*Poster Presented at ENC 2007, Daytona Beach, FL, USA, April 23–26, 2007*

# NMR Chemical Shift Prediction by Atomic Increment-Based Algorithms

*Yegor D. Smurnyy, Kirill A. Blinov, and Mikhail E. Elyashberg,*
*Advanced Chemistry Development, Ltd.,*
*Moscow, Russia*

*Brent A. Lefebvre, and Antony J. Williams*
*Advanced Chemistry Development, Inc.,*
*Toronto, ON, Canada*

## INTRODUCTION

*In-silico* prediction of small molecule properties is widely used today in industry and academia. NMR spectra, in particular, are predicted by a variety of software packages. In this array of software options, two main approaches are used:

- *Database-based.* Compounds are compared against a database, the result is calculated using data for close structural relatives found in the dataset.
- *Regression-based.* An experimental database is used to calculate the parameters for non-linear regression. The chemical shift is calculated by a non-linear function of variables which describe characteristic features of the molecule of interest.

These two outlined approaches require different strategies for implementation and optimization. Database-based results are improved by acquiring larger databases and/or including user specific data into the calculation. Non-linear regression algorithms can be improved through the regression itself, or by improving the structural descriptors.

### *Regression*

Regression itself can be improved. The goal in this case is to make sure that the minimum found by the algorithm is a global minimum, the solution is stable, and available computer resources are enough to process databases of up to one million chemical shifts. Given these goals, two major classes of algorithms are the most popular:

- *Neural networks*—artificial neurons form a network which can be "taught" (regression parameters are adjusted).
- *Least squares algorithms*—starting with partial least squares (PLS).

Neural networks are much more popular these days and are used sometimes because of their effectiveness, sometimes just because they are an exciting area of research. We believe that these two methods are able to produce results of similar quality and the ultimate choice should not be made out of popularity, but after running benchmarks using actual data.

### *Structural Descriptors*

Independent variables used for regression can be extended to precisely describe a chemical structure, including not only chemical topology itself, but also 3D information and experimental conditions (most commonly, solvent).

However, care should be taken not to "over-describe" a structure—a description that is too detailed tends to include structural features that have minor impact on the observed chemical shift. This leads to increased prediction error.

## GOALS

In the current work, we focused on two areas:

- Validation and improvement of the chemical descriptor schemes, with a special emphasis on the level of detail necessary and sufficient for inclusion into the description.
- Comparison of partial least squares and neural network methods. We tried to do our best to ensure that both methods are used in the optimal way. Unlike PLS, neural networks have a number of adjustable parameters, hence we ran a series of calculations separately to ensure that the set of parameters used for comparison with the PLS method was optimal.

## METHODS

Programming was performed in a Borland Delphi 5 environment using MTX libraries for the linear algebra calculations. Adjustment of the neural network parameters was in MATLAB with the neural networks package.

### *Training Database*

The three most important factors in choosing the training database are size, diversity, and quality. This ensures that the algorithms derived are applicable. Fortunately, these requirements are met by using the same database used in ACD/Labs' commercially available products. This database contains approximately 2,160,000 $^{13}$C and 1,440,000 $^{1}$H chemical shifts.

Equally as important is the selection of the test dataset. It should be as independent as possible from the training dataset. To avoid overlaps with the training data, 11,000 new compounds (150,000 chemical shifts) described in the literature in 2005–2006 were chosen as the test dataset. This avoided overlap since the training database included only compounds that were described in the literature before 2005.

### *Structure Description*

Traditionally, chemical structures are described in terms of separate atoms.

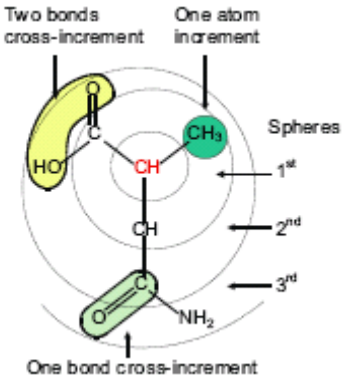


*Figure 1—The Structure Description consists of a central atom (for which the prediction is made) and substituents which are located at different distances from the central atom. All atoms separated from the center by n covalent bonds are called the n-th sphere.*

In total, 66 atom types were used. Generally, atoms are classified based on element number, number of attached hydrogens, hybridization, and valency. Additional descriptors were used to take into account conjugation, stereochemistry, and solvent effects.

For pairs of atoms separated by a few covalent bonds, separate inputs were provided (aka "cross-increments" or "correction factors"). The number and nature of atom types, number of spheres, and number of cross-increments were all subject to optimization.

## PARTIAL LEAST SQUARES

Calculation of the regression coefficients for PLS is generally a faster procedure than neural net training. Moreover, regression is a "deterministic" procedure which leads to the same result every time, unlike neural net learning which relies on stochastically chosen initial weights. Thus regression was chosen as the main method for developing a structure description model.

The first attempt with linear regression gave a significant prediction error, even for the large number of spheres that were used. We concluded that linear regression was not appropriate and non-linear regression should be used. This was accomplished by adding "cross-increments" as a type of correction factor to make the regression model non-linear.

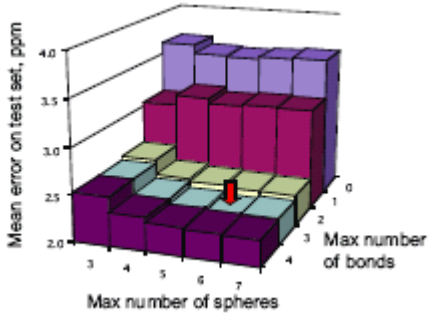


*Figure 2—The dependence of prediction quality on the number of spheres used. The best result is achieved with a total of six spheres taken into account, with cross-increments added for atoms located up to the third sphere.*

## NEURAL NETWORKS

Typically, neural networks are considered to be superior to older least squares-based methods. It is commonly suggested that a trained neuron, given input variables $x_1,\ldots x_n$ automatically takes into account all possible non-linear combinations of variables, such as $x_1x_2$, $x_1x_2x_3$, etc. Given this argument, a neural network should not require cross-increments. We tested this hypothesis using our test data set of approximately 400,000 $^{13}$C chemical shifts.

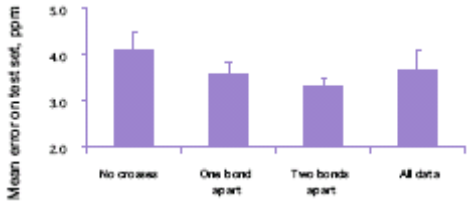


*Figure 3—Mean Error for Neural Net test with Cross-Increments. Three spheres were taken into account; cross-increments were either absent or added for atoms either one or two bonds apart.*

The data clearly show that the neural network still requires cross-increments to be explicitly included into the structure description.

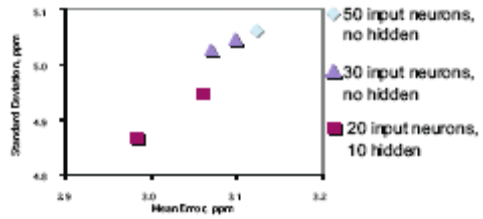


*Figure 4—Mean Error vs. Standard Deviation for different Neural Network topologies. The way inputs were normalized and transfer functions used was varied (not reflected on the graph).*

## COMPARING APPROACHES

After performance of both the neural network and least squares-based methods was optimized using truncated test datasets, the whole database was processed. PLS proved to be much faster than neural networks while maintaining similar performance. Both methods were also compared with the database-based method.

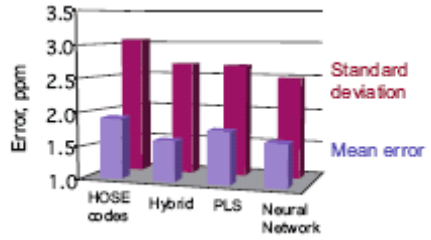


*Figure 5—$^{13}C$ prediction result.*

The best results for $^{13}C$ chemical shift prediction are shown above. The HOSE code approach refers to the database method. The 'Hybrid' method first tries to find a structure in the database and if no close analogs are found, it employs increments-based neural networks and/or PLS. The neural network used in the calculation had 100 neurons in the input layer, and 25 and 5 neurons in two hidden layers; the network was provided with cross-increments up to the 3rd sphere.

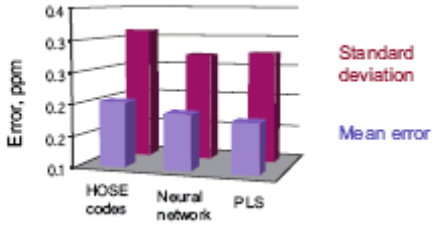


*Figure 6—$^{1}H$ prediction result.*

The best results for proton chemical shift predictions are shown. The best neural network had 30 input neurons and no hidden layer, six spheres were used for structure description and cross-increments were used up to the third sphere for atoms separated by no more than one bond.

## CONCLUSIONS

In the work presented here, two different approaches to chemical shift prediction have been systematically compared. That of least squares-based regression and neural networks.

From this work, we conclude that:

- The usage of neural networks does not help to reduce the number of atomic/cross-increments needed for accurate chemical shift prediction. Neural networks also do NOT achieve accurate results without cross-increments.
- The quality of the best possible results obtained with an optimized least squares scheme and neural network is approximately the same. The mean error can be as low as 1.5 ppm for $^{13}C$ and 0.2 ppm for 1H chemical shift prediction.
- Both PLS and neural networks can be coupled with a database search, resulting in an even more effective hybrid method.