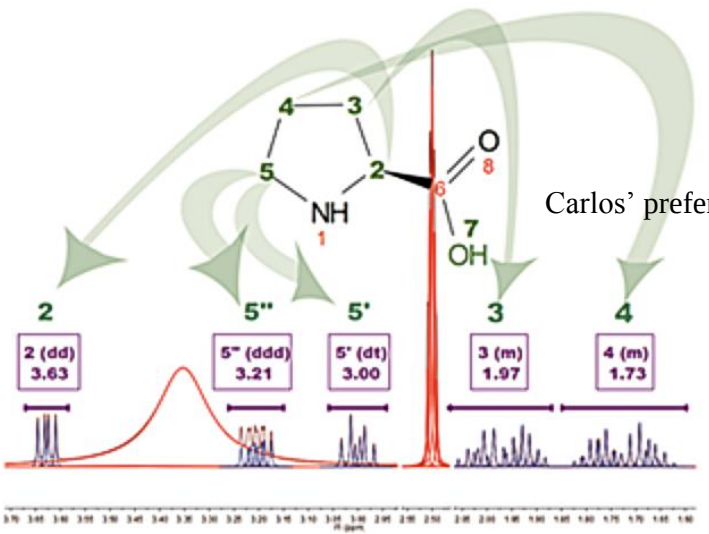
Mestrelab Research
chemistry software solutions

The latest in Automatic Structure Verification (ASV, Mnova Verify)

presented by Stan Sykora (Extra Byte) and by the
Mestrelab team (Carlos Cobas, Felipe Seoane, Ester Vaz, and many others)

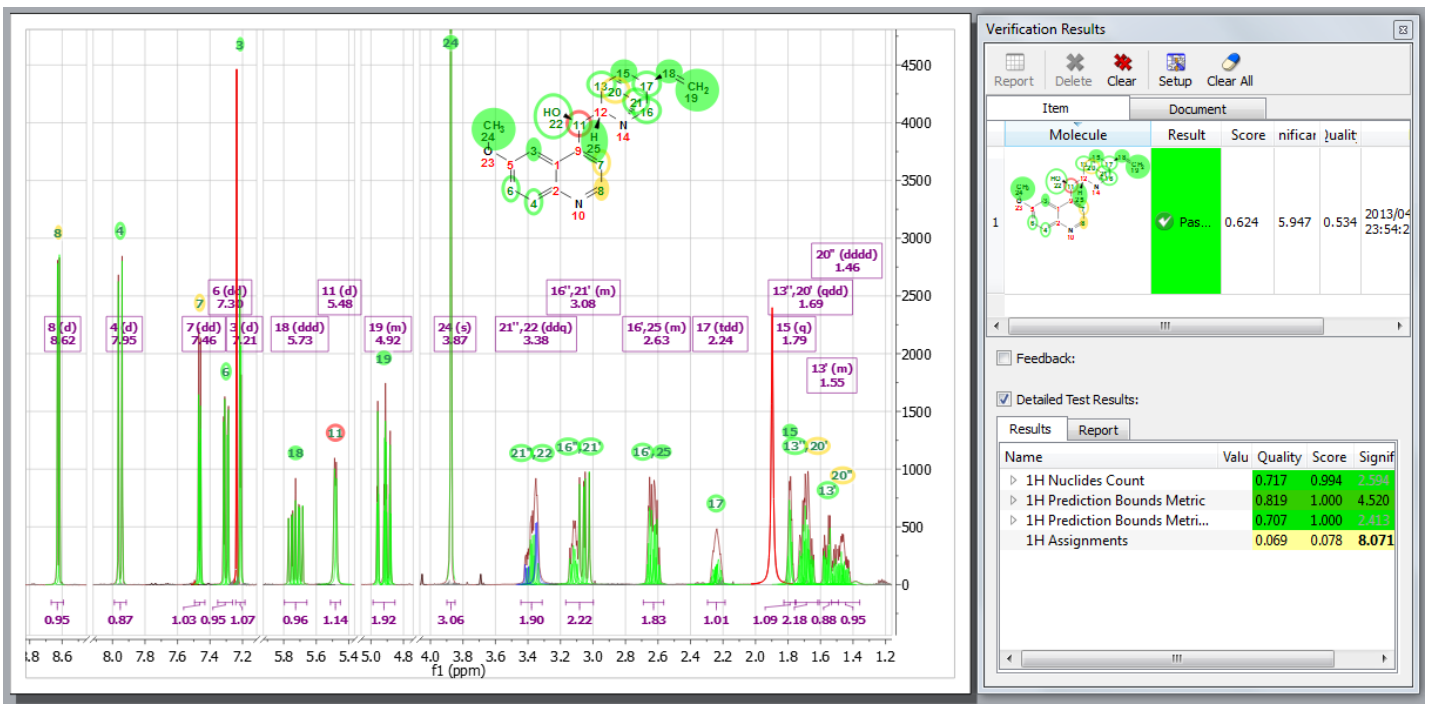


Carlos' preferred front picture

DOI of this online document: 10.3247/SL4Nmr13.003

Mestrelab Research
chemistry software solutions

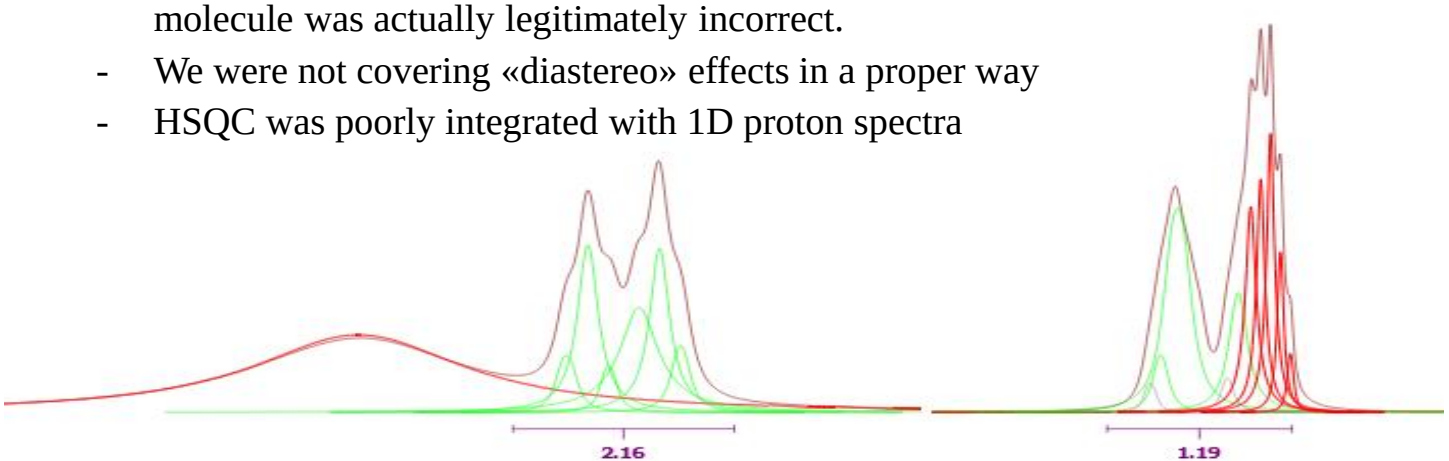
The picture Stan likes more (as opposed to Carlos)



Mestrelab Research
chemistry software solutions

One year ago

- We had a reasonably working ASV, provided the spectrum was nice (20% of cases) and the molecule was "reasonable" (20% as well)
- We had problems with: solvent water recognition, labiles recognition, labiles assignments, and minor problems with main solvent recognition. Though our solvent recognition is the best one available, it was not good enough.
- We had problems with generic proton assignments, especially when the molecule was actually legitimately incorrect.
- We were not covering «diastereo» effects in a proper way
- HSQC was poorly integrated with 1D proton spectra



Mestrelab Research
chemistry software solutions

Some explanations I: The ASV flowchart (a preamble)

1. **GSD**: Identify spectral peaks and list them in a numeric table
2. **Edit the peaks** and carry out **full and greedy analysis** of the data.
3. **Do it for all available data**, both individually and in combinations
4. **Use the results for various tasks**, such as
   - **AA** (Automatic Assignments)
   - **ASV** (Automatic structure verification)
   - ASE (Automatic structure elucidation)
   - ASD (Automatic structure discrimination)
   - ACD (Automatic components detection)
   - etc etc etc…

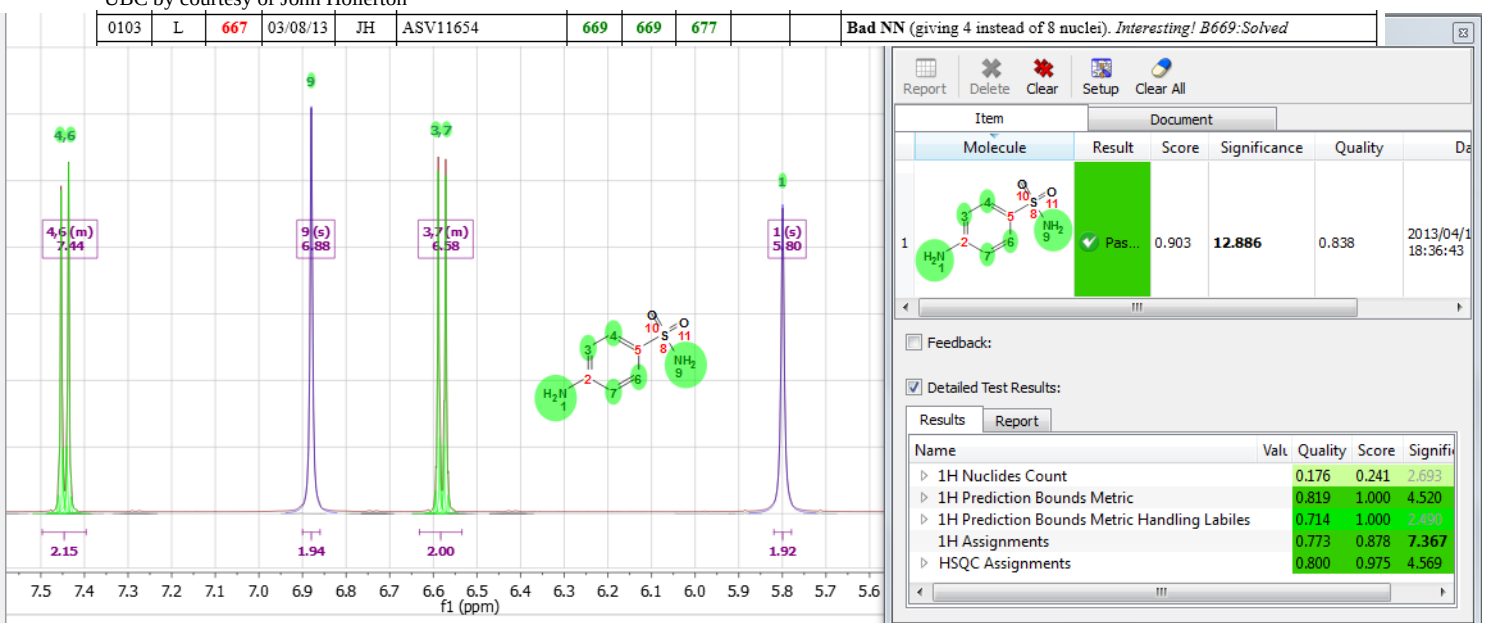
Mestrelab Research
chemistry software solutions

Some explanations II: Troublesome molecules

What does make a molecule «dificult»?

Too simple and too complex molecules are both a challenge!

UBC by courtesy of John Hollerton

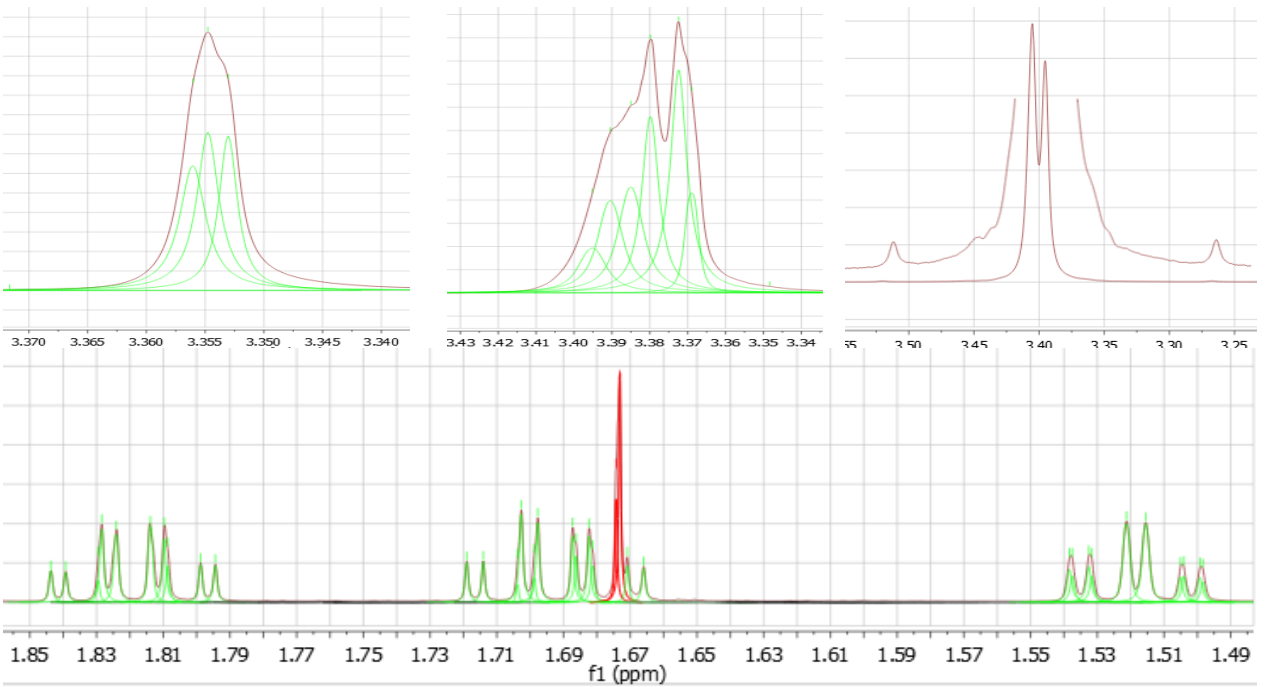


Mestrelab Research
chemistry software solutions

Some explanations III: Solvent blues

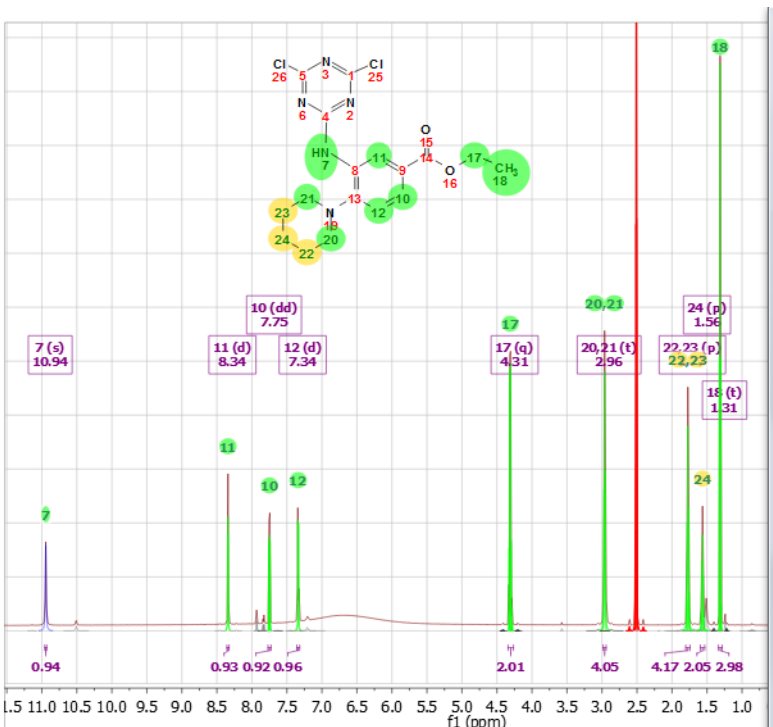
Solvent and solvent-water recognition:

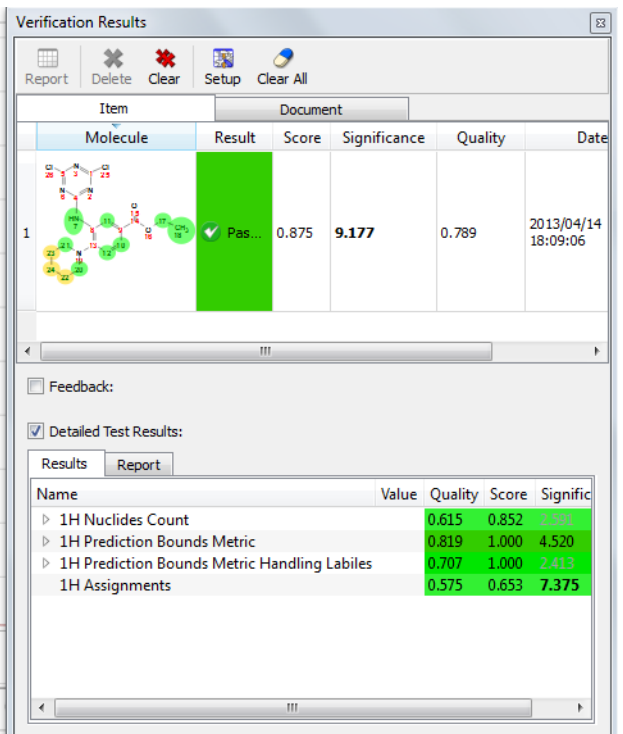
One of the most serious problems are the many «types» of solvent water peaks encountered in real spectra (frequent exceptions to all formal rules)



Mestrelab Research
chemistry software solutions

Some explanations III: more solvent blues





Mestrelab Research
chemistry software solutions

Some explanations IV: Labile blues

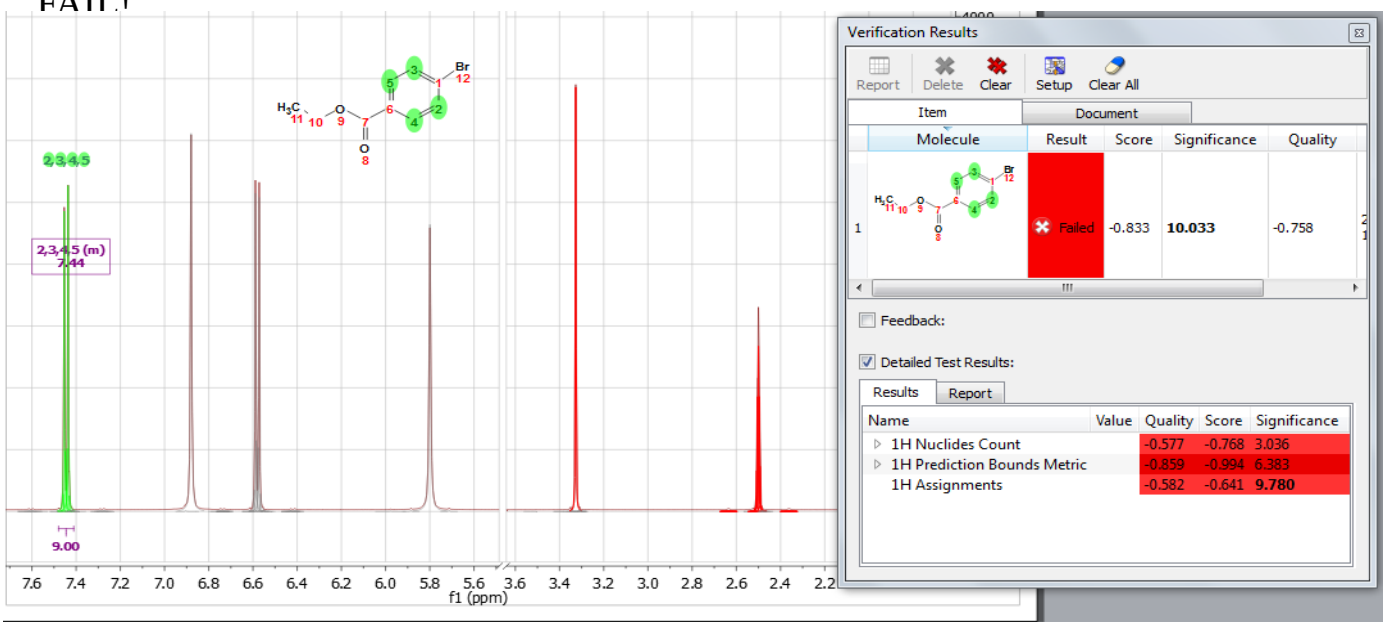
- Labiles shifts are very unpredictable, and so are their shapes.
- They also just love to overlap non-labile multiplets,
- or become too broad to be detected,
- or merge with water and be missing altogether.
- Sometimes they are coupled to other nuclei, but mostly not.
- Sometimes they are humpy (just like water).
- And always they are a headache.

Mestrelab Research
chemistry software solutions

Some explanations V: The lure of Assignments

What's wrong with wrong [elementary] assignments when the structure is anyway incorrect and the whole case is a clear FAIL? Nothing, right? Or NOT … ?!

Chemical spectroscopists think NOT! Their fascination with assignments indecent! Do not dare to assign the methyl to a peak at >5 ppm, even if ASV is the same FAIL!



Mestrelab Research
chemistry software solutions

Some explanations VI: ASV versus AutoAssignments

ASV and AA are two different worlds and two correlated, but not coincident, optimization problems.

Likewise, ASV and ASD are also two different tasks!

Consider a single-bump, low-resolution spectrun and ANY molecule:

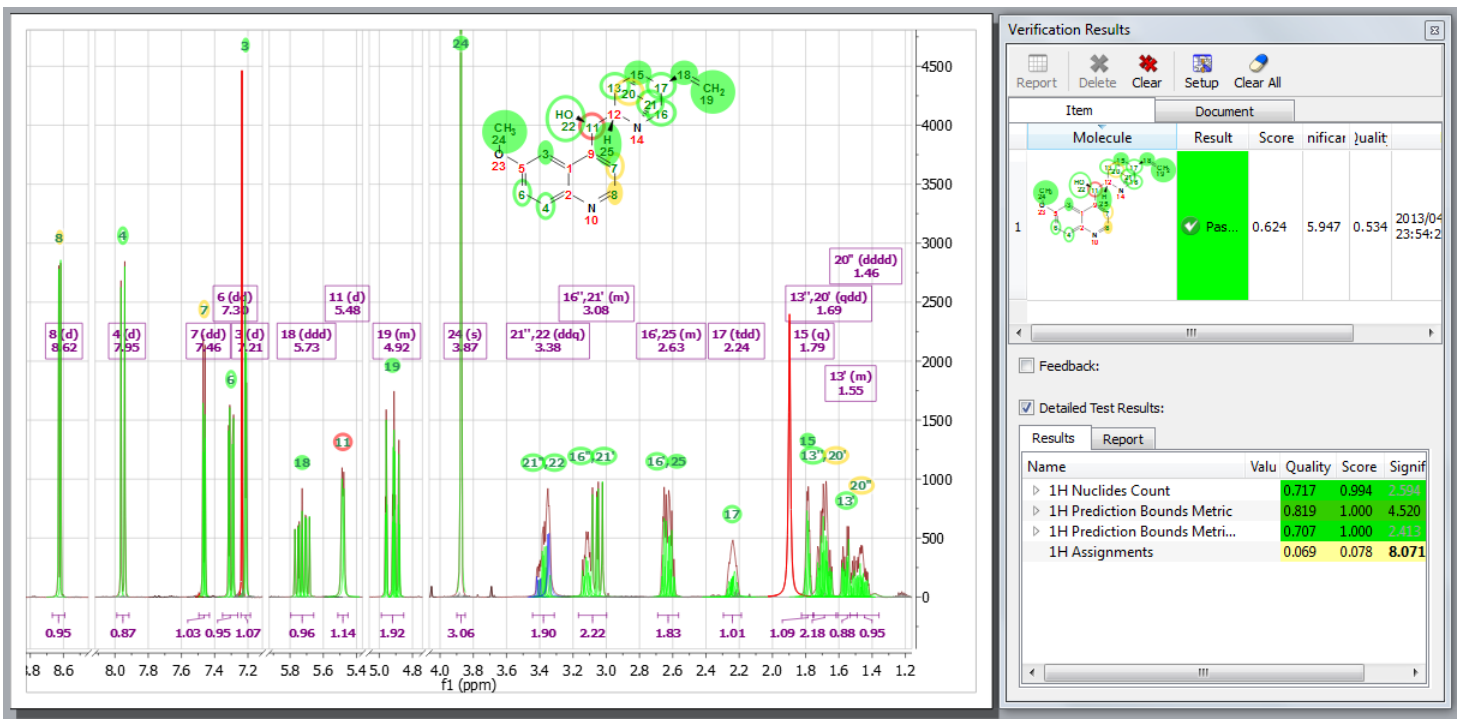
ASV is always a PASS (why), AA is also OK (why),

but both are useless

Mestrelab Research
chemistry software solutions

Some explanations VI: ASV versus AutoAssignments

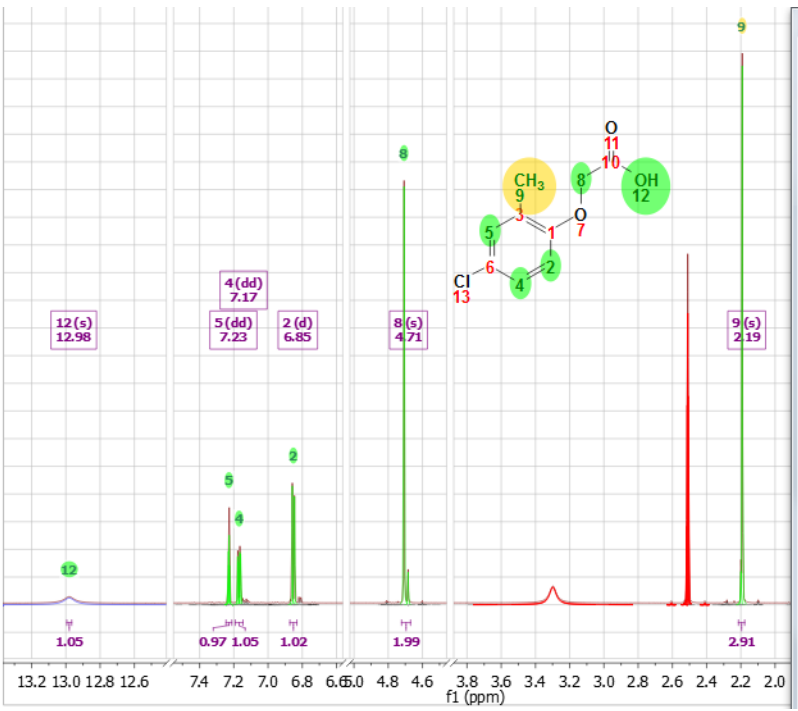
Consider now the Quinine spectrum with some «unstable» assignments:

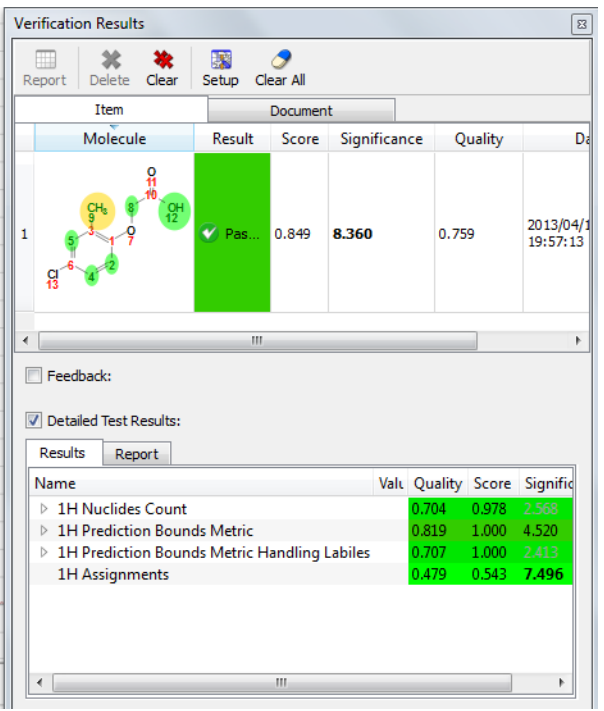


Mestrelab Research
chemistry software solutions

Some explanations VI: ASV versus AutoAssignments

Or this spectrum with all elementary assignments «stable»:





Mestrelab Research
chemistry software solutions

Some explanations VI: ASV versus AutoAssignments

The bottom line:

- When the spectrum has no structural info, ASV and AA suffer equally
- When the spectrum admits unambiguous structural determination, ASV and AA math each other (full correlation)
- When the spectrum admits a number of possible assignments, ASV is greatly advantaged (it is likely that at least one assignment might be correct), while AA id disadvantaged (it is unlikely that we will pick up exactly the correct assignment)

This is because ASV and AA are answers to different questions:

- ASV: could this structure and this spectrum possibly match each other?
- AA: what is the best correcpondence between nuclei in this structure and the multiplets in the spectrum?

Mestrelab Research
chemistry software solutions

Some explanations VII: The many uses of HSQC

HSQC, when available, is used in many places of the Data Analysis.

Care must be taken because its reliability on each of its «strong points» is only about 90%. Once handled properly, it is of considerable help.

Mestrelab Research
chemistry software solutions

Some explanations VIII: NMR is fuzzy to the extreme

- There is the noise and often a limited S/N
- Peaks and multiplets overlap in most unfortunate ways
- Peak shoulders may be real but it can not be taken for granted
- «Singlets», «doublets», «triplets» etc are such only on paper: more often a doublet has 5 GSD peaks and a triplet has 7 (but some triplets have only 2)
- No single rule a chemist ever told me (in particular Mike and Manuel) that would have stood up in practice for more than 2 months
- The Book is great but unreliable

In theory, Theory should agree with Practice,
but in practice, it rarely does

Mestrelab Research
chemistry software solutions

So what have we done in the last 12 months?

1. Instituted the Unjustified Bad Cases database (UBC; this is a truly great development tool. Please, contribute!)
2. Analysed the labiles problem – a huge improvement, based on Gonzalo Hernandez' statistical investigations.
3. Added many new tricks to handle bad and humpy water
4. Added new features in multiplets analysis (intelligent purging and slicing, scoring on each peak's multiplet membership, etc…)
5. Started exploiting in a detailed way the internal JC (J-correlations) table
6. Focused on assignments, both elementary (EA) and global (GA), achieving a huge progress in AutoAssign, though with surprisingly little effect on ASV
7. Added a penalty (veto-based) scoring, complementing the democratic one
8. Introduced the concept of a stable assignment
9. Reached a compromise between AutoAssign and ASV
10. Improved substantially diastereo predictions: a work in progress
11. Developed the concept of 2D *clusters* (a 2D analogy of 1D *multiplets*)
12. Integrated HSQC's (edited and not) with 1H more closely than ever before

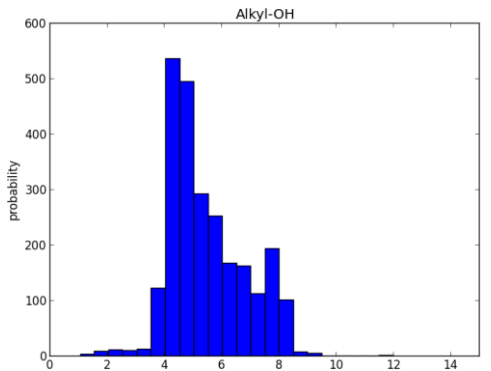
Mestrelab Research
chemistry software solutions

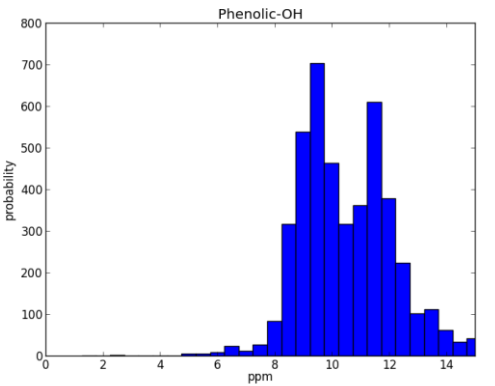
Instituted the Unjustified Bad Cases database (UBC) …

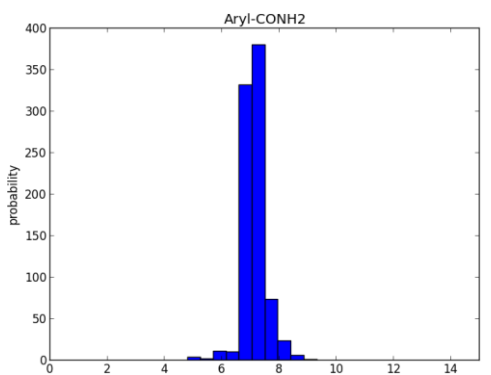
0061	L H	596	12/18/12	EV,CC	XXXXXX	602	602	677 677	LS	679	**Labiles+assignments.** *B602: Improved. B677: the **unmarked labile problem with PBML bad score** (a nice example). B680: Much better. Good solution for H16+H23. H9 not quite clear. Bad diastereo predictions?*
0062	L H	596	12/18/12	EV,CC	XXXXXX	602	602	677 677	HS		**Labiles+assignments**. Needs phasing. *B602: Improved. B677: Solved, but HSQC has unjustified low score (yellow)*
0063	L	598	12/19/12	EV,CC	XXXXXX	602	602	677	LS	680	**Labiles+assignments**. *B602: Assignments are OK. B677: the **unmarked labile problem**. B679: Labiles H8, H13 still not marked! B680: Solved. H11 coupled?*
0064	-	602	12/23/12	CC	Strychnine_400	677	602	677			A returning case of **water misrecognition**. B677: working again.
0065	L H	602	12/23/12	CC	Olanzapine		602	677 677	WM		**Humpy water overlaying a methyl peak**. Very tricky. B677: HSQC alone is OK.
0066	L H	602	12/23/12	CC	Thalidomide	664	667 667	667 667	AD HD	695 695	**Inverted assignments** 6',7'. *B664: it now works ok without E.M, but not with E.M = 0.3 Hz. HSQC does not help, though the wrong assignment is a strong violation. Reclassified as **edited HSQC diastereo***

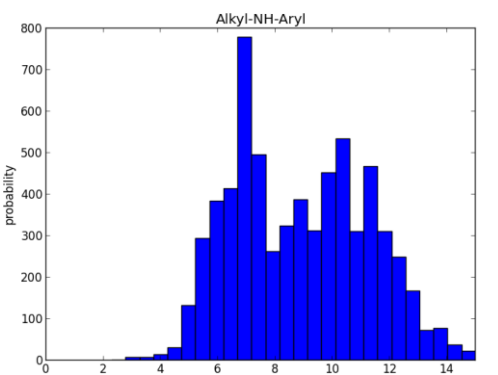
Mestrelab Research
chemistry software solutions

Analysed the labiles problem









C++Builder 6 - EBDataProc

File Edit Search View Project Run

D:\Stan\Soft\Mestrelab\MnEbSources\EbN

bList.cpp | EbNmrLabiles.cpp

```
LabileKindRange LabileKindR
// ------------------------
// The position of the firs
// All other entries can be
{
   {lpg_NotLabile,          1,
   {lpg_Unknown,            0,
   {lpg_AlkylOH,            0,
   {lpg_ArylOH,             0,
   {lpg_AlkylCOOH,          0,
   {lpg_ArylCOOH,           0,
   {lpg_AlkylNH2,           0,
   {lpg_ArylNH2,            0,
   {lpg_AlkylCONH2,         0,
   {lpg_ArylCONH2,          0,
   {lpg_AlkylSH,            1,
   {lpg_ArylSH,             0,

   {lpg_Triazoles,          0,
   {lpg_Pyrazoles,          0,
   {lpg_Imidazoles,         0,
   {lpg_Pyrroles,           0,


   {lpg_AlkylNHAlkyl,       0,
   {lpg_AlkylNHAryl,        0,
   {lpg_ArylNHAryl,         0,
   {lpg_AlkylCONHAlkyl,     1,
   {lpg_AlkylCONHAryl,      0,
   {lpg_ArylCONHAlkyl,      1,
   {lpg_ArylCONHAryl,       0,
   {lpg_AlkylSO3,           0,
   {lpg_ArylSO3,            0,
};
```

Mestrelab Research
chemistry software solutions

… achieving a huge progress in AutoAssign …

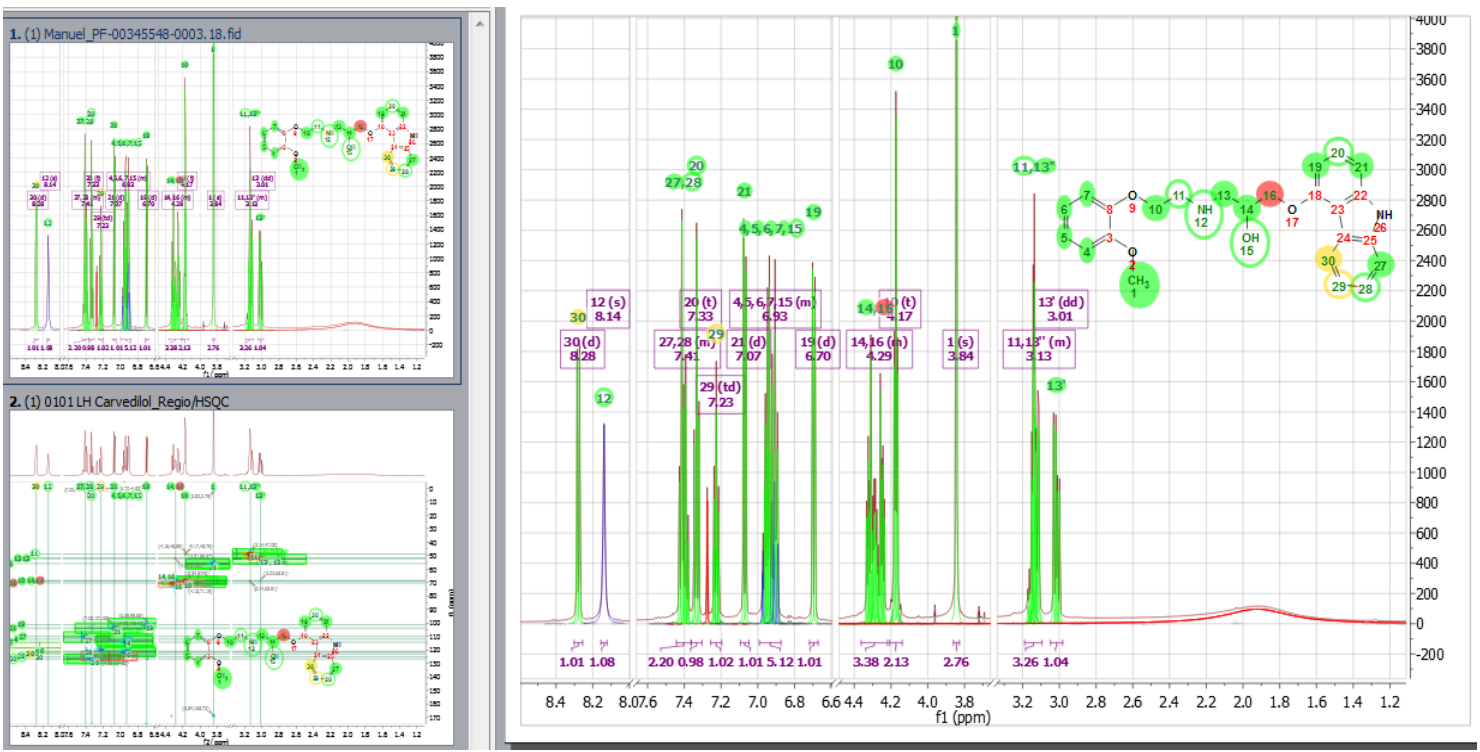
		1H assignments				
		8.1.0-11315	8.2.0-11361BETA	8.2.0-11430BETA	8.2.0-11705BETA	8.2.0-11870BETA
	build	488	602	612	663	690
1H	Right - 1H	861	859	769	901	900
	Wrong - 1H	198	210	307	183	184
	Missing - 1H	24	20	13	5	5
1H&HSQC	Right - 1HHSQC	865	903	809	901	993
	Wrong - 1HHSQC	191	164	259	183	88
	Missing - 1HHSQC	27	22	14	5	8

Build 488 is of 24 June 2012, build 690 of 9 April 2013

Mestrelab Research
chemistry software solutions

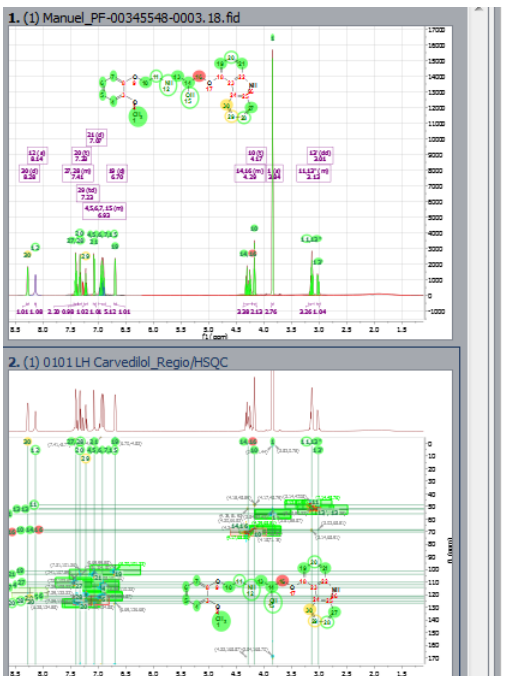
A couple of pretty pictures before the End

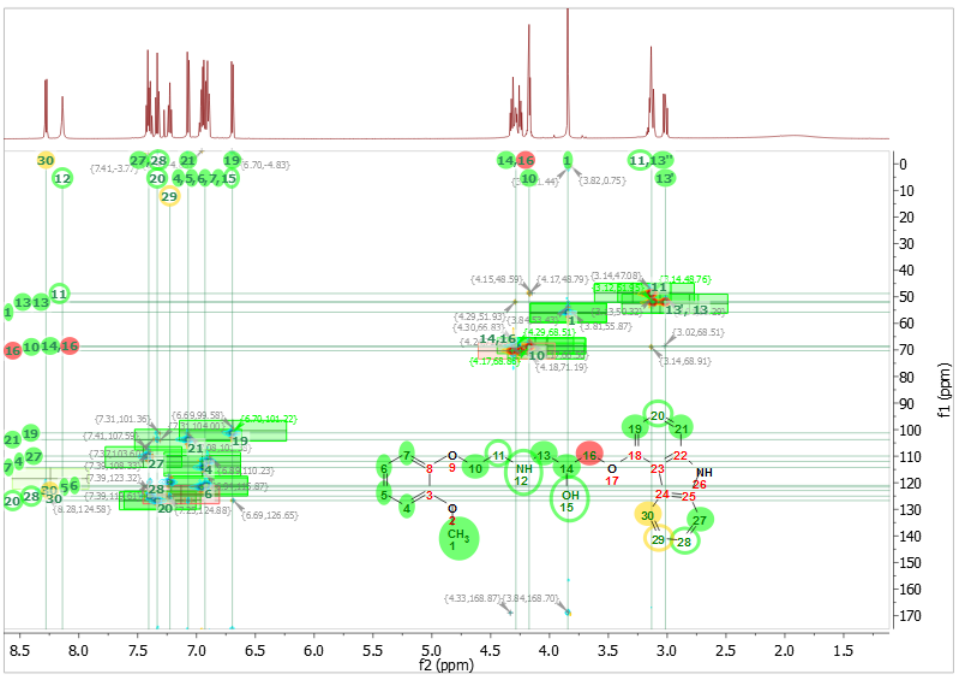
(just Carvediol, don't worry …)



Mestrelab Research
chemistry software solutions

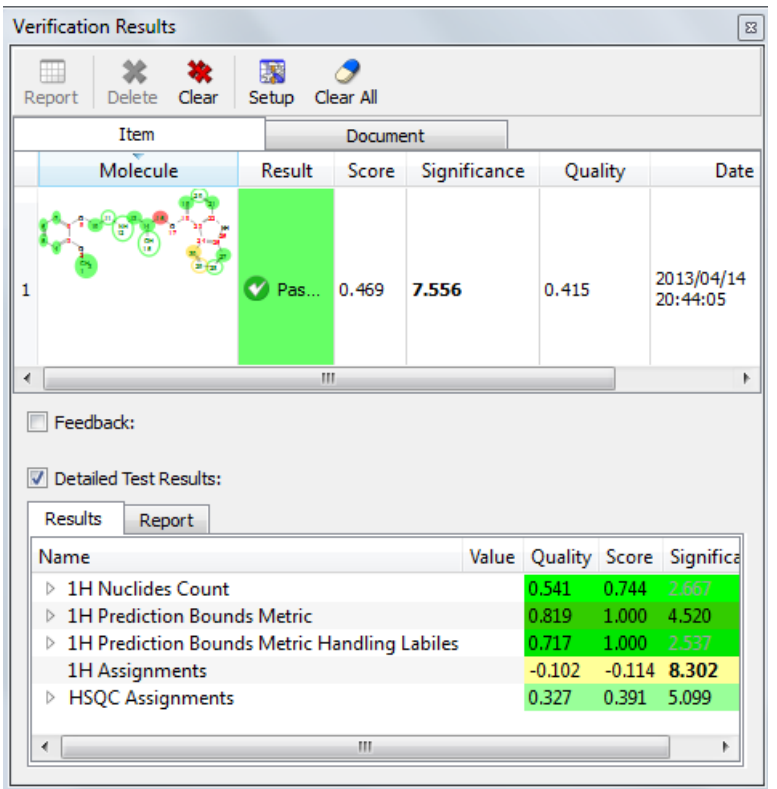
Carvediol edited HSQC





Mestrelab Research
chemistry software solutions

Carvediol ASV results



Mestrelab Research
chemistry software solutions

What's up and coming

- Across-the-board improvements: always, from GSD up
- Support for new kinds of spectra (13C, 19F, 31P, HMBC, JCOR, …)
- ASE: Automatic structure elucidation
- ASD: Automatic structures discrimination
- ACD: Automatic components detection
- User-Wizard interaction (true computer-aided design)

Mestrelab Research
chemistry software solutions

Some of the ASV & AutoAssign & «the next things» developers, testers, tuners, …

- Carlos Juan Cobas (the only *President* ever who understands every new idea on the fly)
- Stan Sykora (under influence, dreams up algorithms and writes number-crunching code)
- Felipe Seoane (defines Mnova interfaces, writes User interfaces, harmonizes the code)
- Esther Vaz, Pable Monje (members of the Testing and Tuning Team, the Triple-T)
- Mike Bernstein, Manuel Perez (VP's who supply iron ASV rules that sometimes work)
- Chen Peng (VP who keeps complaining that every chemist would know better than the AI)
- The Predictors (of **Modgraph**, they predict shifts and couplings)
- Oleg Ovchinnikov (checks on molecular structures and overrides The Predictors)
- Gonzalo Hernández (of **Vis Magnetica** generates structure candidates for false positives, compiles labile shifts histograms, and suggests additional chemical rules that sometimes work)
- Santi Dominguez (well, Santi is Santi is Santi, you know him …)
- + … **many others**, inside and outside Mestrelab: it keeps snow-balling

Warm thanks to all alpha testers!

Mestrelab Research
chemistry software solutions

Welcome to the Mestrelab suite
for more ASV chats and demos
… and some beer