On the Use of ^{1}H and ^{13}C 1D NMR Spectra as QSPR Descriptors

E. L. Willighagen, H. M. G. W. Denissen, R. Wehrens, and L. M. C. Buydens*

Institute for Molecules and Materials, Radboud University Nijmegen, Toernooiveld 1, NL-6525 ED Nijmegen, The Netherlands

Received July 11, 2005

Recently, 1D NMR and IR spectra have been proposed as descriptors containing 3D information. And, as such, said to be suitable for making QSAR and QSPR models where 3D molecular geometries matter, for example, in binding affinities. This paper presents a study on the predictive power of 1D NMR spectra-based QSPR models using simulated proton and carbon 1D NMR spectra. It shows that the spectra-based models are outperformed by models based on theoretical molecular descriptors and that spectra-based models are not easy to interpret. We therefore conclude that the use of such NMR spectra offers no added value.

INTRODUCTION

After several decades, methodological research on quantitative structure activity/property relationship (QSAR and QSPR) modeling still receives much attention.[1] Focus has been both on new modeling methods (e.g., support vector regression[2]) and on describing the molecular structures. Even though many theoretical molecular descriptors have been developed in the past to represent molecular structures in mathematical models, new descriptors are being introduced every day. While some descriptors are more useful in some applications, no general descriptor type is available that can be used for all QSAR/QSPR studies.

Descriptors capture certain features of the molecular structure and are often categorized into descriptor classes according to the information they represent.[3] The first class of descriptors, including the Wiener index and the Kier shape descriptors, represents topological properties of a molecule. These only describe the connectivity and not the geometry. The second class represents descriptors that describe geometrical properties and contains descriptors such as WHIM descriptors and solvent-accessible surface areas. Such descriptors are often named 3D descriptors, while the former are 2D descriptors. The third class of descriptors contains the electronic descriptors, describing the electronic features of the molecules. Examples include the HOMO and LUMO energies and electronegativity. The fourth and last class of descriptors contains features derived from chemical formula, like atom counts.

While such a classification is somewhat artificial, the notion that a descriptor may represent geometrical information instead of just topological information is important. If the modeled activity is highly depending on the 3D geometry of the molecule, which is, for example, the case with binding affinities, the descriptors need to represent geometrical features of the molecules. When the 3D geometry is relatively unimportant, for example, in the case of solubility, then such features need not be present in the descriptor set in order to obtain predictive models.

* Corresponding author tel: +31 24 36 53192; fax: +31 24 36 52653; e-mail: L.Buydens@science.ru.nl.

Recently, IR and 1D NMR spectra have been proposed as 3D molecular descriptors[4] in QSAR modeling. Both spectra types show unique spectra for different compounds. Moreover, these spectra depend on the 3D geometry of the molecules, which can, for example, be seen with the low-temperature NMR spectrum of cyclohexane where the axial and equatorial hydrogens show different chemical shifts. Additionally, the through space spin−spin coupling in proton NMR is used as a restriction in elucidating 3D protein structures. From these examples it can be concluded that spectra indeed contain 3D information, but unlike grid-based representations, such as CoMFA,[5] spectra do not require molecular alignment prior to analysis, simplifying the model building considerably. It is questionable, though, whether this 3D information is useful and relevant for modeling the activities or properties.

QSAR and QSPR models correlate molecular structures with a measured activity or property using numerical descriptors, attempting to capture the relation between the chemical and physical information in the descriptors with that activity. When modeling water solubilities or partition coefficients, the model will focus on descriptors describing features that have a high influence, positively or negatively, on the activity. Consequently, when using NMR spectra as descriptors, the modeling method will find shift areas that correlate with the activity. For example, if ^{1}H NMR are used as descriptor, the peak shift areas where phenyl protons are found are expected to negatively correlate with the water solubility and positively with the octanol/water partition coefficient.

^{1}H NMR and ^{13}C NMR spectra have been used in several QSAR and QSPR studies.[6−16] Three different methods have been used in those studies to include NMR descriptors, although other approaches can be considered too. Most used is the whole spectrum approach.[4,6−14] As explained in the previous paragraph, shift areas will then correlate with the modeled activity. Optionally, specific features of the spectrum can be selected, for example, a few areas where relevant information is found.

A second method that uses NMR spectra uses the chemical shift of an atom common to all compounds.[15,17] The

10.1021/ci050282s CCC: $33.50
Published on Web 01/26/2006

advantage of this method is that it explicitly focuses on information relevant to the problem; for example, when modeling chemical reactivity, one can take the chemical shift of an atom close to the reactive center. Obviously, this method is restricted to homologous compound series, and the peaks need to be assigned, restricting its general use.

A third method that uses NMR spectral information is specific for modeling the logarithm of the partition coefficient between octanol and water (LogP).[16] In this research, advantage was taken from the fact that compounds have different NMR spectra in the two solvents. By summing the differences in chemical shifts for the atoms in two solvents, an estimate is made of the solvent effects on the whole molecule. This difference was used to model the activity, although the influence on the predicted activity is rather small, if significant.

Generally, small data sets were used in these QSAR and QSPR studies, in many cases without independent test sets, making it hard to study the true predictive power of the constructed models. The current paper studies the potential of the proposed use of simulated ^{1}H NMR and ^{13}C NMR spectra as molecular descriptor and compares it with theoretically calculated molecular descriptors, derived from a symbolic representation of the molecules, in this case the connection table. Three data sets are used, of which three contain a diverse set of more than 100 compounds, and have physical properties as end points. For these data sets, any possible 3D information in the descriptor is unlikely to be important. Results for a fourth data set with binding affinities as end point (used in the original NMR QSAR paper[4]), for which such 3D information would be important, have been left out because modeling the activity was unsuccessful with any descriptor used. For all data sets an independent test set is used to be able to estimate the true predictive power. As in most relevant literature, we used full spectra: it does not require peak assignment, nor one or more atoms to be common to all compounds. Other approaches used in the literature did not show clear advantages over the full spectrum approach and are not further considered in this paper.

EXPERIMENTAL SECTION

Methods. 1D NMR spectra have been simulated with ACDs ^{1}H Predictor and ^{13}C Predictor version 7.0. Proton NMR spectra were scaled to a resolution of 0.05 ppm per data point in the range of 0 to 11 ppm using custom scripts, resulting in 220 variables. Likewise, carbon NMR spectra were scaled to a resolution of 1 ppm in the range of 1 to 220 ppm, also giving 220 variables.

Theoretical molecular descriptors are calculated with Dragon 5, although alternatives are abundant including open-source variants such as JOELib and the CDK.[18,19] Binary and constant descriptors are removed, resulting in about 1200 to 1300 descriptors, depending on the data set, from which 220 descriptors were randomly selected to give a descriptor set with the same number of variables are the NMR sets. We used models based on these descriptors for benchmarking only, because it was not our goal to make optimal models based on these descriptors. Therefore, we explicitly did not do featureselection on these descriptors, as is usually done. Columns were autoscaled in order to make each descriptor equally important.

The Dragon 5 program defines 20 different descriptor classes. Replicate random selections for the data sets at least 18 of all descriptor classes represented (not shown). This indicates that the used subset of 220 descriptors has a high diversity in information content including constitutional, topological, connectivity, geometrical descriptors, and many others covering molecular properties that correlate with dipole moment, weight, and hydrophobicity. For completeness, the random selections for the three data sets used to calculate the presented results are found in the Supporting Information.

The amount of information in the **X** matrices for the descriptor sets is first studied by investigating the mathematical ranks of those matrices. The maximum rank equals the lower value of the rows and columns of the matrix. A matrix rank lower than this maximum indicates correlation in the matrix in either the columns or the rows. By comparing ranks for the different descriptor types, the differences can only be caused by correlation between columns.

Partial least squares (PLS)[20] was then used to make mathematical models that relate the molecular descriptor set (**X** matrix, consisting of either spectra or theoretical descriptors) with the activity (**Y** vector). To pick the number of latent variables for the model, we used the root mean square error (RMSE) of leave-one-out cross validation (LOOCV). This is done using an automatic procedure that picks the lowest number of LVs that has a cross validation error that is lower than one standard deviation above the absolute minimum in that error.[21] This might not be the optimal decision, as choosing the best number of LVs is a difficult problem, but at least it is conservative and consistent.

To validate the performance of the different types of descriptors, several statistics are monitored that describe the differences in predicted and real activity: in QSAR/QSPR research commonly used, the R^2 and Q^2 [22] and the root mean square error of cross validation (RMSECV) and of prediction (RMSEP). The RMSEP is used to get a independent estimate of predictive power of the model for unknowns. For each data set, five random divisions in training and test sets have been used to get an estimate on the errors on these statistics due to these divisions.

The RMSE values for the models are compared with a no-information limit, which is calculated from the activities for a data set. It considers a QSPR model where the predicted activity is the mean activity for all compounds in the data set (i.e., $y_{\text{pred}} = \bar{y}$). Obviously, the RMSE of a truly predictive PLS models should be significantly lower than this limit.

Calculations have been performed in the statistical program R 2.1.0[23] on a dual AMD64 processor system running the 64 bits Debian GNU/Linux 3.1 (sarge) operating system. The pls.pcr package was used for building the PLS models.[24]

Data. This paper presents the results of three data sets. These data sets were used to compare the power of NMR spectra in QSPR modeling to theoretical molecular descriptors. The first data set, called WS, contains 431 compounds with aqueous solubilities. This set is a subset of a published test set that was selected on diversity.[25] Models were trained with 400 compounds, and the remaining 31 compounds were used as test set. The second data set, called BP, contains 269 heteroatom-containing compounds excluding nitrogen compounds (data set II from ref 26) with associated boiling points. Eight compounds from the original data set lacking

Table 1. Median Ranks of Five Randomly Chosen Training Sets for Descriptor Types for the Three Data Sets[a]

	^{1}H NMR	^{13}C NMR	Dragon	limit
WS	198	195	219	220
BP	163	157	219	220
LogP	120	117	119	120

[a] The limit is the maximal rank possible for that descriptor type and data set. Clearly, Dragon-based descriptor matrices are always of nearly full rank, which indicates a high amount of uncorrelated information.

any hydrogens were removed. A test set with 42 compounds was used, while training models was done with 227 compounds. The third data set, called LogP, contains 154 compounds with associated LogP values,[16] the partition coefficients between octanol and water. Models were trained with 120 compounds, and the remaining 34 compounds were used as test set. Activities and InChI values for these three data sets can be found in the Supporting Information.

RESULTS

Data Rank. The median ranks of the training X matrices for the five random training/test set divisions are shown in Table 1. For all data sets, the rank for the Dragon descriptor set was found to be equal or close to the minimum of the number of rows and columns of the matrix. For proton NMR the rank was lower, and for carbon NMR the rank was lowest. This indicates that carbon NMR shows most correlation. Dragon descriptors show the least correlation of all three descriptor types. Less correlation does not directly mean better PLS models, though. That the ranks for the NMR spectra are lower than the maximum is not surprising. Spectra normally have shift areas where no peaks are found. Those matrix columns have zero intensity for all compounds, and are obviously correlated.

Predictivity. The RMSECV plots to select the number of latent variables for the three descriptor types typically look like those for the LogP data set shown in Figure 1. Error plots for carbon NMR and Dragon models show that the RMSECV drops with the first few number of latent variables, after which it stabilizes and then increases. This can be explained by assuming that the first few LV's add information to the model, after which the model starts to be overtrained. For Dragon-based models, typically 6 or 7 LV's are chosen and for carbon NMR-based models typically 3 or 4 LV's are chosen. The error plots for proton NMR look different: the error rises from the first or second LV on. For this descriptor type, only one or two LV's are chosen.

The performance of the models is studied using several statistics. Five replicate training/test set divisions are used

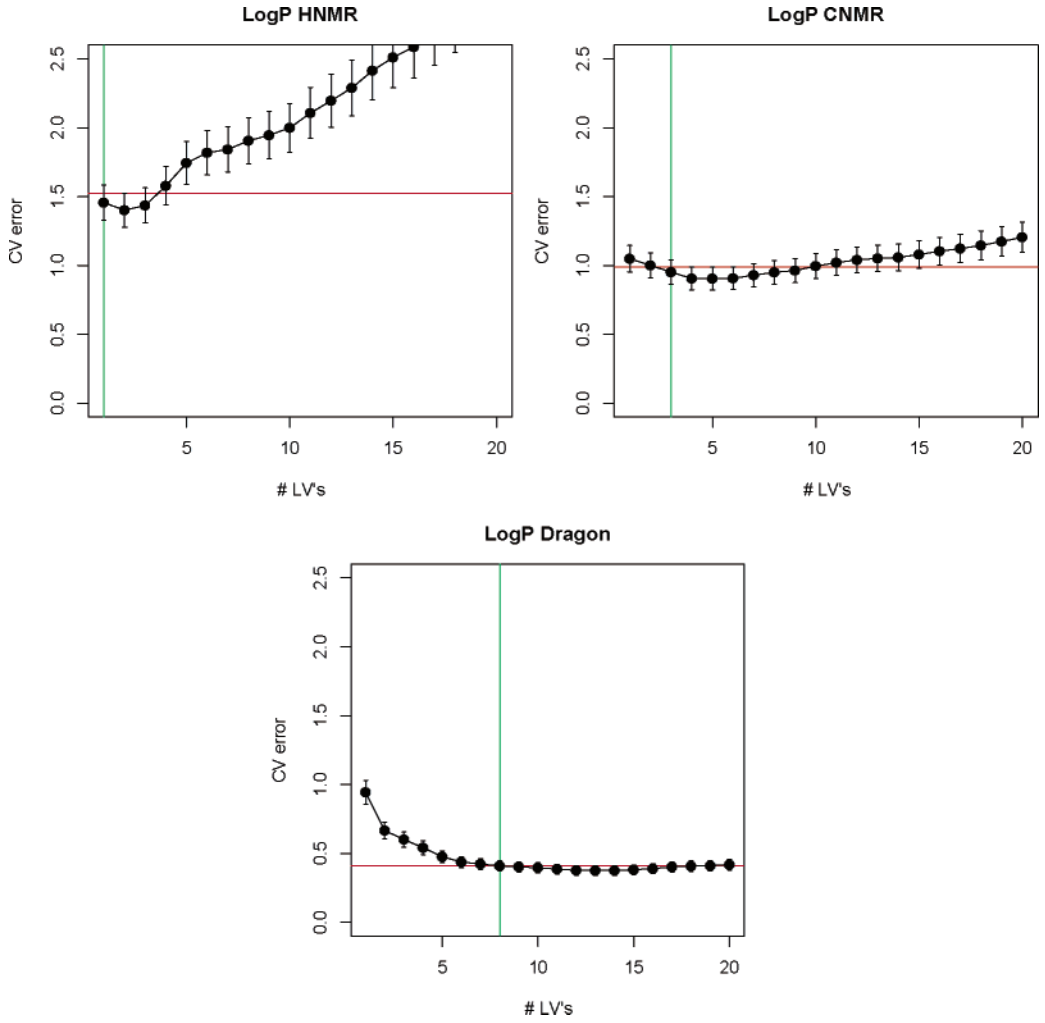

Figure 1. Chosen number of latent variables is based on the LOOCV error. The LogP plots for the three data sets are representative for the other data sets. The red line indicates one standard deviation above the absolute minimum in the LOOCV error used to choose the number of latent variables for the PLS model, which is indicated by the green line.

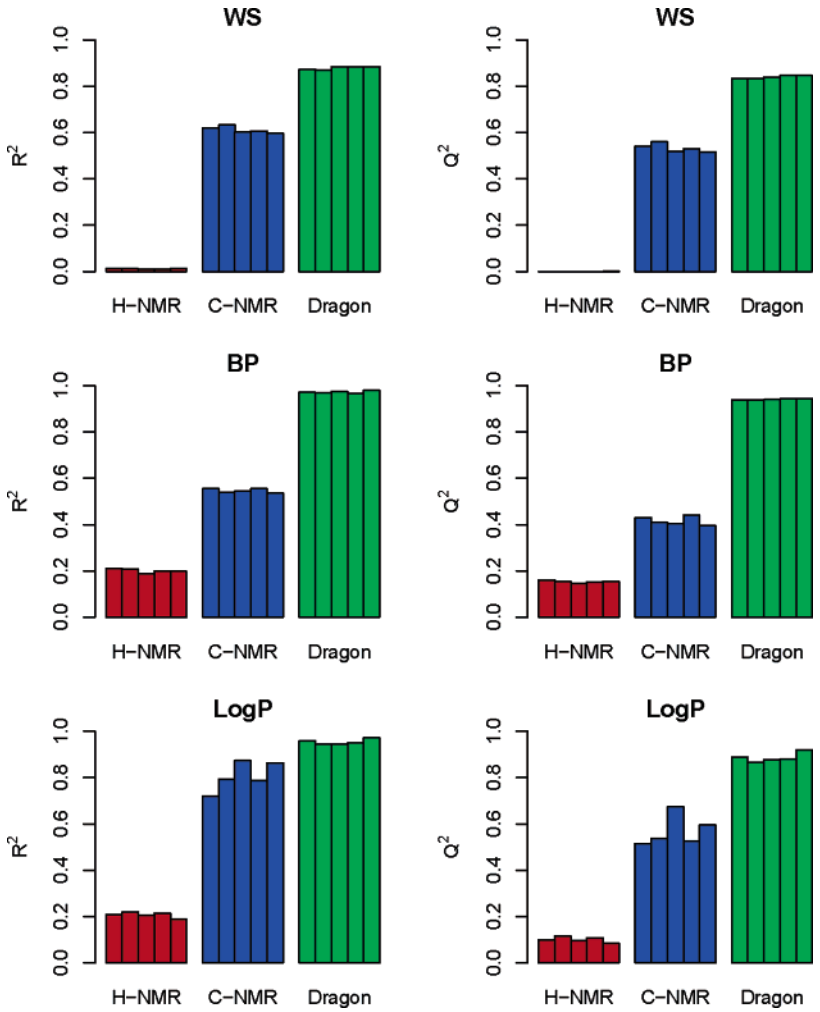

Figure 2. Internal performance statistics R^2 and Q^2 for the three data sets, each with five random training/test set divisions. In the first three cases, the Dragon-based descriptors clearly perform best.

to allow comparing calculated statistics; a small improvement in one of the statistics might not indicate an significant improvement of the model. Taking into account the errors on the statistics is important when picking one model over another.

When looking at the R^2 and Q^2 values (see Figure 2) for the WS, BP and LogP data sets, it is apparent that it was not possible to create acceptable models based on proton NMR, as shown by the low R^2 and Q^2 values. While carbon NMR based models perform reasonably, they are still outperformed by the Dragon-based models which have higher R^2 and Q^2 statistics. Only the Dragon-based models have statistics approaching the optimal value of 1.0. It is also clear that the error due to the choice of the training/test set division is much smaller than the differences between the three descriptor sets. This strengthens our conclusion, that the differences between the descriptor sets are significant.

These results are confirmed by the RMSECV values and the independent RMSEP's for the independent test sets as shown in Figure 3. The RMSE values show that proton NMR in general does not show a prediction performance significantly better than the no-information limit provided by the $y_{pred} = \bar{y}$ model. Also, in agreement with the R^2 and Q^2 statistics, is the observation that carbon NMR performs reasonably, but is outperformed by the Dragon-based models, which clearly have lower prediction errors.

In addition to looking at numerical prediction error differences, one can also look at $y_{measured} - y_{predicted}$ plots. For the WS, BP and LogP data sets, the three plots for the different descriptor sets look similar to those for the LogP data set shown in Figure 4. The recall, i.e., the prediction of the training samples, is plotted with black open circles, and the test set predictions are drawn with red dots. These plots confirm that proton NMR-based models do not improve significantly on the $y_{pred} = \bar{y}$ model. The plot for carbon NMR shows regression around the $y_{pred} = y_{measured}$ line, but the regression is clearly better for the Dragon-based models. The results in Figure 4 are based on one random test set, but are representative for other training/test set divisions.

Model Interpretation. In addition to looking at the predictive power of the models, the explanatory nature of the models is often informative too. In PLS this is done by looking at the regression vectors. In NMR one would expect shift ranges with high positive coefficients, where peaks occur characteristic for molecular fragments, positively affecting the activity; and ranges with high negative coefficients for groups which negatively affect the activity.

Such shift ranges are found for carbon NMR, as shown in Figure 5 for the LogP data set. Chemical shift ranges where peaks are to be expected for molecular fragments with electron withdrawing atoms, like C−O and C═O, have a negative influence on the calculated property. Additionally,

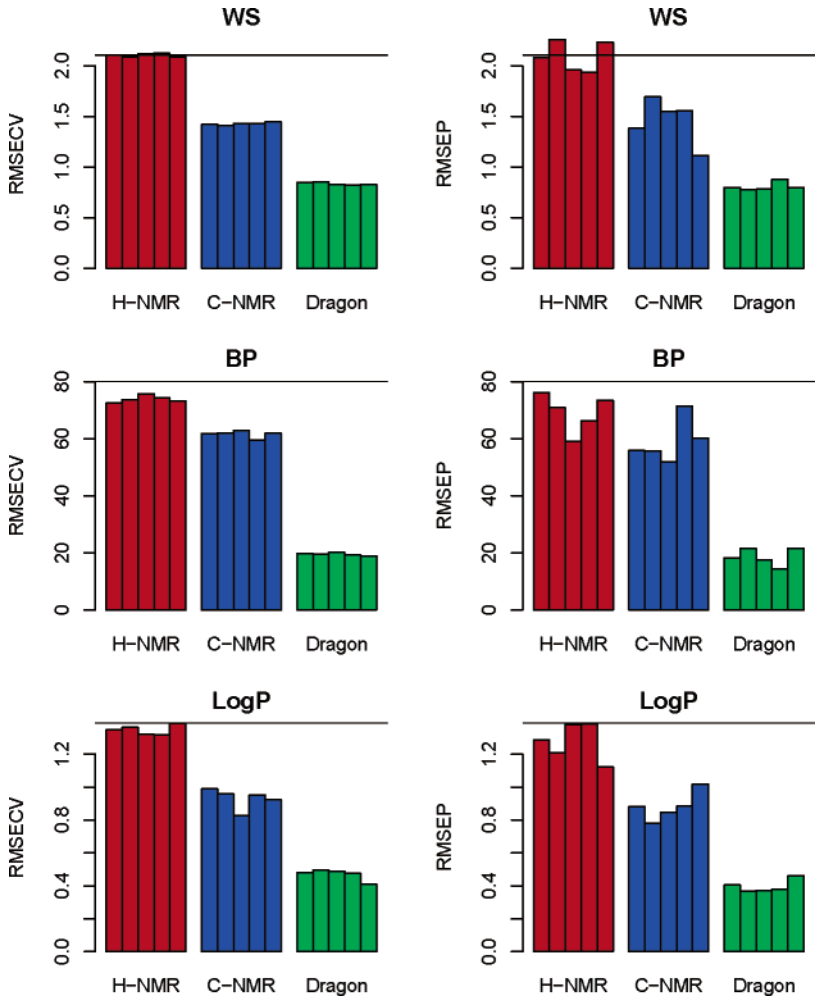

Figure 3. Cross-validation and test set performance statistics RMSECV and RMSEP for the three data sets, each with the same five test sets as in Figure 2. The horizontal line indicates the no-information limit defined by the $y_{pred} = \bar{y}$ model.

ranges where hydrophobic groups, like CHx and C═C, are found, show positive coefficients. The regression coefficients do not seem to provide information beyond the observed influence of these molecular atom groups. The blue lines indicate ± standard deviation for the five random training/test set divisions, and show that the patterns are found for all five replicates.

Proton NMR also seems to show some pattern. Clearly, the area between 3 and 4 ppm has positive contributions. In this area, shifts are expected for protons connected to carbons that bond with heteroatoms, like oxygen and nitrogen, indicating a positive effect of polar groups. This contradicts the interpretation of the PLS coefficients for the carbon NMR models. Moreover, the coefficients are 3 orders of magnitude smaller than those for the carbon NMR models. Even though the regression vector seems to contain information, proton NMR spectra are not predictive.

The regression vector of the Dragon-based model was sorted in ascending order to allow easier interpretation of the significances of the coefficients. As it is not the intention to produce the best possible models based on theoretical molecular descriptors, we will not discuss which individual descriptors had high (positive or negative) coefficients. We do note that for all build models, the 20 descriptors with highest coefficients represent at least 8 different descriptor classes, with an average of 10. It is important to note that all descriptors with high coefficients show this for all five training/test set divisions. From these results, we conclude that by randomly picking 220 descriptors from the larger set, predictive models can be constructed. We anticipate that by carefully selecting descriptors even better predictive models can be built.

DISCUSSION

Important features of a QSAR or QSPR model are its predictive ability and the interpretability. The latter feature is an important tool to help scientists understand the influences of molecular features on the modeled activities. In such cases, the statistical fit is important, and one can focus on training set statistics.[27] An increasingly important application of QSAR and QSPR modeling, however, is virtual screening. For such applications, the predictive power of the model is more important, and just the statistical fit is not enough to characterize the model; an independent test set is then obligatory to estimate the models predictive power. We feel, however, that the use of an independent test set should in both cases be used. It ensures that observed influences of molecular features on the activities are true cause−effect relationships instead of just random correlation.

Spectral areas in NMR spectra are indicative for molecular features but do not offer much information on the important molecular features. This makes the NMR based models not optimal for explanatory purposes. Moreover, the results

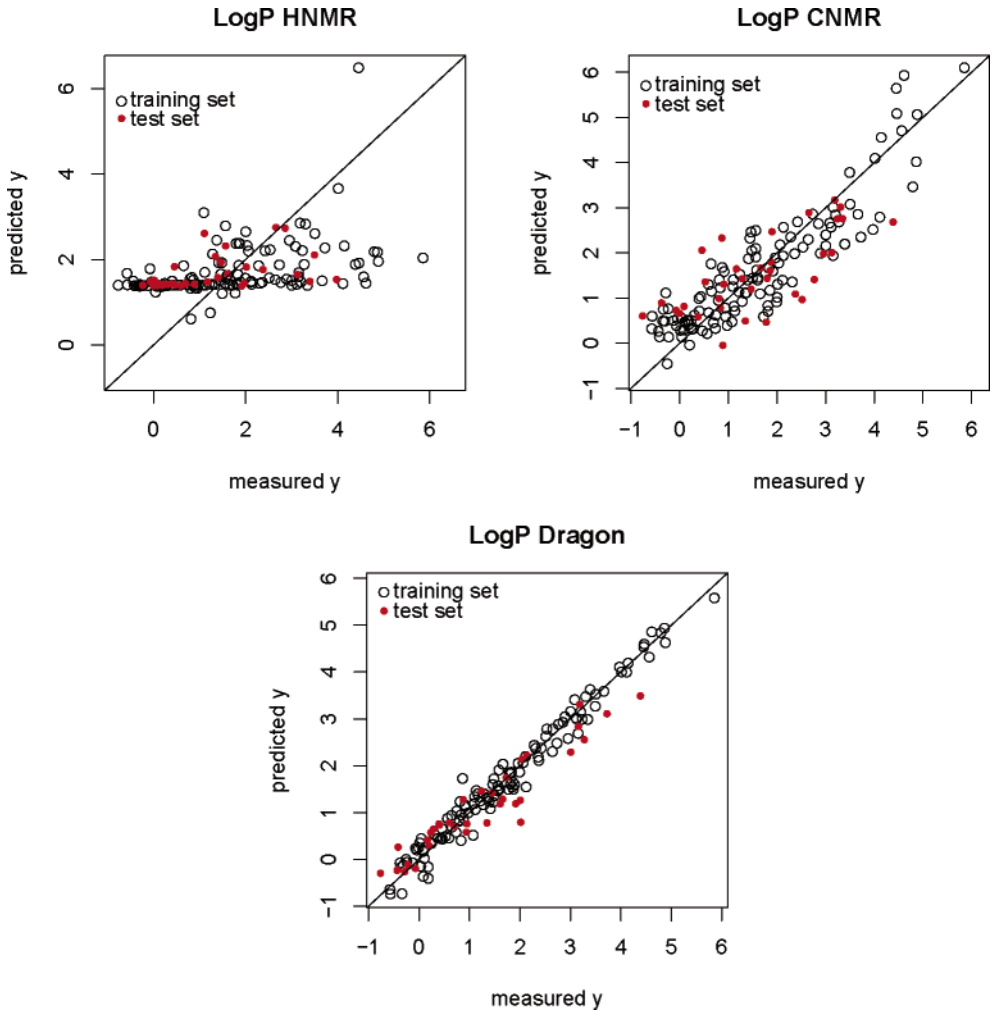

Figure 4. $y_{measured} - y_{predicted}$ plots for the three descriptor types (proton, carbon NMR, and Dragon) for the LogP data set. These plots show that Dragon-based models outperform the NMR-based models: the predicted activities are much closer to the expected values, indicated by the $x = y$ line. These figures are based on one random test set and are typical for other training/test set divisions.

indicate that the predictive power of models based on proton and carbon NMR spectra is not sufficient when compared to models based on theoretical molecular descriptors. For the WS, BP, and LogP data sets, the R^2 and Q^2 statistics and RMSE errors for the Dragon-based models were all favorable as compared to the NMR-based models. The results even indicate that proton NMR-based models do not improve on the null hypothesis model $y_{pred} = \bar{y}$. One possible reason for the inability of PLS to make spectra-based models might be that PLS is a linear regression method unable to model a nonlinear problem well. Unpublished results using support vector machines, classification and regression trees, and wavelength selection did not improve the predictive power of the models.

Comparing the means of the R^2 and RMSEP statistics for the five training/test set divisions with literature values (see Table 2) shows that spectra-based models are inferior to Dragon-based models and models published in the literature. The fact that the statistics for the Dragon-based models are comparable with statistics reported in the literature indicates that PLS in itself is a proper regression method for these data sets.

Although the use of full NMR spectra for proton and carbon nuclei does not give satisfactory results, NMR spectra in general might still be useful. For example, the combination of NMR spectra types has been suggested to improve models,[4] although improvement is not apparent from literature and our own experiments. Moreover, data fusion of two spectra types is not trivial and includes scaling issues. Additionally, NMR spectra of other nuclei (e.g., nitrogen and phosphorus) might be used, but these nuclei are more rarely found in organic compounds and would restrict the applicability of the models, even if they decrease prediction errors. Other approaches are the combination NMR spectra with theoretical descriptors, where scaling issues occur again and the use of spectra derived descriptors, such as the number of chemical shifts or the total sum of shift values. Finally, 2D and 3D methods might provide additional structural

Table 2. R^2 Values for the Carbon NMR-Based and the Dragon-Based PLS Models

		^{13}C NMR	Dragon	reference[a]
WS	R^2	0.61	0.88	0.92[b] [25]
	RMSEP	1.46	0.81	0.59
BP	R^2	0.55	0.97	0.99[26]
	RMSEP	59.0	18.7	7.14
LogP	R^2	0.81	0.95	0.88[16]
	RMSEP	0.88	0.40	*c*

[a] This column is the published R^2 values as reference. [b] The reference value for the WS data set is for a larger data set. [c] No test set was used.

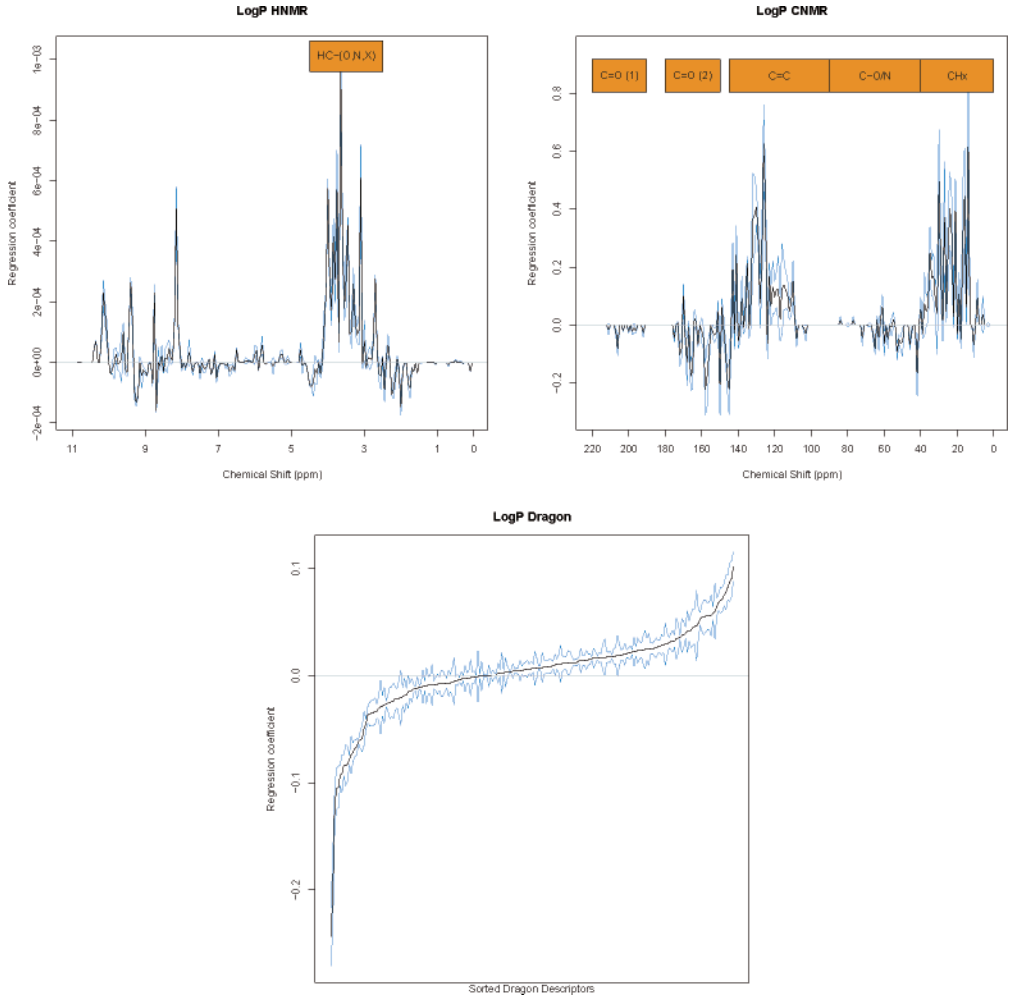

Figure 5. Mean PLS coefficients for the three types of spectra calculated from the five replicates, including ± standard deviation (blue lines). The coefficient vector for the Dragon descriptor set was sorted by size to show more clearly the significance of the most negative and most positive Dragon coefficients.

information that allows better modeling of the activities. Though interesting, such spectra types are, however, beyond the scope of the current QSAR/QSPR literature that uses NMR spectra and will not be further discussed in this paper.

CONCLUSIONS

The predictive powers of the PLS model for the three data sets indicate that proton NMR is not suitable for building QSPR models: the predictive power, as measured by the RMSECV and RMSEP is never better than the $y_{\text{pred}} = \bar{y}$ model, as is clearly visible from the typical $y_{\text{measured}} - y_{\text{predicted}}$ plot of the LogP data set.

Carbon NMR-based models, however, do give acceptable QSPR models as was shown by the prediction errors. Moreover, the regression vectors correlate with areas of relevant molecular fragments, as was exemplified for the LogP data set. However, it was noted that the regression vectors only indicate a few broad chemical shift ranges and do not indicate in detail which molecular features are interesting for modeling the activities.

Importantly, the predictive power of the carbon NMR-based spectra is less than basic Dragon-based models. We did not interpret Dragon descriptors, which were found to be important for the models, but did notice that the training/test set division did not effect the importance of those descriptors. From the fact that Dragon performs better than spectra-based models and that NMR-based models do not offer much information about important molecular features, we conclude that NMR spectra should not be considered as first choice when making predictive models in general and that proton NMR should probably not be used at all.

ACKNOWLEDGMENT

The connection tables of the structures in the LogP data set were kindly provided by Prof. Beger.

Supporting Information Available: The IUPAC International Chemical Identifiers of the molecules in the data sets and the corresponding end points; lists of the used Dragon descriptors for each data set. This material is available free of charge via the Internet at http://pubs.acs.org.

REFERENCES AND NOTES

(1) Livingstone, D. The characterization of chemical structures usign molecular properties. A survey. *J. Chem. Inf. Comput. Sci.* **2000**, *40*, 195−209.
(2) Christianini, N.; Shawe-Taylor, J. *An Introduction to Support Vector Machines*; Cambridge University Press: Cambridge, 2000.
(3) Todeschini, R.; Consonni, V. *Handbook of Molecular Descriptors*; Vol. 11 of Methods and Principles in Medicinal Chemistry; Wiley-VCH: New York, 2000.
(4) Bursi, R.; Dao, Y.; Van Wijk, T.; De Gooyer, M.; Kellenbach, E.; Verwer, P. Comparative spectra analysis (CoSA): spectra as three-

dimensional molecular descriptors for the prediction of biological activities. *J. Chem. Inf. Comput. Sci.* **1999**, *39*, 861−867.
(5) Cramer, R., III; Patterson, D.; Bunce, J. Comparitative molecular field analysis (CoMFA). 1. Effect of shape on binding of steroids to carries proteins. *J. Am. Chem. Soc.* **1988**, *110*, 5959−5967.
(6) Begigni, R.; Passerini, L.; Livingstone, D.; Johnson, M.; Giuliani, A. Infrared spectra information and their correlation with QSAR descriptors. *J. Chem. Inf. Comput. Sci.* **1999**, *39*, 558−562.
(7) Begigni, R.; Giuliani, A.; Passerini, L. Infrared spectra as chemical descriptors for QSAR models. *J. Chem. Inf. Comput. Sci.* **2001**, *41*, 727−730.
(8) Beger, R.; Freeman, J.; Lay, J., Jr.; Wilkes, J.; Miller, D. Use of ^{13}C NMR spectrometric data to produce a predictive model of estrogen receptor binding activity. *J. Chem. Inf. Comput. Sci.* **2001**, *41*, 219−224.
(9) Beger, R.; Wilkes, J. Developing ^{13}C NMR quantitative spectrometric data−activity relationship (QSADR) models of steroid binding to the corticosteroid binding globulin. *J. Comput.-Aided Mol. Des.* **2001**, *15*, 659−669.
(10) Beger, R.; Wilkes, J. Models of polychlorinated dibenzodioxins, dibenzofurans and biphenyls binding affinity to the aryl hydrocarbon recepter developed using ^{13}C NMR data. *J. Chem. Inf. Comput. Sci.* **2001**, *41*, 1322−1329.
(11) Beger, R.; Buzatu, D.; Wilkes, J.; Lay, J., Jr. Comparative structural connectivity spectra analysis (CoSCOSA) models of steroid binding to the corticosteroid binding globulin. *J. Chem. Inf. Comput. Sci.* **2002**, *42*, 1123−1131.
(12) Beger, R.; Buzatu, D.; Wilkes, J. Combining NMR spectral and structural data to form models of polychlorinated dibenzodioxins, dibenzofurans and biphenyls binding to AhR. *J. Comput.-Aided Mol. Des.* **2002**, *16*, 727−740.
(13) 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.
(14) Bailey, N.; Wang, Y.; Sampson, J.; Davis, W.; Whitcombe, I.; Hylands, P.; Croft, S.; Holmes, E. Prediction of anti-plasmodial activity of *Artemisia annua* extracts: application of ^{1}H NMR spectroscopy and chemometrics. *J. Pharm. Biomed. Anal.* **2001**, *41*, 219−224.
(15) Vanderhoeven, S.; Troke, J.; Tranter, G.; Wilson, I.; Nicholson, J.; Lindon, J. Nuclear magnetic resonance (NMR) and quantitative structure−activity relationship (QSAR) studies on the transaceylation reactivity of model 1β-*O*-acyl glucuronides. II: QSAR modelling of the reaction using both computational and experimental NMR parameters. *Xenobiotica* **2004**, *34*, 889−900.
(16) Schnackenberg, L.; Beger, R. Whole-molecule calculation of Log P based on molar volume, hydrogen bonds, and simulated ^{13}C NMR spectra. *J. Chem. Inf. Model.* **2005**, *45*, 360−365.
(17) Khadikar, P.; Sharma, V.; Varma, R. Novel estimation of lipophilicity using ^{13}C NMR chemical shifts as molecular descriptor. *Bioorg. Med. Chem. Lett.* **2005**, *15*, 421−425.
(18) Wegner, J. JOELib. http://joelib.sourceforge.net/, 2005.
(19) Steinbeck, C.; Hoppe, C.; Kuhn, S.; Floris, M.; Guha, R.; Willighagen, E. Recent developments of the chemistry development kit (CDK)−an open-source Java library for chemo- and bioinformatics. *Curr. Pharm. Des.* (in press).
(20) Geladi, P.; Kowalski, B. Partial least-squares regression: a tutorial. *Anal. Chim. Acta* **1986**, *185*, 1−17.
(21) Hastie, T.; Tibshirani, R.; Friedman, J. *The Elements of Statistical Learning*; Springer-Verlag: Heidelberg, 2001.
(22) Golbraikh, A.; Tropsha, A. Beware of q^2!. *J. Mol. Graph. Model.* **2002**, *20*, 269−276.
(23) R Development Core Team: a lanuage and environment for statistical computing. http://www.r-project.org/, 2005.
(24) Wehrens, R. PLS and PCR functions: pls.pcr. http://cran.r-mirror.de/src/contrib/Descriptions/pls.pcr.html, 2005.
(25) Yan, A.; Gasteiger, J. Prediction of aqueous solubility of organic compounds based on a 3D structure representation. *J. Chem. Inf. Comput. Sci.* **2003**, *43*, 429−434.
(26) Goll, E.; Jurs, P. Prediction of the normal boiling points ot organic compounds from molecular structures with a computational neural network model. *J. Chem. Inf. Comput. Sci.* **1999**, *39*, 974−983.
(27) Cronin, M.; Schultz, T. Pitfalls in QSAR. *J. Mol. Struct.* **2003**, *622*, 39−51.

CI050282S