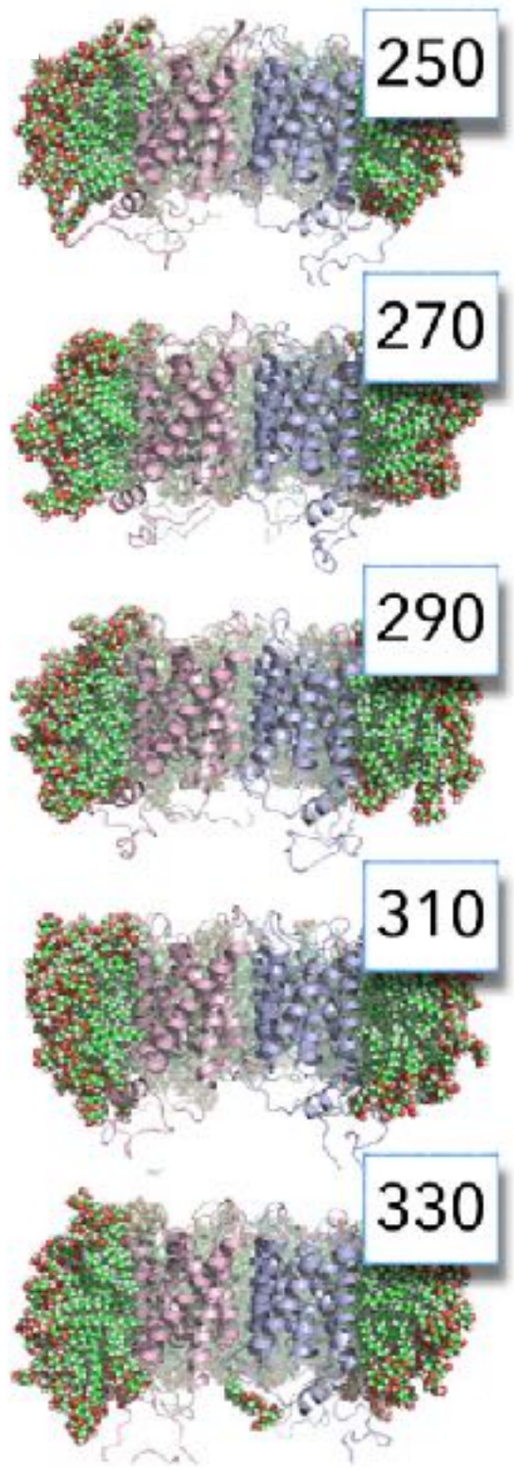


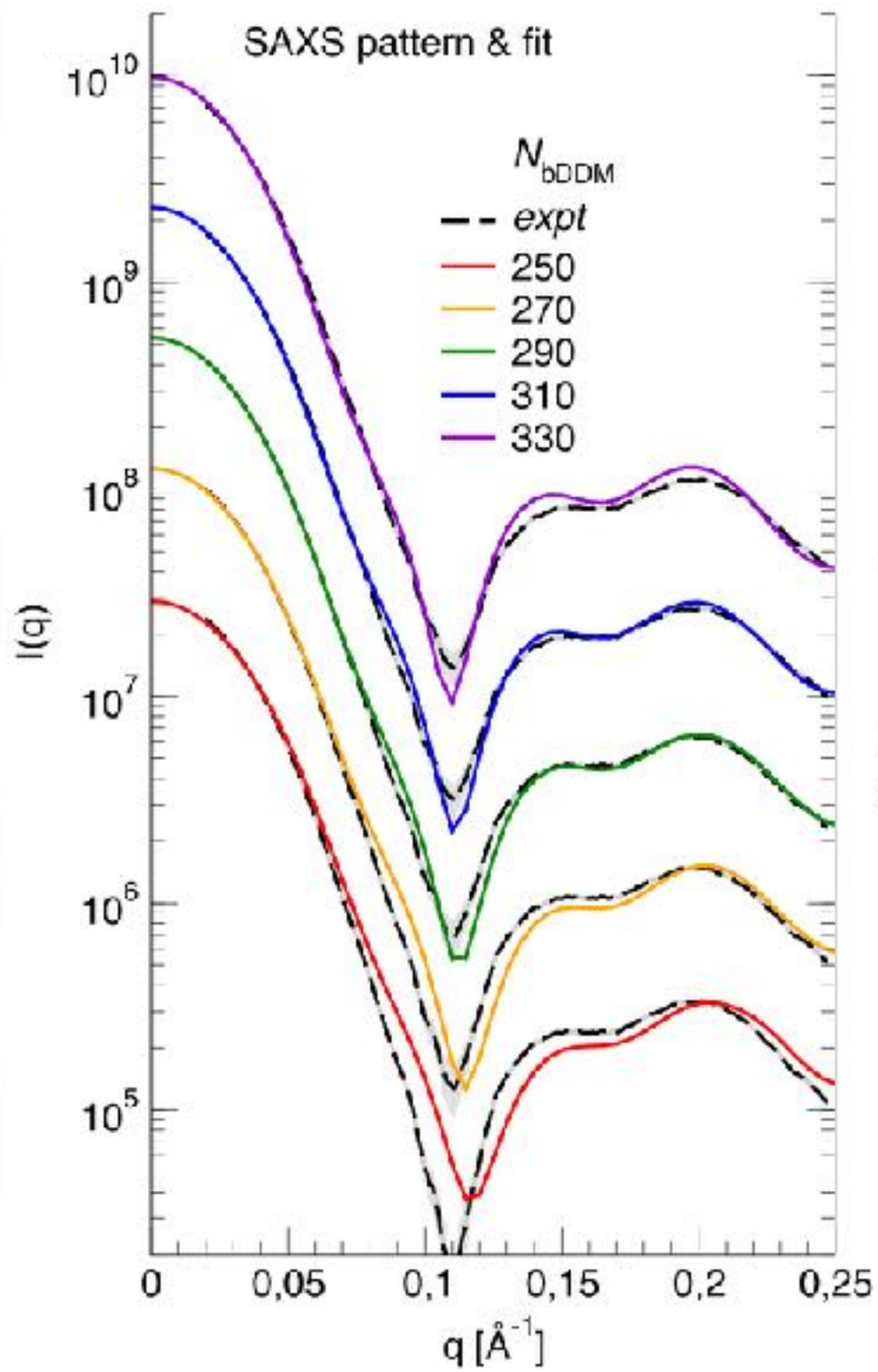


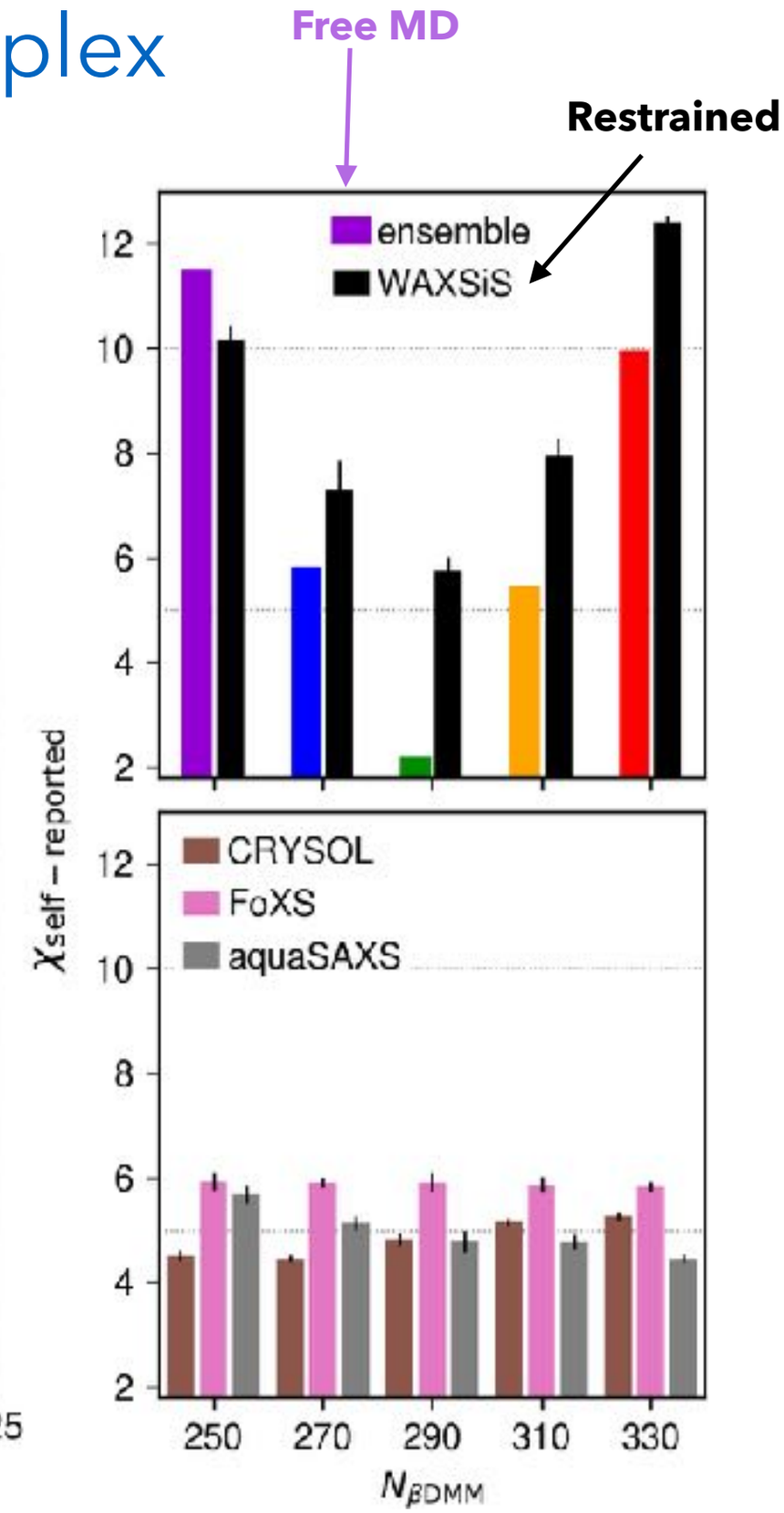
Experimental SEC-SAXS data by Javier Pérez
Berthaud, Manzi, Pérez, Mangenot, *JACS* (2012)

Chen and Hub, *J. Phys. Chem. Lett.* 6:5116-5121 (2015)

Aquaporin protein-detergent complex



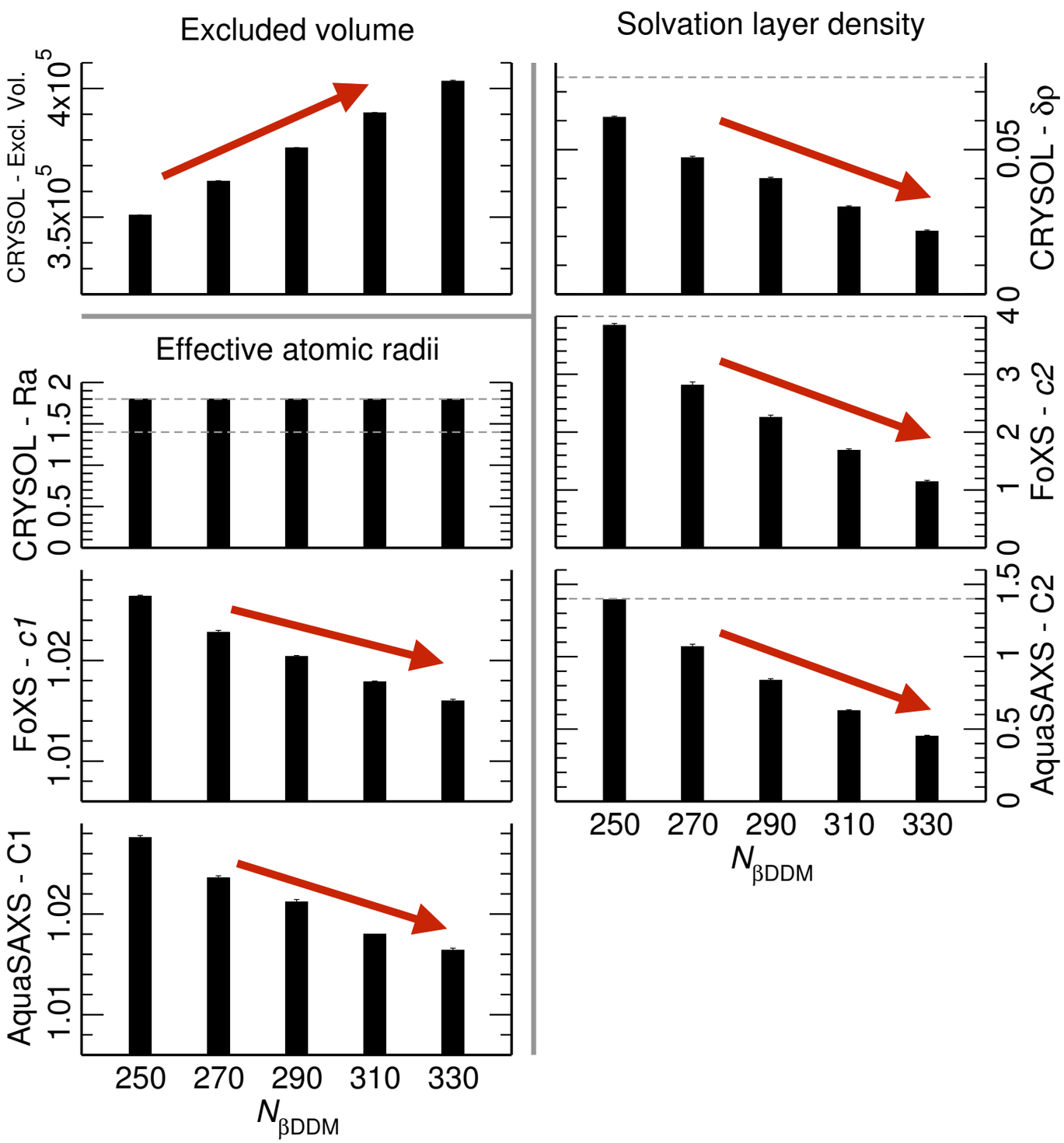




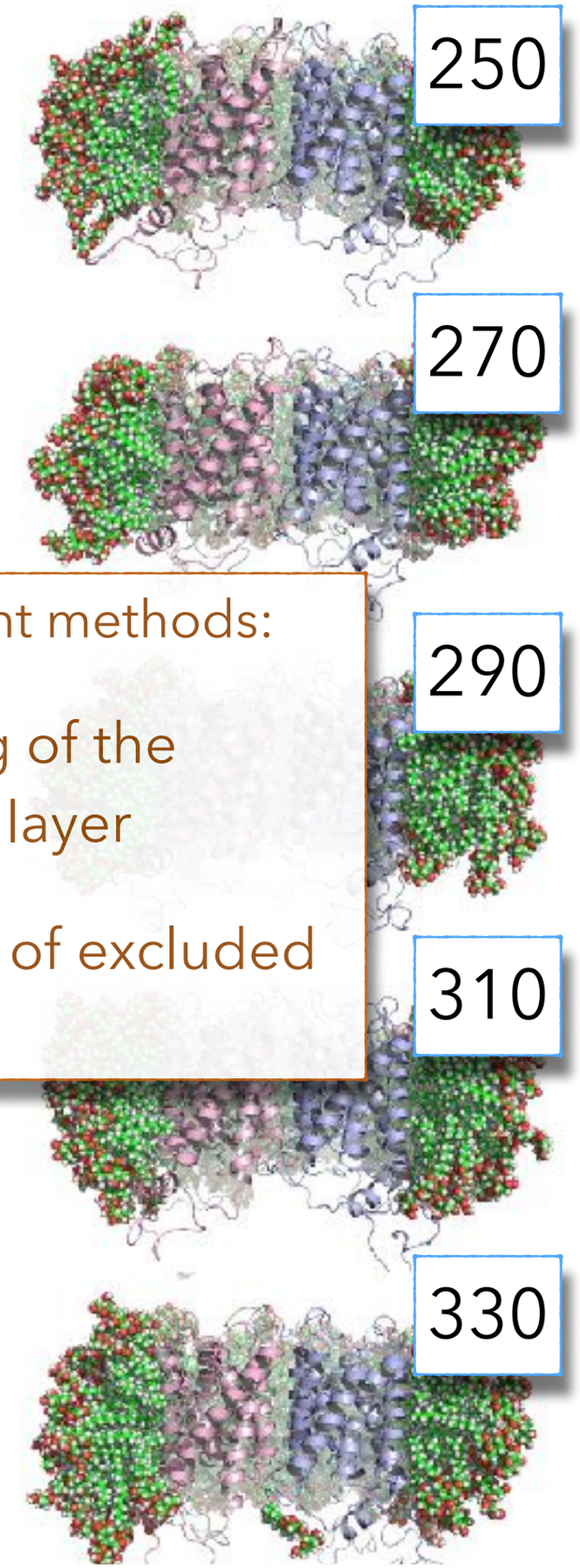
Experimental SEC-SAXS data by Javier Pérez
Berthaud, Manzi, Pérez, Mangenot, *JACS* (2012)

Chen and Hub, *J. Phys. Chem. Lett.* 6:5116-5121 (2015)

Overfitting of solvent-related fitting parameters



implicit solvent

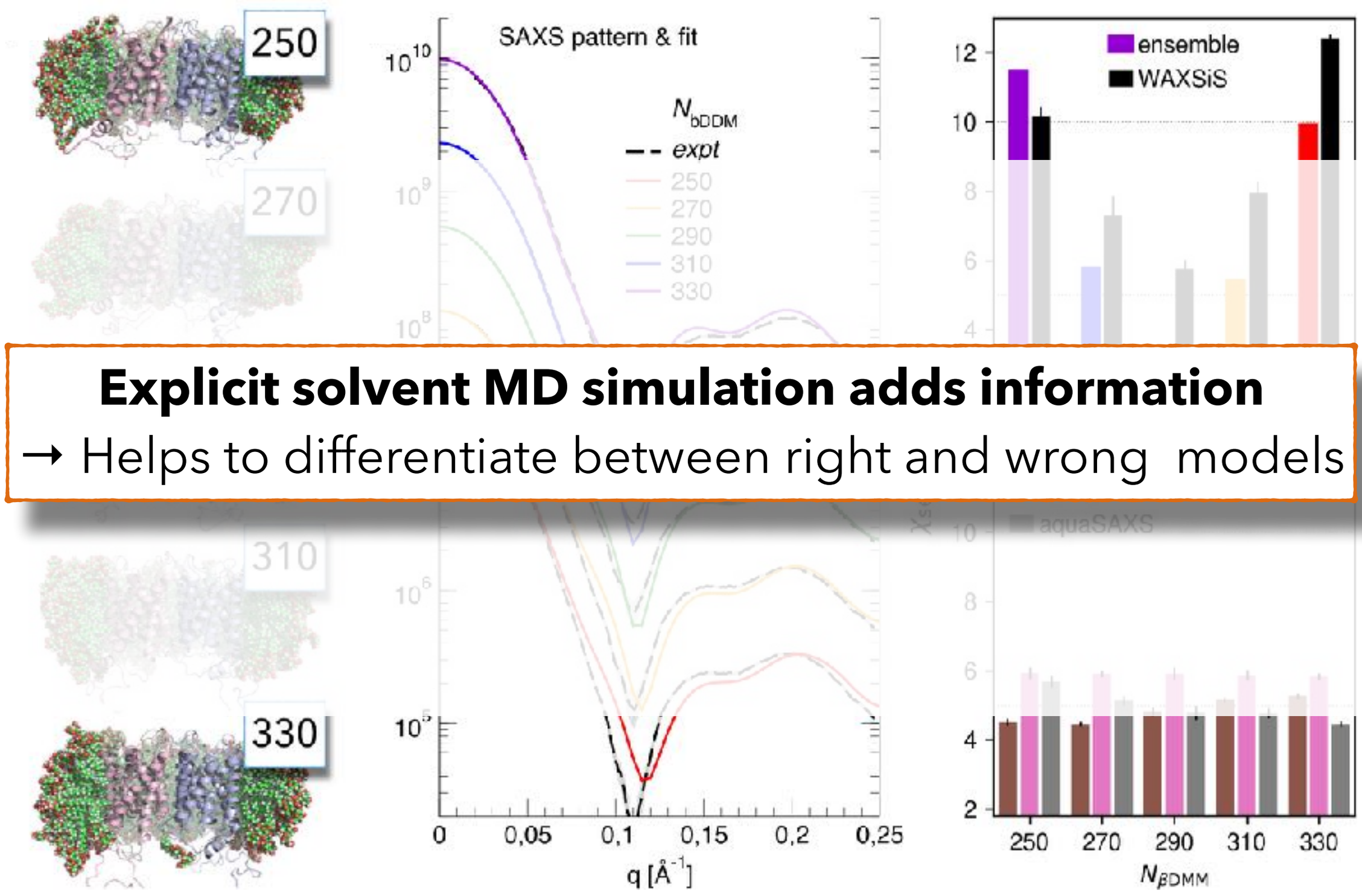


Implicit solvent methods:

- Over-fitting of the hydration layer
- Overfitting of excluded volume

Chen and Hub, *J. Phys. Chem. Lett.* 6:5116-5121 (2015)

Aquaporin protein-detergent complex



Explicit solvent MD simulation adds information

→ Helps to differentiate between right and wrong models

Experimental SEC-SAXS data by Javier Pérez
Berthaud, Manzi, Pérez, Mangenot, *JACS* (2012)

Chen and Hub, *J. Phys. Chem. Lett.* 6:5116-5121 (2015)

Structure refinement against SWAXS data with explicit-solvent MD

Experiment-supported energy

MD force field: physical knowledge

$$E_{\mathrm{hybrid}}(\mathbf{R}; I_{\mathrm{exp}}) = U_{\mathrm{waxs}}(\mathbf{R}, I_{\mathrm{exp}}) + E_{\mathrm{FF}}(\mathbf{R})$$

SAXS curve calculated from simulation

$$U_{\mathrm{WAXS}}(\mathbf{R}, I_{\mathrm{exp}}) \propto -k_B T \sum_i \frac{\left[\langle I_{\mathrm{sim}}(q_i, \mathbf{R})\rangle_t - I_{\mathrm{exp}}(q_i)\right]^2}{\sigma^2_{\mathrm{exp}}(q_i) + \sigma^2_{\mathrm{sim}}(q_i) + \sigma^2_{\mathrm{buf}}(q_i)}$$

Uncertainties

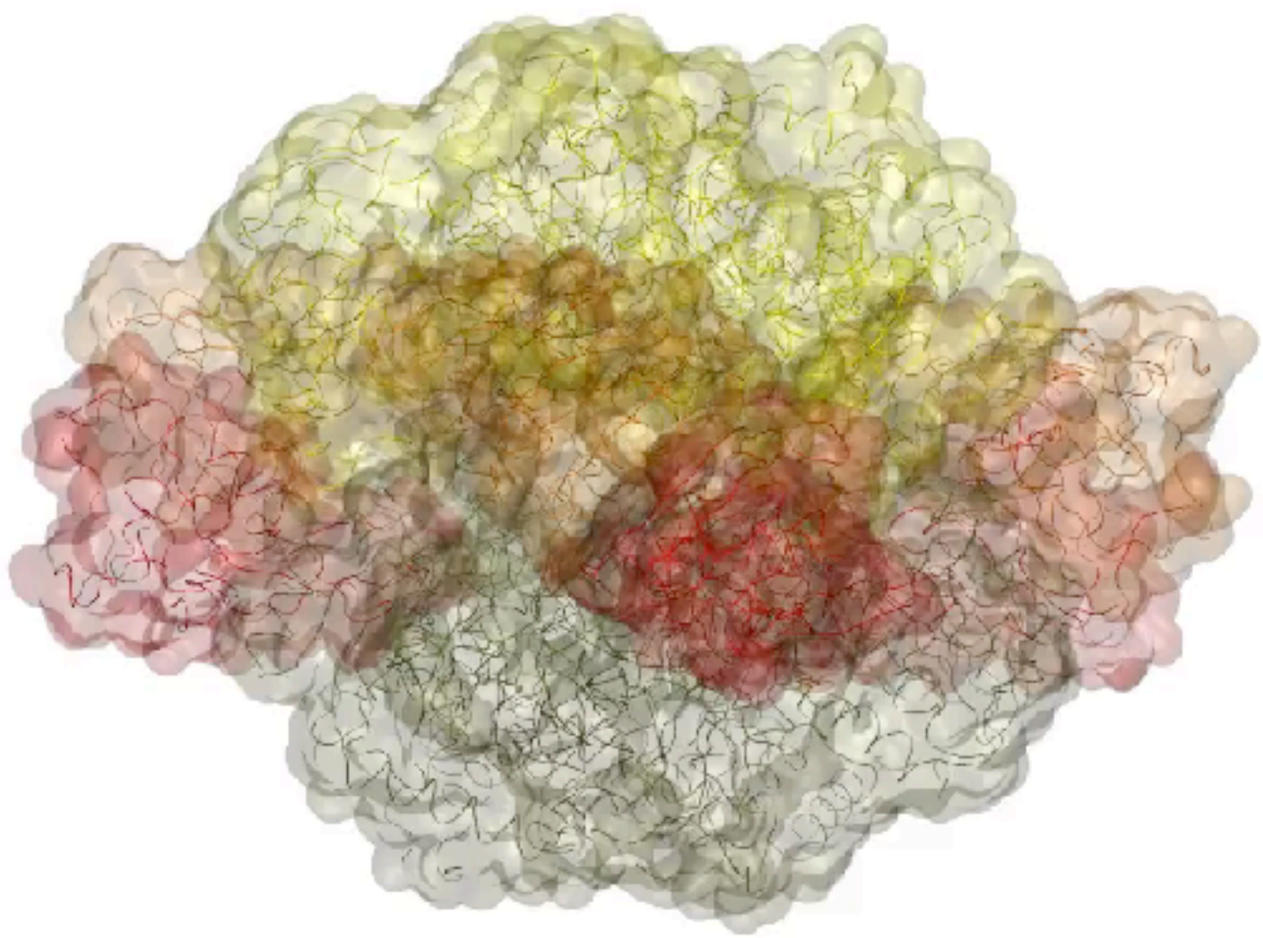
$$\mathbf{F}_k = -\frac{\partial U_{\mathrm{waxs}}}{\partial \mathbf{r}_k}$$

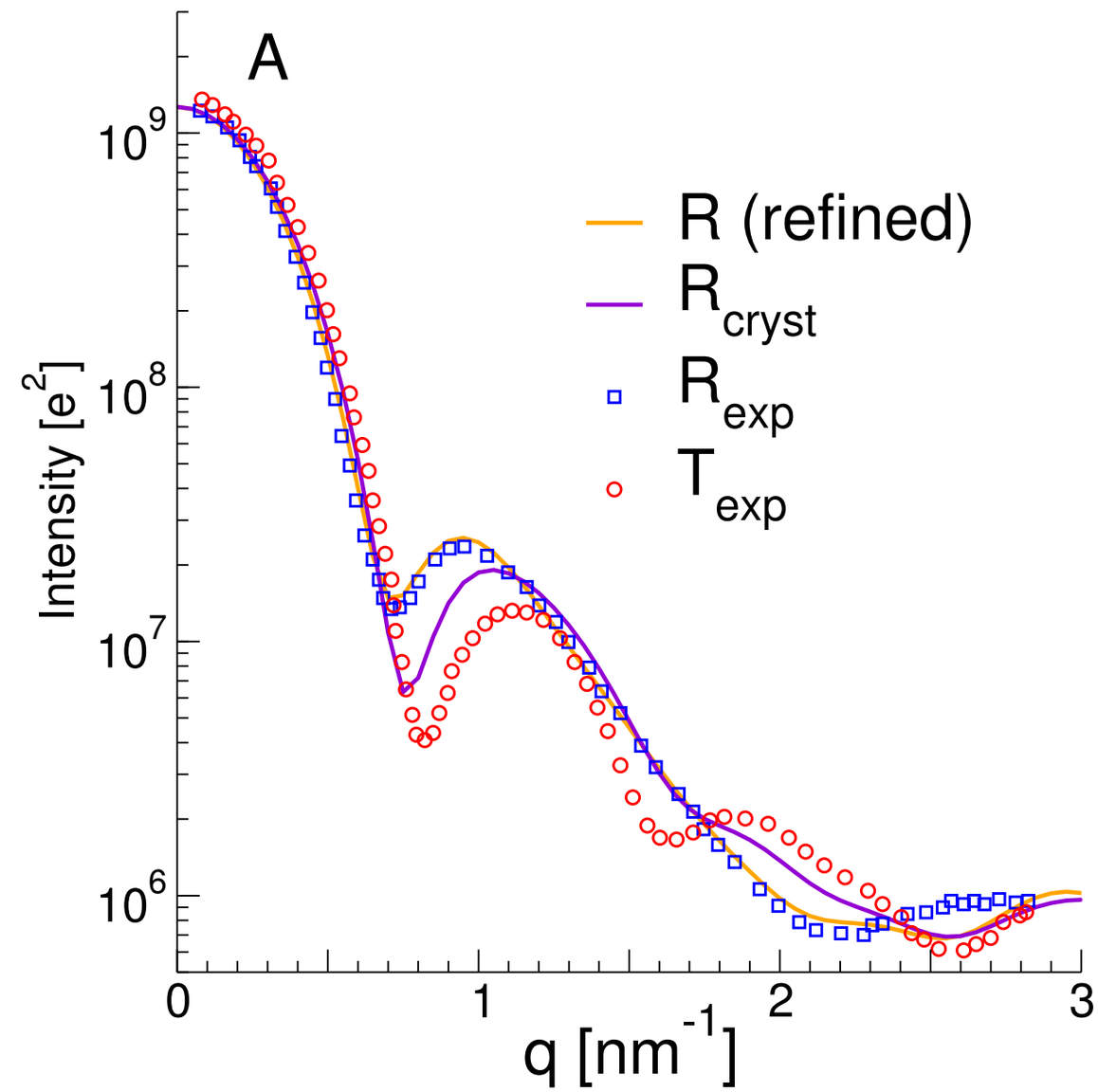
Forces

Chen and Hub, *Biophys. J.* 108, 2573-2584 (2015)

Biophysical Journal Paper of the Year Award

ATCase refinement





- Start: **T** state
- SAXS curve of **R** state

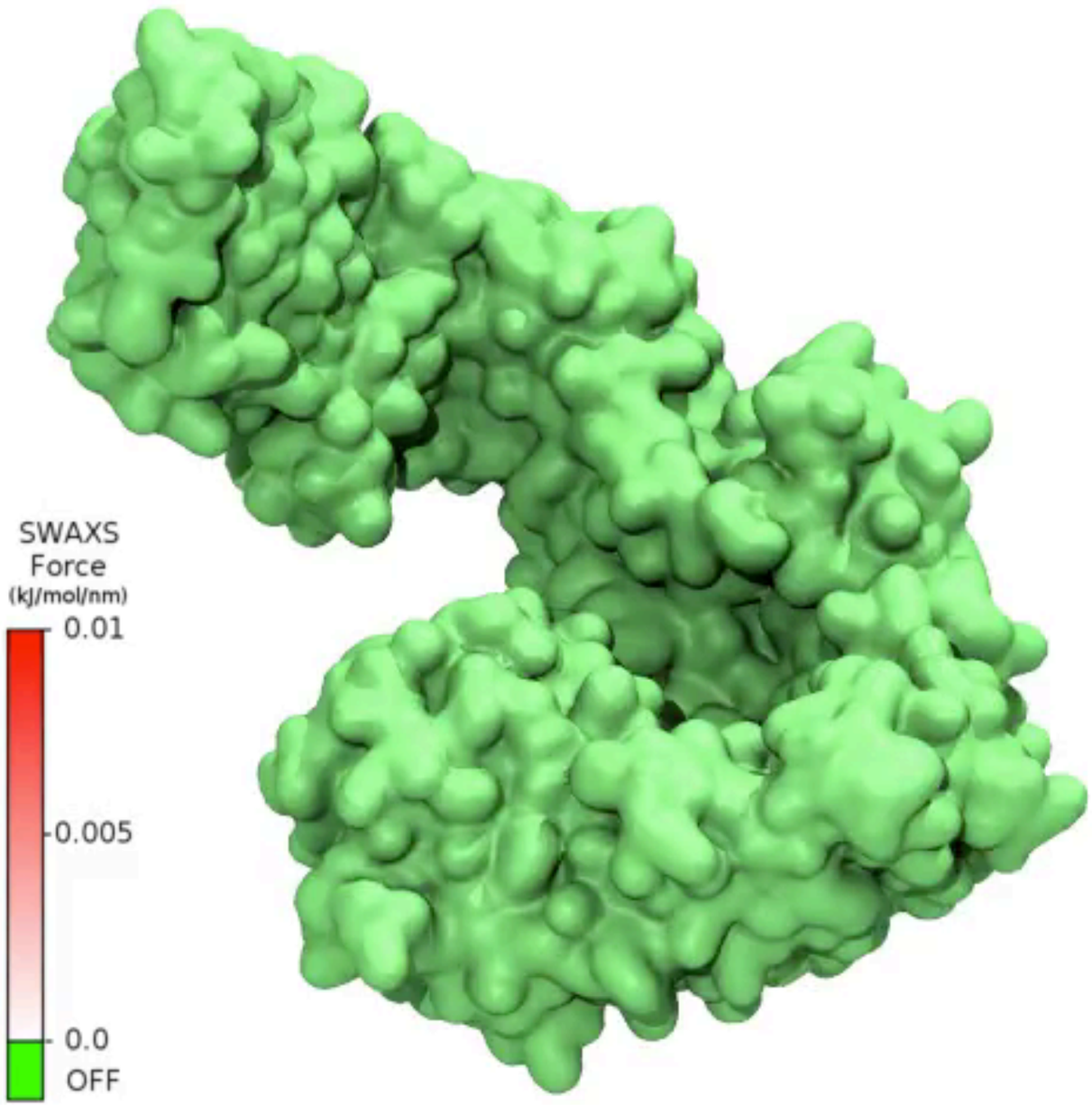
Fetler *et al.*, *J. Mol. Biol.* 2001

Svergun *et al.*, Proteins 1997

Chen and Hub, *Biophys. J.* 108, 2573-2584 (2015)

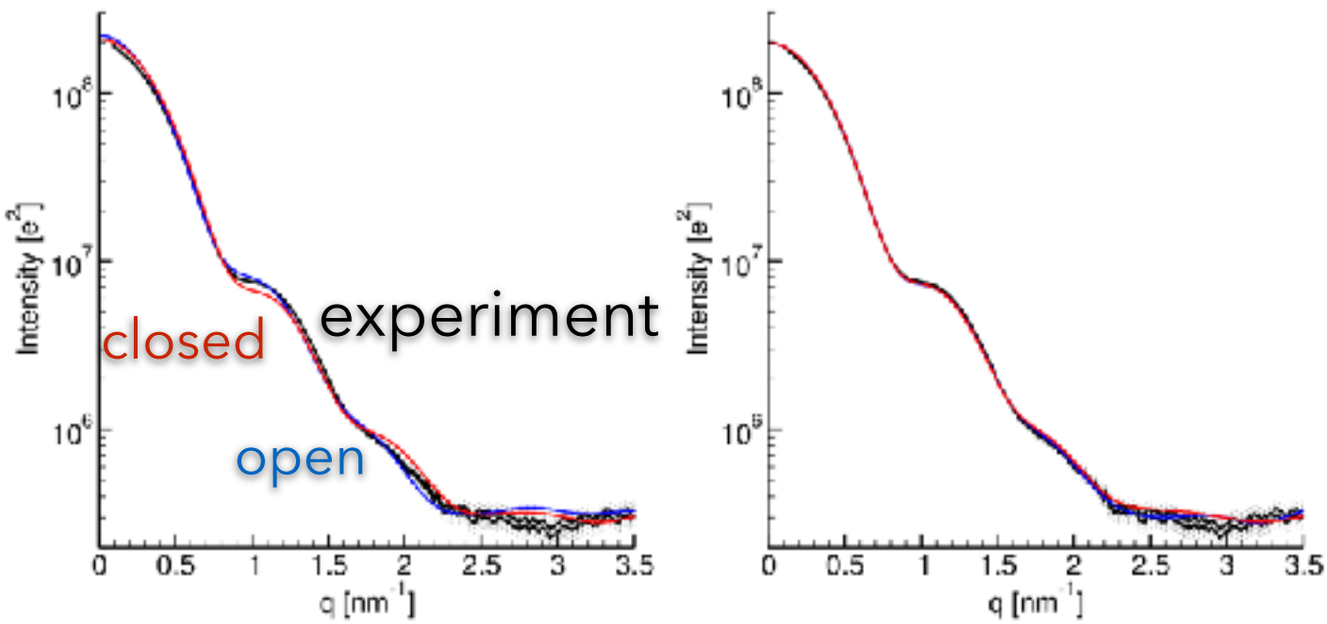
Biophysical Journal Paper of the Year Award

Nuclear exportin CRM1 refinement



Free simulation: slightly too open

SAXS-restrained simulation: more compact



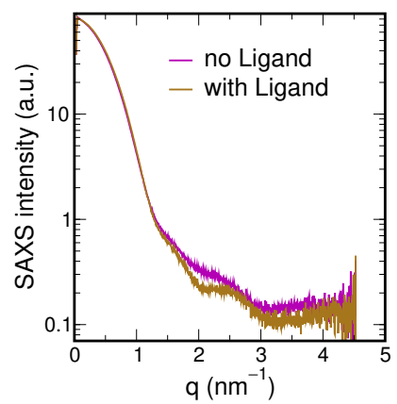
Amber99SB simulation

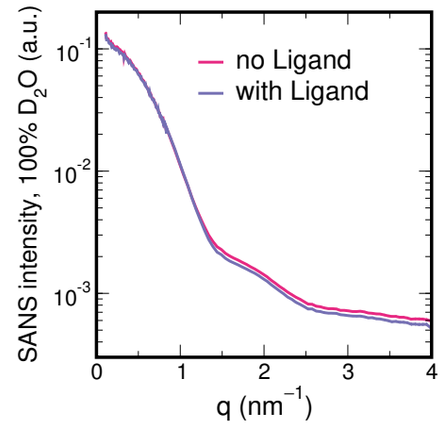
with and w/o SAXS refinement

Chen and Hub, *Biophys. J.* 108, 2573-2584 (2015)
Moneke et al, Science 2009
Moneke et al, PNAS 2013

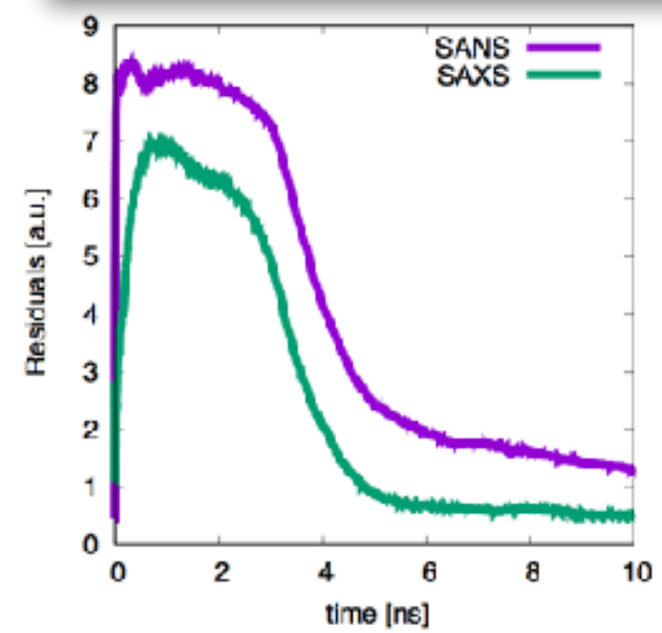
Glimpse on: Cross-validation against **neutron** scattering

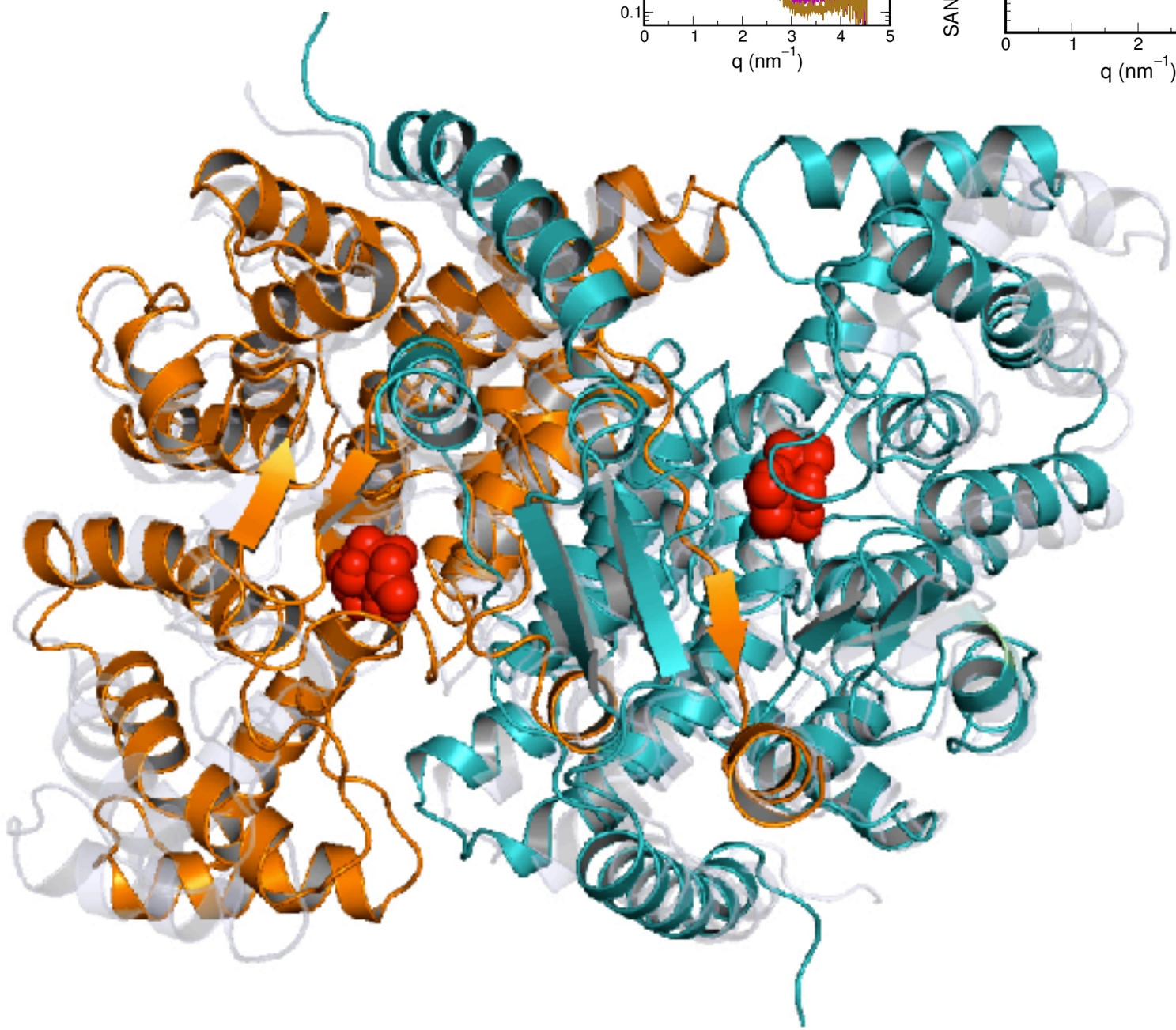
Data by:
Andreas Stadler
(FZ Jülich)



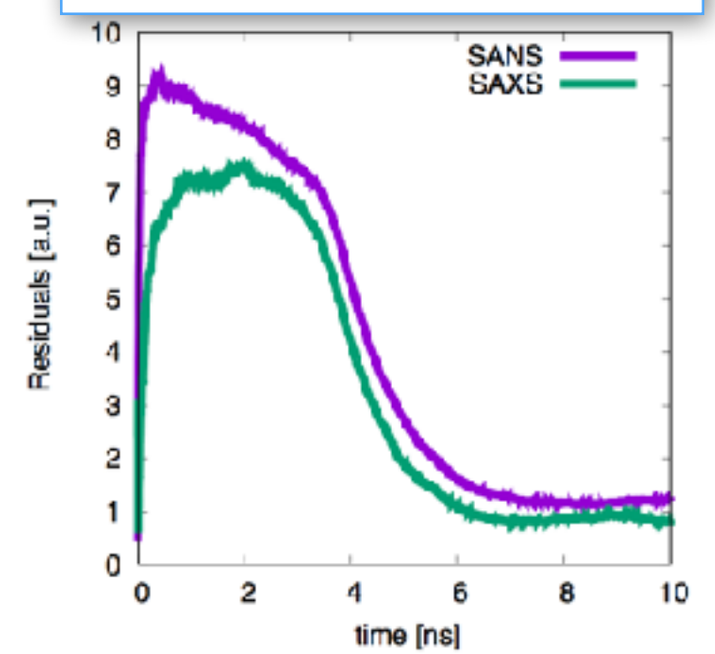


- Refined against SAXS
- Cross-validated against SANS





- Refined against SANS
- Cross-validated against SAXS



Chen, Shevchuk,…, Stadler, Henning, Hub, *JCTC*, 2019

MD-based vs. Rigid-body refinement

	Rigid-body refinement	**MD-based refinement**
Force field, physical model	Simple, little predictive E.g., volume exclusion	Accurate and predictive, all-atom MD forcefield
Sampling dominated by	Experimental data. Risk: overfitting	Force field / physical model Risk: force field bias, sampling
Add-hoc constraint definitions	Rigid domains, linkers	-
Accessibility	Simple, e.g. SASREF	Some MD skills required. https://gitlab.com/cbjh/gromacs-swaxs
Computational cost	Cheap / runs on PC	Fast computer / cluster access needed, sampling problems possible

Single structure vs. ensemble refinement

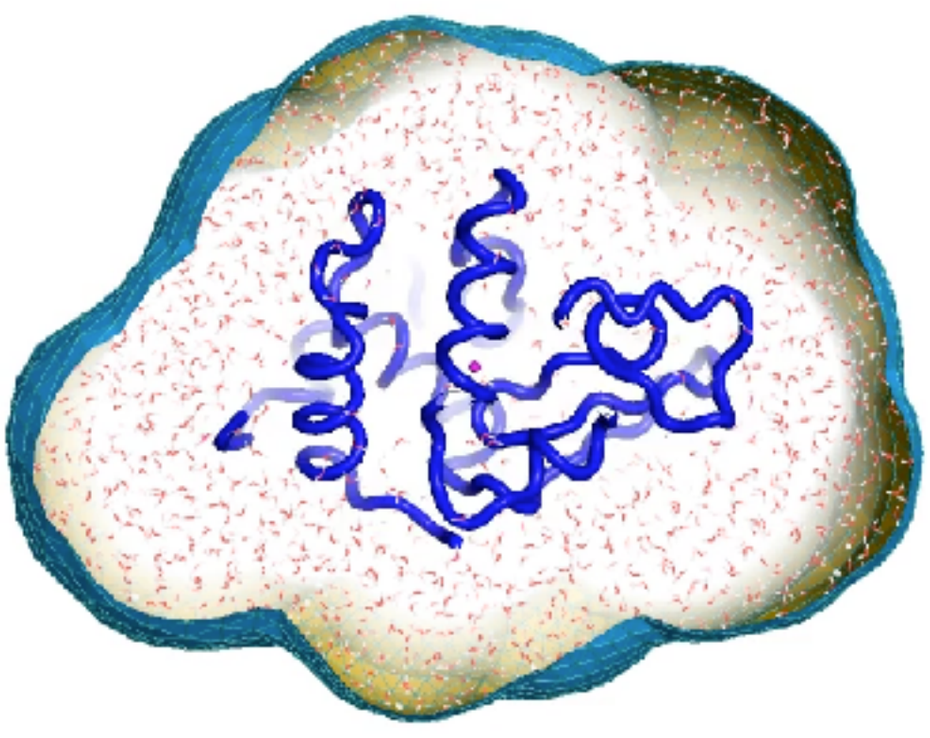
So far ....

Single structure !

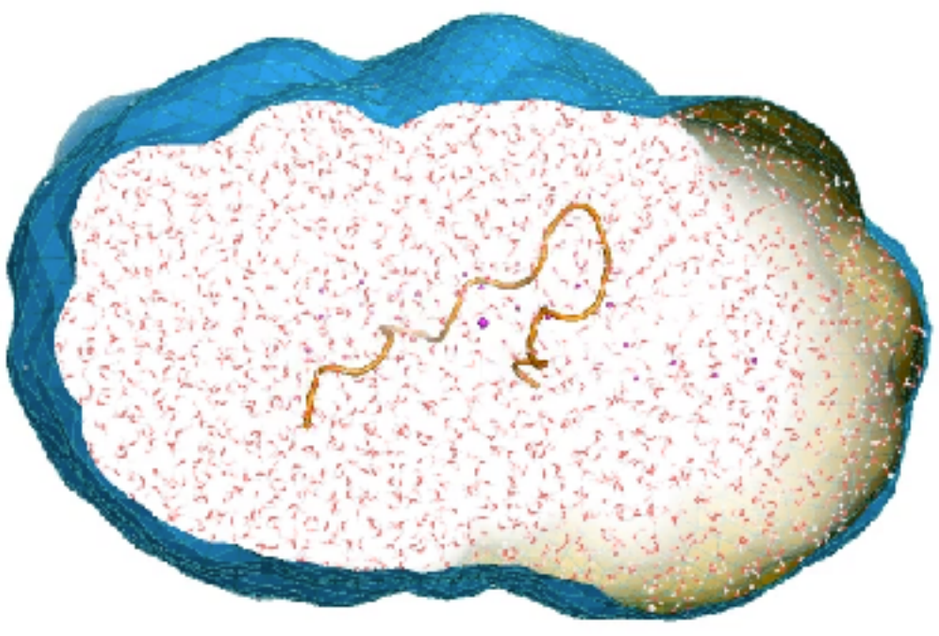
$$E(\mathbf{R}; I_{\text{exp}}) = E_{\text{Force field}}(\mathbf{R}) + w_{\text{exp}} E_{\text{exp}}(\mathbf{R}; I_{\text{exp}})$$

$$E_{\text{exp}}(\mathbf{R}, I_{\text{exp}}) = -k_B T \sum_i \frac{\left[\langle I_{\text{sim}}(q_i, \mathbf{R})\rangle_t - I_{\text{exp}}(q_i)\right]^2}{\sigma^2_{\text{exp}}(q_i) + \sigma^2_{\text{sim}}(q_i) + \sigma^2_{\text{buf}}(q_i)}$$

OK for folded protein



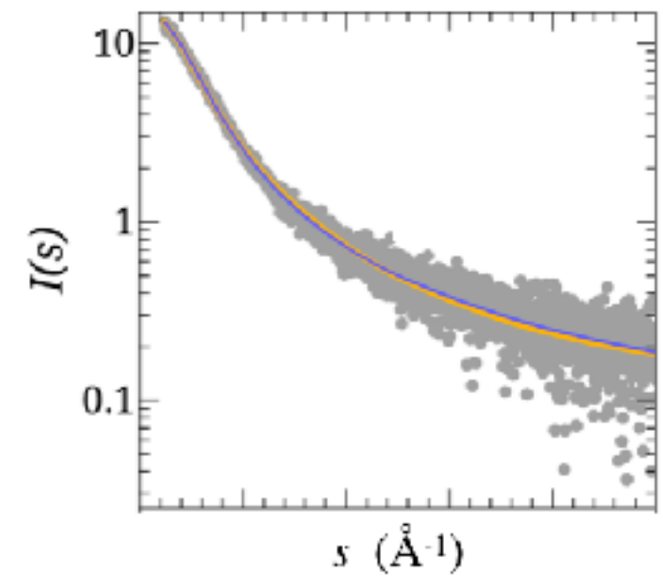
Not OK for intrinsically disordered protein



Ensemble refinement

Goal now: Find **ensemble** $p_1(\mathbf{r})$ that matches the data:

$$\langle I_{\mathrm{calc}}(q_i)\rangle = \int p_1(\mathbf{r}) I_{\mathrm{calc}}(\mathbf{r}, q_i)\mathrm{d}\mathbf{r} = I_{\mathrm{exp}}(q_i)$$



But many distributions explain the data…

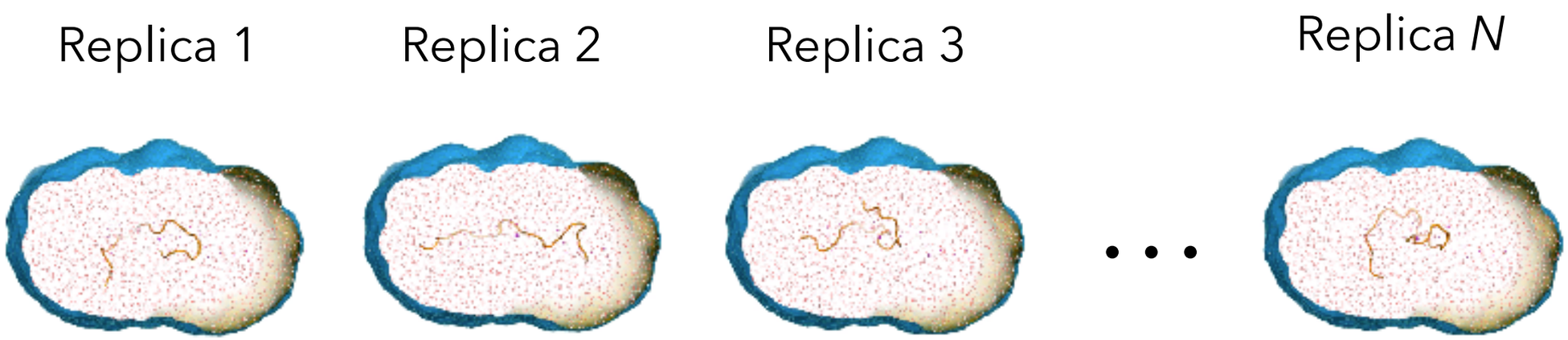
Maximum Entropy Principle (E.T. Jaynes, 1957)

"Use the least informative distribution (distribution with the largest entropy) that is compatible with your constraints / your knowledge."

"Do not add information that you do not have."

Jaynes, *Phys Rev* (1957)

Parallel-replica ensemble-refinement



$I_1(q)$ $I_2(q)$ $I_3(q)$

Average over relipcas: $\bar{I}(q) = N^{-1} \sum_i I_i(q)$

Couple replica-average to experiment:

$$E_{\mathrm{ME}}(\mathbf{r}) = E_{\mathrm{FF}}(\mathbf{r}) + \frac{kN}{2} \sum_{i=1}^{N_q} \left[\bar{I}(q_i) - I_{\mathrm{exp}}(q_i)\right]^2$$

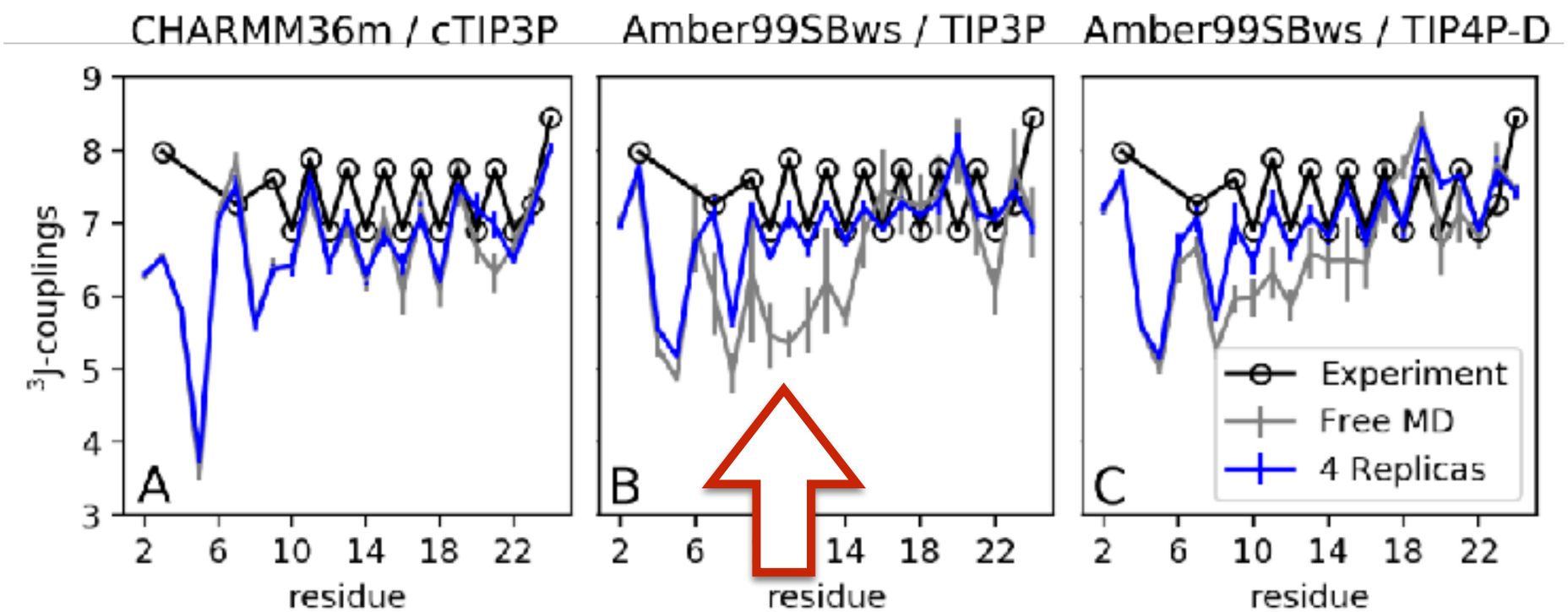
Maximum Entropy ensemble in the $N \longrightarrow \infty$

Pitera & Chodera, *JCTC* (2012)
Roux Weare *J Chem Phys* (2014)
White & Voth, *JCTC* (2014)
Boomsma et al., PLoS Comput Biol (2014)

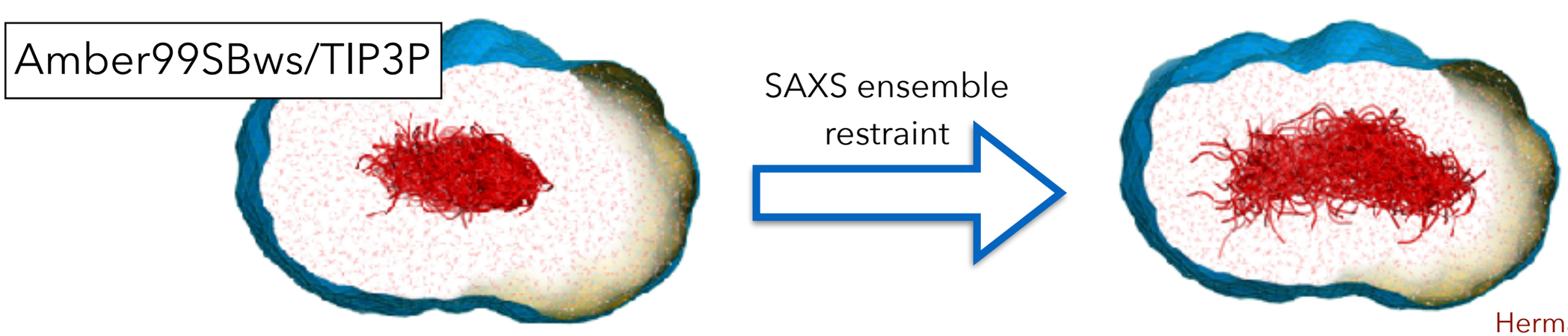
Validation of SAXS-refined ensembles

^{3}J(HN-Hα) couplings

No Overfitting!

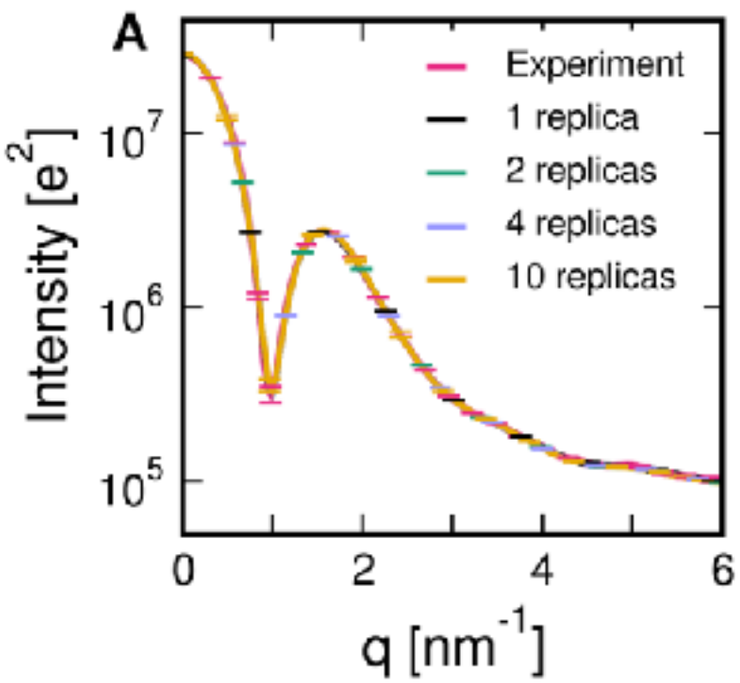


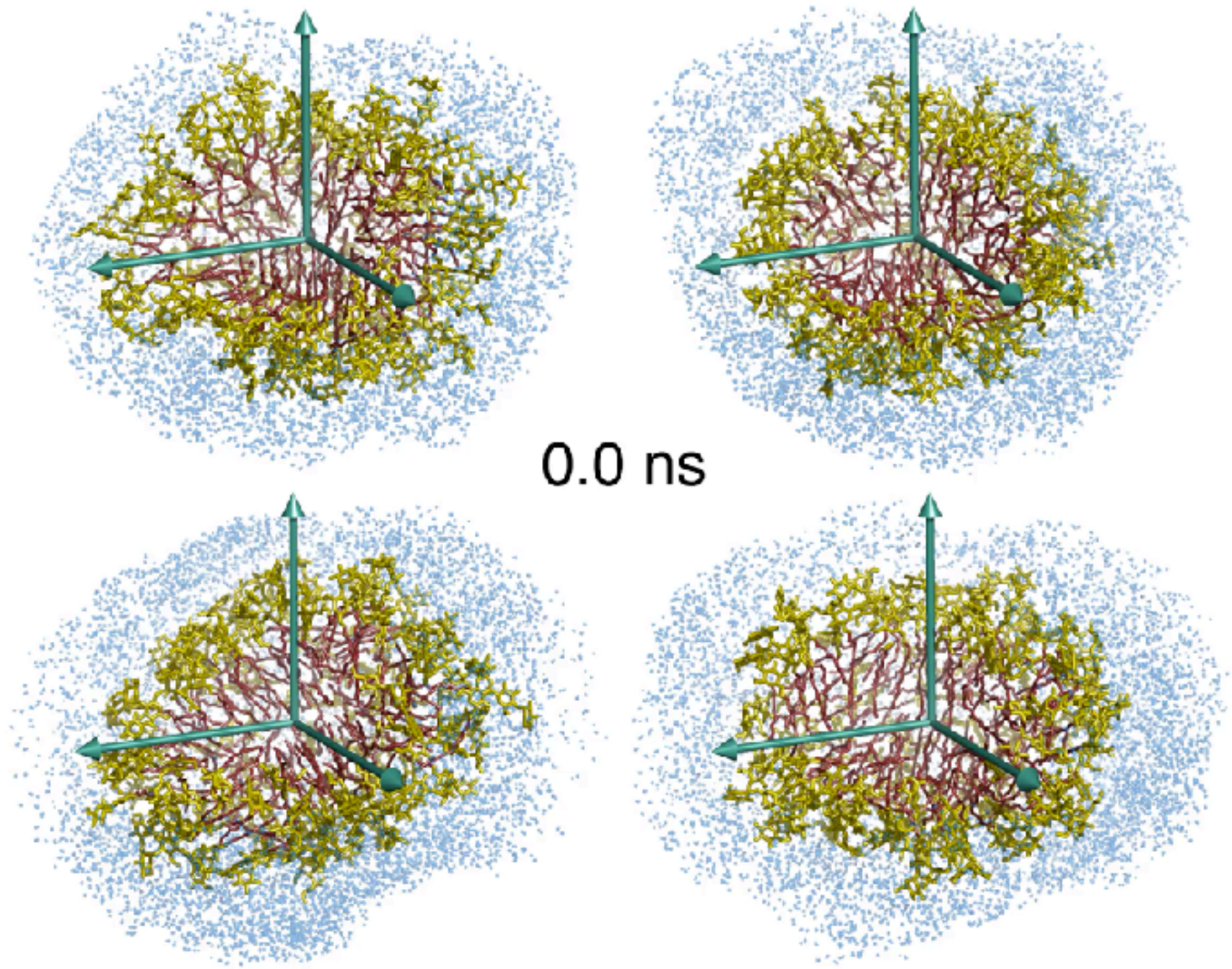
- CHARMM36m: No effect by SAXS restraints on NMR data
- Amber: improvement of NMR data



Hermann and Hub, *JCTC* 2019

Multi-replica ensemble refinement of micelle



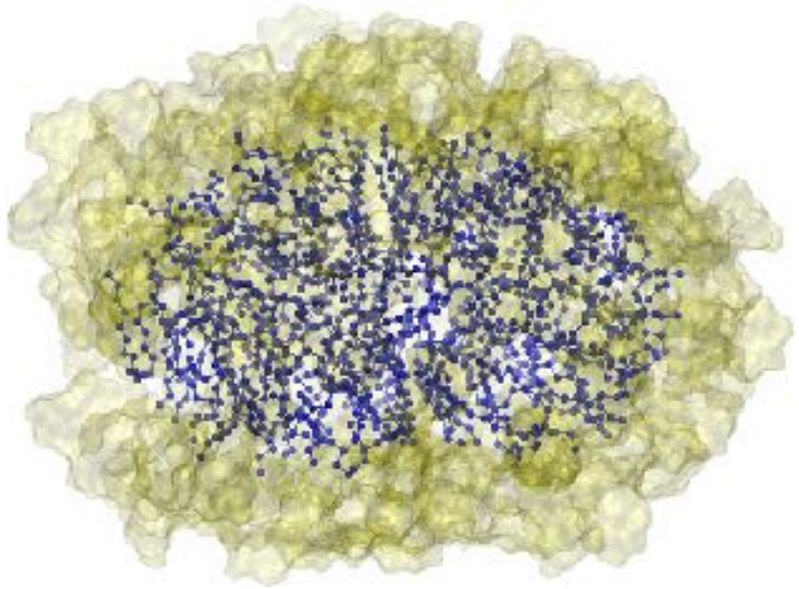


Correct average shape and realistic fluctuations

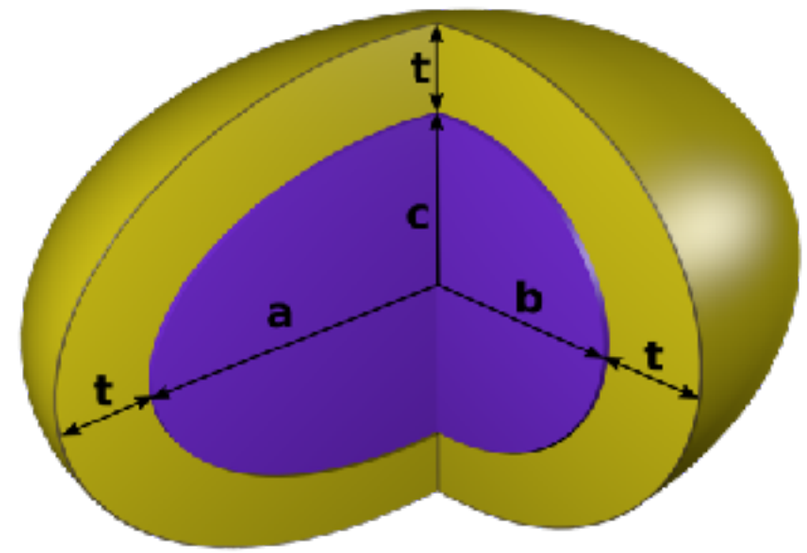
Hermann and Hub, JCTC (2019)

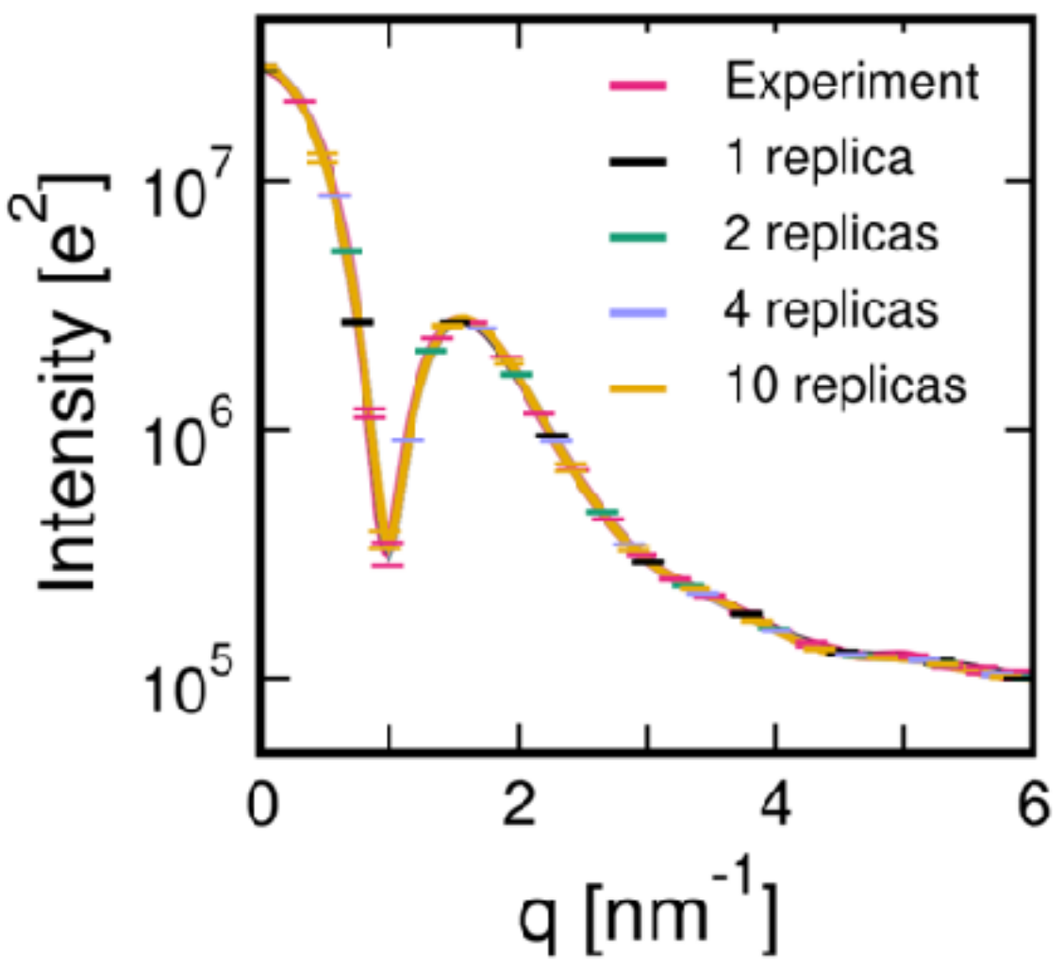
Ivanovic *et al.*, J Phys Chem Lett (2020)

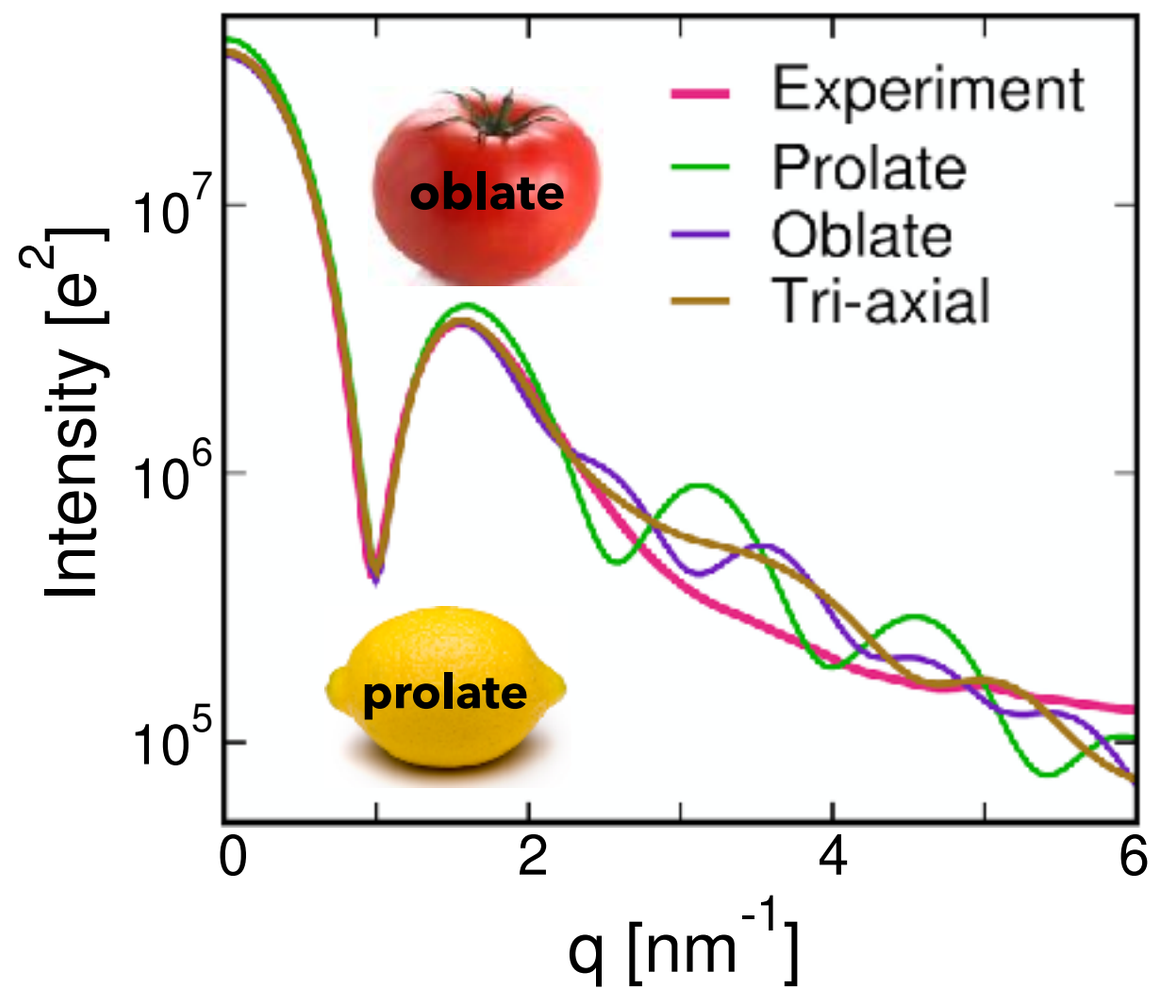
All-atom MD versus analytic continuum models



vs.







Ivanovic, Hermann, Wojcik, Perez, Hub, J Phys Chem Lett (2020)

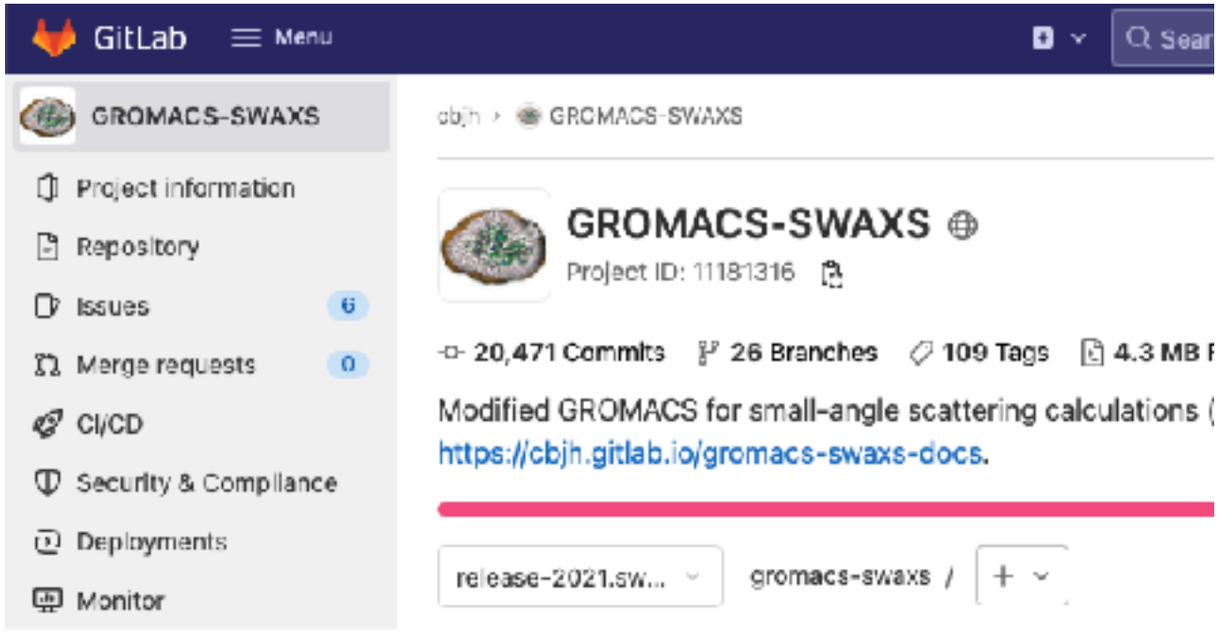
Availability

SWAXS-modified Gromacs code

https://gitlab.com/cbjh/gromacs-swaxs

Spack

```
spack install gromacs-swaxs+cuda
spack install gromacs-swaxs+cuda~mpi ^fftw~mpi
```



Offline use? Please contact us!

WAXSiS - WAXS in Solvent

http://waxsis.uni-saarland.de

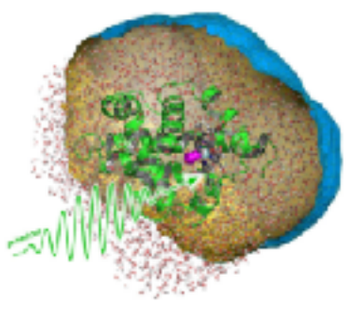
Documentation

https://cbjh.gitlab.io/gromacs-swaxs-docs/

GROMACS-SWAXS

Welcome to GROMACS-SWAXS, a modified GROMACS version for

- predictions of small-angle X-ray and neutron scattering (SAXS/SANS) curves from explicit-solvent MD simulations,
- structure refinement of proteins or soft-matter complexes against SAXS/SANS curves



Documentation

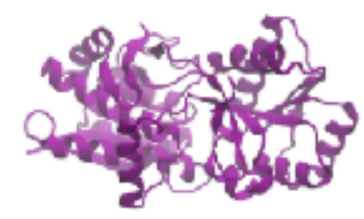
- Usage
  - Modified and added GROMACS modules

Tutorials

https://cbjh.gitlab.io/gromacs-swaxs-docs/tutorials.html

Tutorials

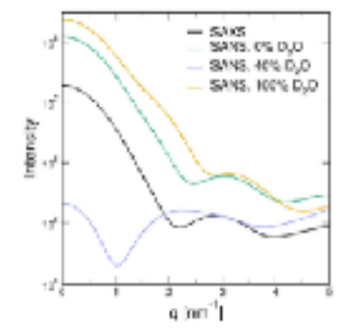
SAXS/SANS predictions, SAXS-driven MD, and multi-replica ensemble refinement



Tutorial carried out at the EMBO SAS workshop in Grenoble

This tutorial show the following:

- Computing SAXS/SANS curves from a given protein trajectory
- SAXS-driven simulations: opening a two-domain protein
- Multi-replica SAXS-restrained ensemble simulations of an intrinsically disordered protein



Getting started with your own MD-based SAXS interpretation

Get a reasonably fast Computer

Example:

- 8-core CPU
- Nvidia GPU, RTX 4080Ti, or RTX 3080Ti @ Ebay
- Small RAM needed

Learn MD basics

GROMACS Tutorials

Justin A. Lemkul, Ph.D.
Virginia Tech Department of Biochemistry

Tutorial 1: Lysozyme in Water

Read two book chapters

Predicting solution scattering patterns with explicit-solvent molecular simulations

Structure and ensemble refinement against SAXS data: combining MD simulations with Bayesian inference or with the maximum entropy principle

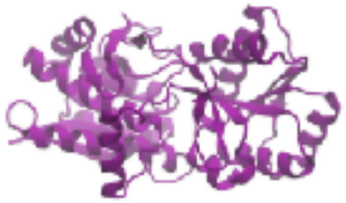
Chatzimagas and Hub, Meth Enzymol (2022, 2023)
or arXiv and bioRxiv (2022)

Do our tutorials at

https://cbjh.gitlab.io/gromacs-swaxs-docs/

Tutorials

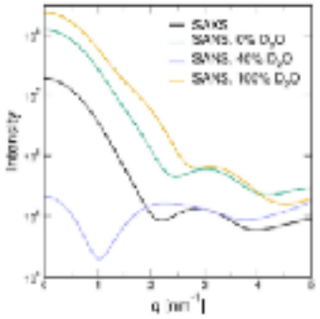
SAXS/SANS predictions, SAXS-driven MD, and multi-replica ensemble refinement



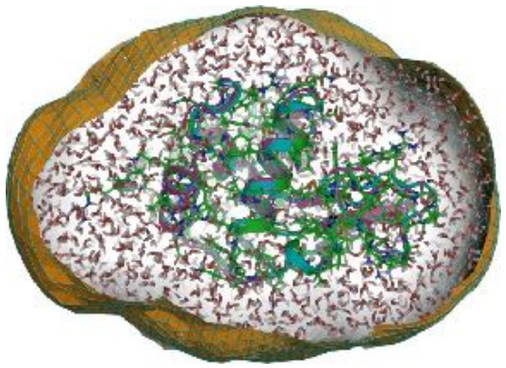
Tutorial carried out at the EMBO SAS workshop in Grenoble

This tutorial show the following:

- Computing SAXS/SANS curves from a given protein trajectory
- SAXS-driven simulations: opening a two-domain protein
- Multi-replica SAXS-restrained ensemble simulations of an intrinsically disordered protein

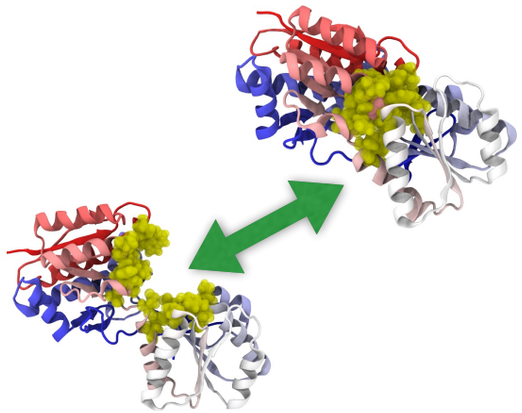


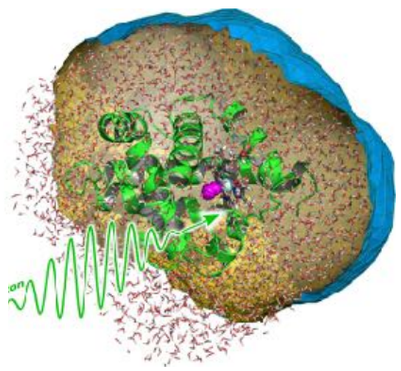
Summary



Explicit-solvent MD may guide the interpretation of experimental SAXS/WAXS data

- Accurate fitting-free SAXS/WAXS predictions
- No solvent-related fitting parameters ☞ highly predictive
- Webserver http://waxsis.uni-saarland.de
- Structure refinement of proteins and soft-matter complexes against SWAXS
- Ensemble refinement with the Maximum Entropy Principle
- All Open Source, documentation and tutorials





Chen and Hub,, ***Biophys J***, 107, 435-447 (2014)
Chen and Hub, ***Biophys J,*** 108, 2573-2584 (2015)
Brinkmann and Hub, ***J. Chem. Phys.*** 143:2897-2899 (2015)
Knight and Hub, ***Nucleic Acids Res.*** 43, W225-W230 (2015)
Chen and Hub,, ***J Phys Chem Lett*** 6, 5116–5121 (2015)
Cordeira et al., ***Nucleic Acids Res*** (2016)
Brinkmann and Hub, ***PNAS*** 113, 10565-10570 (2016)
Ivanovic, Brützel, Liefert, Hub, ***Angew. Chem. Int. Ed.***, (2018)
Ivanovic et al., **PCCP** (2018)
Shevchuk and Hub, ***PLoS Comput. Biol.*** (2017)
Hub, **Curr Opin Struct Biol** (2018)
Chen et al, **JCTC** 2019
Hermann and Hub, **JCTC** 2019

Acknowledgements

Collaborators

Ludwig-Maximilian Universität
Jan Liefert

SOLEIL, Paris
Javier Pérez

Institut Pasteur
Felix Rey

University of Kaiserslautern
Sandro Keller

Gothenburg University
Sebastian Westenhoff

Cincinnati Children's Hospital
Herr lab

University of Göttingen
Marcus Müller, Christina Ting

CNRS Monpellier
Pau Bernado

FZ Jülich
Andreas Stadler

Computational Biophysics Group

Leonie Chatzimagas
Maciej Wojcik
Leonhard Starke
Chetan Poojari
Katharina Scherer
Johanna Linse
Lucas Andersen
Nooa Aho

Former members
"Poker" Po-chia Chen
Levin Brinkmann
Chris Knight
Roman Shevchuk
Kalina Atkovska
Massimliano Anselmi
Robert Becker
Jeremy Lapierre
Igor Ariz
Felix Strand
Tobias Fischbach
Neha Awasthi
Markus Hermann
Milos Ivanovic
Tobias Fischbach
Gari Kasparyan

MARIE CURIE ACTIONS
SFB860
SFB 803
SFB 1027
Heisenberg-Programm DFG
Emmy Noether-Programm Deutsche Forschungsgemeinschaft DFG