Estimation of Melting Points of Brominated and Chlorinated Organic Pollutants using QSAR Techniques

By: Marquita Watkins
Persistent Organic Pollutants

- Do not undergo photolytic, biological, and chemical degradation.
  - low water solubility
  - high lipid solubility
  - semi-volatility
  - high molecular masses

- Found in pesticides
  - Transport by air and water

- Found at extremely low levels in 20% of US Food Supply
  - Continual exposure toxicity and effects are unknown
Previous Research

- Reserved to small sets (10 to 100 compounds) of hydrocarbons, substituted aromatics, aldehydes, amines, and ketones utilizing other descriptor types, regressions, and analysis techniques.

- Larger data sets were only of approximately 400 compounds.
To estimate the melting points of 1436 chloro- and bromo-analogues of dibenzo-p-dioxins, dibenzofurans, biphenyls, naphthalenes, diphenyl ethers, and benzenes by utilizing quantitative structure—property relationship (QSAR) techniques.
Chosen Structure Families

dibenzo-p-dioxins
dibenzofurans
biphenyls
naphthalenes
diphenyl ethers
benzenes
Melting Point

- Specifies transition temperature
- Essential in biochemical and environmental studies
- Solubility estimation
- Difficult QSAR endpoint to predict
QSAR Significance

- Inexpensive
- Time-efficient
- Safe
- Enhances the efficiency of drug design
Basic concept of QSAR modeling

Endpoint (experimentally measured)

QSAR model

\[ y = f(X) \]
(eg. \( y = b_0 + b_1x_1 + b_2x_2 \))

Structural descriptors

\[ X \]

- Linear Regression (LR)
- Multiple Linear Regression (MLR)
- Partial Least Squares (PLS)
- Artificial Neural Networks (ANN)
- Dipole moment
- Polarizability
- HOMO, LUMO
- Topological indexes
- Number of specific atoms/groups
- ...
Software Employed

- Structures were built and optimized using HyperChem 5.0
- Hierarchal clustering in ChemAxon’s MCS Library Software.
- Descriptor Generation Software
  - DRAGON5
- PLS Toolbox
  - Automated Descriptor Selection
    - Genetic Algorithm
  - Model Building
Hierarchical Cluster Results

Minimal MCS Size = 25
MCS Mode = Very Fast
Matching Parameters = Atom Type, Bond Type, Charge
Clustering Options = Keep Rings
Required Cluster Count = 15
Maximum Level Count = 15
# Cluster Analysis

<table>
<thead>
<tr>
<th>Cluster #</th>
<th>Calibration Set</th>
<th>Validation Set</th>
<th>Prediction Set</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>$R^2$ Cal: 0.89</td>
<td>$R^2$ CV: 0.88</td>
<td>$R^2$ Pred: 0.80</td>
</tr>
<tr>
<td></td>
<td>Bias: 0</td>
<td>CV Bias: 0.004</td>
<td>Pred Bias: -1.07</td>
</tr>
<tr>
<td></td>
<td>RMSEC: 8.04</td>
<td>RMSECV: 8.43</td>
<td>RMSEP: 13.24</td>
</tr>
<tr>
<td>3</td>
<td>$R^2$ Cal: 0.76</td>
<td>$R^2$ CV: 0.72</td>
<td>$R^2$ Pred: 0.66</td>
</tr>
<tr>
<td></td>
<td>Bias: -4.26e-14</td>
<td>CV Bias: -0.20</td>
<td>Pred Bias: 5.46</td>
</tr>
<tr>
<td></td>
<td>RMSEC: 21.66</td>
<td>RMSECV: 23.64</td>
<td>RMSEP: 23.72</td>
</tr>
<tr>
<td>4</td>
<td>$R^2$ Cal: 0.89</td>
<td>$R^2$ CV: 0.86</td>
<td>$R^2$ Pred: 0.91</td>
</tr>
<tr>
<td></td>
<td>Bias: -2.84e-14</td>
<td>CV Bias: -0.13</td>
<td>Pred Bias: -5.29</td>
</tr>
<tr>
<td></td>
<td>RMSEC: 17.32</td>
<td>RMSECV: 19.30</td>
<td>RMSEP: 15.71</td>
</tr>
<tr>
<td>7</td>
<td>$R^2$ Cal: 0.94</td>
<td>$R^2$ CV: 0.80</td>
<td>$R^2$ Pred: 0.63</td>
</tr>
<tr>
<td></td>
<td>Bias: -2.84e-14</td>
<td>CV Bias: -5.24</td>
<td>Pred Bias: -3.82</td>
</tr>
<tr>
<td></td>
<td>RMSEC: 18.56</td>
<td>RMSECV: 34.79</td>
<td>RMSEP: 26.26</td>
</tr>
</tbody>
</table>
## Cluster Analysis

<table>
<thead>
<tr>
<th>Cluster #</th>
<th>Calibration Set</th>
<th>Validation Set</th>
<th>Prediction Set</th>
</tr>
</thead>
<tbody>
<tr>
<td>8</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>R² Cal: 0.75</td>
<td>R² CV: 0.56</td>
<td>R² Pred: 0.81</td>
</tr>
<tr>
<td></td>
<td>Bias: 4.26e-14</td>
<td>CV Bias: -1.19035</td>
<td>Pred Bias: 13.38</td>
</tr>
<tr>
<td></td>
<td>RMSEC: 21.12</td>
<td>RMSECV: 28.5243</td>
<td>RMSEP: 34.0649</td>
</tr>
<tr>
<td>9</td>
<td>R² Cal: 0.84</td>
<td>R² CV: 0.54</td>
<td>R² Pred: 0.83</td>
</tr>
<tr>
<td></td>
<td>Bias: 0</td>
<td>CV Bias: 2.41</td>
<td>Pred Bias: -13.43</td>
</tr>
<tr>
<td></td>
<td>RMSEC: 35.51</td>
<td>RMSECV: 67.81</td>
<td>RMSEP: 44.05</td>
</tr>
<tr>
<td>10</td>
<td>R² Cal: 0.85</td>
<td>R² CV: 0.791392</td>
<td>R² Pred: 0.709519</td>
</tr>
<tr>
<td></td>
<td>Bias: 5.68e-14</td>
<td>CV Bias: 1.66</td>
<td>Pred Bias: 6.91</td>
</tr>
<tr>
<td></td>
<td>RMSEC: 22.39</td>
<td>RMSECV: 27.13</td>
<td>RMSEP: 56.34</td>
</tr>
<tr>
<td>11</td>
<td>R² Cal: 0.77</td>
<td>R² CV: 0.73</td>
<td>R² Pred: 0.80</td>
</tr>
<tr>
<td></td>
<td>Bias: 0</td>
<td>CV Bias: -0.14</td>
<td>Pred Bias: -0.059</td>
</tr>
<tr>
<td></td>
<td>RMSEC: 26.61</td>
<td>RMSECV: 29.19</td>
<td>RMSEP: 25.52</td>
</tr>
</tbody>
</table>
Cluster 11 Results/Discussion

- Algorithm
  - \( MP = 0.213208(D/D) + 0.011536(Wap) + 101.442(J) + 2.9733(S1K) + 0.093732(TPC) + 0.000737(pID) + -0.238566(PCR) + 0.063346(PCD) + 1.99058(MATS8m) + 161.158(MATS1v) + 46.4475(MATS2v) + 67.106(MATS3v) + -194.013 \)

- Data Characteristics
  - \( N_{trn} = 80 \quad N_{val} = 28 \quad N_{pred} = 386 \)
Cluster 11 Validation

Applicability Domain

- Standardized Residuals
- Leverages

Training Set
Validation Set
Cluster 11 Validation

Melting Point Experimental vs Predicted

- Training Set
- Validation Set

"1:1"
Missing Clusters

- Cluster 2: still being validated and analyzed
- Cluster 5: has been validated and results are currently being compiled
- Cluster 6: shown to have a poor correlation in multiple aspects
  - Possibly due to data set
  - Endpoint more complex than originally believed
Conclusions

- The melting points of the series of POPs have been modeled using a set of calculated descriptors
  - Suggests that existing cheminformatics descriptors fail to adequately describe interactions in the crystalline solid phase
    - May be a significant cause of error in melting point prediction
- Original data set contains too many similar compounds to produce a valid model
  - Hierarchical clustering method employed via MCS Library
    - Improved predictions significantly
Future Endeavors

- Explore other possible methodology and descriptors
  - Nonlinear methods
  - Other descriptors - for example, quantum-chemical descriptors

- Model other endpoints (biological activity, solubility, etc.) with these data sets
Acknowledgements

- Dr. Jerzy Leszczynski
- Dr. Bakhtiyor Rasulev
- Dr. Tomasz Puzyn and group
- Dr. Hill and group
- NSF-CREST
References


References


- Damstra T, Page S.W., Herrman J.L., Merideth T. *J Epidemiol Community Health*. 2002; 56: 824-825
Questions???