Quantitative Structure-Property Relationship (QSPR) Study of $^{17}$O Carbonyl Chemical Shifts in Substituted Benzaldehydes
Rudolf Kiradj 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
The empirical Li&Li (LL) model:

1. Graphical expression for calculation of $^{17}$O NMR chemical shifts in 68 benzaldehydes (Figure 1)
2. Based on contributions of a) individual X, Y, Z, and W and pre-defined substituents
3. Correction C for pair solvents:

$$\Delta_{ax} \text{ppm} = 164 \times (\alpha + \beta + \gamma + \delta + \varepsilon) + 11 \times \alpha + 3 \times \alpha + 6 \times \beta + 10 \times \gamma + 12 \times \delta + 14 \times \varepsilon + C$$

$$\alpha = \text{inductive effect of } \text{X, Y, Z, and W}$$
$$\beta = \text{ mesomeric effect of } \text{X, Y, Z, and W}$$
$$\gamma = \text{ steric effect of } \text{X, Y, Z, and W}$$
$$\delta = \text{ electronegativity of } \text{X, Y, Z, and W}$$
$$\varepsilon = \text{ electronic polarizability of } \text{X, Y, Z, and W}$$
$$C = \text{ correction for pair solvents}$$

New QSPR methodologies:
1) Substituted benzaldehydes (training set 48, Fig. 1) prediction set 20, 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 (unscrambled data) which were validated by leave-one out cross-validation and additionally externally validated
4) The models were compared with the LL model
5) Additional exploratory analysis (PCA and HCA) and data mining is in the Cambridge Structural Database (CSD) were performed to rationalize the relationships among the samples and variables.

QSPR models versus LL model:

1) The same prediction power for the training set of the LL model
2) Better applicability for more general data sets than the LL model
3) All descriptors with clear chemical background contrary to the LL model:
- fast and simple methodology does not use subterminal constants

QSPR MODELS

1. Table 1: Regression models (PLS, PCR, MLR) and the LL empirical model (LL) with molecular descriptors for the training set.

<table>
<thead>
<tr>
<th>Model</th>
<th>R^2</th>
<th>SEP</th>
<th>RSE</th>
<th>RSE/SEP</th>
<th>RSE/\Delta</th>
<th>RSE/\alpha</th>
<th>RSE/\beta</th>
<th>RSE/\gamma</th>
<th>RSE/\delta</th>
<th>RSE/\varepsilon</th>
<th>RSE/C</th>
<th>RSE/\sigma</th>
</tr>
</thead>
<tbody>
<tr>
<td>PLS</td>
<td>0.54</td>
<td>0.19</td>
<td>0.16</td>
<td>1.00</td>
<td>1.00</td>
<td>1.00</td>
<td>1.00</td>
<td>1.00</td>
<td>1.00</td>
<td>1.00</td>
<td>1.00</td>
<td>1.00</td>
</tr>
<tr>
<td>PCR</td>
<td>0.50</td>
<td>0.18</td>
<td>0.16</td>
<td>1.00</td>
<td>1.00</td>
<td>1.00</td>
<td>1.00</td>
<td>1.00</td>
<td>1.00</td>
<td>1.00</td>
<td>1.00</td>
<td>1.00</td>
</tr>
<tr>
<td>M  L R</td>
<td>0.50</td>
<td>0.19</td>
<td>0.16</td>
<td>1.00</td>
<td>1.00</td>
<td>1.00</td>
<td>1.00</td>
<td>1.00</td>
<td>1.00</td>
<td>1.00</td>
<td>1.00</td>
<td>1.00</td>
</tr>
</tbody>
</table>

2. Proposed QSPR models:

$$\Delta_{ax} \text{ppm} = 164 \times (\alpha + \beta + \gamma + \delta + \varepsilon) + 11 \times \alpha + 3 \times \alpha + 6 \times \beta + 10 \times \gamma + 12 \times \delta + 14 \times \varepsilon + 15 \times C + 16 \times \sigma$$

3. Figure 1: Substituted benzaldehydes 1-58 training set

4. Figure 2: Substituted benzaldehydes 51-68 prediction set

EXPLORATORY ANALYSIS & DATA MINING

1. Figure 3: Scores plot (left) with two clusters (left top) and subclusters (bottom) and HCA dendrogram for samples (right). Clusters I and II contain samples with and without the internal hydrogen bond, respectively
2. Figure 4: Loading plots (top) and HCA dendrogram for variables (bottom) for the training set variables (left) and normalized training set (with negative variables $\Delta_{ax}$, $\Delta_{ay}$ and $\Delta_{az}$, right)

ACKNOWLEDGEMENT: FAPESP