Overview of Current Methods in the Study of Biomolecular Structure by SANS in the Dilute Solution Limit

Susan Krueger
(and many others who I will try
to acknowledge along the way)

Low - q Seminar
October 1, 2014

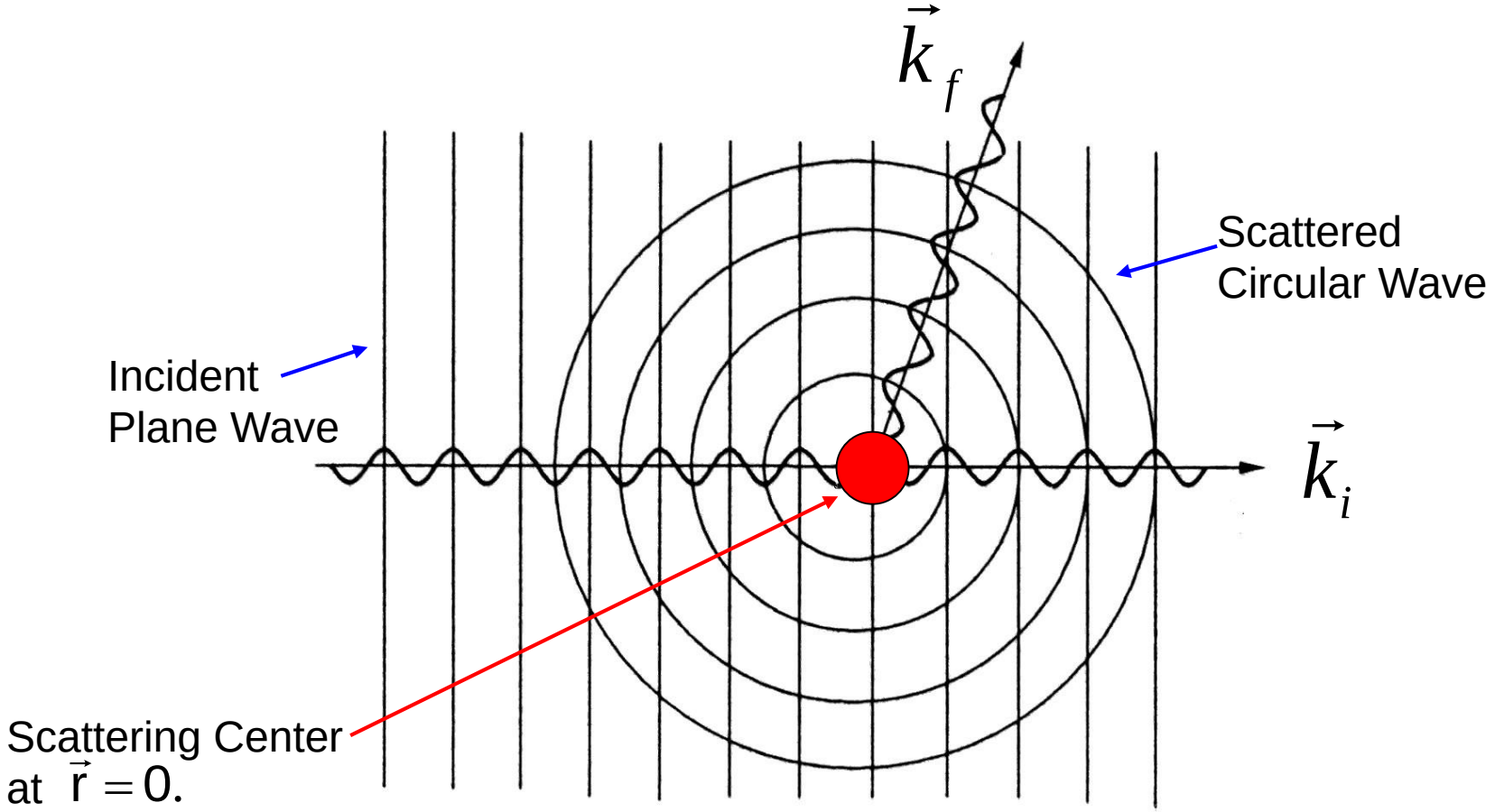
NIST
National Institute of
Standards and Technology

Biology SANS Applications
Dilute Limit

Biological Macromolecules in Bulk Solution

- Proteins
- Nucleic acids
- Protein-nucleic acid complexes
- Multi-subunit protein complexes
- Protein-lipid complexes
- Membranes and membrane components

The Scattering Process: A Review



$$e^{i\vec{k}_i \cdot \vec{r}} \rightarrow e^{i\vec{k}_i \cdot \vec{r}} + f(\Omega)\frac{e^{i\vec{k}_f \cdot \vec{r}}}{r}$$

The Scattering Process

$$\frac{d\sigma}{d\Omega} d\Omega = \frac{\text{neutrons/sec scattered into solid angle } d\Omega}{\text{neutrons/sec/cm}^2 \text{ on target}}$$

$$\Phi \text{ (beam flux)}$$

Differential scattering cross section

$$\sigma = \int \frac{d\sigma}{d\Omega} d\Omega = 4\pi \left| f(\Omega) \right|^2 = 4\pi \left| b \right|^2$$

in units of area (cm^2)
1 barn = 10^{-24} cm^2

Total scattering cross section

Scattering length (per atom)

$$\sigma_{coh} = 4\pi (\bar{b})^2$$

$$\sigma_{inc} = 4\pi \left[(\overline{b^2}) - (\bar{b})^2 \right]$$

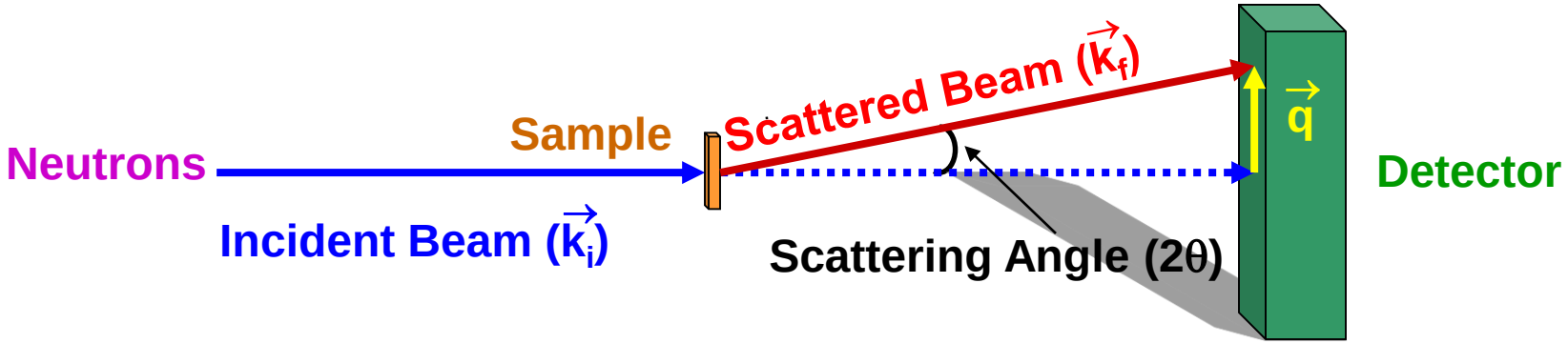
$$\sigma = \sigma_{coh} + \sigma_{inc} + \sigma_{abs}$$

Neutron Cross Sections

Atomic nucleus	b_{coh} (fm)	σ_{coh} (barns)	σ_{inc} (barns)	σ_{abs} (barns)
^{1}H	- 3.741	1.758	80.26	0.333
^{2}D	+ 6.671	5.592	2.05	0.000
^{12}C	+ 6.651	5.559	0.0	0.004
^{14}N	+ 9.37	11.03	0.5	1.91
^{16}O	+ 5.803	4.232	0.0	0.000
^{31}P	+ 5.13	3.307	0.005	0.172
^{32}S	+ 2.804	0.988	0.0	0.54

From the Special Feature section of neutron scattering lengths and cross sections of the elements and their isotopes in *Neutron News*, Vol. 3, No. 3, 1992, pp. 29-37.

Small Angle Scattering



Constructive interference from structures in the direction of $\vec{q}$

Diffraction length scale $d \approx \frac{2\pi}{q}$ where $q = \frac{4\pi}{\lambda}\sin\theta$

$$2q \gg \frac{1}{d} \gg \frac{6\ \text{Å}}{10\ \text{ to }\ 1000\ \text{Å}} \qquad 2q \gg 0.3^\circ\ \text{ to }\ 5^\circ$$

Differential Scattering Cross Section

$$\frac{d\sigma}{d\Omega} = |f(\Omega)|^2 \sum_{ij} e^{-i\vec{q}\cdot(\vec{r}_i - \vec{r}_j)} \qquad \text{where} \quad \vec{q} = \vec{k}_f - \vec{k}_i$$

Coherent: $$\frac{d\sigma}{d\Omega} = \sum_{ij} b_i b_j \; e^{-i\vec{q}\cdot(\vec{r}_i - \vec{r}_j)}$$

All-atom representation

Replace: $$\sum_i^N b_i \rightarrow \int_V r(\vec{r})\, d\vec{r} \qquad \text{where} \quad \rho(\vec{r}) = b_i \delta(\vec{r} - \vec{r}_i)$$

$$\frac{d\sigma}{d\Omega}(\vec{q}) = \frac{1}{N} \left| \int_V r(\vec{r})\, e^{i\vec{q}\cdot\vec{r}}\, d\vec{r} \right|^2$$

Scattering length density

Differential Scattering Cross Section

Normalizing by sample volume :

$$\frac{d\Sigma}{d\Omega}(\vec{q}) = \frac{N}{V}\frac{d\sigma}{d\Omega}(\vec{q}) = \frac{1}{V}\left|\int_V \rho(\vec{r})\, e^{i\vec{q}\cdot\vec{r}}\, d\vec{r}\right|^2$$

Rayleigh-Gans Equation

$$r(\vec{r}) = r + r_f(\vec{r}) \gg r$$

Average scattering length density

Fluctuations about the average

and

$$r = \frac{\sum_i^N b_i}{V}$$

Uniform scattering length density

Neutron Scattering Length Densities

H_2O: $\rho = -0.562 \times 10^{10}$ cm^{-2}
D_2O: $\rho = 6.404 \times 10^{10}$ cm^{-2}
H_2O:D_2O mixture: $\rho = (-0.562 + 6.966y) \times 10^{10}$ cm^{-2},
where y = fraction of D_2O

	ρ in H_2O (10^{10} cm^{-2})	ρ in D_2O (10^{10} cm^{-2})
protein	1.8	3.2
DNA	3.4	4.1
RNA	3.6	4.3
PC Lipid Head Group	1.1	1.1
CH_2	-0.31	-0.31
CH_3	-0.85	-0.85

SLDs in H_2O and D_2O depend on H:D exchange.

The Scattered Intensity

Rayleigh-Gans Equation:

$$I(q) \propto \frac{d\Sigma}{d\Omega}(\vec{q}) = \frac{1}{V}\left|\int_V \rho(\vec{r})\, e^{i\vec{q}\cdot\vec{r}}\, d\vec{r}\right|^2$$

Assume there are N randomly-oriented, homogeneous particles:

$$\frac{dS}{d\Omega}(\vec{q}) = \frac{N}{V} r^2 V^2 \left\langle \left| \frac{1}{V} \int e^{i\vec{q}\cdot\vec{r}}\, d\vec{r} \right|^2 \right\rangle$$

Average over all orientations

Particle volume

$F(\vec{q})$ (depends on particle shape)

Macromolecules in Solution

Reciprocal Space

Real Space

$I(q)$ — Macromolecule in Solvent

μ

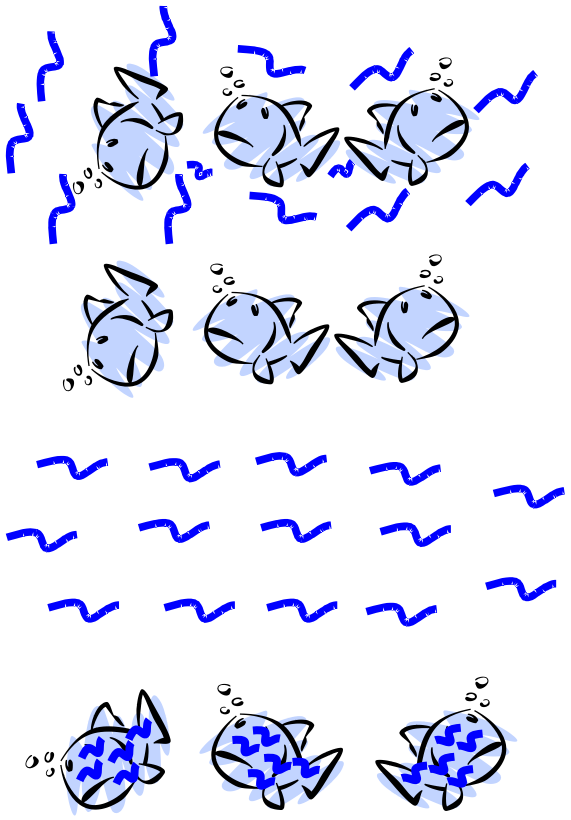
$r \int_V e^{i\vec{q}\cdot\vec{r}}\, d\vec{r}$ — Uniform Scattering Length Density, ρ, in V

+

$\rho_s\, \delta(0)$ — Solvent of Infinite Extent (Not Observed!)

-

$r_s \int_V e^{i\vec{q}\cdot\vec{r}}\, d\vec{r}$ — Solvent Scattering Length Density, ρ_s, in V



$$I(q) = \frac{N}{V}(\Delta r)^2 V^2 \left\langle \left| F(\vec{q}) \right|^2 \right\rangle, \text{ where } \Delta r = r - r_s$$

The Guinier Approximation

$$I(q) @ I(0)\exp(-q^2 R_g^2 / 3)$$

valid when $qR_g \leq 1$

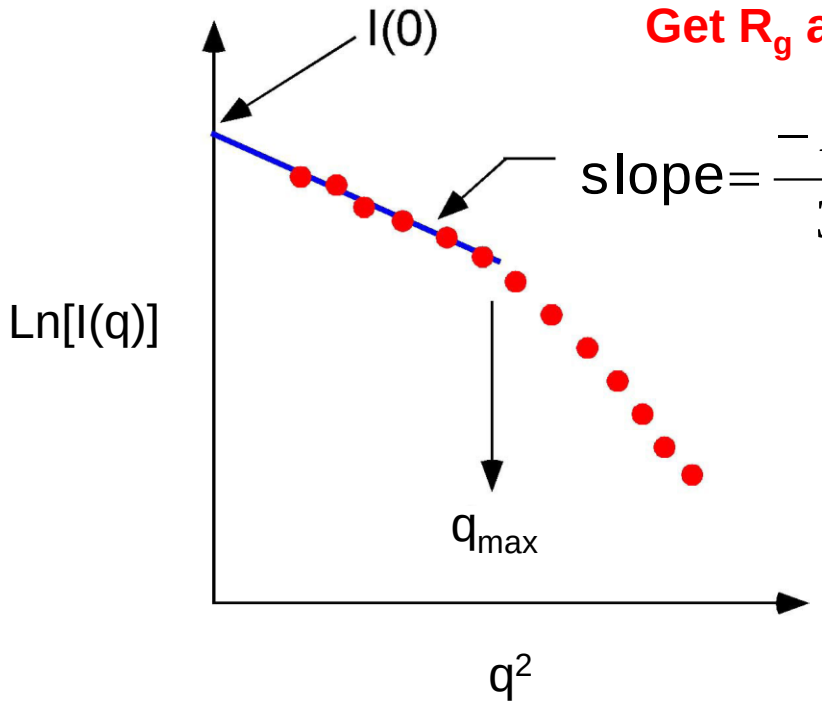
$$I(0) = \frac{d\Sigma(0)}{d\Omega} = \frac{N}{V}(\Delta\rho)^2 V^2$$

Data on absolute scale

R_g: radius of gyration (about the particle C.M.)

$$R_g^2 = \frac{\int \rho(\vec{r}) r^2 d\vec{r}}{\int \rho(\vec{r}) d\vec{r}}$$

Guinier Plot



Get R_g and I(0) from Guinier fit

$q_{max}R_g \leq 1$

$$\ln\left[I(q)\right] = \ln\left[I(0)\right] - \frac{q^2 R_g^2}{3}$$

The Forward Scattering on an Absolute Scale

$$I(0) = \frac{d\Sigma(0)}{d\Omega} = \frac{N}{V}(\Delta\rho)^2 V^2$$

Number density of particles

Concentration of particles

Molecular weight

$$I(0) = \frac{cN_A}{M_w}(\Delta\rho)^2\left(\frac{M_w\bar{v}}{N_A}\right)^2 = \frac{cM_w\bar{v}^2}{N_A}(\Delta\rho)^2$$

Partial Specific Volume

Modified Guinier Analyses

Rod-shaped Particles where l>>r

$$I(q) \cong \frac{1}{q} I_c(0) \exp(-q^2 R_c^2 / 2)$$

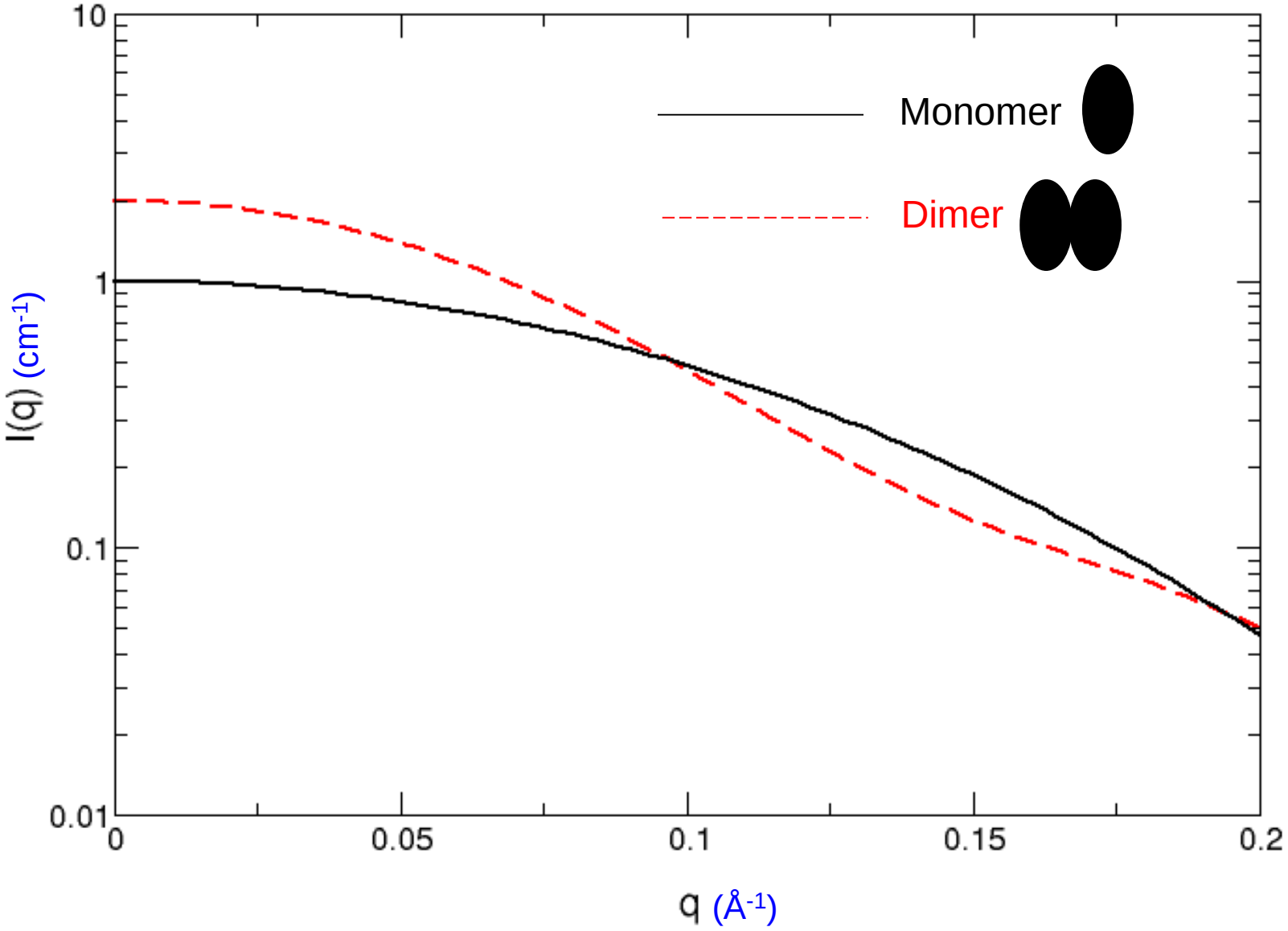
Plot ln[q·I(q)] vs q^2; R_c related to cross-sectional radius

Disk-Shaped particles where r>>t

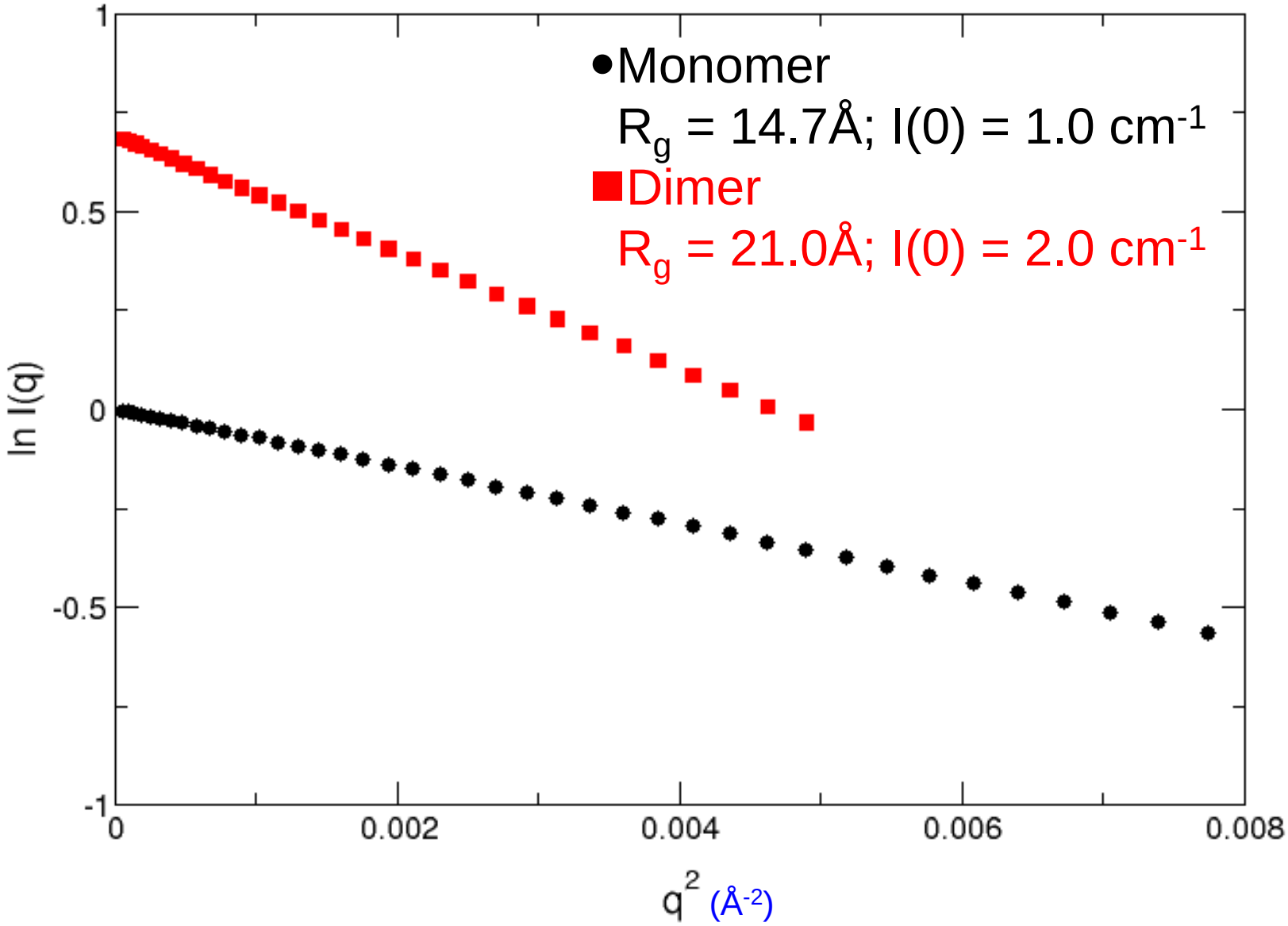
$$I(q) \cong \frac{1}{q^2} I_t(0) \exp(-q^2 R_t^2 / 12)$$

Plot ln[q^2·I(q)] vs q^2; R_t related to thickness

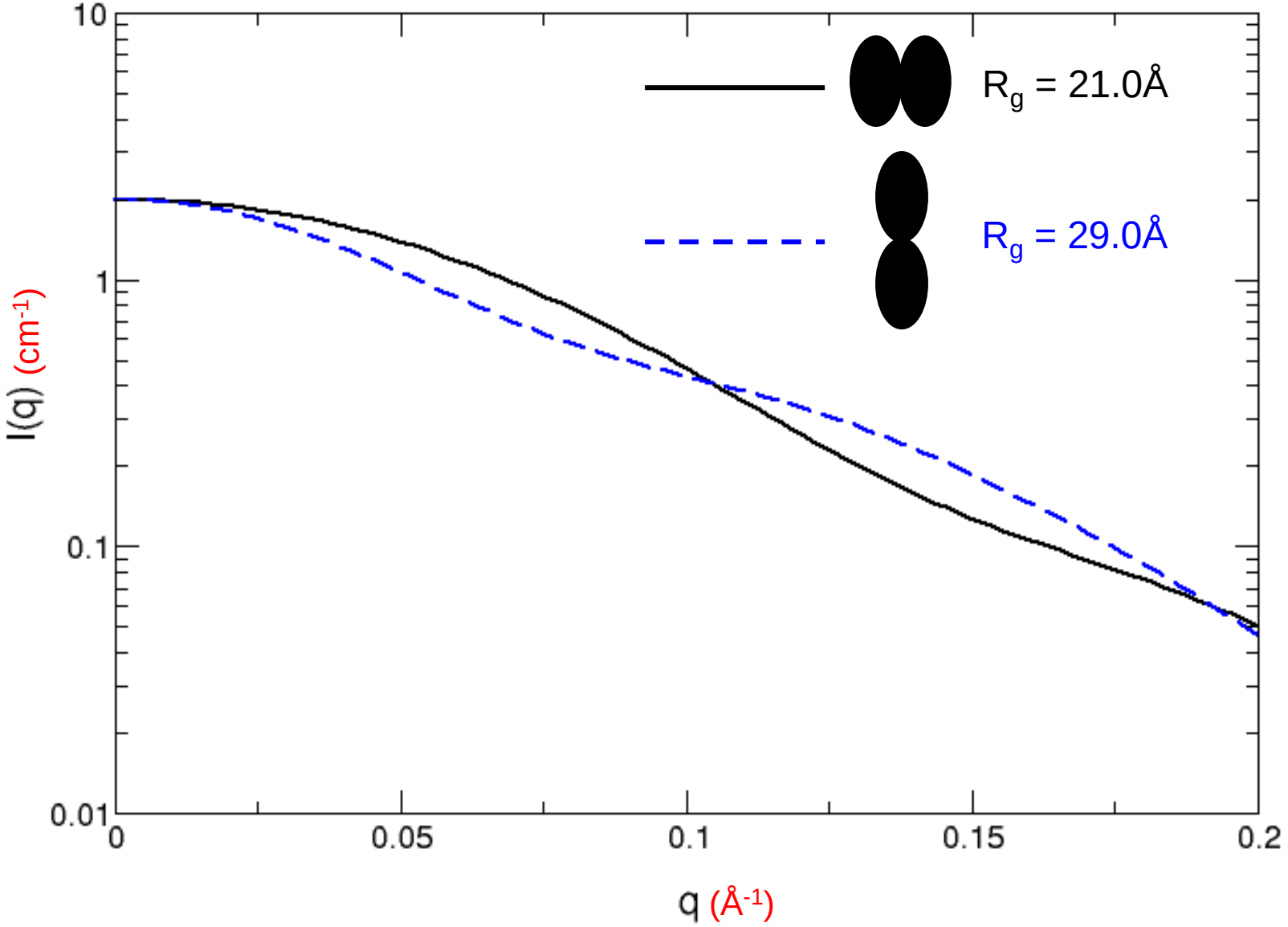
Low Resolution Shape Models



Guinier Fits



Simple Dimer Models



Distance Distribution Function

$$P(r) = r^2\gamma(r) \qquad \gamma(r) = \frac{1}{2\pi^2 r}\int qI(q)\sin(qr)dq$$

Debye-Porod Correlation Function

$4\pi P(r) \equiv$ number of distances within the molecule

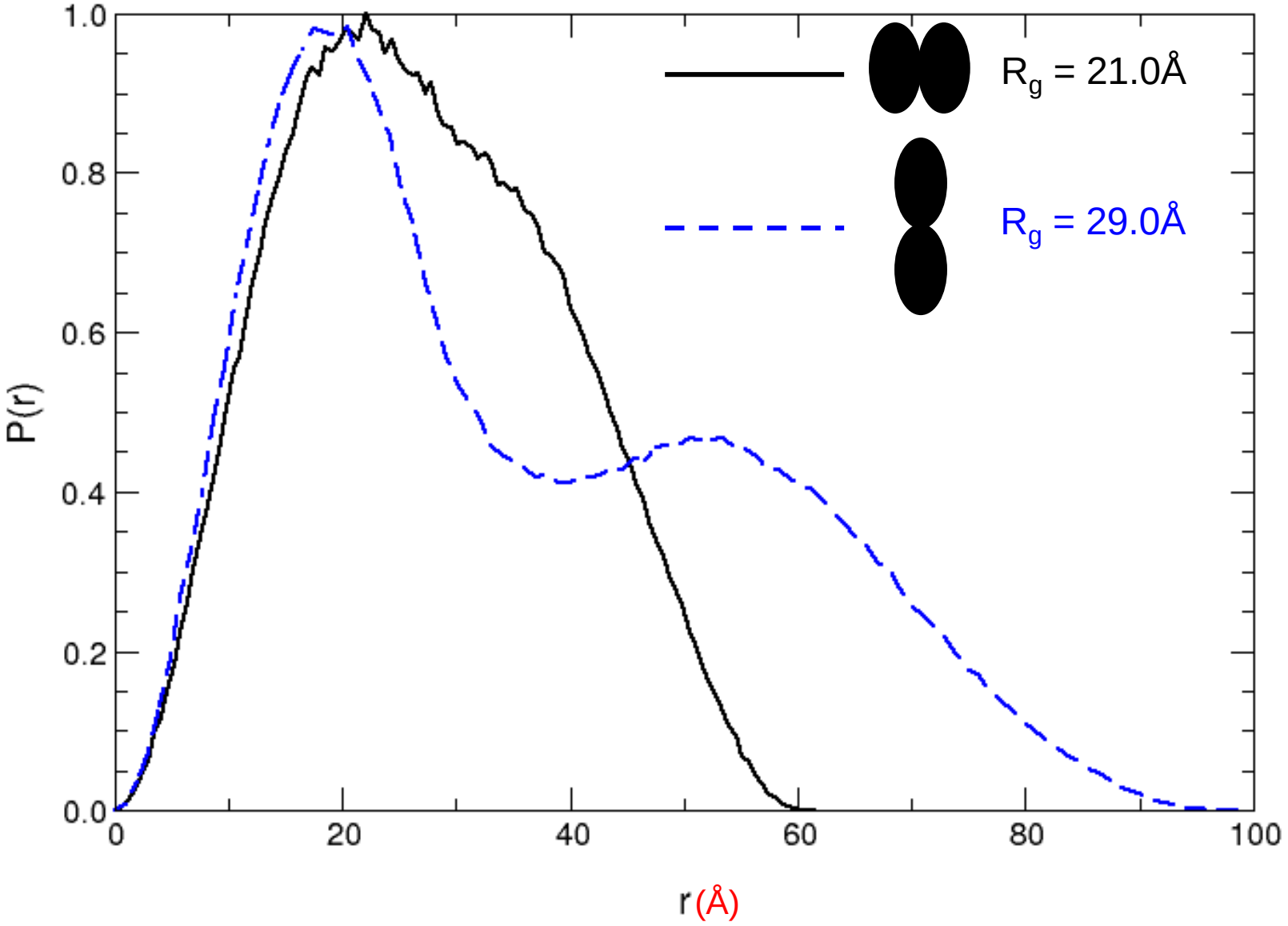
$$I(q) = 4\pi V \int_0^{D_{max}} P(r)\frac{\sin(qr)}{qr}\,dr$$

$D_{max} \equiv$ maximum distance within the molecule

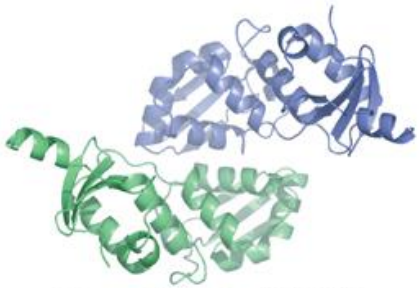
$$P(0) = 0$$

$$P(2r \geq D_{max}) = 0$$

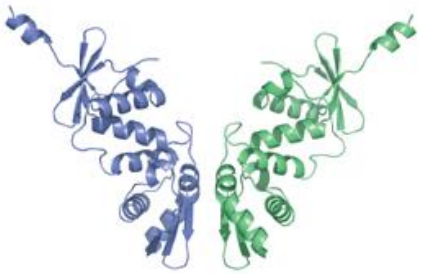
Distance Distribution Function



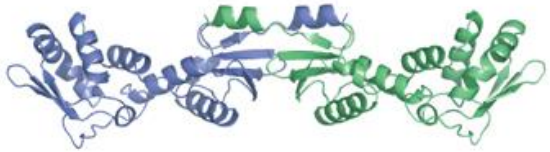
P(r) for All-atom Models



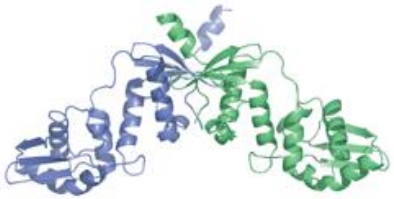
Dimer 1, (B/2 = 755 Å^2)
R_G = 26.08 Å D_{max} = 80 Å



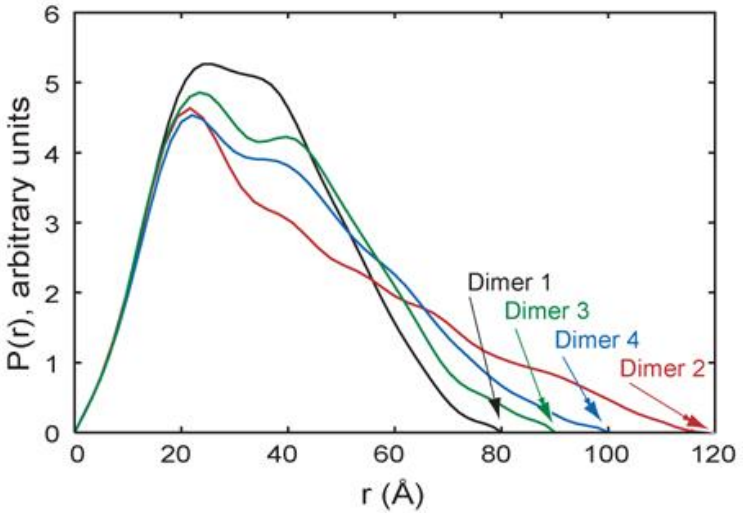
Dimer 3, (B/2 = 406 Å^2)
R_G = 28.3 Å D_{max} = 90 Å



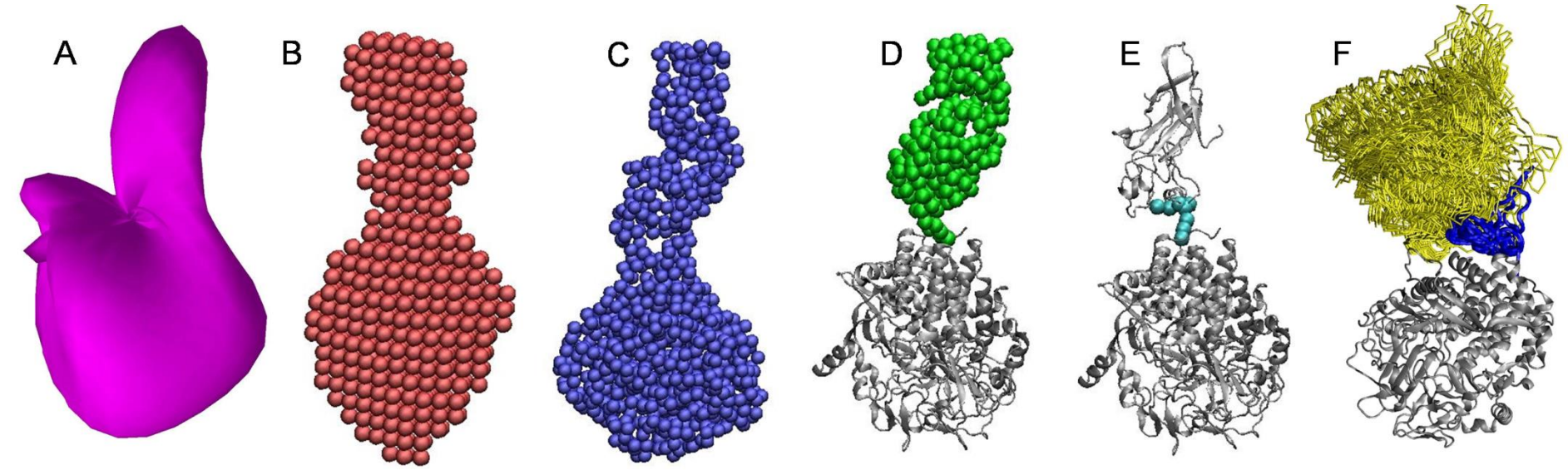
Dimer 2, (B/2 = 923 Å^2)
R_G = 34.04 Å D_{max} = 120 Å



Dimer 4, (B/2 = 255 Å^2)
R_G = 30.4 Å D_{max} = 100 Å



Advanced Modeling Techniques



A	B	C	D	E	F
Envelop representation using spherical harmonics	Envelop from densely packed dummy beads	Envelop from dummy residues forming a chain-compatible model	Missing domain represented by ensemble of dummy residues forming a chain-compatible model.	Rigid body model + missing loop represented by ensemble of dummy residues	Atomic models derived from rigid body modeling applying conformational sampling

Modeling efforts are underway at most neutron and x-ray facilities.

All the modeling in the world doesn't help if your sample isn't good…

- Concentration: 1-5 mg/mL → monodisperse, non-interacting
- Volume: 350-700 μL per sample
- Data collection time: 0.5-6 hrs per sample
- Typical biology experiment: 1-3 days
- Deuterated solvent is highly desirable.
- Multiple concentrations are usually necessary.
- Specific deuteration may be necessary.
- Multiple solvents of different deuteration → contrast variation

Practical Issues

Sample Preparation Challenges

Mondispersity	suitable solvent conditions
Complex formation	under multiple contrast conditions (D_2O effects)
Large quantities	needed for multiple contrasts
Deuterium labeling	many neutron facilities now have labs to support users
Analytical tools	SEC-MALS, AUC (SE and SV), DLS

Experiment Planning Tools

Contrast Calculator (SASSIE)

Starting All-atom and Hybrid hi-res/low-res Structures (for *in silico* contrast variation experiments)

proteins
nucleic acids
lipids
carbohydrates

Intrinsically Disordered Proteins

A Definition:

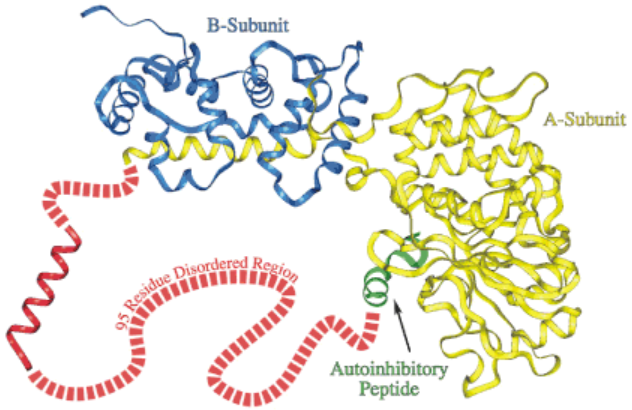
Proteins or regions of proteins that fail to form specific 3-D structure under physiologic conditions *in vitro*.

from Le Gall et al., J. Biomolecular Struct. and Dynamics **24**, 325 (2007)

Backbone Ramachandran angles vary significantly over time with no specific equilibrium values.

Software Needs:

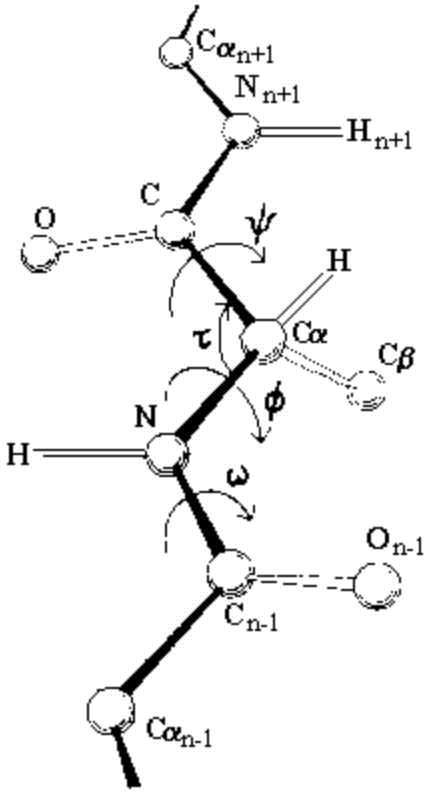
- **>10,000 structures in minutes**
- **"Thumbs up" or "Thumbs down" on model structures**



(Image adapted from: Kissinger CR, et al. 1995. "Crystal structures of human calcineurin and the human FKBP12-FK506-calcineurin complex." Nature 378:641-4.)

http://www.disprot.org/

SASSIE: An Approach to Structure Modeling



Geometric ensemble sampling based on energetics of dihedral angle motion.

Build Starting Structure

- Include "missing" residues

Dihedral Phase Space Search

- Pick regions to vary
- Pick $\Delta\Phi$, $\Delta\Psi$, overlap (basis and cutoff)
- Generate Structures

Calculate SAS

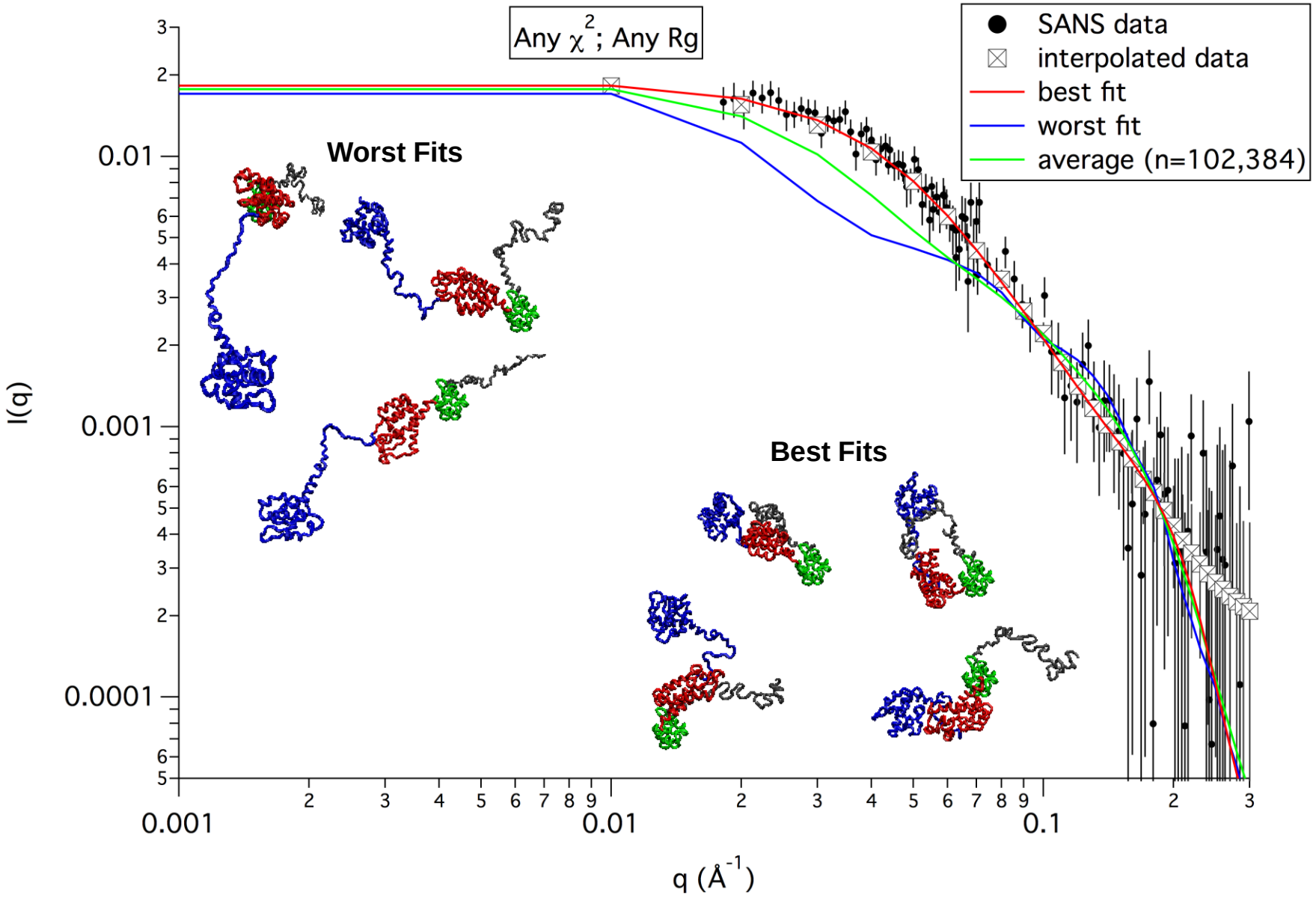
Filter Results

- Compare to SAS data

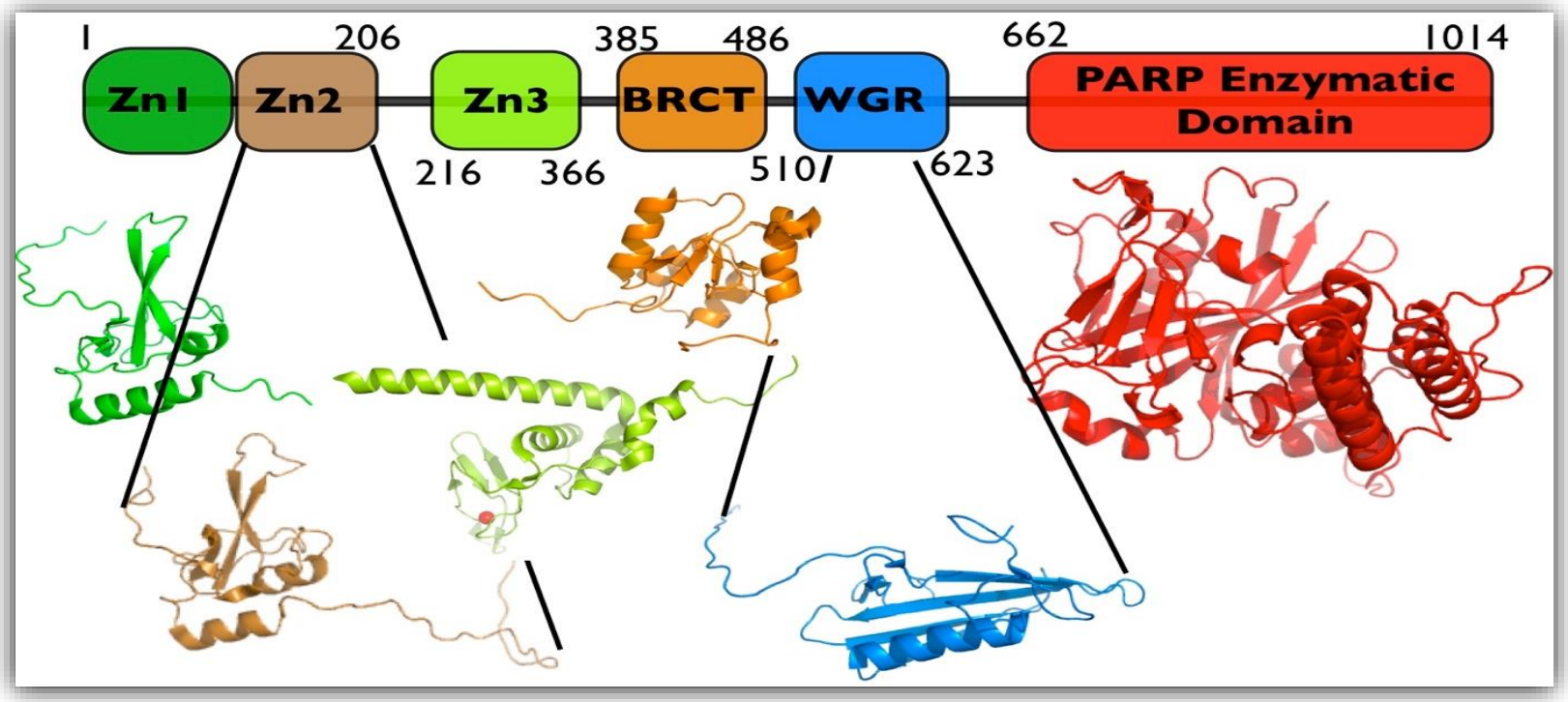
SASSIE Software

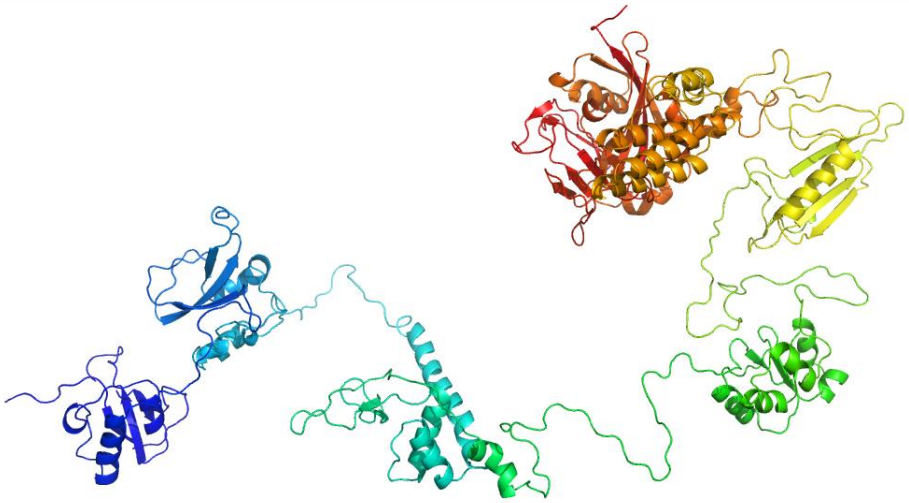
BUILD	Clean up and organize coordinate files Build topology files (force field) All-atom or coarse-grain
TOOLS	Structure alignment, centering, translation, rotation Coordinate manipulation, data interpolation Contrast calculator, experiment planning
INTERACT	Graphically “move” structures Calculate SANS and reflectivity curves in real time
SIMULATE	Protein/RNA dihedral search; grid search Structure minimization; torsion angle MD; normal modes; free energy solvation
CALCULATE	SANS, SAXS, reflectometry EM to SANS; HYDROPRO Spin echo; backscattering, TOF
ANALYZE	Chi-squared filtering to SAS data Density plots of conformation space

All-atom Modeling: SASSIE



PARP-1 Solutions: SASSIE

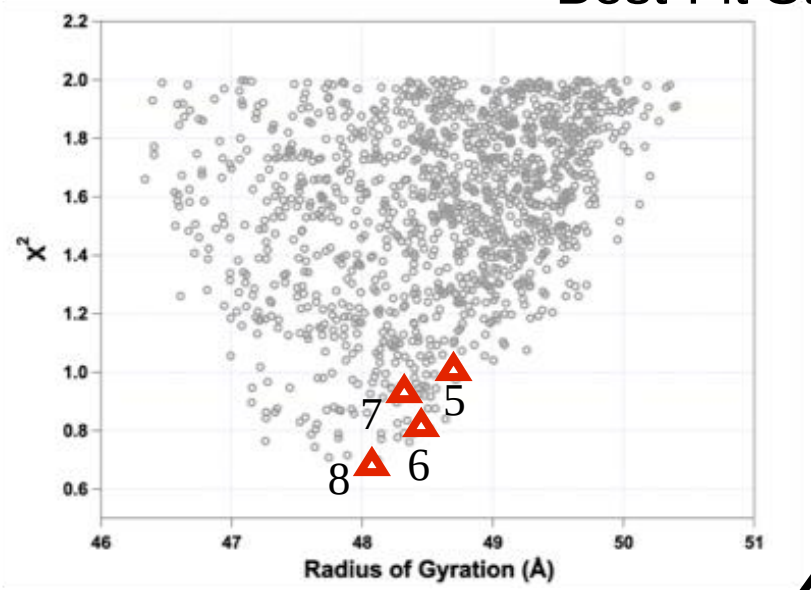


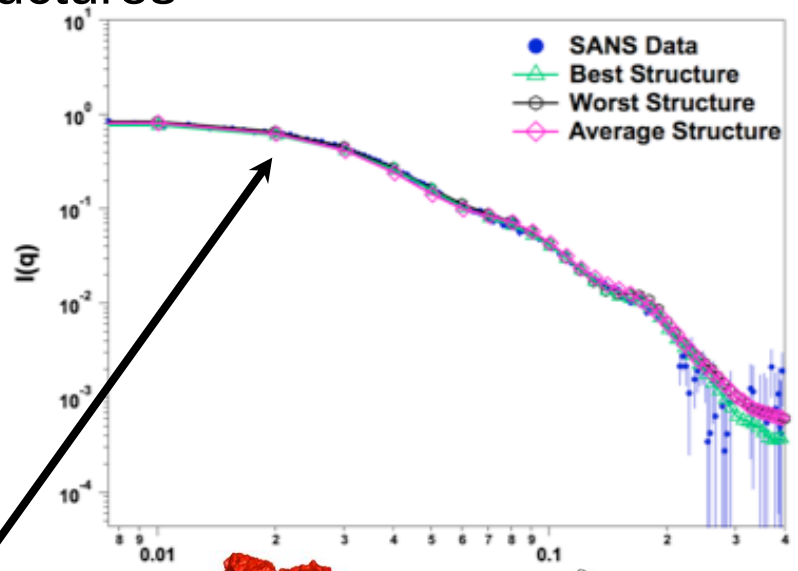


One structure from an ensemble that fits the SAXS and SANS data.

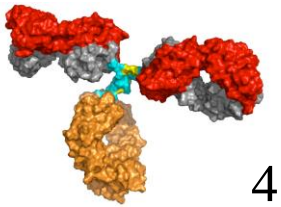
Antibody Ensemble: SASSIE

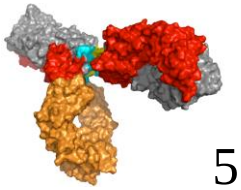
Best-Fit Structures

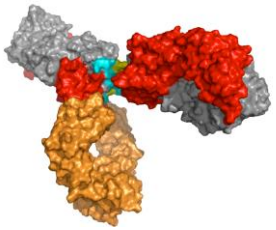


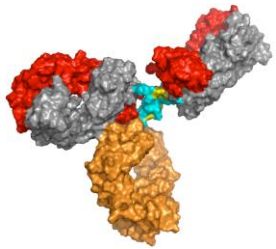


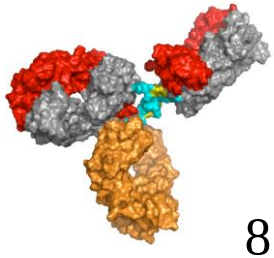
We can use the average fit as a form factor


4


5

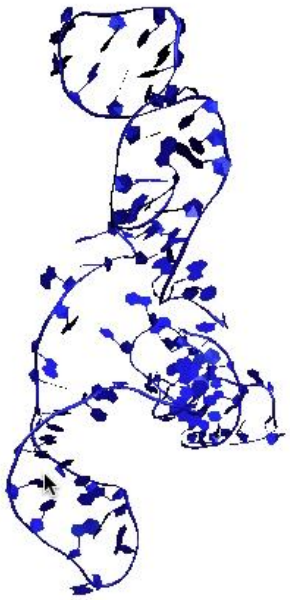

6


7


8

J. Phys. Chem B., 2013, 17 (45):14029-38

Single Stranded RNA: SASSIE



Single stranded RNA Molecular Monte Carlo – Relatively New

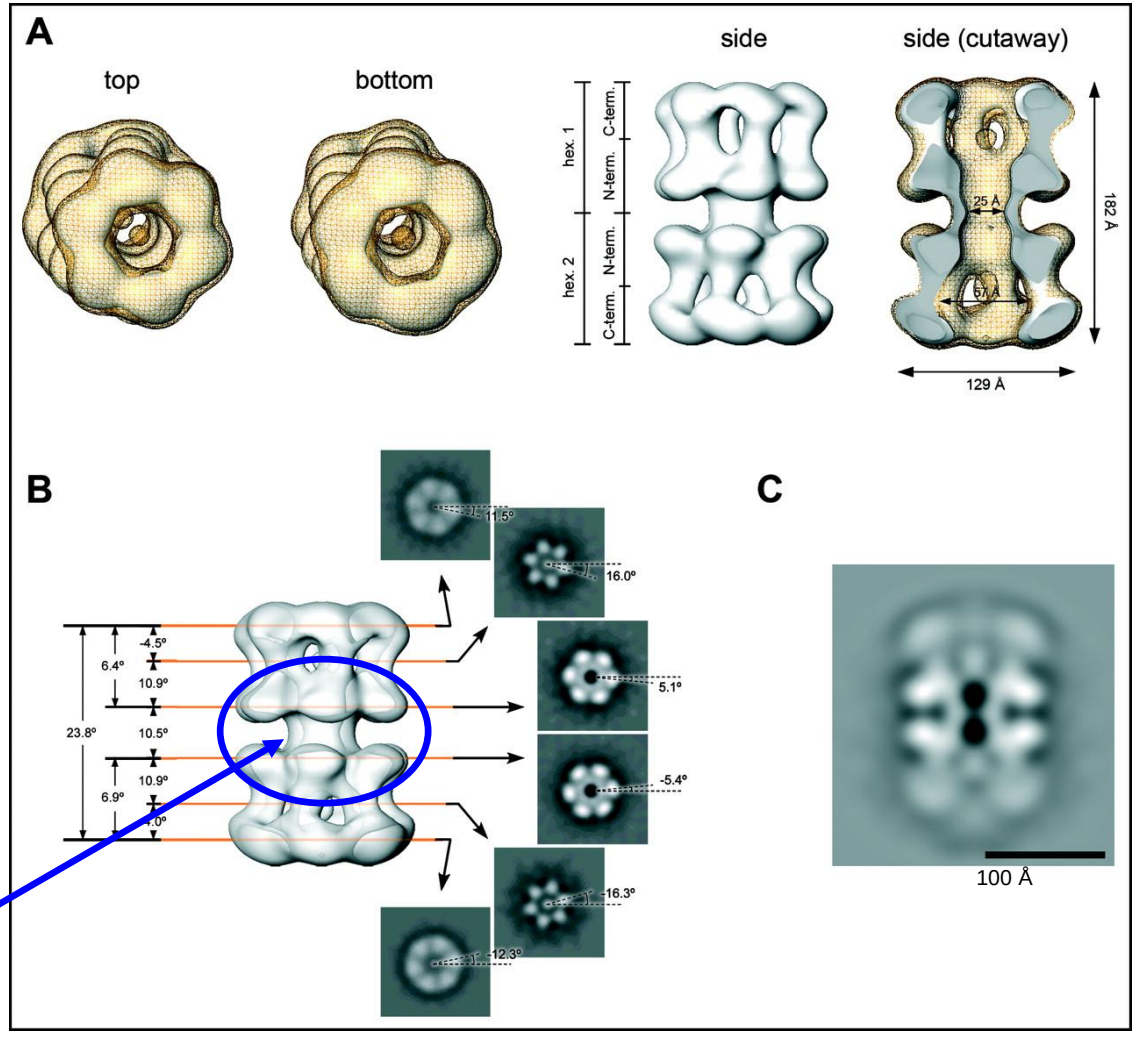
Complexes: MCM DNA Helicase

a

b

```
1   MMKTVDKSKTLTKFEEFFSLQDYKDRVFEAIEKYPNVRSIEVDYLDLEMFDPDLADLLIE
61  KPDDVIRAAQQAIRNIDRLRKNVDLNIRFSGISNVIPLRELRSKFIGKFVAVDGIVRKTD
121 EIRPRIVKAVFECRGCMRHHAVTQSTNMITEPSLCSECGGRSFRLLQDESEFLDTQTLKL
181 QEPLENLSGGEQPRQITVVLEDDLVDTLTPGDIVRVTGTLRTVRDERTKRFKNFIYGNYT
241 EFLEQEFEELQISEEDEEKIKELAGDPNIYEKIIRSTAPSIHGYREVKEAIALQLFGGTG
301 KELDDKTRLRGDIHILIVGDPGIGKSQMLKYVSKLAPRGIYTSGKGTSGVGLTAAAVRDE
361 FGGWSLEAGALVLGDKGNVCVDELDKMREEDRSAIHEALEQQTISIAKAGIMATLNSRCS
421 VLAAANPKFGRFDSYKSIAEQIDLPSTILSRFDLIFVVEDKPDEEKDRELARHILKTHKE
481 DHMPFEIDPELLRKYIAYARKNVRPVLTDEAMQVLEDFYVSMRASAADEDSPVPITARQL
541 EALVRLSEASAKIKLKEHVEAEDARKAIKLSQACLKQVGYDPETGKIDIDKVEGRTPKSE
601 RDKFRLLLELIKEYEDDYGGRAPTNILITEMMDRYNVSEEKVEELIRILKDKGAIFEPAR
661 GYLKIV
```

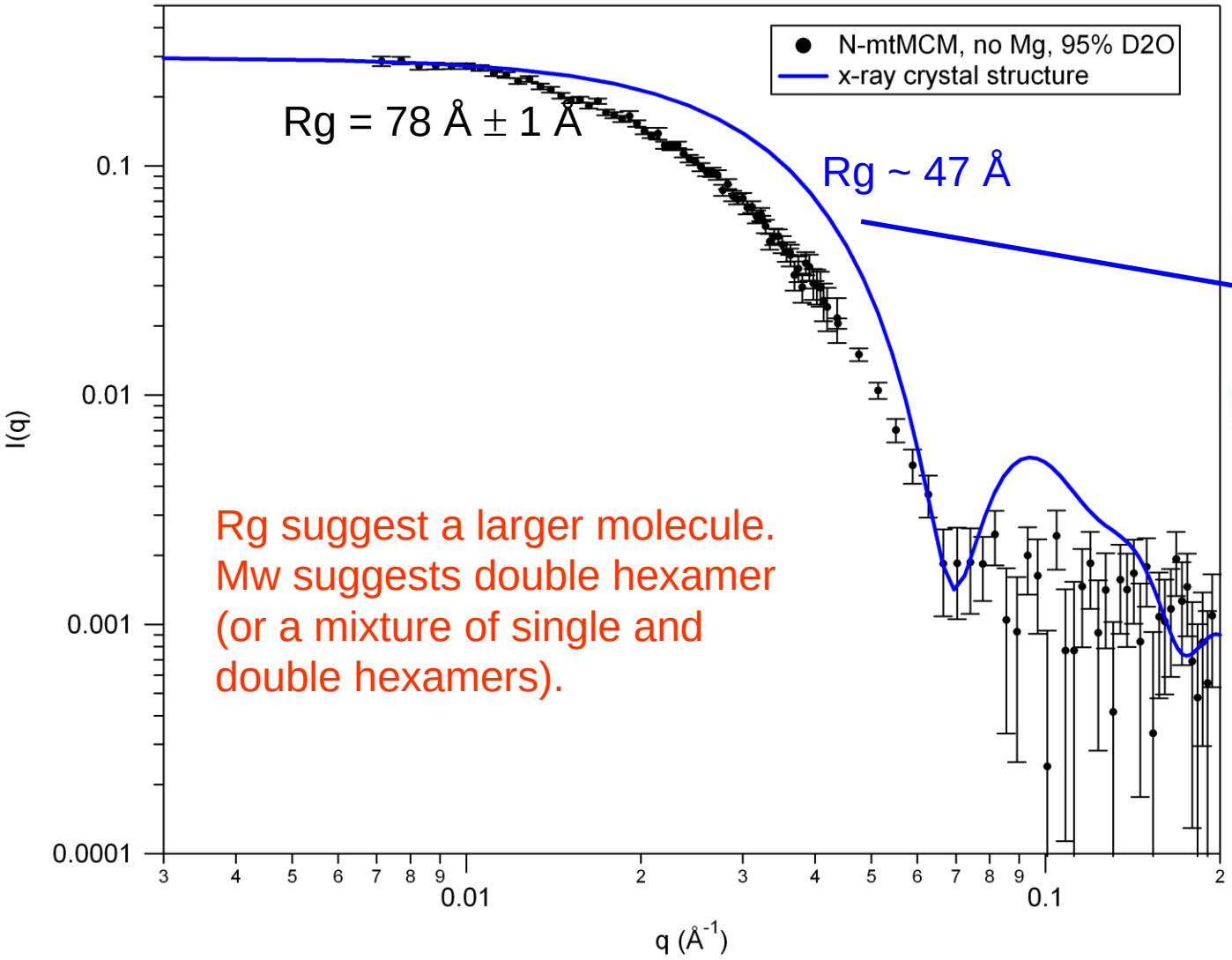
ONE Cryo-EM Structure of MCM



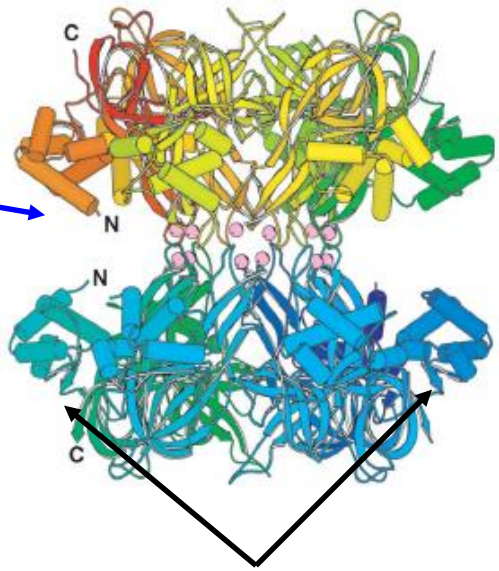
Crystal structure is available for the N-terminal domains.

jbc

Solution Structure of N-terminal Domain of MCM DNA Helicase



N-terminal domain



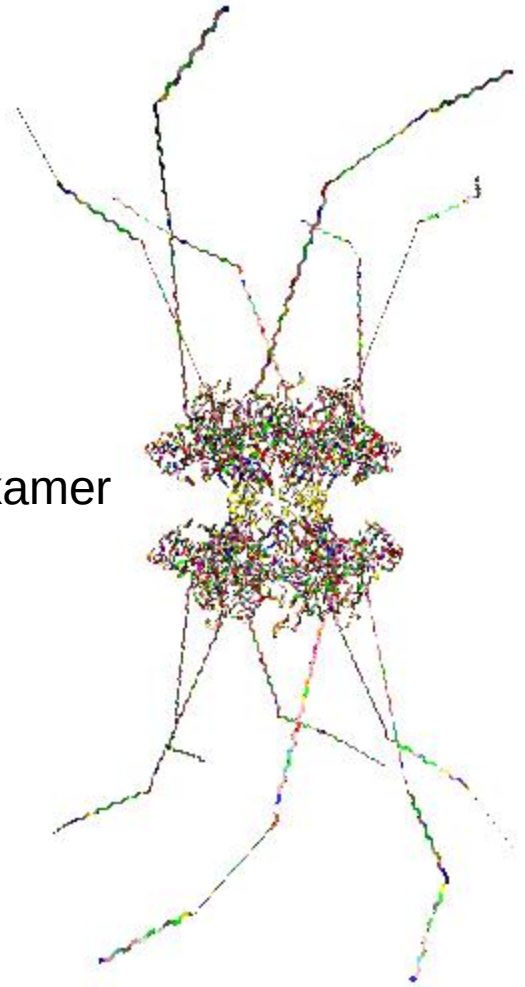
Flexible Regions

>30 missing C-terminal residues

Build a Starting Structure

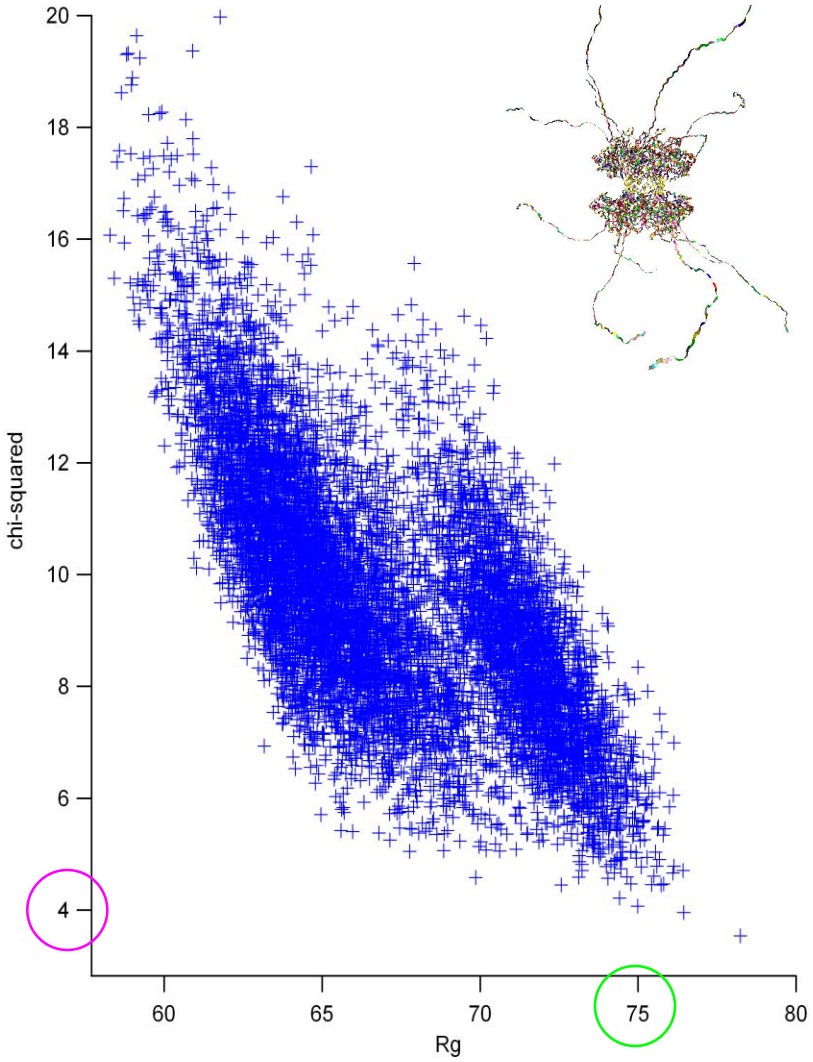
Monomer

Double Hexamer

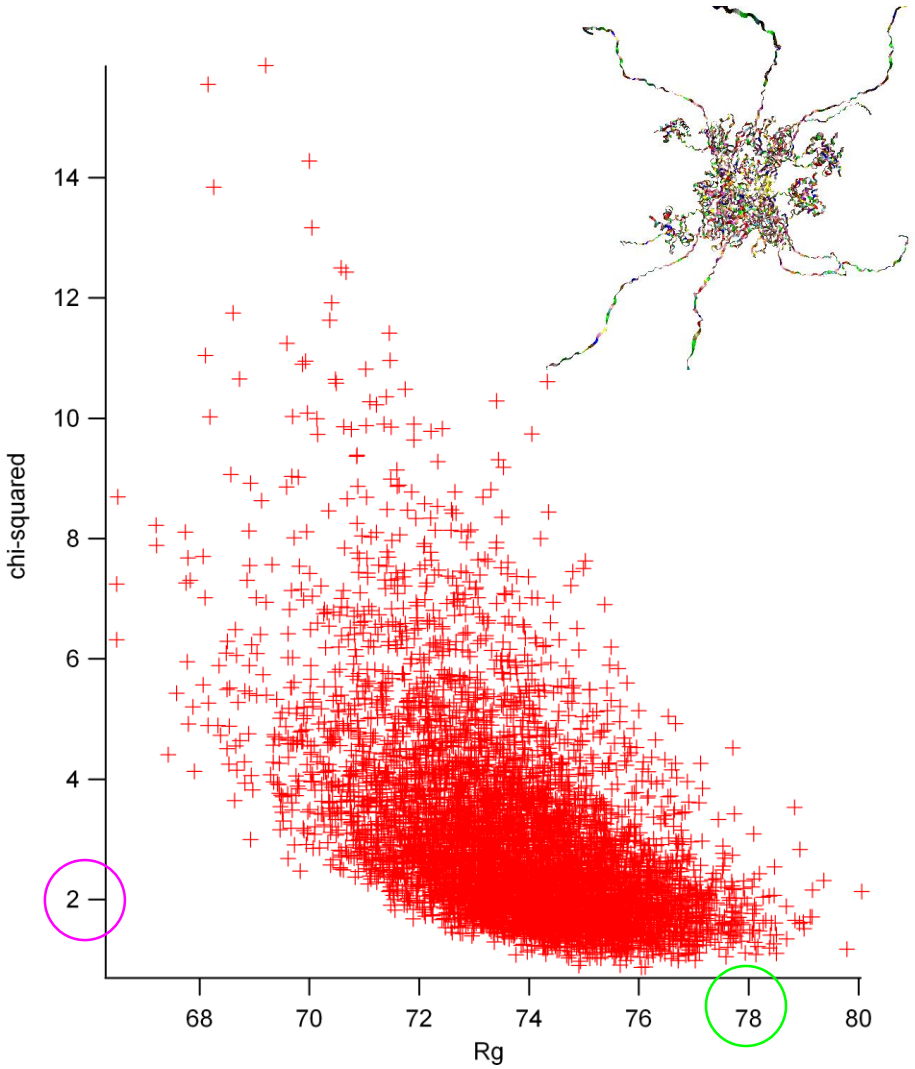


Possible flexible regions:
88-107 and 243-274

Chi-squared Filtering

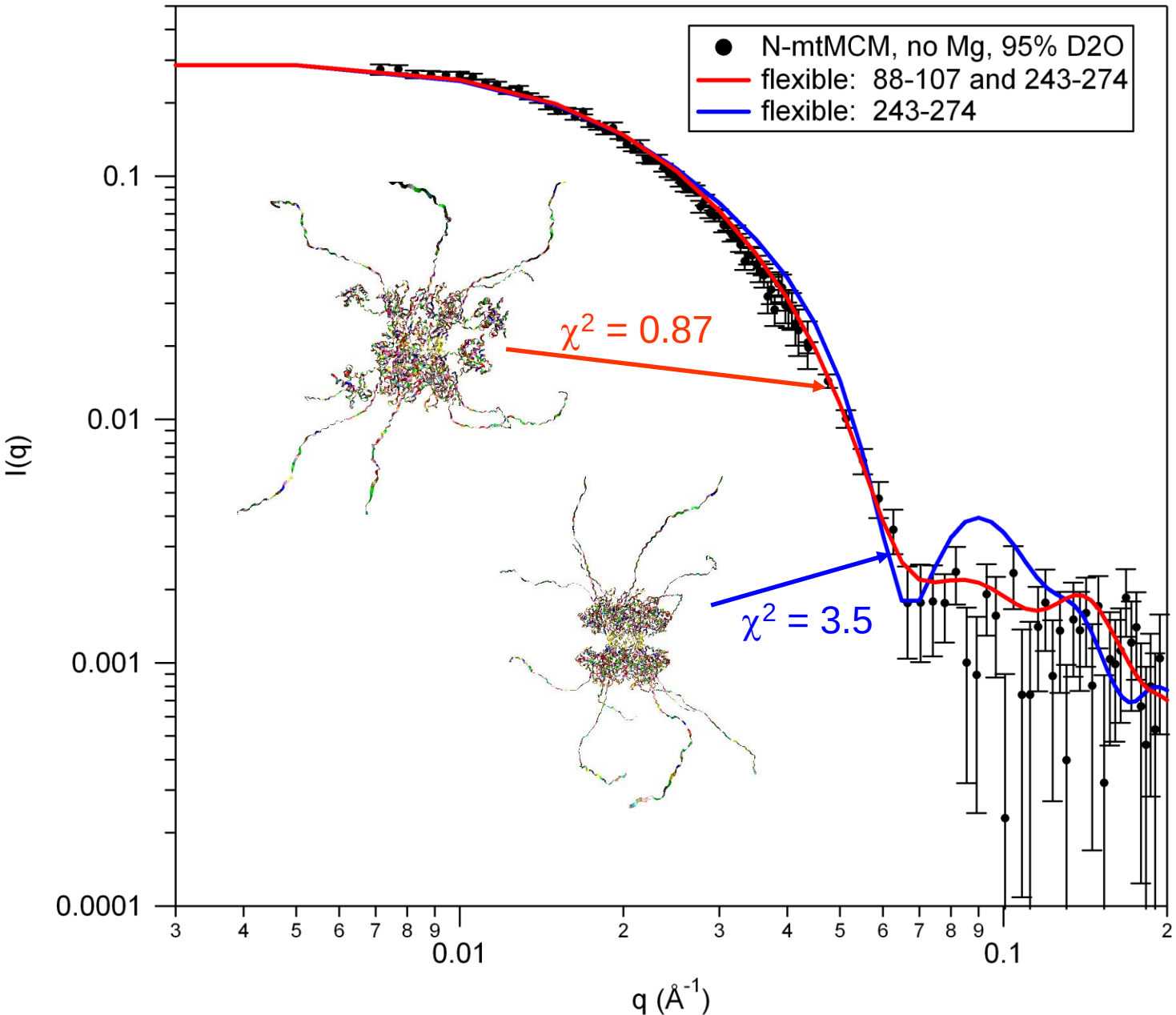


Flexible: 243-274



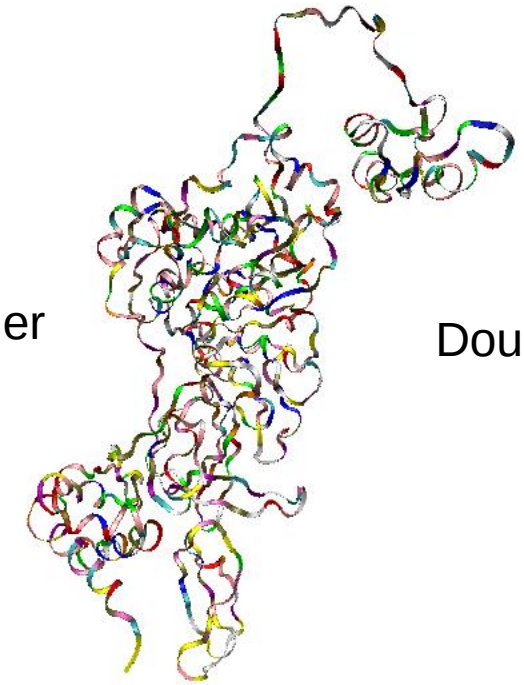
Flexible: 88-107 and 243-274

Best Fits

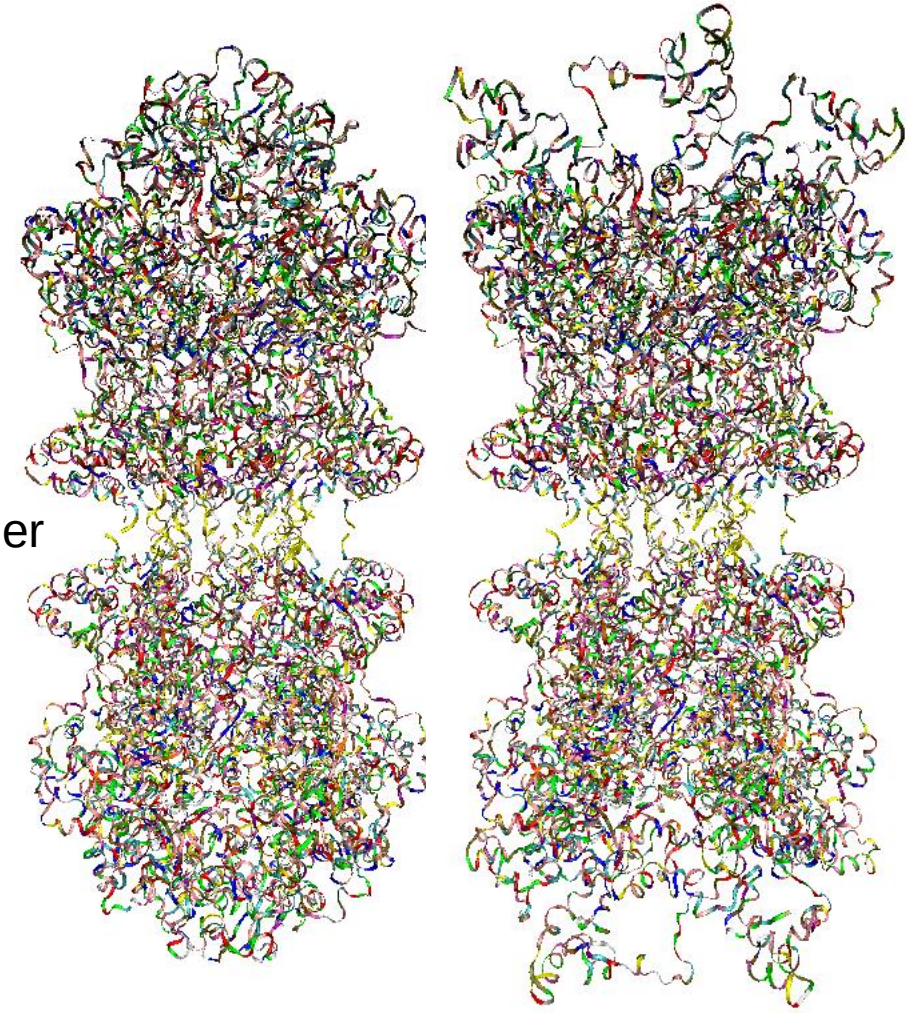

N-mtMCM, no Mg, 95% D2O
flexible: 88-107 and 243-274
flexible: 243-274
χ² = 0.87
χ² = 3.5
I(q)
0.1
0.01
0.001
0.0001
3 4 5 6 7 8 9
0.01
2 3 4 5 6 7 8 9
0.1
2
q (Å⁻¹)

Full-length MCM

Monomer

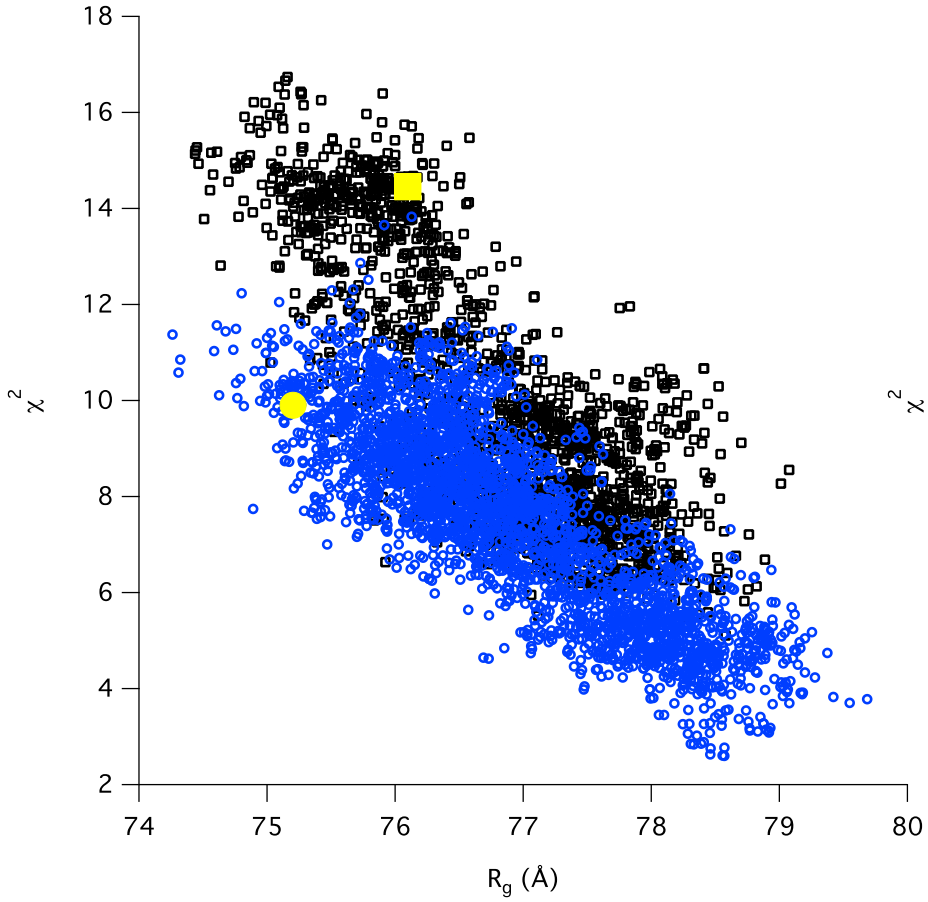


Double Hexamer

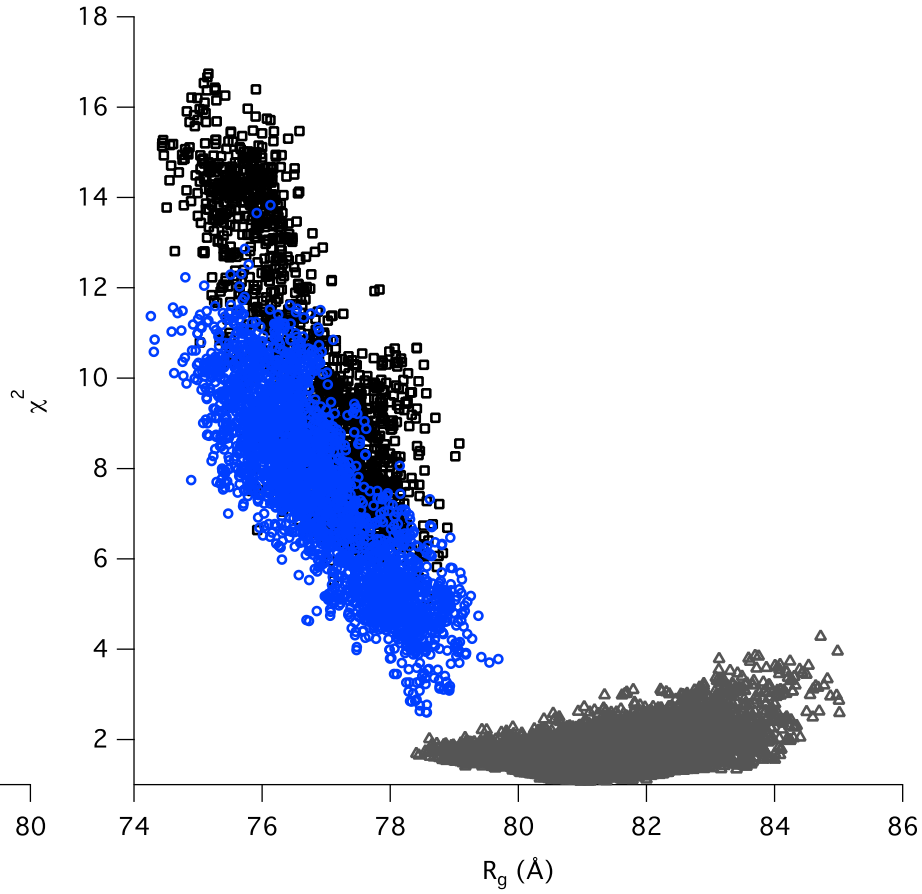


Possible flexible regions:
89-108, 244-246 and 580-600

Chi-squared Filtering

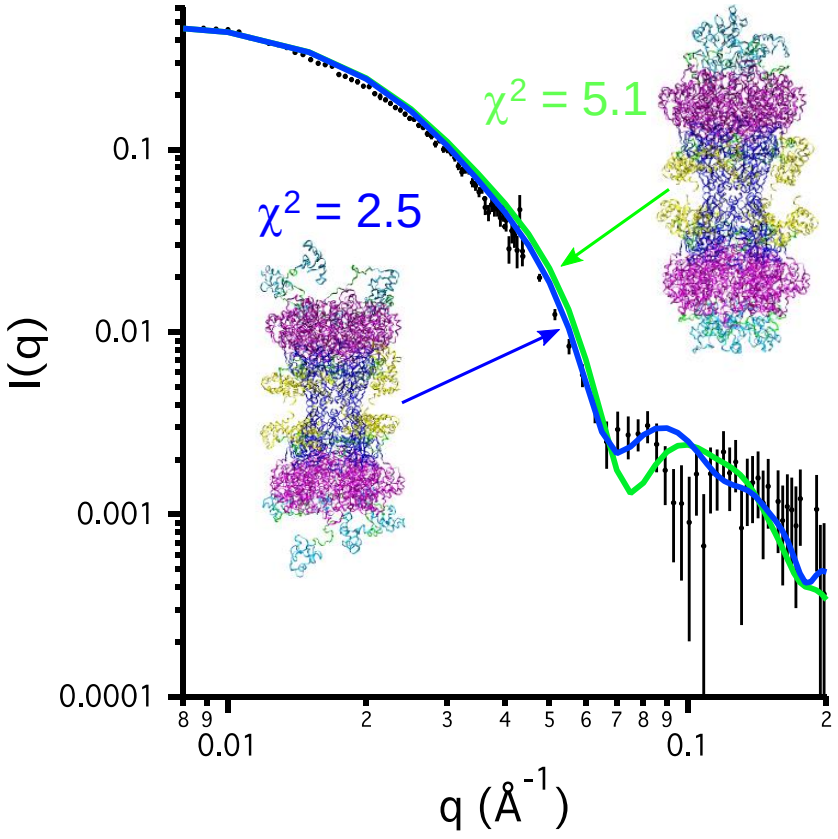


Flexible: 89-108 and 580-600

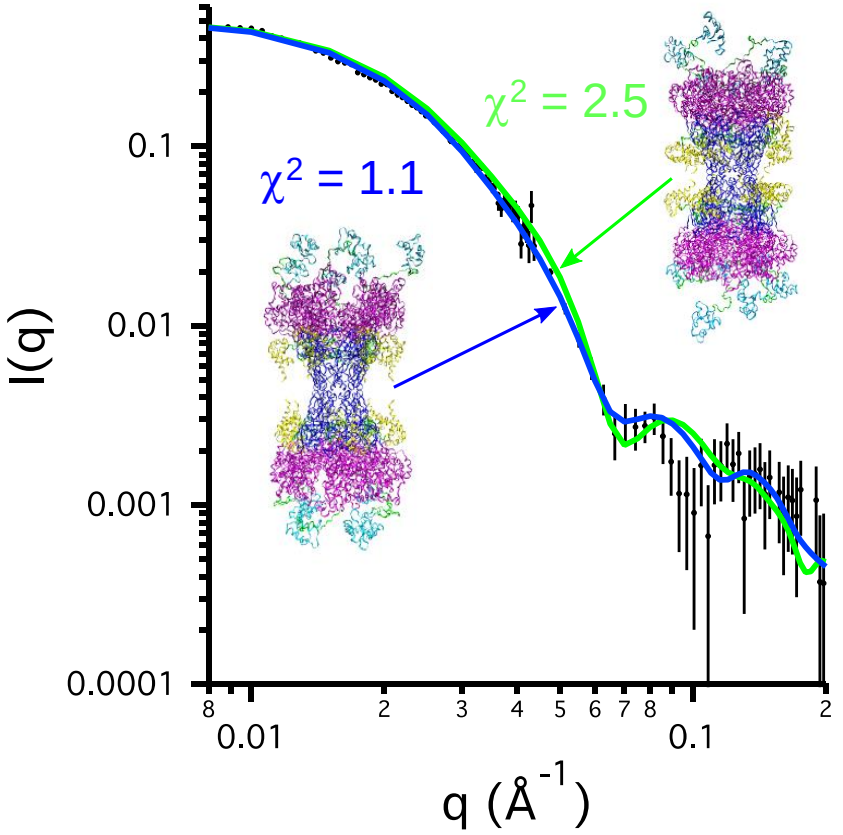


Flexible: 89-108, 244-246 and 580-600

Best Fits

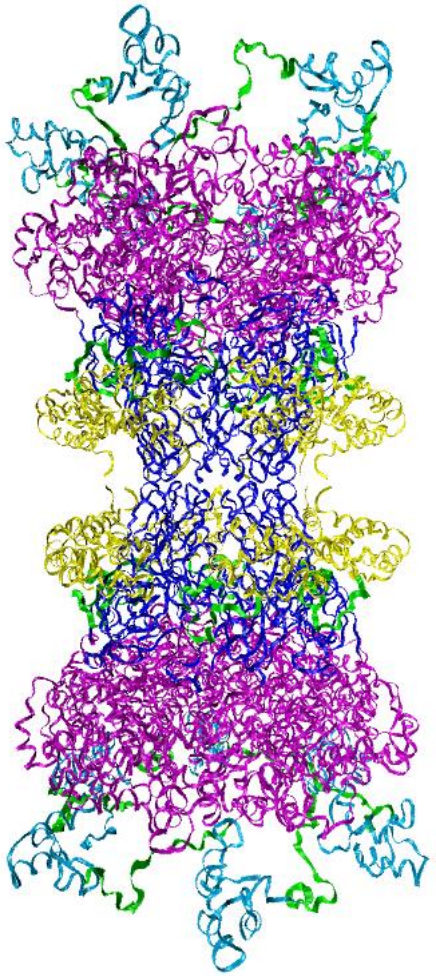


Flexible: 89-108 and 580-600

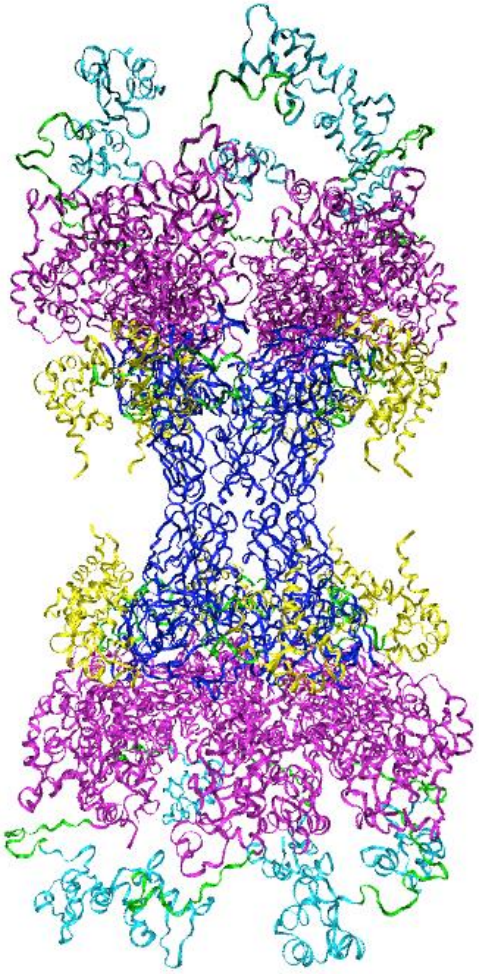


Flexible: 89-108, 244-246 and 580-600

Structure Comparison

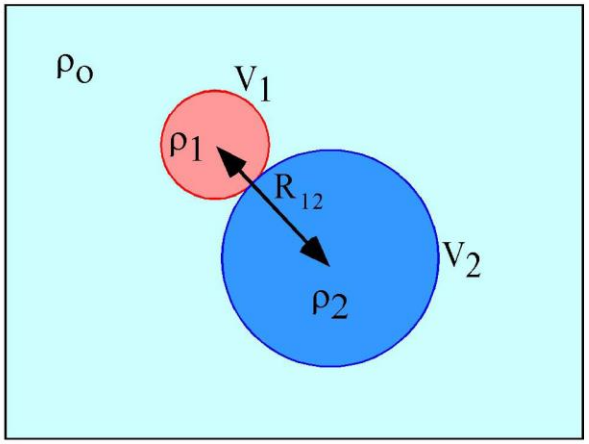


Original Starting Structure



Best Fit Structure Allowing
Three Flexible Regions

Compound Particles

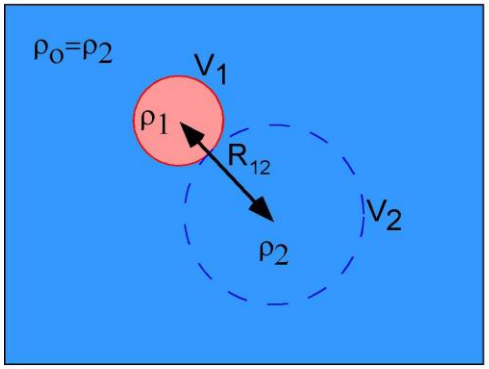


Model as an assembly of uniform particle subunits.

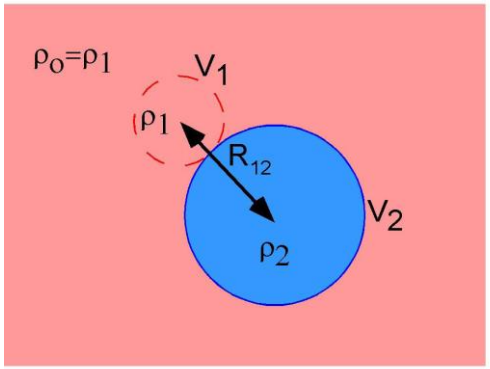
$$I(q) \propto \left\langle \left| (\Delta\rho)_1 \int_{V_1} e^{i\vec{q}\cdot\vec{r}}\, d\,\vec{r}_1 + (\Delta\rho)_2 \int_{V_2} e^{i\vec{q}\cdot\vec{r}}\, d\,\vec{r}_2 \right|^2 \right\rangle =$$

$$(\Delta\rho)_1^2 \left\langle \left|F_1(q)\right|^2 \right\rangle + (\Delta\rho)_2^2 \left\langle \left|F_2(q)\right|^2 \right\rangle + (\Delta\rho)_1(\Delta\rho)_2 \left|F_1\right|\left|F_2\right| \frac{\sin(qr_{12})}{qr_{12}}$$

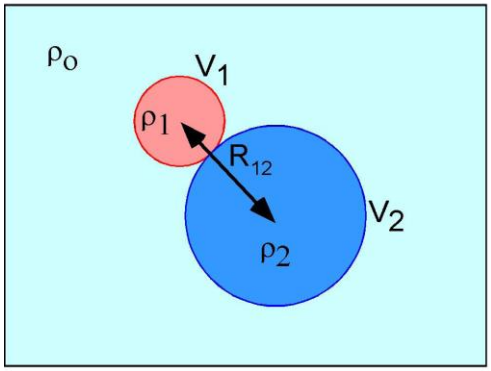
Compound Particles



$$I_1(q) = (\Delta\rho)_1^2 F_1^2$$



$$I_2(q) = (\Delta\rho)_2^2 F_2^2$$



$$I_{12}(q) = 2(\Delta\rho)_1(\Delta\rho)_2 F_1 F_2 \frac{\sin(qr_{12})}{qr_{12}}$$

Separate scattering from subunits using **contrast variation**.

Scattered Intensity
Two-Component System

Scattered intensity from the two components can be separated by solving a set of simultaneous equations.

$$I(q) = \Delta\rho_1^2 I_1(q) + \Delta\rho_1 \Delta\rho_2 I_{12}(q) + \Delta\rho_2^2 I_2(q)$$

$\Delta\rho_1$, $\Delta\rho_2$: contrast for components 1 and 2

recall: $\Delta r = r - r_s$

$I_1(q)$, $I_2(q)$: intensity for components 1 and 2

$I_{12}(q)$: cross-term between components 1 and 2

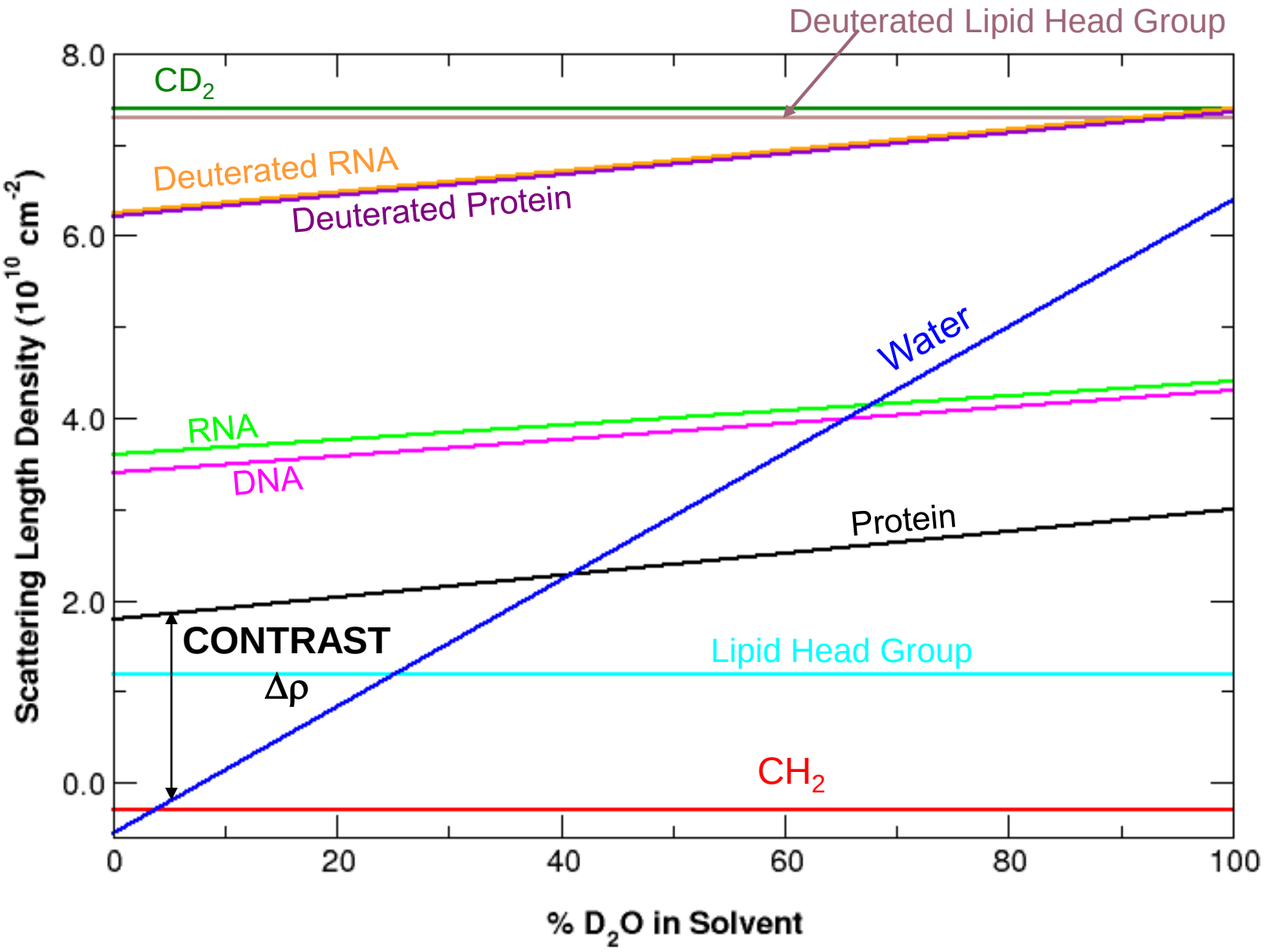
Neutron Contrast Variation in Biology

Used to separate the scattering contribution from the components in a two-component complex.

- Protein-nucleic acid complexes
- Protein-protein complexes*
- Protein-lipid complexes

*One component must be deuterated.

Contrast Variation



Contrast Variation Experiment

- **Model data prior to experiment**
  - Predict contrasts values and match points based on chemical composition (amino acid sequence, bases, type of lipid, etc.)
  - Predict SANS curves at various contrasts if a low or high resolution starting model structure is available.

- **Match point determination**
  - Determine the contrast match point for the complex and individual components.

- **Stuhrmann Analysis and Parallel Axis Theorem**
  - Determine Rg for the individual components and the spatial relationship between the two components.

- **Test Model Structures Against ALL Data**

Match Point Determination

For $q = 0$:

Concentration (g/cm^3)

Molecular Weight (Da)

(cm^{-1})

$$I(0) = n(\Delta\rho V)^2 = \frac{cM_w}{N_A}(\Delta\rho\bar{v})^2$$

Partial Specific Volume (cm^3/g)

Since $\Delta\rho$ varies with the %D_2O in the solvent:

$$\frac{I(0)}{c} = \frac{M_w}{N_A}(\Delta\rho\bar{v})^2 \Rightarrow \frac{I(0)}{c} \propto (\Delta\rho)^2 \Rightarrow \sqrt{\frac{I(0)}{c}} \propto \%D_2O$$

Plot $\sqrt{\frac{I(0)}{c}}$ **vs** %D_2O **to obtain the match point.**

Two Component System: R_g vs $\Delta\rho$

Stuhrmann Analysis

$$R_g^2 = R_o^2 + \frac{a}{\Delta r} + \frac{b}{(\Delta r)^2}$$

$\beta \neq 0 \Rightarrow$ centers of mass of the two components are not concentric

R_g of an equivalent homogeneous complex

Ibel, K. and Stuhrmann, H. B. (1975). *J. Mol. Biol.* **93**, 255–265

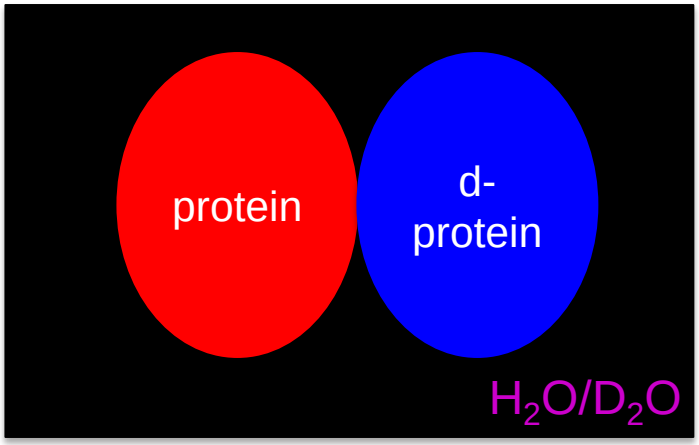
Parallel Axis Theorem

$$R_g^2 = \frac{\Delta r_1 V_1}{\Delta r V} R_1^2 + \frac{\Delta r_2 V_2}{\Delta r V} R_2^2 + \frac{\Delta r_1 V_1 \Delta r_2 V_2}{(\Delta r V)^2} D^2$$

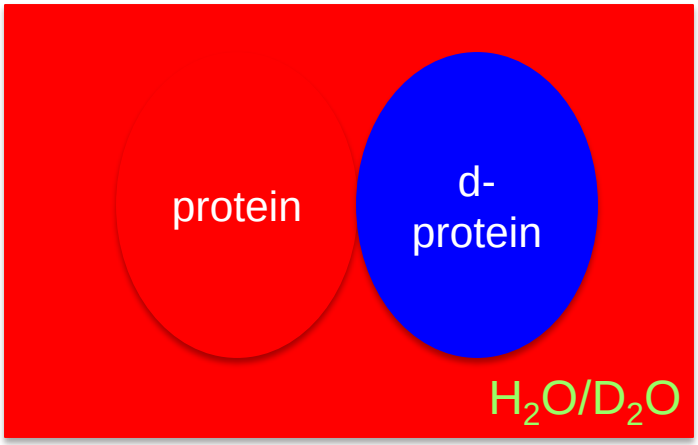
Component 1 | Component 2 | Cross-term | Distance between centers of mass

Moore, P. B. (1982). *Methods Exp. Phys.* **20**, 337–390

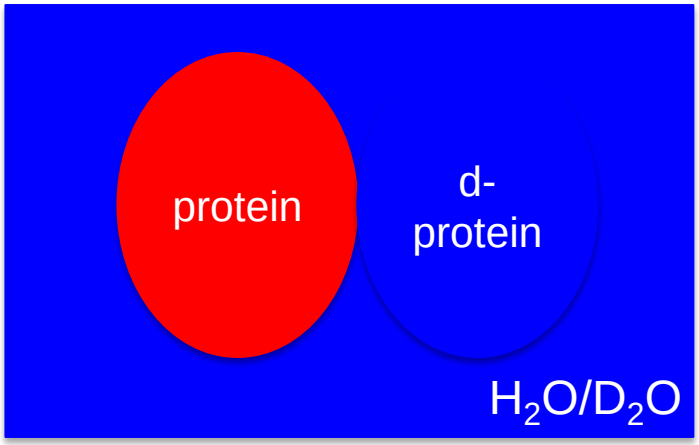
Contrast Variaton: Biology Example



- Two-component protein in H_2O/D_2O solvent.
- One subunit is deuterated (d-protein).
- Scattering length densities of the two components are different from one another.

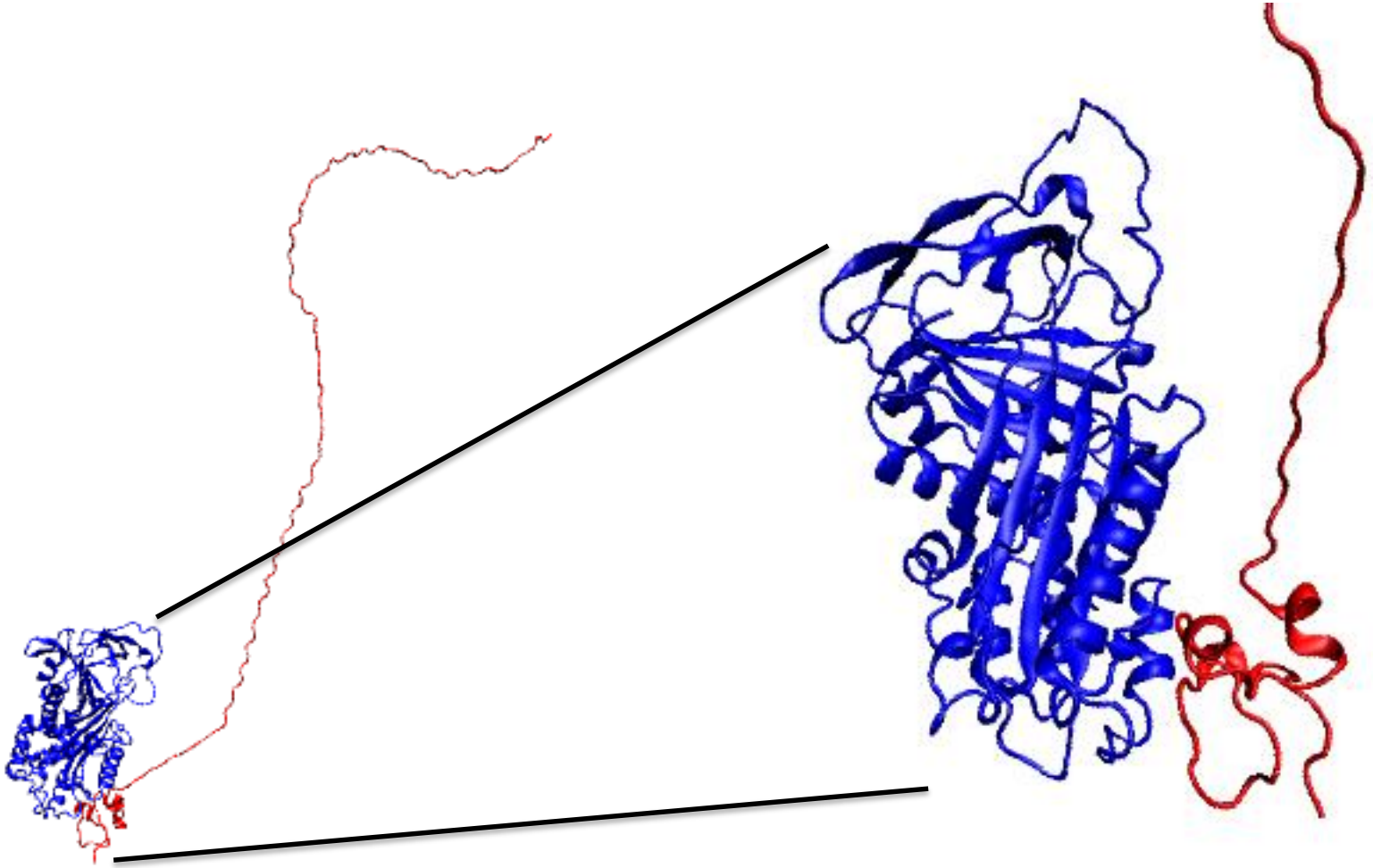


Protein subunit is “matched”.



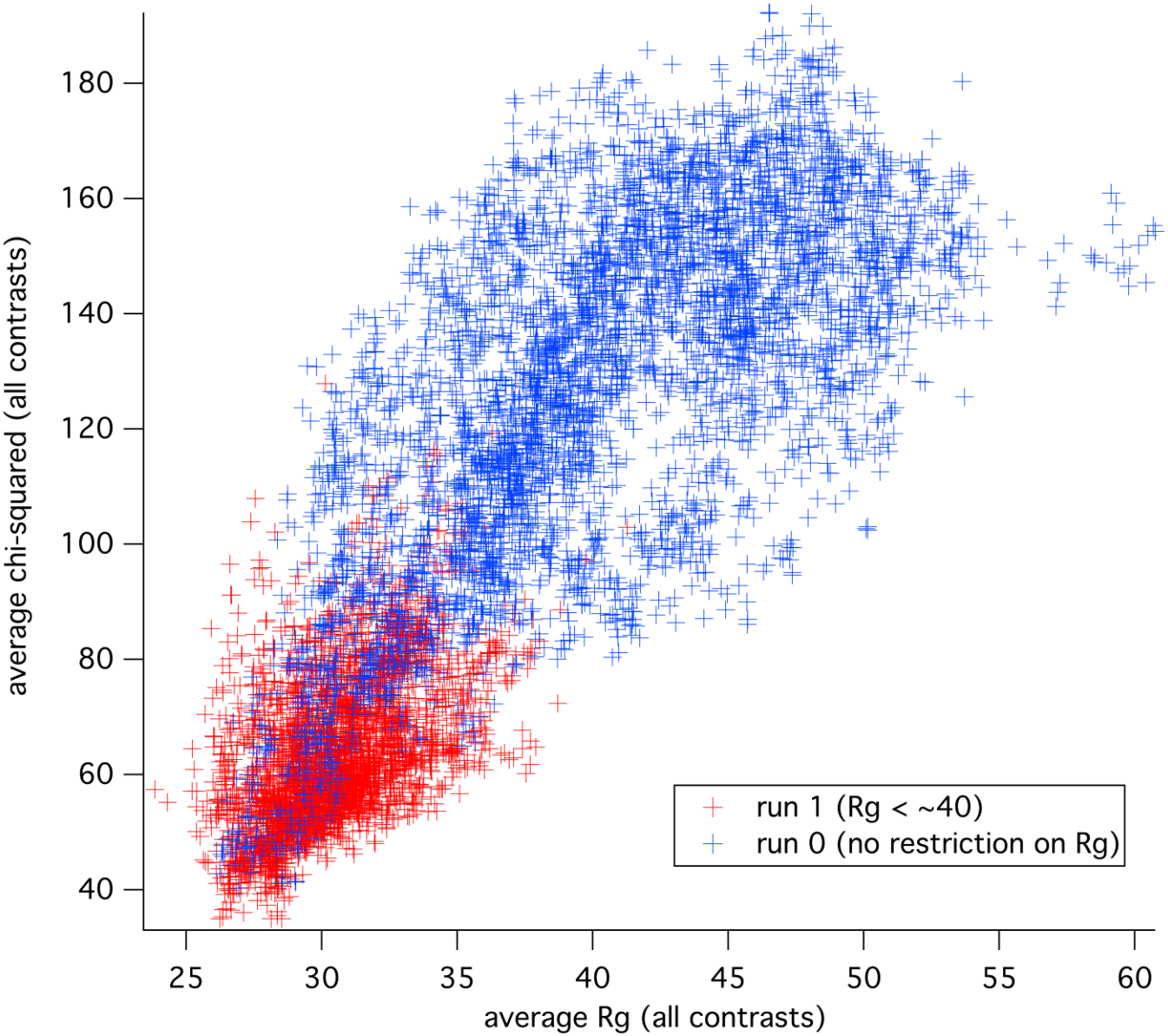
d-protein subunit is “matched”.

SASSIE and Contrast Variation

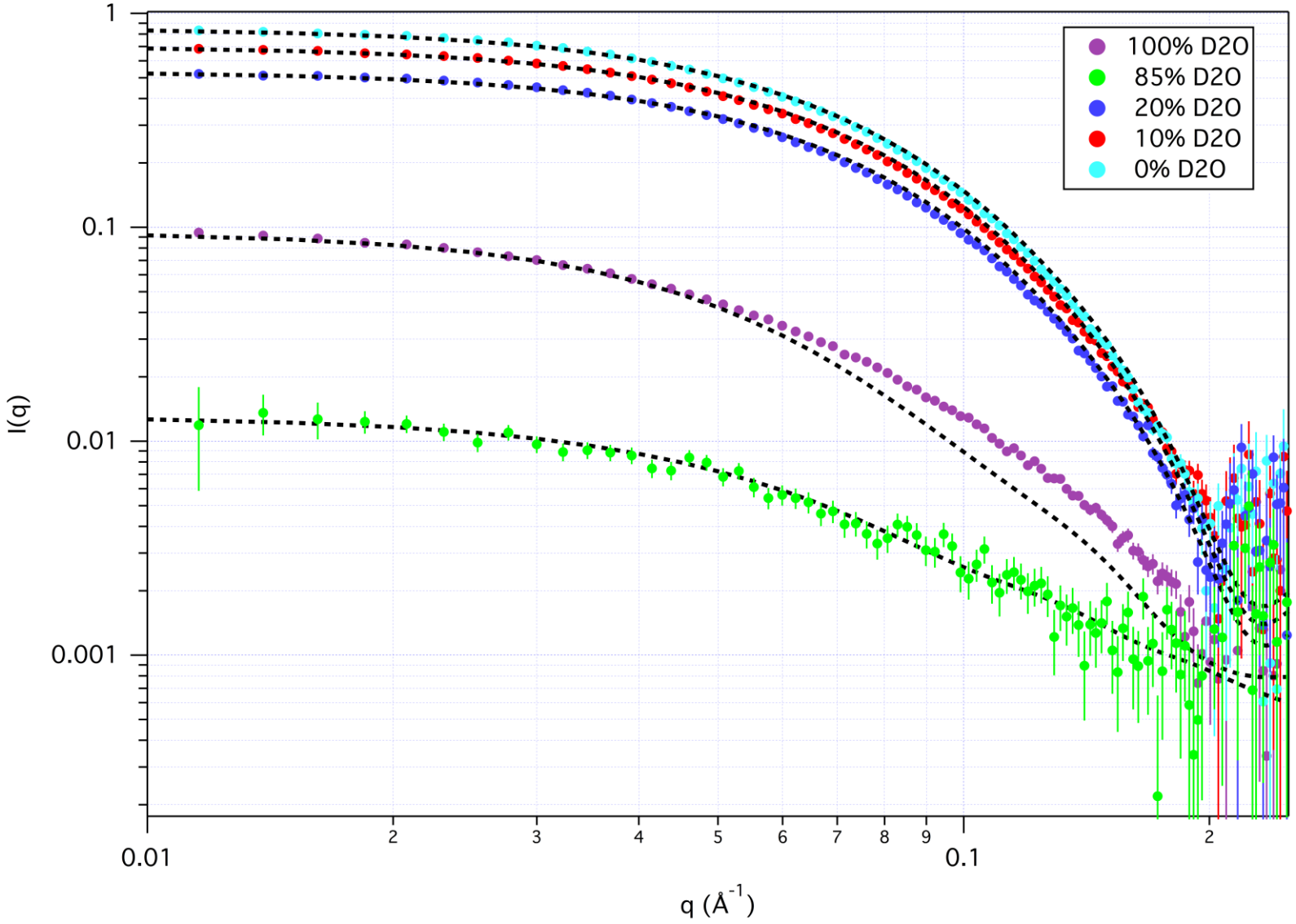


PAI:VN complex with 65% deuterated PAI and non-deuterated VN.

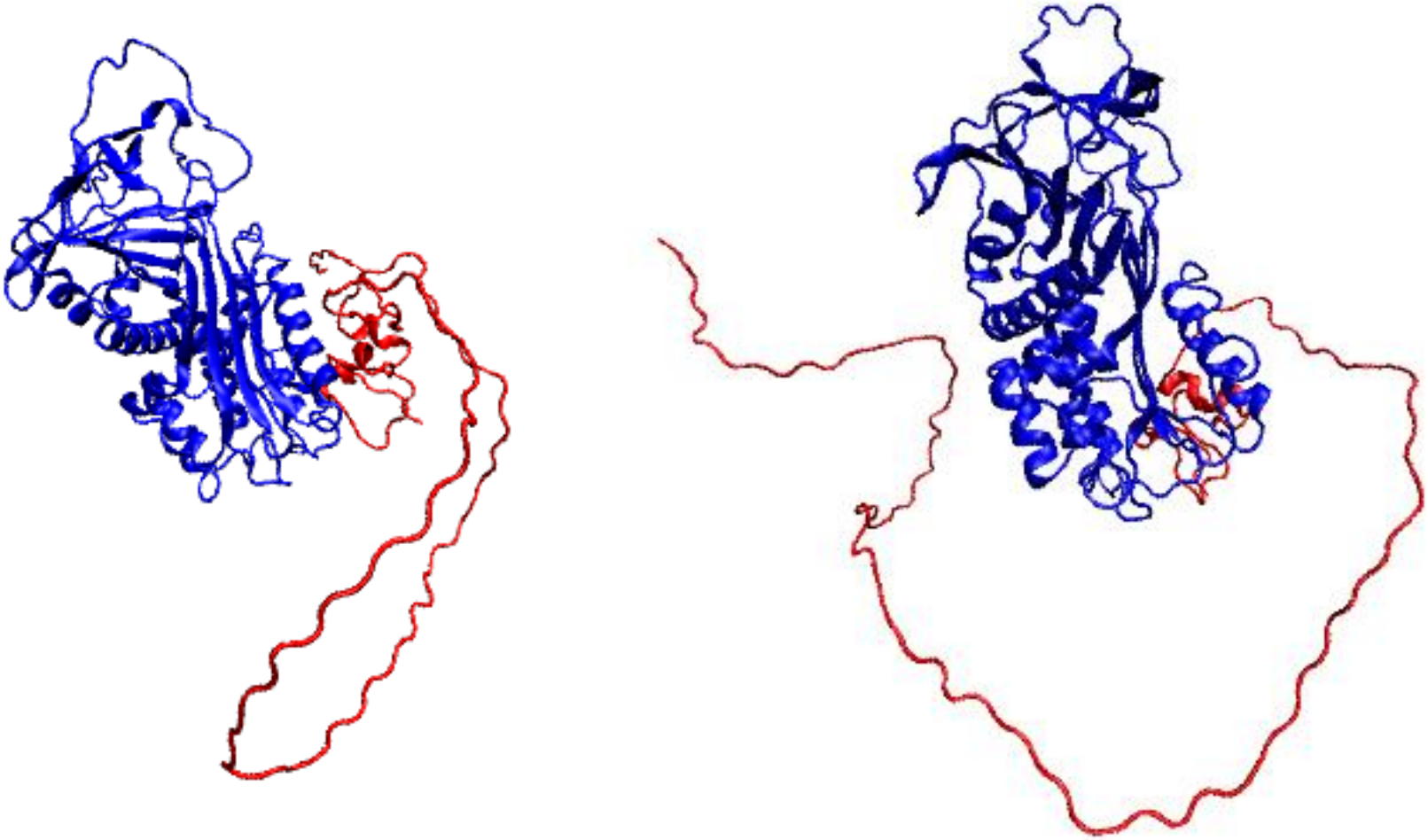
SASSIE and Contrast Variation



SASSIE and Contrast Variation

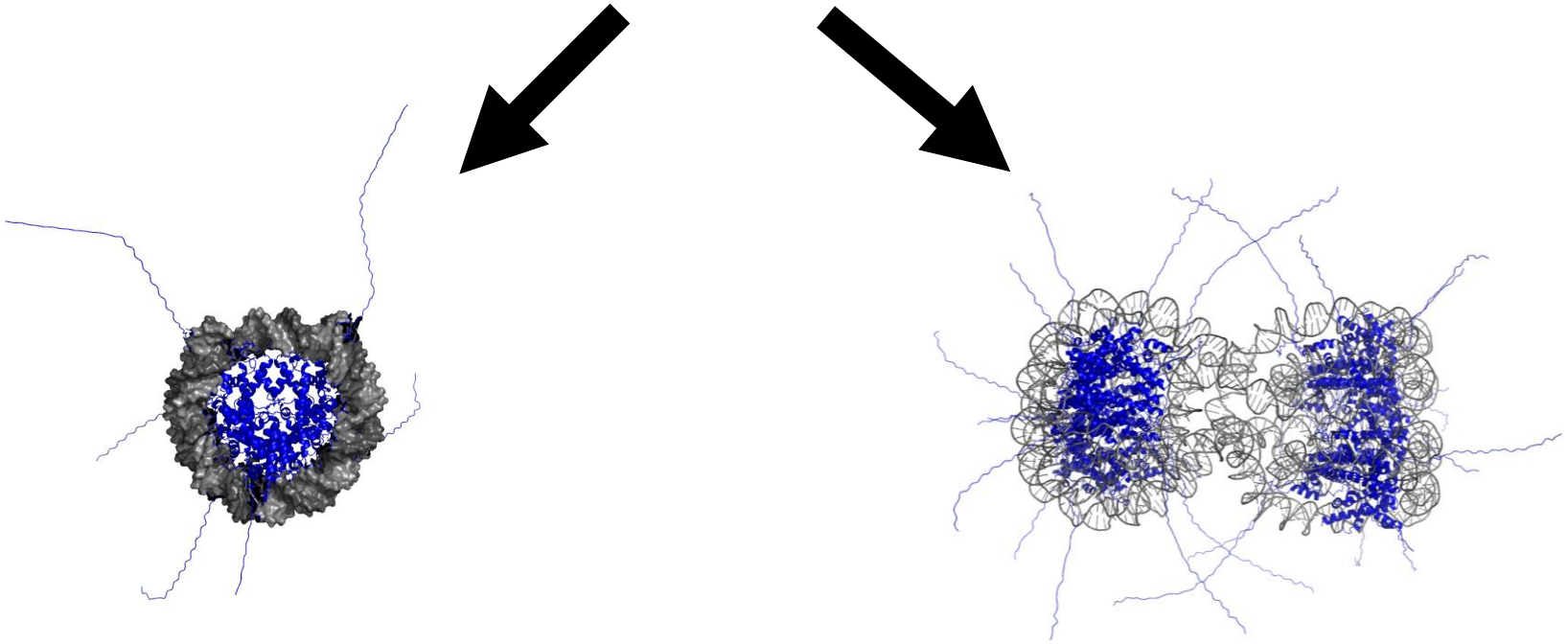


SASSIE and Contrast Variation



Two structures that fit all of the data (except 100% D_2O) well.

Chromatin Solution Structure



Modeling work underway…thanks to new double-stranded DNA Molecular Monte Carlo

Summary: SANS and Biology

- If used correctly, SANS can be a powerful tool to study the structure of biological macromolecules in solution.
- Contrast variation can be easily applied using neutrons.
- Can be combined with SAXS to obtain data at higher q values ($q > 0.3$ Å^{-1}).
- Excellent facilities are available worldwide. Many groups are actively developing software for structure modeling.
- Sample quality is extremely important. Consistency checks must be performed.
- Use information from other techniques to narrow down the possible model structures.