Prediction of ^1H NMR Coupling Constants with Associative Neural Networks Trained for Chemical Shifts

Yuri Binev, Maria M. B. Marques, and João Aires-de-Sousa*

REQUIMTE and CQFB, Departamento de Química, Faculdade de Ciências e Tecnologia,
Universidade Nova de Lisboa, 2829-516 Caparica, Portugal

Received May 19, 2007

Fast accurate predictions of ^1H NMR spectra of organic compounds play an important role in structure validation, automatic structure elucidation, or calibration of chemometric methods. The SPINUS program is a feed-forward neural network (FFNN) system developed over the last 8 years for the prediction of ^1H NMR properties from the molecular structure. It was trained using a series of empirical proton descriptors. Ensembles of FFNNs were incorporated into Associative Neural Networks (ASNN), which correct a prediction on the basis of the observed errors for the *k* nearest neighbors in an additional memory. Here we show a procedure to estimate coupling constants with the ASNNs trained for chemical shifts—a second memory is linked consisting of coupled protons and their experimental coupling constants. An ASNN finds the pairs of coupled protons most similar to a query, and these are used to estimate coupling constants. Using a diverse general data set of 618 coupling constants, mean absolute errors of 0.6−0.8 Hz could be achieved in different experiments. A Web interface for ^1H NMR full-spectrum prediction is available at http://www.dq.fct.unl.pt/spinus.

INTRODUCTION

Prediction of NMR spectra from the molecular structure of organic compounds has a number of applications. Structure validation of large numbers of compounds prepared in parallel syntheses requires the comparison of the experimental spectra with simulated spectra for the target structures.[1] In automatic structure elucidation, the search for a possible structure is often guided by the similarity between the predicted spectra for the candidate structures and the experimental spectrum.[2] Identification of individual signals in NMR spectra of complex mixtures can be assisted by structure-based predictions. Accurate predictions of NMR spectra enable the calibration of spectra-property relationships and chemometric models, exclusively on the basis of simulated data, which can then be applied to experimental data.[3,4]

Quantum chemical methods can predict NMR properties,[5,6] although their computational requirements preclude application to large data sets of compounds. Furthermore, automation of the procedures is often not straightforward. On the other hand, data-driven chemoinformatics approaches have been quite successful within the domain of stable organic compounds for which abundant experimental data are available.[7,8] One family of methods relies on matching a query substructure against a database of substructures with their associated NMR data. HOSE codes,[9] or derivatives, are usually employed. A different strategy involves the training of QSPR methods, most notably neural networks,[10−13] to predict chemical shifts from topological, physicochemical, or geometrical features describing the environment of a nucleus. The two approaches are not mutually exclusive. For example, recent versions of ACD software[14] combine their typical database-centered method with neural networks, while in the SPINUS system the neural network predictions are corrected with a database of experimental data that can be modified without retraining the networks.[15] In a different type of predictions provided by a number of software packages, the estimated chemical shift is computed using tables of additive constants for a list of functional groups or fragments bonded to specific substructures.[16,17] A mixed approach has been proposed by Abraham and co-workers combining semiempirical calculations, parametrized group contributions, and neural networks for the prediction of ^1H NMR chemical shifts.[18] Although primarily developed for the prediction of ^{13}C and ^1H NMR chemical shifts, some of the cited methods have been extended to other nuclei.[14,19]

Currently, Web services powered by chemoinformatics systems are available that provide quick predictions of chemical shifts on input of a molecular structure.[20−22]

Besides chemical shifts, however, a typical ^1H NMR spectrum—the most widely used NMR spectrum—is strongly influenced by coupling constants that define the pattern of signals. Coupling constants are thus required for the simulation of entire spectra. And yet, general empirical methods for the estimation of coupling constants, other than Karplus equation-based,[18] and their integration into full-spectrum predictions, have been rare.[14]

Here we report a method to estimate ^1H NMR coupling constants on top of Associative Neural Networks (ASNN)[23] trained to predict *chemical shifts*. The SPINUS package is a neural network system developed over the last 8 years for the prediction of ^1H NMR properties from the molecular structure.[10−12,15] Ensembles of feed-forward neural networks (FFNN) were trained with databases of experimental NMR chemical shifts, and protons encoded by empirical descrip-

* Corresponding author phone: (+351) 21 2948300; fax: (+351) 21 2948550; e-mail: jas@fct.unl.pt.

10.1021/ci700172n CCC: $37.00 © 2007 American Chemical Society
Published on Web 10/23/2007

tors, and were incorporated into ASNN. These apply a correction to the prediction of a chemical shift obtained by the FFNNs. The correction is based on the observed errors for the *k* nearest neighbors in an additional memory of experimental chemical shifts.

To estimate coupling constants with the same neural networks (trained for chemical shifts) a memory of coupled protons and their experimental coupling constants, *J*, is now used in addition to the memory of chemical shifts. The memory of coupling constants is searched with the ensemble of FFNNs, to find the pairs of coupled protons giving the most similar output profile to a query, and the retrieved values of *J* are used to estimate the coupling constant.

This paper reports the implementation and validation of the approach and illustrates its application to a new organic compound.

METHODS

Data. A data set of 18 122 chemical shifts was used that correspond to 5230 protons bonded to aromatic systems, 872 protons bonded to π nonaromatic systems, 5320 protons bonded to nonrigid aliphatic substructures, and 6700 protons bonded to rigid aliphatic substructures. These include data manually collected in our lab from the literature and proprietary data made available by Molecular Networks GmbH (Erlangen, Germany). They were retrieved from spectra measured not only mostly in $CDCl_3$ but also in DMSO-d_6 and in a very few cases in D_2O, pyridine-d_5, CD_2Cl_2, CD_3OD, and benzene-d_6. A data set of 618 coupling constants was manually retrieved from the literature in our lab or taken from the above-mentioned proprietary collection. Reporting of coupling constants in the organic chemistry literature is usually associated with a high degree of uncertainty. Furthermore, manual retrieval of large numbers of coupling constants from the literature is an error-prone demanding task that often requires interpretation of full papers. The database of retrieved coupling constants was curated by inspecting the outliers of preliminary predictive models, and cross-checking of ca. $^1/_5$ of the database by a second chemist revealed consistency for more than 90% of the cases, with most of the differences not significantly affecting the magnitude of the coupling constant values.

Classification of Hydrogen Atoms. As in our previous studies,[12] four different classes of protons were defined (aromatic, nonaromatic π, rigid aliphatic, and nonrigid aliphatic), and protons belonging to each of them were treated separately. This procedure allowed for the use of more specific descriptors for each class. Protons were classified as (a) “aromatic” when they are bonded to an aromatic system; (b) “nonaromatic π” when they are bonded to a nonaromatic π system; (c) “rigid aliphatic” when a nonrotatable bond is identified in the second sphere of bonds centered on the proton; and (d) “nonrigid aliphatic” when not included in previous classes. A bond was defined as nonrotatable if it belongs to a ring.

Descriptors of Hydrogen Atoms. The same proton descriptors were used as in previous studies, and they are listed in the Supporting Information. These include physicochemical, geometrical, and topological descriptors. Physicochemical descriptors were based on empirical values calculated by the software package PETRA[24] (version 3.2) comprising a variety of published methods[25,26] for the hydrogen atoms and for their neighborhood. Examples of physicochemical descriptors used in this study are the partial atomic charge of the proton, effective polarizability of the proton, average of partial atomic charges of atoms in the second sphere, maximum partial atomic charges of atoms in the second sphere, minimum effective polarizability of atoms in the second sphere, and average of σ electronegativities of atoms in the second sphere. Geometrical descriptors were based on the 3D molecular structure generated by the CORINA software package.[27−29] The global 3D environment of a proton belonging to a rigid substructure was represented by a radial distribution function (RDF).[30] Simple topological descriptors were based on the analysis of the connection table and include, e.g., count of carbon atoms in the second sphere centered on the proton, count of oxygen atoms in the third sphere, or count of atoms in the second sphere that belong to an aromatic system. A topological radial distribution function was also used, where the sum of bond lengths on the shortest possible path between the proton and another atom is used instead of the 3D interatomic distances. All in all, 92 descriptors were initially calculated for aromatic protons, 119 for nonrigid aliphatic protons, 110 for nonaromatic π protons, and 174 for rigid aliphatic protons.

Associative Neural Networks (ASNN) for the Prediction of Chemical Shifts.[23] ASNN integrate an ensemble of feed-forward neural networks (FFNN) with a memory of experimental data. The ensemble consists of 75 independently trained FFNNs, which contribute to a single prediction—the average of the outputs from the 75 individual networks. The ASNN program[31] was used and employed the Levenberg—Marquardt algorithm to train fully connected FFNNs with an input layer (including a bias equal to 1), one hidden layer (also including a bias equal to 1), and one output neuron. The logistic activation function was used: $1/(1+\exp(-x))$. Prior to the training of each network, the program randomly divided the training set into a new cross-validation set and a reduced training set with approximately the same size. In this way, each network of the ensemble was trained with its own training and cross-validation sets, in addition to being seeded with its own (random) initial weights. The number of hidden neurons in each network had been optimized and is as follows: 9 hidden neurons for aromatic protons, 6 hidden neurons for nonaromatic π protons, 10 hidden neurons for rigid aliphatic protons, and 10 hidden neurons for nonrigid aliphatic protons.

The ensemble of FFNNs is combined with a memory into a so-called associative neural network (ASNN).[23] For the prediction of chemical shifts, the memory consists of a list of protons, represented by their output profiles, their uncorrected predictions (the average of its output profile), and the corresponding experimental chemical shifts. The ASNN scheme is employed for composing a prediction of the chemical shift from (a) the outputs produced by the ensemble of NNs and (b) the data in the memory. When a query proton is submitted to an ASNN, the following procedure takes place to obtain a final prediction:

1. The descriptors of the query proton are presented to the ensemble, and a number of output values are obtained from the different NNs of the ensemble—the output profile of the query proton.

2. The average of the values in the output profile is calculated. This is the uncorrected prediction of the chemical shift for the query proton.

3. The memory is searched to find the *k* nearest neighbors of the query proton. The search is performed in the output space, i.e., the nearest neighbors are the protons with the most similar output profiles (stored) to the query proton (calculated in step 1). Similarity is here defined as the Spearman correlation coefficient between output profiles.

4. The uncorrected predictions for the KNN protons (stored) are compared with the experimental chemical shifts. The mean error is computed, weighted with the correlation coefficients of the *k* neighbors.

5. The mean error computed in step 4 is added to the uncorrected prediction of the query proton (computed in step 2) to yield the corrected prediction of the chemical shift for the query proton.

The parameter *k* had been optimized for each class of protons and is as follows: 3 for aromatic protons, 15 for nonaromatic π and rigid aliphatic protons, and 12 for nonrigid aliphatic protons.

Selection of Variables. Proton descriptors were selected by stepwise removal of the descriptors with the least impact in the FFNNs outputs.[32] After the ensemble of FFNNs was trained using all the descriptors, one descriptor was scrambled along the training set, and the transformed data were resubmitted to the trained FFNNs. The outputs were compared to the original outputs. The experiment was performed for all the descriptors, one by one, to find the one whose scrambling resulted in the most similar outputs to the original (measured by the sum of squared differences). The descriptor with the least impact in the output was removed, and the procedure was repeated iteratively to remove descriptors until optimum prediction accuracy was observed.

Prediction of Coupling Constants. A data set of 618 experimental coupling constants was built, and subsets of it were used as the memory of the ASNN for the prediction of coupling constants in different experiments. Part of the data set was used as the memory, and the remaining cases were predicted (test set). In one experiment, the memory consisted of 451 cases taken from random molecules of the data set, and predictions were obtained for a test set of 167 cases from the remaining molecules. Another experiment employed the leave-one-out procedure. In a third experiment, a test set consisting of all the cases (100) from a single source was extracted from the data set.

In the ASNN memory of coupling constants, each entry consists of the experimental value of *J* (in Hz) and the output profile for the two corresponding protons. The output profile of one proton is the list of the 75 output values produced by the ensemble of 75 FFNNs *as prediction of its chemical shift*. More precisely, the output profile is defined as the ranks of the 75 output values. The entries in the memory were categorized into classes according to the types of protons and the number of bonds between them: (a) aromatic−aromatic (at *ortho* or *meta* positions relative to each other), (b) $\pi-\pi$ (geminal, vicinal *cis*, and vicinal *trans*), (c) aliphatic−aliphatic (vicinal), (d) rigid−rigid (geminal), (e) π−rigid (vicinal), (f) π−aliphatic (vicinal), and (g) rigid−aliphatic (vicinal).

To get a prediction for a new pair of coupled protons, output profiles are generated for the two protons, and the memory is searched to retrieve the most similar cases within the same category. As mentioned above, similarity is here defined in terms of the Spearman correlation coefficient between output profiles (average of correlation coefficient for the two protons of the pair). The coupling constant is estimated as the average of the experimental coupling constants for the *k* most similar pairs weighted (or not) by the Spearman correlation coefficients.

For coupling constants between two vicinal protons belonging to rigid substructures, conformation must be taken into account, which could be done in principle by the geometrical descriptors defined in SPINUS. However, a lack of enough data with reliably assigned stereochemistry prevented the investigation of ASNN-based predictions. Instead, the Karplus equation[33] was implemented in the current version of SPINUS to estimate vicinal proton−proton coupling constants in rigid substructures:

$$J = A \cdot \cos^2(\phi) + B \cdot \cos(\phi) + C$$

with

$$A = 7.76,\ B = -1.10,\ \text{and}\ C = 1.4$$

Full-Spectra Generation in the Current Version of SPINUS. Chemical shifts and coupling constants predicted from the molecular structure were incorporated for the simulation of full-spectra, approximated as first-order spectra. In the current version of SPINUS,[22] a spectrum is simulated with 20 000 points covering chemical shifts from 0 to 10 ppm. Coupling constants are translated into patterns of peaks considering a 400 MHz spectrometer. Peaks are generated with the Lorentzian distribution formula and half-width of 1 Hz and integrate proportionally to the number of corresponding protons.

Synthesis of (*S*)-2-Acetylamino-3-(1-but-3-enyl-1*H*-indol-3-yl)propionic Acid Methyl Ester (1). The ability of the spectrum prediction method to be applied to a previously unknown compound was illustrated with compound **1**, newly synthesized as follows. A solution of *N*-acetyl-*L*-tryptophan methyl ester[34] (150 mg, 0.57 mmol) in dry DMF (1.5 mL) was treated portion wise with NaH (60% dispersion, 57.7 mg, 2.40 mmol), and 4-bromo-1-butene (145 μL, 0.8 mmol) was added dropwise. The resulting mixture was stirred under argon atmosphere in an ice bath for 45 min, allowed to warm to room temperature, and stirred until the reaction was complete [16 h, TLC (Et_2O) control]. The resulting mixture was diluted with EtOAc (10 mL) and washed repeatedly with a saturated solution of sodium bicarbonate, followed by brine. Usual workup led to an oil which was purified by column chromatography on silica (Et_2O) to furnish compound **1** (164 mg, 90%) as a colorless oil: $[\alpha]_D^{23}$+44.5 (*c* 1.01, Et_2O); IR (film, cm^{-1}) 3283, 1744 (ester C=O), 1651 (amide C=O); 1H-NMR (400 MHz, $CDCl_3$) δ_H 7.50 (1H, d, $J = 7.8$ Hz, Ar−H), 7.31 (1H, d, $J = 8.2$ Hz, Ar−H), 7.20 (1H, t, $J = 7.4$ Hz, Ar−H), 7.10 (1H, t, $J = 7.4$ Hz, Ar−H), 6.85 (1H, s, H-28), 5.98 (1H, d, $J = 7.2$ Hz, NH, exchangeable with D_2O), 5.79−5.69 (1H, m, $CH{=}CH_2$), 5.08−5.03 (2H, m, $CH{=}CH_2$), 4.96−4.91 (1H, m, $CHCO_2Me$), 4.13 (2H, t, $J = 7.0$ Hz, NCH_2), 3.70 (3H, s, OMe), 3.30 (2H, m, CH_2CHCO_2Me), 2.54 (2H, q, $J = 6.9$, 6.9 Hz, NCH_2CH_2), 1.96 (3H, s, NCOMe); ^{13}C NMR (100 MHz, $CDCl_3$) δ_C 172.34, 169.70, 136.06.134.58, 128.37, 126.36, 121.72,

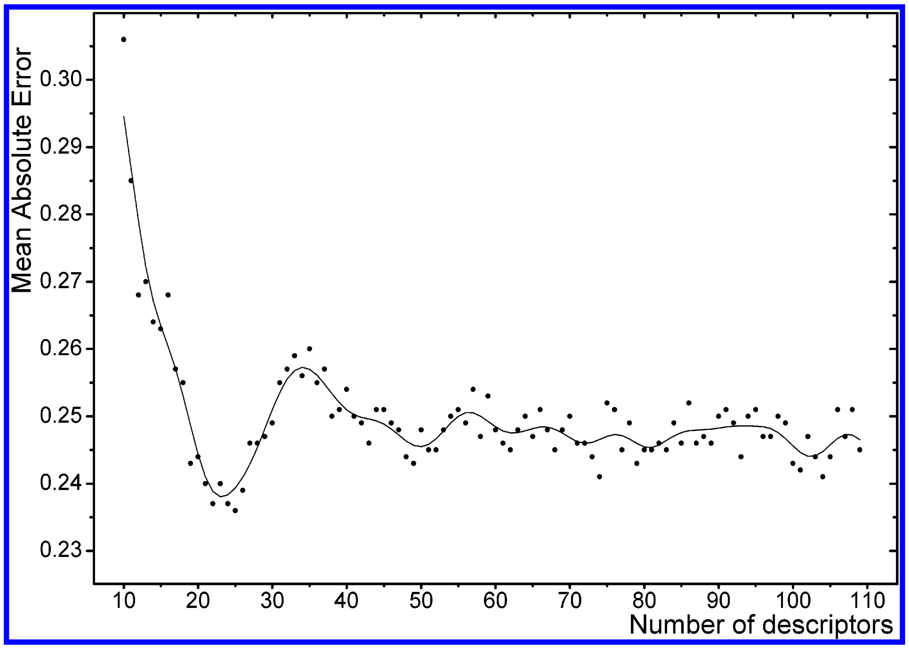

Figure 1. Influence of the number of selected descriptors on the prediction accuracy of FFNN ensembles trained for chemical shifts (protons of the π class).

119.23, 118.75, 117.45, 109.49, 108.48, 53.13, 52.29, 45.86, 34.44, 27.49, 23.18; EI HRMS *m*/*e* calcd for $C_{18}H_{22}N_2O_3$ (M^+) 314.16304, found 314.16301.

Optical rotations were determined in a Perkin−Elmer 241 MC polarimeter at room temperature, and $[\alpha]_D$ values are given in 10^{-1} deg cm^2 g^{-1}. Ordinary mass spectra were recorded on a Fisons TRIO 2000 or AEI MS-9 spectrometers. ^1H- and ^{13}C NMR spectra were recorded in $CDCl_3$ on a Bruker ARX 400 spectrometer. Chemical shifts are reported relative to tetramethylsilane as the internal reference (δ_H 0.00) for ^1H NMR spectra and to $CDCl_3$ (δ_C 77.00) for ^{13}C NMR spectra. High-resolution mass spectra were recorded on an AutoSpecQ spectrometer. IR spectra were run on a FT Perkin−Elmer 683 instrument, with absorption frequencies expressed in reciprocal centimeters. Thin-layer chromatography was performed on Merck silica gel 60 F_{254} plates and PTLC on 0.5 mm thick plates. Column chromatography was carried out on Merck silica gel 60 (70−230 mesh). Usual workup implies drying the water- or brine-washed organic extracts over anhydrous sodium sulfate or magnesium sulfate, followed by filtration and evaporation of the solvent from the filtrate under reduced pressure. Anhydrous solvents were dried as described[35] and freshly distilled.

RESULTS AND DISCUSSION

Training of ASNN for the Prediction of ^1H NMR Chemical Shifts. From the initial data set of 18 122 chemical shifts, reduced balanced data sets were designed for training the neural networks. A subset was chosen for each of the four types of protons. In order to select protons covering the whole diversity available in the data and avoid over-representation of protons with specific features, a Kohonen self-organizing map was trained with all the protons of one type, and then one proton was randomly taken from each neuron of the map into the subset. The procedure yielded training sets with 1186 aromatic protons, 288 π protons, 999 aliphatic protons, and 1201 rigid protons. The remaining protons were used as the additional memory of the ASNN.

Table 1. Performance of the Chemical Shifts Models (Ensembles of FFNN) in Different Tests Using the Training Sets

	no. of cases	LOO (MAE, ppm)	reduced training set (MAE, ppm)	validation set (MAE, ppm)
aromatic	1186	0.240	0.211	0.237
π	288	0.262	0.183	0.256
aliphatic	999	0.186	0.155	0.183
rigid	1201	0.276	0.226	0.272

Selection of descriptors resulted in 41 descriptors for aromatic protons, 26 descriptors for π nonaromatic protons, 45 descriptors for aliphatic protons, and 63 descriptors for rigid protons. The selected descriptors are listed as Supporting Information. Figure 1 illustrates the variation of the mean absolute error (MAE) for the whole training set as a function of the number of selected descriptors, for the π protons.

Table 1 shows the quality of the predictions obtained by the optimized ensembles of FFNNs for the training set. Before training one FFNN, the training set is divided into a reduced training set and a cross-validation set. The training is stopped when the error for the validation set reaches a minimum (point S1). The results in Table 1 for the reduced training set and validation set were obtained with networks trained until point S1. With a validation set of *n* objects, *n* points S1 can be identified, each one based on the error for the validation set leaving one object out. The LOO results were obtained from predictions for the objects of the validation set, each one at his own point S1.[36] The high accuracy of the predictions was confirmed by application to an independent test set with 952 chemical shifts that had been used in our previous study—Table 2. It was observed that the new training sets and the new procedure for selection of descriptors enabled the FFNNs to yield better predictions (MAE=0.24 ppm against 0.29 ppm before). ASNN improved the predictions for aliphatic and rigid protons but not for aromatic and π protons. Compared to our previous study, the global performance of ASNN was now slightly worse (MAE=0.23 ppm against 0.19 ppm), which can be explained

Table 2. Prediction Accuracy of the Chemical Shifts Models for an Independent Test Set

	no. of cases	ensemble FFNN (MAE, ppm)	ASNN (MAE, ppm)
aromatic	247	0.217	0.226
π	93	0.321	0.348
aliphatic	375	0.181	0.159
rigid	237	0.305	0.292
total	952	0.235	0.228

Table 3. Accuracy of Predicted Coupling Constants in the Leave-One-Out and Independent Test Set Experiments[a]

	LOO (excluding above threshold)		test set	
	no. of cases	MAE, Hz	no. of cases	MAE, Hz
aromatic−aromatic	155 (130)	0.38 (0.29)	45	0.44
$\pi-\pi$	117 (103)	0.84 (0.53)	21	0.82
aliphatic−aliphatic	149 (136)	0.41 (0.38)	46	0.54
geminal rigid	49 (30)	1.09 (0.76)	15	1.07
mixed	148 (141)	0.68 (0.57)	40	0.65
all	618 (540)	0.60 (0.46)	167	0.62

[a] In brackets are the results after removing the predictions with correlation coefficient of the most similar example less than 0.53.

by the completely different memory, now coming from a different source than the test set—the probability of highly similar structures in the memory and in the test set was now much less.

Prediction of Coupling Constants. The proposed methodology for the estimation of coupling constants was implemented and evaluated with a data set of 618 cases. Part of the data set was used as the memory, and the remaining cases were predicted (test set). In one experiment, the test set consisted of 167 cases that were left out; another experiment employed the leave-one-out procedure. The results are presented in Table 3 and Figure 2. Mean absolute errors of 0.62 and 0.60 Hz were obtained in the two tests, respectively. This approach critically depends on the availability of similar cases to a query, as it relies exclusively on a KNN search. Figure 3 shows how the MAE decreases by excluding predictions in which the most similar neighbor has a correlation coefficient less than a threshold. It was decided that a threshold of 0.53 was a reasonable compromise between the improvement of the accuracy and the number of cases covered. With such a threshold, the MAE improved from 0.60 to 0.46 Hz still encompassing 87% of the data set. Figure 2 shows that most of the outliers are excluded by this threshold.

The most apparent outliers in the plot of Figure 2 were inspected and revealed to be related to stereochemistry and arbitrary assignments. Three examples are illustrated. In two cases (experimental/predicted 4.5/11.2 Hz and 12.5/5.1 Hz), the coupling was between a π proton and a vicinal rigid proton. It happened that the coupling constants between the π proton and each of the (diastereotopic) rigid protons were significantly different (e.g., 4.5 and 10.5 Hz), but they had been arbitrarily assigned since there was no assignment in the original literature. The same happened with the other clear outlier (experimental 8.9 Hz, predicted 4.75 Hz), now involving coupling constants between a π proton and two vicinal diastereotopic aliphatic protons with quite different values of *J*.

As explained before, the ASNN-based search for the *k* most similar pairs to a query is restricted to the subset of the memory consisting of the coupled protons at the same interatomic distance and of the same types as the query. One could question whether an ASNN-based KNN search within such a restricted memory is really needed, or if simply averaging the *J* values of the entire memory subset (or using the median)—“static” predictions—would produce equally accurate estimations. We compared the MAE of the predictions obtained by the ASNN procedure with those obtained

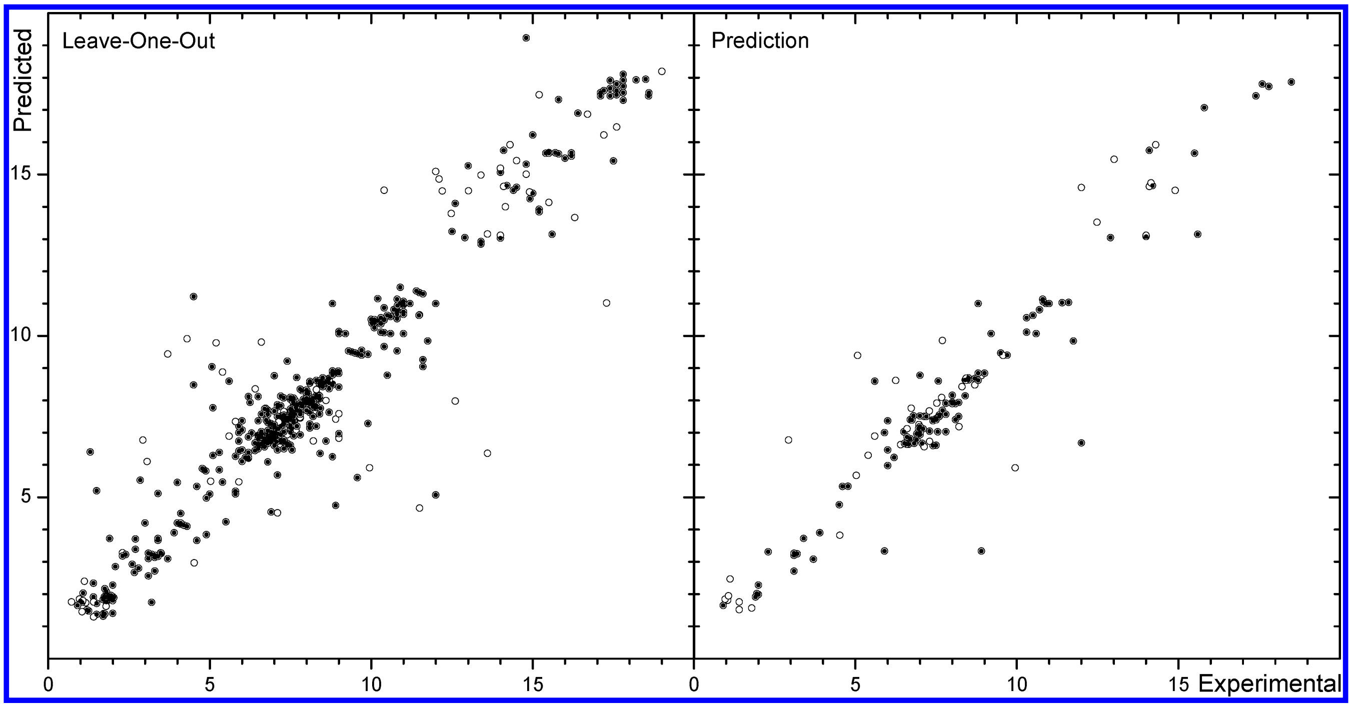

Figure 2. Comparison of the experimental and predicted coupling constants for the leave-one-out (left) and independent test set (right) experiments. The empty dots correspond to cases with similarity (correlation coefficient) lower than 0.53 to its most similar neighbor.

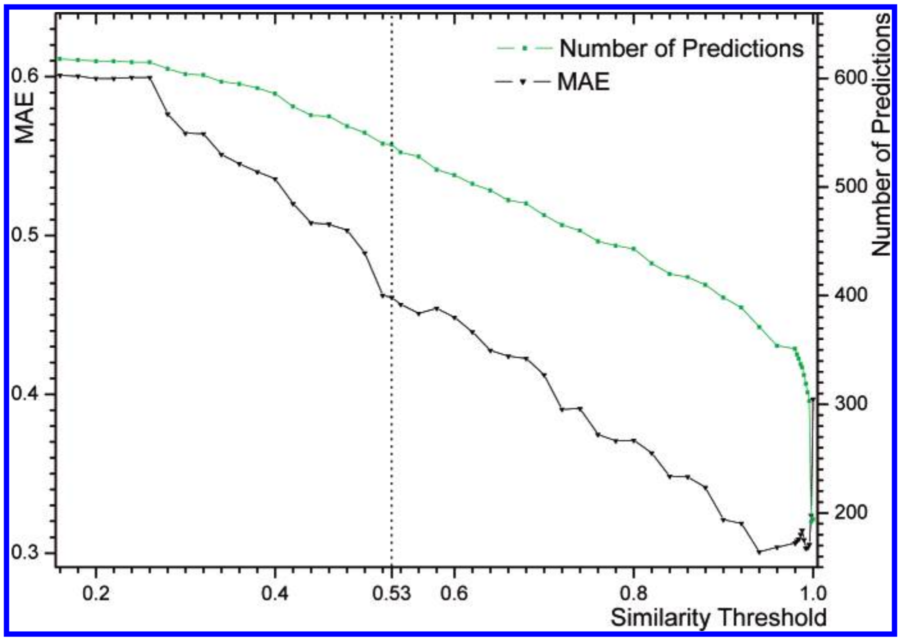

Figure 3. Exclusion of predictions when the correlation coefficient between the query and the most similar case in the memory is less than a threshold. The figure illustrates the impact of the similarity threshold both in the accuracy of predictions and in the number of predictions excluded.

Table 4. ASNN-Based Prediction of Coupling Constants in the LOO Experiment vs "Static" Predictions Based on the Average or Median of Coupling Constants within Homogeneous Subsets of the Memory

no. of cases	subset	av of *J*, Hz	median of *J*, Hz	av-based static predictions (MAE, Hz)	median-based static predictions (MAE, Hz)	ASNN predictions (MAE, Hz)
130	aromatic−aromatic (*ortho*)	7.6	8.1	1.17	1.07	0.38
25	aromatic−aromatic (*meta*)	1.7	1.8	0.42	0.40	0.33
149	aliphatic−aliphatic (vicinal)	6.9	7.0	0.78	0.77	0.41
33	$\pi-\pi$ (*trans*)	16.8	17.2	1.36	1.29	0.93
61	$\pi-\pi$ (*cis*)	10.1	10.5	1.64	1.55	0.81
18	$\pi-\pi$ (geminal)	3.0	2.0	1.98	1.67	0.93
49	rigid−rigid (geminal)	14.0	14.3	2.20	2.17	1.09
64	aliphatic−π	7.4	7.3	0.64	0.64	0.49

by the averages and medians of the *J* values, for the eight largest subsets of the memory—Table 4. The accuracy of the ASNN procedure is considerably higher than the "static" predictions simply based on defined types and distances for all the subsets, which supports the usefulness of the ASNN approach.

Potentially relevant is the parameter *k*, the number of nearest neighbors used for estimating a coupling constant. In the coupling constants experiments described so far, *k* was always set to 3, and the predicted coupling constant was calculated from the experimental *J* values weighted by the Spearman correlation coefficient between the corresponding coupled pairs and the query pair. After screening values of *k* from 1 to 7 and testing predictions by flat and weighted averages (Table 5), we conclude that weighted averages are generally advantageous, and the results are not significantly affected by the value of *k*. Results were obtained (a) by the LOO procedure within the memory of 451 cases, (b) for the independent prediction set (167 cases) on the basis of the memory (451 cases), and (c) by the LOO procedure within the whole data set (618 cases). The best LOO results using only the memory (451 cases) were obtained for $k = 3$ and weighted averages, and the same trend was observed for the independent prediction set. It was therefore decided to employ these parameters in the Web service.

Table 5. Influence of *k* (Number of Neighbors) and Weighting in the Prediction of Coupling Constants by the ASNN *k*-Nearest-Neighbors Approach (MAE in Hz)

	leave-one-out (memory only, 451 cases)		prediction set (167 cases)		leave-one-out (618 cases)	
k	flat	weighted	flat	weighted	flat	weighted
1	0.709	0.709	0.643	0.643	0.662	0.662
2	0.678	0.638	0.707	0.654	0.649	0.602
3	0.704	0.638	0.687	0.624	0.670	0.601
4	0.772	0.666	0.743	0.625	0.711	0.613
5	0.813	0.677	0.787	0.628	0.755	0.631
6	0.858	0.684	0.814	0.653	0.795	0.641
7	0.896	0.701	0.809	0.657	0.814	0.645

In the data set of coupling constants, many values are from groups of structures retrieved from the same paper. Within such groups, structures tend to be highly similar; therefore, many coupling constants in the data set have almost identical neighbors, a situation that can favorably affect the predictions, both in the LOO experiment and with the random test set. Also, in some cases, the three coupling constants involving the three protons of a methyl group were included, which affects the LOO experiments for the aliphatic protons. For a more realistic assessment of the predictive ability of the ASNN approach, a new experiment was performed using as a test set all the coupling constants that were taken from one of the sources, and all the remaining cases were used as

Table 6. Accuracy of the Predicted Coupling Constants in the Independent Test Set Taken from a Single Source[a]

no. of cases	subset	median of J in the memory	median-based "static" predictions (MAE, rms)	ASNN predictions (MAE, rms)
30 (10)	aromatic−aromatic (*ortho*)	8.2	1.48 (1.67) 2.38 (2.92)	0.63 (0.21) 0.86 (0.28)
16 (0)	aromatic−aromatic (*meta*)	2.0	0.56 (-) 0.70	0.60 (-) 0.78
24 (13)	aliphatic−aliphatic (vicinal)	7.1	0.95 (0.34) 1.43 (0.39)	0.91 (0.17) 1.59 (0.28)
9 (1)	rigid−rigid (geminal)	14.5	1.07 (-) 1.41	0.97 (-) 1.14
7 (6)	aliphatic−π	7.4	1.00 (1.55) 1.20 (1.88)	0.94 (0.93) 1.26 (1.32)
14 (0)	aliphatic−rigid	6.6	1.11 (-) 1.56	0.86 (-) 1.30
100 (30)	total	-	1.08 (1.01) 1.70 (1.93)	0.78 (0.33) 1.18 (0.65)

[a] Values are in Hz. In parentheses are the results after removing the cases with correlation coefficient of the most similar example in the memory less than 0.53.

the memory. The results are displayed in Table 6. As expected, the errors were slightly higher (MAE 0.78 Hz) than with the random test set but still lower than those obtained using "static" predictions (the medians of the values within classes of coupled protons in the memory). "Static" predictions performed respectably for classes in which almost all the J values vary within a short range (such as the geminal rigid−rigid, or the m-aromatic classes). For such classes, the ASNN cannot improve predictions much further, and significant correlation between experimental and predicted J values cannot be obtained within the class. Differently, for the o-aromatic class, where J values consistently vary over a range of more than 6 Hz (depending on the size of the ring, and its homo- or heteroaromatic nature) the ASNN procedure dramatically improves the predictions, and high correlations are observed within the class. And again, the accuracy of the predictions was significantly improved when similar cases to the query exist in the memory, as determined from the Spearman correlation coefficients of the output profiles—for the coupled pairs with a neighbor exhibiting a correlation coefficient above 0.53 (now 30% of the cases), the ASSN predictions were always superior to the "static" predictions, and the global MAE dropped to 0.33 Hz. A general trend comparable to Figure 3 was also observed for this test set—the MAE consistently decreased by increasing the similarity threshold for the exclusion of predictions, until stabilizing at a threshold of ca. 0.6.

Full-Spectra Simulation. The ability to predict coupling constants was incorporated into the Web service of SPINUS to generate a full-spectrum simulation on input of a molecular structure. The service is available at http://www.dq.fct.unl.pt/spinus. It is illustrated in Figure 4 with the application to compound **1**, a novel tryptophan derivative synthesized by one of us (M.M.B.M.). The coupling constants that could be measured in the experimental spectrum compare well with the predicted (Hz: $J_{24,25}$ 7.8 exp, 8.4 pred; $J_{25,26}$ 7.4 exp, 7.8 pred; $J_{26,27}$ 8.2 exp, 8.0 pred; $J_{39,41}$ 7.0 exp, 7.4 pred; $J_{40,41}$ 6.9 exp, 7.4 pred; $J_{41,43}$ 6.9 exp, 6.7 pred). The signal

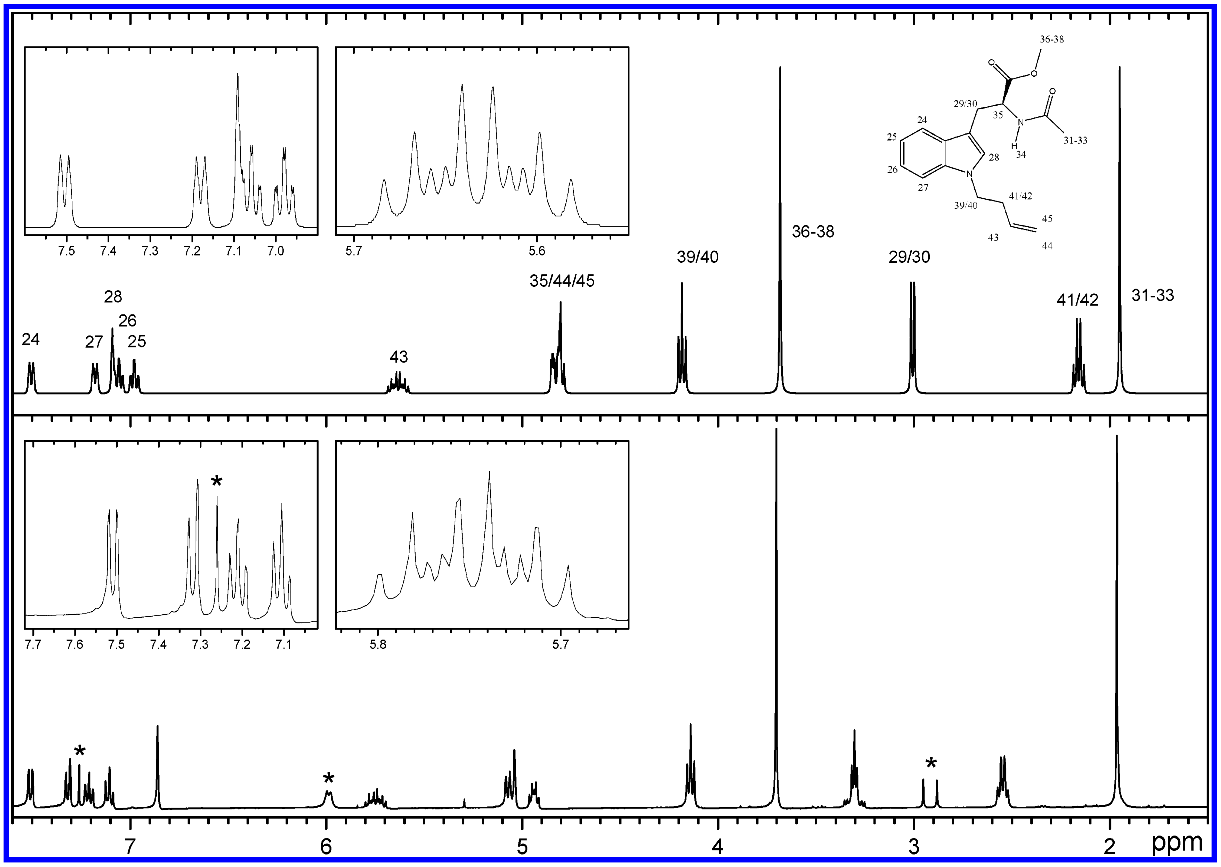

Figure 4. Experimental ^{1}H NMR spectrum of **1** (down) compared to the full spectrum predicted by SPINUS (up) for the same structure. (*) In the experimental spectrum, the signal at 7.26 ppm is from the solvent ($CDCl_3$), the signal at 5.98 ppm is from the exchangeable NH proton, and the peaks at 2.85−2.95 ppm are from residues of DMF.

corresponding to H-43 was split by coupling with protons H-41, H-42, H-44, and H-45. The predicted coupling constants are $J_{43,41} = J_{43,42} = 6.7$ Hz, $J_{43,44} = 10.3$ Hz, and $J_{43,45} = 17.1$ Hz. The zoom in Figure 4 shows that the prediction of the complex splitting pattern is excellent, and the distance between the two most external peaks in the experimental spectrum (41.0 Hz) is in accordance with the sum of the four coupling constants (40.8 Hz). The figure shows the ability of the method to generate full spectra with realistic signal shapes and good accordance between predicted and experimental data.

CONCLUSIONS

This study shows that a similarity between coupling constants of two pairs of protons can be generally inferred from a similarity between output profiles of FFNNS trained for chemical shifts. This fact enables associative neural networks trained for chemical shifts to accurately estimate coupling constants when available experimental data are perceived as similar to the coupled protons to predict. Having ASNNs trained for the prediction of chemical shifts, this approach represents an easy way to run database searches for the prediction of coupling constants without the implementation of substructure search or structural codes.

Particularly good results were obtained for the category of *o*-aromatic protons, even in the absence of highly similar available data.

Notwithstanding, typical ("static") coupling constants for defined classes of coupling pairs provided good estimations of coupling constants for classes in which *J* values vary over a short range. In such cases, and when no data similar to the query are available in the memory, "static" predictions can reach an accuracy comparable to the ASNN predictions.

The good results obtained by ASNN with a relatively small database indicate that general accurate predictions can be expected if the system is mounted on more data. Outliers were related to limitations of the literature sources and emphasize the requirement for high quality data. This work illustrates the ability of ASNNs trained for one property to transfer the acquired knowledge to another related property.

ACKNOWLEDGMENT

The authors acknowledge Molecular Networks GmbH (Erlangen, Germany) for access to experimental NMR data as well as for sharing CORINA and PETRA software packages. Y.B. acknowledges Fundação para a Ciência e Tecnologia (Lisbon, Portugal) for a postdoctoral grant under the POCTI program (SFRH/BPD/7162/2001). Chemical Concepts GmbH is acknowledged for providing the ^{1}H NMR experimental data used to confirm the accuracy of predicted chemical shifts. The authors thank Diogo Latino, Gonçalo Carrera, Qing-You Zhang, and Sunil Gupta for their assistance in the retrieval of NMR data from the literature. M.M.B.M. thanks Luz Fernandes of REQUIMTE Analytical Services Laboratories (Caparica, Portugal) for mass spectra analyses.

Supporting Information Available: Initial pool of proton descriptors, selected descriptors used as input to the final neural networks, and spectroscopic data for compound **1**. This material is available free of charge via the Internet at http://pubs.acs.org.

REFERENCES AND NOTES

(1) Golotvin, S. S.; Vodopianov, E.; Lefebvre, E. B.; Williams, A. J.; Spitzer, T. D. Automated structure verification based on 1H NMR Prediction. *Magn. Reson. Chem.* **2006**, *44*, 524−538.

(2) Meiler, J.; Will, M. Genius: A Genetic Algorithm for Automated Structure Elucidation from 13C NMR Spectra. *J. Am. Chem. Soc.* **2002**, *124*, 1868−1870.

(3) Asikainen, A.; Ruuskanen, J.; Tuppurainen, K. Spectroscopic QSAR methods and self-organizing molecular field analysis for relating molecular structure and estrogenic activity. *J. Chem. Inf. Comput. Sci.* **2003**, *43*, 1974−1981.

(4) Latino, D.; Aires-de-Sousa, A. R. S. Linking Databases of Chemical Reactions to NMR Data: an Exploration of 1H NMR-Based Reaction Classification. *Anal. Chem.* **2007**, *79*, 854−862.

(5) Heine, T.; Corminboeuf, C.; Seifert, G. The Magnetic Shielding Function of Molecules and Pi-Electron Delocalization. *Chem. Rev.* **2005**, *105*, 3889−3910.

(6) Rychnovsky, S. D. Predicting NMR Spectra by Computational Methods: Structure Revision of Hexacyclinol. *Org. Lett.* **2006**, *8*, 2895−2898.

(7) Perez, M.; Peakman, T. M.; Alex, A.; Higginson, P. D.; Mitchell, J. C.; Snowden, M. J.; Morao, I. Accuracy vs Time Dilemma on the Prediction of NMR Chemical Shifts: A Case Study (Chloropyrimidines). *J. Org. Chem.* **2006**, *71*, 3103−3110.

(8) Steinbeck, C.;In *Handbook of Chemoinformatics*; Gasteiger, J., Engel, T., Eds.; Wiley-VCH: 2003; Vol. 3, Chapter 2.2, pp 1368−1377.

(9) Bremser W. HOSE - A Novel Substructure Code. *Anal. Chim. Acta* **1978**, *103*, 355−365.

(10) Aires-de-Sousa, J.; Hemmer, M.; Gasteiger, J. Prediction of H-1 NMR chemical shifts using neural networks. *Anal. Chem.* **2002**, *74*, 80−90.

(11) Binev, Y.; Corvo, M.; Aires-de-Sousa, J. The impact of available experimental data on the prediction of 1H NMR chemical shifts by neural networks. *J. Chem. Inf. Comput. Sci.* **2004**, *44*, 946−949.

(12) Binev, Y.; Aires-de-Sousa, J. Structure-based predictions of 1H NMR chemical shifts using feed-forward neural networks. *J. Chem. Inf. Comput. Sci*. **2004**, *44*, 40−45.

(13) Meiler, J.; Meusinger, R.; Will, M. Fast determination of C-13 NMR chemical shifts using artificial neural networks. *J. Chem. Inf. Comput. Sci.* **2000**, *40*, 1169−1176.

(14) Advanced Chemistry Development, Inc. http://www.acdlabs.com (accessed Aug 9, 2007).

(15) Da Costa, F. B.; Binev, Y.; Gasteiger, J.; Aires-de-Sousa, J. Structure-based predictions of 1H NMR chemical shifts of sesquiterpene lactones using neural networks. *Tetrahedron Lett.* **2004**, *45*, 6931−6935.

(16) Pretsch, E.; Fürst, A.; Robien, W. Parameter set for the prediction of the ^{13}C-NMR chemical shifts of sp^{2}- and sp-hybridized carbon atoms in organic compounds. *Anal. Chim. Acta* **1991**, *248*, 415−428.

(17) Schaller, R. B.; Arnold, C.; Pretsch, E. New parameters for predicting ^{1}H NMR chemical shifts of protons attached to carbon atoms. *Anal. Chim. Acta* **1995**, *312*, 95−105.

(18) Abraham, R. J.; Mobli, M. The prediction of 1H NMR chemical shifts in organic compounds. *Spectrosc. Eur.* **2004**, *4*, 16−22. http://www.spectroscopyeurope.com/NMR_16_4.pdf (accessed Aug 9, 2007).

(19) Gabano, E.; Marengo, E.; Bobba, M.; Robotti, E.; Cassino, C.; Botta, M.; Osella, D. ^{195}Pt NMR spectroscopy: A chemometric approach. *Coord. Chem. Rev.* **2006**, *250*, 2158−2174.

(20) Loss, A.; Stenutz1, R.; Schwarzer, E.; von der Lieth, G.-W. GlyNest and CASPER: two independent approaches to estimate 1H and 13C NMR shifts of glycans available through a common web-interface. *Nucleic Acids Res.* **2006**, *34*, W733−W737.

(21) NMRShiftDB - open nmr database on the web. http://www.nmrshiftdb.org (accessed Aug 9, 2007).

(22) SPINUS-WEB: Prediction of NMR spectra. http://www.dq.fct.unl.pt/spinus (accessed Aug 5, 2007).

(23) Tetko, I. V. Neural Network Studies. 4. Introduction to Associative Neural Networks. *J. Chem. Inf. Comput. Sci.* **2002**, *42*, 717−728.

(24) PETRA can be tested on the Web site http://www2.chemie.uni-erlangen.de (accessed Aug 5, 2007) and is developed by Molecular Networks GmbH. http://www.mol-net.de (accessed Aug 9, 2007).

(25) Gasteiger, J. In *Physical Property Prediction in Organic Chemistry*; Jochum, C., Hicks, M. G., Sunkel, J., Eds.; Springer-Verlag: Heidelberg, Germany, 1988; pp 119−138.

(26) Gasteiger, J.; Marsili, M. Iterative Partial Equalization of Orbital Electronegativity - A Rapid Access to Atomic Charges. *Tetrahedron* **1980**, *36*, 3219−3228.

(27) Sadowski, J.; Gasteiger, J. From Atoms and Bonds to Three-Dimensional Atomic Coordinates: Automatic Model Builders. *Chem. Rev.* **1993**, *93*, 2567−2581.

(28) Gasteiger, J.; Rudolph, C.; Sadowski, J. Automatic Generation of 3D-Atomic Coordinates for Organic Molecules. *Tetrahedron Comput. Method.* **1992**, *3*, 537−547.

(29) Sadowski, J.; Gasteiger, J.; Klebe, G. Comparison of Automatic Three-Dimensional Model Builders Using 639 X-Ray Structures. *J. Chem. Inf. Comput. Sci.* **1994**, *34*, 1000−1008.

(30) Hemmer, M. C.; Steinhauer, V.; Gasteiger, J. The Prediction of the 3D Structure of Organic Molecules from Their Infrared Spectra. *J. Vibrat. Spectrosc.* **1999**, *19*, 151−164.

(31) (a) Tetko, I. V.; Gasteiger, J.; Todeschini, R.; Mauri, A.; Livingstone, D.; Ertl, P.; Palyulin, V. A.; Radchenko, E. V.; Zefirov, N. S.; Makarenko, A. S.; Tanchuk, V. Y.; Prokopenko, V. V. Virtual computational chemistry laboratory - design and description. *J. Comput.-Aided Mol. Des*. **2005**, *19*, 453−463. (b) VCCLAB, Virtual Computational Chemistry Laboratory. http://www.vcclab.org (accessed Aug 9, 2007).

(32) Andersson, F. O.; Åberg, M.; Jacobsson, S. P. Algorithmic approaches for studies of variable influence, contribution and selection in neural networks. *Chemom. Intell. Lab. Syst.* **2000**, *51*, 61−72.

(33) Karplus, M. Contact Electron-Spin Coupling of Nuclear Magnetic Moments. *J. Chem. Phys.* **1959**, *30*, 11−15.

(34) Ohno, M.; Spande, T. F.; Witkop, B. Cyclization of Tryptophan and Tryptamine Derivatives to 2,3-Dihydropyrrolo [2,3-b]Indoles. *J. Am. Chem. Soc.* **1970**, *92*, 343−348.

(35) Perrin, D. D.; Armarego, W. L. F.; Perrin, D. R. In *Purification of Laboratory Chemicals*, 2nd ed.; Pergamon: Oxford, 1980.

(36) Tetko, I. V.; Livingstone, D. J.; Luik, A. I. Neural Network Studies 1. Comparison of Overfitting and Overtraining. *J. Chem. Inf. Comput. Sci.* **1995**, *35*, 826−833.

CI700172N