Integrative structural modeling using SAS data

Dina Schneidman

האוניברסיטה העברית בירושלים
THE HEBREW UNIVERSITY OF JERUSALEM

Integrative Structure Modeling

Information

X-ray crystallography

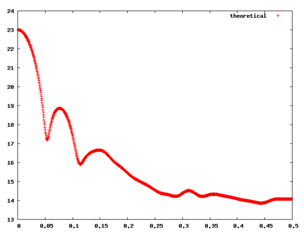
Electron microscopy

NMR spectroscopy

AlphaFold2

RosettaFold

cross-linking mass spectrometry



SAXS/SANS

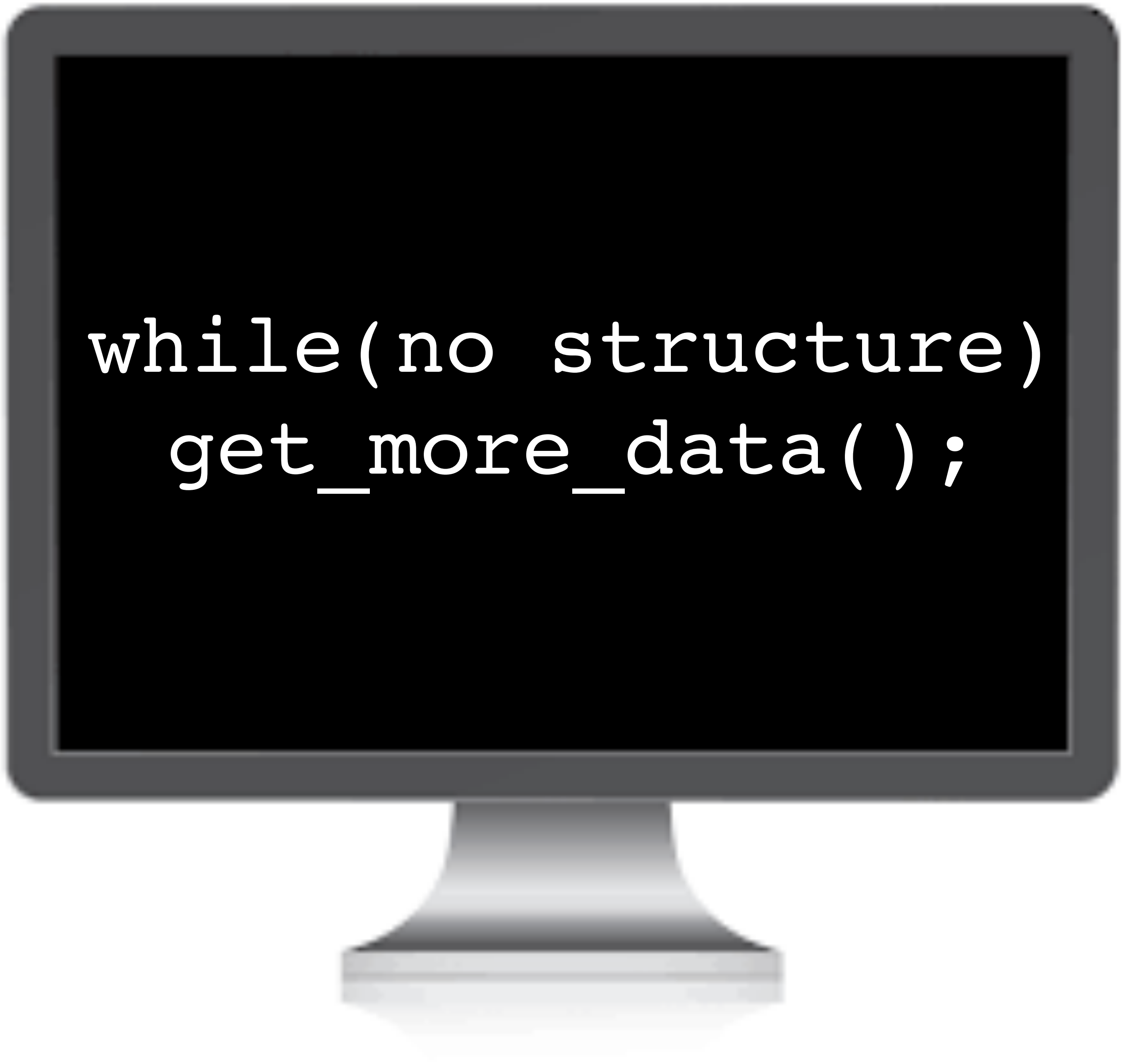
affinity purification

statistical potential

FRET

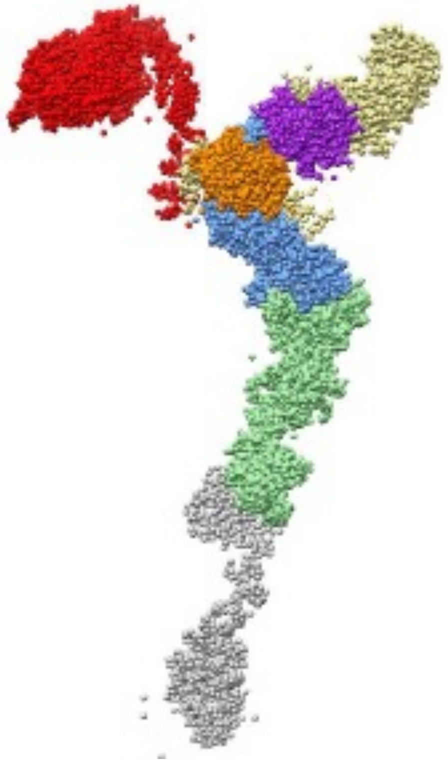
site-directed mutagenesis

Computational methods



```
while(no structure)
 get_more_data();
```

Structural models

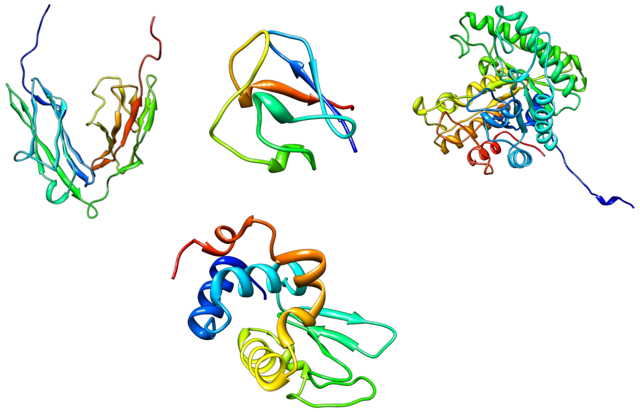


Nup84

Ward A, Sali A, Wilson I. Integrative structural biology. Science 2013.
Rout M, Sali A. Principles for Integrative Structural Biology Studies. Cell 2019.

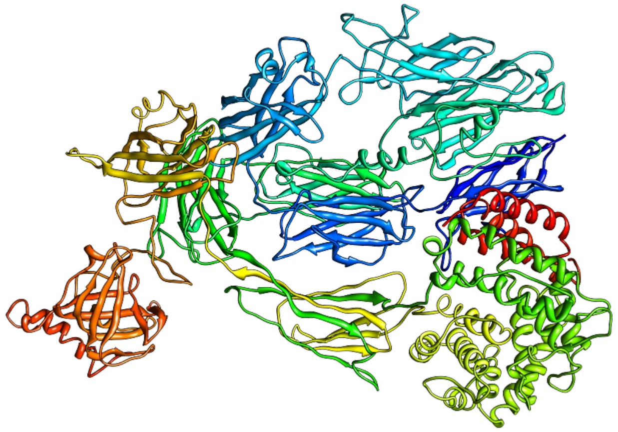
What are the modeling tasks?

protein
folding



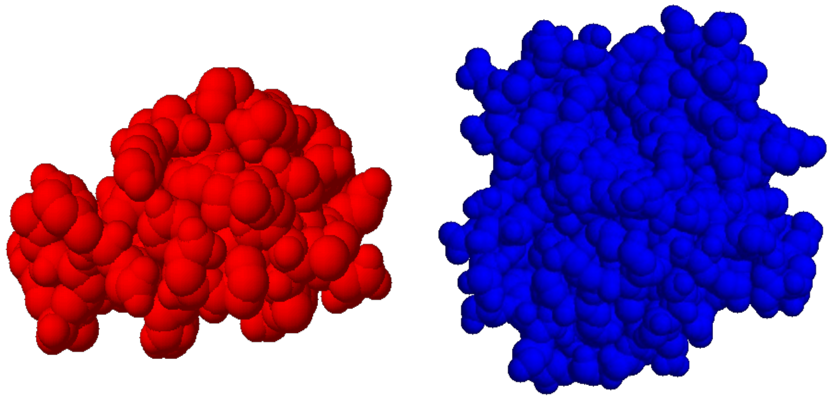
homology modeling
ab initio modeling

assembly of multi
domain proteins

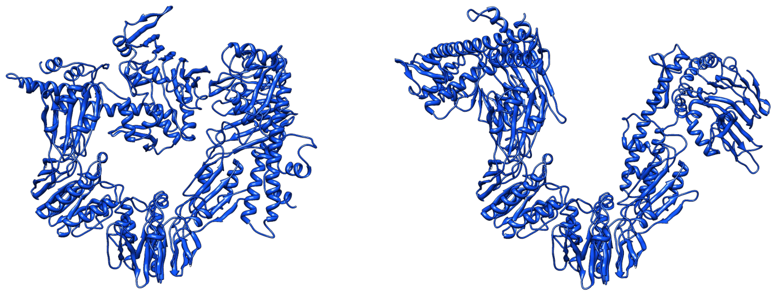


rigid assembly

assembly of multi
protein complexes



structural characterization
of protein dynamics

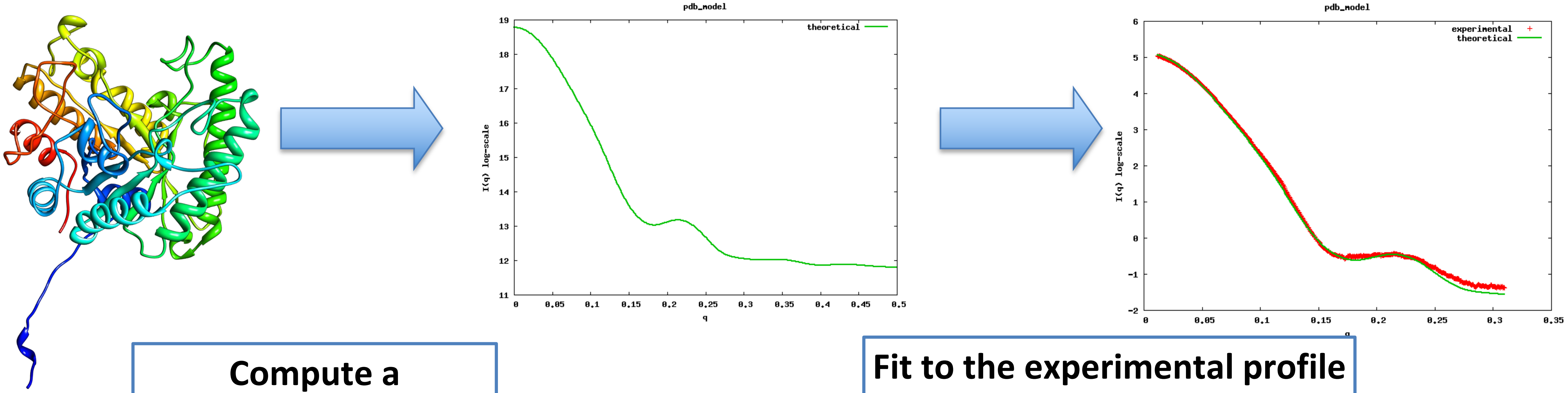


flexible assembly

Scoring: Fast open-source X-ray Scattering

FoXS

A rapid method for computing a SAXS profile of a given structure and for matching of the computed and experimental profiles



Compute a theoretical profile

Fit to the experimental profile and compute the score

Debye formula (1915)

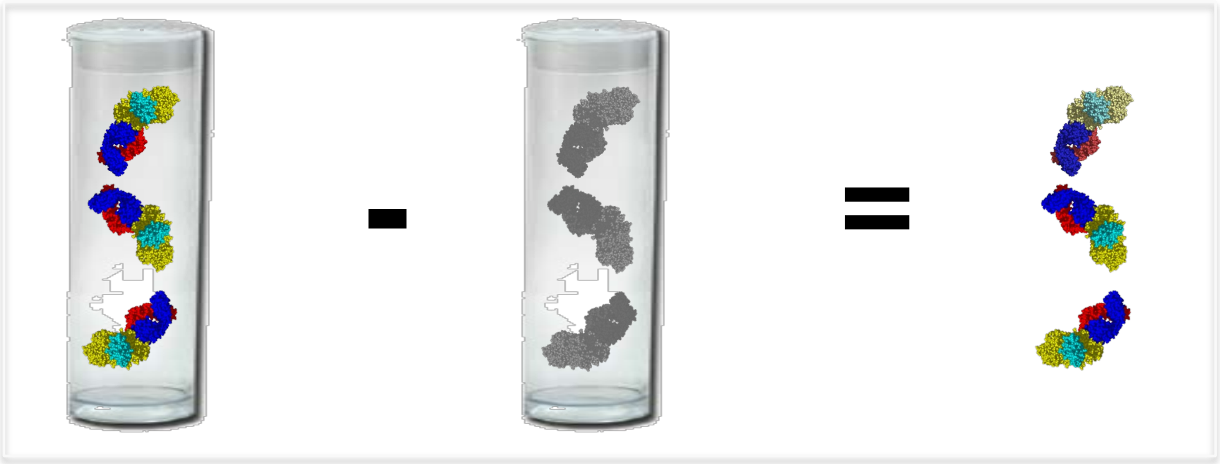
$$I(q)=\sum_{i=1}^{N}\sum_{j=1}^{N} f_i(q) f_j(q)\frac{\sin(qd_{ij})}{qd_{ij}}$$

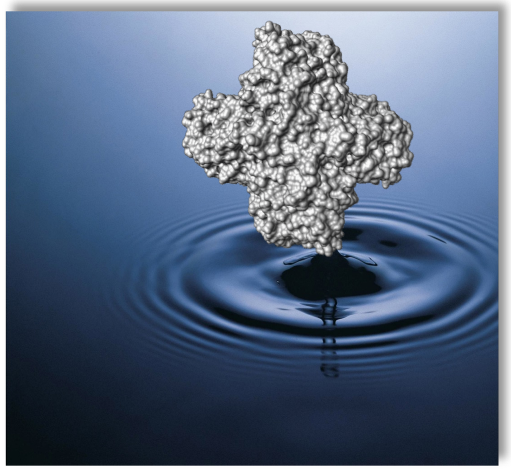
$$\chi=\sqrt{\frac{1}{M}\sum_{i=1}^{M}\left(\frac{I_{exp}(q_i)-cI(q_i)}{\sigma(q_i)}\right)^2}$$

Schneidman-Duhovny D, Hammel M, Sali A. NAR 2010

Schneidman-Duhovny D, Hammel M, Tainer J, Sali A. Biophys J 2013

Scoring: Excluded Volume and Hydration Layer Density





vacuum

solvent excluded volume

hydration layer

$$f_i(q) = f_i^v(q) - \boxed{C_1(q)} f_i^s(q) \quad + \boxed{c_2} s_i f^w(q)$$

Increase/decrease atomic radii to obtain the best fit to the experimental profile

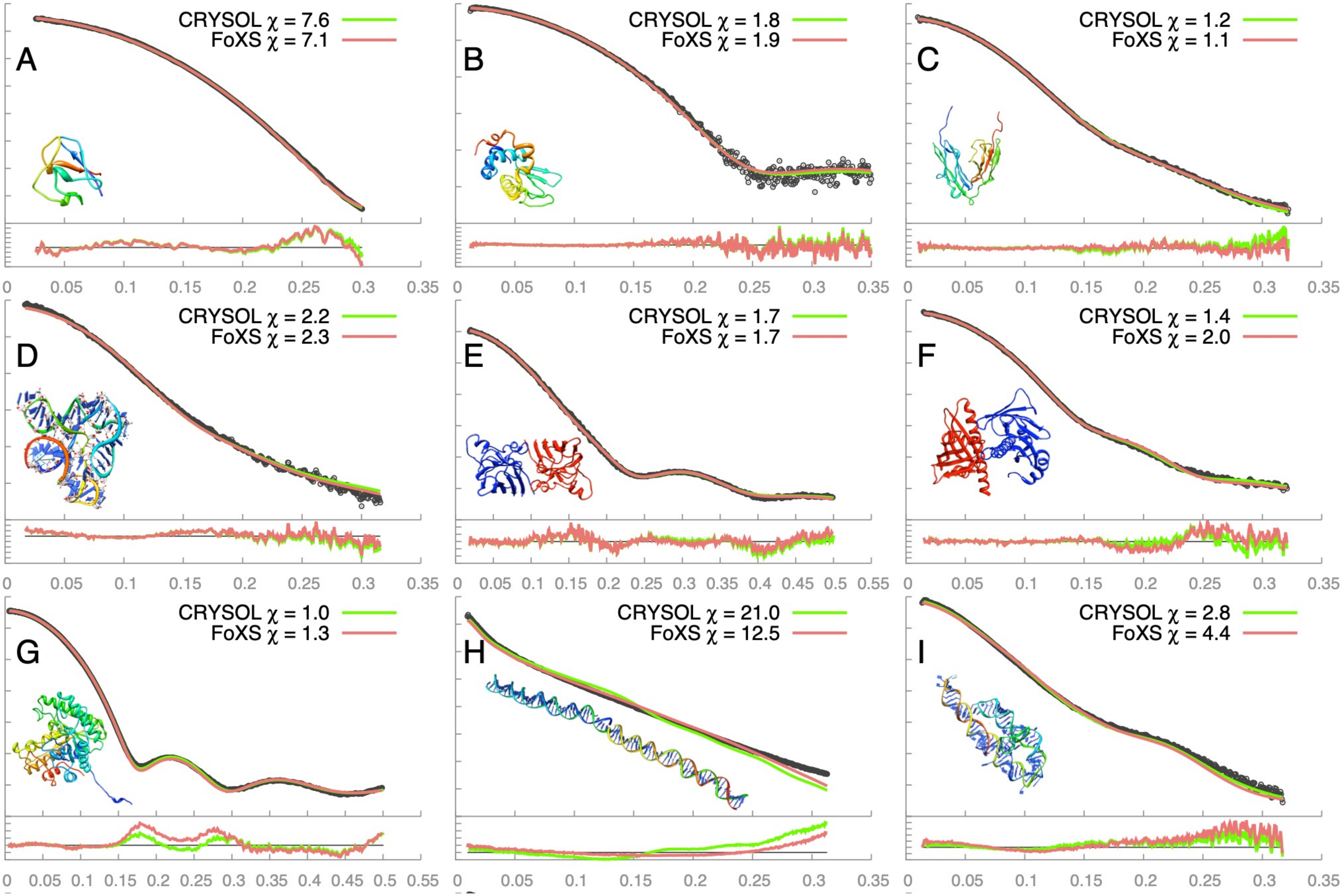
Add water form factor to solvent accessible atoms (s_i measures solvent accessibility [0-1])

5% variance in radius

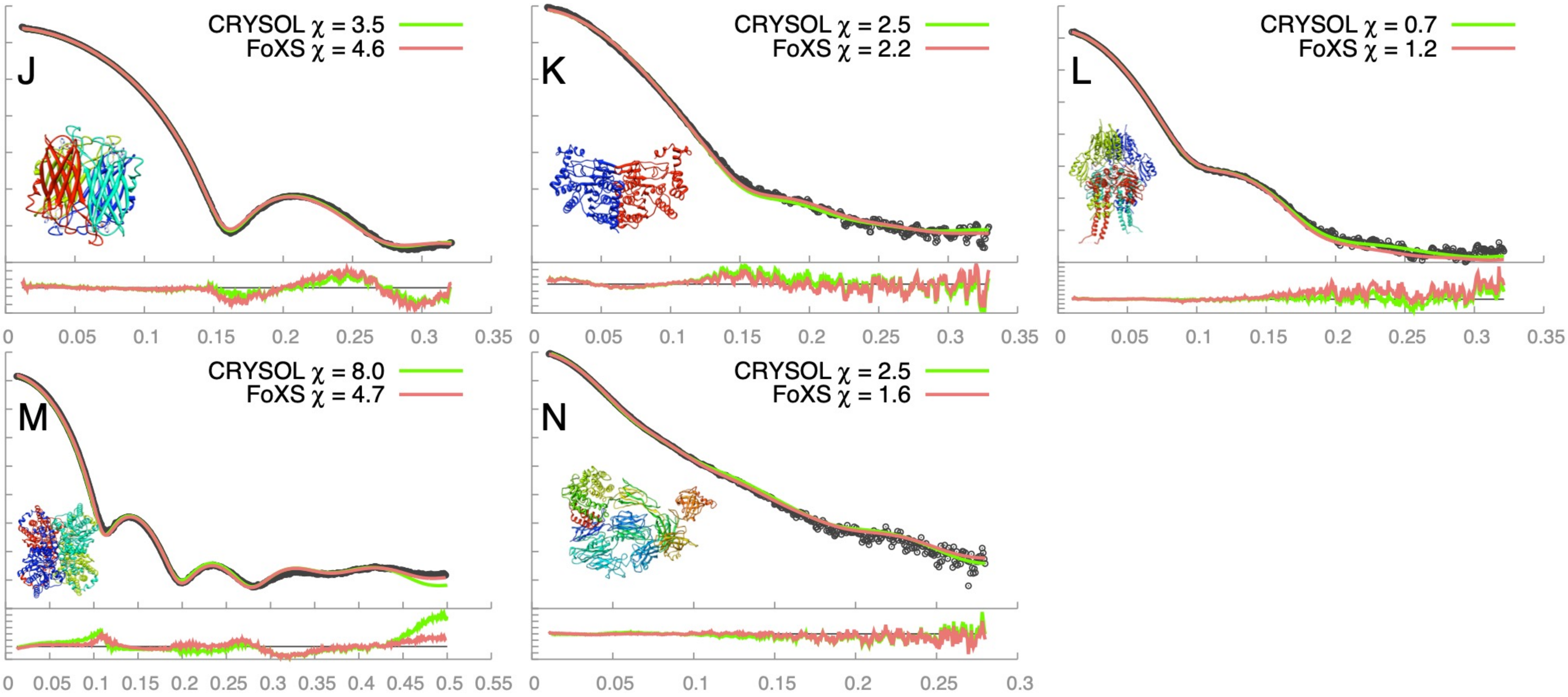
$0.32\ e/Å^3 \leq \rho \leq 0.38\ e/Å^3$

enumeration of 2 fitting parameters: c_1, c_2

X-ray structure vs. SAXS - good fits –> publish



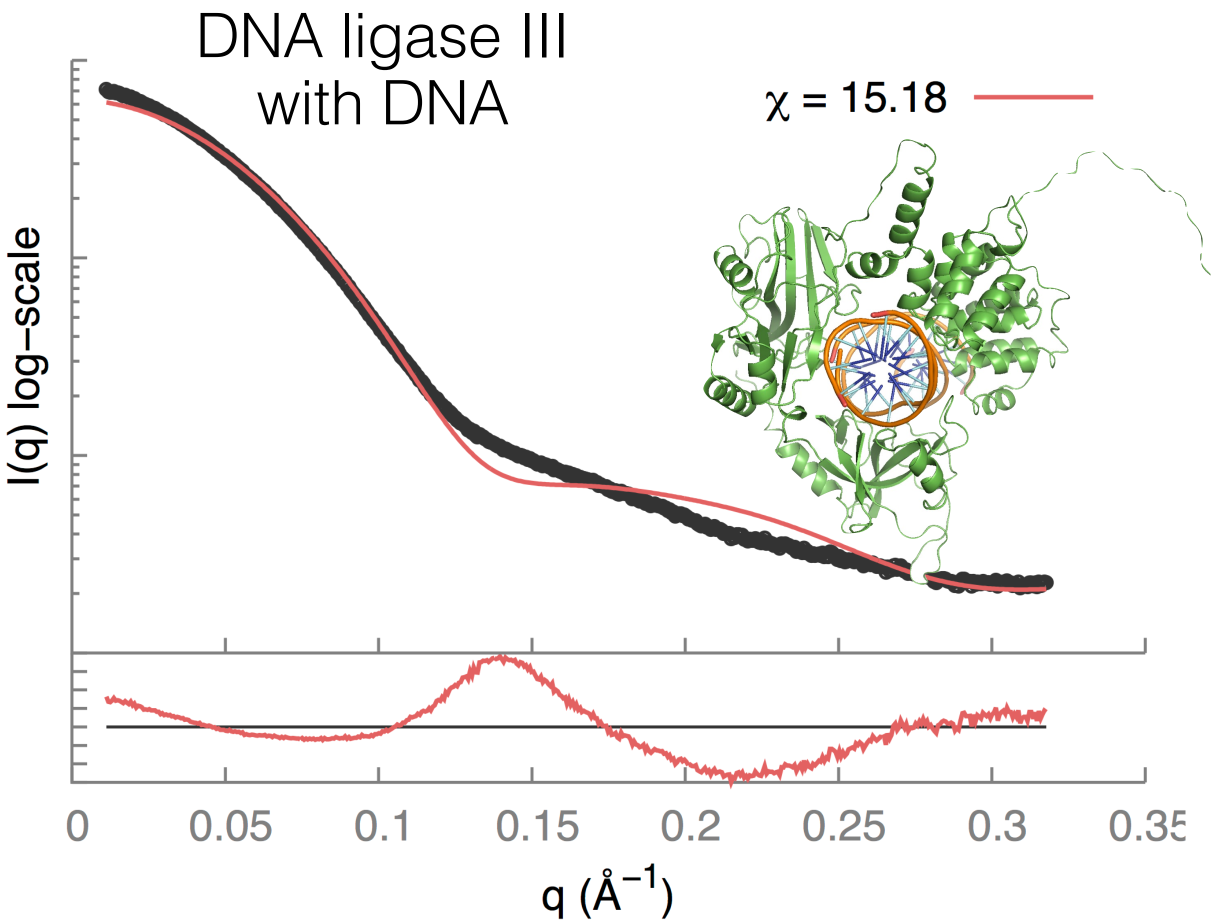
X-ray structure vs. SAXS - good fits –> publish

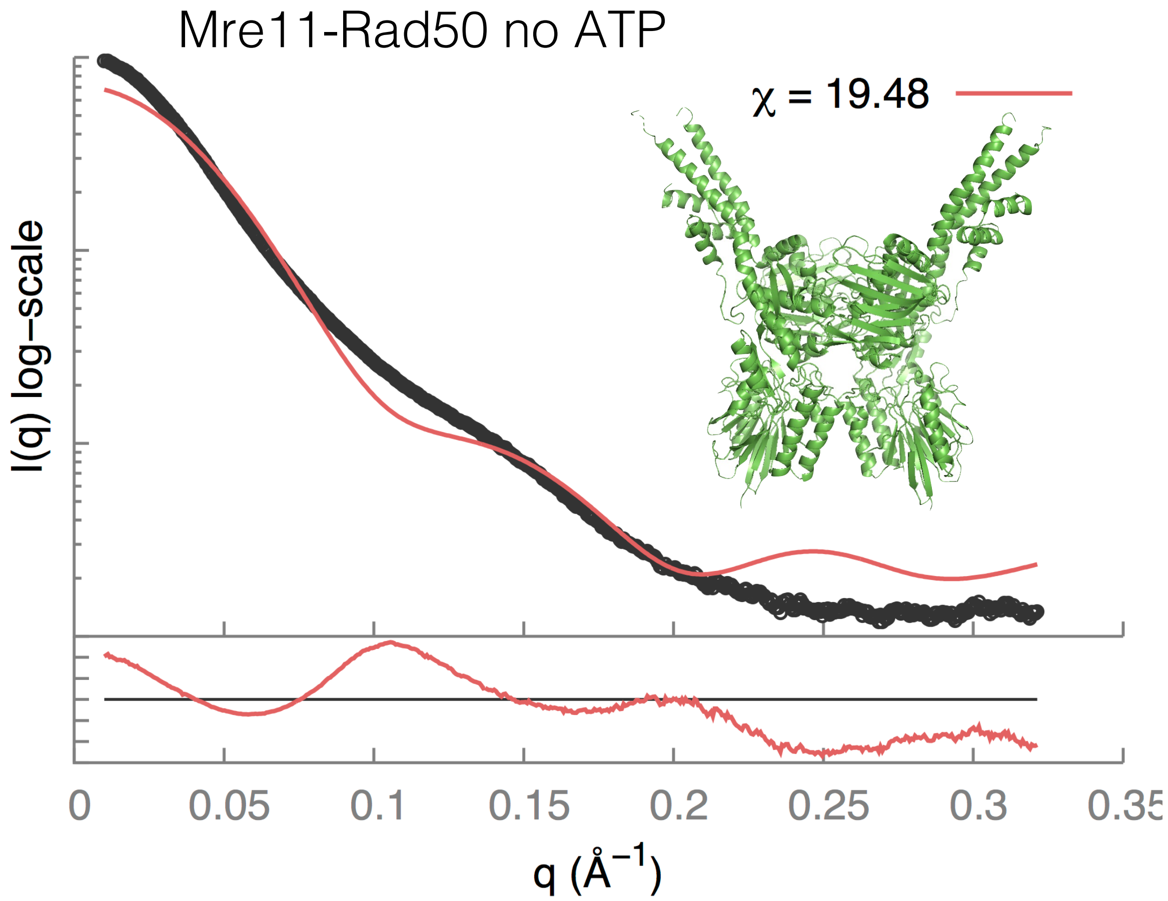


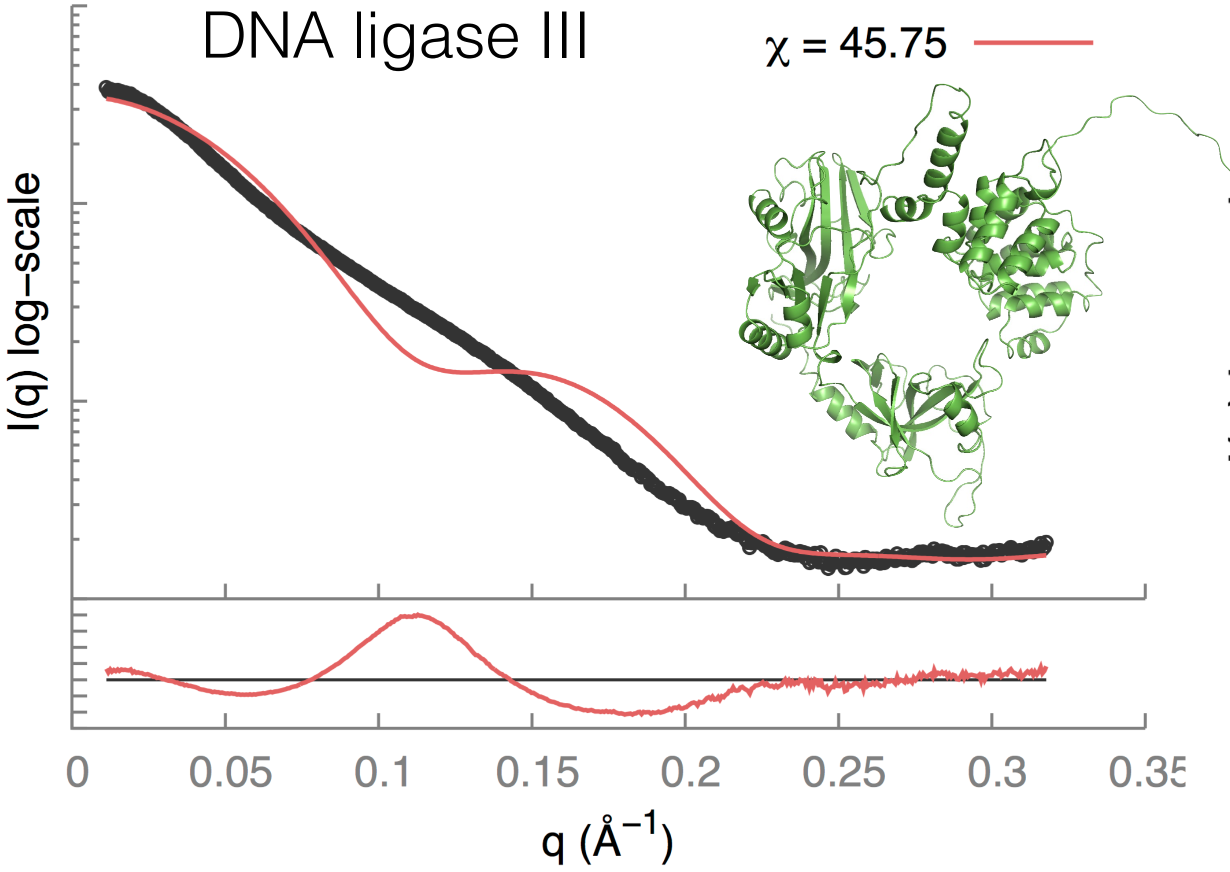
14 experimental datasets with x-ray structures

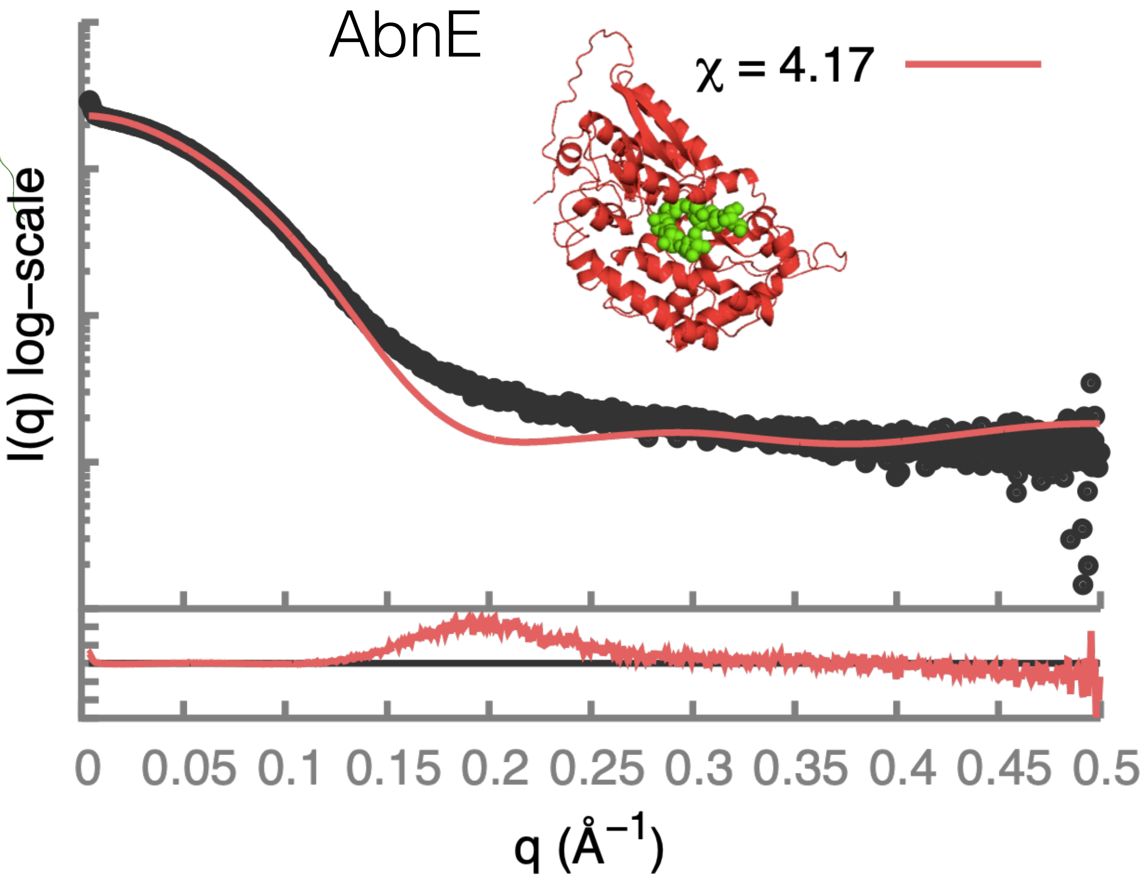
Schneidman-Duhovny D, Hammel M, Tainer J, Sali A. Biophys J 2013

X-ray structure vs. SAXS - they don't fit!


DNA ligase III with DNA

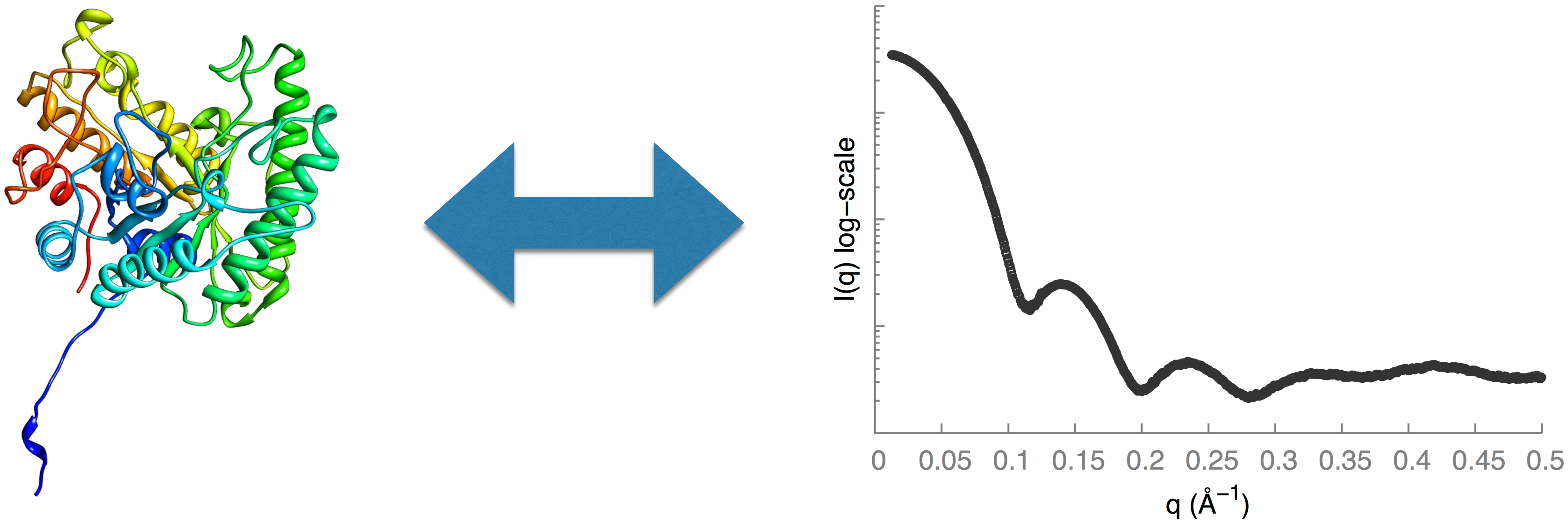

Mre11-Rad50 no ATP

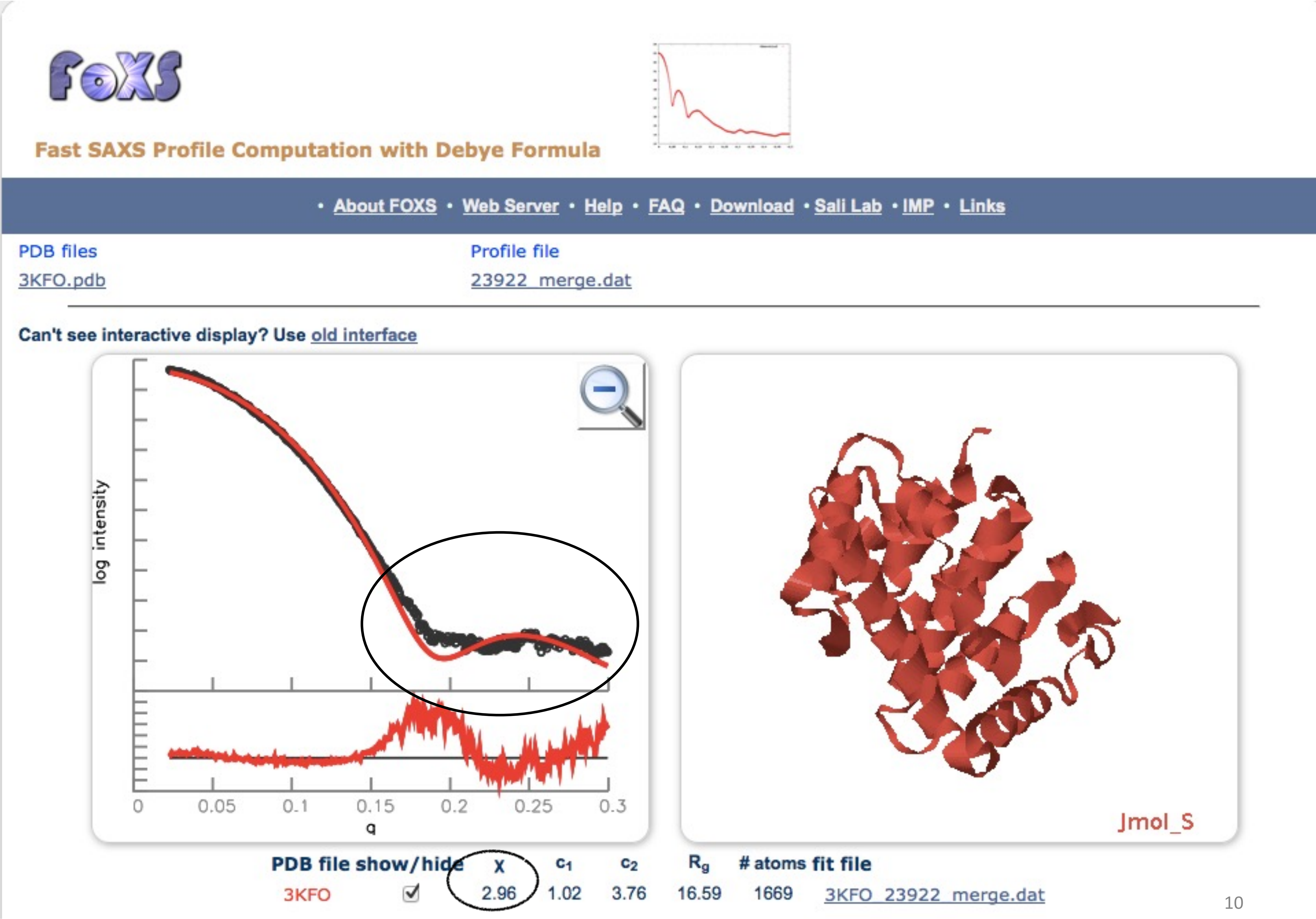

DNA ligase III


AbnE

X-ray structure vs. SAXS

- Data quality
- Missing residues/sugars
- Compositional heterogeneity
- Conformational heterogeneity
- both



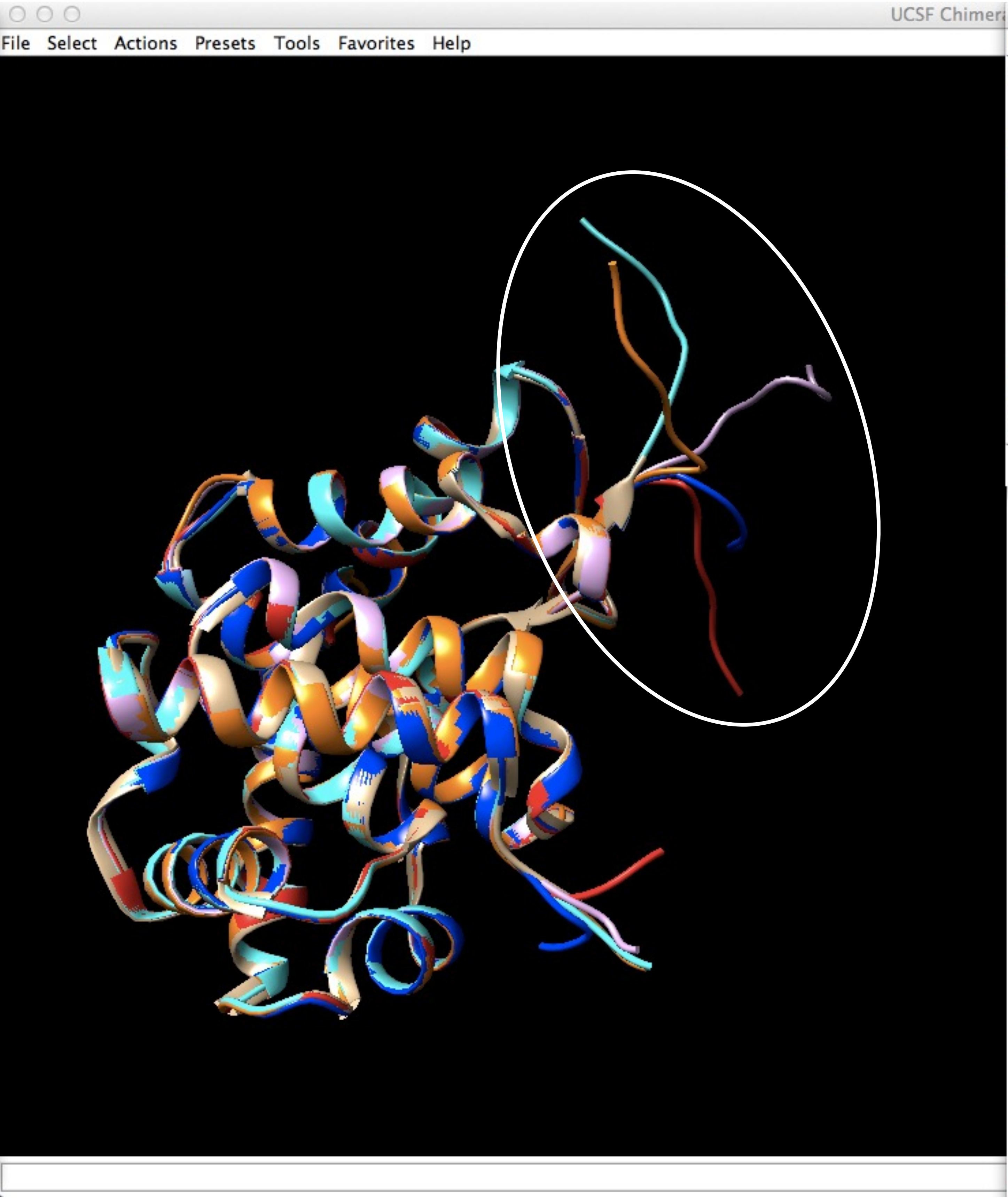


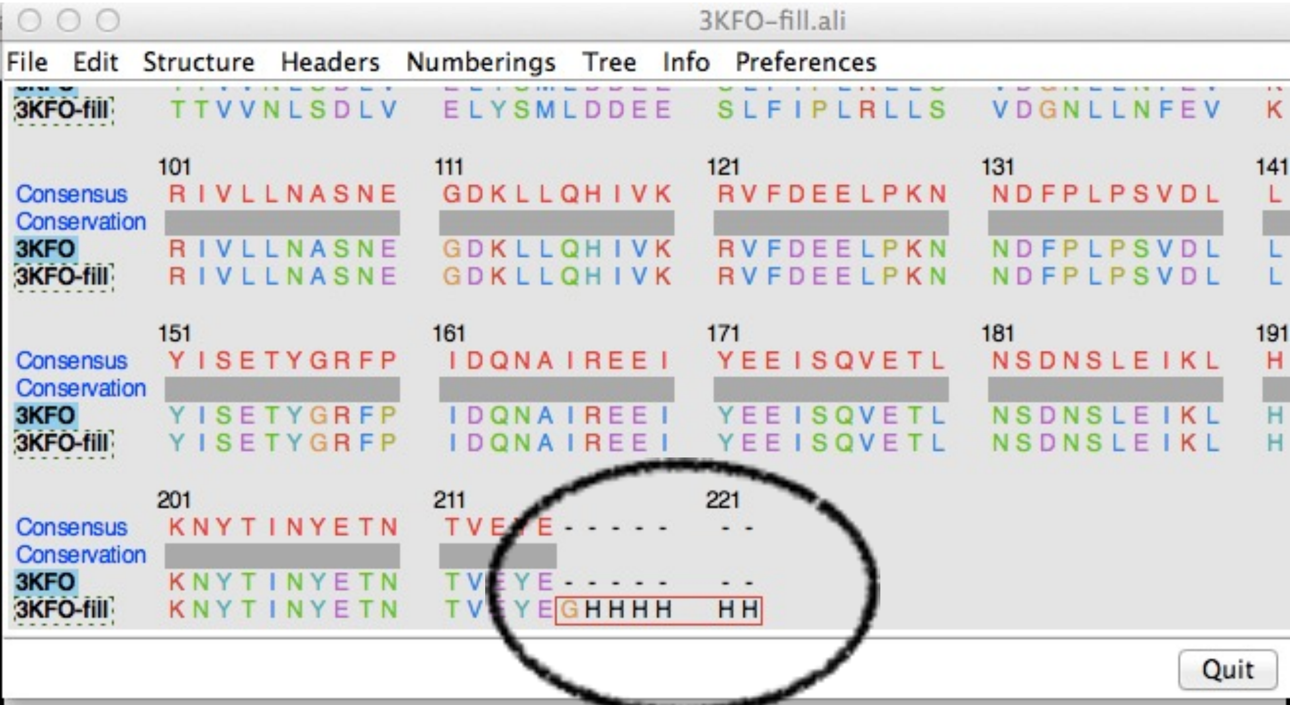
Fast SAXS Profile Computation with Debye Formula

PDB files	Profile file
3KFO.pdb	23922_merge.dat

Can't see interactive display? Use old interface

PDB file	show/hide	χ	c_1	c_2	R_g	# atoms	fit file
3KFO	☑	2.96	1.02	3.76	16.59	1669	3KFO_23922_merge.dat



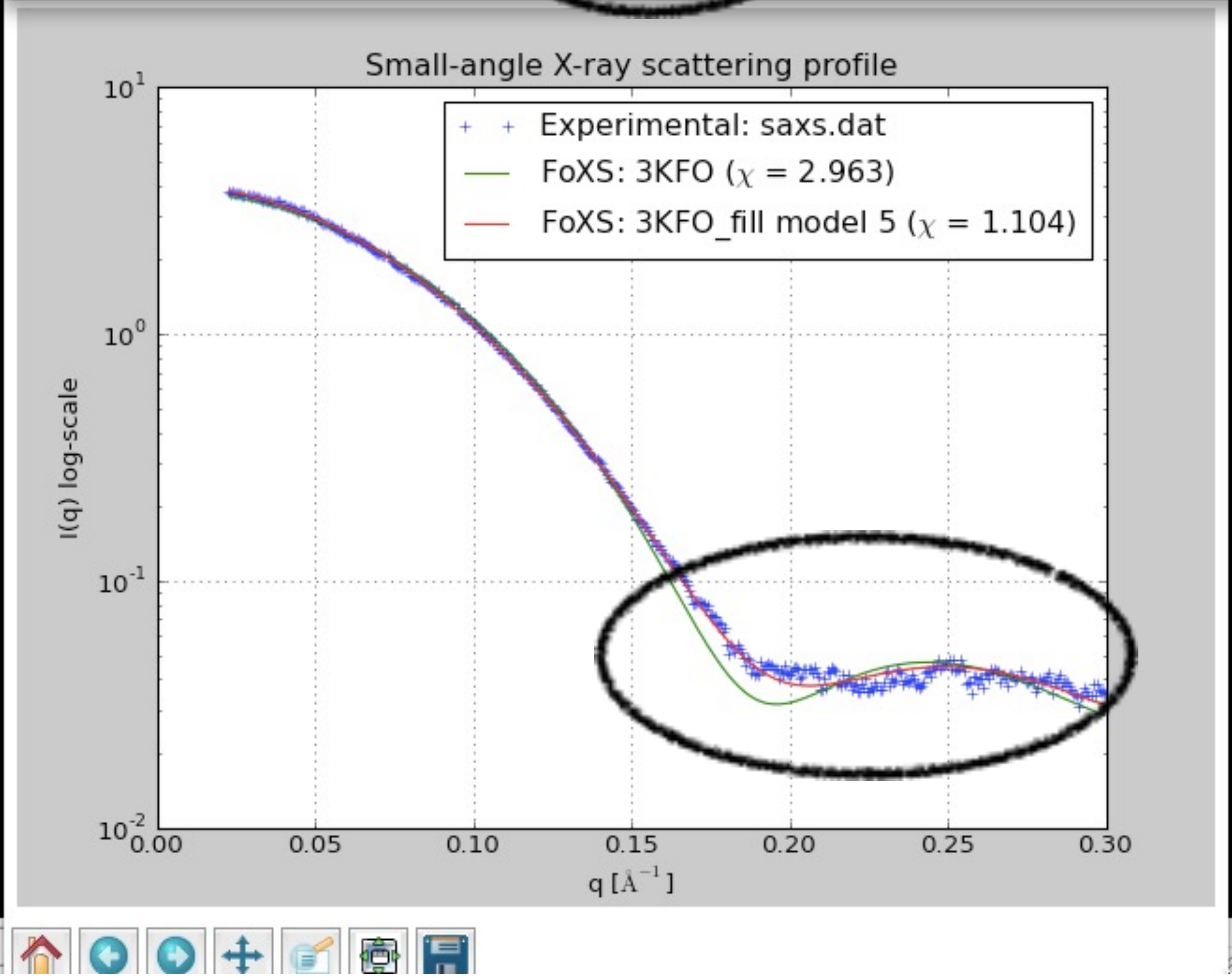


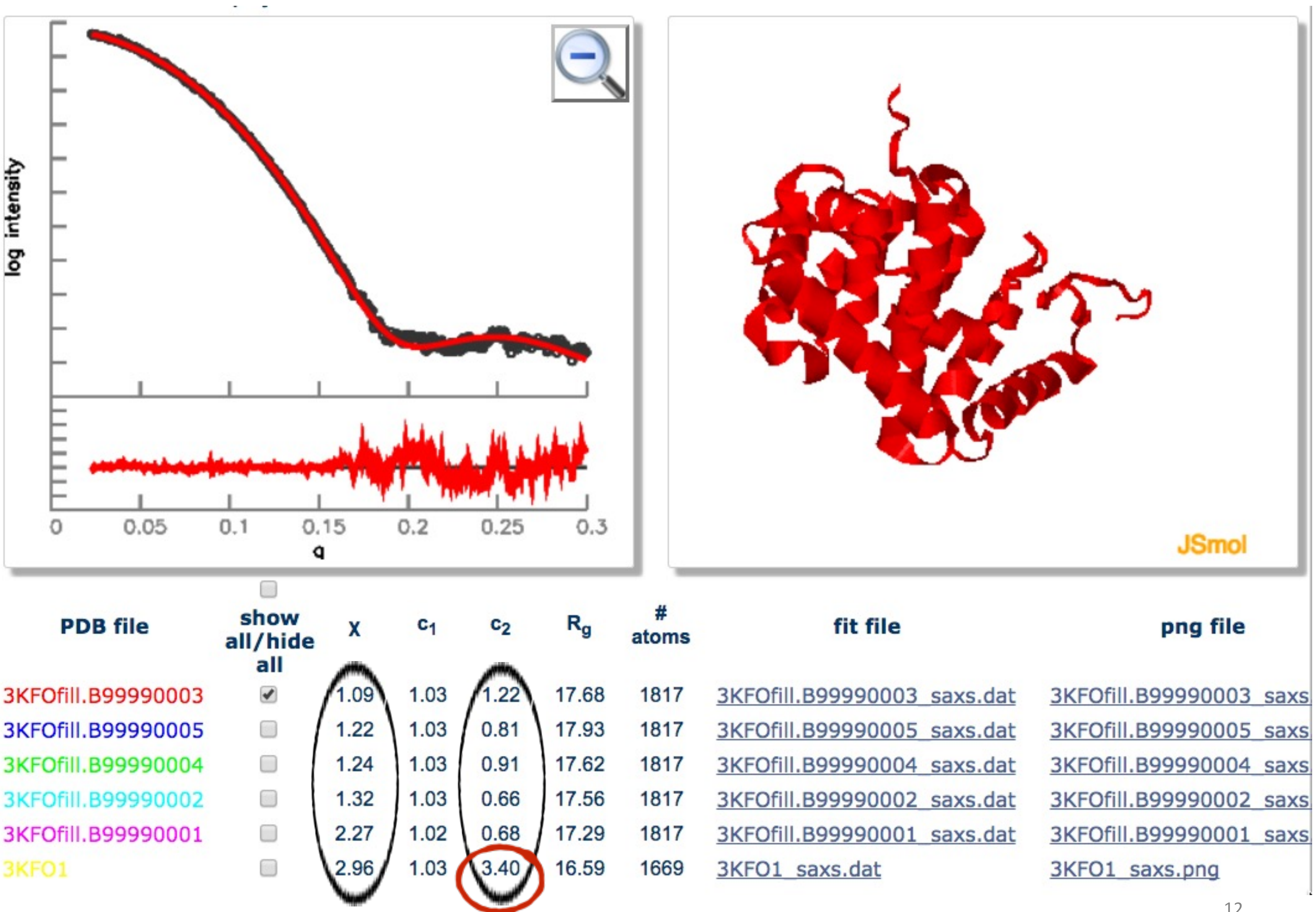
```
3KFO-fill    TTVVNLSDLV  ELYSMLDDEE  SLFIPLRLLS  VDGNLLNFEV  K

             101         111         121         131         141
Consensus    RIVLLNASNE  GDKLLQHIVK  RVFDEELPKN  NDFPLPSVDL  L
Conservation
3KFO         RIVLLNASNE  GDKLLQHIVK  RVFDEELPKN  NDFPLPSVDL  L
3KFO-fill    RIVLLNASNE  GDKLLQHIVK  RVFDEELPKN  NDFPLPSVDL  L

             151         161         171         181         191
Consensus    YISETYGRFP  IDQNAIREEI  YEEISQVETL  NSDNSLEIKL  H
Conservation
3KFO         YISETYGRFP  IDQNAIREEI  YEEISQVETL  NSDNSLEIKL  H
3KFO-fill    YISETYGRFP  IDQNAIREEI  YEEISQVETL  NSDNSLEIKL  H

             201         211         221
Consensus    KNYTINYETN  TVEYE-----  --
Conservation
3KFO         KNYTINYETN  TVEYE-----  --
3KFO-fill    KNYTINYETN  TVEYEGHHHH  HH
```

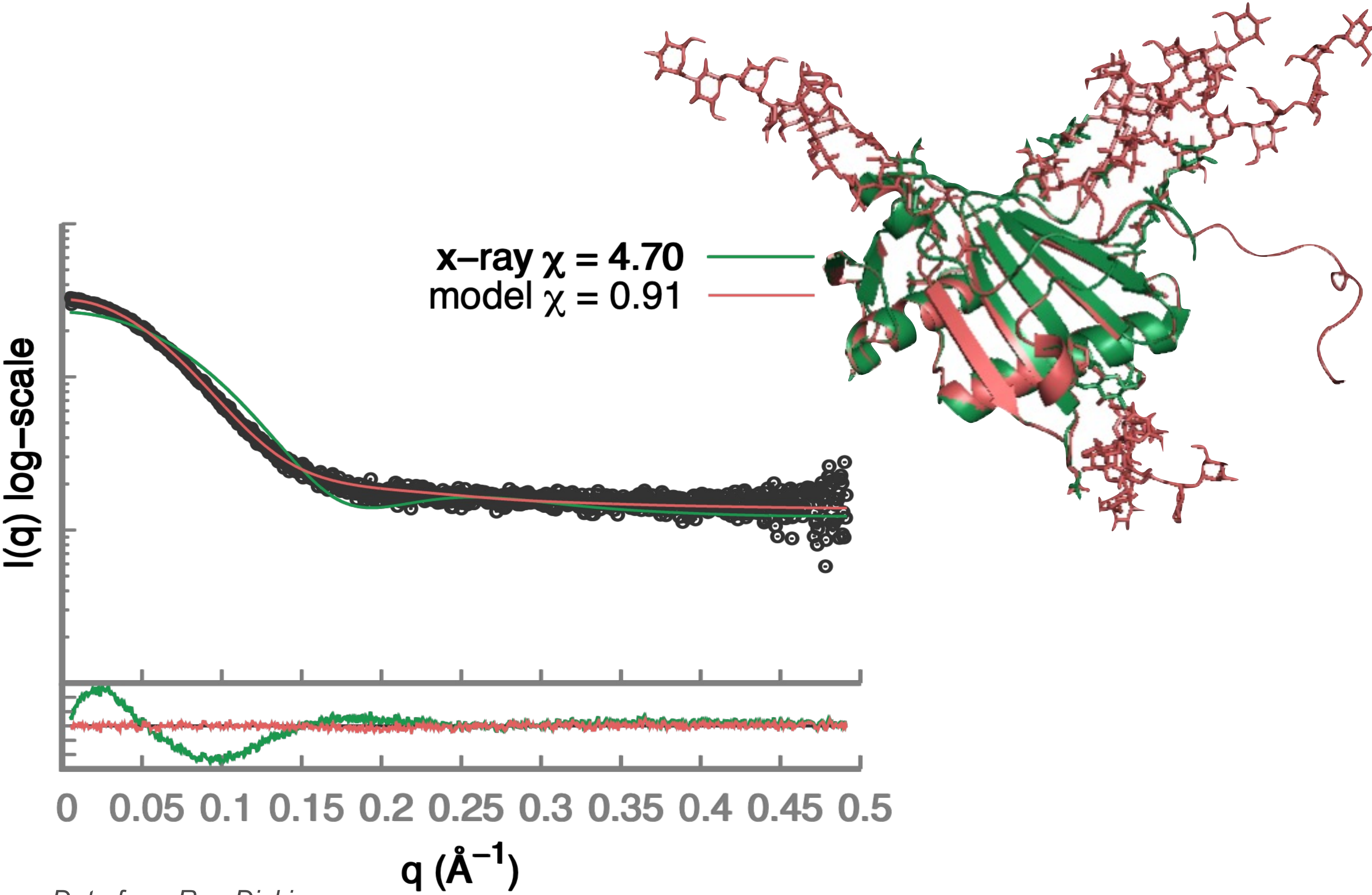




PDB file	☐ show all/hide all	χ	c_1	c_2	R_g	# atoms	fit file	png file
3KFOfill.B99990003	☑	1.09	1.03	1.22	17.68	1817	3KFOfill.B99990003_saxs.dat	3KFOfill.B99990003_saxs
3KFOfill.B99990005	☐	1.22	1.03	0.81	17.93	1817	3KFOfill.B99990005_saxs.dat	3KFOfill.B99990005_saxs
3KFOfill.B99990004	☐	1.24	1.03	0.91	17.62	1817	3KFOfill.B99990004_saxs.dat	3KFOfill.B99990004_saxs
3KFOfill.B99990002	☐	1.32	1.03	0.66	17.56	1817	3KFOfill.B99990002_saxs.dat	3KFOfill.B99990002_saxs
3KFOfill.B99990001	☐	2.27	1.02	0.68	17.29	1817	3KFOfill.B99990001_saxs.dat	3KFOfill.B99990001_saxs
3KFO1	☐	2.96	1.03	3.40	16.59	1669	3KFO1_saxs.dat	3KFO1_saxs.png

overfitting $0.32\ e/Å^3 \leq \rho \leq 0.38\ e/Å^3$ → $c_2 < 4.0$

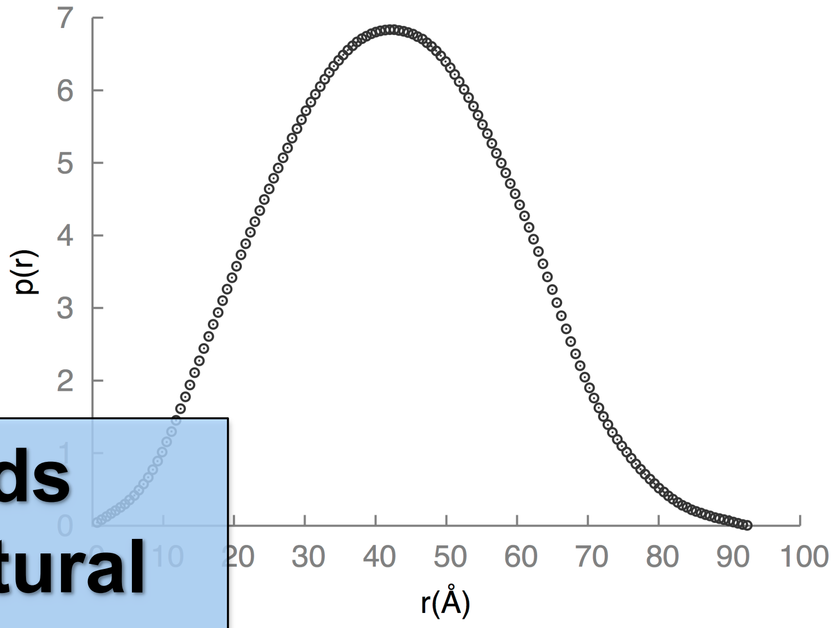
LASSA GP1



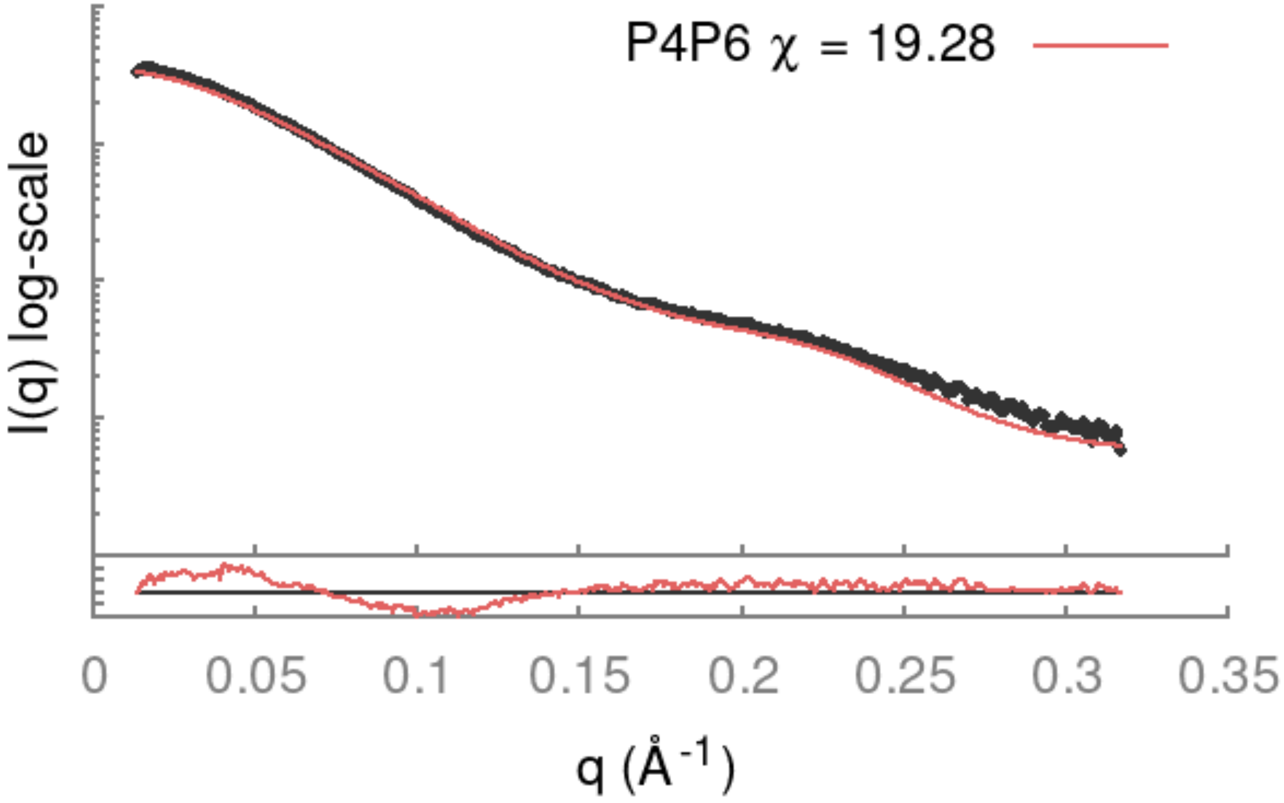
Data from Ron Diskin

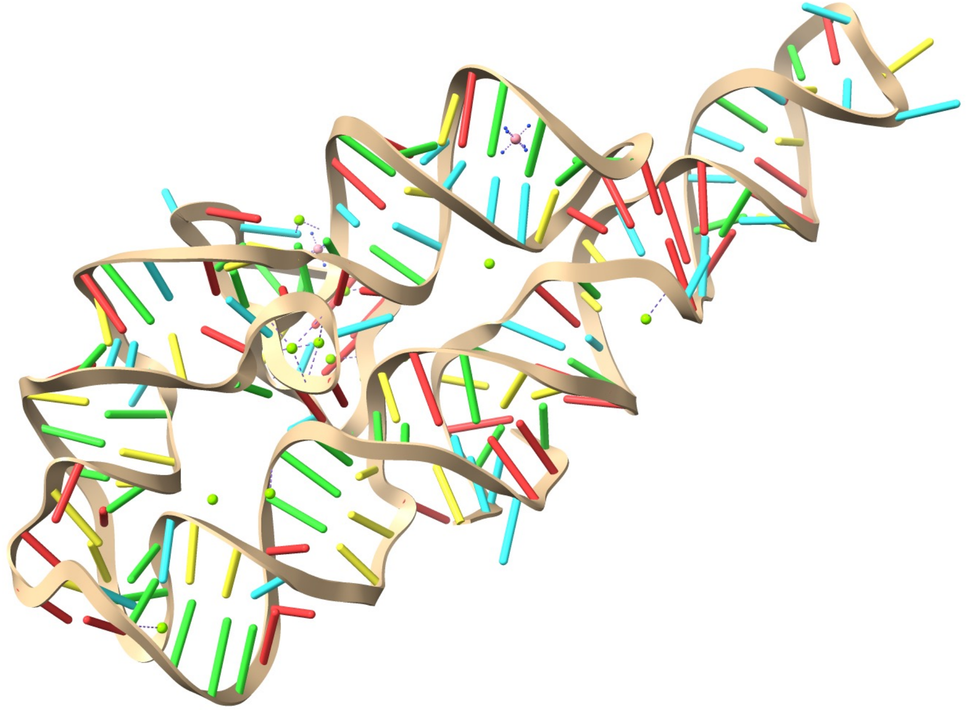
SAXS profile calculator for RNA

- RNA binds Mg ions that are required for proper folding and charge neutralization
- Missing ion density results in incomplete distance distribution, can't fit experimental data



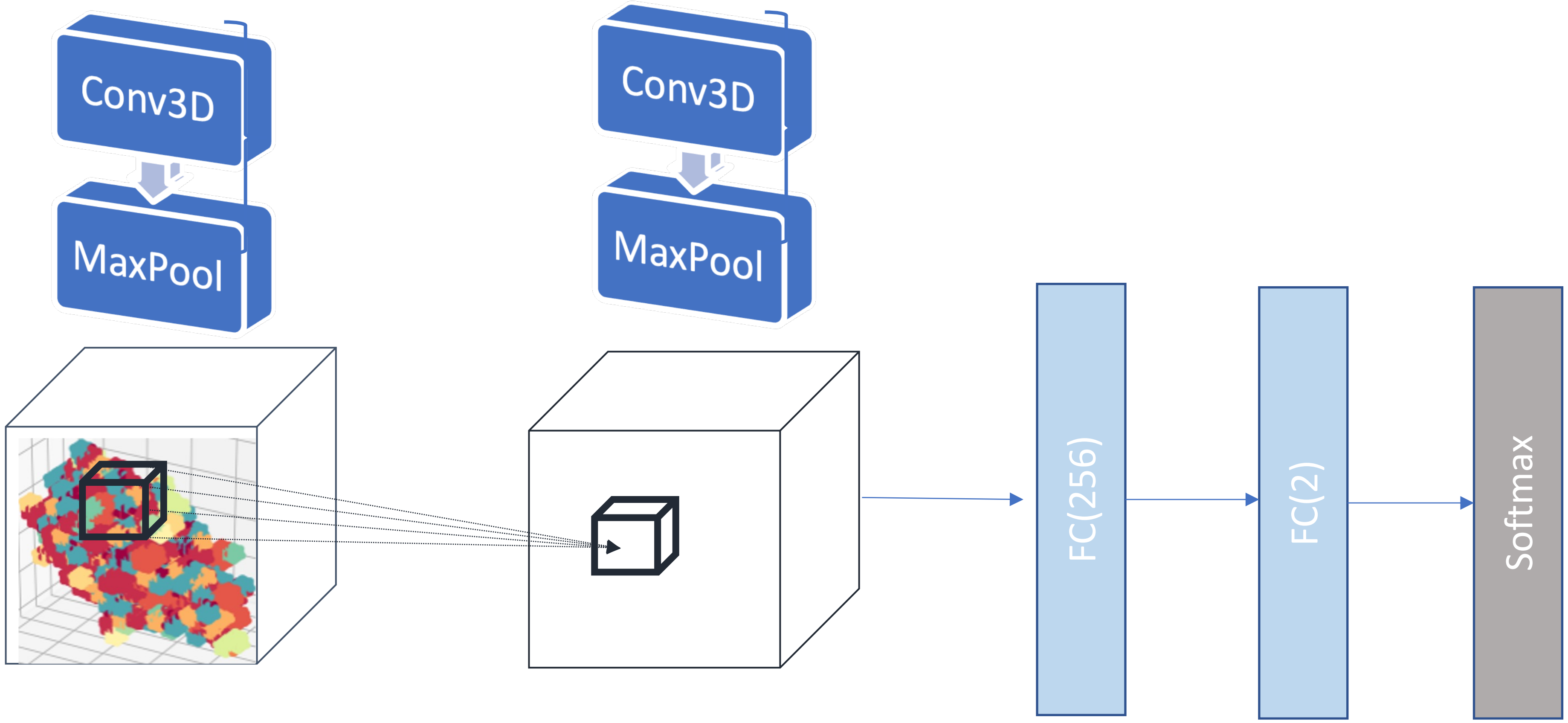
Can we predict Mg ion clouds positions given an RNA structural model?





MGclassifier: predict MG probability in a given 3D position

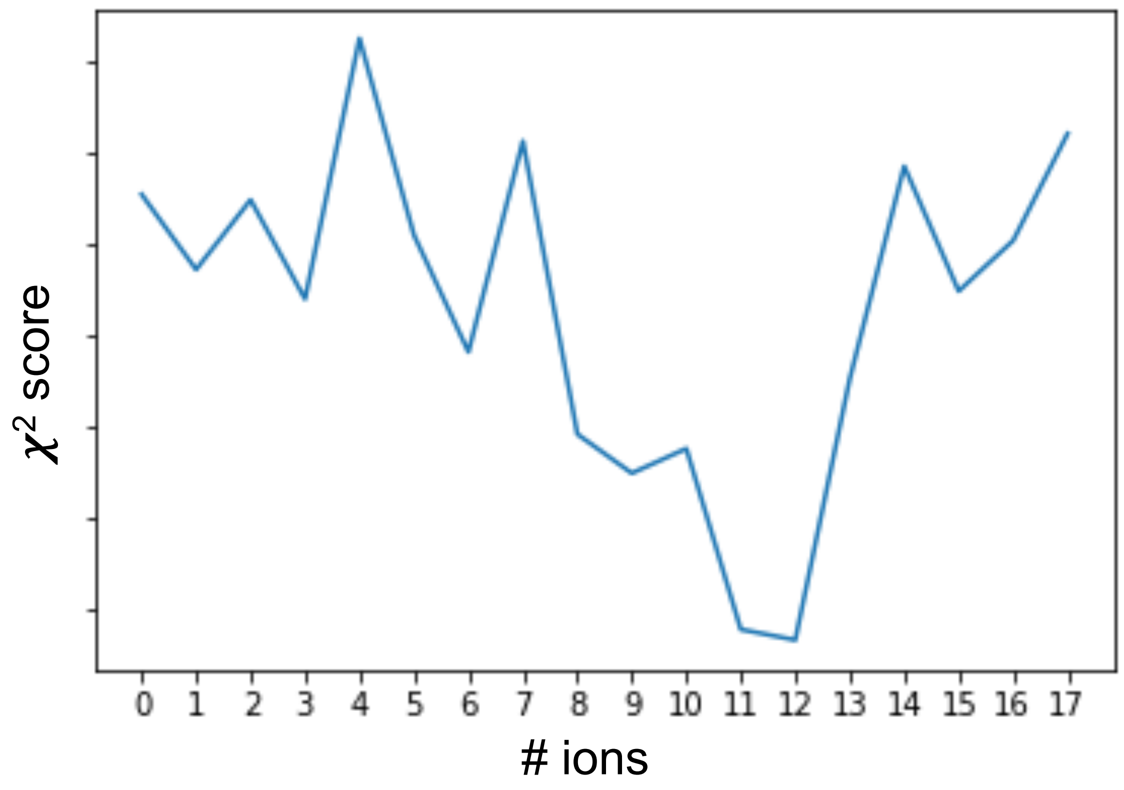
- deep 3D convolutional neural network (CNN), graph neural networks (GNN)
- each MG is represented by 32x32x32 cubes with 16 channels using mol2type representation and charge
- trained on ~100k examples, 50k positive examples and another 50k negative examples (water/probe molecules)

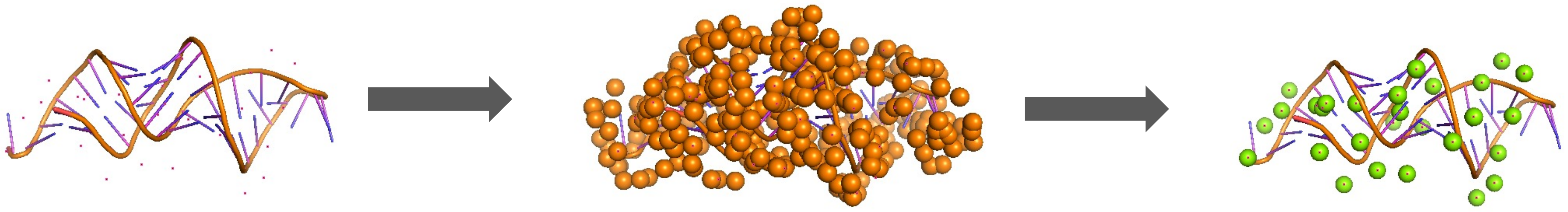


Edan Patt

Using MGclassifier to predict MG positions for an RNA structure

- RNA 3D structure is covered with probes that are classified by the model
- The probes are added to the RNA, starting with the most likely one
- Fit to the experimental SAXS profile is used to select the optimal number of ions

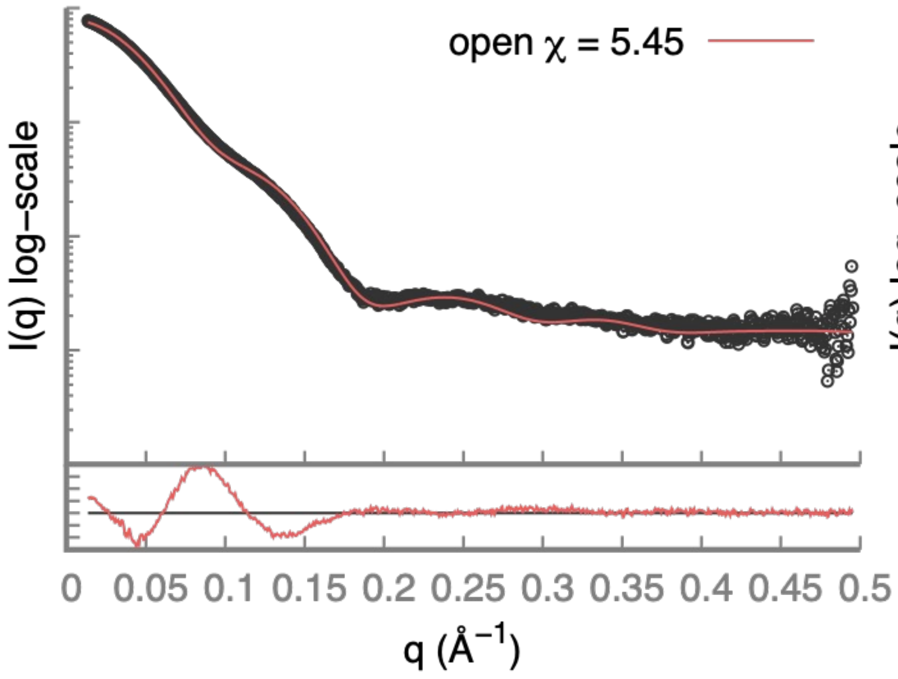


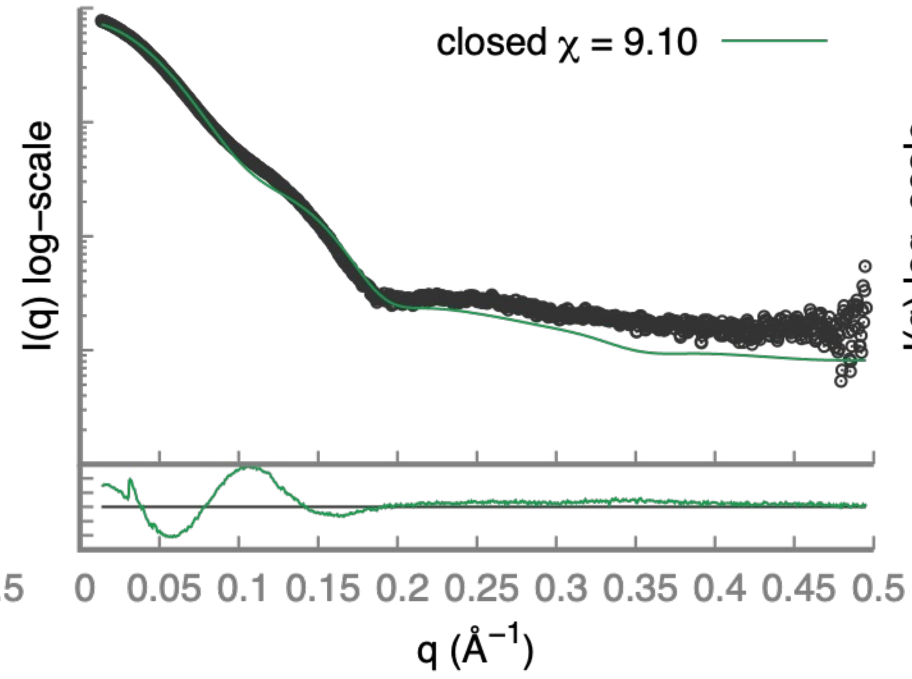


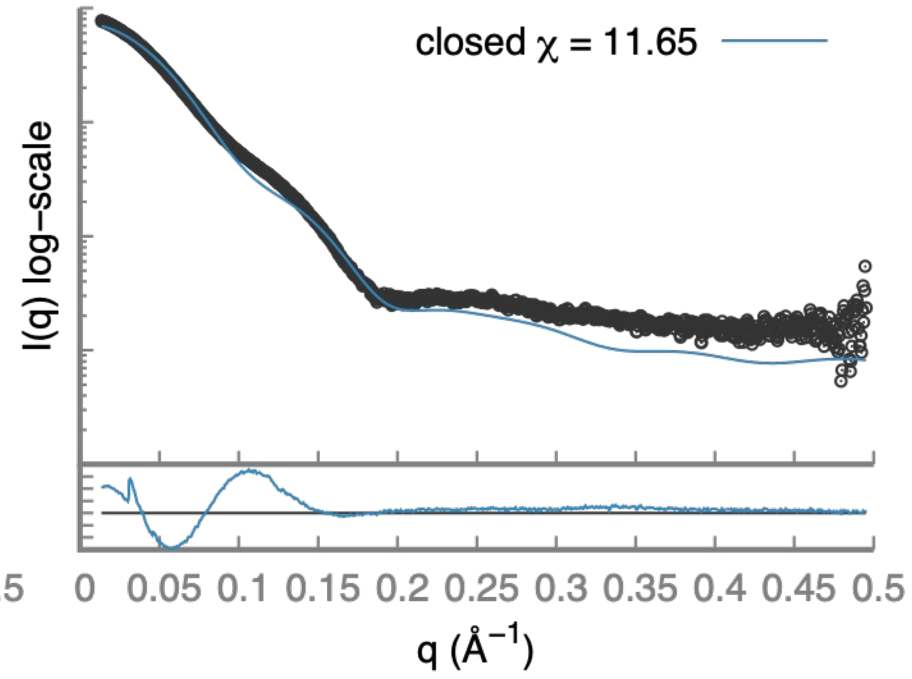
AbnA structures vs. SAXS

?

- 3 X-ray structures in different conformations do not fit the data

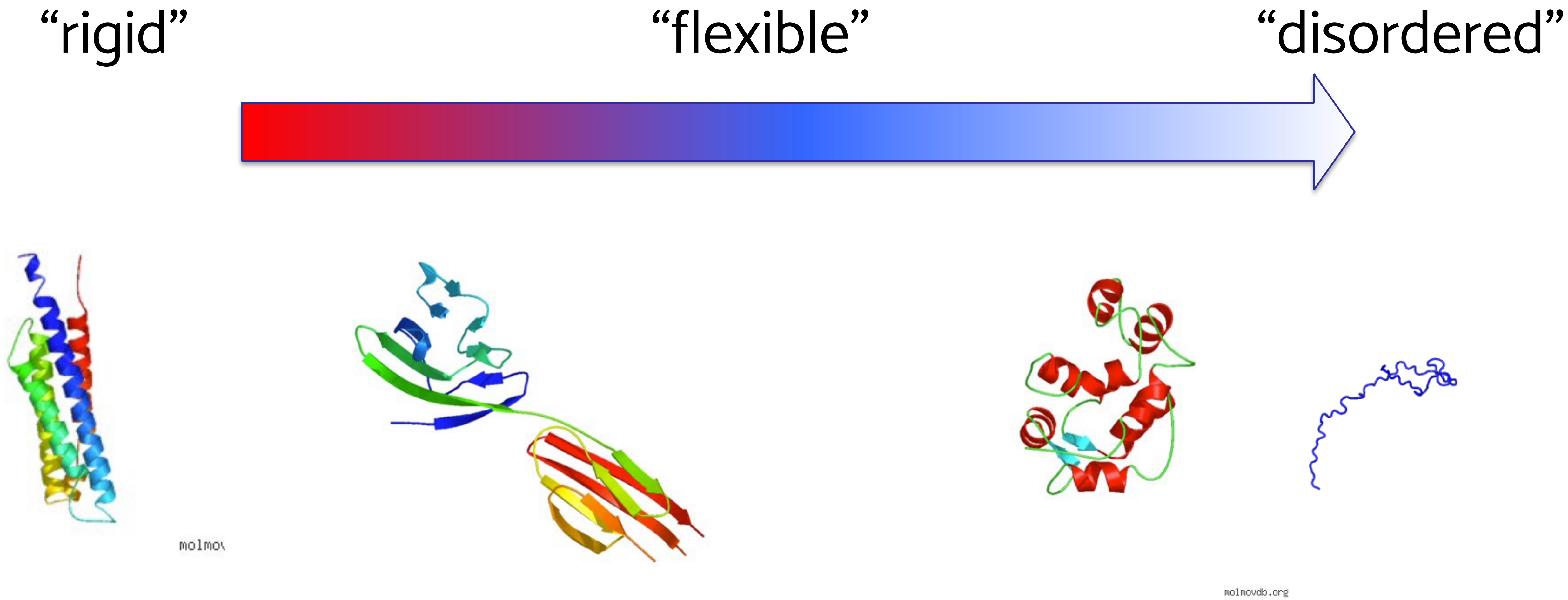






Collaboration with Shifra Lansky and Gil Shoham

Dynamics Comes in Flavors and it is Common



“rigid” “flexible” “disordered”

molmov

molmovdb.org

	short disordered fragments (≥10 and <30 residues)	long disordered fragments (> 40 residues)
PDB	***~ 40% of structures***	***~ 10% of structures***
SwissProt		> 25% of sequences

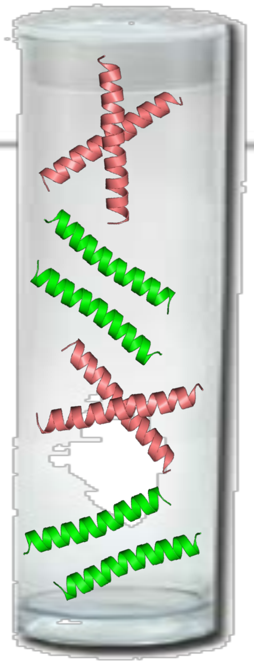
(Romero et al. 2001; Dunker et al 2000; Le Gall 2007)

Dynamics and SAXS

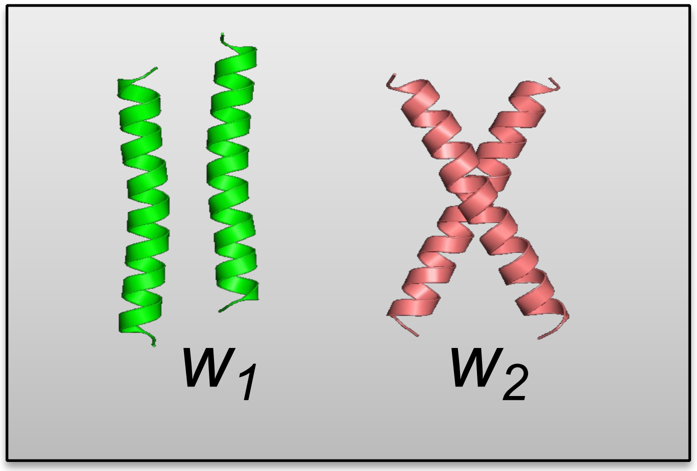
- SAXS data can be easily collected for proteins that include disordered regions
- Data interpretation is challenging

```
while(noSuccess)
{
        tryAgain();
        if(Dead)
                break;
}
```

Heterogeneous Sample Requires Multi-State Model

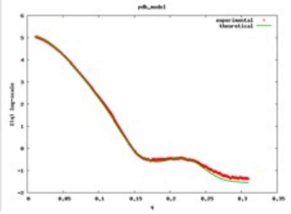


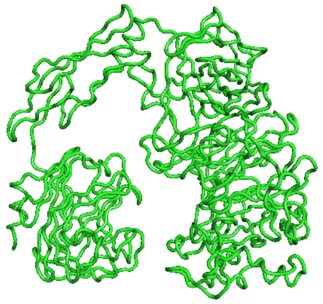
Heterogeneous sample	compositional or conformational heterogeneity in the sample used to generate the data
Multi-state model	a model that specifies two or more co-existing structural states and values for any other parameter



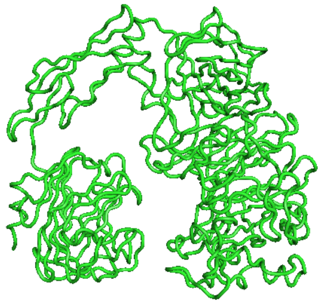
MultiFoXS

salilab.org/multifoxs





flexible residues:
35 A
36 A
...

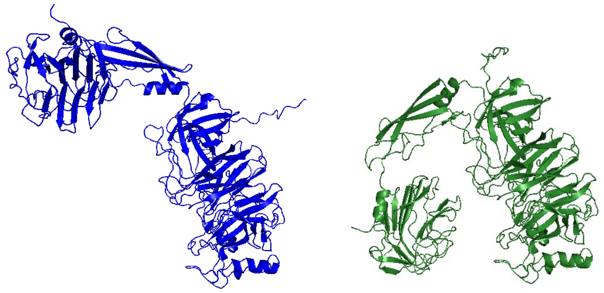


Sampling

Rapidly exploring Random Trees (RRTs)

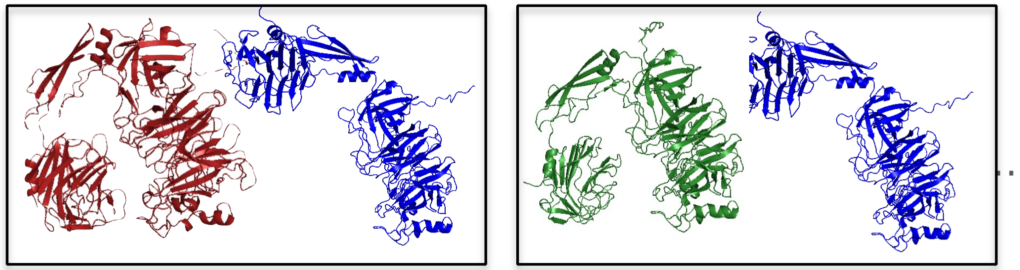
Scoring

Debye formula
salilab.org/foxs



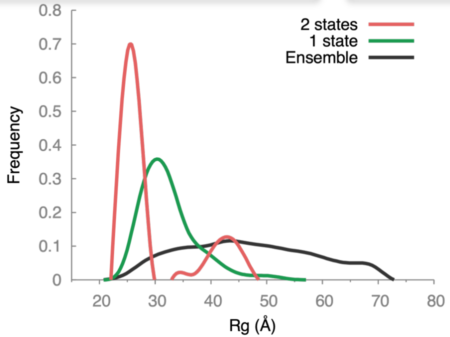
Enumeration

Enumeration of multi-state models that fit the data within noise



Analysis

Quality of fit to data, Rg variance among top scoring models



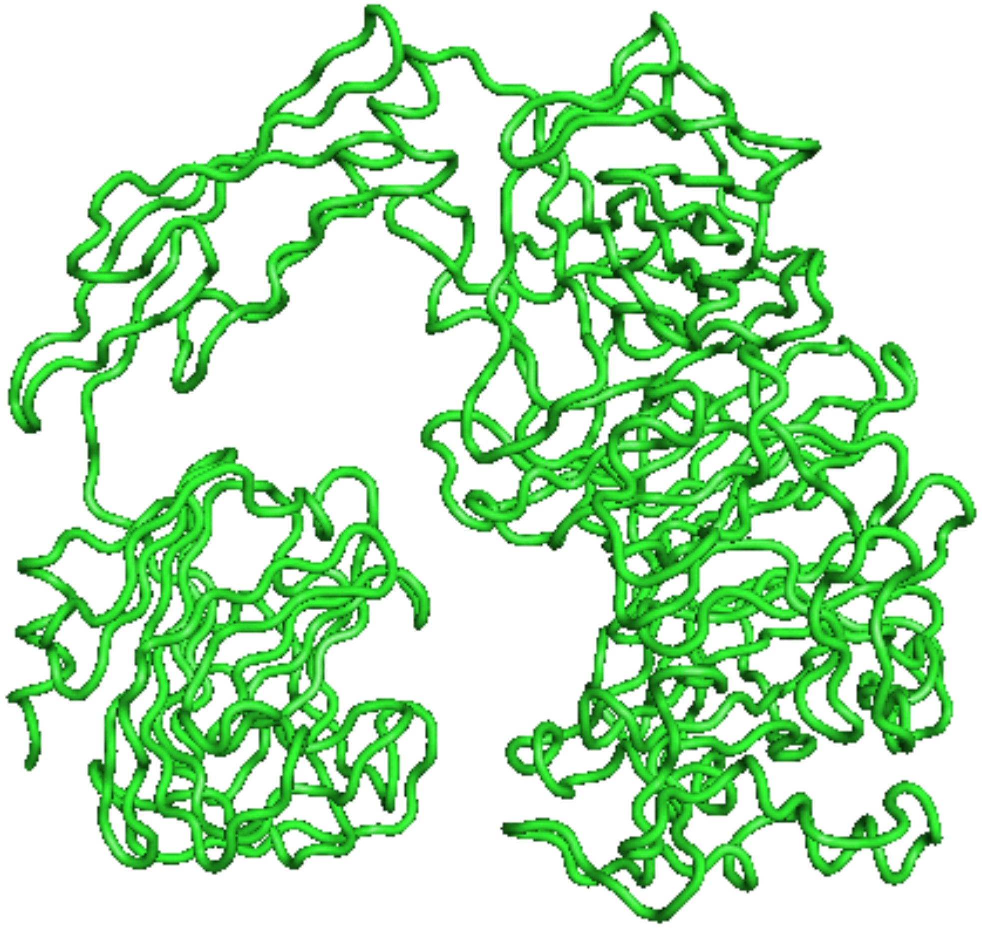
Schneidman-Duhovny, Hammel, Tainer, Sali. NAR 2016

Carter, Kim, Schneidman-Duhovny, Stoehr, Poncet-Montange, Tsuruta, Prusiner, Sali. Biophys J. 2015

Conformational sampling

Proteins and robots have similar degrees of freedom

Robotic arm



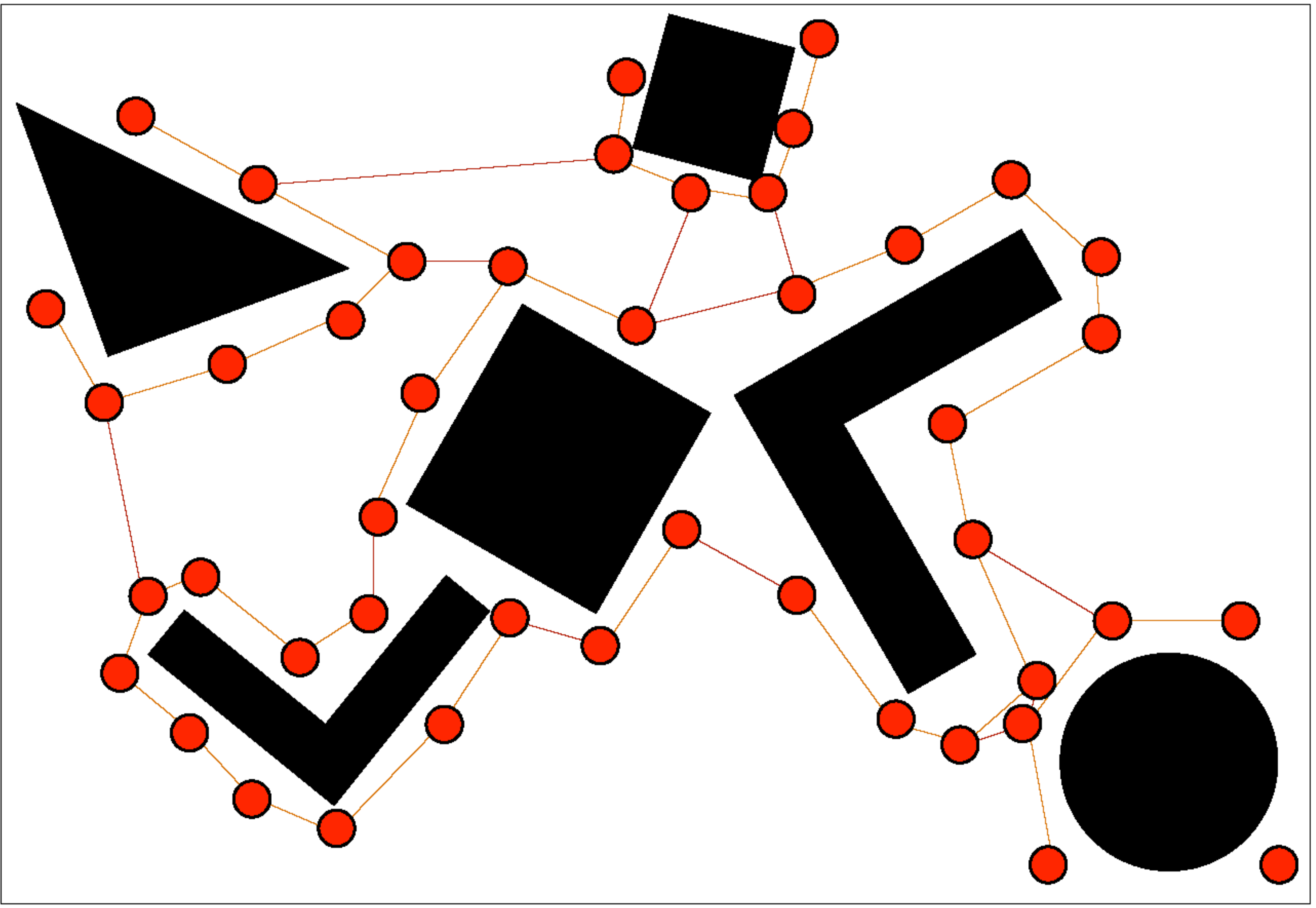
AbnA protein

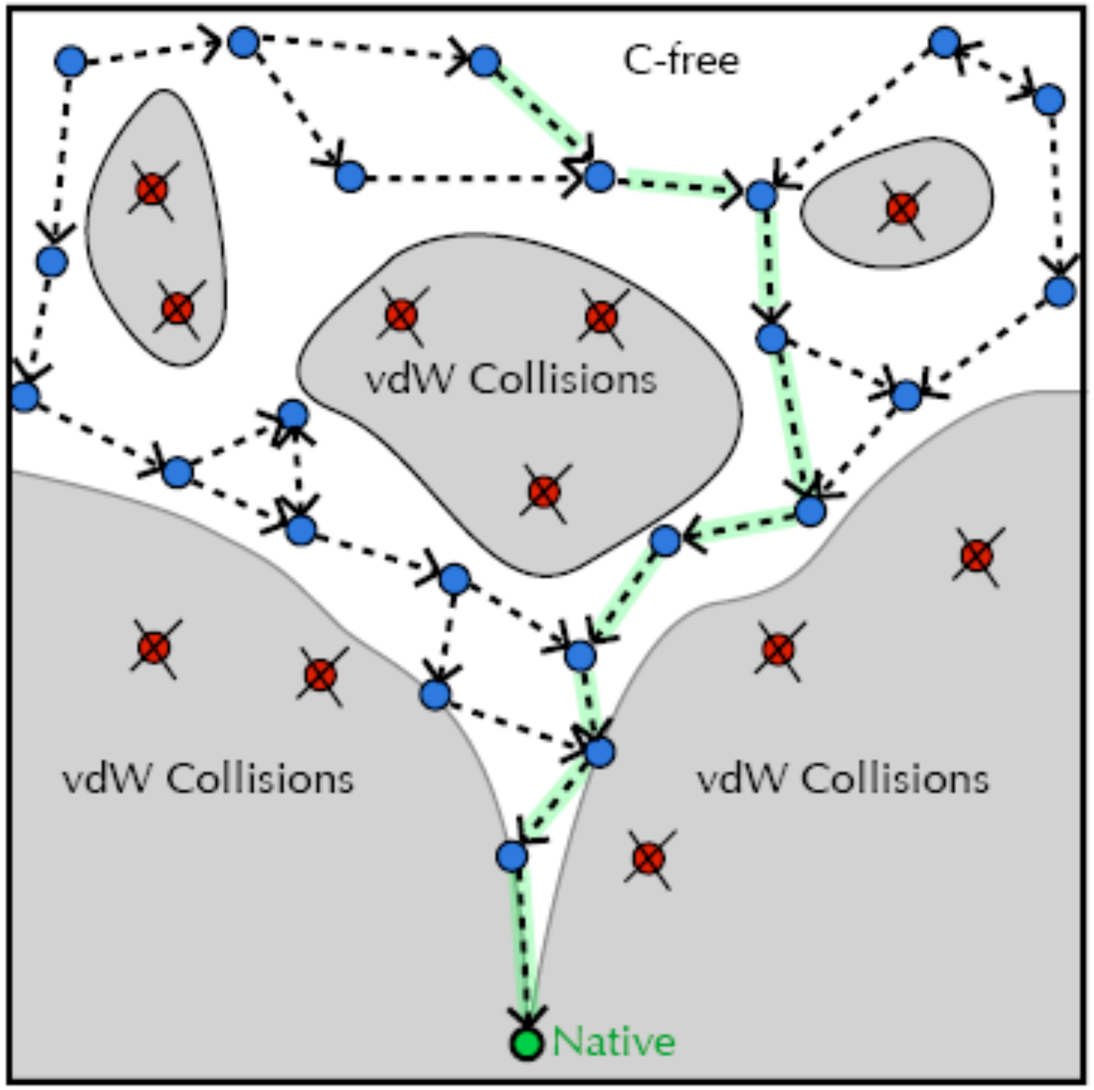
We rely on methods for Motion Planning developed in Robotics *(La Valle, Latomb, Kavraki, Cortes)*

Mapping collision free space with Rapidly exploring Random Tree (RRT)

Collision free space for robot

Collision free space for protein chain

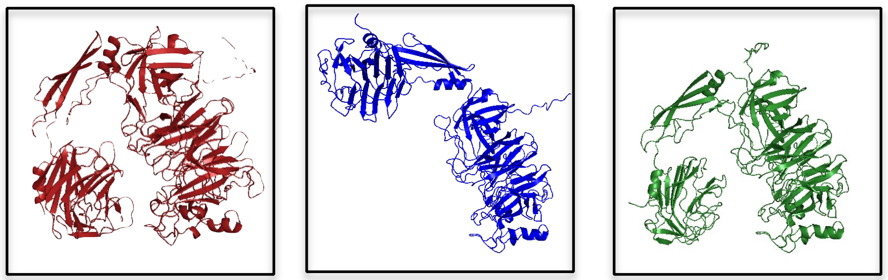




Enumeration of multi-state models

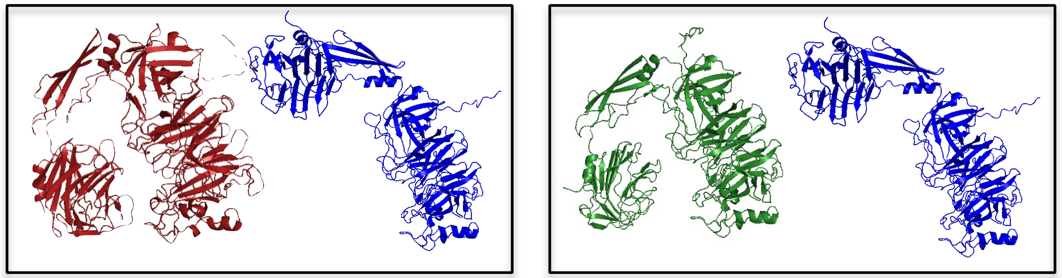
branch & bound deterministic algorithm
Multi-state models of size **i+1** are generated by extending **best K** (=10000) multi-state models of size **i**

best K multi-state models of size 1:



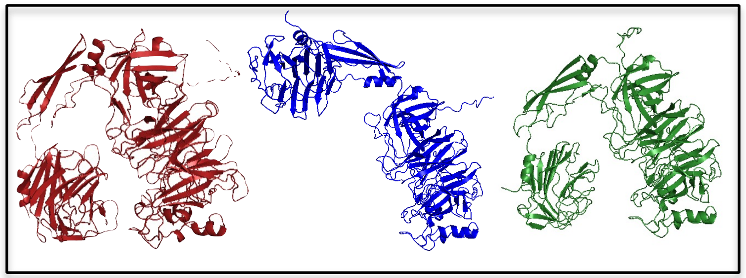
...

best K multi-state models of size 2:



...

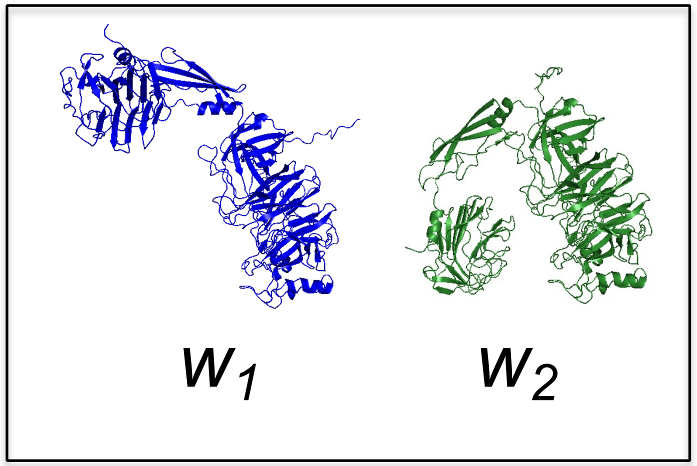
best K multi-state models of size 3:



...

...

Scoring of Multi-State Models



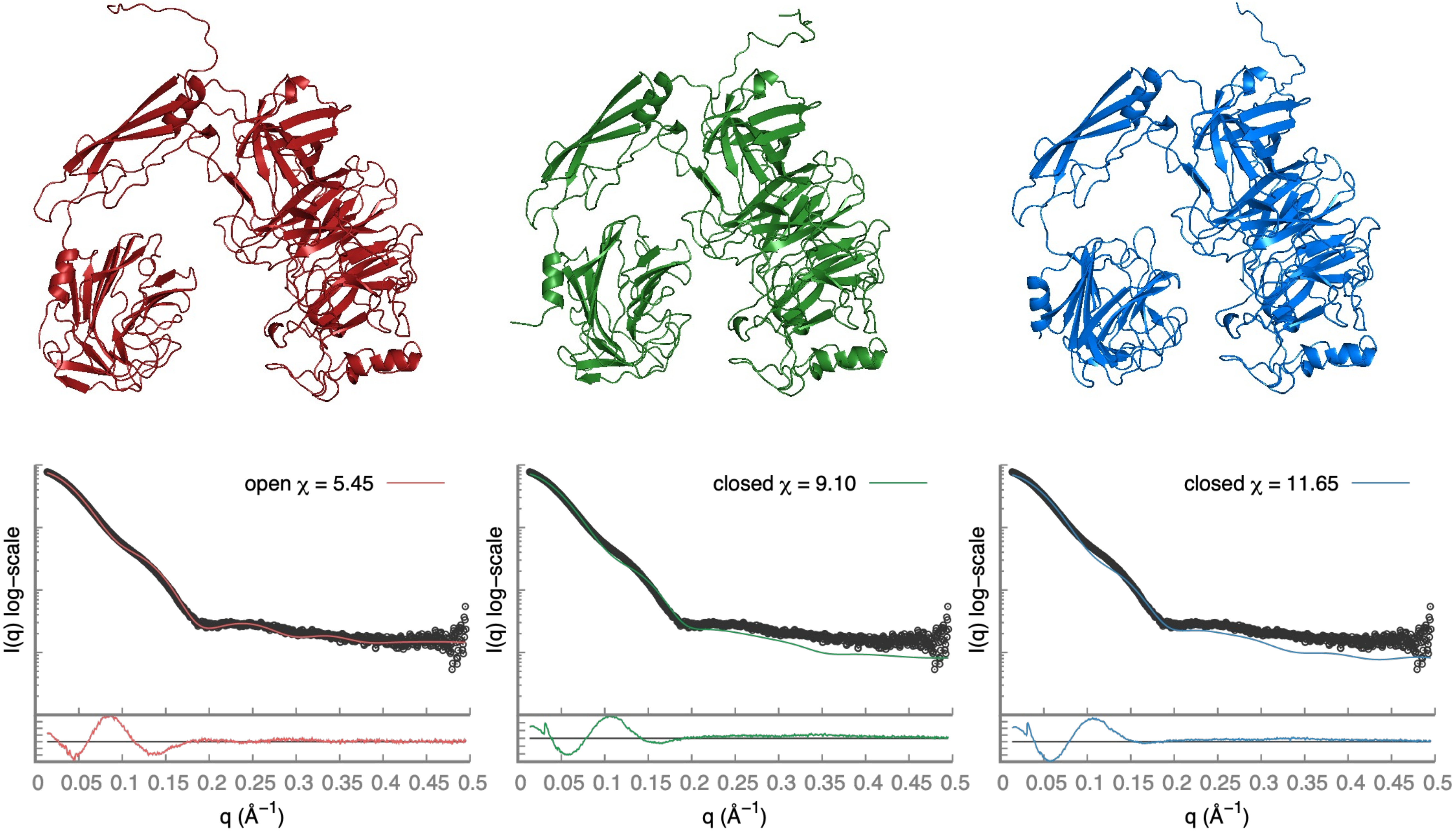
$$\chi = \sqrt{\frac{1}{S}\sum_{i=1}^{S}\left(\frac{I_{exp}(q_i) - c\sum_{n}\boxed{w_n} I_n(q_i, \boxed{c_1, c_2})}{\sigma(q_i)}\right)^2}$$

- weights optimization is needed for each set of structural states
- Non-negative least square fitting (NNLS, Lawson & Hanson 1974)

- c_1 (excluded volume), and c_2 (hydration layer) are enumerated
- a single pair of c_1 and c_2 is used for all states in a multi-state model

AbnA structures vs. SAXS

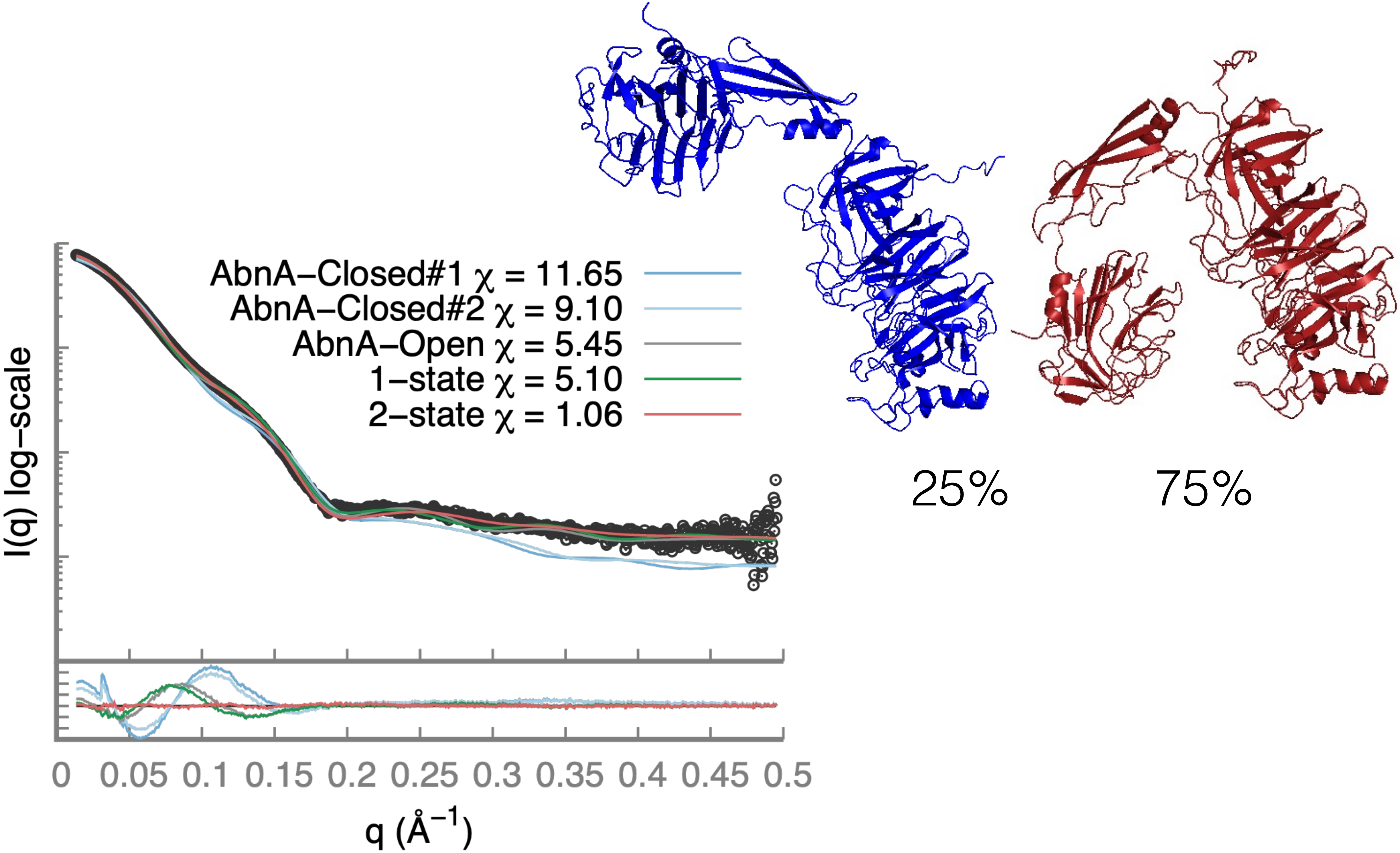
- 3 X-ray structures in different conformations do not fit the data



Collaboration with Shifra Lansky and Gil Shoham

Multi-state Modeling

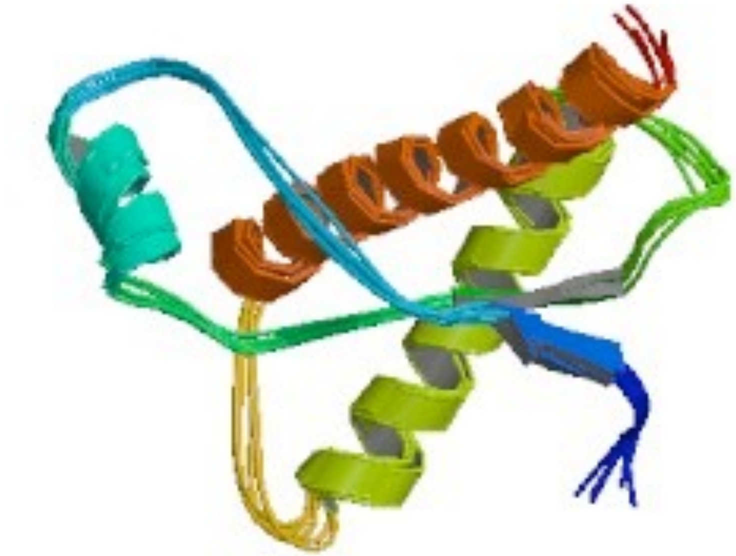
- Good fit to data obtained with open and closed conformations



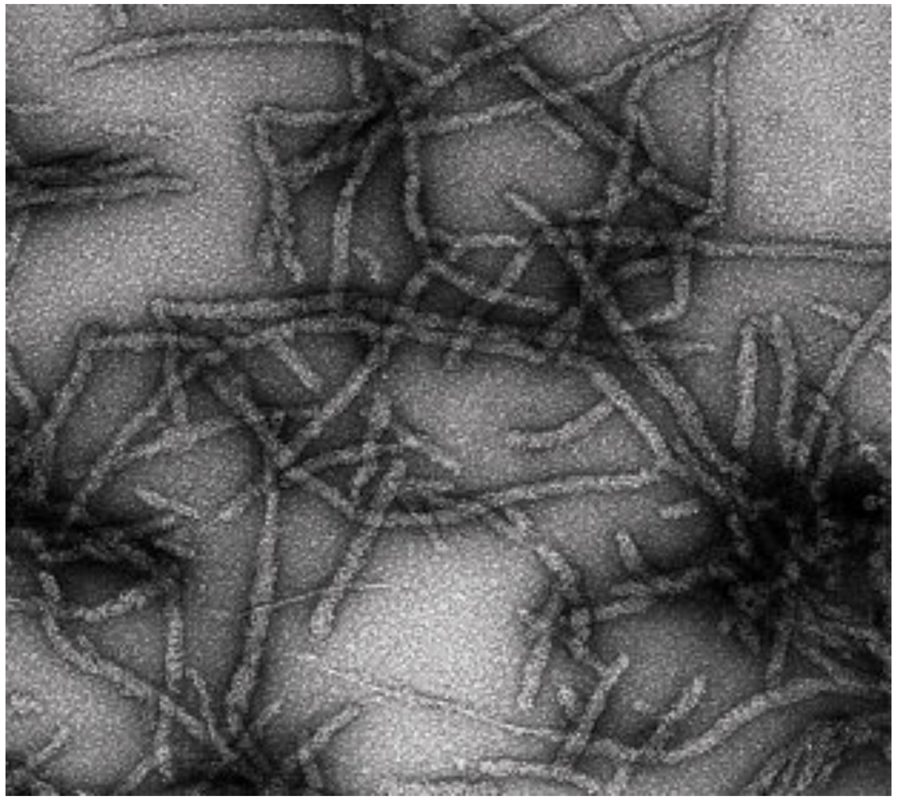
Application to Prion Protein

disordered fragment	structured fragment

47% of the sequence



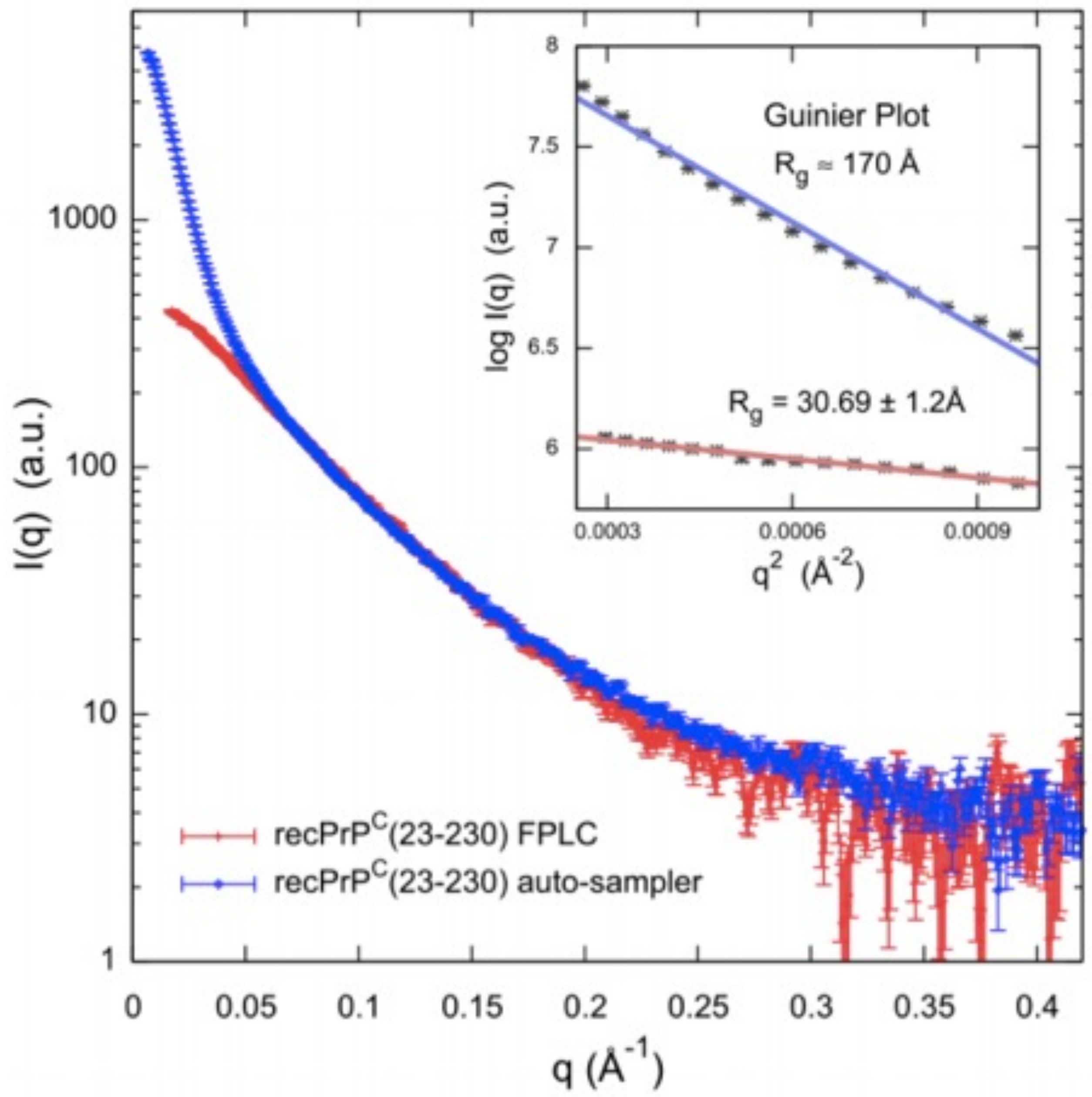
Most structures are solved by NMR for residues 121-230



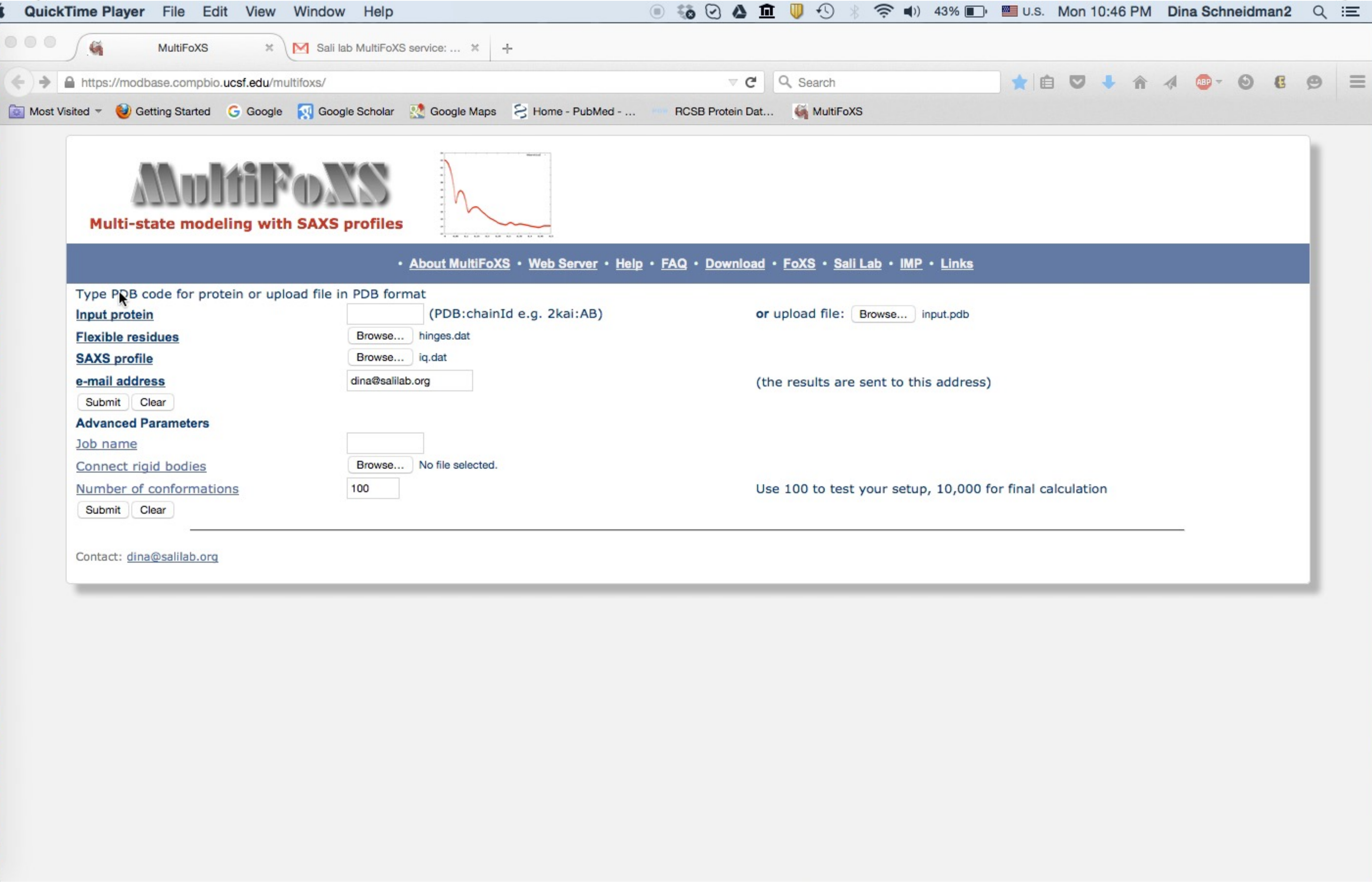
Collaboration with Prusiner lab, UCSF
Carter, Kim, Schneidman-Duhovny, Stoehr, Poncet-Montange, Tsuruta, Prusiner, Sali. Biophys J. 2015

Fast protein liquid chromatography (FPLC) coupled with SAXS

- Using FPLC-SAXS technology we could study the full length prion protein (residues 23-230)



Multi-state modeling with SAXS online



Multi-state modeling with SAXS profiles

Type PDB code for protein or upload file in PDB format

Input protein (PDB:chainId e.g. 2kai:AB) **or** upload file: Browse... input.pdb

Flexible residues Browse... hinges.dat

SAXS profile Browse... iq.dat

e-mail address dina@salilab.org (the results are sent to this address)

Submit Clear

Advanced Parameters

Job name

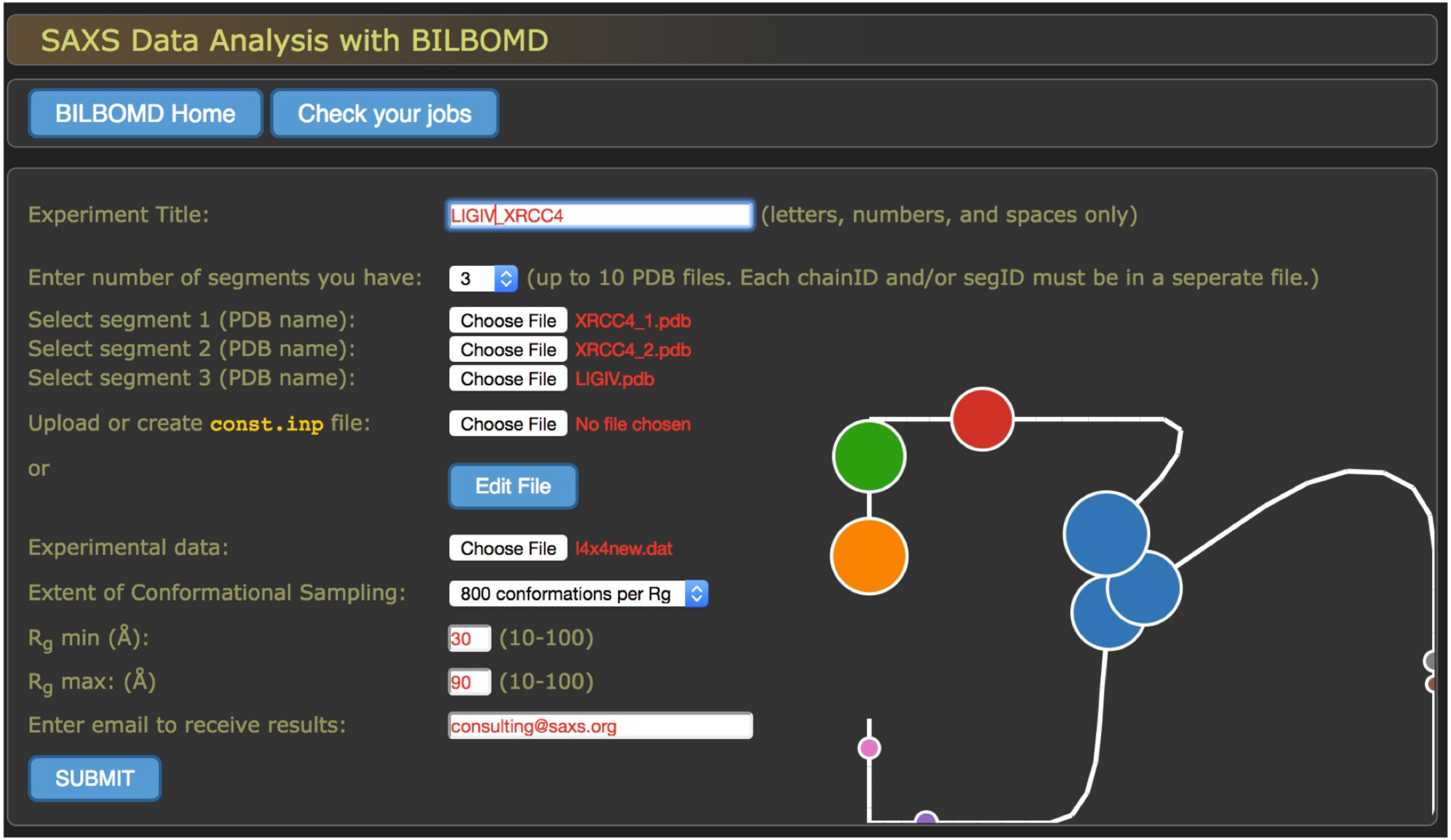
Connect rigid bodies Browse... No file selected.

Number of conformations 100 Use 100 to test your setup, 10,000 for final calculation

Submit Clear

Contact: dina@salilab.org

BilboMD: high-temperature MD for linkers



Experiment Title: LIGIV_XRCC4 (letters, numbers, and spaces only)

Enter number of segments you have: 3 (up to 10 PDB files. Each chainID and/or segID must be in a seperate file.)

Select segment 1 (PDB name): Choose File XRCC4_1.pdb

Select segment 2 (PDB name): Choose File XRCC4_2.pdb

Select segment 3 (PDB name): Choose File LIGIV.pdb

Upload or create `const.inp` file: Choose File No file chosen

or

Experimental data: Choose File l4x4new.dat

Extent of Conformational Sampling: 800 conformations per Rg

R_g min (Å): 30 (10-100)

R_g max: (Å) 90 (10-100)

Enter email to receive results: consulting@saxs.org

https://bl1231.als.lbl.gov/bilbomd

Pelikan M, Hura GL, Hammel M.2009

Did we achieve the balance?

few states – don’t represent the heterogeneous sample

more states –over-interpret the data

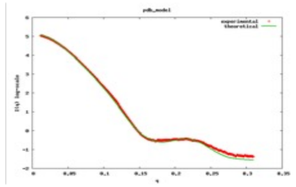
We analyzed a large ensemble of multi-state models (1000 x N) highlighting the conserved features among these models

We computed multi-state models with a low number of states (N=1…5) and set the lower bound on the weight of each state to 5%

Integrative Modeling and SAXS

Information

standards for data collection and validation



Scoring

Debye formula
CRYSOL, FoXS, Pepsi-SAXS

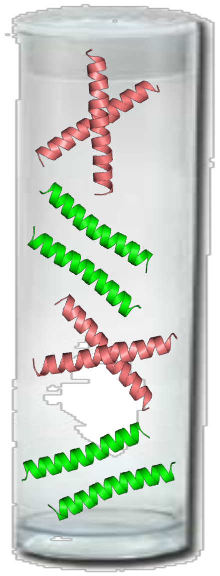
Sampling

AlphaFold, comparative modeling, conformational sampling, docking, assembly

Analysis

Data satisfied?

no



sample heterogeneity

more data is needed

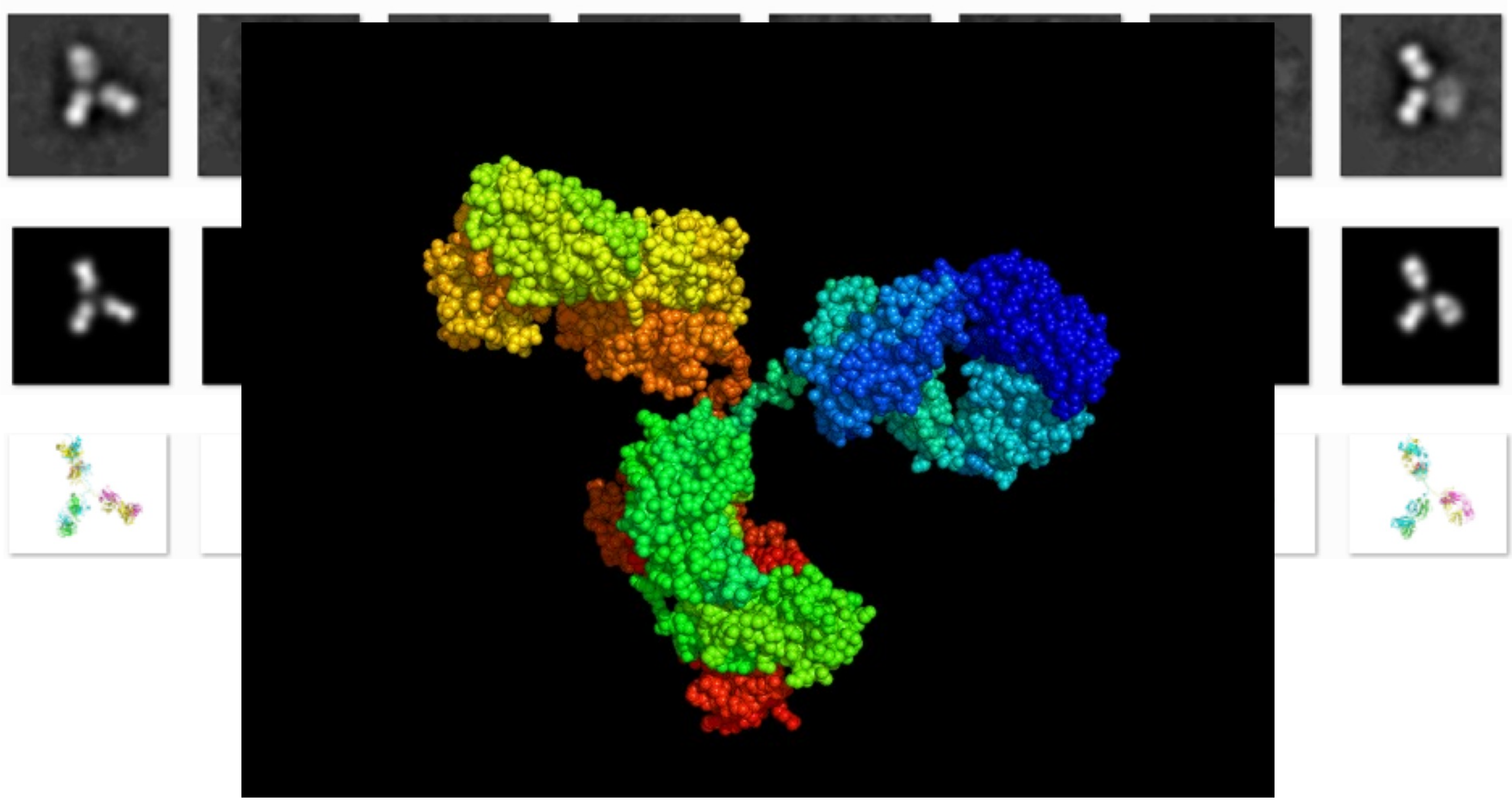
yes

Single structure

Multiple structures

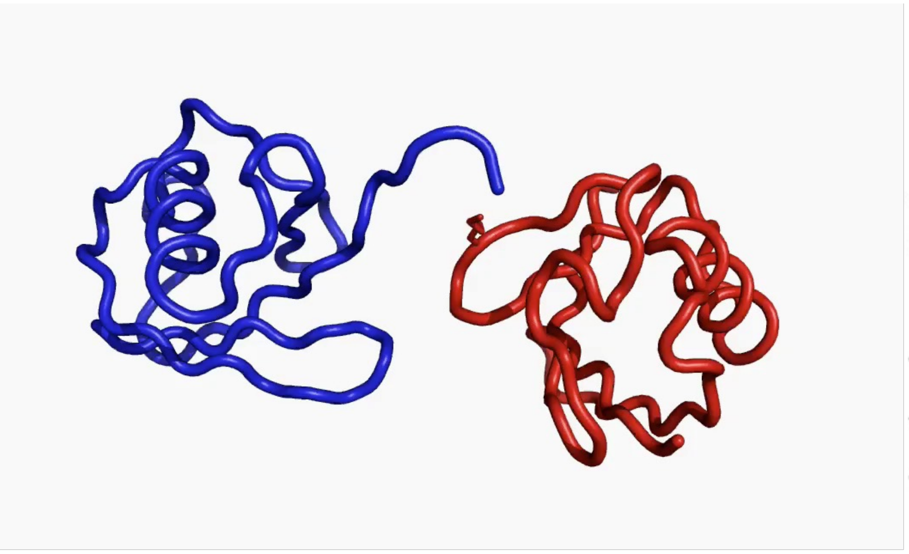
Schneidman-Duhovny, Pellarin, Sali. COSB 2014

Heterogeneity from 2DEM data

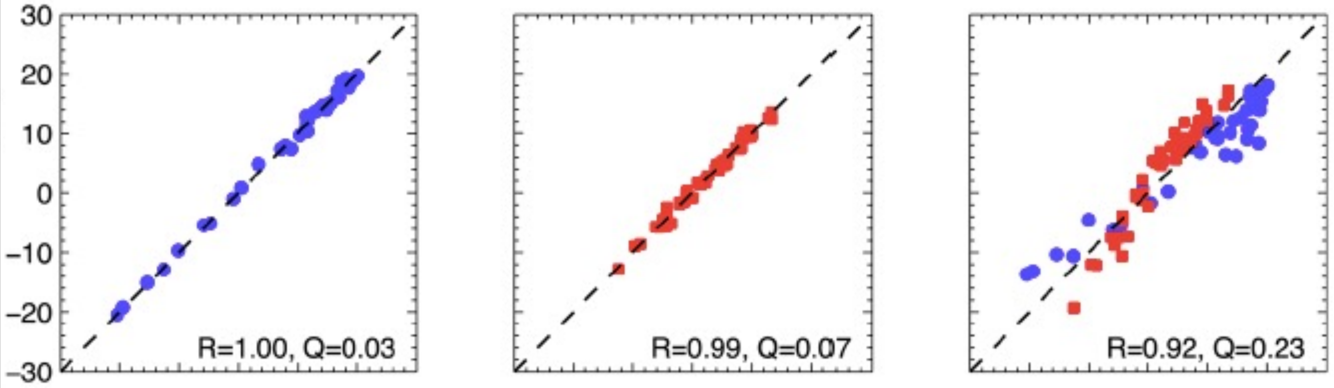


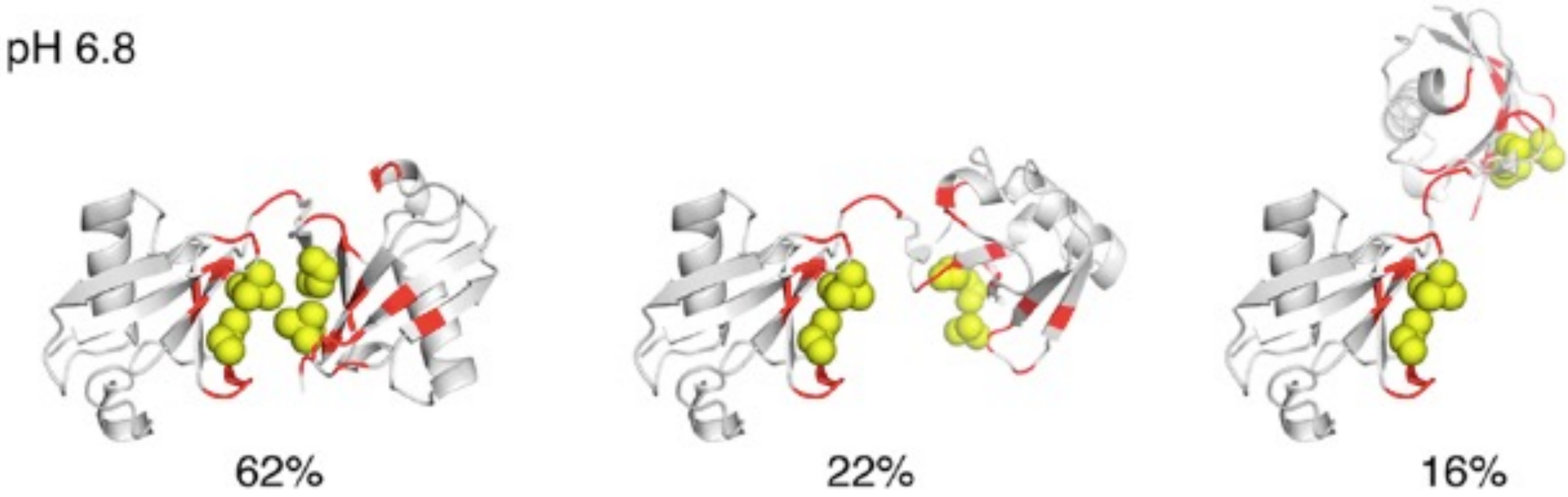
Chen et al. PLoS ONE 2017

Lys48-linked diubiquitin with residual dipolar coupling (RDC) data



correlation between observed and calculated RDCs

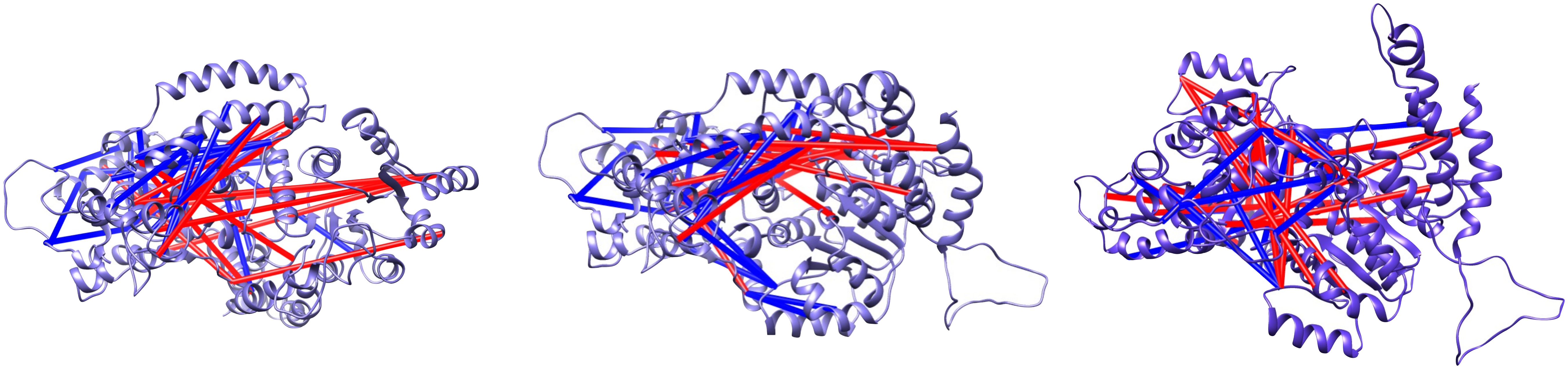


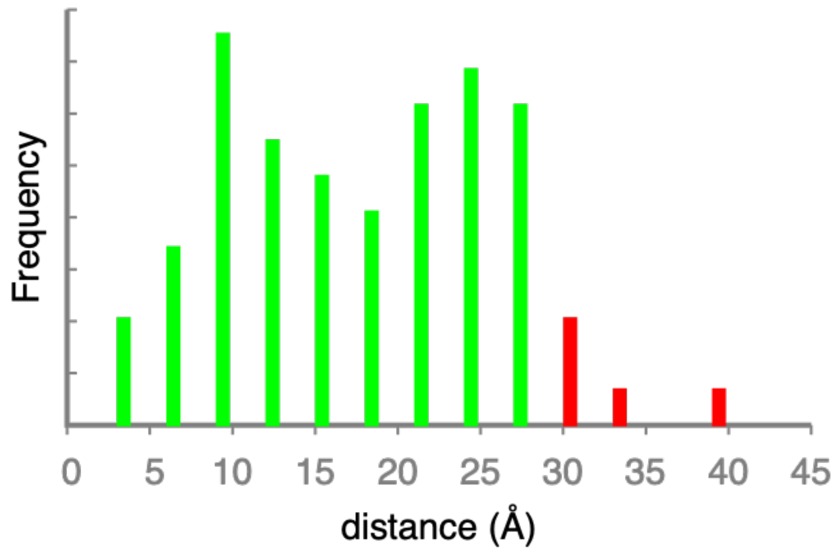


Berlin, Castaneda, Schneidman-Duhovny, Sali, Nava-Tudela, Fushman JACS 2013

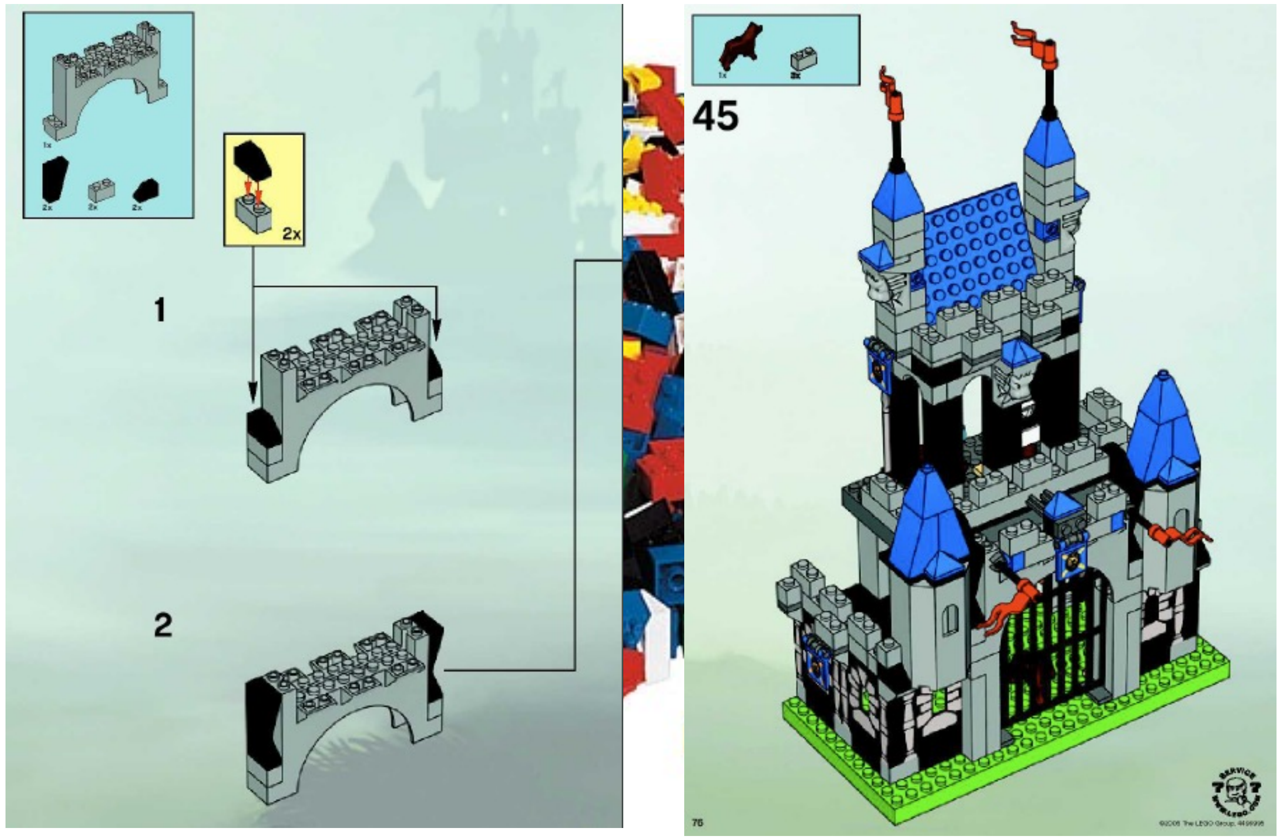
Multi-state model for STH1

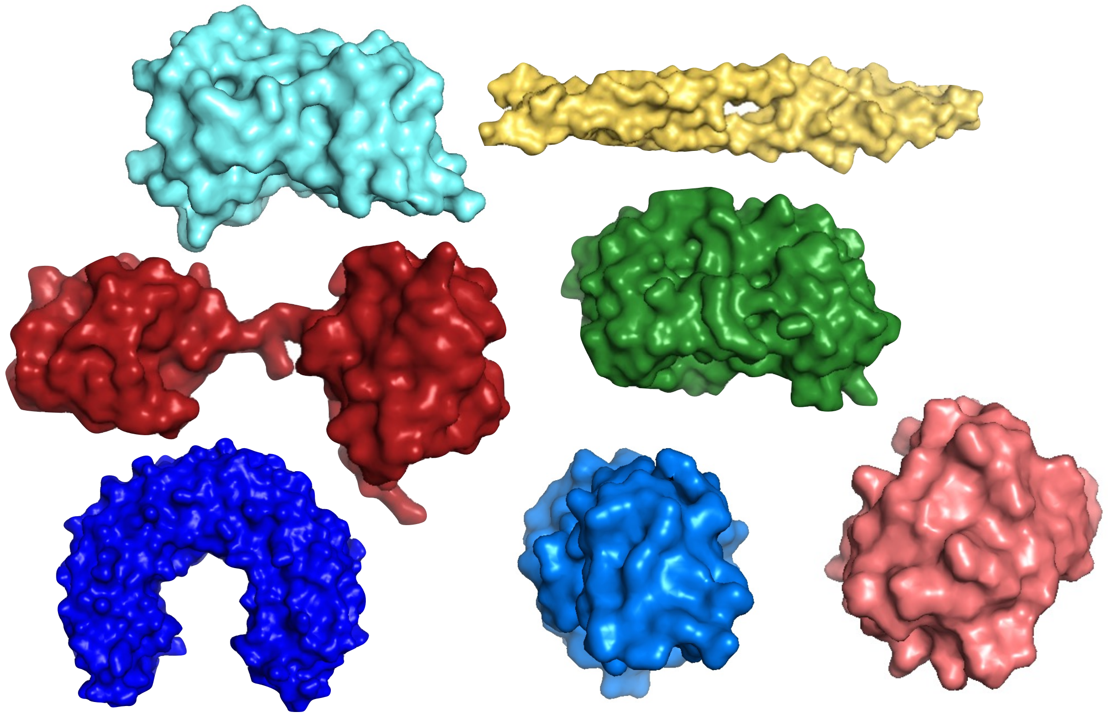
- 3 structures explain all but 1 cross-link using 35Å cutoff
- the remaining cross-link has a 39Å distance

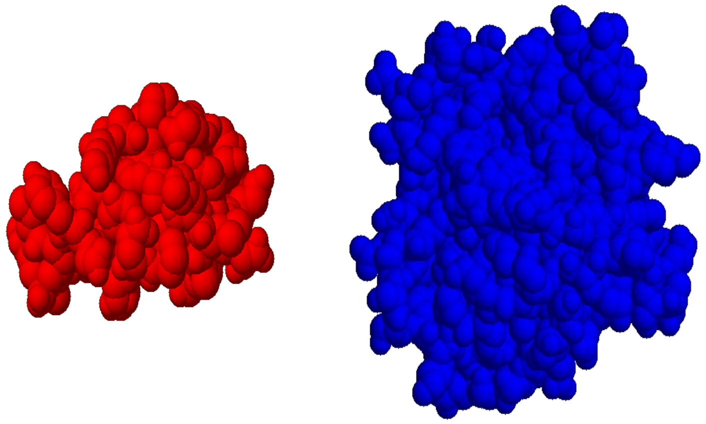




Modeling protein interactions



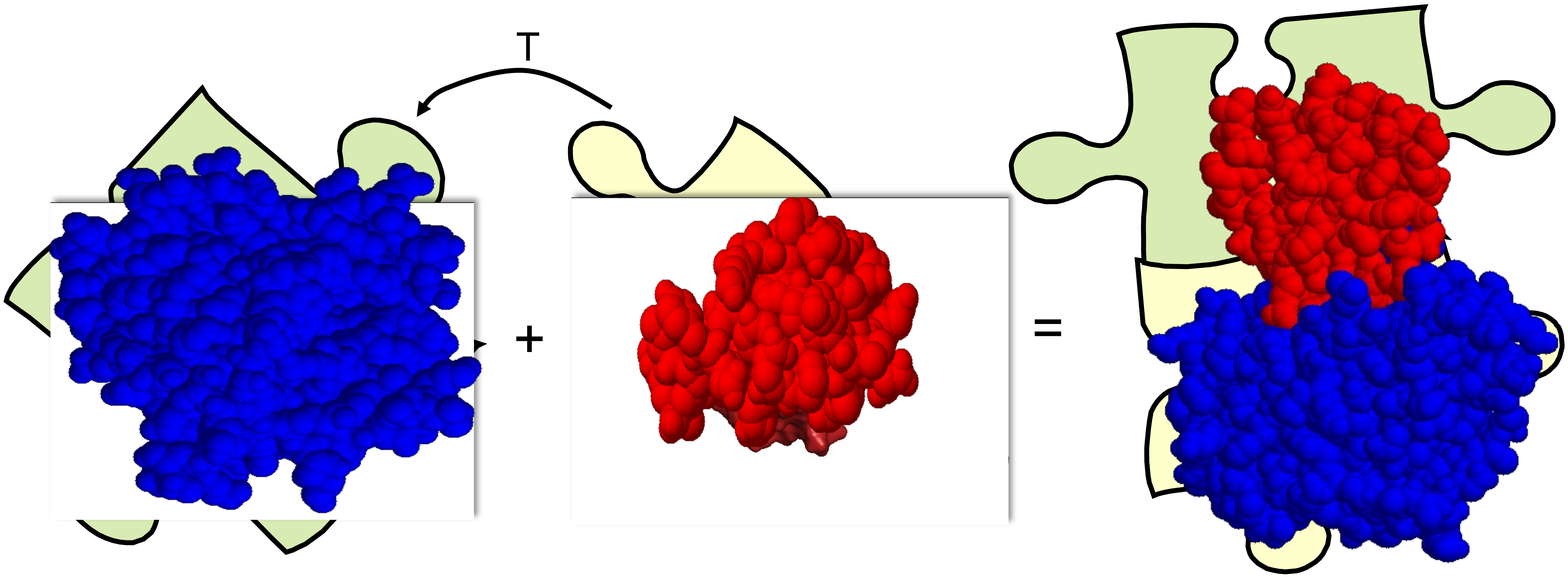




?

Docking problem

Given 2 input molecules in their native conformation, the goal is to find their correct association as it appears in nature.



3D Transformation: 3 rotational and 3 translational parameters

Finding needles in a haystack

- Docking methods generate thousands of models including models that are close to the correct complex

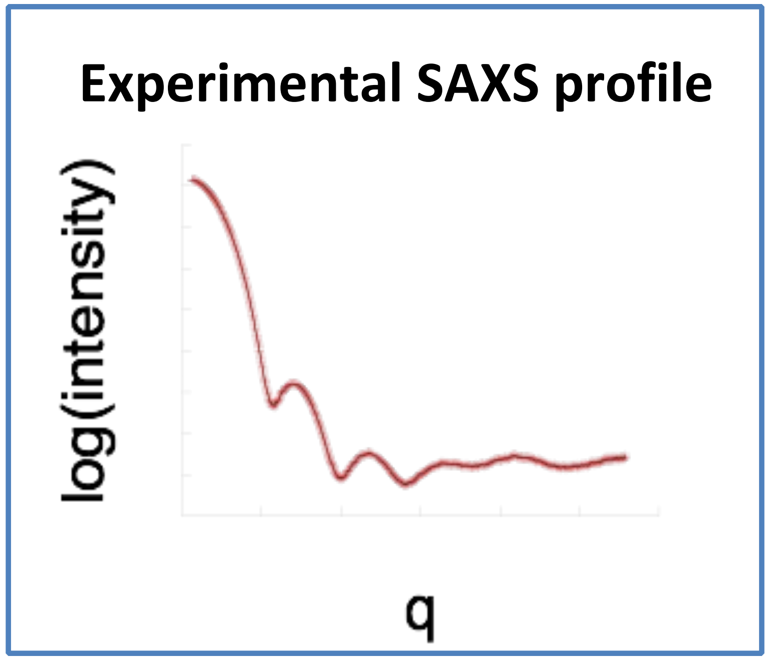
- Only in **20-30%** of cases we find the right model in top10

- Additional information significantly reduces uncertainty

Sampling

Scoring

Information

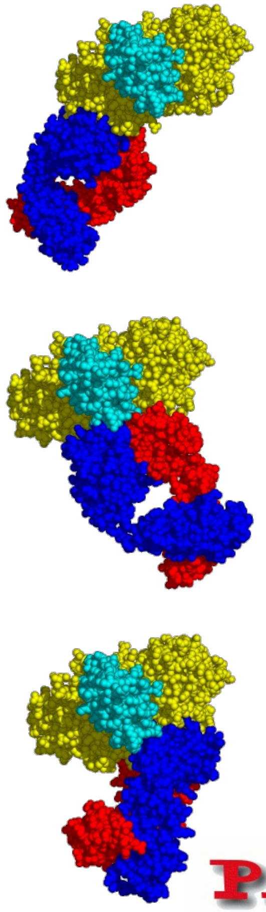


by Brian Jimenez

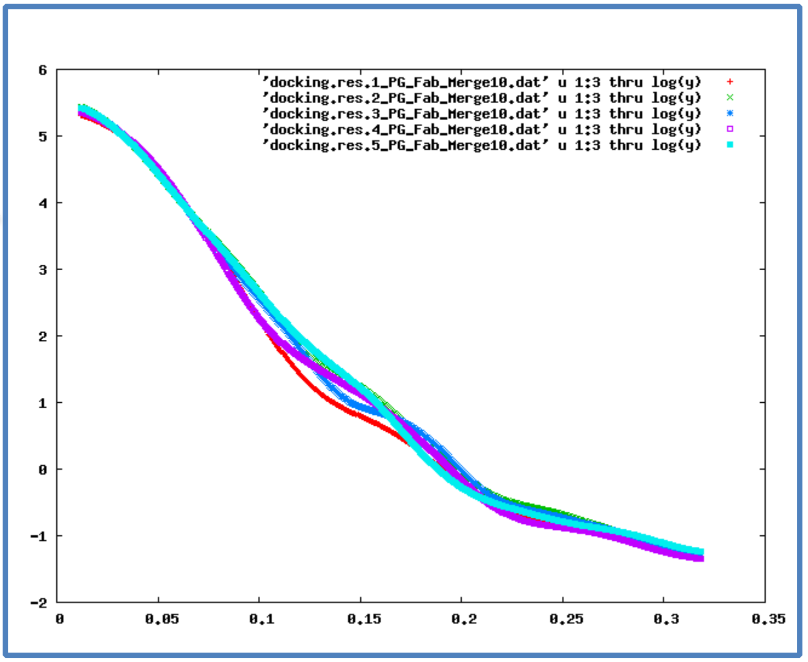
Docking with SAXS profile of the complex

FoXSDock

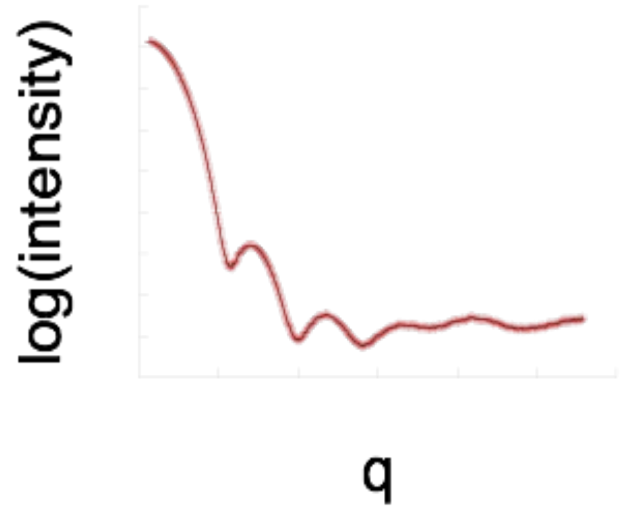
forward model



PATCHDOCK



Experimental SAXS profile



FoXS

Generate docking candidates

Compute theoretical SAXS profiles

Fit experimental profile and compute the score

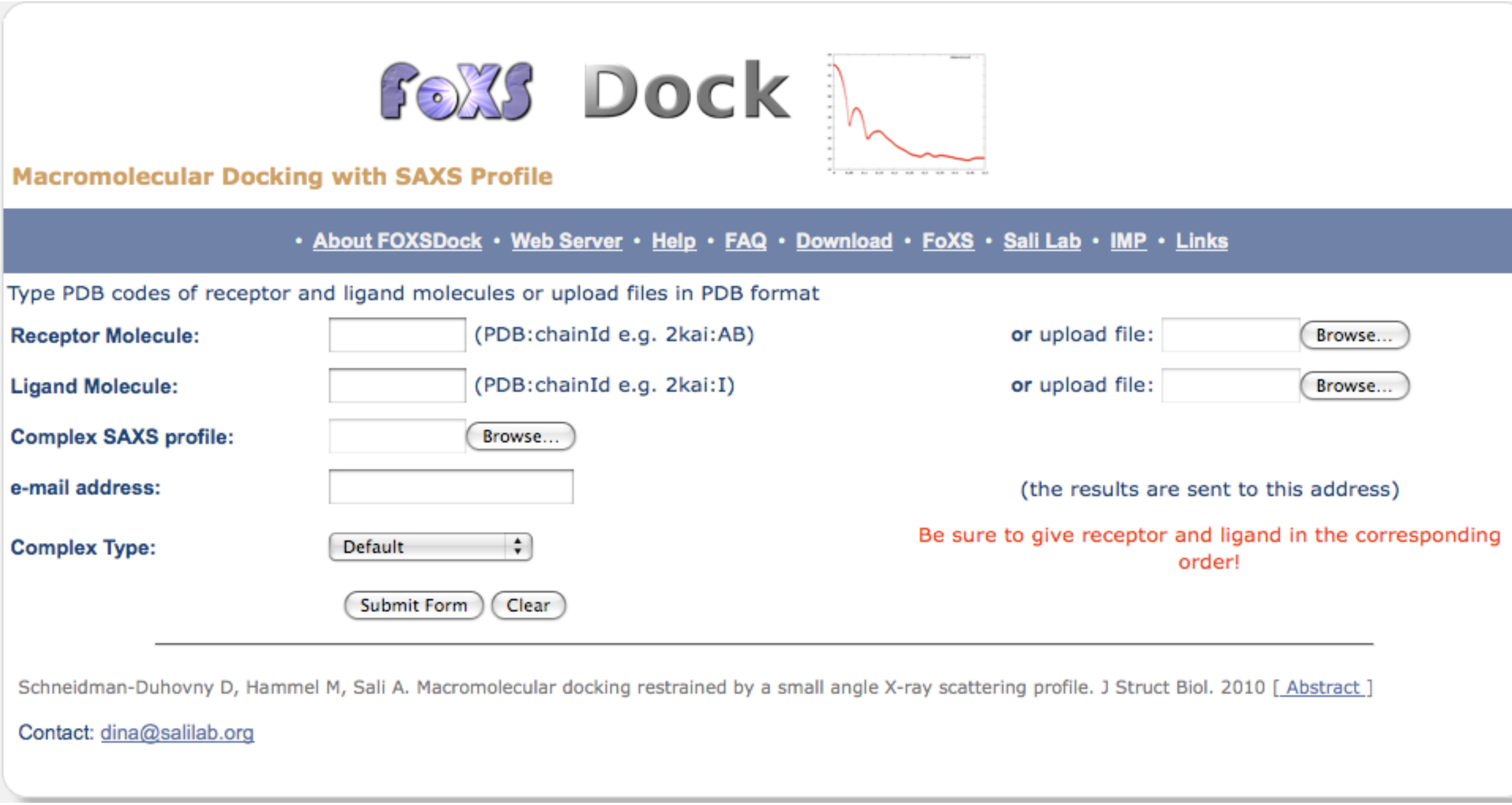
Debye formula

$$I(q)=\sum_{i=1}^{N}\sum_{j=1}^{N} f_i(q)f_j(q)\frac{\sin(qd_{ij})}{qd_{ij}}$$

$$\boxed{\chi=\sqrt{\frac{1}{M-1}\sum_{i=1}^{M}\left(\frac{I_{exp}(q_i)-cI(q_i)}{\sigma(q_i)}\right)^2}}$$

Schneidman-Duhovny D, Hammel M, Sali A. J Struct Biol. 2011
Schneidman-Duhovny D, Hammel M, Tainer J, Sali A. NAR 2016

http://salilab.org/foxsdock



Macromolecular Docking with SAXS Profile

Type PDB codes of receptor and ligand molecules or upload files in PDB format

Receptor Molecule: (PDB:chainId e.g. 2kai:AB) **or** upload file:

Ligand Molecule: (PDB:chainId e.g. 2kai:I) **or** upload file:

Complex SAXS profile:

e-mail address: (the results are sent to this address)

Complex Type: Default

Be sure to give receptor and ligand in the corresponding order!

Schneidman-Duhovny D, Hammel M, Sali A. Macromolecular docking restrained by a small angle X-ray scattering profile. J Struct Biol. 2010 [Abstract]

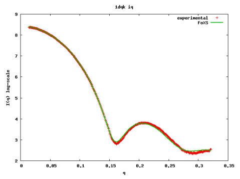
Contact: dina@salilab.org

Experimental data for protein-protein docking

Five data types were selected due to feasibility of data collection

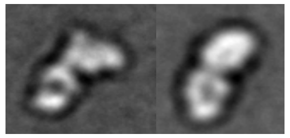
Small **A**ngle **X**-ray **S**cattering profile

SAXS



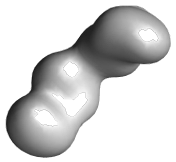
2D class average images from negative stain **EM**

EM2D



3D density map from single particle negative stain **EM**

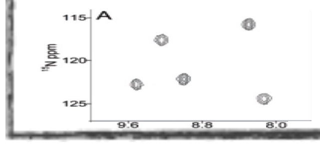
EM3D



shape information

Residue **T**ype **C**ontent from **NMR** spectroscopy

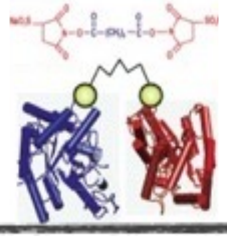
NMR-RTC



interface

Chemical **cross**linking detected by **M**ass **S**pectrometry

CXMS



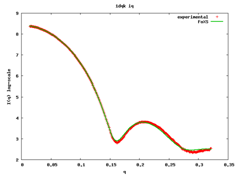
medium range distances

Schneidman-Duhovny D et al. *Bioinformatics* 2012

Benchmark with simulated data

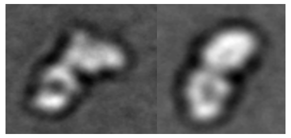
Docking benchmark 4.0: 176 unbound-unbound test cases
Data simulated using co-crystallized structure of the complex

SAXS



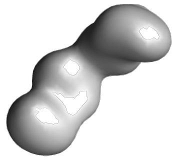
3% noise

EM2D



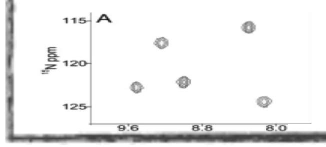
5 class averages with noise

EM3D



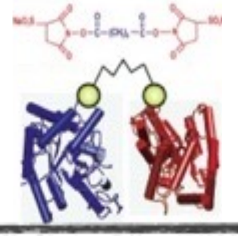
20Å resolution

NMR-RTC



Counts for 4 rarest surface residues

CXMS

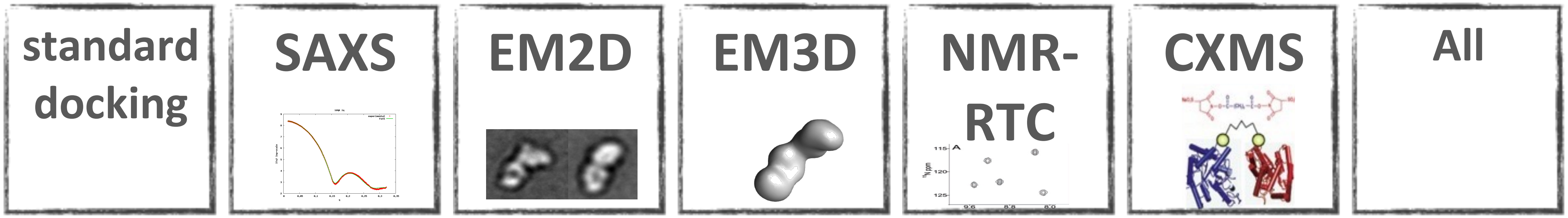


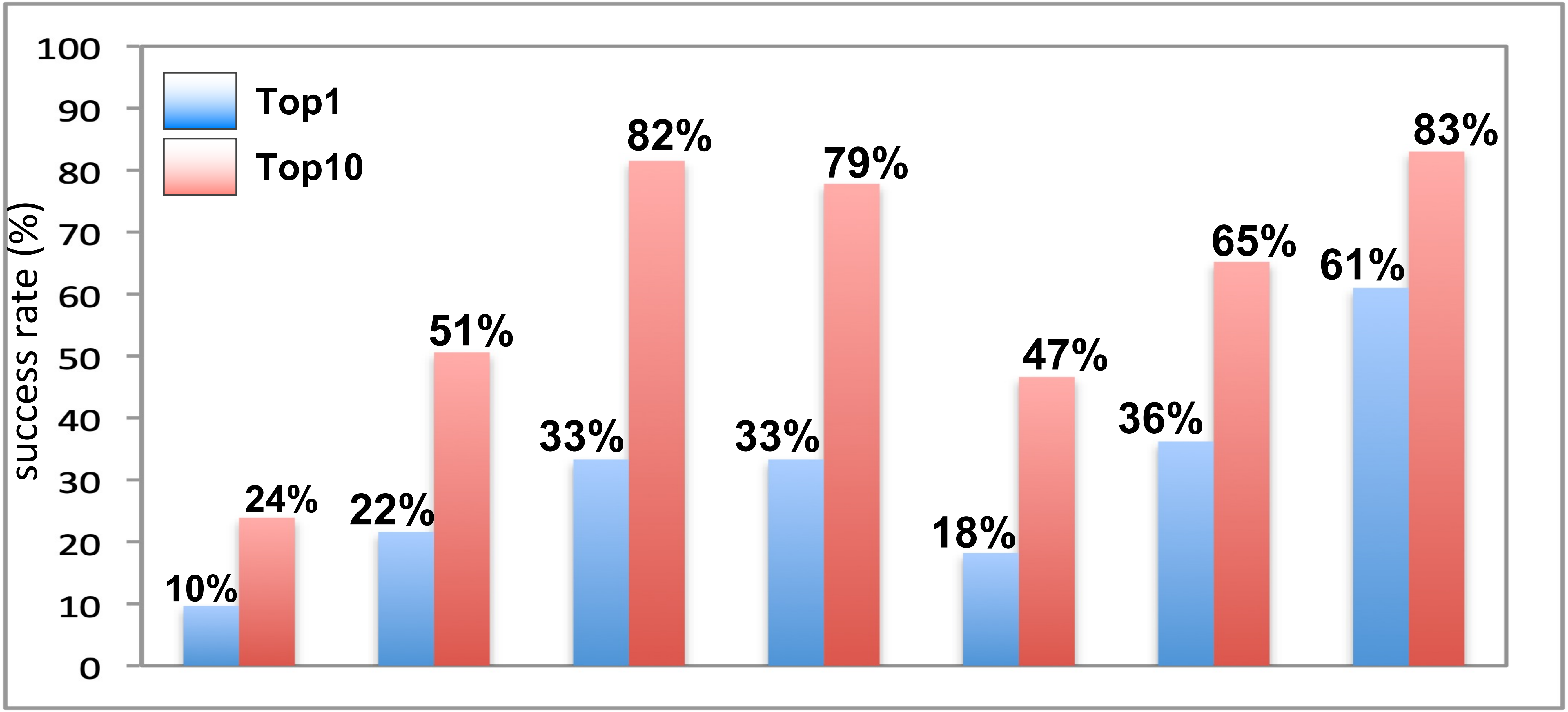
Up to 3 Lys-Lys cross links

- **success rate** = % of benchmark cases with at least one near-native model in the topN predictions
- **near-native** = a model with interface RMSD<4Å or ligand RMSD <10Å

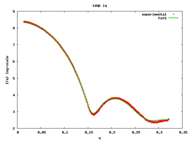
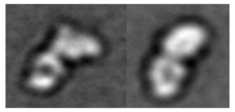
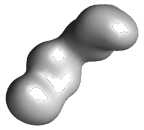
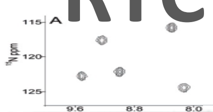
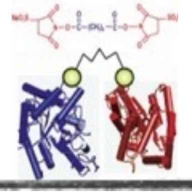
Schneidman-Duhovny D et al. *Bioinformatics* 2012

Success rate for single datasets





Top 1 success rate for combined datasets

	standard docking	SAXS 	EM2D 	EM3D 	NMR-RTC 	CXMS 	All
	10%	22%	33%	33%	18%	36%	61%
SAXS		Shape data				**51%**	
EM2D						**52%**	
EM3D						**48%**	
NMR-RTC						**47%**	

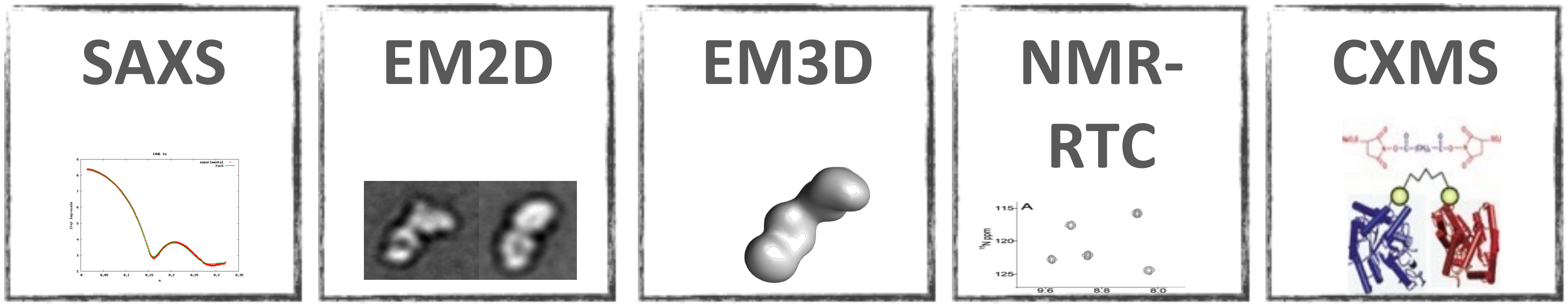
- Cross linking and shape-based datasets (SAXS, EM2D, or EM3D) result in the highest success rate

* Numbers are shown only for dataset pairs with >10% increase in the success rate

Integrative modeling of protein interactions

- component structures
- shape complementarity
- statistical potentials

+



SAXS

EM2D

EM3D

NMR-RTC

CXMS

Single dataset

Two datasets

Blind prediction of protein-protein complex

- Multiple targets with cross-links in the recent CASP
- T0957 was a complex of CdiA_CdiI from E. coli.
- a SAXS profile of the complex
- 7 inter-protein cross-links!!!

- No structures or close homologs for the two proteins
- 10 intra-protein cross-links (7 + 3)

Fold & Dock challenge!

A

```
SNSFEVSSLPDANGKNHITAVKGDAKIPVDKIELYMRGKASGDLDSLQAEYNSLKDARISSQKEFAKDPNNAKRMEVLEKQIHNIERSQDMARVL
EQAGIVNTASNNSMIMDKLLDSAQGATSANRKTSVVVSGPNGNVRIYATWTILPDGTKRLSTVTGTFK
```

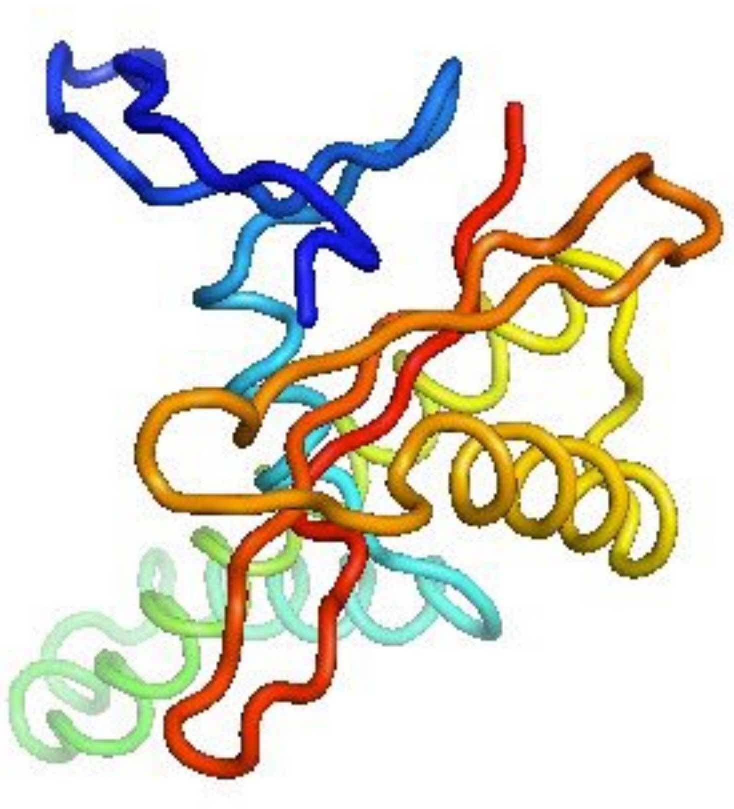
B

```
SNAMINVNSTAKDIEGLESYLANGYVEANSFNDPEDDALECLSNLLVKDSRGGLSFCKKILNSNNIDGVFIKGSALNFLLLSEQWSYAFEYLTSN
ADNITLAELEKALFYFYCAKNETDPYPVPEGLFKKLMKRYEELKNDPDAKFYHLHETYDDFSKAYPLN
```

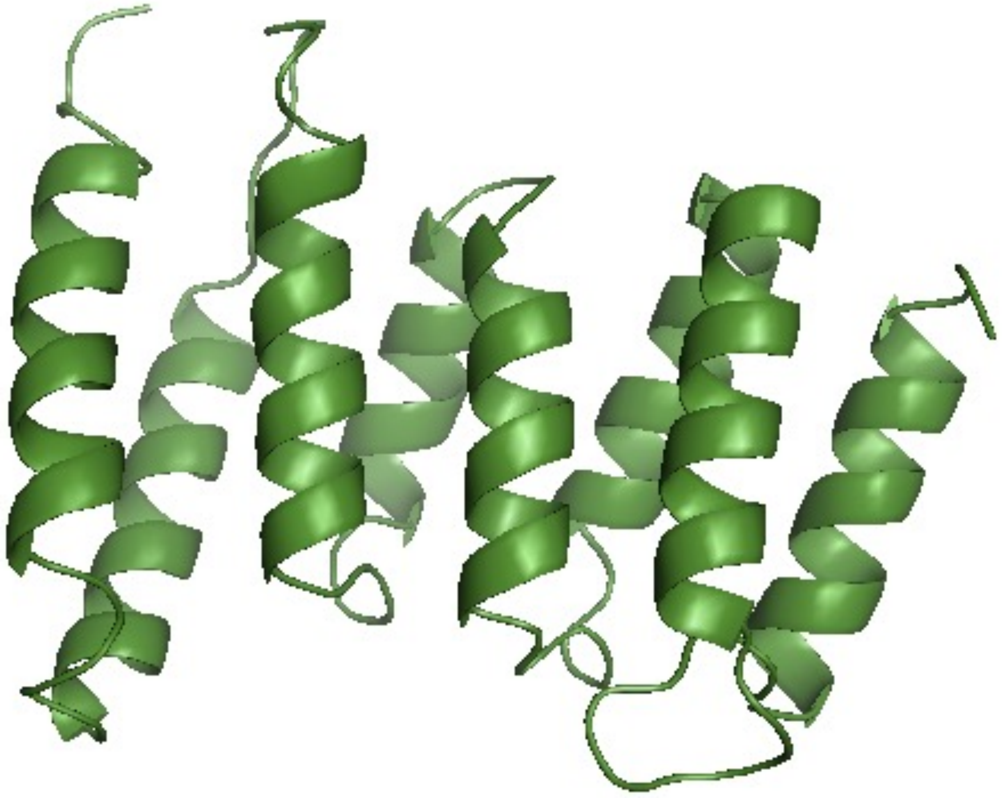
Intra-protein cross-links can help!

- ~450 submitted models for each chain by CASP groups
- Select models that satisfy intra-protein cross-links

chain A:
7 cross-links, no convergence
select 20 best scoring models

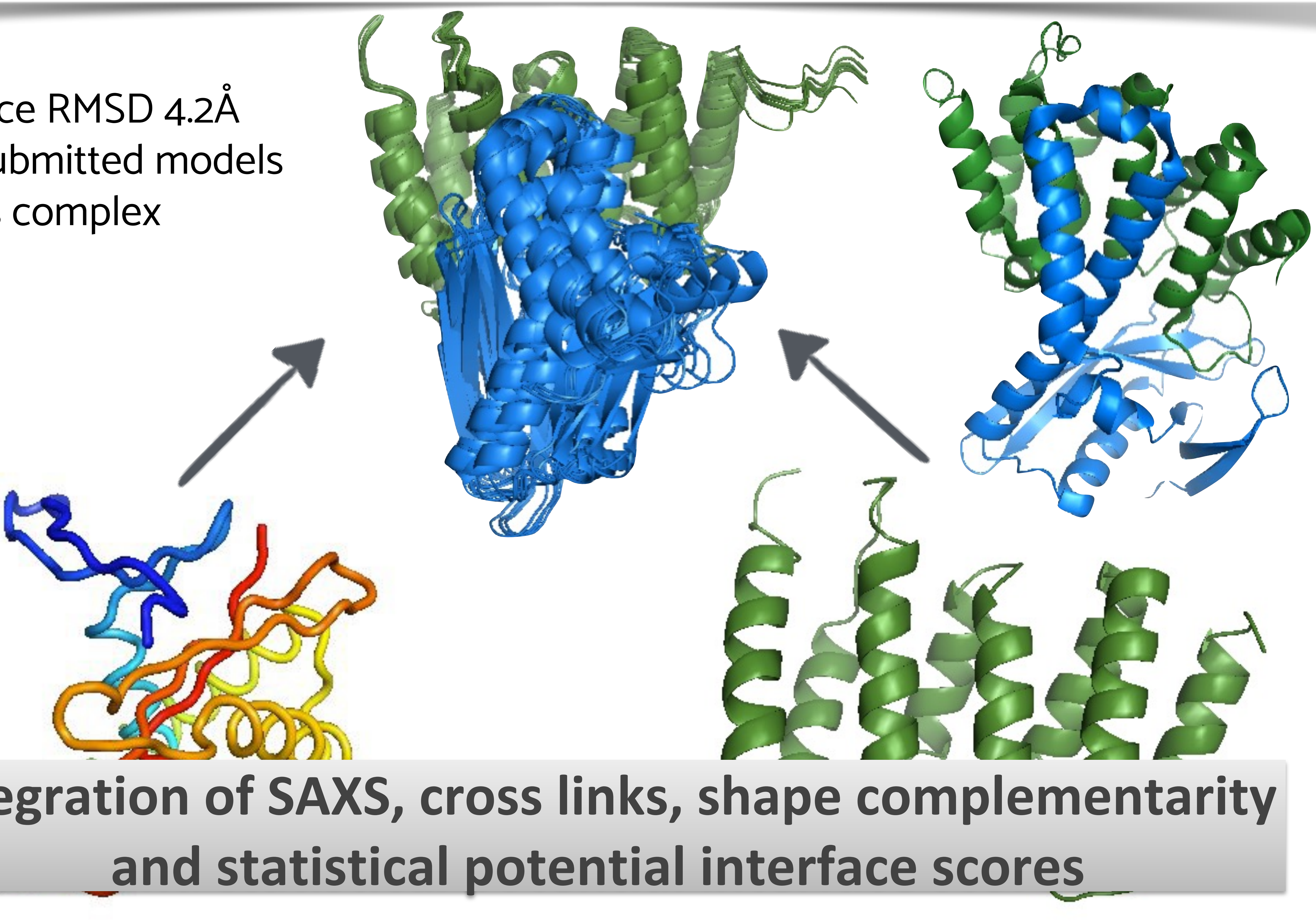


chain B:
3 cross-links, convergence
select 1 model



Integrative modeling finds best models

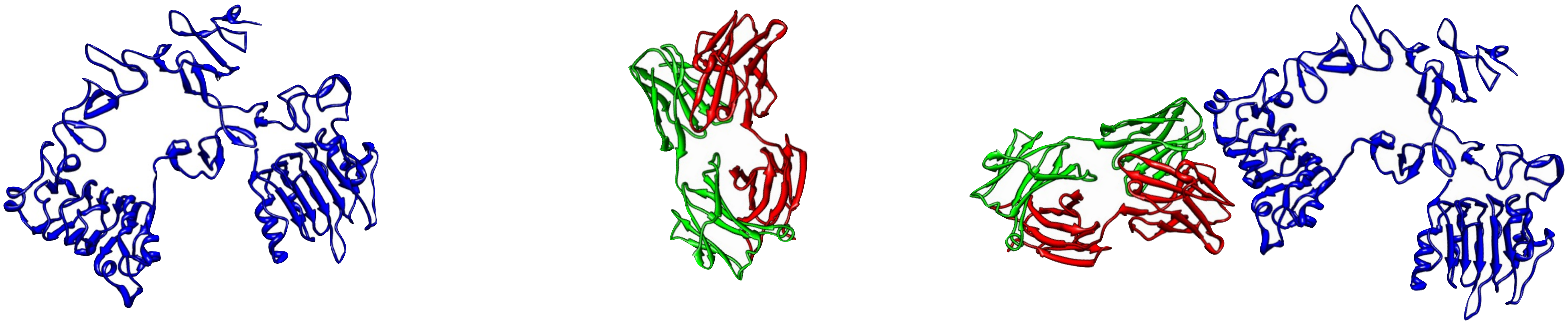
- Interface RMSD 4.2Å
- Best submitted models for this complex

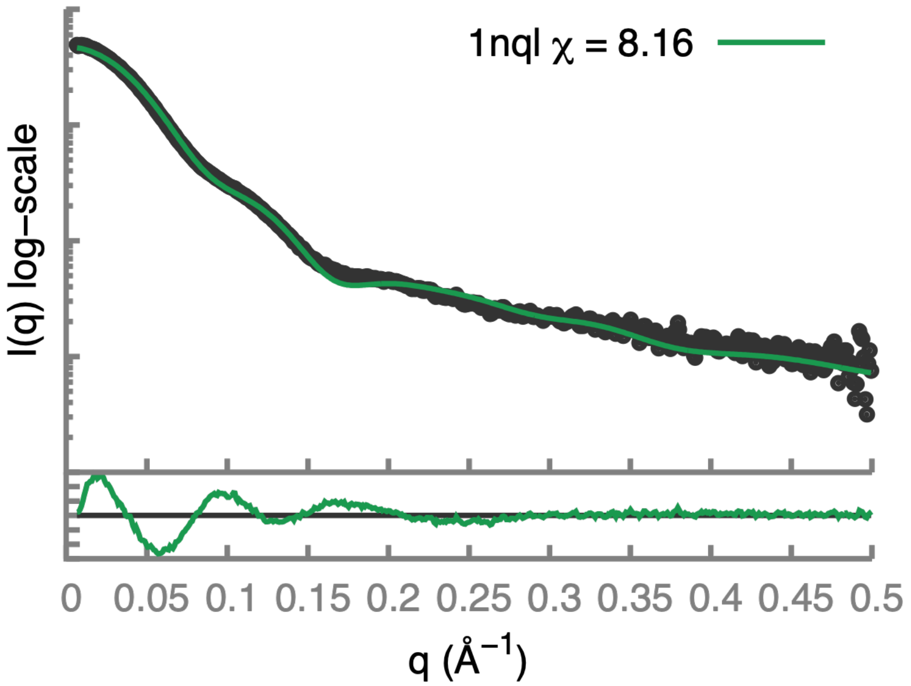


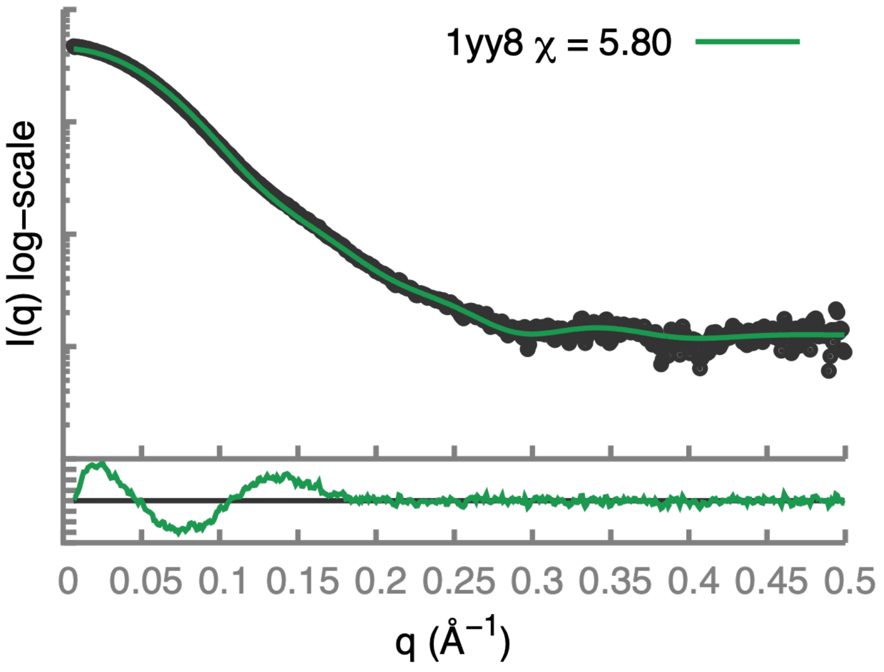
integration of SAXS, cross links, shape complementarity and statistical potential interface scores

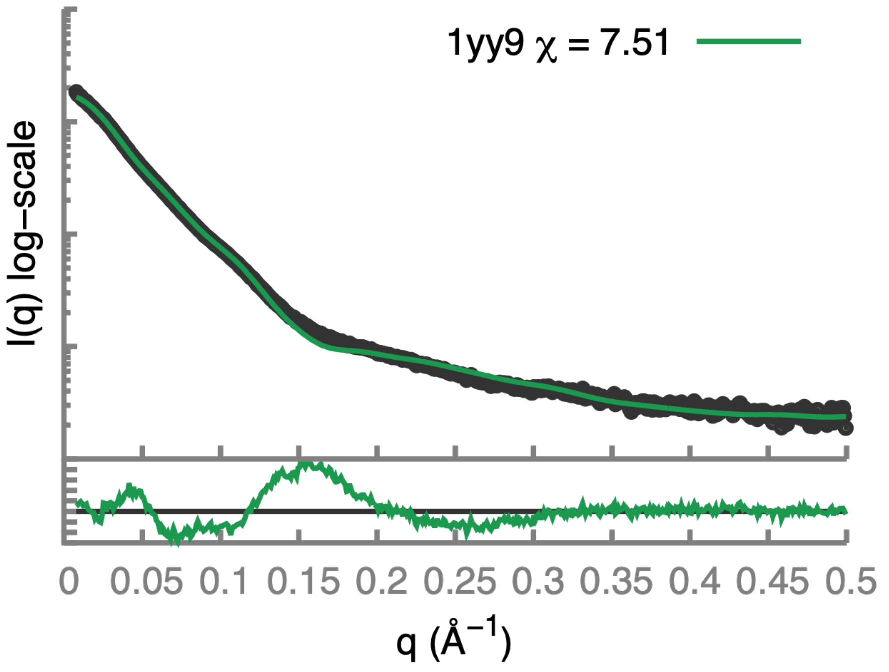
EGFR-antibody complex with SAXS profiles

- SAXS profiles collected for EGFR, antibody, and their complex
- 5 antibodies



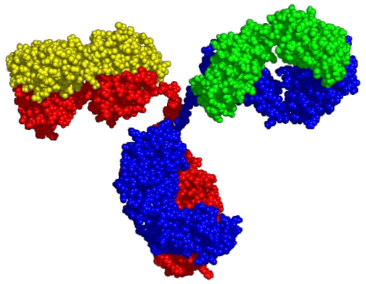






Collaboration with Arvind Sivasubramanian (Adimab)

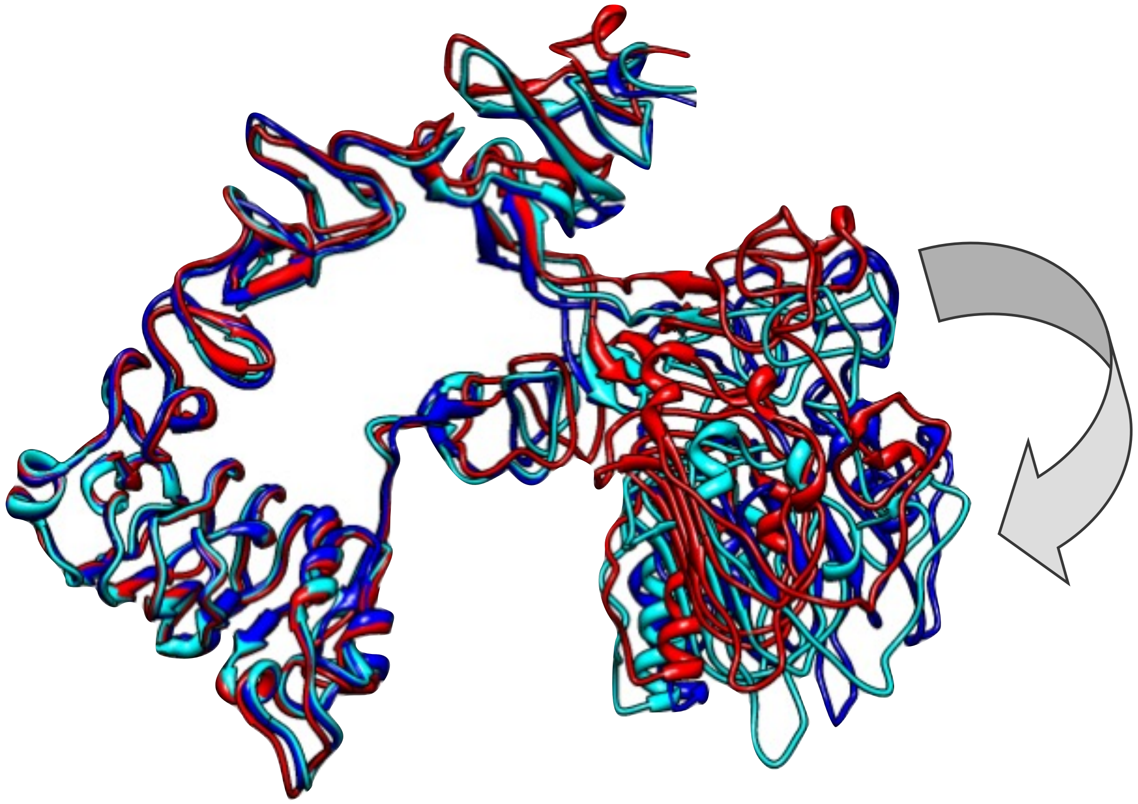
Why antibodies?

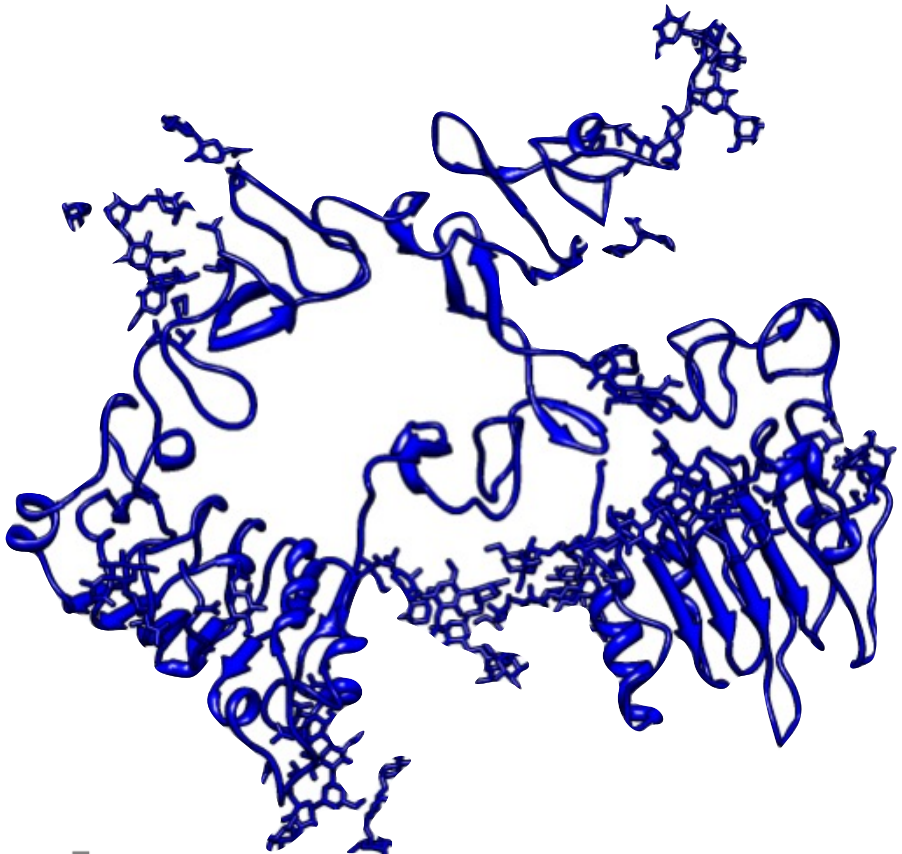


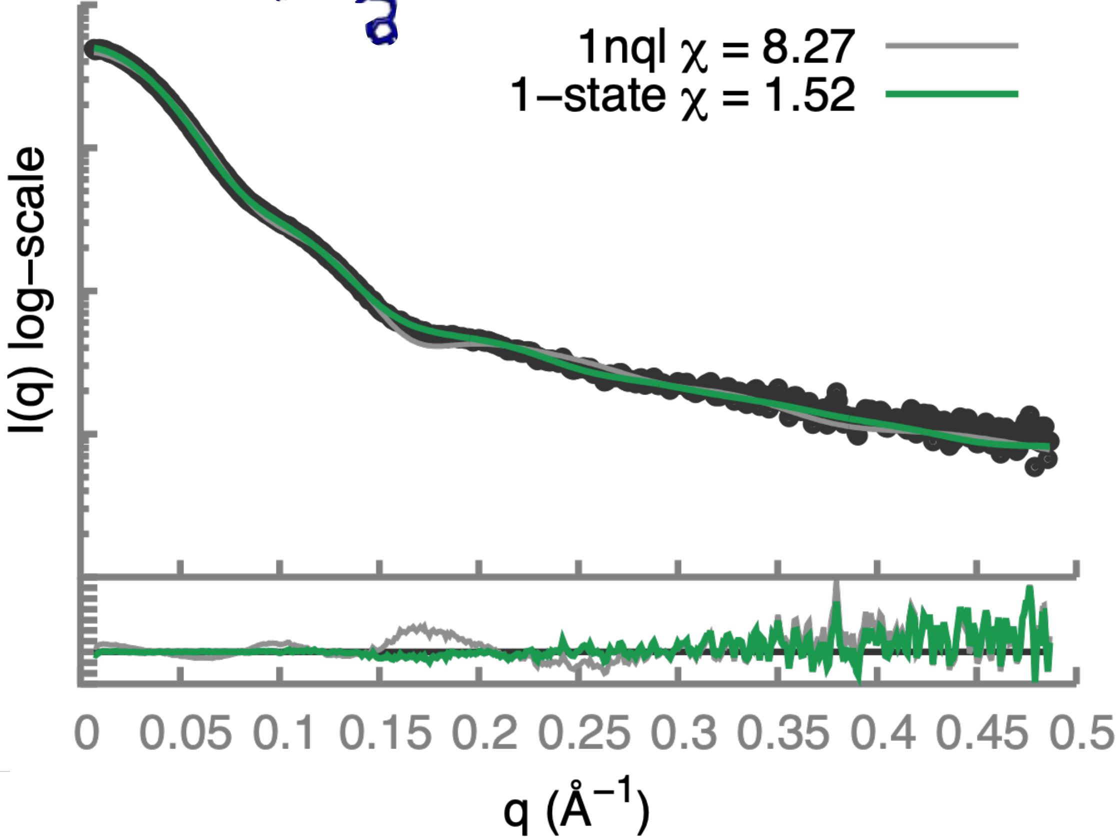
- Antibodies are a key component of the immune system – active immunity
- Antibodies constitute a rapidly growing class of human therapeutics (“mabs”) – passive immunity
- Over 80 approved therapeutic antibodies for cancer, autoimmune diseases, neurological disorders and more
- Nobel prize in Chemistry (2018) - phage display technology
- Nobel prize in Physiology or Medicine (2018) - immune checkpoint modulation in cancer
- SARS-CoV antibodies
- Modeling challenges

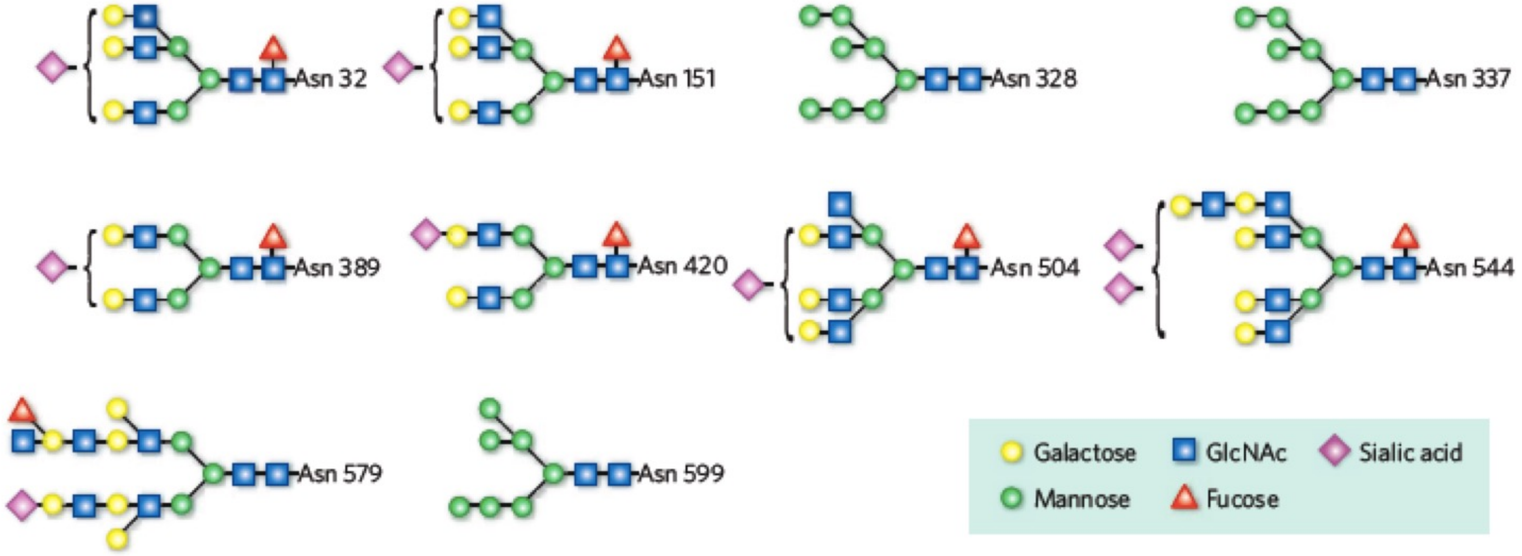
EGFR is flexible and glycosylated

3 crystal structures:



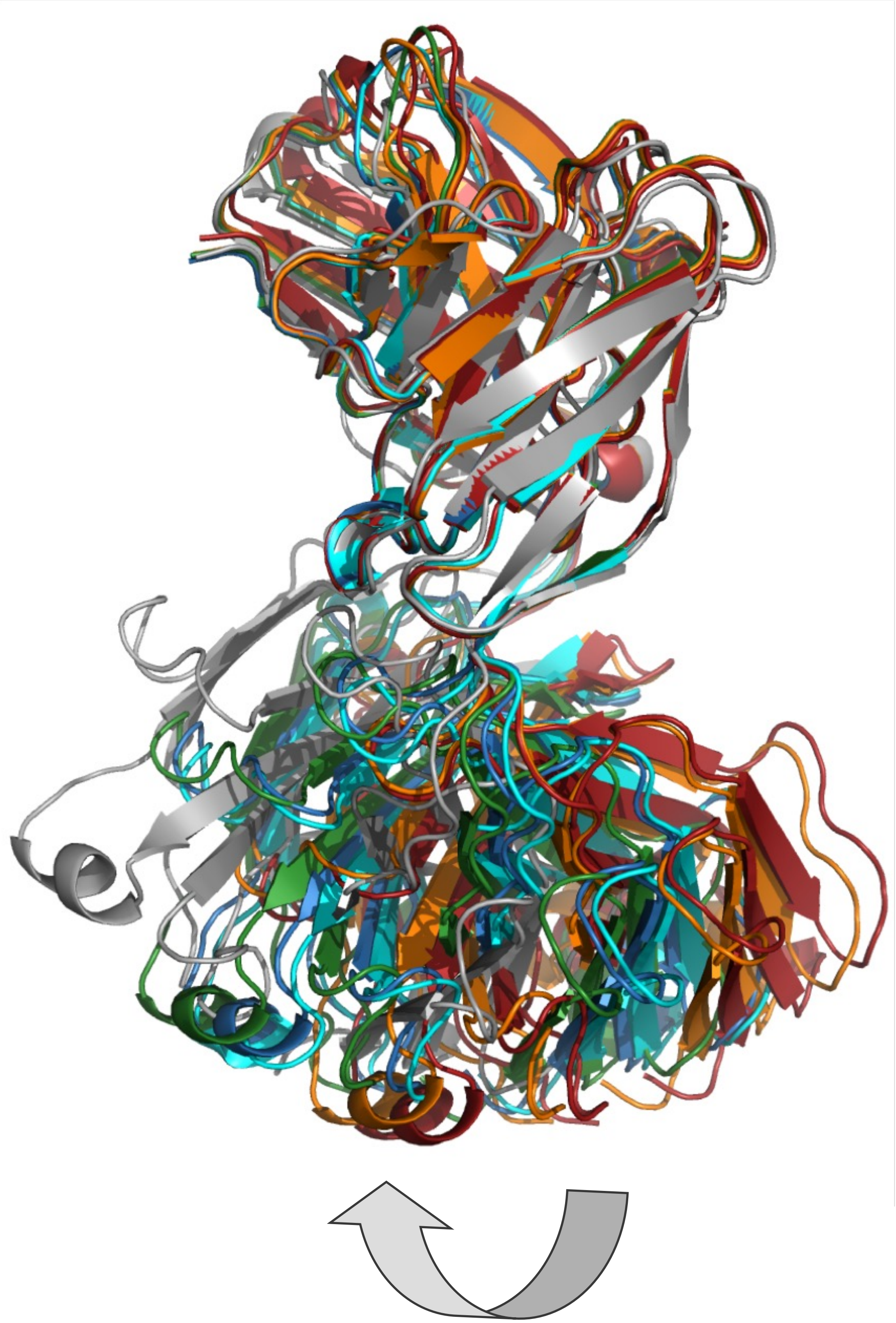




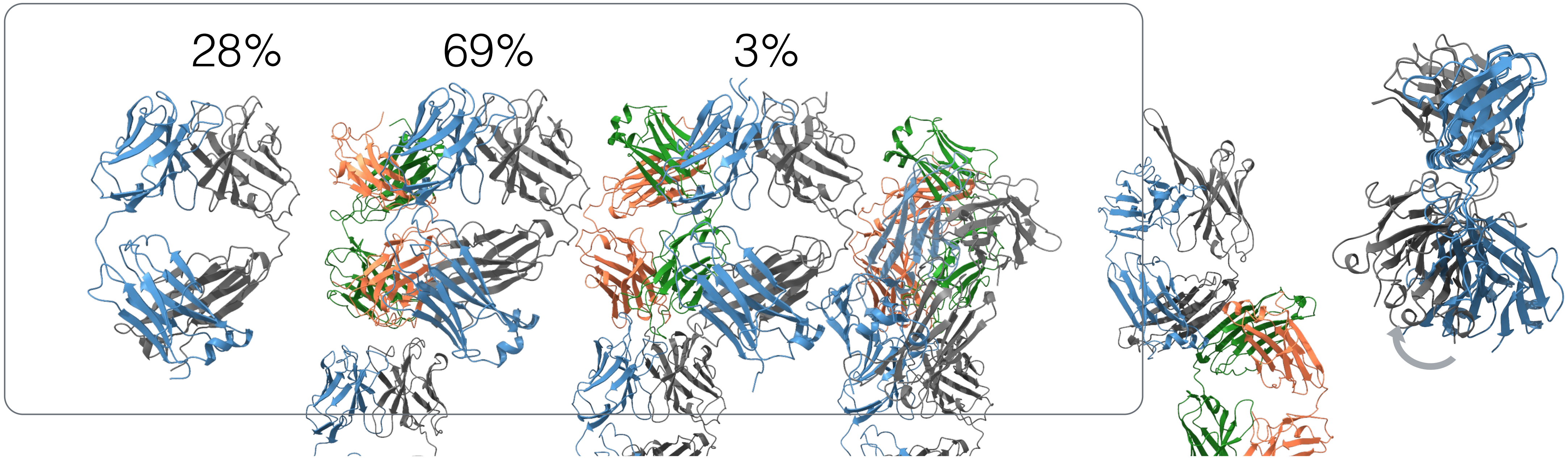


Fabs vary their elbow angle

PDB	χ^2 x-ray Fab	χ^2 single-state	χ^2 multiple elbow angles
1yy9	27.3	9.6	9.6
3b2u	20.4	11.3	11.3
3c09	19.0	4.9	4.9
3p0y	9.4	3.5	3.4
3sqo	69.0	6.5	3.6



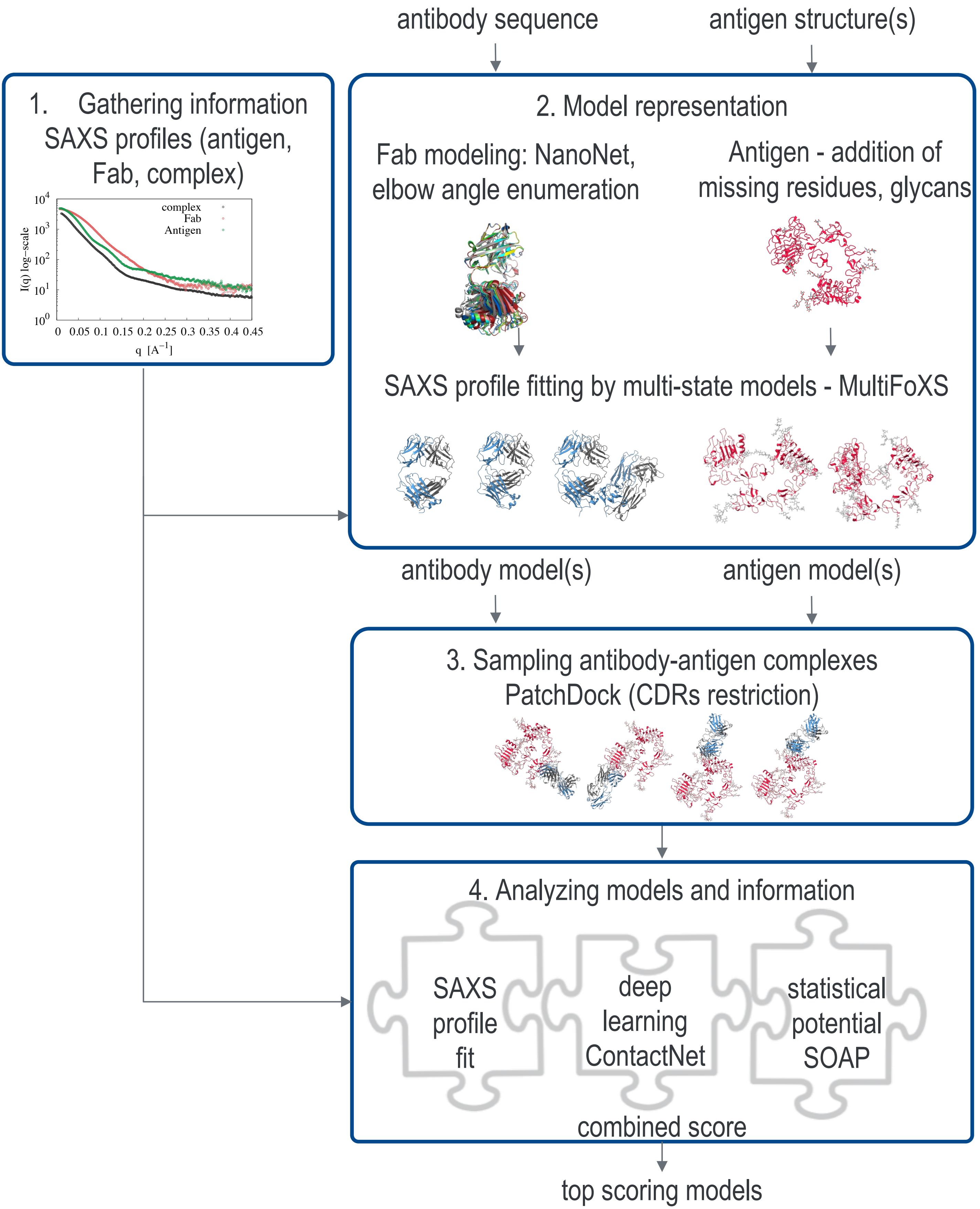
What else?



PDB	min χ^2 without Fab dimers	# states in a final multi-state model	% Fab dimers	χ^2 multiple states
1yy9	9.6	3	3.4	2.3
3b2u	11.3	4	2.4	5.4
3c09	4.9	4	5.5	2.1
3p0y	3.4	3	3.8	2.8
3sqo	3.6	4	2.9	1.6

Let’s dock!

1. SAXS data collection
   - antibody, antigen, complex
2. Antibody and antigen modeling
   - single- or multi-state
3. Docking with all conformations
4. Scoring
   - SAXS – multiple states
   - interaction interface



antibody sequence

antigen structure(s)

1. Gathering information
SAXS profiles (antigen, Fab, complex)

2. Model representation

Fab modeling: NanoNet, elbow angle enumeration

Antigen - addition of missing residues, glycans

SAXS profile fitting by multi-state models - MultiFoXS

antibody model(s)

antigen model(s)

3. Sampling antibody-antigen complexes
PatchDock (CDRs restriction)

4. Analyzing models and information

SAXS profile fit

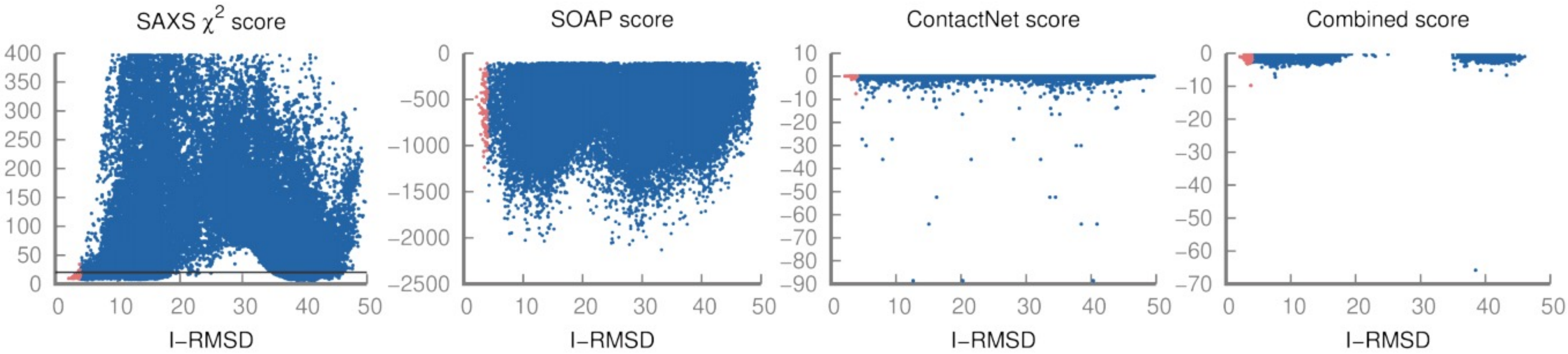
deep learning ContactNet

statistical potential SOAP

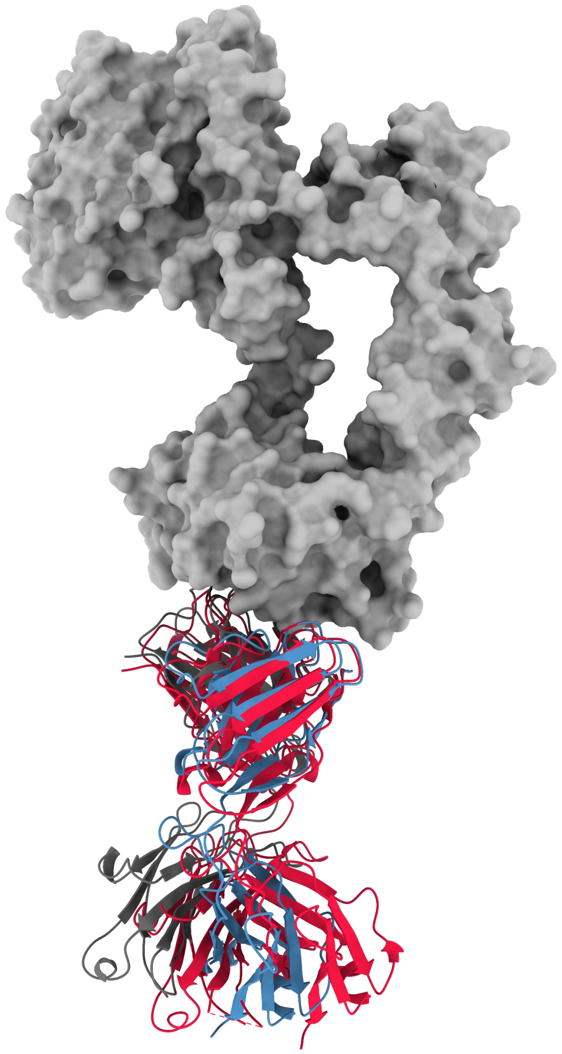
combined score

top scoring models

Docking Results

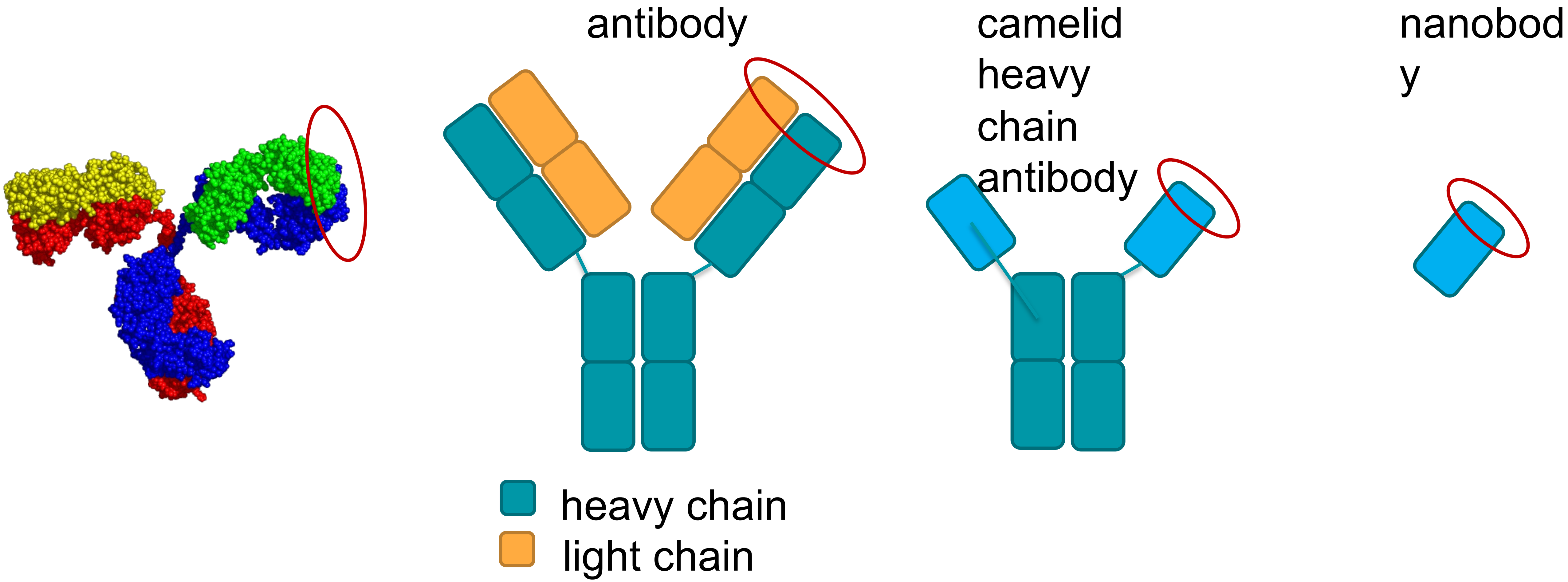


PDB	Rank by SAXS chi²	Rank by SOAP	Rank (IRMSD) by ContactNet	Rank (IRMSD) by combined score
1yy9	1	30	1	1
3b2u	1	2	13	4
3c09	776	192	2122	178
3p0y	139	1228	45	2
3sqo	1022	38	1	1



4/5

Antibody vs. Nanobody (single domain antibody)



antibody

camelid heavy chain antibody

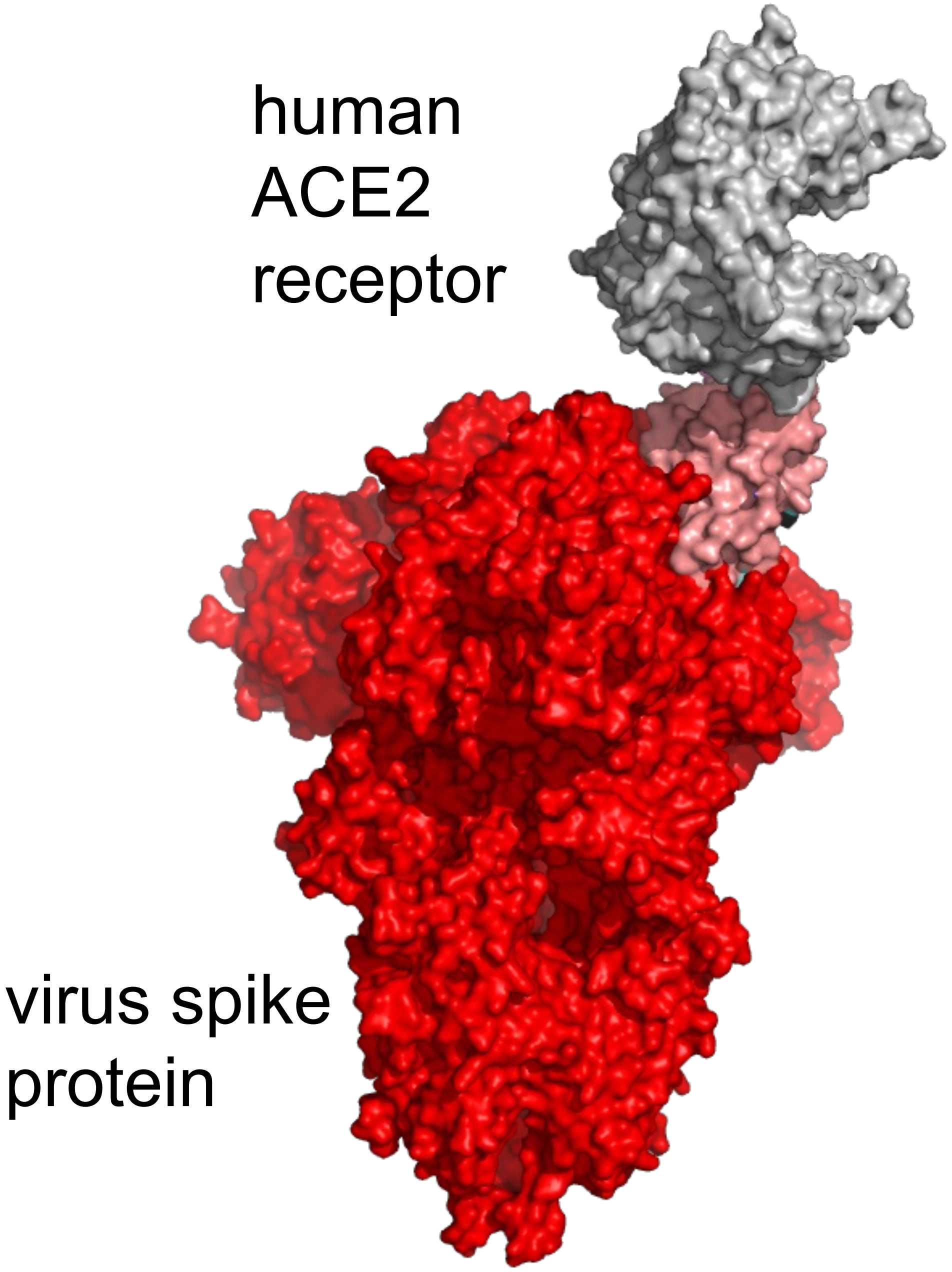
nanobody

🟦 heavy chain

🟧 light chain

- camelids including Llamas, alpacas and camels have a dual-immunoglobulin system

Can nanobodies help?



Nanobody advantages

- Nanobodies can reach **comparable binding affinities** despite smaller size
- small – easy to produce and deliver
- highly stable and soluble
- easily bioengineered
- better tissue penetration
- high sequence similarity to human IgG
- first therapeutic approved recently

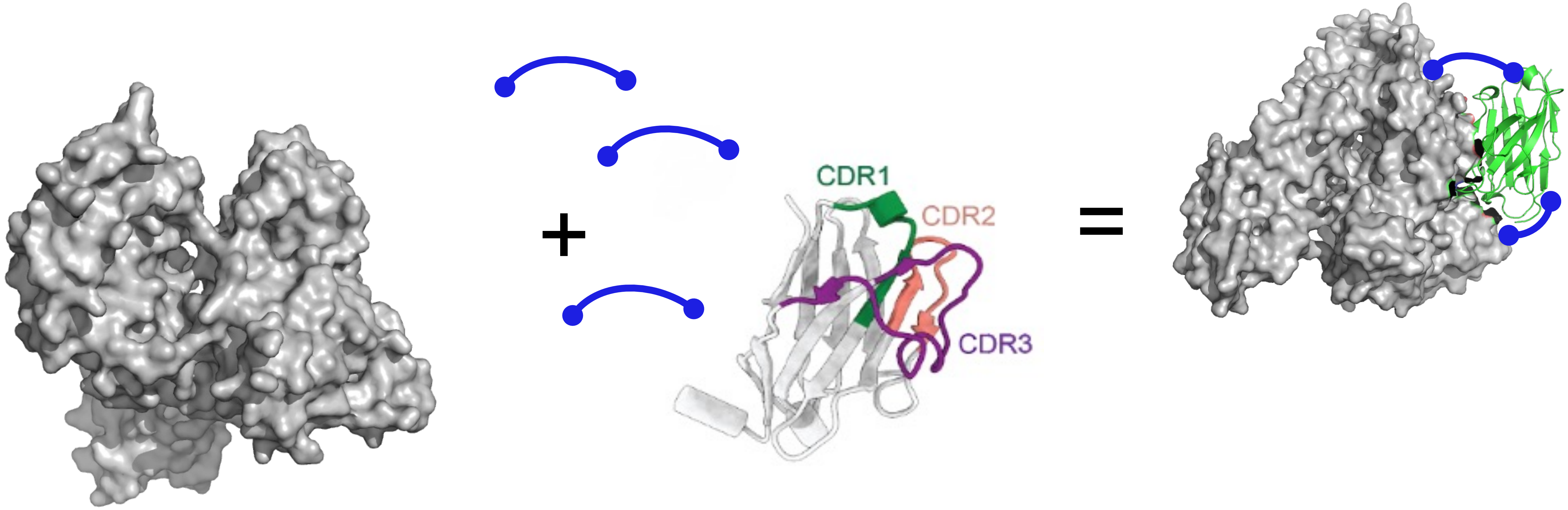
The llama team

Shi lab, University of Pittsburgh

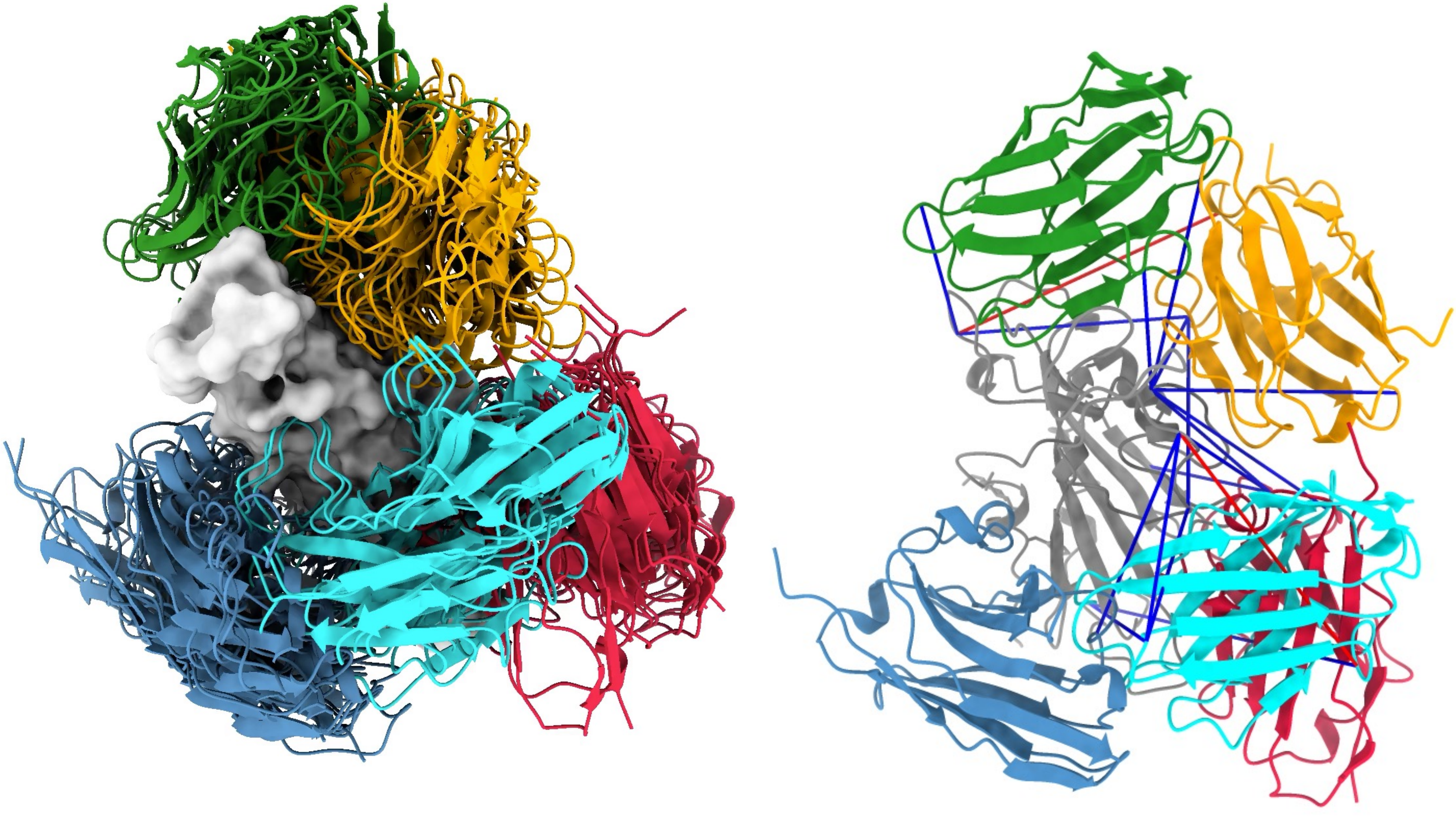
Mass Spectrometry (MS)

Experimental validation by cross-linking MS

- express ~100 nanobodies
- add cross-linker
- use MS to identity residues that were cross-linked by a specific linker



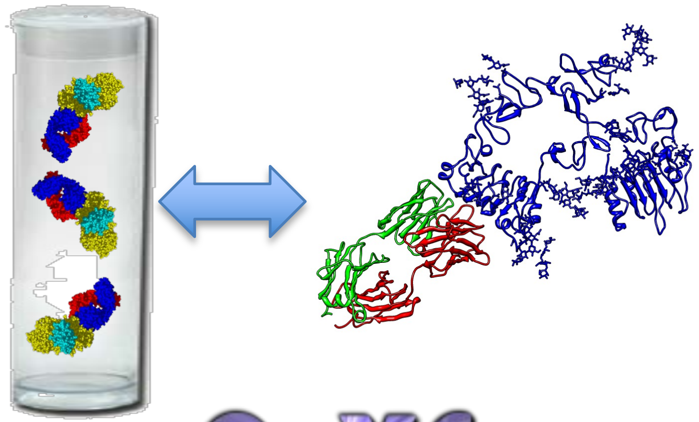
Epitopes of SARS-CoV-2 neutralising nanobodies determined by integrative modeling



Xiang et al. Science 2020

What can we do with the SAXS profile?

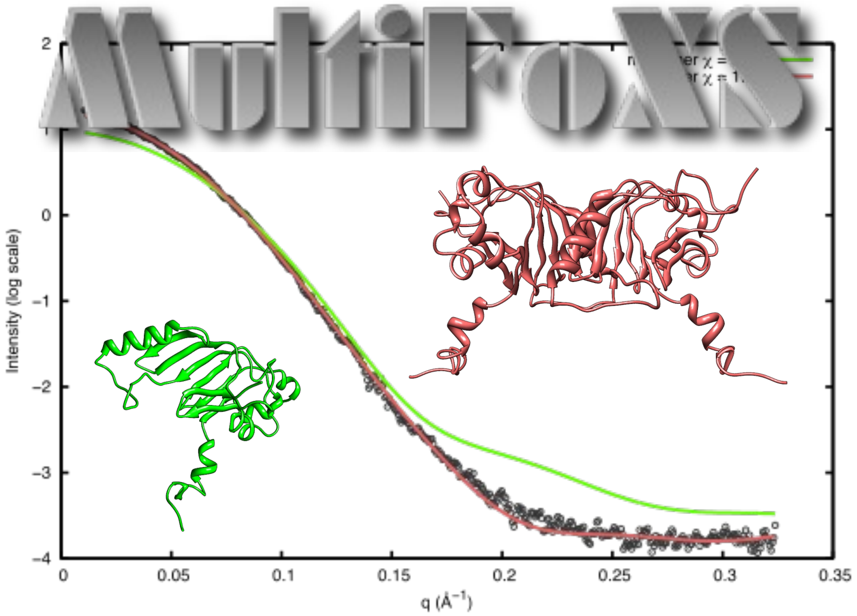
solution structure vs. model



FoXS

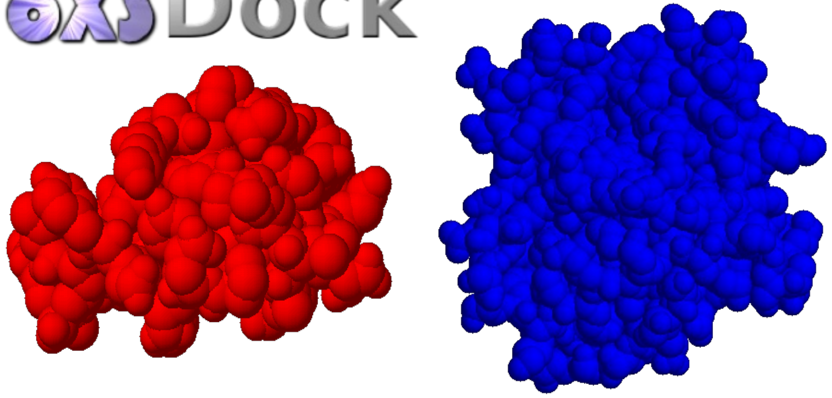
monomer, dimer or mixture?

MultiFoXS

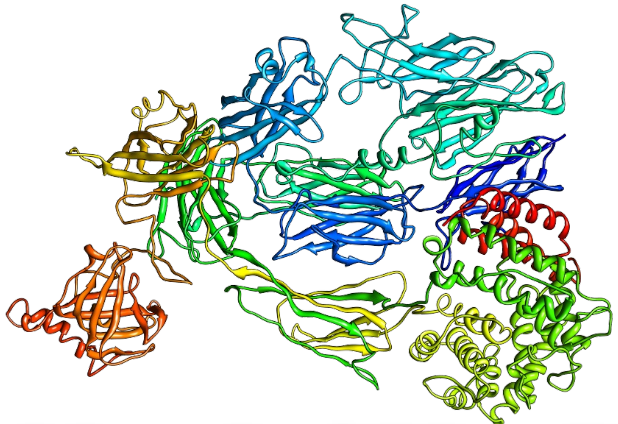


assembly of multi protein complexes

FoXSDock



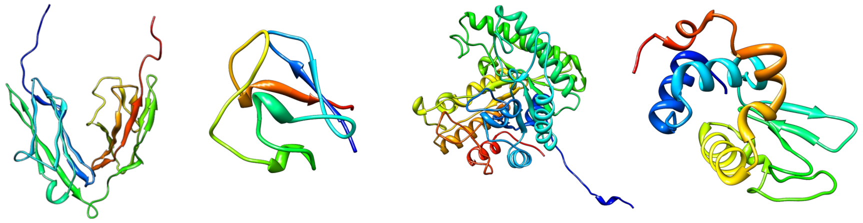
assembly of multi domain proteins



MultiFoXS

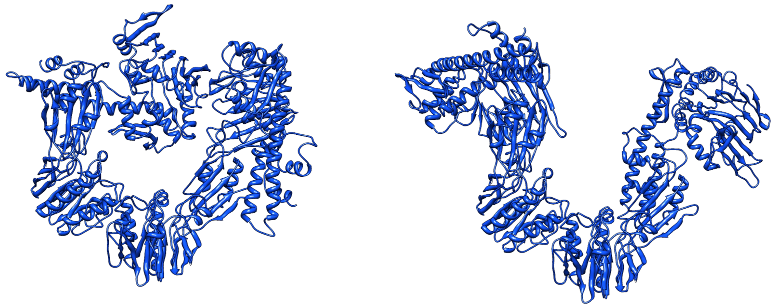
protein folding

Modeller



FoXS

structural characterization of protein dynamics



MultiFoXS

The power of integrative structure modeling

Information

↓

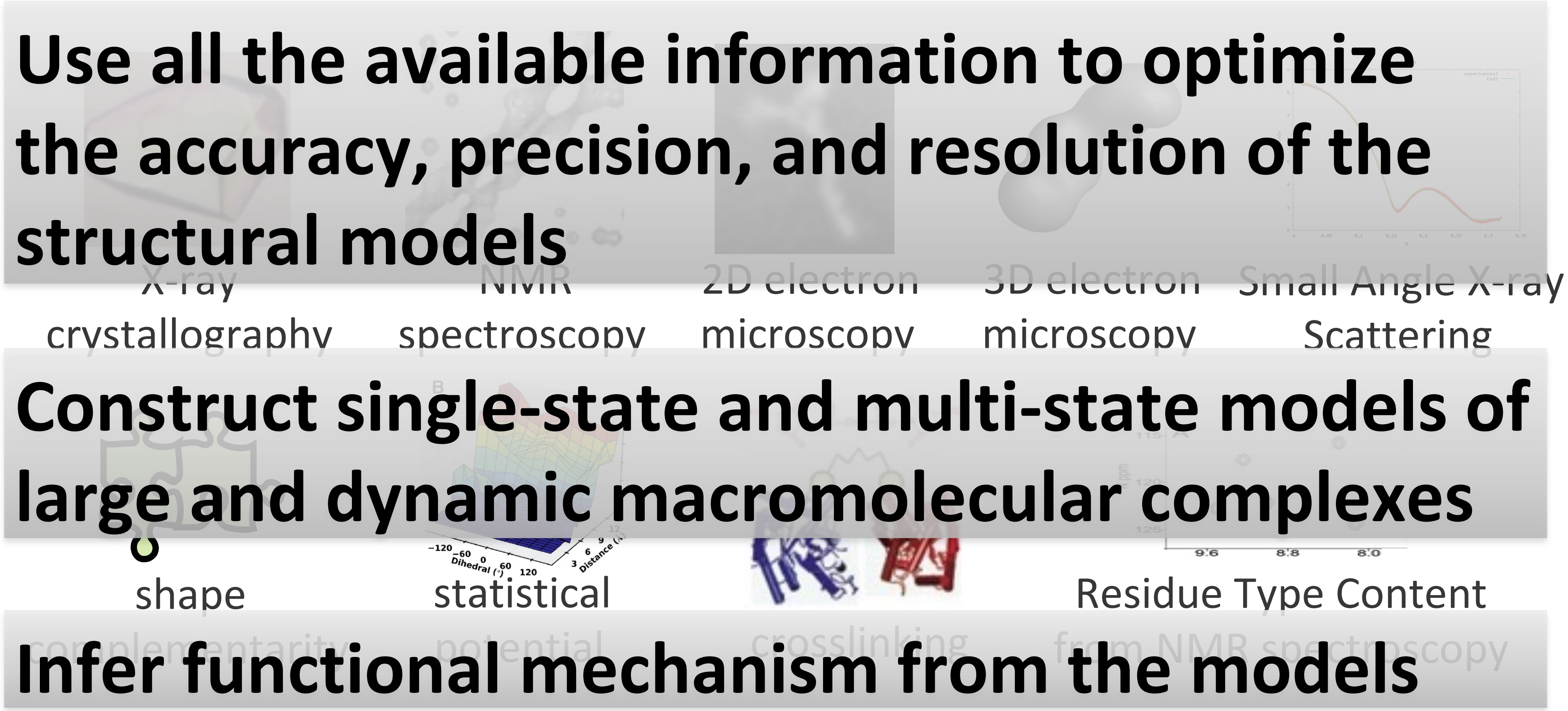
Scoring

↓

Sampling

↓

Analysis



Use all the available information to optimize the accuracy, precision, and resolution of the structural models

X-ray crystallography

NMR spectroscopy

2D electron microscopy

3D electron microscopy

Small Angle X-ray Scattering

Construct single-state and multi-state models of large and dynamic macromolecular complexes

shape complementarity

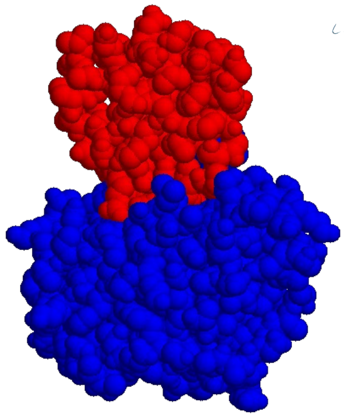
statistical potential

crosslinking

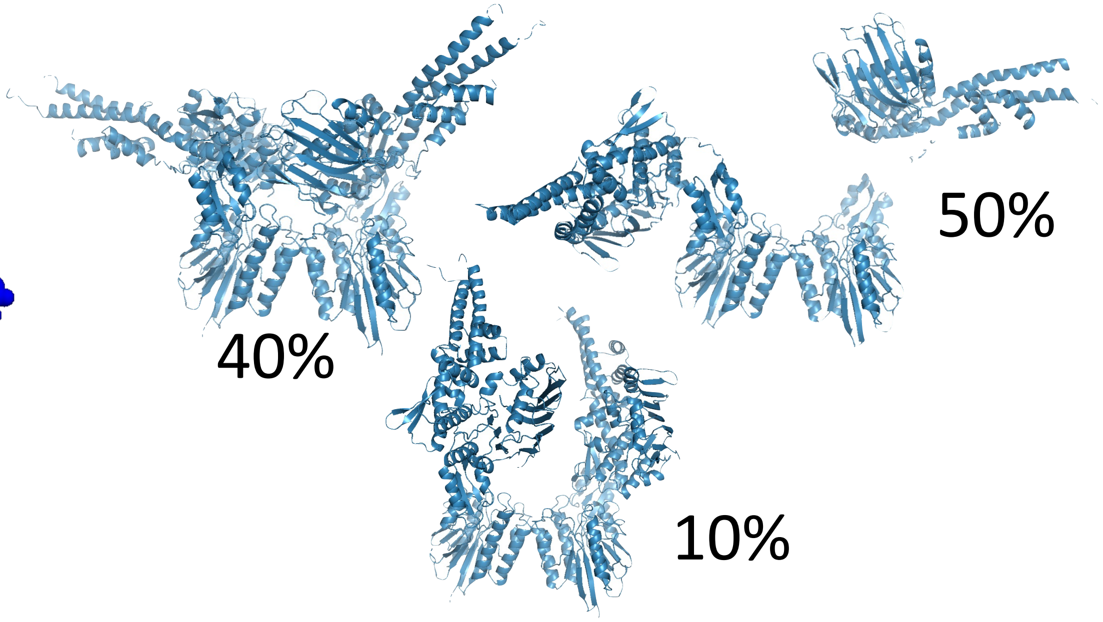
Residue Type Content from NMR spectroscopy

Infer functional mechanism from the models

Protein interactions



Protein dynamics



Thanks!

Lab

Merav Braitbard
Jerome Tubiana
Tomer Cohen
Matan Halfon
Edan Patt
Shon Cohen

BERKELEY LAB
LAWRENCE BERKELEY NATIONAL LABORATORY

Michal Hammel
Greg Hura
Rob Rambo
Susan Tsutakawa
John Tainer

SARS-CoV-2 nanobodies

Yufei Xiang
Zhe Sang
Sham Nambulli
Zhengyun Xiao
Heng Liu
Paul Duprex
Cheng Zhang
Yi Shi

UNIVERSITY OF PITTSBURGH

EGFR antibodies

Arvind Sivasubramanian

Adimab

BSF

FoXS: salilab.org/foxs
FoXSDock: salilab.org/foxsdock
MultiFoXS: salilab.org/multifoxs
source code and executables in IMP: salilab.org/imp