

THE UNBEARABLE FUZZINES OF NMR DATA !?!

... a brief reflection by Stan Sykora (ebyte.it) ...

Given the limited time,
I will present only a few NMR-based illustrations of a broader question that is presently becoming acute due to advances in *Big Data* handling, *Artificial Intelligence*, and other areas:

How *hard* are the *hard Sciences* data, really?

Tentative (but qualified) answers:

- The simulated ones are cute, hard and as sharp as it goes 😊
- The ‘real’ rest is *disgustingly soft* and as fuzzy as it goes 😞

NMR is no exception – at all levels!

Presented at XLVI Annual Congress of GIDRM, Fisciano, Salerno (Italy), 27-29 Sep 2017

My own involvement in these things (just to explain)



In the last 10 years I was basically trying to teach a computer to do the work of a spectroscopist (which is what 'automatic' really means)

Expectations and Reality (in general)

I. Fuzziness in the rules of the game and in imperfect comprehension

Some fellow programmers tell me:

You write software for NMR spectra analysis? OK, so your life is EASY!
Just ask some chemist what the rules are, implement them, and you are done.

Wouldn't that be nice! Alas, they were *just* programmers, poor things.

Some selected problems with coding what spectroscopists do:

- *No spectroscopist can describe how s-he does it, not in a flowchart manner*
- *At best, they provide some (good and useful) examples, and some weak rules*
- *For each such rule, I easily find tens of exceptions!*
- *If I ask two of them why an exception occurred, they rebuke me that, of course, exceptions are to be expected (afterwards, they quarrel among themselves)*
- *Hence, there is no 'NMR Spectroscopy Master Book' that would hold always*
- *About 10% of published assignments (for example) are known to be wrong*

... an interlude ...

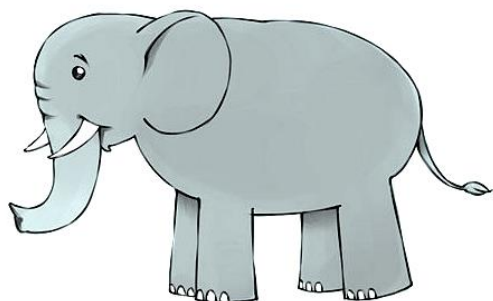
Yet, despite all the odds,
experienced spectroscopists are almost always correct !!!

(That is really vexing!)

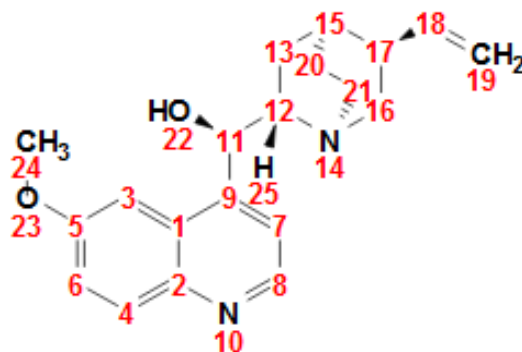
And I am to build software that, statistically, should beat them !??
(no worry, I am in sight of the target, but still a long way to go)

Expectations and Reality (still in general)

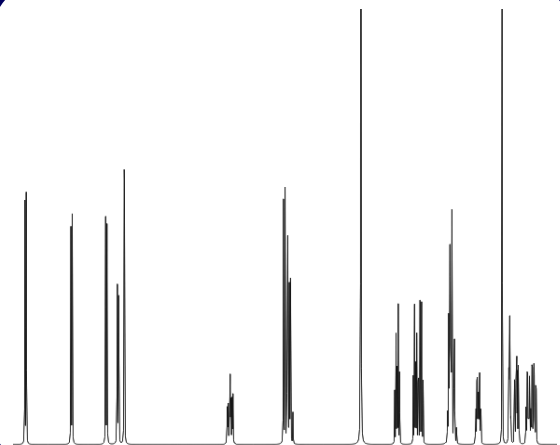
II. Reality is infinitely complex and never totally predictable. Hence, it is fuzzy!



This is NOT a real elephant!
It's just a simple drawing of
an elephant.



This is NOT a real molecule!
It's just a structural sketch of
a molecule.



This is NOT a real spectrum!
It's a naive, simulated NMR
spectrum.

Most of my code (>80%) is about whether what looks like XYZ, is really XYZ.
On the opposite, *true Science* is the trivial part of the code, amounting to < 5%.

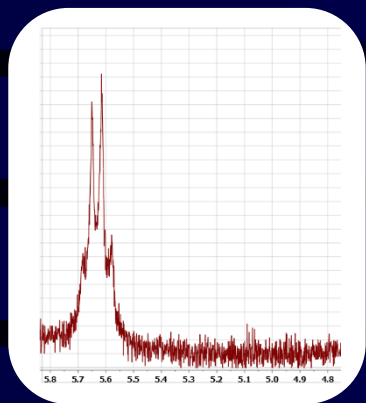
NMR fuzziness at NMR ground level (the Reception Hall)

We would like to 'register' only clean **Peaks** and **Multiplets** (1D) or **Clusters** (2D)
But what enters through the front doors includes:

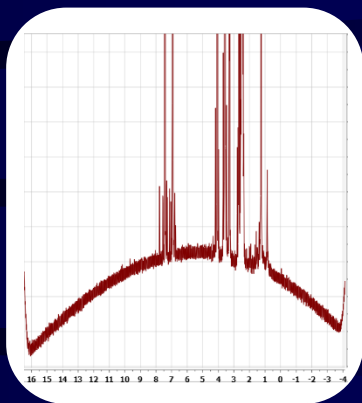
- Noise
(from sample, probe, switching diodes, preamp, external, ...)
- Instrumental artifacts
(dead time, imperfect shimming, rotation sidebands, spikes, ...)
- Sequence artifacts
(ill defined expected intensities, missing and extra clusters, ...)
- Acquisition artifacts
(fast repetition, FID truncation , limited digital resolution, ...)
- Evaluation artifacts
(bad phase or baseline, referencing errors, DFT distortions, ...)
- Expected impurities
(solvent and its water, S- and Q-reference, residual solvents, ...)
- Unexpected impurities
(various contaminants, fingerprints, ...)

It could definitely fill a full **Book** (I already promised to write one ☹)

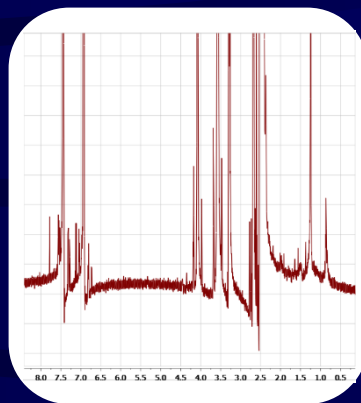
A few examples of undesirable spectral artifacts



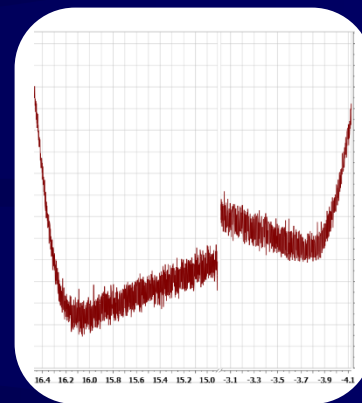
Noise



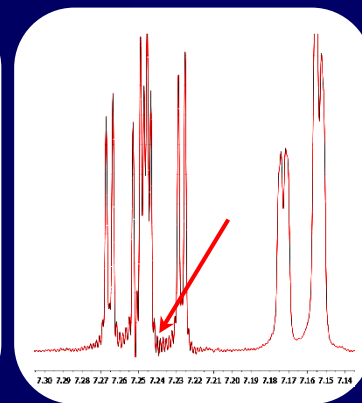
Baseline roll



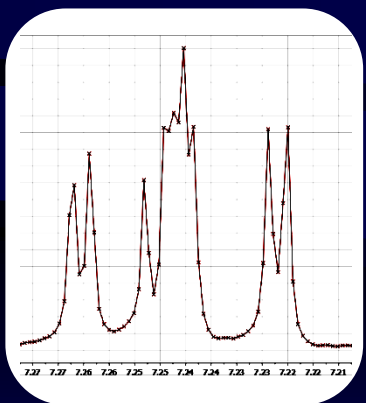
Imperfect phasing



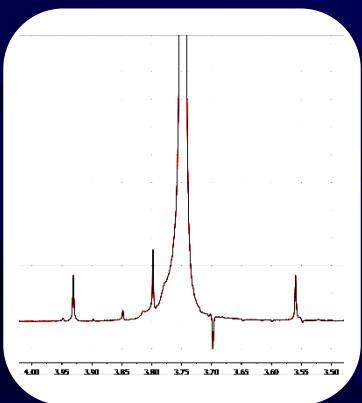
Bruker / Jeol
smileys / brownies



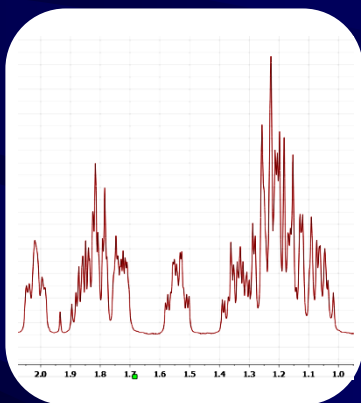
FID truncation
effects



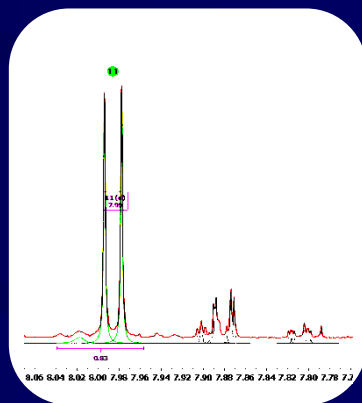
Underdigitization



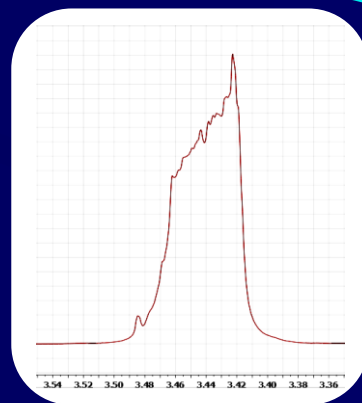
Rotation sidebands



Peaks overlap



Impurities peaks



Sample temperature
drift effects

=> There is so much to get rid of before doing anything serious with a spectrum !!!

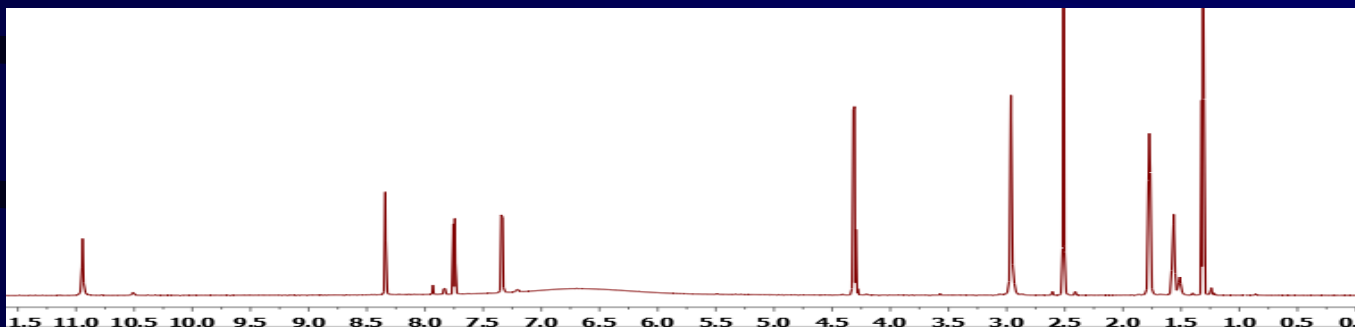
Ground floor data management (continued)

The main task of a Reception is to get rid of all the bullshit that enters.
In my world, this is one of the tasks of GSD (**Global Spectral Deconvolution**).

Why was GSD born? The basic idea, going back to 2006:

What does a spectroscopist see?

Peaks, multiplets (singlets, doublets, AB quartets, triplets, quadruplets, ...), labiles, ^{13}C satellite peaks, aromatic peaks, d-solvent and, reference peaks, water peaks, impurities, reaction solvent residuals, spinning sidebands, ...



What does a programmer see?

Just an unexciting array of complex-valued data!
He can't understand what is the chemist talking about!

Implication: there is a big communication problem

GSD: Historic notes

- I have started pushing the basic idea of Peaks List in 2006
- The rather complex algorithm was finalized in Summer 2008
- Since then, GSD was well tested and proved to be very robust
- Imitators and potential competitors appear around 2013 (CRAFT) ☺
- Today, GSD was applied in many different contexts, including structure verification, drug discovery, NMR quantitation, metabolomics, and others, and often turns out to be an enabling key to various novel avenues.

References:

DOI: 10.3247/SL2Nmr08.011

DOI: 10.3247/SL3Nmr09.003

Global Spectral Deconvolution (GSD) of 1D-NMR spectra

Carlos Cobas¹, Felipe Seoane¹, Stanislaw Sykora²

¹Medislab Research, Xosé Pousa 6, Santiago de Compostela, 15706 Spain; carbas@medislab.es
²Extra Byta, Via Raffaello Sanzio 22/C, Cuneo Perno (MI), Italy; f.seoane@extrabyta.it

INTRODUCTION to the GSD concept

No matter how "noisy" a typical 1D NMR spectrum might be, it always contains the following "matter":

1. A set of spectral "peaks" with NMR characteristics (such as T₂-rho dependent position angle and phase)
2. Desirable distortions due to random artifacts (distortions L and V effects, center clipping, and acquisition distortions)
3. Spikes and other "non-relevant" 1D signals entering the detector but not generated by the nuclei in the sample
4. The dispersion receiver noise which, at least in first approximation, is completely independent of the sample

We are used to cope daily with the co-presence of all these parts, even though only those indicated are of certain meaningful information. Using a number of distinct algorithms, an experienced operator cope with the co-presence meanwhile well when processing the data manually, but if other plays a "noisy" with automatic routine, such as fitting theoretical spectra of coupled spin systems to a experimental data or, for that matter, just obtaining a simple integration. Hence,

the first goal of GSD is to extract from the spectrum only part 1

The second problem one encounters regards the interpretation of part 1) itself. In particular:

- a. Rather than being "continuous", peaks are distributed by their chromatography and molecular dynamics
- b. In an for medical molecules, there are many other more theoretical distortions than experimentally recognizable peaks. This may affect the apparent peak width and shapes even than chromatography
- c. Due to their finite bandwidth, peaks overlap heavily giving rise to hump-like peak combinations
- d. Some peaks are "overlapping" effects (overlapping and others, though structure) meaningful, are undesirable (solvent and impurity peaks) and, in some contexts, undesirable signals such as "C₁₃ satellites". All such peaks should be "washed" out, something that is very difficult to do directly in an analog spectrum.

Again, a complete, automated procedure dealing with all these problems at once has so far never been tried. At best, there are separate procedures for reference deconvolution, resolution enhancement algorithms, multiple deconvolution, etc) which address some of these problems one at a time. Hence,

the second goal of GSD is to set up a numerical Peaks List such that a synthetic spectrum generated from it differs from part 1) by less than the spectral real noise.

The way the GSD algorithm works, the two goals are closely intertwined and can not be intended as separate phases. One can view GSD as an extension to the whole spectrum of the classical multiple deconvolution (since the model list is not a priori known) and as an automatic application of the algorithm. The individual steps will be described on a spectrum of Decoupled Aromatic Amino Acids (DAAA) which is completely automatically and on the whole spectrum but, for reasons of clarity, the performance is illustrated on selected narrower windows.

Step 1: Noise and noise levels estimation

The pre-requisite for a correct operation of GSD is a reliable estimate noise and noise bandwidth. To do that, we have developed the low and some high resolution algorithms which will be published and discussed elsewhere.

Step 2: Automatic calculation of first and second derivatives. This is done by means of the Savitzky-Golay convolution algorithm (SGD). The task, of course, is a correct setting of the SGD parameters (number of points and order), based on the estimated noise and bandwidth. The aim is to reach an any possible dependence. Moreover, the use of the first and second derivatives enhances resolution (like in our Resolution DecoderTM) and converts base fluctuations into distinct peaks. This is also in line with the derivative spectra of each, first and second derivatives can be used for subsequent evaluation.

Figure 1a) shows the 4-4.4 ppm portion of the spectrum (bottom) and its first (middle) and second (top) derivatives.

Step 3: Special points identification

An efficient peak-picking algorithm based on correct noise estimate is applied to the original spectrum as well as to the first and second derivatives spectra. It searches recognizable local maxima and maxima occur in each of the three spectra are approximately flagged in case of a complete spectrum, this is done for both the real and imaginary parts.

Figure 1b) shows the special points as before. Negative marks correspond to local minima, positive ones to local maxima.

Figure 1c) shows the 4-4.4 ppm portion of the spectrum (bottom) and its first (middle) and second (top) derivatives.

Step 4: Peaks recognition

All recognizable peaks are detected and marked, using the possible maxima in the second derivative spectrum (realization and total baseline independent) and the maxima in the original spectrum for local maxima. Peaks with estimated bandwidth smaller than 2 points are discarded as spikes.

Figure 1d) shows the peak recognition algorithm at work. The red and red dots marks are as before but the red dots show each peak (positive-negative) are replaced by a red circle and rectangle. Notice that some small peaks (some small peaks) were correctly picked up, some red derivative rectangles are tagged by more than one red circle and some peaks and some are, etc. There is not enough space on this slide to explain all the points of this peak recognition method.

Step 5: Fine setting of each peak parameter (frequency, bandwidth, height)

To avoid baseline dependence, the parameter values are based on the knowledge of only the special points (maxima and minima) of the first and derivative spectra. The result is a new Peaks List. The latter can be used to generate a synthetic spectrum which reliably reproduces quite well the experimental one but is not of any baseline and noise.

Figure 1e) shows the fit to a 4.4 ppm portion of the experimental spectrum (top) and of the synthetic spectrum (bottom) computed from the new Peaks List. Notice that the comparison, though not perfect, is certainly as defined as a starting point for fit.

Step 6: Refinement of the parameters of all the peaks in the fit

There are several routines how to do this task after successful to find quickly and characteristically in speed (but in progress). We will eventually choose two routines - fast and slow - among which the user of the software will be able to choose. The goal is, of course, a perfect match of the spectrum - noise and spikes - instead. Indeed, this step is not distributed here because the fit is perfect - distributed only by the baseline artifacts and noise. The steps we are about to describe are shown in Figure 1e).

What is it good for?

We have already mentioned several of the advantages of having a spectrum through the GSD procedure and reducing it to an equivalent and clear list of peaks. There are other advantages which are related to having peak lists with a total exclusion of peak overlap errors. Of them about resolution enhancement. Of, for that matter, baseline correction - this is the only method as far that can potentially find out the baseline fit both real and imaginary parts.

For us, however, this is primarily a necessary step towards automated and/or computer aided molecule structure verification and elucidation. This "clean" list, rather than the spectrum, becomes the input to further editing and analysis aimed at the determination of all chemical shifts, J-couplings, and, eventually, molecules.

©2010 MEDISLAB
©2010 EXTRA BYTA

Developed in collaboration with



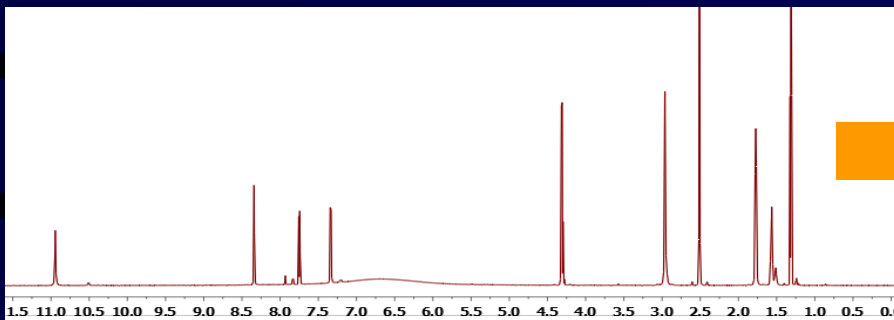
Ground floor data management (continued)

The GSD receptionist turns the input into a table of peaks that appear to be meaningful. They get room keys and proceed to higher floors. All the rest is turned out!

What does a spectroscopist see?

Peaks,

(solvent, reference, impurity,...), multiplets, ...



GSD

What does a programmer see?

An array of ... **peaks**, finally! That's **GREAT**

	ppm Δ	Intensity	Width	Area	Type	Flags	Kurtosis
44	2.612	0.474	1.224	8.289	Compound	None	0.000
45	2.637	0.437	11.624	239.109	Compound	None	-0.019
46	2.955	150.983	4.326	9158.016	Compound	None	0.132
47	2.963	243.975	5.934	19624.604	Compound	None	0.376
48	2.970	132.463	3.810	6486.636	Compound	None	0.744
49	3.058	1.449	11.493	236.917	Compound	None	0.022
50	3.302	1.667	43.038	Compound	None	1.600	
51	4.189	0.393	3.473	14.592	Compound	None	1.866
52	4.198	1.292	2.248	35.925	Compound	None	1.000
53	4.208	1.254	2.166	33.616	Compound	None	0.994
54	4.218	0.425	2.243	12.427	Compound	None	0.639
55	4.419	2.366	210.333	Compound	None	1.200	
56	4.289	22.854	2.197	595.927	Compound	None	1.259

Implications: language synchronization => better communication

GSD peaks-list example (a 400 MHz strychnine spectrum)

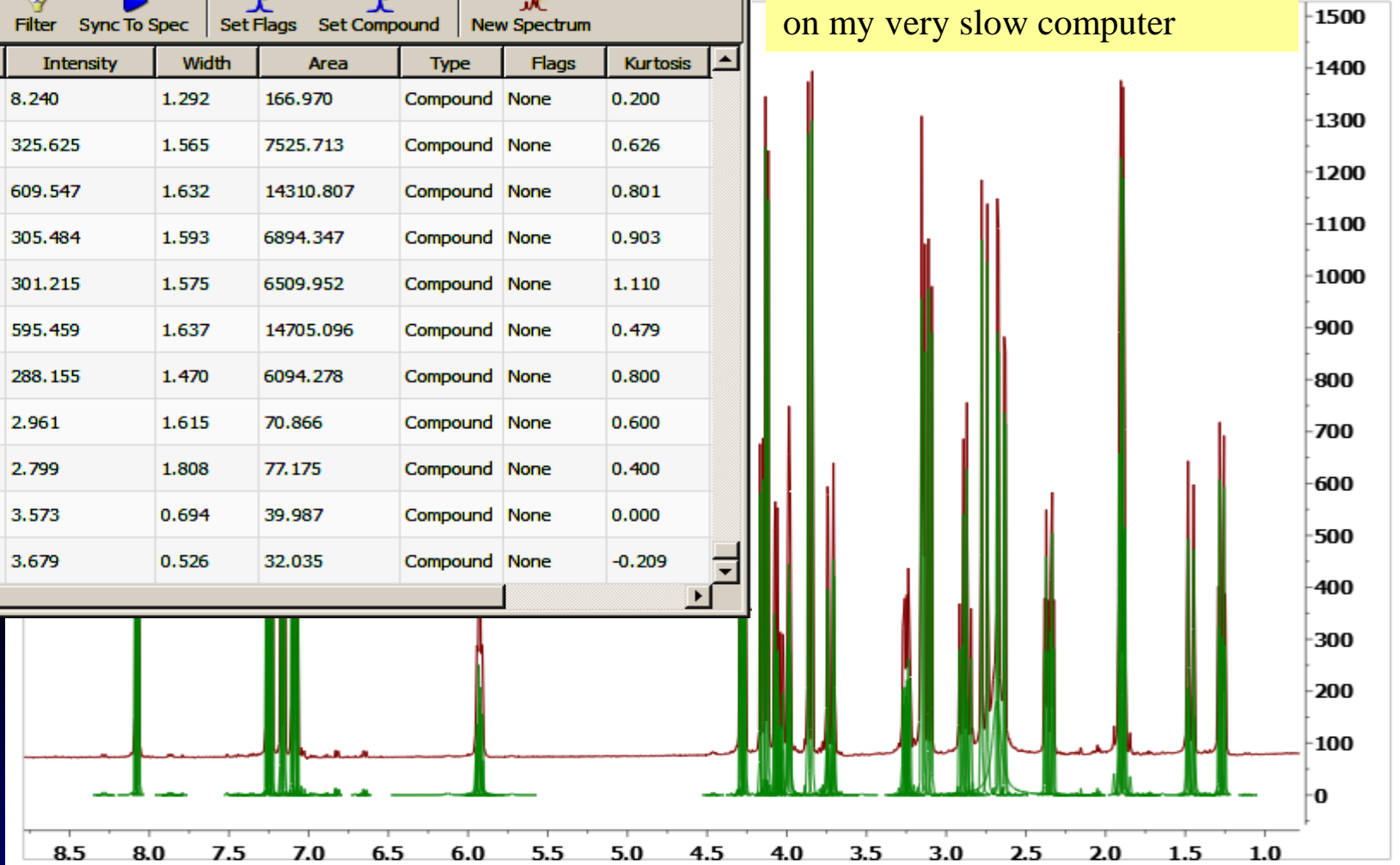
Peaks

Report Peaks Copy Peaks Setup Report Delete Select Peaks

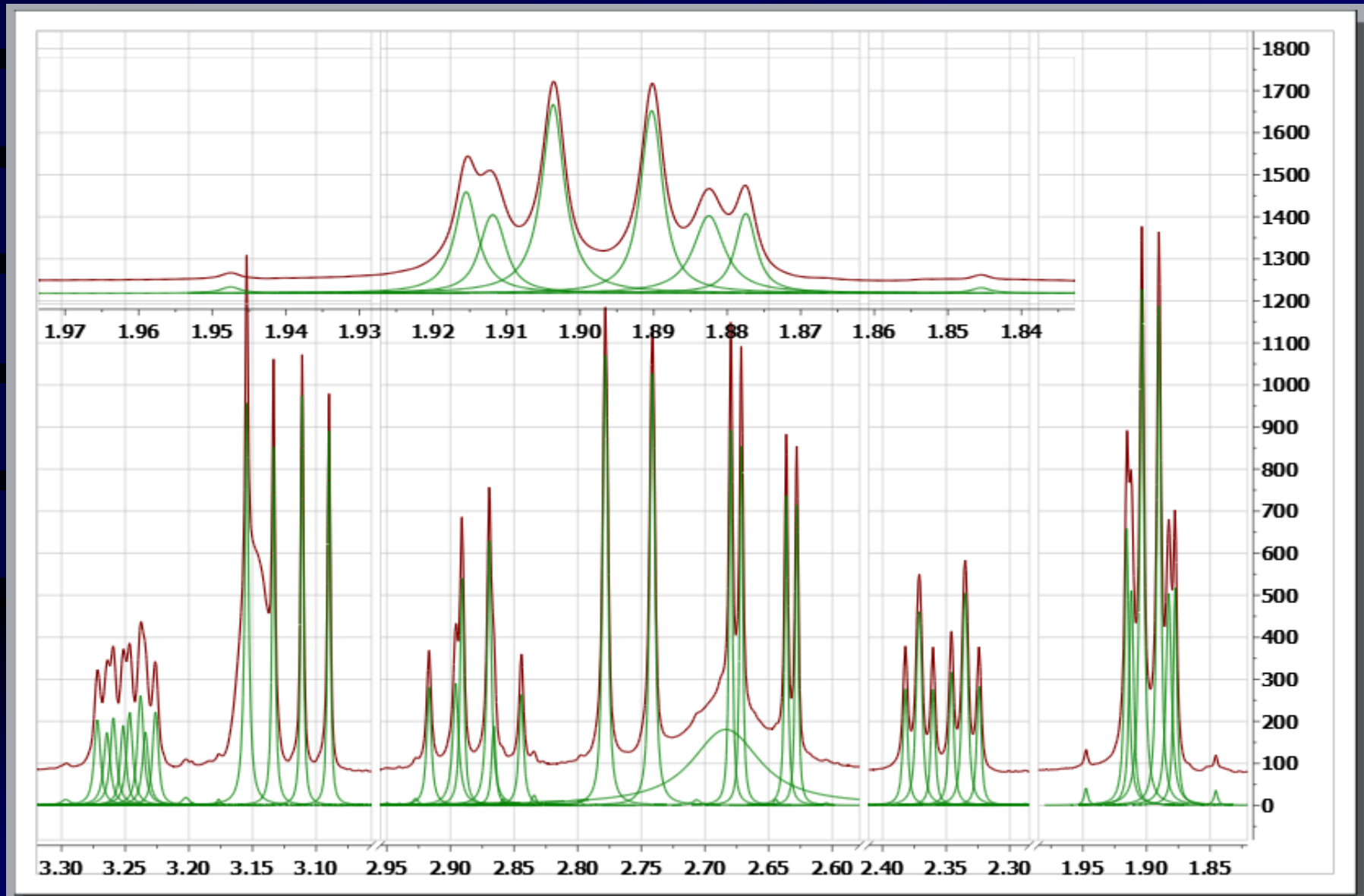
Sync From Spec Filter Sync To Spec Set Flags Set Compound New Spectrum

	ppm	Intensity	Width	Area	Type	Flags	Kurtosis
240	1.301	8.240	1.292	166.970	Compound	None	0.200
241	1.292	325.625	1.565	7525.713	Compound	None	0.626
242	1.284	609.547	1.632	14310.807	Compound	None	0.801
243	1.275	305.484	1.593	6894.347	Compound	None	0.903
244	1.265	301.215	1.575	6509.952	Compound	None	1.110
245	1.257	595.459	1.637	14705.096	Compound	None	0.479
246	1.249	288.155	1.470	6094.278	Compound	None	0.800
247	1.125	2.961	1.615	70.866	Compound	None	0.600
248	1.100	2.799	1.808	77.175	Compound	None	0.400
249	0.052	3.573	0.694	39.987	Compound	None	0.000
250	0.047	3.679	0.526	32.035	Compound	None	-0.209

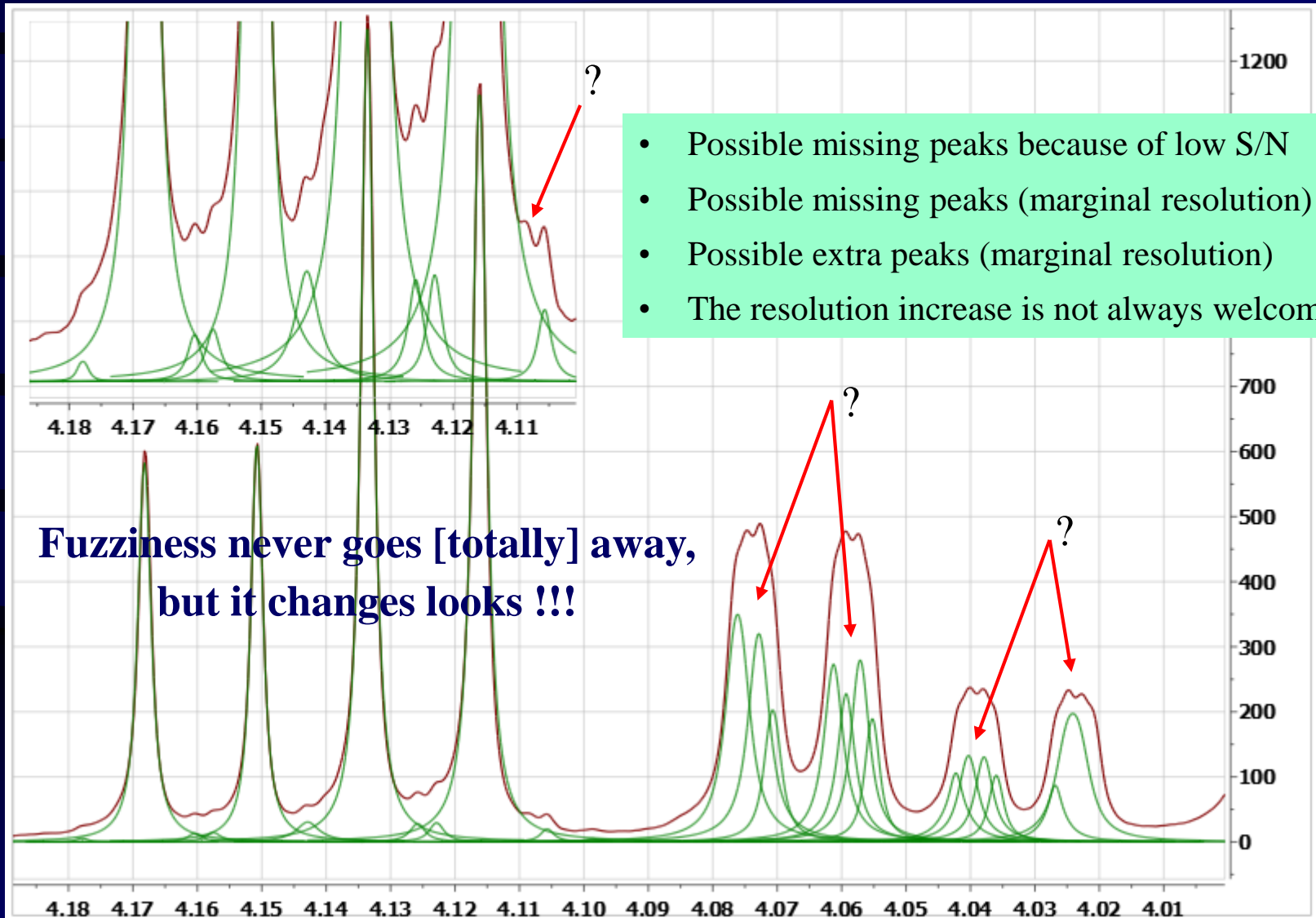
All done in 15 sec
on my very slow computer



Examples of peaks detection (WOW)



Examples of peaks detection (GSD limits and artifacts)



Automatic Peaks Editing

Having identified and tabulated all peaks, what more needs to be done ?

GSD by itself does not address issues like classifying each peak as:

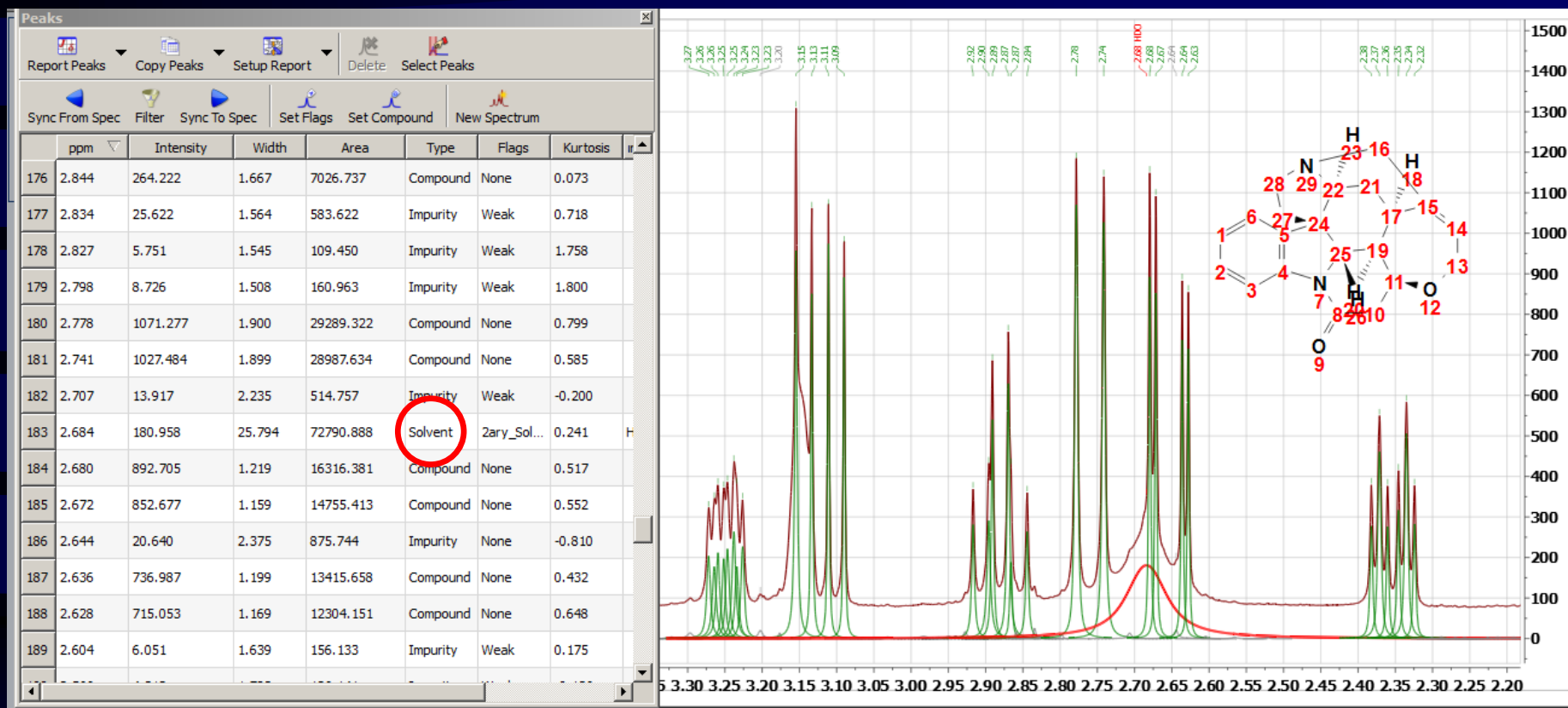
- compound,
- primary or secondary solvent,
- potential labile,
- ^{13}C satellite,
- valid member of a multiplet,
- impurity,
- S or Q reference,
- artifact,
- etc.

Nor does GSD group 1D peaks into multiplets and/or clusters and classify those.

All these are tasks which, unavoidably, bring-in fuzziness of their own.

Automatic peaks editing example

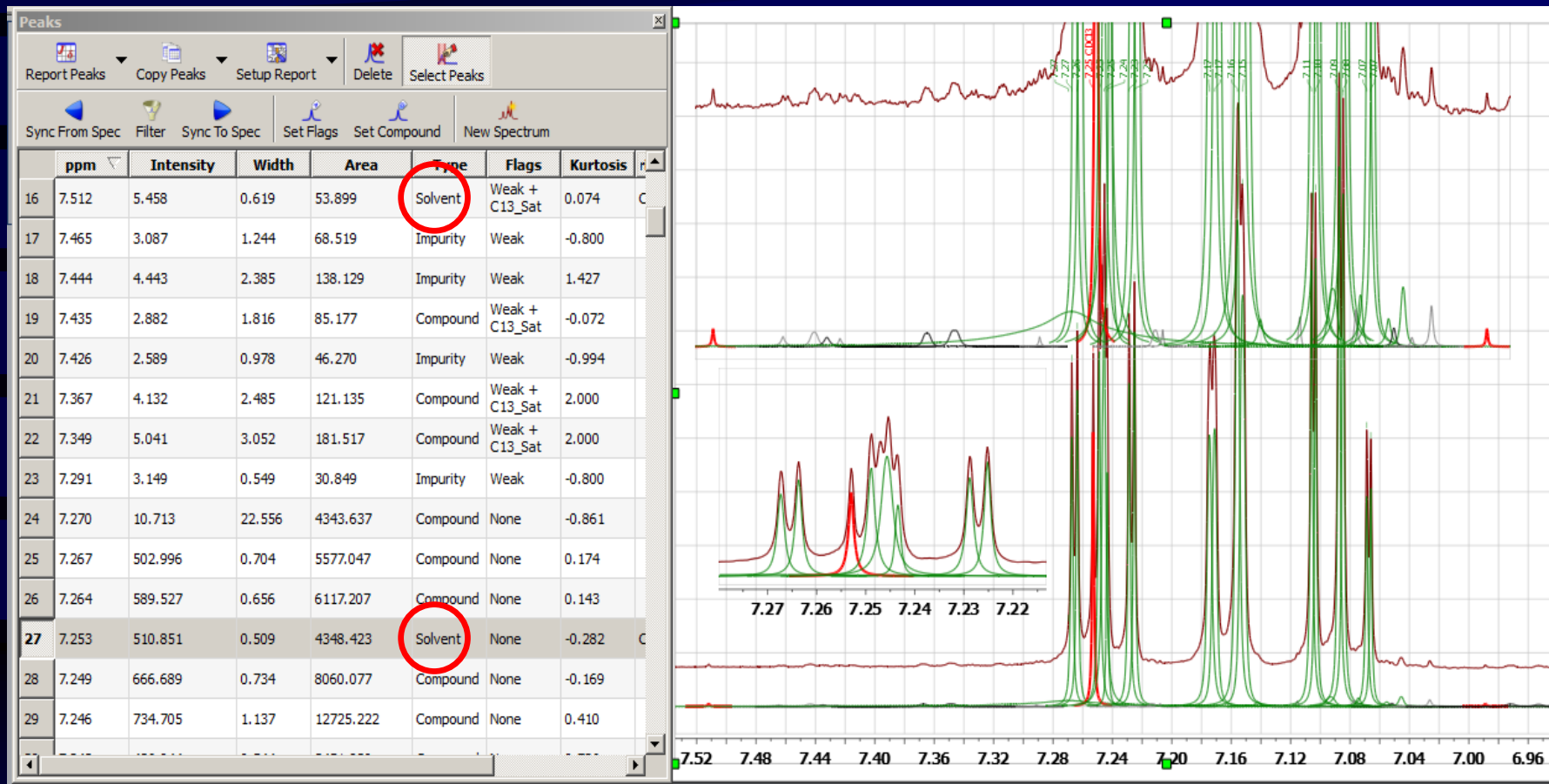
water peak recognition (primary and secondary solvent peaks are drawn in red)



Peaks editing uses greedily any available info (1H, HSQC, molecular predictions, ...). It is the first plank in the evaluation hierarchy where specific NMR know-how is used.

Automatic peaks editing example

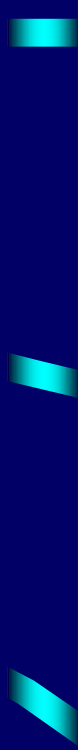
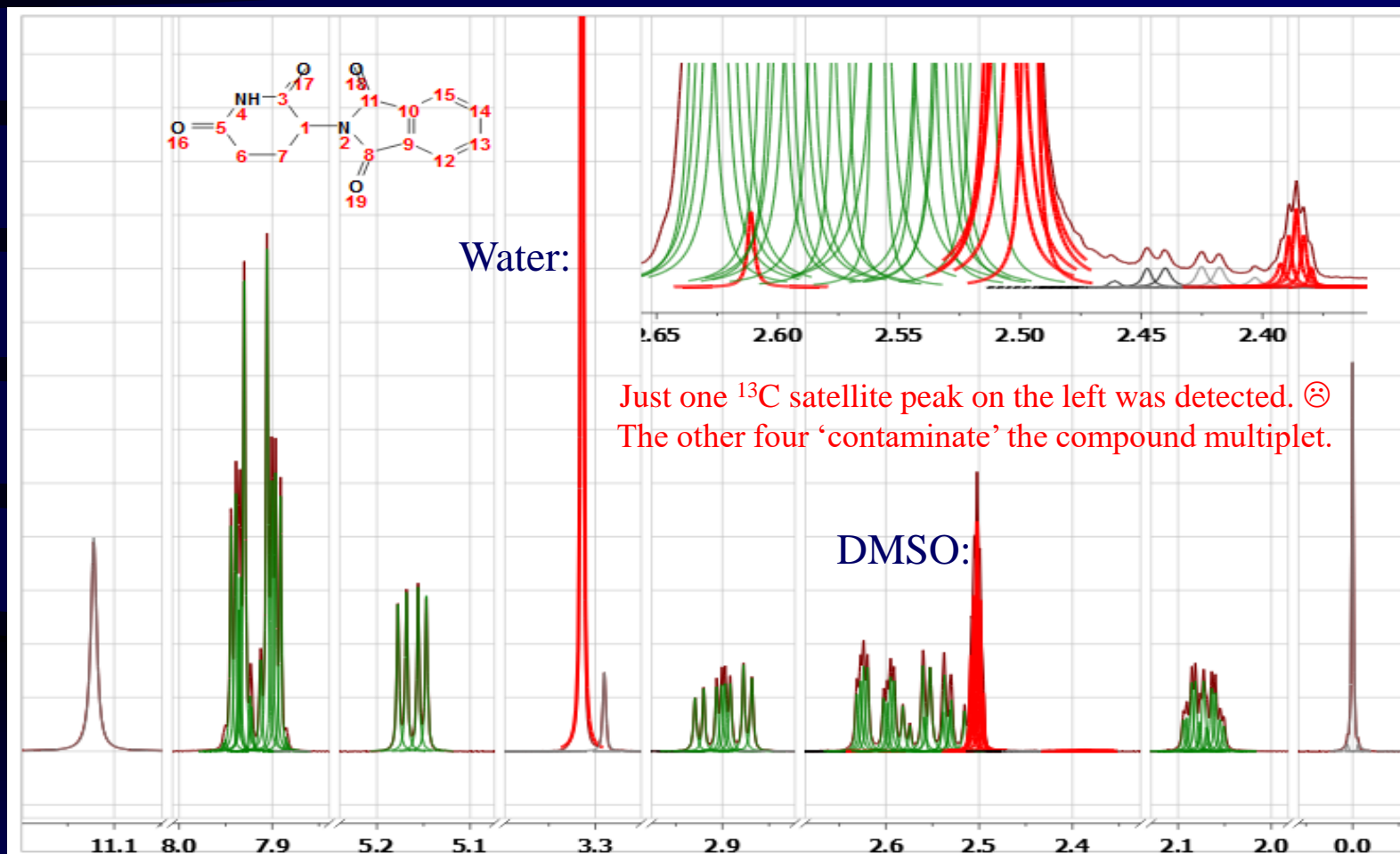
CHCl_3 identification in an overlapping aromatic multiplet



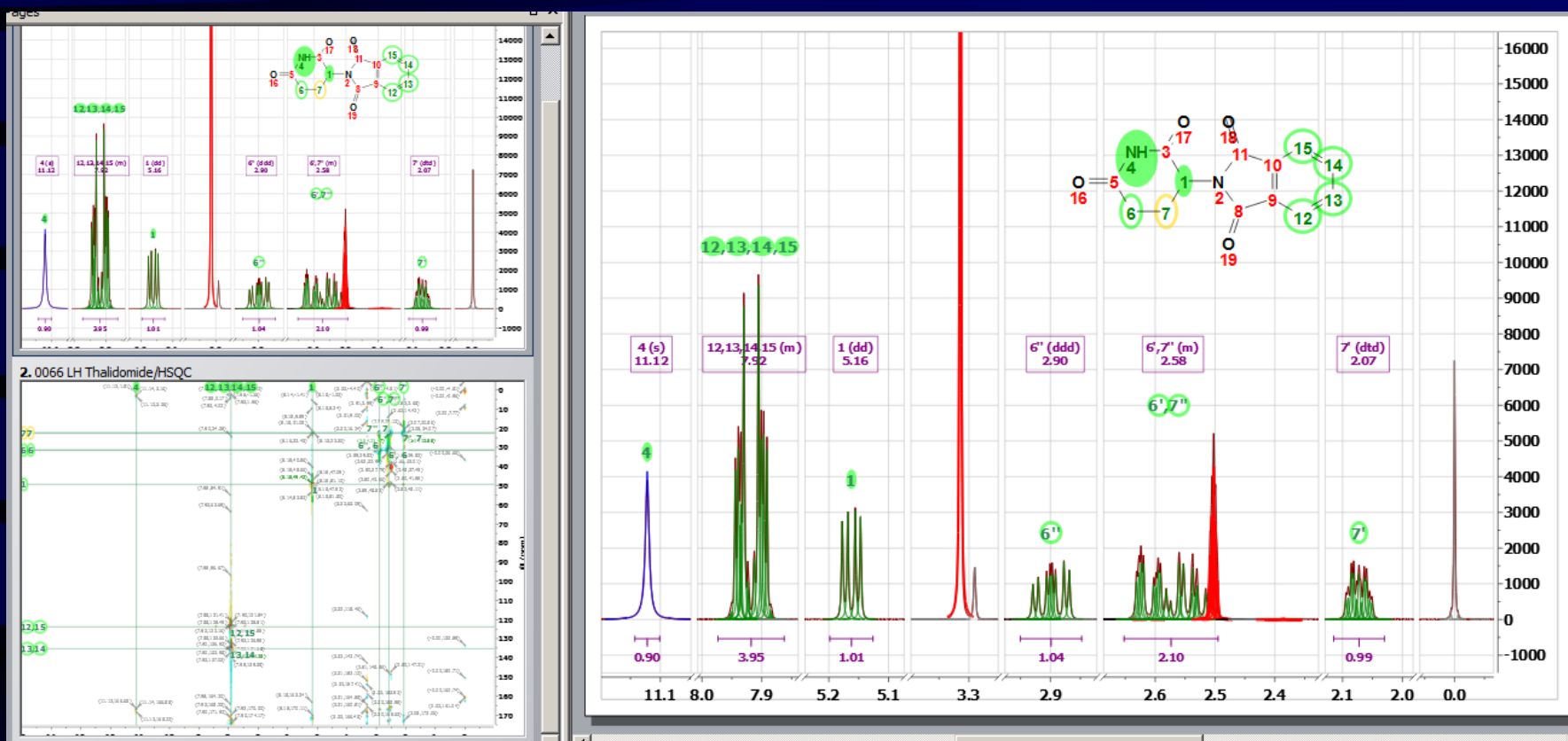
It uses even the ^{13}C satellites (209.25 Hz apart) and their isotopic shift (the satellite pair center is -2.67 ppb from the main peak)

Automatic peaks editing example

DMSO identification (Thalidomid 600 MHz)

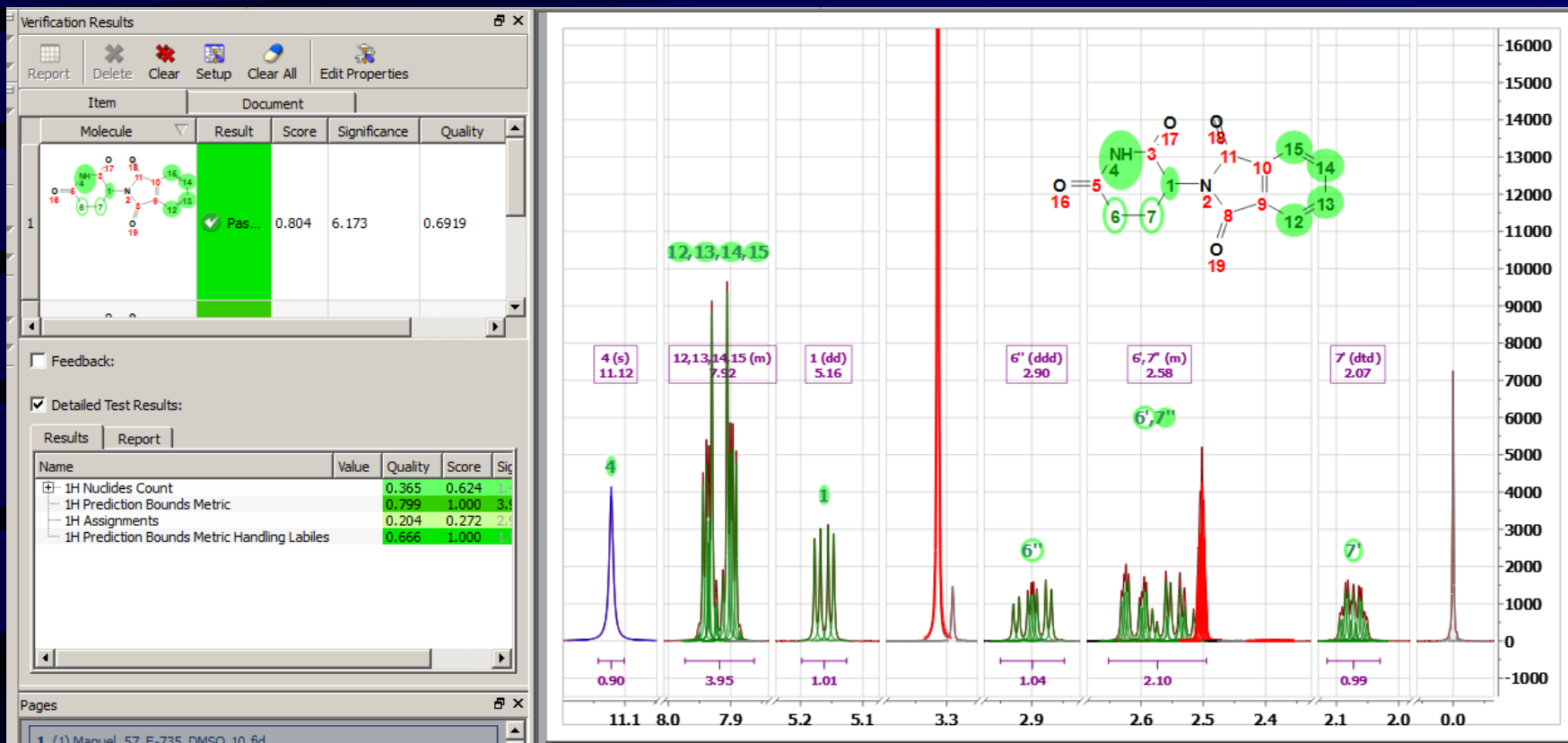


Higher level NMR data evaluation tasks: Automatic Assignments (Thalidomid 600 MHz)



Notice the correctly identified labile peak.
Labiles are often a big pain in ... They can be fuzzier than hell!

Higher level NMR data evaluation tasks (continued): Automatic Structure Verification



Etc.... **data evaluation tasks never finish** (they grow in a fractal way)

... data evaluation tasks never finish ...

OK, But !!!

At every new step new uncertainties arise and get added.
And the 'NMR master books' never tell the whole story!

Since, at each stage even the input data are fuzzy, what
about **fuziness quantification** and **fuzziness propagation**.

This is an a huge emerging area, needing more research.
NMR is just a little niche in it, but a nice example, too.

The Scoring System

Among a dozen or so algorithms that I have developed (so far unpublished) a prominent role is held by the *mathematical concept of scoring system*.

About votes and their Significances

A *scoring system entry* is a pair of real numbers associated with a query which was ‘voted’ upon by a ‘voter’ (or simply a test function):

$$\{\textit{vote}, \textit{significance}\}$$

where -1.0 (total rejection) \leq *vote* \leq $+1.0$ (total acceptance).

In general, the *significance* is an attribute of the ‘voter’
(0.0 ~ idiot, 2.0 ~ normal, 10.0 ~ expert).

The comparison of a scoring system to a committee set-up to vote on an issue is very appropriate (and even practically viable).

In such a context the *entries* are the votes cast by the voters, each vote accompanied by the particular voter’s ‘rating’.

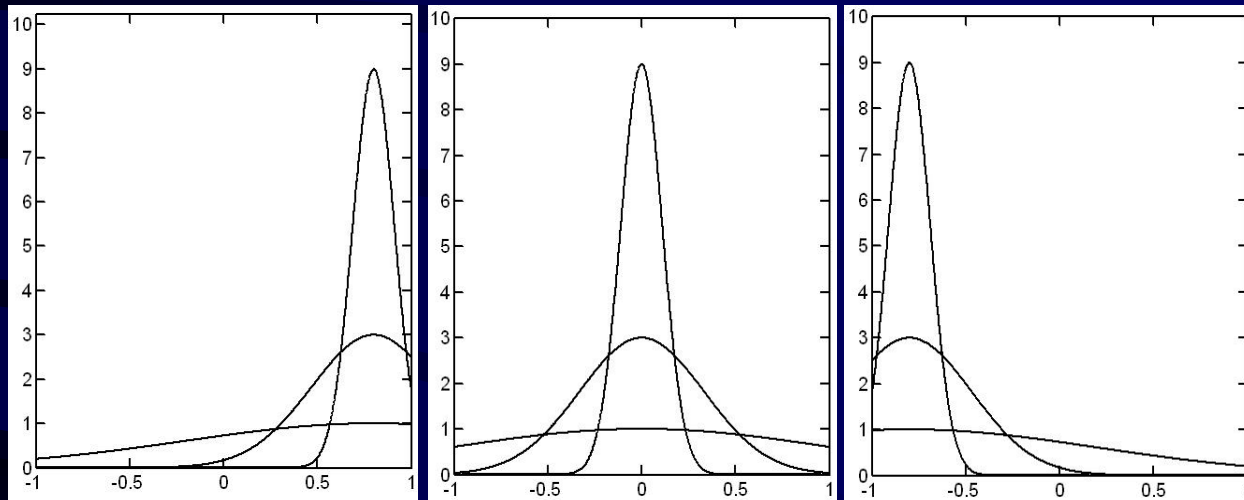
A graphical representation of a Scoring System

The graph below shows nine different entries into a scoring system, represented as peaks.

For each peak, its location corresponds to the *vote*, and its height to its *significance*.

The peak height is inversely proportional to its width since its integral must be 1.0

(*some* vote must be cast, anyway).



Positive votes

Undecided:

Insufficient data
and/or
Incapable voter

Negative votes

Combining votes and a Scoring System output

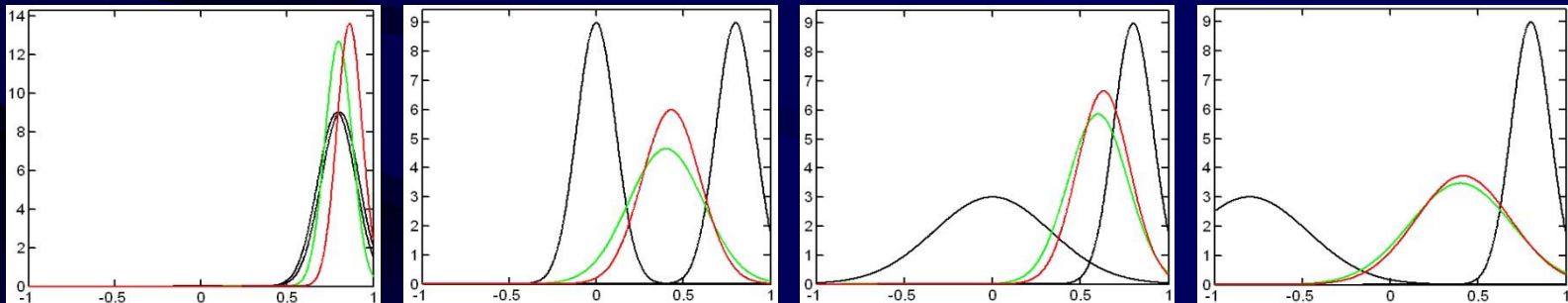
A central idea is that a scoring system output is again a $\{vote, significance\}$ pair.

This means that any number of such pairs can be combined into a single one,
(and N committees can form a super-committee).

To enable this game, we need a mathematical @ operator of the type
 $\{vote_1, significance_1\} @ \{vote_2, significance_2\} = \{vote, significance\}$

The operator must give reasonable results and also satisfy several constraints.
The mathematical problem was solved [almost*] satisfactorily.

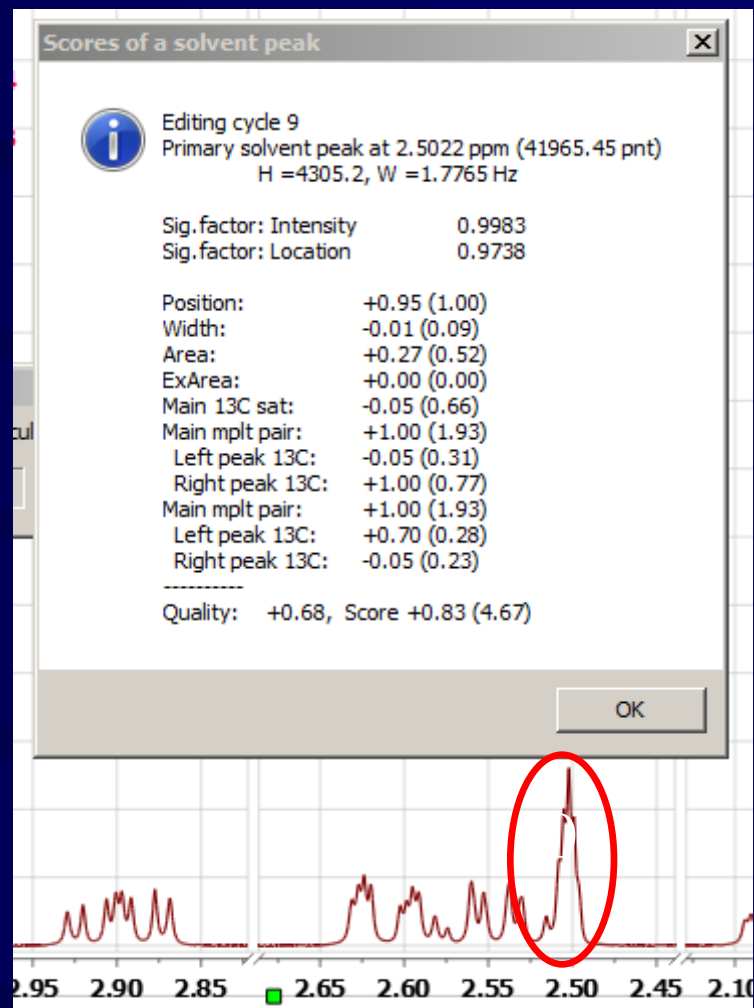
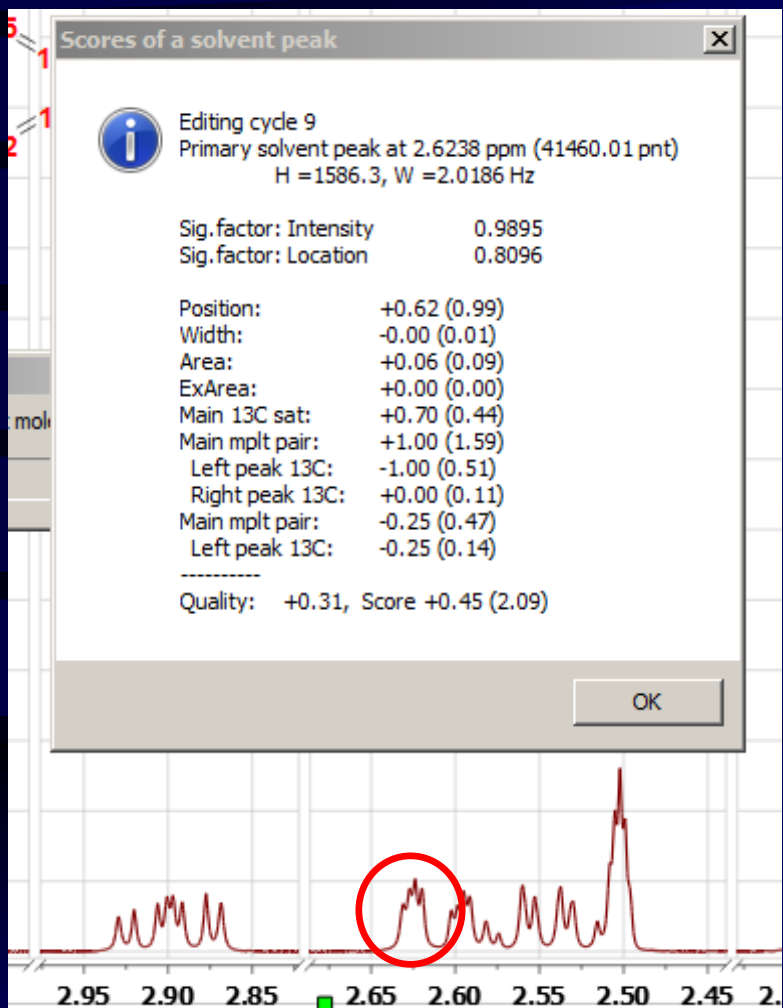
Examples:



*Note: A mathematically perfect, non-approximate solution does not exist

Example: Using a Scoring System to locate a solvent

Each peak is scored on being the pivot peak of the primary solvent (DMSO):



Usage of Scoring Systems: a summary

Mnova Data Processing package uses Scoring systems pervasively, starting from Peaks List Editing up to more complex tasks.

Typically over 1000 scoring systems are set-up and over 10000 votes are cast in a single run of any advanced NMR data evaluation task.

It often looks almost as a magic to witness how the scoring-system algorithm can arrive at a correct deduction from very fuzzy data.

This type of technologies will no doubt flourish in near future also also outside NMR, especially in the social area and big data areas. There are other such emerging ideas (e.g., artificial intuition).

Side remarks

In a recent talk I have concluded with a slide like this
The NMR Spectroscopist in XXII Century ==>



1) In 10 years time, given a set of spectra (say ^1H , ^{13}C , HSQC, COSY, HMBC) and five possible structures, will anybody go through the tedious 'manual' analysis of the thing? Obviously not, if a click will be enough to do it.

Hence, why should anybody be interested to laboriously study how to do it?

2) And will anybody still study spin Hamiltonians and product operators if absolutely anything you can do with them is in the reach of a click?

Few are versed in these things even today, and their numbers are dropping!

Thank You for Your Attention

Coworkers:

The whole of Mestrelab team, in particular:

Carlos Cobas, Felipe Seoane, Esther Vaz,
Santiago Dominguez, Maruxa Sordo,
Cristina Geadá, Pablo Monje,

...

Getting on-board:

Ester Maria Vasini

(Extra Byte might yet take a new turn 😊)

External collaborators and/or ‘collaborators’:

far too many to name (this is partially fuzzy, too)

