

# 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](http://www.ebyte.it))

This online document's DOI : [10.3247/SL4Nmr13.007](https://doi.org/10.3247/SL4Nmr13.007)

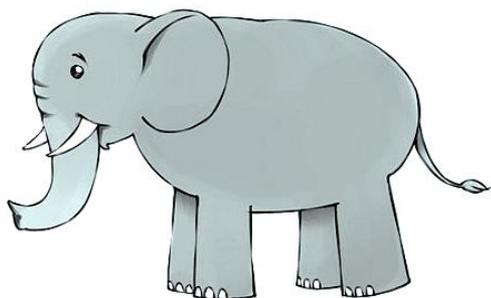
# Why do I work with the Mnova guys?

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

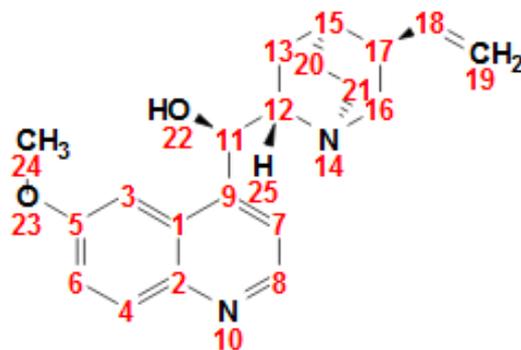


... and also because **Small Molecules Are Still** so Hot !

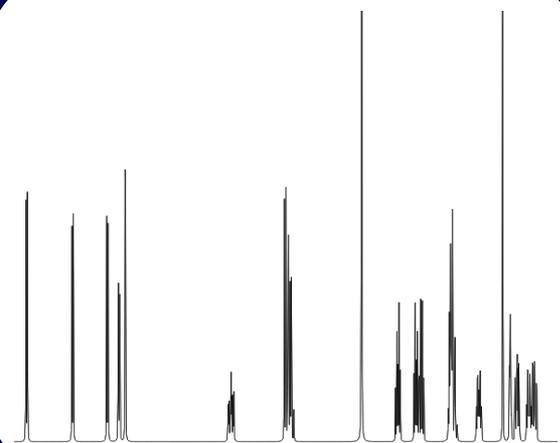
# No REAL thing can be totally predictable! Everything real is infinitely complex and fuzzy!



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



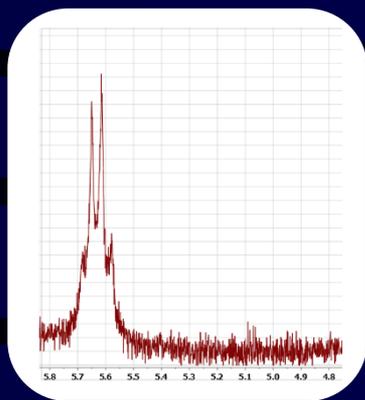
This is NOT a real molecule!  
It's just a structural sketch of  
a molecule.



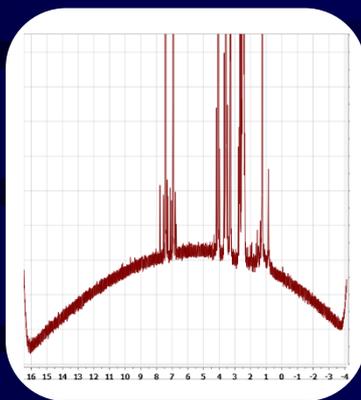
This is NOT a real spectrum!  
It's a naive, simulated NMR  
spectrum.

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

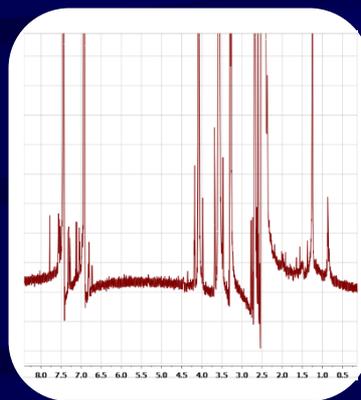
# Few examples of undesirable spectral artifacts



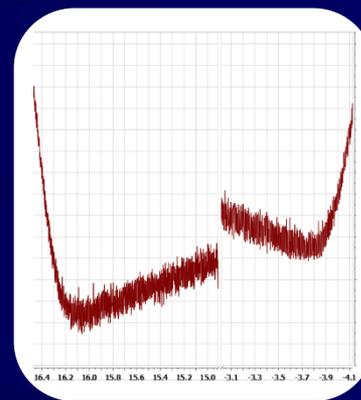
Noise



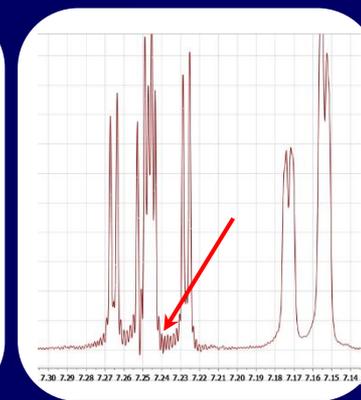
Baseline roll



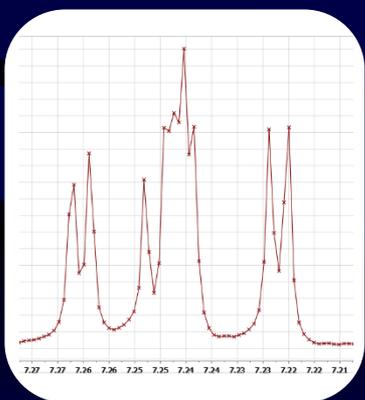
Imperfect phasing



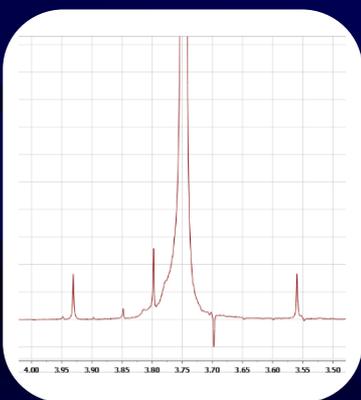
Bruker & Jeol  
smileys & brownies



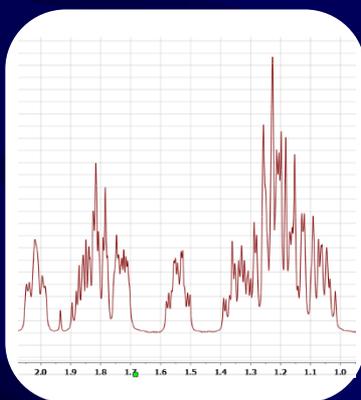
FID truncation  
effects



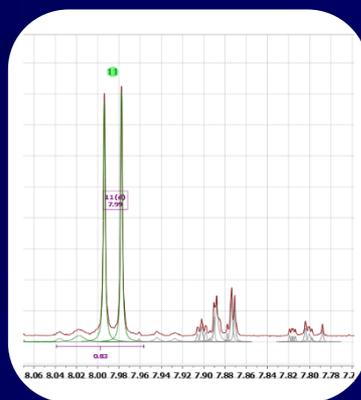
Underdigitization



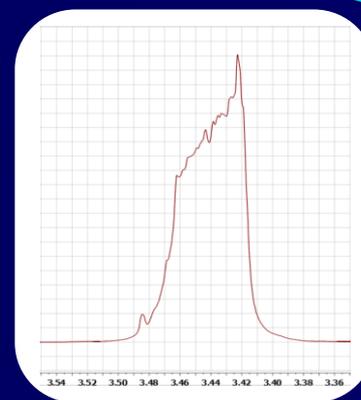
Rotation sidebands



Peaks overlap



Impurities peaks



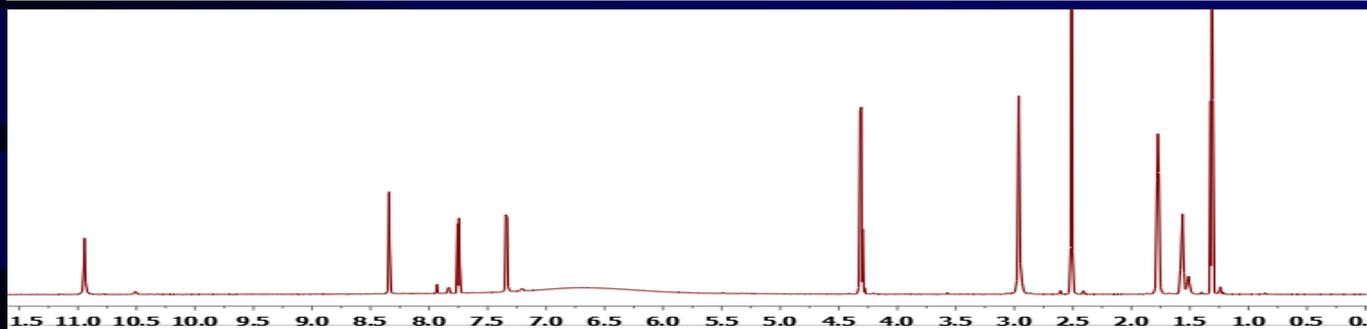
Sample temperature  
drift effects

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

# 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,  $^{13}\text{C}$  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!

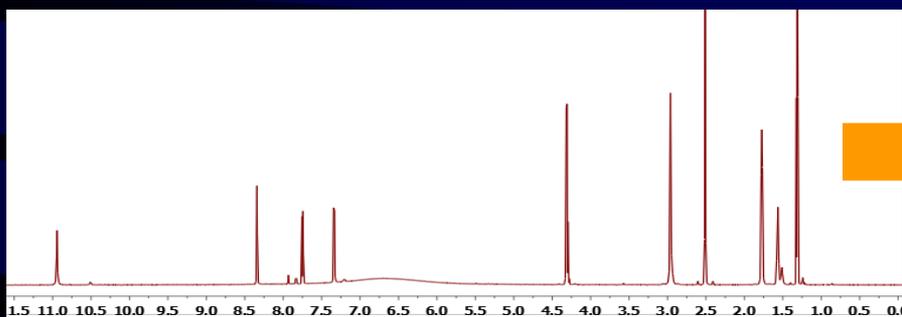
**=> there is a big communication problem**

# 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

What does a programmer see?

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

	ppm $\Delta$	Intensity	Width	Area	Type	Flags	Kurtosis
44	2.612	0.474	1.224	8.289	Compound	None	0.000
45	2.437		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	2.302		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

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 1<sup>st</sup> 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 appearing:

- Internal (Padé based)
- External (CRAFT)

## References:

DOI: 10.3247/SL2Nmr08.011

DOI: 10.3247/SL3Nmr09.003

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

Carlos Cobas<sup>1</sup>, Felipe Seoane<sup>1</sup>, Stanislaw Szykora<sup>2</sup>

<sup>1</sup>Mestrelab Research, Xosha Pousán 6, Santiago de Compostela, 15706 Spain; carlos@mestrelab.com  
<sup>2</sup>Enza NMR, Via Feltrino Veneto 20/C, Cusiano Primo (MI), Italy 1-20022; szykora@enza.it

**INTRODUCTION to the GSD concept**

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

- A set of spectral "peaks" with NMR characteristics (such as T<sub>2</sub>-pulse dependent relaxation angle and phase).
- Baseline distortions due to receiver artifacts (channels U and V offset), probe ringing and acquisition dead time.
- Distortions and other non-linearities (RF amplifiers saturating the detector but not generated by the nuclei in the sample).
- The ubiquitous noise in 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 the one indicated as "NMR" carries meaningful information. Using a number of distinct algorithms, an experienced operator copes with the co-presence reasonably well when processing the data manually, but if other plays a base-wide automatic routine, such as fitting theoretical spectra to experimental data or, for that matter, just attempting a simple integration, it fails.

The first goal of GSD is to extract from the spectrum only part (d):

The second problem one encounters regards the imperfections of part (f) itself. In particular:

- Peaks that being Lorentzian, peaks are distorted by field inhomogeneity and molecular dynamics.
- Due to their finite linewidths, peaks overlap heavily giving rise to complex correlations.
- Some peaks are overlapping and/or partially obscured and, in some contexts, isotropic signals such as -CH<sub>3</sub> (acetate).

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

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 (hence the name) but actually it goes far beyond it. For one thing, it is fully automatic and superior in terms in deciding the number of peaks and their starting parameters (peak recognition). Another way of looking at GSD is as an information filter which discards the undesirable parts of the spectrum. The peaks list it produces contains all the desirable information present in the original spectrum but none of the undesirable one. All subsequent data processing tasks (such as integration, digital J-Compensation, structure verification and/or elucidation) can work as cleanly on this numeric information without following any more usual the expected original spectrum. Moreover, the list is very easy to edit both by the User and automatically.

### UNDER the HOOD

GSD operates in a fully automatic mode on either real or complex 1D spectra. At present, an approximate prior phase adjustment is required. What follows is of course just a brief description of the algorithm. The individual steps will be described on a spectrum of lactose. All operations were carried out automatically and on the whole spectrum but, for reasons of clarity, the performance is illustrated on selected narrow windows.

**Step 1: Noise and mean linewidth estimate.**

The pre-requisite for a correct operation of GSD is a reliable estimate noise and mean linewidth. To do that, we have developed new and smart trigonometric 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 (1). The trick, of course, is a correct setting of the GSD parameters (number of points and order), based on the estimated noise and linewidth. The use of derivatives is not done in any "passive" dependence. Moreover, the use of the 2nd and 3rd derivatives is not done in a "passive" dependence. Moreover, the use of the 2nd and 3rd derivatives is not done in a "passive" dependence. Moreover, the use of the 2nd and 3rd derivatives is not done in a "passive" dependence.

Figure 1a shows the 4-4.2 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. Points where recognizable local minima and maxima occur in each of the three arrays are approximately flagged in case of a complex spectrum, this is done for both the real and imaginary parts.

Figure 1b shows the same portion as before. Negative marks correspond to local minima, positive ones to local maxima. The 2-3 divisions long marks refer to the original spectrum, the shorter ones that a cluster to the 1st derivative and the one division tall marks to the negative second derivative (in case of approximation, the mark lengths are adjusted).

**Step 4: Peaks recognition.**

All except noise peaks are detected and marked, using the peaks maxima in the second derivative spectrum (real and imaginary) as starting points. Peaks with estimated linewidth smaller than 2 points are discarded as spurs.

Figure 1c shows the peak recognition algorithm at work. The 1st and 2nd derivative marks are as before, but the 1st derivative marks points (positive/negative) are replaced by a division tall rectangles. Notice that some very small peaks barely above noise were correctly picked up, some 1st derivative rectangles are topped by more than one and 2nd derivative peak and some by none, etc. There is not enough room on the plot to include all the points of the peak recognition result.

**Step 5: Fine setting of each peak parameter (frequency, linewidth, height).**

To avoid baseline dependence, the parameter estimate is based on the knowledge of only the apical points (maxima and minima) of the 1st and 2nd derivatives spectra. The result is a new Peaks List. The latter can be used to generate a synthetic spectrum which already incorporates in it both the experimental one but in kind of any baseline and noise.

Figure 1d shows the 3.6-4.2 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 excellent as a starting point for fit.

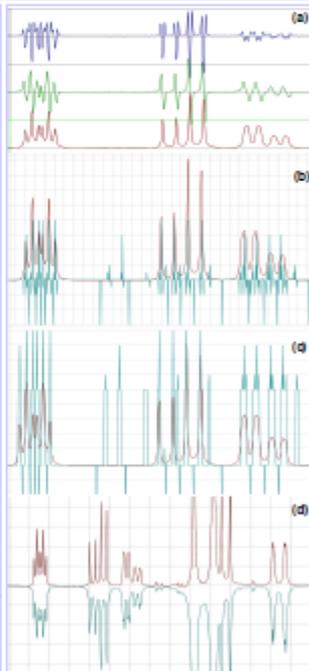
**Step 6: Refinement of the parameters of all the peaks in the list.**

There are several modes how to do this which differ somewhat in final quality and dramatically in speed (work in progress). We will eventually choose two modes - 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 (for baseline and noise) and noise (noise). This step is not illustrated here because the fit is perfect - disturbed only by the baseline artifacts in the experimental spectrum which are absent in the synthetic one as shown in Figure 1e.

### What is it good for?

We have already mentioned several of the advantages of passing a spectrum through the GSD procedure and reducing it to an equivalent and a list of peaks. There are several other advantages which are replaced by summing peak areas with a total evaluation of peak overlap errors. Do think about resolution enhancement, etc, but that matter baseline correction - the only method so far that can potentially find out the baseline of both real and imaginary parts.

For us, however, this is primarily a technique also towards automated and/or computer aided molecular structure verification and elucidation. The Peaks List, rather than the spectrum, becomes the input to further editing and analysis aimed at the determination of all compatible spin systems and, eventually, molecules.



ENZA NMR  
MESTRELAB RESEARCH

DOI: 10.3247/SL2Nmr08.011

Developed in collaboration with



# 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

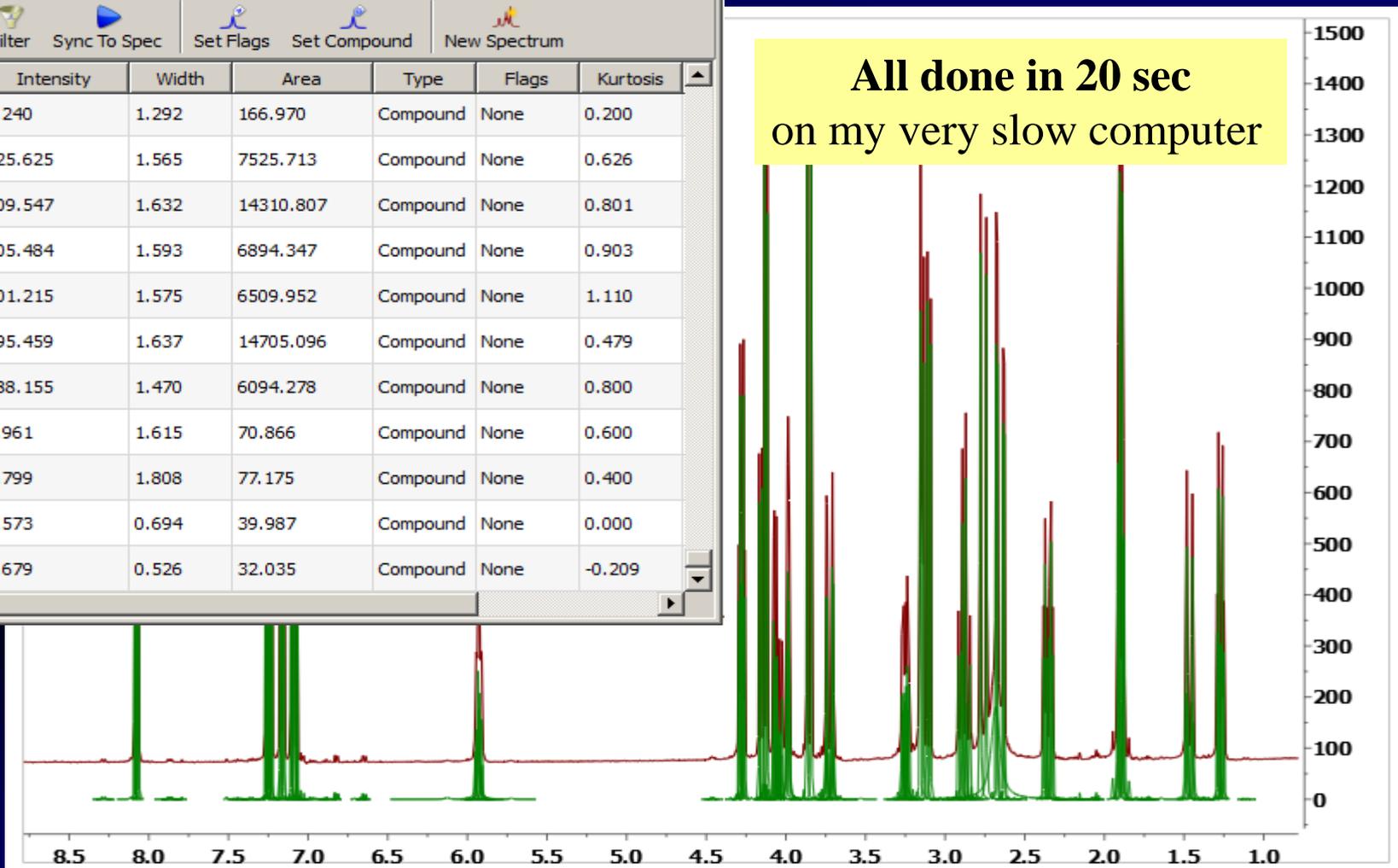
Peaks

Report Peaks Copy Peaks Setup Report Delete Select Peaks

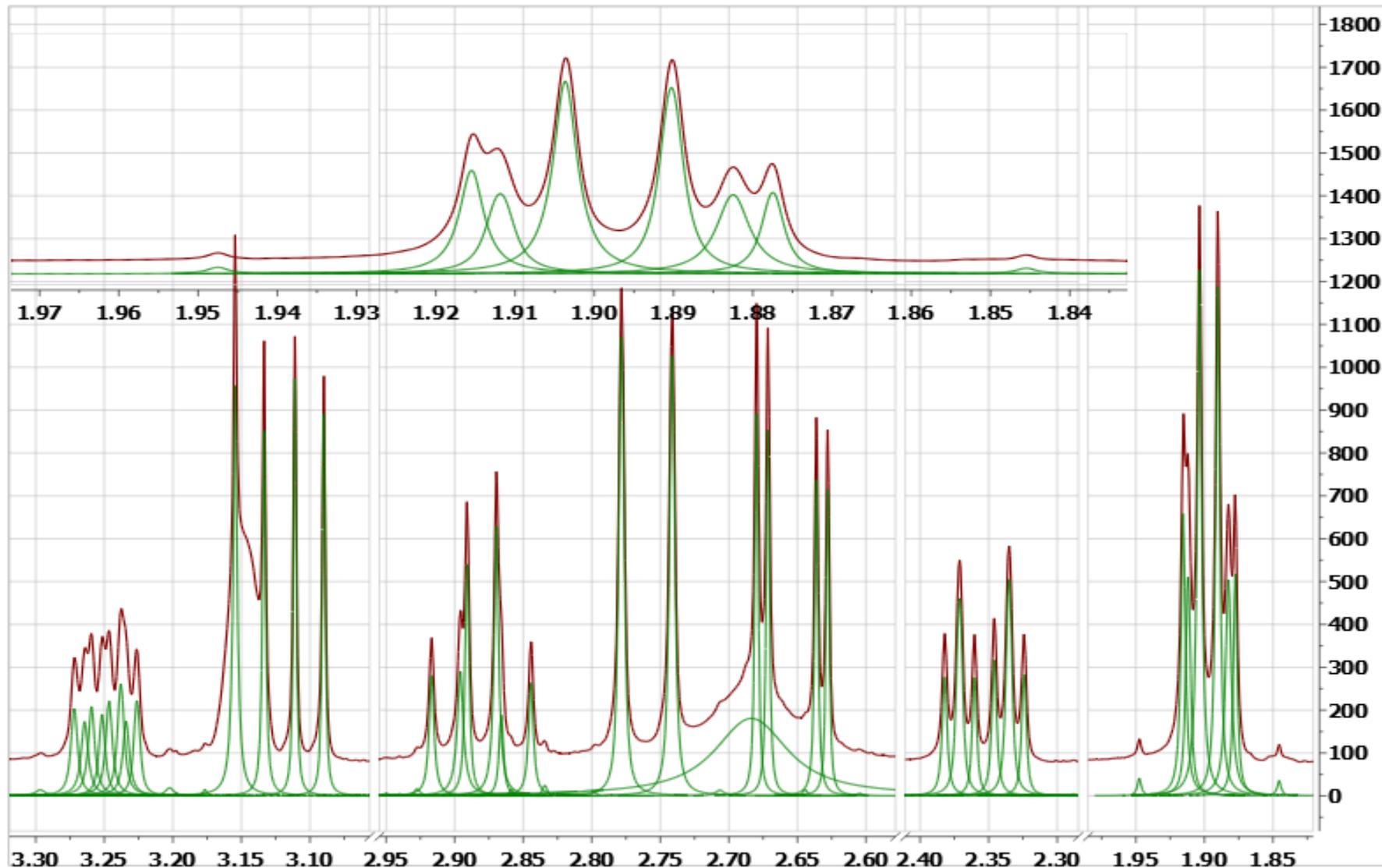
Sync From Spec Filter Sync To Spec Set Flags Set Compound New Spectrum

	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

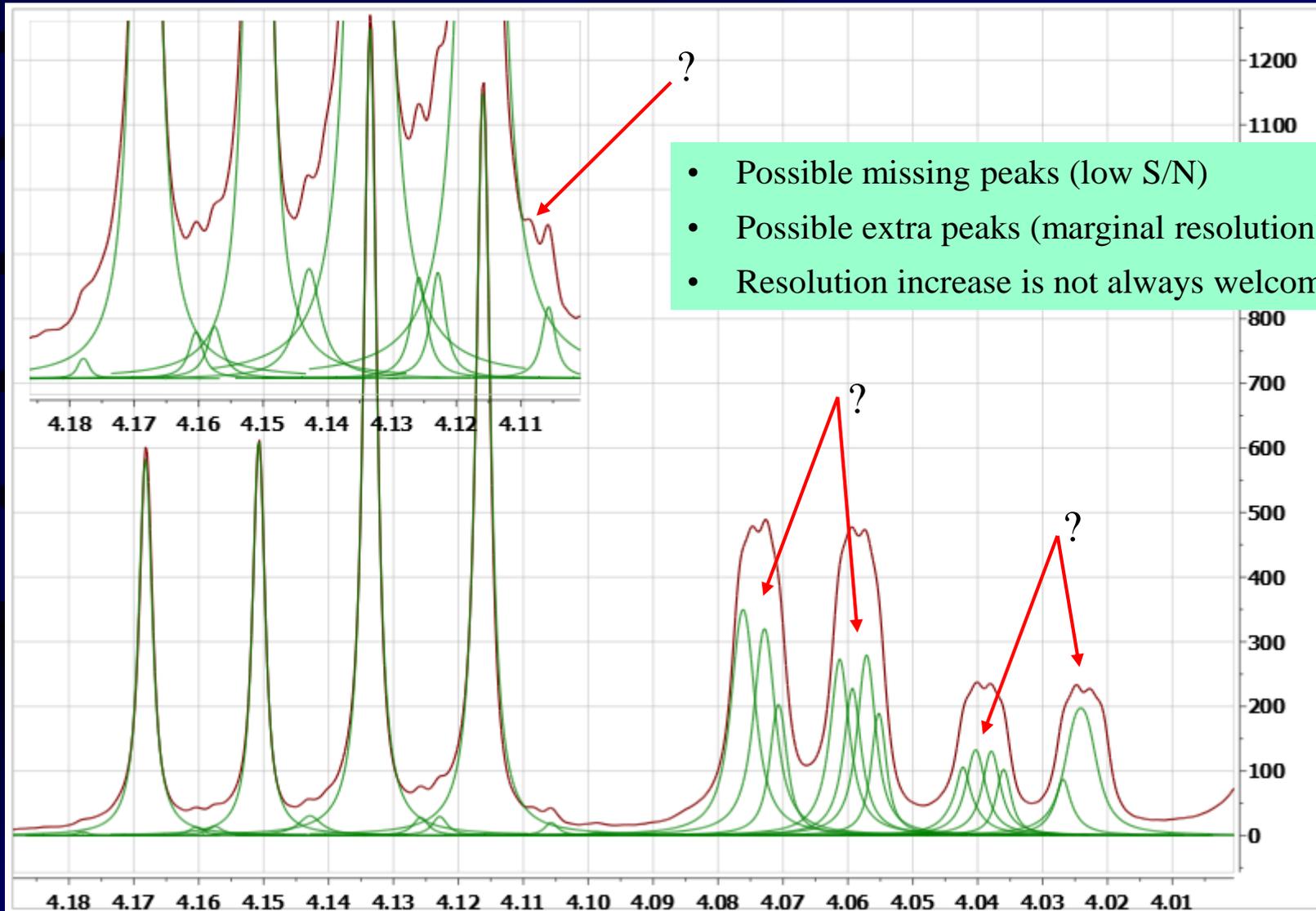
All done in 20 sec  
on my very slow computer



# Examples of peaks detection



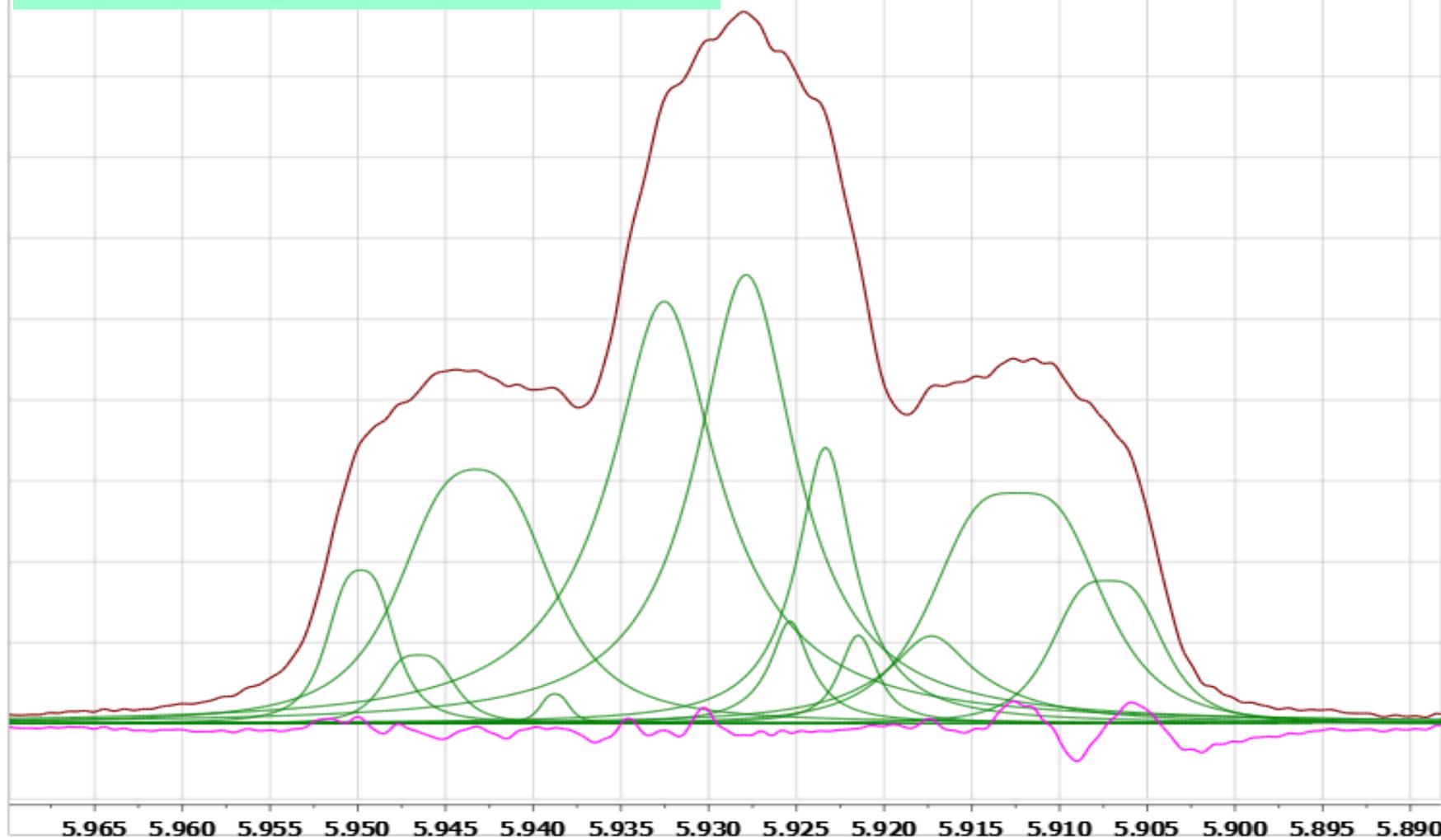
# Limits of GSD and GSD artifacts



# Limits of GSD and GSD artifacts:

Example of a «broken» symmetry in the GSD peaks

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



# Limits and uncertainties 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 II<sup>nd</sup> 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):

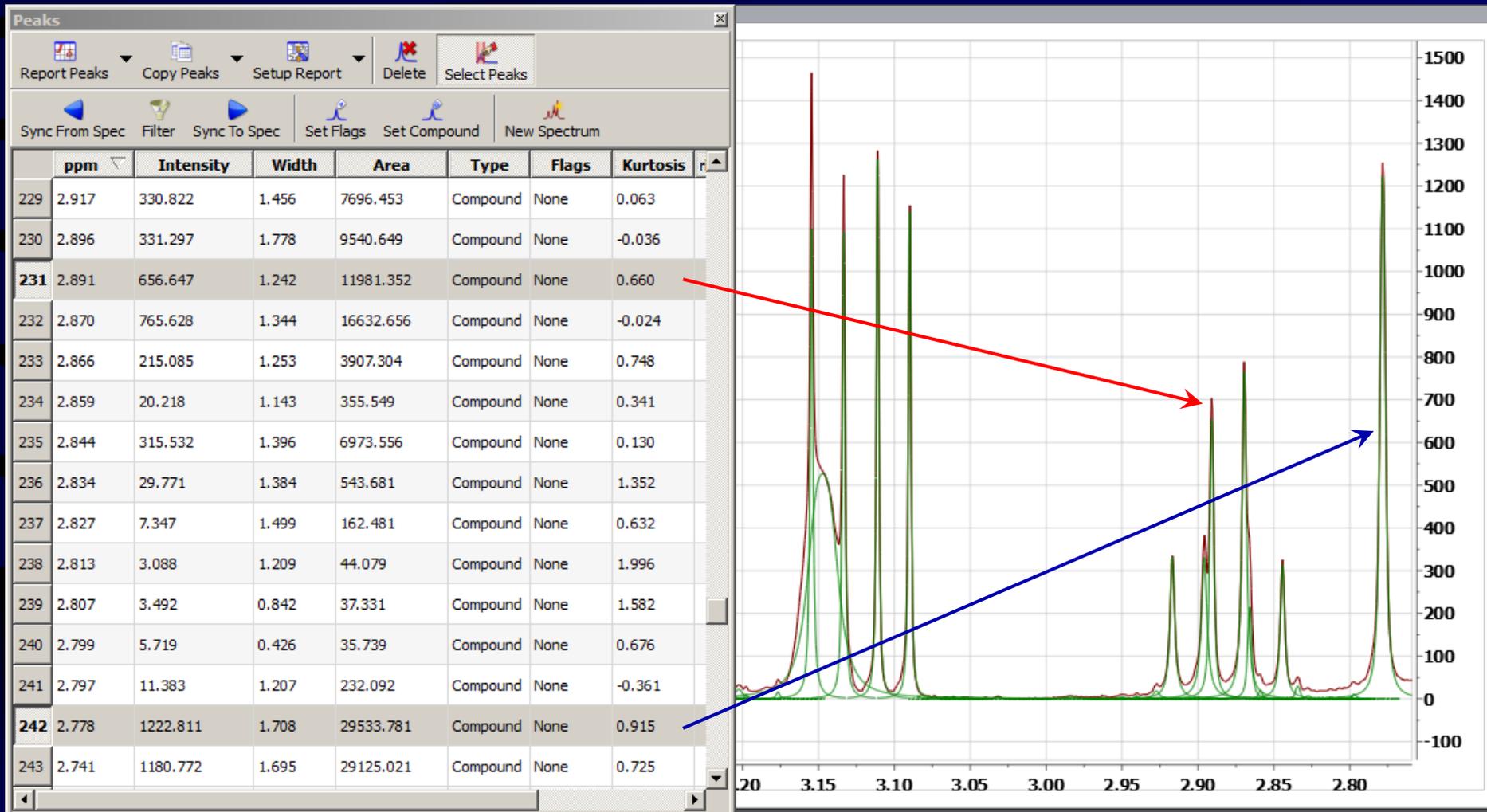
- How much have we gained and how much have we lost?
- Is the balance positive?

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 III<sup>rd</sup> Law of Data Evaluation:  
**Nothing useful comes for free!**

# GSD peaks linewidths and shapes



# 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  $\frac{1}{2}$  nuclei there are  $N \cdot 2^{N-1}$  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 IV<sup>th</sup> 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](http://ebyte.it/library/docs/nmr06a/NMR_FieldNoise_Fid.html))
3. Sample spinning (dtt0)
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 ...

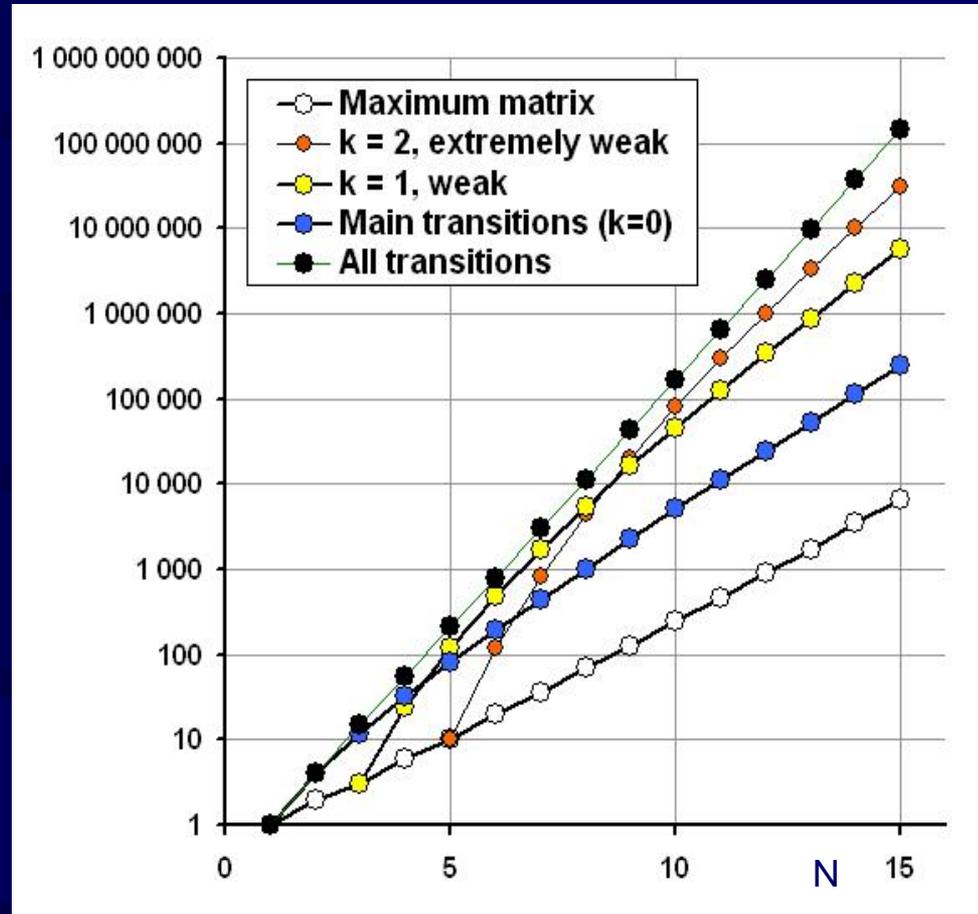
# Overlap of transitions

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.

⇒

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

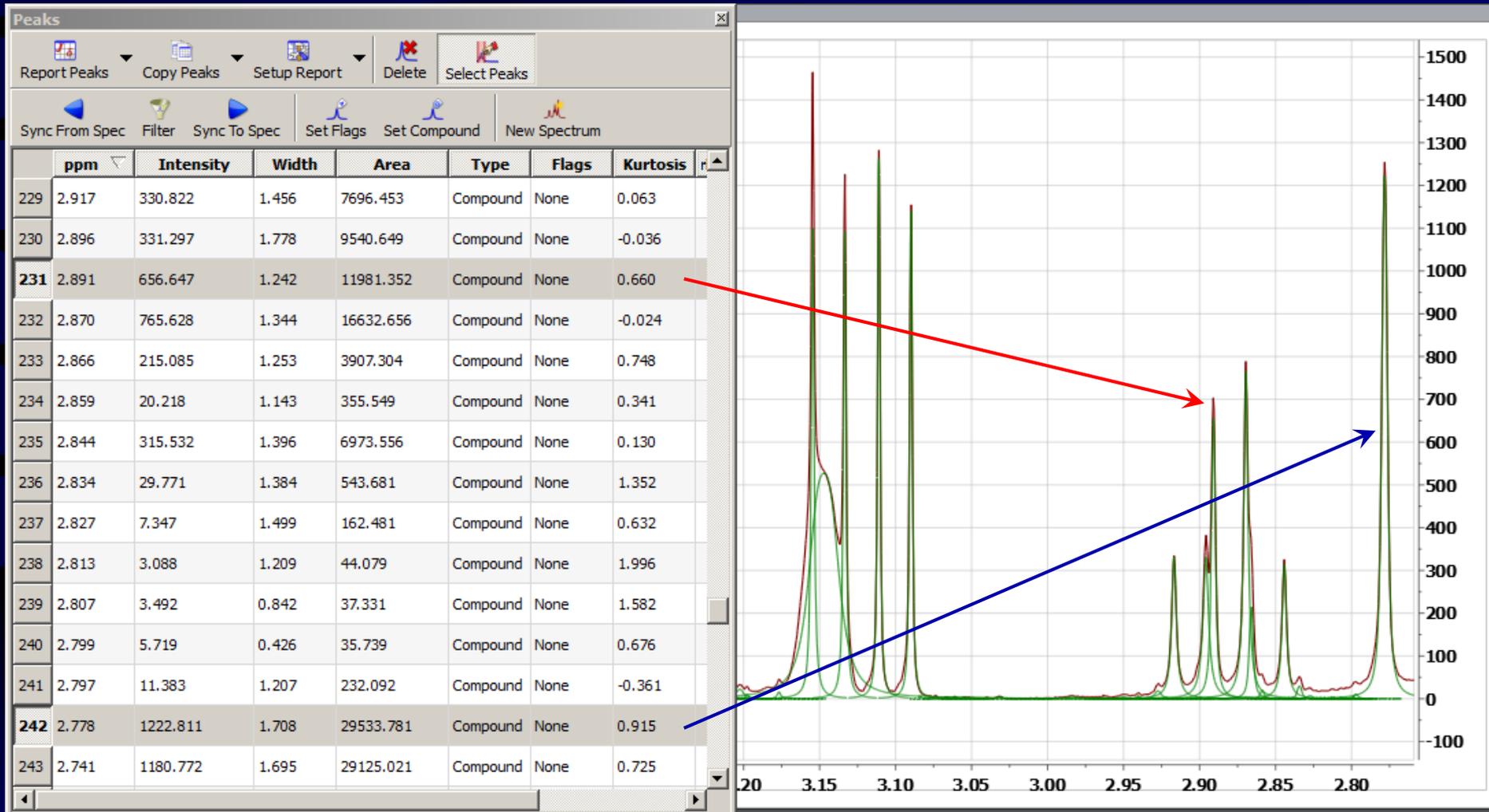
$$GL(z) = (1-k) L(z) + k G(z),$$

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

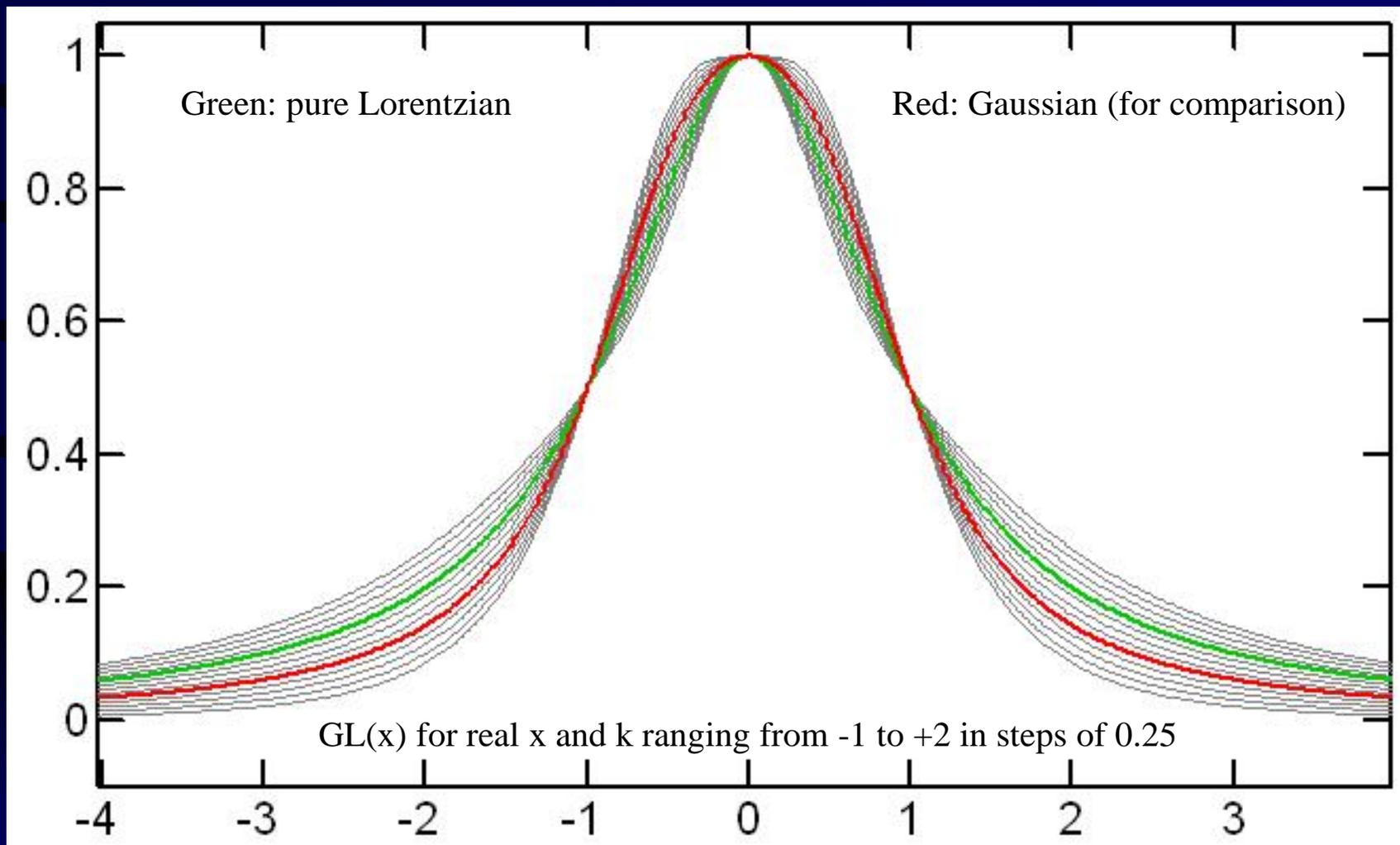
The V<sup>th</sup> Law of Data Evaluation:

**Keep an ace up your sleeve and cheat without shame! It's Science!**

# GSD peaks linewidths and shapes



# 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 VI<sup>th</sup> 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:

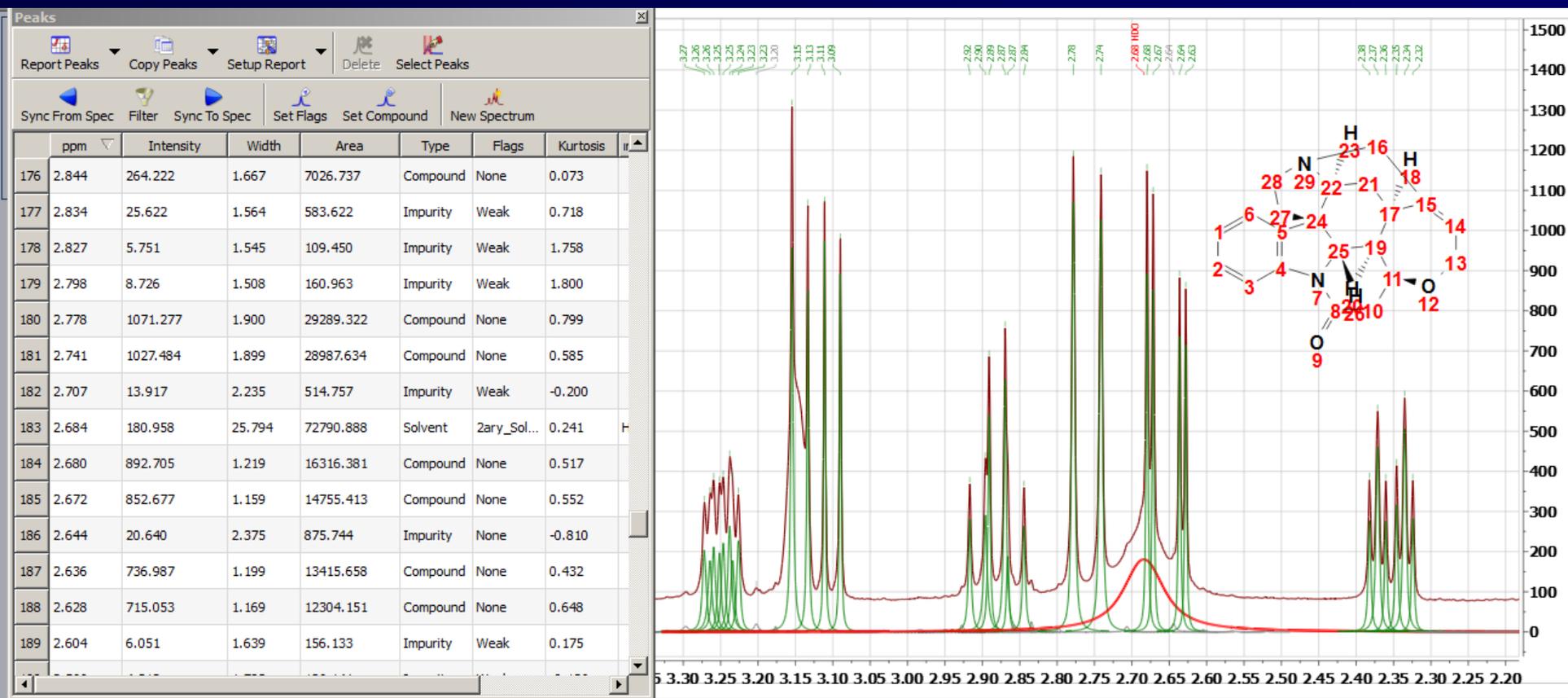
- compound,
- primary or secondary solvent,
- potential labile,
- $^{13}\text{C}$  satellite,
- valid member of a multiplet,
- impurity,
- S or Q reference,
- artifact,
- etc.

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

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

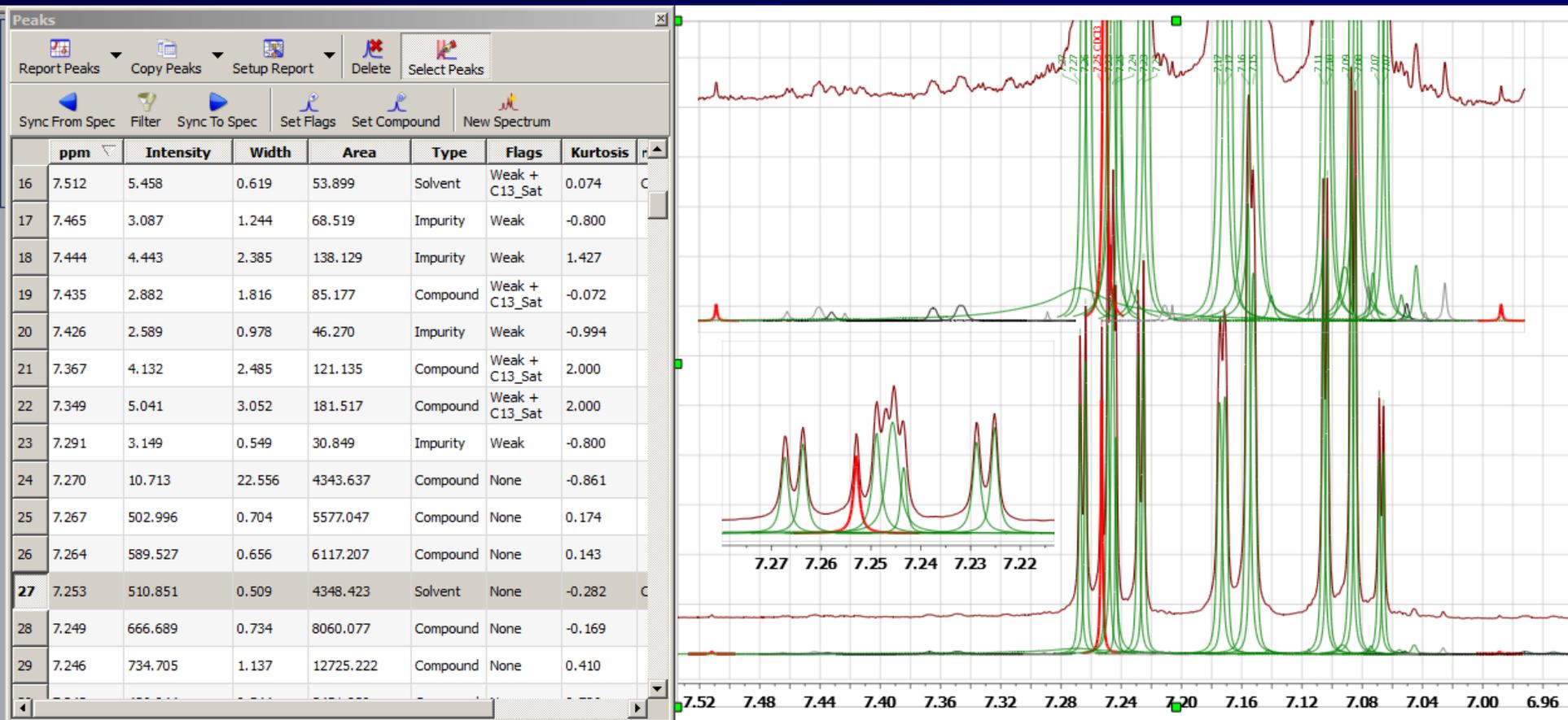
# Peaks editing is primarily fully automatic

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



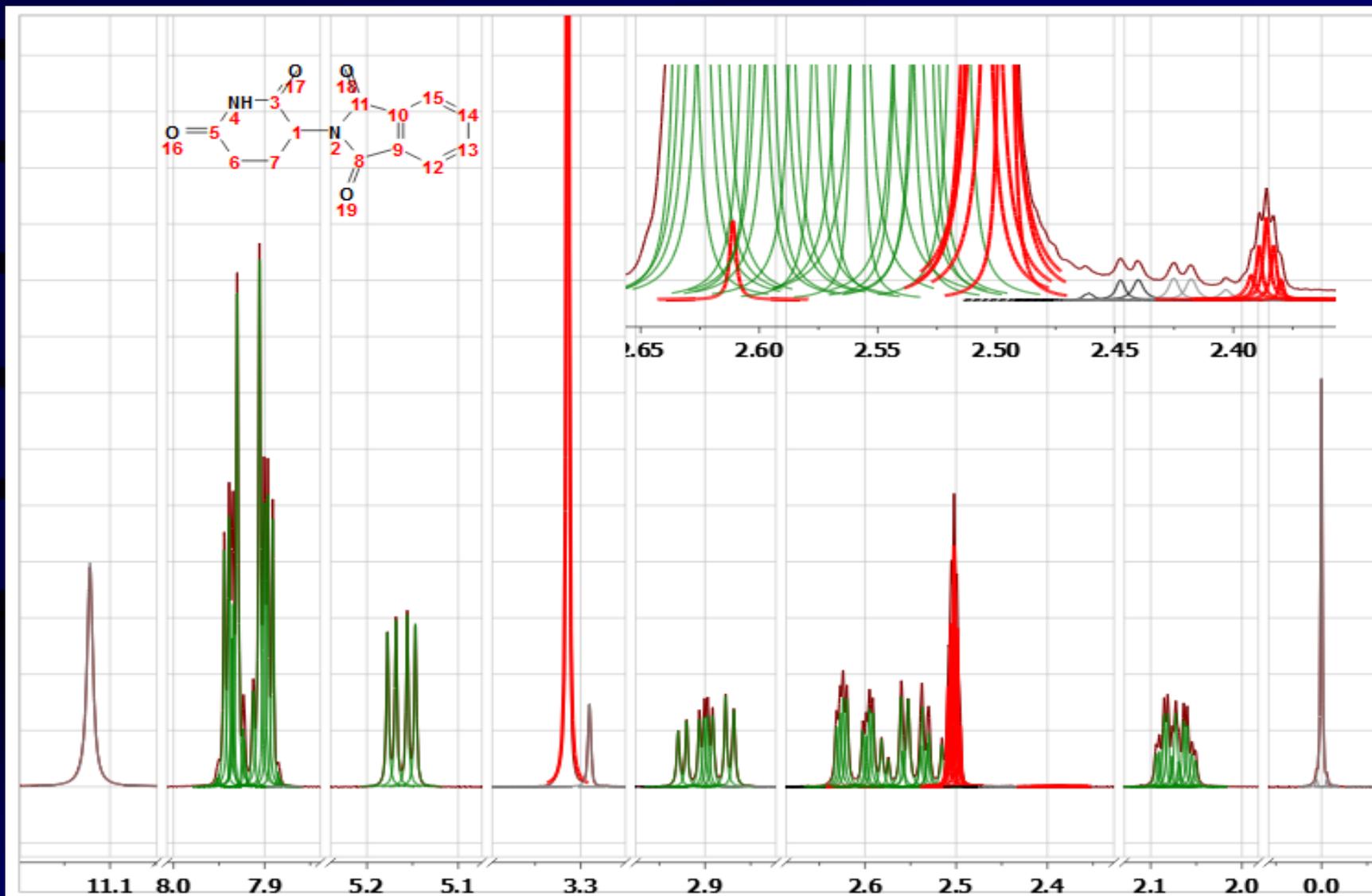
It is also the first plank in NMR spectra evaluation hierarchy  
where specific NMR know-how is used

# CHCl<sub>3</sub> identification in an aromatic multiplet

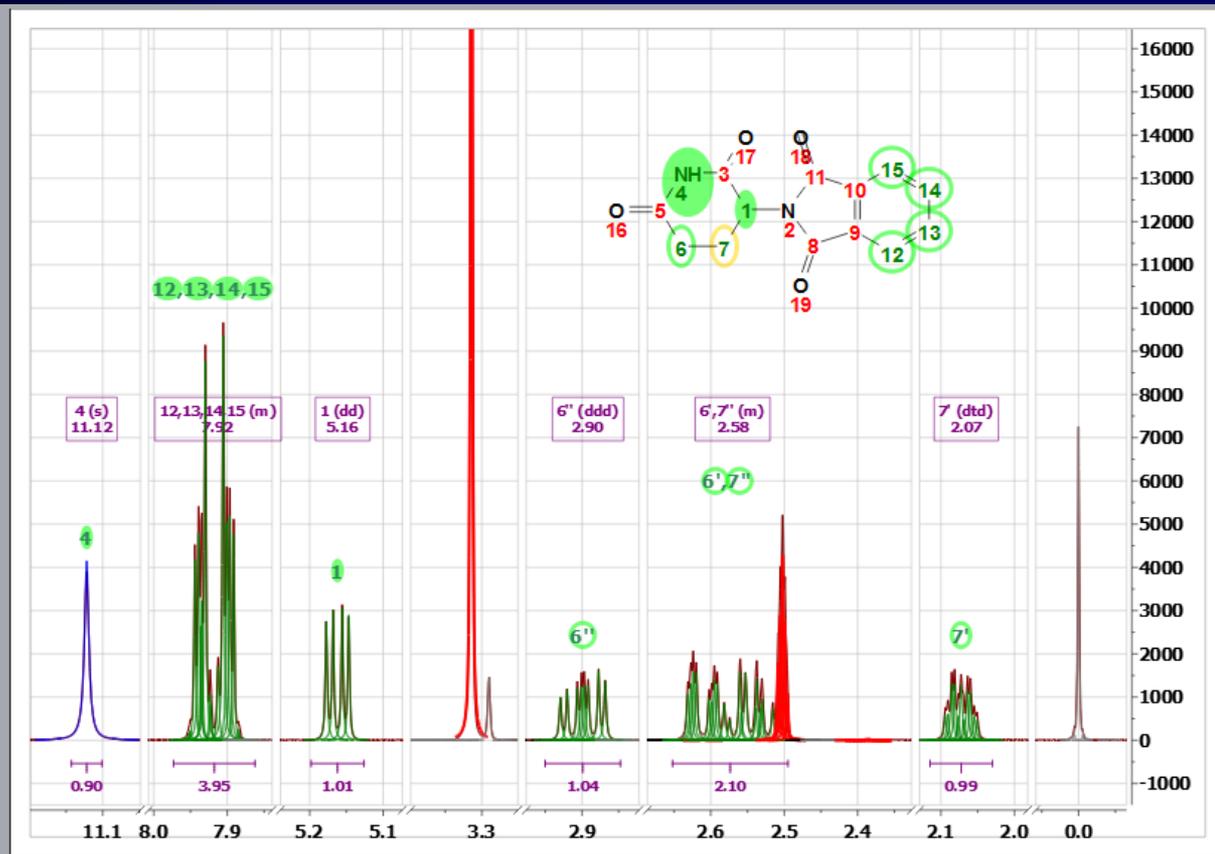
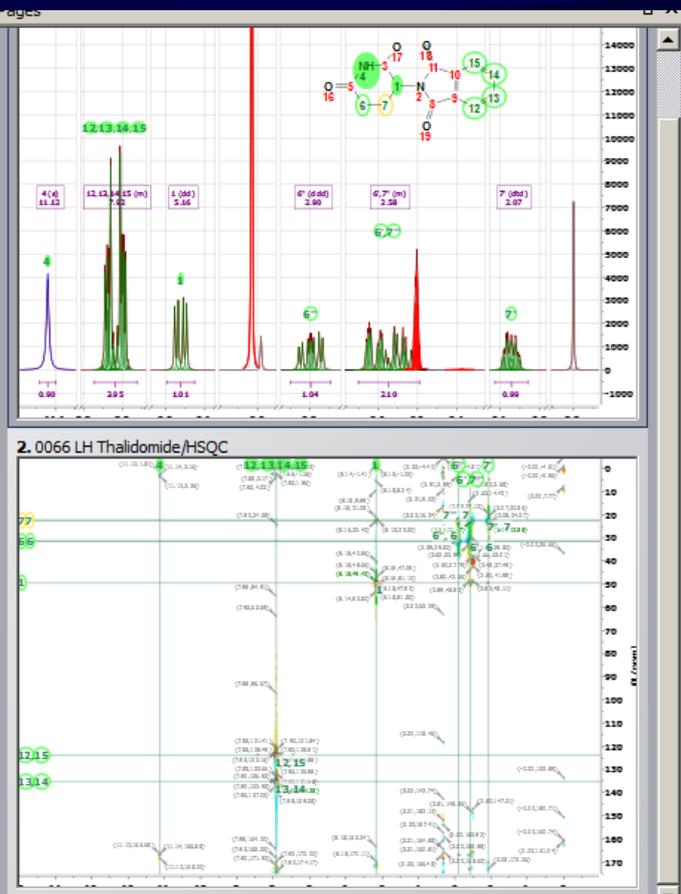


Uses even the <sup>13</sup>C satellites (209.25 Hz apart) and their isotopic shift (the satellite pair center is -2.67 ppb from the main peak)

# 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)

Verification Results

Report Delete Clear Setup Clear All Edit Properties

Item	Document	Molecule	Result	Score	Significance	Quality
1			Pas...	0.804	6.173	0.6919

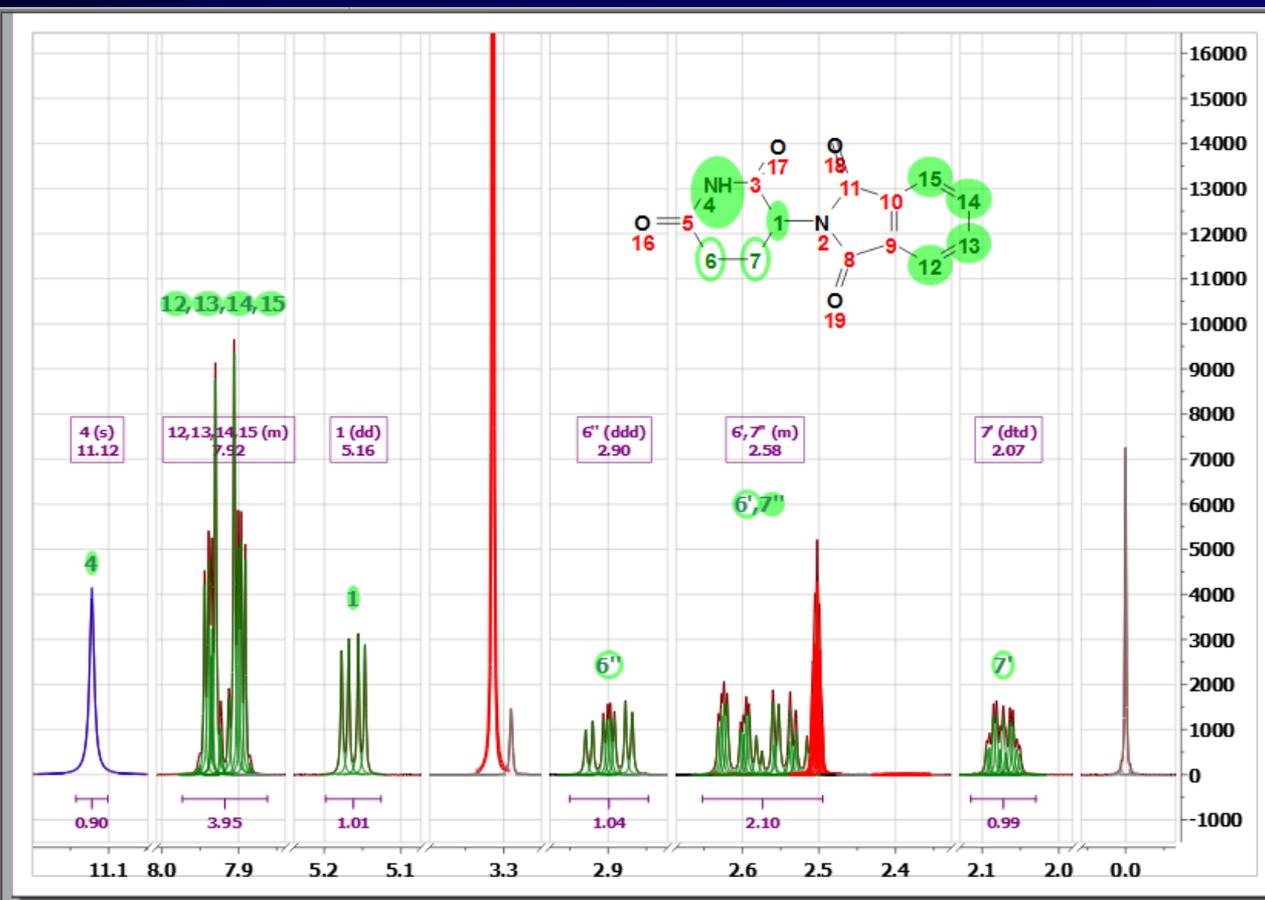
Feedback:

Detailed Test Results:

Name	Value	Quality	Score	Significance
1H Nudides Count	0.365	0.624		
1H Prediction Bounds Metric	0.799	1.000	3.0	
1H Assignments	0.204	0.272	2.0	
1H Prediction Bounds Metric Handling Labiles	0.666	1.000	1.0	

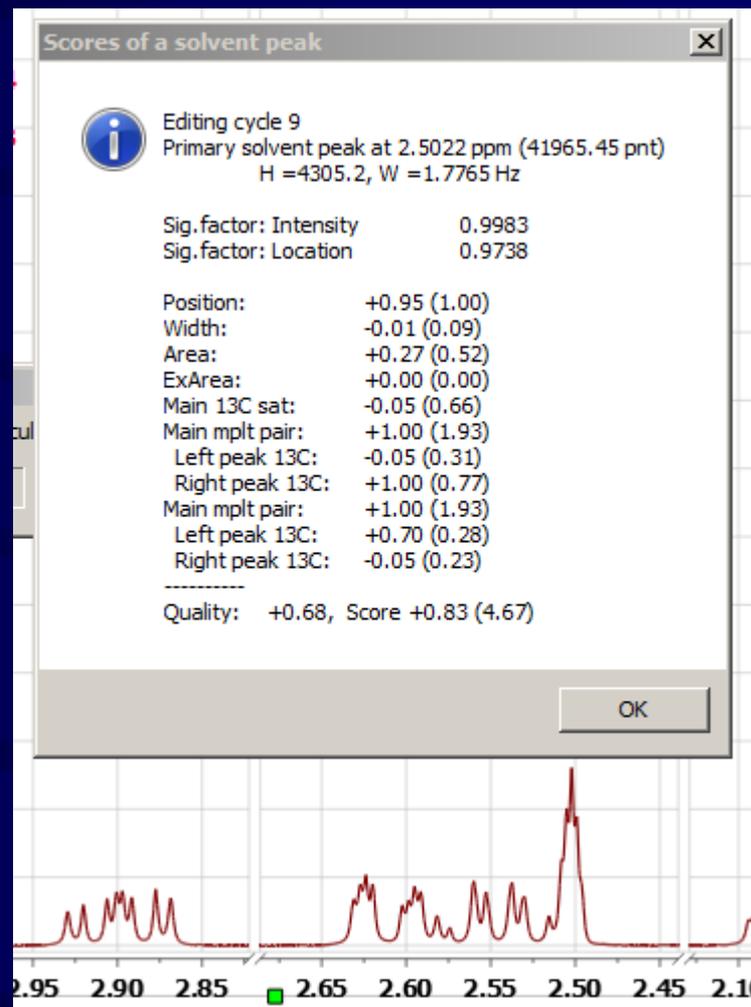
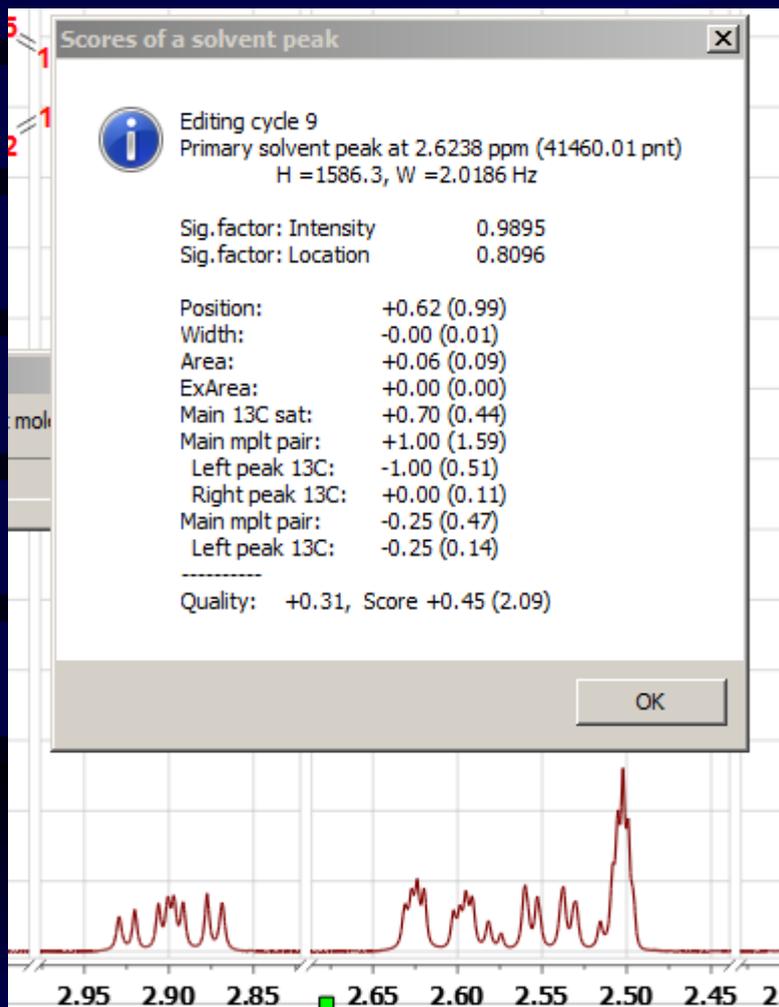
Pages

1, (1) Manuel\_57\_E-735\_DMSO\_10.fid



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  $^{13}\text{C}$  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

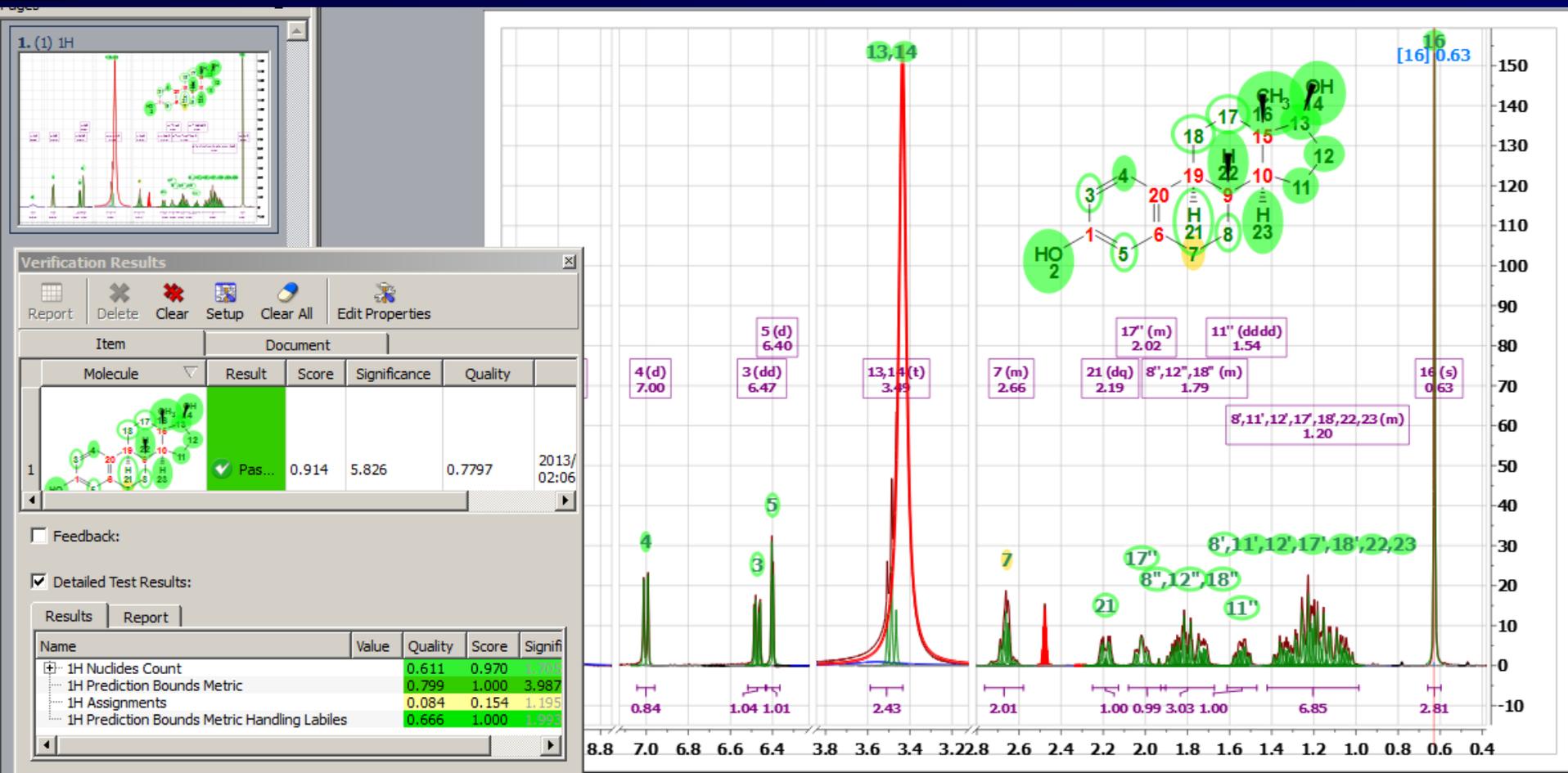
# Once Auto-Editing is finished,

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

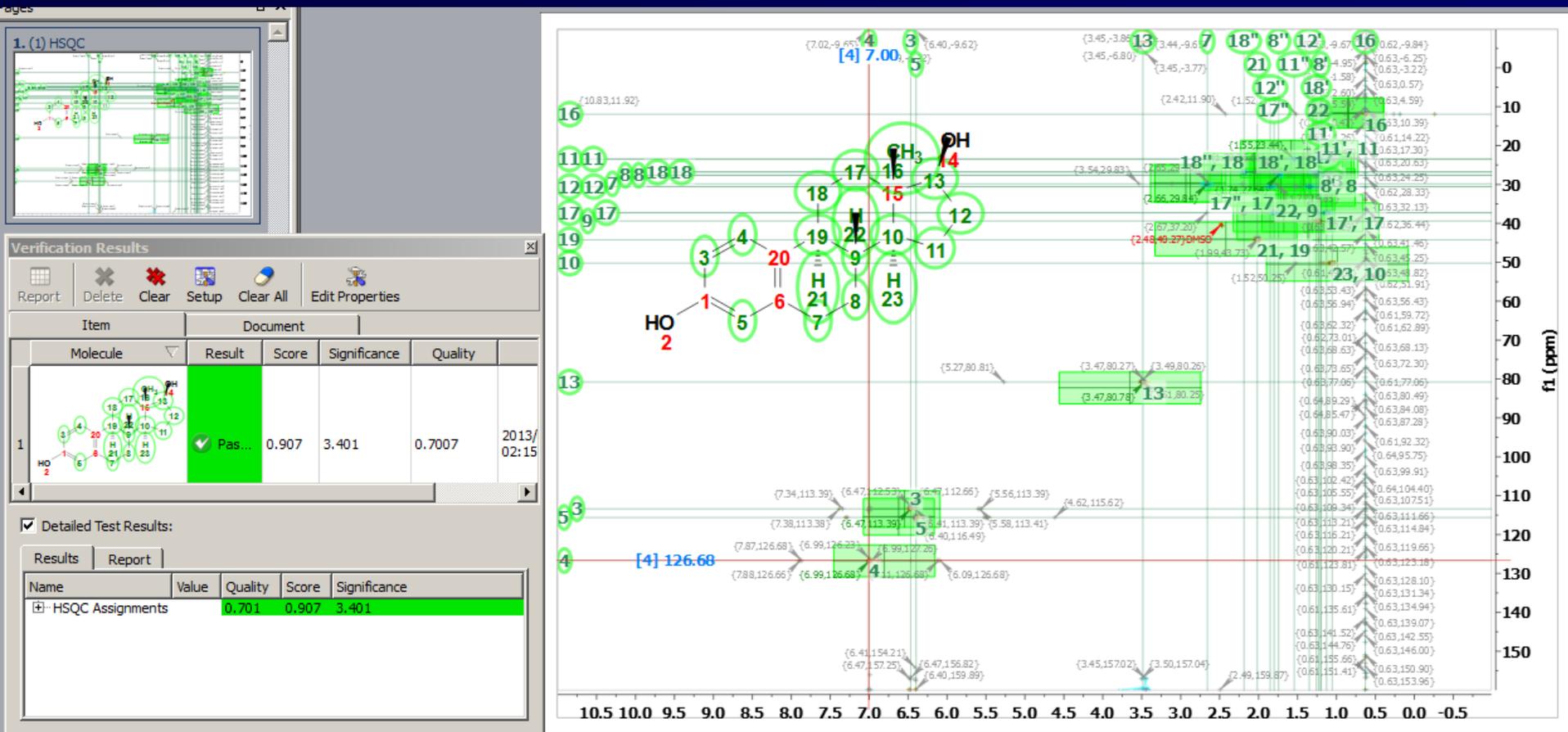
- ASV Automatic Structure Verification
- ASPV Automatic Structure Presence Verification
- ACD Automatic Component Detection
- ASD Automatic Structure Discrimination
- ASE Automatic Structure Elucidation
- ADBV Automatic DataBase Validation
- etc

# Final example(s)

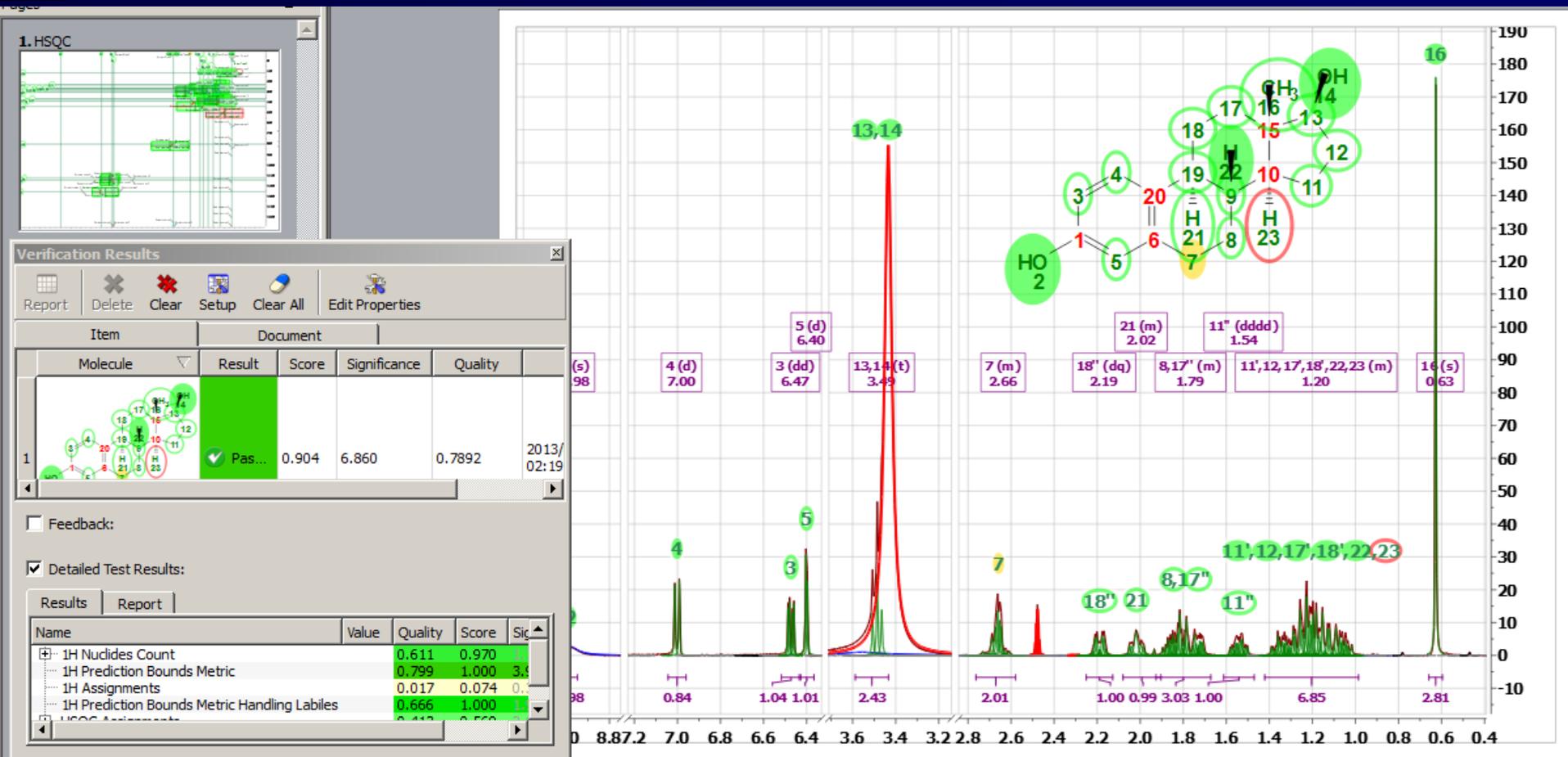
## Estradiol 400 MHz, using 1H only



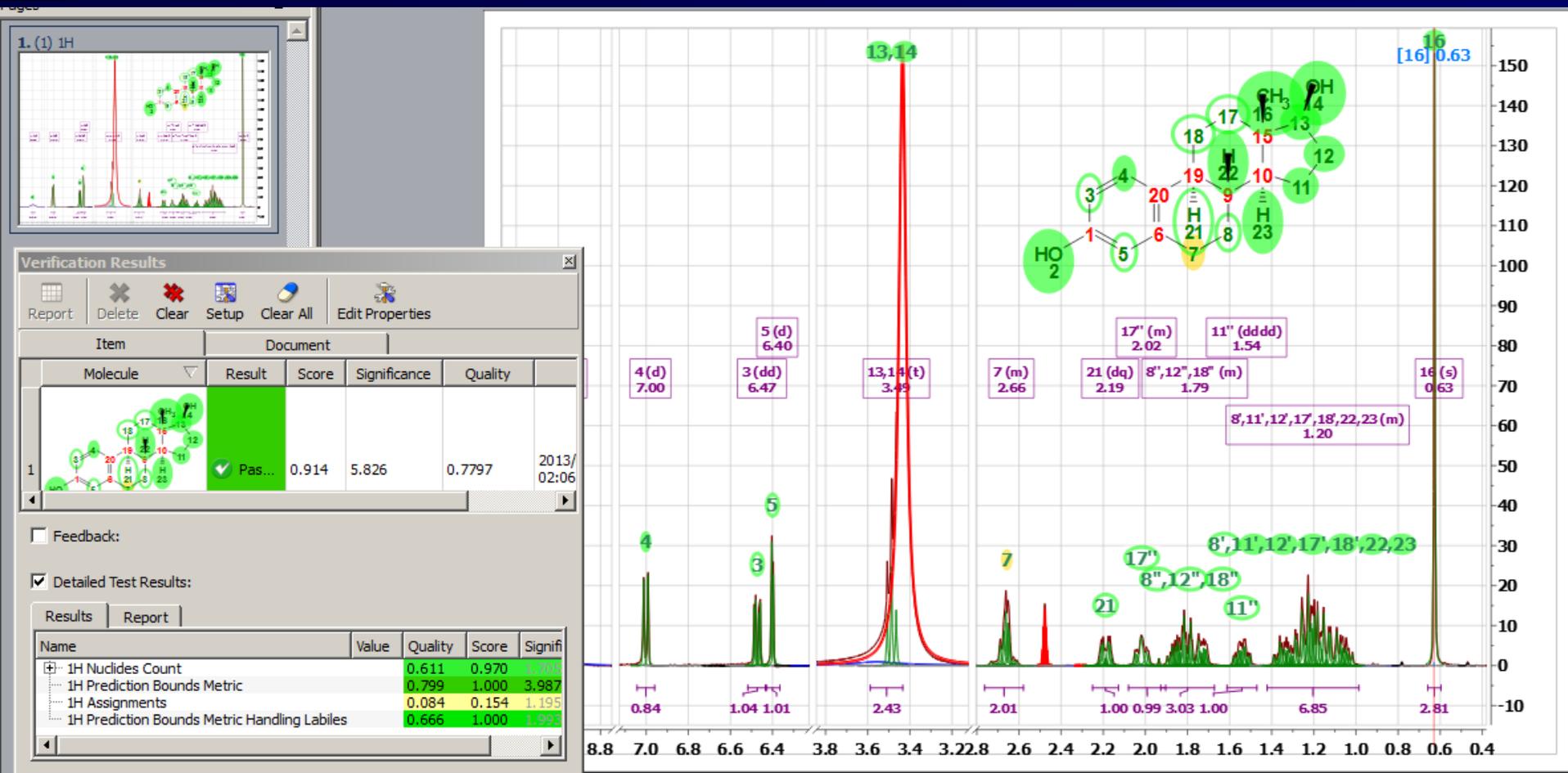
# Estradiol 400 MHz, using HSQC only



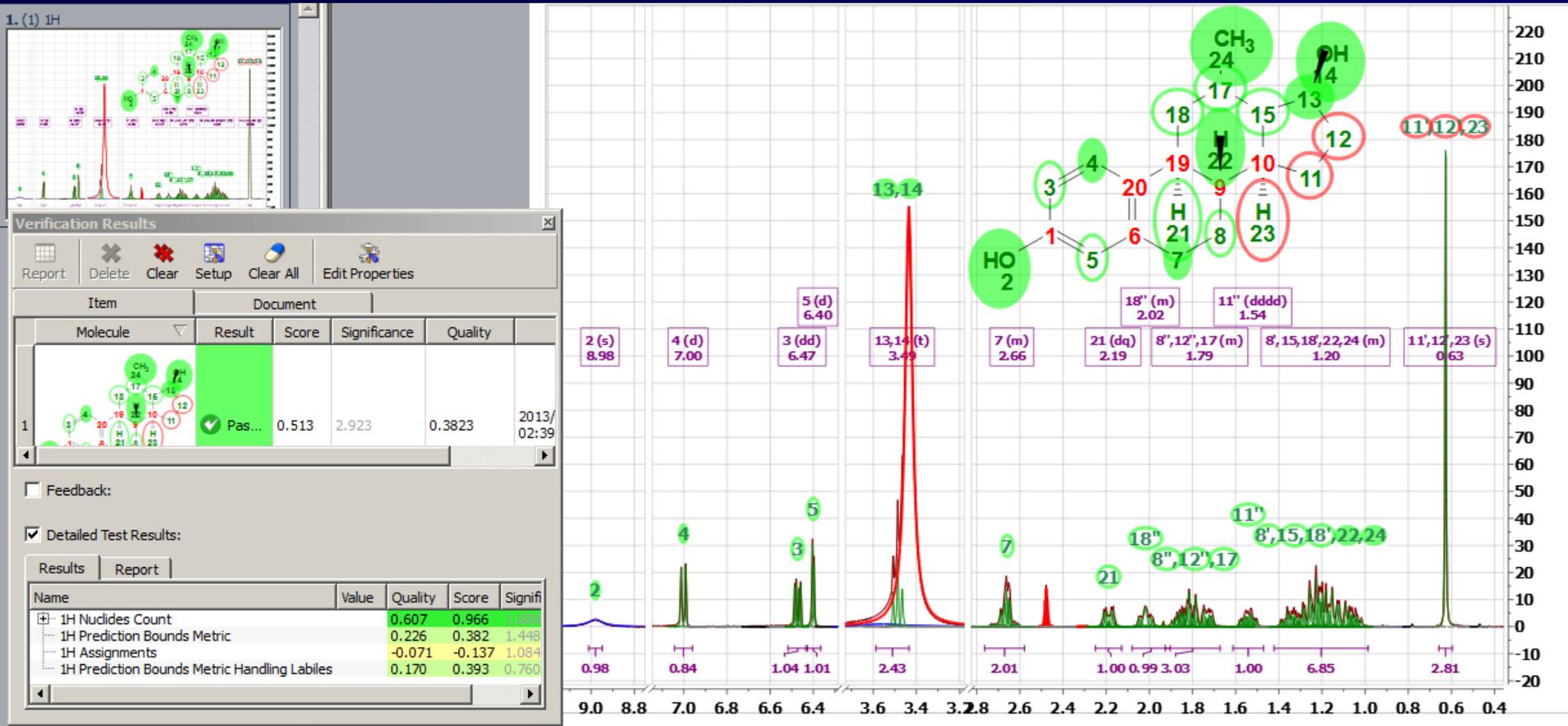
# Estradiol 400 MHz, using 1H and HSQC



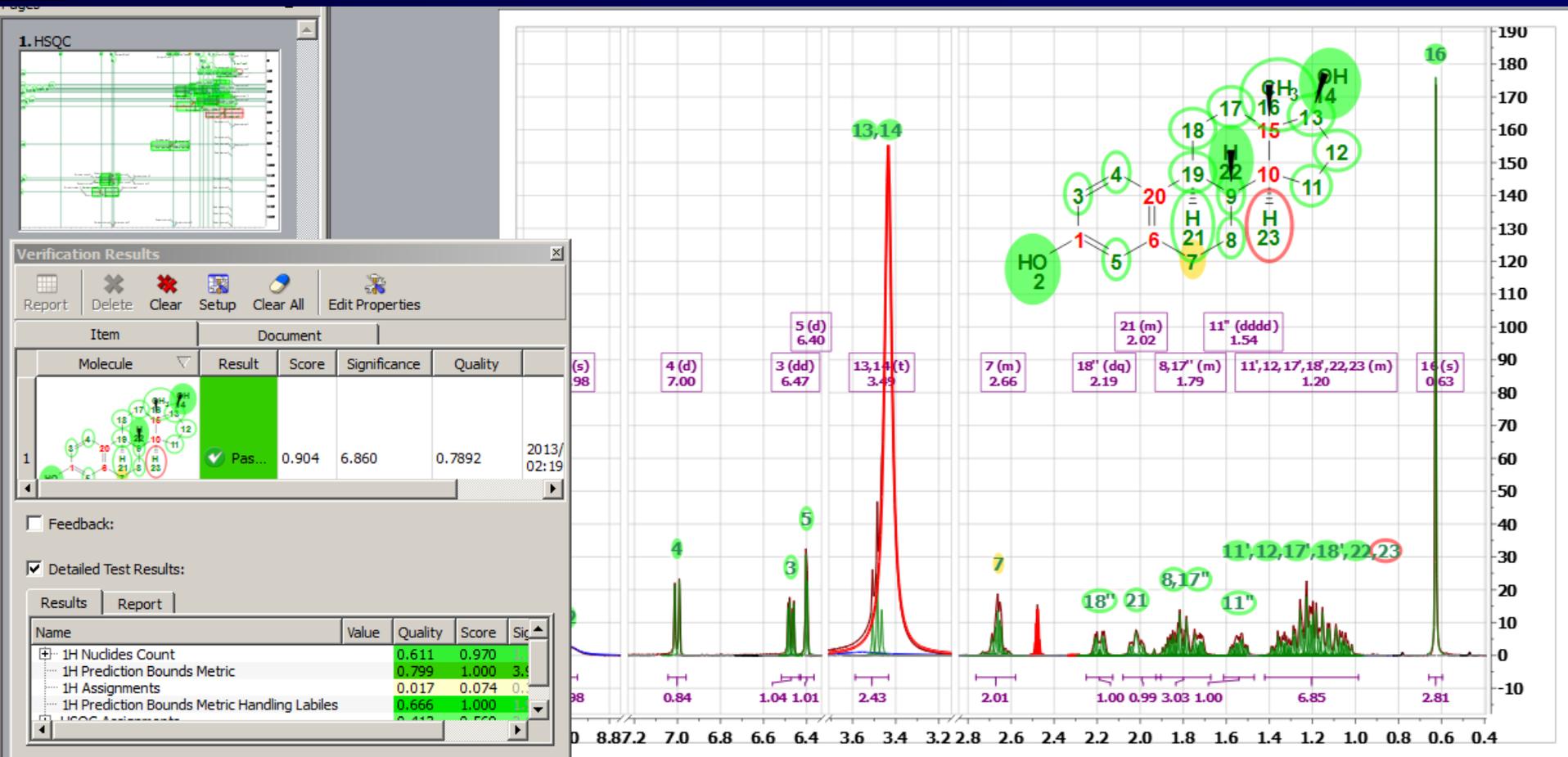
# Estradiol 400 MHz, using 1H only



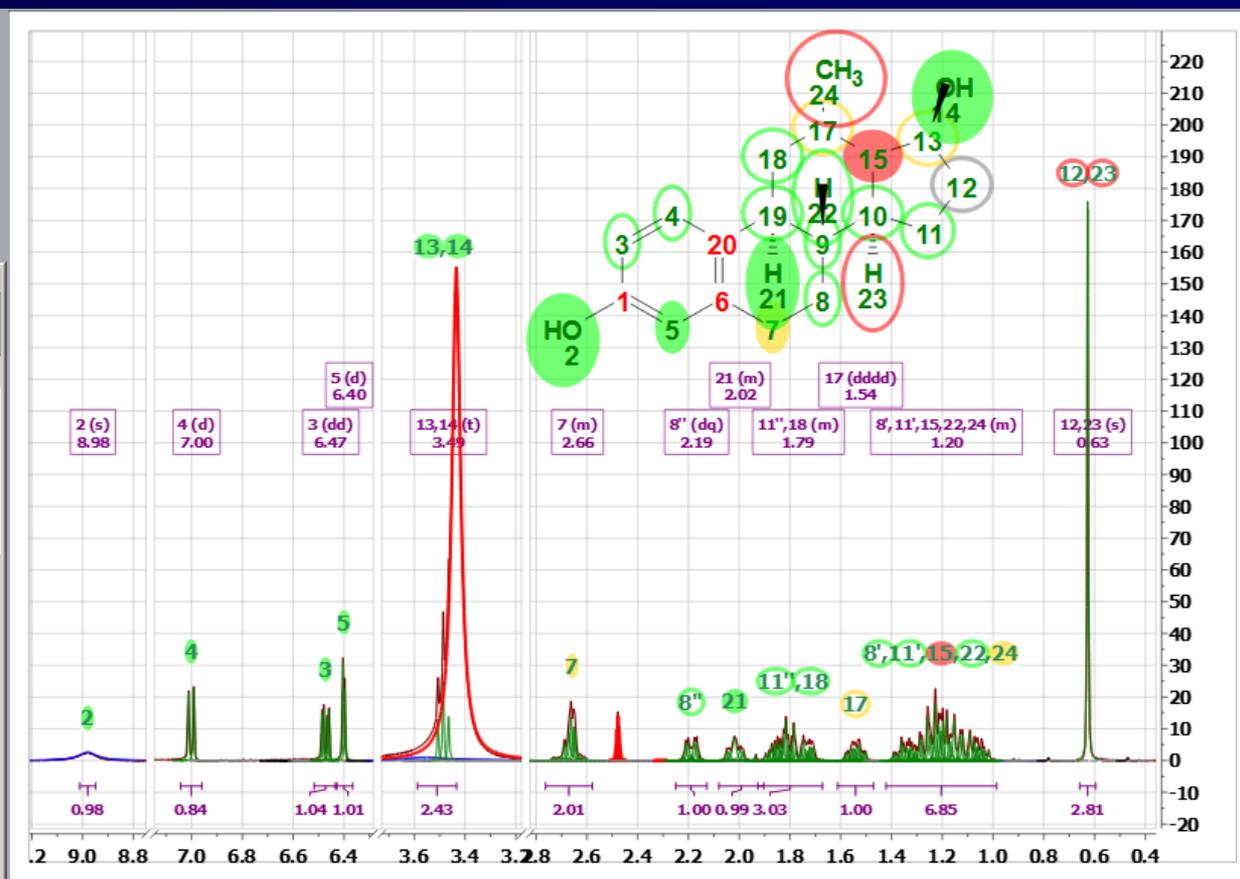
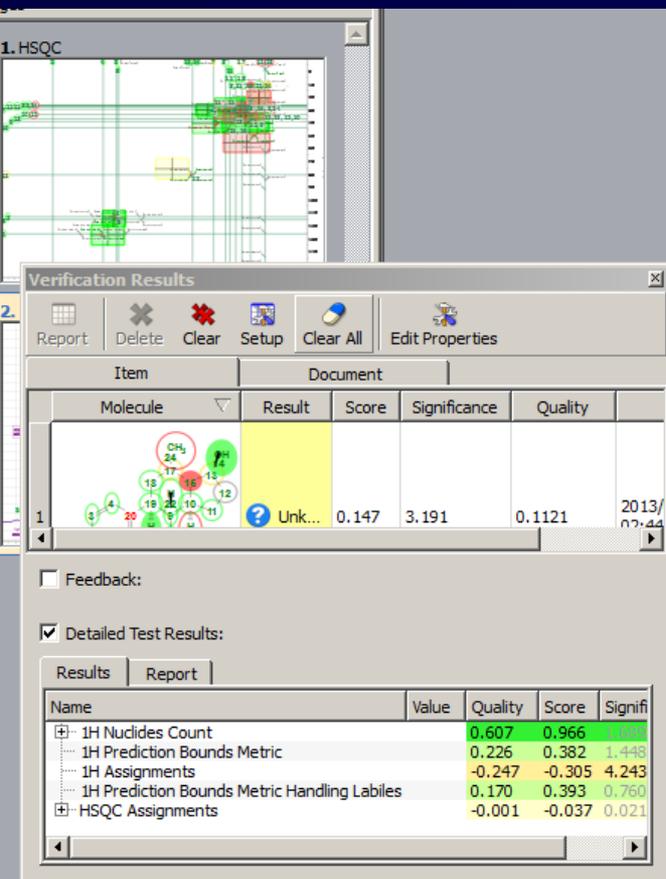
# Estradiol, modified mol, 400 MHz, using 1H only



# Estradiol 400 MHz, using 1H and HSQC



# 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<sup>3</sup>,  
the *Testing and Tuning Team*

+ MODGRAPH (the «predictors»)

# Thank You for your Attention

All slides will appear on [ebyte.it](http://ebyte.it) under

DOI: [10.3247/SL4Nmr13.007](https://doi.org/10.3247/SL4Nmr13.007)