Recent Advances in ASV: from Math to NMR

Stan Sykora, Carlos Cobas, Felipe Seoane, et al

Prior steps

Current status

Work – in - progress divide

Future steps

Towards ASE

NMR wizard

Legend:

ASV: Automatic Structure Verification

ASE: Automatic Spectrum Elucidation

MESTRELAB RESEARCH

Chemistry Software Solutions

www.mestrelab.com

www.ebyte.it

ASV: all the steps, up to the horizon …

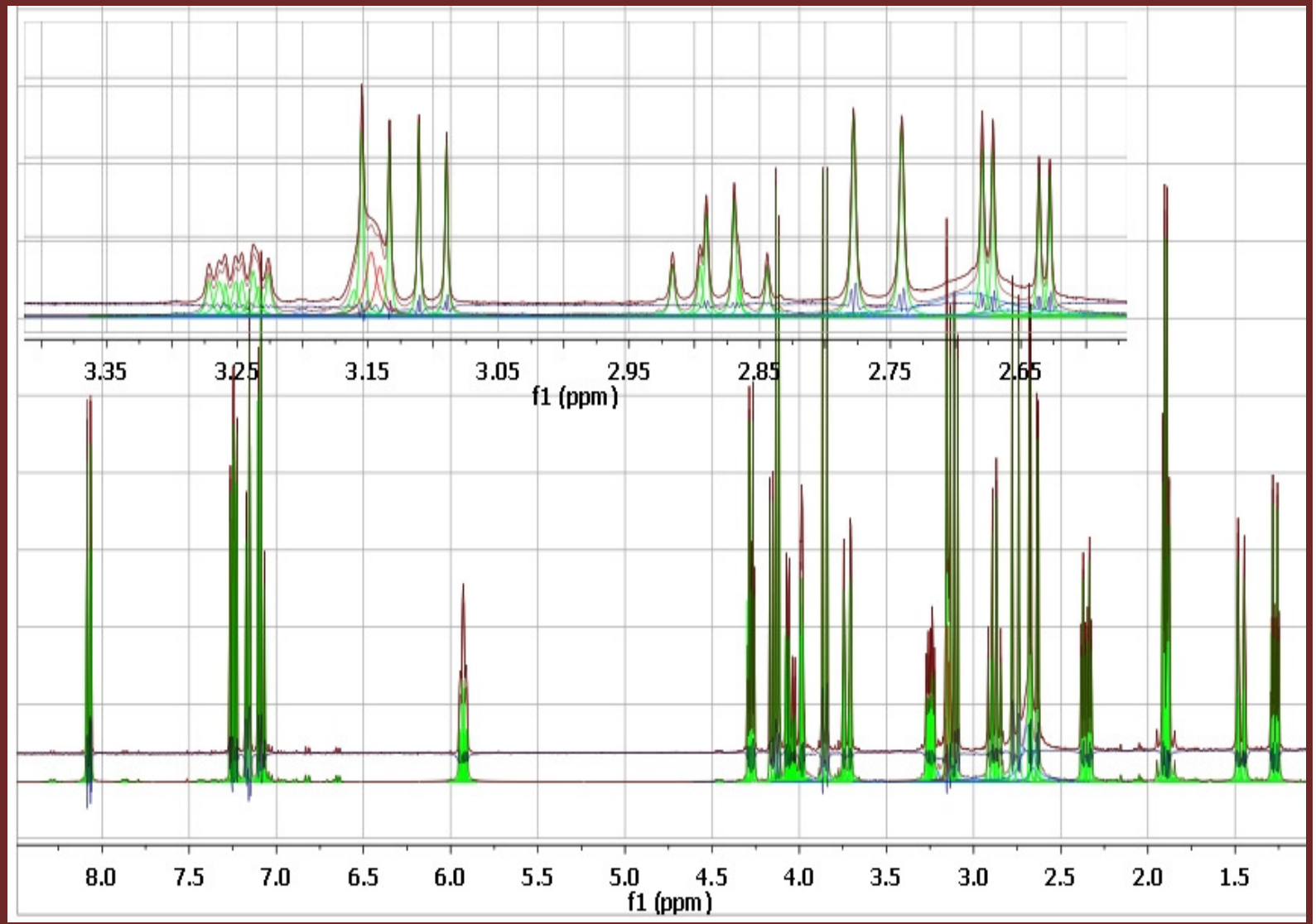
- ✓ GSD: Global Spectral Deconvolution
- ✓ Scoring systems: a new mathematical concept
- ✓ ASV structure in Mnova: Tasks & Tests
- ✓ Comparing spectra: NMR data elements of metric sets
- ✓ GSD peaks (auto)editing: the concept
- ✓ Solvent recognition: simple masking & AI approaches
- ✓ Labiles, the pesky outcasts: 3 ways to handle them
- ✓ Multiplets: recognition & characterization
- ✓ Counting the nuclei I: global & regional

- ✓ Prediction errors: definition of prediction regions
- ✓ Counting the nuclei II: prediction regions
- ✓ Coupling patterns: using JC algorithm & predictions
- ✓ Assignments: enumeration and scoring
- ✓ etc …

MESTRELAB RESEARCH
Chemistry Software Solutions

GSD: a functional definition

A fully automatic de-convolution of a whole spectrum

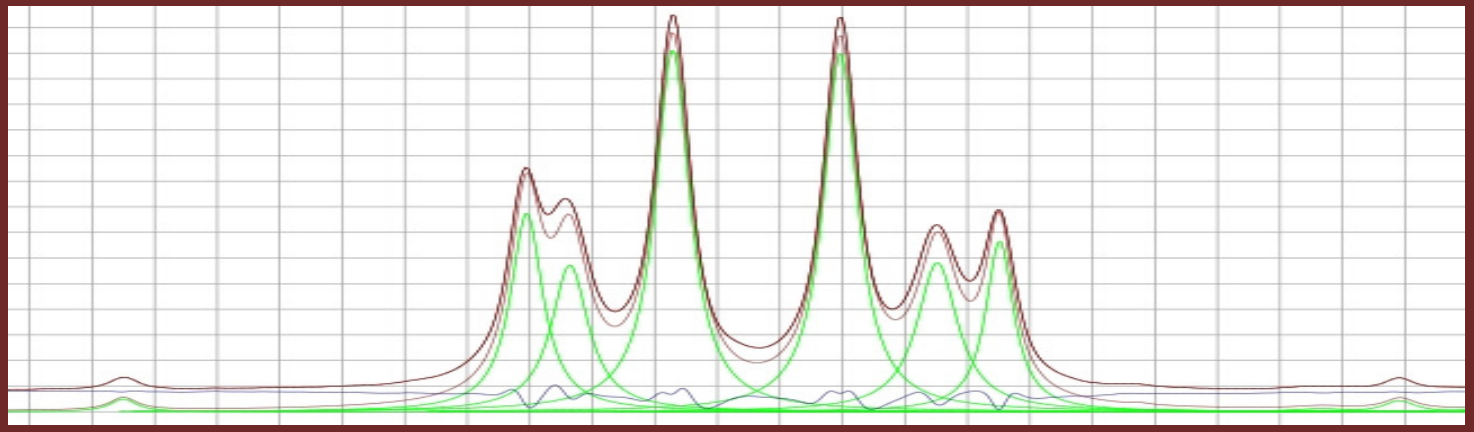


MESTRELAB RESEARCH
Chemistry Software Solutions

GSD history

- Introduction of the idea: SMASH 2007 (Sep, Metrelab User meeting)
- Presentation of first alpha results: 2008 (talks in Italy, UK, China)
- Declared to be fully operative: MMCE 2009 (Feb, a talk)
- Detailed presentation: 50th ENC in 2009 (Mar, poster)
- Significance of GSD for ASV: SMASH 2009 (Sep, User Meeting)
- First official release within Mnova: Autumn 2009
- First major revision: Jan-Feb 2009 (released in March)
- Applications to ASV: coming out now; ENC 2010 (Apr, User meeting)

- Applications to qNMR: in the works (Dr.Peng will present alpha results)
- Lineshape generalization: coming very soon
- …

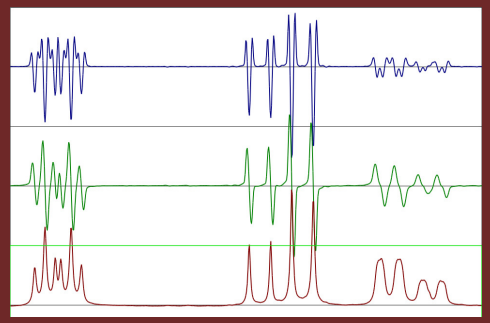


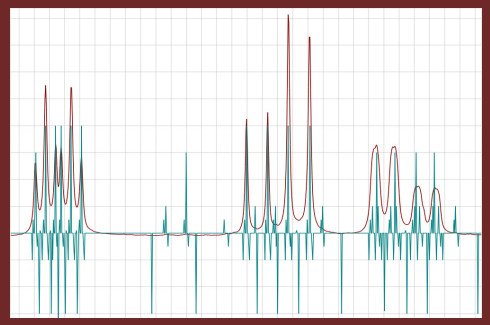
MESTRELAB RESEARCH
Chemistry Software Solutions

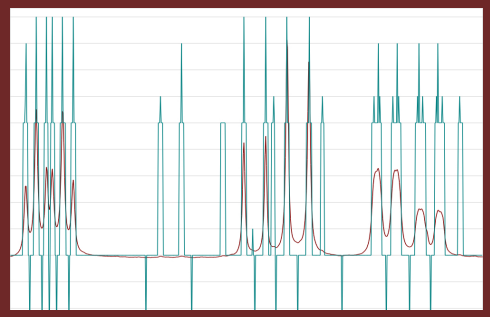
GSD algorithm

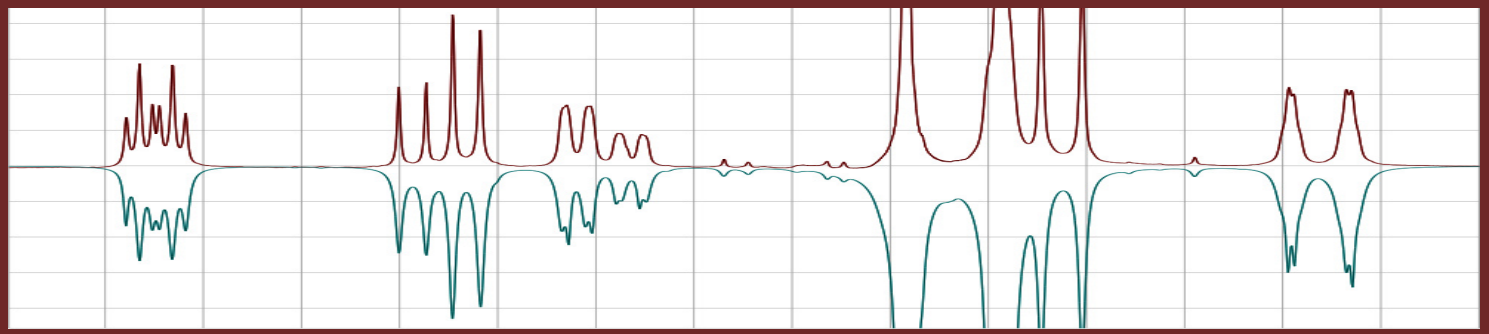
Four major innovative steps:

1. Derivatives (0th, 1st, 2nd)
2. Special points mark-up
3. Peaks ‘boxing’ (raw GSD)
4. Fast peak fitting









MESTRELAB RESEARCH
Chemistry Software Solutions

GSD output

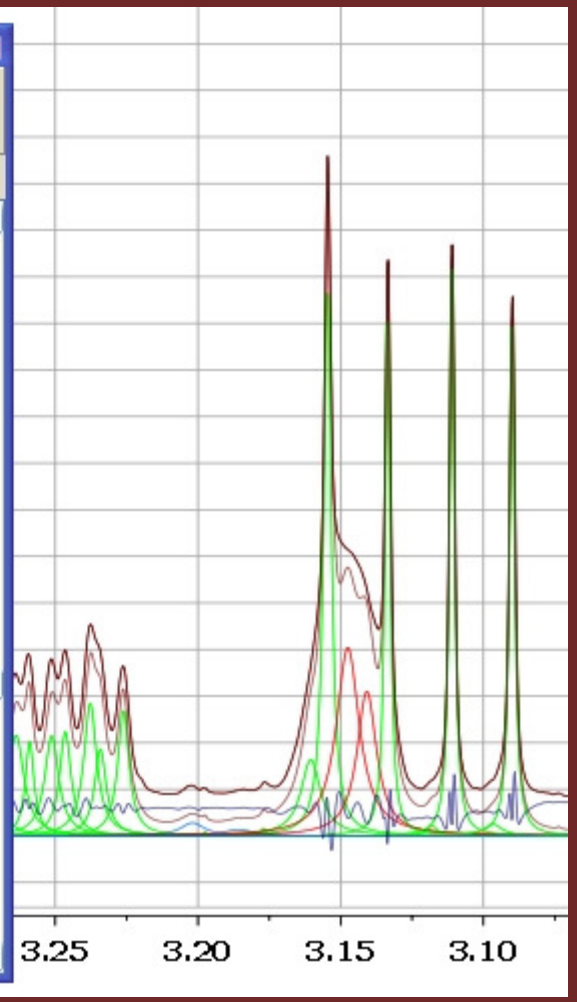
An editable Peaks List of all objectively detectable peaks to be used for any subsequent evaluation, including ASV

GSD Peaks

Report | Copy | Set Flags | Select Peaks | Reveal Peaks | Setup

Resolution: 1.00 [Normal]; Refinement: 3; AutoEdit: yes

	ppm	Width	Height	Area	J(CH)	Type	Flags
193	3.246	1.943	223.483	6997.780	-	Compound	None
194	3.238	2.180	284.980	10012.591	-	Compound	None
195	3.234	1.977	186.083	5927.819	-	Compound	None
196	3.226	1.899	268.806	8226.090	-	Compound	None
197	3.220	0.364	4.671	27.435	-	Compound	Weak
198	3.202	4.053	27.971	1827.027	-	Compound	Labile
199	3.197	0.528	8.930	75.960	-	Compound	Weak
200	3.186	5.302	13.343	1140.155	-	Compound	Weak + Labile
201	3.177	0.827	15.777	210.369	-	Compound	None
202	3.160	3.338	163.805	8810.688	-	Compound	None
203	3.154	1.285	1169.192	24216.081	-	Compound	None
204	3.148	3.914	404.380	25504.989	-	Solvent	Labile
205	3.141	3.319	310.539	16608.620	-	Solvent	None
206	3.133	0.877	1110.694	15702.071	-	Compound	None
207	3.118	1.677	21.319	576.206	-	Compound	None
208	3.111	0.836	1217.435	16408.012	-	Compound	None
209	3.097	3.071	24.940	1234.446	-	Compound	None



MESTRELAB RESEARCH
Chemistry Software Solutions

Scoring system

A novel mathematical concept (as well as a software class) devised to take decisions based on a number of tests, each having its own intrinsic significance

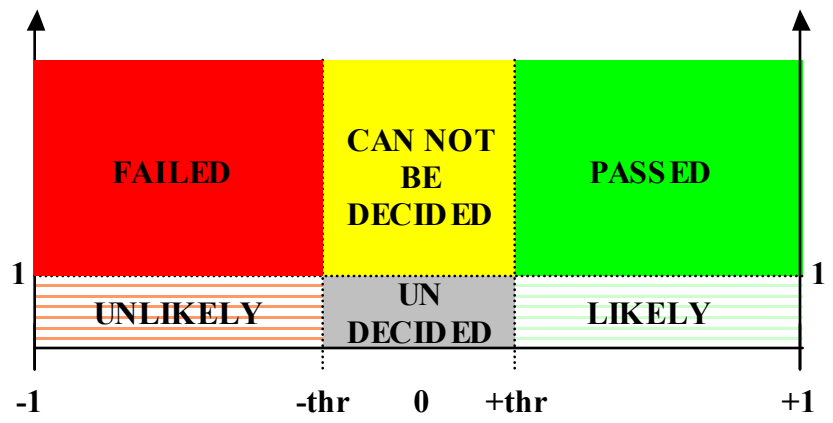
A situation to which it might apply:

- ✓ A committee of 20 votes on a querry.
- ✓ Each member casts a 'vote' or **score** ranging from –1 (false) to +1 (true).
- ✓ Each member, based on his prior record and independently of his vote, is assigned an intrinsic 'reliability' or **significance** (0 to 10).
- ✓ The final decision takes into account ALL the (score, significance) pairs.
- ✓ The final result is again a pair (score, significance).
- ✓ The way the pairs are combined is subject to a set of carefully modeled mathematical constraints (scoring system axioms).

Note: Significance can depend in a predefined way upon the sign of the score (some experts may be better on false scores, others on true scores)

MESTRELAB RESEARCH
Chemistry Software Solutions

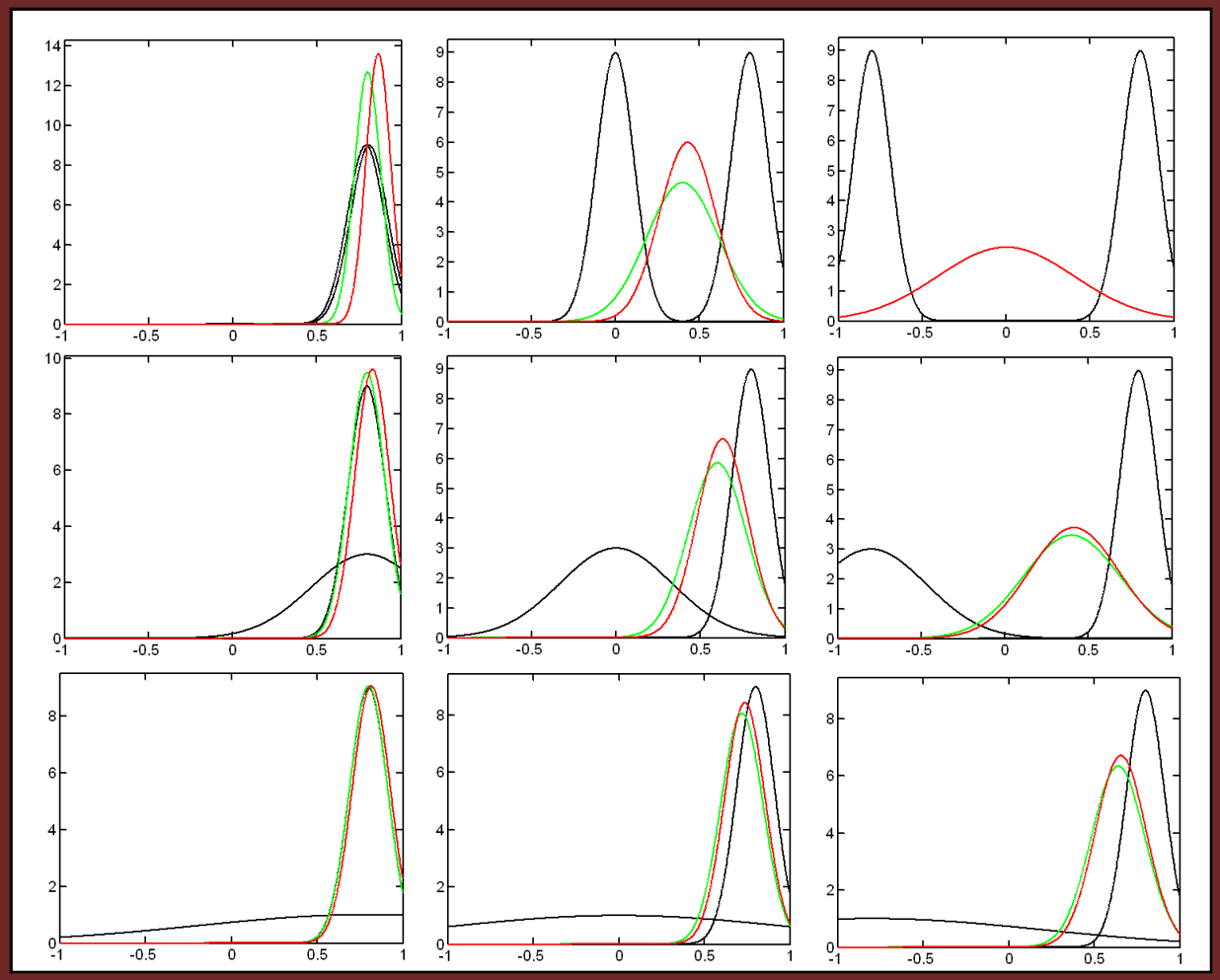
Some scoring system notes



Horizontal axis:	Score value (-1 to +1)
Vertical axis:	Significance (0 to INF, 1 unreliable, 10 expert)
PASSED and FAILED:	Areas of **high-quality** decisions
CAN NOT BE DECIDED:	All tests indicate a **decision is impossible**
UNDECIDED:	Tests give **contradictory results**
LIKELY:	PASSED but with very low significance
UNLIKELY:	FAILED but with very low significance

MESTRELAB RESEARCH
Chemistry Software Solutions

Scoring system: Matlab tests of compliance with *scoring system axioms* (a few examples)



… This is pure applied math; no NMR at all !

MESTRELAB RESEARCH
Chemistry Software Solutions

ASV structure

Technically, NMR ASV is a software structure which is embodies:

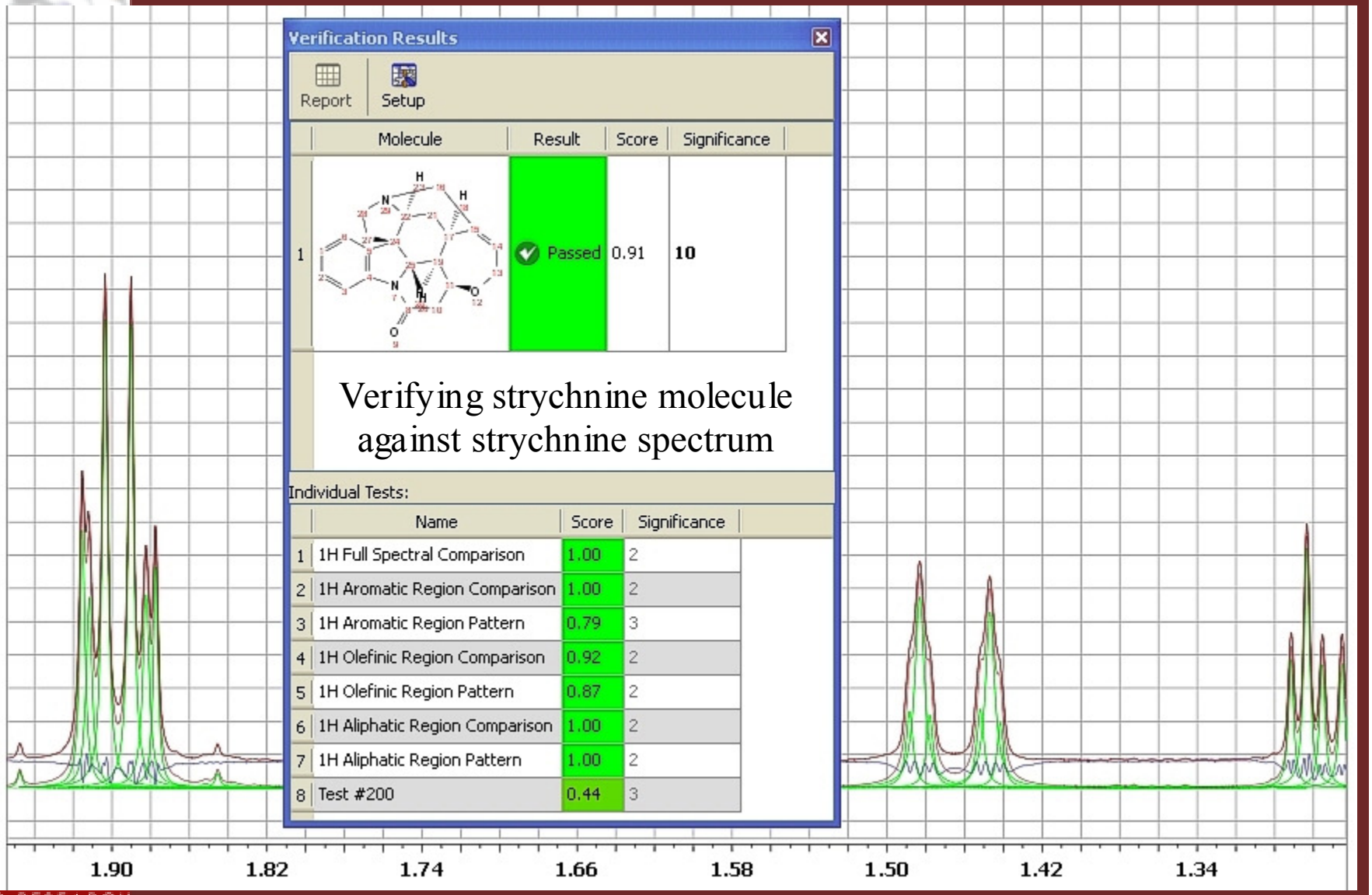
1) A scoring system which is its mathematical basis
2) A pool of tests (experts, voters)
3) A set of tasks (committees) each specialized for some purpose. Each task is composed of a number of tests, each with its own positive and negative significance (tunable parameters)

At present, just one task is implemented: a generic ASV wizard using 8 distinct tests. We plan to support a number of tasks drawing on a large pool of tests (hundreds?).

MESTRELAB RESEARCH
Chemistry Software Solutions

ASV example

(this is an anticipation, just to give you a feeling)



MESTRELAB RESEARCH

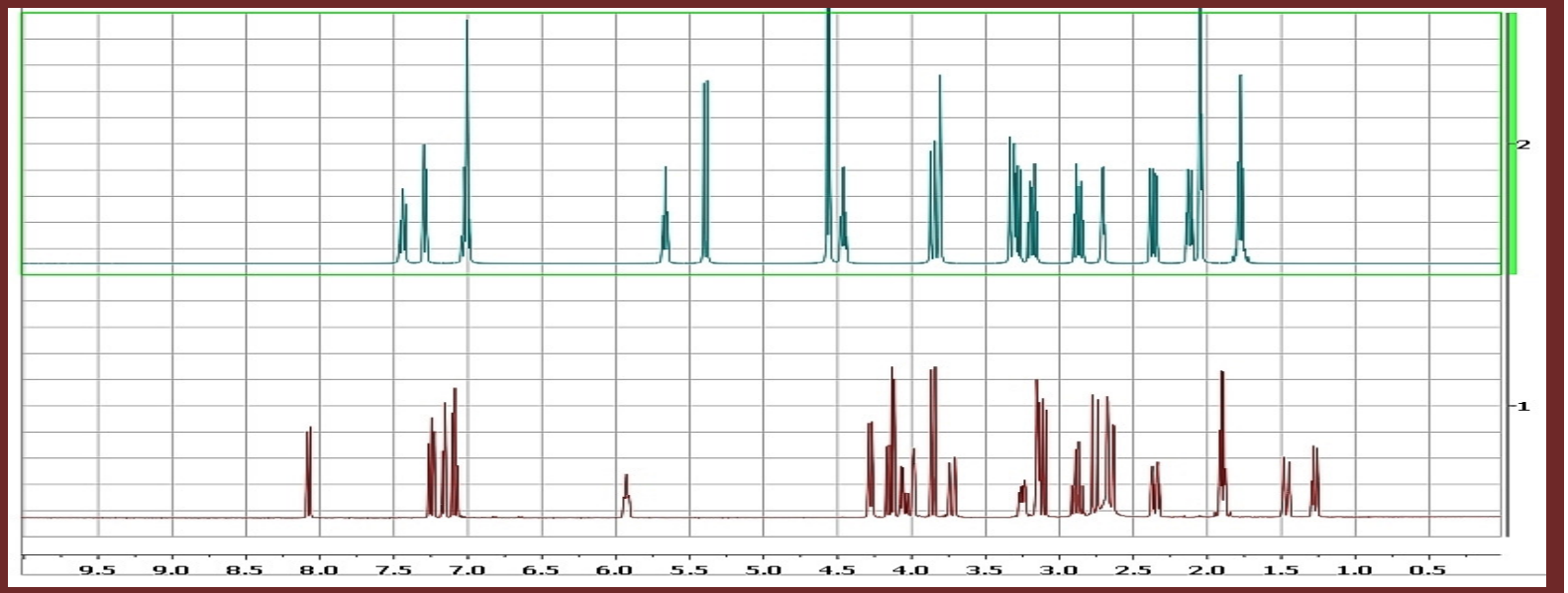
Chemistry Software Solutions

Comparing spectra

Another math concept which is very useful as a tool is that of a distance between spectra

This is because we want to use black-box predictions and just compare the experimental spectrum with a predicted one

But the concept of a distance could come handy also in other situations (for example finding similar spectra in a data base)



MESTRELAB RESEARCH
Chemistry Software Solutions

Spectral metric

There were prior proposals of distance-like functionals on pairs of spectra (Bodis, Ross, Pretsch), but they are lacking in some desirable aspects (irregular behavior upon sharp peak overlaps, excessive sensitivity to lineshape, etc.)

We have found a real-valued functional on a pair of spectra which has all the mathematical properties of a metric, avoids the drawbacks of the BRP distance, and is algorithmically compatible with GSD (can be computed directly from the two peak tables).

Distance $d(S_1,S_2)$:

- ✓ Is always non-negative: $d(S_1,S_2) \geq 0$
- ✓ Is 0 if and only if $S_1 = S_2$ $d(S_1,S_2) \Longleftrightarrow S_1 \equiv S_2$
- ✓ Is symmetric $d(S_1,S_2) = d(S_2,S_1)$
- ✓ Satisfies triangular inequality $d(S_1,S_3) \leq d(S_1,S_2) + d(S_2,S_3)$

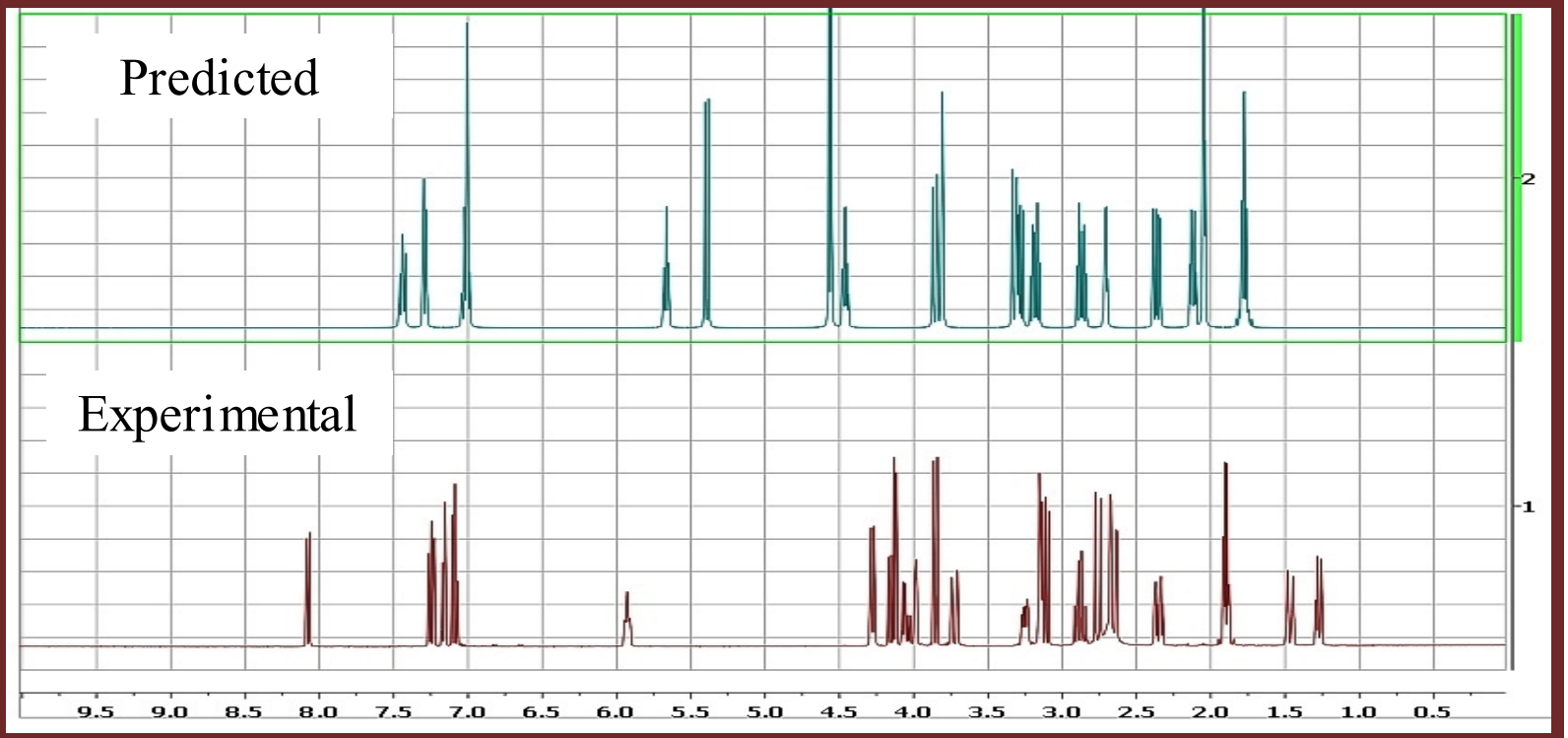
MESTRELAB RESEARCH
Chemistry Software Solutions

Metric tests

d(experimental,predicted) > upper_threshold	**FAILED**
d(experimental,predicted) < lower_threshold	**PASSED**
Between the thresholds	UNDECIDED

One can do it on the whole spectrum or separately on the aliphatic (-0.5 – 2.5), olefinic (2.5 – 5.5) or aromatic (5.5 – 12) regions using two different modes of normalization (one accentuating local quantity, the other local structure).

Hence our first 1+2*3 = 7 tests!

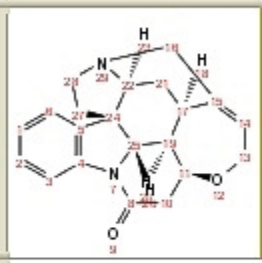


MESTRELAB RESEARCH
Chemistry Software Solutions

Testing the tests

Verification Results

Report | Setup

	Molecule	Result	Score	Significance
1		Unknown	0.18	9

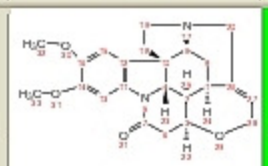
Verifying strychnine molecu
against brucine spectrum

Individual Tests:

	Name	Score	Significance
1	1H Full Spectral Comparison	0.69	3
2	1H Aromatic Region Comparison	0.37	4
3	1H Aromatic Region Pattern	-0.59	7
4	1H Olefinic Region Comparison	0.33	4
5	1H Olefinic Region Pattern	0.20	4
6	1H Aliphatic Region Comparison	0.52	3
7	1H Aliphatic Region Pattern	0.31	4
8	Test #200	0.31	3

Verification Results

Report | Setup

	Molecule	Result	Score	Signific
1		Passed	0.96	10

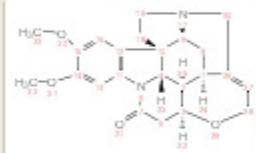
Verifying brucine molecule
against brucine spectrum

Individual Tests:

	Name	Score	Significance
1	1H Full Spectral Comparison	1.00	2
2	1H Aromatic Region Comparison	1.00	2
3	1H Aromatic Region Pattern	0.76	3
4	1H Olefinic Region Comparison	1.00	2
5	1H Olefinic Region Pattern	1.00	2
6	1H Aliphatic Region Comparison	1.00	2
7	1H Aliphatic Region Pattern	1.00	2
8	Test #200	0.78	2

Verification Results

Report | Setup

	Molecule	Result	Score	Significance
1		Unknown	0.29	7

Verifying brucine molecule
against strychnine spectrum

Individual Tests:

	Name	Score	Significance
1	1H Full Spectral Comparison	1.00	2
2	1H Aromatic Region Comparison	1.00	2
3	1H Aromatic Region Pattern	-0.22	6
4	1H Olefinic Region Comparison	0.22	4
5	1H Olefinic Region Pattern	0.71	3
6	1H Aliphatic Region Comparison	1.00	2
7	1H Aliphatic Region Pattern	0.75	3
8	Test #200	-0.59	4

MESTRELAB RESEARCH
Chemistry Software Solutions

GSD peaks (auto)editing

When comparing experimental and predicted spectra, there are some aspects which need to be addressed before ANY tests can be applied.

Predicted spectra, for example, do not contain any **solvent peaks** while experimental spectra contain them in unpredictable amounts. The same applies to the **reference peaks** (such as those of TMS)

Solvent and reference peaks must be located and labeled in the GSD Peaks List of the experimental spectrum. Once labeled, they can be ignored by all subsequent processing algorithms.

Other types of peaks whose recognition and labeling is desirable, such as ^{13}C satellites, **impurities**, **rotational sidebands**, etc.

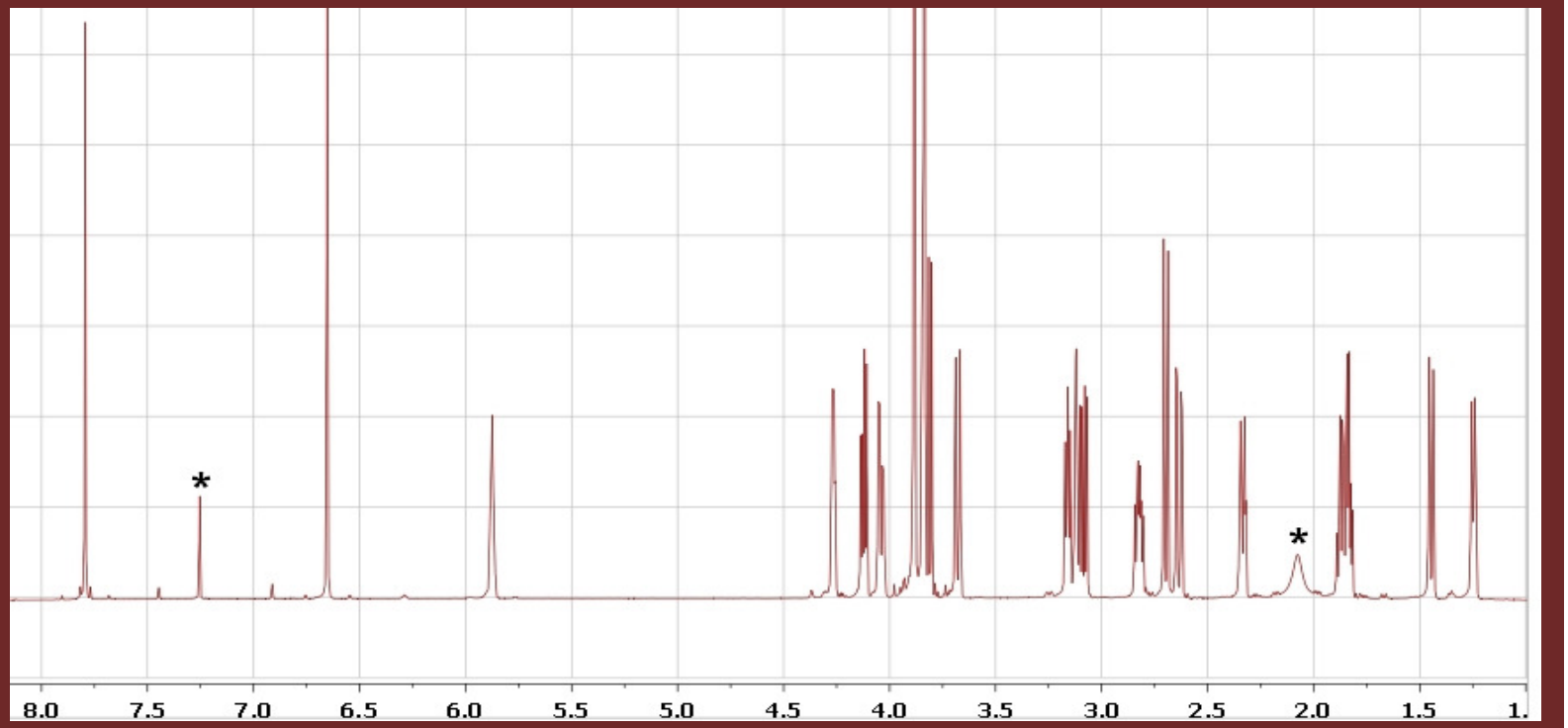
This often tricky peaks editing process can be done both **automatically and/or manually**.

MESTRELAB RESEARCH
Chemistry Software Solutions

Solvent recognition: basics

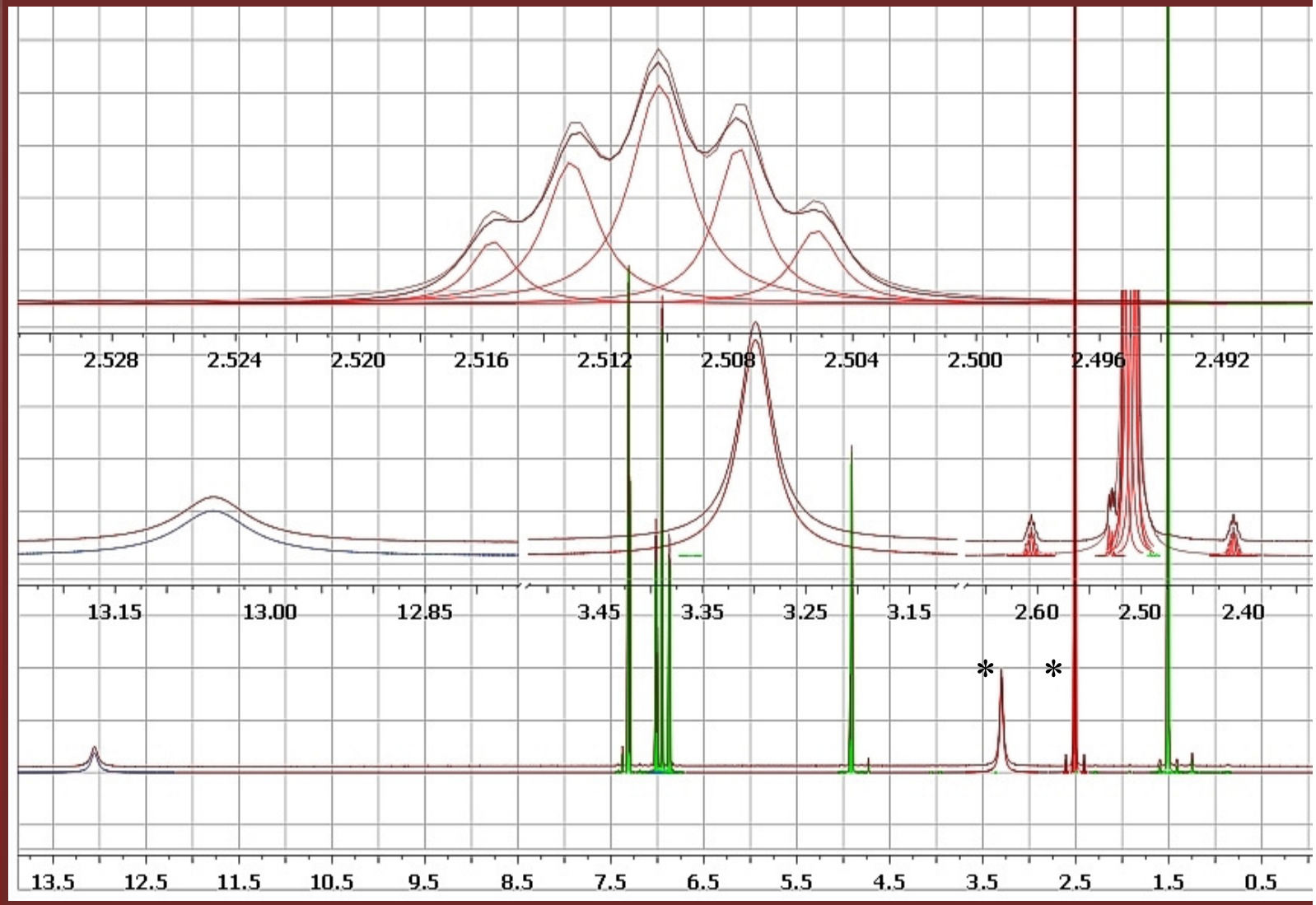
One way to handle the solvent is zone masking: one labels every peak in more or less ample region around each expected solvent peak. This is quite drastic and often discards a lot of useful info!

A better way is to apply the same know-how an experienced chemist uses (often instinctively) to selectively pick out the solvent peaks even when they are in a crowded region. This means a kind of AI wizard or, more simply, just a pretty clever software.



MESTRELAB RESEARCH
Chemistry Software Solutions

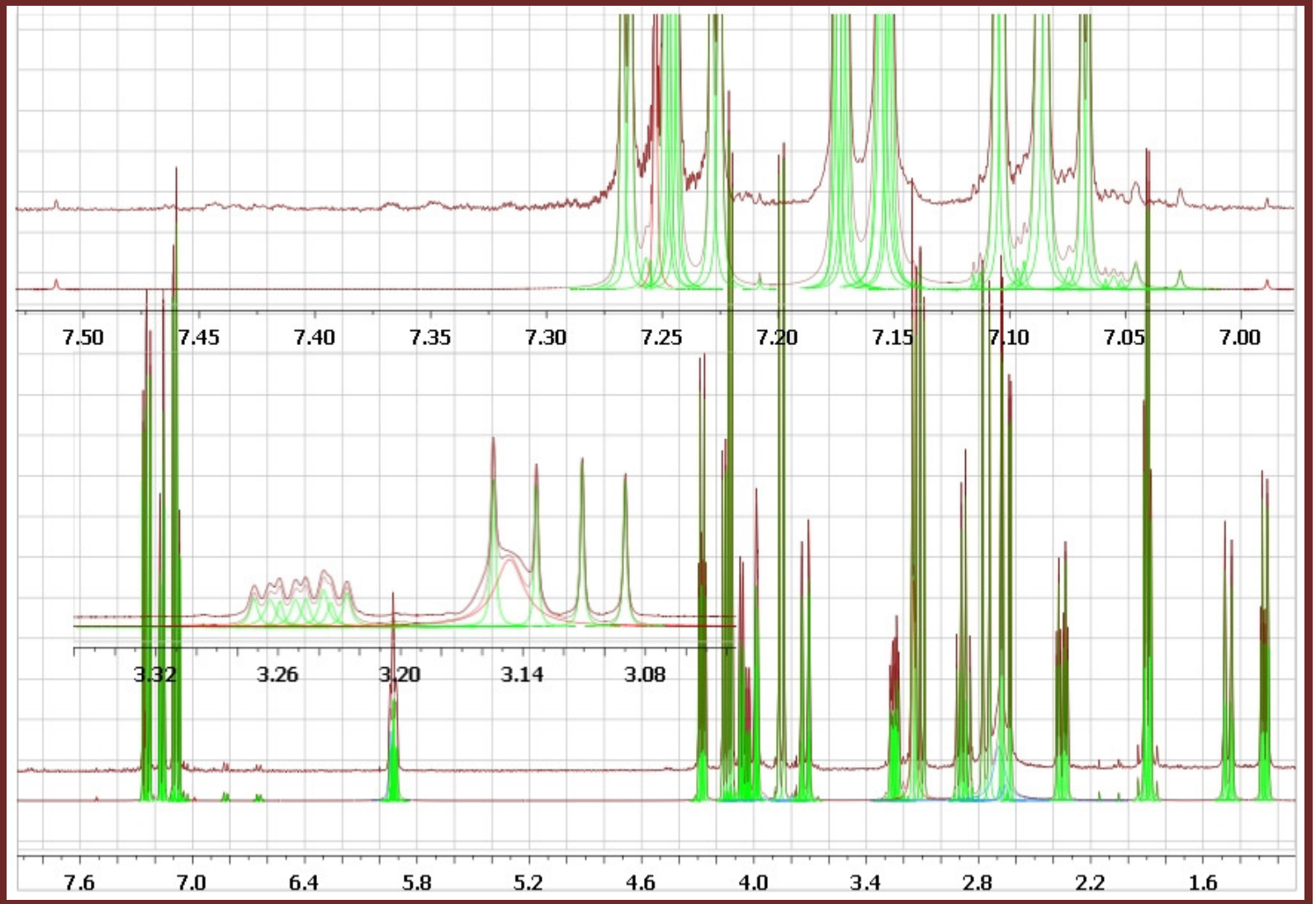
Solvent recognition:
masking versus AI wizard



MESTRELAB RESEARCH
Chemistry Software Solutions

Solvent recognition AI:

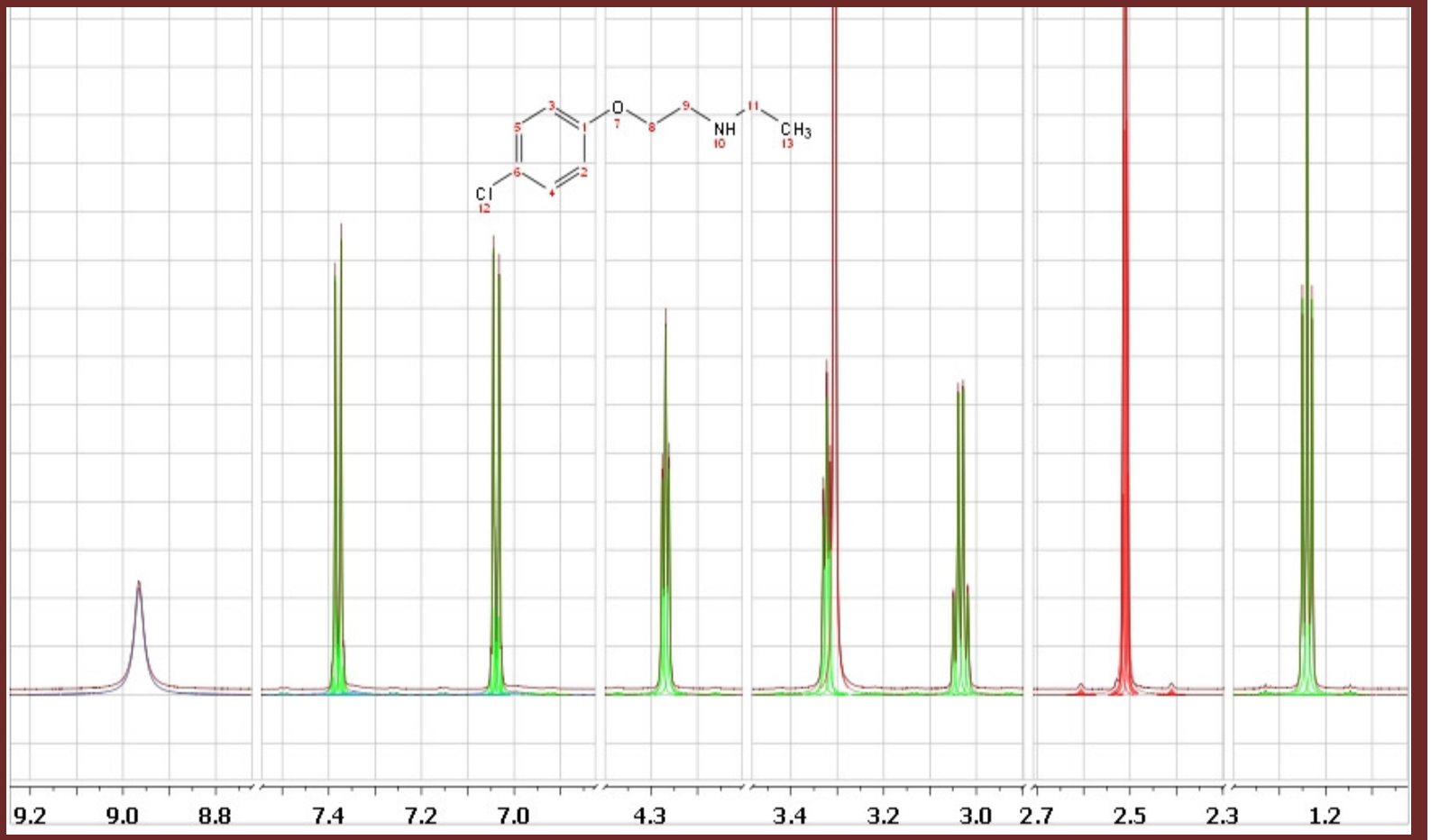
let there be a *scoring system* for every line!



MESTRELAB RESEARCH
Chemistry Software Solutions

Labile protons peaks

This is much the same story,
but a more difficult one since there is much less to build upon



MESTRELAB RESEARCH
Chemistry Software Solutions

Recognition of other special peaks

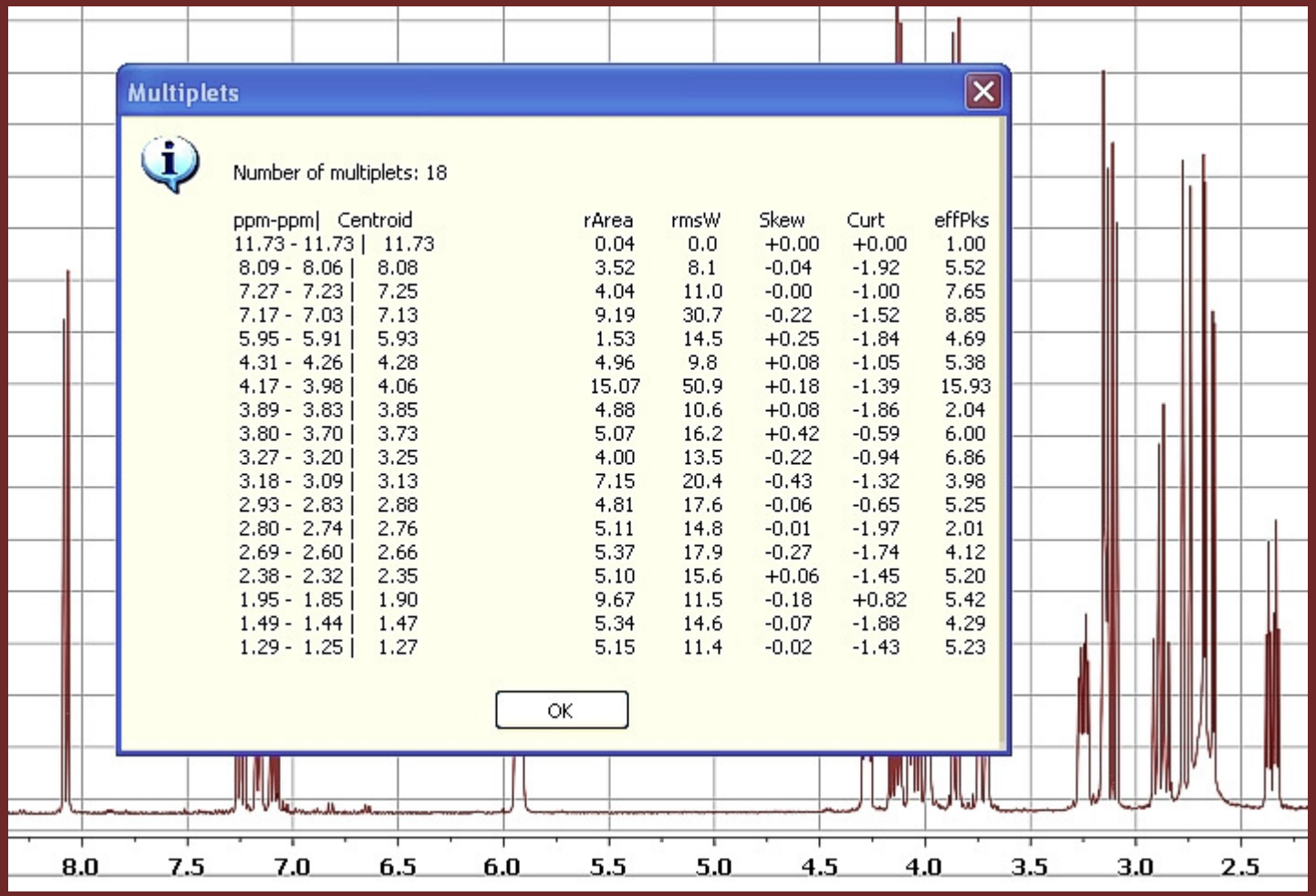
- ✓ ^{13}C satellites
- ✓ Rotation sidebands
- ✓ Impurities

This is work-in-progress involving extensive application of scoring systems associated with individul spectral peaks

MESTRELAB RESEARCH
Chemistry Software Solutions

Multiplets recognition

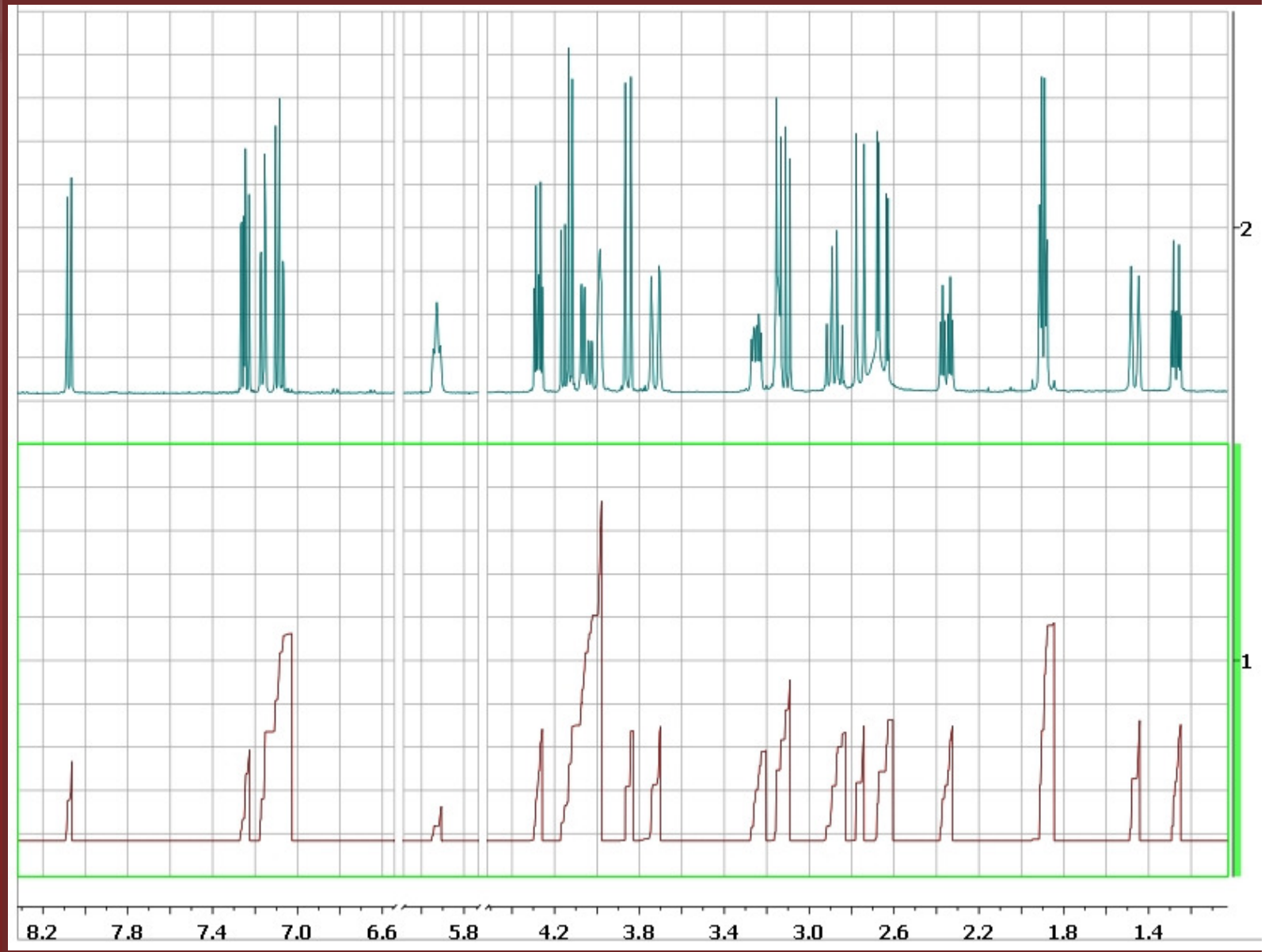
Work - in - progress



ppm-ppm	Centroid	rArea	rmsW	Skew	Curt	effPks
11.73 - 11.73	11.73	0.04	0.0	+0.00	+0.00	1.00
8.09 - 8.06	8.08	3.52	8.1	-0.04	-1.92	5.52
7.27 - 7.23	7.25	4.04	11.0	-0.00	-1.00	7.65
7.17 - 7.03	7.13	9.19	30.7	-0.22	-1.52	8.85
5.95 - 5.91	5.93	1.53	14.5	+0.25	-1.84	4.69
4.31 - 4.26	4.28	4.96	9.8	+0.08	-1.05	5.38
4.17 - 3.98	4.06	15.07	50.9	+0.18	-1.39	15.93
3.89 - 3.83	3.85	4.88	10.6	+0.08	-1.86	2.04
3.80 - 3.70	3.73	5.07	16.2	+0.42	-0.59	6.00
3.27 - 3.20	3.25	4.00	13.5	-0.22	-0.94	6.86
3.18 - 3.09	3.13	7.15	20.4	-0.43	-1.32	3.98
2.93 - 2.83	2.88	4.81	17.6	-0.06	-0.65	5.25
2.80 - 2.74	2.76	5.11	14.8	-0.01	-1.97	2.01
2.69 - 2.60	2.66	5.37	17.9	-0.27	-1.74	4.12
2.38 - 2.32	2.35	5.10	15.6	+0.06	-1.45	5.20
1.95 - 1.85	1.90	9.67	11.5	-0.18	+0.82	5.42
1.49 - 1.44	1.47	5.34	14.6	-0.07	-1.88	4.29
1.29 - 1.25	1.27	5.15	11.4	-0.02	-1.43	5.23

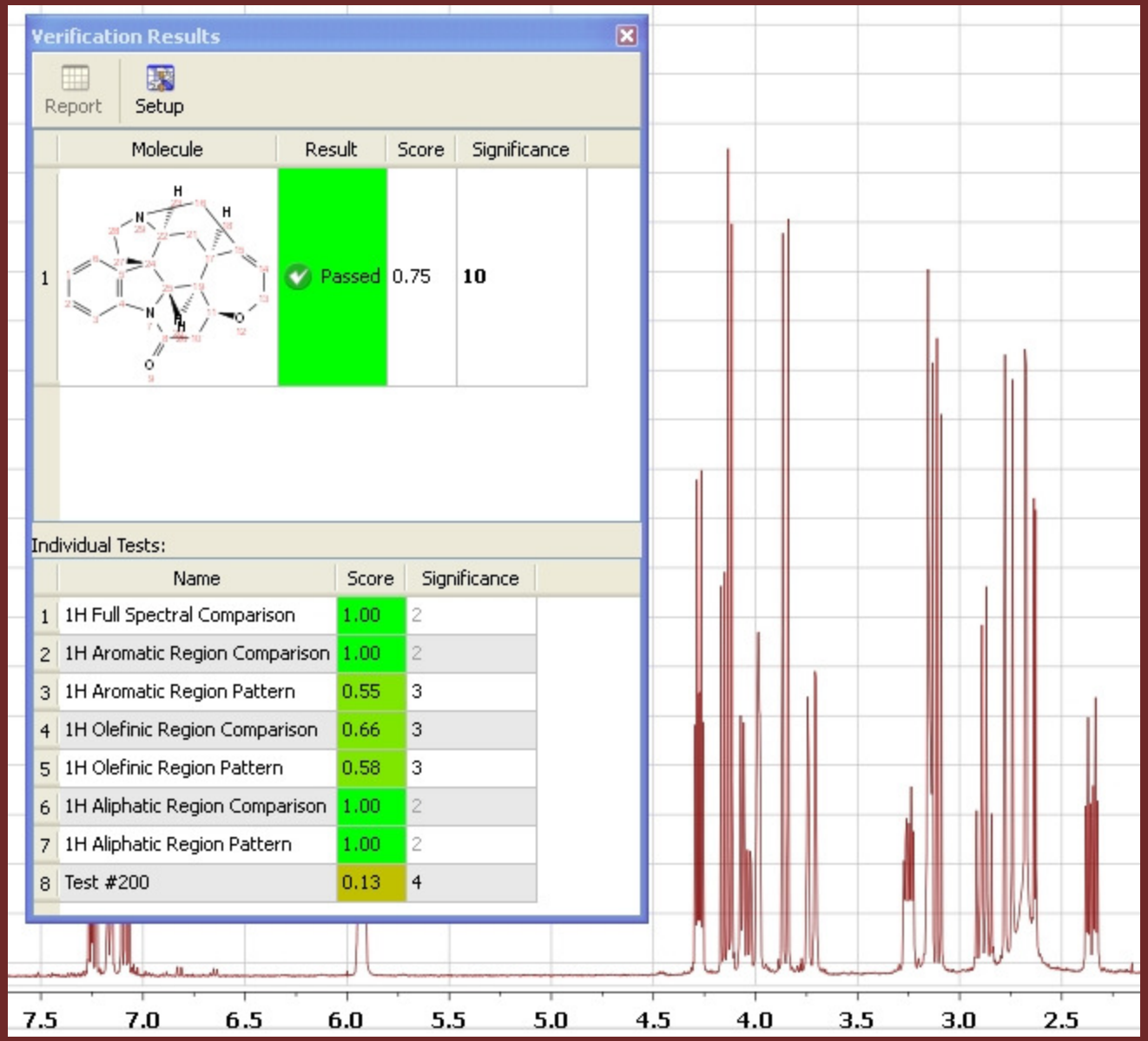
MESTRELAB RESEARCH
Chemistry Software Solutions

Scoring on number of nuclei



MESTRELAB RESEARCH
Chemistry Software Solutions

Scoring on number of nuclei



MESTRELAB RESEARCH
Chemistry Software Solutions

ASV: all the steps, up to the horizon …

- ✓ GSD: Global Spectral Deconvolution
- ✓ Scoring systems: a new mathematical concept
- ✓ ASV structure in Mnova: Tasks & Tests
- ✓ Comparing spectra: NMR data elements of metric sets
- ✓ GSD peaks (auto)editing: the concept
- ✓ Solvent recognition: simple masking & AI approaches
- ✓ Labiles, the pesky outcasts: 3 ways to handle them
- ✓ Multiplets: recognition & characterization
- ✓ Counting the nuclei (I): global & regional

- ✓ Prediction regions as defined by prediction error bounds
- ✓ Counting the nuclei (II) within prediction regions
- ✓ Coupling patterns: using JC algorithm & predictions
- ✓ Assignments: enumeration and scoring
- ✓ etc …

Now, just let the blue line (work-in-progress divide) move down !

MESTRELAB RESEARCH
Chemistry Software Solutions