

Size distribution analysis with Peter Schuck's SEDFIT



1. least-squares $g(s)$
2. high resolution $c(s)$ method

Least-squares $g^*(s)$ from SEDFIT

1. Assume no diffusion, fitting functions for each species are infinitely sharp boundaries (step functions)
2. Fit to a grid of possible sedimentation coefficients to determine the concentration at that sedimentation coefficient
3. Impose smoothing (regularization) to minimize spikiness and false peaks in the distribution

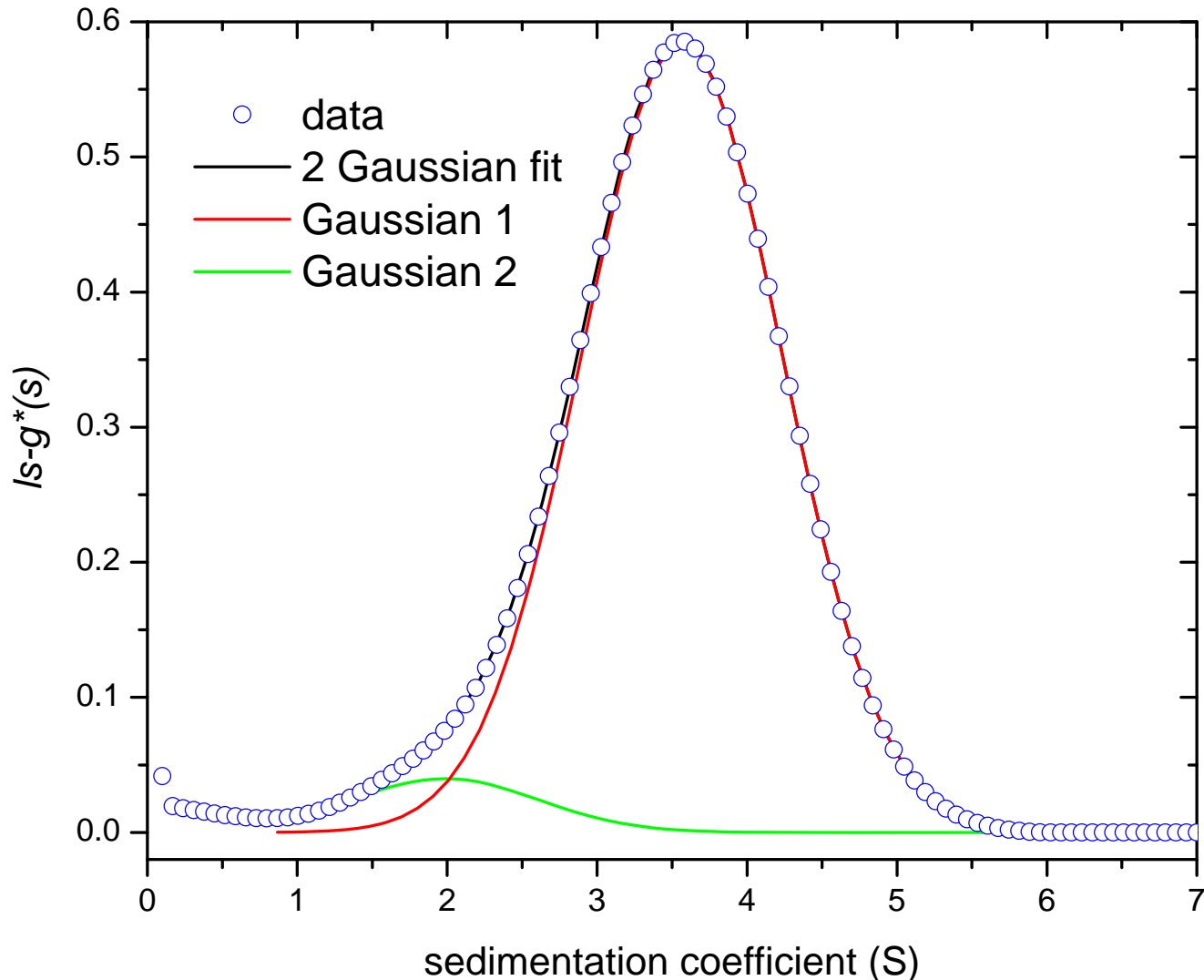
Advantages of $ls-g^*(s)$ over standard dc/dt

1. In principle data from the entire run can be used, greatly increasing the signal/noise
 - ★ Peter Schuck has stopped advocating this, and now recommends using no more than 2-3 times the time span given by the dc/dt 'rule of thumb'
2. A larger time span means a larger range of sedimentation coefficients can be covered in one analysis
3. The peak resolution tends to be somewhat higher, giving each species the resolution it has just before it reaches the cell base

The big problem with $1/s-g^*(s)$ [as I see it]

- ★ The model is fundamentally wrong because it assumes zero diffusion
- ★ Therefore usually the model cannot fit the data well both early in the run (when diffusion has the largest relative effect) and late in the run (when diffusion has the least relative effect)
- ★ Consequently this method can easily produce false peaks, so if you use it, be very careful

For example, even if you stay within the recommended time span to compute the $I_s-g(s)$ for simulated ovalbumin data, you get false peaks



My opinion about $1/s-g(s)$

- ★ Can be useful for highly heterogeneous samples where diffusion is relatively unimportant
 - ☆ viruses and virus-like particles used for vaccines
 - ☆ works for species > 6000 S where $c(s)$ blows up
- ★ overall if you stay away from danger there is little advantage over the standard dc/dt analysis, and none over the improved $g(s^*)$ fitting method

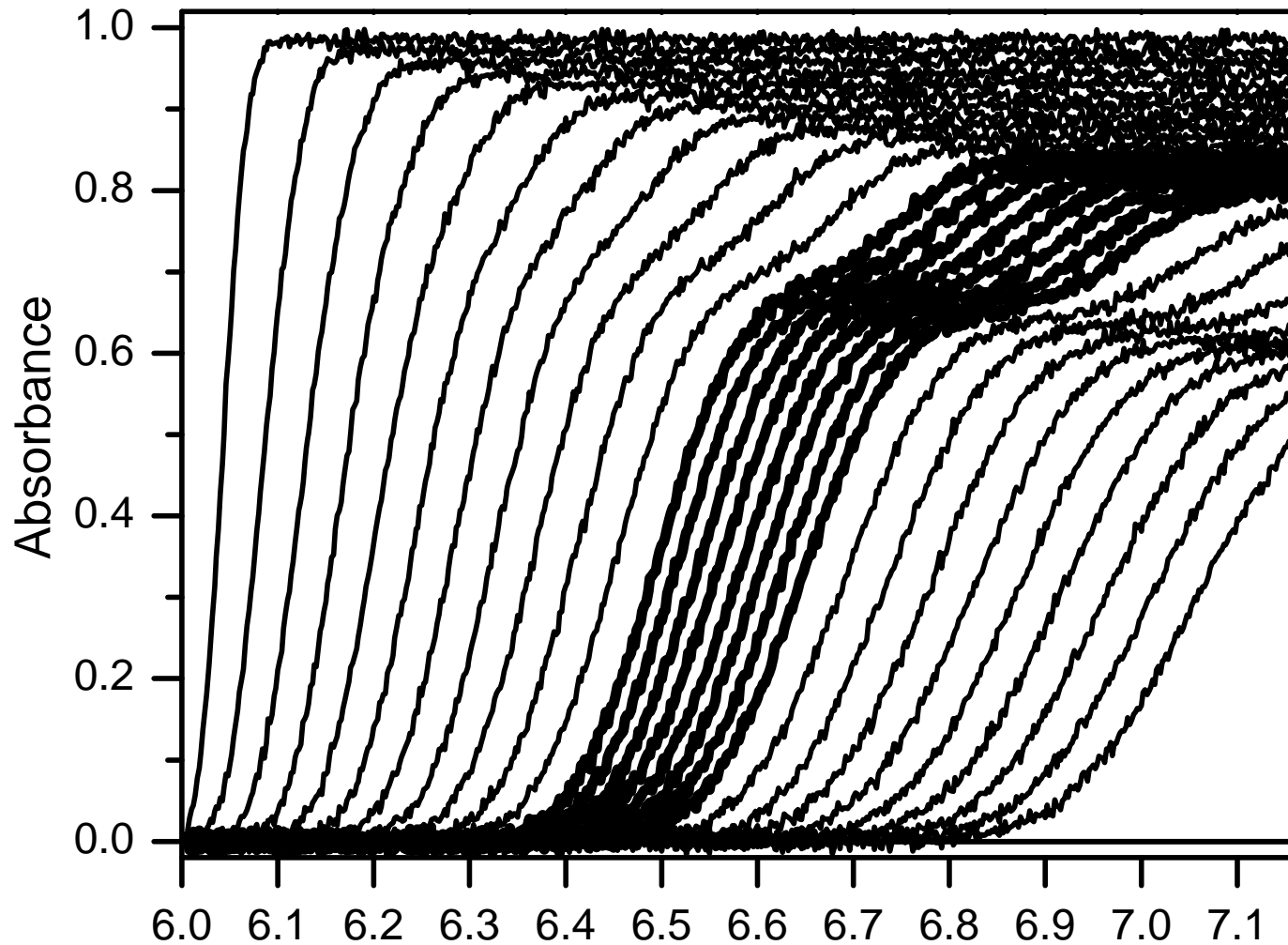
The $c(s)$ method



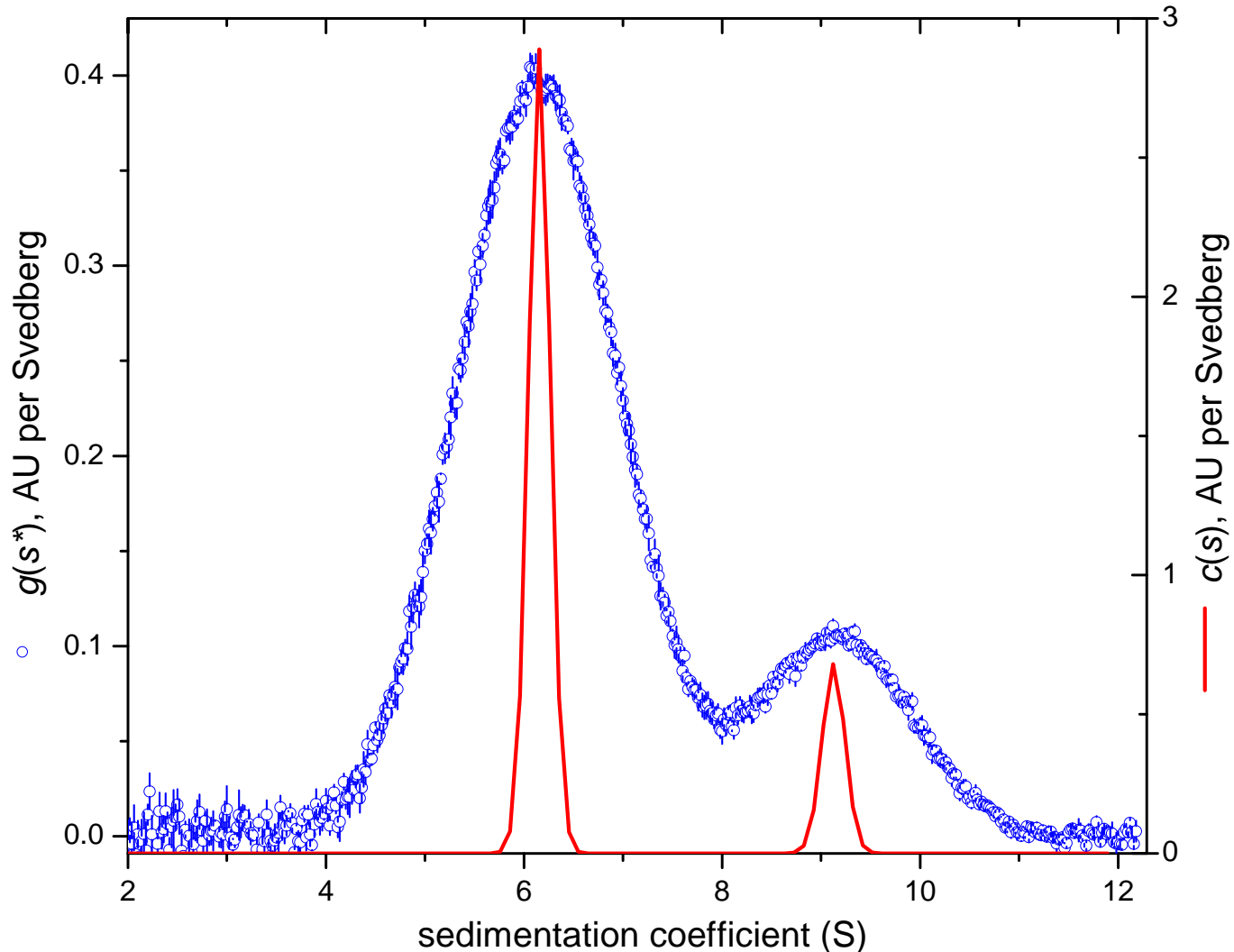
The basic idea behind the $c(s)$ method

- ★ Establish a grid of possible sedimentation coefficients
- ★ To each sedimentation coefficient assign a diffusion coefficient based on an assumed hydrodynamic shape (f/f_0 ratio)
- ★ Calculate the theoretical boundaries for each s, D pair (by numerical integration in this case)
- ★ Fit the data to find the concentration of each potential species, using maximum entropy smoothing ("regularization") on the distribution

To illustrate the advantages of this approach, consider these simulated data for an antibody sample containing 20% dimer. The dark traces mark the optimal range for dc/dt analysis

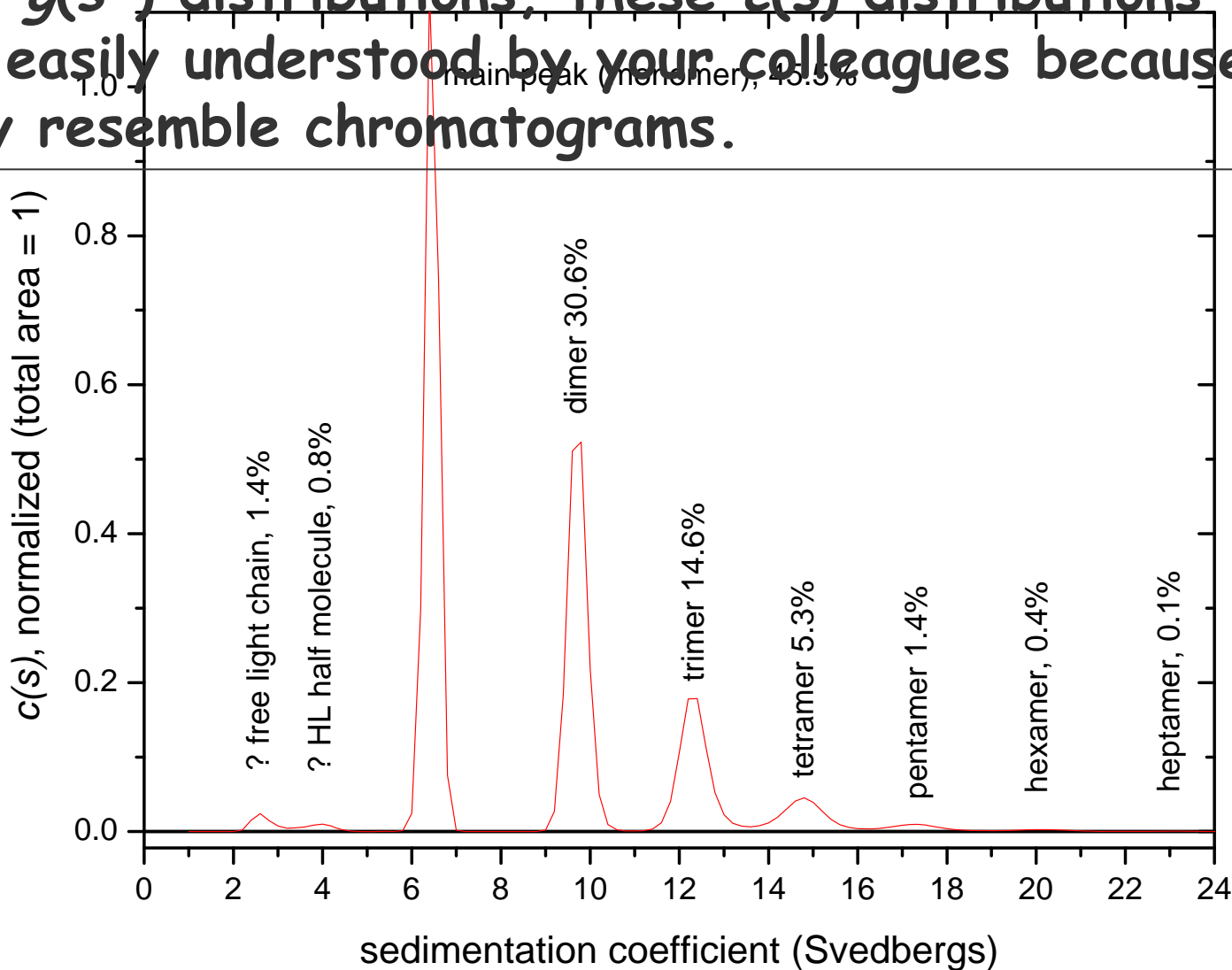


This compares the $g(s^*)$ and $c(s)$ distributions derived from those data



This data for a highly stressed antibody sample illustrates the promise of excellent resolution and sensitivity

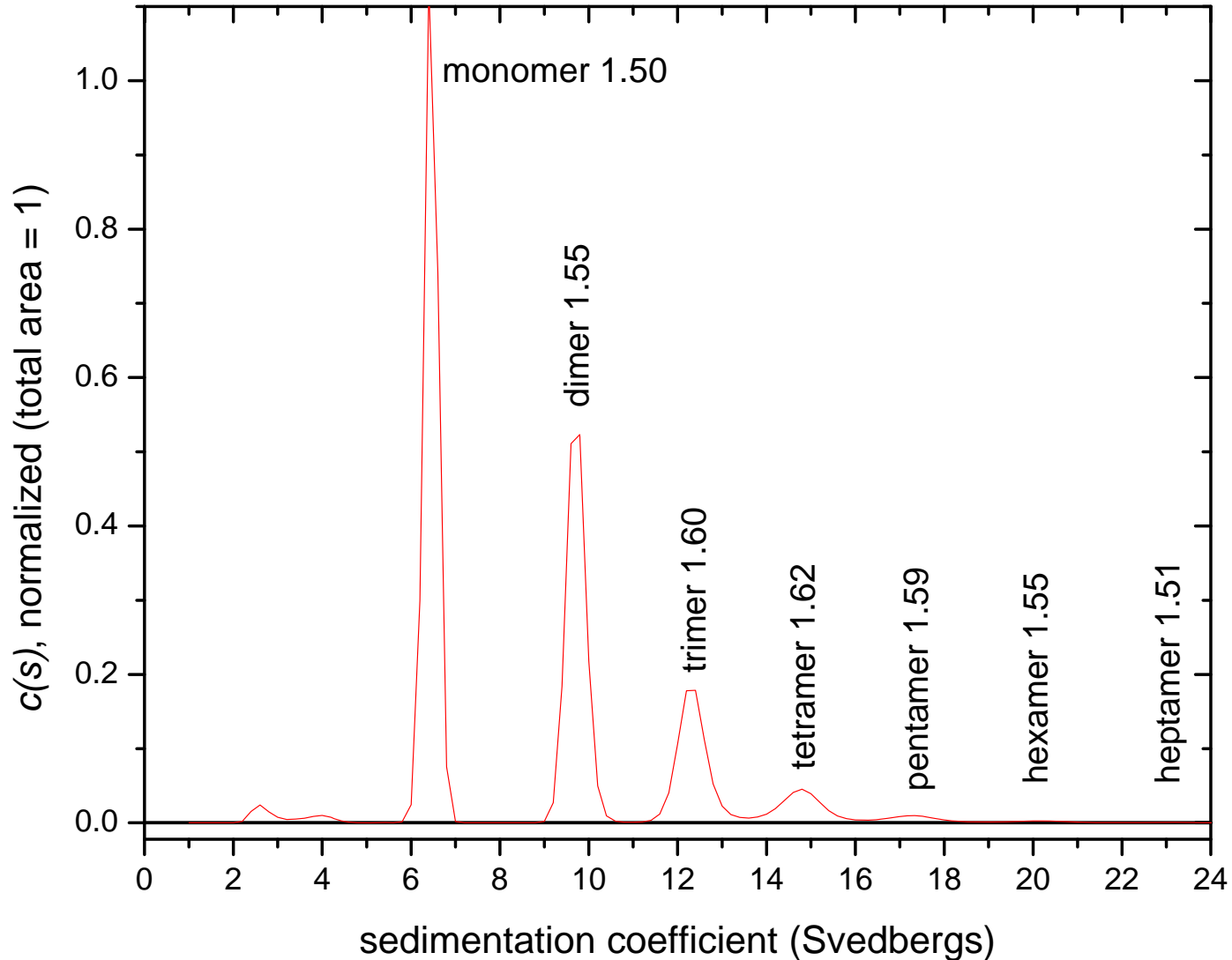
Like $g(s^*)$ distributions, these $c(s)$ distributions are easily understood by your colleagues because they resemble chromatograms.



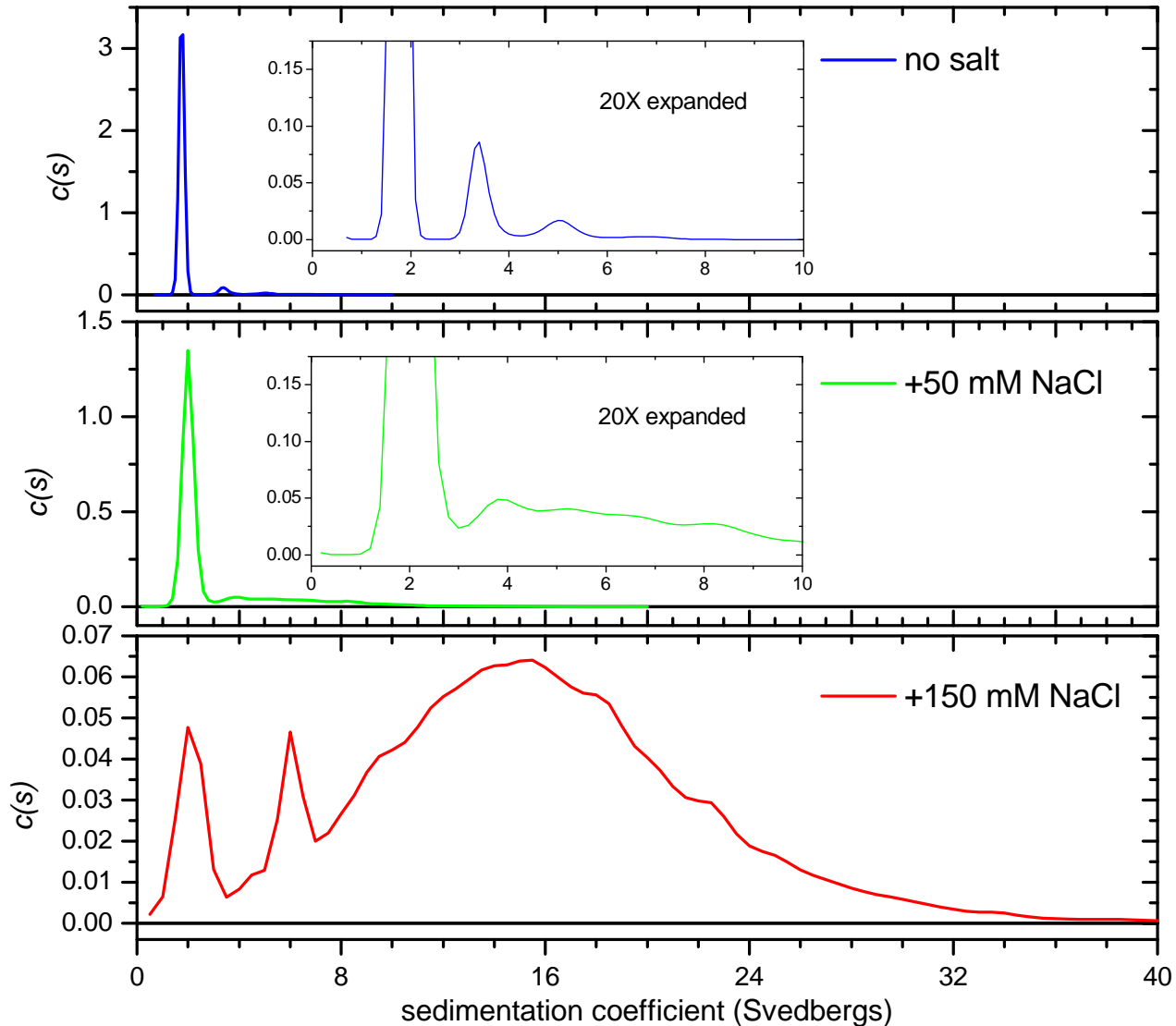
The peril: $c(s)$ distributions are also often misunderstood

1. the effective resolution goes down as the fraction of minor peaks goes down
2. the resolution you can achieve for a 150 kDa antibody is much greater than for a 20 kDa cytokine
3. the nature of the noise (variability) is very different than in chromatography, and false peaks are possible
4. for interacting systems the peaks probably do not represent individual molecular species
5. the best-fit f/f_0 ratio should be viewed as a fitting parameter; it may or may not have real meaning as a hydrodynamic property of any particular species
 - ✧ it is much more accurate to determine f/f_0 from the sedimentation coefficient and known mass

For antibody samples the assumption that all species have similar f/f_0 ratios is typically fairly good for the oligomers

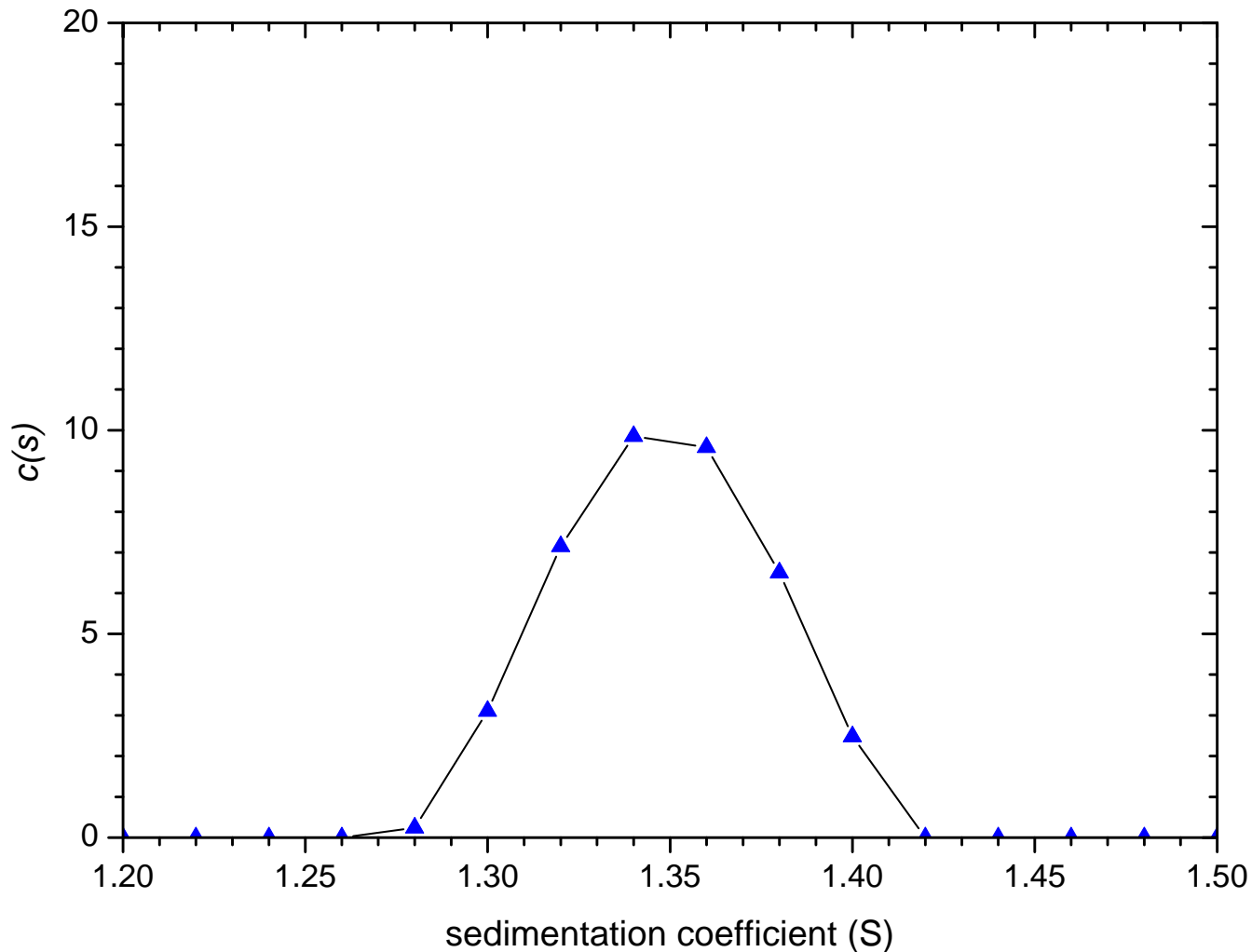


This study of the effects of NaCl on aggregate content illustrates the enormous size range that can be covered in one analysis with $c(s)$

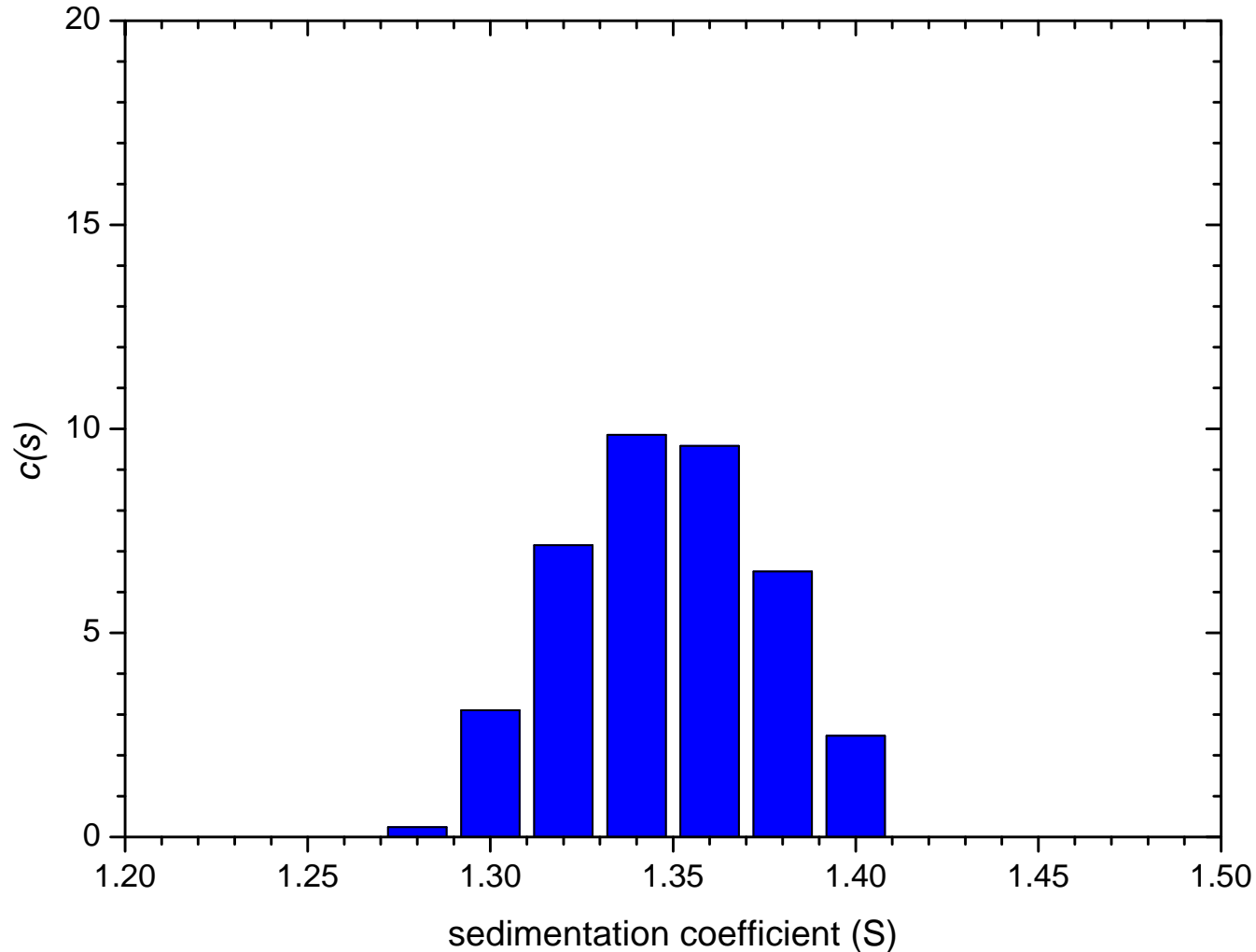


Why do the $ls-g(s)$ and $c(s)$ methods use smoothing (regularization) of the distributions?

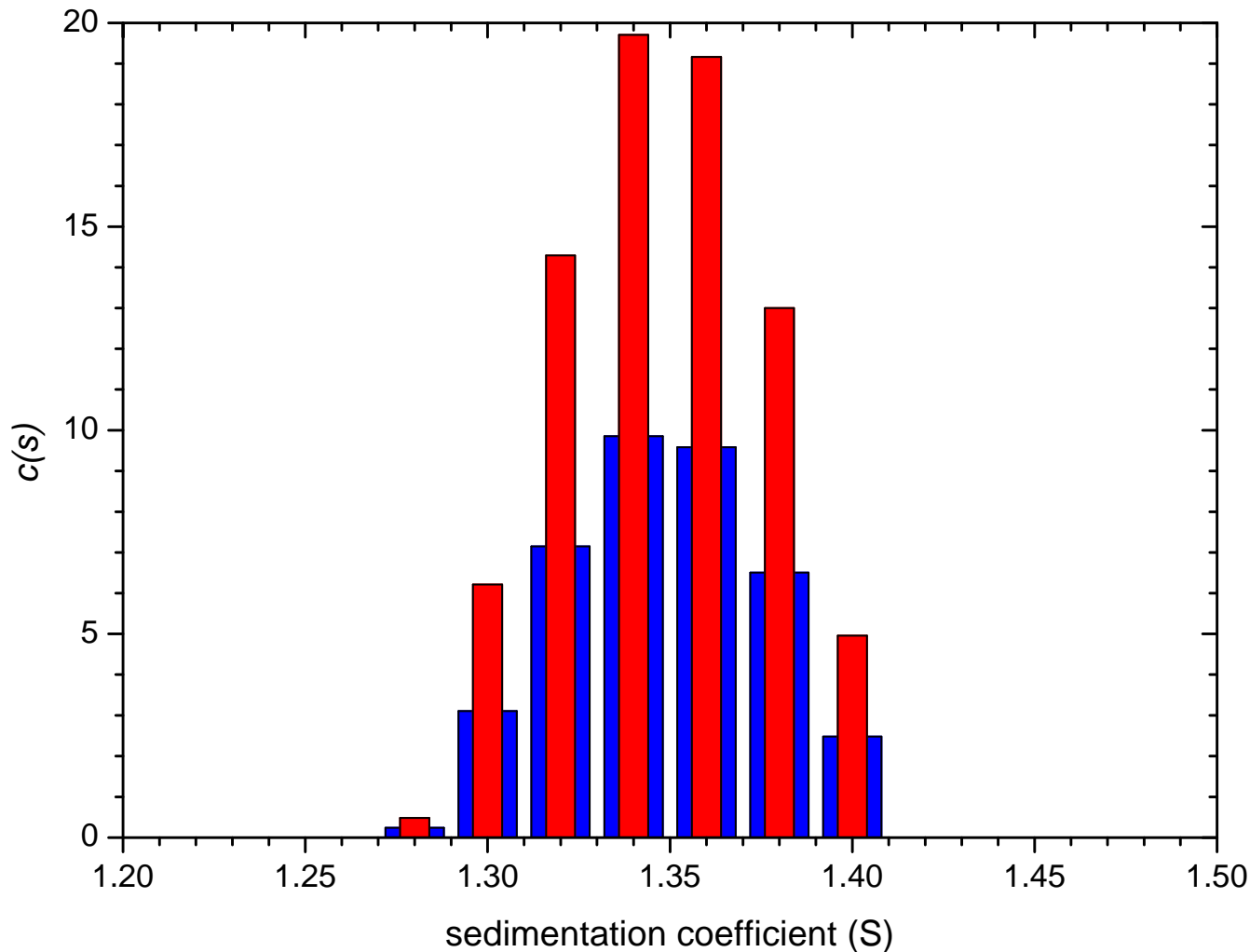
Let's look closely at this single peak:



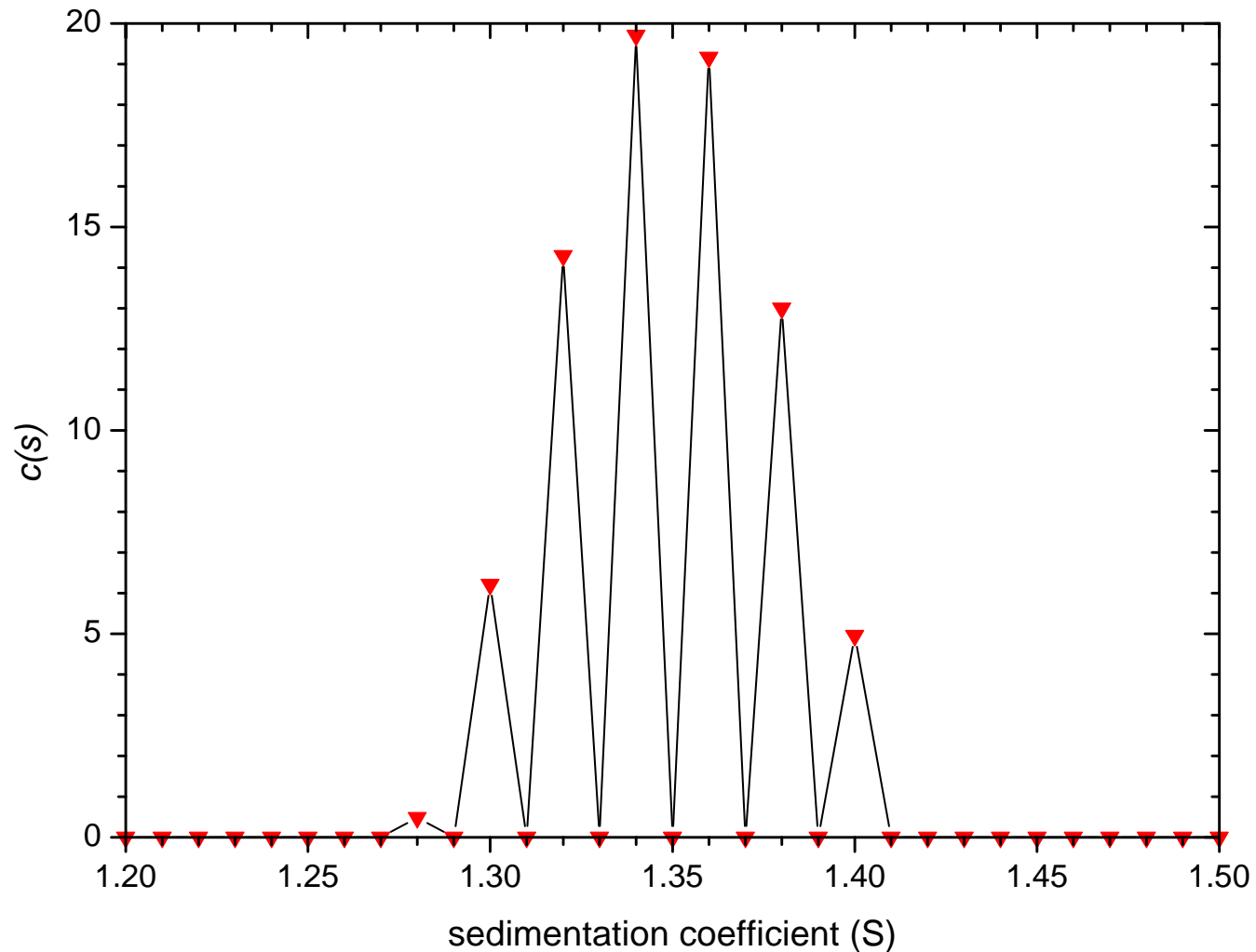
While we usually draw these distributions as line graphs, they really are histograms (bins of finite width)



But suppose we reduce the bin width by half and set all the new bins to zero: the red histogram produces the identical theoretical curves (boundaries) as the blue one



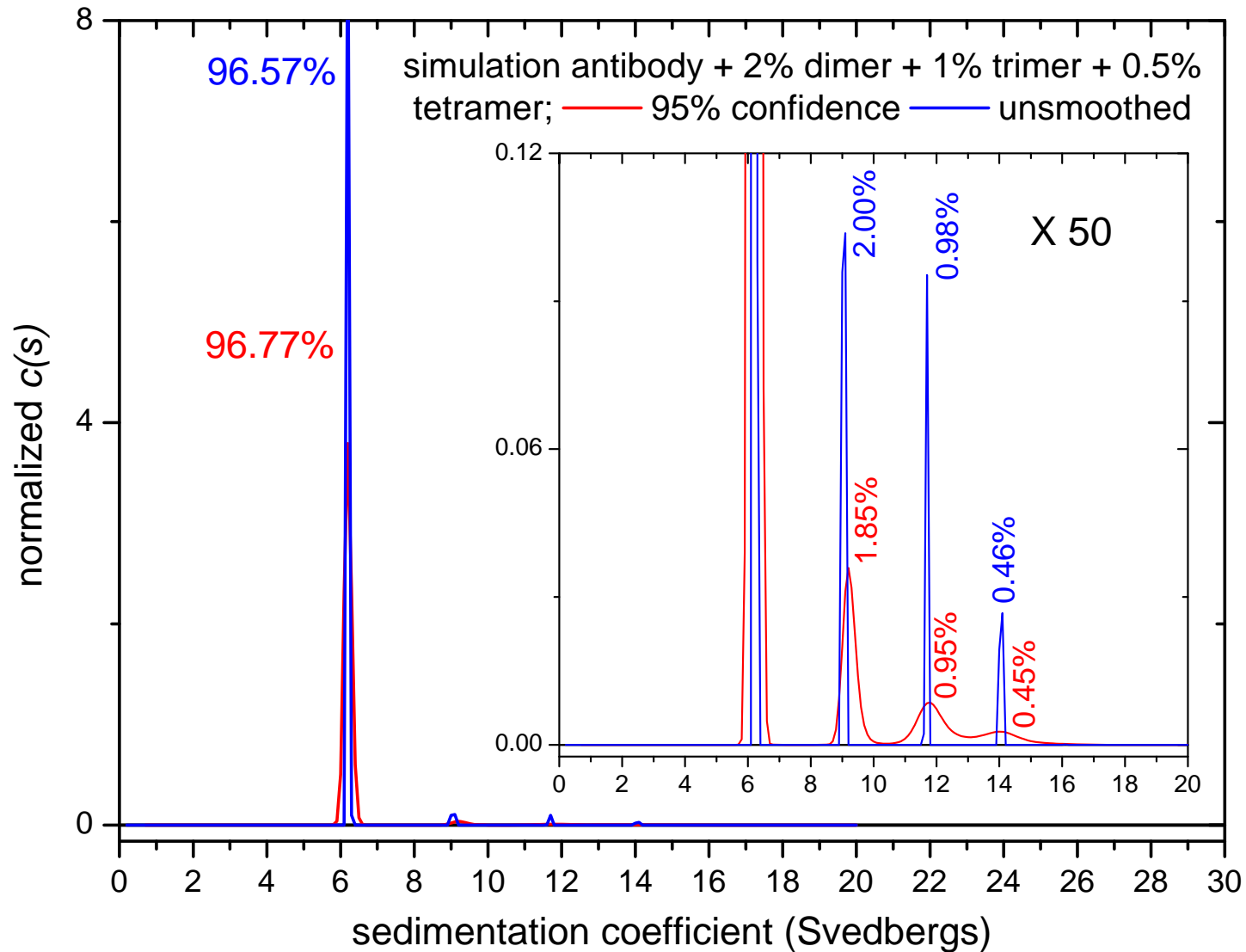
But the red histogram really represents 7 peaks, not one!



Therefore regularization/smoothing is used to try to prevent false peaks

- ★ It is assumed that the resolution of peaks and information content of the distribution is limited, and at some level of detail the distributions should be smooth
- ★ The degree of smoothing is set via a confidence level. The smoothed distribution does not fit quite as well (the r.m.s. deviation is slightly higher), but at some confidence level one is sure it contains all the correct information content of the distribution

Caution---maximum entropy normalization does not preserve peak areas!



The $c(M)$ method also exists but I would discourage its use except for samples that truly are a single species

- ★ The M axis is calculated based on
 1. the f/f_0 ratio that best fits the raw data
 2. the assumption that all species have the same shape
- ★ However the minor components contribute correspondingly little to the overall weight-average f/f_0 ratio, so there is almost no information about their true mass in the raw data
- ★ In my opinion, this method throws away the robust data (the sedimentation coefficients) and substitutes a molecular mass scale that is largely fantasy

Your best guide to assigning masses to aggregates at low levels are their s values

oligomer	s/s_{monomer}	oligomer	s/s_{monomer}
dimer	1.45	pentamer (pentagon)	2.45
trimer (linear)	1.75	pentamer (bipyramid)	2.60
trimer (triangle)	1.86	hexamer (hexagon)	2.67
tetramer (linear)	2.00	hexamer (trigonal prism)	2.90
tetramer (square planar)	2.20	hexamer (octahedron)	2.97
tetramer (tetrahedron)	2.26	octamer (cube)	3.46

Garcia de la Torre, J. and V. A. Bloomfield (1981). Hydrodynamic properties of complex, rigid, biological macromolecules: Theory and applications. *Q. Rev. Biophys.* 14: 81-139

Some advantages and drawbacks of the $c(s)$ method

Advantages:

1. No *a priori* assumptions about how many species are present
2. Excellent resolution and signal/noise
3. Excellent for assessing homogeneity vs. heterogeneity
4. Good quantitation of amount of various species
5. Graphs easily understood by colleagues

Drawbacks:

1. Results are model-dependent and vary with assumed f/f_0 and regularization settings
2. May produce false peaks, or real peaks may be merged by smoothing
3. Minor peaks ($\sim 1\%$ or less) may appear or disappear with small changes in parameters or data included in fit
4. No error bars
5. Can be computationally intensive

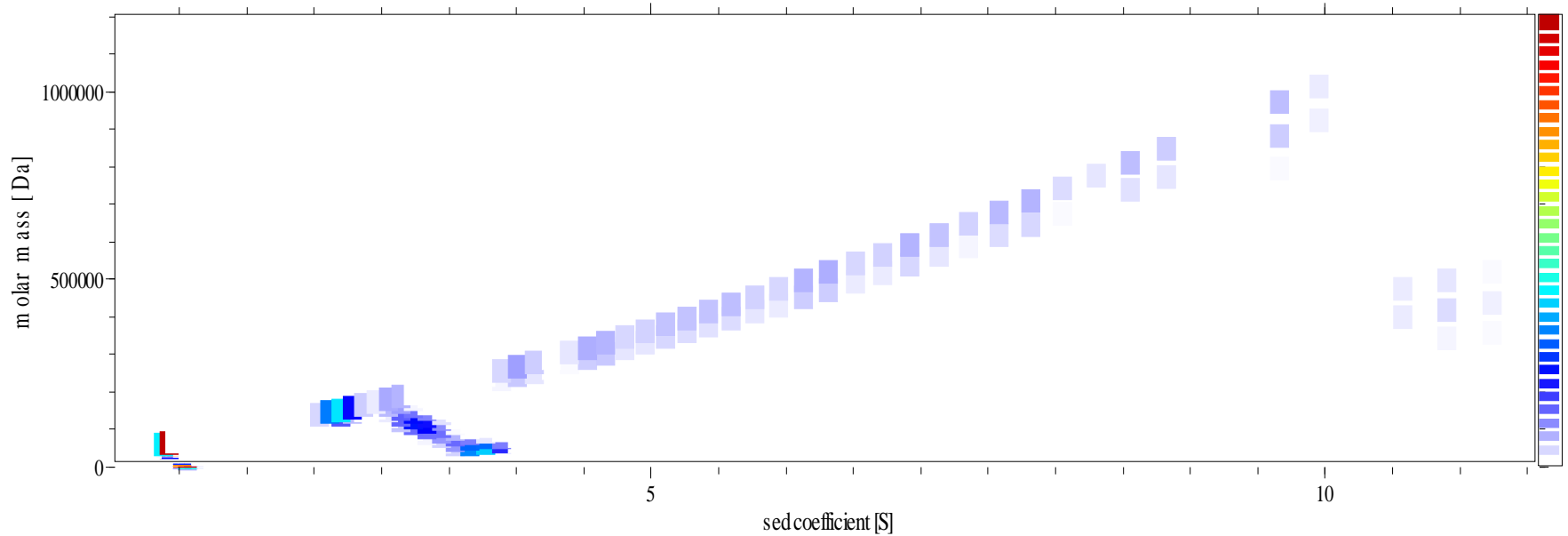
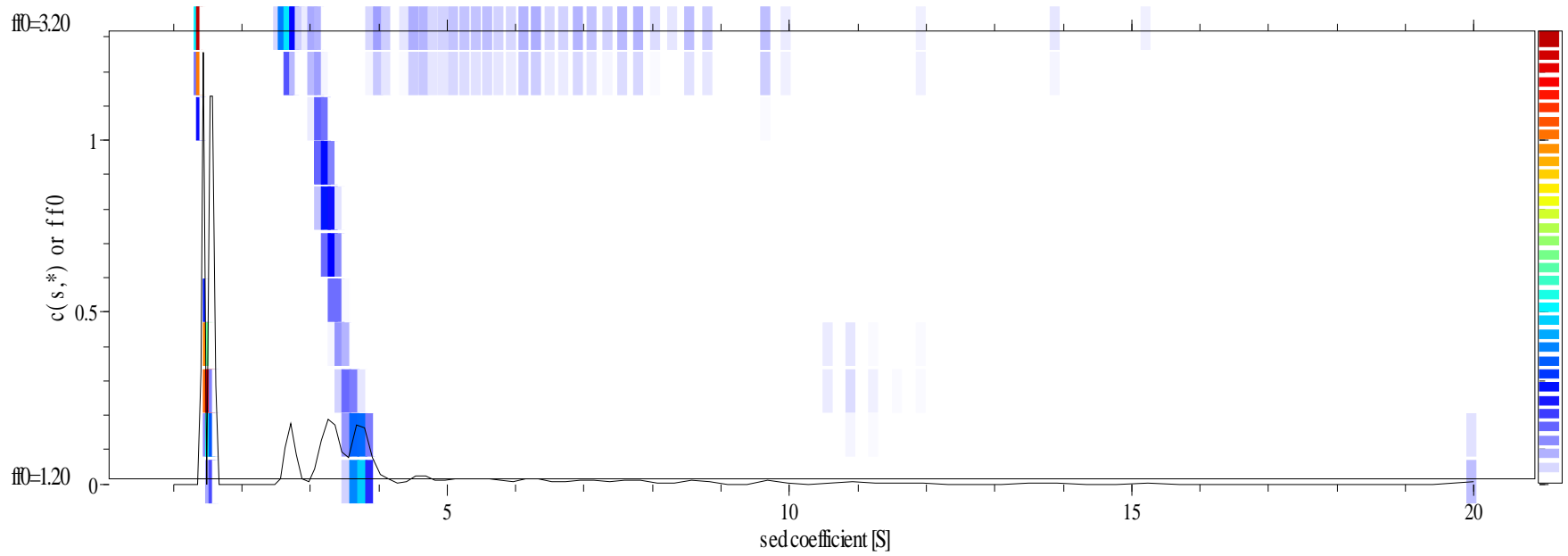
$c(s)$ has become very popular, but remember no single approach is optimum for every sample

1. if $c(s)$ indicates only a few species are present, then whole boundary analysis with a discrete species model will give better quantitative results
2. watch out for false peaks (usually at levels of a few percent of the total area), especially if you don't get a good fit of the data

Kick it up a notch---the $c(s, f/f_0)$ distribution

- ★ Rather than assigning one f/f_0 ratio to all species, this method allows a set of possible f/f_0 ratios at each sedimentation coefficient (typically 10 values)
- ★ Therefore there is a 2-dimensional grid of possible species
- ★ Because the number of species is much higher than in standard $c(s)$ the computations take significantly longer
 - ☆ but there is no fitting of an optimum single f/f_0 ratio so it is all done in one step

Example: a protein in lots of detergent



When it comes to data analysis software, let the user beware!

- ★ The 'conventional wisdom' about these methods may be wrong
- ★ You really should test whether the method gives correct results for your application
- ★ Remember, new features and new models provide ample opportunities for the law of unintended consequences
 - ☆ things that used to work correctly might not any more
- ★ The fact that a method has been published doesn't necessarily mean it has been thoroughly tested or evaluated

Is my analysis right?

A light blue brushstroke underline that is slightly wavy and tapers at both ends, positioned directly below the text.

Remember, all estimates of D or M from velocity experiments assume the boundary width is due only to diffusion

- ★ these estimates assume there is no extra spreading from variations in sedimentation coefficient; thus heterogeneity leads to overestimates of D and underestimates of M
- ★ these estimates assume there is no significant interconversion of species during the experiment (*e.g.* association/dissociation)
 - ★ oligomers or complexes should, however, give the correct apparent mass if the concentration is above ~ 300 times K_d

Beyond that, just how do you know when you have it right?

1. Be sure the analysis method is appropriate for your sample, based on
 1. what you know about its biochemistry
 2. what information you are trying to obtain
2. Select the least model-dependent method that will give you the information you need
3. Be sure the data you've selected is appropriate for the analysis method
 1. scans cover appropriate time range
 2. data only in appropriate regions of the cell

Do I have it right? (continued)

4. Do the results make hydrodynamic sense?
 1. Is it sedimenting faster than is theoretically possible for the mass I've assigned ($f/f_0 < 1$)?
 2. Is it sedimenting so slowly it implies it is unfolded or unreasonably asymmetrical ($f/f_0 > 2$)?
5. If more than one method is appropriate, do they give the same answer?
6. Repeat the experiment!
 - ★ that may take less time than fussing with more data analysis, and it is the only true way to evaluate the robustness of the result

Do I have it right? (continued)

7. Simulate your experiment with a finite-element simulator + realistic noise; then fit the simulation to find out what can theoretically be done
 - ★ can I really resolve that many species?
 - ★ what accuracy can be obtained?
 - ★ use simulators in SEDANAL, SEDFIT, SVEDBERG, DCDT+, or ULTRASCAN
8. Repeat the experiment again!
9. Do the results change (or not change) as expected if I repeat the experiment under different conditions?