A QSPR Study of the ^{31}P NMR Chemical Shifts of Phosphines

Ramon Bosque and Joaquim Sales*

Departament de Química Inorgànica, Universitat de Barcelona, Martí i Franquès 1, 08028-Barcelona, Spain

Received August 7, 2000

Quantitative structure−property relationships (QSPR) on a large set of descriptors are developed for the ^{31}P NMR chemical shifts of a large set of phosphines. The data set was composed of 291 primary, secondary, and tertiary phosphines, $PH_{3-n}R_n$, including substituents with different steric and electronic characteristics. Multiple linear regression and computational neural networks (CNN) were employed to create the models best suited for the prediction of ^{31}P NMR chemical shifts. A correlation equation including seven descriptors ($R^2 = 0.8619$) is reported. A 7-5-1 CNN was developed that produced a root-mean-error of 9.6 ppm ($R^2 = 0.9513$) for the training set, of 11.7 ppm ($R^2 = 0.8986$) for the cross-validation set, and of 11.3 ppm ($R^2 = 0.9527$) for an external prediction set. The CNN methods give significantly better predictions of ^{31}P NMR chemical shifts for phosphines than the multiple linear regression approach.

INTRODUCTION

Phosphines are probably the most important ligands in coordination chemistry, especially in organometallic compounds. They stabilize low oxidation states of the metal centers and can be used to modify both the electronic and steric properties of their derived coordination compounds. The number of known transition-metal complexes containing phosphine ligands is truly immense and includes examples of monodentate, didentate, tridentate, and higher chelating posphines as well as a great variety of substituents on phosphorus.[1] The behavior of phosphines as ligands is dependent on the electron-donating and electron-accepting properties of the phosphorus atom and of the steric effects of the substituents. To quantify these electronic and steric effects, many ligand parameters have been proposed.[2,3] The early parameters introduced by Tolman[4] in 1977, the cone angle, θ, and the electronic parameter, χ, are the most popular.

These parameters have been used in many structure−activity relationships of coordination and organometallic compounds containing phosphines as ligands. The properties studied are thermodynamic (pK_a values, equilibrium constants, redox potential values), kinetic (log k, activation parameters), spectroscopic (IR, NMR of ^{1}H, ^{13}C, and some transition metals), and structural (bond length). Drago has extended his well-known electrostatic−covalent model (ECW)[5] to these phosphine-containing coordination compounds and has derived the E_B and C_B parameters for nearly 40 phosphines.[6−9] The QALE model (quantitative analysis of ligand effects) has also been applied to this kind of study. In this model, the physicochemical property studied is correlated with the cone angle and the electronic parameter defined previously by Tolman and with an additional parameter called the aryl effect, E_{ar}.[10,11] In general, good correlations are obtained with these models, despite the complexity of the M−P bonds present in these coordination compounds. The nature of the M−P bond, mainly its separation in σ and π electronic and steric components, is a classic controversy in coordination chemistry.[12]

Close correlation[13] has been found between the basicity of free phosphine ligands−expressed as the pK_a in nitromethane of the phosphonium derivatives, HPR_3^+- and the Hammett, σ, and Taft, σ^*, parameters, as well as the Kabachnik constants, σ^ϕ. Close correlation was also found between the pK_a and the phosphorus lone pair ionization potential of several phosphines.[13] However, there is no correlation of ^{31}P NMR chemical shifts with Taft σ^* parameters or with meta and para Hammett parameters.[14] The absence of a correlation between the chemical shift and basicity shows that the inductive effect that governs basicity is not the main influence on the chemical shift of phosphines.[13,14] It should be noted that in both cases, coordination compounds and free phosphines, the correlations studied contain, in general, only a small group of compounds (about 25 entries).

The interpretation of the NMR chemical-shift values for elements different from hydrogen is known to be difficult. For the phosphorus derivatives several factors have been proposed. Three factors, two electronic and one steric, seem to be the most important: the distribution of electron density in the σ bonds between phosphorus and its substituents; the extent to which phosphorus participates in π bonding; the bond angles about the phosphorus.[14,15] For phosphines the electronegativity of substituents on phosphorus and the angles between them are very important variables determining ^{31}P NMR chemical shifts and coupling constants. One empirical method, proposed in the 1960s,[16,17] is available for the calculation of ^{31}P NMR chemical shifts of phosphines by summation of increments that are characteristic of each ligand, using additive constants σ^P for the various groups involved. Different equations are needed for primary, secondary, and tertiary phosphines.

Recently, some work has been published[18,19] dealing with the calculation of phosphorus chemical shifts, using DFT (density functional theory) and the IGLO[20] (individual gauge

* Corresponding author. Fax: 34 934907725. E-mail: joaquim.sales@qi.ub.es.

10.1021/ci000458k CCC: $20.00
Published on Web 12/28/2000

for localized orbital) or GIAO[21] (gauge-including atomic orbitals) approaches. However, due to the complexity of the calculations involved, they have been performed only for a short number of small phosphines.

The quantitative structure−property relationship approach (QSPR) has become very useful in the prediction and interpretation of several physical and chemical properties. The basis of such relationships is the assumption that the variation of behavior of the compounds, as expressed by any measured physical or chemical properties, can be correlated with changes in molecular features of the compounds termed descriptors. While the traditional approach often needs some intuitive vision to derive the relevant mathematical relationship, QSPR methods are based on statistically determined linear or nonlinear functional forms that relate the property of interest with descriptors.[22] Descriptors are numerical values used to describe different characteristics about certain structure in order to yield information about the property being studied.

QPSR studies have been successfully applied to the correlation of many diverse physicochemical properties of chemical compounds. Recently, Katritzky et al. have reviewed the applications of the QSPR approach to technologically relevant physical properties.[23,24] Reaction rates[25,26] of several chemical processes have been also analyzed by QSPR. Related to NMR chemical shifts, Jurs et al. using multiple linear regression and neural networks have predicted ^{13}C NMR chemical shifts of several organic compounds.[27−30]

The ^{31}P NMR chemical shift of phosphines of the type $PH_{3-n}R_n$ is an appropriate property to be studied by QSPR methodology because (i) its numerical value depends on the molecular structure, (ii) the compounds studied have similar molecular structures and the calculated descriptors reflect their changes, and (iii) the ^{31}P NMR chemical shift is very clear, since the NMR spectra of the studied phosphines have only one signal.

In this paper we study a QSPR model for ^{31}P NMR chemical shift prediction for a large variety of phosphines, of the type $PH_{3-n}R_n$. We selected a set of 291 monophosphines with a wide variety of electronic and steric properties. This set is rather wide since it contains primary, PH_2R, secondary, PHR_2, and tertiary phosphines, PR_3. These phosphines contain alkylic, arylic, and alkenyl groups; the substituents present different electronic properties: some are electron donors (OR, NR_2, etc.), and others electron acceptors (F, Cl, Br, OH, NO_2, etc.). These substituents also cause very different steric effects, since some of the alkylphosphines contain groups as bulky as the *tert*-butyl substituent and the arylphosphines contain several different substituents in ortho-, meta-, and para-positions (some examples are depicted in Figure 1). This set contains the most popular monophosphines used in coordination and organometallic chemistry, with a wide range in their electronic and steric properties.

METHODOLOGY

The QSPR studies were developed using the Microsoft Windows version of the CODESSA[31] program and the software system entitled Automated Data Analysis and Pattern recognition Toolkit (ADAPT).[32,33]

Three types of models are generated in this study: type 1, 2, and 3 models, following the terminology introduced by Jurs.[25] Type 1 model is a linear model developed by multiple linear regression analysis; it is the least computationally intensive of the three models. Type 2 model is generated through the use of a computational neural network (CNN); it is a linear/nonlinear hybrid model because it is based on a set of descriptors chosen by multiple linear regression but a nonlinear CNN model is developed from these descriptors. It is important to note that the best set of descriptors chosen to support a linear relationship is not necessarily the best for a nonlinear one. It is advantageous to create type 2 models to study the errors associated with various network architectures. In a type 3 model, the reduced descriptor pool is analyzed with a genetic algorithm to seek the best subset of descriptors to support a nonlinear CNN model. A CNN is used as a fitness evaluator for each subset of descriptors that is selected by the genetic algorithm. A small set of nonlinear models is reported, and further analysis is necessary to differentiate among them. The use of neural networks usually produces the most effective models; they are also the most computationally intensive of the three types.

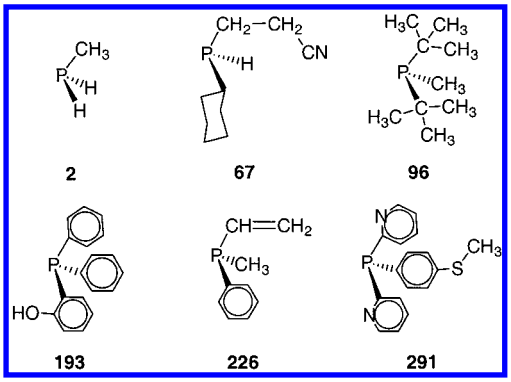

Figure 1. Example structures of phosphines studied.

EXPERIMENTAL SECTION

The CODESSA program[31] was run on a Pentium III personal computer. The computations with the ADAPT program,[32,33] including the feature selection routines (genetic algorithm[34] and simulated annealing)[35,36] and CNN routines,[37] were performed using a Silicon Graphics Indigo 2 workstation, running under UNIX. The generation of the descriptors and the multilinear regression analysis, type 1 model, were performed with the CODESSA program. Neural networks analysis, type 2 and 3 models, were performed with ADAPT routines.

Data Sets. The data set was comprised of 291 phosphines: 29 primary; 38 secondary; 224 tertiary. The ^{31}P NMR chemical shifts data set ranged from −183 to +61 ppm. They were referred to 85% phosphoric acid. The experimental data were obtained mainly from refs 17 and 38. It needs to be emphasized that these values were not obtained under the same experimental conditions of solvent and temperature in all cases. Since the ^{31}P NMR chemical shifts depend on the experimental conditions−concentration, temperature, and solvent−the wide diversity in the origin of the measures may result in an increase of the error between observed and calculated values.

The 291 compound data set was split randomly into a 261 member training set (tset) and an external prediction set (pset) of 30 compounds. The tset was used for the development of

type 1 models. The tset was further divided into a 231-member tset and 30-member cross-validation set (cvset) for the neural network analysis. The cvset was used to determine when to stop training the neural network so that the network would have good, general predictive ability. Table 1 presents a complete listing of the data set compounds and their corresponding experimental and calculated ^{31}P NMR chemical shifts.

Structural Descriptors. The structures were drawn with HyperChem Lite (Hypercube, Inc.) and exported in a file format suitable for MOPAC. The geometry optimization was performed with the semiempirical quantum method AM1[39] using the MOPAC 6.0 program.[40] The MOPAC output files were used by the CODESSA program to calculate 708 descriptors. CODESSA computes six classes of structural descriptors: constitutional (number of various types of atoms and bonds, number of rings, molecular weight, etc.); topological (Wiener index, Randic indices, Kier−Hall shape indices, etc.); geometrical (moments of inertia, molecular volume, molecular surface area, etc.); electrostatic (minimum and maximum partial charges, polarity parameter, charged partial surface area descriptors, etc.); quantum (reactivity indices, dipole moment, HOMO and LUMO energies, etc.); thermodynamic (vibrational and translational enthalpies and entropies of the molecule, etc.).

Multilinear Regression Model (Type 1 Model). The heuristic multilinear regression procedures available in the framework of the CODESSA program were used to find the best correlation models. These procedures provide collinearity control (i.e., any two descriptors intercorrelated above 0.8 are never involved in the same model) and implement heuristic algorithms for the rapid selection of the best correlation, without testing all possible combinations of the available descriptors. After the heuristic reduction the pool of descriptors was reduced to 200. A variety of subset sizes were investigated to determine the optimum number of descriptors in a model. When addition of another descriptor did not improve significantly the statistics of a model, it was determined that the optimum subset size had been achieved. The optimum model size in this study was seven descriptors. In addition, the data were treated by best multilinear regression method, but the results were very similar to those obtained from the heuristic approach and the quality of the final regression equation with the same number of descriptors was not significantly improved. The stability of every potential model was tested against the cross-validated coefficient, R_{cv}^2. The R_{cv}^2 describes the stability of a regression model obtained by focusing on the sensitivity of the model to the elimination of any single data point.

Type 2 Model. The seven descriptors chosen for the type 1 model were imported to the ADAPT program and submitted to CNN for improvement as a type 2 model. A fully connected, feed-forward, three layer neural network was used. The input layer consisted of as many neurons as there were descriptors in the type 1 model. The number of neurons in the hidden layer was considered to be optimized when addition of another neuron did not decrease the training set root-mean-square, rms, error significantly. A single neuron in the output layer provided the ^{31}P NMR chemical shift. A quasi-Newton method BFGS (Broyden−Flectcher−Golfarb−Shanno)[41] was used to train the network. As mentioned previously, a cvset was used to prevent overtraining. The cvset was a small subset of compounds randomly drawn from the tset that was not included during the training but was tested periodically during training. When the cvset error was minimized, training was stopped since beyond this point the network was fitting characteristics specific to individual tset compounds rather than general characteristics of the entire data set. The use of a cvset increased the confidence with which external predictions using the trained network could be made.

The CNN results were very dependent upon the starting weights and biases, which were randomly selected. To sample a variety of starting points, an automated CNN program was run that trained the neural network from a user-determined number of random starting weights and biases. The results from these trials were then examined to find a good starting point for the neural network training.

Type 3 Model. The 200 descriptors calculated by CODESSA were imported to the ADAPT program. These descriptors were subjected to the objective feature selection routines of ADAPT, and a reduced pool of 105 descriptors was obtained and used in type 3 models. Fully nonlinear CNN type 3 models were developed using a genetic algorithm descriptor selection routine with a CNN for evaluating the fitness of each subset of descriptors selected. The fitness of descriptor subsets was calculated as $\mathrm{COST} = \mathrm{TSET} + 0.4|\mathrm{TSET} - \mathrm{CVSET}|$, where TSET and CVSET denote rms errors for the training and cross-validation sets, respectively. Models chosen with this quality factor performed better than models chosen with just training set rms error as the quality factor. That is, CNNs that produce training and cross-validation set errors that are low and similar in magnitude tend to perform well in predicting properties of interest for compounds not used in the training process. Unlike the development of type 1 models, it was extremely computationally intensive to investigate a large number of different type 3 models. The architecture for the CNN was retained from the type 2 models, providing a basis for comparison and to save the computational expense involved in determining the best network architecture. Type 3 models were found that had lower rms errors than the corresponding type 2 models, proving that the descriptors chosen on the basis of linear criteria were not the best subset for use with CNN.

RESULTS AND DISCUSSION

Multiple linear regression was used to find the best type 1 model; it consisted of the seven descriptors shown in Table 2. The model had a training set rms error of 16.1 ppm and a correlation coefficient, R^2, of 0.8619, with all *t*-values greater than 4. The cross-validated correlation coefficient, $R_{cv}^2 = 0.8499$, as compared with the correlation coefficient R^2, indicates the stability of the regression equation. The external prediction set, pset, of 30 phosphines had a rms error of 16.5 ppm and a correlation coefficient $R^2 = 0.9142$.

Two of the seven descriptors used in the linear model were topological, three were electrostatic, one was quantum-chemical, and the other one was thermodynamic. Pairwise correlations for the seven descriptors ranged from 0.149 to 0.717, with an average value of 0.437. The two topological descriptors are the Wiener index[42] and Kier & Hall (order 2).[43] The Wiener number is the sum of the graph-theoretical

Table 1. Experimental and Predicted ^{31}P NMR Chemical Shifts

no.	compds[a]	exptl[b]	type 1	type 2	type 3	no.	compds[a]	exptl[b]	type 1	type 2	type 3
1	$PH_2(p\text{-}OMeC_6H_4)$	−125.8	−108.1	−130.3	−138.6	75	PMe_2Et	−48.5	−50.6	−43.5	−45.1
2	PH_2Me[p]	−163.5	−159.2	−165.8	−134.7	76	$PMe_2(t\text{-}Bu)$	−28.7	−19.1	−19.0	−17.7
3	PH_2CF_3[cv]	−129	−137.0	−139.4	−107.6	77	$PMe_2CH(NMe)_2$	−46.2	−34.8	−30.9	−23.0
4	PH_2Et[cv]	−128	−125.7	−140.7	−128.4	78	PMe_2Cy	−43.1	−27.9	−31.5	−36.8
5	$PH_2(n\text{-octyl})$[p]	−128.5	−94.1	−126.5	−108.3	79	$PMe_2(p\text{-}MeC_6H_{10})$	−42.5	−20.2	−22.8	−34.6
6	$PH_2(s\text{-}Bu)$	−120.5	−93.6	−112.9	−109.3	80	PMe_2Ph	−46.9	−45.8	−35.3	−38.3
7	PH_2Cy[p]	−110	−82.4	−106.5	−108.1	81	$PMe_2(o\text{-}ClC_6H_4)$	−49.4	−46.3	−30.2	−27.4
8	PH_2Bz	−120.9	−111.7	−127.5	−130.7	82	$PMe_2(o\text{-}BrCH_2C_6H_4)$[cv]	−43.85	−45.1	−26.7	−23.4
9	$PH_2(n\text{-}Bu)$[p]	−140	−109.4	−130.9	−133.7	83	$PMe_2(o\text{-}NMe2C_6H_4)$	−52.2	−32.5	−28.8	−31.2
10	$PH_2(i\text{-}Bu)$	−151	−100.2	−120.5	−119.3	84	$PMe_2(p\text{-tol})$	−38.6	−38.6	−31.9	−35.3
11	$PH_2(n\text{-pentyl})$	−139	−105.6	−129.4	−127.4	85	$PMe_2(p\text{-}ClC_6H_4)$	−37.8	−46.5	−30.1	−32.5
12	$PH_2(C_6F_5)$	−183	−173.4	−184.4	−173.2	86	$PMe_2(p\text{-}MeOC_6H_4)$	−39.1	−41.8	−33.6	−36.0
13	$PH_2(o\text{-}BrC_6H_4)$	−130.6	−118.2	−124.9	−120.4	87	$PMe_2(C_6F_5)$	−47.8	−82.4	−46.4	−45.1
14	$PH_2(p\text{-}BrC_6H_4)$	−124	−116.4	−123.4	−121.5	88	$PMe_2(o\text{-}NC_5H_4)$	−41.6	−48.7	−34.1	−35.1
15	$PH_2(o\text{-}ClC_6H_4)$	−127.5	−115.0	−114.0	−126.1	89	$P(CH_2NEt_2)_2Ph$	−51.3	−29.2	−38.7	−43.0
16	$PH_2(CH_2CH_2CN)$	−135	−125.4	−137.4	−141.6	90	$PMe(CH_2OH)_2$[p]	−36.9	−54.4	−33.2	−39.8
17	$PH_2(p\text{-}ClC_6H_4)$	−124.1	−117.5	−123.7	−136.8	91	$PMe(CH_2Cl)_2$[p]	−26.9	−53.5	−12.8	−13.1
18	$PH_2(p\text{-}FC_6H_4)$	−125.4	−114.5	−125.1	−116.2	92	$PMe(CF_3)_2$	−5.76	3.6	−10.8	−17.6
19	$PH_2(o\text{-}OHC_6H_4)$	−147.4	−113.5	−134.8	−148.2	93	$PMe(CN)_2$	−81.4	−82.1	−81.4	−86.5
20	$PH_2(p\text{-tol})$	−124.5	−102.9	−116.5	−133.7	94	$PMeEt_2$	−34	−45.4	−41.7	−39.7
21	$PH_2(p\text{-}OEtC_6H_4)$	−125.5	−102.4	−123.6	−125.7	95	$PMe(n\text{-Pr})_2$	−45	−40.4	−42.6	−41.7
22	$PH_2(p\text{-}NMe_2C_6H_4)$	−126.1	−90.3	−100.6	−121.1	96	$PMe(t\text{-Bu})_2$	10.8	14.1	23.1	16.1
23	$PH_2(CH_2CH_2NH_2)$	−150.4	−121.1	−144.8	−158.0	97	$PMeBz_2$	−30	−28.2	−23.5	−24.2
24	PH_2Ph	−118.7	−111.7	−125.3	−141.3	98	$PMePh_2$	−28	−28.5	−25.1	−26.5
25	$PH_2(i\text{-Pr})$[p]	−106	−96.2	−112.7	−102.0	99	$PMe(p\text{-tol})_2$[p]	−21	−27.2	−25.3	−22.9
26	$PH_2(t\text{-Bu})$	−82	−71.6	−88.3	−78.1	100	$PMe(p\text{-}ClC_6H_4)_2$[cv]	−20.2	−34.7	−16.9	−18.7
27	$PH_2(CH{=}CH_2)$	−135.7	−144.0	−157.5	−142.4	101	$P(CH_2NEt_2)Ph_2$	−27.3	−21.2	−27.4	−29.3
28	$PH_2(o\text{-tol})$	−130.9	−101.5	−113.3	−127.4	102	$PMe(i\text{-Pr})(t\text{-Bu})$	1.5	−2.8	1.4	−0.9
29	$PH_2(o\text{-}SHC_6H_4)$	−131.7	−127.9	−141.5	−118.5	103	$PMe(i\text{-Pr})Ph$	−37.9	−23.9	−22.9	−20.4
30	$PHMe_2$	−99.5	−108.6	−122.9	−105.1	104	$PMe(CH_2CH_2Cl)Ph$	−37.9	−25.3	−16.8	−25.7
31	$PH(CF_3)_2$	−50.7	−55.9	−55.3	−51.3	105	$PMe(t\text{-Bu})Ph$[p]	−11.2	−7.9	−3.1	−22.9
32	$PHEt_2$	−55.5	−80.1	−84.7	−73.1	106	$PMeCyPh$[cv]	−25.4	−13.7	−18.3	−21.0
33	$PH(n\text{-Bu})_2$	−69.5	−66.7	−78.0	−67.4	107	$PMeBzPh$[cv]	−30.9	−28.8	−25.8	−23.9
34	$PH(CH_2CH_2CN)_2$	−75	−78.3	−66.9	−74.0	108	$PMe(o\text{-}BrCH_2C_6H_4)Ph$	−29.8	−21.2	−11.4	−13.7
35	$PHPh_2$	−41	−53.0	−38.0	−56.7	109	$P(CH_2CN)_3$	−33.1	−78.6	−54.7	−32.5
36	$PHMeEt$[p]	−77	−92.6	−101.0	−89.2	110	$PEt_2(CH_2CN)$	−24.4	−40.3	−35.1	−18.7
37	$PHMe(n\text{-Pr})$[p]	−87	−86.2	−93.7	−89.7	111	$PEt_2CH(NMe)_2$	−23.6	−34.6	−29.8	−20.8
38	$PHMe(n\text{-Bu})$	−86	−81.0	−89.4	−82.6	112	PEt_2Ph	−15.1	−32.7	−32.4	−27.5
39	$PHMePh$	−72.3	−81.4	−65.8	−76.6	113	$PEt_2(p\text{-}FC_6H_4)$	−16.1	−30.9	−21.1	−25.9
40	$PH(i\text{-Bu})_2$	−82.5	−52.6	−62.1	−61.8	114	$PEt_2(p\text{-}ClC_6H_4)$	−15.9	−31.2	−23.7	−22.5
41	$PH(i\text{-Pr})_2$	−26.4	−44.7	−49.4	−44.7	115	$PEt_2(p\text{-}BrC_6H_4)$	−16	−25.9	−16.3	−15.8
42	$PHEtPh$	−43.7	−73.8	−61.6	−65.4	116	$PEt_2(p\text{-}MeOC_6H_4)$	−17.6	−34.7	−31.9	−30.9
43	$PH(n\text{-Pr})_2$	−73	−72.0	−78.4	−68.9	117	$PEt_2(p\text{-}NMe_2C_6H_4)$	−19.1	−32.4	−32.3	−26.1
44	$PH(n\text{-pentyl})_2$	−69.4	−60.1	−72.2	−65.8	118	$PEt_2(p\text{-}NHEtC_6H_4)$	−17.79	−24.1	−27.5	−33.5
45	$PHPh(CH_2CH_2NH_2)$	−52	−69.2	−61.1	−62.0	119	$PEt_2(\text{mesityl})$	−21.9	−17.9	−20.6	−24.3
46	$PH(n\text{-Pr})Ph$	−53.4	−62.6	−50.9	−54.9	120	$PEt_2(C_6F_5)$	−23.4	−55.4	−26.9	−19.6
47	$PH(i\text{-Bu})Ph$	−60.2	−56.6	−49.9	−49.4	121	$PEt(t\text{-Bu})_2$	33.7	21.0	28.6	23.8
48	$PH(n\text{-Bu})Ph$[cv]	−52.4	−62.2	−53.4	−55.9	122	$PEt(Cy)_2$	−0.56	−5.7	−6.1	−9.1
49	$PH(i\text{-Pr})Ph$	−26.4	−53.3	−45.3	−42.2	123	$PEtPh_2$	−12	−27.3	−25.1	−22.4
50	$PH(s\text{-Bu})Ph$	−30.1	−48.6	−44.4	−40.6	124	$PEt(C_6F_5)_2$	−44	−71.3	−62.4	−48.7
51	$PH(\text{cyclopentyl})Ph$[p]	−37.1	−45.5	−43.4	−42.8	125	$PEt(\text{mesityl})_2$	−18.1	−2.7	−9.2	−14.1
52	$PH(t\text{-Bu})Ph$[p]	−5.7	−26.0	−21.6	−22.0	126	$P(CH_2CN)Ph_2$[p]	−18	−19.5	−22.4	−18.7
53	$PH(C_6F_5)_2$	−143	−116.2	−120.4	−138.2	127	$PEtCyPh$	−8.1	−13.7	−19.0	−13.8
54	$PHPh(C_6F_5)$	−92.2	−81.4	−82.2	−80.0	128	$P(n\text{-Pr})_3$	−33	−34.4	−39.9	−39.4
55	$PH(m\text{-}ClC_6H_4)_2$	−41.5	−72.8	−36.3	−38.3	129	$P(i\text{-Pr})_3$[cv]	19.3	−1.1	2.8	−3.3
56	$PH(m\text{-}FC_6H_4)_2$	−40.9	−75.5	−51.8	−37.5	130	$P(CH_2CH_2CN)_3$	−23	−24.5	−17.0	−25.9
57	$PH(p\text{-}FC_6H_4)_2$	−44.01	−61.2	−49.5	−34.9	131	$P(n\text{-Pr})_2(t\text{-Bu})$	−8.7	−6.5	−6.2	−6.1
58	$PH(m\text{-}ClC_6H_4)(p\text{-}CF_3C_6H_4)$	−42	−52.4	−25.3	−35.9	132	$PPr(p\text{-}ClC_6H_4)_2$[cv]	−17.5	−21.7	−12.4	−14.3
59	$PH(p\text{-}CF_3C_6H_4)_2$	−42.7	−40.4	−24.3	−38.9	133	$P(i\text{-Pr})_2(CH_2CN)$	0.7	−11.1	−7.7	−3.0
60	$PH(p\text{-}MeOC_6H_4)(o\text{-tol})$	−44.4	−46.8	−48.5	−33.6	134	$P(i\text{-Pr})_2Ph$	6.8	−6.2	−6.0	−5.9
61	$PH(p\text{-}MeOC_6H_4)_2$	−45.4	−56.7	−57.6	−28.7	135	$P(n\text{-Pr})(t\text{-Bu})_2$	26.3	22.1	28.6	22.0
62	$PH(o\text{-tol})_2$	−59.1	−45.7	−42.8	−40.1	136	$P(n\text{-Pr})Ph_2$	−17.6	−24.2	−25.7	−22.4
63	$PH(m\text{-tol})_2$	−40.2	−46.1	−42.5	−45.0	137	$P(n\text{-Pr})Cy_2$	−6.96	−6.6	−7.6	−8.5
64	$PH(p\text{-tol})_2$[cv]	−42.9	−50.5	−42.7	−45.7	138	$P(i\text{-Pr})(t\text{-Bu})_2$[p]	45.9	29.5	36.5	32.1
65	$PH(p\text{-}t\text{-}BuC_6H_4)_2$	−43.9	−11.3	−9.5	−41.2	139	$P(i\text{-Pr})Ph_2$	0.7	−9.7	−10.7	−11.5
66	$PHBu(CH_2CH_2NH_2)$[p]	−37.4	−67.9	−77.3	−64.0	140	$P(n\text{-Pr})CyPh$	−13.5	−10.9	−17.0	−15.0
67	$PHCy(CH_2CH_2CN)$	−37.4	−42.6	−41.1	−42.0	141	$P(n\text{-Bu})_3$	−32.3	−28.6	−36.9	−37.9
68	PMe_3	−62	−52.5	−41.0	−51.5	142	$P(i\text{-Bu})_3$[p]	−40	−6.8	−6.1	−24.1
69	$P(CF_3)_3$	−2.5	12.4	−0.4	4.8	143	$P(t\text{-Bu})_3$	61.1	47.6	44.8	43.1
70	$P(CH_2OH)_3$	−24.8	−51.0	−21.1	−29.5	144	$P(n\text{-Bu})_2(t\text{-Bu})$	−4.6	−5.1	−3.1	−5.4
71	$P(CH_2NEt_2)_3$	−65.3	−39.0	−49.9	−66.7	145	$P(n\text{-Bu})_2Ph$	−25.8	−25.4	−31.9	−27.9
72	$P(CN)_3$	−135.7	−115.5	−136.0	−129.0	146	$P(i\text{-Bu})_2Ph$	−34.2	−8.3	−11.2	−11.9
73	$PMe_2(CF_3)$	−26.9	−19.0	−17.3	−22.3	147	$P(s\text{-Bu})_2Ph$	1.8	−9.2	−11.6	−11.2
74	$PMe_2(CH_2Cl)$	−42	−55.4	−25.6	−25.5	148	$P(t\text{-Bu})_2Bz$[cv]	33.7	29.5	20.3	30.1

Table 1 (Continued)

no.	compds[a]	exptl[b]	type 1	type 2	type 3	no.	compds[a]	exptl[b]	type 1	type 2	type 3
149	P(*t*-Bu)$_2$Ph	40.2	29.1	26.4	30.9	221	PMe(CH$_2$CH$_2$SH)$_2$	−42.3	−68.0	−50.0	−32.9
150	P(*t*-Bu)$_2$(*p*-tol)^p	37.02	28.4	24.4	32.2	222	PMe(C$_6$F$_5$)$_2$	−52.2	−82.9	−76.9	−62.3
151	P(*n*-Bu)(*t*-Bu)$_2^{cv}$	26.6	22.6	28.6	22.0	223	PMeEtPh	−32.8	−34.0	−30.1	−31.4
152	P(*n*-Bu)Ph$_2$	−17.1	−21.4	−24.5	−22.1	224	PMe(*n*-Pr)Ph^{cv}	−37.9	−35.2	−34.7	−32.7
153	P(*i*-Bu)Ph$_2$	−21	−16.1	−18.0	−18.2	225	PMe(CH$_2$CH$_2$CH$_2$Cl)Ph^p	−36.3	−33.1	−24.3	−29.2
154	P(*s*-Bu)Ph$_2$	−2.8	−15.9	−18.0	−14.6	226	PMe(CH=CH$_2$)Ph	−18.2	−45.0	−35.3	−37.0
155	P(*t*-Bu)Ph$_2$	17.2	7.6	2.7	12.1	227	PEt$_2$(*t*-Bu)	−6.6	−8.5	−9.0	−5.7
156	P(*n*-pentyl)$_3$	−34	−26.4	−33.8	−34.7	228	P(CHOHCH$_3$)$_2$Ph	−12	−13.1	−7.0	7.9
157	P(cyclopentyl)$_3$	4.7	8.7	8.4	1.4	229	P(CHOHCCl$_3$)$_2$Ph^p	−10	15.0	13.9	8.0
158	P(CMe$_2$Et)Ph$_2^{cv}$	15.4	4.8	−2.6	5.5	230	P(CH$_2$CH$_2$NH$_2$)$_2$Ph^{cv}	−25.6	−30.5	−30.5	−31.7
159	P(CMeEt$_2$)Ph$_2^p$	11.4	0.1	−7.6	5.4	231	P(CH$_2$CN)$_2$Ph	−28.9	−58.3	−34.2	−23.9
160	P(*n*-octyl)$_3^{cv}$	−31.8	−35.2	−12.2	−25.4	232	PEt(CH=CH$_2$)$_2$	−20.8	−55.6	−41.7	−41.8
161	P(*o*-MeOC$_6$H$_4$)$_3$	−37.1	−34.0	−28.8	−30.1	233	P(CH$_2$CH=CH$_2$)$_3$	−34.5	−38.6	−30.8	−35.8
162	PBz$_3$	−10.4	−16.3	−11.6	−11.1	234	P(*n*-Pr)$_2$Ph	−27.7	−27.3	−32.9	−30.2
163	PBzPh$_2^p$	−10.4	−19.8	−15.9	−15.1	235	P(*i*-Pr)$_2$(CN)	−7.5	−1.7	−3.3	−5.9
164	PCy$_3$	7	6.2	2.2	9.9	236	P(*i*-Pr)$_2$(CH$_2$CH$_2$CH$_2$Cl)	1.4	−0.8	9.9	−9.2
165	PCy$_2$Ph	2.5	−0.8	−6.0	−1.7	237	P(*i*-Pr)$_2$CO$_2$Me	16.9	4.9	12.1	13.8
166	PCyPh$_2^p$	−4.4	−3.6	−11.2	−7.1	238	P(*i*-Pr)$_2$COCF$_3$	41.2	28.9	41.4	32.6
167	PPh$_3$	−4.7	−18.8	−17.1	−16.2	239	P(*i*-Pr)$_2$COCH$_2$F	32	18.2	30.8	14.6
168	P(*o*-FC$_6$H$_4$)$_3$	−44.6	−13.4	−18.5	−28.6	240	P(*i*-Pr)$_2$COMe	38.7	16.6	21.2	30.9
169	P(*o*-MeOC$_6$H$_4$)$_2$(*p*-MeOC$_6$H$_4$)	−27.7	−15.8	−21.0	−24.3	241	P(*i*-Pr)$_2$CO-*t*-Bu	15.1	34.8	36.6	35.8
170	P(*o*-tol)$_3$	−30.5	−6.6	−11.3	−12.5	242	P(*i*-Pr)$_2$COPh	26.5	24.9	20.0	33.1
171	P(*m*-FC$_6$H$_4$)$_3$	−6.5	−2.9	−8.5	−16.5	243	P(*i*-Pr)$_2$CO(*p*-BrC$_6$H$_4$)	26.1	21.5	18.8	41.7
172	P(*m*-ClC$_6$H$_4$)$_3$	−4.4	−8.4	−8.2	−6.3	244	P(CH$_2$CH$_2$CN)$_2$Ph	−23.8	−26.6	−23.2	−25.4
173	P(*m*-tol)$_3$	−5.7	−12.3	−11.7	−13.3	245	P(CH$_2$CH$_2$CH$_2$CH$_2$Cl)Ph$_2^p$	−17.9	−7.7	−8.1	−19.6
174	P(*m*-MeOC$_6$H$_4$)$_3$	−2.1	−36.8	−20.9	−16.2	246	P(*o*-NC$_5$H$_4$)$_3$	−1.3	−10.9	−24.0	−17.1
175	P(*p*-ClC$_6$H$_4$)$_3$	−9.2	−16.6	−28.7	−4.1	247	P(*i*-Pr)(*t*-Bu)(CH$_2$CH=CH$_2$)^{cv}	14.4	4.1	9.8	8.7
176	P(*p*-BrC$_6$H$_4$)$_3$	−8.2	−9.1	−6.6	−6.5	248	P(*i*-Pr)BzCOMe	27.9	13.5	9.6	24.4
177	P(*p*-tol)$_3$	−8	−8.9	−11.6	−13.7	249	P(*t*-Bu)$_2$(CH$_2$COPh)	30.6	53.8	29.8	33.6
178	P(*p*-PrC$_6$H$_4$)$_3^p$	−8.2	−14.2	−10.7	−12.3	250	P(*n*-Bu)(*n*-octyl)$_2^{cv}$	−33.3	−32.4	−19.0	−28.8
179	P(*p*-*i*-PrC$_6$H$_4$)$_3$	−7.8	3.9	−9.3	−10.8	251	P(*t*-Bu)(CH$_2$C=CH$_2$)$_2^{cv}$	−6	−7.5	−3.7	−6.3
180	P(*p*-BuC$_6$H$_4$)$_3$	−7.6	−24.6	−10.4	−11.1	252	P(*t*-Bu)(*p*-tol)$_2$	15.2	9.5	−0.3	15.0
181	P(*p*-*t*-BuC$_6$H$_4$)$_3^{cv}$	−9.1	22.9	−8.2	−4.8	253	P(*t*-Bu)BzCO$_2$Me	19.5	26.4	20.0	25.9
182	P(*p*-MeOC$_6$H$_4$)$_3^p$	−10.8	−30.2	−20.1	−15.0	254	P(*t*-Bu)BzCOTf	36.1	37.6	24.0	35.7
183	P(*p*-SMeC$_6$H$_4$)$_3$	−8.3	−2.5	−8.7	−9.6	255	P(*t*-Bu)PhCOMe	40.8	29.3	27.1	35.9
184	P(*p*-NO$_2$C$_6$H$_4$)$_3$	5.1	11.6	−12.5	−12.5	256	P(*t*-Bu)PhCOCF$_3$	39.2	48.1	39.5	36.4
185	P(*p*-Me$_2$NC$_6$H$_4$)$_3$	−11.5	−22.4	−11.9	−11.1	257	P(neopentyl)Ph$_2$	−23.9	4.0	−3.9	−10.3
186	P(2,6-F$_2$C$_6$H$_3$)$_3$	−78.5	−45.6	−55.5	−72.1	258	P(cyclopentyl)$_2$Ph^p	1.6	1.1	−3.8	−2.0
187	P(2,3,6-F$_3$C$_6$H$_2$)$_3$	−78.5	−72.6	−91.4	−83.0	259	P(cyclopentyl)Ph$_2$	−3.9	−9.8	−12.9	−10.0
188	P[2,6-(MeO)$_2$C$_6$H$_3$]$_3$	−65.7	−50.4	−54.8	−64.8	260	P(*o*-MeSC$_6$H$_4$)$_3$	−30.2	3.5	−8.0	−0.7
189	P(C$_6$F$_5$)$_3$	−76	−80.7	−70.7	−72.6	261	P(*n*-octyl)(CH$_2$CH$_2$CN)$_2$	−25.7	−4.7	−14.5	−29.4
190	P(mesityl)$_3^{cv}$	−39.5	7.5	−9.1	−7.2	262	PBz$_2$Ph	−12.1	−12.6	−11.5	−13.6
191	PPh$_2$(*o*-FC$_6$H$_4$)	−19.8	−25.3	−10.7	−14.6	263	PBz(CH$_2$CH=CH$_2$)Ph	−20.7	−19.3	−17.1	−21.4
192	PPh$_2$(*o*-ClC$_6$H$_4$)	−10.9	−27.6	−11.2	−11.5	264	P(CHOHPh)$_2$Ph^{cv}	−6	2.1	−14.3	14.7
193	PPh$_2$(*o*-OHC$_6$H$_4$)	−18.1	−22.8	−10.3	−10.0	265	PCy$_2$(CH$_2$CH$_2$Cl)	−6.2	14.8	15.4	1.2
194	PPh$_2$(*o*-tol)	−13	−14.3	−15.6	−14.8	266	PCy$_2$COCCl$_3$	20.5	43.3	30.1	13.2
195	PPh$_2$(*o*-MeOC$_6$H$_4$)^{cv}	−16	8.9	−8.5	−6.0	267	PCy$_2$COCHCl$_2$	26.4	35.5	28.3	28.1
196	PPh$_2$(*o*-EtOC$_6$H$_4$)	−15.2	−16.9	−11.9	−15.8	268	PCy$_2$COCH$_2$Cl	29.1	26.9	20.8	20.9
197	PPh$_2$(*o*-NMe$_2$C$_6$H$_4$)	−13.5	−16.1	−14.3	−13.6	269	P(*m*-hexylC$_6$H$_4$)$_3$	−7.7	−45.4	−10.4	−9.5
198	PPh$_2$(*m*-FC$_6$H$_4$)^{cv}	−10.9	−10.6	−13.1	−14.2	270	P(*m*-CF$_3$C$_6$H$_4$)$_3$	−4.3	−2.6	−16.4	−12.4
199	PPh$_2$(*m*-CNC$_6$H$_4$)^p	−8.3	1.5	−10.3	−14.4	271	P(*p*-CF$_3$C$_6$H$_4$)$_3$	−7	−21.3	−18.2	−11.3
200	PPh$_2$(*p*-ClC$_6$H$_4$)^{cv}	−6.9	−15.0	−12.4	−13.8	272	P[2,3-(MeO)$_2$C$_6$H$_3$]$_3$	−40.8	−66.1	−56.7	−55.2
201	PPh$_2$(*p*-BrC$_6$H$_4$)	−4.8	−13.8	−10.0	−12.2	273	P[2,3-(EtO)$_2$C$_6$H$_3$]$_3$	−64.6	−79.6	−53.3	−55.9
202	PPh$_2$(*p*-MeOC$_6$H$_4$)	−7	−23.4	−17.8	−14.9	274	P(*o*-CF$_3$C$_6$H$_4$)$_3$	−18.4	−13.0	−19.9	−16.0
203	PPh$_2$(*p*-NMe$_2$C$_6$H$_4$)^{cv}	−7	−17.2	−14.8	−13.8	275	P(CH$_2$CH$_2$CN)Ph$_2$	−17.6	−9.4	−13.8	−20.8
204	PPh$_2$(*p*-BuC$_6$H$_4$)	−6.5	−12.9	−13.1	−14.2	276	PPh$_2$(*o*-NEt$_2$C$_6$H$_4$)	−14.1	−6.0	−13.3	−14.4
205	P(CH$_2$CH$_2$CH=CH$_2$)$_3^p$	−30.2	−24.1	−27.3	−35.9	277	PPh$_2$(*m*-CF$_3$C$_6$H$_4$)	−10.9	−9.8	−12.2	−12.9
206	PPh(*o*-FC$_6$H$_4$)$_2$	−32.5	−15.4	−16.2	−24.3	278	PPh(CH$_2$CH=CH$_2$)$_2$	−27.5	−31.2	−28.2	−30.5
207	PPh(*o*-ClC$_6$H$_4$)$_2$	−18.8	−20.3	−12.5	−7.9	279	PPh(CH$_2$NEt$_2$)$_2$	−51.3	−23.4	−35.4	−47.1
208	PPh(*o*-MeOC$_6$H$_4$)$_2$	−27.2	−16.8	−21.2	−22.4	280	PPh(*o*-tol)$_2^{cv}$	−21.4	−10.6	−13.1	−13.7
209	PPh(C$_6$F$_5$)$_2$	−48.7	−57.3	−49.5	−46.7	281	PPh(*o*-CF$_3$C$_6$H$_4$)$_2$	−14.5	−6.1	−15.9	−11.6
210	P(*o*-tol)$_2$(*p*-tol)	−22.1	−1.3	−10.7	−13.1	282	PPh(*p*-ClC$_6$H$_4$)$_2$	−8	−18.8	−11.4	−11.8
211	P(*p*-tol)$_2$(*p*-ClC$_6$H$_4$)	−8.2	−8.5	−11.0	−12.4	283	PPh(*p*-tol)$_2$	−7	−14.8	−13.7	−13.8
212	P(*p*-tol)$_2$(*p*-MeOC$_6$H$_4$)$_2$	−9.2	−16.4	−13.4	−13.3	284	P(*m*-ClC$_6$H$_4$)$_2$(*p*-CF$_3$C$_6$H$_4$)	−7.5	−13.8	−14.5	−8.7
213	P(*p*-ClC$_6$H$_4$)(*p*-MeOC$_6$H$_4$)$_2$	−8.4	−6.2	−13.0	−12.8	285	P(*p*-MeOC$_6$H$_4$)$_2$(*p*-NMe$_2$C$_6$H$_4$)^p	−9.5	−20.9	−15.2	−13.8
214	P(*p*-BrC$_6$H$_4$)(*p*-MeOC$_6$H$_4$)$_2$	−8.6	2.5	−10.2	−11.5	286	P(*m*-ClC$_6$H$_4$)(*p*-CF$_3$C$_6$H$_4$)$_2$	−6.3	−11.9	−14.3	−8.0
215	P(*p*-tol)(*p*-MeOC$_6$H$_4$)$_2$	−10	−8.7	−14.5	−14.3	287	P(*o*-MeSC$_6$H$_4$)$_2$(*p*-MeOC$_6$H$_4$)*	−23.1	7.9	−8.9	−8.0
216	PEt$_3^{cv}$	−20.4	−39.5	−39.0	−34.9	288	P(*o*-MeSC$_6$H$_4$)$_2$(*p*-MeSC$_6$H$_4$)*	−21.1	11.6	−7.7	−2.7
217	P(CHOHCCl$_3$)$_3^{cv}$	18	20.7	29.0	34.2	289	P(*o*-MeOC$_6$H$_4$)$_2$(*p*-MeSC$_6$H$_4$)*^{cv}	−25.9	−13.7	−13.7	−14.8
218	PMe$_2$(COMe)	−19.7	−32.0	−25.5	−25.4	290	P(*o*-NC$_5$H$_4$)$_2$(*p*-MeOC$_6$H$_4$)*	−2.3	−23.4	−25.8	−17.9
219	PMe$_2$(CN)	−63	−59.3	−46.0	−52.0	291	P(*o*-NC$_5$H$_4$)$_2$(*p*-MeSC$_6$H$_4$)*	−1.6	−17.4	−12.5	−7.3
220	P(CN)$_2$(C$_6$F$_5$)	−121.3	−92.8	−119.6	−134.9						

[a] Cv = cross-validation set members; p = prediction set members. [b] All values were taken from ref 17, except those marked with asterisks, which were taken from ref 38.

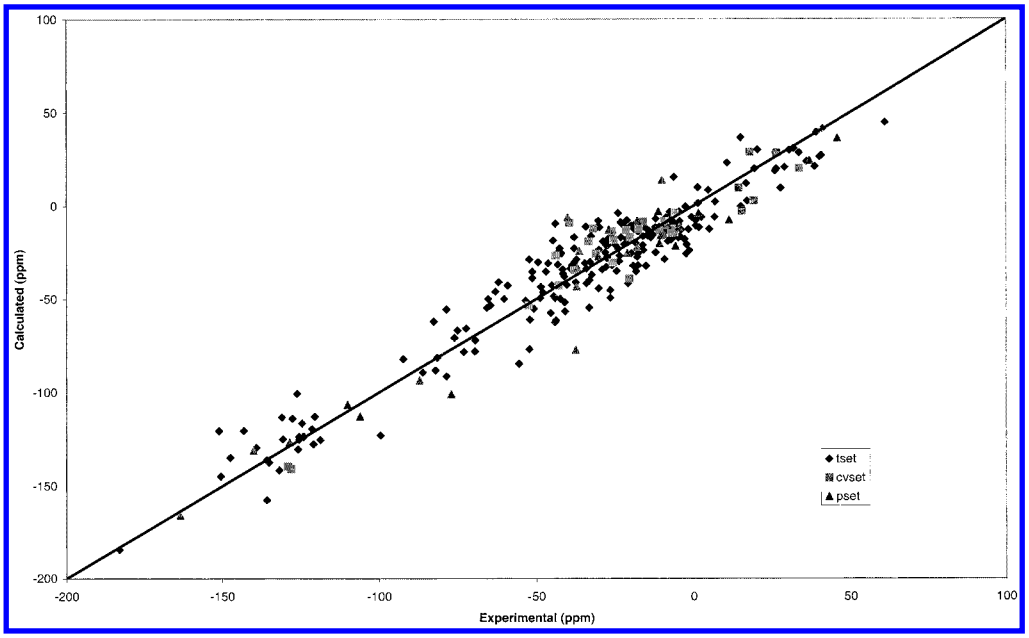

Figure 2. Plot of calculated vs experimental ^{31}P NMR chemical shifts (ppm) of the training, cross-validation, and prediction set compounds using type 2 model.

Table 2. Seven-Parameter Correlation Equation for the tset of Compounds[a]

descriptor	X	error	*t*-test
intercept	−457.62	19.52	−23.44
Kier & Hall (order 2)	13.46	0.71	18.91
Wiener index	−0.02	0.003	−7.07
PNSA-3 atomic charge weighted PNSA (quantum-chemical PC)	−1.32	0.141	−9.35
zero point vibrational energy/no. of atoms	54.39	3.49	15.58
min partial charge for a P atom (Zefirov's PC)	1063.3	89.13	11.928
tot. point-charge comp. of the molecular dipole	8.83	1.21	7.25
FPSA-2 fractional PPSA (PPSA-2/TMSA)(Zefirov's PC)	−91.38	12.99	−7.01

[a] $R^2 = 0.8619$; $F = 225.6$; $s = 16.1$; $n = 261$; $R_{cv}^2 = 0.8499$.

distances between pairs of atoms on the structure. The Kier & Hall descriptors, $^{v}\chi^{v}$, contain atomic connectivity and electronic identity of the atoms. They are related to the steric-branching effects of the molecules. The electrostatic descriptors reflect the characteristics of the charge distribution of the molecule. The Minimum partial charge for a P atom is the empirical partial charge calculated by Zefirov's approach.[44,45] The other two electrostatic descriptors are PNSA-3 atomic charge weighted PNSA and FPSA-2 fractional PPSA (PPSA-2/TMSA). They are related to the charged partial surface area (CPSA) descriptors defined by Jurs,[46,47] which are derived from the net atomic charge distribution on the solvent-accessible surface. The model contains one quantum-chemical descriptor, the total point-charge component of the molecular dipole; it represents the component of the molecular dipole deriving from the point-charge term and is obtained from the MOPAC output. The thermodynamic descriptor is zero point vibrational energy divided by the number of atoms and is also obtained from MOPAC calculations.

The seven descriptors forming the type 1 linear model were used to build a nonlinear type 2 CNN model. The input layer of the network consisted of seven neurons, one for each descriptor. The output layer contained a single neuron representing the predicted ^{31}P NMR chemical shift. The number of hidden layer neurons was varied to find the optimal architecture. To begin with, three hidden neurons were used, and the network was trained. An additional hidden layer neuron was added, the network was trained, and networks were compared using rms errors for the training and cross-validation sets. The process was repeated until additional hidden layer neurons did not enhance network performance. The optimal neural network architecture is that which produces the lowest rms errors with the fewest adjustable parameters. The number of adjustable parameters is computed as $AP = (IL + 1) \times HL + (HL + 1) \times OL$, where AP represents the number of adjustable parameters and IL, HL, and OL denote the number of neurons in the input layer, hidden layer, and output layer, respectively. It should be noted that the ratio of training set observations to adjustable parameters should be kept above 2.0 to avoid overtraining.

A 7−5−1 neural network architecture was selected as optimal. This network contains 46 adjustable parameters, corresponding-to-a ratio of 5 for training set observations (231) to adjustable parameters, well above the minimum acceptable ratio of 2.0.

With this model, the training set rms error was reduced to 11.7 ppm ($R^2 = 0.9277$; $F = 2940$), and the cross-validation set rms error was 11.4 ppm ($R^2 = 0.9035$; $F = 262$). The prediction set rms error was 13.9 ppm ($R^2 = 0.9253$; $F = 346.8$), showing the ability of the CNN to generalize to compounds not used in model formation. Compared to the type 1 model, rms errors are significantly reduced, showing that the type 2 is a superior model. Figure 2 shows a plot of calculated versus observed chemical shift values for the training, cross-validation, and prediction sets for results generated with the type 2 model.

A type 3 model was also built. The 200 descriptors calculated by CODESSA were imported to the ADAPT program, and the objective selection was carried out. The

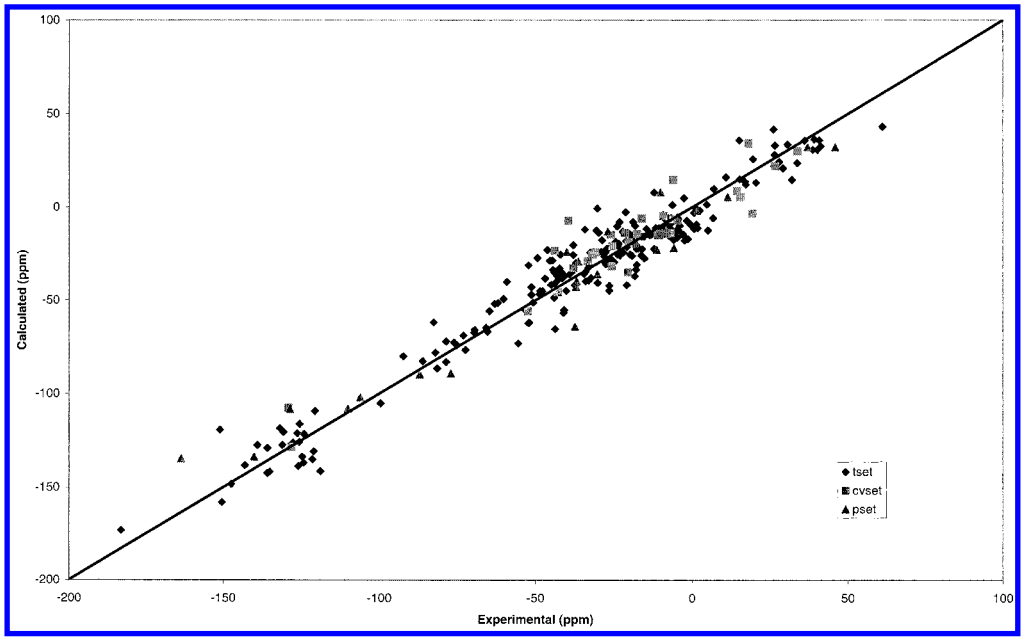

Figure 3. Plot of predicted vs experimental ^{31}P NMR chemical shifts (ppm) of the training, cross-validation, and prediction set compounds using type 3 model.

Table 3. Seven Descriptors Forming the Type 3 Model

descriptor	group
Kier & Hall index (order 2)	topological
Max Partial charge (Q_{max})	electrostatic
FPSA-2 fractional PPSA (PPSA-2/TMSA) (Zefirov's PC)	electrostatic
FPSA-3 fractional PPSA (PPSA-3/TMSA) (Zefirov's PC)	electrostatic
min *n*−*n* repulsion for a C−P bond	quantum-chemical
min tot. interaction for a C−P bond	quantum-chemical
tot. molecular 1-center E−E repulsion	quantum-chemical

105 descriptors in the reduced pool were submitted to a genetic algorithm with a 7−5−1 CNN as the fitness evaluator. The seven descriptors forming the best type 3 model are shown in Table 3. Pairwise correlations for the seven descriptors ranged from 0.029 to 0.701, with an average value of 0.223. This model contains one topological, three electrostatic, and three quantum-chemical descriptors. The topological descriptor, the Kier & Hall[43] index (order 2), was also present in the multilinear regression type 1 model. Of the three electrostatic descriptors, two are related to CPSA descriptors,[46,47] FPSA-2 and FPSA-3. The descriptor FPSA-2 is calculated from the total charge weighted (PPSA-2) and the total molecular surface area (TMSA), and FPSA-3, from the atomic charge weigthed (PPSA-3) and the TMSA. The FPSA-2 descriptor also appeared in the type 1 model. The other electrostatic descriptor is the maximum partial charge in the molecule, Q_{max}.[44,45] The three quantum-chemical descriptors involved are related to the intramolecular energy distribution in the molecules. Two of them are related to two atomic species, the C−P bonds, and the other descriptor, total molecular 1-center E−E repulsion energy, is related to all the atoms in the molecule.

With this type 3 model, training set error was reduced to 9.5 ppm ($R^2 = 0.9513$; $F = 4478$), the prediction set error reduced to 11.3 ($R^2 = 0.9527$; $F = 564$), and the cross-validation set error was 11.7 ($R^2 = 0.8996$; $F = 248$). The errors of the training and the prediction sets improved nearly 20% over the type 2 model. Figure 3 shows a plot of calculated vs observed ^{31}P NMR chemical shifts for the training, cross-validation, and prediction sets using the type 3 model. The one-to-one correlation line clearly demonstrates that the model is accurate in predicting ^{31}P NMR chemical shifts.

CONCLUSIONS

Multiple linear regression and computational neural networks were used to develop type 1, type 2, and type 3 models that can be used for the estimation of ^{31}P NMR chemical shifts of phosphines of the type PH_{3-n} R_n. The type 3 nonlinear model with a 7−5−1 CNN had the lowest errors for the training and prediction sets. The large set of compounds studied contains primary, secondary, and tertiary phosphines. The phosphines are very different from their steric and electronic characteristics, since they have substituents with several heteroatoms, O, N, F, Cl, Br, and S, and different functionalities. The obtained results show that ^{31}P NMR chemical shifts can be predicted on the basis of molecular structure alone. Similar studies of simple phosphine derivatives, such as oxides, OPR_3, sulfides, SPR_3, and phosphites, $P(OR)_3$, for example, can give information about what kind of descriptors can play an important role in the prediction of ^{31}P NMR chemical shifts in these types of compounds.

ACKNOWLEDGMENT

The authors thank Prof. Peter C. Jurs for giving us access to the ADAPT program. We also thank Greg A. Bakken and Dr. David T. Stanton for helping us in the compilation of the program. Financial support from the Catalan Government (Grant SGR99-00044) is gratefully acknowledged.

REFERENCES AND NOTES

(1) Colman, J. P.; Hegedeus, L. S.; Norton, J. R.; Finke, R. G. *Principles and Applications of Organotransition Metal Chemistry*; University Science Books: Mill Valley, CA, 1987; Chapter 3.

(2) Alyea, E. C.; Song, S. Reexamination of the Metal Carbonyl Complex Infrared Parameter, ν(CO), and Phosphorus Ligand Parameter, pK_a, $\Sigma\chi_i$ and $\Sigma\sigma^{ph}$, in relation to an Evaluation of σ and π components of M−P bonds. *Comments Inorg. Chem.* **1996**, *18*, 189−221.
(3) White, D.; Coville, N. J. Quantification of Steric Effects in Organometallic Chemistry. *Adv. Organomet. Chem.* **1994**, *36*, 95−158.
(4) Tolman, C. A. Steric Effects of Phosphorus Ligands in Organometallic Chemistry and Homogeneous Catalysis. *Chem. Rev.* **1977**, *77*, 313−348.
(5) Drago, R. S. *Applications of Electrostatic-Covalent Models in Chemistry*; Surfside: Gainesville, FL, 1994.
(6) Drago, R. S. $\Delta E-\Delta C$ Analysis of Phosphine Basicity. *Organometallics* **1995**, *14*, 3408−3417.
(7) Drago, R. S.; Joerg, S. Phosphine E_B and C_B Values. *J. Am. Chem. Soc.* **1996**, *118*, 2654−2663.
(8) Joerg, S.; Drago, R. S.; Sales, J. Reactivity of Phosphorus Donors. *Organometallics* **1998**, *17*, 589−599.
(9) Joerg, S.; Webster, C. E.; Drago, R. S.; Sales, J. Extension of the electronic-covalent model to 2:1 adducts. *Polyhedron* **1999**, *18*, 1097−1106.
(10) Wilson, M. R.; Woska, D. C.; Prock, A.; Giering, W. P. The Quantitative Analysis of Ligand Effects (QALE). The Aryl Effect. *Organometallics* **1993**, *12*, 1742−1752.
(11) Fernandez, A. L.; Lee, T. Y.; Reyes, C.; Prock, A.; Giering, W. P.; Haar, C. M.; Nolan, S. P. A thermodynamic method based on isoequilibrium behaviour to determine the values of stereoelectronic parameters of phosphines. *J. Chem. Soc., Perkin Trans. 2* **1999**, 2631−2639.
(12) Song, S.; Alyea, E. C. An Assessment of the Parameters Relevant to the Subdivision of σ and π Electronic Effects in M−P Bonds. *Comments Inorg. Chem.* **1996**, *18*, 145−164.
(13) Allman, T.; Goel, R. G. The basicity of phosphines. *Can. J. Chem.* **1982**, *60*, 716−722.
(14) Derencsenyi, T. T. Phosphorus-31 Nuclear Magnetic Resonance as a Method of Predicting Ligand Basicity and Rates of Homogeneous Catalysis. *Inorg. Chem.* **1981**, *20*, 665−670.
(15) Quin, L. D.; Bree, J. J. Steric Effects in ^{31}P NMR Spectra: "Gamma" Shielding in Aliphatic Phosphorus compounds. *Org. Magn. Reson.* **1973**, *5*, 17−19.
(16) Grim, S. O.; McFarlane, W. Phosphorus-31 Chemical Shifts in Secondary and Tertiary Phosphines. *Nature* **1965**, *208*, 995−996.
(17) Maier, L.; Diel, P. J.; Tebby, J. C. In *CRC Handbook of Phosphorus-31 Nuclear Magnetic Resonace Data*; Tebby, J. C., Ed.; CRC Press: Boca Raton, FL, 1991; Chapter 6.
(18) Kaupp, M. Interpretation of ^{31}P NMR Coordination Shifts for Phosphane Ligands. Ab Initio ECP/DFT Study of Chemical Shift Tensors in $M(CO)_5L$ [M = Cr, Mo, W; L = $P(CH_3)_3$, PF_3, PCl_3]. *Chem. Ber.* **1996**, *129*, 535−544.
(19) Ruiz-Morales, Y.; Ziegler, T. A Theoretical Study of ^{31}P and ^{95}Mo NMR Chemical Shifts in $M(CO)_5PR_3$ (M = Cr, Mo; R = H, CH_3, C_6H_5, F, and Cl) Based on Density Functional Theory and Gauge-Including Atomic Orbitals. *J. Phys. Chem. A* **1998**, *102*, 3970−3976.
(20) Kutzelnigg, W.; Fleischer, U.; Schlinder, M. In *NMR−Basic Principles and Progress*; Springer-Verlag: Berlin, 1990; Vol. 23, p 165.
(21) Ditchfield, R. Self-consistent perturbation theory of diamagnetism. I. A gauge-invariant LCAO (linear combination of atomic orbitals) method for NMR chemical shifts. *Mol. Phys.* **1974**, *27*, 789−807.
(22) Katritzky, A. R.; Lobanov, V. S.; Karelson, M. QSPR: The correlation and Quantitative Prediction of Chemical and Physical Properties from structure. *Chem. Soc. Rev.* **1995**, 279−287.
(23) Katritzky, A. R.; Maran, U.; Lobanov, V. S.; Karelson, M. Structurally Diverse Quantitative Structure−Property Relationship Correlations of Technologically Relevant Physical Properties. *J. Chem. Inf. Comput. Sci.* **2000**, *40*, 1−18.
(24) Karelson, M.; Maran, U.; Wang, Y.; Katritzky, A. R. QSPR and QSAR models derived using large molecular descriptors spaces. A review of CODESSA applications. *Collect. Czech. Chem. Commun.* **1999**, *64*, 1551−1571.
(25) Bakken, G.; Jurs, P. C. Prediction of Methyl Radical Addition Rate Constants from Molecular Structure. *J. Chem. Inf. Comput. Sci.* **1999**, *39*, 508−514.
(26) Bakken, G.; Jurs, P. C. Prediction of Hydroxyl Radical rate Constants from Molecular Structure. *J. Chem. Inf. Comput. Sci.* **1999**, *39*, 1064−1075.
(27) Anker, L. S.; Jurs, P. C. Prediction of Carbon-13 Nuclear Magnetic Resonance Chemical Shifts by Artificial Neural Networks. *Anal. Chem.* **1992**, *64*, 1157−1164.
(28) Ball, J. W.; Jurs, P. C. Simulation of Polysaccharide ^{13}C Nuclear Magnetic Resonance Spectra Using Regression Analysis and Neural Networks. *Anal. Chem.* **1993**, *65*, 3615−3621.
(29) Clouser, D. L.; Jurs, P. C. Simulation of ^{13}C nuclear magnetic resonance spectra of tetrahydropyrans using regression analysis and neural networks. *Anal. Chim. Acta* **1994**, *295*, 221−231.
(30) Clouser, D. L.; Jurs, P. C. The simulation of ^{13}C nuclear magnetic resonance spectra of dibenzofurans using multiple regression analysis and neural networks. *Anal. Chim. Acta* **1996**, *321*, 127−135.
(31) Katritzky, A. R.; Lovanov, V. S.; Karelson, M. *CODESSA, Reference Manual V 2.13*; Semichem and the University of Florida: Gainsville, FL, 1997
(32) Stuper, A. J.; Brugger, W. E.; Jurs, P. C. *Computer-Assisted Studies of Chemical Structure and Biological Function*; Wiley: New York, 1979.
(33) Jurs, P. C.; Chow, J. T.; Yuan, M. In *Computer-Assisted Drug Design*; Olson, E. C., Christoffersen, R. E., Eds.; The American Chemical Society: Washington, DC, 1979.
(34) Luke, B. T.; Evolutionary Programming Applied to the Development of Quantitative Structure−Activity Relationships. *J. Chem. Inf. Comput. Sci.* **1994**, *34*, 1279−1287.
(35) Kalivas, J. H.; Roberts, N.; Sutter, J. M. Global Optimization by Simulated Annealing with Wavelength Selection for Ultraviolet−Visible Spectrophotometry. *Anal. Chem.* **1989**, *61*, 2024−2030.
(36) Kalivas, J. H. Generalized Simulated Annealing for Calibration Sample Selection from an Existing Set and Orthogonalization of Undesigned Experiments. *J. Chemom.* **1991**, *5*, 37−48.
(37) Sutter, J. M.; Dixon, S. L.; Jurs, P. C.; Automated Descriptor Selection for Quantitative Structure−Activity Relationships Using Generalized Simulated Annealing. *J. Chem. Inf. Comput. Sci.* **1995**, *35*, 77−84.
(38) Laitinen, R. H.; Riihimäki, H.; Haukka, M.; Jääskelänien, S.; Pakkanen, T. A.; Pursiainen, J. Syntheses and Characterization of New Tertiary Phosphane Ligands Prepared from p-Anisyl- and p-Thioanisyldichlorophosphanes. *Eur. J. Inorg. Chem.* **1999**, 1253−1258.
(39) Dewar, M. J. S.; Zoebisch, E. G.; Healy, E. F.; Stewart, J. J. P. AM1: A New General Purpose Quantum Mechanical Molecular Model. *J. Am. Chem. Soc.* **1985**, *107*, 3902−3909.
(40) Stewart, J. P. P. *MOPAC 6.0, Quantum Chemistry Program Exchange;* QCPE No. 455; Indiana University: Bloomington, IN, 1989.
(41) Wessel, M. D.; Jurs, P. C. Prediction of Reduced Ion Mobility Constants from Structural Information Using Multiple Linear Regression Analysis and Computational Neural Networks. *Anal. Chem.* **1994**, *66*, 2480−2487.
(42) Wiener, H. J. Structural Determination of Paraffin Boiling Points. *J. Am. Chem. Soc.* **1947**, *69*, 17−20.
(43) Kier, L. B.; Hall, L. H. *Molecular Connectivity and Drug Research*; Academic Press: New York, 1976.
(44) Zefirov, N. S.; Kirpichenok, M. A.; Izmailov, F. F.; Trofimov, M. I. Calculation schemes for atomic electronegativities in molecular graphs within the framework of Sanderson principle. *Dokl. Akad. Nauk SSSR* **1987**, *296*, 883−887.
(45) Kirpichenok, M. A.; Zefirov, N. S. Electronegativity and molecular geometry. General bases of the developed approach and determination of closer electrostatic interactions on bond length in organic molecules. *Zh. Org. Khim.* **1987**, *23*, 673−691.
(46) Stanton, D. T.; Jurs, P. C. Development and Use of Charge Partial Surface Area Structural descriptors in Computer assisted quantitative structure−property relationship studies. *Anal. Chem.* **1990**, *62*, 2323−2329.
(47) Stanton, D. T.; Egolf, L. M.; Jurs, P. C.; Hicks, M. G. Computer-Assisted Prediction of Normal Boiling Points of Pyrans and Pyrroles. *J. Chem. Inf. Comput. Sci.* **1992**, *32*, 306−311.

CI000458K