Simple Quantitative Structure−Property Relationship (QSPR) Modeling of ^{17}O Carbonyl Chemical Shifts in Substituted Benzaldehydes Compared to DFT and Empirical Approaches

Rudolf Kiralj* and Márcia M. C. Ferreira

Instituto de Química, Universidade Estadual de Campinas, Campinas, SP, 13083-970, Brazil

Received: January 29, 2008; Revised Manuscript Received: March 27, 2008

The geometry of 50 substituted benzaldehydes was optimized at the semiempirical PM3 level, and various electronic and steric descriptors accounting for properties of the benzene ring, aldehyde group, and their connecting carbon−carbon bond were calculated. Quantitative structure−property relationships (QSPR) between ^{17}O carbonyl chemical shifts and these descriptors were established using partial least-squares regression and principal component regression. These two parsimonious QSPR models were comparable with the literature empirical model and DFT (density functional theory) and capable of predicting ^{17}O chemical shifts for 10 benzaldehydes. Principal component analysis, hierarchical cluster analysis, and crystal structure data retrieved from the Cambridge Structural Database were additional methods for chemical verification of the regression models. The QSPR models are recommended as being more reliable than and superior to the empirical and DFT models due to the results of all validations, simplicity, and short time that regressions need for ^{17}O shift prediction.

Introduction

Li and Li[1] studied ^{17}O NMR chemical shifts of 50 substituted benzaldehydes (Figure 1) and established an empirical relationship via parametric eq 1

$$\delta_{LL}/\text{ppm} = 564.0 + \delta_o + \delta_{o'} + \delta_m + \delta_{m'} + \delta_p + C \quad (1)$$

where δ_o, $\delta_{o'}$, δ_m, $\delta_{m'}$, and δ_p are contributions (increments) to chemical shifts that account for ortho (*o*), ortho′ (*o*′), meta (*m*), meta′ (*m*′), and para (*p*) substituents, respectively, *C* is a correction constant for polar solvents, and the free coefficient accounts for ^{17}O shift in formaldehyde. The authors determined previously the δ increments for 11 *o*/*o*′-, *m*/*m*′-, and *p*-substituents by multiple linear regression (MLR).[2–4] Intramolecular hydrogen bonds and steric and substituent (inductive and conjugation) effects on ^{17}O shifts were reported as the chemical basis of eq 1.

Another way to calculate ^{17}O chemical shifts are quantum-chemical calculations such as DFT (density functional theory) of solutes or solvent−solute complexes using at least the 6-311+G(d,p) basis set.[5,6] The nuclear shielding tensor is calculated for each atom[5,7] via the GIAO (gauge-independant atomic orbital) approach.[8] The tensor's diagonal elements give the isotropic shielding σ_{iso} (eq 2)

$$\sigma_{iso} = \sigma_{xx} + \sigma_{yy} + \sigma_{zz} \quad (2)$$

The σ_{iso} values can be used for calculation of the chemical shift δ of a compound[5,7] when the shift δ_{ref} and the shielding $\sigma_{iso\text{-}ref}$ of the referent compound are known (eq 3)

$$\delta = \delta_{ref} + \sigma_{iso\text{-}ref} - \sigma_{iso} \quad (3)$$

While the empirical (eq 1) and quantum-chemical (eqs 2 and 3) models are based on increments of the same property, a quantitative structure−property relationship (QSPR) approach linearly combines distinct properties (molecular descriptors).[9–16] There is a progressive demand to apply rigorous validation procedures for regression models in QSPR and related areas,[17–22] while various empirical equations and quantum chemical calculations are simply taken "as is". This statistical injustice can easily provoke confusion when the validity of various calculation approaches is questioned by means of comparative statistics with the aim to identify the easiest, simplest, and most economic procedure. For example, in spite of the predictive ability of the parametric model (eq 1) as claimed by the authors,[1] it cannot be validated as a regression model and has no errors for parameters. Equation 1 is limited to benzaldehydes with 11 substituents: $-CH_3$, $-OH$, $-N(CH_3)_2$, $-F$, $-Cl$, $-Br$, $-CN$, $-NO_2$, $-OCH_3$, $-COCH_3$, and $-OCOCH_3$. Furthermore, the quantum chemical model (eqs 2 and 3) does not report cumulative errors of geometry optimization and property calculations and is very sensitive to molecular conformation and intramolecular and intermolecular interactions.

Previous experiences in modeling of NMR[23] and ESCA[24] shifts by chemometric methods, QSPRs,[9–16] and multivariate quantitative structure correlations[25–31] have encouraged the authors of this work to develop a simple and fast QSPR methodology for prediction of ^{17}O carbonyl chemical shifts in substituted benzaldehydes (Figure 1, training set; Figure 2, prediction set). The methodology includes a semiempirical procedure for molecular modeling and calculation of molecular descriptors that are then quantitatively correlated to experimental shifts via regression methods, partial least squares (PLS) and principal component regression (PCR).[32–35] Chemical validation of the regressions is carried out by exploratory analysis (principal component analysis, PCA, and hierarchical analysis, HCA)[32–35] and structural investigations of benzaldehydes in the crystalline state. Attention is also paid to electron delocalization of the benzaldehyde system and intramolecular hydrogen bonds which are coupled to this system and thus significantly affect ^{17}O shifts of the carbonyl and hydroxyl groups[36] (resonance-assisted moderately strong hydrogen bonds[26,37,38]). Several validation procedures are

* To whom correspondence should be addressed. Phone: +55 19 3521 3102. Fax: +55 19 3521 3023. E-mail: rudolf@iqm.unicamp.br.

10.1021/jp800864s CCC: $40.75
Published on Web 06/12/2008

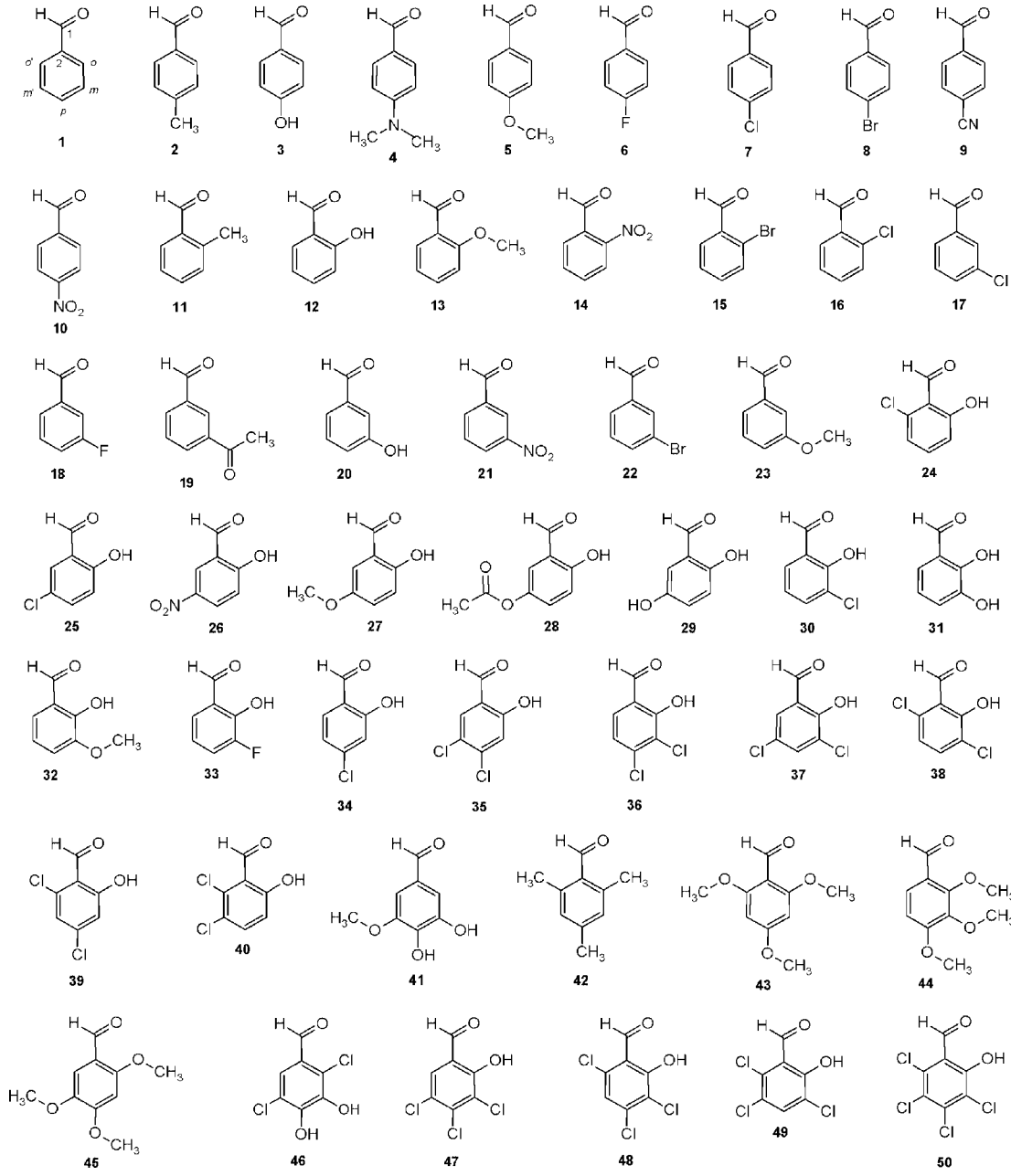

Figure 1. Molecular structures of substituted benzaldehydes **1**−**50** (training set) with marked substitution positions and partial atomic numbering.

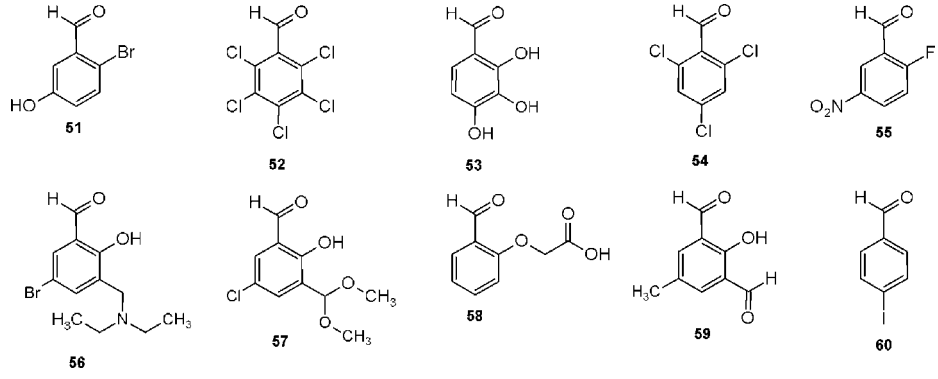

Figure 2. Molecular structures of substituted benzaldehydes **51**−**60** (prediction set).

performed, and additional statistical parameters are established with the aim to compare the regression models with the Li and Li (LL) model (eq 1) and a DFT model (eqs 2 and 3) for **1**−**60**.

Methodology

Experimental Chemical Shifts. Experimental ^{17}O chemical shifts δ_{exp} for carbonyl oxygen atoms in 50 substituted benzaldehydes (Figure 1) were as collected from the literature by Li and Li.[1] No more recent data for benzaldehydes have been found in the current literature; thus, **1**−**50** was defined as the training set. The original experimental data were from measurements under different experimental conditions: at two temperatures (room and 75 °C), in three solvents (acetonitrile, 1,4-dioxane, and $CDCl_3$) or in the liquid state of a pure substance. In the case of multiple data for 13 benzaldehydes (**1**−**3**, **5**, **7**, **8**, **10**, **13**, **17**, **20**, **21**, **23**, and **25**), the differences between particular data due to variations in experimental procedures or repetitions were reasonable (at most 15.9 ppm). Hence, the data were averaged for these compounds in order to establish QSPR models without degenerated samples. Li and Li did not average the data, which virtually increased the prediction ability of their model.[1] The prediction set **51**−**60** (Figure 2) was defined to compare the predictive ability of the QSPR and DFT (eqs 2 and 3) models and show limitations of the LL model (eq 1; no adequate group contributions for **55**−**60**).

Reconstructed LL Model. Using group increments defined by Li and Li,[1] the original LL model (eq 1) was formally simplified (eq 4a) due to the orientation ambiguity in distinguishing *o*- from *o*′- and *m*- from *m*′-sites (see Figures 1 and 2)

$$O = \delta_o + \delta_{o'} M = \delta_m + \delta_{m'} P = \delta_p \quad (4a)$$

$$\delta_o = \delta_{o'} = \delta_m = \delta_{m'} = \delta_p = 0$$

for H at positions *o*, *o*′, *m*, *m*′, and *p* (4b)

$C = -14.7$ for **24**,**34**−**40**,**47**−**50**; otherwise, $C = 0.0$ (4c)

$$\delta_{LL}/\text{ppm} = 564.0 + O + M + P + C \quad (4d)$$

Finally, the reconstructed LL data set consisted of a matrix with dimensions 54 × 4: four empirical molecular descriptors (*O*, *M*, *P*, and *C*) were calculated for 54 benzaldehydes.

Structural Studies. A series of searches for crystal structures containing the general (partially hydrogen-depleted) benzaldehyde fragment (GBF), i.e., C_6−C(O)H was performed in the November 2007 version 5.29 of the Cambridge Structural Database (CSD,[39,40]) supported by ConQuest 1.10[41,42] for data retrieval, Vista 2.1[42,43] to visualize numerical data, and Mercury CSD 2.0[44,45] to analyze intermolecular interactions. Qualitative searches without filters were directed to identification of structures of **1**−**50** or similar molecules, formation of the set **51**−**60**, and intermolecular interactions involving these and similar molecules in the crystalline state. Quantitative searches included calculations of certain geometric parameters of GBF and *o*-hydroxybenzaldehyde with these filters: crystallographic factor $R < 0.05$, no disorders or errors, experimental errors on C−C bond lengths ≤ 0.05 Å, and no chemical bonds between the −C(O)H group and other species. Finally, a semiquantitative search for calculations of torsion angles in *o*-hydroxybenzaldehydes with no filters was carried out in order to explore the conformational features of these benzaldehydes.

Semiempirical Procedures and Calculation of Molecular Descriptors for QSPR. Modeling of **1**−**60** was greatly aided by the existence of crystal structures of 20 benzaldehydes: **1**−**10**, **12**−**14**, **16**, **20**, **23**, **25**, **31**, **32**, **43**, **44**, **51**, and **56**−**60**. Several hundreds of crystal structures of other substituted benzaldehydes and more complex systems where GBF was a substituent were useful in defining **52**−**55** and molecular modeling of other benzaldehydes. **1**−**60** were modeled by Chem3D Ultra[46] using the crystal structures and chemical knowledge (the lowest total electronic energy of a molecule) in such a way that *o*-hydroxyl and aldehyde groups established −(H)C═O···HO− hydrogen bonds. Other neighboring OH groups were connected to each other and to the aldehyde group through hydrogen bonds whenever possible. A molecular dynamics conformation search was performed for the modeled molecules under default conditions (step interval 2 fs, frame interval 10 fs, 10 000 steps, heating/cooling rate 1 kcal/atom/ps, and 300 K), and obtained minimum energy structures were optimized by molecular mechanics MM2[47] in Chem3D Ultra. The new geometries were further energy minimized by semiempirical PM3 method in the Titan software.[48] Several global and local molecular descriptors (110 in total) for QSPR study, mainly of electronic and geometric nature, were calculated using Titan, MOPAC 6.0 for Windows[49] (single point-calculation), and Matlab 5.2.[50] The data were organized into a matrix with dimensions 60 × 110.

DFT Calculations. Geometries of **1**−**59** from PM3 calculations were optimized at the B3LYP 6-311+G(d,p) level, and nuclear shielding tensors were calculated by Gaussian 98[51] according to the literature.[5] Only **60**, due to the presence of the iodine atom and limitations in Gaussian basis sets, was treated at the B3LYP 3-21G(d,p) level. Obtained shielding tensors of oxygen atoms σ_{iso} were used for calculation of corresponding chemical shifts by selecting **1** for the referent compound and modifying eqs 2 and 3 into

$$\delta_{DFT} = \delta_{exp}(\mathbf{1}) + \sigma_{iso}(\mathbf{1}) - \sigma_{iso} = 250.5 \text{ ppm} - \sigma_{iso} \quad (5)$$

Respective diagonal elements σ_{xx}, σ_{yy}, and σ_{zz} of the shielding tensors were organized into a matrix, the DFT data set, with dimensions 60 × 3. To find out how much the DFT procedure for **60** is reliable, the same procedure was carried out for **1** and **6**−**8**. The shifts for these molecules from B3LYP 6-311+G(d,p) and B3LYP 3-21G(d,p) calculations were compared, which revealed systematically lower shifts for the lower basis set by 4.0−7.1 ppm. This means that the δ_{DFT} value for **60** can be used for qualitative purposes once that the basis set effect is probably less than 10 ppm.

Modeling of Additional Systems for *o*-Hydroxybenzaldehyde (12). Additional molecular systems were modeled from *o*-hydroxybenzaldehyde (**12**), the simplest benzaldehyde with an intramolecular CHO···HO hydrogen bond, in order to evaluate the influence of intramolecular and intermolecular hydrogen bonds on the ^{17}O shifts in benzaldehydes. The modeled structure of **12** already possesses the hydrogen bond. Conformers of **12**, with different interactions between the OH and CHO groups, were also modeled and studied at the DFT level in the same way as **12**. Additionally, **12** and its conformers were further treated at the same DFT level, and the corresponding oxygen chemical shifts were calculated, including solvent effects (simulating the experimental conditions: acetonitrile, chloroform, ether, and cyclohexane) by the polarizable continuum model (PCM).[52]

A dimer containing two molecules of *o*-hydroxybenzaldehyde via two CHO···HO hydrogen bonds was modeled, its conformational space was studied under default molecular dynamics conditions, and the resulting geometry was optimized at the MM2 level,[46] all inside the Chem3D Ultra platform. Due to the weak nature of intermolecular interactions, the complex was treated first at the B3LYP 6311+G(d,p) level in Gaussian 98 and posteriori at the PM3 level. To complete the data matrix for QSPR with **12d1** and **12d2**, the dimer was first optimized by PM3 in Titan, and then this was repeated for each monomer with fixed conformation. The monomer properties were then

TABLE 1: List of Statistical Parameters for Evaluation of Regression (PLS and PCR) and Parametric (DFT and LL) Models

parameter	definition and recommendations[a]	lit.[b]
number of LVs, PCs, or original descriptors in a model	p	GCK
number of samples (molecular systems) in a test or validation set	n	GCK
standard error of validation	$SEV = [\Sigma_i (y_{ei} - y_{vi})^2/n]^{1/2}$	GCK
standard error of calibration	$SEC = [\Sigma_i (y_{ei} - y_{ci})^2/(n - p - 1)]^{1/2}$	GCK
standard error of prediction	$SEP = [\Sigma_i (y_{ei} - y_{pi})^2/(n - p - 1)]^{1/2}$	GCK
correlation coefficient of validation[c]	$Q^2 = 1 - [\Sigma_i (y_{ei} - y_{vi})^2]/[\Sigma_i (y_{ei} - \langle y_{ei}\rangle)^2]$, $Q^2 > 0.5$	GCK
correlation coefficient of calibration or prediction[c]	$R^2 = 1 - [\Sigma_i (y_{ei} - y_{ci})^2]/[\Sigma_i (y_{ei} - \langle y_{ei}\rangle)^2]$, $R^2 > 0.6$ for calibration $R^2 = 1 - [\Sigma_i (y_{ei} - y_{pi})^2]/[\Sigma_i (y_{ei} - \langle y_{ei}\rangle)^2]$, $R^2 > 0.6$ for prediction	GCK
linear correlation coefficient of validation	$Q = [\Sigma_i (y_{ei} - \langle y_{ei}\rangle)(y_{vi} - \langle y_{vi}\rangle)]/[\Sigma_i (y_{ei} - \langle y_{ei}\rangle)^2]^{1/2} [\Sigma_i (y_{vi} - \langle y_{vi}\rangle)^2]^{1/2}$	GCK
linear correlation coefficient of calibration or prediction	$R = [\Sigma_i (y_{ei} - \langle y_{ei}\rangle)(y_{pi} - \langle y_{pi}\rangle)]/[\Sigma_i (y_{ei} - \langle y_{ei}\rangle)^2]^{1/2} [\Sigma_i (y_{pi} - \langle y_{pi}\rangle)^2]^{1/2}$ for calibration $R = [\Sigma_i (y_{ei} - \langle y_{ei}\rangle)(y_{pi} - \langle y_{pi}\rangle)]/[\Sigma_i (y_{ei} - \langle y_{ei}\rangle)^2]^{1/2} [\Sigma_i (y_{pi} - \langle y_{pi}\rangle)^2]^{1/2}$ for prediction	GCK
absolute deviation (error) of calibration or prediction	$\Delta_i = \lvert y_{ei} - y_{ci}\rvert$ or $\Delta_i = \lvert y_{ei} - y_{pi}\rvert$	GCK
relative deviation (error) of calibration or prediction	$\Delta_{rel,i} = \Delta/88.6$, where 88.6 ppm is the maximum variation of y_e (δ_{exp})	TW
maximum of Δ_i and $\Delta_{rel,i}$ values	Δ_{max} and $\Delta_{max\text{-}rel}$	GCK
minimum of Δ_i values	Δ_{min}	GCK
"pure" maximum error	$\Delta_{max} - \Delta_{min}$	TW
average of Δ_i and $\Delta_{rel,i}$ values	$\langle\Delta\rangle = [\Sigma_i \Delta_i]/n$ and $\langle\Delta_{rel}\rangle = [\Sigma_i \Delta_{rel,i}]/n$	GCK
weighted $\langle\Delta\rangle$ and $\langle\Delta_{rel}\rangle$	$w\langle\Delta\rangle$ and $w\langle\Delta_{rel}\rangle$, where $w = n/(n - p - 1)$	TW
number of samples with significant errors	$N_{rel>10\%}$: number of samples with relative errors $\Delta_{rel} > 10\%$	GCK
total overprediction	$T_{over} = \Sigma_i y_{ei} - \Sigma_i y_{ci}$ or $T_{over} = \Sigma_i y_{ei} - \Sigma_i y_{pi}$	TW
number of overpredictions, underpredictions, and zero errors	N_{over}, N_{under} and N_{zero}	GCK
average of Q^2 from leave-N-out cross-validations	$\langle Q^2_{LNO}\rangle$,$\langle Q^2_{LNO}\rangle$ over 0.5	LIT1
maximum Q^2 and R^2 from **y**-randomizations	Q^2_{yrand}, R^2_{yrand}, $Q^2_{yrand} < 0.3$, $R^2_{yrand} < 0.3$	LIT1
maximum Q^2 and R^2 from HCA-based bootstrappings	Q^2_{bstr}, R^2_{bstr}, $Q^2_{bstr} > 0.5$, $R^2_{bstr} < 0.6$	LIT1
number of HCA clusters with $\Delta_{rel,i} > 10\%$	$C_{HCA>10\%}$	TW
slope of the y_e against y_p regression line	$k = [\Sigma_i y_{ei} y_{pi}]/[\Sigma_i (y_{pi})^2]$, $0.85 \le k \le 1.15$	LIT2
slope of the y_p against y_e regression line	$k' = [\Sigma_i y_{ei} y_{pi}]/[\Sigma_i (y_{ei})^2]$, $0.85 \le k' \le 1.15$	LIT2
R^2 for the y_e against y_p regression line shifted to zero intercept	$R_0^2 = 1 - [\Sigma_i (y_{pi} - y_{ei})^2]/[\Sigma_i (y_{pi} - \langle y_{pi}\rangle)^2]$, $y_{ei} = k\, y_{pi}$	LIT2
R^2 for the y_p against y_e regression line shifted to zero intercept	$R_0'^2 = 1 - [\Sigma_i (y_{ei} - y_{pi})^2]/[\Sigma_i (y_{ei} - \langle y_{pi}\rangle)^2]$, $y_{pi} = k'\, y_{ei}$	LIT2
$R_0^2 - R_0'^2$ absolute difference for external validation	$\lvert R_0^2 - R_0'^2\rvert < 0.3$ and close to zero	LIT3
ratio parameter for external validation[d]	$(R_0'^2 - R_0^2)/R^2_{ev} < 0.1$ and close to zero	LIT1
ratio parameter for external validation[d]	$(R_0^2 - R_0'^2)/R^2_{ev} < 0.1$ and close to zero	LIT1

[a] Basic definitions: i = the summation index and index of a particular shift value for the ith sample; y_e = experimental values of **y**, i.e., δ_{exp}; y_c = calculated values of **y**, i.e., δ_{cal} (δ_{PLS}, δ_{PCR}, δ_{DFT}, or δ_{LL}) values from calibration (for the training set of the proposed model); y_p = predicted values of **y**, i.e., δ_{cal} (δ_{PLS}, δ_{PCR}, δ_{DFT}, or δ_{LL}) values from external validation (for the training set of the external validation); y_v = calculated values of **y** from an internal validation (leave-one-out cross-validation, leave-N-out cross-validation, **y**-randomization, and HCA-based bootstrapping). [b] Literature sources LIT1 (ref 18), LIT2 (refs 18 and 19), LIT3 (ref 19), general chemometric knowledge (GCK), and this work (TW). [c] Recommendations according to the literature.[18,19] [d] Correlation coefficient R^2_{ev} is R^2 for the training set from external validation.

calculated by MOPAC 6.0 (single-point calculations). The molecules in the dimer with approximate C_i symmetry were named
12d1 and **12d2**. Molecule **12**, its conformers, and **12d1** and **12d2** were additionally geometry optimized, and isotropic ^{17}O shielding tensors were calculated at the B3LYP 6-311++G(d,p) level. The differences in obtained δ_{DFT} values for these systems obtained using the 6-311+G(d,p) and 6-311++G(d,p) basis set were negligible (from 0.0 to 0.7 ppm), confirming that the former basis set was an appropriate computational approach.

Regression Models (QSPR). The initial data matrix for the training set had dimensions 50 × 110 (50 = number of benzaldehydes, 110 = number of molecular descriptors) and was autoscaled prior to chemometric analyses. The initial set of 110 descriptors was inspected by calculating the correlation coefficients of the **y** vector (δ_{exp} data) with all descriptors and then applying the cutoff value 0.60. Then, the variable selection was manually directed to satisfy several criteria: (1) satisfactory descriptor−**y** correlations and PLS statistics (eliminating false nonlinearity, chance correlation, and pronounced nonuniform distribution or dispersion of points in the scatterograms), (2) minimization of the number of descriptor intercorrelations, (3) selection of descriptors of different nature, origin, and clusterings in HCA and PCA, (4) computationally simple descriptors, and (5) chemically interpretable and understandable descriptors. In other words, besides correlation analysis (scatterograms and correlation coefficients), chemometric methods[34,53] for variable selection were used: leave-one-out cross-validation parameters (SEV, SEP, Q, and R, defined in Table 1), HCA dendograms for descriptors, PCA loadings plots, and regression vectors from PLS. Finally, eight selected molecular descriptors (QSPR data set, matrix **X**) were used to build the final PLS and PCR models. No nonlinear relationships between **y** and molecular descriptors have been detected.

The regression models were validated as recommended in the literature for QSPR and related areas[17–22] by carrying out (1) leave-one-out cross-validation, (2) leave-N-out cross-validation, where N varied from 1 to 10, 3) 10 **y**-randomizations according to Wold and Eriksson,[54] (4) external validation with 40/10 benzaldehydes in the new training and external validation sets, respectively, based on a HCA analysis with complete linkage, and (5) 10 bootstrappings based on the same HCA. The **X** and **y** data were randomized prior to leave-N-out cross-validation and **y**-randomization. The HCA analysis has identified five clusters at the similarity index 0.70, which was useful in defining a representative external validation set. The HCA results

were also used to carry out a simple bootstrapping procedure in which 10 molecules were randomly excluded each time. Due to different size of the clusters (from 1 to 22 benzaldehydes), the number of excluded molecules was fixed for each cluster (from 0 to 4). Several statistical parameters (Table 1), common in QSPR or defined in this work, were calculated for regression models and their external validations. The MLR model was not considered in all further analyses due to its extremely unacceptable statistics ($Q^2 = -5.851$).

Comparison of QSAR Models with LL and DFT Models. Although the LL and DFT models cannot be validated, correlation coefficients, errors, and other parameters, according to their definitions in Table 1, were calculated for these models whenever possible. This way, all the models could be compared in terms of several statistical parameters. HCA analyses were performed for LL and DFT data sets to see the clustering patterns at the similarity index 0.70. Posterior HCA-based bootstrapping procedures were carried out as for the QSPR data set.

Exploratory Analysis. HCA with incremental linkage and principal component analysis (PCA) of the QSPR data set (matrix **X**) that had been used in construction of the final regression models were carried out in order to aid in interpretation and chemical validation of these models. All chemometric analyses were performed using the Pirouette program[55] on autoscaled data. Histogram plots were obtained using OriginPro 7.[56]

Results and Discussion

The QSPR data set is in Table 2, and the LL and DFT data sets are in Table 3. Experimental and calculated chemical shifts with deviations are shown in Table 4. Correlations between descriptors from the three data sets as well as correlations between the descriptors and experimental shifts are presented in Table 5. DFT predictions of ^{17}O shifts of *o*-hydroxybenzaldehyde systems with solvent effects are in Table 6. Detailed statistical comparison of the four models (PLS, PCR, LL, and DFT) is in Table 7, and the corresponding regression and parametric equations are in Table 8. External validation details for PLS and PCR are in Table 9. Types of intermolecular interactions involving selected benzaldehyde systems in the crystalline state are presented in Table 10. Figure 3 shows a simple frequency distribution of experimental shifts. Figures 4–10 present chemometric, computational, and structural results in relation to ^{17}O shifts in **1**−**60**. The samples for the external validation set were selected from each cluster, as shown in Figure 8.

y Data (Experimental Chemical Shifts). The δ_{exp} data (vector **y**), as can be seen in Figure 3 and Table 4, vary from 505.0 (**48**) to 593.6 ppm (**9**), which is a quite large variation (88.6 ppm). The histogram in Figure 3 shows that there are three main regions of benzaldehydes concentration. The region of low shifts or strong shielding (from 505.0 to 522.8 ppm) contains only compounds with *o*-OH which established hydrogen bonds with the aldehyde oxygen atom: **12**, **24**−**40**, and **47**−**50** (see Figure 1). The local maximum corresponds mostly to molecules with a *m*- and/or *m*′-substituent. The region of medium shift or weak shielding is the valley between two peaks (from 526.9 to 563.2 ppm): **1**−**5**, **13**, **20**, **23**, **41**, and **44**−**46**. The lowest shifts are caused by typically weak electron-donating *p*-substituents in **3** (−OH) and **4** (−NMe_2). The region of high shift or strong deshielding (from 565.0 to 593.6 ppm) relative to **1** includes molecules with one or more groups that have a strong electron-withdrawal effect on the aromatic ring: **6**−**11**, **14**−**19**, **21**, **22**, **42**, and **43**. The local maximum is mostly for molecules with halogens, −Me and −NO_2 groups. Shifts above 580 ppm are due to strong withdrawing groups such as *o*,*o*′-Me_2

TABLE 2: Selected Molecular Descriptors for Regression Models for 1−60 (QSPR Data Set)

no.	E_e/eV	E_{CC}/eV	Δ_{HL}/eV	σ_b/Å	σ_r/Å	D_{CC}/Å	Q_{C2mul}	Q_{Omul}
1	70.679	122.756	−9.567	0.071	0.003	1.484	−0.201	−0.317
2	71.061	122.671	−9.284	0.071	0.004	1.483	−0.209	−0.319
3	72.252	122.749	−9.047	0.071	0.007	1.481	−0.244	−0.322
4	72.092	122.819	−8.336	0.071	0.008	1.480	−0.240	−0.326
5	72.150	122.790	−9.012	0.071	0.006	1.481	−0.241	−0.322
6	65.144	122.617	−9.335	0.072	0.005	1.485	−0.215	−0.313
7	70.884	122.591	−8.930	0.072	0.003	1.485	−0.204	−0.313
8	70.633	122.540	−9.315	0.072	0.007	1.486	−0.196	−0.311
9	70.181	122.431	−9.140	0.072	0.003	1.487	−0.182	−0.305
10	68.864	122.302	−9.138	0.073	0.004	1.490	−0.143	−0.294
11	70.690	122.526	−9.298	0.072	0.005	1.486	−0.198	−0.319
12	74.926	123.365	−8.695	0.067	0.011	1.471	−0.324	−0.360
13	71.304	122.324	−9.336	0.073	0.006	1.489	−0.217	−0.307
14	67.203	121.716	−9.367	0.076	0.005	1.499	−0.093	−0.282
15	69.963	122.470	−9.366	0.072	0.006	1.487	−0.176	−0.309
16	70.688	122.499	−9.013	0.072	0.003	1.486	−0.198	−0.310
17	70.367	122.491	−8.962	0.072	0.003	1.486	−0.188	−0.312
18	69.915	122.457	−9.310	0.072	0.004	1.487	−0.175	−0.309
19	71.291	122.541	−9.483	0.072	0.003	1.485	−0.216	−0.311
20	69.495	122.448	−8.905	0.072	0.005	1.486	−0.162	−0.313
21	71.519	122.396	−9.437	0.073	0.004	1.488	−0.223	−0.303
22	70.580	122.597	−9.324	0.071	0.007	1.485	−0.195	−0.311
23	69.529	122.508	−8.894	0.072	0.006	1.486	−0.163	−0.313
24	74.885	123.200	−8.548	0.068	0.011	1.473	−0.323	−0.353
25	74.580	123.258	−8.341	0.068	0.011	1.473	−0.314	−0.355
26	75.864	123.104	−8.915	0.069	0.013	1.476	−0.351	−0.349
27	73.652	123.179	−8.188	0.068	0.009	1.474	−0.286	−0.355
28	74.466	123.250	−8.539	0.068	0.011	1.473	−0.310	−0.356
29	73.572	123.241	−8.215	0.068	0.012	1.473	−0.284	−0.355
30	74.560	123.263	−8.347	0.068	0.010	1.473	−0.313	−0.355
31	73.360	123.193	−8.353	0.068	0.011	1.474	−0.279	−0.357
32	74.192	123.283	−8.292	0.069	0.015	1.473	−0.302	−0.357
33	74.230	123.166	−8.570	0.069	0.011	1.475	−0.303	−0.351
34	75.069	123.302	−8.603	0.068	0.011	1.472	−0.327	−0.357
35	74.719	123.219	−8.217	0.068	0.011	1.474	−0.317	−0.353
36	74.722	123.230	−8.629	0.068	0.010	1.473	−0.318	−0.354
37	74.194	123.138	−8.139	0.068	0.010	1.475	−0.303	−0.351
38	74.517	123.116	−8.155	0.069	0.010	1.475	−0.313	−0.349
39	75.052	123.188	−8.509	0.068	0.011	1.474	−0.328	−0.351
40	74.628	123.142	−8.202	0.069	0.012	1.475	−0.315	−0.350
41	70.289	122.517	−8.677	0.073	0.007	1.485	−0.186	−0.313
42	70.690	122.507	−9.224	0.072	0.005	1.486	−0.199	−0.315
43	73.273	122.264	−9.205	0.074	0.005	1.490	−0.275	−0.313
44	71.796	122.477	−9.004	0.071	0.008	1.486	−0.231	−0.322
45	70.738	122.356	−8.761	0.074	0.011	1.489	−0.200	−0.307
46	70.328	122.471	−8.461	0.073	0.008	1.487	−0.188	−0.305
47	74.358	123.155	−8.050	0.069	0.010	1.475	−0.307	−0.350
48	74.739	123.148	−8.062	0.069	0.010	1.475	−0.318	−0.348
49	74.237	123.024	−7.968	0.069	0.010	1.476	−0.304	−0.346
50	74.405	123.092	−7.888	0.069	0.010	1.476	−0.309	−0.346
51	68.475	122.459	−8.754	0.073	0.008	1.487	−0.132	−0.288
52	69.939	122.196	−8.193	0.074	0.002	1.491	−0.176	−0.285
53	74.926	123.426	−8.487	0.068	0.011	1.471	−0.324	−0.364
54	70.652	122.358	−8.703	0.073	0.004	1.489	−0.197	−0.290
55	72.600	122.323	−9.352	0.074	0.007	1.489	−0.255	−0.283
56	74.922	123.262	−8.268	0.069	0.016	1.473	−0.326	−0.356
57	75.235	123.212	−8.261	0.069	0.013	1.474	−0.332	−0.352
58	71.611	122.645	−9.146	0.071	0.006	1.484	−0.226	−0.323
59	75.362	123.276	−8.459	0.068	0.010	1.473	−0.336	−0.357
60	70.728	122.529	−8.565	0.072	0.007	1.486	−0.199	−0.312
12c1	72.068	122.563	−9.052	0.071	0.007	1.484	−0.239	−0.328
12c2	72.123	122.637	−8.976	0.073	0.009	1.484	−0.241	−0.302
12d1	71.546	122.408	−9.236	0.073	0.007	1.487	−0.224	−0.301
12d2	71.542	122.408	−9.236	0.073	0.007	1.487	−0.224	−0.301

TABLE 3: Molecular Descriptors for the LL (*C*, *O*, *M*, and *P*) and DFT (σ_{xx}, σ_{yy}, and σ_{zz}, including σ_{iso}) Models

no.	*C*/ppm	*O*/ppm	*M*/ppm	*P*/ppm	σ_{xx}/ppm	σ_{yy}/ppm	σ_{zz}/ppm	σ_{iso}/ppm
1	0.0	0.0	0.0	0.0	−703.2	−610.0	375.1	−312.7
2	0.0	0.0	0.0	−7.0	−659.7	−629.0	374.6	−304.7
3	0.0	0.0	0.0	−23.6	−646.6	−605.1	372.8	−292.9
4	0.0	0.0	0.0	−31.2	−600.4	−592.5	373.3	−273.2
5	0.0	0.0	0.0	−11.6	−672.1	−576.2	373.8	−291.5
6	0.0	0.0	0.0	4.9	−663.6	−633.8	373.7	−307.9
7	0.0	0.0	0.0	10.5	−856.6	−205.4	374.4	−313.4
8	0.0	0.0	0.0	13.5	−639.9	−679.6	375.2	−314.8
9	0.0	0.0	0.0	29.6	−433.2	−947.6	377.5	−334.4
10	0.0	0.0	0.0	36.1	−692.3	−714.7	375.7	−343.8
11	0.0	11.0	0.0	0.0	−670.5	−678.0	397.8	−316.9
12	0.0	−54.9	0.0	0.0	−714.4	−333.5	352.1	−231.9
13	0.0	−4.5	0.0	0.0	−915.4	−425.6	413.3	−309.2
14	0.0	12.0	0.0	0.0	−586.7	−695.8	296.6	−328.6
15	0.0	9.0	0.0	0.0	−560.0	−806.2	393.4	−324.3
16	0.0	9.0	0.0	0.0	−415.3	−959.0	402.2	−324.0
17	0.0	0.0	6.5	0.0	−872.8	−462.8	372.6	−321.0
18	0.0	0.0	6.8	0.0	−587.6	−747.7	372.4	−320.9
19	0.0	0.0	4.4	0.0	−919.0	−404.8	373.4	−316.8
20	0.0	0.0	−0.6	0.0	−573.6	−753.0	371.9	−318.2
21	0.0	0.0	13.0	0.0	−943.6	−416.1	376.1	−327.9
22	0.0	0.0	2.0	0.0	−552.9	−783.8	373.5	−321.1
23	0.0	0.0	0.5	0.0	−903.3	−403.7	370.1	−312.3
24	−14.7	−45.9	0.0	0.0	−413.3	−656.5	371.9	−232.6
25	0.0	−54.9	6.5	0.0	−494.3	−574.7	353.2	−238.6
26	0.0	−54.9	13.0	0.0	−672.0	−416.5	356.0	−244.2
27	0.0	−54.9	0.5	0.0	−633.7	−438.5	350.5	−240.6
28	0.0	−54.9	5.6	0.0	−703.4	−328.3	327.4	−234.8
29	0.0	−54.9	−0.6	0.0	−604.6	−446.6	352.8	−232.8
30	0.0	−54.9	6.5	0.0	−337.9	−718.5	352.4	−234.7
31	0.0	−54.9	−0.6	0.0	−447.9	−598.4	349.9	−232.1
32	0.0	−54.9	0.5	0.0	−337.0	−702.7	349.5	−230.0
33	0.0	−54.9	6.8	0.0	−459.2	−610.0	351.7	−239.2
34	−14.7	−54.9	0.0	10.5	−335.4	−707.3	353.9	−229.6
35	−14.7	−54.9	6.5	10.5	−358.7	−711.2	355.1	−238.3
36	−14.7	−54.9	6.5	10.5	−706.2	−337.9	355.0	−229.7
37	−14.7	−54.9	13.0	0.0	−349.3	−727.4	352.4	−241.5
38	−14.7	−45.9	6.5	0.0	−702.0	−374.1	373.5	−234.2
39	−14.7	−45.9	0.0	10.5	−589.0	−473.7	374.1	−229.5
40	−14.7	−45.9	6.5	0.0	−537.6	−539.4	372.6	−234.8
41	0.0	0.0	−0.1	−23.6	−832.5	−425.5	367.1	−297.0
42	0.0	22.0	0.0	−7.0	−772.4	−615.4	377.1	−336.9
43	0.0	−9.0	0.0	−11.6	−732.4	−574.8	365.8	−313.8
44	0.0	−4.5	0.5	−11.6	−558.2	−727.4	404.4	−293.7
45	0.0	−4.5	0.5	−11.6	−895.9	−372.1	338.0	−310.0
46	0.0	9.0	5.9	−23.6	−823.4	−504.9	398.8	−309.8
47	−14.7	−54.9	13.0	10.5	−608.5	−462.3	352.4	−239.5
48	−14.7	−45.9	6.5	10.5	−339.0	−720.9	376.4	−227.8
49	−14.7	−45.9	13.0	0.0	−558.9	−523.6	372.2	−236.8
50	−14.7	−45.9	13.0	10.5	−350.2	−721.2	370.4	−233.7
51	0.0	9.0	−0.6	0.0	−930.4	−455.6	258.8	−232.6
52	0.0	18.0	13.0	10.5	−735.1	−332.1	−59.8	−375.6
53	0.0	−54.9	−0.6	−23.6	−460.2	−505.2	351.8	−204.6
54	0.0	18.0	0.0	10.5	−828.0	−597.2	329.6	−365.2
55					−1023.0	−432.0	356.3	−366.2
56					−506.0	−450.7	258.8	−232.6
57					−458.3	−602.0	350.7	−236.5
58					−804.9	−419.6	345.3	−293.1
59					−498.4	−550.3	351.7	−232.3
60					−599.0	−713.9	385.0	−309.3
12c1	0.0	−54.9	0.0	0.0	−687.7	−307.7	411.6	−311.2
12c2	0.0	−54.9	0.0	0.0	−967.1	−423.6	346.4	−348.1
12d1	0.0	−54.9	0.0	0.0	−718.1	−350.1	333.2	−245.0
12d2	0.0	−54.9	0.0	0.0	−718.1	−350.1	333.2	−245.0

(**42**), *p*-NO_2 (**10**), and *p*-CN (**9**). Molecular structures from PM3 and DFT calculations are consistent in showing that, besides the hydrogen-bonding structures in the low shifts region, there are other nonbonding intramolecular interactions between the benzaldehyde group and ortho substituents as well as between other adjacent substituents including formation of additional hydrogen bonds. This way, substituents containing −Me groups may react with O or H from the aldehyde group. The structure of **14** indicates that, besides the electron-withdrawing nature of *o*-NO_2, there is an

TABLE 4: Experimental and Calculated ^{17}O NMR Shifts with Absolute Deviations[a]

no.	δ_{exp}/ppm	δ_{PLS}/ppm	Δ_{PLS}/ppm	δ_{PCR}/ppm	Δ_{PCR}/ppm	δ_{DFT}/ppm	Δ_{DFT}/ppm	δ_{LL}/ppm	Δ_{LL}/ppm
1	563.2	572.1	**8.9**	571.7	8.5	563.2	0.0	564.0	0.8
2	561.4	564.6	3.2	564.2	2.8	555.2	6.2	557.0	4.4
3	526.9	549.0	**22.1**	548.6	**21.7**	543.4	**16.5**	540.4	**13.5**
4	532.8	535.9	3.1	535.9	3.1	523.7	**9.1**	532.8	0.0
5	545.7	551.1	5.4	550.6	4.9	542.0	3.7	552.4	6.7
6	568.9	572.3	3.4	575.5	6.6	558.4	**10.5**	568.9	0.0
7	570.1	565.3	4.8	564.4	5.7	563.9	6.2	574.5	4.4
8	570.3	562.1	8.2	562.6	7.7	565.3	5.0	577.5	7.2
9	593.6	572.9	**20.7**	572.3	**21.3**	584.9	8.7	593.6	0.0
10	590.1	578.0	**12.1**	578.2	**11.9**	594.3	4.2	600.1	**10.0**
11	575.0	565.5	**9.5**	565.6	**9.4**	567.4	7.6	575.0	0.0
12	505.8	512.1	6.3	512.4	6.6	482.4	**23.4**	509.1	3.3
13	555.0	566.3	**11.3**	565.5	**10.5**	559.7	4.7	559.5	4.5
14	576.0	592.6	**16.6**	593.6	**17.6**	579.1	3.1	576.0	0.0
15	573.0	568.2	4.8	568.8	4.2	574.8	1.8	573.0	0.0
16	573.0	568.3	4.7	567.4	5.6	574.5	1.5	573.0	0.0
17	569.3	568.2	1.1	567.6	1.7	571.5	2.2	570.5	1.2
18	570.8	572.6	1.8	572.7	1.9	571.4	0.6	570.8	0.0
19	568.4	572.2	3.8	571.1	2.7	567.3	1.1	568.4	0.0
20	555.2	564.9	**9.7**	565.6	**10.4**	568.7	**13.5**	563.4	8.2
21	574.5	572.0	2.5	570.5	4.0	578.4	3.9	577.0	2.5
22	566.0	561.2	4.8	561.8	4.2	561.6	4.4	566.0	0.0
23	562.3	561.8	0.5	562.8	0.5	562.7	0.5	564.5	2.2
24	507.0	513.4	6.4	513.3	6.3	483.1	**23.9**	503.4	3.6
25	516.2	510.8	5.4	511.0	5.2	489.1	**27.1**	515.6	0.6
26	522.8	513.3	**9.5**	512.8	**10.0**	494.7	**28.1**	522.1	0.7
27	512.1	517.1	5.0	517.3	5.2	491.1	**21.0**	509.6	2.5
28	514.7	513.7	1.0	514.0	0.7	485.3	**29.4**	514.7	0.0
29	511.8	509.5	2.3	510.7	1.1	483.3	**28.5**	508.5	3.3
30	509.0	513.5	4.5	513.3	4.3	485.2	**23.8**	515.6	6.6
31	510.0	514.6	4.6	515.8	5.8	482.6	**27.4**	508.5	1.5
32	513.9	501.2	**12.7**	502.8	**11.1**	480.5	**33.4**	509.6	4.3
33	518.2	517.2	1.0	517.4	0.8	489.7	**28.5**	515.9	2.3
34	507.0	512.2	5.2	512.1	5.1	480.1	**26.9**	504.9	2.1
35	515.0	509.7	5.3	509.6	5.4	488.8	**26.2**	511.4	3.6
36	509.0	517.1	8.1	516.9	7.9	480.2	**28.8**	511.4	2.4
37	520.0	513.6	6.4	513.4	6.6	492.0	**28.0**	507.4	**12.6**
38	512.0	513.8	1.8	513.3	1.3	484.7	**27.3**	509.9	2.1
39	507.0	513.1	6.1	512.8	5.8	480.0	**27.0**	513.9	6.9
40	513.0	508.9	4.1	508.9	4.1	485.3	**27.7**	509.9	3.1
41	550.0	554.4	4.4	554.9	4.9	547.5	2.5	540.3	**9.7**
42	585.0	565.2	**19.8**	565.1	**19.9**	587.4	2.4	579.0	6.0
43	565.0	561.8	3.2	559.1	5.9	564.3	0.7	543.4	**21.6**
44	545.0	550.2	5.2	550.2	5.2	544.2	0.8	548.4	3.4
45	538.0	547.8	**9.8**	548.8	**10.8**	560.5	**22.5**	548.4	**10.4**
46	560.0	551.0	**9.0**	551.4	8.6	560.3	0.3	555.3	4.7
47	518.0	512.5	5.5	512.2	5.8	490.0	**28.0**	517.9	0.1
48	505.0	512.0	7.0	511.3	6.3	478.3	**26.7**	520.4	**15.4**
49	517.0	513.4	3.6	512.9	4.1	487.3	**29.7**	516.4	0.6
50	513.0	511.5	1.5	510.9	2.1	484.2	**28.8**	526.9	**13.9**
51		562.9		564.5		594.8		572.4	
52		570.0		568.1		626.1		605.5	
53		508.9		509.1		455.1		484.9	
54		567.1		565.7		615.7		592.5	
55		564.7		562.6		616.7			
56		496.5		497.8		483.1			
57		504.5		504.5		487.0			
58		556.0		555.7		543.6			
59[b]		512.4		511.7		482.8			
60		551.6		551.7		559.8			
12c1	505.8	550.5	**44.7**	550.3	**44.5**	561.7	**55.9**	509.1	3.3
12c2	505.8	549.1	**43.3**	548.6	**42.8**	598.6	**92.8**	509.1	3.3
12d1[c]	504.5	561.6	**57.1**	560.9	**56.4**	495.5	9.0	509.1	4.6
12d2[c]	504.5	561.6	**57.1**	560.9	**56.4**	495.5	9.0	509.1	4.6

[a] Experimental shifts (δ_{exp}) and shifts calculated by the PLS (δ_{PLS}), PCR (δ_{PCR}), DFT (δ_{DFT}), and LL (δ_{LL}) models with absolute deviations Δ_{PLS}, Δ_{PCR}, Δ_{DFT}, and Δ_{LL}, respectively, as defined in Table 1. [b] It is assumed that the chemical shifts refer to the benzaldehyde group which has established a hydrogen bond with the hydroxyl group. The other benzaldehyde group is not considered in modeling the chemical shift for **59**. [c] Experimental shifts for pure liquid.

additional electron-withdrawal effect via a weak hydrogen bond which is established between the H atom from the aldehyde group and an O atom from the nitro group.

The main connection between the general benzaldehyde fragment's (GBF) structure and its ^{17}O shift lies in the fact that higher electron density at the carbonyl oxygen atom is directly related to lower chemical shift. In other words, more electrons produce a stronger induced magnetic field that opposes the external field (more intense screening or shielding of the oxygen nucleus), which results in smaller differences between the

TABLE 5: Correlation Matrices Including Experimental Chemical Shifts and Selected Molecular Descriptors, Shielding Tensor Components and LL Parameters

	E_e	E_{CC}	Δ_{HL}	σ_b	σ_r	D_{CC}	Q_{C2mul}	Q_{Omul}	δ_{exp}
E_e	1	0.847	0.705	−0.830	0.772	−0.859	−0.930	−0.884	−0.856
E_{CC}		1	0.750	−0.976	0.791	−0.997	−0.912	−0.976	−0.892
Δ_{HL}			1	−0.710	0.768	−0.768	−0.712	−0.782	−0.827
σ_b				1	−0.745	0.978	0.880	0.964	0.862
σ_r					1	−0.797	−0.803	−0.839	−0.891
D_{CC}						1	0.916	0.981	0.907
Q_{C2mul}							1	0.936	0.892
Q_{Omul}								1	0.928
	σ_{xx}	σ_{yy}	σ_{zz}	δ_{exp}					
σ_{xx}	1	0.689	−0.183	−0.468					
σ_{yy}		1	−0.214	−0.254					
σ_{zz}			1	0.391					
	C	*O*	*M*	*P*	δ_{exp}				
C	1	0.572	−0.475	−0.295	0.586				
O		1	−0.445	−0.243	0.910				
M			1	0.121	−0.326				
P				1	0.066				

TABLE 6: Calculated ^{17}O Carbonyl Chemical Shifts (in ppm) with Respective Deviations[a] (in brackets, in ppm), As Obtained from DFT Calculations for *o*-Hydroxybenzaldehyde Species

medium	**12**	**12c1**	**12c2**	**12d1−12d2**	experimental[b]
vacuum	**482.4 (23.4)**	561.7 (−55.9)	598.6 (−92.8)		505.8[c]
acetonitrile (PCM)	465.8 (43.3)	**531.4 (−22.3)**	545.2 (−36.1)		509.1[d]
chloroform (PCM)	469.2 (36.8)	539.3 (−33.3)	558.4 (−52.4)		506.0[e]
diethyl ether (PCM)	471.9 (33.1)	538.2 (−33.2)	559.3 (−54.3)		505.0[f]
cyclohexane (PCM)	465.8 (39.2)	548.9 (−43.9)	577.6 (−72.6)		505.0[f]
pure liquid				**2 × 495.5 (9.0)**[g]	504.5[h]

[a] Calculated values that represent an acceptable approximation to respective experimental values are in bold. [b] From the literature.[1] [c] Average value of all experimental data. [d] Experimental data measured in acetonitrile. [e] Experimental data measured in $CDCl_3$. [f] Experimental data measured in dioxane, which could not be calculated for this solvent using the PCM method incorporated in the Gaussian software. Therefore, the calculations were performed for two structurally closest solvents to dioxane, diethyl ether and cyclohexane. [g] Values for the two molecules **12d1** and **12d2** in the dimer. [h] Compound **12** in the liquid state.

ground and excited states of the oxygen nucleus. The electron density at the carbonyl oxygen, as shown in some cases, can be elevated by electron donation of the *o*-hydrogen donor group (*o*-OH), coplanar benzene ring, and substituent electron donation. On the contrary, electron-withdrawing substituents as well as ortho substituents that sterically hinder the aldehyde−benzene coplanarity weaken the electron content of the aldehyde oxygen (deshielding effects).

X Data (QSPR data set) Compared to the DFT and LL Data Sets. Several molecular descriptors (110 in total) were calculated, mostly electronic, structural, and combined descriptors from quantum-chemical calculations. From this initial set eight molecular descriptors (Table 2) were selected via variable selection methods to build the final regression models: E_e, electron−electron repulsion energy at C_2 (one-center term), calculated by MOPAC; E_{CC} − C_1−C_2, nuclear−nuclear repulsion energy, calculated by MOPAC; Δ_{HL}, HUMO−LUMO gap, the difference between energies of the frontier orbitals HOMO (the highest occupied molecular orbital) and LUMO (the lowest unoccupied molecular orbital), calculated by Titan; σ_b, standard deviation of six C−C bond lengths in the benzene ring; σ_r, standard deviation of eight delocalized bond lengths (C_1−C_2, C_1−O, and six ring C−C bond lengths); D_{CC}, C_1−C_2 bond length; Q_{C2mul}, Mulliken partial atomic charge of C_2, as obtained by Titan; and Q_{Omul}, Mulliken partial atomic charge of the carbonyl oxygen O, calculated by Titan.

These descriptors exhibit high correlation with experimental shifts δ_{exp} (Table 2) with absolute correlation coefficients being 0.83−0.93 and satisfactory bivariate plots (plots not shown). Taking into account the signs of the correlation coefficients, the following explanation of shift-descriptor relationships can be given. The cumulative effect of substituents which account for electron donation to the benzene ring is further transferred via C_1−C_2 and C_1−O bonds to O, whose electron density is enriched, and consequently, its chemical shift is lowered. Therefore, increased negative charge Q_{Omul} at O, increased negative charge Q_{C2mul} at C_2 (C_2 mediates electron transfer between O and the ring), and enhanced repulsion between electrons located at C_2 (increase of repulsion energy E_e) are directly related to the decrease of δ_{exp}. The aldehyde O and the ring are electron-withdrawing and -donating systems, respectively, when there are no substituents to change this relationship. In this sense, shortening of the bond C_1−C_2 (decrease of D_{CC}) or weakening of the nuclear repulsion between these atoms (decrease of E_{CC}) means intensified electron delocalization between O and the ring. High electron delocalization in the benzene ring is a measure of its aromaticity[25–27,38,57–59] and its electron-donating ability to O, which is visible through the regular hexagonal structure with small bond length variation σ_b and reduced HOMO−LUMO gap Δ_{HL}. The other bond length variation σ_r is determined predominantly by C_1−C_2 and C_1−O bond lengths, which do not tend to equalize with the ring bond lengths since O is a much better electron-withdrawing than -donating system. The final electron delocalization effect results in elevated σ_r values at low ^{17}O shifts.

Descriptor intercorrelations (Table 5) are moderate to very high. It is recommended in QSPR[18] that such intercorrelations have correlation coefficients below 0.90. However, it is not always possible to follow this recommendation. In the present case, all eight descriptors positively contribute to the quality of

TABLE 7: Comparison of the Regression and Parametric Models by Means of Various Statistical Parameters[a]

parameters	PLS	PCR	DFT	LL
model statistics				
training set size[b]	50	50	[50]	[50]
LVs or PCs (%Var)[c]	2 (92.3%)	3 (96.2%)		
p	2	3	6	6
SEV/ppm[d]	9.1	9.1		
SEC/ppm[e]	8.4	8.6	**20.5**	6.9
Q, Q^{2}[f]	0.946, 0.895	0.946, 0.894		
R, R^{2}[g]	0.957, 0.915	0.956, 0.913	0.973, **0.545**	0.975, 0.948
Δ_{max}/ppm ($\Delta_{rel\text{-}max}$)[h]	22.1 (24.9%)	21.7 (24.5%)	**33.4 (37.7%)**	**21.6 (24.4%)**
$(\Delta_{max} - \Delta_{min})$/ppm[h]	21.6	21.2	**33.4**	**21.6**
$\langle\Delta\rangle$/ppm ($\langle\Delta_{rel}\rangle$)[h]	6.6 (7.5%)	6.7 (7.6%)	**14.9 (16.8%)**	4.3 (4.9%)
$w\langle\Delta\rangle$/ppm ($w\langle\Delta_{rel}\rangle$)[h]	7.0 (7.9%)	7.3 (8.2%)	**17.3 (19.5%)**	5.0 (5.6%)
$N_{rel>10\%}$[h]	13	11	**27**	8
T_{over}/ppm[i]	0.3	0.4	**588.5**	**−34.1**
$N_{over}/N_{under}/N_{zero}$[i]	24/26/0	25/25/0	**35/14/1**	20/20/10
leave-N-out cross-validation[j]				
$\langle Q^{2}_{LNO}\rangle$	0.888	0.888		
y-randomization[k]				
Q^{2}_{yrand}	−0.593	−0.169		
R^{2}_{yrand}	0.069	0.001	[−0.321]	[−0.555]
HCA-based bootstrapping ($S = 0.70$)[l]				
Q^{2}_{bstr}	0.887	0.886		
R^{2}_{bstr}	0.914	0.912	**[0.557]**	[0.956]
$C_{HCA>10\%}$	5 out of 5	5 out of 5	**5 out of 9**	**4 out of 8**
external validation				
training set/ext. valid. set sizes[b]	40/10	40/10	[40/10]	[40/10]
LVs (% Var)[c]	2 (92.6%)	2 (93.0%)		
p	2	2	**[6]**	**[6]**
training set statistics				
SEV_{ev}/ppm[d]	9.6	9.7		
SEP/ppm	8.8	9.1	**[21.3]**>	[7.3]
Q_{ev}, Q^{2}_{ev}[f]	0.942, 0.886	0.940, 0.883		
R_{ev}, R^{2}_{ev}[g]	0.954, 0.911	0.952, 0.906	[0.969, **0.534]**	[0.973, 0.946]
$\Delta_{max\text{-}ev}$/ppm ($\Delta_{rel\text{-}max\text{-}ev}$)[h]	21.6 (24.4%)	22.3 (25.2%)	**[33.4 (37.7%)]**	**[21.6 (%24.4)]**
$(\Delta_{max} - \Delta_{min})$/ppm[h]	20.9	22.0	**33.4**	**21.6**
$\langle\Delta_{ev}\rangle$/ppm ($\langle\Delta_{rel\text{-}ev}\rangle$)[h]	6.7 (7.6%)	6.9 (7.8%)	**[15.3 (17.2%)]**	[4.3 (4.9%)]
$w\langle\Delta_{ev}\rangle$/ppm ($w\langle\Delta_{rel\text{-}ev}\rangle$)[h]	7.2 (8.2%)	7.5 (8.4%)	**[18.5 (20.9%)]**	[5.2 (5.9%)]
$N_{rel>10\%\text{-}ev}$[h]	11	9	**[24]**	[6]
$T_{over\text{-}ev}$/ppm[i]	0.1	0.1	**464.0**	**−33.0**
$N_{over\text{-}ev}/N_{under\text{-}ev}/N_{zero\text{-}ev}$[i]	21/19/0	20/20/0	**26/13/1**	14/16/10
k	1.00	1.00	**[1.02]**	**[1.00]**
k'	1.00	1.00	**[0.98]**	**[1.00]**
R_0^{2}	1.000	1.000	**[0.936]**	**[0.999]**
$R_0'^{2}$	1.000	1.000	**[0.851]**	**[0.999]**
$\lvert R_0^{2} - R_0'^{2}\rvert$	2.3×10^{-5}	2.1×10^{-5}	**[0.085]**	**[1.8×10^{-4}]**
$(R_0'^{2} - R_0^{2})/R^{2}_{ev}$	−0.098	−0.104	**[−0.753]**	[−0.057]
$(R_0^{2} - R_0'^{2})/R^{2}_{ev}$	−0.098	−0.104	**[−0.593]**	[−0.057]
external validation set statistics				
SEC_{ext}/ppm[e]	8.1	7.8	**[31.5]**	**[9.8]**
R_{ext}[g], Q^{2}_{ext}[f]	0.970, 0.937	0.976, 0.943	[0.991, **0.595]**	[0.987, 0.961]
$\Delta_{max\text{-}ext}$/ppm ($\Delta_{rel\text{-}max\text{-}ext}$)[h]	12.8 (14.4%)	13.7 (15.5%)	**[29.7 (33.5%)]**	[10.0 (11.3%)]
$(\Delta_{max} - \Delta_{min})$/ppm[h]	9.5	11.4	**[27.2]**	**[10.0]**
$\langle\Delta_{ext}\rangle$/ppm ($\langle\Delta_{rel\text{-}ext}\rangle$)[h]	6.1 (6.9%)	5.8 (6.5%)	**[13.3 (15.0%)]**	[4.1 (4.6%)]
$w\langle\Delta_{ext}\rangle$/ppm ($w\langle\Delta_{rel\text{-}ext}\rangle$)[h]	6.5 (7.3%)	6.1 (6.9%)	**[19.0 (21.4%)]**	[5.9 (6.6%)]
$N_{rel>10\%\text{-}ev}$[h]	2	1	**[4]**	**[2]**
$T_{over\text{-}ext}$/ppm[i]	13.5	16.4	**124.5**	−1.1
$N_{over\text{-}ext}/N_{under\text{-}ext}/N_{zero\text{-}ext}$[i]	5/5/0	5/5/0	**9/1/0**	6/3/1

[a] Statistical parameters from Table 1. Additional details can be found in this table. The values of parameters in square brackets for the DFT and LL models means that these values are only numerical equivalents to the corresponding values for the regression models, based on calculations for the same sets of benzaldehydes. Among these values, those in bold show where the DFT and LL models are not better than the regression model. [b] Number of samples (molecular systems) in training and external validation sets. [c] Number of latent variables (LVs) and principal components (PCs) in the PLS and PCR models, respectively, with the respective contents of the total variance (Var %). [d] SEV parameters are distinguished by index: no index for the proposed model, and "ev" index for external validation. [e] SEC parameters are distinguished by index: no index for the proposed model, "ev" index for the training set in external validation, and "ext" index for the external validation set. Q^{2}_{ext} is calculated by using the expression for Q^{2} where the mean of experimental values is for the training set. [f] Correlation coefficients Q and Q^{2} are distinguished by index: no index for the proposed model, and "ev" index for the external validation. [g] Correlation coefficients R and R^{2} are distinguished by indices: no index for calibration, "ev" index for the training set from external validation, and "ext" for the external validation set. [h] Parameters based on absolute or relative errors are distinguished by additional indices: no indices for the proposed model, indices "ev" for the training set of external validation, and indices "ext" for the external validation set. [i] Parameters of overprediction and underprediction are distinguished by additional indices: no indices for the proposed model, indices "ev" for the training set of external validation, and indices "ext" for the external validation set. [j] Leave-N-out cross-validations with $N = 1, 2,..., 10$. [k] Parameters for 10 **y**-randomizations. [l] Parameters for ten HCA-based bootstrapping using the similarity index $S = 0.70$.

TABLE 8: Details of Regression[a] (PLS and PCR) and Parametric (DFT and LL) Equations

PLS	PCR	DFT	LL
-0.104 $[E_e]_{au}$	-0.155 $[E_e]_{au}$	1/3 σ_{xx}(**1**)	C
-0.064 $[E_{CC}]_{au}$	-0.064 $[E_{CC}]_{au}$	1/3 σ_{yy}(**1**)	δ_o
-0.228 $[\Delta_{HL}]_{au}$	-0.234 $[\Delta_{HL}]_{au}$	1/3 σ_{zz}(**1**)	$\delta_{o'}$
0.037 $[\sigma_b]_{au}$	0.034 $[\sigma_b]_{au}$	$-1/3$ σ_{xx}	δ_m
-0.288 $[\sigma_r]_{au}$	-0.254 $[\sigma_r]_{au}$	$-1/3$ σ_{yy}	$\delta_{m'}$
0.088 $[D_{CC}]_{au}$	0.072 $[D_{CC}]_{au}$	$-1/3$ σ_{zz}	δ_p
0.112 $[Q_{C2mul}]_{au}$	0.128 $[Q_{C2mul}]_{au}$	250.5	564.0
0.122 $[Q_{Omul}]_{au}$	0.104 $[Q_{Omul}]_{au}$		

[a] Autoscaled descriptors E_{CC}, Q_{Oesp}, σ_d, d_{CC}, and Q_{C2mul} are marked with brackets $[\ ]_{au}$.

TABLE 9: Predictions of External Validation of the Regression Models (PLS and PCR)[a]

no.	δ_{exp}/ppm	δ_{PLS}/ppm	$\%\Delta_{PLS}$	δ_{PCR}/ppm	$\%\Delta_{PCR}$
2	561.4	564.7	3.7	563.7	2.6
5	545.7	551.0	6.0	550.5	5.4
7	570.1	565.8	4.9	563.5	7.4
10	590.1	577.3	14.4	576.4	15.5
22	566.0	560.2	6.5	561.1	5.5
26	522.8	511.9	12.3	515.5	8.2
27	512.1	518.0	6.7	516.2	4.6
34	507.0	512.2	5.9	513.0	6.8
41	550.0	553.9	4.4	553.4	3.8
49	517.0	513.7	3.7	512.5	5.1

[a] Experimental shifts (δ_{exp}) and shifts calculated by the PLS (δ_{PLS}) and PCR (δ_{PCR}) models for the external validation set with respective relative deviations in percent (%) ($\%\Delta_{PLS}$ and $\%\Delta_{PCR}$).

the models. Exclusion of any descriptor has always resulted in worsened models. On the other hand, descriptors of different nature (energies, charges, geometrical) and origin may aid in understanding the electronic structure of the benzaldehyde system. It is worth noting that after establishing a regression model for **1**−**50** a trained professional needs less than 20 min to predict the carbonyl ^{17}O shift for a new benzaldehyde.

The LL model (eq 1) is extremely fast in predicting the ^{17}O shift for a benzaldehyde (1−2 min). Positive and negative values of its descriptors correspond directly to deshielding and shielding effects of substituents and solvents. However, this model and its data set (Table 3) are questioned in this work because of three reasons. First, the LL model is not general since it is limited to a small number of simple substituents at some substituent positions. Second, Li and Li[1] claimed very low average deviation of calculated from experimental data $\Delta = 2.9$ ppm. It was an artificial effect because the experimental data were not averaged for multiple measurements, but instead, the most suitable experimental data were used to report minimum deviations. Furthermore, C is a negative correction in cases when the ^{17}O shifts were measured in polar solvents ($CDCl_3$). In some cases, besides using $CDCl_3$, a measurement in another solvent was performed, yielding similar values. For such cases, Li and Li have not used any correction. In some other cases, measurements were performed using only pure compounds in the liquid state, some of them being rich in hydrogen-bonding groups (relatively polar compounds). No correction C was applied for these substances. The empirical LL model is also questioned in terms of statistical evaluation. The model has been constructed using parameters from a series of previous MLR studies but with no reported errors for the increment values. Although the model seems to be simple, presenting a virtually univariate problem (δ_{LL} is understood as a unique variable), in fact, it needs six variables and a constant (eq 1, Table 8). Table 4 shows that O, M, P, and C are poorly intercorrelated. Only O (o-substituent effects) is highly correlated with δ_{exp}, C (having only two distinct values, see eq 4c) is moderately correlated with δ_{exp}, and the other descriptors are poorly correlated with δ_{exp}. Descriptor-shift bivariate plots are not satisfactory for O and C (plots not shown). Hence, these descriptors cannot be used in regressions.

Results of different DFT treatments of o-hydroxybenzaldehyde are presented in Figures 4 and 5 and Table 6. According to the structures of o-benzaldehyde fragments from the CSD database, there are only three modes of interaction between the aldehyde and hydroxyl groups, where the OH groups are not necessarily coplanar with the ring. These are **type 12**, **type 12c1**, and **type 12c2** in decreasing order of respective frequencies (Figure 5), characterized well in terms of the distance between the aldehyde O and hydroxyl H. Therefore, three conformers of o-hydroxybenzaldehyde in vacuum were modeled, having OH coplanar with the ring (Figure 4): **12** (resonance-assisted hydrogen bond), **12c1** (weak H···O interaction), and **12c2** (O···O interaction), yielding electronic energies and ^{17}O shifts in excellent agreement with the frequencies of the respective o-benzaldehyde fragments in the crystalline state (Figure 5). Inclusion of solvent effects via the PCM method (Table 6) resulted in substantial shielding effects up to 40.4 ppm and two acceptable models with deviations below 30 ppm: **12** in vacuum and **12c1** in acetonitrile. In another modeling approach, the dimer **12d1**−**12d2** was obtained via two intermolecular hydrogen bonds which released -14 kcal mol^{-1}, which is less than twice the stabilization caused by the hydrogen bond in **12** (-11 kcal mol^{-1}). One can notice that all benzaldehydes with an intramolecular hydrogen bond (Figure 2: **12**, **24**−**40**, and **47**−**50**), modeled in absence of solvents have underpredicted shifts with deviations above 20 ppm (Table 4).

The DFT model (eqs 2, 3, and 5) is questioned in this work because of four reasons. First, as illustrated (Table 6) and according to the literature,[5–7] an optimum modeling has to be selected to calculate ^{17}O carbonyl shifts: solute conformers with inclusion of solvent effects, solvent−solute, and solute−solute complexes. Second, DFT procedures for **1**−**60** in vacuum or solvent need hours and for solvent−solute and solute−solute complexes even days. Third, DFT procedures are limited by the size and complexity of systems under study. The fourth disadvantage is the lack of error propagation data and statistical validation of calculated shifts. Table 5 shows that the DFT data (σ_{xx}, σ_{yy}, and σ_{zz} variables) are characterized by low to moderate intercorrelations and correlations with δ_{exp}. The bivariate plots are not satisfactory (plot not shown). This data set is not useful for a regression analysis.

QSPR Regression Models. Comparative statistics for the PLS and PCR models is presented in detail in Table 7, which is the companion of Table 1. The models are almost the same in most parameters, although small differences as correlation coefficients and deviations (errors) for the models and their validations show a little favor for PLS. The main advantage of PLS lies in using two latent variables, while PCR needs three principal components. On the other hand, PCR results in fewer chemical shifts with significant errors ($N_{rel>10\%}$) than PLS, both in calibration and external validation. PCR has a slightly better balance of overpredictions and underpredictions (N_{over}/N_{under}) than PLS. PLS and PCR predictions for **1**−**50** differ at most by 2.1 ppm, and for **51**−**60** the differences are not greater than 3.2 ppm (Table 4). No chance correlations have been found by **y**-randomizations. Internal validations and bootstrappings showed the robustness of the models. Both models can be considered

TABLE 10: Types of Directional Intermolecular Interactions[a] between Selected Benzaldehyde Systems and Other Species in the Crystalline State[b]

type of interaction[c]	*o*-hydroxybenzaldehyde (**12**) + other species[d]	benzaldehydes (GBF)[e] + solvents[f]
moderate hydrogen bonds	HO···HO C=O···HO C=O···HN C(O)H···O_2N	none
weak and very weak hydrogen bonds	CH···OH HO···HC CH···O=C C=O···HC CH···HC CH···π π···HC OH···Cl	C=O···HC (an, cf) π···HC (an) C(O)H···Cl (cf) CH···Cl (cf) *o*-OH···HC (cf) *m*-OH···$O(CH_2)_2$ (do)
orbital interactions	π···π	π···π (an) π···Cl (cf) C=O···Cl (cf)

[a] Directional intermolecular interactions are characterized by measurable geometric parameters defined by the atoms involved in these interactions. [b] Crystal structures retrieved from the Cambridge Structural Database (CSD) include, besides benzaldehyde species, other molecular and ionic species, such as metals, solvents, and organic species. These structures are selected for analysis because of reported atomic coordinates, the absence of disorder or errors that could make the analysis doubtful, and the presence of intermolecular interactions of interest. [c] All interactions are written in the order of interacting fragments, benzaldehyde···other species, in order to distinguish interactions in which the benzaldehyde fragments behave as hydrogen donors from those in which the fragments are hydrogen acceptors. The benzaldehyde group is shortly written as C(O)H to be distinguished from aromatic CH in benzaldehydes. [d] Intermolecular interactions involving **12** in the neutral state or complexed to metals, as found in five selected CSD crystal structures which contained atomic coordinates of two or more different chemical species. [e] GBF or general benzaldehyde fragment is the hydrogen-depleted aromatic C_6−C(O)H fragment which is the skeleton of a substituted benzaldehyde or part of a larger chemical system (organic or organometallic ion/molecule or a metallic complex). [f] Intermolecular interactions involving GBF and three solvents: an, acetonitrile; cf, chloroform; do, 1,4-dioxane. On the basis of 17 selected CSD crystal structures with atomic coordinates of at least two chemical species.

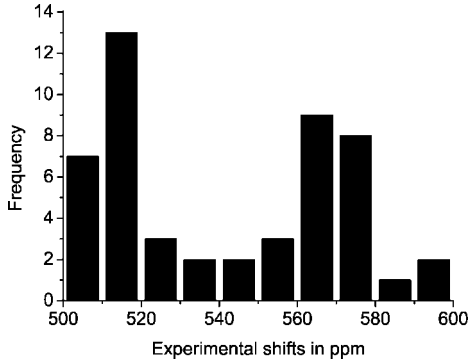

Figure 3. Histogram for experimental shifts δ_{exp} for carbonyl ^{17}O atoms in benzaldehydes **1**−**50**.

parsimonious and used for practical purposes for substituted benzaldehydes.

The similarity between the models can be seen also in the respective regression vectors (Table 8), which are equal in the signs of their components, and do not differ in more than 50% of absolute values. The positive/negative signs of the components are equal to those of the corresponding descriptor-shift correlation coefficients (Table 5). This means that the chemical background of the regression vectors has been already explained by the correlation analysis. The regression vectors provide additional information: relationships between the absolute values of the regression coefficients (i.e., their importance to the models) and the corresponding molecular fragments. These values are above 0.2 for global properties (Δ_{HL}, a global descriptor; σ_r, a quasi-global descriptor for GBF), between 0.1 and 0.2 for very local features (Q_{C2mul} and E_e for the C_2 atom and Q_{Omul} for the O atom), and below 0.1 for fragmental descriptors (E_{CC} and D_{CC} for the C_1–C_2 bond and σ_b for the ring).

PLS and PCR models are also similar in external validations (Table 9) with reasonable predictions that differ up to 3.6 ppm. Two molecules with the largest deviations (**10** and **26**) contain one of the strongest electron-withdrawing substituents (−NO_2). Several correlation coefficients, errors, and other parameters were calculated for the training and external validation sets (Table 7), proving the robustness of the models.

Regression Models Compared to the DFT and LL Models. Sets of benzaldehydes in external validation of the regression models as well as additional sets that were defined in HCA-bootstrappings were very useful in checking the validity of the two parametric models DFT and LL. According to Table 7, the DFT model is obviously worse or not better than PLS and PCR in most parameters. This is noticeable especially for errors (SEV, SEC, and deviation parameters) and correlation coefficients (Q^2 and R^2 parameters). The DFT model totally fails in most parameters for external validation and bootstrappings (R^2_{bstr}). The LL model is not better than the regression models in maximum deviation parameters (Δ_{max}, $\Delta_{rel\text{-}max}$, and $\Delta_{max} - \Delta_{min}$ for the training sets) and several external validation parameters (SEC_{ext}, k, k', R_0^2, $R_0'^2$, and derived parameters).

The models are compared in Figure 6 in terms of frequency distribution of deviations $y_e - y_c$ for **1**−**50**. Three types of distribution can be seen: PLS/PCR, DFT, and LL type. Among these, only the DFT type is not symmetric and not centered about the zero deviation. As already discussed, most DFT deviations are either small or large (due to internal hydrogen bonds, see Table 4). The LL type is well centered and symmetric around the zero deviation, while the PLS/PCR type retains these features but the maxima are at −5 and 5 ppm. The most reasonable statistics of errors seems to be that of the PLS/PCR type.

Figure 7 shows frequency distributions of predicted shifts y_c for **1**−**50** as obtained by the four models. When these distributions are compared with the experimental distribution (Figure 3) it becomes clear that the PLS/PCR distribution profile is most similar to the experimental one with some underpopulation in the ranges 500−510 and 520−540 ppm and overpopulation in the range 510−520 ppm. The DFT distribution is more radical in this sense, while the LL distribution shows substantial overpopulation in the ranges 500−510 and 520−560 ppm. The PLS and PCR models are superior to the DFT and LL models in terms of additional parameters accounting for overprediction

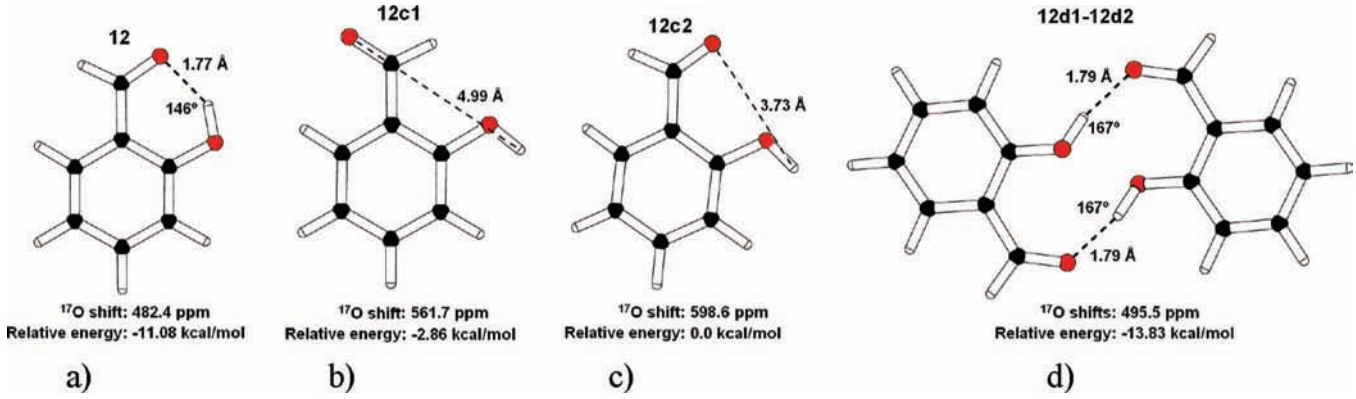

Figure 4. Molecular structures of four *o*-hydroxybenzaldehyde systems as obtained from DFT calculations, showing hydrogen-bonding geometry and/or —(H)C═O···HO— distance, respective relative electronic energy together with calculated ^{17}O chemical shift(s): (a) **12** with an internal hydrogen bond, (b) **12c1** with a weak —C(O)H···OH interaction, (c) **12c2** with a —(H)C═O···OH interaction, and (d) **12d1**—**12d2** dimer where the benzaldehyde and hydroxyl groups are not coplanar with the benzene rings.

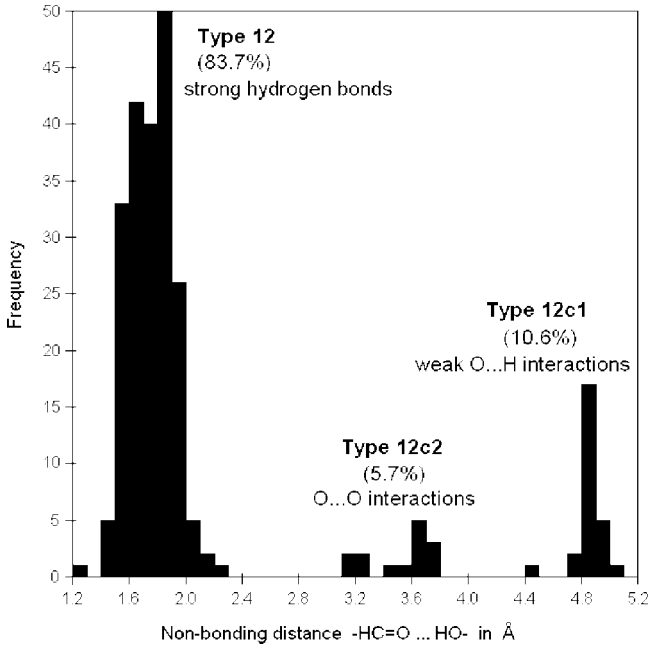

Figure 5. Absolute and relative frequencies of —(H)C═O···HO— distances in *o*-hydroxybenzaldehyde fragments as retrieved from the Cambridge Structural Database. The three types of fragments basically correspond to the three conformers of **12** presented in Figure 4 with the only structural difference that *o*—O—H bonds in crystal structures are not necessarily coplanar with the benzene ring.

(Table 7) of the training sets. Although the LL model seems to be superior when the external validation set is considered, since this set is small, the observed trends are not so statistically reliable. The overprediction parameters are given by total overprediction (T_{over}), number of overpredictions, underpredictions, and zeros ($N_{over}/N_{under}/N_{zero}$), and the number of HCA clusters containing predictions with elevated deviations ($C_{HCA>10\%}$, which accounts for a uniform distribution of such samples, Figure 8). The number of parameters p (Tables 7 and 9) shows that the DFT and LL models are not simpler than the regression models. Weighted errors $w\langle\Delta\rangle$ and $w\langle\Delta_{rel}\rangle$ take into account p and show that the LL model is not substantially better than the regression models in terms of deviations.

Shift predictions and deviations for **1**—**60** (Table 4) are a useful way to compare the regression and parametric models. PLS and PCR, although formally without solvent effects, take into account these effects indirectly: overall molecular topology and electronic features of the benzaldehyde group already incorporate information about possible interactions with solvents. All four models have only two samples in common with significant errors (**3** and **45**), meaning that other samples with errors above 10% follow distinct trends. When considering predictions for **51**—**60** and δ_{exp} for the most similar bezaldehydes, it seems that PLS and PCR offer reasonable predictions. DFT probably overpredicts the shift for **52** and underpredicts shifts for benzaldehydes with hydrogen bonds **53**, **57**, and **59**. LL predicts shifts only for **51**—**54**, where the shift for **53** may be underpredicted and shifts for **52** and **54** overpredicted. Shift predictions for conformers of **12** (**12c1** and **12c2**) and its dimer (**12d1**—**12d2**) show reasonable differences between the monomers and dimer in PLS/PCR, large differences in DFT, and negligible differences in LL.

It can be concluded that the two regression models are of equal quality and recommendable for prediction of ^{17}O carbonyl chemical shifts in substituted benzaldehydes. The DFT model is poor mainly due to hydrogen-bonding and conformational effects. The LL model generally overpredicts the shifts and results in an artificially large number of predictions with zero deviations.

Exploratory Analysis. HCA with complete linkage for **1**—**50** (QSAR data set) shows that there are five clusters at the similarity index $S = 0.70$ (Figure 8): C1, C2, C3, C4, and C5. Two clusters can be further divided into two subclusters: C1 into C1-A and C1-B and C4 into C4-A and C4-B. All benzaldehydes in C1 possess ^{17}O shifts with strong shielding due to the internal hydrogen bond —(H)C═O···HO—, as noticed in all previous analyses. Other clusters contain benzaldehydes with weak shielding or deshielding because of electron-withdrawing effects of substituents and substituent positions. Figure 8 illustrates a rather uniform cluster distribution of benzaldehydes from external validation and those with significant relative errors.

When the QSAR data set is analyzed by PCA, other visual characteristics that complement the HCA dendogram become rather apparent (Figure 9). The first two principal components describe 96.2% of the total variance. Similarly to the HCA plot, the scores plot shows that the C1 cluster is separated from the conglomeration of C2—C4 clusters along PC1, while C5 is isolated at the opposite side. In fact, C3 can be defined as a unique cluster, while C2 and C4 are partially mixed.

Previous correlation and regression analyses have explained relationships between benzaldehydes **1**—**50** and respective descriptors. However, these analyses are not able to give insight into such relationships within clusters, which HCA and PCA

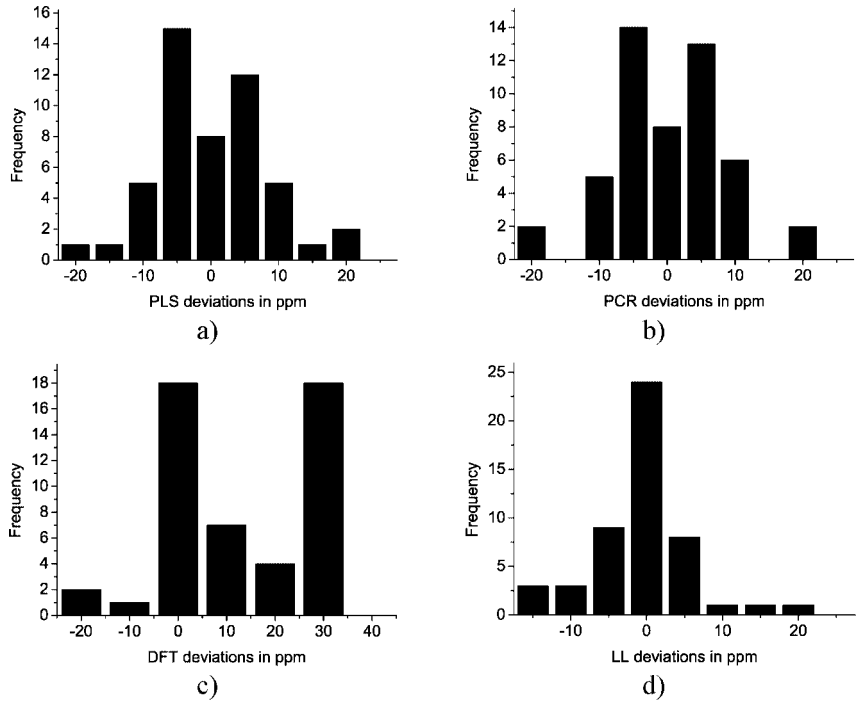

Figure 6. Frequency distribution of shift deviations $y_e - y_c$ for **1**−**50** as obtained from the four models: (a) PLS, (b) PCR, (c) DFT, and (d) LL.

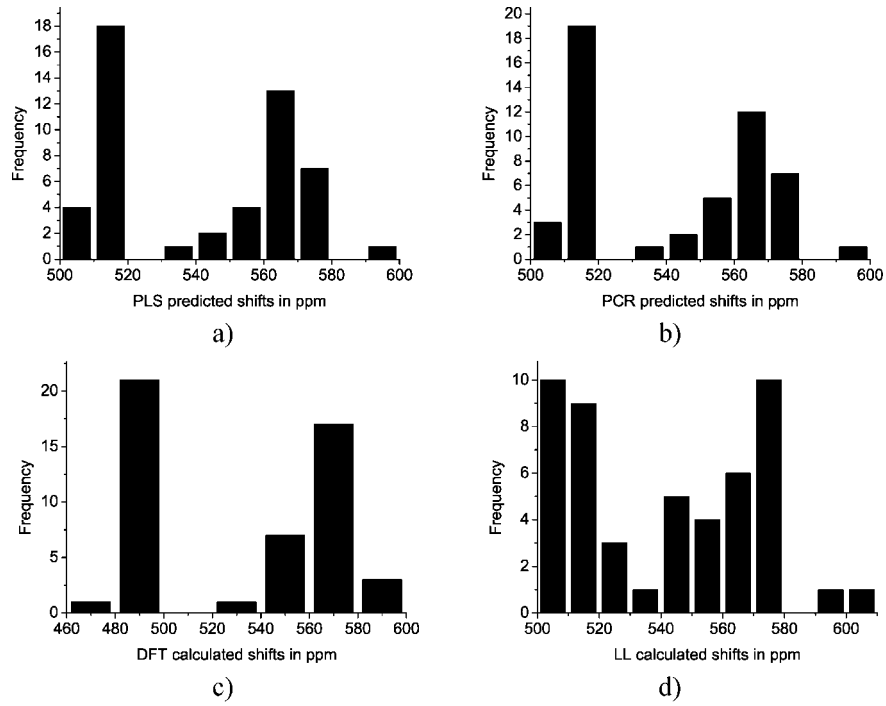

Figure 7. Frequency distribution of calculated shifts y_c for **1**−**50** as obtained from the four models: (a) PLS, (b) PCR, (c) DFT, and (d) LL.

can easily do. When the signs of descriptor-shift correlation coefficients (Table 5) or regression vector components (Table 8) are taken into account, knowing that three descriptors always have negative values (Δ_{HL}, Q_{C2mul}, and Q_{Omul}) and other descriptors are always positive, the following interpretation of PC1 can be given. Two groups of descriptors, placed at the negative and positive ends of PC1 (Figure 9b), correspond to negative and positive signs of the descriptor-shift correlation coefficients, respectively. Hence, shielding effects are pronounced when the gap Δ_{HL} decreases and the repulsion energies (E_e and E_{CC}) and aromaticity index σ_r increase. On the contrary, deshielding effects follow the PC1 increase, which is due to the increase of the aromaticity parameter σ_b and bond length D_{CC}, while the absolute values of negative charges at O (Q_{Omul}) and C_2 (Q_{C2mul}) decrease.

PC1 can be considered as a measure of deshielding (when positive) or shielding (when negative). Accordingly, samples at high PC1 > 2.72 (**6**, **9**, **10**, **14**, **18**, and **21**) contain strong electron-withdrawing groups −F, −CN, and $-NO_2$. Samples at low PC1 < −3.15 (**12**, **32**, and **34**) as well as other samples in C1 are well characterized by hydrogen-bonding effects of *o*-OH: partial electron transfer from hydroxyl H to the carbonyl O. Samples in C2−C4 occupy the central space of the scores

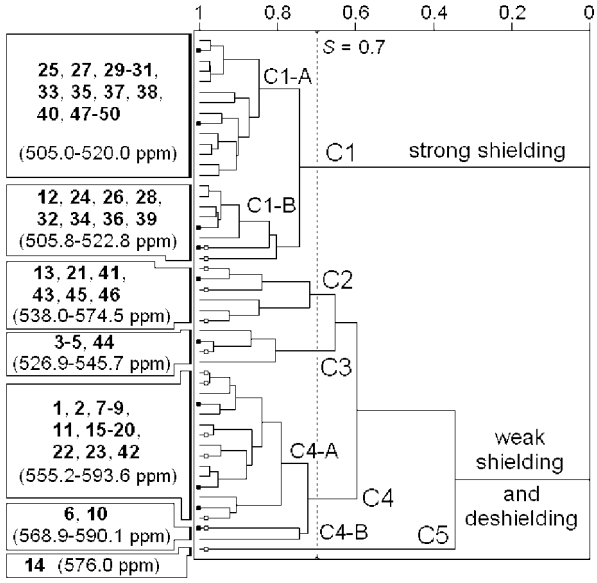

Figure 8. HCA dendogram with complete linkage for **1**−**50** from the QSAR data set. Clusters C1−C5 can be distinguished at similarity index $S = 0.70$ (dashed vertical line), and additional subclusters of C1 and C4 can be well noticed also. Particular (sub)cluster composition, corresponding experimental shift ranges, and shielding character are marked in the dendogram. Samples with solid squares are those selected for the external validation set. Samples with white squares are those having errors above 10% as obtained from the PLS and/or PCR model.

plot as a continuous transition between systems with electron-donating and systems with electron-withdrawing effects.

Final Structural Considerations. There are some conceptual differences between the QSPR and DFT/LL models. First, the DFT/LL models deal with increments of the same property, and therefore, such models are interesting mainly for researchers which measure and/or calculate chemical shifts. The QSPR models correlate chemical shifts with various molecular features, rooted in basic chemical concepts, making them understood by a relatively large community of researchers. One may question why not use molecular descriptors from the DFT calculations and construct regression models. Indeed, this may be an interesting but not a practical approach for a large set of molecules within a reasonable time. Besides, one should deal with convergence problems and the sensitivity of DFT calculations to initial molecular geometry (molecular mechanics and/or semiempirical pretreatment is recommended). The second advance of the QSPR models is that a rather detailed chemical verification is possible via correlation analysis, exploratory analysis, and interpretation of regression vectors. The third important advantage of the QSPR models is the geometrical verification supported by a large number of related experimental geometries from the Cambridge Structural Database. The following discussion shows how several crystal structures from the CSD are related to the QSPR models.

Intramolecular interactions in substituted benzaldehydes and to a lesser extent intermolecular interactions involving these molecules cause variations in the electron density at the carbonyl oxygen. The increase of this electron density is well quantified via the increase of the absolute value of Q_{Omul} (which is always negative), resulting in deshielding effects (decrease of δ_{exp}). These effects (see Table 5 for correlation coefficients) are followed by the lengthening of the C1═O (D_{CO}) and mean C_2-C_{ortho} ($\langle D_{ort}\rangle$) bonds, while the C_1-C_2 bond (D_{CC}) shortens. Correlations of D_{CO}, $\langle D_{ort}\rangle$, D_{CC} with δ_{exp} are high (−0.913, −0.797, and 0.907, respectively), and the same is observed for analogue correlations with Q_{Omul} (−0.976, −0.829, and 0.981, respectively) when **1**−**50** are analyzed. DFT geometries of **1**−**50** exhibit similar high correlations of the C1═O, mean C_2-C_{ortho}, and C_1-C_2 bond lengths with δ_{exp} (−0.925, −0.688, and 0.934, respectively) and Q_{Omul} (−0.967, −0.822, and 0.967, respectively). Consequently, there are high correlations between the bond lengths both at the PM3 and DFT level (absolute correlation coefficients are 0.84−0.95). Moderate to high correlations involving these bond lengths (Figure 10) and the oxygen charge are maintained both at the PM3 and DFT levels for **1**−**60** (absolute correlation coefficients are 0.69−0.97). These facts confirm the reliability of PM3 calculations and indicate the weakening of electron delocalization within GBF at low δ_{exp} (single bond−double bond alterations are more pronounced). More than 300 GBF fragments from the CSD, although including diverse intramolecular and crystal packing effects, still show moderate correlations between D_{CO} and D_{CC} and between D_{CC} and $\langle D_{ort}\rangle$ when compared to the PM3 and DFT results (Figure 10a and 10b). Hydrogen bonds between the carbonyl O and *o*-hydroxyl H are another interesting structural example about how hydrogen bonding is intimately related to electron delocalization in (hetero)aromatic systems, i.e., resonance-assisted hydrogen bonds.[26,37,38] Twenty-seven of the 60 studied benzaldehydes (**12**, **24**−**40**, **48**−**50**, **53**, **56**−**59**, where **58** is a pseudoexample, see Figure 2) and more than 50 fragments from the CSD contain such hydrogen bonds. When the hydrogen-acceptor distance O···H (D_{oh}) decreases, the bond $\langle D_{ort}\rangle$ increases. The respective correlation coefficients are moderate for the CSD and DFT data and low for PM3 data. Semiempirical methods are not highly efficient in reproducing hydrogen-bond geometries, but the PM3 data are correctly placed in the area defined by the CSD data (Figure 10c).

Intermolecular interactions are another type of structural verification of the QSPR models, involving benzaldehyde self-associations and interactions with other species in several crystal structures from the CSD. There are 17 structures of crystals of pure benzaldehyde studied in this work (**3**, **4**, **7**, **8**, **10**, **13**, **14**, **20**, **31**, **32**, **43**, **51**, **56**−**60**) and another 111 pure substituted benzaldehydes (without ring-containing substituents). All these structures, without any exception, clearly show that benzaldehydes establish $\pi\cdots\pi$ stacking interactions[30,31,38,59] between mutually parallel neighboring molecules as sandwiches or infinite stacks, similarly to polycyclic aromatic hydrocarbons whose aromaticity is known to be responsible for such crystal packing patterns. Crystals of pure benzaldehydes show that hydrogen-mediated interactions involving the aldehyde group and aromatic hydrogen atoms, as well as hydrophobic interactions, are always present in the crystals. Over two hundred crystal structures containing GBF and other fragments/species confirm these observations and, furthermore, point out the importance of moderately strong hydrogen bonds as well as other intermolecular interactions involving substituents of the benzene ring, for crystal lattice stabilization. Self-association of benzaldehyde fragments is frequently observed in these crystals. Table 10 is a natural continuation of the discussion about solvation effects on **12** in DFT calculations. The table shows that **12** established various types of moderately strong, weak, and very weak hydrogen bonds as well as other interactions involving its π-electron system (which includes both the benzene ring and the benzaldehyde group). On the other hand, three solvents of interest in this work (acetonitrile, chloroform, and 1,4-dioxane) were found forming crystals with benzaldehydes (GBF) via various weak interactions. In the

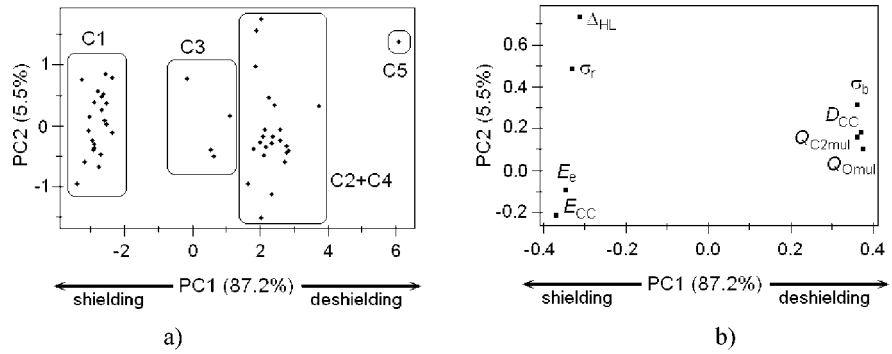

Figure 9. PCA plots for **1**−**50** from the QSAR data set: (a) scores plot denoting the HCA clusters along PC1 and (b) loadings plot.

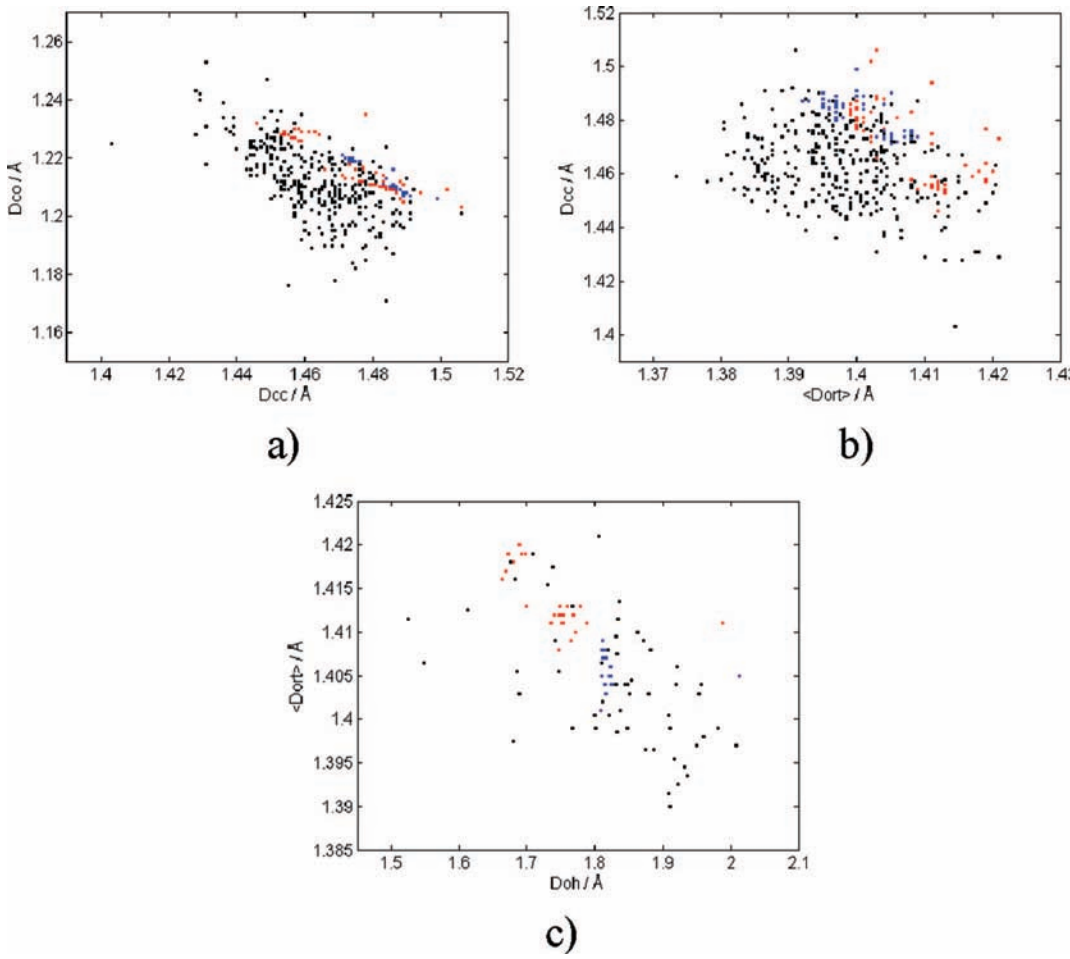

Figure 10. CSD or crystal structure data (black), PM3 data (blue), and DFT data (red) showing important intramolecular effects. (a) Linear correlation between the lengths of the $C_1{=}O$ (D_{CO}) and $C_1{-}C_2$ (D_{CC}) bonds: correlation coefficients are −0.625 (CSD 323 samples), −0.959 (PM3 60 samples), and −0.905 (DFT 60 samples). (b) Linear correlation between the lengths of the $C_1{-}C_2$ (D_{CC}) and mean $C_1{-}C_{ortho}$ (D_{ort}) bonds: correlation coefficients are −0.280 (CSD 323 samples), −0.752 (PM3 60 samples), and −0.683 (DFT 60 samples). (c) Linear correlation between the lengths of the mean $C_1{-}C_{ortho}$ (D_{ort}) bond and hydrogen-bond distance $C_1{=}O\cdots HO{-}$ (D_{oh}) where the OH group is an *ortho*-substituent: correlation coefficients are −0.554 (CSD 57 samples), −0.125 (PM3 27 samples), and −0.594 (DFT 27 samples).

absence of moderately strong hydrogen bonds, other weak interactions appear such as chlorine-containing hydrogen bonds and orbital interactions involving the benzaldehyde π system and carbonyl oxygen's lone pairs. It can be concluded that benzaldehyde heteroaromaticity and hydrogen-bonding features are essential for self-associations and interactions with different species in the crystalline state and probably for the liquid state and solutions. Self-associations may occur in solutions, especially at high concentrations and in nonpolar solvents. A good logical parallelism of QSPR with the CSD-based structural observations is achieved by molecular descriptors used in the regression models (Tables 2 and 5). Three descriptors are typical geometrical measures of (hetero)aromaticity[27,37,38] (D_{CC}, σ_r, and σ_b), while other electronic descriptors are indirect indices of local (Q_{Omul}, Q_{C2mul}, E_e, and E_{CC}) and overall (Δ_{HL}) electron delocalization in π systems. It is known[30,31,59] that electronic descriptors are important quantitative determinants of $\pi\cdots\pi$ stacking geometry in crystals of heteroaromatics. Furthermore, topological, electronic, and other molecular descriptors for (hetero)aromatic fragments are quantitatively related to biological activities[60–64] and physicochemical properties[13–15] of diverse classes of organic compounds.

Conclusions

Chemometric, QSPR, and structural studies applied to fifty and ten substituted benzaldehydes with known and unknown carbonyl ^{17}O shifts, respectively, lead to the following conclusions. (1) Parsimonious QSPR models employing PLS and PCR regression were comparable with the literature empirical model LL and the DFT model. The regression models were validated externally and internally and compared with the LL and DFT via several statistical parameters. Chemical validity of the models was verified through correlation and exploratory analyses, interpretation of the regression vectors, and additional qualitative and quantitative structural analyses supported by the Cambridge Structural Database. (2) These validations, simple and fast calculations and good statistics for prediction of chemical shifts, are the reason to recommend the QSPR models as superior to the LL and DFT models for practical purposes.

Acknowledgment. FAPESP is gratefully acknowledged for financial support.

References and Notes

(1) Li, L.-D.; Li, L.-S. *Magn. Reson. Chem.* **2004**, *42*, 977–982.
(2) Li, L.-D.; Li, L.-S. *Chin. J. Magn. Reson.* **2002**, *19*, 115–123.
(3) Li, L.-D.; Li, L.-S. *Chin. J. Magn. Reson.* **2002**, *19*, 289–292.
(4) Li, L.-D.; Li, L.-S. *Chin. J. Magn. Reson.* **2002**, *19*, 385–390.
(5) Benzi, C.; Crescenzi, O.; Pavone, M.; Barone, V. *Magn. Reson. Chem.* **2004**, *42*, S57–S67.
(6) Kowalewski, D. G.; Kowalewski, V. J.; Contreras, R. H.; Díez, E.; Casanueva, J.; San Fabián, J.; Esteban, A. L.; Galanche, M. P. *J. Magn. Reson.* **2001**, *148*, 1–10.
(7) Dahn, H.; Carrupt, P.-A. *Magn. Reson. Chem.* **1997**, *35*, 577–588.
(8) Wolinski, K.; Hilton, J. F.; Pulay, P. *J. Am. Chem. Soc.* **1990**, *112*, 8251–8260.
(9) Ferreira, M. M. C. *Chemosphere* **2001**, *44*, 125–146.
(10) Nigovic, B.; Antolic, S.; Kojic-Prodic, B.; Kiralj, R.; Magnus, V.; Salopek-Sondi, B. *Acta Crystallogr., Sect. B* **2000**, *56*, 94–111.
(11) Cirino, L. R.; Ferreira, M. M. C. *Quim. Nova* **2003**, *26*, 312–318.
(12) Ribeiro, F. A. L.; Ferreira, M. M. C. *THEOCHEM-J. Mol. Struct.* **2003**, *663*, 109–126.
(13) Kiralj, R.; Ferreira, M. M. C. *Book of Abstracts of the IX Brazilian Symposium of Theoretical Chemistry*, Caxambu MG, Brazil, Nov 16−19, 1997; p 123.
(14) Ferreira, M. M. C.; Kiralj, R. *Book of Abstracts of the XI Brazilian Symposium of Theoretical Chemistry*, Caxambu MG, Brazil, Nov 18−21, 2001; p 227.
(15) Kiralj, R.; Ferreira, M. M. C. *Book of Abstracts of the XI Brazilian Symposium of Theoretical Chemistry*, Caxambu MG, Brazil, Nov 18−21, 2001; p 301.
(16) Ferreira, M. M. C.; Kiralj, R. *Book of Abstracts of the 25th Annual Meeting of the Brazilian Chemical Society*, Poços de Caldas - MG, Brazil, May 20−23, 2002; p QT-040.
(17) Golbraikh, A.; Tropsha, A. *J. Mol. Graph. Mod.* **2002**, *20*, 269–276.
(18) Tropsha, A.; Gramatica, P.; Gombar, V. K. *QSAR Comb. Sci.* **2003**, *22*, 69–77.
(19) Zhang, S.; Wei, L.; Bastow, K.; Zheng, W.; Brossi, A.; Lee, K.-H.; Tropsha, A. *J. Comput.-Aided Mol. Des.* **2007**, *21*, 97–112.
(20) Rücker, C.; Rücker, G.; Meringer, M. *J. Chem. Inf. Model.* **2007**, *47*, 2345–2357.
(21) Gramatica, P. *QSAR Comb. Sci.* **2007**, *26*, 694–701.
(22) Gramatica, P.; Giani, E.; Papa, E. J. *Mol. Graph. Mod.* **2007**, *25*, 755–766.
(23) Tasic, L.; Rittner, R.; Ferreira, M. M. C. *An. Ressonancia Magn. Nucl.* **1999**, *6*, 107–114.
(24) Kiralj, R.; Takahata, Y. *Struct. Chem.* **2005**, *17*, 525–538.
(25) Kiralj, R.; Ferreira, M. M. C. J. *Chem. Inf. Comput. Sci.* **2002**, *42*, 508–523.
(26) Kiralj, R.; Ferreira, M. M. C. *J. Chem. Inf. Comput. Sci.* **2002**, *43*, 787–809.
(27) Kiralj, R.; Ferreira, M. M. C. *Hem. Pregled* **2003**, *44*, 82–88.
(28) Kiralj, R.; Ferreira, M. M. C.; Donate, P. M.; da Silva, R.; Albuquerque, S. *J. Phys. Chem. A* **2007**, *111*, 6316–6333.
(29) Kiralj, R.; Ferreira, M. M. C. *Book of Abstracts of the XIV Annual Meeting of the Brazilian Crystallographic Society*, São Carlos - SP, Brazil, Nov 20−22, 1997; p 6.
(30) Kiralj, R. Ph.D. thesis, University of Zagreb, Zagreb, Croatia, 1999.
(31) Kiralj, R.; Ferreira, M. M. C. *Book of Abstracts of the XII Brazilian Symposium of Theoretical Chemistry*, Caxambu - MG, Brazil, Nov 23−26, 2003; p 151.
(32) Martens, H.; Naes, T. *Multivariate Calibration*, 2nd ed.; Wiley: New York, 1989.
(33) Beebe, K. R.; Pell, R.; Seasholtz, M. B. *Chemometrics: a practical guide*; Wiley: New York, 1998.
(34) Ferreira, M. M. C. *J. Braz. Chem. Soc.* **2002**, *13*, 742–753.
(35) Ferreira, M. M. C.; Antunes, A. M.; Melo, M. S.; Volpe, P. L. O. *Quim. Nova* **1999**, *22*, 724–731.
(36) Boykin, D. W.; Baumstrak, A. L.; Besson, M. *J. Org. Chem.* **1991**, *56*, 1969–1971.
(37) Bertolasi, V.; Gilli, P.; Ferreti, V.; Gilli, G. *Acta Crystallogr., Sect. B* **1995**, *51*, 1004–1015.
(38) Kiralj, R.; Ferreira, M. M. C. *Int. J. Quantum Chem.* **2003**, *95*, 237–251.
(39) Cambridge Structural Database 5.29, Release Nov 2007; Cambridge Structural Data Centre, University of Cambridge: Cambridge, U.K., 2007.
(40) Allen, F. H. *Acta Crystallogr., Sect. B* **2002**, *58*, 380–388.
(41) *ConQuest 1.10*; Cambridge Structural Data Centre, University of Cambridge: Cambridge, U.K., 2007.
(42) Bruno, I. J.; Cole, J. C.; Edgington, P. R.; Kessler, M.; Macrae, C. F.; McCabe, P.; Pearson, J.; Taylor, R. *Acta Crystallogr., Sect. B* **2002**, *58*, 389–397.
(43) *Vista 2.1*; Cambridge Structural Data Centre, University of Cambridge: Cambridge, U.K., 2007.
(44) *Mercury CSD 2.0*; Cambridge Structural Data Centre, University of Cambridge: Cambridge, U.K., 2007.
(45) Macrae, C. F.; Edgington, P. R.; McCabe, P.; Pidcock, E.; Shields, G. P.; Taylor, R.; Towler, M.; van de Streek, J. *J. Appl. Crystallogr.* **2006**, *39*, 453–457.
(46) *Chem3D Ultra 6.0*; CambridgeSoft.Com: Cambridge, MA, 2000.
(47) Allinger, N. L. *J. Am. Chem. Soc.* **1977**, *99*, 8127–8134.
(48) *Titan 1.0.8*; Wavefunction, Inc.: Irvine, CA, 2001.
(49) Lobanov, V. S. *MOPAC 6.0 for Microsoft Windows*; University of Florida: Tallahassee, FL, 1996.
(50) *Matlab 5.2*, Version 6.1.0.450; MathWorks, Inc.: Natick, MA, 2001.
(51) Frisch, M. J.; Trucks, G. W.; Schlegel, H. B.; Scuseria, G. E.; Robb, M. A.; Cheeseman, J. R.; Zakrzewski, V. G.; Montgomery, J. A.; Stratmann, R. E.; Burant, J. C.; Dapprich, S.; Millam, J. M.; Daniels, A. D.; Kudin, K. N.; Strain, M. C.; Farkas, O.; Tomasi, J.; Barone, V.; Cossi, M.; Cammi, R.; Mennucci, B.; Pomelli, C.; Adamo, C.; Clifford, S.; Ochterski, J.; Petersson, G. A.; Ayala, P. Y.; Cui, Q.; Morokuma, K.; Malick, D. K.; Rabuck, A. D.; Raghavachari, K.; Foresman, J. B.; Cioslowski, J.; Ortiz, J. V.; Stefanov, B. B.; Liu, G.; Liashenko, A.; Piskorz, P.; Komaromi, I.; Gomperts, R.; Martin, R. L.; Fox, D. J.; Keith, T.; Al-Laham, M. A.; Peng, C. Y.; Nanayakkara, A.; Gonzalez, C.; Challacombe, M.; Gill, P. M. W.; Johnson, B. G.; Chen, W.; Wong, M. W.; Andres, J. L.; Head-Gordon, M.; Replogle E. S.; Pople, J. A. *Gaussian 98W*, version 5.4; Revision A.1; Gaussian, Inc.: Pittsburgh, PA, 1998.
(52) Barone, V.; Cossi, M *J. Phys. Chem. A* **1998**, *102*, 1995–2001.
(53) Ferreira, M. M. C.; Montanari, C. A.; Gáudio, A. G. *Quim. Nova* **2002**, *25*, 439–448.
(54) Wold, S.; Eriksson, L. In *Chemometric Methods in Molecular Design*; van de Waterbeemd, H., Ed.; VCH: Weinheim, 1995; pp 309−318..
(55) *Pirouette 3.02*; Infometrix, Inc.: Woodinville, WA, 2001.
(56) *OriginPro 7.0*; OriginLab Corp.: Northampton, MA, 2002.
(57) Kiralj, R.; Kojic-Prodic, B.; Zinic, M.; Alihodzic, S.; Trinajstic, N. *Acta Crystallogr., Sect. B* **1996**, *52*, 823–837.
(58) Kiralj, R.; Kojic-Prodic, B.; Nikolic, S.; Trinajstic, N. *J. Mol. Struct.-THEOCHEM* **1998**, *427*, 25–37.
(59) Kiralj, R.; Kojic-Prodic, B.; Piantanida, I.; Zinic, M. *Acta Crystallogr., Sect. B* **1999**, *55*, 55–69.
(60) Kiralj, R.; Ferreira, M. M. C.; Takahata, Y. *QSAR Comb. Sci.* **2003**, *22*, 430–448.
(61) Kiralj, R.; Ferreira, M. M. C. *J. Mol. Graph. Mod.* **2003**, *22*, 435–448.
(62) Kiralj, R.; Ferreira, M. M. C. *J. Mol. Graph. Mod.* **2003**, *22*, 499–515.
(63) Kiralj, R.; Ferreira, M. M. C. *QSAR Comb. Sci.* **2008**, *27*, 289–301.
(64) Kiralj, R.; Ferreira, M. M. C. *QSAR Comb. Sci.* **2008**, *27*, 314–329.

JP800864S