Statistics in the experimental practice of SAXS

The Basics

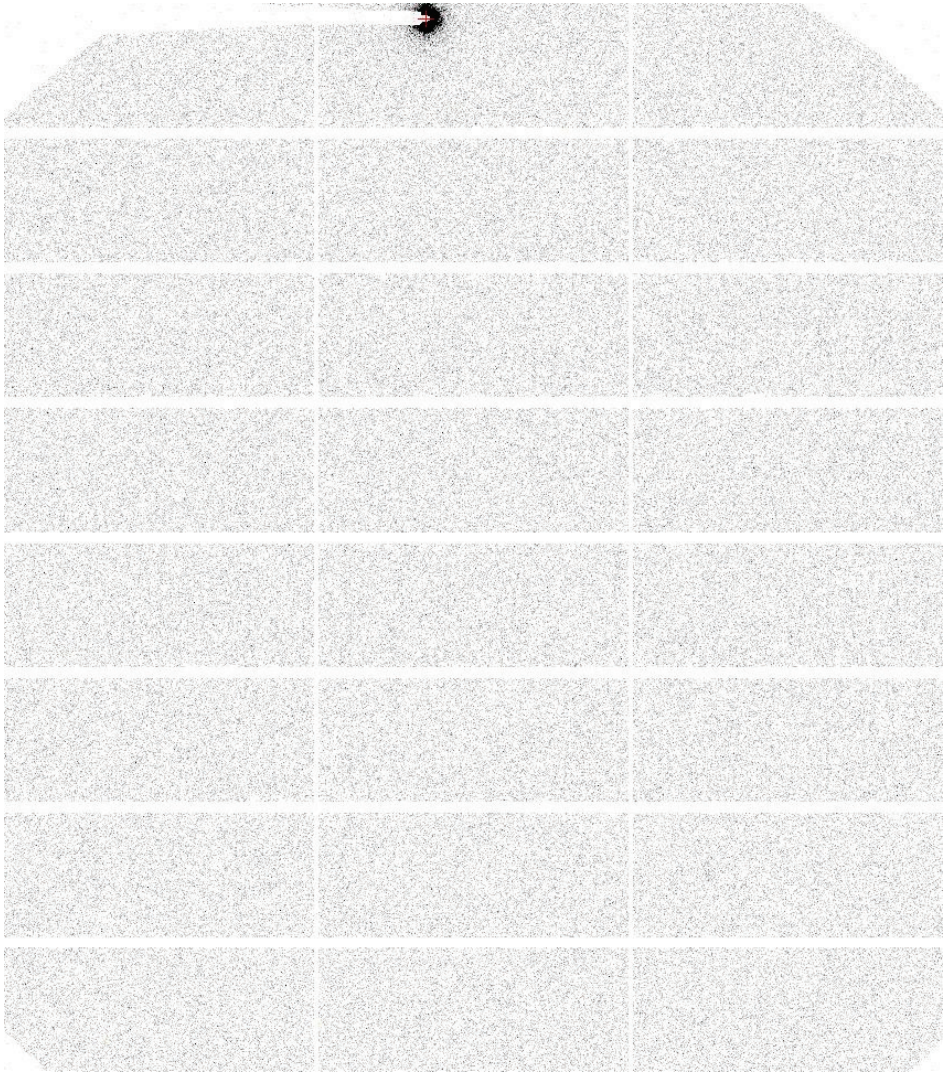
EMBL

Requirements for Data Collection

- sample must be pure and monodisperse
- sufficient volumes for multiple concentrations
- do not make up buffers from scratch, use dialysis buffer
- know your sample concentrations

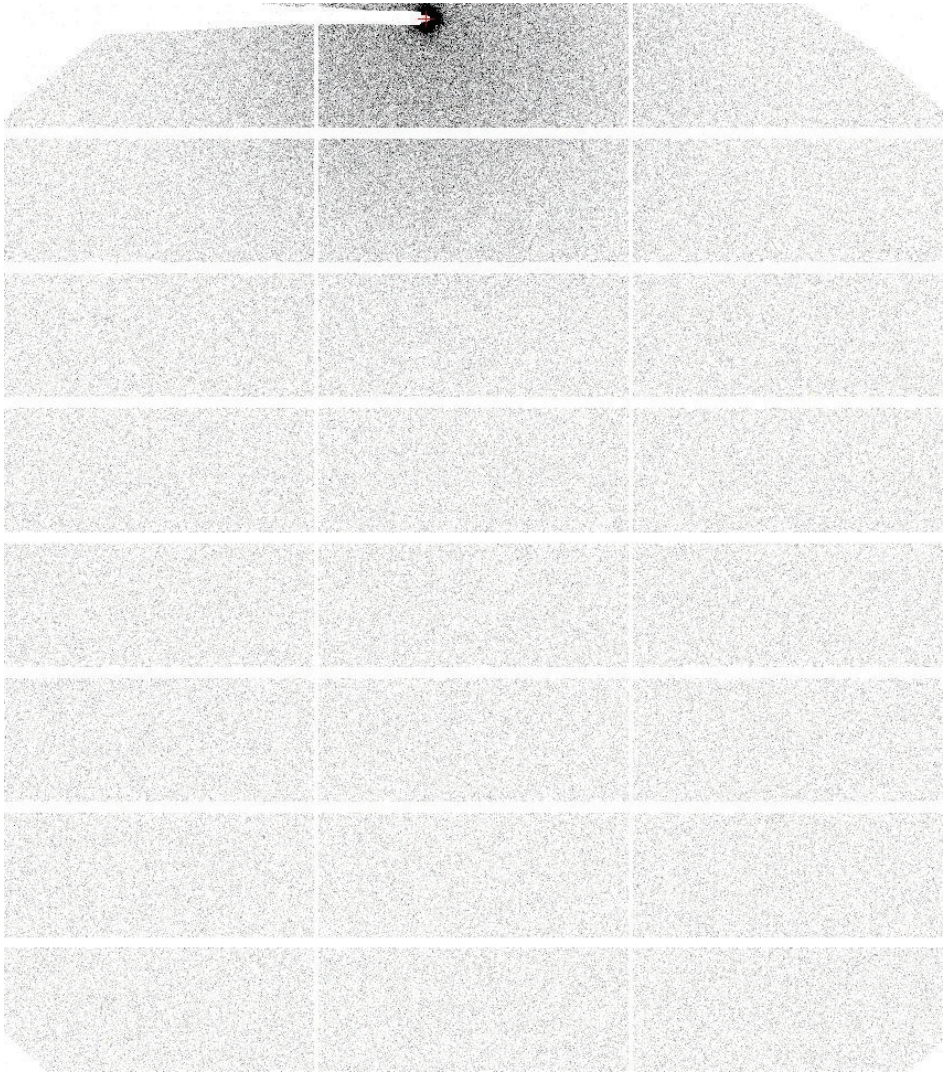
Know what you do not know and
what you want to learn from SAXS!

Small Angle X-Ray Scattering Pattern



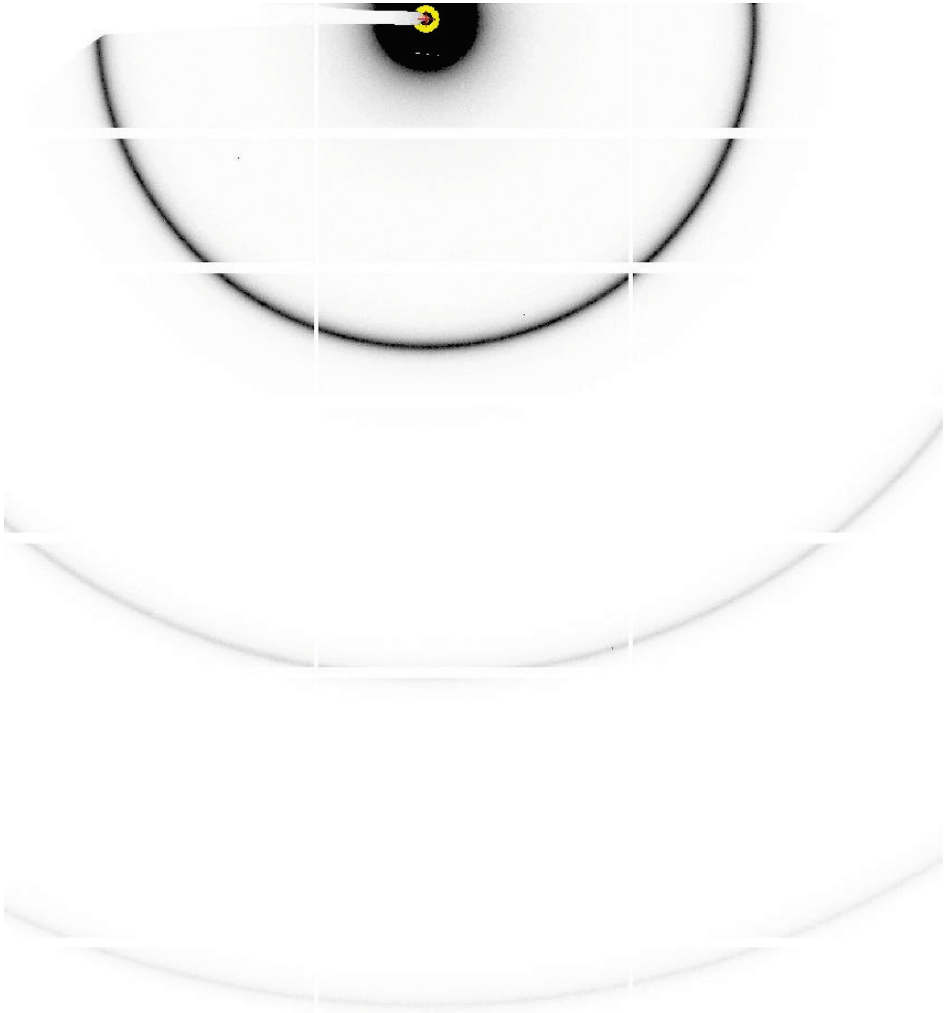
BSA Buffer (HEPES)

Small Angle X-Ray Scattering Pattern



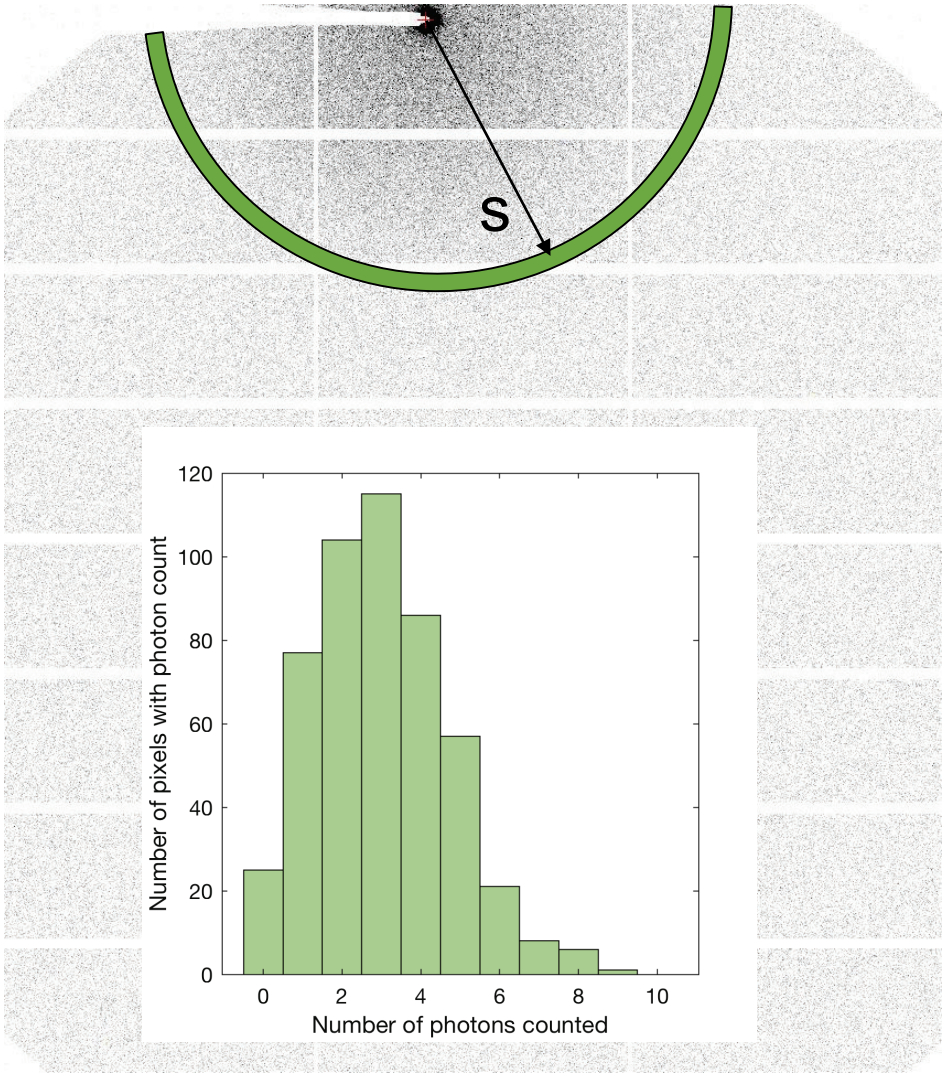
BSA at 2.3 mg/ml

Small Angle X-Ray Scattering Pattern



AgBh Standard

Basics of Counting Statistics



$$I(s) = \frac{1}{n}\sum_{k}^{n} c_k$$

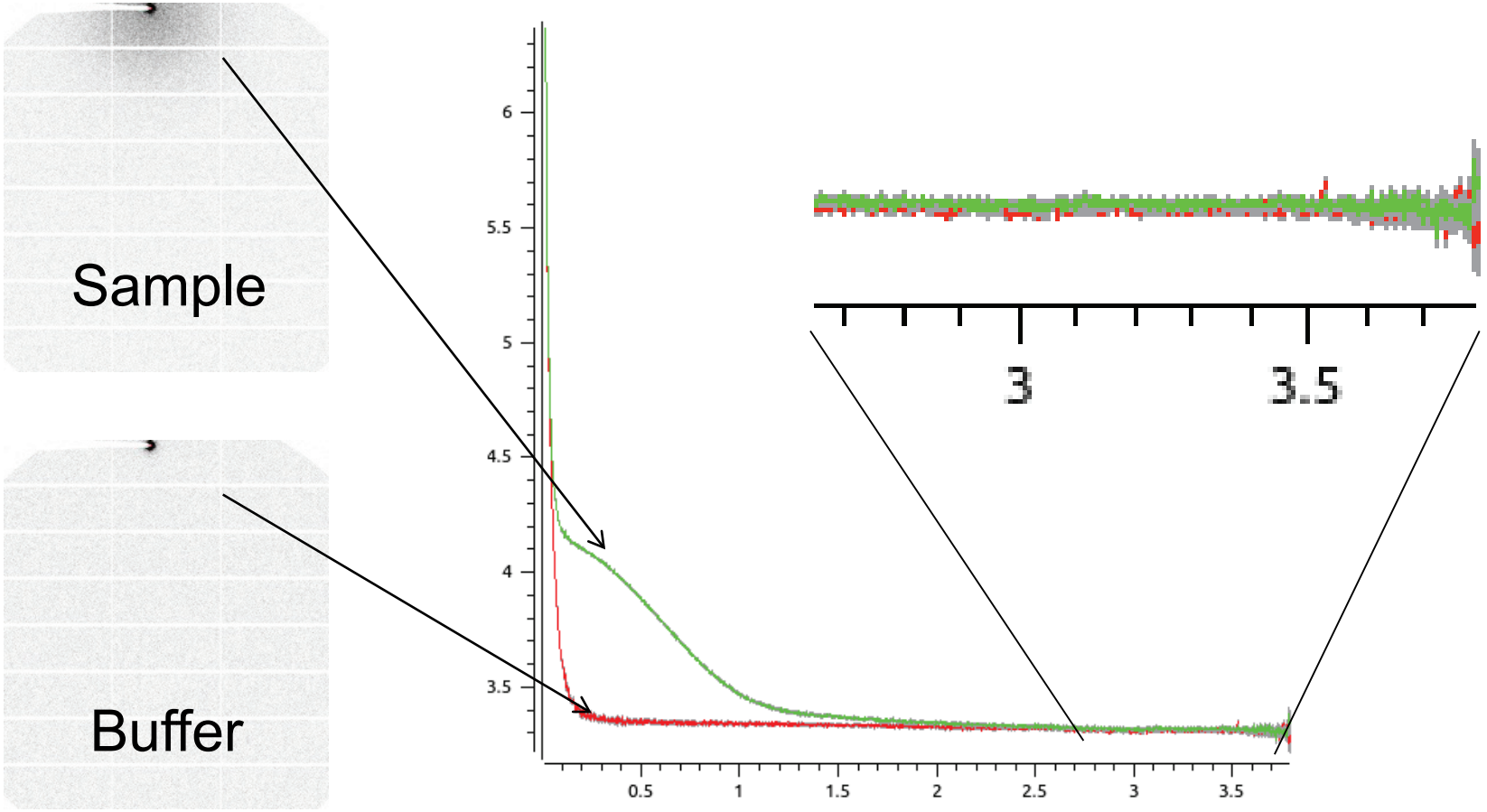
$$SE\big(I(s)\big) = \sqrt{I(s)/n}$$

Central Limit Theorem: the sampling distribution of the mean of independent and identically distributed random variables I(s) will follow a normal distribution.

Note on outlier detection: use Anscombe Transform and modified Z-score

EMBL

Radial Averaging

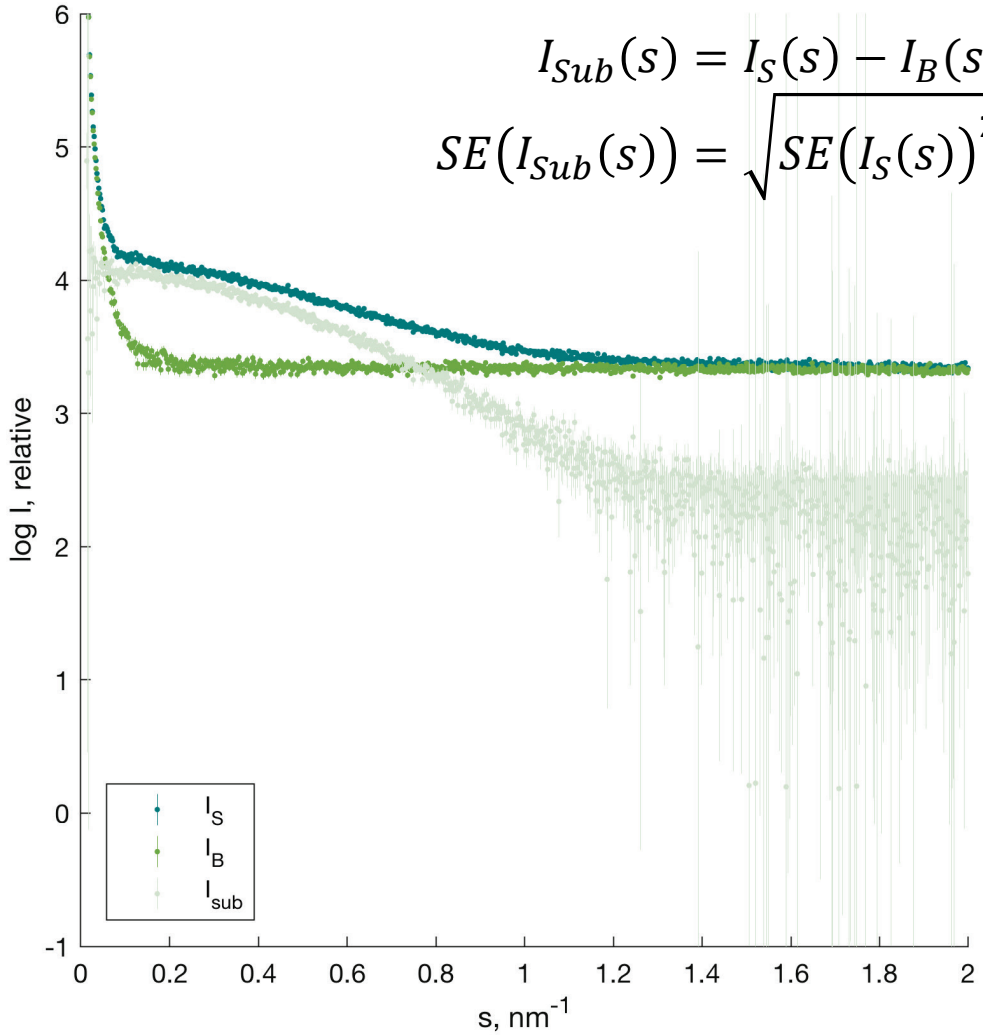


EMBL

Propagation of Estimated Errors

$$I_{Sub}(s) = I_S(s) - I_B(s)$$

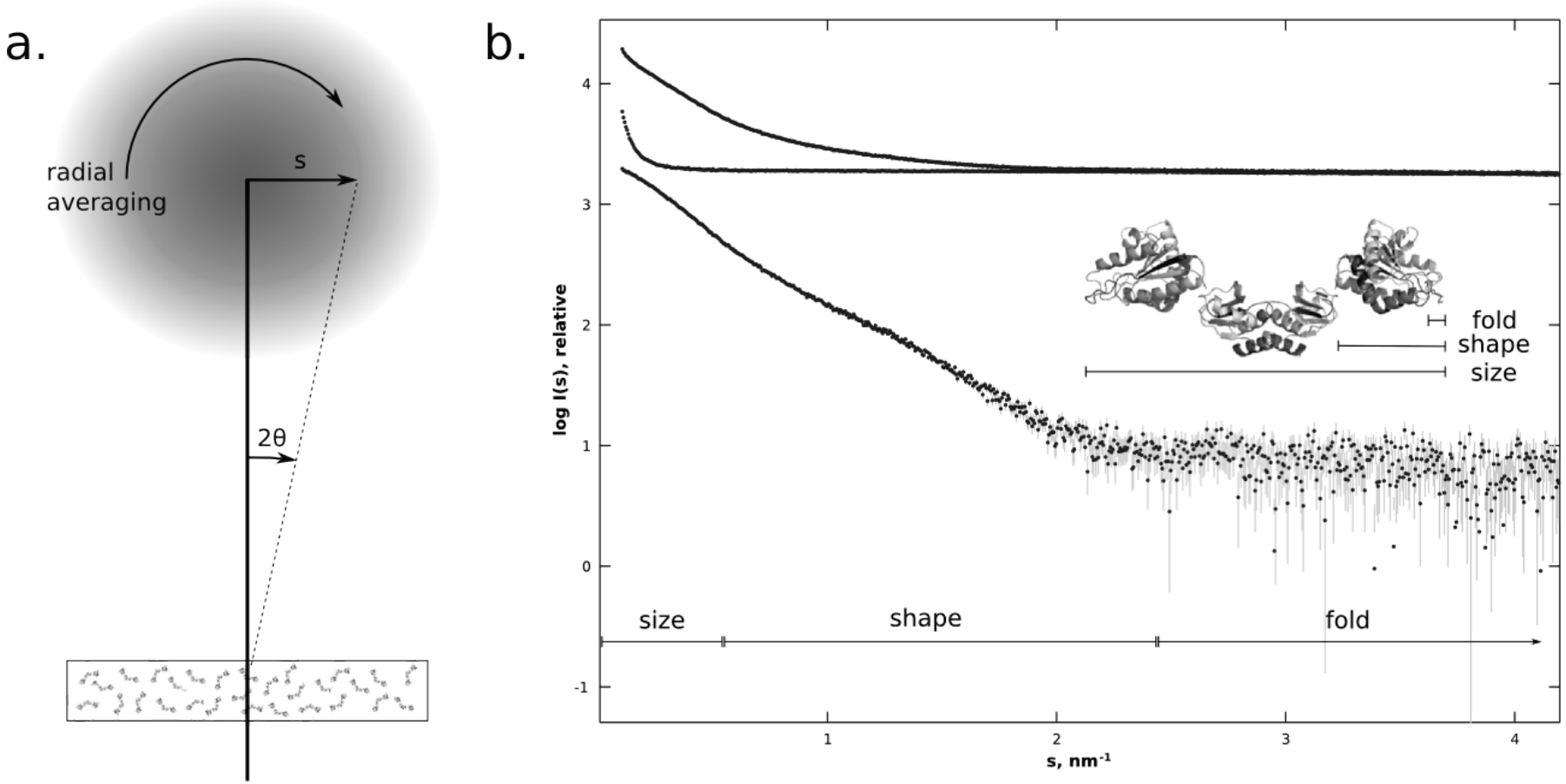
$$SE\left(I_{Sub}(s)\right) = \sqrt{SE\left(I_S(s)\right)^2 + SE\left(I_B(s)\right)^2}$$



Correlations are generally not considered and usually assumed to not exist!

EMBL

Small Angle X-Ray Scattering

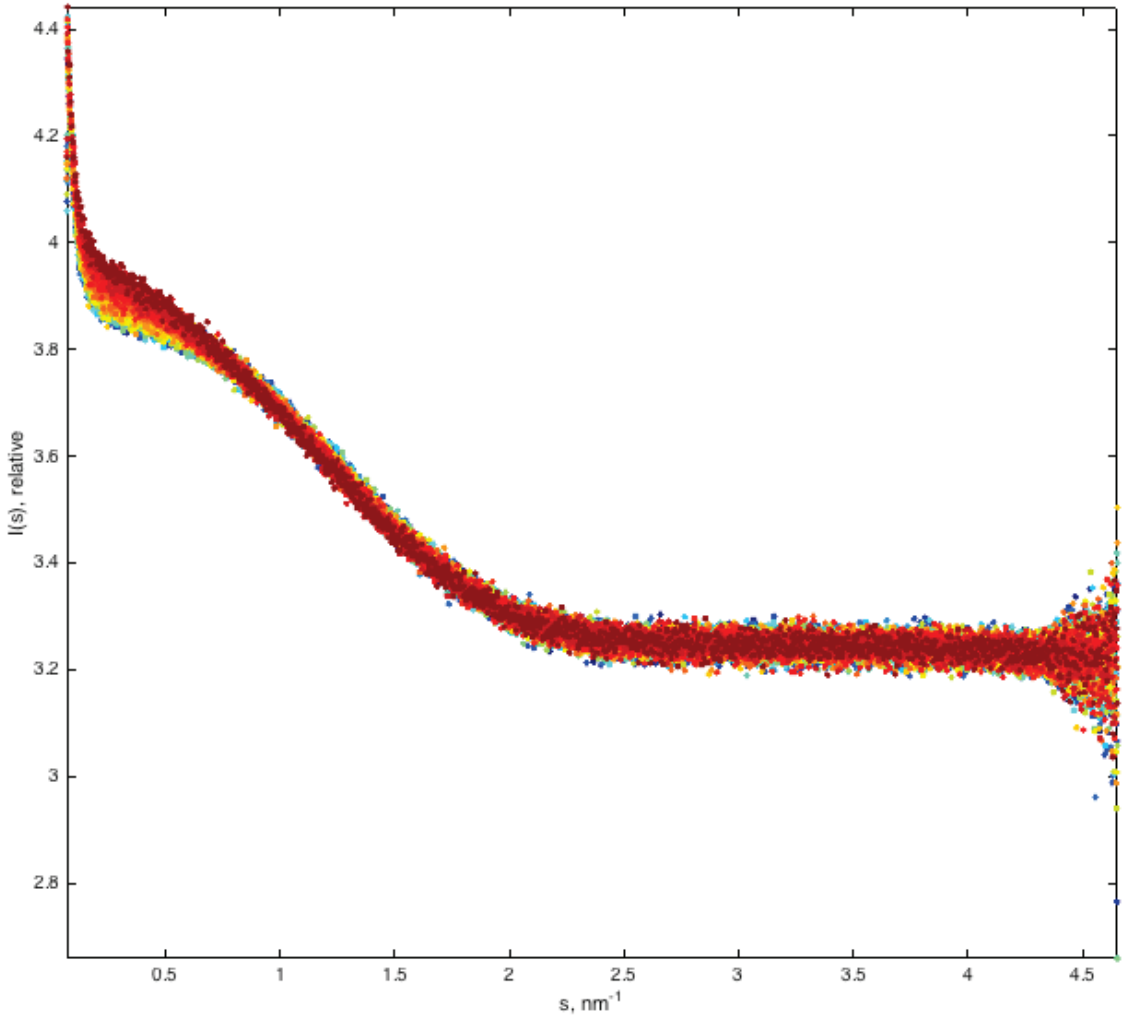


SASBDB: SASDAM5

EMBL

What could possibly go wrong?

- absorbance
- fluorescence
- radiation damage
- inter particle effects
- residue buildup



Radiation damage, 20 frames of Lysozyme

EMBL

Contingency Plans

- Radiation Damage: multiple frames per data collection
- Systematic Errors: repeat buffers
- Interparticle Effects: multiple concentrations of samples

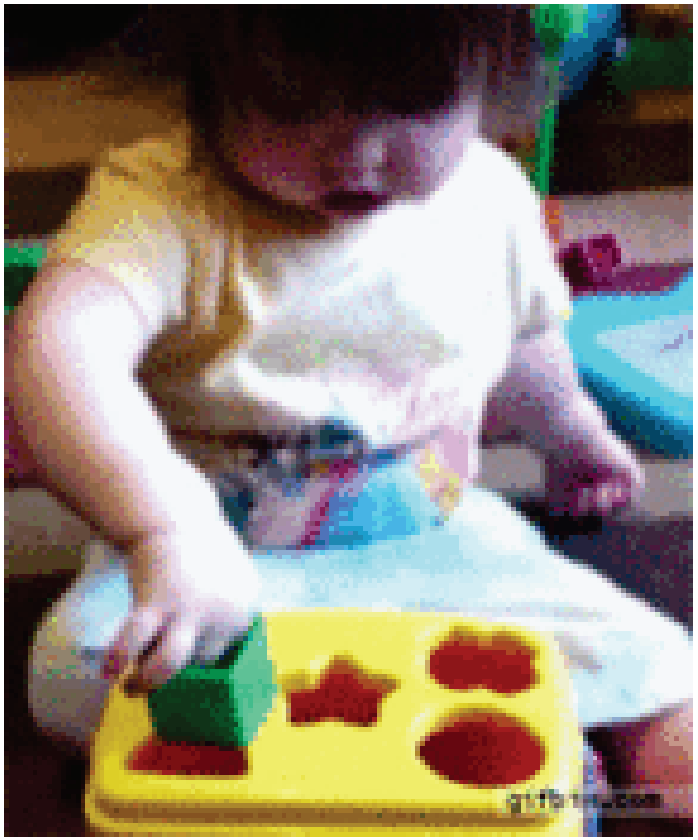
Data Comparison

- compare empty capillaries to monitor residue buildup
- compare frames of a single collection to monitor radiation damage
- compare buffers to monitor stability
- compare concentrations to monitor inter particle effects

But also:

- does the model correspond to the experimental data?

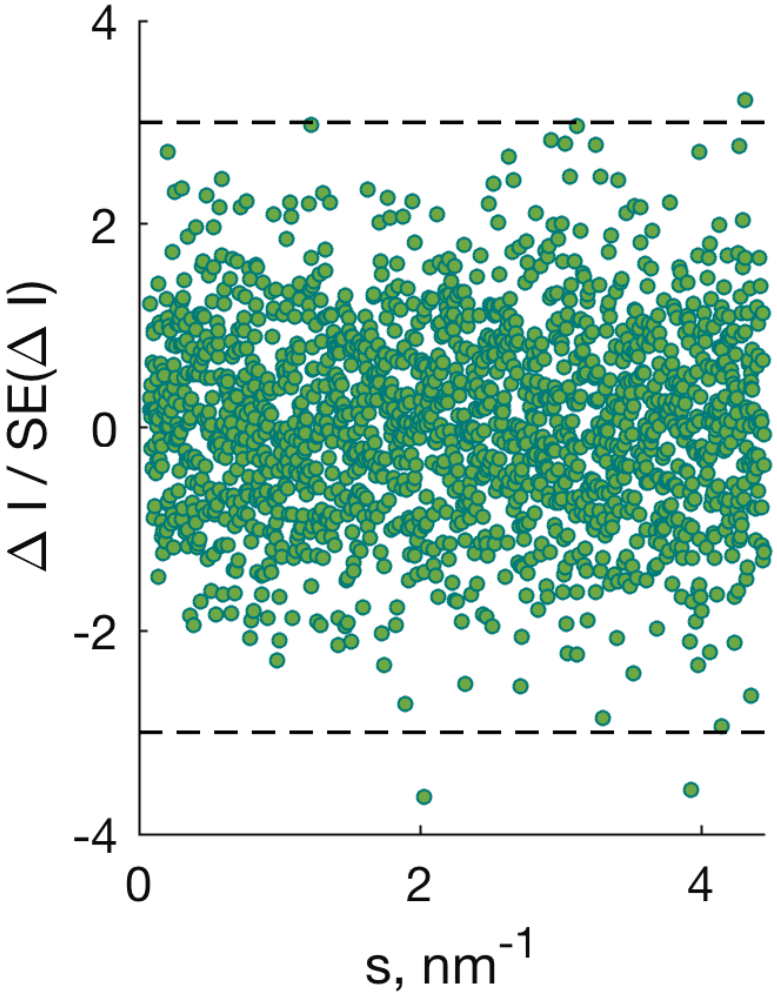
In other words: does it fit?



Outline

- Pairwise Similarity Tests
  - Standardized Residuals
  - Reduced χ^2 test
  - CorMap test
  - Anderson-Darling test
- Multiple Testing
- If errors are available
  - Utility of *correct* error estimates
  - How to validate error estimates of raw data
  - Requirements of error propagation
  - Implications to buffer subtraction

Goodness of Fit: Standardized Residuals



Data Data or Model

$$\tilde{r}(s) = \frac{I_1(s) - I_2(s)}{\sqrt{SE\left(I_1(s)\right)^2 + SE\left(I_2(s)\right)^2}}$$

Standardized Residuals:

- Standard Normal Distribution
- Centered on zero
- Symmetric in the range of [-3; +3]

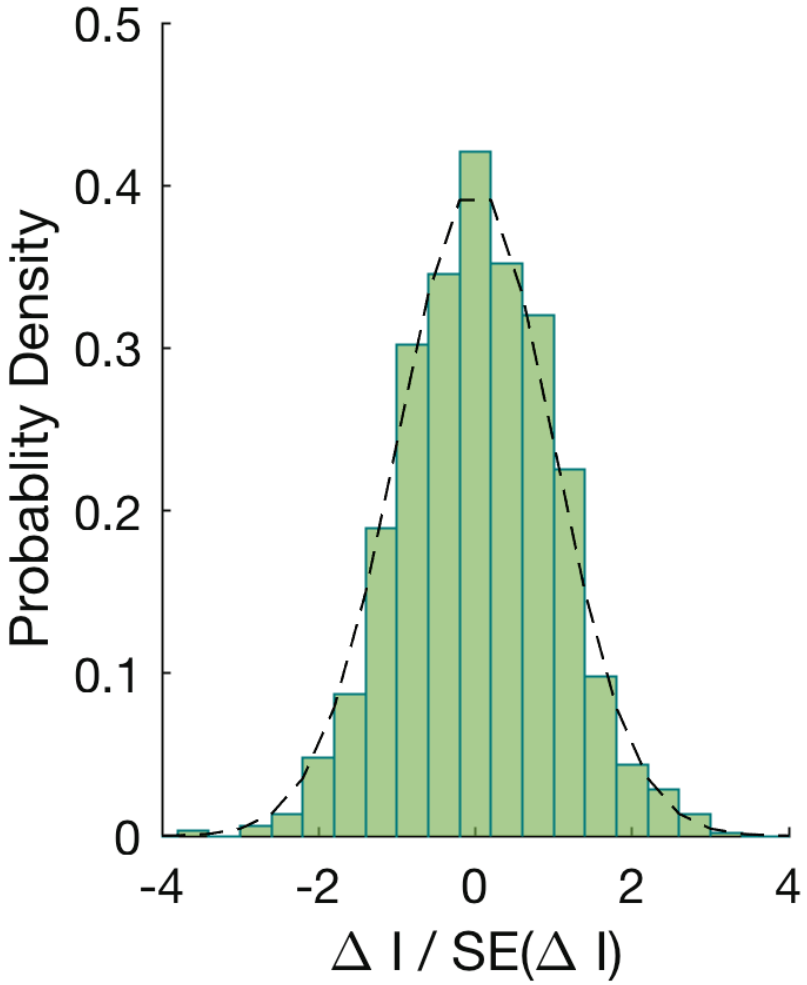
Standardizes Residual Plot

- Visual assessment of randomness

See also: Trewhella et al. (2017): Publication Guidelines

EMBL

Goodness of Fit: Residual Histogram



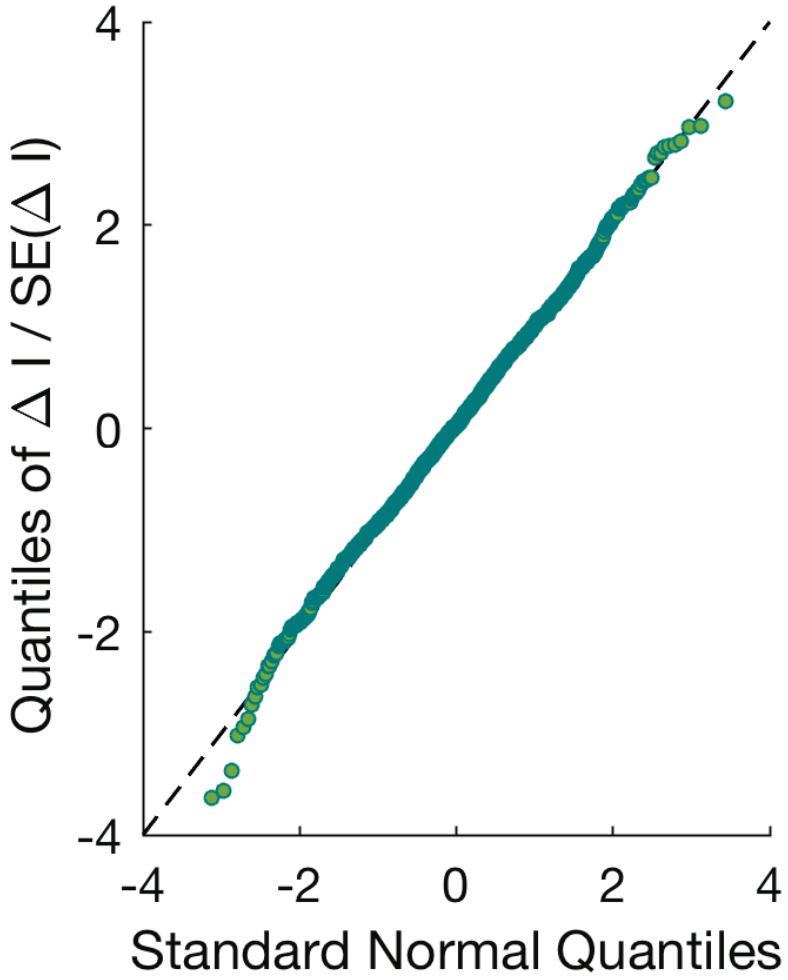
Histogram of standardized residuals:

- Visual assessment of standard normal distribution

- Centered on zero
- Symmetric in the range of [-3; +3]

EMBL

Goodness of Fit: Normal-Probability Plot



Normal Probability plot of standardized residuals:

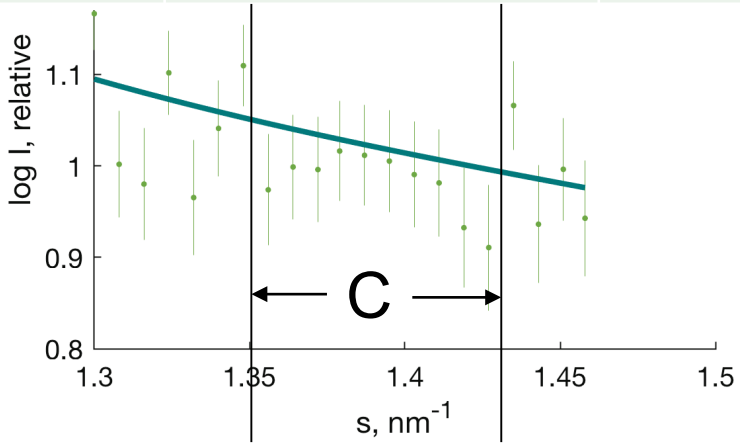
- Visual assessment of randomness and normal distribution

- Non-zero offset: scale error
- Non-unity slope: errors over/under estimated
- Not linear: systematic deviations

EMBL

Statistical Tests

	CorMap Test	Red. χ^2 Test	A-D Test
Purpose	Randomness of residuals	Randomness of standardized residual	Standard normal distribution of residuals
Uses	Sign of residuals	Standardized residuals	Ordered standardized residuals
Properties	Indicates location of non-fit	Requires accurate error estimates	
Test value	C	χ^2	A^2

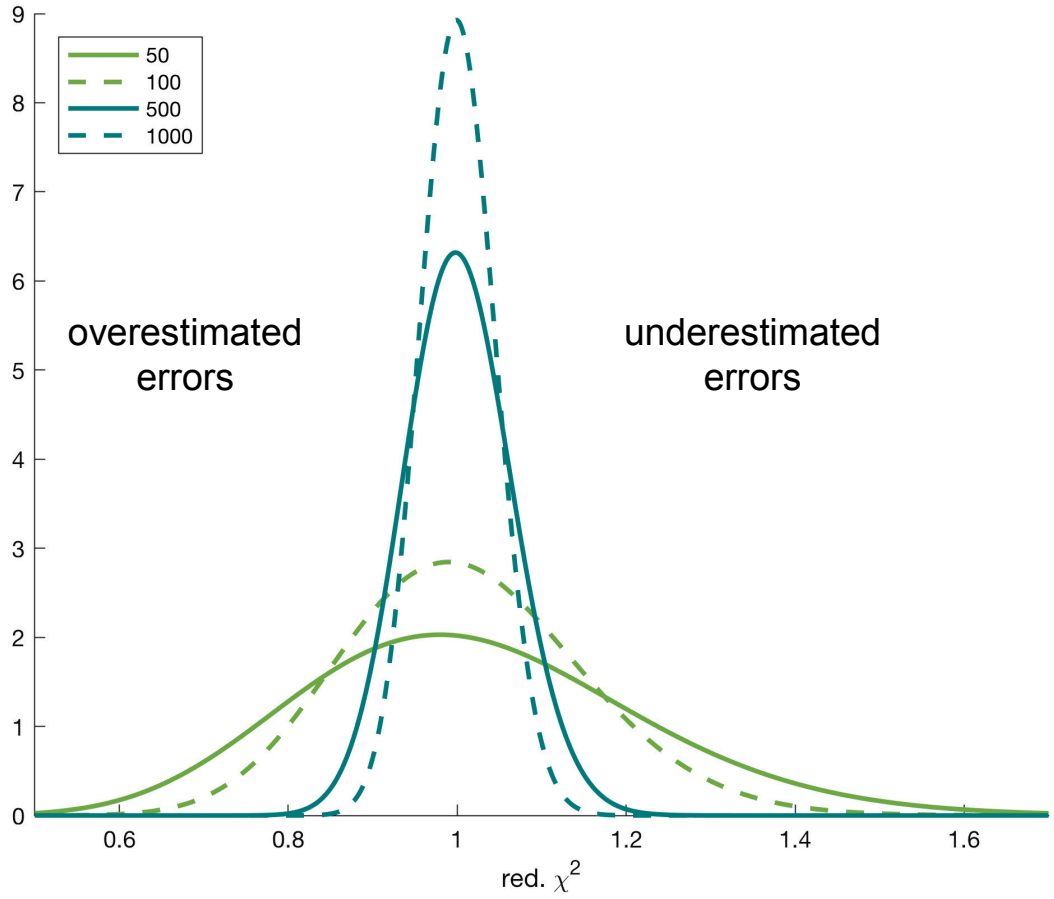


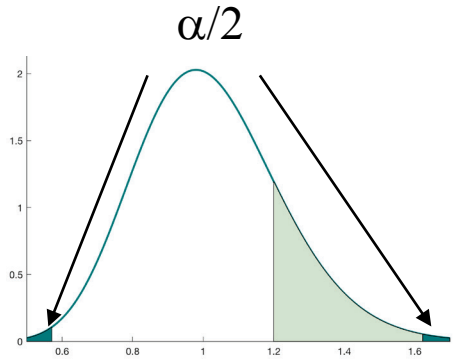
ATSAS: datcmp and primus/qt

EMBL

Caveat of the reduced χ^2 test

$$\chi^2_{n-1} = \frac{1}{n-1}\sum_{k=1}^{n}\left(\frac{I_1(s_k) - I_2(s_k)}{\sqrt{SE\left(I_1(s_k)\right)^2 + SE\left(I_2(s_k)\right)^2}}\right)^2$$





1%

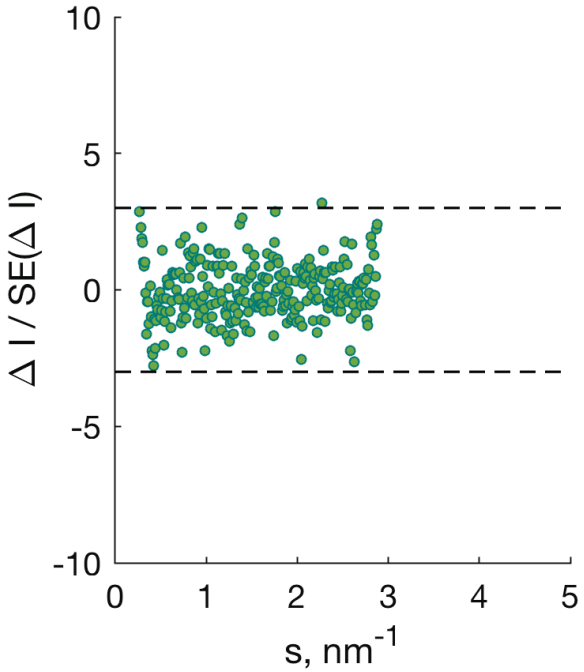
n	p=α/2	p=1-α/2
50	0.57	1.62
100	0.68	1.41
500	0.85	1.17
1000	0.89	1.12
2000	0.92	1.08
5000	0.95	1.05

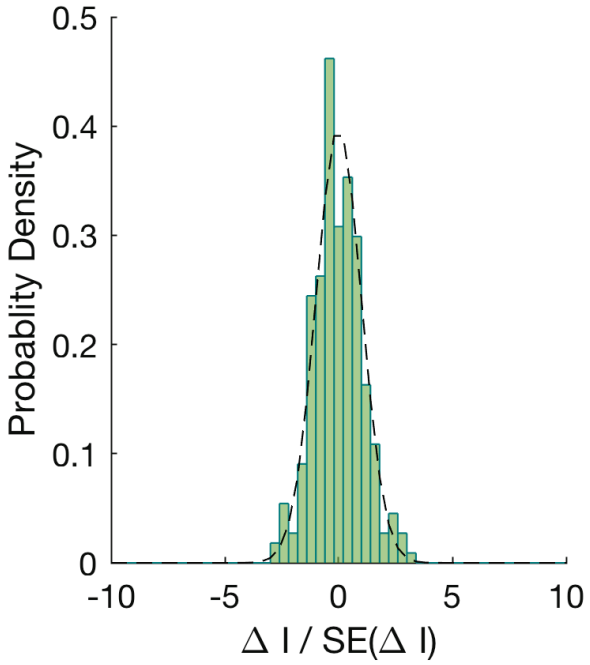
Matlab: chi2inv
Python: scipy.stats.chi2.ppf

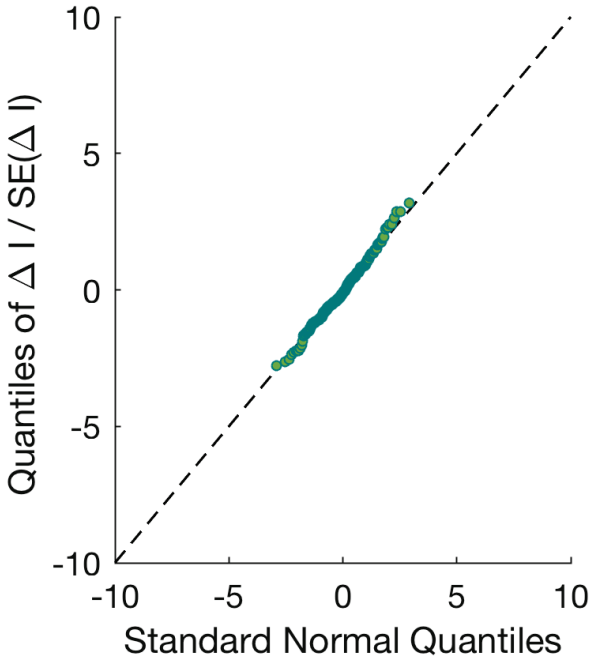
EMBL

Examples

SASBDB
Small Angle Scattering Biological Data Bank







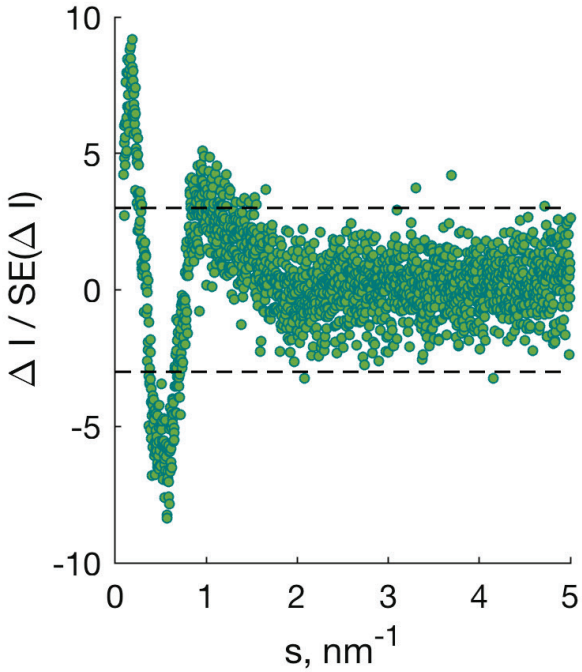
N=276	CorMap Test	Red. χ^2 Test	A-D Test
Test Value	C = 13	χ^2 = 1.105	A^2 = 0.785
p-value	P(>C) = 0.032	P(>χ^2) = 0.120	P(>A^2) = 0.492

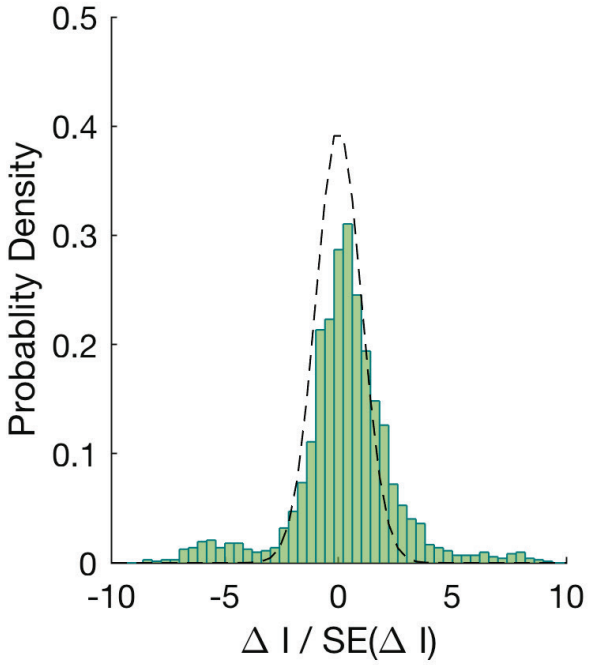
(ab initio model fit)

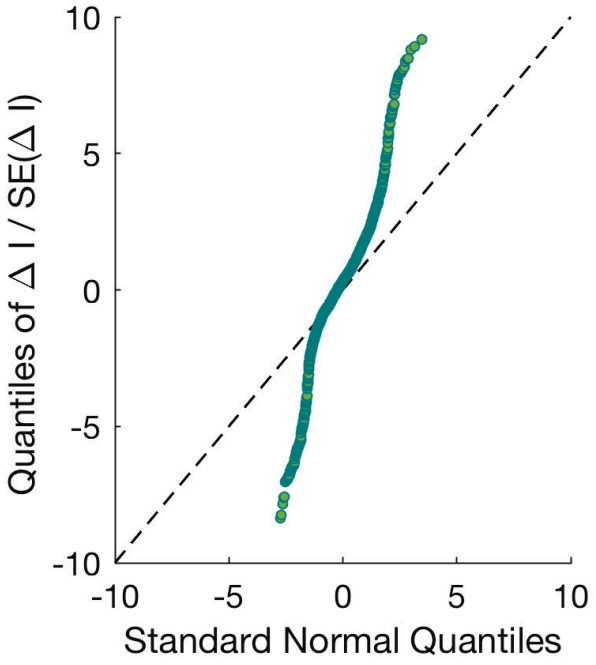
EMBL

Examples

SASBDB
Small Angle Scattering Biological Data Bank







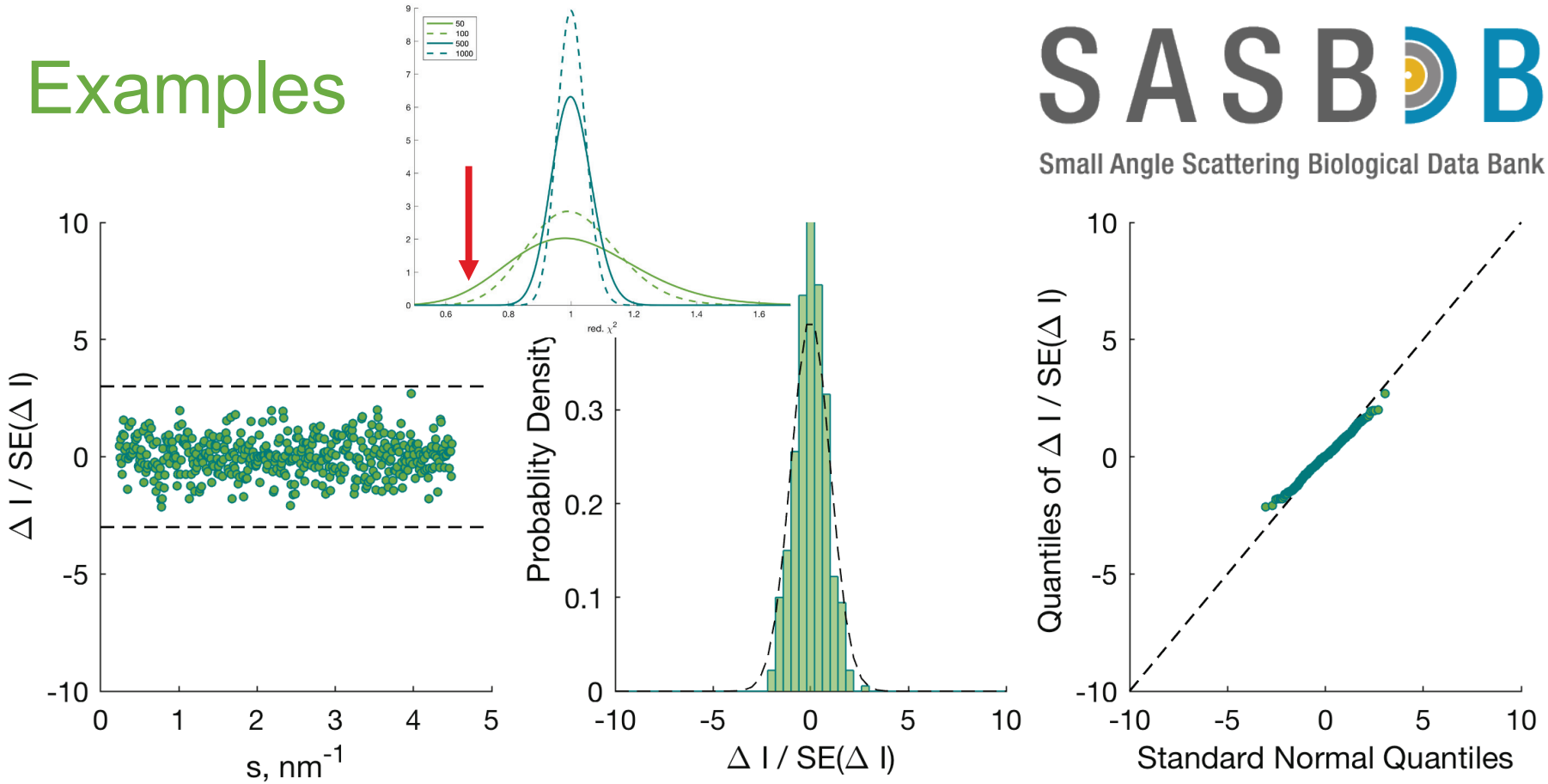
N=1799	CorMap Test	Red. χ^2 Test	A-D Test
Test Value	C = 147	χ^2 = 5.112	A^2 = 307.8
p-value	P(>C) = 0.000	P(>χ^2) = 0.000	P(>A^2) = 0.000

(rigid body model fit)

EMBL

Examples

SASBDB
Small Angle Scattering Biological Data Bank



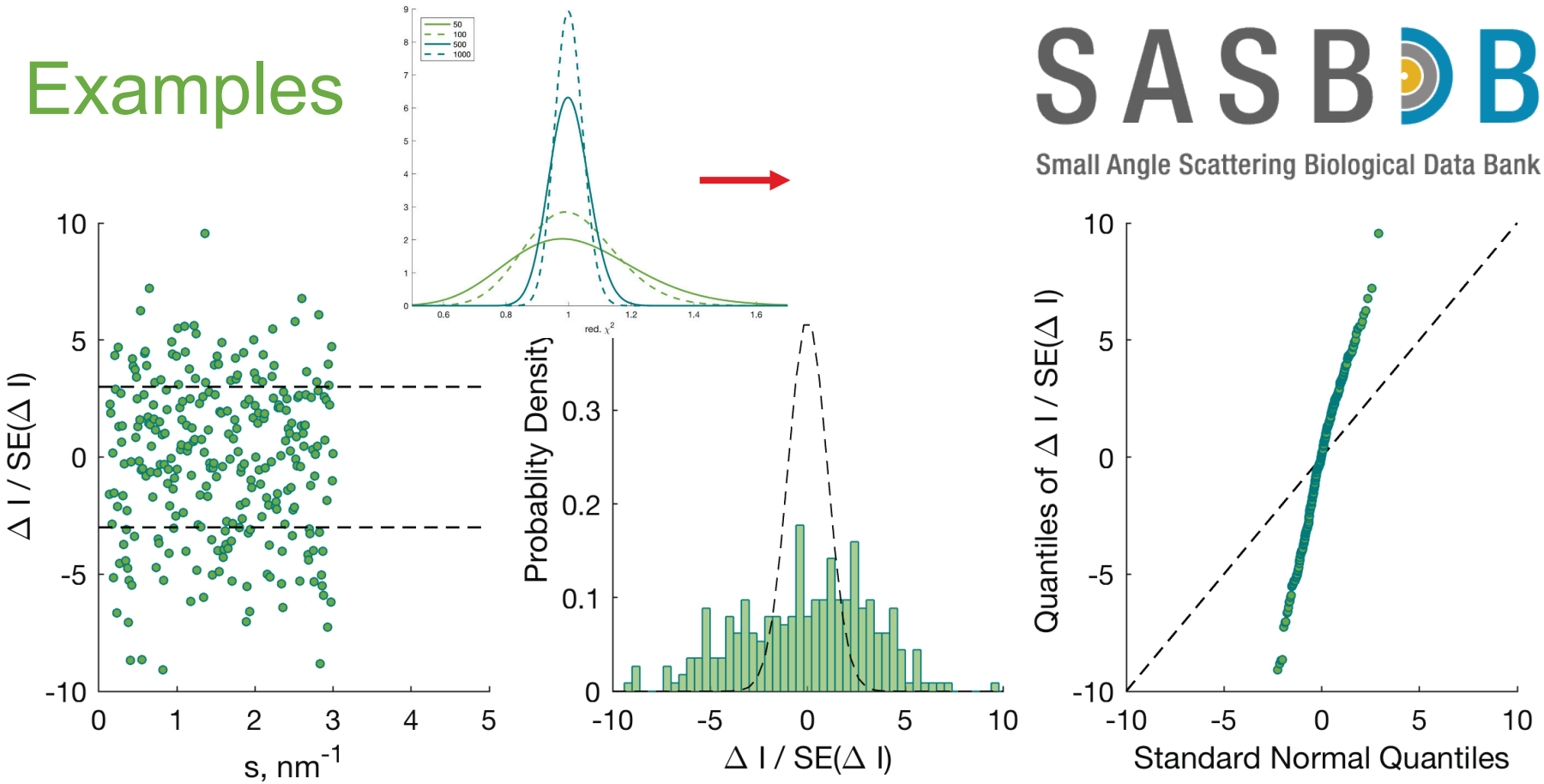
N=450	CorMap Test	Red. χ^2 Test	A-D Test
Test Value	C = 13	χ^2 = 0.636	A^2 = 5.963
p-value	P(>C) = 0.052	P(>χ^2) = 1.000	P(>A^2) = 0.001

(ab initio model fit)

EMBL

Examples

SASBDB
Small Angle Scattering Biological Data Bank



N=282	CorMap Test	Red. χ^2 Test	A-D Test
Test Value	C = 12	χ^2 = 12.58	A^2 = 303.75
p-value	P(>C) = 0.064	P(>χ^2) = 0.000	P(>A^2) = 0.000

(ab initio model fit)

EMBL

Outline

- Pairwise Similarity Tests
  - Standardized Residuals
  - Reduced χ^2 test
  - Cormap test
  - Anderson-Darling test
- Multiple Testing
- If errors are available
  - Utility of *correct* error estimates
  - How to validate error estimates of raw data
  - Requirements of error propagation
  - Implications to buffer subtraction

Journal of Serendipitous and Unexpected Results

Neural Correlates of Interspecies Perspective Taking in the Post-Mortem Atlantic Salmon: An Argument For Proper Multiple Comparisons Correction

Craig M. Bennett [1*], Abigail A. Baird [2], Michael B. Miller [1] and George L. Wolford [3]

"[…] we completed an fMRI scanning session with a post-mortem Atlantic Salmon as the subject. The salmon was shown the same social perspective taking task that was later administered to a group of human subjects. Statistics that were uncorrected for multiple comparisons showed active voxel clusters in the salmon's brain cavity and spinal column."

Multiple Comparisons

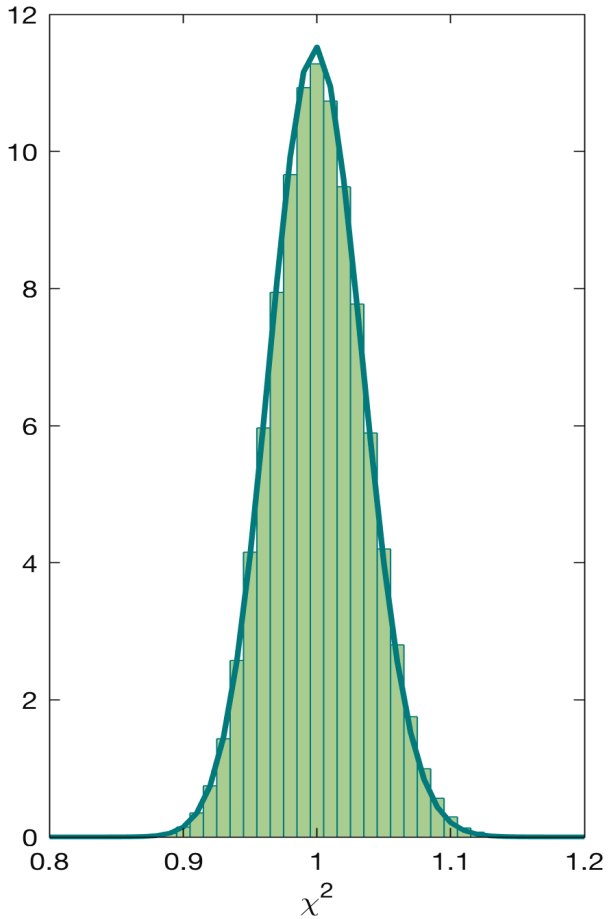
The probability of obtaining at least one false positive in *n* (independent) tests:

$$P(\text{false positive}) = 1 - (1 - \alpha)^n$$

Let α = 0.01 and *n* = 190, then the probability for a false positive is about 85%.

Translation: with 20 frames per data collection, it is almost guaranteed that at least one frame gets rejected as different – but it may not be!

Adjusting for Multiple Testing



1,390 water frames
965,355 pairwise comparisons

Single Test: $\alpha = 0.01$

p	crit n=1682	p x 965,355	# < crit
$\alpha/2$	0.9140	4,827	4,992
1-$\alpha/2$	1.0917	960,530	958,818

Pair-wise Tests: $\alpha = 0.01/965,355$

p	crit n=1682	p x 965,355	# < crit
$\alpha/2$	0.8155	0.005	0
1-$\alpha/2$	1.2109	965,354.995	965,355

(Bonferroni Correction)

EMBL

Outline

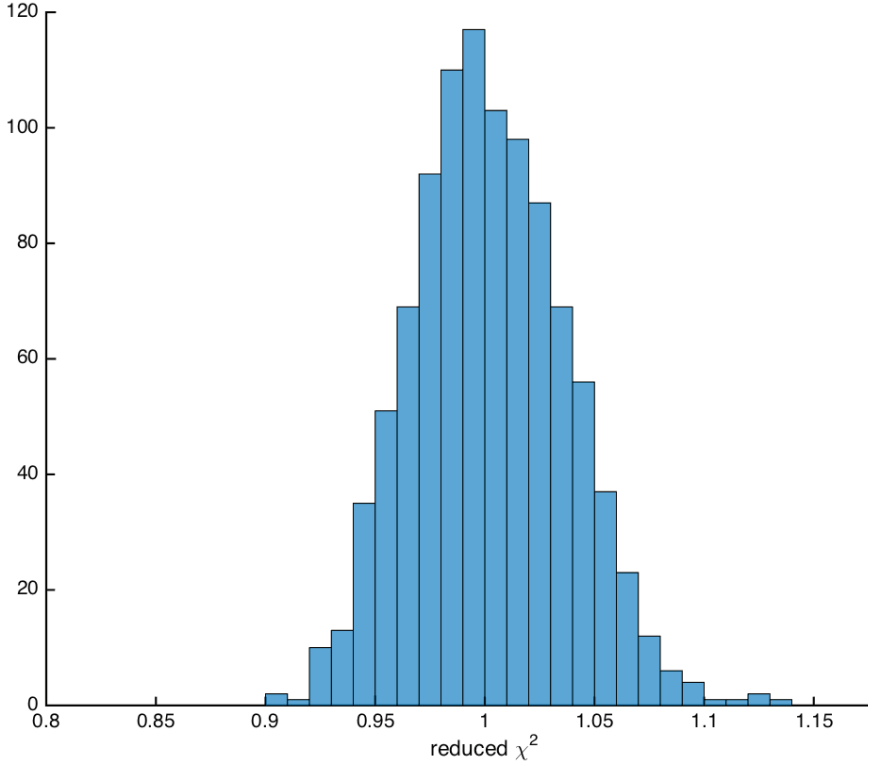
- Pairwise Similarity Tests
  - Standardized Residuals
  - Reduced χ^2 test
  - Cormap test
  - Anderson-Darling test
- Multiple Testing
- If errors are available
  - Utility of *correct* error estimates
  - How to validate error estimates of raw data
  - Requirements of error propagation and implications to buffer subtraction

Utility of *correct* error estimates

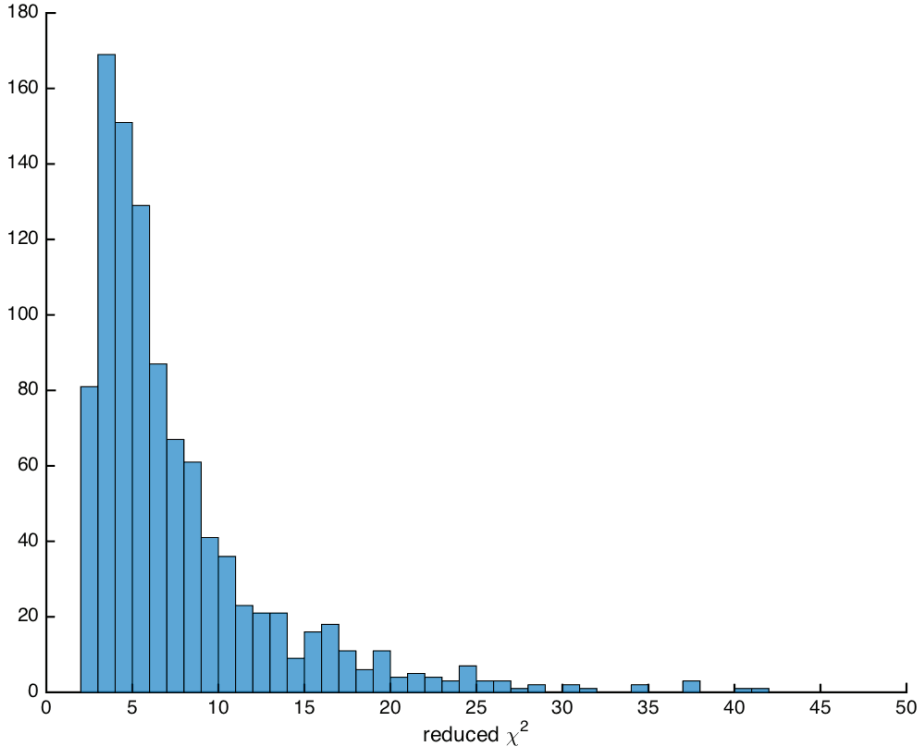
- Reduced χ^2 test of similarity
- Comparability of χ^2 values
- Weighted Least Squares
- Minimize χ^2 during modeling
- Standardized residuals

Validity of Reduced χ^2 Test

1.000 comparisons of simulated BSA without systematic differences



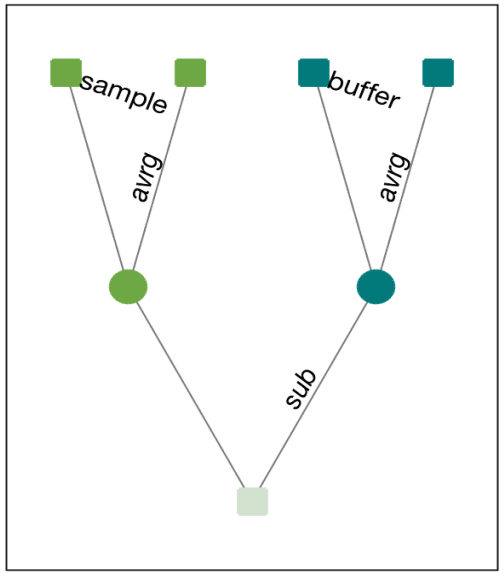
correct error estimates, valid results

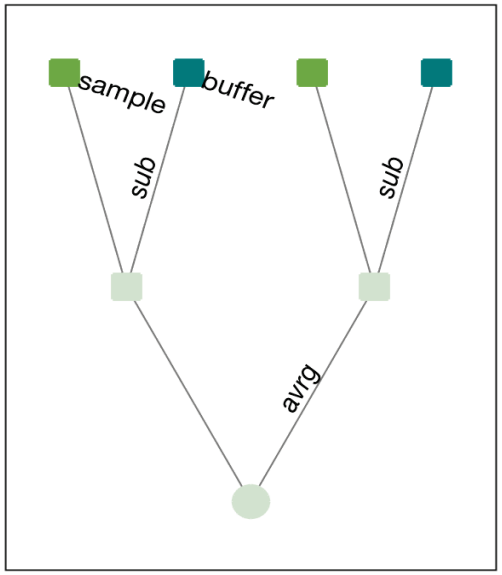


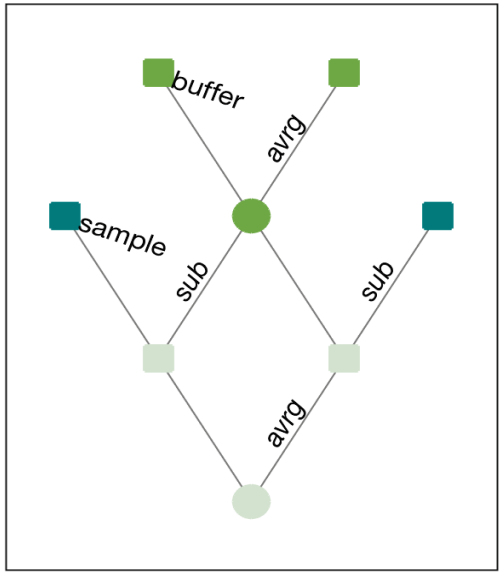
incorrect error estimates, invalid results

EMBL

Validating Error Propagation







$$\frac{S_1 + S_2}{2} - \frac{B_1 + B_2}{2}$$

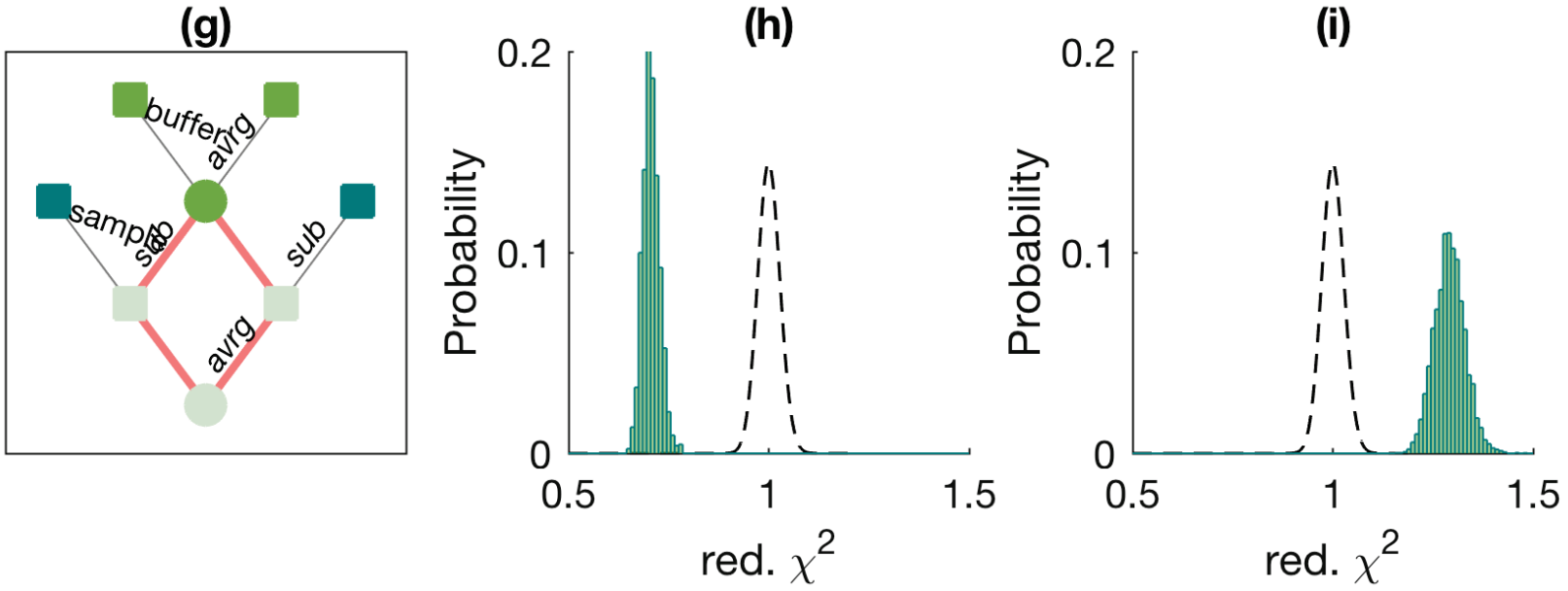
$$\frac{(S_1 - B_1) + (S_2 - B_2)}{2}$$

$$\frac{\left(S_1 - \frac{B_1 + B_2}{2}\right) + \left(S_2 - \frac{B_1 + B_2}{2}\right)}{2}$$

Intensities: associative (ab)c == a(bc) and commutative (ab == ba)

EMBL

Error Propagation: Order Matters!



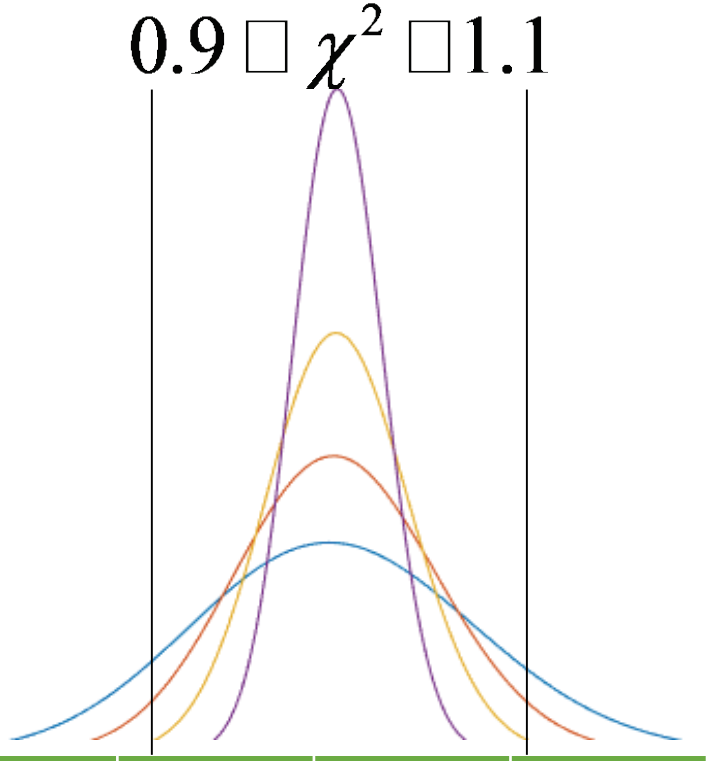
$$SE\left(I_{sub}(s)\right) = \sqrt{SE\left(I_1(s)\right)^2 + SE\left(I_2(s)\right)^2 - 2 \cdot COV\left(I_1(s), I_2(s)\right)}$$

Avoid Correlations by not re-using anything twice!

Error propagation is NOT associative

EMBL

Verifying Error Estimates of Raw Data



m	1 not in range	2 not in range	3 not in range
500	11.3%	1.2%	0.1%
1000	2.5%	0.1%	<0.1%
2000	0.2%	<0.1%	<0.1%
5000	<0.1%	<0.1%	<0.1%

- Take >= 6 consecutive frames, e.g. of water
- Compare pairwise with cormap to verify similarity
- Compare with χ^2 frames 1-2, 3-4, 5-6, …
- If all are in [0.9 – 1.1], ok
- If there is one outside range, repeat procedure once
- If there are two or more outside range (counting repeat), error estimates are incorrect

EMBL

Error Propagation

- General purpose tools assume independent data
- If your data are not independent, it is still possible, but much harder (custom solution necessary)!

- Implications:
  - May need as many buffer frames as there are sample frames
  - data collection on lab sources: who wants to measure blanks for the same amount of time as samples?
  - SEC-SAXS: buffer regions are often hard to define and there may not be sufficient buffer frames for all peaks

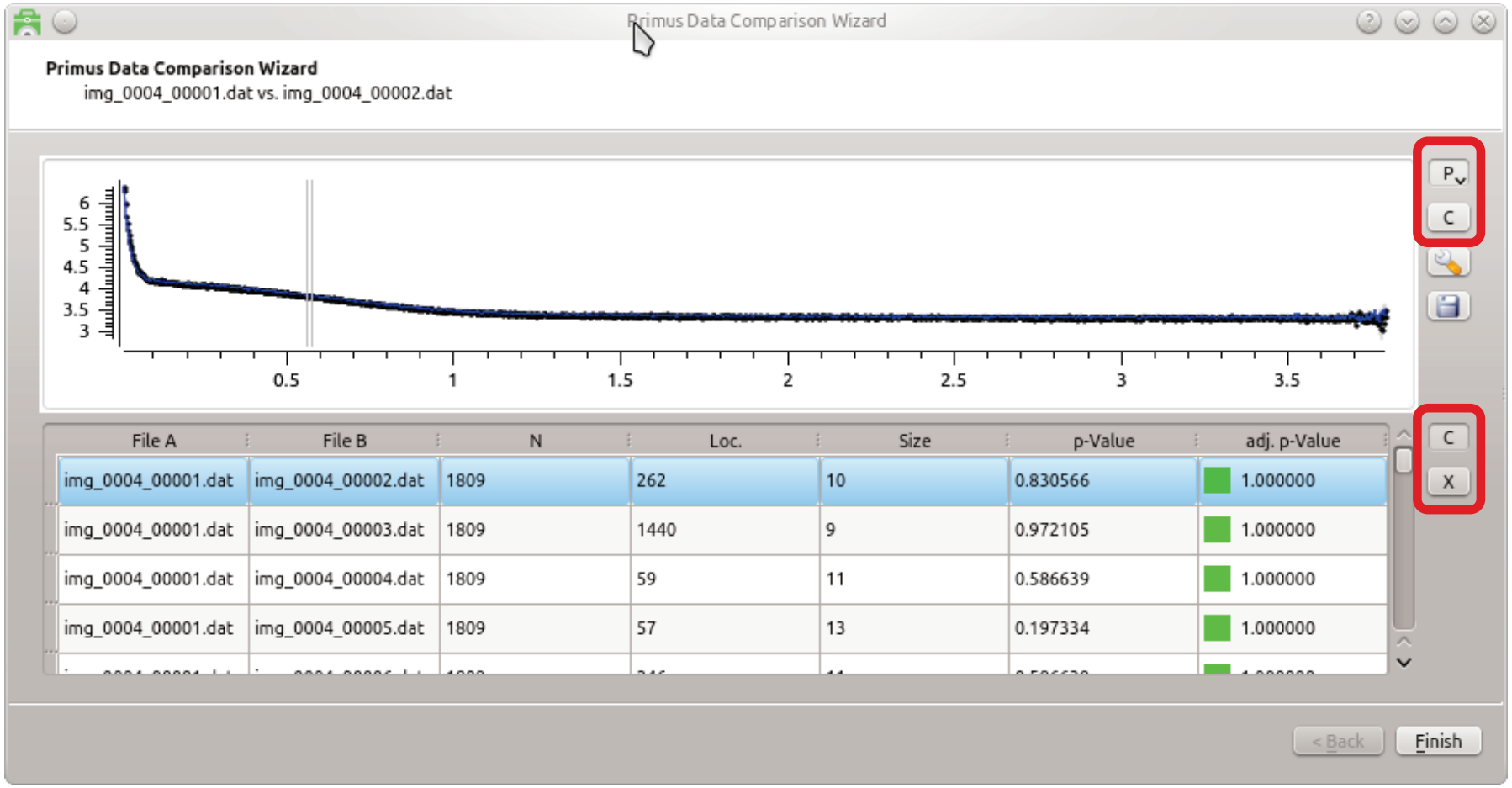
ATSAS

- `datcmp --test=cormap [...]`
- `datcmp --test=chi-square [...]`

- `datop add|sub|mul|div [...] --output [...]`
- `dataver [...] --output [...]`
- `datmerge [...] --output [...]`

- `primus/qt`

Manual processing: Comparing Frames



EMBL

Questions?