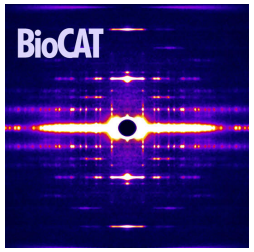
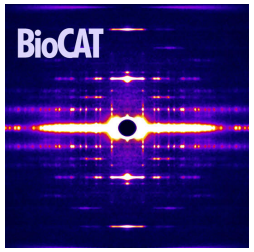


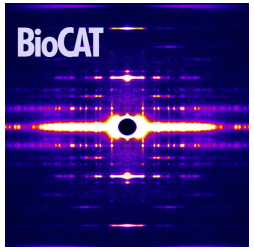
Basic data validation and analysis

Jesse Hopkins, PhD
IIT/CSRRI
Staff Scientist, BioCAT
Sector 18, Advanced Photon Source



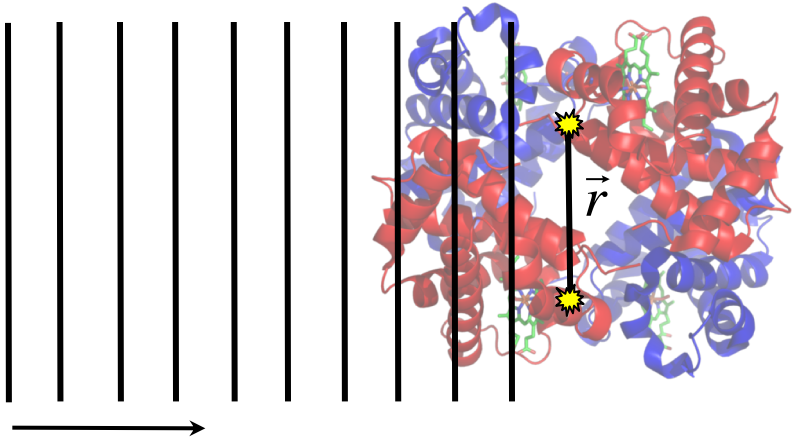
Overview

- The scattering profile
- What can go wrong with your data
- Guinier analysis
- Molecular weight analysis
- Porod and Kratky analysis
- Indirect Fourier Transforms
- Summary



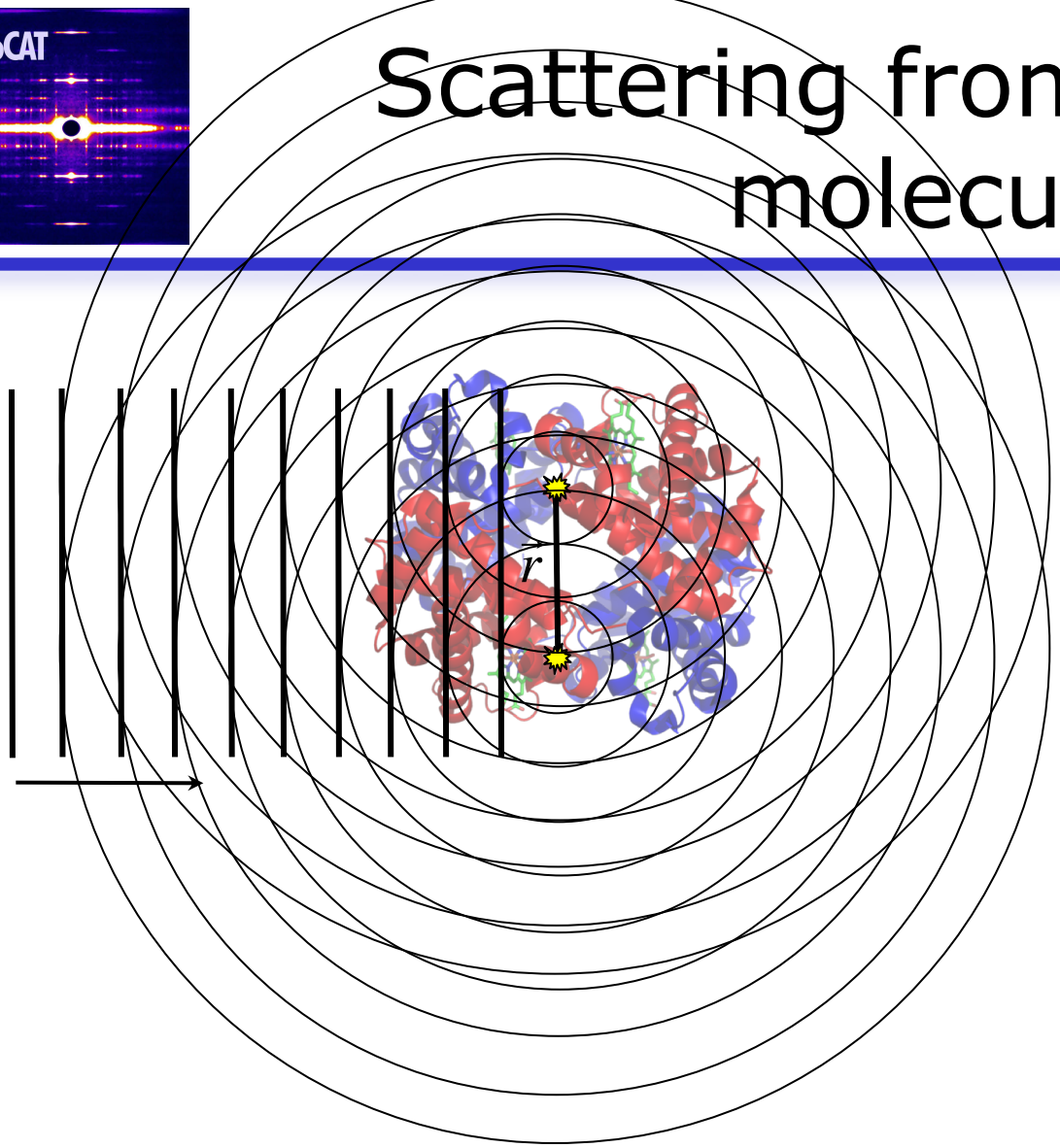
The scattering profile

Scattering from a single molecule

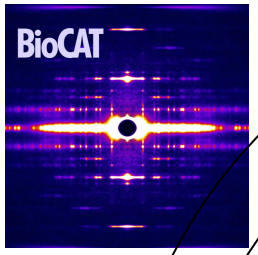


Images from Richard Gillilan's BioSAXS Essentials presentation

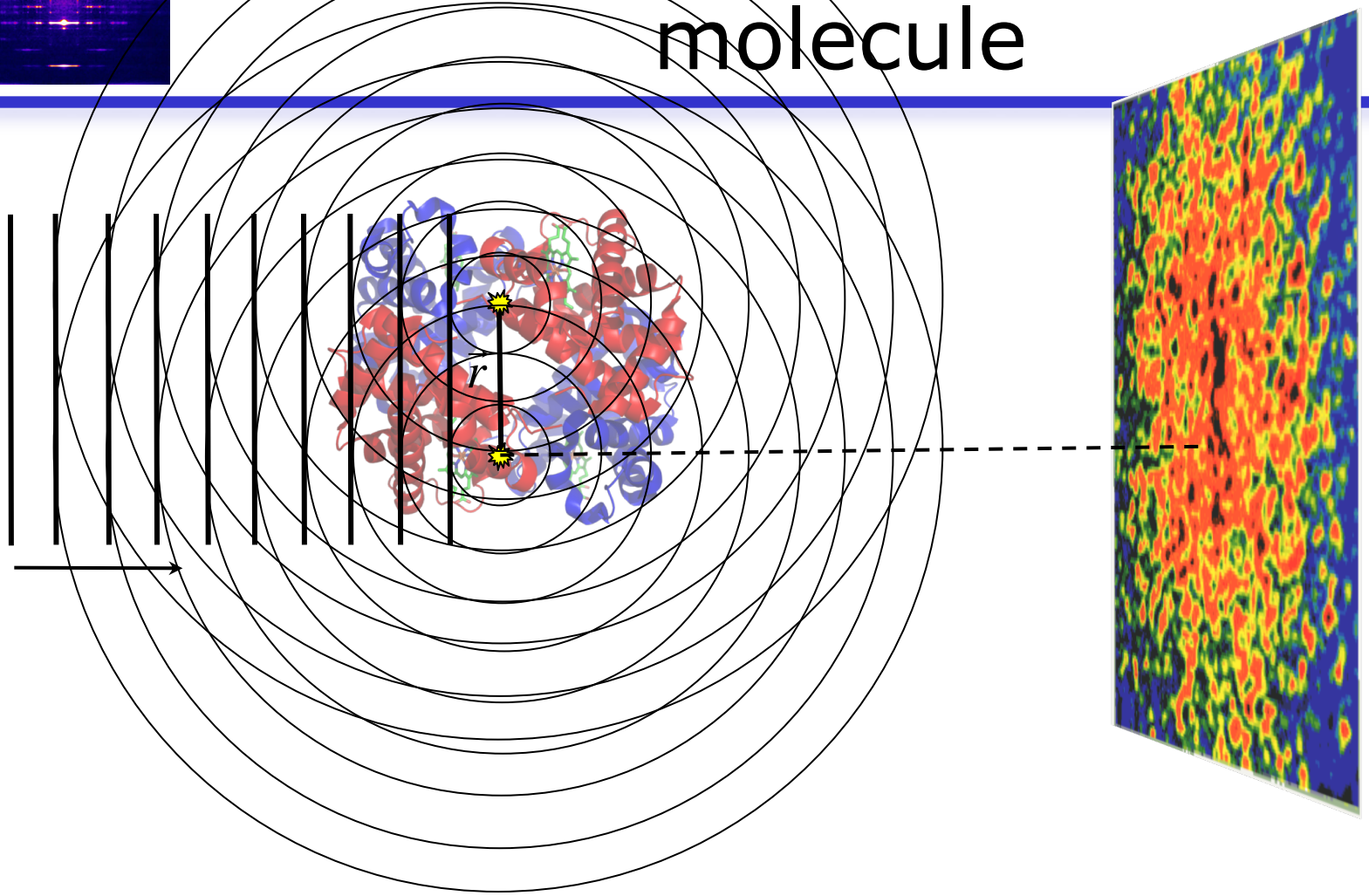
Scattering from a single molecule



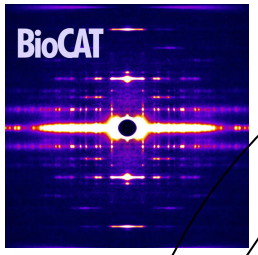
Images from Richard Gillilan's BioSAXS Essentials presentation



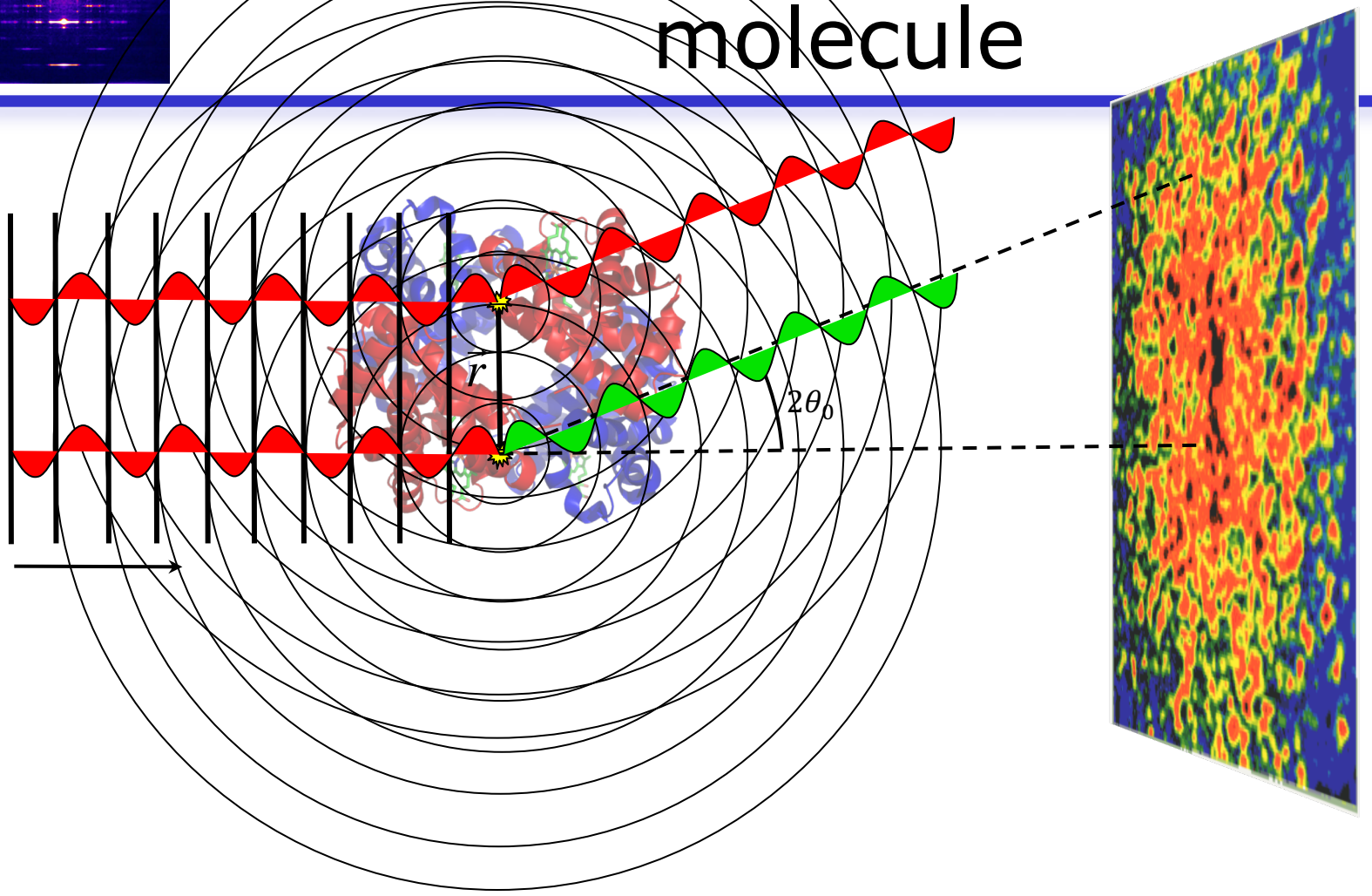
Scattering from a single molecule



Images from Richard Gillilan's BioSAXS Essentials presentation

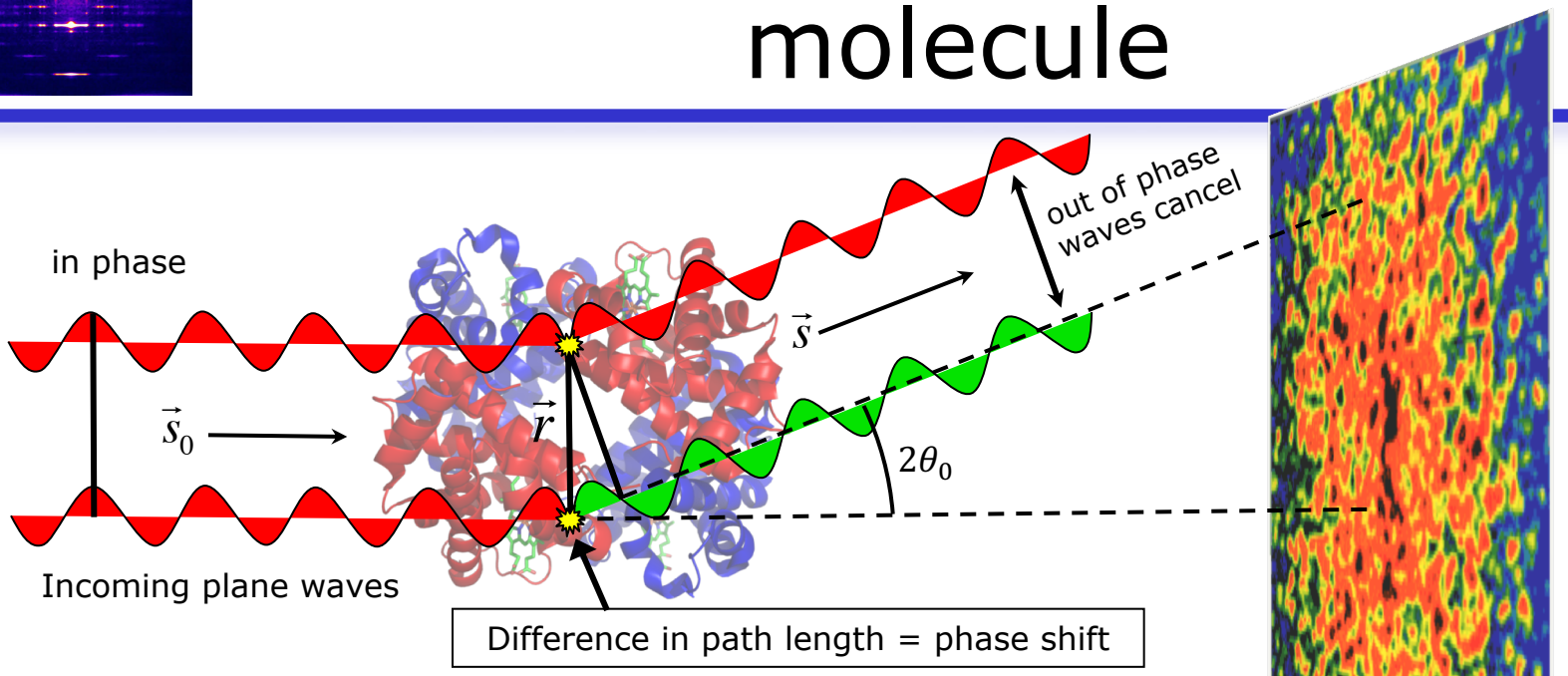


Scattering from a single molecule



Images from Richard Gillilan's BioSAXS Essentials presentation

Scattering from a single molecule

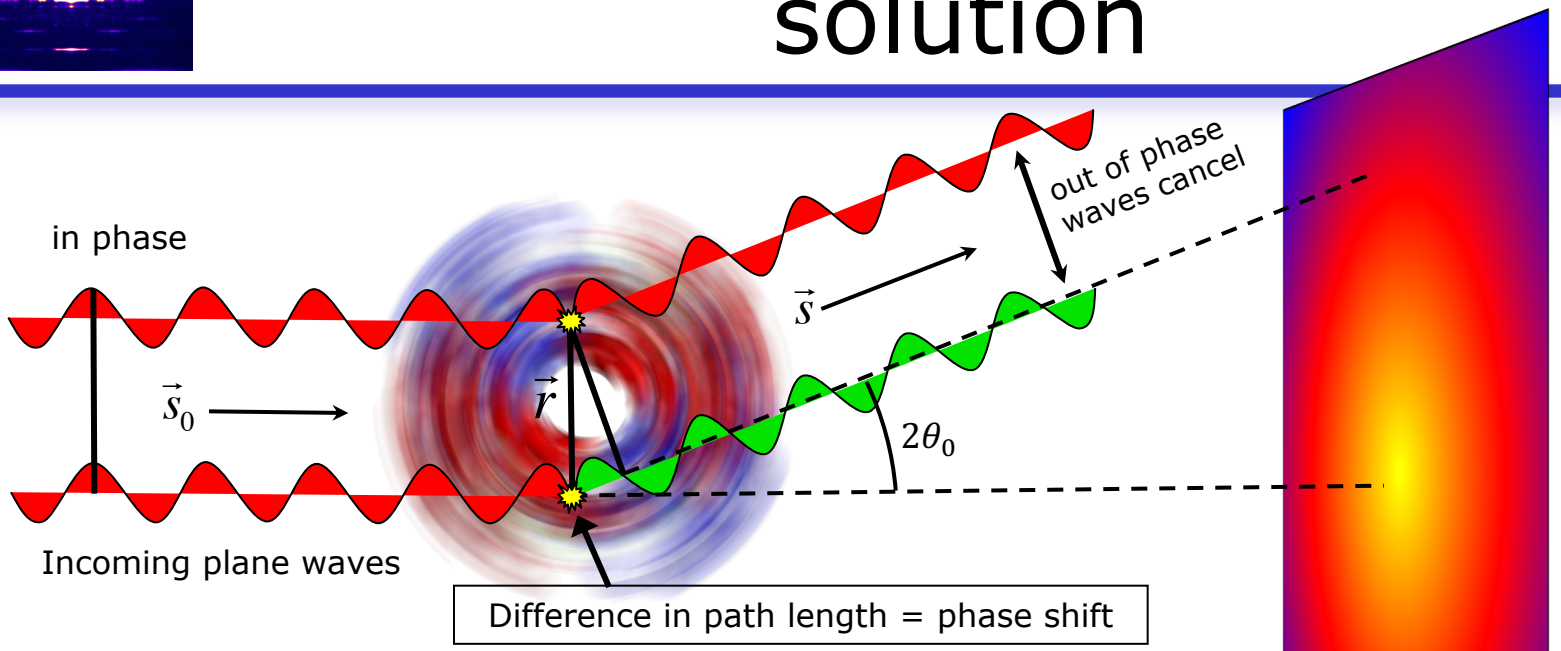


Far-field diffraction pattern

Waves scattered from different parts of the molecule result in phase shifts – a speckled intensity pattern on detector

Images from Richard Gillilan's BioSAXS Essentials presentation

Scattering from molecules in solution

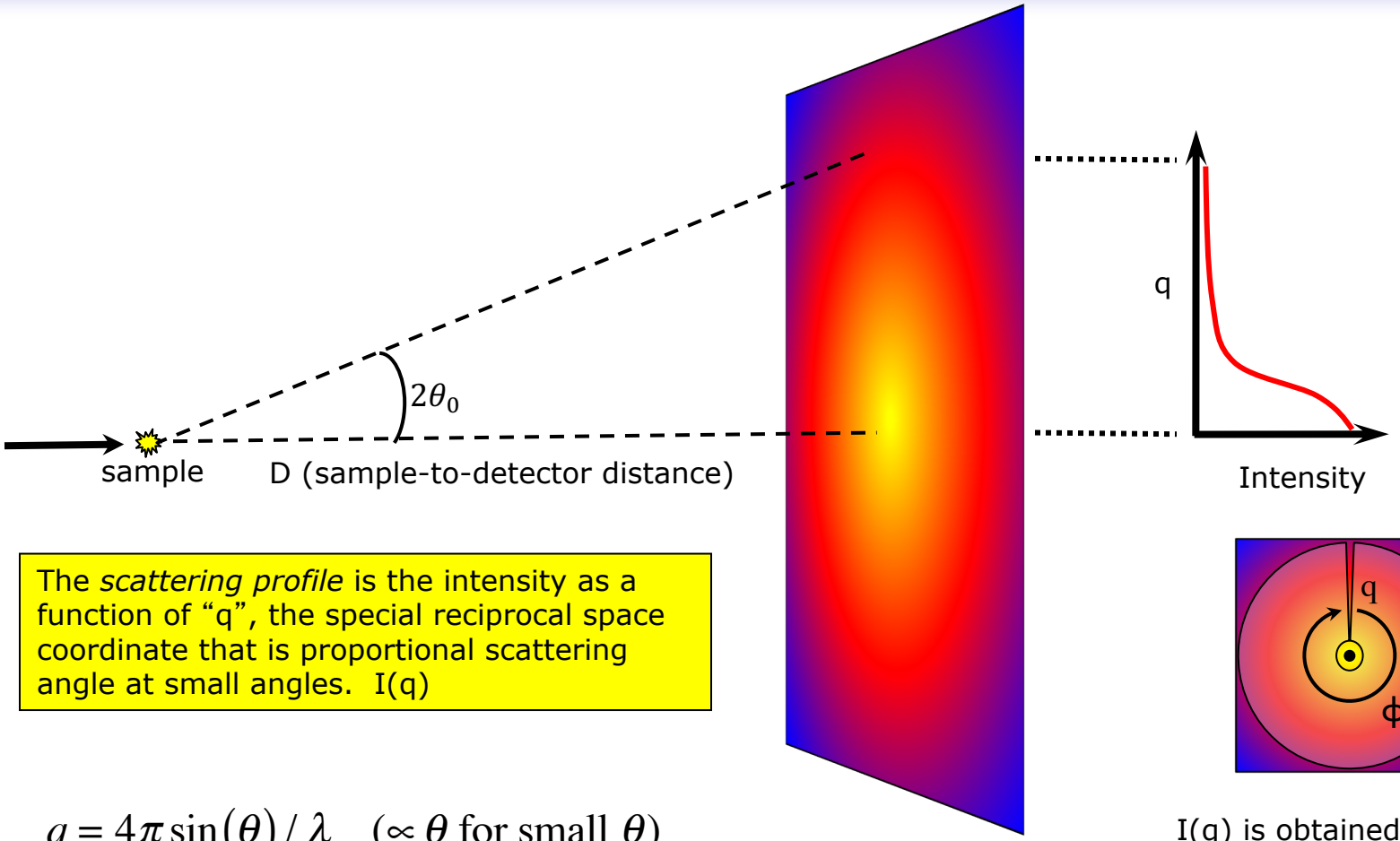


Since molecules are far apart in dilute solutions, the idealized scattering pattern of the solution is the same as the scattering pattern for a single rotationally-averaged molecule.

Far-field diffraction pattern

Images from Richard Gillilan's BioSAXS Essentials presentation

The scattering profile

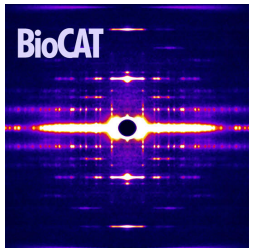


The *scattering profile* is the intensity as a function of "q", the special reciprocal space coordinate that is proportional scattering angle at small angles. I(q)

$$q = 4\pi \sin(\theta) / \lambda \quad (\propto \theta \text{ for small } \theta)$$

I(q) is obtained by integrating around the circle. For detectors, the standard deviation of signal σ(q) is also calculated.

Slide from Richard Gillilan's BioSAXS Essentials presentation



The scattering profile

$$I(q) \propto Mc(\rho_1 - \rho_2)^2 |F(q)|^2 S(q)$$

$I(q)$ – Experimental intensity

M – molecular weight

c – concentration

ρ – scattering density (electrons per unit volume)

ρ_1 - particle

ρ_2 - solvent

$F(q)$ – Form factor, i.e. molecular shape

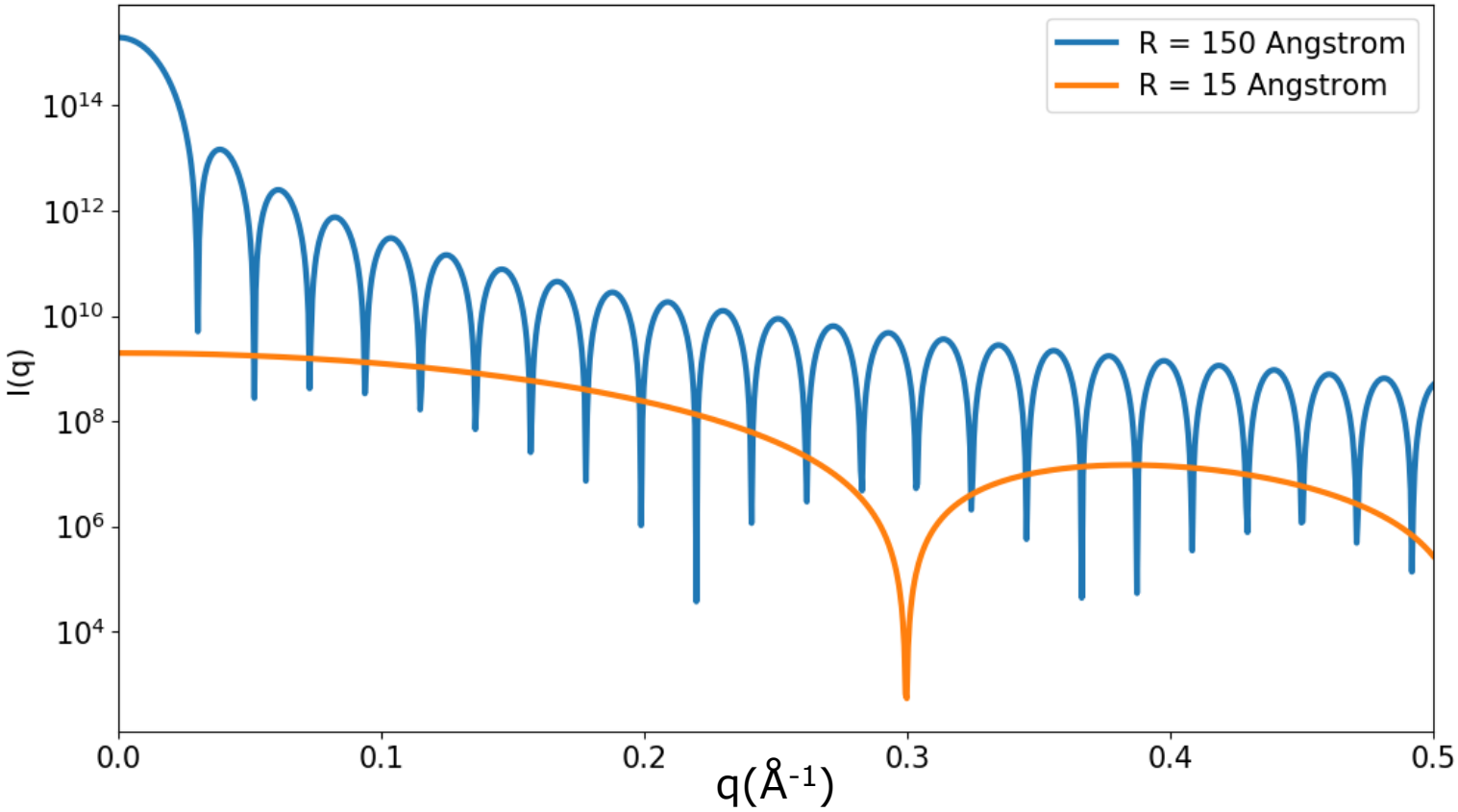
$S(q)$ – Structure factor, i.e. inter-molecular interaction

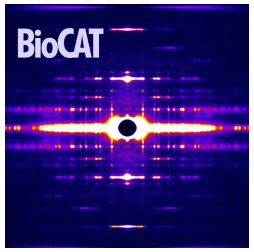
≈ 1 for dilute solutions

Scattering from a sphere

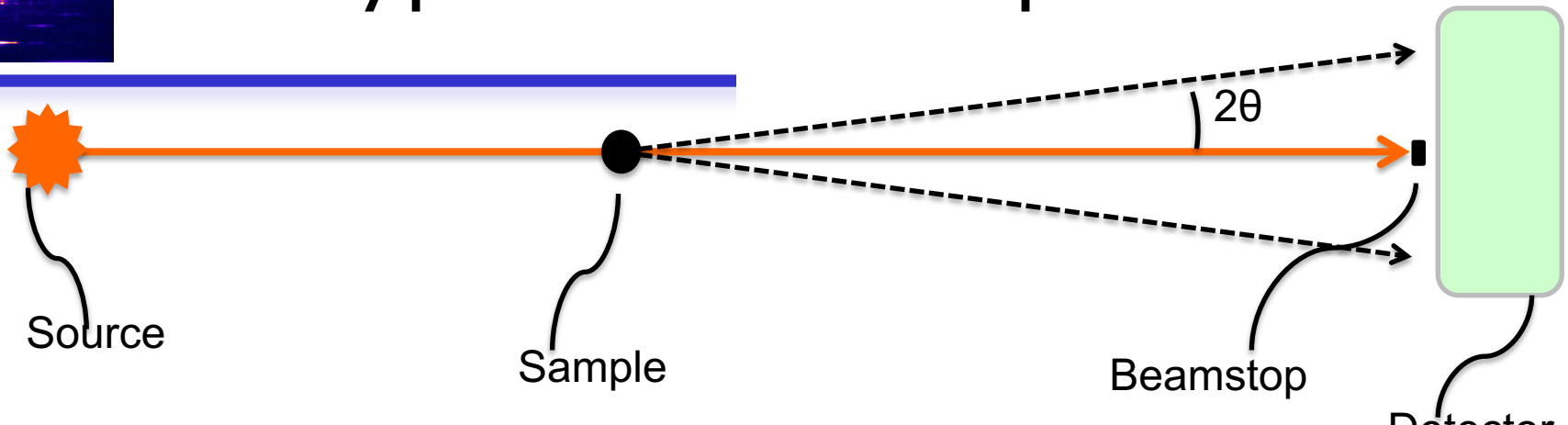
Scattering from a uniform density sphere with radius R:

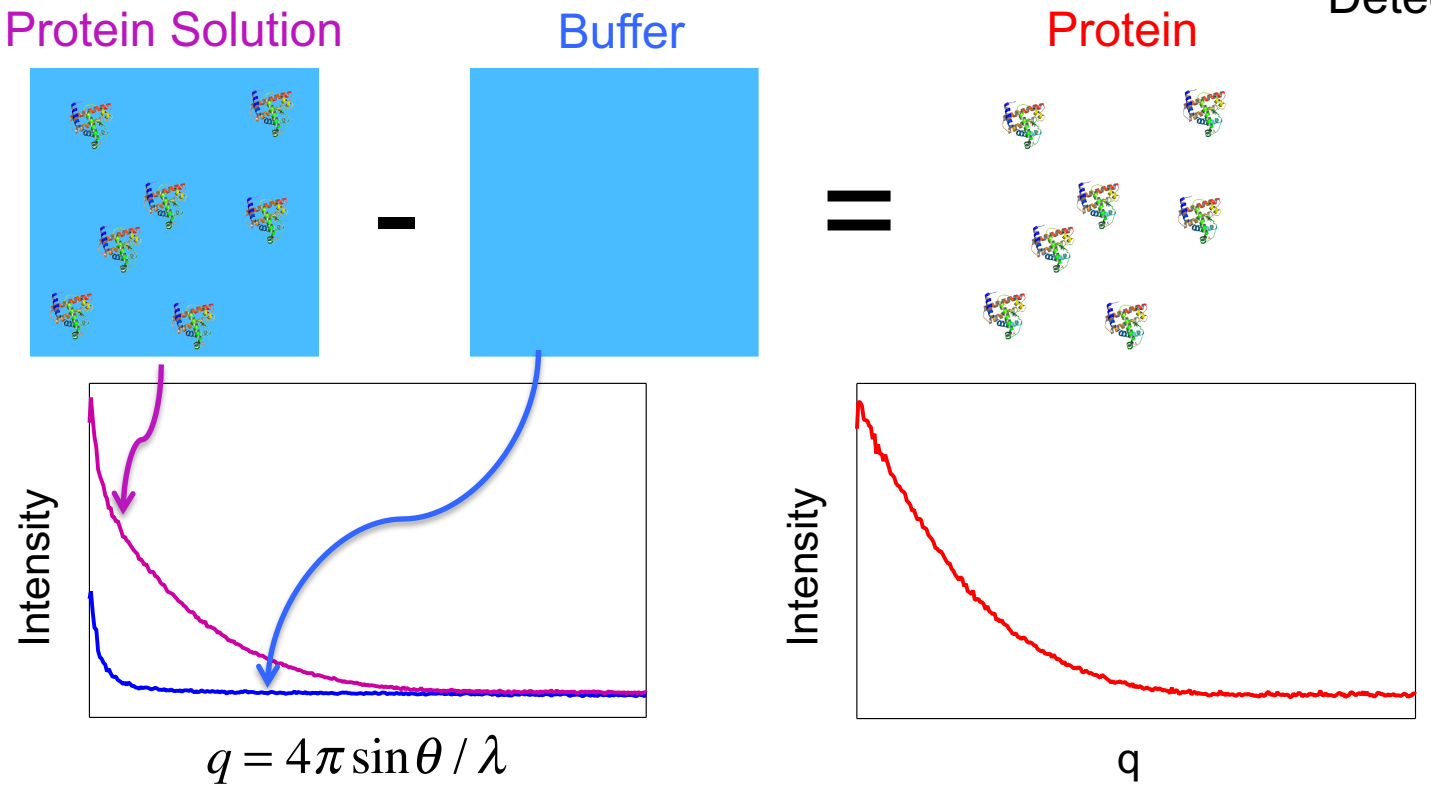
$$I(q) \propto \left(\frac{4\pi}{3}R^3\right)^2 \left(3\frac{\sin(qR) - qR\cos(qR)}{(qR)^3}\right)^2$$

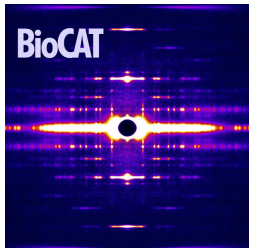




A typical SAXS experiment

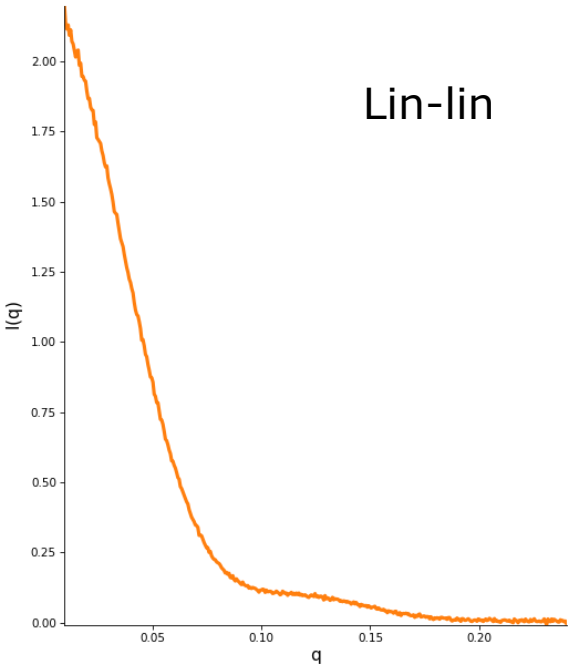


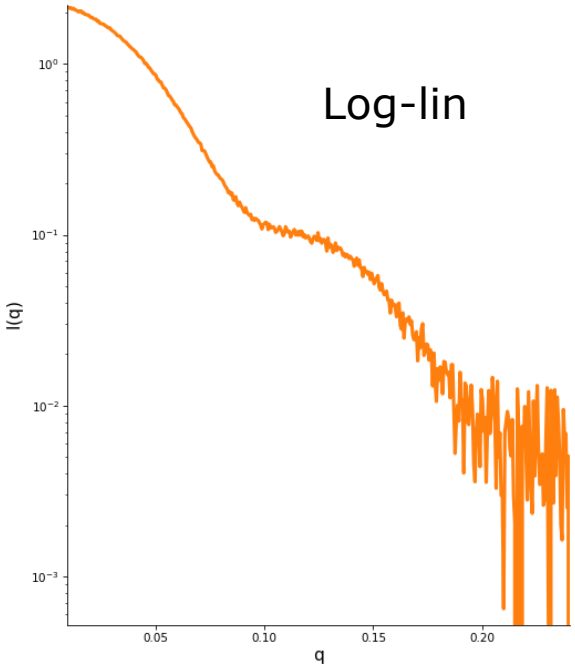


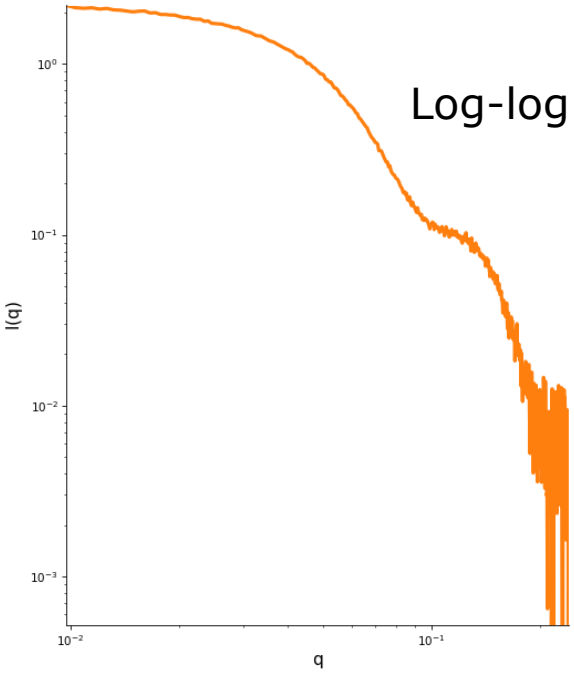


Plotting the scattering profile

Same profile, three different plots

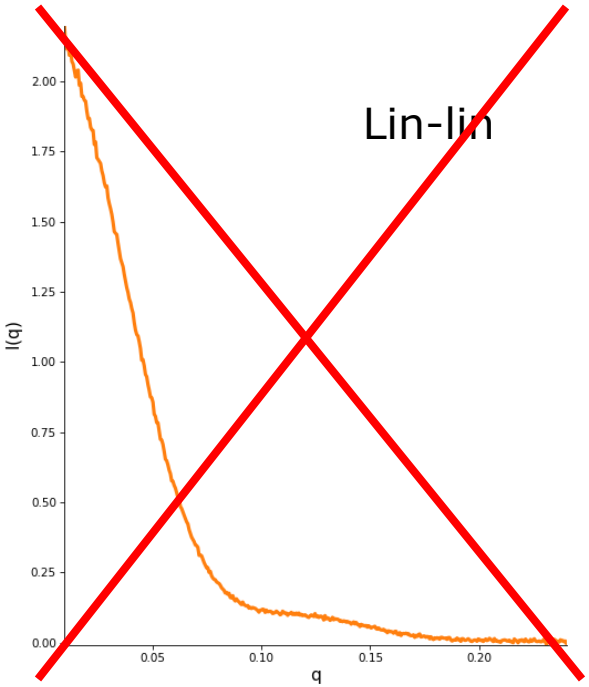


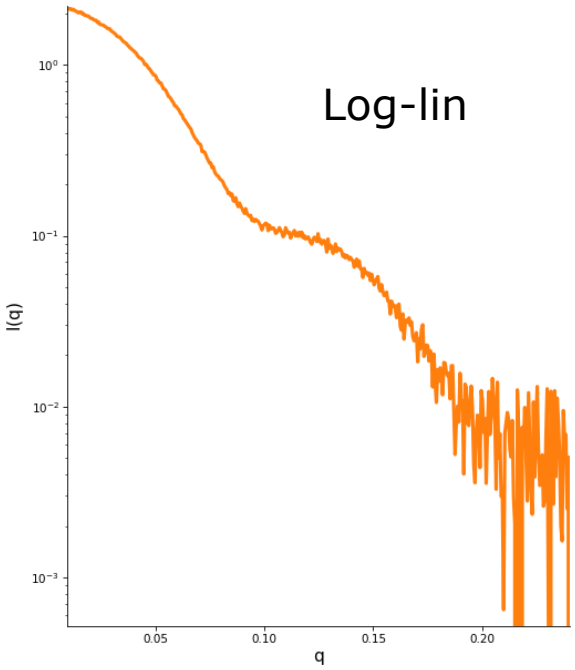


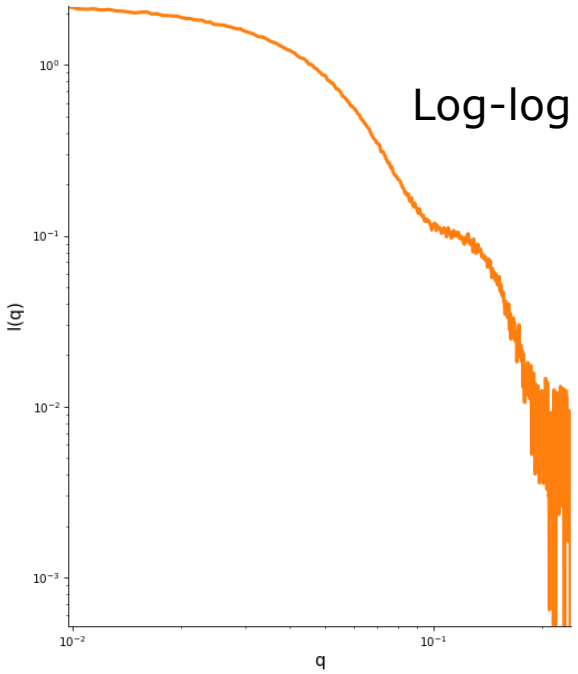


Plotting the scattering profile

Same profile, three different plots

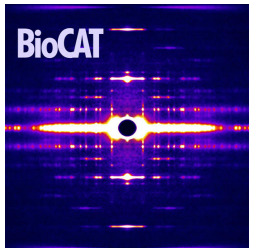




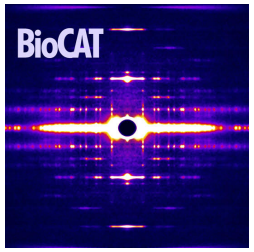


Profile covers 3-4 orders of magnitude. A linear y axis hides significant features

Log-lin emphasizes mid to high q (shape), log-log emphasizes low q (size)



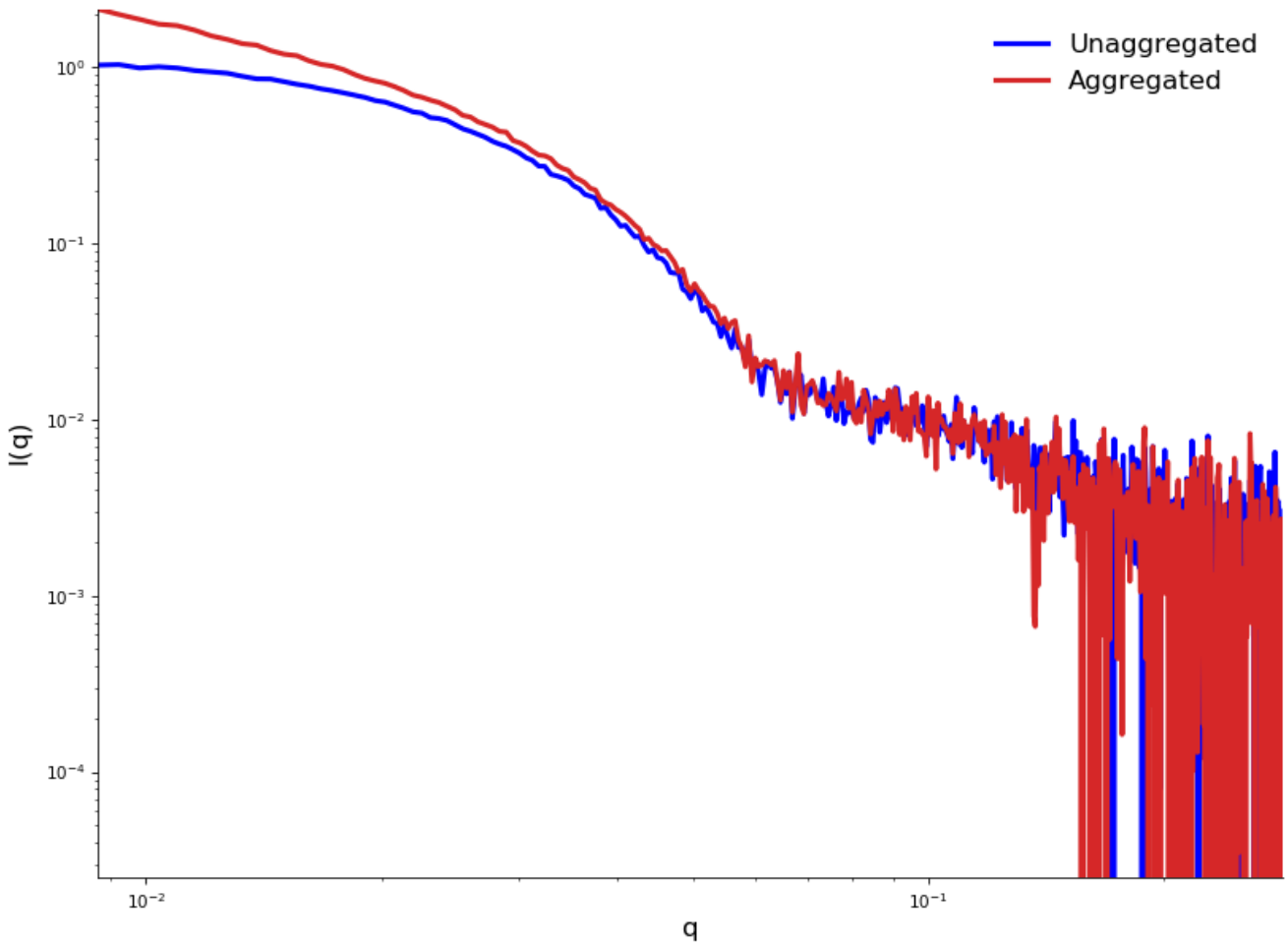
What can go wrong with your data



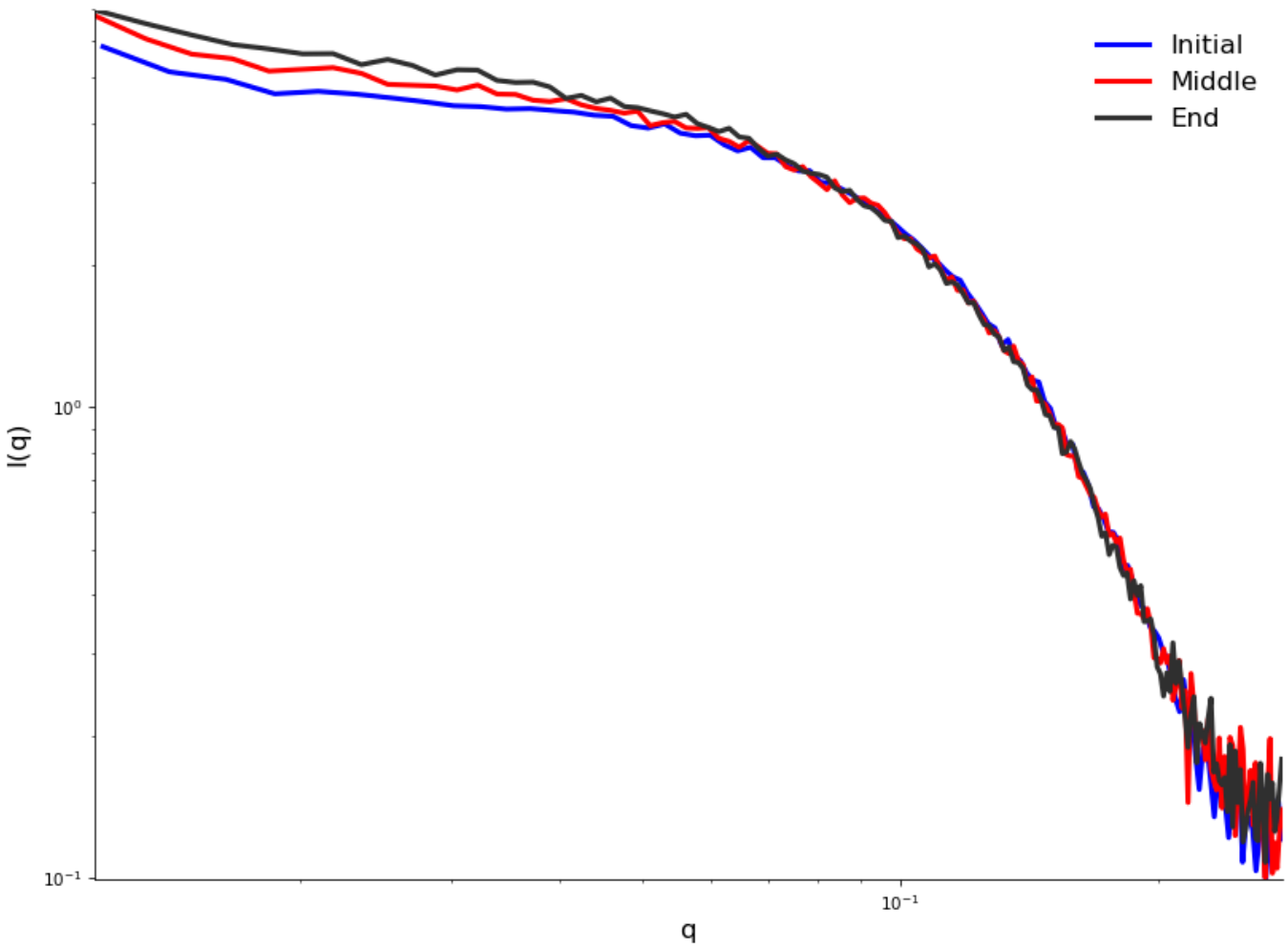
What can go wrong with your data

- Poor quality sample
  - Aggregates or unexpected oligomers in solution
- Radiation damage
  - Time dependent changes in the measured profile
- Concentration effects (structure factor)
  - Concentration dependent changes in the measured profile
  - Uptick (attraction) or downturn (repulsion) at low q
- Bad buffer subtraction
  - Profile going negative at high q or low q (over subtraction)
  - Profile offset at high q, uptick at low q (under subtraction)

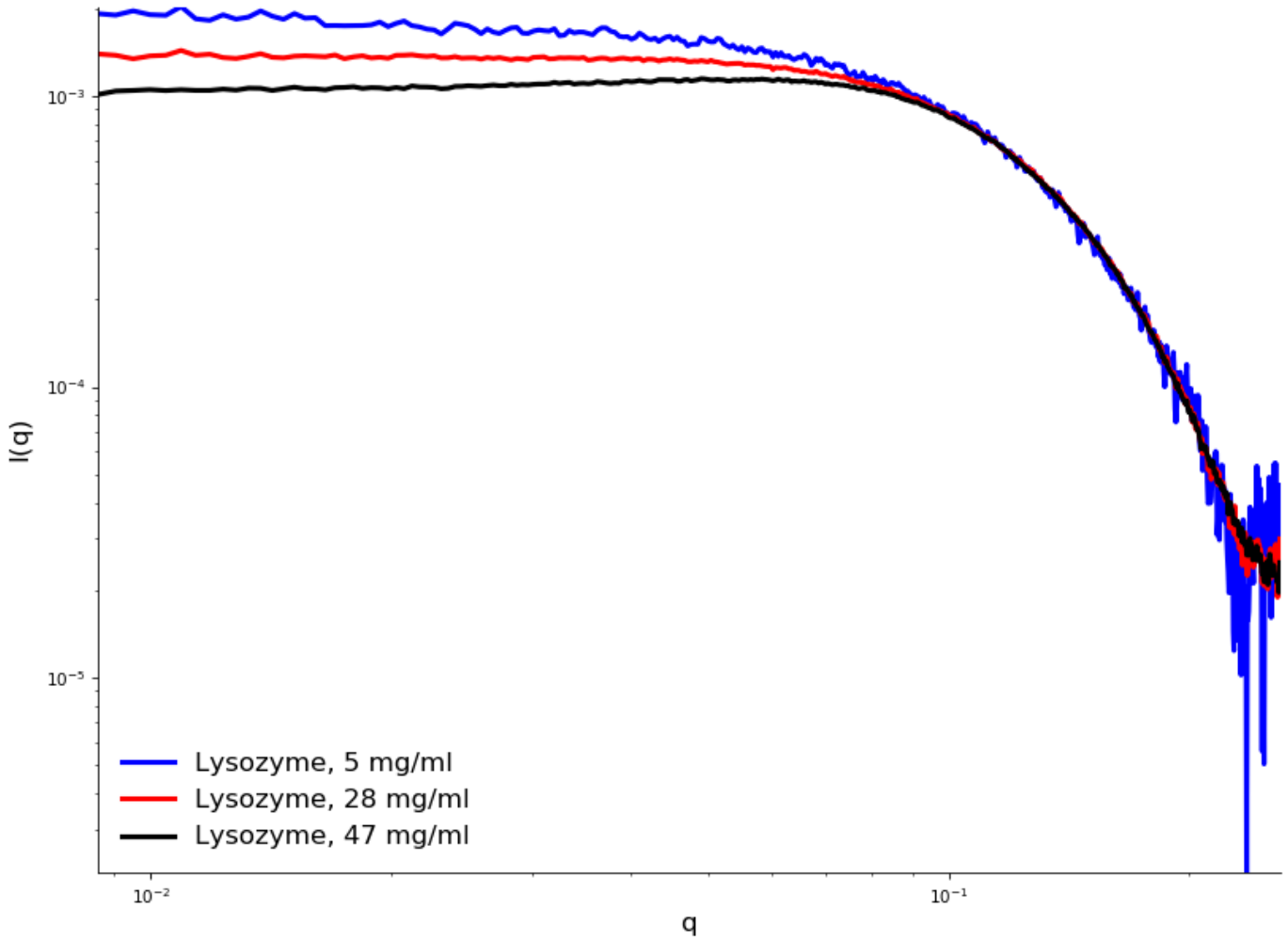
Aggregation

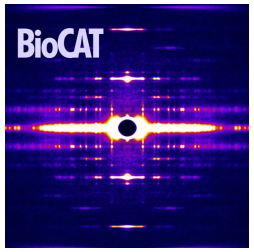


Radiation damage



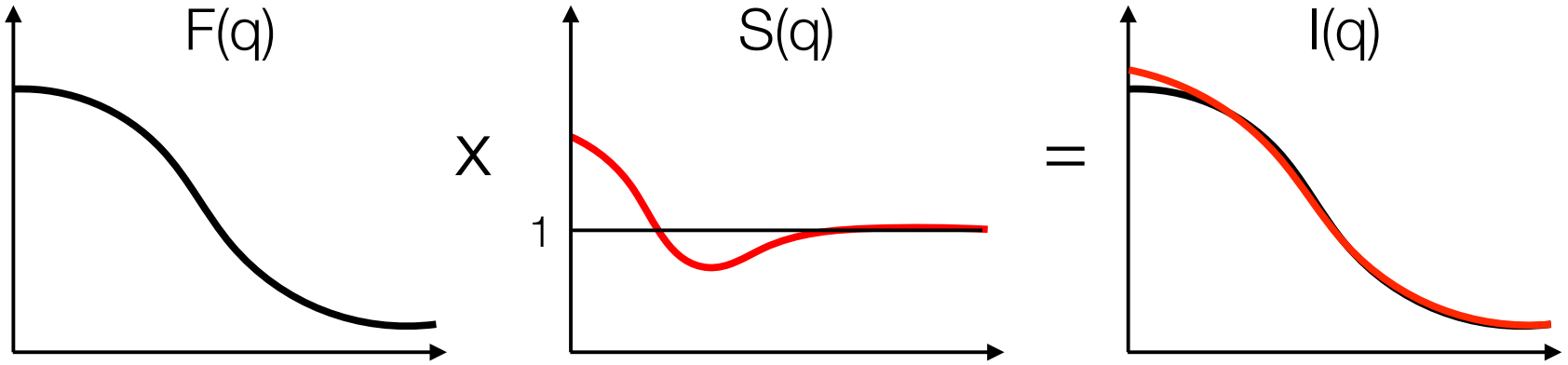
Interparticle Interaction





Interparticle Interactions

Attraction



Repulsion

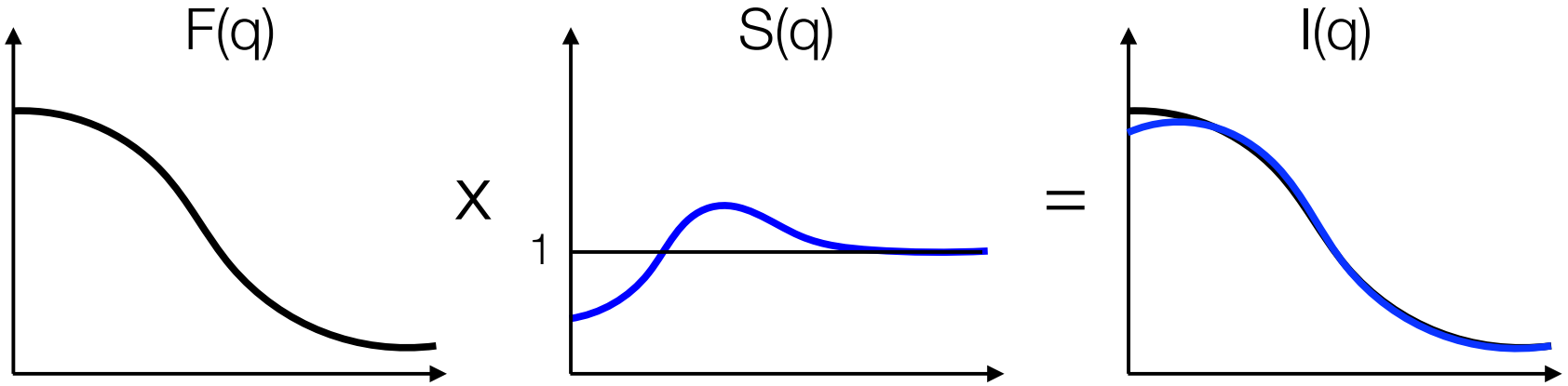
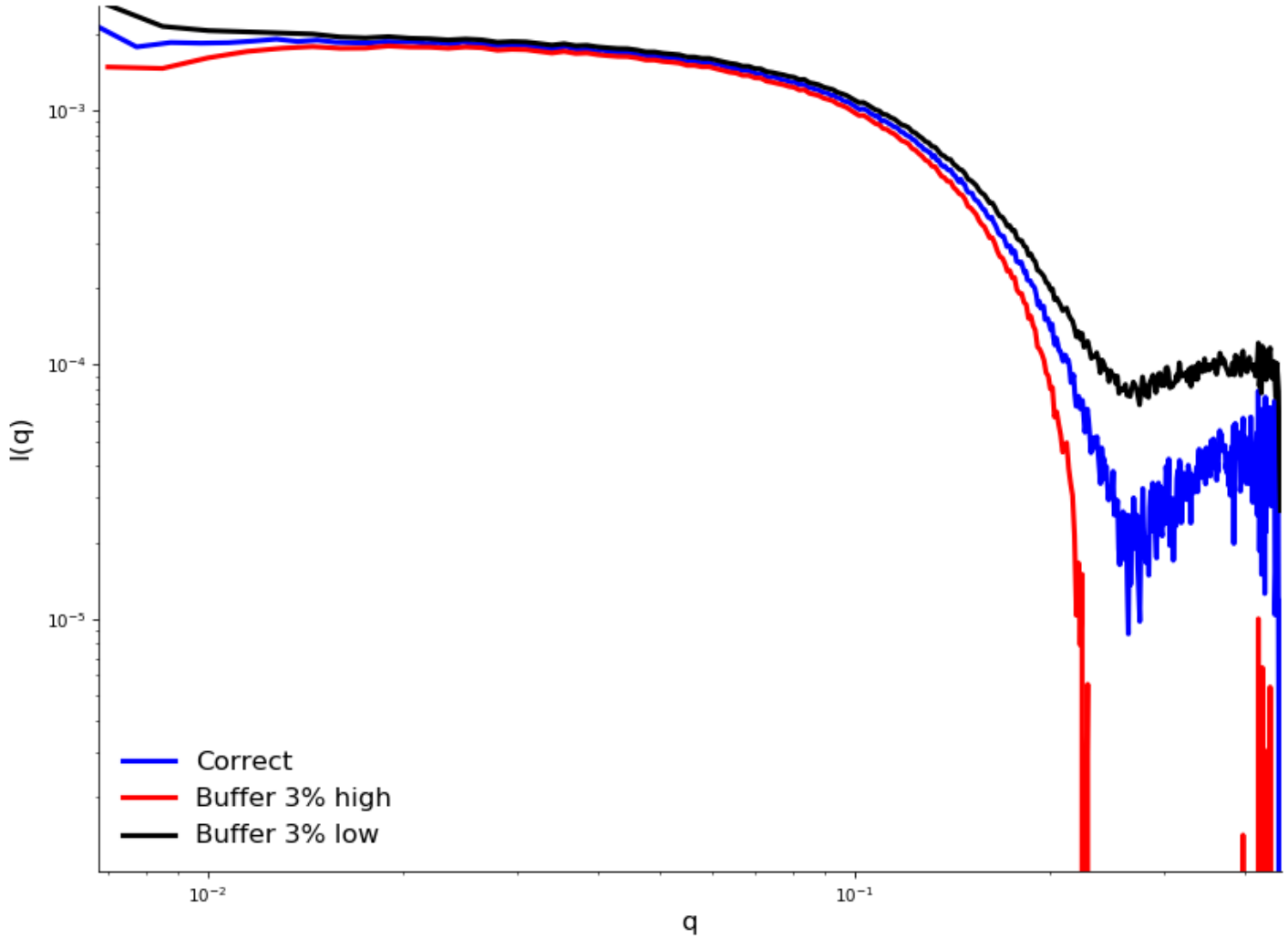
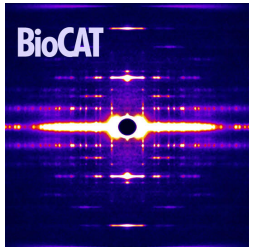


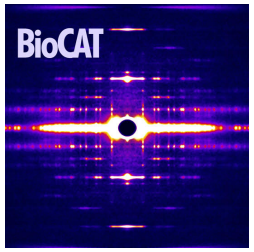
Image from a talk by Thomas Grant

Subtraction errors





Guinier analysis



Guinier analysis

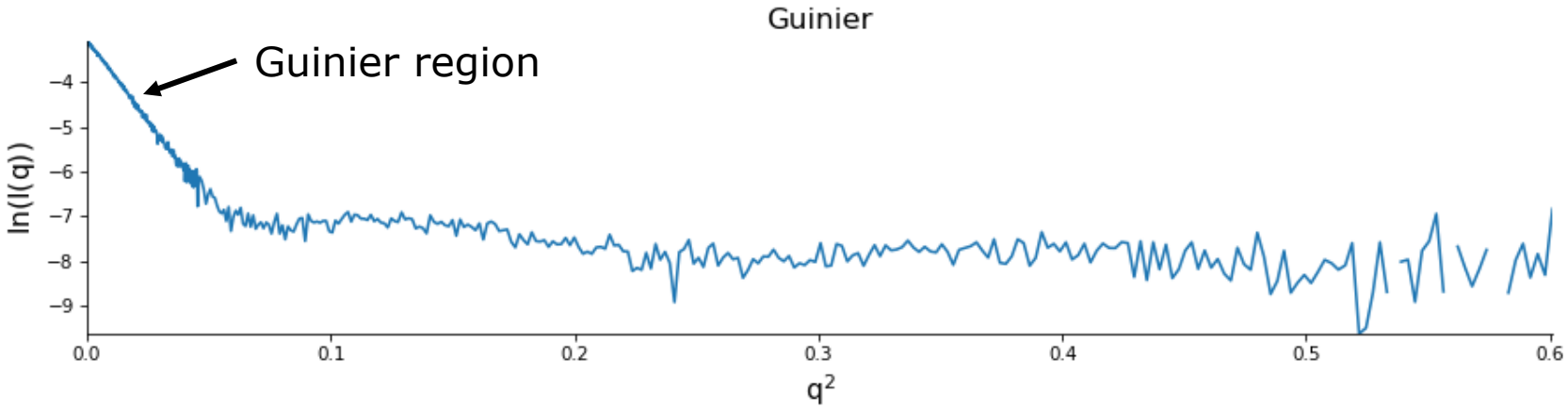
- Developed by Andre Guinier in 1939

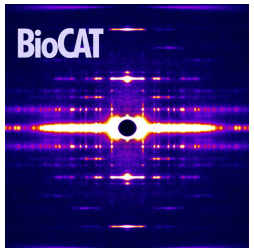
- As $q \to 0$, intensity can be approximated by:

$$I(q) = I(0)e^{-q^2 R_g^2/3}$$

R_g = "radius of gyration"

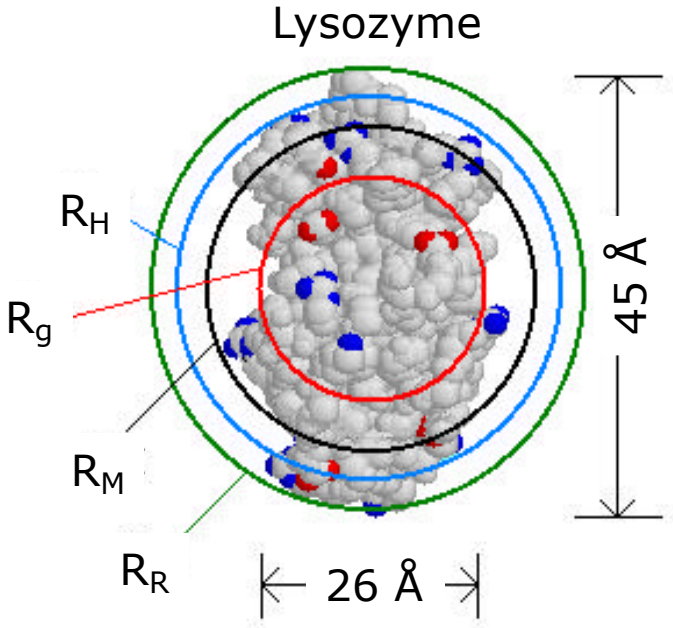
- Plot $\log(I)$ vs. q^2: slope = $-R_g^2/3$, intercept = $\log(I(0))$





Guinier analysis

- Radius of gyration:
  - RMS distance from center of mass



R_g – radius of gyration
R_H – hydrodynamic radius
R_M – radius of mass-equivalent sphere
R_R – maximum hard sphere radius

Useful definitions of R_g

$$R_g^2 = \frac{1}{N}\sum \|\vec{r_i} - \overrightarrow{r_{COM}}\|^2$$ by atoms

$$R_g^2 = \int_V r^2\rho(r)dr \Big/ \int_V \rho(r)dr$$ by electron density

$$R_g^2 = \frac{1}{2N(N-1)}\sum_i\sum_j \|\vec{r_i} - \vec{r_j}\|^2$$ by atom pairs

$$R_g^2 = \frac{1}{2}\int_V r^2p(r)dr \Big/ \int_V p(r)dr$$ by pair distribution

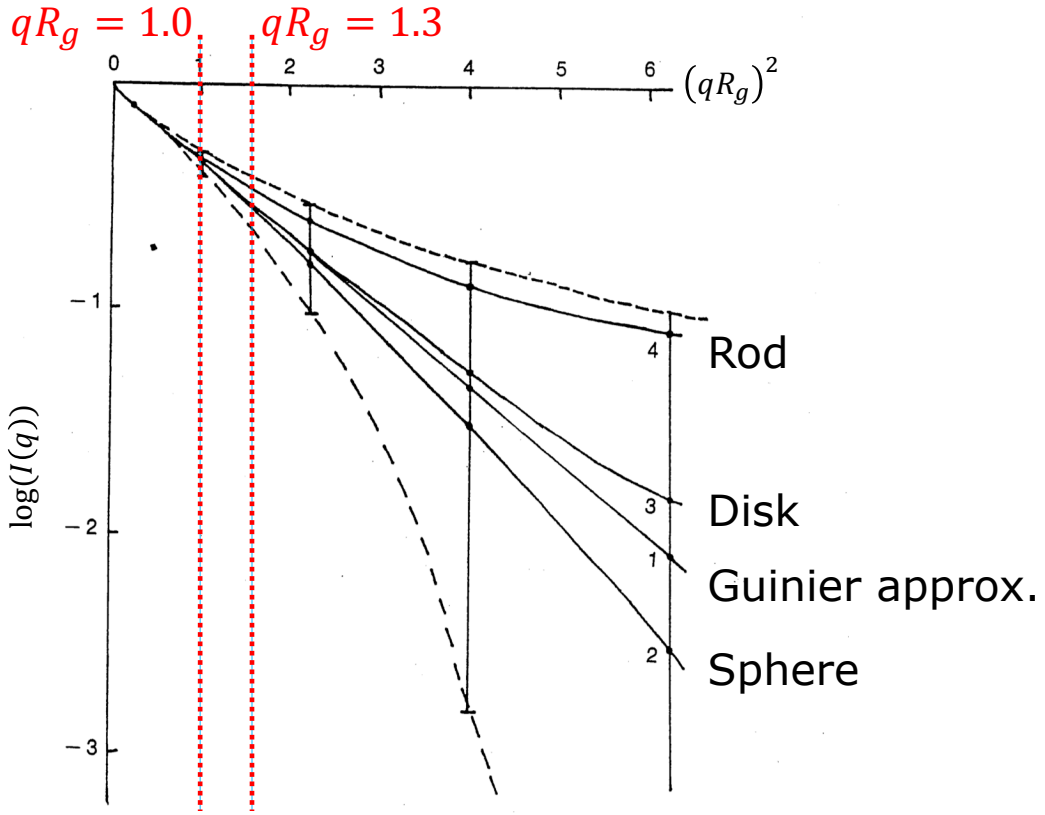
Image from:
www.silver-colloids.com/Papers/hydrodynamic-radius.pdf

Guinier analysis

- The Guinier approximation is only accurate at low q. How do you pick you fit endpoints?
  - It depends on particle shape and size!

$$I(q) = I_0 e^{-q^2 R_g^2/3}$$

Need qR_g sufficiently small that this approximate holds



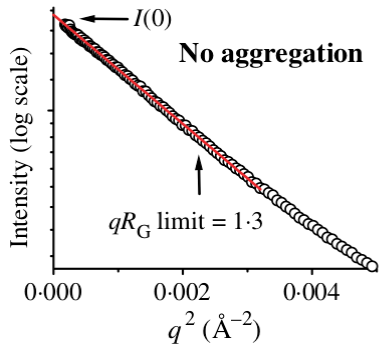
- Conventionally, we fit the Guinier region out to $qR_g \approx 1.3$
  - This works for globular molecules
- Rods need to be fit to $qR_g \approx 1$
- Guinier region should be fit to as low q as your data
  - Do not cut out low q data!
- Need $q_{min}R_g < 1$, preferably $q_{min}R_g < 0.65$

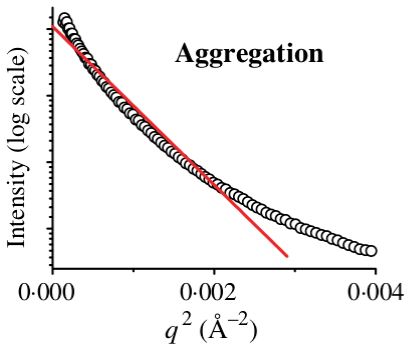
Image from Feigun and Svergun, Structure Analysis by Small-Angle X-ray and Neutron Scattering (1987).

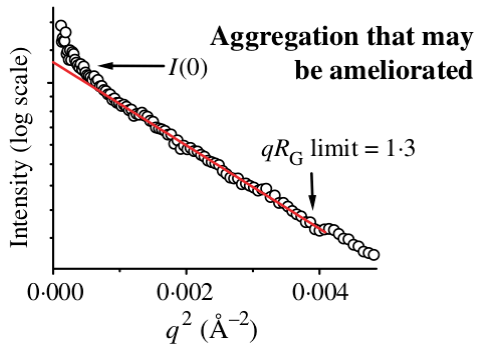
Guinier analysis

- Non-linearities in Guinier analysis are indicative of problems with your sample

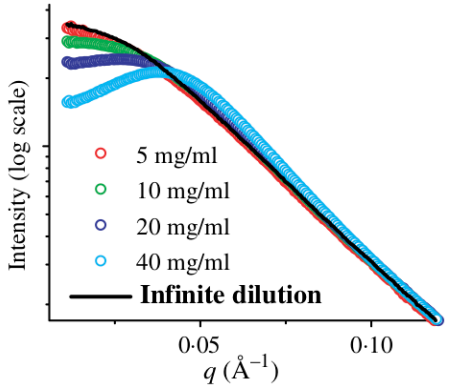
Aggregation causes a characteristic upturn at low q
- Could be caused by aggregates in the sample, or by radiation induced aggregation (radiation damage)







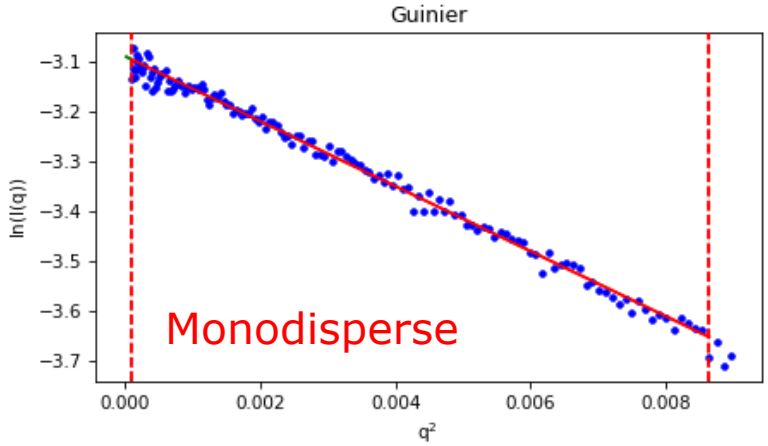
Downturns at low q are characteristic of structure factor, also show up in a Guinier fit

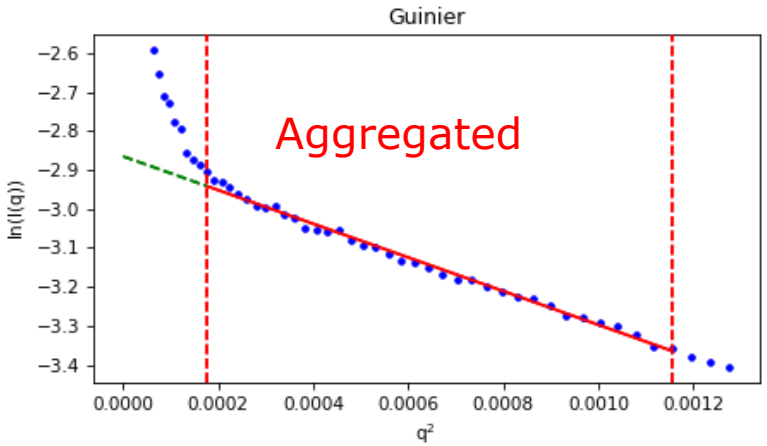


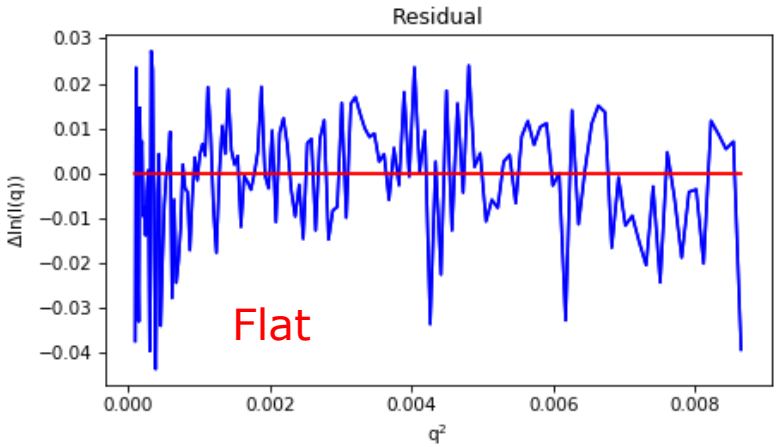
Images from Putnam et al. Quarterly reviews of biophysics, 40(3) 2007.

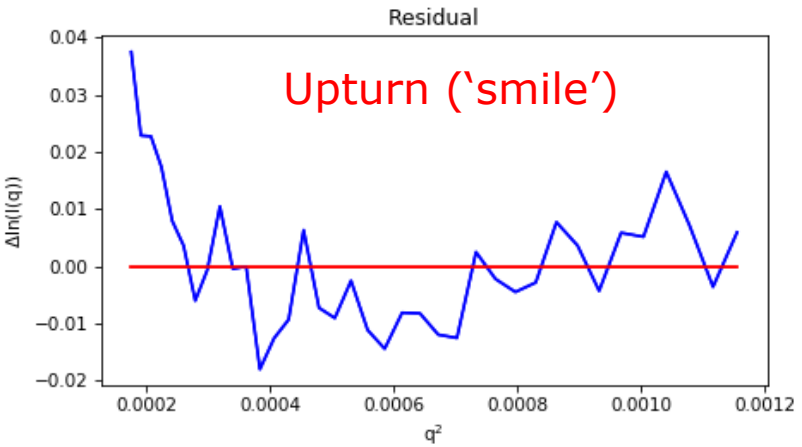
Guinier analysis

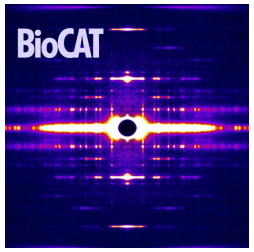
- Fit residual can help you see problems





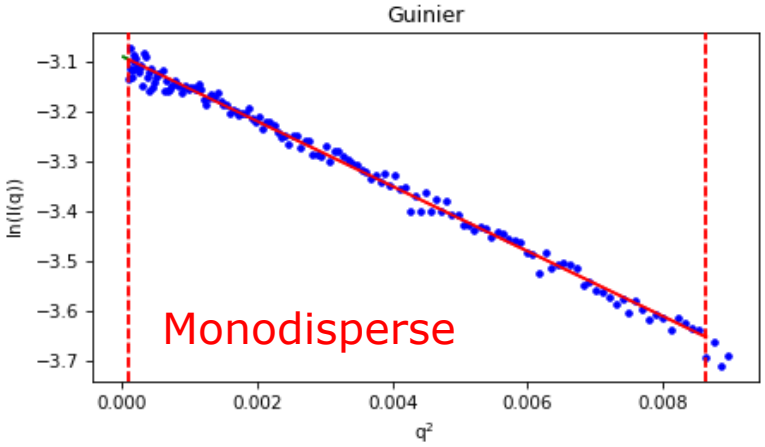


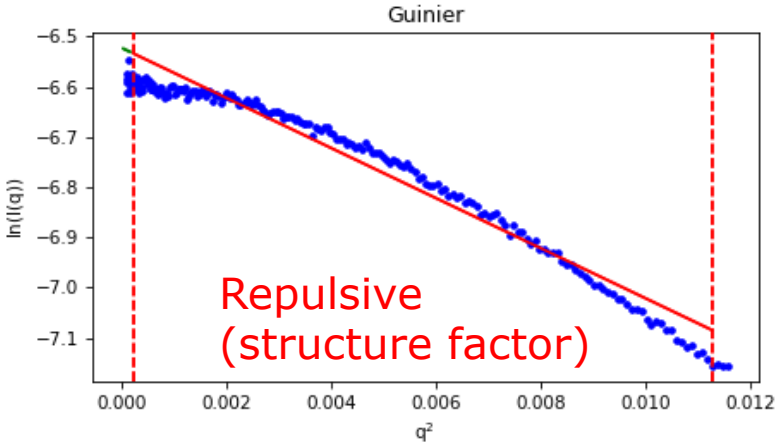


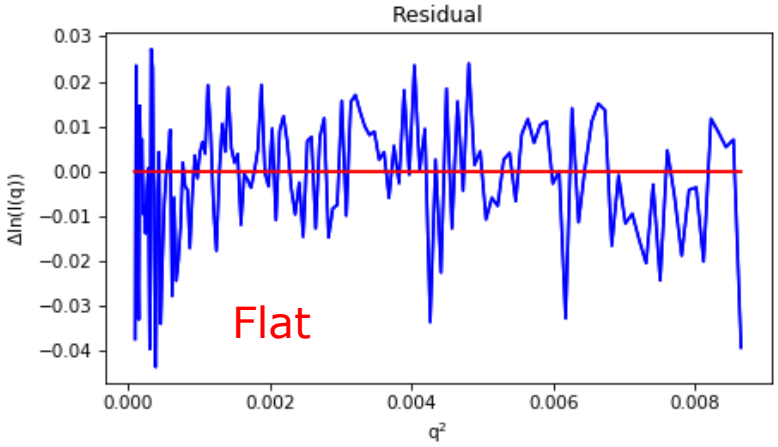


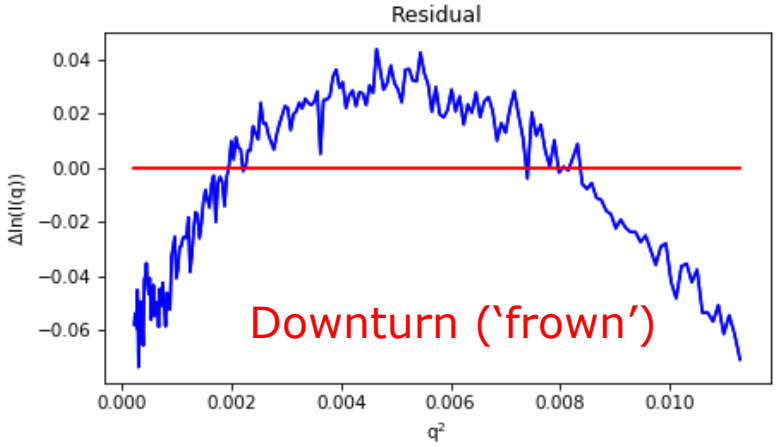
Guinier analysis

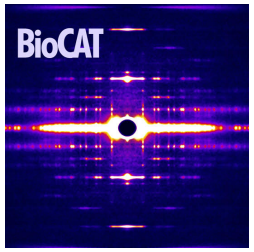
- Fit residual can help you see problems





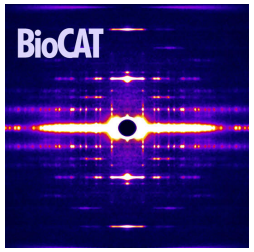




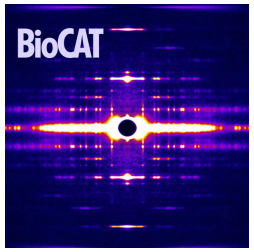


Guinier analysis summary

- Guinier analysis sensitive to low q
- Most problems with your data will show up here!
  - Aggregation
  - Radiation damage
  - Interparticle interactions
  - Some buffer subtraction issues
- Guinier region should be linear, with flat fit residuals
  - Upturn in profile or residuals usually aggregation
  - Downturn in profile or residuals usually repulsion
- Gives R_g, informs on particle size
- Gives $I(0)$, informs on particle mass



Molecular weight analysis



Molecular weight from SAXS

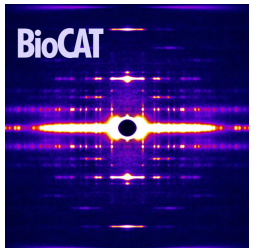
Molecular weight estimates from SAXS are ~10% accurate at best. Despite this, it is important to estimate MW to verify it matches what you expect

1. **I(0) in absolute units (water/glassy carbon standard)**
   Scattering intensity actually has "absolute" units of cm^{-1} when properly calibrated with a known standard such as water. *Once I(0) is expressed in absolute units*,

$$\text{Mol. Wt.} = \frac{N_A I(0)/c}{(\Delta\rho_M)^2}$$

$N_A = 6.02 * 10^{23}$ c = concentration $\Delta\rho_M$ = "scattering contrast"

Reference: Mylonas, E. & Svergun, D. I. (2007). *J. Appl. Cryst. 40, S245-S249*



Molecular weight from SAXS

1. **I(0) in absolute units (water/glassy carbon standard)**
   Scattering intensity actually has "absolute" units of cm^{-1} when properly calibrated with a known standard such as water. *Once I(0) is expressed in absolute units*,

$$\text{Mol. Wt.} = \frac{N_A I(0)/c}{(\Delta\rho_M)^2}$$

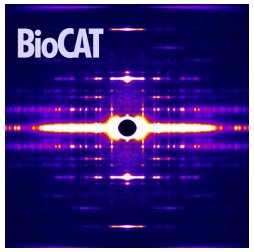
$N_A = 6.02 * 10^{23}$ c = concentration $\Delta\rho_M$ = "scattering contrast"

Reference: Mylonas, E. & Svergun, D. I. (2007). *J. Appl. Cryst. 40, S245-S249*

2. **Protein standards**
   Unknown molecular weights can be determined by comparison with known protein standards such as lysozyme or glucose isomerase:

$$\text{Mol. Wt.} = \frac{I(0)/c}{I(0)_{std}/c_{std}}(MW_{std})$$

Reference: Mylonas, E. & Svergun, D. I. (2007). *J. Appl. Cryst. 40, S245-S249*



Molecular weight from SAXS

1. **I(0) in absolute units (water/glassy carbon standard)**
   Scattering intensity actually has "absolute" units of cm^{-1} when properly calibrated with a known standard such as water. *Once I(0) is expressed in absolute units*,

$$\text{Mol. Wt.} = \frac{N_A I(0)/c}{(\Delta\rho_M)^2}$$

$N_A = 6.02 * 10^{23}$ c = concentration $\Delta\rho_M$ = "scattering contrast"

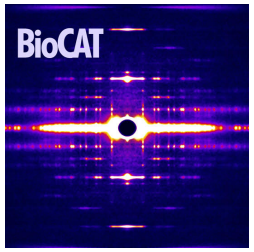
Reference: Mylonas, E. & Svergun, D. I. (2007). *J. Appl. Cryst. 40, S245-S249*

Both of these methods require accurate concentration measurements!

2. **Protein standards**
   Unknown molecular weights can be determined by comparison with known protein standards such as lysozyme or glucose isomerase:

$$\text{Mol. Wt.} = \frac{I(0)/c}{I(0)_{std}/c_{std}}(MW_{std})$$

Reference: Mylonas, E. & Svergun, D. I. (2007). *J. Appl. Cryst. 40, S245-S249*



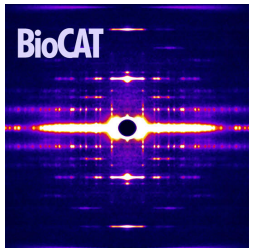
Molecular weight from SAXS

3. **Porod volume and relative scale methods**
   Mass (in kDa) can calculated as the density times the volume of the particle. The Porod volume of the particle is used, and is calculated:

$$V = 2\pi^2 I(0) / \int_0^\infty q^2 I(q) dq$$

   The density used is typically $0.83*10^{-3}$ kDa/Å^3, but can be adjusted for the particular application.

   More advanced techniques based on this idea can give quite accurate
   Reference: Fischer et al. (2009). *J. Appl. Cryst.*, 43, 101-109



Molecular weight from SAXS

3. Porod volume and relative scale methods

Mass (in kDa) can calculated as the density times the volume of the particle. The Porod volume of the particle is used, and is calculated:

$$V = 2\pi^2 I(0) / \int_0^\infty q^2 I(q) dq$$

The density used is typically 0.83*10^{-3} kDa/Å^3, but can be adjusted for the particular application.

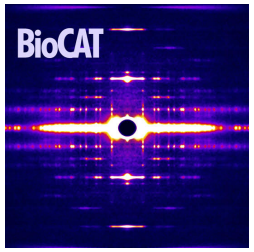
More advanced techniques based on this idea can give quite accurate
Reference: Fischer et al. (2009). *J. Appl. Cryst.*, 43, 101-109

4. Volume of correlation method

Molecular weight can be estimated using the empirically relation:

$$\text{MW} = \left(\frac{Q_R}{c}\right)^{1/k} \quad \text{where} \quad Q_R = \frac{V_c^2}{R_g} \quad \text{and} \quad V_c = \frac{I(0)}{\int qI(q)dq}$$

The values of *k* and *c* depend on the type of macromolecule. For proteins *k* = 1 and *c* = 0.1231, for RNA *k* = 0.808 and *c* = 0.00934.
Reference: Rambo and Tainer (2013). *Nature*, 496, 477-481



Molecular weight from SAXS

3. Porod volume and relative scale methods

Mass (in kDa) can calculated as the density times the volume of the particle. The Porod volume of the particle is used, and is calculated:

$$V = 2\pi^2 I(0) / \int_0^\infty q^2 I(q) dq$$

The density used is typically 0.83*10^{-3} kDa/Å^3, but can be adjusted for the particular application.

Both of these methods do not rely on the concentration of the sample, making them useful as checks for methods 1 and 2, and in cases where the concentration may not be known (such as SEC-SAXS).

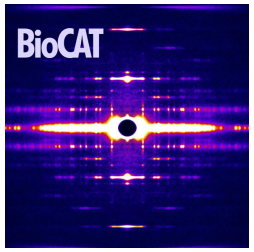
4. Volume of correlation method

Molecular weight can be estimated using the empirically relation:

$$\text{MW} = \left(\frac{Q_R}{c}\right)^{1/k} \quad \text{where} \quad Q_R = \frac{V_c^2}{R_g} \quad \text{and} \quad V_c = \frac{I(0)}{\int qI(q)dq}$$

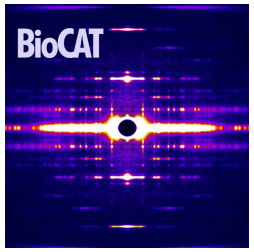
The values of *k* and *c* depend on the type of macromolecule. For proteins *k* = 1 and *c* = 0.1231, for RNA *k* = 0.808 and *c* = 0.00934.

Reference: Rambo and Tainer (2013). *Nature*, 496, 477-481



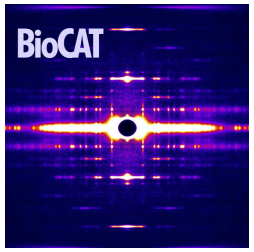
Molecular weight from SAXS

- Four methods, each fails in different ways
  - Absolute scale – Requires accurate calculation of macromolecule contrast, partial specific volume. Depends on accuracy of concentration, absolute scale calibration
  - Reference to known standard: Reference standard must be in a buffer with similar contrast as your sample. Depends on accuracy of concentration for both reference and your standard
  - Porod volume: Works best for compact, globular, rigid molecules. Requires accurately knowing the macromolecule density.
  - Volume of correlation: Fails for protein-nucleic acid complexes. Requires the integral to converge. Sensitive to noisy high q data. Fails for molecules $\lesssim$ 20 kDa.
- All methods will fail if your Guinier fit is bad
- Integral methods are sensitive to accurate background subtraction

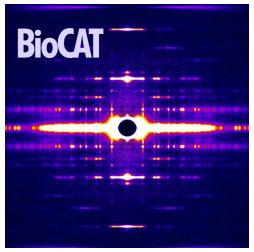


Molecular weight in SAXS

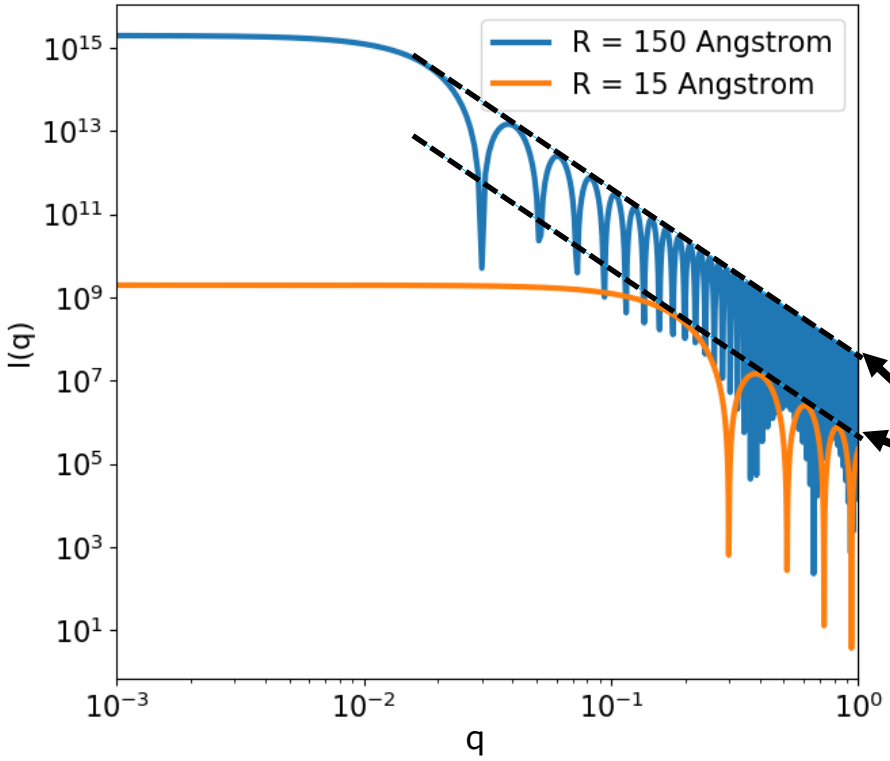
- Be aware of different failure modes
- Use the method(s) that should work best for your data, not the one that best matches your expectations
- Verify that MW matches expected oligomeric state
- If MW doesn't match expected, don't assume you know what's going on. Could be an error in MW calculation, could be a sample problem.
  - Test with another method (e.g. MALS)



Porod and Kratky analysis



Porod Analysis



Objects with sharp boundaries, like ideal spheres, have scattering that follow Porod's law at wide angles:

$$I(q) \propto q^{-4} \quad \text{with} \quad qR \gg 1$$

Slope = -4

Hard sphere: $I(q) \propto q^{-4}$
Unfolded 'random walk' polymer: $I(q) \propto q^{-2}$
Fully extended chain: $I(q) \propto q^{-1}$

The Porod exponent can be interpreted in terms of particle shape and porosity (usually for materials)

Be careful: requires perfect background subtraction

Kratky analysis

Unfolded proteins have Porod exponents near -2, folded generally near -4 (if globular)

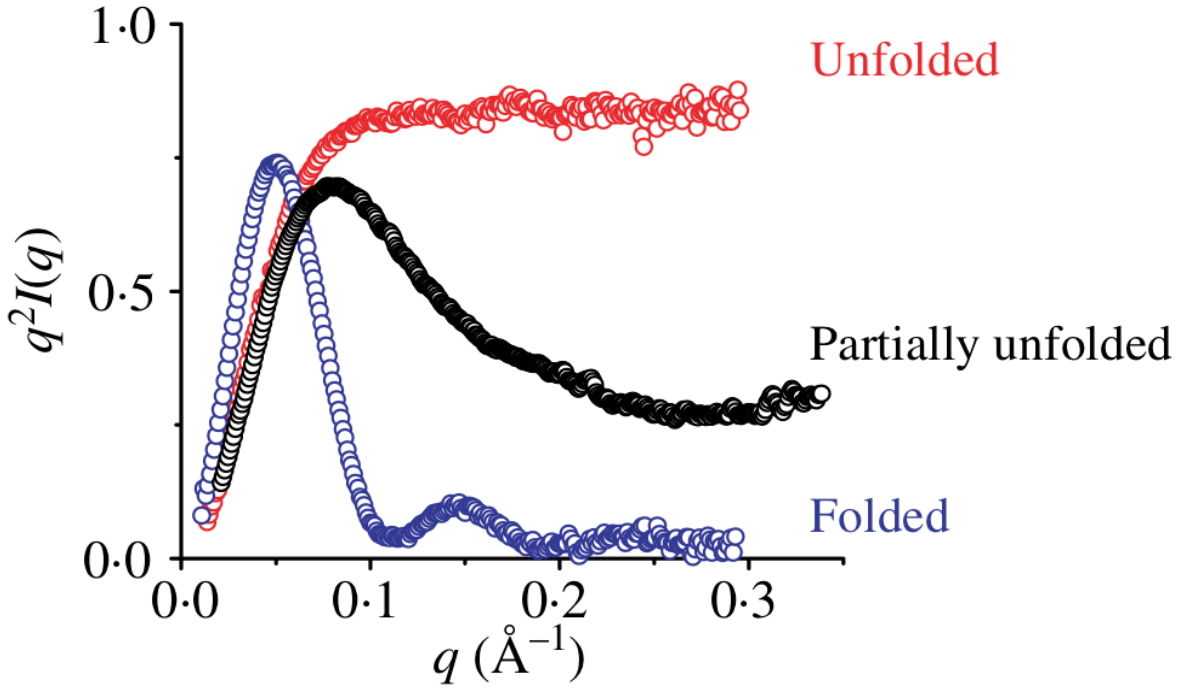
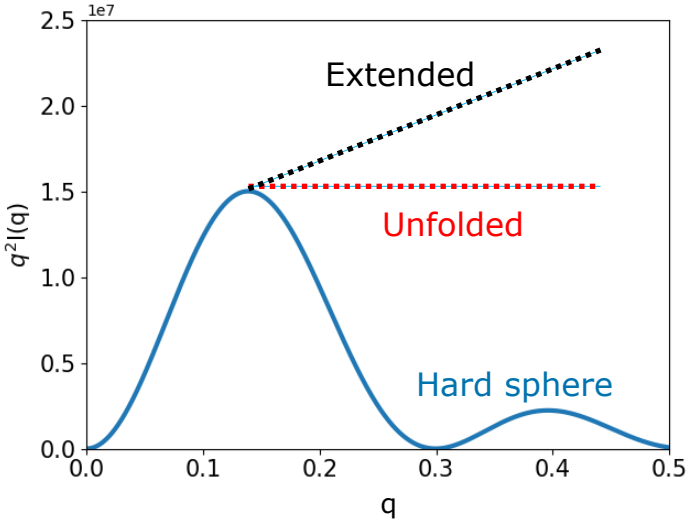


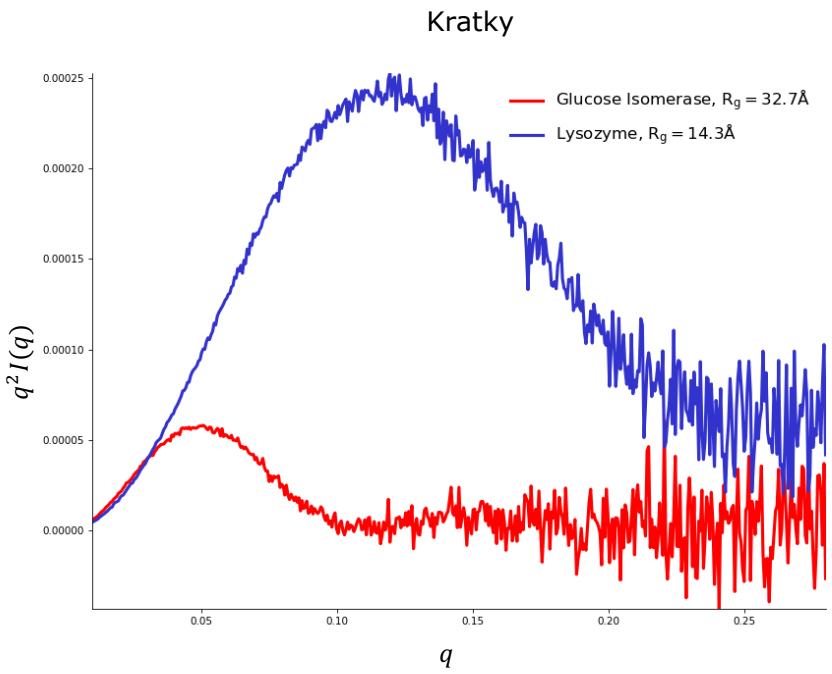
Image from Putnam et al. Quarterly reviews of biophysics, 40(3) 2007.

Kratky plot:
$q^2I(q)$ vs q



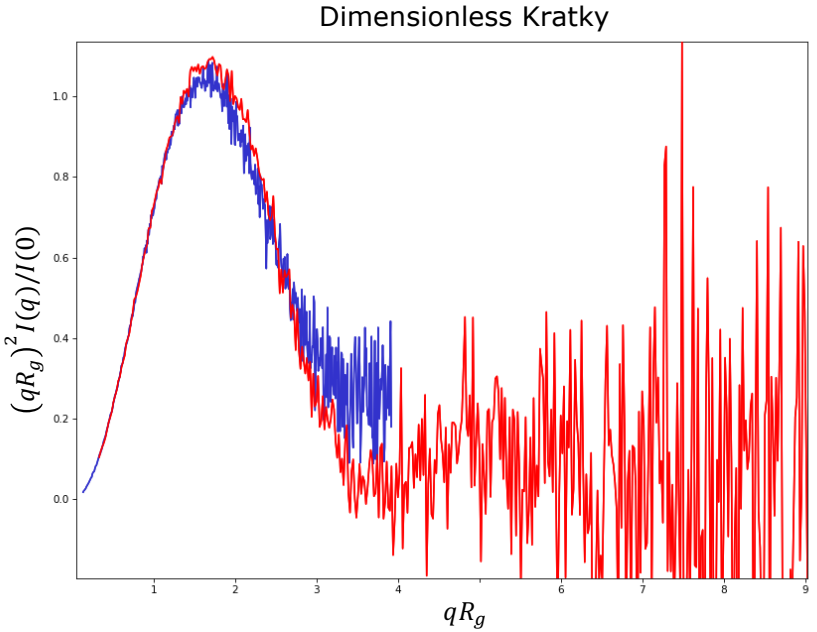
Kratky analysis

Problem: Kratky plot depends on size of an object, scaling of scattering profiles



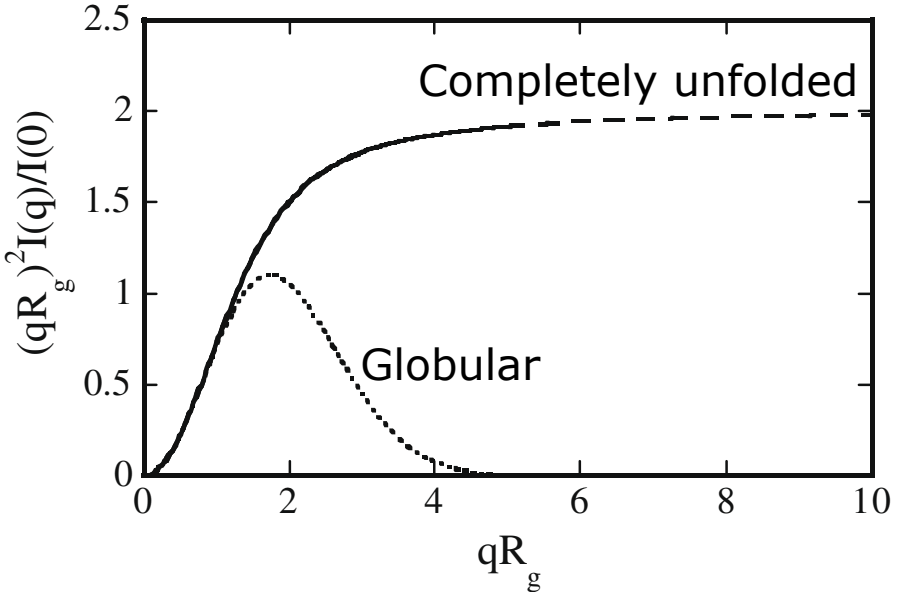
Solution: normalize by R_g and $I(0)$

Dimensionless Kratky plot:
$(qR_g)^2 I(q)/I(0)$ vs. qR_g



Kratky analysis

Globular particles all have the same shape.
Deviations inform on flexibility/extendedness



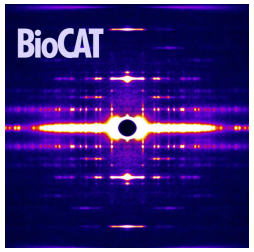
An ideal random chain rises to a plateau of 2

A fully extended chain continues to slope upward without a plateau (not shown)

Globular particles have a maximum of 1.1 at $qR_g = \sqrt{3} \approx 1.73$

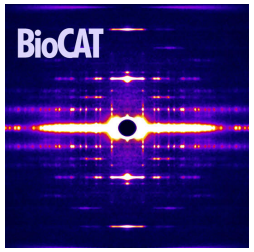
Image from Durand et al. J. Struct. Biol. 169, 2010

Shifts in peak location to the right of 1.73, or a partial plateau, indicate more flexibility in a system. Changes in size/shape are directly comparable because the curves are dimensionless.

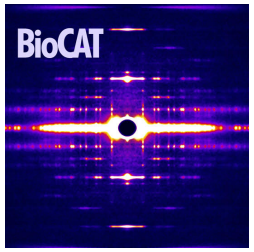


Kratky analysis

- Kratky plots inform on flexibility and shape
- Kratky plots are relatively insensitive to a small amount of aggregates or radiation damage
- Kratky plots are extremely sensitive to buffer subtraction issues
- Dimensionless Kratky plots can provide semi-quantitative assessment of flexibility



Indirect Fourier transforms



Indirect Fourier Transform (IFT)

So the scattering profile is the Fourier transform of the electron density. Can we just Fourier transform it back to get the molecular shape?

No.

- The scattering profile is a radial average of the intensities of a rotationally averaged molecule
- We've lost too much information, including phases (which is also an issue in crystallography)

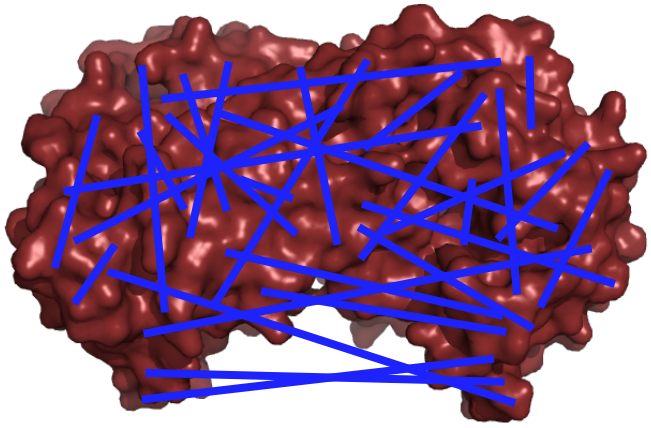
However . . .

We can do an Indirect Fourier Transform (IFT) and get the pair distance distribution function, P(r)

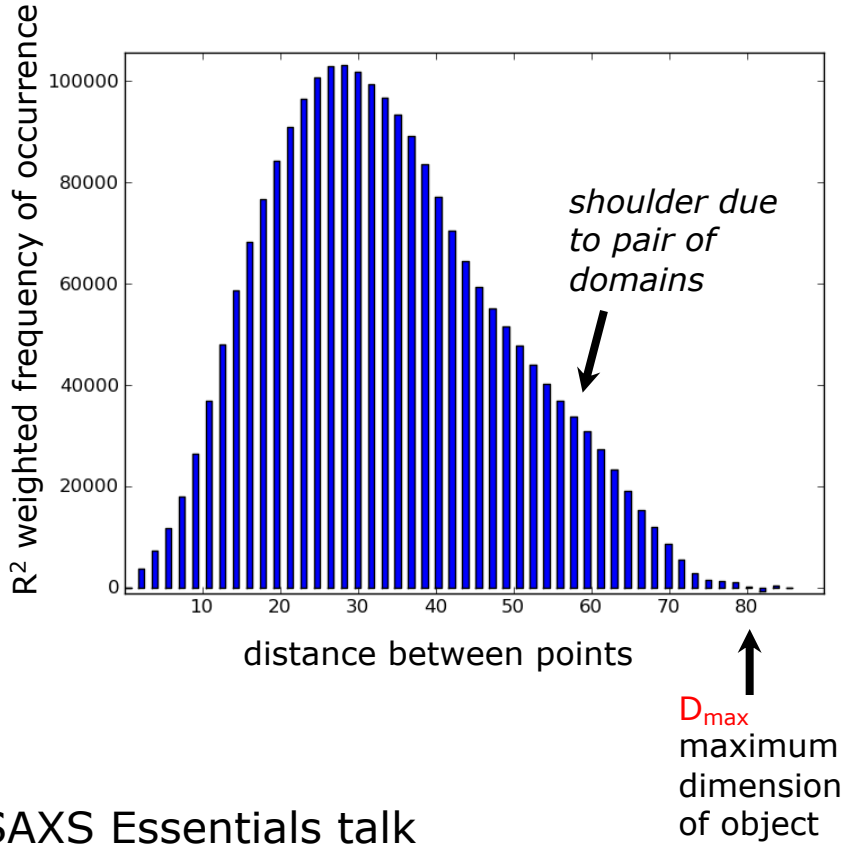
$$I(q) = 4\pi \int_0^{D_{max}} P(r) \frac{\sin(qr)}{qr} dr \quad \longleftrightarrow \quad P(r) = \frac{r^2}{2\pi^2} \int_0^{\infty} q^2 I(q) \frac{\sin(qr)}{qr} dq$$

Physical interpretation of P(r)

P(r) is the r^2 weighted histogram of all possible pairs of electrons: the pair distance distribution function



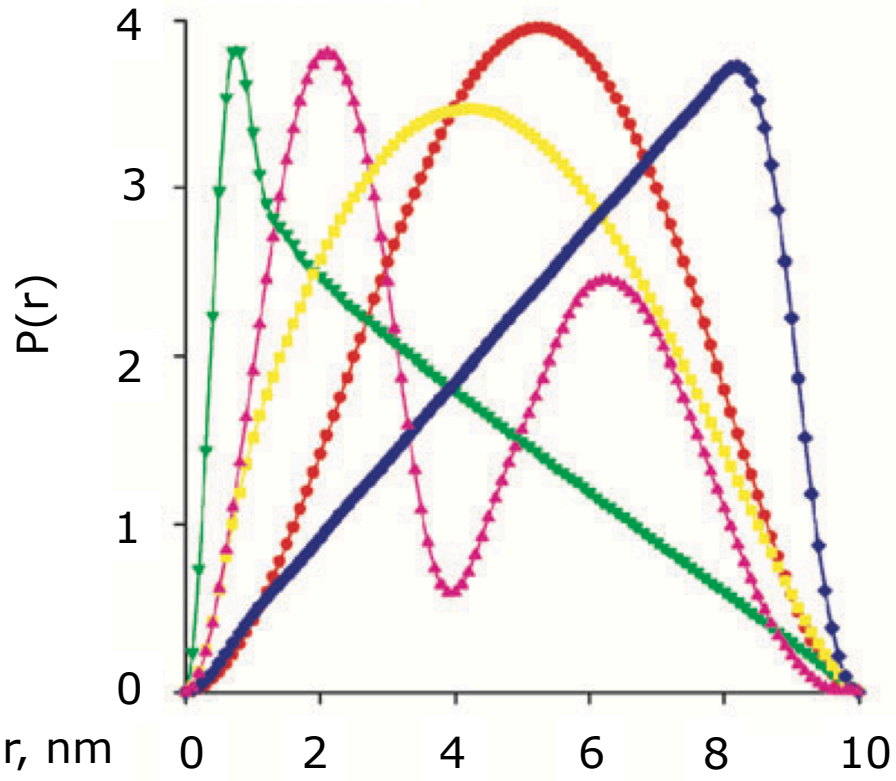
1TIM.pdb

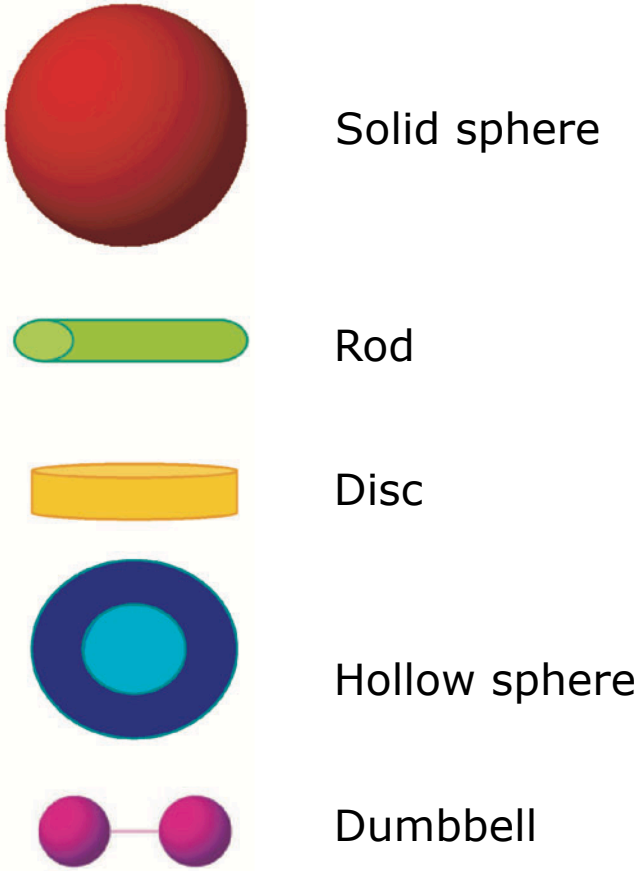


Images from Richard Gillilan's BioSAXS Essentials talk

Physical interpretation of P(r)

The shape of the P(r) function can tell you a lot about the shape of your particle





Images from Svergun & Koch, Reports on the progress of physics, 66, 2003.

Physical interpretation of P(r)

The shape of the P(r) function can tell you a lot about the shape of your particle

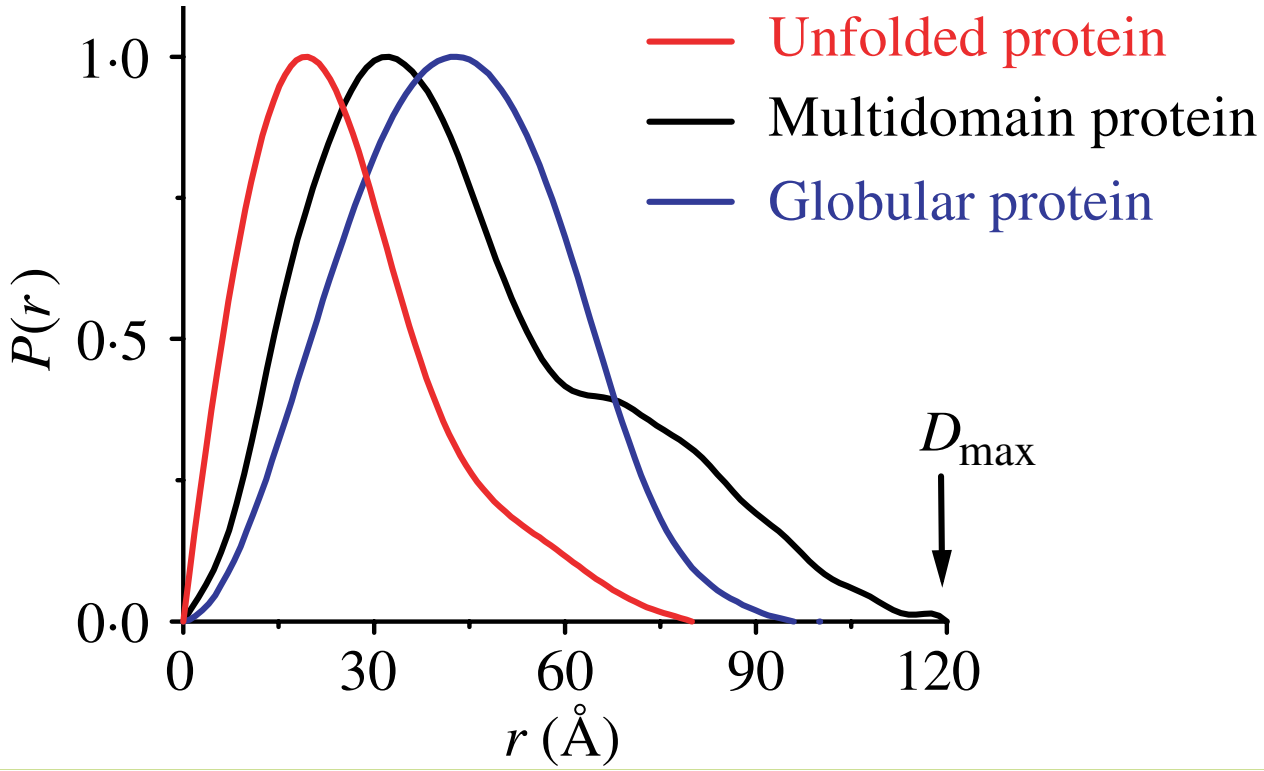
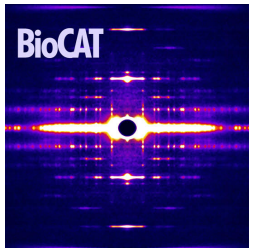


Image from Putnam et al. Quarterly reviews of biophysics, 40(3) 2007.

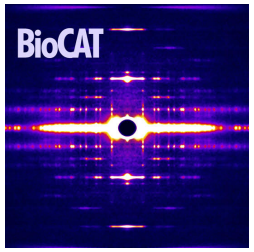


Physical interpretation of P(r)

The P(r) function can be used to calculate the R_g and $I(0)$ values of the curve.

- Uses entire curve
- Less sensitive to interparticle interactions
- Less sensitive to aggregates
- Automatic extrapolation to $q = 0$
- Especially useful for large particles with small Guinier regions and for noisy data
- Good check against Guinier analysis

$$R_g^2 = \frac{\int_0^{D_{max}} r^2 P(r) dr}{2\int_0^{D_{max}} P(r) dr} \qquad I(0) = 4\pi \int_0^{D_{max}} P(r) dr$$



How to calculate a P(r) function

Why can't you directly do a Fourier transform (why the I in IFT)?

$$P(r) = \frac{r^2}{2\pi^2}\int_0^{\infty} q^2 I(q)\frac{\sin(qr)}{qr}\,dq$$

The finite extent of our measurement (and measurement noise) means that a direct Fourier transform distorts the true P(r) function. You get 'truncation artifacts'.

You generate a P(r) with a given D_{max} by fitting against the data

- Fitting criteria include both 'fit' (χ^2) and 'regularization' parameters
- Regularization include 'perceptual' criteria such as
  - Smoothness of the P(r)
  - Systematic deviations from I(q)
  - Stability of the solution when changing parameter weighting
  - Positivity of the solution

How to calculate a P(r) function

Most commonly we use a program called GNOM to do the IFT, though others exist.

- Requires estimate of D_{max} for IFT

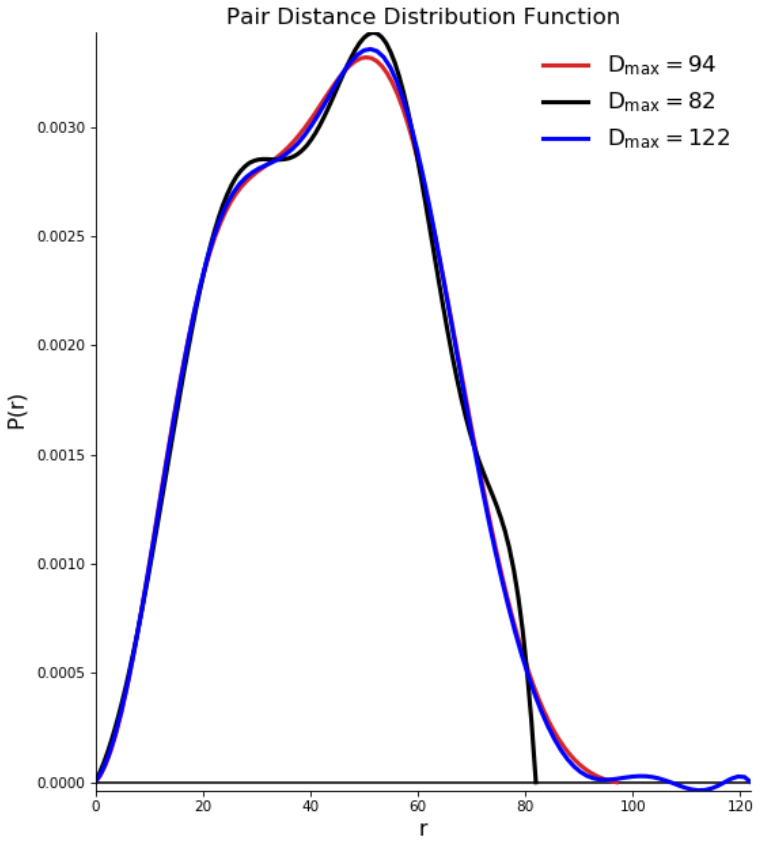
Criteria for judging a good D_{max} based on P(r) function:

- P(r) falls gradually to zero at D_{max}
- Underestimated D_{max} has an abrupt descent
- Overestimated D_{max} usually shows oscillation about zero

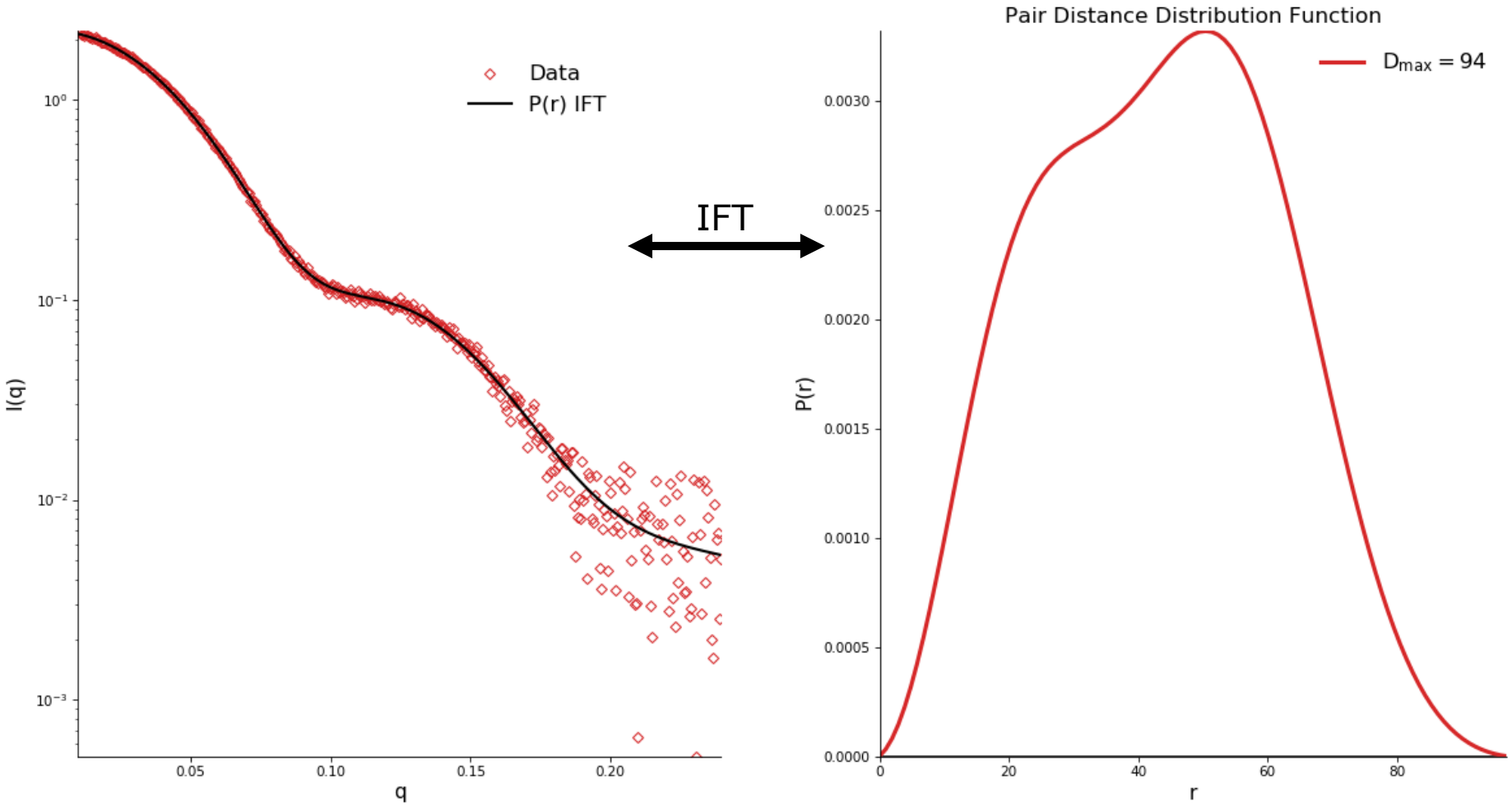
Additional P(r) criteria:

- R_g and $I(0)$ from Guinier and P(r) should agree well
- P(r) goes to zero at $r = 0$ and $r = D_{max}$
- The transform of P(r) fits your data

Even for good data, uncertainty in determining D_{max} can be >10%

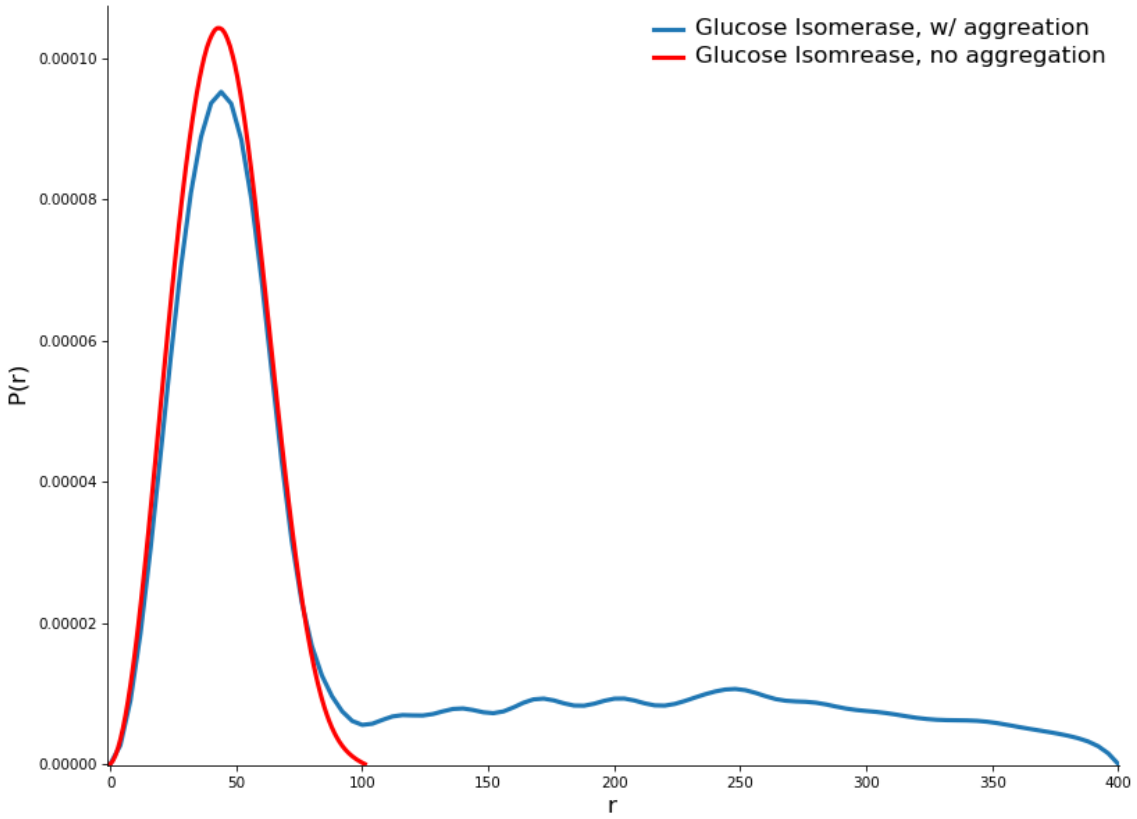


How to calculate a P(r) function



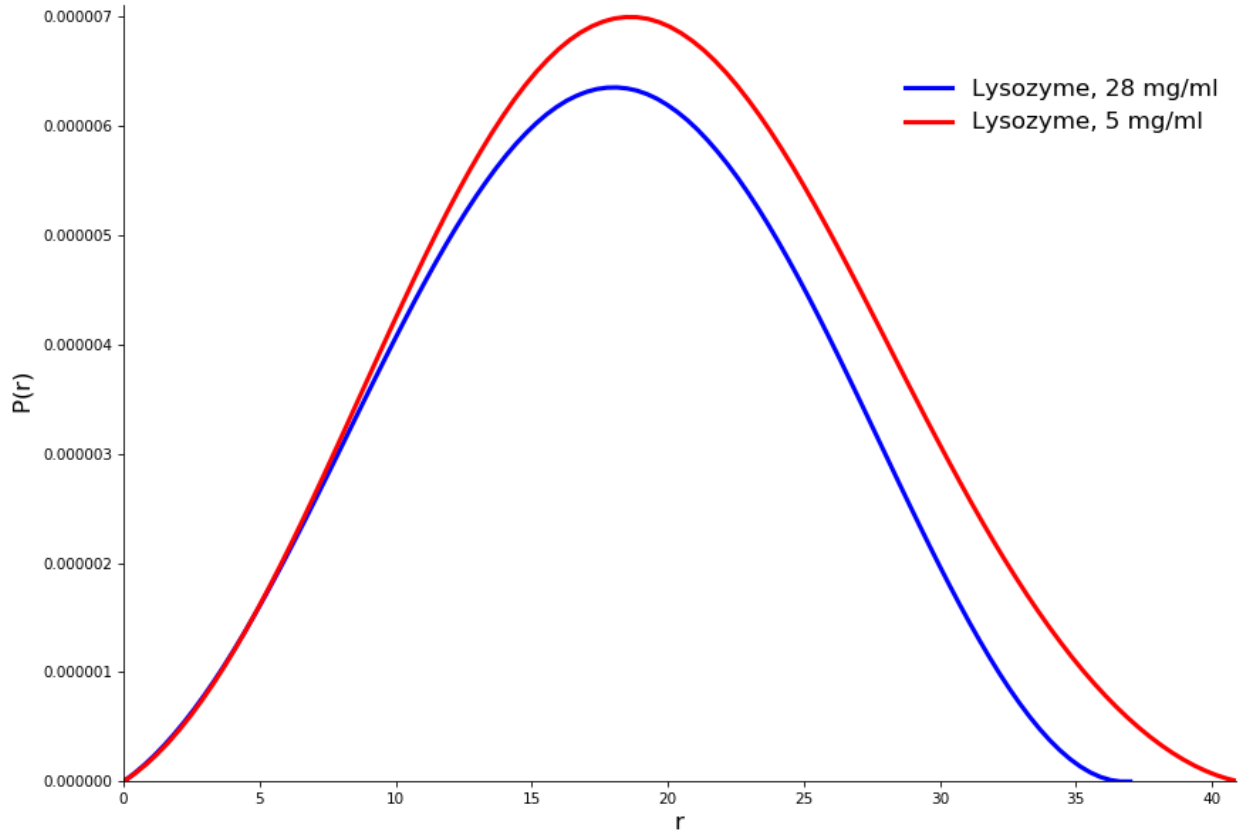
Aggregation and the P(r)

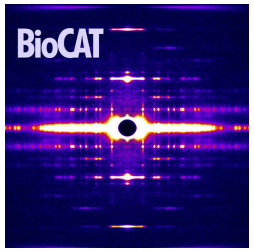
When doing an IFT, if you are unable to find a reasonable D_{max}, may indicate aggregation



Interparticle interference and P(r)

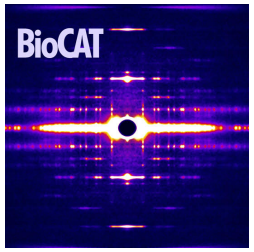
Interparticle interference that leads to a downturn in the low q (repulsion) leads to an artificially small D_{max}





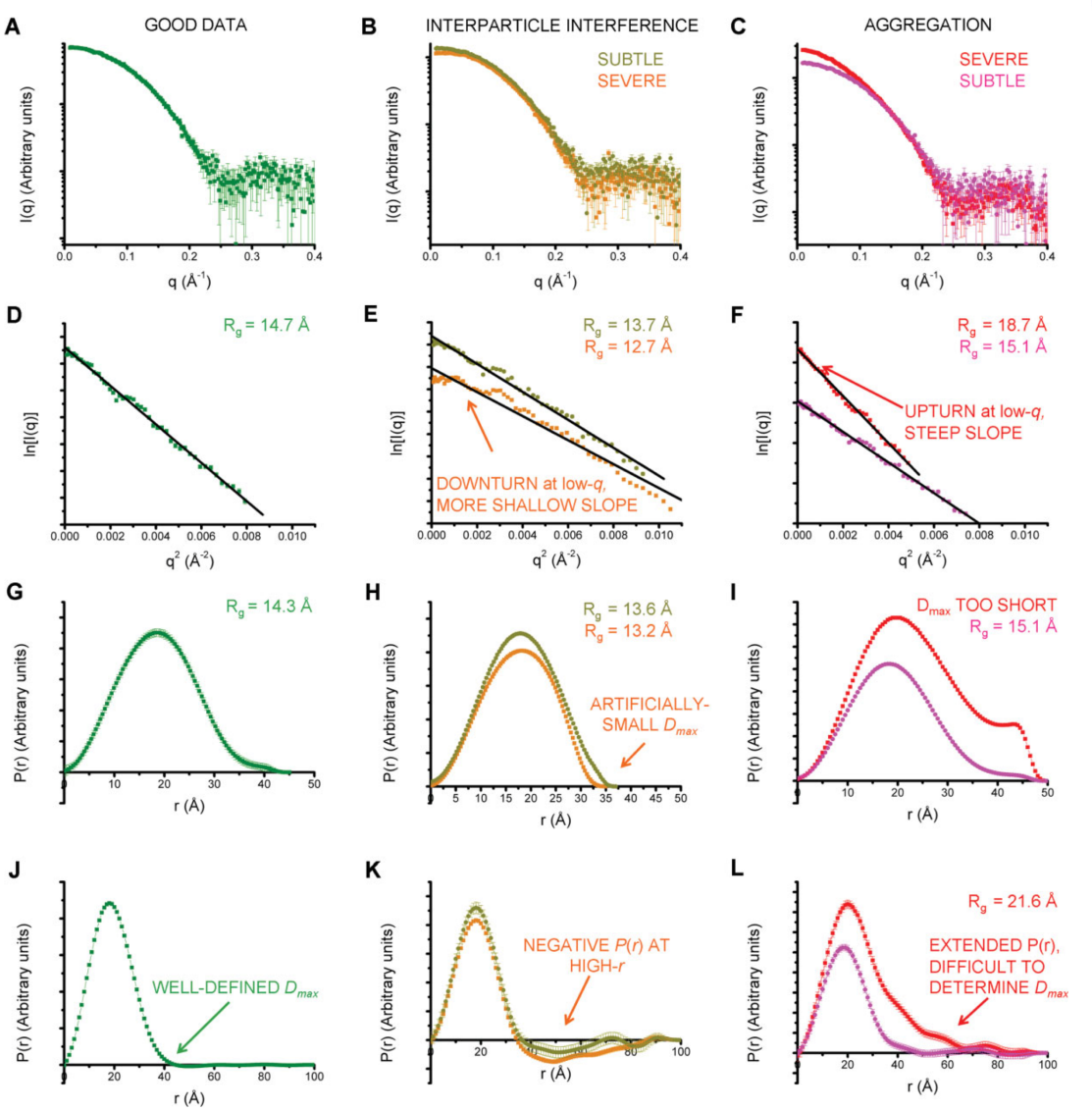
The P(r) function

- Provides real space structural information about the shape of the macromolecule
- Provides an estimate of D_{max}, and more accurate determination of R_g and $I(0)$
- Sensitive to aggregation and interparticle interference
- Generally required before moving on to more advanced analysis

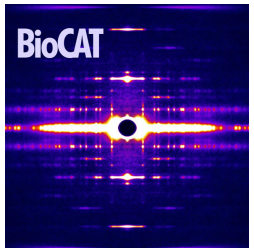


Summary

Summary of data validation

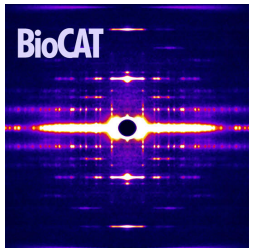


Jacques and Trewhella, Protein Science Review 2010.



Summary of data validation

- Guinier fit will show most issues
- P(r) function good for catching aggregation, interparticle interference
- MW validates what you have in solution
  - Use appropriate method(s)
- Kratky plot particularly sensitive to background subtraction



Summary of data analysis

- Guinier plot gives estimates of R_g and $I(0)$
  - Sensitive to data quality issues
- MW is relatively unreliable from SAXS, but required to validate what state/sample you have in solution
  - Pick the right calculation method
- Kratky and dimensionless Kratky plots provide analysis of flexibility and shape
- P(r) function provides real space shape information, estimate of D_{max}, and more accurate determination of R_g and $I(0)$
  - Also sensitive to data quality issues
- P(r) is generally required before moving to advanced analysis techniques