chem-bla-ics

Inter- and Extrapolation: the NMR shift prediction debate

Egon Willighagen

Published July 13, 2007

Citation

Willighagen, E. (2007). Inter- and Extrapolation: the NMR shift prediction debate. In *chem-bla-ics*. chem-bla-ics. https://doi.org/10.59350/tbd0q-67564

Keywords

Nmr, Nmrshiftdb

Chemical blogspace has seen a lengthy discussion on the quality of a few NMR shift prediction programs , and Ryan wanted to make a final statement. Down his blog item he had this quote from Jeff, discussing the use of the NMRShiftDB as external test set:

> *"Of course customers are really interested in how accurately a prediction program can predict THEIR molecules - not a collection of external data such as NMRShiftDB."*

I'm sure none of us knows what weird chemistry people are doing; we will never know what the overlap of the NMRShiftDB test set with the customer data set is. The quote suggests it is low, but we simply do not know.

Interpolation and Extrapolation

The accuracy of prediction models is very difficult to grasp, and one can only estimate it; using a test set. If few data is available, one may opt for using the training set as test set too, and gives an estimate if the modeling method is able to predict at all. However, the outcome of this exercise is the worst possible estimate you can make. So, when possible you use an independent test set, which does not contain any molecules that were present in the training set. (Actually, one could even suggest that this must happen on a shift level, but that gives problems with HOSE-code based prediction.)

Now, what Ryan stresses in his latest blog item is that prediction test results for the various available methods does not explicitly state the amount of overlap between the training and test set, one cannot draw any conclusions. Agreed. I would, however, like to tune this even a bit further, after reading the stupid quote (of course, taking out of context). What Jeff probably aimed at, is that the prediction accuracy is only meaningful to a customer if there is considerable between the customers data set and the test set, which is what the model makers do not know.

And the overlap actually goes beyond the overlap in terms of molecular identity. It is really the overlap in terms of molecular substructures that matters: a database with alkanes but no phenyl rings will more accurately predict other alkanes not present in the training set (interpolation), but will not accurately predict compounds with phenyl rings (extrapolation). What the customer needs is that his personal data set does not require extrapolation. That is what matters.

It is interesting to realize, however, that the NMRShiftDB allows you to upload your molecules, or alternatively, you download the software (it's open source) and the data (it's open data) if you don't want to send your molecules over the internet, and the NMRShiftDB software will automatically take into account your own data set.

Thus, if you are working on a series of related molecules, you can extend the NMRShiftDB data set with already elucidated structures, reducing the prediction error for your yet related unknowns derivatives. It is that easy to include prior/expert knowledge in the NMRShiftDB. I

believe the ACD/Labs software allows this too, so the quote is really meaningless. Not correct, not wrong, simply says nothing.

Open Data, Open Source, Open Standards

Now, the various releases of the ACD/Labs software show a simple, understandable trend that increasing the number of data you use for the training set, reduces the prediction error. That's because of various reasons I will not go into in this item. The ACD/Labs NMR databases are expensive, because they have to manually extract and validate the data from literature (see The Purgatory Database); so, during my PhD I only bought the CNMR and HNMR prediction packages. (Off topic: two weeks after I received my copies of the software, ACD/Labs released a new version, which they kindly sent me a copy of too. Common in opensource, but much appreciated at that time. Cheers, ACD/Labs!)

The ACD/Labs databases are likely expensive because of various reasons. And this is where the ODOSOS concept of the Blue Obelisk comes in. **Open Data**: if publishers would not copyright their data, NMR databases would be much cheaper to set up (see this thread in Peter's blog); assuming ACD/Labs has to pay publishers for actually setting up their database. **Open Source**: the various Blue Obelisk projects provide the tools to automatically create a purgatory NMR database ; no humans needed for that any more. **Open Standards**: the data from the NMRShiftDB can be downloaded in various formats, among which CMLSpect. Being able to easily read the data, made it possible that we actually have this discussion. Sure, the open data part of the NMRShiftDB is crucial too! But the database could have used an obscure, binary, undocumented, with many software tweaks and special cases, `.doc`-like format, which no one could support.

Clearly, ODOSOS gives all, even proprietary, NMR prediction tools a boost, and I am very happy to see that happen. It is the point that we, the Blue Obelisk Movement, are trying to make for some time now.