



Improving the “Best” Predictions in NMRPredict

ENC

April 18th 2010



Prediction and Verification

- Automatic structure verification (ASV) is the “Holy Grail” for NMR spectroscopists
- Accurate and reliable prediction is a crucial element in automatic structure verification
- We are constantly trying to improve the accuracy of NMRPredict



The “Best” Prediction

- In ASV you need one prediction result for each nucleus
- But what if different prediction methods are used?
- It is our responsibility to apply rules to select the “Best” prediction from the various methods



NMRPredict Components

- **Carbon prediction** from Dr Wolfgang Robien, University of Vienna, developed since 1981
- **Conformer proton prediction** from Dr Ray Abraham, University of Liverpool, developed since 1984
- **Increment proton prediction** from Dr Ernő Pretsch, ETH Zurich, developed since 1969
- **3D conformers generation** using GMMX from Dr Kevin Gilbert, Serena Software



C13 Prediction

- A HOSE code database with up to 379,000 highly verified spectra
- A sophisticated Neural Network
- A real understanding of stereochemistry
- A choice of the “Best” prediction for each atom from the HOSE code and the Network



Improving C13 Prediction

- Adding more data to the HOSE code database
- Correcting errors in the HOSE code database
- Training the Network on more data
- Improving the selection of the “Best” value



New Data

- Wolfgang continues to add ~20,000 new data each year -
<http://nmrpredict.orc.univie.ac.at/csearchlite/update.htm>
- He also corrects about 50,000 data each year

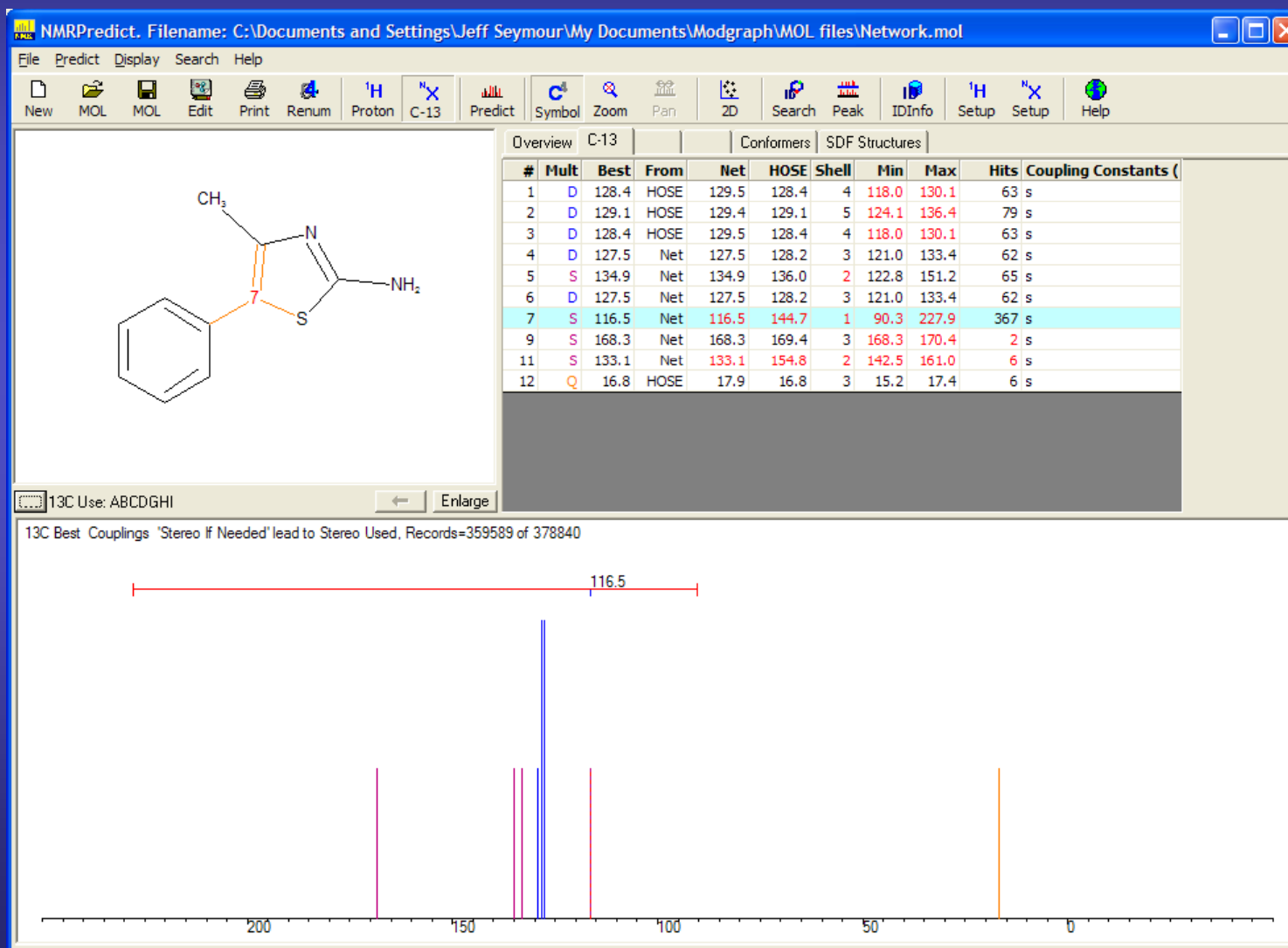


Improving the Neural Network

- The Network is now trained on 379,000 rather than 132,000 spectra



"Best" C13 Predicted Spectrum





“Best” (‘til now) C13 Prediction

- | | |
|---------------------------------|---------|
| – 1 shell | Network |
| – 2 shells | Network |
| – 3 shells - Singlet or doublet | Network |
| – 3 shells - Triplet or quartet | HOSE |
| – 4 shells | HOSE |
| – 5 shells | HOSE |



Catch 22 of improving predictions

- For us to improve predictions we need to know what customers are working on in the real world
- But a customer's data is usually confidential and cannot be shared



Case study

- We recently signed an NDA with a major pharmaceutical company in the USA
- They let us have 25 of their molecules with experimental C13 data
- This allowed us to check whether our “Best could be better”



Case study

- Our RMS for the data was:
 - HOSE code 5.38
 - Network 3.77
 - Best in NMRPredict 3.67
- The HOSE code alone is not enough. Data is unusual. Average shell only 2.7 with 380,000 records!
- The Best was better than the Network or the HOSE
- Could we change our rules to improve our current best?



Case study

- Wolfgang came up with some new rules based on the 25 molecules
- He then tested, modified and verified the rules with 5,000 different molecules which were not contained in NMRPredict
- All molecules contained at least two rings and at least one sulphur or nitrogen atom



New rules for “Best” C13

- | | |
|-------------------------------|--------------------------|
| – 1 shell | Mean of Network and HOSE |
| – 2 shells | Mean of Network and HOSE |
| – 3 shells (sp ²) | Mean of Network and HOSE |
| – 3 shells (sp ³) | 25% Network 75% HOSE |
| – 4 shells | 20% Network 80% HOSE |
| – 5 shells | HOSE |



Results

- Old RMS for the data was:
 - HOSE code 5.38
 - Network 3.77
 - Best in NMRPredict 3.67
 - **New RMS 3.14**
- Our “Best got Better” in half a day thanks to just 25 real life examples



Proton Prediction

- CHARGE conformer predictions from Ray Abraham
- Increment predictions from Ernő Pretsch
- 3D conformer generation from Kevin Gilbert



Improving Proton Prediction

- Ray improving his accuracy
- Ernő improving his accuracy
- Kevin's improving his reliability
- Improving the selection of the “Best” value



Ray Improving his Accuracy

- Ray has been reviewing all of his functional groups with significant improvements in some cases
- A new version of CHARGE is included in NMRPredict 4.7



Ernö Improving his Accuracy

- Ernö now uses a hand selected database to apply a “correction” to his native value. This database is being expanded
- Ernö has started to make use of the 3D conformers from GMMX and will expand its use

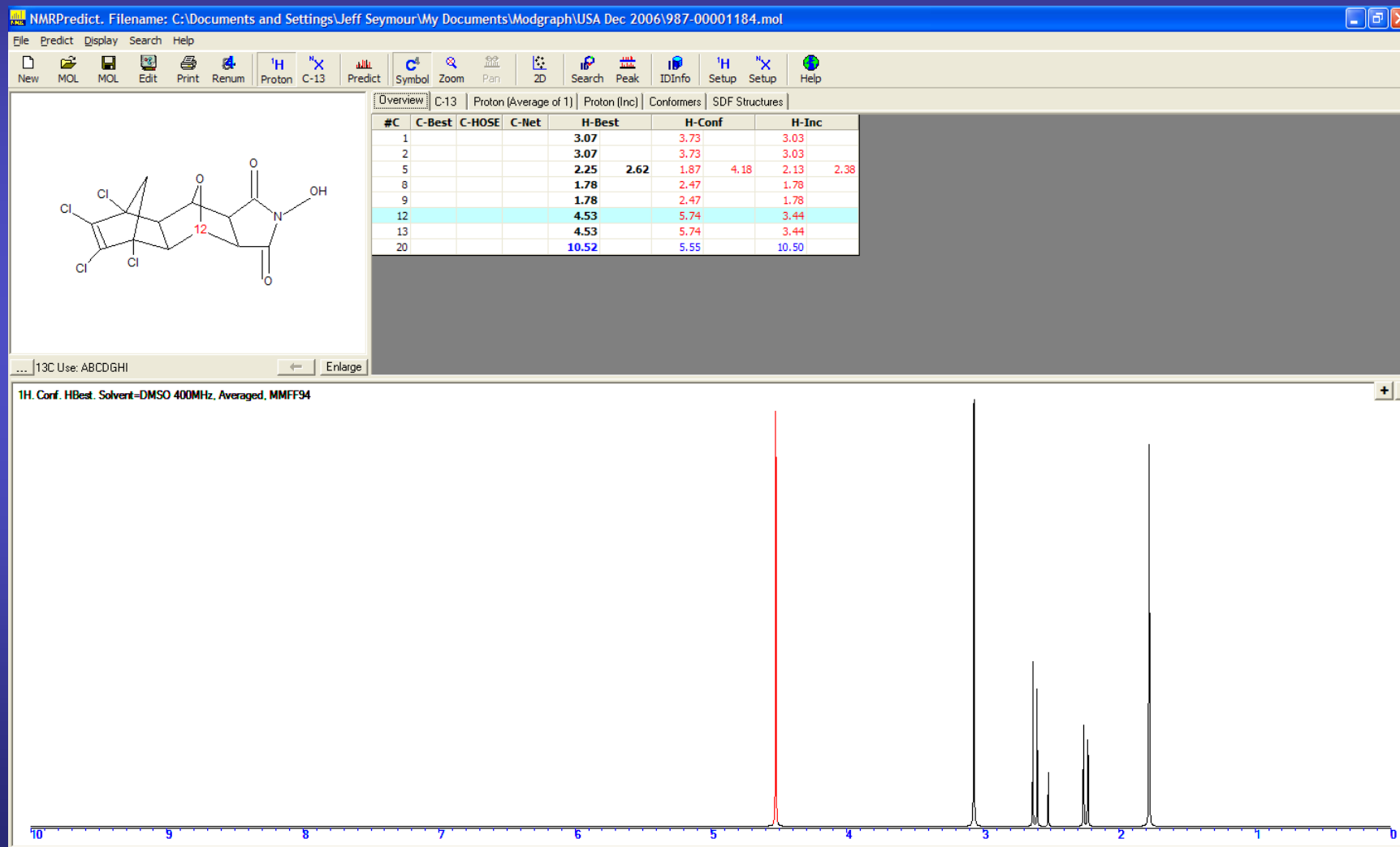


Kevin Improving his Reliability

- Kevin has re-written the MMX (MMFF94) routine within GMMX
- This is both faster and more reliable than before
- This is included in NMRPredict 4.7



"Best" Proton Prediction





The “Best” Proton Prediction

- In order to determine how to select a “best” proton value we ran predictions against 91,077 assigned structures from Wiley in 2007
- 961,922 non-labile shifts were predicted
- 6,500 one bond HOSE codes were found in the data



How did we select the “Best” in 2007?

- We applied a correction factor to the two native values based on the 6,500 chemical environments around each atom
- We took a weighted average of the two values, again based on the chemical environment around each atom
- The result was totally statistical based on the 90,000 molecules



Average Deviation

0.13ppm – “Best Possible”

0.19ppm – Our 2007 “Best”

0.21ppm – Ernö

0.27ppm – Ray

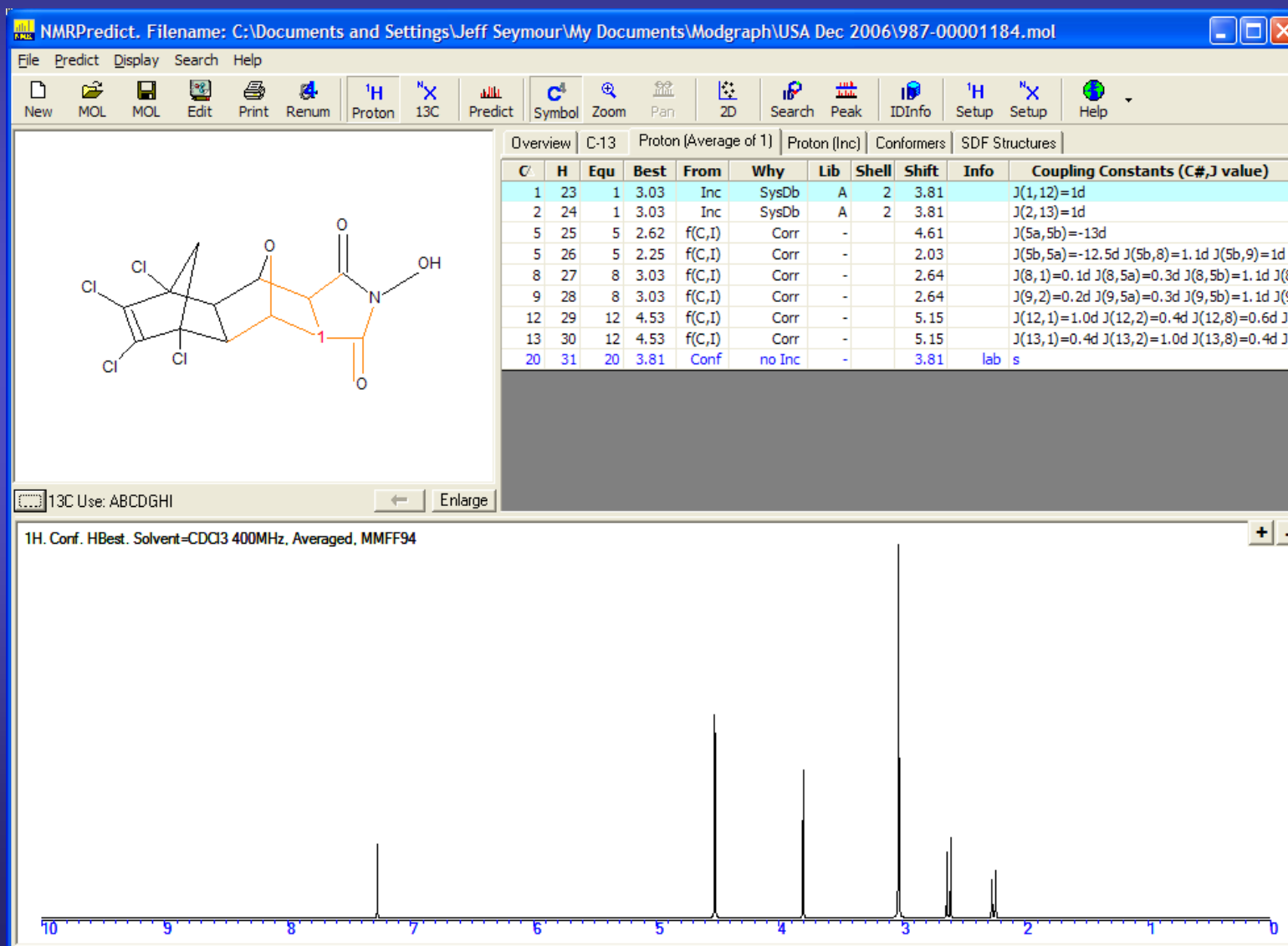


2009 results against the 90,000

- We re-ran our predictions against the 90,000 in July 2009
- We now have a column to say when Ernő is able to use his own “correction database”
- We wanted to know if results differed when the database was used or not



Ernö's Database





2009 results against the 90,000

- 90% of the atoms could be predicted using the database
- Prediction with database
 - Ray 0.25 ppm
 - Erno 0.18 ppm
- Prediction without database
 - Ray 0.40 ppm
 - Erno 0.46 ppm

Whether or not the database is used is critical



Change to selection of “Best”

- If Ernö’s database has been used we will take his native value and not apply a “correction to a correction”
- We will only use Wolfgang’s correction if Ernö’s database was not used
- When no correction is possible we will take Ray’s value not Ernö’s



New “Best”

- Prediction with database
 - Ray 0.25 ppm
 - Erno 0.18 ppm
 - Best 0.18 ppm
- Prediction without database
 - Ray 0.40 ppm
 - Erno 0.47 ppm
 - Best 0.36 ppm

New “Best” now optimized on smaller more specific dataset



Is the 90,000 representative?

- We recently received 28 assigned proton spectra from another major US pharmaceutical company
- We are now able to compare the results of the 90,000 and “industry data”



Comparing results

	90,000	Industry data
• Best possible	0.13	0.17
• Used database	90%	60%
• Ernö	0.21	0.29
• Ray	0.27	0.27
• Best	0.19	0.22



Improving the Best in proton

- Do we need to modify our rules depending on the number of shells, as in carbon?
- Can we apply chemical selection rules when through-space effects should be considered?



Ernö Accuracy at Different Shell Levels

- 5 shells 0.09
- 4 shells 0.14
- 3 shells 0.20
- 2 shells 0.28
- As with carbon, there is a big difference in accuracy depending on the shell level



Mean value at different shells

Ernö

- 5 shells 0.09
- 4 shells 0.14
- 3 shells 0.20
- 2 shells 0.28

Mean

- 5 shells 0.12
- 4 shells 0.15
- 3 shells 0.19
- 2 shells 0.26



Proton Chemical Selection Rules

- What if we have taken Ernö's value but...
- Ray and Ernö differ "a lot" (by >0.4 ppm)
- Ernö is "bad" (>0.3 ppm from experimental)
- Ray is "good" (<0.2 ppm from experimental)
- Are there chemical rules we can apply to find these circumstances?
- This only applies with 2.8% of the 950,000 atoms



NMRPredict version 4.8 – July 2010

- Weekly or monthly automatic synchronisation with Wolfgang's master databases
- Implementation of new carbon "Best" rules
- Implementation of new proton "Best" rules
- Proton "Best" based on chemistry and through-space effects?



Conclusion

- Accurate prediction is central to successful ASV
- A prediction must come from at least two different complementary methods
- The vendor must supply a “Best” prediction from the various methods
- The selection of the Best must take into account the number of shells reached by the database method
- If customers can supply us with even a small quantity of data it can make a significant difference



Trials

- Please try NMRPredict or NMRPredict Desktop at your own facility
- See me at ENC to arrange an in-house trial of NMRPredict
- Or email me at:

Jeff@modgraph.co.uk



Free service for the next 3 months

- Register with Wolfgang by email
- You will be sent to a web page with a full prediction from NMRPredict at no charge
- http://nmrpredict.orc.univie.ac.at/wr/0e0e7863ed222875e3774d71400236e3_summary.html
- You are limited to three predictions per day



New service for publications

- Use Wolfgang's new service to verify that the data you are publishing is correct
- Enter your molecule and shift values and get a reply telling you whether we recommend the data to be accepted or not



Trials

- See me at ENC to arrange to try this service
- Or email me at:

Jeff@modgraph.co.uk