# A Comparison of NMR Predictors Using Tree-based Similarity of Spectra

**Andrés M. Castilloa, Andrés Bernalb, Luc Patinyc, Julien Wistd,\***

*a Universidad Nacional de Colombia, Facultad de Ingeniería; b Grupo de Química Teórica, Universidad Nacional de Colombia, c Institute of Chemical Sciences and Engineering, Ecole Polytechnique Fédérale de Lausanne (EPFL), CH-1015 Lausanne, Switzerland; d Chemistry Department, Universidad del Valle, Cali, Valle, Colombia;*

## Abstract

A tree-based method for NMR spectra comparison developed recently by the authors enables the design of a methodology for the comparison and evaluation of NMR prediction algorithms. First, each prediction tool is used to predict a spectrum for each molecule of the dataset and to construct matrices by evaluating the similarity with their corresponding experimental spectra. In a second step, the molecules corresponding to the predicted spectra that show most similarity with a selected experimental spectrum are considered as probable candidate structures. The best candidate is checked against the true structure before being considered successful, if both match. This query is repeated for each experimental spectrum allowing to rank NMR predictors according their ability to produce accurate, thereby similar, spectra. This methodology dispenses us with the need of assigning each signal with its nucleus, making the task of benchmarking NMR predictors less tedious. This approach allowed to compare four popular NMR predictors (ACD/Labs, Modgraph, Chemdraw, Spinus) using a dataset of 1000 molecules and their corresponding experimental spectra. We found that ACD/Labs provides the best results by far, while the free Spinus platform’s performance is comparable to that of the other commercial alternatives. These results are consistent with those obtained by directly comparing deviations between predicted versus experimental shifts.

## Keywords

Nuclear Magnetic Resonance, spectrum prediction, 1H chemical shift prediction, spectral similarity, trees.

### Introduction

Cheminformatics plays an increasingly important role in structure validation by NMR spectroscopy, providing methods and algorithms for computer-assisted assignment [1-3,19], elucidation [4-19], and prediction [19-26] of NMR spectra. Along with the introduction of such novel methods comes the need to compare and evaluate the available alternatives, a task that can be performed using the techniques of cheminformatics itself.

In a recent article the authors presented a tree-based method for measuring similarity between NMR spectra [29]. It was shown to produce results comparable to those of the binning method [30], with significant improvement in efficiency by focusing on the regions of the spectrum with most information. Furthermore, this new approach also directly operates on raw spectra, i.e., without the need of a peak-picking subroutine. This features turn it into an attractive tool for the comparison and evaluation of NMR predictors, as it allows to measure the similarity of predictions and experiments without requiring to assign peaks to nuclei. This article presents a such methodology and shows the comparison of four popular NMR prediction platforms.

### Methodology

The success of an NMR prediction algorithm is determined by its ability to reproduce the experimental chemical shifts. Determining the adequacy of a prediction thus implies having assigned the experimental spectra, and having their chemical shifts compared with predicted ones. Peak-picking and assignment are troublesome and time-demanding tasks, however. As an alternative, we propose to evaluate the success of a prediction algorithm by its ability to produce, by means of a proper simulation algorithm, a spectrum that is *sufficiently* similar to the one given by the experiment. This similarity can be measured directly on the spectra using the tree-based method developed by the authors [29], thus avoiding the need for peak-picking and assignment.

**Evaluation of prediction tools:** To see what is understood by *sufficiently* similar, we consider the experimental and simulated spectra for each element of a collection of molecules and build the matrix of similarities between each experimental and each simulated spectrum (see Figure 1). An accurate prediction algorithm ensures that the highest similarity values lay on the diagonal of such matrix, i.e. the experimental spectrum of any given molecule would be more similar to its simulation than to predicted spectra of other molecules. The performance of a predictor can thus be measure by the Mean Reciprocal Rank (MRR),

where *n* is the number of queries and *ranki* is the rank of the correct answer to the i-th query.

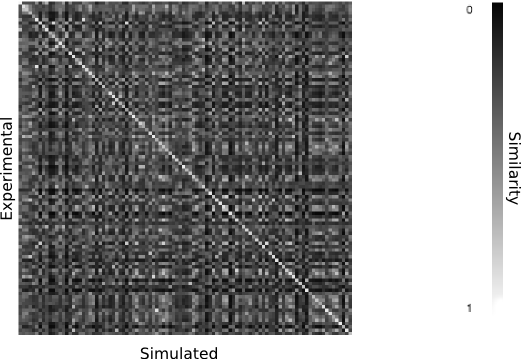


Figure 1: Example of a spectra similarity matrix. Rows correspond to experimental spectra and columns to simulated spectra of a 100 molecule data set, matrix elements give the similarity between the corresponding experimental a simulated spectrum. The thin light gray line on the diagonal shows a trend towards higher similarity between the matching spectra, as desired of a NMR predictor.

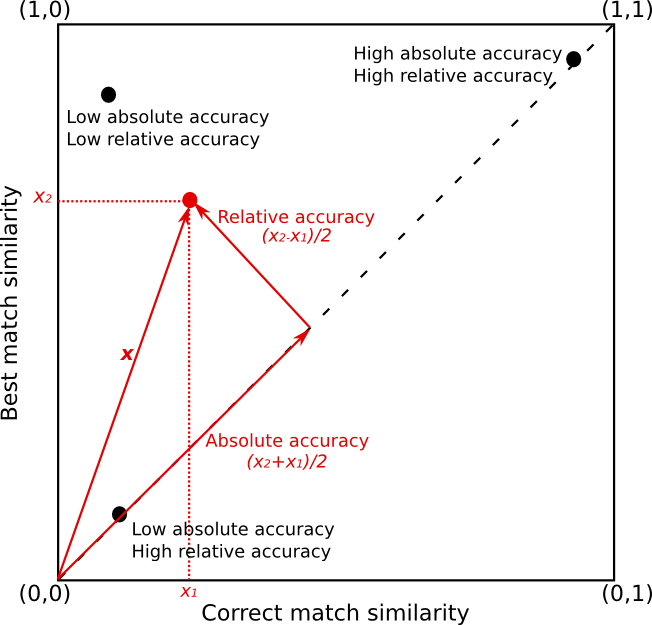


Figure 2 Position of a query of a simulated spectra to the experimental database on the best match vs. correct match similarity plane works as a descriptor of the accuracy of the prediction. The component of the position vector on the direction of the identity (dashed line) describes the accuracy in absolute terms: the larger the component, the greater the similarity between the simulated spectra and the correct experimental match. The component orthogonal to the identity describes the accuracy relative to the database: the smaller the component, the greater the similarity between the correct match and the best match among all spectra in the database.

A complementary approach consists in identifying a query with a point in the correct-hit-similarity vs. best-match-similarity plane (Figure 2). Note that in this plane:

* All practically realizable points are located over the identity function (dashed line going from (0,0) to (1,1)), as the correct match cannot be more similar to the experimental spectrum than the best match.
* Points located on this line correspond to those cases where the best match is the correct match.
* The accuracy of the prediction in absolute terms (i.e. in terms of the similarity between the correct match and the experimental spectrum) increases as we move up to the extreme at (1,1), where the correct match and the experimental spectrum are identical.
* The accuracy of the prediction relative to the data set (i.e. the ratio between the experimental spectrum’s similarity to the correct match and its similarity to the best match in the data set) decreases as we move away from the identity line.

This observations motivate the decomposition of each point ***x=****(x1,x2)* of the plane into its component along the identity line,

and the orthogonal component,

The first component describes the accuracy of the prediction in absolute terms: the closest to 1, the most accurate the prediction. The second component, on the other hand, describes the accuracy of the prediction relative to the data set: as it approaches zero, the correct match moves towards the best match. Unlike MRR, this technique allows us to look into the actual similarity values between experimental and simulated spectra, distinguishing between low-ranking queries due to poor prediction and low-ranking queries due to high spectrum similarity.

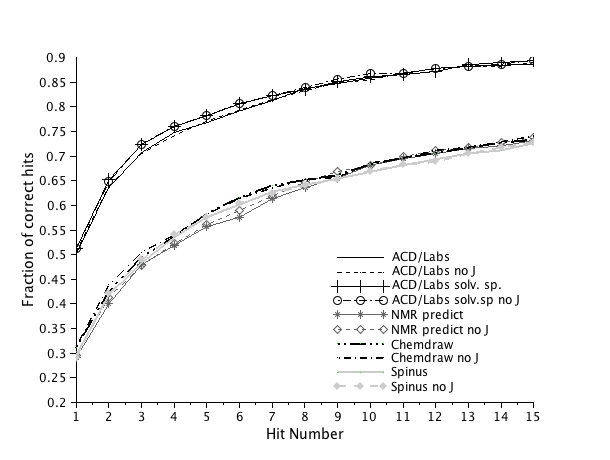
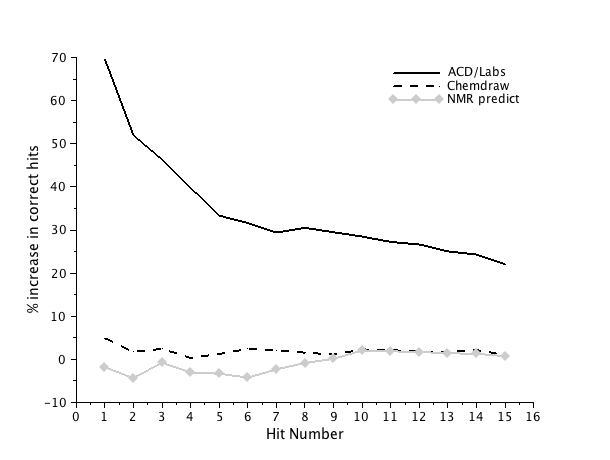


Figure 3: Fraction of correct hits obtained that fall in the n-first bins (n = Hit Number). Each point corresponds to the fraction of correct matches within the *n* highest ranking hits of a query of 1000 simulated 1H spectra to the data base of the corresponding experimental spectra. For example, using ACD/Labs solvent specific predictor with no coupling constant simulation, around 75% of the correct matches are found within the 4 highest ranked hits.

|  |  |
| --- | --- |
| **Method** | **MRR** |
| ACD/Labs | 0.59359 |
| ACD/Labs no J | 0.60013 |
| ACD/Labs solvent specific | 0.61047 |
| ACD/Labs solvent specific no J | 0.61504 |
| NMRPredict | 0.36927 |
| NMRPredict no J | 0.37289 |
| Chemdraw | 0.40388 |
| Chemdraw no J | 0.4091 |
| Spinus | 0.37565 |
| Spinus no J | 0.37521 |

Table 1: Comparison of the MRR values obtained for four prediction methods, computed from 1000 queries of simulated spectra to a database of corresponding experimental spectra. Higher values indicate more accurate predictions.



**Figure 4: Comparison of commercial algorithms with the free online Spinus system regarding the amount of correct hits.**

**Comparison of NMR spectra:** The tree-based methodology for the measurement of spectrum similarity applied here has been described in detail elsewhere [29]; it consists in the construction of a tree representation of each spectrum that summarizes key information on its signal-rich regions using fewer data points, followed by the computation of a similarity measure between these trees. This technique is similar to the traditional binning technique [28] but presents the advantage of focusing on regions with high signal intensity, avoiding large blank or merely noisy zones.

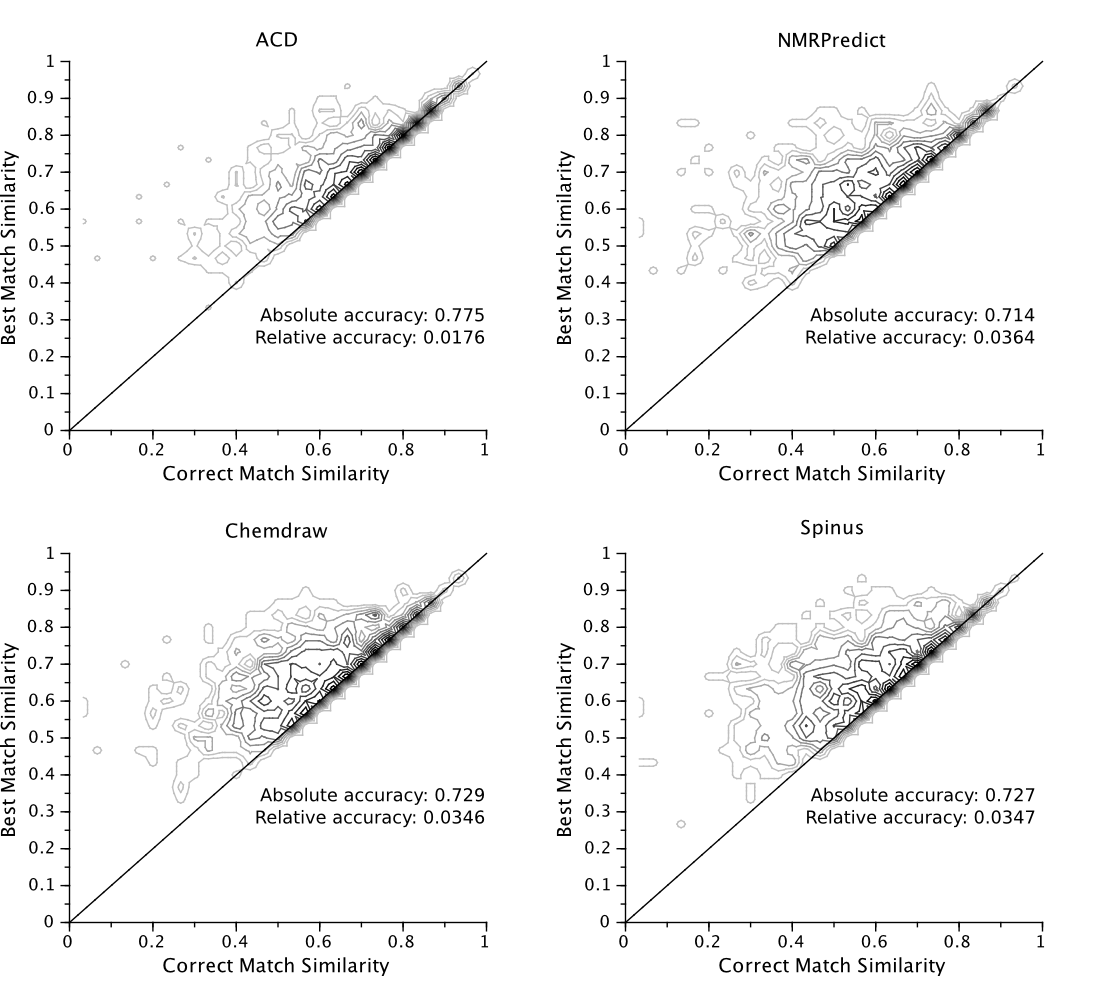


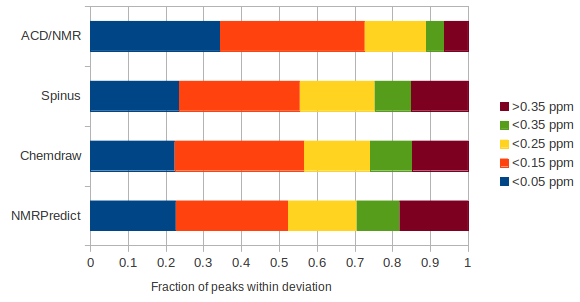
Figure 5 Contour plots on the best match vs. correct match similarity plane of the query distributions. An ideal prediction tool would have all the density packed along the diagonal.

## Experimental section

A set of 1000 molecules of up to 33 heavy atoms was chosen. For each of them, the 1H-NMR chemical shifts were predicted using ACD/Labs NMR predictor [22], Modgraph NMRPredict [23], Chemdraw 9 [24], and the Spinus [25-28] online platform. Each prediction was used to simulate a spectrum at a frequency of 400 MHz and 1024 points with the algorithm that we described elsewhere [31]. As not all the tested algorithms predict coupling constants, two data sets were generated for each system: the first one by using the coupling constants predicted by Spinus in the simulation, and the second by ignoring them. Furthermore, both the regular and the solvent-specific versions of ACD/Labs’ algorithm were tested. Similarity matrices between simulated and experimental spectra, MRR, and average absolute and relative prediction accuracy were computed for each data set using the methodology described on the previous section. The molecules chosen in this works come from Maybridge’s website [32] and the assignment of their experimental spectra was done in our lab.

## Results and Discussion

Figure 3 shows the distributions of correct hits within the *n* highest ranking hits for each prediction algorithm. It can be seen that ACD/Labs predictor performed significantly better than all other algorithms. This result is confirmed by the MRR values reported in Table 1. The effect of ACD/Labs’ solvent correction and Spinus predicted coupling constants were small overall, providing a 2% improvement on results at best. It was expected that the inclusion of coupling constants would have little to no effect: typical H-H coupling constants range from 2-12 Hz, which is too small compared to the size of the spectral regions considered during construction of the tree to produce a significant shift on the location of their mass centers as a consequence of peak splitting. As these corrections were not found to be much significant, they were omitted on posterior analyses.



## Figure 6: Evaluation of predictors by direct comparison of predicted and experimental shifts using the same data set employed for evaluation through the tree-based similarity method. This approach required the manual peak-picking and the assignment of all experimental spectra.

Figure 5 displays the predictions of the four predictors on the correct match similarity vs. best match similarity plane. Clearly, it appears that ACD/Labs produces predictions more closely packed along the identity line, which is associated with better relative accuracy as discussed above. This suspicion is confirmed by computation of the absolute and relative accuracy indices. It is found that ACD/Labs improves absolute accuracy by ~10% and relative accuracy by ~100% compared to the other algorithms (See Figure 5). Once again, the remaining three predictors were found to perform similarly.

To validate our approach, we repeated the ranking of the four prediction tools but now using a *traditional* approach: experimental signals were assigned to their corresponding nuclei and the differences between experimental and predicted shifts were computed. These chemical shift errors were partitioned on 0.1 ppm intervals up to 0.35 ppm, the last interval comprising errors over 0.35 ppm. The resulting histograms are shown in Figure 6. Again, it was found that ACD/Labs predictor’s performance is clearly superior, producing around 10% more predictions on the two lower error intervals, while the other systems performed comparatively, which matches the results obtained through our method.

## Conclusions

The direct comparison of simulated and experimental spectra was shown to be a suitable tool for the evaluation of NMR prediction algorithms, allowing for an efficient and fully automatic methodology that yields results that are consistent with those obtained by the traditional expert-driven method based on peak picking and assignment.

## References

[1] Nuzillard, J.-M. & Massiot, G., Computer-aided spectral assignment in nuclear magnetic resonance spectroscopy, *Analytica chimica acta*, 1991, **242**, 37-41

[2] Vitek, O.; Vitek, J.; Craig, B. & Bailey-Kellogg, C., Model-based assignment and inference of protein backbone nuclear magnetic resonances, *Statistical applications in genetics and molecular biology*, 2004, **3**, 6

[3] Vitek, O.; Bailey-Kellogg, C.; Craig, B.; Kuliniewicz, P. & Vitek, J., Reconsidering complete algoriths for protein backbone NMR assignment, *Bioinformatics*, 2005, **21**, ii230-ii236

[4] Bangov, I. & Simova, S., Computer-assisted structure generation from a gross formula: 6 Reducing the structural redundancy by the employment of 2D NMR spectral information, *Journal of chemical information and computer science*, 1994, **34**, 546-557

[5] Christie, B. & Munk, M., The role of two-dimensional nuclear magnetic resonance spectroscopy in computer-enhanced structure elucidation, *Journal of the american chemical society*, 1991, **113**, 3750-3757

[6] Christie, B. & Munk, M., Structure generation by reduction: a new strategy for computer-assisted structure elucidation *Journal of chemical information and computer science*, 1988, **28**, 87-93

[7] Christie, B. & Munk, M., The application of two-dimensional nuclear magnetic resonance spectroscopy in computer-assisted structure elucidation, *Analytica chimica acta*, 1987, **200**, 347-361

[8] Funatsu, K.; del Carpio, C. & ichi Sasaki, S., Automated structure elucidation system -CHEMICS *Fresenius Z Anal Chem*, 1986, **324**, 750-759

[9] Funatsu, K. & ichi Sasaki, S., Recent advances in the automated structure elucidation system, CHEMICS. Utilization of two-dimensional NMR spectral information and development of peripheral functions for examination of candidates, *Journal of chemical information and computer science*, 1996, **36**, 190-204

[10] Lindel, T.; Junker, J. & Köck, M. COCON: From NMR correlation data to molecular constitutions, *Journal of molecular modeling*, 1997, **3**, 364-368

[11] Junker, J.; Maier, W.; Lindel, T. & Köck, M., Computer-assisted constitutional assignment of large molecules: COCON analysis of ascomycin, *Organic letters*, 1999, **1**, 737-740

[12] Lindel, T.; Junker, J. & Köck, M., 2D-NMR guided constitutional analysis of organic compounds employing the computer program COCON, *European journal of organic chemistry*, 1999, 573-577

[13] Nuzillard, J.-M. & Massiot, G., Logic for structure determination, *Tetrahedron*, 1991, **47**, 3655-3664

[14] Ley, S.; Doberty, K.; Massiot, G. & Nuzillard, J.-M., "Connectivist" approach to organic structure determination LSD-program assisted NMR analysis of the insect antifeedant Azadirachtin, *Tetrahedron*, 1994, **50**, 12267-12280

[15] Masui, H. & Hong, H., Spec2D: A structure elucidation system based on 1H NMR and H-H COSY spectra in organic chemistry, *Journal of chemical information and modelling*, 2006, **46**, 775-787

[16] Munk, M. & Christie, B., The characterization of structure by computer, *Analytica chimica acta*, 1989, **216**, 57-68

[17] Schaller, R.; Munk, M. & Pretsch, E., Spectra estimation for computer-aided structure determination, *Journal of chemical information and computer science*, 1996, **36**, 239-243

[18] Liu, X.; Balasubramanian, K. & Munk, M., Computer-assisted graph-theoretical construction of 13C NMR signal and intensity patterns, *Journal of magnetic resonance*, 1990, **87**, 457-474

[19] Elyashberg, M. E.; Williams, A. J. & Martin, G. E., Computer-assited structure verification and elucidation tools in NMR-based structure elucidation, *Progress in Nuclear Magnetic Resonance Spectroscopy*, 2008, **53**, 1-104

[20] Golotvin, S. S.; Vodopianov, E.; Lefebvre, B. A.; Williams, A. J. & Spitzer, T. D., Automated structure verification based on 1H NMR prediction, *Magnetic resonance in chemistry*, 2006, **44**, 524-538

[21] Golotvin, S. S.; Vodopianov, E.; Pol, R.; Lefebvre, B. A.; Williams, A. J.; Rutkowse, R. D. & Spitzer, T. D., Automated structure verification based on a combination of 1D 1H NMR and 2D 1H-13C HSQC spectra, *Magnetic resonance in chemistry*, 2007, **45**, 803-813

[22] ACD/HNMR Predictor v.9.0, Advanced Chemistry Development, Inc., 110 Yonge Street, 14th floor, Toronto, Ontario, Canada M5C1T4.

[23] <http://www.modgraph.co.uk/product_nmr.htm>

[24] <http://www.cambridgesoft.com/Ensemble_for_Chemistry/ChemDraw/>

[25] J. Aires-de-Sousa, M. Hemmer, J. Gasteiger, Prediction of 1H NMR Chemical Shifts Using Neural Networks, *Analytical Chemistry*, **2002**, *74*(1), 80-90

[26] Y. Binev, J. Aires-de-Sousa, Structure-Based Predictions of 1H NMR Chemical Shifts Using Feed-Forward Neural Networks, *J. Chem. Inf. Comp. Sci.*,**2004**, *44*(3), 940-945

[27] Y. Binev, M. Corvo, J. Aires-de-Sousa, The Impact of Available Experimental Data on the Prediction of 1H NMR Chemical Shifts by Neural Networks, *J. Chem. Inf. Comp. Sci.*, **2004**, *44*(3), 946-949

[28] Y. Binev, M.M. Marques, J. Aires-de-Sousa, Prediction of 1H NMR coupling constants with associative neural networks trained for chemical shifts, *J. Chem. Inf. Model.* **2007**, *47*(6), 2089-2097.

[29] Castilo, A., Uribe, L., Patiny, L., Wist, J., Similarity between NMR spectra, *Chemometrics and Intelligent Laboratory Systems* (2013), doi:10.1016/j.chemolab.2013.05.009

[30] Bodis, L., Ross, A. & Pretsch, E. A novel spectra similarity measure. *Chemometr. Intell. Lab.* **85**, 1–8 (2007).

[31] Castillo, A. M., Patiny, L. & Wist, J., Fast and accurate algorithm for the simulation of NMR spectra of large spin systems., *Journal of magnetic resonance* **209**, 123–130 (2011).

[32] [http://www.maybridge.com](http://www.maybridge.com/)