Journal of
CHEMICAL INFORMATION
AND MODELING

Article

pubs.acs.org/jcim

Comprehensive Strategy for Proton Chemical Shift Prediction: Linear Prediction with Nonlinear Corrections

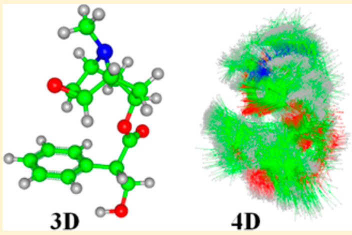
Reino Laatikainen,*,[†] Tommi Hassinen,[†] Juuso Lehtivarjo,[†] Mika Tiainen,[†] Juha Jungman,[†] Tuulia Tynkkynen,[†] Samuli-Petrus Korhonen,[‡] Matthias Niemitz,[‡] Pekka Poutiainen,[†] Olli Jääskeläinen,[†] Topi Väisänen,[†] Janne Weisell,[†] Pasi Soininen,[†] Pekka Laatikainen,[§] Henri Martonen,[§] and Kari Tuppurainen[†]

[†]University of Eastern Finland, School of Pharmacy, POB 1627, FI-70211 Kuopio, Finland

[‡]PERCH Solutions Ltd., Puijonkatu 24B, FI-71100 Kuopio, Finland

[§]Department of Chemistry, University of Jyväskylä, POB 35, FI-40014 Jyväskylä, Finland

ABSTRACT: A fast 3D/4D structure-sensitive procedure was developed and assessed for the chemical shift prediction of protons bonded to sp^3carbons, which poses the maybe greatest challenge in the NMR spectral parameter prediction. The LPNC (Linear Prediction with Nonlinear Corrections) approach combines three well-established multivariate methods viz. the principal component regression (PCR), the random forest (RF) algorithm, and the k nearest neighbors (kNN) method. The role of RF is to find nonlinear corrections for the PCR predicted shifts, while kNN is used to take full advantage of similar chemical environments. Two basic molecular models were also compared and discussed: in the MC model the descriptors are computed from an ensemble of the conformers found by conformational search based on Metropolis Monte Carlo (MMC) simulation; in the 4D model the conformational space was further expanded to the fourth dimension (time) by adding molecular dynamics to the MC conformers. An illustrative case study about the application and interpretation of the 4D prediction for a conformationally flexible structure, scopolamine, is described in detail.



1. INTRODUCTION

1.1. Background. ^1H NMR spectra could be used to explore solution structures of molecules - much like Fourier synthesis is used in X-ray spectroscopy for solid state structures. However, this requires that proton NMR spectral parameters, in particular chemical shifts, can be predicted with high accuracy as a function of molecular structure. The major bottleneck here is that the molecular structures in solution are at least four dimensional (4D, i.e. 3D structures are averaged over the time) or possibly even 5D because the chemical environment (solvation) should also be taken into account. The proton chemical shifts are determined by covalent and through-space interactions, together with the electronic environment surrounding each nucleus, which also means that the shifts may be sensitive to solvent, conformation, and even intra- and intermolecular dynamics. Consequently, the exact relationship between the 4D/5D molecular structure and the proton chemical shifts has remained largely unsolved. A major problem is that the effects of structural features are nonadditive, in opposite to the ^{13}C shifts. While the ^{13}C chemical shift prediction is well established,[1] a good ^1H chemical shift prediction forms a persistent problem.

In principle, proton chemical shifts can be predicted from molecular structures using three main approaches: (i) database methods, (ii) incremental methods employing descriptors derived from the molecular topology and structure, or (iii) quantum mechanical (QM) calculations. In the current literature, different strategies for ^1H chemical shift prediction from molecular structure have been reported.[2−12] Incremental techniques are implemented in many commercial software packages such as ACD,[13] ChemDraw,[14] MNOVA,[15] NMRPredict,[16] and PERCH Software.[17] The incremental predictors are often supplemented with advanced machine learning methods such as neural networks[5−8] or random forests.[9] There is also a publicly available database nmrshiftdb[18] with an associated machine learning (RF) method for predictions.[9] Alternatively, the parametrized methods can provide good results using functional group identification and information about the 3D structure of the compounds to be predicted, as implemented in the CHARGE program.[11] Recently, the best features of these methods have been fused together. This approach resulted in an average prediction error of 0.18 ppm against the 90,000 Wiley ^1H NMR database; this value likely represents a typical performance of the current incremental methods,[12] although recently 0.14 ppm has been proposed for the predictor of PERCH Software.[19]

New theoretical advances, together with progress in computer technology and algorithms, have eventually led to a situation where also the QM methods have become an option for chemical shift prediction. The gauge-invariant atomic orbitals (GIAO) method employing density functional theory (DFT) forms nowadays a competent method for proton chemical shift calculation especially if the structure is unusual and there is not much

Received: November 4, 2013
Published: January 23, 2014

experimental data available from similar chemical environments. However, the chemical shifts calculated by the DFT/GIAO methods need to be empirically scaled to achieve good numerical agreement with observed chemical shifts. For recent comprehensive reviews, see Jain et al.[20] and Lodewyk et al.[21] A compilation of empirical scaling factors for different QM prediction models can be found in a chemical shift repository Cheshire.[22] After scaling and when the solvent effects are taken into account using the rather economical self-consistent field procedure, good results (rrms of 0.103 ppm) have been reported for small molecules and within reasonable computing time,[21] especially when the chemical diversity of the data is considered. However, perhaps the most important bottleneck in the ab initio chemical shift calculations is again molecular flexibility, the treatment of which would require an exploration of conformational space employing either systematic or stochastic methods. At the moment, QM models are restricted to small- or medium-size structures and to the problems in which the key question is not the conformational freedom, like in our example below, become tedious with the QM methods. On the other hand, the QM methods may be used to support the development of the incremental methods by estimation of effects which cannot be characterized otherwise.

As far as we know, all the previous approaches except PERCH Software[17] are mainly based on rigid structures or only a limited conformational search has been included. The approach proposed here is based on a combination of an incremental model with MC or 4D structural descriptors, derived from a MC conformational analysis or molecular dynamics, and localized PCR supplemented with nonlinear RF and kNN corrections. The notation MC means that the conformational space is mapped as discrete conformers, whereas in the 4D model (= MC+MD) the full molecular dynamics (the fourth dimension, i.e. time) is included.

The chemical shift prediction of aliphatic protons attached to sp^3 carbon obviously poses the largest challenge for chemical shift prediction methods: the structural diversity of the sp^3 chemical environment is wide, the structures can be flexible, and the quantitative description of conformational free energies is not easy. A further complication is formed by solvent effects. A major bottleneck in the multivariate models is that substituent effects are often very nonadditive so that there are numerous unique structural fragments which may have shifts which cannot be understood by using a linear additive model or the number of the term for special structural conditions grows rapidly. This makes the problem strongly nonlinear. In this study, we used the aliphatic sp^3 proton chemical shifts to develop a general approach that is able to account for both the large trends, like the substituent electronegativity effects, and also the effects arising from special electronic environments without being sensitive to poorly described conformations. In particular, a comprehensive study of the impact of the use of RF machine learning technique, different predictor variable sets, and large training test sets on structure-based chemical shift prediction accuracy is carried out and discussed.

As to the applications of NMR parameter prediction, as recently reviewed,[21] there are many cases in the literature where the given assignments or even the structure is wrong. This could have been avoided by spectral parameter prediction. One may also envisage that the ultimate goal of NMR spectral analysis is a fully automated derivation of the molecular 4D structure from its NMR spectra. It has been proposed that the residual root-mean-square (rrms) error of ^{1}H shift prediction should be lower than ~0.1 ppm for this purpose.[23]

2. MATERIALS AND METHODS

2.1. Chemical Shift Data and Molecular Modeling. For this work we siphoned 32126 chemical shifts of 8404 compounds, collected by the project team from the vast NMR literature and publicly available databases such as SDBS.[24] Some data is common with the database of PERCH Solutions Ltd.[17] The data set consists of the molecules with protons attached to sp^3 carbon, as measured in nine solvents (carbon tetrachloride, chloroform, dimethyl sulfoxide (DMSO), acetone, acetonitrile, methanol, water, dichloromethane, and 50%:50% DMSO/chloroform). Cationic and anionic compounds were also included, but their counterions were not specified and only data from conditions where one form predominates were included. Moreover, we deliberately excluded large molecules (like peptides), exotic molecules (like large cyclophanes), and compounds with rare elements other than H, C, N, O, F, Cl, Br, I, Si, S, and P.

The analyses reported below and comparison of the chemical shifts found from different sources suggest that the standard deviation of the experimental shifts in our data is not better than 0.03 ppm (= s_{exp}, see below), at least if we include concentration effects and accept that a number of the assignments can be wrong. An illustrative example of problems in the literature data is given below, see the scopolamine example. One common problem is that the assignation of CH_2-protons is often missing, uncertain, or even incorrect. Therefore, in the cases where the interchange of the CH_2 protons shifts significantly improved the result, we allowed them to be interchanged in the first phase of the model building, before the Leave-Some-Out (LSO) validation cycles. The interchange was not allowed for the molecules for which the order was judged to be correct.

Previous prediction models for proton chemical shifts except the predictor of PERCH Software[17] have been based, more or less, on large databases and/or rigid molecular structures. In our model the conformational space of molecule was mapped with an MMC (Metropolis Monte Carlo) type method so that the maximum of 32 3D conformations (assumed to be rigid) were used in averaging the structural descriptors. The conformers were weighted using Maxwell–Boltzmann statistics. In the 4D model the descriptors were obtained after molecular dynamics averaging, which also enabled that the conformational entropies were taken into account, employing the method we have described for n-alkanes:[25] for the ratio of populations of two conformers can be written $n_i/n_0 = \exp(-\Delta G_i/RT)$, where the entropic contribution ($T\Delta S$) to the free energy ΔG_i was estimated from the torsional freedom of the MC conformers assuming that the ΔS between the conformers is equal to $R \ln\langle\Delta\phi_i\rangle/\langle\Delta\phi_0\rangle$, where $\langle\Delta\phi\rangle$'s represent the mean deviations of the dihedral angles of the conformers, obtained from an MD simulation.

The present critical settings, the maximum number of the conformers (32), and the total number of MD structures (1024 + 64 × no. of conformers) were chosen to keep the 4D computing times reasonable. Increase of the defaults had no significant effects.

The approximate nature of the force field (FF) and the MMC and MD protocols leads to more or less systematic bias in the descriptors. For example, it is obvious that the ring conformations of 5-membered rings are not well described with the present modeling protocol and, in any case, not comparable with the 6-membered systems. These biases are the reason for the local and targeted corrections in the chemical shift model: we assume that the systematic FF+MMC+MD errors can be partly compensated by dividing the whole data set into subsets (for example, the protons in the 5- and 6-membered rings) on the basis of structure and the properties like the solvent, anisotropy, and flexibility, each producing their characteristic bias.

2.2. Programming. The data (3D structures and chemical shift data) used in this work were prepared using the PERCH modeling tool.[17] The force field of the modeling tool is an extended version of the general molecular mechanics force field MMFF94,[26] as implemented in the ghemical[27] program. The MC and MCMD protocols were developed from the corresponding protocols used in the PERCH modeling software by varying the MD run lengths of the conformers on the basis of their relative energies. The conformational solvent effects are taken into account through dielectric constant, which is varied as a function of the distance between the atomic charges. The atomic partial charges for the prediction are calculated by an algorithm based on the Lewis-Langmuir[28] and Gasteiger–Marsili[29]charges, as roughly optimized using the present data. The prediction module, its testing platform and a tool for statistical analyses and kNN parametrization were written by employing Compaq Visual FORTRAN under Microsoft Developer Studio. All the calculations were performed in a microcomputer environment, and the CPU times, referred later, were obtained with an IntelXeon CPU E3-1245 V2 (3.40 GHz) processor.

2.2.1. Random Forest. The original random forest[30] FORTRAN code provided by Breiman and Cutler[31] was implemented in the prediction tool as a subroutine with necessary case-specific modifications. Preliminary test calculations were performed with the MATLAB program package (The MathWorks, Inc., version 7.6) in a PC environment using an RF package provided by Jaiantilal[32] and in-house scripts written by the authors.

2.3. Statistical Performance Indicators and Leave-Some-Out (LSO) Protocol. The most critical estimate for the quality of the prediction (δ_{pred}) is the residual root-mean-square (rrms), defined as in eq 1:

$$\mathrm{rrms} = [\sum((\delta_{obs}(\mathrm{i}) - \delta_{pred}(\mathrm{i}))^2)/N]^{1/2} \qquad (1)$$

The mean absolute errors (MAE = $|\delta_{obs} - \delta_{pred}|/N$), favored by many authors, are usually much smaller than the rrms values and may be considered as overoptimistic estimates of prediction. For the sake of comparison both are reported here. Another useful estimate is the percentage of poor predictions; in the following the observed-predicted differences >0.25 or >0.50 ppm are defined as poor and bad, respectively. On the other hand, we define as hits the predictions that are within 0.05 ppm.

The rrms values were determined using a Leave-Some-Out (LSO) protocol in which P% of the compounds (the probe set) was randomly excluded from the model, and the procedure was repeated 20 times so that the total number of predictions varies from 67589 to 116681, depending on the model (Table 2). The very basic chemical structures like n-alkanes, cyclohexane, cyclopentane, bornane, etc., and those with less than five non-hydrogen atoms were always kept in the training set and, thus, not used in estimating the LSO rrms. Unique structures like methane, chloroform, and pentafluorobenzene were also excluded from the LSO analysis as their chemical shifts can be easily picked by a search algorithm from the database. Because our database contains clusters of similar compounds (for example, steroids, carbohydrates, and derivatives of bornane) and they are clustered also in the file list, we excluded compounds in blocks of M compounds (where M = 1, 10, or 100, i.e. the files from n to $n+M-1$ in the list). The clustering forms a major challenge and, at the same time, an additional dimension for measuring the goodness of the prediction. The rrms depends strongly on the type of the proton; therefore, the statistics are reported also for the local groups (see below).

Table 1. Proton Types

proton type	explanation	N[a]
aCH_3	CH_3 protons on C attached to aromatic ring	1655
aCH_2	CH_2 protons on C attached to aromatic ring	1122
aCH	CH protons on C attached to aromatic ring	350
XCH_3	CH_3 protons on C attached to O, N, Si, P or S	2446
CH_3	C−CH_3 protons	5440
CH_2	C−CH_2 protons	6646
CH	C−CH protons	1358
5CH_2	CH_2 protons in 3−5 membered rings	3152
5CH	CH protons in 3−5 membered rings	1814
6CH_2	CH_2 protons in 6 membered ring	5165
6CH	CH protons in 6 membered ring	2210

[a] The numbers of the protons in the database.

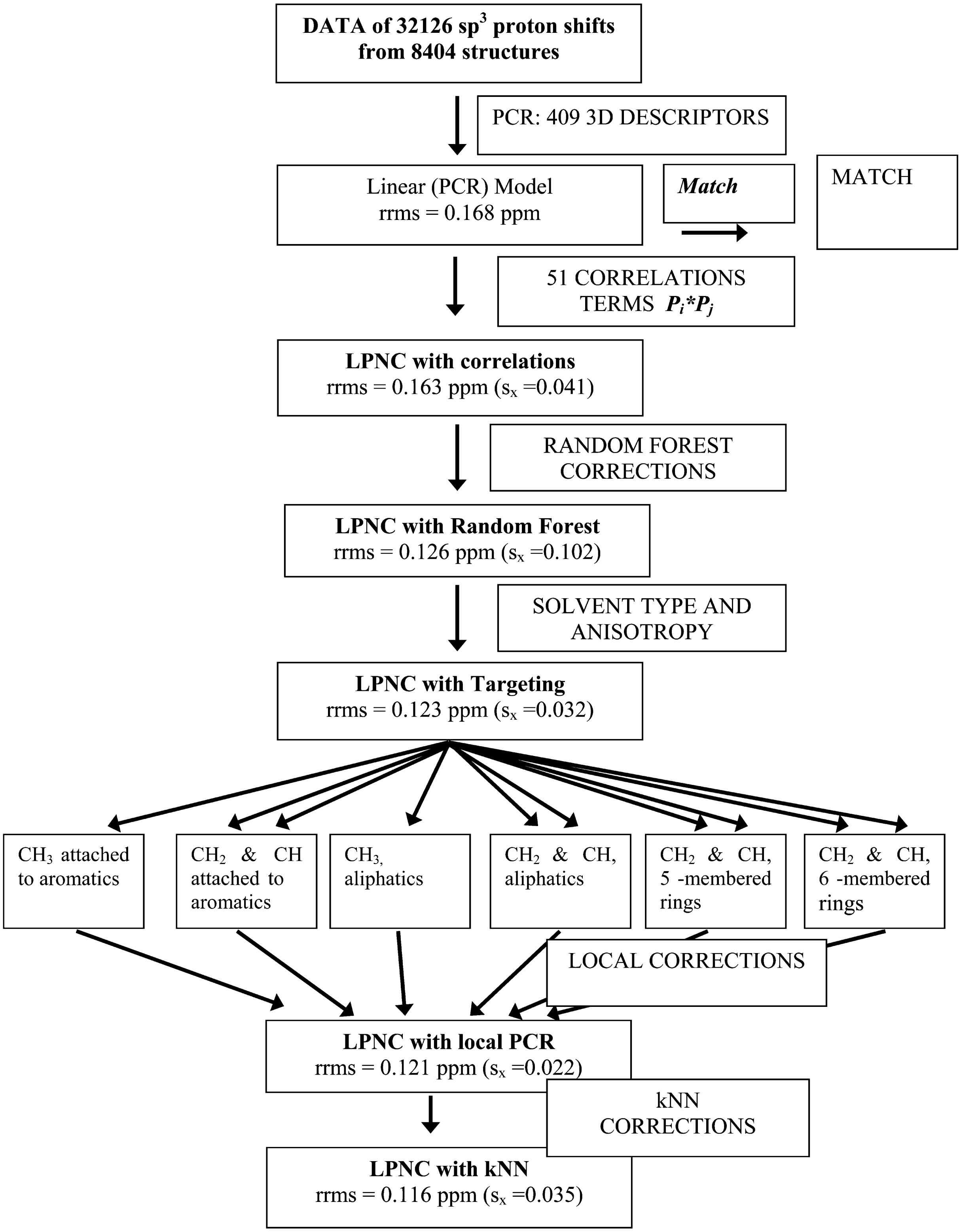
Table 2. Comparison of the LSO rrms (in ppm) Values for the MC and 4D Models, with and without RF Corrections[g]

model[a]	P%[c]	block[d]	N[e]	rrms/MC[f]	rrms/4D[f]
NORF0/0/INC[b]	0	0	31793	0.0166	0.0157
NORF0/0	0	0	31793	0.1092	0.1088
RF0/0	0	0	31793	0.1088	0.1073
NORF0/0/NIC	0	0	31793	0.1124	0.1117
RF0/0/NIC	0	0	31793	0.1125	0.1110
NORF0/0/av.CH_2	0	0	27032	0.1049	0.1041
RF0/0/av.CH_2	0	0	27032	0.0989	0.0980
NORF10/1	10	1	67589	0.1137	0.1129
NORF33/1	33	1	116681	0.1382	0.1336
NORF10/10	**10**	**10**	**68668**	**0.1231**	**0.1216**
NORF10/100	10	100	79745	0.1361	0.1360
NORF10/10/av.CH_2	10	10	58112	0.1214	0.1208
RF10/1	10	1	67589	0.1100	0.1049
RF33/1	33	1	116681	0.1188	0.1175
RF10/10	**10**	**10**	**68668**	**0.1180**	**0.1160**
RF10/100	10	100	79745	0.1314	0.1294
RF10/10/av.CH_2	**10**	**10**	**57395**	**0.1113**	**0.1113**

[a] RF = RF phase included, NORF = no RF phase, INC = the current chemical shift included at the kNN phase, NIC = the CH_2 shifts were Not Inter Changed automatically, av.CH_2 = the CH_2 shifts were averaged. [b] In the ideal case, in the absence of experimental effects like concentration shifts, these rrms values should be zero. [c] P% is the percentage of compounds in the testing set which was randomly extracted from the database. The LSO protocol was repeated 20 times, and the predictions were combined and then used for developing the 14 terms kNN model. Some (ca. 180) key compounds and all the small molecules with less than 5 carbons were always kept in the teaching set and not predicted. [d] The size of the blocks of molecules which were excluded in LSO. Zero means that the model was done without LSO. [e] The total number of predictions. [f] The 90% confidence limits for the rrms values are <0.001 ppm. [g] The bolded models are compared more in Tables 3−6.

In order to estimate the importance and magnitudes of different contributions to the chemical shifts we also report the corresponding standard deviations which are obtained assuming

Scheme 1. Flow Chart for the LPNC Procedure


DATA of 32126 sp3 proton shifts from 8404 structures
PCR: 409 3D DESCRIPTORS
Linear (PCR) Model rrms = 0.168 ppm
Match
MATCH
51 CORRELATIONS TERMS Pi*Pj
LPNC with correlations rrms = 0.163 ppm (sx =0.041)
RANDOM FOREST CORRECTIONS
LPNC with Random Forest rrms = 0.126 ppm (sx =0.102)
SOLVENT TYPE AND ANISOTROPY
LPNC with Targeting rrms = 0.123 ppm (sx =0.032)
CH3 attached to aromatics
CH2 & CH attached to aromatics
CH3, aliphatics
CH2 & CH, aliphatics
CH2 & CH, 5 -membered rings
CH2 & CH, 6 -membered rings
LOCAL CORRECTIONS
LPNC with local PCR rrms = 0.121 ppm (sx =0.022)
kNN CORRECTIONS
LPNC with kNN rrms = 0.116 ppm (sx =0.035)

that the total variance (s_{total}^2) can be divided into independent terms, first

$$s_{\text{total}}^2 = s_{\text{exp}}^2 + s_x^2 + s_M^2 \tag{2}$$

where s_{exp}^2 represents the experimental variance, s_x^2 represents the variance arising from a special feature x of the model (for example changing the solvent to DMSO), and s_M^2 represents the variance arising from the rest of the model. If we assume that s_{exp} is ca. 0.03 ppm (the estimate obtained below) and that s_{total} of 0.118 ppm (model RF10/10 for the MC prediction), the standard deviation of the model (s_M, when $s_x = 0$) is 0.114 ppm. When the s_{total} of 0.118 ppm is reduced by 0.002 ppm by adding a new descriptor or property (like the fourth dimension to the RF10/10 model, for example) to the model, the average contribution (s_x) of the property to the shifts would be as large as 0.022 ppm (11 Hz at 500 MHz). These numbers mean that even the 0.002 ppm improvement in s_{total} means a significant contribution to the shifts, especially if the new property concerns only a limited fraction of the shifts, like the fourth dimension.

2.4. LPNC (Linear Prediction with Nonlinear Corrections).

The major problem in any multivariate regression type approach is that there are numerous chemically different substituent groups. In addition, their effects on the chemical shifts depend on solvent and on the pathway between the substituent and nucleus. In practice it is impossible to gather information about all the possible chemical systems, and there is no such an easily computed general physicochemical descriptor (like atomic charge) which would correlate completely with chemical shifts and which could be easily computed for any chemical system. If the prediction is needed to be done on the fly or just for obtaining reasonable starting values for QM spectral analysis (QMSA, see below), the GIAO/DFT methods do not fulfill this condition either.

One way to avoid the explosion of the number of parameters is to calculate corrections to them arising from different conditions like the primary local environments of the nucleus (for example, CH, CH_2, CH_3 protons), "targeted" chemical classes (like steroids and carbohydrates), or solvents. The basic idea of the present LPNC approach is that it consists of a linear phase based on several "local and targeted" PCR steps, together with an RF analysis for nonlinear corrections to the PCR model. Further, a nonlinear correction and a database search type feature, based on a kNN type algorithm, is applied in order to find the hits for the case that the target molecule is found in the database or there is another chemical environment which resembles it very closely. The entire prediction procedure is illustrated in Scheme 1.

2.5. The General Strategy for 3D and 4D Structure Sensitive Chemical Shift Prediction.

In the present model, a proton chemical shift δ_i is expressed by a multivariate equation as follows

$$\delta_i = \delta_k{}^{o} + \sum P_n\left\langle X_n\right\rangle + \sum C_{nm}\left\langle X_nX_m\right\rangle + \Delta\delta_i{}^{TARGET} + \Delta\delta_i{}^{RF} + \Delta\delta_i{}^{LOCAL} + \Delta\delta_i{}^{kNN} \quad (3)$$

where $\delta_k{}^{o}$ is the basic chemical shift of the class k proton (see Table 1), and the term $P_n\,\langle X_n\rangle$ adds the contribution of the descriptor X_n ($\langle X_n\rangle$ = the numerical value of the descriptor averaged over the conformational space) to the chemical shift. The last four terms represent the linear and nonlinear corrections to the basic model and are discussed below.

The contribution of each correction to the total variance can be written as follows:

$$s_{total}{}^2 = s_{exp}{}^2 + s_{PCR}{}^2 + s_{correlations}{}^2 + s_{target}{}^2 + s_{RF}{}^2 + s_{local}{}^2 + s_{kNN}{}^2 + s_M{}^2 \quad (4)$$

One should note that the contributions depend on the order in which they are applied: for example, s_{kNN} represents the improvement which is obtained when the kNN correction is applied after the local corrections.

The experimental $s_{exp}{}^2$ term can be decomposed further

$$s_{exp}{}^2 = s_{ref}{}^2 + s_{sol}{}^2 \quad (5)$$

where the former term ($s_{ref}{}^2$) represents the effect arising mainly from the interactions between the solute and the reference compound (TMS or TSP). The interactions, arising for example from aromatic anisotropy of the solute, change all the solute chemical shifts with the same amount; this assumption is applied below in estimating its magnitude. The latter term $s_{sol}{}^2$ represents all the other solvent and concentration effects, experimental errors, and also uncertainties arising from structure (conformation) which may be even wrong or poorly defined, like the structures of acids and bases at certain pH.

2.6. Chemical Shift Descriptors with Correlation, Local, and Targeted Corrections.

The descriptors of the model have been described (for protein chemical shift prediction) in detail elsewhere,[33,34] and only some guidelines will be given here. The number of the basic descriptors was here 409. Each descriptor is based on some sound physical property like atomic charges or dipolar terms commonly related to the chemical shifts. However, the importance and contributions of the individual terms are not essential for the present evaluation and therefore not widely discussed.

Broadly, the basic descriptors (X_n) can be divided into two classes. First, in addition to the basic chemical shift $\delta_k{}^{o}$ of the "local" class k proton (CH_3, CH_2, CH in 3−5- or 6-membered ring, etc., see Table 1), the static descriptors include the atomic charges, obtained as above-described, and the substituent effects of atoms. For example, all the oxygen atoms are divided into only six classes (ROC, ROX, C═O, N═O, furane, and X═O, where R = H or C and X = Si, P, or S), and the substituent terms are formed for these classes over up to five bonds. The difference between the substituent effects of the COOH and COO^- oxygens, for example, is done by defining the corresponding carbons to different classes. In addition, the groups are separated because of the different atomic charges. The total number of the oxygen species, 47, defined for the kNN similarity analyses, is much larger. All these terms are independent of the 3D geometry. Second, the terms characterizing the MC and 4D structures include conformational and stereochemical properties all averaged over the conformational space. These terms include Coulombic and van der Waals interactions, bond and ring anisotropies, and solvent accessibility parameters.

The correlation parameters C_{nm} describe corrections arising from the correlation of the X_n and X_m descriptors: for example, the two-bond substituent effect of substituent X can be different in the presence of the three-bond effect of Y. The significant correlation parameters are searched by a straightforward linear regression analysis. The number of C_{nm} terms was 51 so that the total number of the regression variables was 460. Also this part of the model can be considered as a nonlinear correction.

The "local" and "targeted" corrections (ΔP) are linear corrections to the P parameters and obtained using weighted PCR analysis:

$$\Delta\delta_i{}^{LOCAL} = \sum \Delta P_n{}^{LOCAL}\left\langle X_n\right\rangle + \sum \Delta C_{nm}{}^{LOCAL}\left\langle X_nX_m\right\rangle \quad (6)$$

$$\Delta\delta_i{}^{TARGET} = \sum \Delta P_n{}^{TARGET}\left\langle X_n\right\rangle + \sum \Delta C_{nm}{}^{TARGET}\left\langle X_nX_m\right\rangle \quad (7)$$

The use of these corrections is a way to avoid expanding the dimension of the PCR analysis. This approach can be considered analogous to the neural network with associated networks.[1]

For the local corrections the sp^3 protons are divided into eleven local subclasses defined in Table 1. Separate corrections to the regression parameters are derived for each of them so that the parameters are solved for each of the subclasses by underweighting the data of the other subclasses on the basis of their “heuristic” similarities. The model could still be made more localized by focusing the classification range (for example to 3- and 4-membered rings), but it would demand more data.

If the substituent effects are different in different solvents, the effects are not easily described by the above model. This is not a serious problem for sp^3 protons, but it may be significant for aromatic systems in which the width of the aromatic part depends on the solvent polarity. For the sp^3 protons, the poorly predicted conformational equilibriums and flexibility form the problem. In this study we divided the data into three categories on the basis of solvent (CCl_4, $CDCl_3$, and CD_2Cl_2; DMSO, acetone-d_6, and acetonitrile-d_3; D_2O and CD_3OD) and into three categories on the basis of flexibility and anisotropy (flexible structure with anisotropic groups close to the proton; flexible but no anisotropic groups close to the proton; rigid structures or no anisotropic groups close to the proton). This kind of targeted prediction can be added on the basis of any property: for example, the regression parameters could be calculated for compound groups like steroids, carbohydrates, or compounds containing only H, C, and O. The targeting was realized in the same way as the local corrections, i.e. by using weight parameters.

The number of the nonorthogonal corrections related to the target and local terms is 18 × 460, where 18 is the number of target and local classes and 460 (= 409 + 51) is the total number of P_n and C_{nm} terms, respectively (see below). However, the effective number of the independent nonzero correction parameters is very much smaller, because the important corrections are selected by PCR analysis and the statistically insignificant corrections are forced to zero. For the RF and kNN terms the question is not relevant. The division of the corrections serves also the theoretical question about the importance of the different solvents and structural properties. The details of the entire algorithm for these corrections are out of the scale of this presentation.

2.6.1. Correction with Random Forest (RF).

The methods such as PCR cannot deal properly with nonlinearities and complex interactions between descriptors. As a starting point of this study, we assumed that the residuals obtained from the PCR models include nonlinear information about the relationships between chemical shifts and descriptors and that the RF regression should be particularly suited to model them. Previously, the same strategy has been applied in the field of QSAR by Devillers[35] employing, however, supervised artificial neural networks (ANN) for nonlinear modeling. At the early stage of this work, ANNs were tested, but they were abandoned in favor of RFs due to problems in tuning, overfitting, and performance.

The RF method,[30] introduced by Breiman, combines two machine learning techniques viz. bagging (an acronym for bootstrap aggregating) and random feature subset selection using a large collection of unpruned decision trees for predictions. Each of these trees predicts a real value by querying a set number of variables and instances within the regression model. Each regression tree is thus trained on a different bootstrap sample of both training samples and features. The RF then averages the predictions made by the trees in the forest to produce the final output. The RF method is thus an example of an ensemble method of machine learning. In general, the performance of random forests compares favorably with modern machine learning methods such as support vector regression or neural networks, with very little tuning required.[36] As a consequence, the RF models have become increasingly popular in both classification and regression problems.

Among multivariate methods, RFs are particularly resistant to overfitting, including only two adjustable parameters; the number of trees (Ntree) and the number of variables to be tried in each split (Mtry). Ntree is typically set at 500−1000, as a larger number will not provide any gain. In regression problems, much smaller values (100−200, or even 50 as observed in this work) may work properly.[37] For Mtry, a value of Nvar/3 is recommended for regression; here Nvar is the number of variables. The optimization of Mtry usually provides only a slight improvement.

Distinctively, RFs come with a build-in cross-validation using out-of-bag data (usually 30% of the samples). An OOB (Out-Of-Bag) estimate is almost identical to that obtained by N-fold cross-validation.[36] In fact, it has been proved that if the number of bootstrap samples gets large, the OOB error estimate for a RF approaches its N-fold cross-validated error estimate, and that, in the limit, the identity is exact.[36] This implies that RFs can be fit in one sequence, without the use of separate training and test sets. In fact, it has been established that the OOB error can overestimate the true error in regression problems, i.e. RF actually performs better than indicated by the OOB error.[38]

The RF corrections $\Delta\delta_i^{RF}$ were obtained using 50 trees and setting Mtry to its recommended value (Nvar/3). Some of the original descriptors, like the solvent index and local types were transformed into categorical variables.

2.6.2. kNN Algorithm.

In order to detect the cases where the database contains similar magnetic environment as the target proton, we developed a kNN type algorithm, which allows an exact prediction if an identical or very similar compound is found in the database. The kNN correction can also be considered as a nonlinear correction. The recognition of the identity can also be done after the first step of the prediction (see Scheme 1). Because the objective of our assessment was to explore the MC and 4D algorithms, the “match” option was turned off in our predictions. For less similarity it can be applied as the last step when the prediction errors are at the smallest for the training set.

In our protocol, the $\Delta\delta_i^{kNN}$ correction is estimated on the basis of the distance between the target and the nearest k database points. It is assumed that the prediction error (E_i, which equals to the kNN correction) is the same if the chemical environments are identical. If an identical or very similar structure is not found from the database, k (<10) nearest neighbors (with E_k) are selected and assumed that $E_i = \sum W_k * E_k$. The weights (W_k) were estimated from the squared ‘distances’ $d_{ik}^2 = \sum(\Delta\delta_{in} - \Delta\delta_{kn})^2$, $d_{ik}^{2'} = \sum(\Delta\delta_{in}^2 - \Delta\delta_{in}\Delta\delta_{kn})^2$, $d_{ik}^{2''} = (\sum |\Delta\delta_{in} - \Delta\delta_{kn}|)^2$, $d_{ik}^{2'''} = (\delta_i - \delta_k)^2$ and a ‘distance’ where the differences between the shift contributions ($\Delta\delta_{in} = P_n\langle X_{in}\rangle$, from eq 3) are summed and squared for each descriptor class. When the weights $W_k(d_{ik}^m)$ are set equal to $1/(d_{ik}^m + b)$, a linear equation is formed:

$$E_i = P_1W_k(d_{ik}^2)E_k + P_2W_k(d_{ik}^3)E_k + P_3W_k(d_{ik}^{2\prime})E_k + P_4W_k(d_{ik}^{3\prime})E_k... \quad (8)$$

The constant b was set to depend on the standard deviation of the proton class. In addition to the second and third order terms of the distances, we added four second order terms weighted by the distance of the nearest neighbor, in order to enhance the importance of very similar environments. The 14 parameters (P) describing the importance of the different distances were solved from the group of linear eqs (27032−116681 equations) where the prediction errors (E_i) were obtained from the LSO analyses (see below).

2.6.3. Principal Components Regression (PCR). Because some parameters used in the model may be strongly correlated, all the linear regression equations were solved by PCR. The analysis can be controlled by the threshold parameter, which defines the minimum eigenvalue of the principal components included into the model. The threshold value is rather critical in this application: if the value is too large, it may remove a contribution that is important and well-defined for a few data points in the teaching data, which is the reason using PCR instead of the partial least-squares (PLS) method. To prevent the problem of the very rare contributions, we overweighted these descriptors in the covariance matrix. The PCR analysis reduced the number of the original descriptors from 460 to ca. 420 (depends on the LSO probe set) independent (orthogonalized) parameters. To avoid overfitting, the threshold values were set larger in the calculation of the local and targeted corrections.

2.7. Computation Times. The CPU time of chemical shift prediction consists of two parts. The first part represents the conformational analysis and the computation of the descriptors, and it depends much on the molecular structure and the values of the MC and MD parameters. For the MC (where only conformational search is performed, allowing at most 32 conformers and 256 MC-trials) the CPU time for scopolamine (see below) with six rotating bonds was 14 s, but when the fourth dimension was added, it was increased to 64 s. For testosterone, which has only one rotating bond, the corresponding times were 1 and 55 s, respectively. Anyhow, according to our experience, the improvement obtained with a longer run does not often pay the efforts. With the second part, the actual prediction time is <1 s for one chemical shift. Thus, the total CPU time burden is negligible in comparison with GIAO/DFT calculations.

3. RESULTS AND DISCUSSION

The stepwise prediction protocol is illustrated in Scheme 1, and the numbers given there are based on the LSO analysis and correspond to the RF10/10 model of Table 2. The protocol contains elements common with the PERCH predictor[17] and the 4DSPOT protein chemical shift predictor,[33,34] although there are also some novel features and the prediction protocols differ in several aspects: the shift descriptors and atom typing were checked and revised (especially for the ionic species), the databases differ in volume and content, the targeted models and the RF corrections are not available in PERCH predictor (RF was tested in 4DSPOT with minor success), and the kNN and error prediction algorithms are based on the extensive LSO analysis. Furthermore, the modeling tool and the protocols contain modifications to those used in the PERCH molecular modeling tool.

Each step in the protocol corrects effects which correspond from 0.02 to 0.10 ppm contribution (s_x in eq 3) on shifts. For example, s_x = 0.035 ppm for the targeted correction means the solvents and molecular structure type has such an average effect on the shifts. The largest effect is obtained by the RF model.

3.1. Chemical Shift Prediction Simulating Real-World Conditions: Leave-Some-Out (LSO) Analyses of the Predictor. Straightforward statistical tests (randomized training and test sets, y-scrambling tests) were employed at the early stage of this study, mainly in order to establish the applicability of the chosen strategy in general and the reliability of the OOB estimate for the prediction error in particular. Both tests were successful (data not shown), and thus OOB estimate could be used as a surrogate for the true prediction error. However, in order to obtain a sound picture of the reliability of the predictions in real situations, we calculated statistics based on several LSO protocols, where up to 33% of the compounds were left randomly out of the model. For all the models where the block size and P% are nonzero in Table 2, the LSO protocol was repeated 20 times which seems to be sufficient for the convergence of the statistical performance indicators.

The results of the LSO analyses are summarized in Table 2 for both the MC and 4D models. For example, in the model RF10/100 10% of the molecules were raffled into the testing set in blocks of 100 molecules (molecules M to M+99 in the database). The LSO protocol was repeated 20 times so that the total number of predictions was even 116682, which means that some molecules were predicted several times. The statistics in Scheme 1 and Tables 3−6 correspond to the model RF10/10 for the 4D prediction.

Table 3. Classification of the LSO Predictions for the Corresponding RF10/10 Protocols in Table 2

	MC		4D		4D/av.CH_2	
deviation	rrms	%	rrms	%	rrms	%
hit (0−0.05)	0.014	33.3	0.014	35.8	0.014	35.6
good (0.05−0.10)	0.058	43.1	0.057	42.6	0.057	43.2
fair (0.10−0.25)	0.159	17.9	0.158	16.7	0.158	16.4
poor (0.25−0.50)	0.340	4.9	0.337	4.2	0.340	4.2
bad (>0.50)	0.645	0.8	0.656	0.7	0.649	0.6

The total statistics appeared to be sensitive to the block size. The reason is that similar structures (like steroids, carbohydrates, and derivatives of other basic structures) are in blocks in our database, where the molecules are arranged according to their origin and type. This forms an extra dimension for the LSO analyses but, on the other hand, offers a way to explore the situations where the similarity between the current structure and the teaching database structures varies from good to poor. If molecules were selected to the testing set completely randomly (block size 1), rrms falls to 0.1073 ppm (model RF0/0). However, this number can be considered too optimistic because the database contains many rather similar molecules and therefore kNN is able to locate the similarities, while in a real situation there may not be such relatives in the database. On the other hand, if block size is increased to 100, the rrms jumps to 0.129 ppm (model RF10/100) which may be considered somewhat biased, too, because the exclusion of a block of 100 compounds may lead to the situation where, for example, all the small carbohydrates are excluded from the training set but then predicted. The model RF10/10 can be considered as a compromise between these two tendencies.

An alternative for increasing the block size is to increase the fraction of the molecules extracted from the database to the LSO probe set. This is done in the models NORF33/1 and RF33/1. The effect is surprisingly small in the latter model, proving the intelligence of the RF algorithm.

Leave-Some-Out Cross Validation (CV) is a widely applied standard method to estimate the true predictive ability of a regression model. Also CV may lead to too optimistic statistics, with small data sets and/or if used in the feature selection in classification problems.[39] However, it seems that these items, although important per se, are not an issue in our regression problem because our data set is large and CV has not been applied for feature selection. Instead, our variables are preselected on physical grounds and then orthogonalized with PCR in order to enhance the stability of the models. Because our database contains data from very different sources, it is difficult to imagine a large high-quality data set which would represent larger structural diversity than our database. As found, we needed more than 10 different sets of ca. 3000 chemical shifts before the LSO statistics converged; therefore, the external data set should be very large to guarantee that the good or poor result is not accidental.

3.2. Concentration Shift and Experimental Errors. A part of the rrms error arises from experimental accuracy (s_{exp}) which has been divided into two parts in eq 5. When each data point is included in kNN prediction while it is predicted, rrms should be close to zero instead of ca. 0.016 ppm which is obtained with the optimized kNN algorithm (NORF/INC in Table 2). This means that the chemical shifts of even very similar chemical environments seem to vary with 0.016 ppm, which could be considered as an estimate of the experimental standard deviation (s_{exp} in eq 2) of the shift data. However, an important contribution to this value arises from the contingency of the MC and MD protocols, which can be denoted with s_{MCMD}, an estimate of which (>0.04 ppm) was obtained also in the case of scopolamine, discussed below, for which the large effect stems from the dynamic nature of the structure.

Our data allows us to examine also the bias (s_{ref} in eq 5) arising from solute−solute and solute-reference interactions. For example, even moderate concentrations of aromatic solute may increase significantly the difference between the solute and the reference signals. Because the concentrations of the teaching database spectra are virtually unknown and varying, also the concentration shifts vary between the samples. If we assume that there is a shift which arises only from the effects of solute on the reference shifts, the shift should be equal for all the solute shifts. When we selected the molecules with at least ten sp^3 proton shifts (totally 35667 shifts from the LSO analysis) and applied a linear model (deviation = A × predicted shift + B) for each structure, the rrms dropped from 0.1109 to 0.1081 ppm, which corresponds to s_{sol} of 0.025 ppm. Without the term A (which reflects the assumption that the concentration shift could be larger for protons having large shifts), the rrms dropped to 0.1085 ppm. Because the s_{sol} term does not affect the order of the chemical shifts, it can be easily corrected in computerized spectral analysis, in which the predicted wrong order of shifts is a fundamental problem (see the discussion below). When we combine s_{ref} and s_{sol} and assume them independent of each other, we obtain an estimate for s_{exp} (ca. 0.03 ppm) and the quality of the data. This number does not include all the uncertainties which are not related to the prediction model, like those arising from erroneous data, assignments, and even structures, so that even 0.05 ppm could be a realistic value for s_{exp}.

In order to obtain an rrms estimate which measures the goodness of the prediction when it is intended for a starting point of QMSA or for a similarity analysis (where due to concentration effects, the relative shifts are more important and diagnostic than the absolute shifts), the numbers in Table 2 can be "cleaned" by removal of the experimental variance (s_{exp}^2): this decreases the rrms of 0.116 ppm to 0.113−0.105 ppm, if assuming s_{exp} of 0.03−0.05 ppm.

3.3. Local Predictions and Diastereotopic CH_2 Protons. The rrms values depend strongly on proton type (Table 4), and they are best for the CH_3 protons and poorest for the CH protons, varying from 0.087 ppm of the CH_3 protons in aliphatic system to 0.216 ppm of the aCH-protons with CH carbons attached to aromatic system. The large variation between the proton types makes the total rrms a poor estimate of the goodness of the prediction. With the exception of the aCH protons the estimate of MAE was <0.10 ppm. The number of the aCH protons in the database was 350 (Table 1), and rrms and MAE were only 0.121 and 0.087 ppm in the "pseudo kNN" model (RF0/0), respectively, indicating that the number of similar aCH chemical environments is too small and therefore the rrms fit could be improved by adding new high-quality data.

Table 4. Comparison of the Local Predictions for the Corresponding RF10/10 Protocols in Table 2

		MC	4D		4D/av.CH_2
type[a]	N[b]	rrms	rrms	MAE	rrms
aCH_3	4563	0.093	0.100	0.065	0.092
aCH_2	2726	0.138	0.134	0.088	0.131
aCH	806	0.202	0.216	0.138	0.201
XCH_3	5166	0.098	0.093	0.057	0.096
CH_3	10830	0.080	0.077	0.045	0.080
CH_2	14325	0.110	0.101	0.066	0.099
CH	2904	0.159	0.154	0.097	0.162
5CH_2	5482	0.147	0.131	0.084	0.130
5CH	3050	0.163	0.161	0.099	0.161
6CH_2	11836	0.128	0.122	0.073	0.107
6CH	4993	0.152	0.135	0.084	0.148

[a] See Table 1. [b] The number of the predictions.

Diastereotopic CH_2 protons[40] pose a special challenge for the sp^3 proton prediction. It seems that there are often errors in their assignments or they are not given at all, and, second, they are sensitive to the local conformation and, therefore, not so well predicted by the present modeling. For the LSO validation described in this work, to remove the most obvious errors in the assignments, we interchanged the chemical shifts of CH_2 protons automatically if it improved significantly their prediction. The interchange was not allowed for those molecules for which the order was judged to be correct. This improves the rrms statistics by 3.3% (RF0/0 vs RF0/0/NIC) and was set to default for the other LSO models in Table 2.

If the shifts of the diastereotopic CH_2 protons are averaged, the total rrms drops from 0.116 to 0.111 ppm (RF10/10 vs RF10/10/av.CH_2). For the average of the 5CH_2 and 6CH_2 protons rrms drops from 0.129 to 0.111 ppm and MAE from 0.080 to 0.069 ppm. These numbers can be compared also to the MAE (0.19 ppm) reported by Aires-de-Sousa et al.[5−7] Unexpectedly, the rrms for the nonring CH_2 protons (mostly not diastereotopic) increases by ca. 0.01 ppm. The predicted average shifts can be used as constraints in QMSA based on spectral prediction.

3.4. Targeted Prediction and Molecular Size. Table 5 gives statistics when the protons are classified to groups on the basis of solvent type, flexibility, and anisotropy. Not surprising, the DMSO data gave a larger rrms than the $CDCl_3$ data. Unexpectedly, the fourth dimension had only a small effect on this statistics. The rrms for the rigid structures was larger than for the flexible ones but was improved more with the fourth dimension. The effects of the molecular size (Table 6) were more logical: the increase in the size increased rrms, but the 4D model improved more the predictions of the larger molecules.

3.5. MC vs 4D. Unexpectedly, the 4D prediction is not much better than the MC protocol (Tables 3−6): the reduction of rrms from 0.118 to 0.116 ppm (for RF10/10, see the speculation after eq 2) means that the average contribution of the fourth dimension is only 0.022 ppm, which is usually insignificant for prediction but, on the other hand, not negligible when given in

Table 5. Comparison of the Targeted Predictions for the Corresponding RF10/10 Protocols in Table 2[I]

		3D	4D		4D/av.CH_2[h]
type	N[h]	rrms	rrms	MAE	rrms
$CDCl_3$[a]	46937	0.116	0.111	0.066	0.108
DMSO[b]	6757	0.152	0.143	0.097	0.145
HOX[c]	14974	0.124	0.117	0.072	0.115
rigid[d]	9379	0.132	0.119	0.075	0.118
flexible[e]	4438	0.102	0.094	0.052	0.096
anisotropy+flexible[f]	39487	0.123	0.118	0.071	0.114
anisotropy[g]	15004	0.120	0.115	0.072	0.120

[a] CCl4, $CDCl_3$, CD_2Cl_2, and polysolv ($CDCl_3$+DMSO) solutions. [b] DMSO, acetone-d6, and CD_3CN solutions. [c] D_2O and CD_3OD solutions. [d] No rotating groups (except symmetric fragments like C−CX_3 or phenyl groups). [e] Rotating groups but no strong aromatic or double bond anisotropy. [f] Rotating groups and remarkable aromatic or double bond anisotropy. [g] Aromatic or double bond anisotropy but no rotating groups. [h] Number or shifts in the LSO validation. The number is smaller when the CH_2 shifts were averaged. [I] A proton can belong simultaneously to two targeted groups.

Table 6. Effect of Molecular Size (Number of Carbons) on the LSO Prediction for the Corresponding RF10/10 Protocols in Table

		MC	4D		4D/av.CH_2
carbons	N[a]	rrms	rrms	MAE	rrms
1−5	8303	0.111	0.114	0.071	0.109
6−10	21195	0.115	0.111	0.069	0.110
11−20	19096	0.129	0.124	0.074	0.122
21−40	20020	0.128	0.114	0.070	0.114
>40	54	0.074	0.085	0.055	0.068

[a] Number or shifts in the LSO validation. The numbers are smaller when the CH_2 shifts were averaged.

frequency units (10 Hz in 500 MHz proton spectra). One reason is that the number of the molecules for which the fourth dimension is essential is not large and that the fourth dimension may be partly compensated by the RF and kNN algorithms. Therefore, because the 4D prediction is often much slower than the MC protocol, it pays to consider its use only in the cases where the phrasing of the question is related to the fourth dimension. The following example is meant to illustrate the special features and limitations related to the 4D ^1H prediction and its applications.

3.6. A Case Study: Scopolamine. Figure 1 shows the 3D, MD, MC, and 4D structures of scopolamine.[41] The structure has five rotating bonds, when the symmetric methyl and phenyl group rotations are not counted, which means that there are plenty of potential conformers. The MC search typically found seven conformers within 10 kJ/mol above the minimum energy. The entropic contributions ($T\Delta S$) to the free energies varied from −0.5 to 1.2 kJ/mol when compared to the minimum energy conformation.

The lowest energy conformer (Figure 1) has the phenyl ring in a position which leads to an aromatic group induced shift that breaks the symmetry of the system and explains the observed nondegeneracy of the H6 and H7 shifts. The shown 3D conformation of the −CH−CH_2OH moiety is in accordance with the $^3J(CH_2OH,CH)$ (9.2 and 5.7 Hz); the intramolecular hydrogen-bond (which is an attractive alternative) would lead to smaller and more degenerate $^3J(CH_2OH,CH)$ couplings. Notable is that our geometry optimization calculations do not favor the hydrogen bonded conformation but gives an energy which is ca. 4 kJ above the best structure.

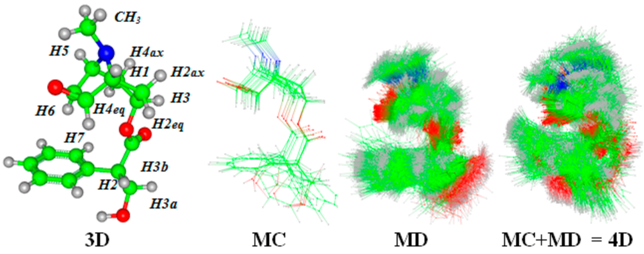


Figure 1. The 3D, MC, MD, and 4D (= MC+MD) structures of scopolamine. The observed and predicted chemical shifts are compared in Table 7.

The observed shifts and observed-predicted differences are given in Table 7. The most diagnostic feature in this structure is the strong nondegeneracy of the H6 and H7 chemical shifts. Although the tropane system itself is fully symmetrical, the chiral center locating behind 7 bonds removes the symmetry and leads to a 0.67 ppm shift difference for those protons. This kind of effect would be impossible to explain on the basis of 2D structures. Due to the symmetry of the tropane system, the original assignments[41] of the H6 and H7, H1 and H5, and H2 and H4 protons could be − and in fact − interchanged in Table 7 because the assignment (where the predicted H6 shift is smaller than that of H7) gives clearly better (rrms 0.18 ppm) prediction than the original one (0.34 ppm).

Table 7. Observed Chemical Shifts[41] and the Observed − Predicted Differences for the MC and 4D Predictions for Scopolamine in $CDCl_3$

proton	obsd	3D/ MC[a]	MC/ MC[a,b]	MD/ 4D[a,c]	4D/4D[a,d]	Pred.SD[a,e]
H2	3.72	3.69	3.68(2)	3.69(2)	3.67(1)	0.11
H3a	4.30	4.14	4.10(7)	4.22(5)	4.18(2)	0.14
H3b	3.77	4.10	4.03(5)	4.05(6)	4.00(2)	0.14
H3	4.98	4.98	5.00(5)	5.05(4)	5.05(2)	0.10
H2ax	2.08	1.83	1.84(1)	1.97(2)	1.94(1)	0.20
H4ax	1.99	1.83	1.82(1)	2.02(3)	1.97(1)	0.22
H2eq	1.55	1.63	1.60(7)	1.57(8)	1.55(2)	0.21
H4eq	1.31	1.71	1.75(4)	1.69(3)	1.73(1)	0.19
H1	3.08	3.07	3.07(1)	3.08(2)	3.05(1)	0.14
H5	2.94	2.88	2.93(5)	2.93(3)	2.93(1)	0.20
H6	2.70	1.70	2.15(51)	1.96(31)	2.34(4)	0.75
H7	3.37	3.11	3.17(9)	3.10(9)	3.15(12)	0.55
CH_3	2.42	2.36	2.38(1)	2.42(1)	2.33(1)	0.07
rrms		0.34	0.28(14)[g]	0.28(9)[g]	0.18(4)[g]	0.30[f]

[a] The prediction was done excluding scopolamine and its derivatives from the teaching database. The predicted values are average values of 20 runs, each started with different random numbers for MC and MD. The numbers in the parentheses give the standard deviations (in the last digit) of the 20 predictions. [b] The structures were an ensemble of MC structures, and the MC model was used. [c] The structures were an ensemble of MD structures, and the 4D model was used. [d] MD was added to the MC structures, and the 4D model was used. [e] Estimated predictions errors (standard deviations) for the 4D/4D model. [f] Average SD. [g] The numbers in parentheses give the standard deviation (s_{MCMD}) arising from the MC and MD protocols.

The trend in Table 7 is clear: the more dynamic the model, the better rrms. However, doubling the efforts in the 4D model did not pay the costs, but the rrms was almost identical with shorter calculations (results not shown). In the rigid 3D model (3D/MC) the H6 chemical shift obtains its minimum value, and the addition of the fourth dimension improves the prediction indicating that the scopolamine structure is strongly four-dimensional.

In order to obtain an estimate for the variance arising from the stochastic nature of the MCMD algorithm, we performed the prediction 20 times with different random seed numbers. The prediction errors are rather small (MAE 0.025 ppm and standard deviation 0.041 ppm for 4D/4D, see Table 7), with the exception of the H7 proton (0.12 ppm), which is easily explained by the tilting phenyl group. Moreover, the estimated prediction standard deviations (see below) are in qualitative agreement with the deviations between the observed and predicted shifts; the poor predictions of H6 and H7 are even overestimated. In this case the variation (s_{MCMD}) between different predictions is fairly large (0.04−0.14 ppm), mainly because of the two protons.

3.7. Some Remarks about Outliers.

The lesson of the above example is that in the cases with many rotatable bonds and chemical shift anisotropy, it is difficult to get an unambiguous prediction for the shifts. Because the solvent effects are taken into account in a rather primitive way, both in the modeling and prediction, it is obvious that the prediction is more or less qualitative. This is, however, useful and sufficient in this kind of cases and helps to understand the conformational nature and behavior of the system in solution, and, anyway, the prediction can be used to estimate the chemical shifts in the energetically feasible conformations which can then be compared to the observed ones. A similar conclusion was reached in a work for flavonoid glycosides,[42] for which the most probable conformer could be concluded much in the same manner as here.

The statistics reported above is likely near-to-correct for relatively small molecules with common structural features. On the other hand, the rrms of 0.116 ppm is obviously optimistic for structures with exotic functional groups, with many flexible moieties, and ring systems having several stable conformations, for example. A poor prediction may reflect some unusual special features in the chemistry of the system. As a rule, proton chemical shifts with the observed-predicted difference larger than 0.25 ppm should be checked and the difference explained: in many cases the reason for the difference is interesting, for example a special solvent effect, molecular dimerization equilibrium, or conformational/tautomeric equilibrium which is not reproduced by prediction or is not obvious by the molecular model.

3.8. Spectral Parameter Prediction and Quantum Mechanical Spectral Analysis (QMSA).

The modern software for QMSA allows the analysis of accurate spectral parameters of almost any ^{1}H NMR spectrum if the spectrum is not too overcrowded and the structure does not contain long $(CH_2)_n$ systems, the spectral parameters of which would not be interesting anyway.[43] One might say that the symbiosis of the NMR parameter prediction and QMSA represents another landmark in this history of QMSA initiated by the program LAOCOON3.[44]

It has been recently suggested[21,42] that the assignments of novel compounds to be published in the future should be checked by chemical shift prediction. As to the above scopolamine example,[41] it can be appended to the gallery of examples[21] where chemical shift prediction would have been useful. We want to add here that also the spectral parameters analysis using computerized QMSA would reduce the number of erroneous spectral parameters considerably and clarify interpretations in the literature. In practice, the QMSA protocol can be faster, after it is adopted properly, than the traditional way where the chemical shifts are estimated manually and the signals classified into singlets, doublets, and other types of multiplets. The chemical shift could then be given with an accuracy of at least 0.001 ppm, instead of the 0.01 ppm which is common nowadays. Although the solvent and concentration effects may be considerably larger than 0.01 ppm, the accuracy with three digits is relevant because the medium effects are smaller on the relative shifts. It should be noted that QMSA routinely gives proton chemical shifts with an accuracy of 0.01 Hz or better,[43] corresponding to 0.0002 ppm at 500 MHz. If the concentration and temperature are known, even this kind of accuracy becomes relevant.

For coupling constants the QMSA accuracy is usually better than 0.05 Hz, if the line width is normal and even in the presence of many long-range couplings.[43] The good accuracy of the couplings makes them diagnostic, for example, in analysis of complex mixtures. The spectral parameters, together with predicted parameters and molecular coordinates, can be stored in compact text files and then used to simulate the spectrum at any field and with the current line-shape.[43] Moreover, the effects of the conditions like solvent, concentration, and simple chemical modifications like methylation could then be predicted with a fair accuracy. The spectral libraries obtained collecting, for example, such metabolite spectral parameter files, have been called Adaptive Spectral Libraries (ASL).[43,45]

3.9. Estimation of Prediction Errors and QMSA.

A fundamental property of the traditional QMSA iterator algorithms is that they are not able to change the order of chemical shifts, especially if the signals are strongly coupled to each other. They converge to the correct solution usually only if the trial order of the shifts is correct. From the point of the iteration, the prediction errors (arising from concentration shifts) which do not affect the order of the predicted shifts are not serious. In the PERCHit iterator and the Automated Consistency Analysis (ACA) of PERCH Software[17] the potential orders of the shifts are generated by a cost analysis in which the confidence limits of the shifts become a critical measure and the number of the potential solutions may grow quickly when the limits grow. Therefore, the reliability of the prediction confidence limits becomes an essential property for the structure verification based on QMSA.

Because of the nature of the present LPNC algorithm, there is no simple explicit expression for an estimate of the accuracy of each prediction. By developing a linear model based on the original descriptors and the same similarity parameters as used in the kNN corrections and the LSO data we obtained a model, which estimated the accuracy of the prediction with R^2 of 0.58. As shown above, in the case of scopolamine the predicted errors are in fair agreement with the observed deviations.

If the range of a predicted chemical shift is large, it means from the point of view of QMSA that in a crowded spectral part its place may be virtually unknown. Therefore it would be important to be able to recognize such predictions. In MODEL RF10/10 LSO analysis gives ca. 0.7% predictions in which the prediction error is larger than 0.50 ppm, and its estimated value is larger than 0.40 ppm (which is considered a success for use in QMSA). On the other hand, the number of false positives for which the prediction error was >0.50 ppm, and its estimated value smaller than 0.40 ppm was 0.6% of the total. One could say also that more than 50% of the poor predictions are discovered by the error analysis.

4. CONCLUSION

In this work LPNC (Linear Prediction with Nonlinear Corrections) was developed and assessed to the chemical shift prediction of protons on sp^3 carbons. Our results indicate that the RF method (for nonlinear effects) and the kNN algorithm (for recognizing of common chemical environments) offer a good alternative for the nonlinear corrections. The LSO rrms of the plain linear model (corresponding to RF10/10) is 0.168 ppm, and after the corrections, which all can be termed more or less nonlinear corrections to the regression model, the rrms is reduced to 0.116 ppm. The large number of the required terms and corrections reflects the nonlinearity of the proton chemical shift models. A part of the corrections have a real physical correspondence (like the terms related to proton types) which they represent, but a part obviously just compensates the bias in the modeling, conformational analysis, and dynamics. A somewhat unexpected result was that, with a few exceptions, the 4D model was not superior to the MC model. On the other hand, this means that the much faster procedure can be recommended in most of the cases, excluding those in which the fourth dimension forms a key question like in our example, scopolamine.

As turned up in this work, it is difficult to obtain unambiguous measures for the quality of the prediction algorithm. Although the total rrms may look good, the variation between the proton types renders the total rrms a poor estimate of the goodness of the prediction. For example, the rrms and MAE estimates (0.148−0.201 ppm and 0.084−0.138 ppm, for the RF10/10 model) for the four CH classes cannot be considered satisfactory, yet. On the other hand, if the same structure or a very similar chemical environment is found in the predictor database one can expect that the shifts are predicted within 0.02−0.03 ppm standard deviation, depending on concentration of the sample from which the spectrum has been recorded, or better if the prediction is based on a match of structures. One should note that while a prediction for a rigid structure or a single conformation can be accurate, due to the contingency of the MC and MD protocols, even the different MC and 4D predictions for a structure may vary much more, as for scopolamine. Anyhow, the single conformer predictions can be valuable in mapping the conformational alternatives of a conformationally complex system.

To summarize the statistics, while the prediction rrms was <0.10 ppm for 78% of the LSO shifts, the rrms was >0.25 in 5% and >0.50 ppm in 0.7% of the cases (Table 3). In our LSO validation we obtained total rrms which varied from 0.105 ppm (RF0/0 in Table 2) to 0.129 ppm for the model (RF10/100), in which we allowed even complete removal of some types of structures from the teaching data but which were then predicted. The corresponding MAEs were 0.068 and 0.083 ppm. We were also able to estimate the standard deviations arising from the experimental data and, separately, the concentration shifts (ca. 0.024 ppm). When we add the other sources of errors (wrong assignments, etc., discussed above) in our database, we may conclude that the “effective” rrms estimates (thinking QMSA which is an important application of the ^1H chemical shift prediction) are ca. 0.01 ppm smaller than the numbers given above. For QMSA, it is important that also the confidence limits for the predictions are given with a fair reliability ($R^2 = 0.58$).

The major limiting factors for further improvement of the proton chemical shift prediction are obviously related to the diversity and quality of the available experimental data and the conformational mapping in solution. In general, the proton chemical shift prediction, as combined with QMSA, offers an invaluable aid for ^1H NMR spectral analysis, structure verification, and conformation analysis.

■ AUTHOR INFORMATION

Corresponding Author

*E-mail: reino.laatikainen@uef.fi.

Notes

The authors declare no competing financial interest.

■ ACKNOWLEDGMENTS

This work was supported by the National Graduate School of Organic Chemistry and Chemical Biology.

■ REFERENCES

(1) Blinov, K. A.; Smurnyy, Y. D.; Churanova, T. S.; Elyashberg, M. E.; Williams, A. J. Development of a Fast and Accurate Method of ^{13}C NMR Chemical Shift Prediction. *Chemom. Intell. Lab. Syst.* **2009**, *97*, 91−97.

(2) Bürgin-Schaller, R.; Arnold, C.; Pretsch, E. New Parameters for Predicting ^1H NMR Chemical Shifts of Protons Attached to Carbon Atoms. *Anal .Chim. Acta* **1995**, *312*, 95−105.

(3) Abraham, R. J. A Model for the Calculation of Proton Chemical Shifts in Non-Conjugated Organic Compounds. *Prog. Nucl. Magn. Reson. Spectrosc.* **1999**, *35*, 85−152.

(4) Griffits, L. Towards the Automatic Analysis of ^1H NMR Spectra. *Magn. Reson. Chem.* **2000**, *38*, 444−451.

(5) Aires-de-Sousa, J.; Hemmer, M. C.; Gasteiger, J. Prediction of ^1H NMR Chemical Shifts Using Neural Networks. *Anal. Chem.* **2002**, *74*, 80−90.

(6) 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*, 940−945.

(7) Binev, Y.; Corvo, M.; Aires-de-Sousa, J. The Impact of Available Experimental Data on the Prediction of the ^1H NMR Chemical Shift by Neural Networks. *J. Chem. Inf. Comput. Sci.* **2004**, *44*, 946−949.

(8) Smurnyy, Y. D.; Blinov, K. A.; Churanova, T. S.; Elyashberg, M. E.; Williams, A. J. Toward More Reliable ^{13}C and ^1H Chemical Shift Prediction: A Systematic Comparison of Neural-Network and Least-Squares Regression Based Approaches. *J. Chem. Inf. Model.* **2008**, *48*, 128−134.

(9) Kuhn, S.; Egert, B.; Neumann, S.; Steinbeck, C. Building Blocks for Automated Elucidation of Metabolites: Machine Learning Methods for NMR prediction. *BMC Bioinf.* **2008**, *9*, 400.

(10) Pretsch, E.; Bühlmann, P.; Affolter, Ch. *Structure Determination of Organic Compounds: Tables of Spectral Data*, 3rd Completely Revised and Enlarged English ed.; Springer-Verlag: Berlin, 2000.

(11) Abraham, R. J.; Mobli, M. The Prediction of ^1H NMR Chemical Shift in Organic Compounds. *Spectrosc. Eur.* **2004**, *16*, 16−22.

(12) Davies, A. N. Who Has the Best Proton NMR Crystal Ball? *Spectrosc. Eur.* **2008**, *20*, 21−23.

(13) Advanced Chemistry Development, Inc. http://www.acdlabs.com (accessed Jan 27, 2014).

(14) CambridgeSoft Corp. http://www.cambridgesoft.com (accessed Jan 27, 2014).

(15) Mestrelab Research. http://www.mestrelab.com (accessed Jan 27, 2014).

(16) Modgraph Consultants. http://www.modgraph.co.uk (accessed Jan 27, 2014).

(17) PERCH Solutions Ltd. http://www.perchsolutions.com (accessed Jan 27, 2014).

(18) nmrshiftdb2. http://nmrshiftdb.nmr.uni-koeln.de (accessed Jan 27, 2014).

(19) Mihaleva, V. V.; te Beek, T. A. H.; van Zimmeren, F.; Moco, S.; Laatikainen, R.; Niemitz, M.; Korhonen, S.-P.; van Drie, M. A.; Vervoort, J. MetIDB: A Publicly Accessible Database of Predicted and Experimental ^1H NMR Spectra of Flavonoids. *Anal. Chem.* **2013**, *85*, 8700−8707.

(20) Jain, R.; Bally, T.; Rablen, P. R. Calculating Accurate Proton Chemical Shifts of Organic Molecules with Density Functional Methods and Modest Basis Sets. *J. Org. Chem.* **2009**, *74*, 4017−4023.

(21) Lodewyk, M. W.; Siebert, M. R.; Tantillo, D. J. Computational Prediction of ^{1}H and ^{13}C Chemical Shifts: A Useful Tool for Natural Product, Mechanistic, and Synthetic Organic Chemistry. *Chem. Rev.* **2012**, *112*, 1839−1862.

(22) Cheshire. www.cheshirenmr.info (accessed Jan 27, 2014).

(23) Spanton, S. G.; Whittern, D. The Development of an NMR Chemical Shift Prediction Application with the Accuracy Necessary to Grade Proton NMR Spectra for Identity. *Magn. Reson. Chem.* **2009**, *47*, 1055−1061.

(24) SDBS database. http://www.sdbs.riodb.aist.go.jp/ (accessed Jan 27, 2014).

(25) Tynkkynen, T.; Hassinen, T.; Tiainen, M.; Soininen, P.; Laatikainen, R. ^{1}H NMR Spectral Analysis and Conformational Behavior of *n*-Alkanes. *Magn. Reson. Chem.* **2012**, *50*, 598−607.

(26) Halgren, T. A. Merck Force Field. 1. Basis, Form, Scope, Parameterization, and Performance of MMFF94. *J. Comput. Chem.* **1996**, *17*, 490−519.

(27) ghemical. http://bioinformatics.org/ghemical (accessed Jan 27, 2014).

(28) Allen, L. C. Lewis-Langmuir Atomic Charges. *J. Am. Chem. Soc.* **1989**, *111*, 9115−9116.

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

(30) Breiman, L. Random Forests. *Machine Learning* **2001**, *45*, 5−32.

(31) Random Forests. http://www.stat.berkeley.edu/~breiman/RandomForests/ (accessed Jan 27, 2014).

(32) Random Forests for MATLAB. http:///code.google.com/p/randomforest-matlab/ (accessed Jan 27, 2014).

(33) Lehtivarjo, J.; Hassinen, T.; Korhonen, S.-P.; Peräkylä, M.; Laatikainen, R. 4D Prediction of ^{1}H Chemical Shifts. *J. Biomol. NMR* **2009**, *45*, 413−426.

(34) Lehtivarjo, J.; Tuppurainen, K.; Hassinen, T.; Laatikainen, R.; Peräkylä, M. Combining NMR Ensembles and Molecular Dynamics Simulations Provides More Realistic Models of Protein Structures in Solution and Leads to Better Chemical Shift Prediction. *J. Biomol. NMR* **2012**, *52*, 257−267.

(35) Devillers, J. A New Strategy for Using Supervised Artificial Neural Networks in QSAR. *SAR QSAR Environ. Res.* **2005**, *16*, 433−442.

(36) Hastie, T.; Tibshirani, R.; Friedman, J. *The Elements of Statistical Learning*; Springer: New York, 2008; Chapter 15.

(37) Zhang, H.; Wang, M. Search for the Smallest Random Forest. *Statistics and Its Interface* **2009**, *2*, 381−388.

(38) Mitchell, M. W. Bias of the Random Forest Out-of-Bag (OOB) Error for Certain Input Parameters. *Open J. Statistics* **2011**, *1*, 205−211.

(39) Rao, R. B.; Fung, G.; Rosales, R. On the dangers of cross-validation. An experimental evaluation. In *Proceedings of the 2008 SIAM International Conference on Data Mining*; 2008; pp 588−596.

(40) Tormena, C. F.; Freitas, M. P.; Rittner, R.; Abraham, R. J. A ^{1}H NMR and Molecular Modelling Investigation of Diastereotopic Methylene Hydrogen Atoms. *Magn. Reson. Chem.* **2002**, *40*, 289−283.

(41) Sarazin, C.; Goethals, G.; Seguin, J.-P.; Barbotin, J.-N. Spectral Reassignment and Structure Elucidation of Scopolamine Free Base Through Two-Dimensional NMR Techniques. *Magn. Reson. Chem.* **1991**, *29*, 291−300.

(42) Riihinen, K. R.; Mihaleva, V. V.; Gödecke, T.; Soininen, P.; Laatikainen, R.; Vervoort, J. M.; Lankin, D. C.; Pauli, G. F. ^{1}H-NMR Fingerprinting of *Vaccinium vitis-idaea* Flavonol Glycosides. *Phytochem. Anal* **2013**, *24*, 476−483.

(43) Laatikainen, R.; Tiainen, M.; Korhonen, S. P.; Niemitz, M. Computerized Analysis of High-resolution Solution-State Spectra. In *Encyclopedia of Magnetic Resonance*; Harris, R. K., Wasylishen, R. E., Eds.; John Wiley: Chichester, 2011.

(44) Castellano, S.; Bothner-By, A. A. Analysis of NMR Spectra by Least Squares. *J. Chem. Phys.* **1964**, *41*, 3863−3869.

(45) Tiainen, M.; Maaheimo, H.; Niemitz, M.; Soininen, P.; Laatikainen, R. Spectral Analysis of ^{1}H Coupled ^{13}C Spectra of the Amino Acids: Adaptive Spectral Library of Amino Acid ^{13}C Isotopomers. *Magn. Reson. Chem.* **2008**, *46*, 125−137.