Sharing chemical structures with peer-reviewed publications. Are we there yet?

Antony Williams

National Center for Computational Toxicology, U.S. Environmental Protection Agency, RTP, NC

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

March 2018
ACS Spring Meeting, New Orleans
Disclaimer of Endorsement

- Mention of or referral to commercial products or services, and/or links to non-EPA sites **does not imply official EPA endorsement** of or responsibility for the opinions, ideas, data, or products presented at those locations, or guarantee the validity of the information provided.
National Center for Computational Toxicology established in 2005 to integrate:

- High-throughput and high-content technologies
- Modern molecular biology
- Data mining and statistical modeling
- Computational biology and chemistry

Outputs: a lot of data, models, algorithms and software applications

Open Data – we want scientists to interrogate it, learn from it, develop understanding
We publish a lot...
And lots of chemistry...

<table>
<thead>
<tr>
<th>Publications</th>
<th>Research Project</th>
<th>Altmetric</th>
<th>PlumX</th>
<th>Kudos</th>
</tr>
</thead>
<tbody>
<tr>
<td>Patlewicz, Grace; Casati, Silvia; Basketter, David A.; Asturio, David; Roberts, David W.; Lepoittevin, Jean-Pierre; Worth, Andrew P.; Aschberger, Karin. Can currently available non-animal methods detect pre and pro-haptens relevant for skin sensitization?. <em>REGULATORY TOXICOLOGY AND PHARMACOLOGY</em>, 82(), (2016). doi:10.1016/j.yrtph.2016.08.007</td>
<td></td>
<td>🌈</td>
<td>🌈</td>
<td>🦋</td>
</tr>
</tbody>
</table>
The project I work on... https://comptox.epa.gov
We Curated Public Data to Create the Models – STANDARDS!

Public data should be curated prior to modeling

Covalent Halogens

Identical Chemicals

Mismatches
Workflow Details and Data

SAR and QSAR in Environmental Research
Volume 27, 2016 - Issue 11: 17th International Conference on QSAR in Environmental and Health Sciences (QSAR 2016) - Part II. Guest Editors: C.G. Barber and G.J. Myatt

Articles

An automated curation procedure for addressing chemical errors and inconsistencies in public datasets used in QSAR modelling

K. Mansouri, C. M. Grulke, A. M. Richard, R. S. Judson & A. J. Williams
Pages 911-937 | Received 03 Sep 2016, Accepted 24 Oct 2016, Published online: 25 Nov 2016

Download citation
http://dx.doi.org/10.1080/1062936X.2016.1253611

OPERA Models: https://github.com/kmansouri/OPERA
OPEN Models for Prediction

Model Performance with full QMRF

Nearest Neighbors from Training Set
OPERA models for predicting physicochemical properties and environmental fate endpoints

Kamel Mansouri\textsuperscript{1,2,3}, Chris M. Grulke\textsuperscript{1}, Richard S. Judson\textsuperscript{1} and Antony J. Williams\textsuperscript{1}

• The best way to publish our data as Supplementary Information Files?
Supplementary Information

- Supplementary information files generally Word Docs and Spreadsheets to PDF
- PDF conversions can be problematic. A recent example of interest

High-throughput, computer assisted, specific MetID. A revolution for drug discovery
## Table 1. Metabolites found in the different incubations tested

<table>
<thead>
<tr>
<th>Name</th>
<th>RT</th>
<th>m/z</th>
<th>Formula</th>
<th>m/z Diff (ppm)</th>
<th>Mass score</th>
<th>SMILES</th>
</tr>
</thead>
<tbody>
<tr>
<td>Parent</td>
<td>3.13</td>
<td>455.2926</td>
<td>C27H38N2O4</td>
<td>−3.48</td>
<td>429</td>
<td>N#CC(CCCNCN(C)CCc1cc(OC)c(OC)c1)(C(C)C)c2ccc(OC)c(OC)c2</td>
</tr>
<tr>
<td>M6 −164</td>
<td>2.26</td>
<td>291.2077</td>
<td>C17H26N2O2</td>
<td>−1.37</td>
<td></td>
<td>N(C)CCCCC(C#N)(C(C)C)c1ccc(OC)c(OC)c1</td>
</tr>
<tr>
<td>M16 −14</td>
<td>3.06</td>
<td>441.2743</td>
<td>C26H36N2O4</td>
<td>2.41</td>
<td>534</td>
<td>c1cc(CCNCCNC(C#N)(C(C)C)c2ccc(OC)c(OC)c2)cc(OC)c1OC</td>
</tr>
<tr>
<td>M14 +16</td>
<td>2.92</td>
<td>471.2866</td>
<td>C27H38N2O5</td>
<td>−1.59</td>
<td>476</td>
<td>N#CC(CCCNCN(C)CCc1ccc(OC)c(OC)c1)(C(C)C)c2ccc(OC)c(OC)c2</td>
</tr>
<tr>
<td>M9 −14</td>
<td>2.78</td>
<td>441.2761</td>
<td>C26H36N2O4</td>
<td>−1.7</td>
<td>590</td>
<td>C(#N)C(CCCNCN(C)CCc1ccc(OC)c(OC)c1)(C(C)C)c2ccc(O)c(OC)c2</td>
</tr>
<tr>
<td>M11 −14</td>
<td>2.84</td>
<td>441.2742</td>
<td>C26H36N2O4</td>
<td>2.57</td>
<td>473</td>
<td>Oc1ccc(CCN(C)CCCCC(C#N)(C(C)C)c2ccc(OC)c(OC)c2)cc1OC</td>
</tr>
<tr>
<td>M12 +2</td>
<td>2.87</td>
<td>457.2707</td>
<td>C26H36N2O5</td>
<td>−0.93</td>
<td>570</td>
<td>O(C)c1cc(ccc1OC)c(C#N)(CCNCCOC)c2ccc(OC)c(OC)c2c(C(C)C</td>
</tr>
<tr>
<td>M5 −178</td>
<td>2.2</td>
<td>277.1894</td>
<td>C16H24N2O2</td>
<td>7.84</td>
<td>419</td>
<td>C(C)(C)(C#N)(CCCN)c1ccc(OC)c(OC)c1</td>
</tr>
<tr>
<td>M8 +2</td>
<td>2.67</td>
<td>457.2708</td>
<td>C26H36N2O5</td>
<td>−1.18</td>
<td>581</td>
<td>OC(CN(C)CCCCC(C#N)(C(C)C)c1ccc(OC)c(OC)c1)c2ccc(O)c(OC)c2</td>
</tr>
<tr>
<td>M15 −14</td>
<td>2.92</td>
<td>441.2743</td>
<td>C26H36N2O4</td>
<td>2.23</td>
<td>614</td>
<td>Oc1ccc(CCN(C)CCCCC(C#N)(C(C)C)c2ccc(OC)c(OC)c2)cc1OC</td>
</tr>
<tr>
<td>M2 −259</td>
<td>0.73</td>
<td>196.1326</td>
<td>C11H17NO2</td>
<td>5.91</td>
<td>618</td>
<td>c1(CCNC)c(ccc(OC)c(OC)c1)OC</td>
</tr>
<tr>
<td>M10 −28</td>
<td>2.8</td>
<td>427.2617</td>
<td>C25H34N2O4</td>
<td>−4.79</td>
<td>487</td>
<td>c1(OC)c(ccc1OC)c(C#N)(CCNCCc2ccc(OC)c(OC)c2)C(C)C</td>
</tr>
<tr>
<td>M7 +2</td>
<td>2.46</td>
<td>457.2717</td>
<td>C26H36N2O5</td>
<td>−3.23</td>
<td>492</td>
<td>c1cc(CCN(O)CCCCC(C#N)(C(C)C)c2ccc(OC)c(OC)c2)cc(OC)c1OC</td>
</tr>
<tr>
<td>M17 +16</td>
<td>3.21</td>
<td>471.2853</td>
<td>C27H38N2O5</td>
<td>1.19</td>
<td>534</td>
<td>N#CC(CCCNCN(C)Ccc1ccc(OC)c(OC)c1)(c2ccc(OC)c(OC)c2)C(C)(O)O</td>
</tr>
<tr>
<td>M4 −178</td>
<td>1.86</td>
<td>277.1927</td>
<td>C16H24N2O2</td>
<td>−3.8</td>
<td>444</td>
<td>COc1cc(ccc1O)c(C#N)(CCNCC)C(C)C</td>
</tr>
<tr>
<td>M1 −289</td>
<td>0.44</td>
<td>166.0858</td>
<td>C9H11NO2</td>
<td>5.93</td>
<td>136</td>
<td>c1(CCNC)c(ccc(OC)c(OC)c1)=O</td>
</tr>
<tr>
<td>M13 −16</td>
<td>2.88</td>
<td>439.2603</td>
<td>C26H34N2O4</td>
<td>−1.43</td>
<td>367</td>
<td>c1cc(CcccNCN(C#N)(C(C)C)c2ccc(OC)c(OC)c2)cc(OC)c1OC</td>
</tr>
</tbody>
</table>
How do I extract structures?

• Copy and paste into Excel as a start point
• Assume no loss of formatting!
• Convert SMILES to structures
How do I extract structures?

- Copy and paste into Excel as a start point
- Assume no loss of formatting!
- Convert SMILES to structures
- But Copy-Paste doesn’t work

```
c1(CCNC)ccc(=O)c(c1)=O
```
```
c1cc(CC=NC)CCC(C#N)(C(C)C)2ccc(OC)c(OC)c2cc(OC)c1OC
```
```
c1(CCNC)ccc( 0)c(c1) 0
```
```
c1cc(CC NCCC(C#N)(C(C)C)2ccc(OC)c(OC)c2cc(OC)c1OC
```
How do I extract structures?

• Copy and paste into Excel as a start point
• Assume no loss of formatting!
• Convert SMILES to structures
• But Copy-Paste doesn’t work
### Table 2. Selection of fragments that help in the M16-16 metabolite structure elucidation

<table>
<thead>
<tr>
<th>Sub. obs. m/z</th>
<th>Sub. cal. m/z</th>
<th>Sub. m/z diff. ppm</th>
<th>Substrate</th>
<th>Metabolite</th>
<th>Δ</th>
<th>Met. obs. m/z</th>
<th>Met. calc. m/z</th>
<th>Met. m/z diff. ppm</th>
</tr>
</thead>
<tbody>
<tr>
<td>150.0664</td>
<td>150.0681</td>
<td>11.42</td>
<td>[Substrate Drawing]</td>
<td>[Metabolite Drawing]</td>
<td>+0</td>
<td>150.0670</td>
<td>150.0681</td>
<td>7.25</td>
</tr>
<tr>
<td>165.0869</td>
<td>165.0916</td>
<td>28.22</td>
<td>[Substrate Drawing]</td>
<td>[Metabolite Drawing]</td>
<td>+0</td>
<td>165.0892</td>
<td>165.0916</td>
<td>14.30</td>
</tr>
<tr>
<td>260.1637</td>
<td>260.1651</td>
<td>5.33</td>
<td>[Substrate Drawing]</td>
<td>[Metabolite Drawing]</td>
<td>+0</td>
<td>260.1652</td>
<td>260.1651</td>
<td>-0.50</td>
</tr>
</tbody>
</table>
Try hand-drawing Algal Toxins!
Think of files in multiple formats!

- SMILES are hyper-dependent on good layout algorithms. It’s not easy!
Names and CASRNs are NOT structures

- In our domain most chemicals are text – chemical names and CAS Numbers

<table>
<thead>
<tr>
<th>Name of Compound</th>
<th>CAS No.</th>
<th>Quantitation</th>
<th>Ion</th>
<th>Qualifier Ions</th>
</tr>
</thead>
<tbody>
<tr>
<td>Phenol-d6 (SS)</td>
<td>13187-88-3</td>
<td>99</td>
<td></td>
<td>71, 42</td>
</tr>
<tr>
<td>Phenol</td>
<td>108-95-2</td>
<td>94</td>
<td></td>
<td>66</td>
</tr>
<tr>
<td>1,4-Dichlorobenzene</td>
<td>106-46-0</td>
<td>146</td>
<td></td>
<td>111, 75, 50</td>
</tr>
<tr>
<td>Acetophenone</td>
<td>98-86-2</td>
<td>105</td>
<td></td>
<td>77, 51, 120</td>
</tr>
<tr>
<td>Acenaphthene-d10 (IS)</td>
<td>15067-26-2</td>
<td>162</td>
<td></td>
<td>160, 80</td>
</tr>
<tr>
<td>p-Cresol</td>
<td>106-44-5</td>
<td>107</td>
<td></td>
<td>108, 77</td>
</tr>
<tr>
<td>Isophorone</td>
<td>78-59-1</td>
<td>82</td>
<td></td>
<td>138, 54</td>
</tr>
<tr>
<td>Camphor</td>
<td>76-22-2</td>
<td>95</td>
<td></td>
<td>81, 108, 152</td>
</tr>
<tr>
<td>Isoborneol</td>
<td>124-76-5</td>
<td>95</td>
<td></td>
<td>110, 121, 136</td>
</tr>
<tr>
<td>Menthol</td>
<td>89, 78, 1</td>
<td>71</td>
<td></td>
<td>81, 123, 138</td>
</tr>
<tr>
<td>Naphthalene</td>
<td>91-20-3</td>
<td>128</td>
<td></td>
<td>102, 51</td>
</tr>
<tr>
<td>Methyl salicilat</td>
<td>119-36-8</td>
<td>120</td>
<td></td>
<td>92, 152, 65</td>
</tr>
</tbody>
</table>
And generally problematic...

<table>
<thead>
<tr>
<th>Name of Compound</th>
<th>CAS No.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Phenol-d6 (SS)</td>
<td>13187-88-3</td>
</tr>
<tr>
<td>Phenol</td>
<td>108-95-2</td>
</tr>
<tr>
<td>1,4-Dichlorobenzene</td>
<td>106-46-0</td>
</tr>
<tr>
<td>Acetophenone</td>
<td>98-86-2</td>
</tr>
<tr>
<td>Acenaphthene-d10 (IS)</td>
<td>15067-26-2</td>
</tr>
<tr>
<td>p-Cresol</td>
<td>106-44-5</td>
</tr>
<tr>
<td>Isophorone</td>
<td>78-59-1</td>
</tr>
<tr>
<td>Camphor</td>
<td>76-22-2</td>
</tr>
<tr>
<td>Isoborneol</td>
<td>124-76-5</td>
</tr>
<tr>
<td>Menthol</td>
<td>89, 78, 1</td>
</tr>
<tr>
<td>Naphthalene</td>
<td>91-20-3</td>
</tr>
<tr>
<td>Methyl salicilate</td>
<td>119-36-8</td>
</tr>
</tbody>
</table>
When publishing data consider..

• Licenses should be an important part of your data distribution considerations
• Digital Object Identifiers for versioning
• Consider your greater contribution to science outside of the publication

Findable Accessible Interoperable Reusable
Consider Distribution Repositories: Examples

- Zenodo
- Figshare
- Harvard Dataverse
- Open Science Framework
- Dryad
Repositories cover it all…

**CMAQ**

US EPA Office of Research and Development;

The Community Multiscale Air Quality (CMAQ) model is an active open-source development project of the U.S. EPA that consists of a suite of programs for conducting air quality model simulations. CMAQ combines current knowledge in atmospheric science and air quality modeling, multi-proc

Uploaded on February 6, 2018

*4 more version(s) exist for this record*

**VERSIONING**

**DOI**

10.5281/zenodo.1167892

**DOIs**

**Keyword(s):**

- EPA
- National Exposure Research Laboratory
- air quality model
- ozone
- particulate matter
- acid deposition
- toxics

**License (for files):**

Creative Commons Attribution 4.0

**LICENSING**
When we publish now...

- Generally store files on our FTP site PLUS copies in a repository (or two)
- Multiple formats of data as appropriate
  - Can be as supplementary data or DOI’ed data files
- DOI’ed data gives altmetrics also..
File(s) stored somewhere else

ftp://newftp.epa.gov/COMPTOX/NCCT_Publication_Data/Williams_A/Opera_Model_Paper/

Please note: Linked content is NOT stored on figshare and we can't guarantee its availability, quality, security or accept any liability.

This item is shared privately

OPERA Model Paper Data
06.03.2018, 19:33

Data associated with the OPERA Model paper authored by Kamel Mansouri and Tony Williams.
OPERA models for predicting physicochemical properties and environmental fate endpoints

Kamel Mansouri\textsuperscript{1,2,3}, Chris M. Grulke\textsuperscript{1}, Richard S. Judson\textsuperscript{1} and Antony J. Williams\textsuperscript{1}

Availability of data and materials

ftp://newftp.epa.gov/COMPTOX/Sustainable_Chemistry_Data/Chemistry_Dashboard/OPERA. The dataset supporting the conclusions of this article are available within the article and its additional file as well as OPERA’s Github repository (https://github.com/kmansouri/OPERA) and the EPA’s ftp site [150]: https://figshare.com/s/6fa1babbc9a0e9560317.
Thirty Years of Experience

• What has changed?
  – Cheminformatics has progressed
  – The internet proliferates data access
  – Standards have progressed (InChI) and are improving
  – Anybody can access/download millions of structures!

• What hasn’t changed?
  – Minor progress presenting structures in publications
  – Delivery via PDFs still dominates
  – Mandates on scientists are very unlikely
Antony Williams
US EPA Office of Research and Development
National Center for Computational Toxicology (NCCT)
Williams.Antony@epa.gov
ORCID: https://orcid.org/0000-0002-2668-4821