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:

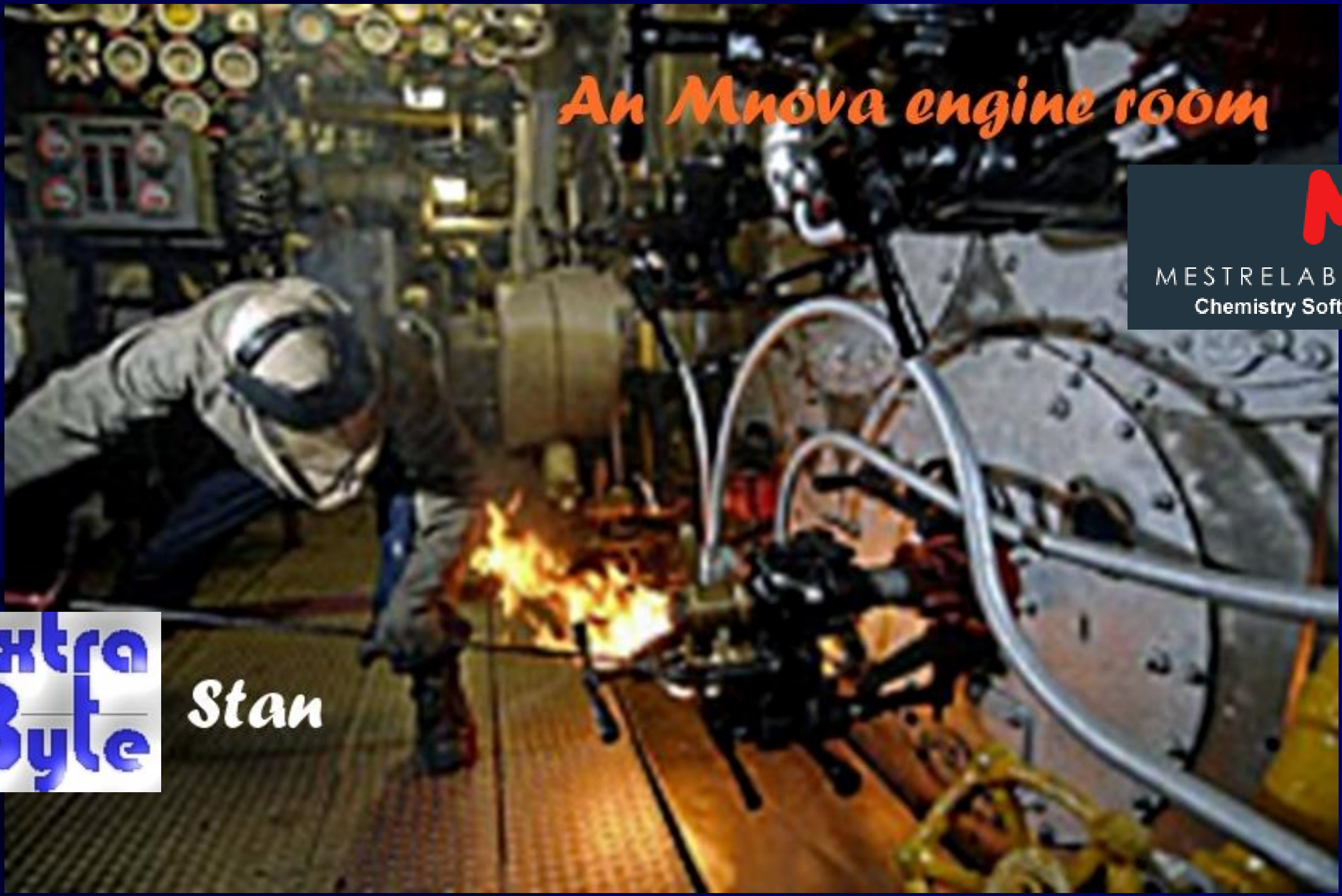
- The simulated ones are cute, hard and as sharp as it goes ☺
- The ‘real’ rest is *disgustingly soft* and as fuzzy as it goes ☹

NMR is no exception – at all levels!

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

DOI: 10.3247SL6Nmr17.006

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 occured, 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 !!!

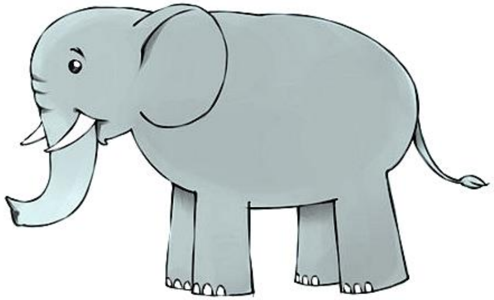
(That is really vexing!)

And I am to build software that, statistically, should beat them !??

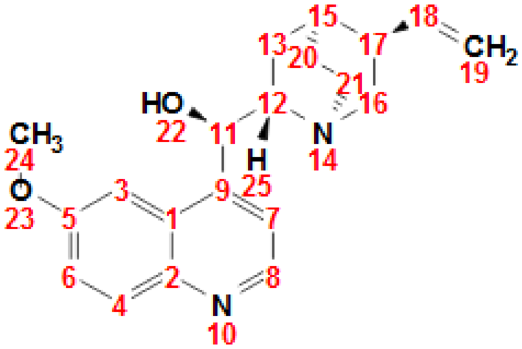
(no worry, I am in sight of the target, but still a long way to go)

Expectations and Reality (still in general)

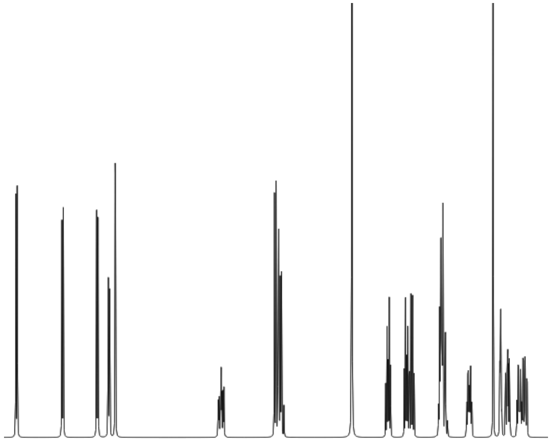
II. Reality is infinitely complex and never totally predictable. Hence, it is fuzzy!



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%.

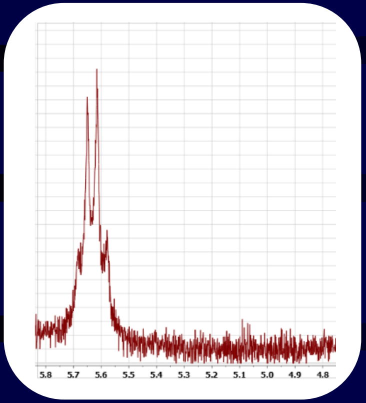
NMR fuzziness at NMR ground level (the Reception Hall)

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

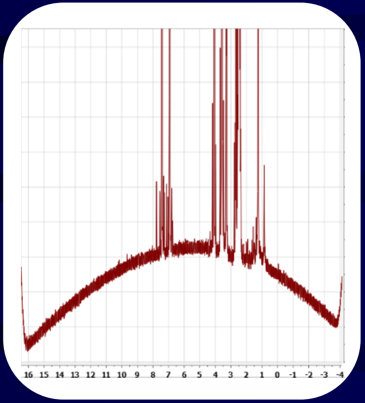
- Noise
  (from sample, probe, switching diodes, preamp, external, ...)
- Instrumental artifacts
  (dead time, imperfect shimming, rotation sidebands, spikes, ...)
- Sequence artifacts
  (ill defined expected intensities, missing and extra clusters, ...)
- Acquisition artifacts
  (fast repetition, FID truncation , limited digital resolution, ...)
- Evaluation artifacts
  (bad phase or baseline, referencing errors, DFT distorsions, ...)
- 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 ☹)

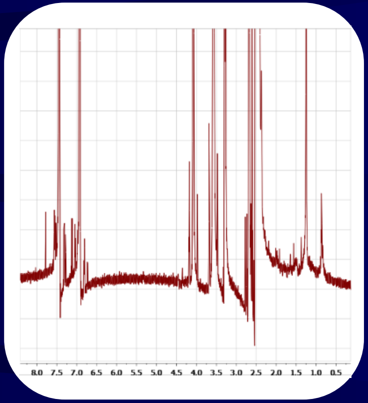
A few examples of undesirable spectral artifacts



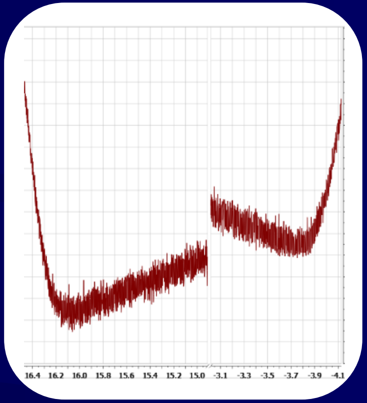
Noise



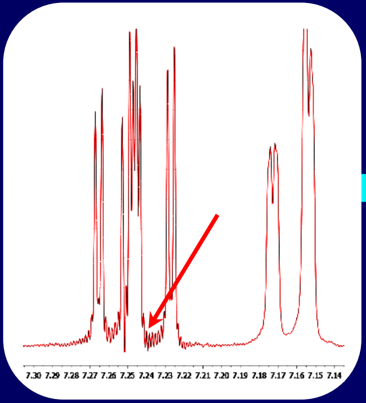
Baseline roll



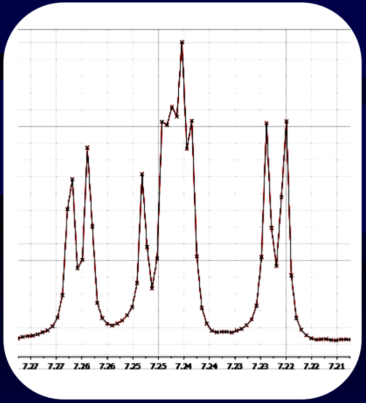
Imperfect phasing



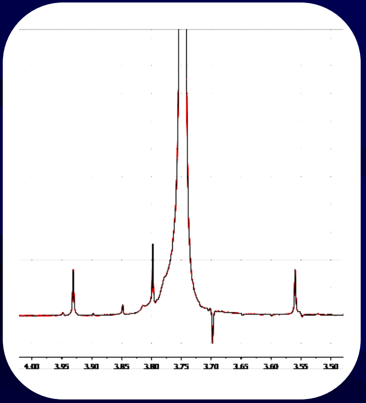
Bruker / Jeol smileys / brownies



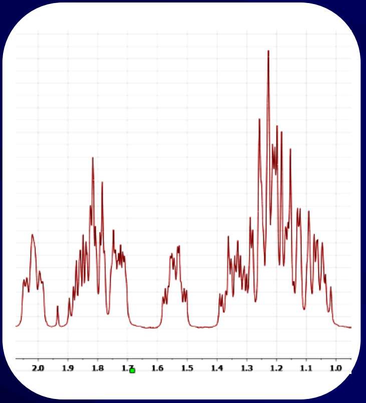
FID truncation effects



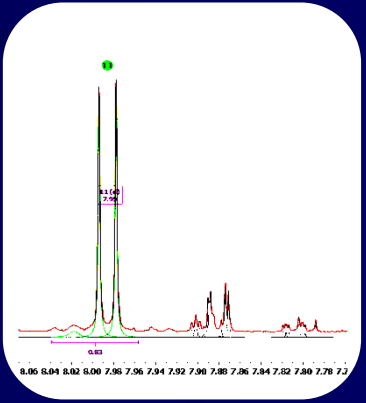
Underdigitization



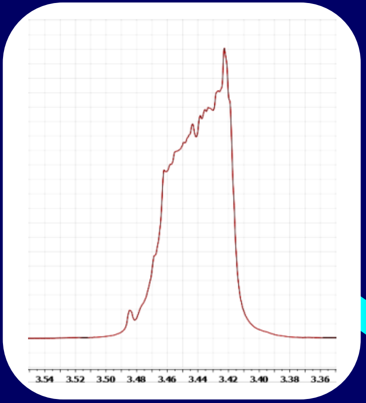
Rotation sidebands



Peaks overlap



Impurities peaks



Sample temperature drift effects

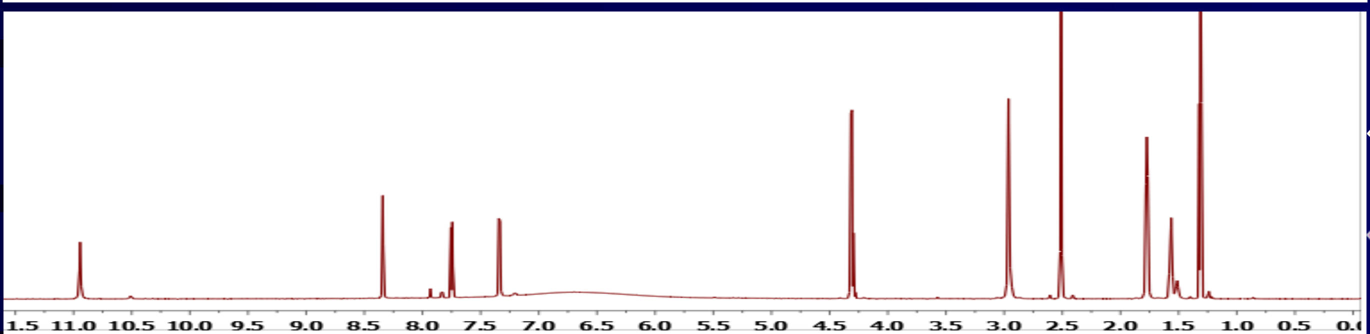
=> There is so much to get rid of before doing anything serious with a spectrum !!!

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, 13C 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!

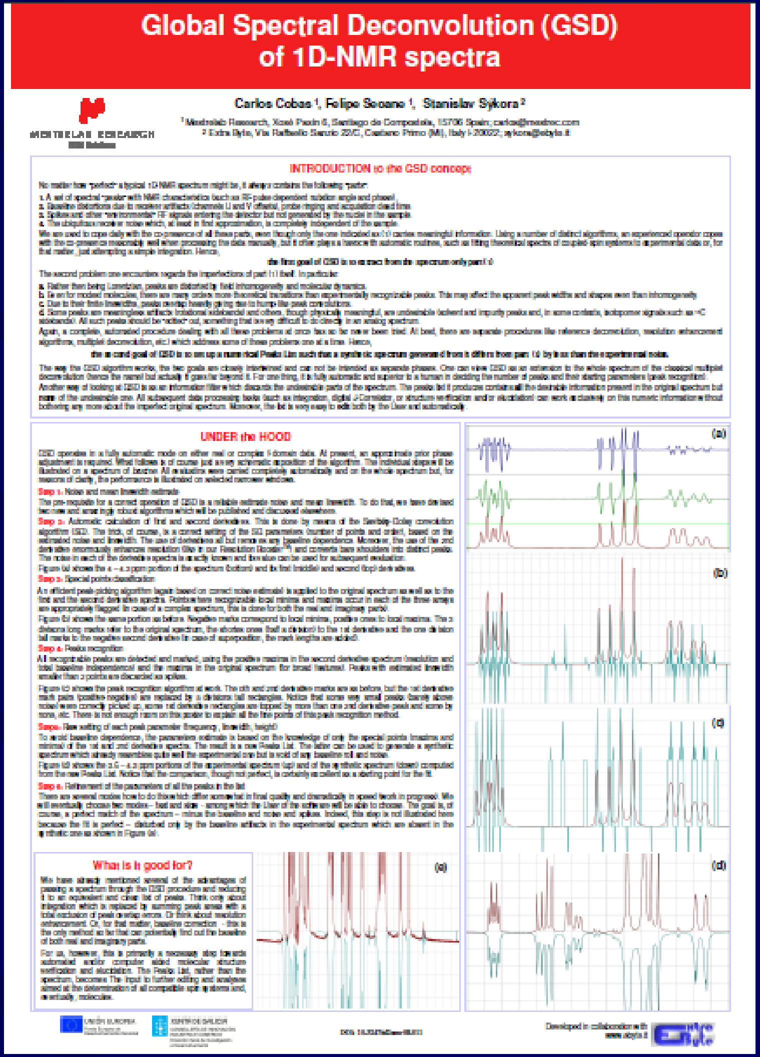
Implication: there is a big communication problem

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 Cobas [1], Felipe Seoane [1], Stanislav Sýkora [2]

[1] Mestrelab Research, Xosé Pasín 6, Santiago de Compostela, 15706 Spain; carlos@mestrec.com
[2] Extra Byte, Via Raffaello Sanzio 22/C, Castano Primo (MI), Italy I-20022; sykora@ebyte.it

INTRODUCTION to the GSD concept

[illegible]

UNDER the HOOD

[illegible]

What is it good for?

[illegible]

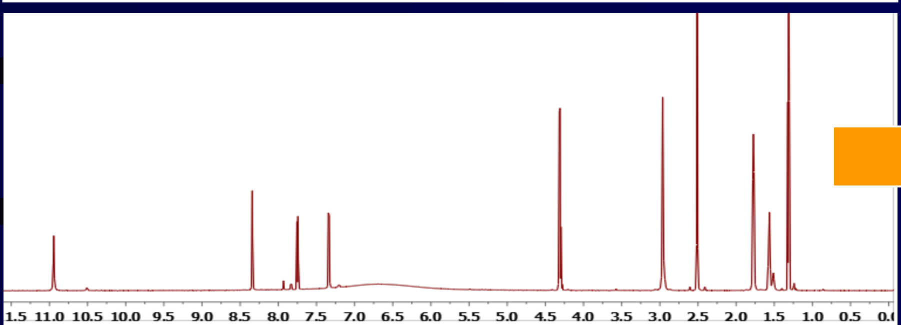
Ground floor data management (continued)

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,

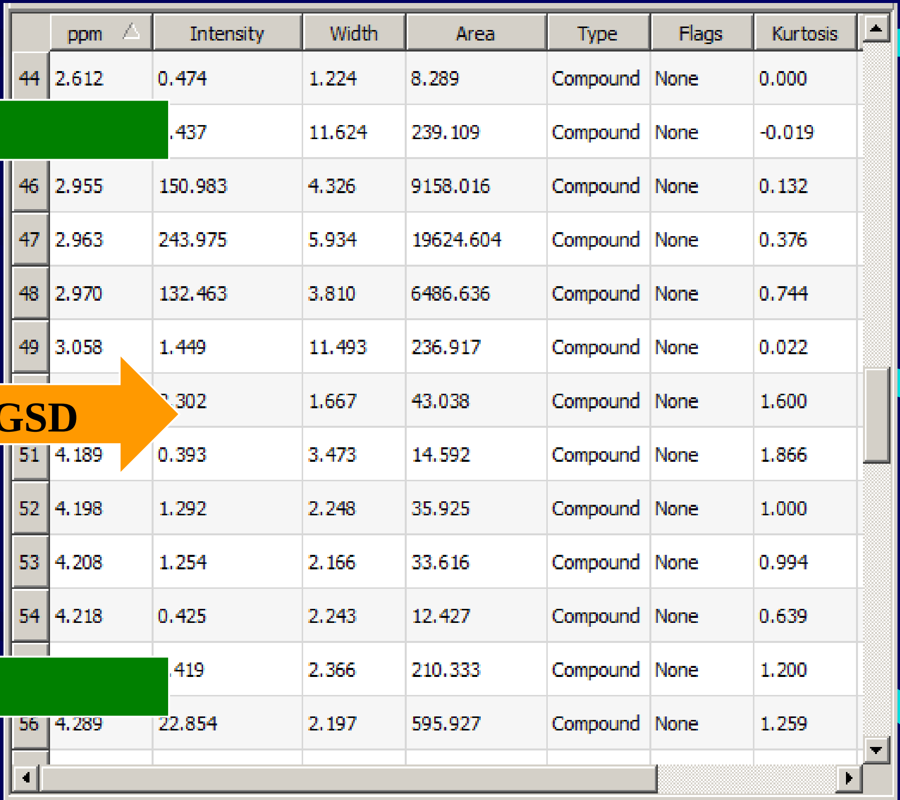
(solvent, reference, impurity,…), multiplets, ...



GSD

What does a programmer see?

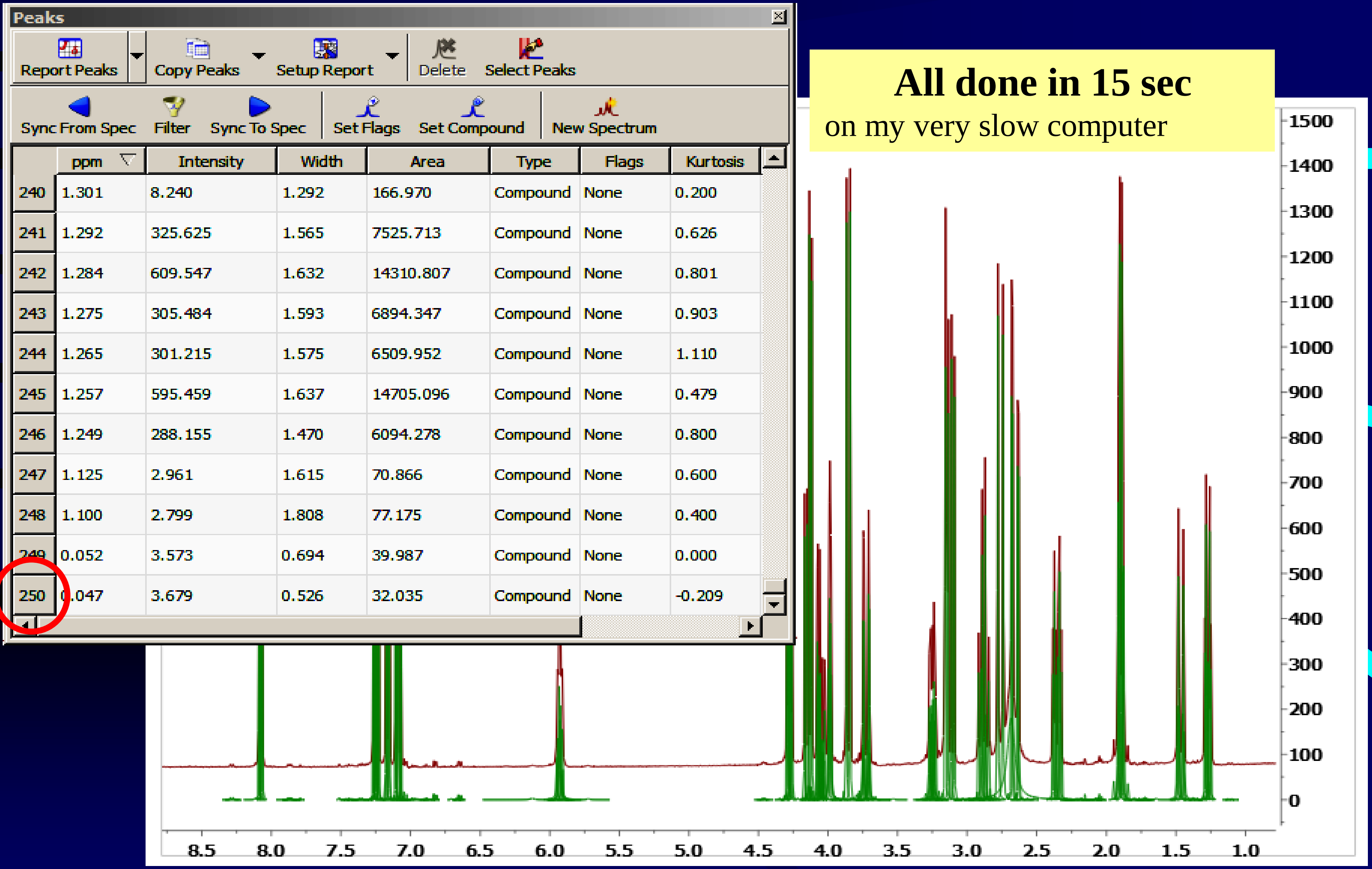
An array of … **peaks**, finally! That's **GREAT**



	ppm	Intensity	Width	Area	Type	Flags	Kurtosis
44	2.612	0.474	1.224	8.289	Compound	None	0.000
		.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
		.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
		.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)



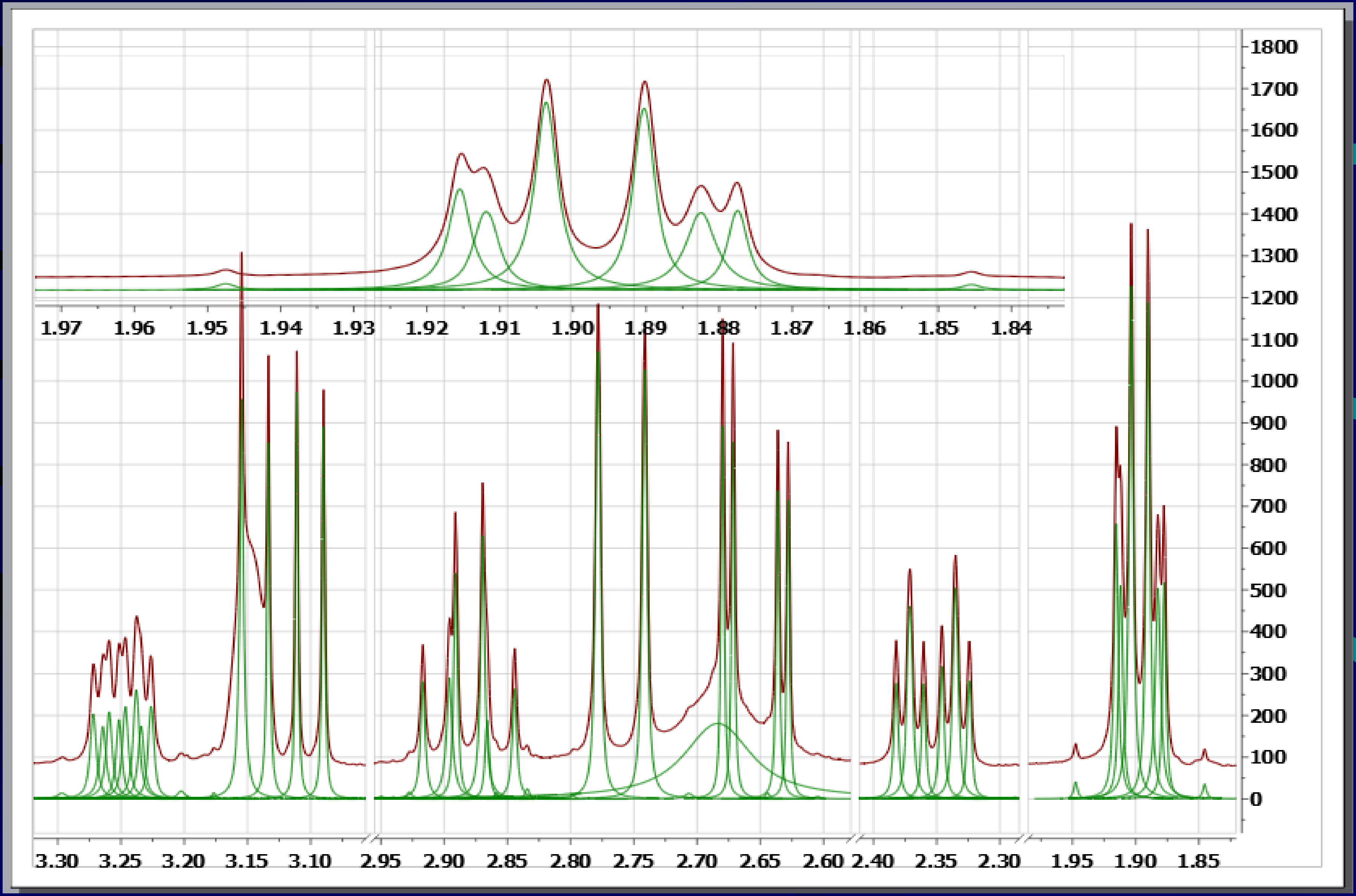
	ppm	Intensity	Width	Area	Type	Flags	Kurtosis
240	1.301	8.240	1.292	166.970	Compound	None	0.200
241	1.292	325.625	1.565	7525.713	Compound	None	0.626
242	1.284	609.547	1.632	14310.807	Compound	None	0.801
243	1.275	305.484	1.593	6894.347	Compound	None	0.903
244	1.265	301.215	1.575	6509.952	Compound	None	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

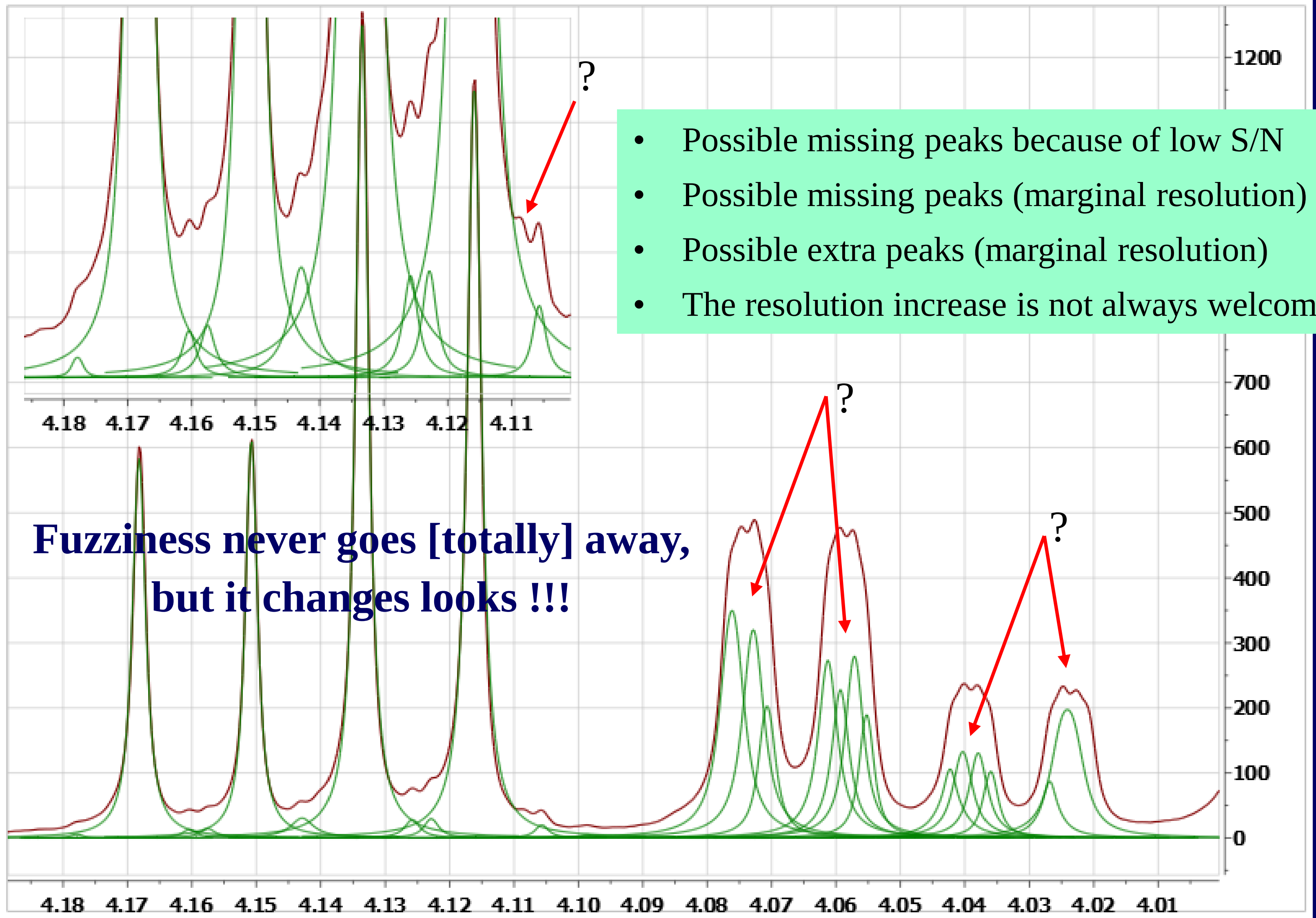
F

Examples of peaks detection (WOW)



F

Examples of peaks detection (GSD limits and artifacts)



- Possible missing peaks because of low S/N
- Possible missing peaks (marginal resolution)
- Possible extra peaks (marginal resolution)
- The resolution increase is not always welcome

Fuzziness never goes [totally] away, but it changes looks !!!

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:

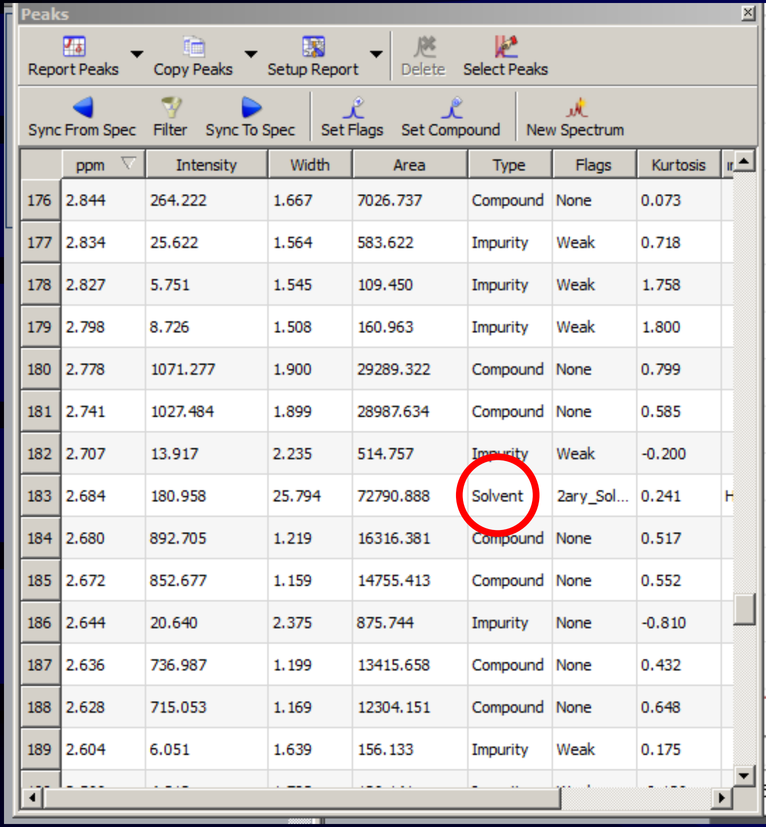
- compound,
- primary or secondary solvent,
- potential labile,
- 13C satellite,
- valid member of a multiplet,
- impurity,
- S or Q reference,
- artifact,
- etc.

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

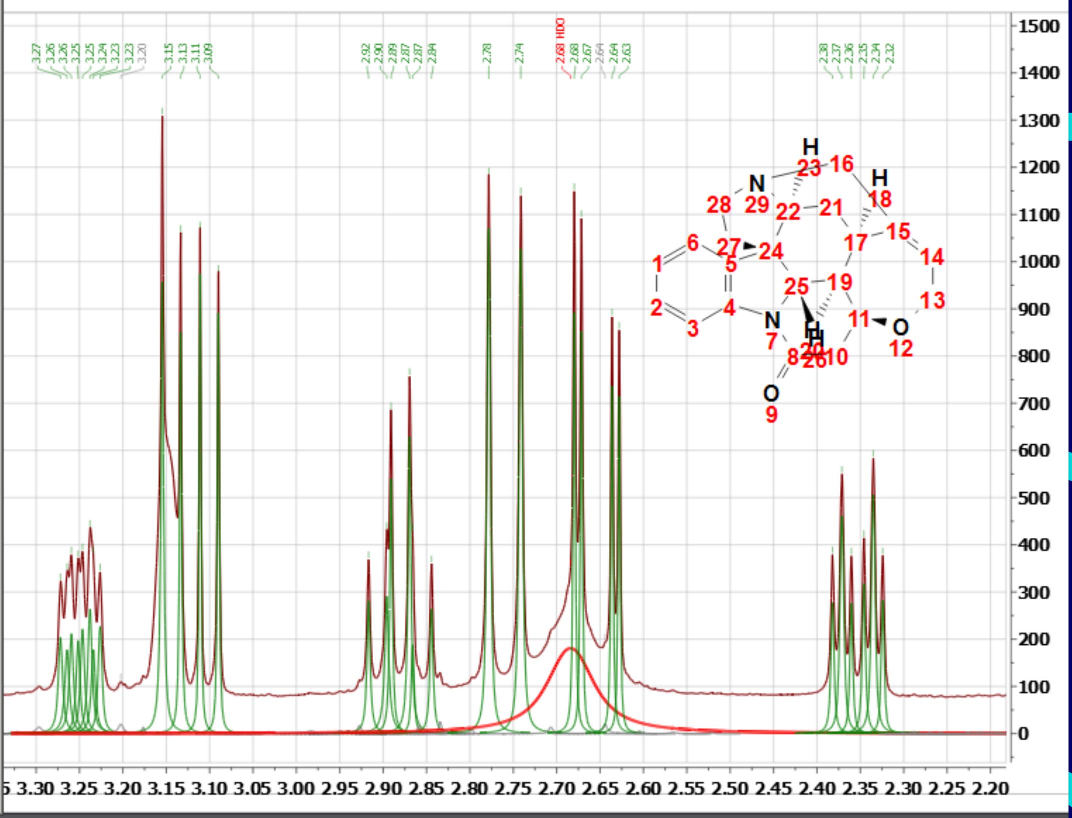
All these are tasks which, unavoidably, bring-in fuzziness of their own.

Automatic peaks editing example

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



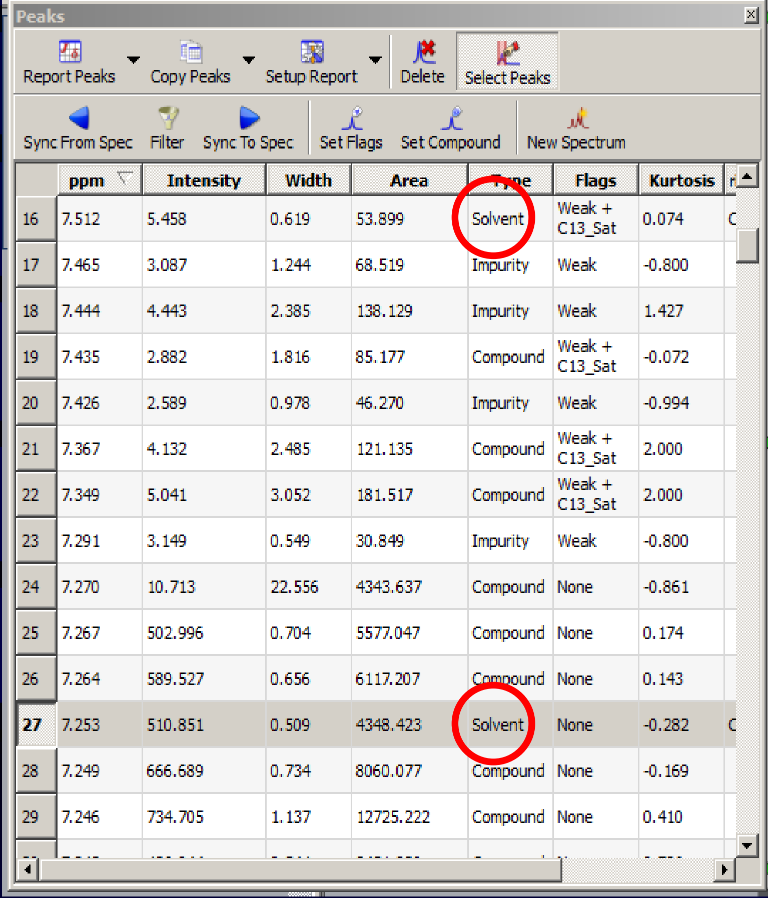
	ppm	Intensity	Width	Area	Type	Flags	Kurtosis	ir
176	2.844	264.222	1.667	7026.737	Compound	None	0.073	
177	2.834	25.622	1.564	583.622	Impurity	Weak	0.718	
178	2.827	5.751	1.545	109.450	Impurity	Weak	1.758	
179	2.798	8.726	1.508	160.963	Impurity	Weak	1.800	
180	2.778	1071.277	1.900	29289.322	Compound	None	0.799	
181	2.741	1027.484	1.899	28987.634	Compound	None	0.585	
182	2.707	13.917	2.235	514.757	Impurity	Weak	-0.200	
183	2.684	180.958	25.794	72790.888	Solvent	2ary_Sol...	0.241	H
184	2.680	892.705	1.219	16316.381	Compound	None	0.517	
185	2.672	852.677	1.159	14755.413	Compound	None	0.552	
186	2.644	20.640	2.375	875.744	Impurity	None	-0.810	
187	2.636	736.987	1.199	13415.658	Compound	None	0.432	
188	2.628	715.053	1.169	12304.151	Compound	None	0.648	
189	2.604	6.051	1.639	156.133	Impurity	Weak	0.175	



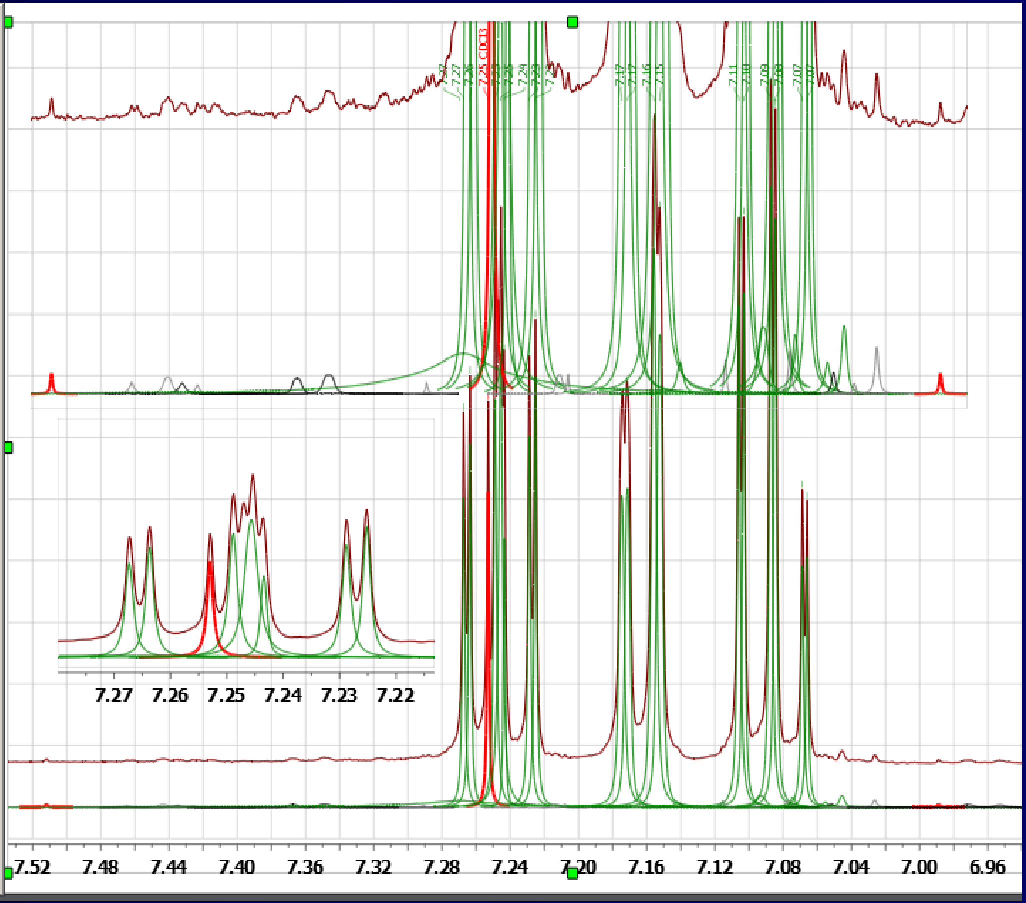
Peaks editing uses greedily any available info (1H, HSQC, molecular predictions, ...).
It is the first plank in the evaluation hierarchy where specific NMR know-how is used.

Automatic peaks editing example

$CHCl_3$ identification in an overlapping aromatic multiplet



	ppm	Intensity	Width	Area	Type	Flags	Kurtosis	r
16	7.512	5.458	0.619	53.899	Solvent	Weak + C13_Sat	0.074	C
17	7.465	3.087	1.244	68.519	Impurity	Weak	-0.800	
18	7.444	4.443	2.385	138.129	Impurity	Weak	1.427	
19	7.435	2.882	1.816	85.177	Compound	Weak + C13_Sat	-0.072	
20	7.426	2.589	0.978	46.270	Impurity	Weak	-0.994	
21	7.367	4.132	2.485	121.135	Compound	Weak + C13_Sat	2.000	
22	7.349	5.041	3.052	181.517	Compound	Weak + C13_Sat	2.000	
23	7.291	3.149	0.549	30.849	Impurity	Weak	-0.800	
24	7.270	10.713	22.556	4343.637	Compound	None	-0.861	
25	7.267	502.996	0.704	5577.047	Compound	None	0.174	
26	7.264	589.527	0.656	6117.207	Compound	None	0.143	
27	7.253	510.851	0.509	4348.423	Solvent	None	-0.282	C
28	7.249	666.689	0.734	8060.077	Compound	None	-0.169	
29	7.246	734.705	1.137	12725.222	Compound	None	0.410	



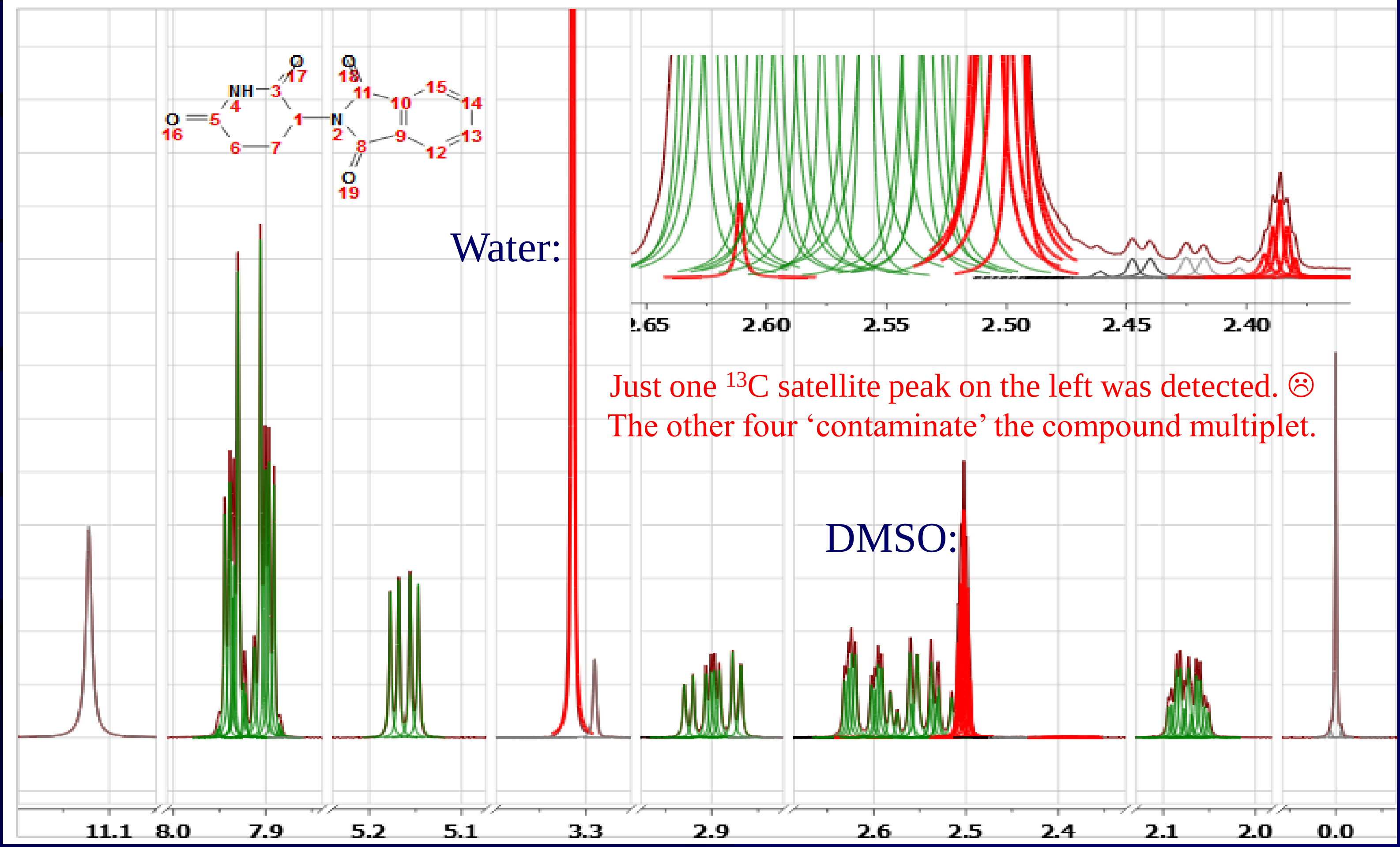
It uses even the ^{13}C satellites (209.25 Hz apart) and their isotopic shift

(the satellite pair center is -2.67 ppb from the main peak)

F

Automatic peaks editing example

DMSO identification (Thalidomid 600 MHz)



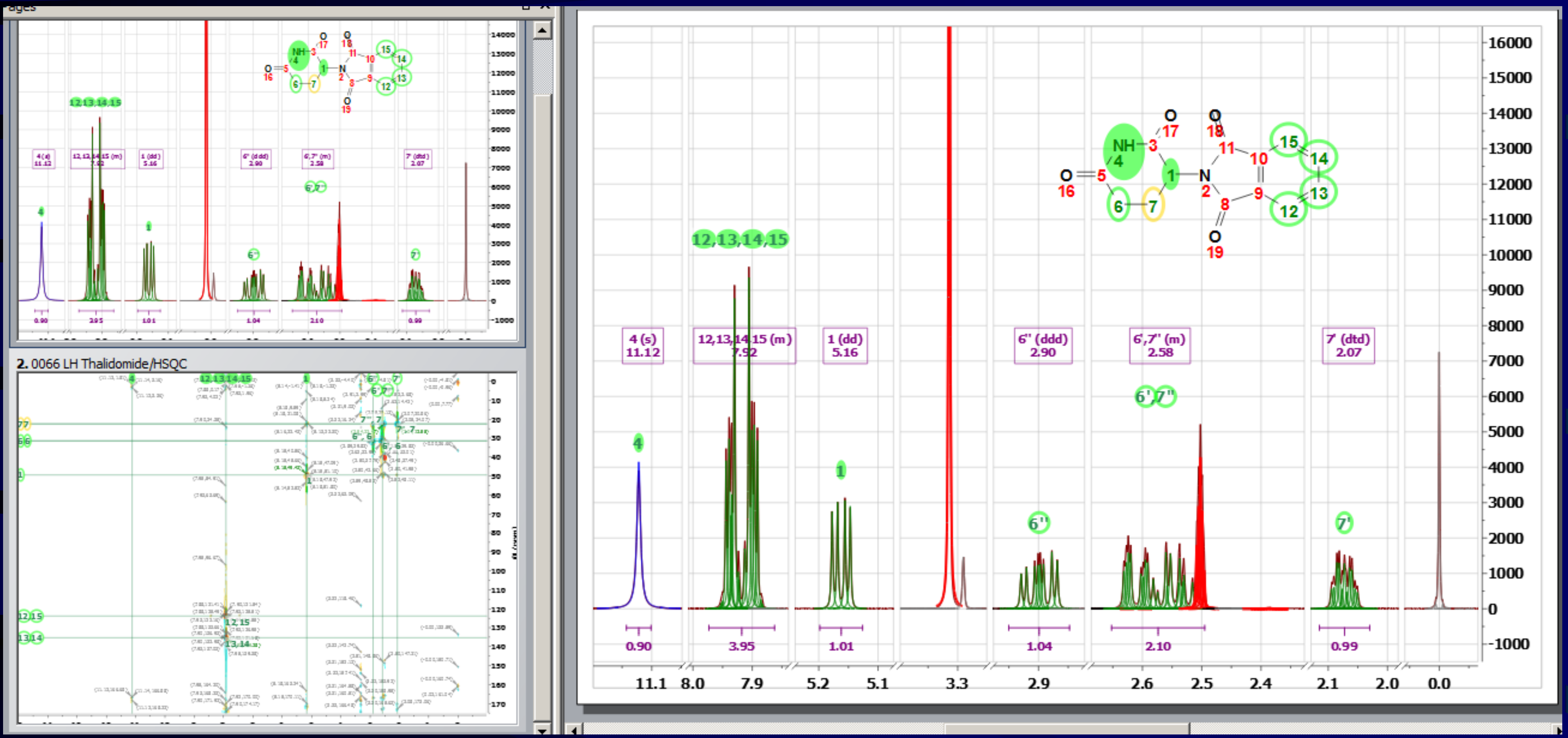
Water:

DMSO:

Just one ^{13}C satellite peak on the left was detected. ☹
The other four ‘contaminate’ the compound multiplet.

Higher level NMR data evaluation tasks:

Automatic Assignments (Thalidomid 600 MHz)

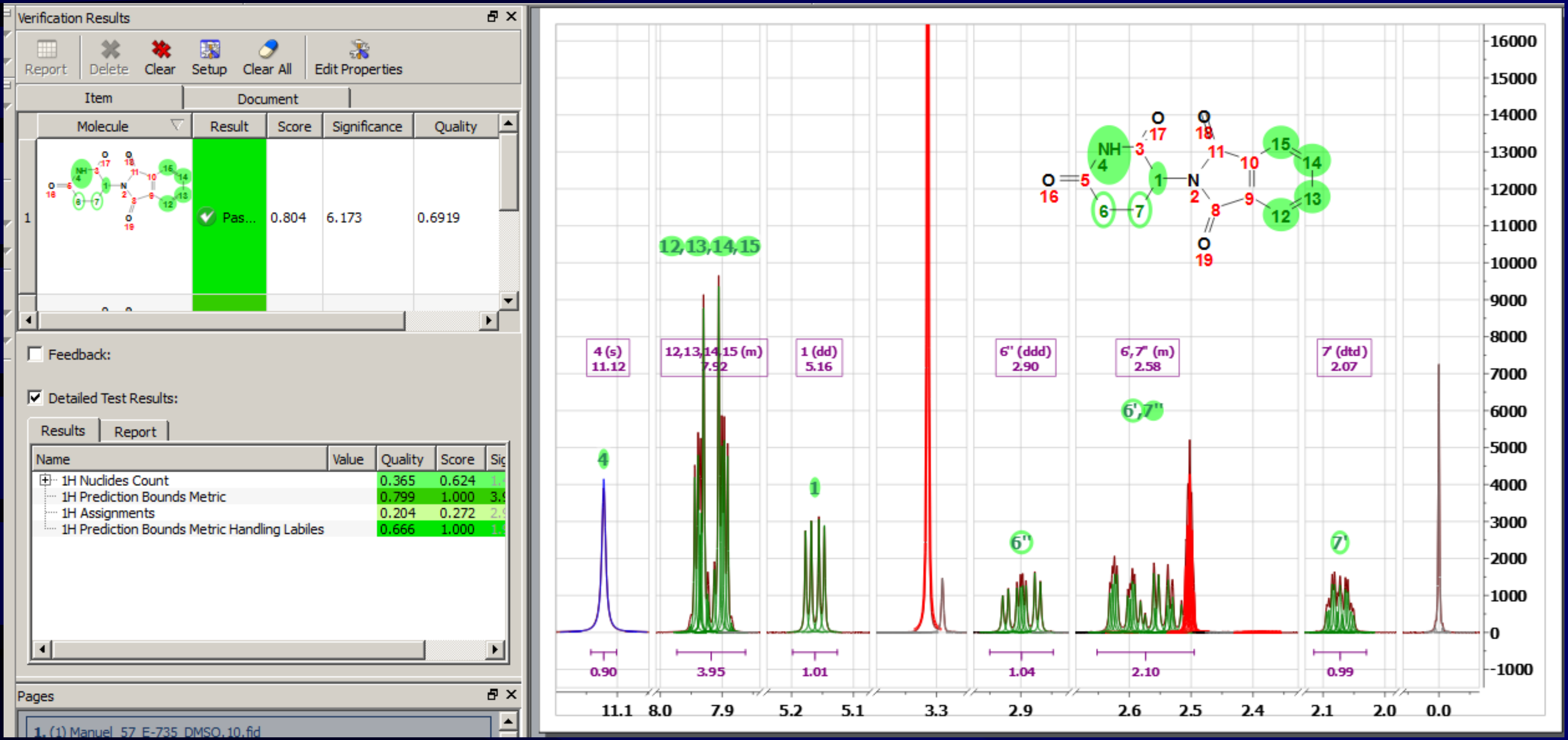


2. 0066 LH Thalidomide/HSQC

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



Verification Results

	Molecule	Result	Score	Significance	Quality
1		✓ Pas...	0.804	6.173	0.6919

☐ Feedback:

☑ Detailed Test Results:

Name	Value	Quality	Score	Sig
⊞ 1H Nuclides Count		0.365	0.624	1.
1H Prediction Bounds Metric		0.799	1.000	3.9
1H Assignments		0.204	0.272	2.9
1H Prediction Bounds Metric Handling Labiles		0.666	1.000	1.

Pages

1. (1) Manuel_57_E-735_DMSO.10.fid

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

F

… 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):

{*vote*, *significance*}

where -1.0 (total rejection) ≤ vote ≤ +1.0 (total acceptance).

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

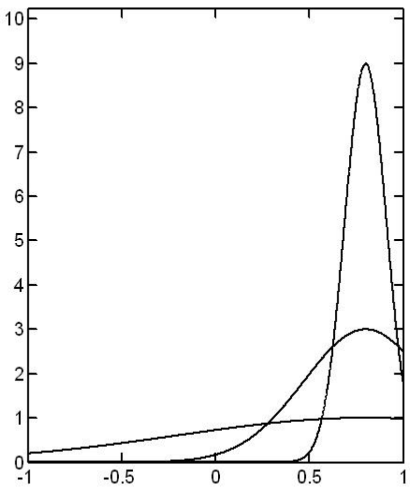
The comparison of a scoring system to a committee set-up to vote on an issue is very appropriate (and even practically viable).

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

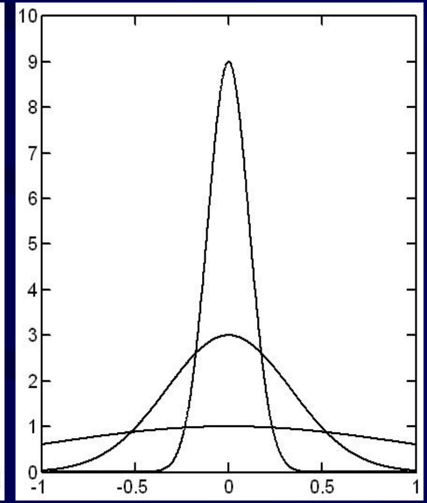
A graphical representation of a Scoring System

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).

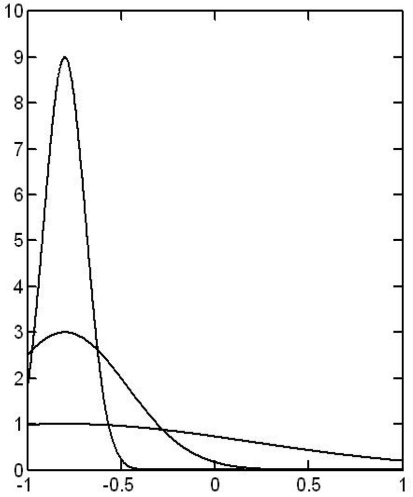


Positive votes



Undecided:

Insufficient data
and/or
Incapable voter



Negative votes

Combining votes and a Scoring System output

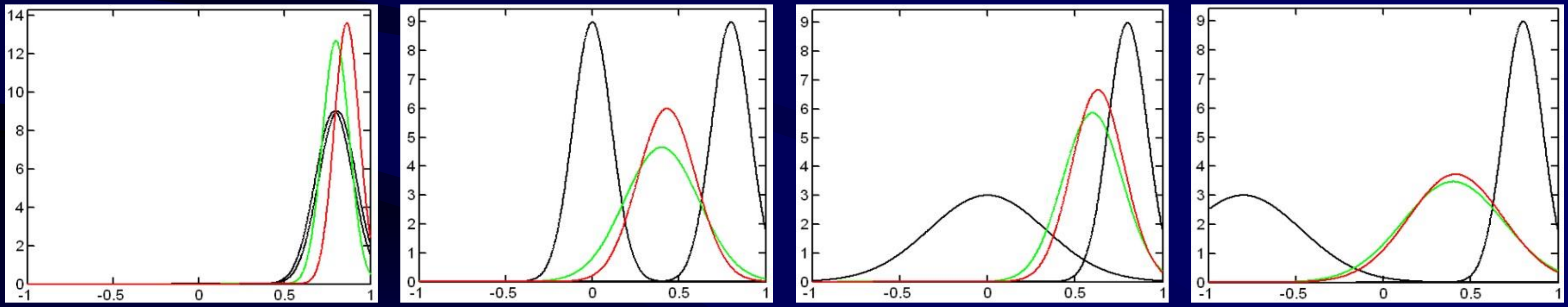
A central idea is that a scoring system output is again a {*vote*, *significance*} pair.

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

$$\{vote_1, significance_1\} @ \{vote_2, significance_2\} = \{vote, significance\}$$

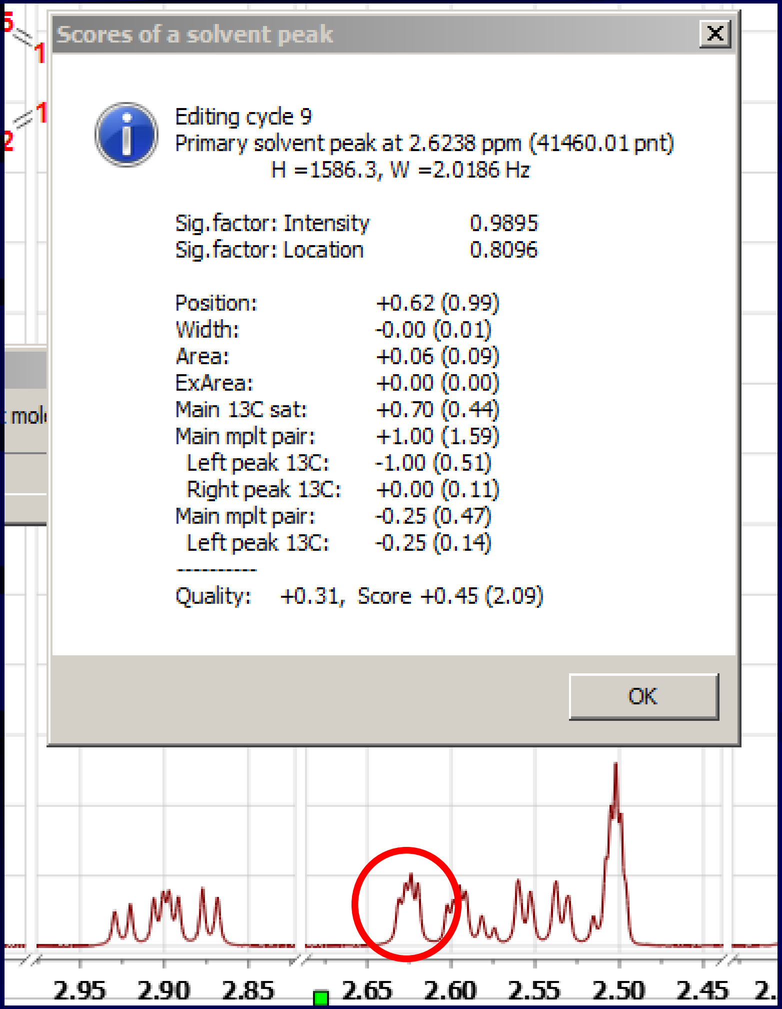
The operator must give reasonable results and also satisfy several constraints.
The mathematical problem was solved [almost*] satisfactorily.
Examples:



*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):

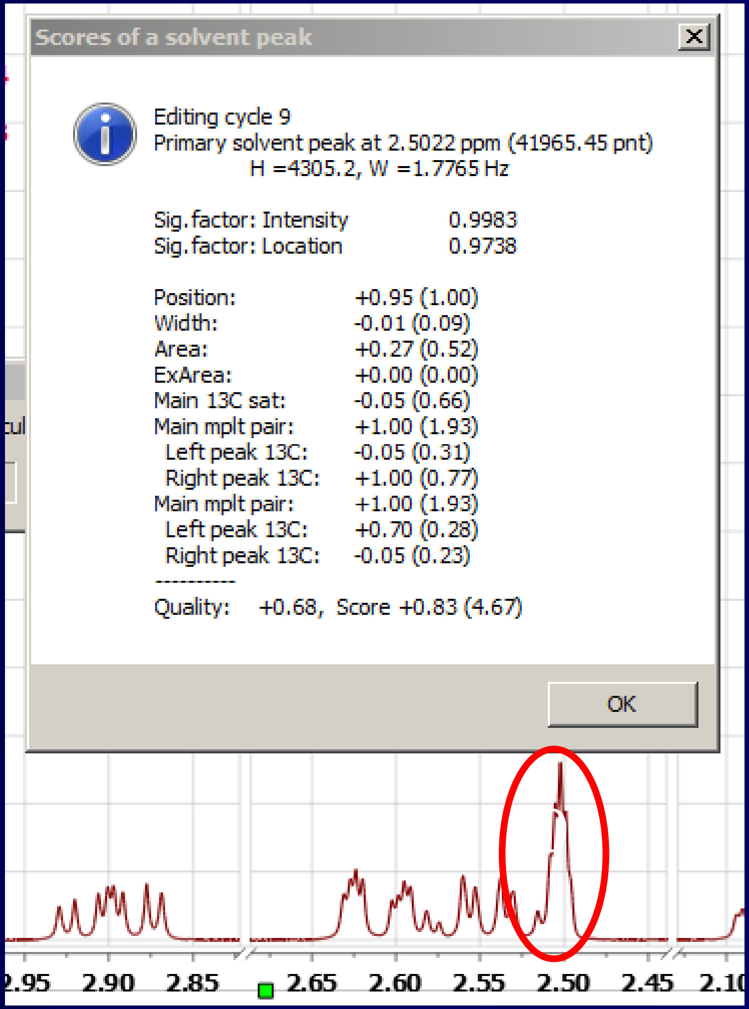


Scores of a solvent peak

```
Editing cycle 9
Primary solvent peak at 2.6238 ppm (41460.01 pnt)
        H =1586.3, W =2.0186 Hz

Sig.factor: Intensity          0.9895
Sig.factor: Location           0.8096

Position:               +0.62 (0.99)
Width:                  -0.00 (0.01)
Area:                   +0.06 (0.09)
ExArea:                 +0.00 (0.00)
Main 13C sat:           +0.70 (0.44)
Main mplt pair:         +1.00 (1.59)
  Left peak 13C:        -1.00 (0.51)
  Right peak 13C:       +0.00 (0.11)
Main mplt pair:         -0.25 (0.47)
  Left peak 13C:        -0.25 (0.14)
----------
Quality:   +0.31,  Score +0.45 (2.09)
```



Scores of a solvent peak

```
Editing cycle 9
Primary solvent peak at 2.5022 ppm (41965.45 pnt)
        H =4305.2, W =1.7765 Hz

Sig.factor: Intensity          0.9983
Sig.factor: Location           0.9738

Position:               +0.95 (1.00)
Width:                  -0.01 (0.09)
Area:                   +0.27 (0.52)
ExArea:                 +0.00 (0.00)
Main 13C sat:           -0.05 (0.66)
Main mplt pair:         +1.00 (1.93)
  Left peak 13C:        -0.05 (0.31)
  Right peak 13C:       +1.00 (0.77)
Main mplt pair:         +1.00 (1.93)
  Left peak 13C:        +0.70 (0.28)
  Right peak 13C:       -0.05 (0.23)
----------
Quality:   +0.68,  Score +0.83 (4.67)
```

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 ^{1}H, ^{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 Geada, 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)

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