Quantitative Structure-Property Relationship (QSPR) Study of ^{17}O Carbonyl Chemical Shifts in Substituted Benzaldehydes

Rudolf Kiralj and Márcia Miguel Castro Ferreira. *rudolf@iqm.unicamp.br*, *marcia@iqm.unicamp.br, http://lqta.iqm.unicamp.br*
Laboratório de Quimiometria Teórica e Aplicada (LQTA), Instituto de Química, Universidade Estadual de Campinas, Campinas SP, 13083-970, Brazil

THE OBJECTIVES OF THIS WORK

1) To develop a fast and simple QSPR methodology for prediction of ^{17}O carbonyl chemical shifts in substituted benzaldehydes, comparable to the empirical model of Li&Li (LL)

2) To show that this methodology is based on well understandable chemical concepts and that the QSPR models can be validated unlike the LL model

3) To use the QSPR models for general substituted benzaldehydes, in advance of the LL model

THE STORY

1 **The empirical Li&Li (LL) model[1]:**

Empirical equation for calculation of ^{17}O NMR chemical shifts in 50 benzaldehydes (Figure 1), based on contributions Δ of individual *o*-, *o*'-, *m*-, *m*'- and *p*-positioned substituents and correction *C* for polar solvents:

$$\delta_{LL}\ /\ \mathrm{ppm} = 564.0 + \Delta o + \Delta o' + \Delta m + \Delta m' + \Delta p + C$$

$$O = \Delta o + \Delta o' \qquad M = \Delta m + \Delta m' \qquad P = \Delta p$$

$$\Delta o = \Delta o' = \Delta m = \Delta m' = \Delta p = 0 \text{ for H at positions } o, o', m, m' \text{ and } p$$

$$C = -14.7 \text{ ppm for } \mathbf{24}, \mathbf{34\text{-}40}, \mathbf{47\text{-}50}, \text{ otherwise } C = 0.0 \text{ ppm}$$

[1] Li LD, Li LS (2004) Magn Reson Chem 42:977

New QSPR methodology:

1)Substituted benzaldehydes (training set: 50, Fig. 1; prediction set: 10, Fig. 2) were modeled and optimized at semi-empirical PM3 level.
2) Various global and local molecular descriptors of electronic and steric nature were generated.
3) Variable selection was performed for PLS, PCR and MLR models (autoscalled data) which were validated by leave-one out crossvalidation and additionally externally validated.
4) The models were compared with the LL model and used to predict ^{17}O carbonyl shifts in the prediction set.
5) Additional exploratory analysis (PCA and HCA) and data mining in the Cambridge Structural Database (CSD) were performed to rationalize the relationships among the samples and variables.

QSPR models versus LL model:

-the same prediction power for the training set as of the LL model
-better applicable for more general data sets than the LL model
-can be validated and the LL cannot
-all descriptors with clear chemical background, contrary to the LL model
-fast and simple methodology, does not need substituent constants

2

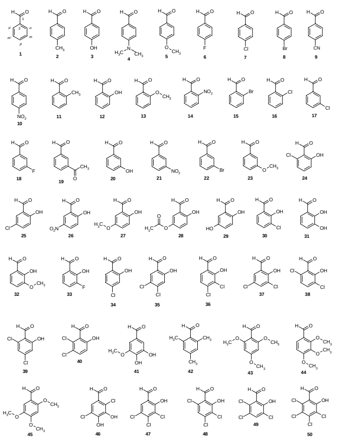


Figure 1. Substituted benzaldehydes 1-50 (training set).

3

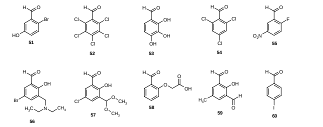


Figure 2. Substituted benzaldehydes 51-60 (prediction set).

4

Selected molecular descriptors:
1) E_{CC} – C_1-C_2 nuclear-nuclear repulsion energy
2) Q_{Oesp} – electrostatic potential-based partial atomic charge of the carbonyl oxygen O
3) σ_d – standard deviation of the six C-C bond lengths in the benzene fragment
4) d_{CC} – C_1-C_2 bond length
5) Q_{C2mul} – Mulliken partial atomic charge of C_2

Table 1. Selected molecular descriptors for benzaldehydes 1-50

No.	E_{CC}/eV	Q_{Oesp}	σ_d/Å	d_{CC}/Å	Q_{C2mul}	δ_{exp}/ppm
1	122.756	-0.483	0.003	1.484	-0.201	563.2
2	122.671	-0.487	0.004	1.483	-0.209	561.4
3	122.749	-0.501	0.007	1.481	-0.244	526.9
4	122.819	-0.500	0.008	1.480	-0.240	532.8
5	122.790	-0.504	0.006	1.481	-0.241	545.7
6	122.617	-0.494	0.005	1.485	-0.215	568.9
7	122.591	-0.480	0.003	1.485	-0.204	570.1
8	122.540	-0.473	0.007	1.486	-0.196	570.3
9	122.431	-0.467	0.003	1.487	-0.182	593.6
10	122.302	-0.457	0.004	1.490	-0.143	590.1
11	122.526	-0.498	0.005	1.486	-0.198	575.0
12	123.365	-0.570	0.011	1.471	-0.324	505.8
13	122.324	-0.495	0.006	1.489	-0.217	555.0
14	121.716	-0.444	0.005	1.499	-0.093	576.0
15	122.470	-0.454	0.006	1.487	-0.176	573.0
16	122.499	-0.480	0.003	1.486	-0.198	573.0
17	122.491	-0.480	0.003	1.486	-0.188	569.3
18	122.457	-0.481	0.004	1.487	-0.175	570.8
19	122.541	-0.482	0.003	1.485	-0.216	568.4
20	122.448	-0.493	0.005	1.486	-0.162	555.2
21	122.396	-0.470	0.004	1.488	-0.223	574.5
22	122.597	-0.474	0.007	1.485	-0.195	566.0
23	122.508	-0.490	0.006	1.486	-0.163	562.3
24	123.200	-0.552	0.011	1.473	-0.323	507.0
25	123.258	-0.564	0.011	1.473	-0.314	516.2
26	123.104	-0.550	0.013	1.476	-0.351	522.8
27	123.179	-0.567	0.009	1.474	-0.286	512.1
28	123.250	-0.582	0.011	1.473	-0.310	514.7
29	123.241	-0.574	0.012	1.473	-0.284	511.8
30	123.263	-0.565	0.010	1.473	-0.313	509.0
31	123.193	-0.572	0.011	1.474	-0.279	510.0
32	123.283	-0.577	0.015	1.473	-0.302	513.9
33	123.166	-0.561	0.011	1.475	-0.303	518.2
34	123.302	-0.559	0.011	1.472	-0.327	507.0
35	123.219	-0.554	0.011	1.474	-0.317	515.0
36	123.230	-0.567	0.010	1.473	-0.318	509.0
37	123.138	-0.569	0.010	1.475	-0.303	520.0
38	123.116	-0.553	0.010	1.475	-0.313	512.0
39	123.188	-0.554	0.011	1.474	-0.328	507.0
40	123.142	-0.550	0.012	1.475	-0.315	513.0
41	122.517	-0.508	0.007	1.485	-0.186	550.0
42	122.507	-0.502	0.005	1.486	-0.199	585.0
43	122.264	-0.514	0.005	1.490	-0.275	565.0
44	122.477	-0.518	0.008	1.486	-0.231	545.0
45	122.356	-0.509	0.011	1.489	-0.200	538.0
46	122.471	-0.486	0.008	1.487	-0.188	560.0
47	123.155	-0.558	0.010	1.475	-0.307	518.0
48	123.148	-0.557	0.010	1.475	-0.318	505.0
49	123.024	-0.548	0.010	1.476	-0.304	517.0
50	123.092	-0.538	0.010	1.476	-0.309	513.0

5

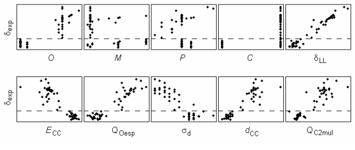


Figure 3. Correlation of experimental ^{17}O NMR shifts with independent variables from the LL model (upper plots) and calculated in this work (lower plots).
Samples with (lower shifts) and without (higher shifts) internal -HC=O … HO- hydrogen bond are separated by a dashed horizontal line in all plots.

6

Table 2. Correlation matrix including experimental chemical shifts and selected molecular descriptors

	E_{CC}	Q_{Oesp}	σ_d	d_{CC}	Q_{C2mul}	δ_{exp}
E_{CC}	1	-0.911	0.791	-0.997	-0.912	-0.892
Q_{Oesp}		1	-0.867	0.920	0.912	0.927
σ_d			1	-0.797	-0.803	-0.891
d_{CC}				1	0.916	0.907
Q_{C2mul}					1	0.892
δ_{exp}						1

QSPR MODELS

7

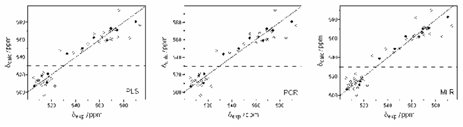


Figure 4. Experimental against calculated ^{17}O NMR shifts as obtained from PLS (left), PCR (middle) and MLR (right) QSPR model. Solid squares account for the samples from the external validation set. The dashed line separates samples with internal -HC=O … HO- hydrogen bond from those without it.

8

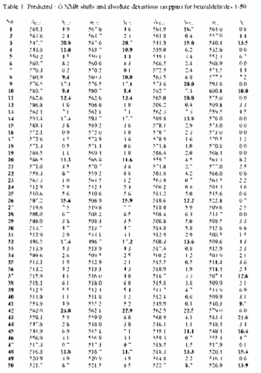
Table 3. Regression models (PLS, PCR, MLR) and the Li-Li empirical model (LL) with basic statistics and regression coefficients

Model	PCs(%)[a]	SEV[b]	SEP[c]	Q[d]	R[e]	<Δ>[f]	N_Δ[g]	E_{CC}[h]	Q_{Oesp}[h]	σ_d[h]	d_{CC}[h]	Q_{C2mul}[h]
PLS	2(96.1)	9.4	8.8	0.942	0.953	6.6	12	-0.106	0.239	-0.359	0.150	0.154
PCR	2(96.2)	9.3	8.8	0.944	0.953	6.6	12	-0.127	0.225	-0.359	0.131	0.165
MLR	5(100)	9.8	8.3	0.939	0.961	5.5	11	1.395	0.173	-0.360	1.731	0.131
LL	-	-		-	0.975	4.3	8	-	-	-	-	-

[a]Number of used principal component with % of the variance. [b]Standard error of validation. [c]Standard error of prediction. [d]Correlation coefficient of validation. [e]Correlation coefficient of prediction. [f]Average absolute deviation. [g]Number of samples with average absolute deviation >10%. [h]Autoscalled regression coefficients for selected variables E_{CC}, Q_{Oesp}, σ_d, d_{CC}, and Q_{C2mul}.

9

Table 4. Predicted ^{17}O NMR shifts and absolute deviations (in ppm) for benzaldehydes 1-50



10

External validation of the regression models

Table 5. External validation of the regression models with basic statistics

No.	δ_{exp}/ppm	δ_{PLS}/ppm	%Δ_{PLS}/ppm	δ_{PCR}/ppm	%Δ_{PCR}/ppm	δ_{MLR}/ppm	%Δ_{MLR}/ppm
4	532.8	543.9	12.5	543.9	12.5	538.9	6.9
5	545.7	550.4	5.3	550.3	5.2	549.0	3.7
10	590.1	580.3	11.1	580.5	10.8	581.9	9.3
12	505.8	507.2	1.6	507.3	1.7	506.1	0.3
13	555.0	562.2	8.1	561.9	7.8	562.8	8.8
17	569.3	572.9	4.1	572.9	4.1	570.6	1.5
21	574.5	571.2	3.7	570.7	4.3	571.6	3.3
22	566.0	559.1	7.8	559.2	7.7	560.8	5.9
25	516.2	511.0	5.9	511.0	5.9	511.0	5.9
49	517.0	521.3	4.9	521.2	4.7	515.6	1.6
Model	PCs(%)[a]	SEV[b]	SEP[c]	Q[d]	R[e]	<Δ>[f]	N_Δ[g]
PLS	2(96.1)	10.2	9.4	0.934	0.948	6.8	11
PCR	2(96.2)	10.0	9.4	0.935	0.947	6.9	13
MLR	5(100)	11.0	9.1	0.924	0.955	5.9	11

[a]Number of used principal component with % of the variance. [b]Standard error of validation. [c]Standard error of prediction. [d]Correlation coefficient of validation. [e]Correlation coefficient of prediction. [f]Average absolute deviation. [g]Number of samples with average absolute deviation >10%. [h]Autoscalled regression coefficients for selected variables E_{CC}, Q_{Oesp}, σ_d, d_{CC}, and Q_{C2mul}

11

QSPR and LL predictions

Table 6. Molecular descriptors and predicted ^{17}O NMR shifts for benzaldehydes 51-60

No.	E_{CC}/eV	Q_{Oesp}	σ_d/Å	d_{CC}/Å	Q_{C2mul}	δ_{PLS}/ppm	δ_{PCR}/ppm	δ_{MLR}/ppm	δ_{LL}/ppm
51	122.459	-0.443	0.008	1.487	-0.132	568.1	568.0	566.6	572.4
52	122.196	-0.428	0.002	1.491	-0.176	591.0	590.7	587.0	605.5
53	123.426	-0.586	0.011	1.471	-0.324	503.7	503.7	510.5	484.9
54	122.358	-0.444	0.004	1.489	-0.197	578.1	577.6	579.5	590.0
55	122.323	-0.443	0.007	1.489	-0.255	565.0	561.4	563.2	-
56	123.262	-0.549	0.016	1.473	-0.326	496.6	496.4	497.0	-
57	123.212	-0.570	0.013	1.474	-0.332	503.1	503.0	505.7	-
58	122.615	-0.540	0.006	1.484	-0.226	548.2	548.5	552.6	-
59	123.276	-0.565	0.010	1.473	-0.336	512.0	511.8	514.8	-
60	122.529	-0.469	0.007	1.486	-0.199	561.1	560.8	562.5	-

12

Proposed QSPR models:

PLS:

$$\delta = 683.413 - 7.762\,E_{CC} + 168.153\,Q_{Oesp} - 3188.046\,\sigma_d + 635.022\,d_{CC} + 69.290\,Q_{C2mul}$$

PCR:

$$\delta = 986.390 - 9.298\,E_{CC} + 158.568\,Q_{Oesp} - 3189.821\,\sigma_d + 555.309\,d_{CC} + 74.429\,Q_{C2mul}$$

MLR:

$$\delta = -22814.206 + 102.528\,E_{CC} + 121.992\,Q_{Oesp} - 3194.259\,\sigma_d + 7338.557\,d_{CC} + 58.967\,Q_{C2mul}$$

EXPLORATORY ANALYSIS & DATA MINING

13

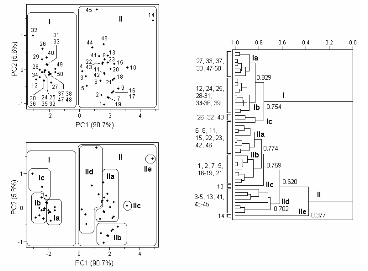


Figure 5. Scores plot (left) with two clusters (left top) and subclusters (bottom) and HCA dendogram for samples (right). Clusters I and II contain samples with and withouth the internal hydrogen bond, respectively.
PC1 – related to cumulative electron withdrawal/donation effects felt by the carbonyl oxygen
PC2 – related to variations in the benzaldehyde heteroaromatic character

14

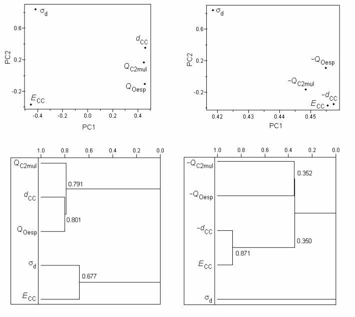


Figure 6. Loading plots (top) and HCA dendogram for variables (bottom) for the training set variables (left) and modified training set (with negative variables $-d_{CC}$, $-Q_{Oesp}$ and $-Q_{C2mul}$, right).

15

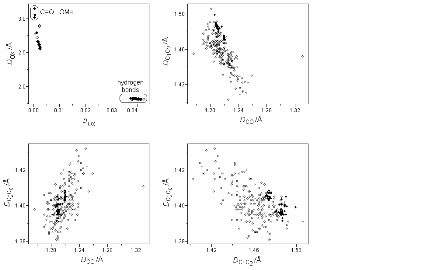


Figure 7. Relationships between structural variables demonstrating electron delocalization in substituted benzaldehydes. Left top: Bond length-bond order relationship for interaction of the carbonyl oxygen with the closest *o*-hydrogen or closest atom from the *o*-substituent. Relationships between bond lengths C=O and C_1-C_2 (right top), C=O and mean C_2-C_o (C_o – *o*-carbon atoms, left bottom), and C_1-C_2 and mean C_2-C_o (right bottom) with experimental data (from the CSD database, white squares) and calculated data (for 1-60 in this work, solid squares). These findings agree with QSPR models and exploratory analysis and can explain intramolecular interactions affecting ^{17}O carbonyl shifts.

ACKNOWLEDGEMENT: FAPESP