The importance of data curation on QSAR Modeling: PHYSPROP open data as a case study

Kamel Mansouri, Christopher Grulke, Ann Richard, Richard Judson, Antony Williams

NCCT, U.S. EPA

QSAR 2016

14 June 2016, Miami, FL

The views expressed in this presentation are those of the author and do not necessarily reflect the views or policies of the U.S. EPA
Recent Cheminformatics development at NCCT

- We are building a new cheminformatics architecture
- PUBLIC dashboard gives access to curated chemistry
- Focus on integrating EPA and external resources
- Aggregating and curating data, visualization elements and "services" to underpin other efforts
  - RapidTox
  - Read-across
  - Predictive modeling
  - Non-targeted screening
Developing “NCCT Models”

• Interest in physicochemical properties to include in exposure modeling, augmented with ToxCast HTS *in vitro* data etc.

• Our approach to modeling:
  – Obtain high quality training sets
  – Apply appropriate modeling approaches
  – Validate performance of models
  – Define the applicability domain and limitations of the models
  – Use models to predict properties across our full datasets

• Work has been Initiated using available physicochemical data
PHYSPROP Data: Available from:
http://esc.syrres.com/interkow/EpiSuiteData.htm

- Water solubility
- Melting Point
- Boiling Point
- LogP (KOWWIN: Octanol-water partition coefficient)
- Atmospheric Hydroxylation Rate
- LogBCF (Bioconcentration Factor)
- Biodegradation Half-life
- Ready biodegradability
- Henry's Law Constant
- Fish Biotransformation Half-life
- LogKOA (Octanol/Air Partition Coefficient)
- LogKOC (Soil Adsorption Coefficient)
- Vapor Pressure
Data Files

- The data files have **FOUR** representations of a chemical, plus the property value.

http://esc.syrres.com/interkow/EpiSuiteData.htm
The Approach

• To build models we need the set of chemicals and their property series

• Our curation process
  – Decide on the “chemical” by checking levels of consistency
  – We did NOT validate each measured property value
  – Perform initial analysis manually to understand how to clean the data (chemical structure and ID)
  – Automate the process (and test iteratively)
  – Process all datasets using final method
General Observations from LogP dataset

- CAS Numbers not matching structure
- Some SMILES won’t convert (non-standard SMILES)
- Valence and charge imbalance issues
- Stereochemistry poorly depicted if not totally absent
- Multiple duplicate pairs for a particular chemical compound
- Majority of duplicates from structure representations not matching the chemical.
KNIME workflow to evaluate the dataset
LogP dataset: 15,809 chemicals (structures)

- CAS Checksum: 12163 valid, 3646 invalid (>23%)
- Invalid names: 555
- Invalid SMILES 133
- Valence errors: 322 Molfile, 3782 SMILES (>24%)
- Duplicates check:
  - 31 DUPLICATE MOLFILES
  - 626 DUPLICATE SMILES
  - 531 DUPLICATE NAMES
- SMILES vs. Molfiles (structure check)
  - 1279 differ in stereochemistry (~8%)
  - 362 “Covalent Halogens”
  - 191 differ as tautomers
  - 436 are different compounds (~3%)
### Invalid CASRNs

<table>
<thead>
<tr>
<th>CASRN</th>
<th>Name</th>
</tr>
</thead>
<tbody>
<tr>
<td>SRC000-01-7</td>
<td>Ethanaminium, 2-((chlorosetyl)oxy)-N,N,N-trimet</td>
</tr>
<tr>
<td>SRC000-02-3</td>
<td>2-Fluorocarboxamide-picrate</td>
</tr>
<tr>
<td>SRC000-02-7</td>
<td>Ethanaminium, N,N,N-trimethyl-2-((1-oxo-2-propyl</td>
</tr>
<tr>
<td>SRC000-04-3</td>
<td>Guanidine, N-hydroxy-N&quot;-(4-(methylthio)benzenesulfonamide</td>
</tr>
<tr>
<td>SRC000-04-4</td>
<td>Hydrazinecarboximidamide, N&quot;-(4-(methylthio)benz</td>
</tr>
<tr>
<td>SRC000-04-5</td>
<td>NNN5-TeMe-N&quot;-(3Furan)NMe  Br</td>
</tr>
<tr>
<td>SRC000-04-6</td>
<td>Benzenamine, 4-bromo-N,N-bis(2,2,2-trifluoroethyl</td>
</tr>
<tr>
<td>SRC000-04-7</td>
<td>2-Propenoic acid, 3-(2-chlorophenoyl)-, methyl e</td>
</tr>
<tr>
<td>SRC000-05-1</td>
<td>9H-Purine-9-oxaldehyde, a-(1-formyl)-2-hydroxy</td>
</tr>
<tr>
<td>SRC000-05-2</td>
<td>N1-Pr-N2-CN-N3-Me guanidine</td>
</tr>
<tr>
<td>SRC000-05-3</td>
<td>1-(2-OH)-2-Me imidazole HCL</td>
</tr>
<tr>
<td>SRC000-06-3</td>
<td>Propanoic acid, 3-[(4-cyanophenyl)methyl]salene</td>
</tr>
</tbody>
</table>

### Truncated names

### Missing SMILES
Examples of errors

- 362 Halogens bonded to nitrogen

<table>
<thead>
<tr>
<th>CAS</th>
<th>Name</th>
<th>Structure</th>
</tr>
</thead>
<tbody>
<tr>
<td>000036-93-9</td>
<td>BENZYL TRIMETHYL AMMONIUM CHLORIDE</td>
<td><img src="structure1.png" alt="Benzyl Trimethyl Ammonium Chloride" /></td>
</tr>
<tr>
<td>000069-05-3</td>
<td>TETRAETHYL AMMONIUM IOIDE</td>
<td><img src="structure2.png" alt="Tetraethyl Ammonium Iodide" /></td>
</tr>
<tr>
<td>000071-91-0</td>
<td>TETRAETHYL AMMONIUM BROMIDE</td>
<td><img src="structure3.png" alt="Tetraethyl Ammonium Bromide" /></td>
</tr>
</tbody>
</table>
Examples of errors

• 191 Valence errors
Examples of errors

- 463 completely different compounds
Examples of errors

- Duplicate Structures

<table>
<thead>
<tr>
<th>Structure</th>
<th>Formula</th>
<th>FW</th>
<th>CAS</th>
<th>NAME</th>
<th>MP</th>
<th>EstMP</th>
<th>ErrorMP</th>
</tr>
</thead>
<tbody>
<tr>
<td><img src="image1" alt="Structure" /></td>
<td>C₃H₆O₃</td>
<td>90.0779</td>
<td>000055-21-5</td>
<td>LACTIC ACID</td>
<td>1.6800000000000000</td>
<td>00e+001</td>
<td>2.2660000000000000</td>
</tr>
<tr>
<td><img src="image2" alt="Structure" /></td>
<td>C₃H₆O₃</td>
<td>90.0779</td>
<td>000079-33-4</td>
<td>L-LACTIC ACID</td>
<td>5.3000000000000000</td>
<td>00e+001</td>
<td>2.2660000000000000</td>
</tr>
<tr>
<td><img src="image3" alt="Structure" /></td>
<td>C₃H₆O₃</td>
<td>90.0779</td>
<td>000598-82-3</td>
<td>A-HYDROXYPROPRONIC ACID</td>
<td>1.8000000000000000</td>
<td>00e+001</td>
<td>2.2660000000000000</td>
</tr>
<tr>
<td><img src="image4" alt="Structure" /></td>
<td>C₃H₆O₃</td>
<td>90.0779</td>
<td>010220-41-7</td>
<td>D-LACTIC ACID</td>
<td>5.2900000000000000</td>
<td>00e+001</td>
<td>2.2660000000000000</td>
</tr>
<tr>
<td><img src="image5" alt="Structure" /></td>
<td>C₃H₆O₃</td>
<td>90.0779</td>
<td></td>
<td></td>
<td>5.0140000000000000</td>
<td>00e+001</td>
<td>2.2660000000000000</td>
</tr>
</tbody>
</table>
Quality flags: 1-4 STARs

4 levels of consistency exists between:

• The Molblock
• The SMILES string
• The chemical name (based on ACD/Labs dictionary)
• The CAS Number (based on a DSSTox lookup)
Quality FLAGS into LogP data

- 4 Stars ENHANCED: 4 levels of consistency with stereo information
- 4 Stars: 4 levels of consistency, stereo ignored.
- 3 Stars Plus: 3 out of 4 levels. The 4th is a tautomer.
- 3 Stars ENHANCED: 4 levels of consistency with stereo information
- 3 Stars: 3 levels of consistency, stereo ignored.
- 2 Stars PLUS: 2 out of 4 levels. The 3th is a tautomer.
- 1 Star - What's left.
Improved structures and updated flags

- 3 STAR and 2 STAR Plus are "upgraded" to a higher level of consistency
  Done by correcting the mismatching field(s), or by generating a name or smiles string when missing or unreadable.

- 3 STAR to 4 Star:
  • Available: Molblock, Name, CAS: Smiles generated from Molblock (DSSTOX)
  • Available: Molblock, Smiles, CAS: Name retrieved from DSSTOX
  • Available: Name, Smiles, CAS: Molblock retrieved from DSSTOX
  • Available: Molblock, Smiles, Name: CAS retrieved when available in DSSTOX (no stereoisomers)

- 2 Star Plus with Unreadable Smiles, name or CAS

- Total upgraded chemicals for LogP data: 1740 chemicals
- Total chemicals with 3 STAR levels of consistency for LogP data: 7910 chemicals
- Total chemicals with 4 STAR levels of consistency for LogP data: 6525 chemicals

Only part considered For QSAR
Structure standardization

Initial structures

- Remove inorganics and mixtures
- Cleaning salts and counterions
- Normalization of tautomers
- Removal of duplicates
- Final inspection

QSAR-ready structures
**Aim of the workflow:**

- Combine (not reproduce) different procedures and ideas
- Minimize the differences between the structures used for prediction by different groups
- Produce a flexible free and open source workflow to be shared

Mansouri et al. (http://ehp.niehs.nih.gov/15-10267/)
### Summary:

<table>
<thead>
<tr>
<th>Property</th>
<th>Initial file flagged</th>
<th>Updated 3-4 STAR</th>
<th>Curated QSAR ready</th>
</tr>
</thead>
<tbody>
<tr>
<td>AOP</td>
<td>818</td>
<td>818</td>
<td>745</td>
</tr>
<tr>
<td>BCF</td>
<td>685</td>
<td>618</td>
<td>608</td>
</tr>
<tr>
<td>BioHC</td>
<td>175</td>
<td>151</td>
<td>150</td>
</tr>
<tr>
<td>Biowin</td>
<td>1265</td>
<td>1196</td>
<td>1171</td>
</tr>
<tr>
<td>BP</td>
<td>5890</td>
<td>5591</td>
<td>5436</td>
</tr>
<tr>
<td>HL</td>
<td>1829</td>
<td>1758</td>
<td>1711</td>
</tr>
<tr>
<td>KM</td>
<td>631</td>
<td>548</td>
<td>541</td>
</tr>
<tr>
<td>KOA</td>
<td>308</td>
<td>277</td>
<td>270</td>
</tr>
<tr>
<td>LogP</td>
<td>15809</td>
<td>14544</td>
<td>14041</td>
</tr>
<tr>
<td>MP</td>
<td>10051</td>
<td>9120</td>
<td>8656</td>
</tr>
<tr>
<td>PC</td>
<td>788</td>
<td>750</td>
<td>735</td>
</tr>
<tr>
<td>VP</td>
<td>3037</td>
<td>2840</td>
<td>2716</td>
</tr>
<tr>
<td>WF</td>
<td>5764</td>
<td>5076</td>
<td>4836</td>
</tr>
<tr>
<td>WS</td>
<td>2348</td>
<td>2046</td>
<td>2010</td>
</tr>
</tbody>
</table>
Development of a QSAR model

- Curation of the data
  - *Flagged and curated files available for sharing*
- Preparation of training and test sets
  - *Inserted as a field in SDFiles and csv data files*
- Calculation of an initial set of descriptors
  - *PaDEL 2D descriptors and fingerprints generated and shared*
- Selection of a mathematical method
  - *Several approaches tested: KNN, PLS, SVM…*
- Variable selection technique
  - *Genetic algorithm*
- Validation of the model’s predictive ability
  - *5-fold cross validation and external test set*
- Define the Applicability Domain
  - *Local (nearest neighbors) and global (leverage) approaches*
The conditions for the validity of QSARs

<table>
<thead>
<tr>
<th>Principle</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>1) A defined endpoint</td>
<td>Any physicochemical, biological or environmental effect that can be measured and therefore modelled.</td>
</tr>
<tr>
<td>2) An unambiguous algorithm</td>
<td>Ensure transparency in the description of the model algorithm.</td>
</tr>
<tr>
<td>3) A defined domain of applicability</td>
<td>Define limitations in terms of the types of chemical structures, physicochemical properties and mechanisms of action for which the models can generate reliable predictions.</td>
</tr>
</tbody>
</table>
| 4) Appropriate measures of goodness-of-fit, robustness and predictivity   | a) The internal fitting performance of a model  
   b) the predictivity of a model, determined by using an appropriate external test set.                                                                                                                    |
<p>| 5) Mechanistic interpretation, if possible                               | Mechanistic associations between the descriptors used in a model and the endpoint being predicted.                                                                                                           |</p>
<table>
<thead>
<tr>
<th>Prop</th>
<th>Vars</th>
<th>5-fold CV (75%)</th>
<th>Training (75%)</th>
<th>Test (25%)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>Q2</td>
<td>RMSE</td>
<td>N</td>
</tr>
<tr>
<td>BCF</td>
<td>10</td>
<td>0.84</td>
<td>0.55</td>
<td>465</td>
</tr>
<tr>
<td>BP</td>
<td>13</td>
<td>0.93</td>
<td>22.46</td>
<td>4077</td>
</tr>
<tr>
<td>LogP</td>
<td>9</td>
<td>0.85</td>
<td>0.69</td>
<td>10531</td>
</tr>
<tr>
<td>MP</td>
<td>15</td>
<td>0.72</td>
<td>51.8</td>
<td>6486</td>
</tr>
<tr>
<td>VP</td>
<td>12</td>
<td>0.91</td>
<td>1.08</td>
<td>2034</td>
</tr>
<tr>
<td>WS</td>
<td>11</td>
<td>0.87</td>
<td>0.81</td>
<td>3158</td>
</tr>
<tr>
<td>HL</td>
<td>9</td>
<td>0.84</td>
<td>1.96</td>
<td>441</td>
</tr>
</tbody>
</table>
### NCCT models

<table>
<thead>
<tr>
<th>Prop</th>
<th>Vars</th>
<th>5-fold CV (75%)</th>
<th>Training (75%)</th>
<th>Test (25%)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>Q2</td>
<td>RMSE</td>
<td>N</td>
</tr>
<tr>
<td>AOH</td>
<td>13</td>
<td>0.85</td>
<td>1.14</td>
<td>516</td>
</tr>
<tr>
<td>BioHL</td>
<td>6</td>
<td>0.89</td>
<td>0.25</td>
<td>112</td>
</tr>
<tr>
<td>KM</td>
<td>12</td>
<td>0.83</td>
<td>0.49</td>
<td>405</td>
</tr>
<tr>
<td>KOC</td>
<td>12</td>
<td>0.81</td>
<td>0.55</td>
<td>545</td>
</tr>
<tr>
<td>KOA</td>
<td>2</td>
<td>0.95</td>
<td>0.69</td>
<td>202</td>
</tr>
<tr>
<td></td>
<td></td>
<td>BA</td>
<td>Sn-Sp</td>
<td>BA</td>
</tr>
<tr>
<td>R-Bio</td>
<td>10</td>
<td>0.8</td>
<td>0.82-0.78</td>
<td>1198</td>
</tr>
</tbody>
</table>
LogP Model: Weighted kNN Model, 9 descriptors

Weighted 5-nearest neighbors
9 Descriptors
Training set: 10531 chemicals
Test set: 3510 chemicals

5 fold Cross-validation:
Q2=0.85  RMSE=0.69
Fitting:
R2=0.86   RMSE=0.67
Test:
R2=0.86   RMSE=0.78
Standalone application:

**Input:**
- MATLAB .mat file, an ASCII file with only a matrix of variables
- SDF file or SMILES strings of QSAR-ready structures. In this case the program will calculate PaDEL 2D descriptors and make the predictions.

**Output**
- Depending on the extension, the can be text file or csv with
  - A list of molecules IDs and predictions
  - Applicability domain
  - Accuracy of the prediction
  - Similarity index to the 5 nearest neighbors
  - The 5 nearest neighbors from the training set: Exp. value, Prediction, InChi key
The iCSS Chemistry Dashboard at https://comptox.epa.gov
The iCSS Chemistry Dashboard
NCCT Models: Melting point (MP)

Model Performance

Weighted KNN model
15 molecular descriptors

Model Results

Predicted value: 143 °C
Observed value in training set: Not available
Global applicability domain: Inside the AD
Local applicability domain index: 0.88
Confidence level: 0.70

5 nearest neighbors from the training set:

- Benzilic acid
  - Observed: 163 C
  - Predicted: 141 C

- 4'-Methylbenzanilide
  - Observed: 158 C
  - Predicted: 141 C

- 2-Acetamidobiphenyl
  - Observed: 121 C
  - Predicted: 150 C

- 3'-Methylbenzanilide
  - Observed: 125 C
  - Predicted: 150 C

- 2'-Methylbenzanilide
  - Observed: 145 C
  - Predicted: 143 C
QMRF for LogP model

1. QMRF identifier

1.1. QSAR identifier (title):
   LogP: Octanol-water partition coefficient prediction from the NCCT_Models Suite.

1.2. Other related models:
   No related models

1.3. Software coding the model:
   NCCT_models V1.02
   Suite of QSAR models to predict physicochemical properties and environmental fate of organic chemicals
   Kamel Mansouri (mansouri.kamel@epa.gov; mansourikamal@gmail.com);
   https://comptox.epa.gov/dashboard/

PaDEL descriptors V2.21
   Open source software to calculate molecular descriptors and fingerprints.
   Chun Wei Yap (phayapc@nus.edu.sg)
Conclusion

- QSAR prediction models (kNN) produced for all properties
- 700k chemical structures pushed through NCCT_Models
- Supplementary data will include appropriate files with flags – full dataset plus QSAR ready form
- Full performance statistics available for all models
- Models will be deployed as prediction engines in the future – one chemical at a time and batch processing (to be done after RapidTox Project)
Thank you for your attention