A new paradigm streamlining the analysis of NMR spectra of small molecules:

We demand rigidly defined areas of doubt and uncertainty!

(A citation from Douglas Adams' *The Hitchhikers Guide to the Galaxy*)

Stanislav Sykora

(google «stan nmr»)

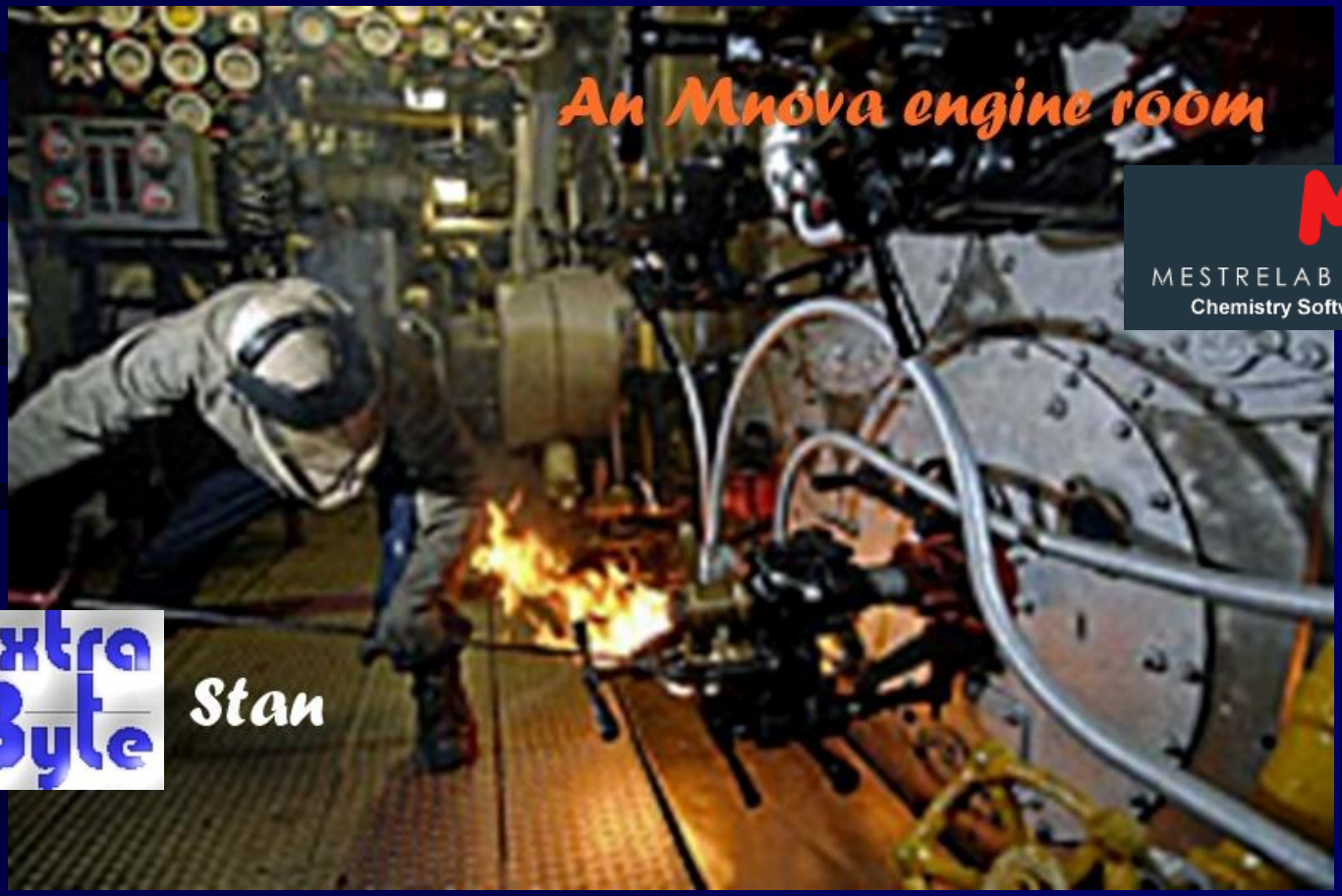
Extra Byte (www.ebyte.it)

This online document's DOI : **10.3247/SL4Nmr13.007**

Presented at the meeting of GERMN, the Spanish NMR Discussion Group, Santiago de Compostela, Spain, 27 Sep 2013

Why do I work with the Mnova guys?

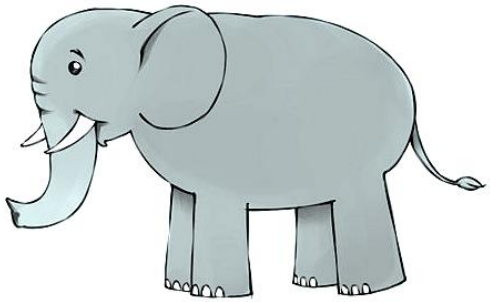
I guess it's because they are fun to work with ...



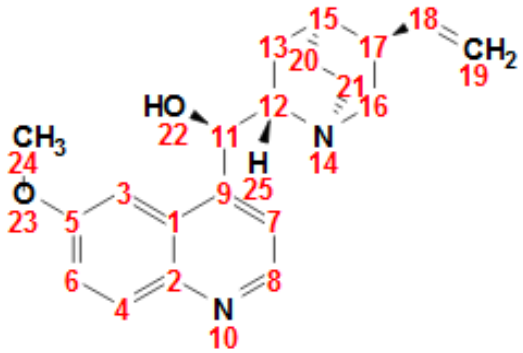
... and also because Small Molecules Are Still so Hot !

No REAL thing can be totally predictable!

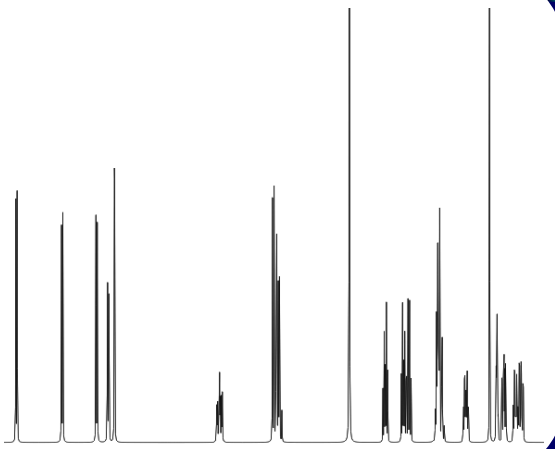
Everything real is infinitely complex and fuzzy!



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



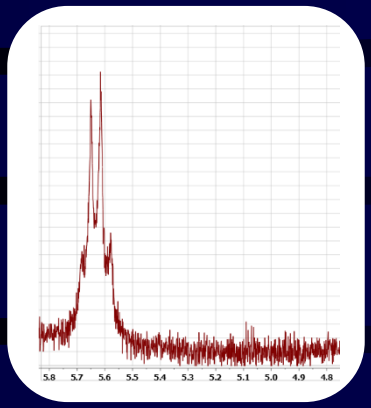
This is NOT a **real** molecule! It's just a structural sketch of a molecule.



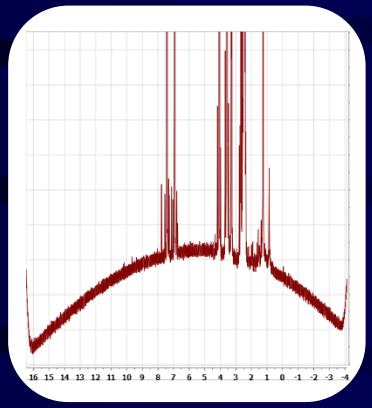
This is NOT a **real** spectrum! It's a naive, simulated NMR spectrum.

I am going to discuss these things, applied to NMR spectra analysis!

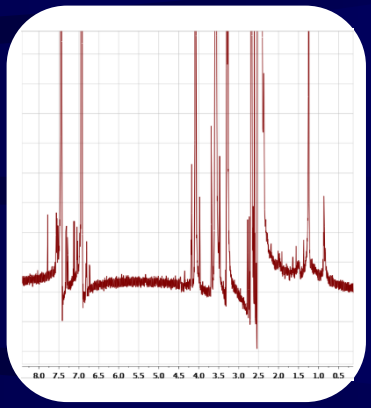
Few examples of undesirable spectral artifacts



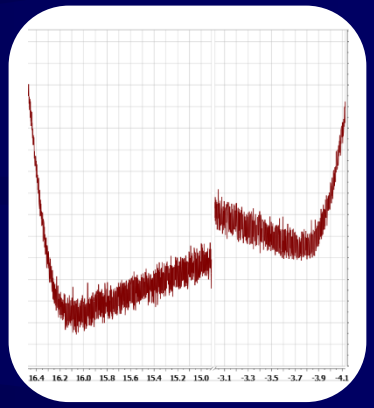
Noise



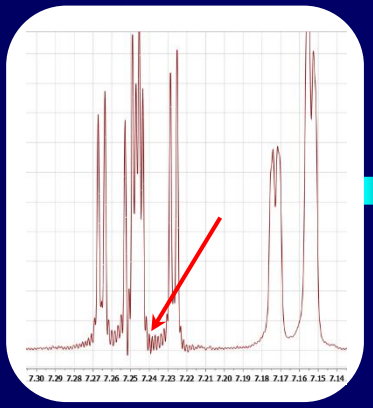
Baseline roll



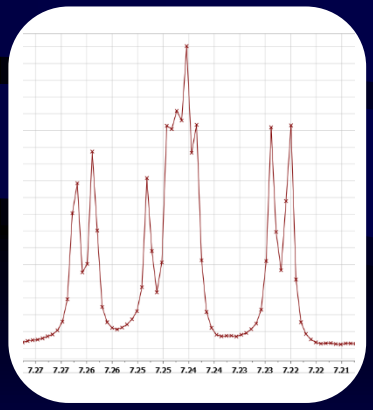
Imperfect phasing



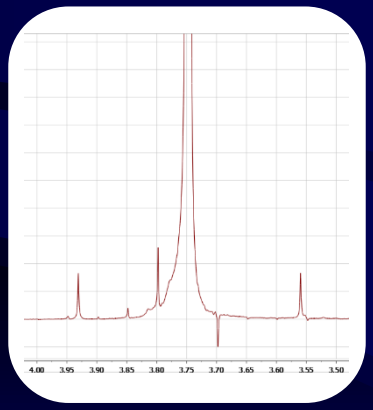
Bruker & Jeol smileys & brownies



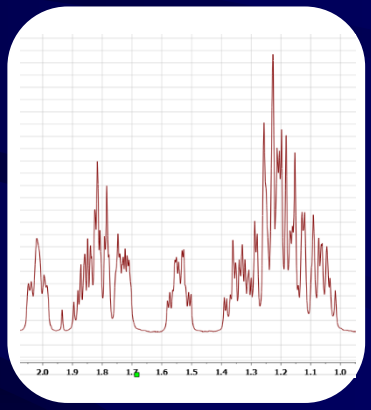
FID truncation effects



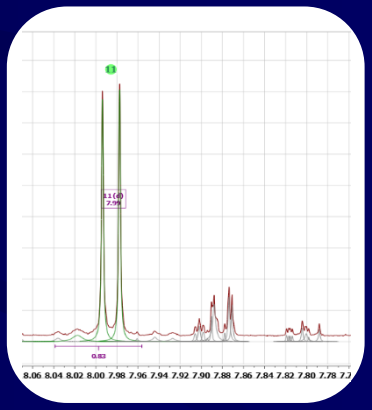
Underdigitization



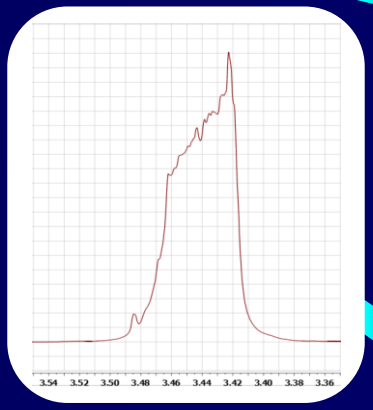
Rotation sidebands



Peaks overlap



Impurities peaks



Sample temperature drift effects

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

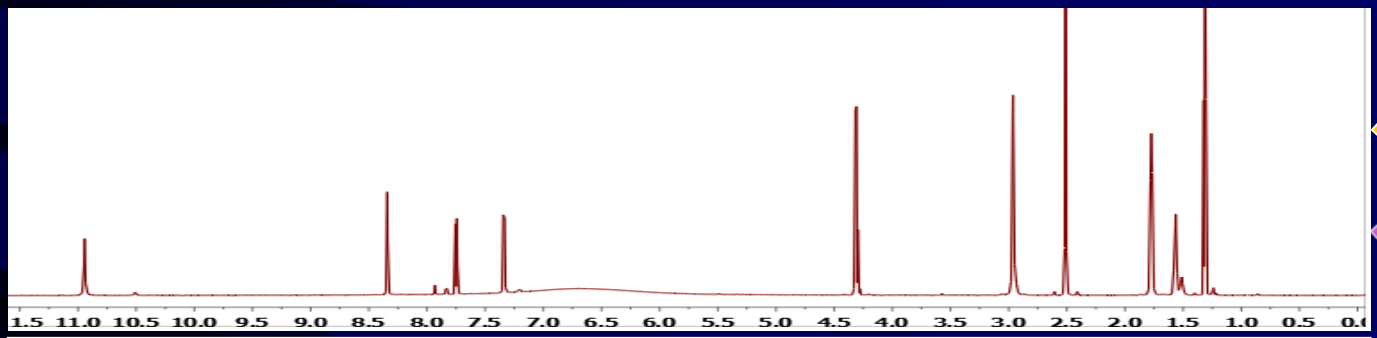
Meeting of GERMN, the Spanish NMR Discussion Group, Santiago de Compostela, Spain, 27 Sep 2013, held just after SMASH

A basic idea:
extract all pertinent information into a table

... and forget the rest!

What does a spectroscopist see?

Peaks, multiplets (singlets, doublets, AB quartets, triplets, quadruplets, …), labiles, 13C stellite peaks, aromatic peaks, d-solvent peaks, 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!

=> there is a big communication problem

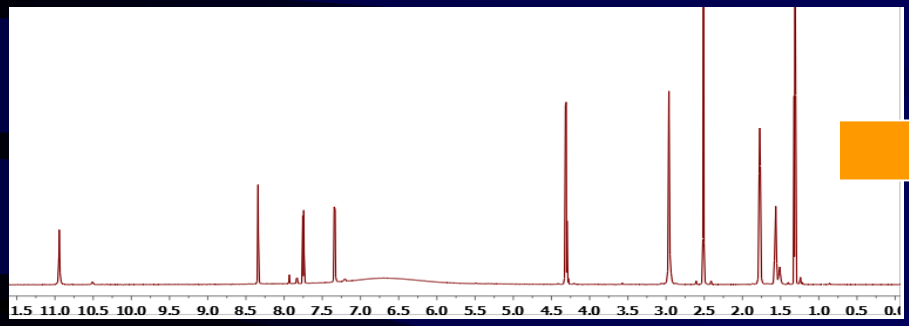
A basic idea:
extract all pertinent information into a table

... and forget the rest!

What does a spectroscopist see?

Peaks,

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



GSD

	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

What does a programmer see?

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

Language synchronization => better communication => better software

The basic idea:
extract all pertinent information into a table

... and forget the rest!

I have started insisting on this approach since 2006

but the NMR community did not pay any attention

- until it was implemented and working in Mnova !!

Now every software vendor feels obliged to do it!

The Ist Law of Data Evaluation:

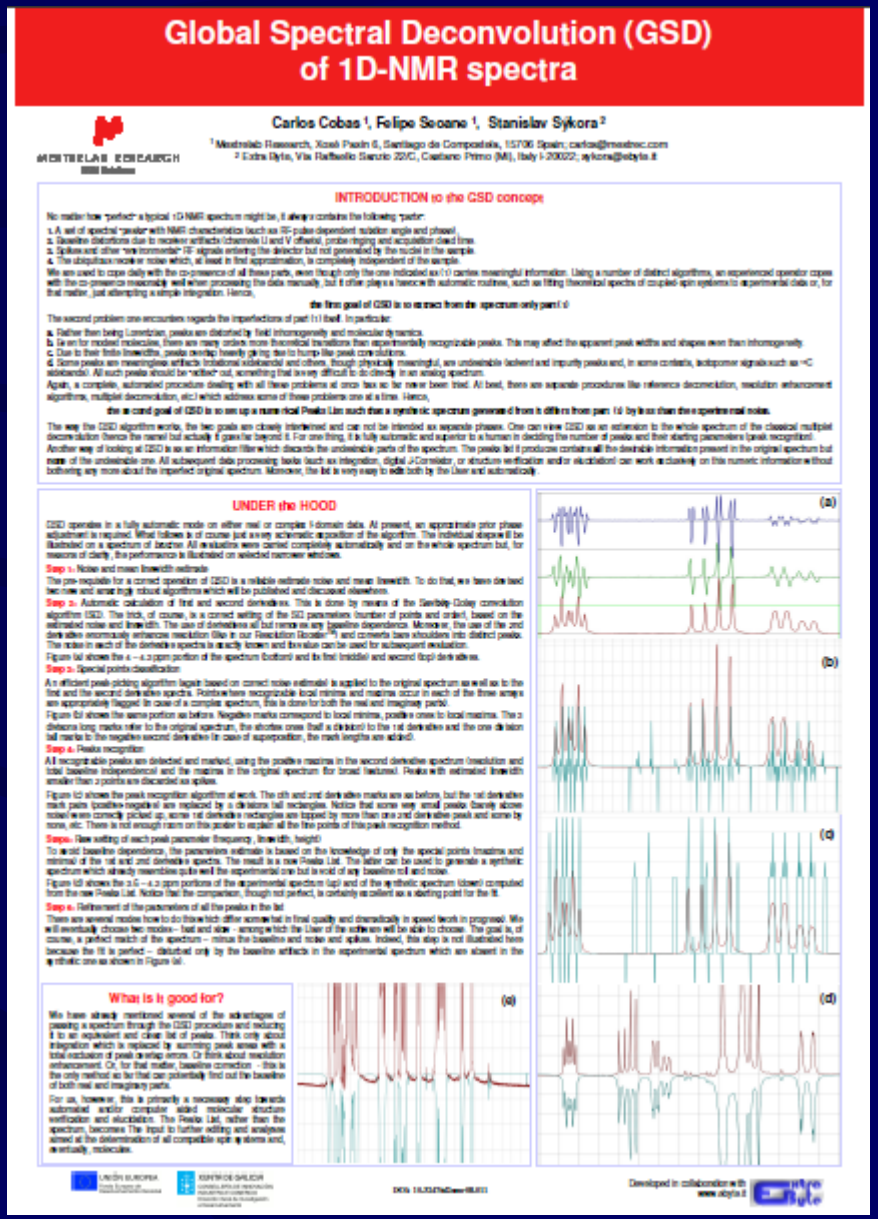
Don't talk about it! Do it!

GSD: Global Spectral Deconvolution

- Born in the summer of 2008,
- so it just celebrated 5 years
- Paved the way, and showed that the basic idea can work
- Very robust and well tested
- 2013: Potential competitors start appeaearing:
  - Internal (Padé based)
  - External (CRAFT)

References:

DOI: 10.3247/SL2Nmr08.011
DOI: 10.3247/SL3Nmr09.003



Meeting of GERMN, the Spanish NMR Discussion Group, Santiago de Compostela, Spain, 27 Sep 2013, held just after SMASH

Why must peaks be recognized and boxed-in prior to any fitting ?

Spectral peaks have *(very approximately)* Lorentzian shapes:

$$P(h,\Omega,\Delta;\nu) = h\, L((\nu-\Omega)/(\Delta/2))$$

$$L(x) = 1/(1-ix)$$

Well-known fact: all nearly complete sets of Lorentzian-shaped functions are approximately linearly dependent

A trivial but painful consequence:

Lorentzian-type deconvolutions are numerically ill defined

A Lorentzian peak can be approximated very well by three or five different Lorentzian peaks (=> acute danger of **peak spawning**).

GSD example: a 400 MHz strychnine spectrum

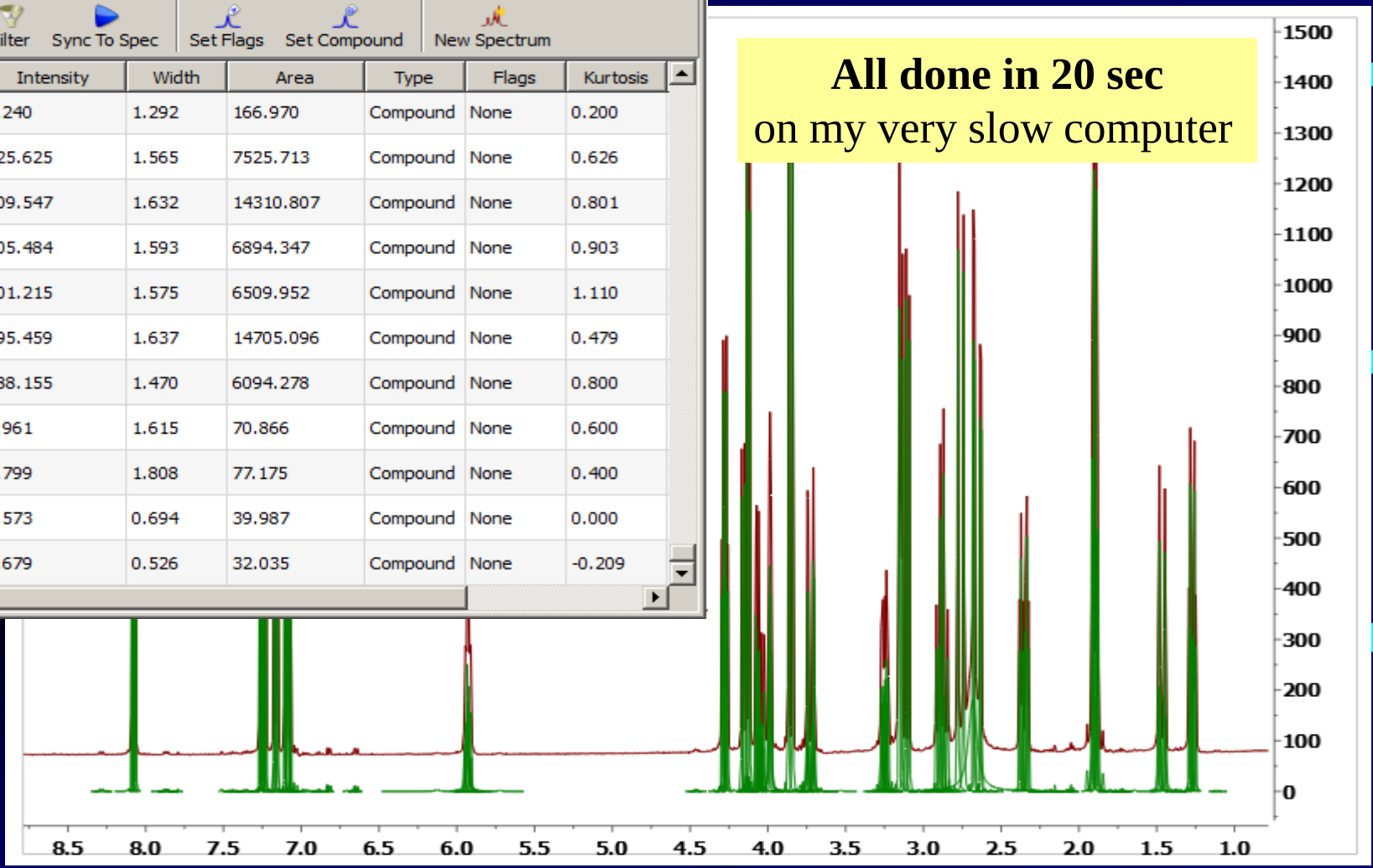
Peaks

Report Peaks | Copy Peaks | Setup Report | Delete | Select Peaks

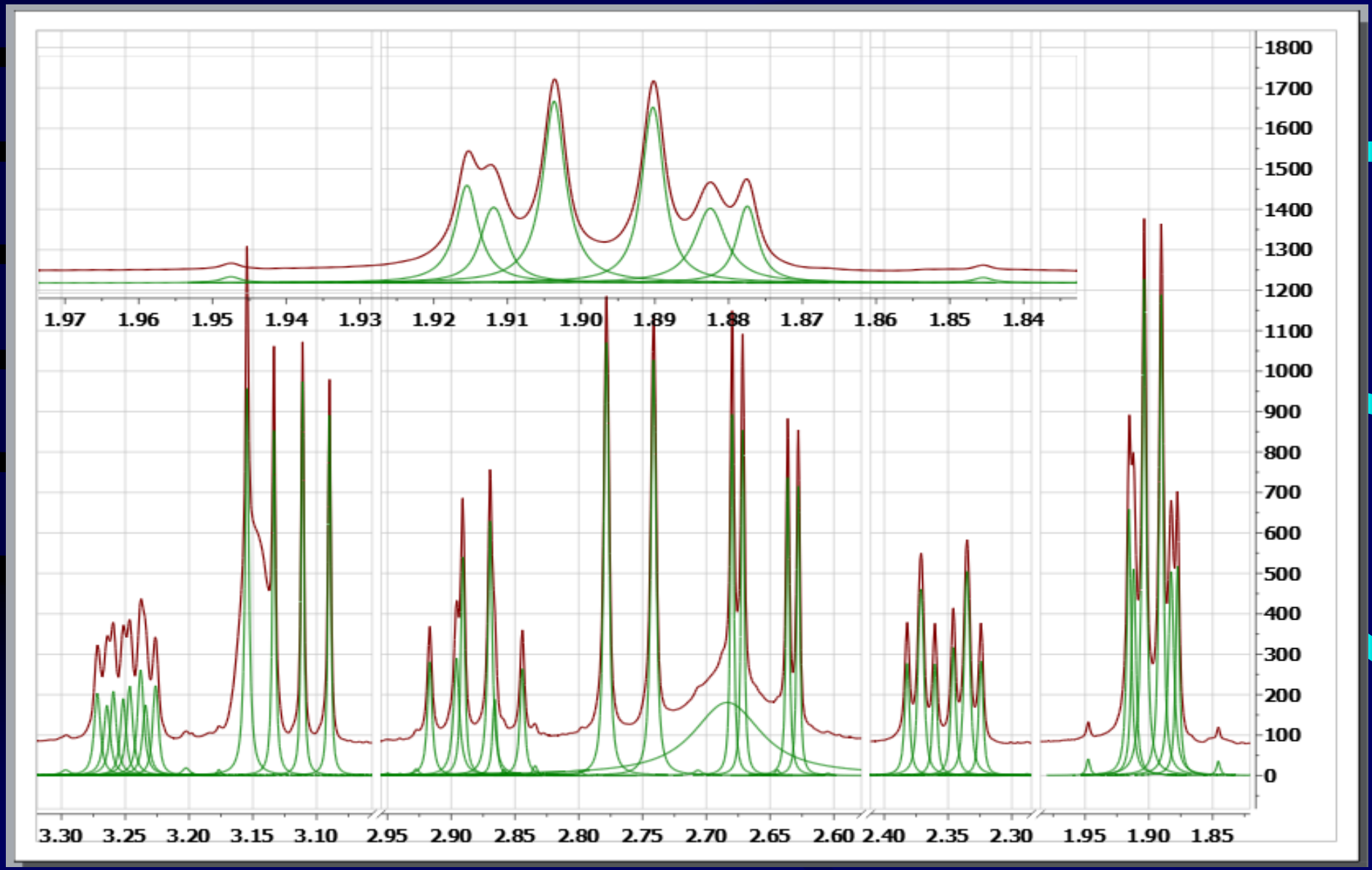
Sync From Spec | Filter | Sync To Spec | Set Flags | Set Compound | New Spectrum

	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

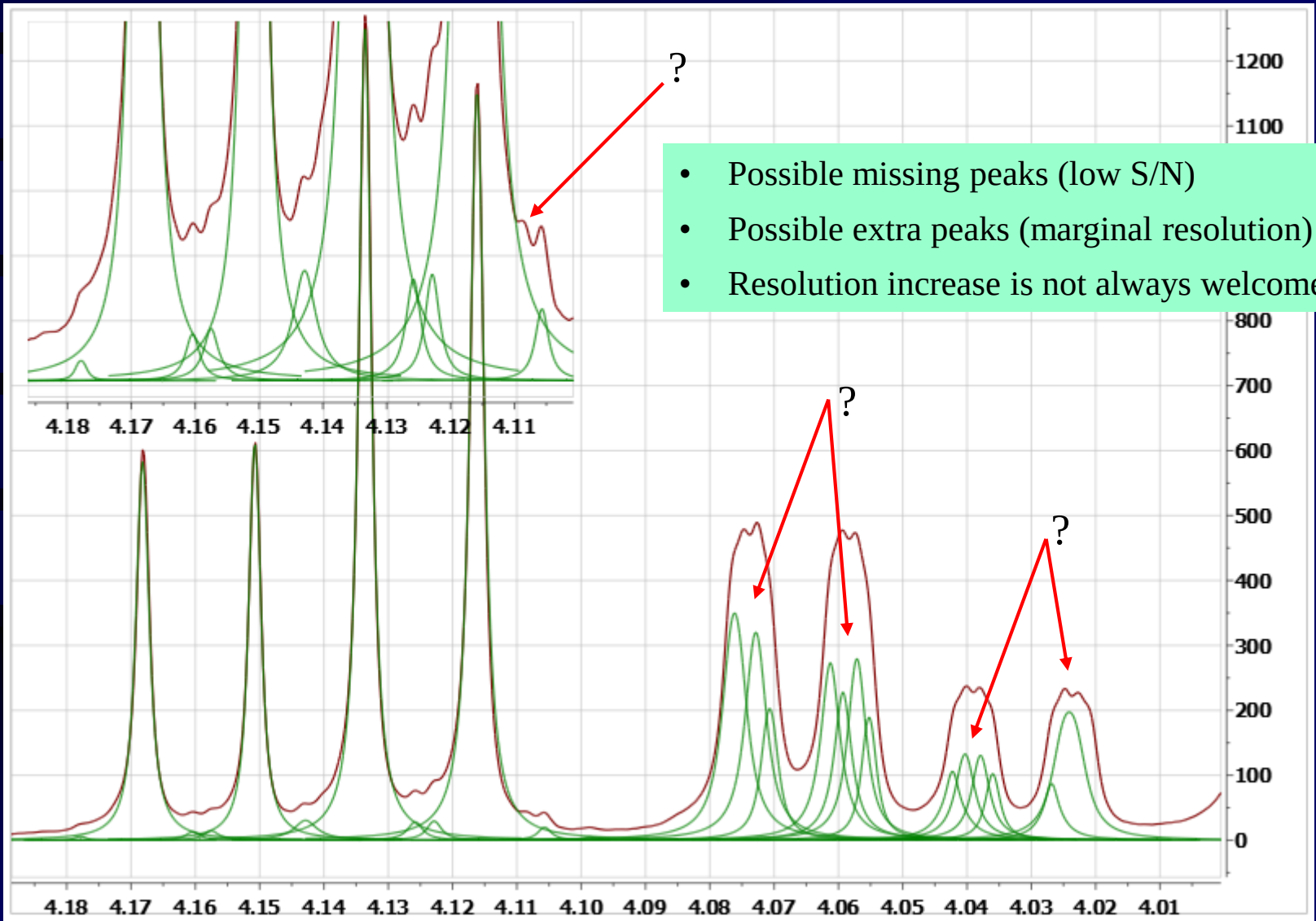
All done in 20 sec
on my very slow computer



Examples of peaks detection



Limits of GSD and GSD artifacts

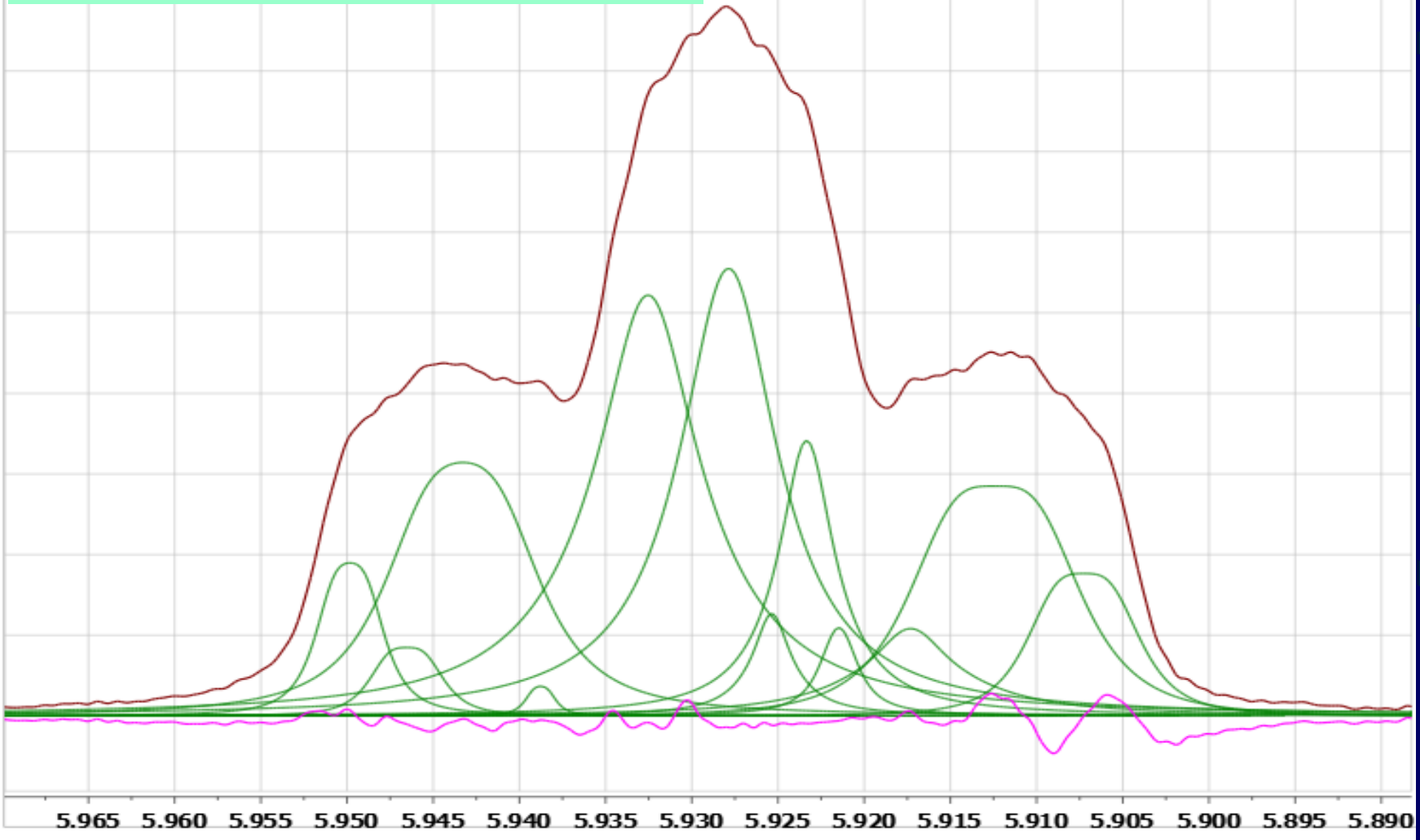


- Possible missing peaks (low S/N)
- Possible extra peaks (marginal resolution)
- Resolution increase is not always welcome

Limits of GSD and GSD artifacts:

Example of a «broken» symmetry in the GSD peaks

The «famous» triplet in strychnine (400 MHz).



Limits and uncertaintaies of GSD
... or *where did the fuzziness go ?*

On one hand, great many original artifacts and uncertainties were eliminated (noise, baseline) and some were reduced (mis-phasing). Moreover, effective resolution was markedly enhanced, and most multiplets get nicely matched quantitatively.

On the other hand, some new potential problems were introduced:

- A real weak peak may be detected or not, depending upon the particular noise sample.
- A nonexistent peak may get «invented» due to an unusually strong noise fluctuation.
- Symmetric multiplet patterns may get «broken» (very annoying).

The IInd Law of Data Evaluation:

Uncertainties don't go away, they just change looks!

So, did we gain anything with GSD?

The questions which all this raises are (as always):

- How much have we gained and how much have we lost?
- Is the balance positive?

Only ample statistical testing, with actual applications, can answer that.

We did it and we know for sure that the answer is very positive.

The IIIrd Law of Data Evaluation:

Nothing useful comes for free!

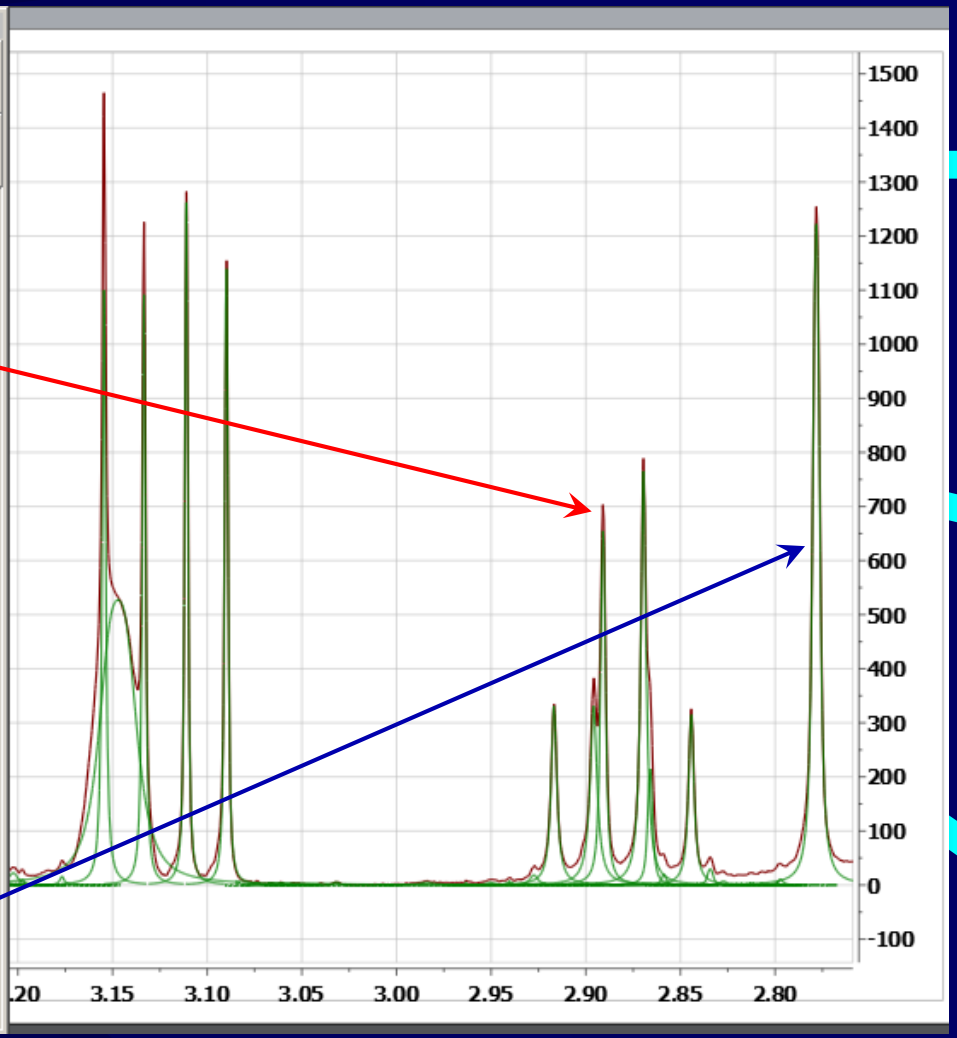
GSD peaks linewidths and shapes

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
229	2.917	330.822	1.456	7696.453	Compound	None	0.063
230	2.896	331.297	1.778	9540.649	Compound	None	-0.036
231	2.891	656.647	1.242	11981.352	Compound	None	0.660
232	2.870	765.628	1.344	16632.656	Compound	None	-0.024
233	2.866	215.085	1.253	3907.304	Compound	None	0.748
234	2.859	20.218	1.143	355.549	Compound	None	0.341
235	2.844	315.532	1.396	6973.556	Compound	None	0.130
236	2.834	29.771	1.384	543.681	Compound	None	1.352
237	2.827	7.347	1.499	162.481	Compound	None	0.632
238	2.813	3.088	1.209	44.079	Compound	None	1.996
239	2.807	3.492	0.842	37.331	Compound	None	1.582
240	2.799	5.719	0.426	35.739	Compound	None	0.676
241	2.797	11.383	1.207	232.092	Compound	None	-0.361
242	2.778	1222.811	1.708	29533.781	Compound	None	0.915
243	2.741	1180.772	1.695	29125.021	Compound	None	0.725



Why are peak shapes so different
even in the same spectrum?

Reminder of the path from a Spin System to Spectrum:

Quantum transitions => Peaks => Multiplets => Spectrum

In a spin system with N spin ½ nuclei there are $N.2^{N-1}$ transitions

Small molecule example: when n = 4 there are only 32 transitions. With a whif of luck, we might distinguish 32 peaks in its spectrum, each of which would therefore contain a single quantum transition. Physical theory tells us that transitions are of Lorentzian shape (though their linewidth can vary – another story).

GREAT! How simple! Or not???

Why are peak linewidths and shapes so different? (even in the same spectrum)?

Counting the main transitions in somewhat larger molecules:

N = 15: 245’760

N = 30: 16’106’127’360

N = 45: 791’648’371’998’720

But in a typical spectrum of such molecules we rarely distinguish more than 200 peaks. That, for N = 30 makes it well over 1000 quantum transitions per resolved peak!

What we see is an envelope of a distribution of Lorentzians

The IVth Law of Data Evaluation:

Don’t loose time trying to beat combinatorics! It’s hopeless! Can’t be done!

Sources of peak-shape deviations from the Lorentzian

1. Magnetic field inhomogeneity (shimming)
2. Magnetic field noise (ebyte.it\library\docs\nmr06a\NMR_FieldNoise_Fid.html)
3. Sample spinning (dtto)
4. Sample temperature gradients (up to 0.01 ppm/deg)
5. FID weighting before FT (Voight and other profiles)
6. Distorsions due to Discrete Fourier Transform (cyclic condition)
7. Overlap of miriads of transitions in coupled spin systems
8. Relaxation effects (e.g., methyl lines contain 3 transitions of different widths)
9. Molecular dynamics effects (chemical exchange, limited mobility)
10. etc …

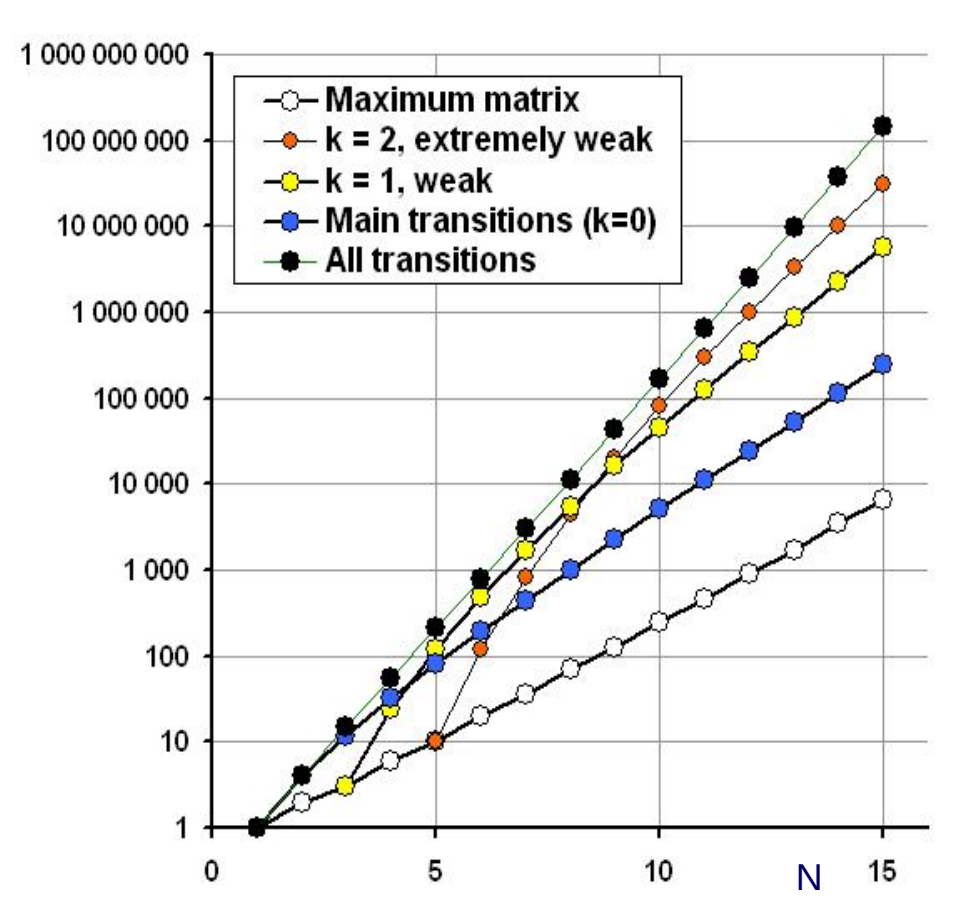
Overlap of transitions

Spectral peaks are in reality envelopes of many transitions:

Even in molecules of modest size the number of distinct peaks is thousands times smaller than that of quantum transitions.

$\Rightarrow$

Every peak is an envelope of a large number of transitions and its shape is dominated by the coupling pattern of the spin system. The general characteristics of such distributions can be analyzed and exploited.



The Generalized Lorentzian lineshape

The complex-valued Lorentzian lineshape, $L(x) = 1/(z+i)$, is a rational function which for large real x behaves as $O[1/x^2]$ and satisfies $L(1/x) = 1–L^*(x)$.
There are other rational function which possess these properties.
The simplest such «successor» of a Lorentzian is

$$G(z) = [(2+z^2)+iz^3\sqrt{3}] / [2(1+z^2+z^4)].$$

Since any linear combination of L(z) and G(z) also has the desired properties, we use the Generalized Lorentzian lineshape defined as

$$GL(z) = (1-k)\, L(z) + k\, G(z),$$

Where k is a real «kurtosis parameter»,
so called because it affects the peak's kurtosis.

The Vth Law of Data Evaluation:
Keep an ace up your sleeve and cheat without shame! It's Science!

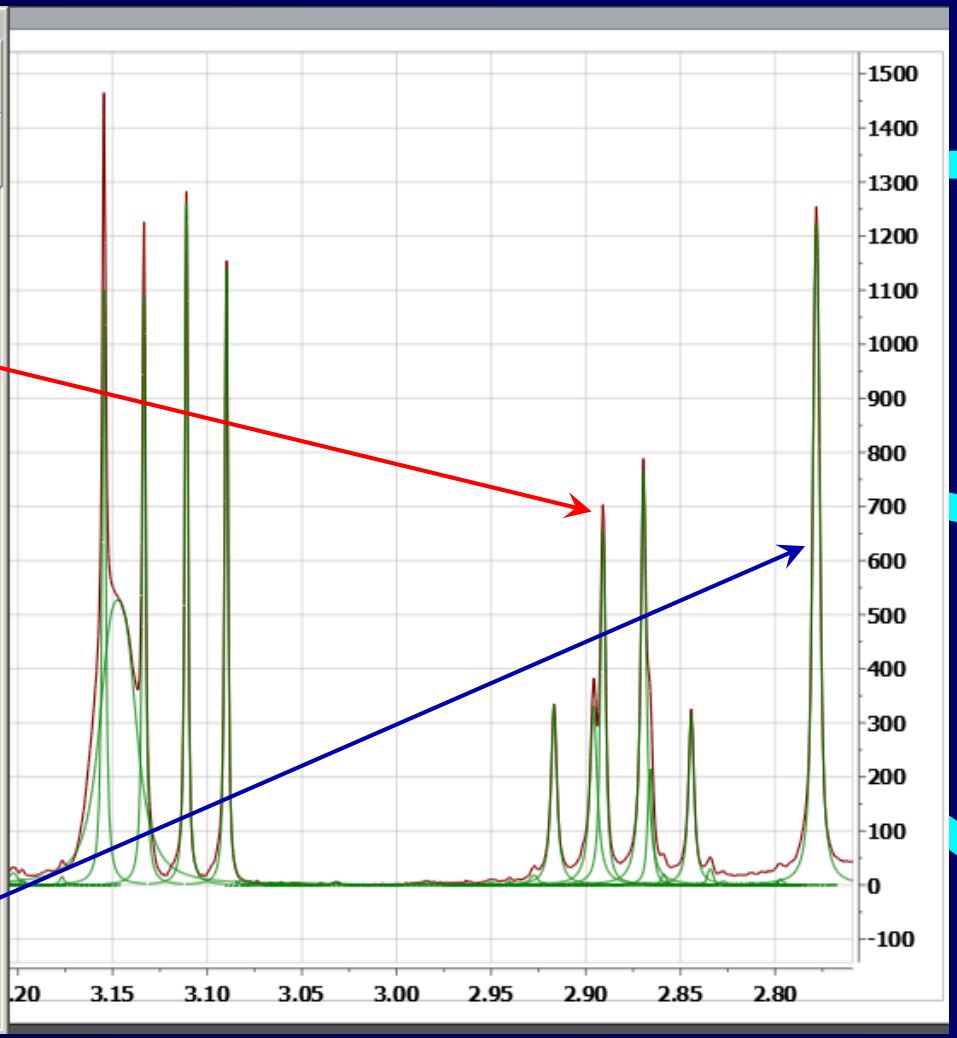
GSD peaks linewidths and shapes

Peaks

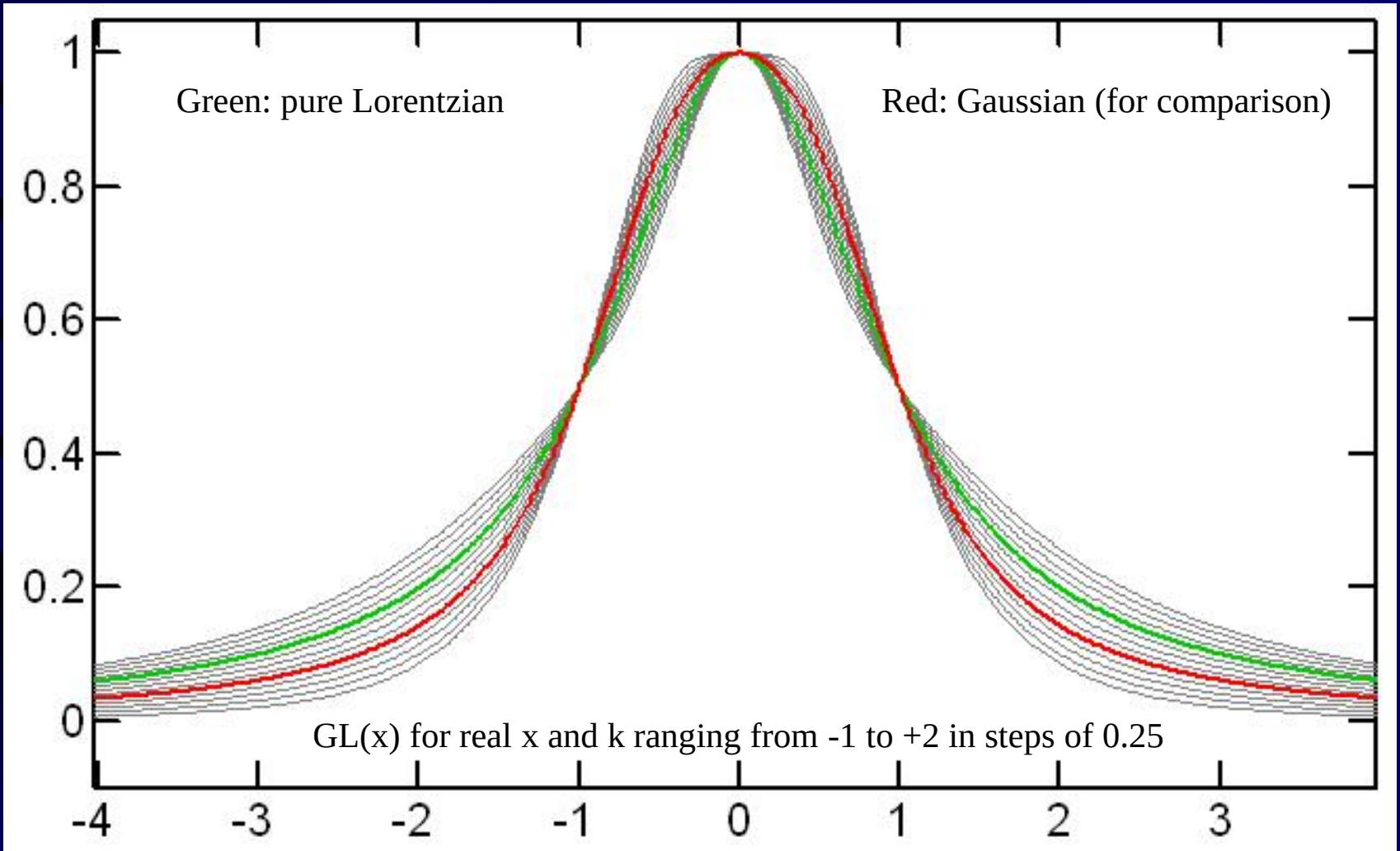
Report Peaks | Copy Peaks | Setup Report | Delete | Select Peaks

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

	ppm	Intensity	Width	Area	Type	Flags	Kurtosis
229	2.917	330.822	1.456	7696.453	Compound	None	0.063
230	2.896	331.297	1.778	9540.649	Compound	None	-0.036
231	2.891	656.647	1.242	11981.352	Compound	None	0.660
232	2.870	765.628	1.344	16632.656	Compound	None	-0.024
233	2.866	215.085	1.253	3907.304	Compound	None	0.748
234	2.859	20.218	1.143	355.549	Compound	None	0.341
235	2.844	315.532	1.396	6973.556	Compound	None	0.130
236	2.834	29.771	1.384	543.681	Compound	None	1.352
237	2.827	7.347	1.499	162.481	Compound	None	0.632
238	2.813	3.088	1.209	44.079	Compound	None	1.996
239	2.807	3.492	0.842	37.331	Compound	None	1.582
240	2.799	5.719	0.426	35.739	Compound	None	0.676
241	2.797	11.383	1.207	232.092	Compound	None	-0.361
242	2.778	1222.811	1.708	29533.781	Compound	None	0.915
243	2.741	1180.772	1.695	29125.021	Compound	None	0.725



Graph of the Generalized Lorentzian lineshape



A final word on peak shapes

While plain Lorentzian shape is basically sound, without a generalization going beyond simple Gaussian-Lorentzian, it could never provide good universal fits, particularly when quantitation is an issue.

GSD works satisfactorily on typical pharma spectra, but also on metabolomic spectra, protein spectra, etc.
It is a universal workhorse.

The VIth Law of Data Evaluation:
Generalize, but not too much!

Part II: Peaks Auto-Editing

Having identified and tabulated all the peaks,

what more can we do ?

GSD by itself does not address issues like what might each peak be:

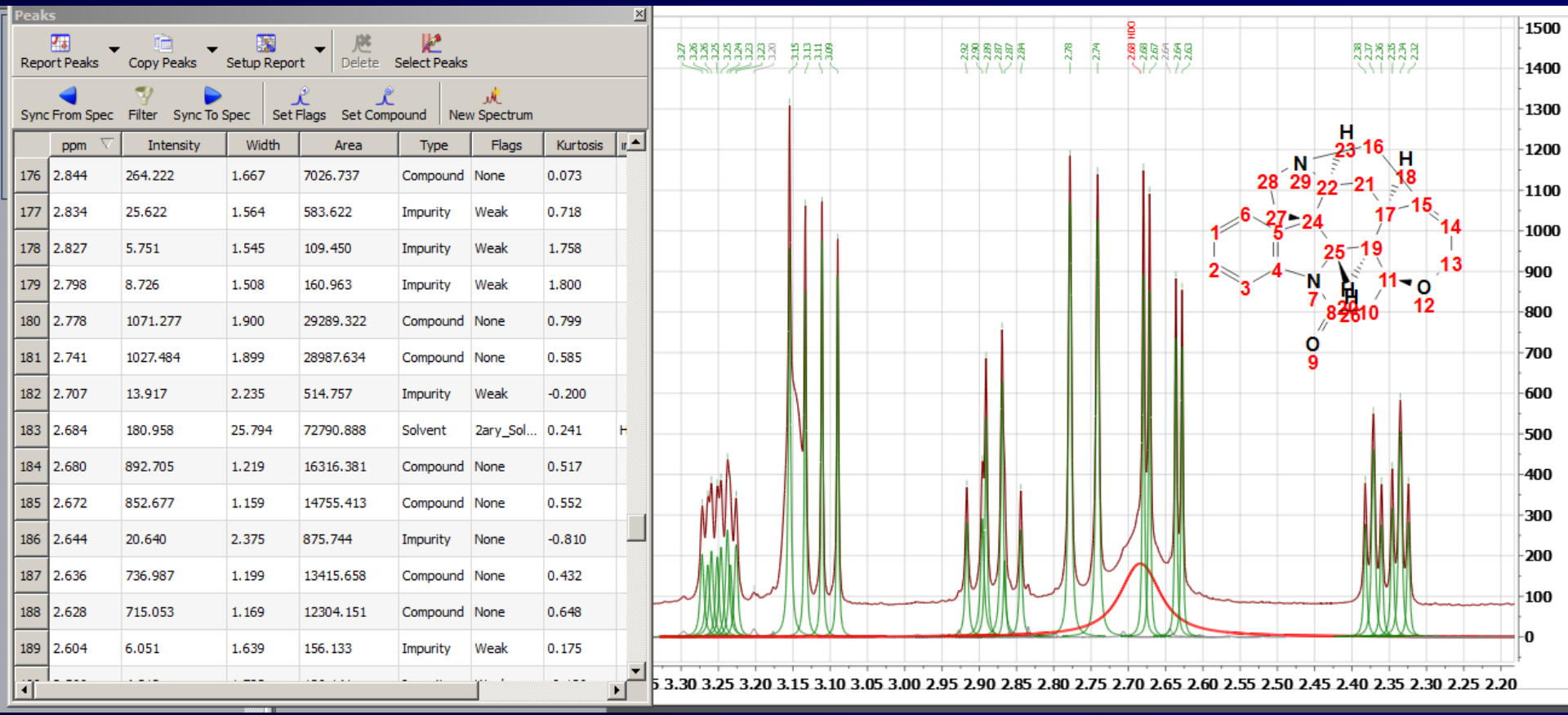
- compound,
- primary or secondary solvent,
- potential labile,
- 13C satellite,
- valid member of a multiplet,
- impurity,
- S or Q reference,
- artifact,
- etc.

Nor does GSD group the peaks into multiplets and classify those.

All these are the tasks refered to generically as peaks auto-editing.

Peaks editing is primarily fully automatic

and uses greedily whatever information is available (1H, HSQC, molecule, ...)



It is also the first plank in NMR spectra evaluation hierarchy
where specific NMR know-how is used

Meeting of GERMN, the Spanish NMR Discussion Group, Santiago de Compostela, Spain, 27 Sep 2013, held just after SMASH

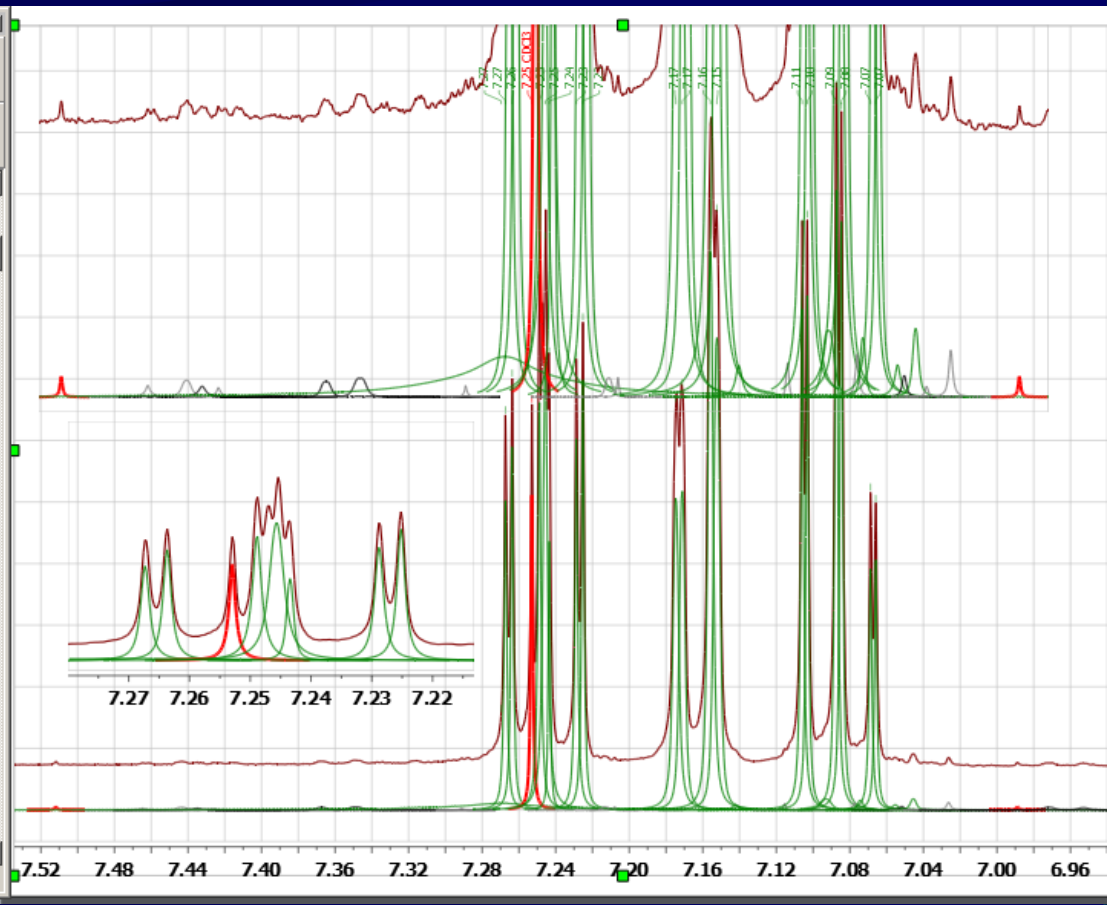
CHCl3 identification in an aromatic multiplet

Peaks

Report Peaks | Copy Peaks | Setup Report | Delete | Select Peaks

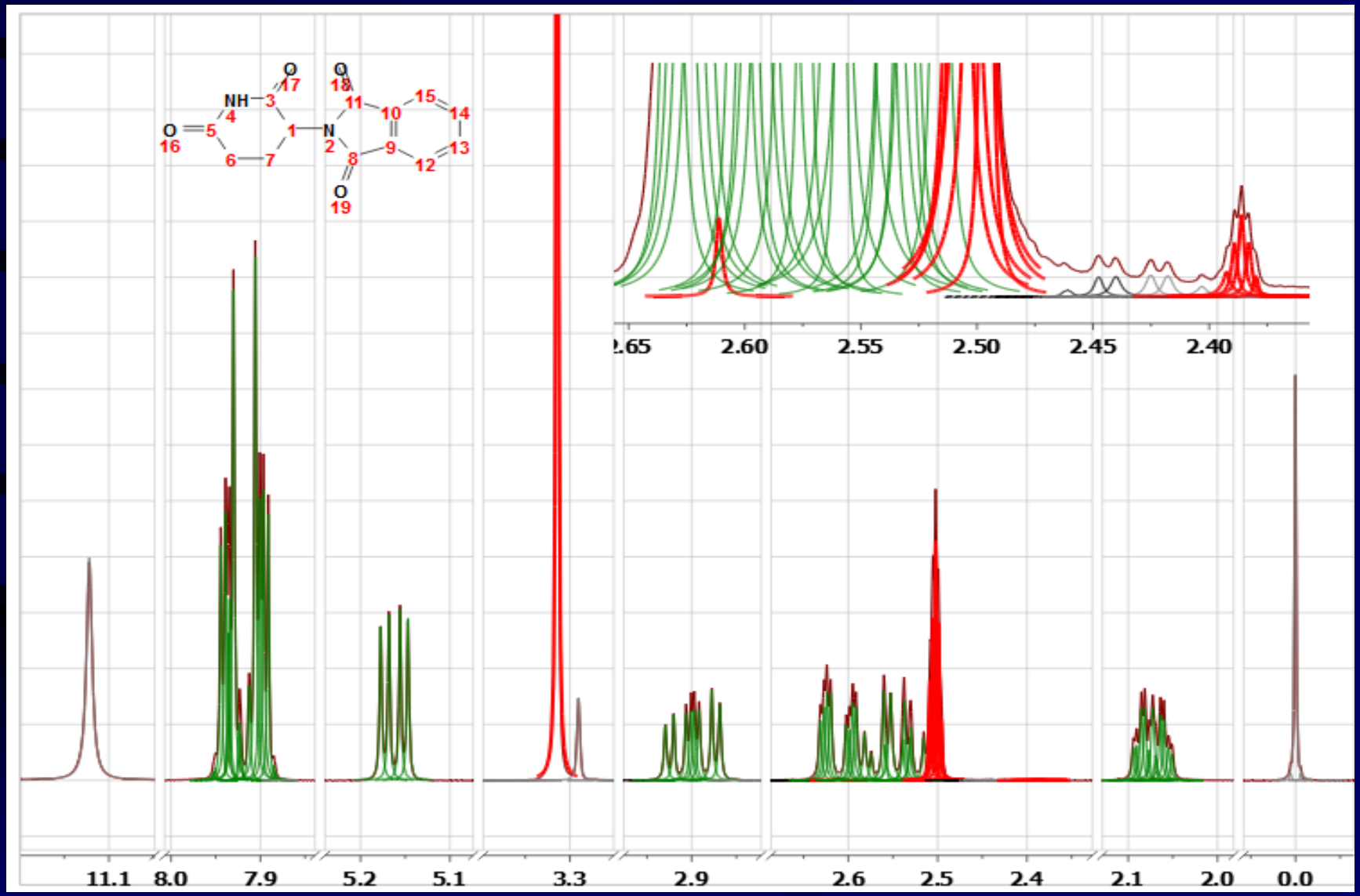
Sync From Spec | Filter | Sync To Spec | Set Flags | Set Compound | New Spectrum

	ppm	Intensity	Width	Area	Type	Flags	Kurtosis
16	7.512	5.458	0.619	53.899	Solvent	Weak + C13_Sat	0.074
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
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

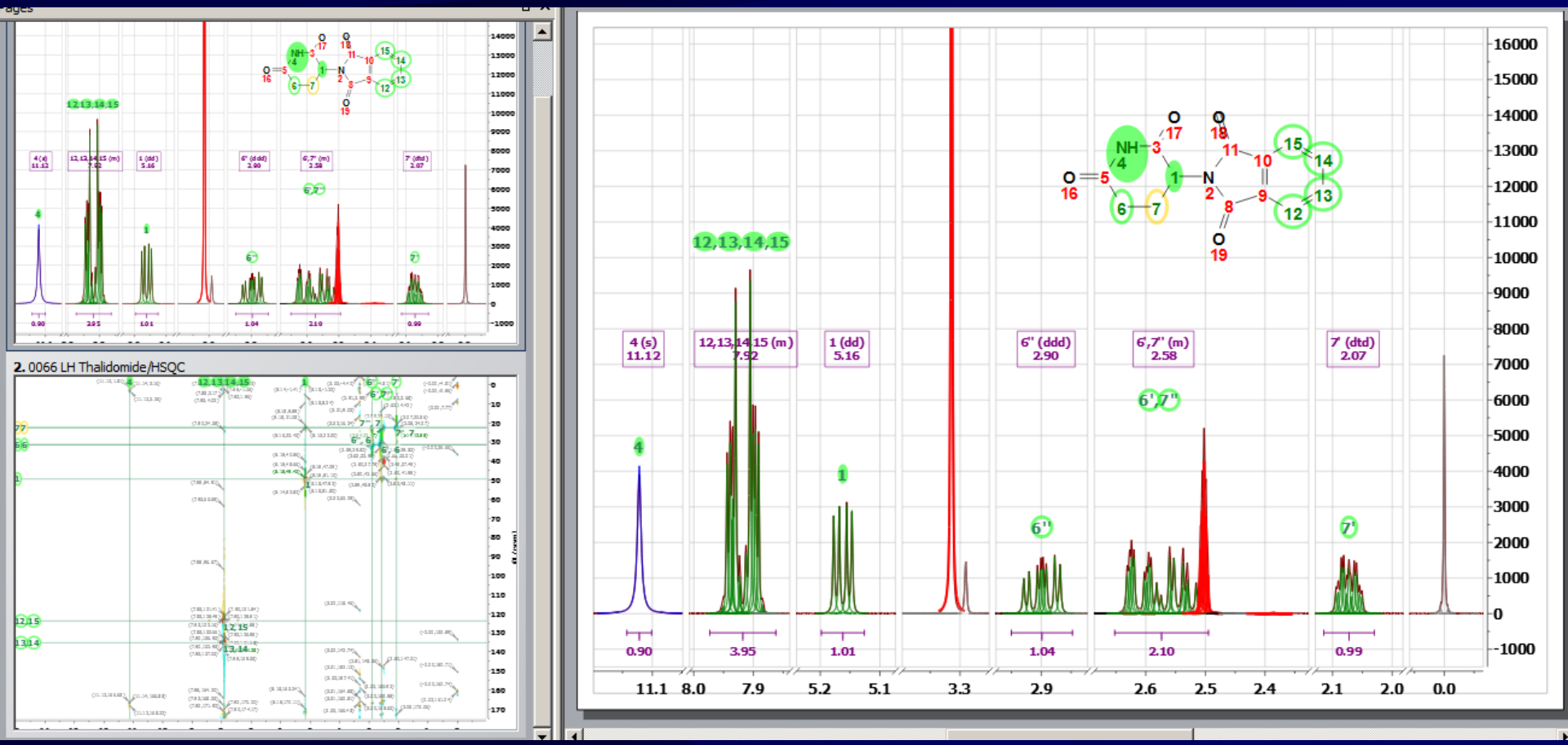


Uses even the 13C satellites (209.25 Hz apart) and their isotopic shift (the satellite pair center is -2.67 ppb from the main peak)

DMSO identification example (Thalidomid 600 MHz)

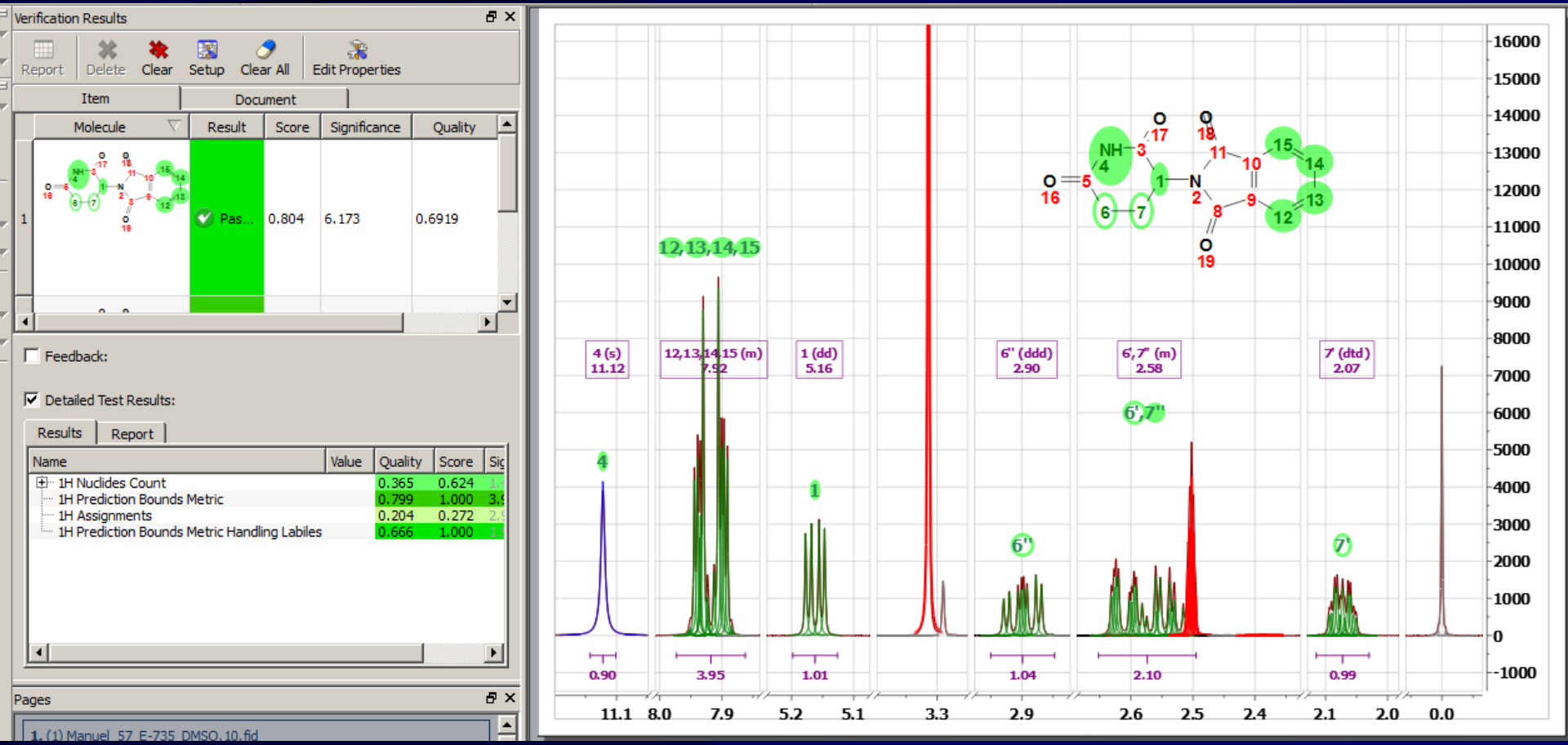


Labile identification example (Thalidomid 600 MHz)



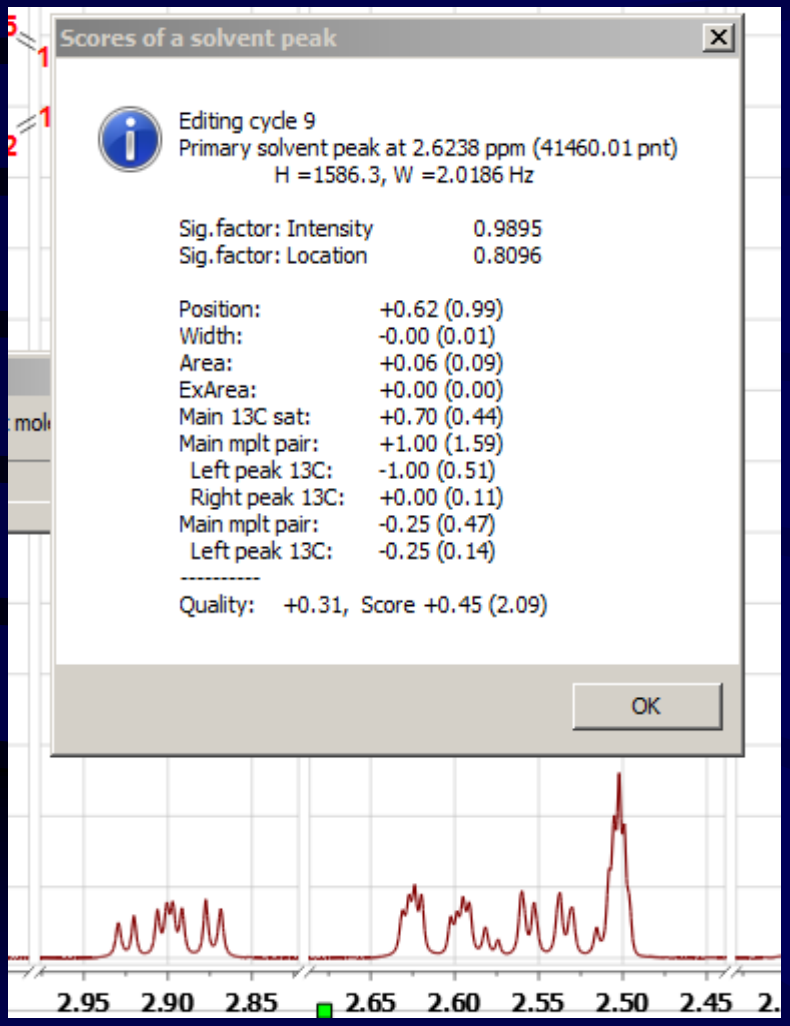
Assignments analysis was used to correctly label the labile peak:
a simple example of «loopbacking»

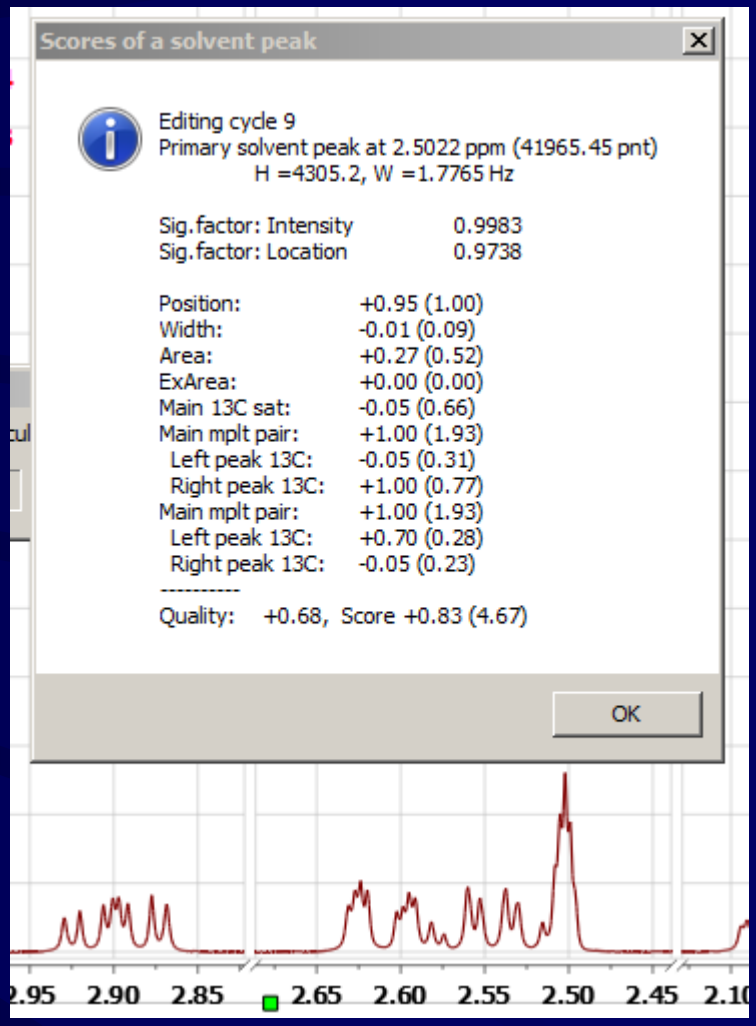
Labile identification example (Thalidomid 600 MHz)



An anticipation of the ASV application (Automatic Structure Verification)

Scoring, scoring, and – what was the third one?





Here, each peak is scored for «being the pivot peak of the primary solvent»

Scoring intended as a way of life

Every question that a spectroscopist is asking himself when inspecting a spectrum becomes a scoring procedure in the software. Examples:

Could this peak be the main solvent? (up to 15 votes)
Could this peak be a labile? (up to 12 votes)
Does this splitting exist somewhere else in the spectrum? (6 votes)
Is this peak an essential member of its multiplet? (12 votes)
Etc etc etc etc

Except that the algorithm does it brutally for all peaks and all multiplets, and all assignments which have the slightest of chances to pass.
It is as setting up a voting committee on every little query.

In a typical ASV run on an average pharma spectrum, for example, the number of «votes» cast is around 10000!
It is much like a voting day in Santiago.

Various types of Scoring

Within the AI wizard (which is what the software is becoming),
we use several types of voting approaches:

- Democratic voting with predefined voter significances
- Quadratic voting with significance proportional to the cast score
- Penalty voting for things that better should be ok
- Veto voting (extreme case of penalty voting)

The long list of Auto-Editing tasks

- Detection of reference peaks
- Detection of 13C satellites
- Detection of potential labiles
- Formation of multiplets
- Multiplets purging and slicing
- Detection of primary solvent
- Detection of secondary solvent (water)
- Detection of non-deuterated solvent (if requested)
- Detection of known impurities (e.g, residual reaction solvent)
- Identification of potential unknown impurities
- Enumeration of feasible assignments (if molecule is known)
- Identification of best-scoring assignment (if molecule is known)
- Enumeration of feasible matching spin systems
- Identification of actual detectable labiles
- Identification of actual unknown impurities

However ...

Do not think about this as a linear process !!!!

Emulating human intelligence on a sequential computer is difficult, but that is what we have to do. Some tricks which help are:

- Iterative alternation of various steps
- Loop-back and look-ahead strategies

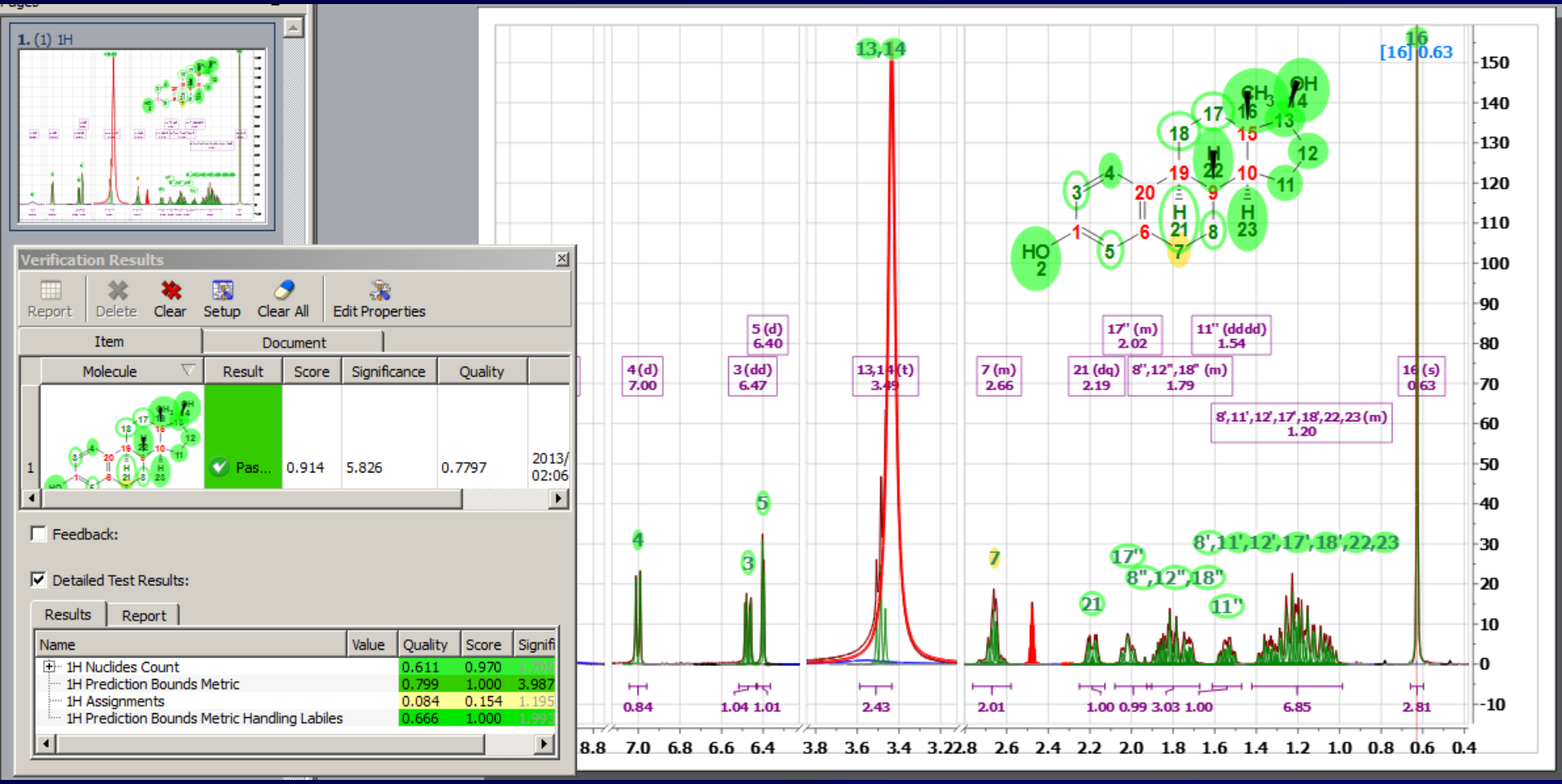
Once Auto-Editing is finished,

the information becomes available
for a number of «applications» which can use it in various ways:

- ASV Automatic Structure Verification
- ASPV Automatic Structure Presence Verification
- ACD Automatic Component Detection
- ASD Automatic Structure Discrimination
- ASE Automatic Structure Elucidation
- ADBV Automatic DataBase Validation
- etc

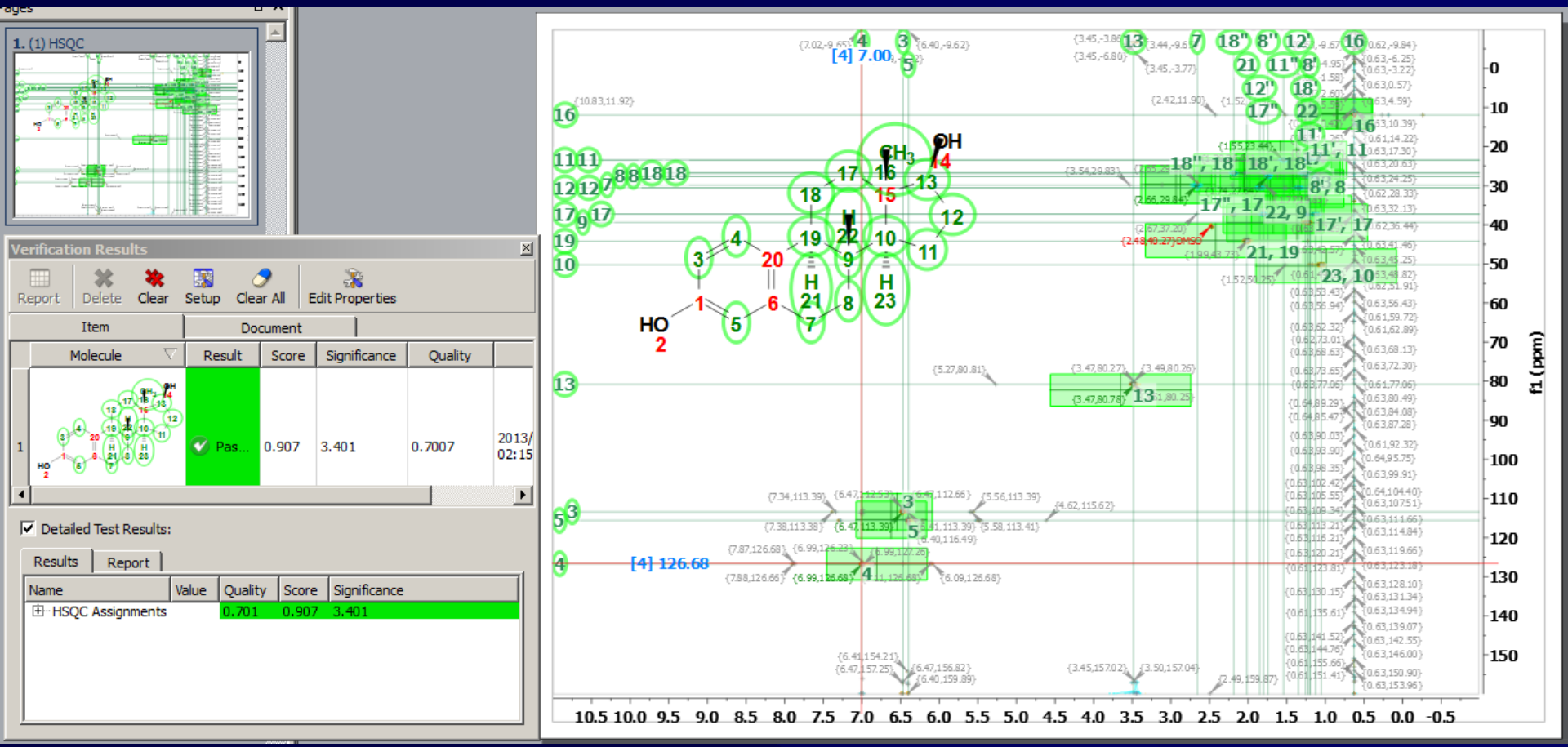
Final example(s)

Estradiol 400 MHz, using 1H only



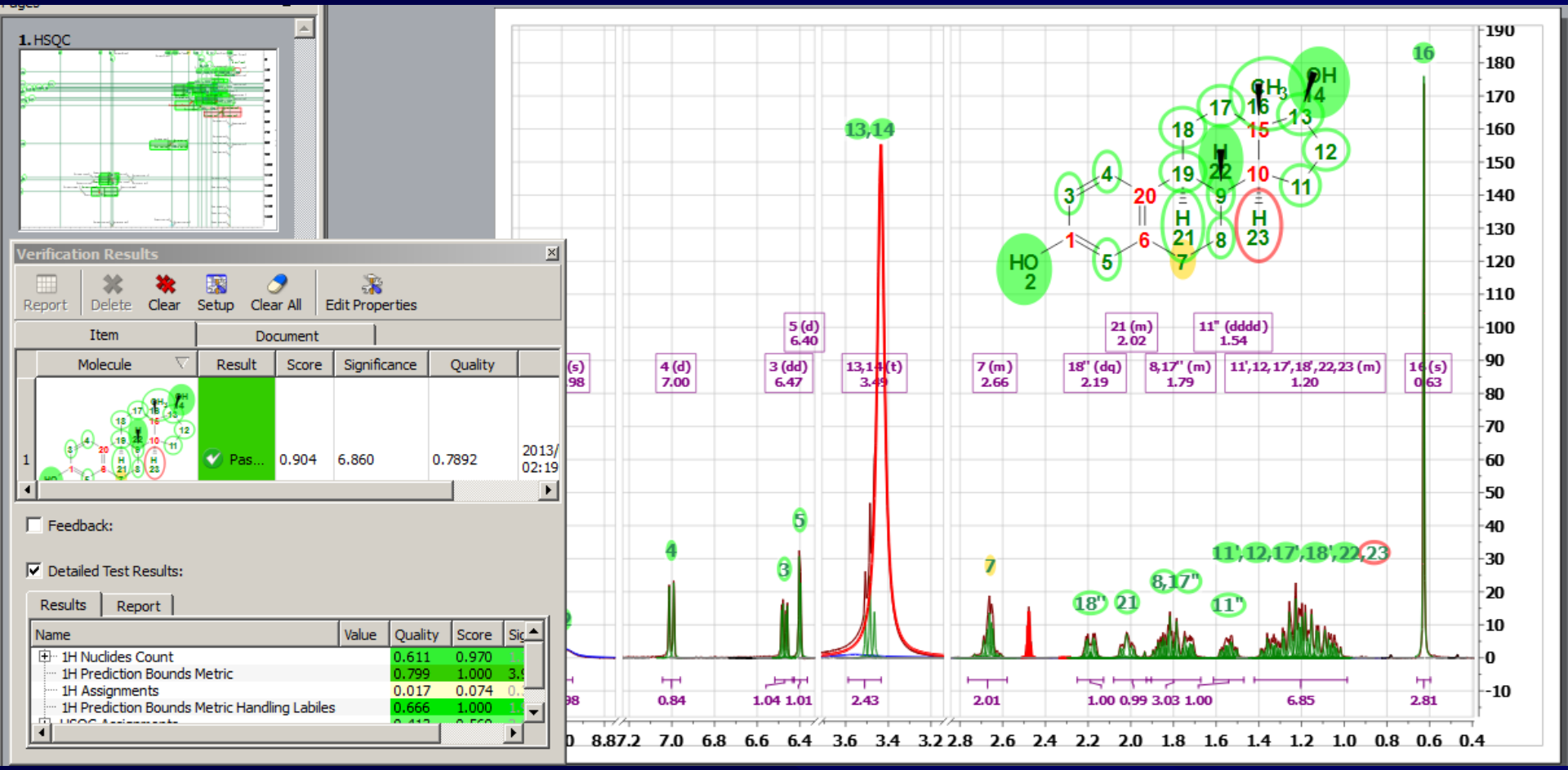
Meeting of GERMN, the Spanish NMR Discussion Group, Santiago de Compostela, Spain, 27 Sep 2013, held just after SMASH

Estradiol 400 MHz, using HSQC only



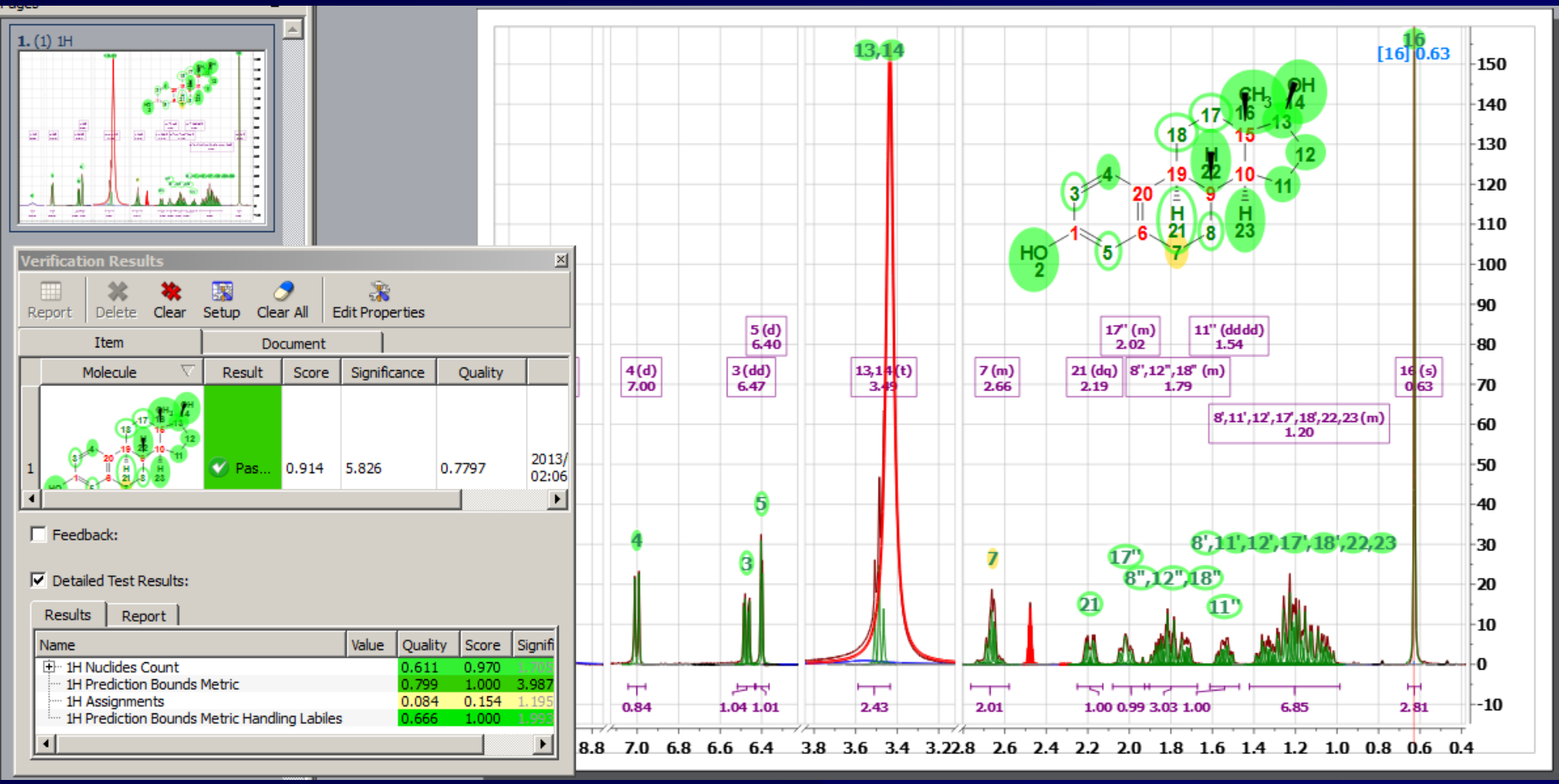
Meeting of GERMN, the Spanish NMR Discussion Group, Santiago de Compostela, Spain, 27 Sep 2013, held just after SMASH

Estradiol 400 MHz, using 1H and HSQC



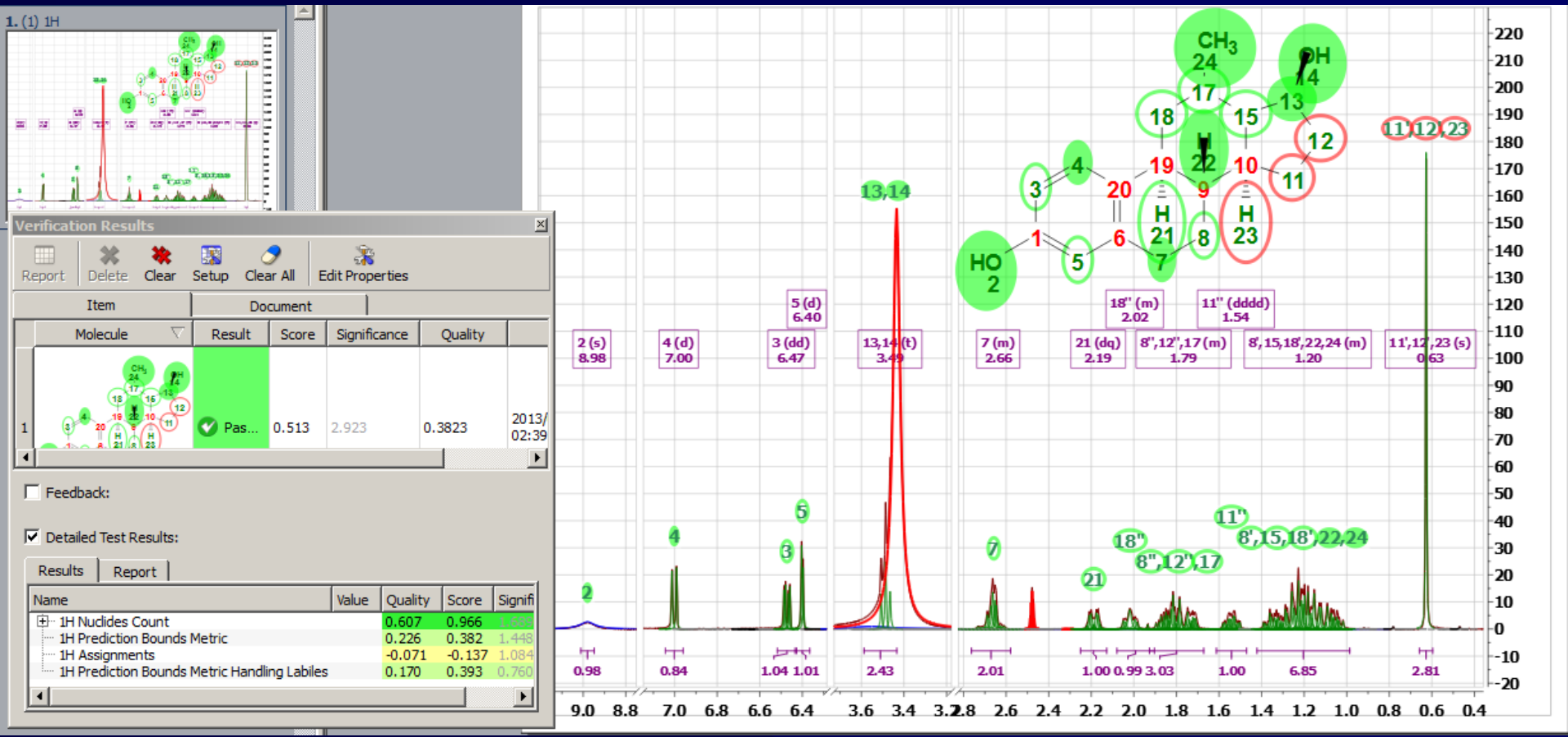
Meeting of GERMN, the Spanish NMR Discussion Group, Santiago de Compostela, Spain, 27 Sep 2013, held just after SMASH

Estradiol 400 MHz, using 1H only



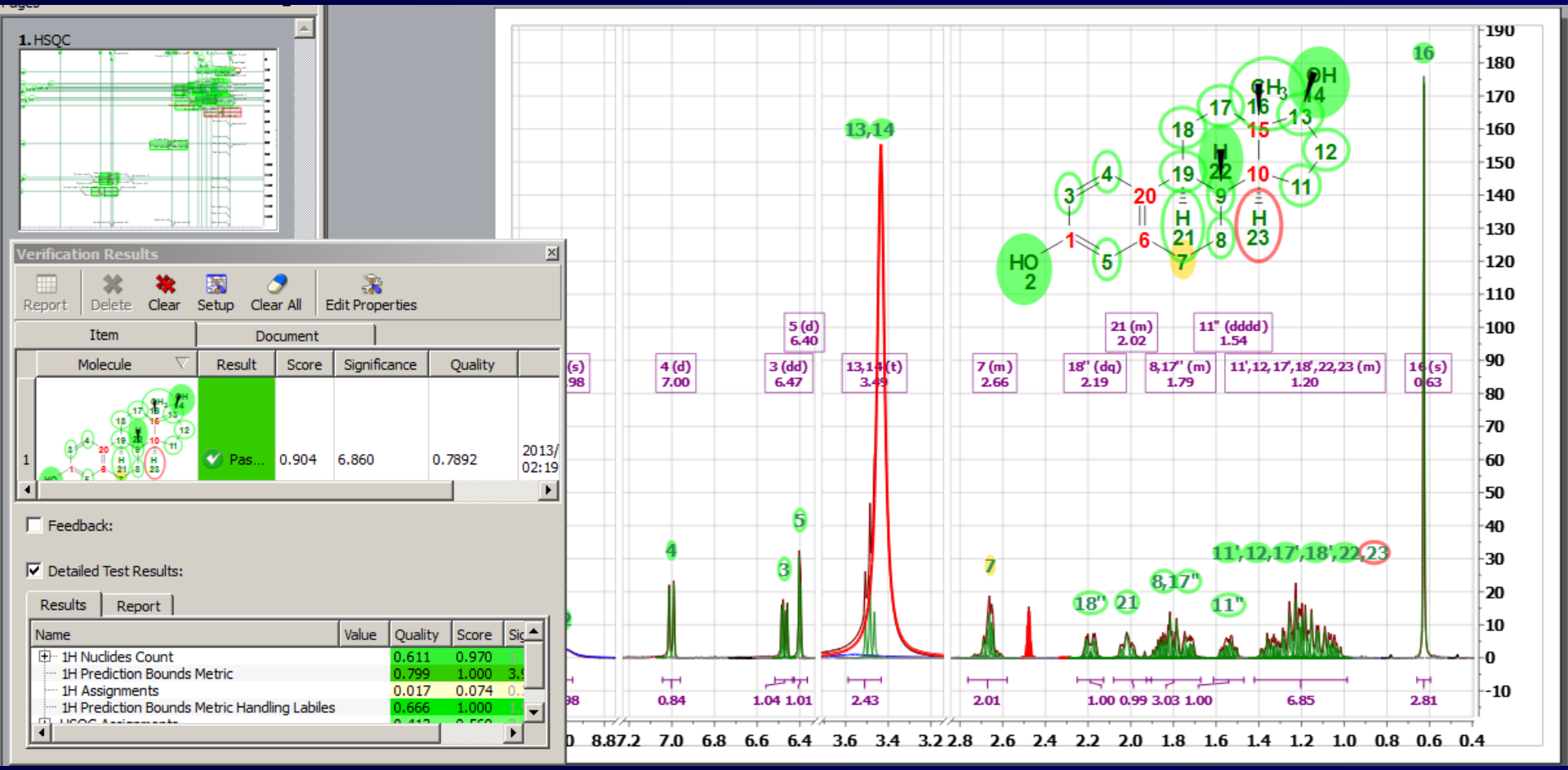
Meeting of GERMN, the Spanish NMR Discussion Group, Santiago de Compostela, Spain, 27 Sep 2013, held just after SMASH

Estradiol, modified mol, 400 MHz, using 1H only



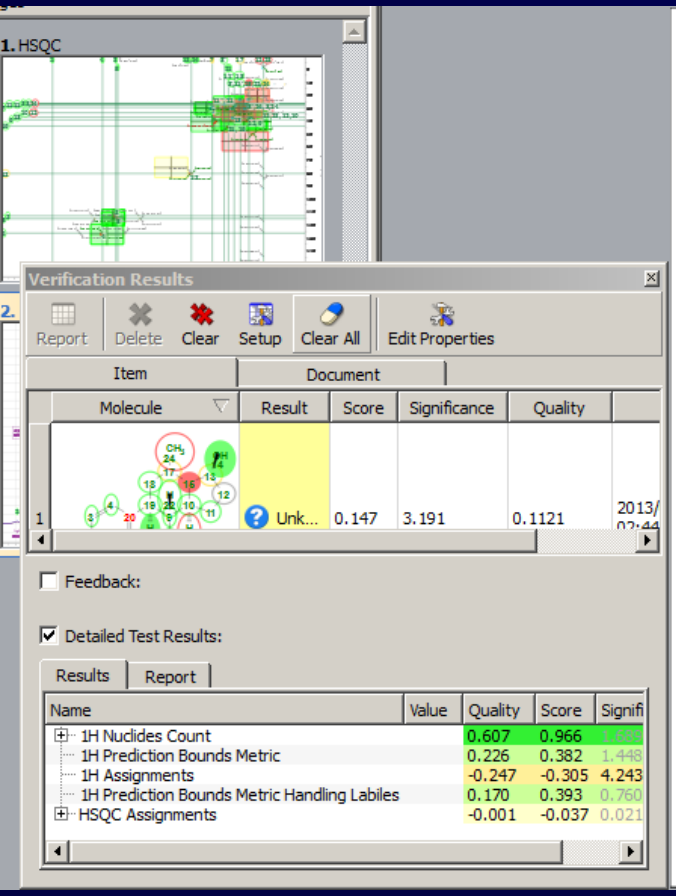
Meeting of GERMN, the Spanish NMR Discussion Group, Santiago de Compostela, Spain, 27 Sep 2013, held just after SMASH

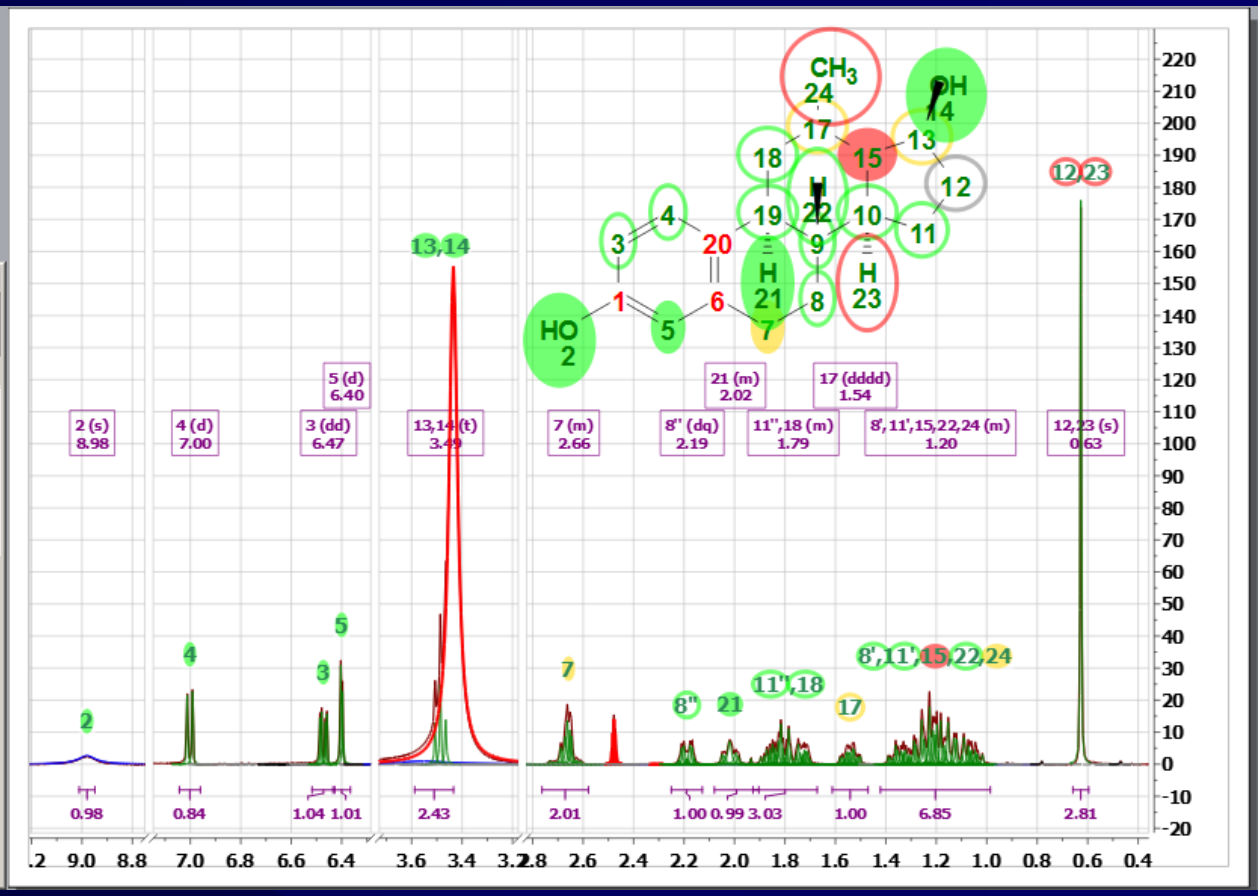
Estradiol 400 MHz, using 1H and HSQC



Meeting of GERMN, the Spanish NMR Discussion Group, Santiago de Compostela, Spain, 27 Sep 2013, held just after SMASH

Estradiol, modified mol, 400 MHz, using 1H and HSQC





Acknowledgements

All the Mestrelab Research people,
particularly

Felipe Seoane

Carlos Cobas

Esther Vas

Mike Bernstein

Manuel Peres

……

the TTT = T^3,
the *Testing and Tuning Team*

+ MODGRAPH (the «predictors»)

Thank You for your Attention

All slides will appear on ebyte.it under

DOI: 10.3247/SL4Nmr13.007