An examination of the impact of data quality on QSAR Modeling in regards to the environmental sciences

Antony Williams & Kamel Mansouri
National Center for Computational Toxicology

April 12, 2016

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 have been building a new cheminformatics architecture
• The first “app” provides PUBLIC access to curated chemistry
• Focus on integrating agency and external resources
• Aggregating and curating data, visualization elements and “services” to underpin other internal efforts including:
  • Read-across
  • Predictive modeling
  • Non-targeted screening
QC Levels

- **DSSTox_High**: Hand curated and validated
- **DSSTox_Low**: Hand curated and confirmed using multiple public sources
- **Public_High**: Extracted from EPA SRS and confirmed to have no conflicts in ChemID and PubChem
- **Public_Medium**: Extracted from ChemID and confirmed to have no conflicts in PubChem
- **Public_Low**: Extracted from ACToR or PubChem
- **Public_Untrusted**: Postulated, but found to have conflicts in public sources

QC Level Totals (12Jun2015)

- **DSSTox_High**: 4535 validated
- **DSSTox_Low**: 16K
- **Public_High**: 33K
- **Public_Medium**: 101K
- **Public_Low**: ~310K pending
- **Public_Untrusted**: ~150K pending
Chemistry Software Architecture

**User Interfaces**
- Non-Targeted Screening
- External Component Users
- RapidTox
- ToxQP
- iCSS Chemistry Dashboard

**UI Components**
- Data Annotation
- Graphing Views
- Tile and Table Result Views
- Structure Visualization
- Identifier Search, Structure Search

**APIs and Web Services**
- Calculated properties (temp), Models
- Calculated Properties
- Experimental Properties
- Similarity and substructure searches
- Structures, Identifiers

**Data Layer**
- TOXQP Database
- PhysChem Database
- ACTOR v1 and v2
- DSSTox Compounds
- Prototyping Data Collections

**Chemical-Biology Platforms**
The iCSS Chemistry Dashboard at [https://comptox.epa.gov](https://comptox.epa.gov)

Chemistry Dashboard

Search a chemical by systematic name, synonym, CAS number, or InChIKey

- Single component search
- Ignore isotopes

Need more? Use advanced search.

720 Thousand Chemicals
The iCSS Chemistry Dashboard

### Bisphenol A

**80-05-7 | DTXD1020182**

- **Molecular Formula:** C15H16O2
- **Average Mass:** 228.291 g/mol
- **Monoisotopic Mass:** 228.11563 g/mol

## Intrinsic Properties

### Structural Identifiers

## Citation

### Chemical Properties

<table>
<thead>
<tr>
<th>Property</th>
<th>Average (Exp.)</th>
<th>Range (Exp.)</th>
<th>Average (Pred.)</th>
<th>Range (Pred.)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Solubility</td>
<td>0.0001 (1)</td>
<td>0.00005257 to 0.0005257</td>
<td>0.38 (2)</td>
<td>0.003675 to 0.7565</td>
</tr>
<tr>
<td>Melting Point</td>
<td>154.929 (7)</td>
<td>153.0 to 158.0</td>
<td>144.033 (3)</td>
<td>131.8 to 158.0</td>
</tr>
<tr>
<td>Boiling Point</td>
<td>200.0 (1)</td>
<td>200.0 to 200.0</td>
<td>348.95 (2)</td>
<td>334.4 to 363.5</td>
</tr>
<tr>
<td>LogP</td>
<td>3.357 (3)</td>
<td>3.32 to 3.431</td>
<td>3.524 (3)</td>
<td>3.226 to 3.727</td>
</tr>
<tr>
<td>Atmospheric Hydroxylation Rate</td>
<td>N/A</td>
<td>N/A</td>
<td>0.0 (1)</td>
<td>4.237e-11 to 4.237e-11</td>
</tr>
<tr>
<td>LogKOC</td>
<td>1.64 (1)</td>
<td>1.64 to 1.64</td>
<td>1.376 (1)</td>
<td>1.376 to 1.376</td>
</tr>
<tr>
<td>Biodegradation Half-life</td>
<td>N/A</td>
<td>N/A</td>
<td>15.11 (1)</td>
<td>15.11 to 15.11</td>
</tr>
<tr>
<td>Henry's Law Constant</td>
<td>N/A</td>
<td>N/A</td>
<td>0.0 (1)</td>
<td>0.972e-07 to 0.972e-07</td>
</tr>
</tbody>
</table>
The iCSS Chemistry Dashboard
The iCSS Chemistry Dashboard
The iCSS Chemistry Dashboard

Bisphenol A
80-05-7 | DTXSID7020182

Molecular Formula: C15H16O2
Average Mass: 228.291 g/mol
Monoisotopic Mass: 228.11903 g/mol

Structural Identifiers

Citation

General
- ToxCast Dashboard 2
- EPA Substance Registry Service
- PubChem
- ChemSpider
- EDSP Dashboard
- CPCat
- DrugBank
- Google Patents
- National Environmental Methods Index
- ChemView

Publications
- Google Scholar
- PubMed

Biochemistry
- CTD
- Gene-Tox

Prediction
- Chemicalize
Simple Link Integration

Edit Link

**Name**
National Environmental Methods Index

**Url**
http://www.nem.gov/methods/keyword/?keyword_search_field=%s

**Parameter**
CAS

**Category**
General

**Select icon**
Search Icon

**Description**
NEMI is a searchable database that allows scientists and managers to find and compare analytical and field methods for all phases of environmental monitoring.

Use this database to get information on chemical health and safety data received by EPA and EPA's assessments and regulatory actions for specific chemicals under the Toxic Substances Control Act (TSCA).

ChemView

Information and data sources made available via ChemView are currently undergoing system upgrades or maintenance. During this time, some pages of the site (searchable databases, PDFs, etc.) may not be available. We apologize for the inconvenience and thank you for your patience and understanding.

Use this database to get information on chemical health and safety data received by EPA and EPA's assessments and regulatory actions for specific chemicals under the Toxic Substances Control Act (TSCA).

If you do not receive results for a particular chemical, it does not mean EPA does not have information on that chemical; the data may not be posted yet but will be available in the future as EPA continues to populate the database:

- Learn more and find additional information about EPA’s efforts in assessing and managing chemicals
- Read the ChemView User’s Guide and Web Service Information
- To continuously improve ChemView, contact us with your feedback.
Data sharing and access

Cholesterol
57-88-5 | DTXSID3022401

Searched by Synonym: Found 1 result for 'cholesterol'.

Chemical Properties: LogP

<table>
<thead>
<tr>
<th>Property</th>
<th>Raw Result</th>
<th>Mean Result</th>
<th>Minimum Result</th>
<th>Maximum Result</th>
<th>Result Unit</th>
<th>Result Type</th>
<th>Source</th>
</tr>
</thead>
<tbody>
<tr>
<td>Estimated Log Kow</td>
<td>3.64</td>
<td>3.64</td>
<td>3.64</td>
<td>3.64</td>
<td>predicted</td>
<td>experimental</td>
<td>EPI SUITE</td>
</tr>
<tr>
<td>Log P</td>
<td>3.431</td>
<td>3.431</td>
<td>3.431</td>
<td>3.431</td>
<td>experimental</td>
<td>EPI SUITE</td>
<td></td>
</tr>
<tr>
<td>Log P</td>
<td>3.727</td>
<td>3.727</td>
<td>3.727</td>
<td>3.727</td>
<td>predicted</td>
<td>ACD/Labs</td>
<td></td>
</tr>
<tr>
<td>Measured Log Kow</td>
<td>3.32</td>
<td>3.32</td>
<td>3.32</td>
<td>3.32</td>
<td>experimental</td>
<td>EPI SUITE</td>
<td></td>
</tr>
<tr>
<td>Octanol-water partition coefficient</td>
<td>3.32</td>
<td>3.32</td>
<td>3.32</td>
<td>3.32</td>
<td>experimental</td>
<td>CURATED_PHYSP</td>
<td></td>
</tr>
</tbody>
</table>

Found 75 synonyms

Bisphenol A
4,4'-Propane-2,2-diyldiphenol
phenol, 4,4'-[1-methylethylidene]bis-
BPA
4,4'-Propane-2,2-diyldiphenol
Phenol, 4,4'-[1-methylethylidene]bis-
80-05-7 CAS RN
4-06-00-00717 Registered Registry Number
UNII-MLT3645109 FDA Registry Number
(4,4'-Dihydroxydiphenyl)dimethylmethane

Office of Research and Development
### Chemical Properties: LogP

<table>
<thead>
<tr>
<th>Property</th>
<th>Raw Result</th>
<th>Mean Result</th>
<th>Minimum Result</th>
<th>Maximum Result</th>
<th>Result Type</th>
<th>Source</th>
</tr>
</thead>
<tbody>
<tr>
<td>Estimated Log Kow</td>
<td>2.62</td>
<td>2.62</td>
<td>2.62</td>
<td>2.62</td>
<td>predicted</td>
<td>EPI SUITE</td>
</tr>
<tr>
<td>LogP</td>
<td>2.632</td>
<td>2.632</td>
<td>2.632</td>
<td>2.632</td>
<td>experimental</td>
<td>Vitas-M</td>
</tr>
<tr>
<td>LogP</td>
<td>2.674</td>
<td>2.674</td>
<td>2.674</td>
<td>2.674</td>
<td>predicted</td>
<td>ACD/Labs</td>
</tr>
<tr>
<td>Measured Log Kow</td>
<td>2.61</td>
<td>2.61</td>
<td>2.61</td>
<td>2.61</td>
<td>experimental</td>
<td>EPI SUITE</td>
</tr>
<tr>
<td>Octanol-water partition</td>
<td>2.61</td>
<td>2.61</td>
<td>2.61</td>
<td>2.61</td>
<td>experimental</td>
<td>CURATED_PHYSP</td>
</tr>
<tr>
<td>Octanol-water partition</td>
<td>2.6669</td>
<td>2.67</td>
<td>2.67</td>
<td>2.67</td>
<td>predicted</td>
<td>NCCT_Models</td>
</tr>
</tbody>
</table>

### Additional Properties

<table>
<thead>
<tr>
<th>Property</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Henry’s Law Constant</td>
<td>N/A</td>
</tr>
<tr>
<td>Fish Bioconversion-Half-life</td>
<td>0.089 (1)</td>
</tr>
<tr>
<td>LogKOA</td>
<td>N/A</td>
</tr>
<tr>
<td>LogKOC</td>
<td>2.24 (1)</td>
</tr>
<tr>
<td>Vapor Pressure</td>
<td>0.0 (1)</td>
</tr>
</tbody>
</table>

The PHYSPROP data sets are the publicly available data files underpinning the EPISuiteTM prediction models (http://www.epa.gov/tsca-screening-tools/episuite-estimation-program-interface). The data were curated using a combination of manual and automated processing routines with only the highest quality data reported in this database. These data were utilized, in combination with other datasets where available, in the development of the NCCT Models provided via this iCSS Chemistry Dashboard. This partition coefficient is a ratio of concentrations of un-ionized compound between the two liquid phases. This property is the Octanol-Water partition coefficient, log Kow.
Developing “NCCT Models”

• Interest in physicochemical properties to include in exposure modeling, augmenting with Toxcast data etc.

• Our approach to modeling:
  – Obtain high quality training sets
  – Apply appropriate modeling approaches (modern machine learning)
  – Validate performance of models and use to predict properties across our full datasets

• Initiated work with available “EPI Suite” data
EPI Suite Properties

- 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
Looking at EPISuite KOWWIN: 2447 Training Compounds

- KOWWIN model was trained on 2447 chemicals

- Available experimental dataset is 15,809 chemicals

- Previous experience of the data necessitated review
Predicted vs. Experimental considering >15k chemicals
Data Files

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

- **CAS Number**: 80-05-7
- **Name**: Bisphenol A
- **SMILES**: CC(C)(C1=CC=C(O)C=C1)C1=CC=C(O)C=C1

- **Molfile**:
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
  – Automate the process (and test iteratively)
  – Feed through all datasets
General Observations

• MANY mismatches across datasets – comparing structure molfile, SMILES, Names and CAS Numbers
• Duplicate pairs for chemicals – same structures but different values for properties. Some wildly different.
• Some SMILES cannot convert – what software was used??
• Invalid CAS Numbers (can be validated with a CheckSum)
• Some chemical names are neither systematic nor trivial. Totally ambiguous and truncated
• “Stereochemistry” representations are not appropriate for computers and commonly not even included
<table>
<thead>
<tr>
<th>Chemical Name</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ethanaminium, N,N,N-trimethyl-2-[(1-oxo-2-propen</td>
</tr>
<tr>
<td>Guanidine, N-hydroxy-N&quot;-[4-(methylthio)benzeneme</td>
</tr>
<tr>
<td>Hydrazinecarboximidamide, N'-[4-(methylthio)benz</td>
</tr>
<tr>
<td>NNN5-TeMe-N-(3FuranMe),ammon Br</td>
</tr>
<tr>
<td>Benzenamine, 4-bromo-N,N-bis(2,2,2-trifluoroethy</td>
</tr>
<tr>
<td>2-Propenoic acid, 3-(2-chlorophenoxy)-, methyl e</td>
</tr>
<tr>
<td>9H-Purine-9-acetaldehyde, a-(1-formyl-2-hydroxye</td>
</tr>
<tr>
<td>N1-Pr-N2-CN-N3-Me guanidine</td>
</tr>
<tr>
<td>1-(2-OHet)-2-Me imidazoline HCL</td>
</tr>
</tbody>
</table>
# “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 1" /></td>
<td>C₂H₅O₃</td>
<td>90.0779</td>
<td>060050-21-5</td>
<td>LACTIC ACID</td>
<td>1.6800000000000</td>
<td>2.2660000000000</td>
<td>5.8600000000000</td>
</tr>
<tr>
<td><img src="image2" alt="Structure 2" /></td>
<td>C₂H₅O₃</td>
<td>90.0779</td>
<td>060079-33-4</td>
<td>L-LACTIC ACID</td>
<td>5.3000000000000</td>
<td>2.2660000000000</td>
<td>-3.0340000000000</td>
</tr>
<tr>
<td><img src="image3" alt="Structure 3" /></td>
<td>C₂H₅O₃</td>
<td>90.0779</td>
<td>0600598-82-3</td>
<td>A-HYDROXYPRGPICNIC ACID</td>
<td>1.8000000000000</td>
<td>2.2660000000000</td>
<td>4.6500000000000</td>
</tr>
<tr>
<td><img src="image4" alt="Structure 4" /></td>
<td>C₂H₅O₃</td>
<td>90.0779</td>
<td>010326-41-7</td>
<td>D-LACTIC ACID</td>
<td>5.2800000000000</td>
<td>2.2660000000000</td>
<td>-3.0140000000000</td>
</tr>
</tbody>
</table>
# Large Differences in Values?

<table>
<thead>
<tr>
<th>Structure</th>
<th>Formula</th>
<th>FW</th>
<th>CAS</th>
<th>NAME</th>
<th>MP</th>
<th>EsMP</th>
<th>ErrorMP</th>
</tr>
</thead>
<tbody>
<tr>
<td><img src="image1.png" alt="Structure 1" /></td>
<td>C_{11}H_{14}O_{2}</td>
<td>178.2277</td>
<td>000093-15-2</td>
<td>METHYLEUGENOL</td>
<td>-4.000000000000000000e+000</td>
<td>3.285000000000000000e+001</td>
<td>3.685000000000000000e+001</td>
</tr>
<tr>
<td><img src="image2.png" alt="Structure 2" /></td>
<td>C_{11}H_{14}O_{2}</td>
<td>178.2277</td>
<td>006380-24-1</td>
<td>4-Acetyl-1,2-dimethoxy-benzene</td>
<td>7.000000000000000000e+001</td>
<td>3.285000000000000000e+001</td>
<td>-3.715000000000000000e+001</td>
</tr>
</tbody>
</table>

6380-24-1

93-15-2
Salts

PHYSPROP database

Appropriate Depiction
Some statistics: 15,809 chemicals (KOWWIN dataset)

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

• 4 Star ENHANCED Stereochemistry – chemical name, CAS Number, Molfile and SMILES all consistent PLUS enhanced with correct stereochemistry
Quality FLAGS

• 4 Star – chemical name, CAS Number, Molfile and SMILES all consistent. Stereo ignored.

• 3 Star ENHANCED Stereochemistry – consistency between 3 of 4: CAS Number (DSSTox lookup), Molfile, SMILES, Tautomer - PLUS enhanced with correct stereochemistry

• 3 Star– consistency between 3 of 4: CAS Number, Molfile, SMILES, Tautomer - stereo ignored
Quality FLAGS

Treated as “Junk Data” in terms of using as training data

• 2 Star PLUS – consistency between 2 of 4: chemical name, CAS Number (DSSTox lookup), Molfile, SMILES, Tautomer. Molfile is changed to be consistent if necessary
• 2 Star – only two fields internally consistent
• 1 star – what’s left

• AUTOMATION procedures used across all data – “KNIME” pipelining software for processing SDF files
Cleaning the data

• Models built on “QSAR ready” forms of the curated structures – standardize tautomers, remove stereochemistry, no salts

• Compare 3 STAR and BETTER models with entire dataset

• Exclusion of 1 and 2 star data removed SOME “outliers” only

• Definite cluster not being modeled appropriately - “Applicability domain” – initial model based on ~2500 chemicals
How did it get this way?

• Data was assembled over many years (I assume), by various people and likely heterogeneous processes

• Cheminformatics tools of today are leaps and bounds ahead of tools from the 80s period.

• Data errors are COMMONPLACE! It took 3 years and a team of 6 people to clean up Wikipedia of the 8% errors!

• Just because its commonplace doesn’t mean we shouldn’t improve it!
Errors in Databases are Common Place

KEGG

DrugBank

DailyMed

Wolfram Alpha

Merck Index

Wikipedia

ChEBI

Common Chemistry
KNIME workflow to evaluate the dataset
## 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
QSARs for regulatory purposes

- Scientifically valid QSAR model
- Reliable QSAR results
- Adequate QSAR results
- QSAR model applicable to chemical query
- QSAR results relevant to regulatory purposes
The 5 OECD principles:
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 <strong>physicochemical, biological or environmental</strong> effect that can be measured and therefore modelled.</td>
</tr>
<tr>
<td>2) An unambiguous algorithm</td>
<td><strong>Ensure transparency</strong> in the description of the model algorithm.</td>
</tr>
<tr>
<td>3) A defined domain of applicability</td>
<td><strong>Define limitations</strong> in terms of the types of chemical structures, physicochemical properties and mechanisms of action for which the models can generate <strong>reliable predictions</strong>.</td>
</tr>
</tbody>
</table>
| 4) Appropriate measures of goodness-of-fit,    | a) The internal **fitting** performance of a model  
b) **predictivity** of a model, determined by using an appropriate **external test set**. |
| robustness and predictivity                    |                                                                                                                                              |
| 5) Mechanistic interpretation, if possible     | Mechanistic **associations** between the **descriptors** used in a model and the **endpoint being predicted**.                                |
## 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>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>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
Applicability Domain of original EPI Suite LogP Model

*accuracy of the global model is an issue*

Example 280 compound cluster outside applicability domain of EPI Suite compared to new model predictions
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.

• The program will extract the molecules names from the input csv or SDF (or assign arbitrary names if not) As IDs for 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
NCCT Model Predictions

• We want to provide WEB-access to the cleaned data, the predicted data for all chemicals and ultimately real time models. CLEAN data once and use many times!

• QSAR prediction models (kNN) produced for all properties

• 700k chemical structures pushed through NCCT_Models

• Each property now available on the dashboard
  – Original EPISuite predicted value where available
  – 3 and 4 star quality experimental data displayed
  – Predicted data displayed
### Chemical Properties: LogP

<table>
<thead>
<tr>
<th>Property</th>
<th>Raw Result</th>
<th>Mean Result</th>
<th>Minimum Result</th>
<th>Maximum Result</th>
<th>Result Unit</th>
<th>Result Type</th>
<th>Source</th>
</tr>
</thead>
<tbody>
<tr>
<td>Measured Log Kow</td>
<td>-3.4</td>
<td>-3.4</td>
<td>-3.4</td>
<td>-3.4</td>
<td></td>
<td>experimental</td>
<td>EPI SUITE</td>
</tr>
<tr>
<td>Octanol-water partition coefficient</td>
<td>-3.4</td>
<td>-3.4</td>
<td>-3.4</td>
<td>-3.4</td>
<td></td>
<td>experimental</td>
<td>CURATED_PHYSPRC</td>
</tr>
<tr>
<td>Estimated Log Kow</td>
<td>-4.47</td>
<td>-4.47</td>
<td>-4.47</td>
<td>-4.47</td>
<td></td>
<td>predicted</td>
<td>EPI SUITE</td>
</tr>
<tr>
<td>Octanol-water partition coefficient</td>
<td>-3.143</td>
<td>-3.144</td>
<td>-3.144</td>
<td>-3.144</td>
<td></td>
<td>predicted</td>
<td>NCCT_Models</td>
</tr>
</tbody>
</table>
### Chemical Properties: Melting Point

<table>
<thead>
<tr>
<th>Melting Point</th>
<th>245-250 °C</th>
<th>247.5</th>
<th>245.0</th>
<th>250.0</th>
<th>°C</th>
<th>experimental</th>
</tr>
</thead>
<tbody>
<tr>
<td>SynQuest</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Melting Point</th>
<th>245-250 °C</th>
<th>247.5</th>
<th>245.0</th>
<th>250.0</th>
<th>°C</th>
<th>experimental</th>
</tr>
</thead>
<tbody>
<tr>
<td>Alfa Aesar</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Melting Point</th>
<th>242-247 °C</th>
<th>244.5</th>
<th>242.0</th>
<th>247.0</th>
<th>°C</th>
<th>experimental</th>
</tr>
</thead>
<tbody>
<tr>
<td>Merck Millipore</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Melting Point</th>
<th>245-250 °C</th>
<th>247.5</th>
<th>245.0</th>
<th>250.0</th>
<th>°C</th>
<th>experimental</th>
</tr>
</thead>
<tbody>
<tr>
<td>SynQuest</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Melting Point</th>
<th>240.5 °C</th>
<th>240.5</th>
<th>240.5</th>
<th>240.5</th>
<th>°C</th>
<th>experimental</th>
</tr>
</thead>
<tbody>
<tr>
<td>Jean-Claude Bradl Open Melting Point Dataset</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Melting Point</th>
<th>248 °C</th>
<th>248.0</th>
<th>248.0</th>
<th>248.0</th>
<th>°C</th>
<th>experimental</th>
</tr>
</thead>
<tbody>
<tr>
<td>Jean-Claude Bradl Open Melting Point Dataset</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Melting Point</th>
<th>245 °C</th>
<th>245.0</th>
<th>245.0</th>
<th>245.0</th>
<th>°C</th>
<th>experimental</th>
</tr>
</thead>
<tbody>
<tr>
<td>Biosynth</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Melting Point</th>
<th>240.5 °C</th>
<th>240.5</th>
<th>240.5</th>
<th>240.5</th>
<th>°C</th>
<th>experimental</th>
</tr>
</thead>
<tbody>
<tr>
<td>CURATED_PHYS</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**Estimated MP (°C)**

<table>
<thead>
<tr>
<th>Estimated MP (°C)</th>
<th>176.41</th>
<th>176.4</th>
<th>176.4</th>
<th>176.4</th>
<th>°C</th>
<th>predicted</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>EPI SUITE</td>
</tr>
</tbody>
</table>

**Melting Point**

<table>
<thead>
<tr>
<th>Melting Point</th>
<th>247.68</th>
<th>247.7</th>
<th>247.7</th>
<th>247.7</th>
<th>°C</th>
<th>predicted</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>NCCT_Models</td>
</tr>
</tbody>
</table>
Reporting Outcome

• Presently authoring: “Examining the impact of Quality vs Quantity on Predictive Models for PhysChem”

• 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)
Small Data Sets limit Applicability Domain

• Some EPI Suite data sets are quite small…a few thousand points. There is a lot of Open Data now

• EXAMPLE: 200,000 melting points and 13,000 pyrolysis data points

The development of models to predict Melting and Pyrolysis point data associated with 300k compounds mined from patents

Igor V. Tetko,1,* Daniel Lowe2 and Antony J. Williams3

• Modeling this much data is a challenge!
Adding MORE Data

• Collect “Open Data” for various endpoints from:
  ➢ PubChem
  ➢ eChemPortal
  ➢ Open Data Sources

• Data already harvested and awaiting processing

• New models to be produced from expanded data sets
Conclusions

• Componentized architecture approach will be iterated as/if we hit limitations based on projects

• Much work to do with key activities on the API to access data and algorithms

• Our Focus is “Build Once, Use Many Times” – for our projects but made available to others to use also

• This will allow rapid prototyping as we progress

• We will rebuild existing prototypes on this new architecture based on priorities