Modeling of physicochemical properties for perfluorinated compounds using a data integration approach

Wolfram Teetz

Helmholtz Zentrum München - German Research Center for Environmental Health (GmbH)
Institute of Bioinformatics & Systems Biology

Idstein, Fresenius, 19. Juni 2010
HelmholtzZentrum münchen
German Research Center for Environmental Health
- Efficiently use scarce and scattered information available

- Reduce animal testing - use in silico and in vitro information on related compounds.
• Practical guidance to integrated risk assessment
• Full risk assessment for four compound classes (incl. PFC)
• Explicit uncertainty in data and in models
• Applicability domain
• Exemplify integration of information for risk assessments for large numbers of substances.
• How to use non-testing information for regulatory decision

FP7-funded 2009-01-01 end: 2012-12-31
## ITS

<table>
<thead>
<tr>
<th>Method</th>
<th>Costs</th>
<th>Time</th>
<th>Quality</th>
</tr>
</thead>
<tbody>
<tr>
<td>Blind <em>in-vitro</em> testing of all compounds</td>
<td><img src="image" alt="Costs" /> <img src="image" alt="Time" /> <img src="image" alt="Quality" /></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Blind <em>in-silico</em> testing of all compounds</td>
<td><img src="image" alt="Costs" /> <img src="image" alt="Time" /> <img src="image" alt="Quality" /></td>
<td></td>
<td></td>
</tr>
<tr>
<td><em>in-silico</em> testing plus <strong>intelligent testing</strong> of selected compounds</td>
<td><img src="image" alt="Costs" /> <img src="image" alt="Time" /> <img src="image" alt="Quality" /></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
What compound to measure?
What property to measure?
What property to predict?

What property can be predicted from another one? -> Knowledge

The accuracy + AD has to be estimated in the end!
Properties of molecules

- Data are lost after publication of an article
- The original sources of data are difficult to track
- The conditions of experiments are frequently not provided
- The conversion between different units is error prone
- Current databases do not allow community correction of errors
- The tracking of changes (by users) is required

Models

- Most published models are never used
- Implementation can be as difficult as new model development
- Different implementations can produce different results
“Wiki” approach to data handling
users can add, modify and delete data

Mandatory reference to an origin of information
each record in a database should contain a reference to a source (article, book, proceeding or even personal communications), where the data were published

Storing rich information
we store measurement conditions to increase data quality

Several tools to support decision making
integration with other web-services (validation of molecule names against PubChem database, automatic fetching of article information from PubMed), duplicate records management

Aimed at model building
convenient to build training sets from data - filter by property, article and export data either to internal modelling tools or download as Excel file
Model name: BarunMPBP_MP-PERFORCE-28apr10v1
Public ID is 5866871
Predicted property: Melting Point
Training method: ANN
measured in °C

<table>
<thead>
<tr>
<th>Data Set</th>
<th>Original R²</th>
<th>Original RMSE</th>
<th>Original MAE</th>
<th>Recalculated R²</th>
<th>Recalculated RMSE</th>
<th>Recalculated MAE</th>
</tr>
</thead>
<tbody>
<tr>
<td>Training set: BarunMPBP_MP-28apr10v1-93 (93 records)</td>
<td>0.84</td>
<td>36.98</td>
<td>30.59</td>
<td>0.85</td>
<td>38.69</td>
<td>30.38</td>
</tr>
<tr>
<td>Test set: BarunMPBP_MP-PERFORCE-28apr10v1-15 (15 records)</td>
<td>0.21</td>
<td>35.09</td>
<td>29.43</td>
<td>0.22</td>
<td>34.12</td>
<td>28.34</td>
</tr>
</tbody>
</table>

Correl. limit: 0.95
Supersab, 1000 iterations, 3 neurons
5-fold cross-validation

87 filtered descriptors
Supersab, 1000 iterations, 3 neurons
Calculated in 35 seconds
Model name: MP BP Model - Perforce
Training method: ANN

**Property: Melting Point measured in °C**

<table>
<thead>
<tr>
<th>Data Set</th>
<th>R²</th>
<th>RMSE</th>
<th>MAE</th>
</tr>
</thead>
<tbody>
<tr>
<td>Barun MPBP-MPBP Train(93)</td>
<td>0.85</td>
<td>36.65</td>
<td>29.20</td>
</tr>
<tr>
<td>PERFORCE MPBP(15)</td>
<td>0.32</td>
<td>31.91</td>
<td>25.89</td>
</tr>
</tbody>
</table>

**Property: Boiling Point measured in °C**

<table>
<thead>
<tr>
<th>Data Set</th>
<th>R²</th>
<th>RMSE</th>
<th>MAE</th>
</tr>
</thead>
<tbody>
<tr>
<td>Barun MPBP-MPBP Train(93)</td>
<td>0.85</td>
<td>31.76</td>
<td>20.52</td>
</tr>
<tr>
<td>PERFORCE MPBP(25)</td>
<td>0.78</td>
<td>22.78</td>
<td>17.02</td>
</tr>
</tbody>
</table>
Save the model

Please enter your model’s name: Melting Point 10->108

Model name: Melting Point, set 3731, 72 [rename]
Public ID is 2344045
Predicted property: Melting Point
Training method: ANN
measured in °C

<table>
<thead>
<tr>
<th>Data Set</th>
<th>R²</th>
<th>RMSE</th>
<th>MAE</th>
</tr>
</thead>
<tbody>
<tr>
<td>Training set: NewMP10 (10 records)</td>
<td>0.63</td>
<td>62.42</td>
<td>51.63</td>
</tr>
<tr>
<td>Test set: NewMP108 (98 records)</td>
<td>0.21</td>
<td>84.70</td>
<td>65.21</td>
</tr>
</tbody>
</table>

Graph showing data points and a trend line.
Save the model

Please enter your model's name: Melting Point from BP

Model name: Melting Point, set 3730, 67
Training method: ANN

Property: Melting Point measured in °C (Details..)

<table>
<thead>
<tr>
<th>Data Set</th>
<th>R²</th>
<th>RMSE</th>
<th>MAE</th>
</tr>
</thead>
<tbody>
<tr>
<td>NewBP118(10)</td>
<td>0.73</td>
<td>53.63</td>
<td>48.15</td>
</tr>
<tr>
<td>NewMP108(98)</td>
<td>0.49</td>
<td>65.63</td>
<td>56.27</td>
</tr>
</tbody>
</table>

Property: Boiling Point measured in °C (Details..)

<table>
<thead>
<tr>
<th>Data Set</th>
<th>R²</th>
<th>RMSE</th>
<th>MAE</th>
</tr>
</thead>
<tbody>
<tr>
<td>NewBP116(118)</td>
<td>0.91</td>
<td>27.04</td>
<td>16.55</td>
</tr>
<tr>
<td>NewMP108(0)</td>
<td>NaN</td>
<td>NaN</td>
<td>NaN</td>
</tr>
</tbody>
</table>
**Open Positions**

Marie Curie Initial Training Networks (ITN) are aimed at improving the career perspectives of European researchers who are in the first five years of their career by offering structured training in well defined scientific and technological areas as well as providing complementary skills and exposing the researchers to other sectors including private companies.

A number of open positions is available within two conceptual work packages listed below.

The positions are available within two workpackages:

- **WP 1 - Urgent problems of QSAR/QSPR modeling for REACH**
- **WP 2 - Computational and in vitro screening methods for nano-materials**

There are:

- one 1-year duration **ER (Experienced Researchers)**, postdoc position, position,
- still 2 of 11 **Long Term Fellowships (LTF)** (ESR positions, Early Stage Researcher, PhD positions), each one of 36 months duration, and
- 37 **Short Term Fellowships (STF)** (ESR positions), each one of 3-12 months duration.
Acknowledgements

Our group
Iurii Sushko
Sergii Novotarskyi
Anil K. Pandey
Robert Körner
Stefan Brandmaier
Matthias Rupp
Wolfram Teetz
Eva Schlosser

Collaborators:
Dr. G. Poda (Pfizer)
Dr. C. Ostermann (Nycomed)
Dr. C. Höfer (DMPKore)
Prof. A.S. Dimoglo (Turkey)
Prof. A. Tropsha (NC, USA)
Prof. T. Oprea (New Mexico, USA)
Prof. A. Varnek (Strasbourg, France)
Prof. R. Mannhold (Düsseldorf, Germany)
Prof. R. Todeschini (Milano, Italy)
+ many other colleagues

Funding
GO-Bio BMBF http://qspr.eu
Germany-Ukraine grant UKR 08/006
FP7 Marie Curie ITN ECO http://eco-itn.eu
FP7 CADASTER http://www.cadaster.eu