In Silico Prediction of Toxicology - One Can't Embrace Unembraceable

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

Praga, 4 September, marcus evans Predictive Toxicology 2009
Layout of presentation

Introduction:
• Why accuracy of prediction is important?

Methods:
• What is a Distance to Model? How can we estimate it? What is a property-based space?

Case study 1: Prediction of environmental toxicity
Case study 2: Benchmarking of lipophilicity (logP) predictions
Case study 3: AMES test prediction
Case study 4: CYP450 prediction
Case study 5: Prediction of in vivo acute rodent toxicity

Conclusions
Which common challenges do they face?
Possible: $10^{60} - 10^{100}$ molecules theoretically exist
( > $10^{80}$ atoms in the Universe)

Achievable: $10^{20} - 10^{24}$ can be synthesized now by companies (weight of the Moon is $10^{23}$ kg)

Available: $2 \times 10^{7}$ molecules are on the market

Measured: $10^{2} - 10^{5}$ molecules with ADME/T data

Problem: To predict ADME/T properties of just molecules on the market we must extrapolate data from one to 1,000 - 100,000 molecules!

There is a need for methods which can estimate the accuracy of predictions!
Representation of Molecules

Can be defined with calculated properties (logP, quantum-chemical parameters, etc.)

Can be defined with a set of structural descriptors (toxicophores, 2D, 3D, etc).

The descriptors are used to define the applicability domain.

Distance to model:
Examples of distances to models (DM) in descriptor space

1) Only two descriptors are used.

2) Colors refer to the same values.

3) More complex DMs (property-based DMs) also include the target property.²

The descriptor space challenge

We need to know the target property and select correct descriptors!
Property-based space illustration

*Do they agree in their votes (STD)?*
*Do they have the same pattern of votes (CORREL)?*
Morphinan-3-ol, 17-methyl-Levallorphan

STD - standard deviation of ensemble predictions

CORREL - correlation between vectors of predictions

1: Estimation of toxicity against *T. pyriformis*

The overall goal is to predict and to assess the reliability of predictions for toxicity against *T. pyriformis* for chemicals directly from their structure.

Dataset: 1093 molecules

CAse studies on the development and application of in-silico techniques for environmental hazard and risk assessment

Challenge (deadline is Sep. 10) is co-organized with the European Neural Network Society
### Analyzed QSARs (Quantitative Structure Activity Relationship) and distances to models (DM)

<table>
<thead>
<tr>
<th>country</th>
<th>modeling techniques</th>
<th>descriptors</th>
<th>abbreviation</th>
<th>distances to models (in space)</th>
<th>property-based</th>
</tr>
</thead>
<tbody>
<tr>
<td>ensemble of 192 kNN models</td>
<td>ensemble of 542 kNN models</td>
<td>MolconnZ</td>
<td>kNN-MZ</td>
<td>EUCLID</td>
<td>STD</td>
</tr>
<tr>
<td>SVM</td>
<td>SVM</td>
<td>Dragon</td>
<td>kNN-DR</td>
<td>EUCLID</td>
<td>STD</td>
</tr>
<tr>
<td>SVM</td>
<td>kNN</td>
<td>Dragon</td>
<td>SVM-MZ</td>
<td>SVM-DR</td>
<td></td>
</tr>
<tr>
<td>SVM</td>
<td>Fragments</td>
<td>SVM-FR</td>
<td>EUCLID, TANIMOTO</td>
<td></td>
<td></td>
</tr>
<tr>
<td>MLR</td>
<td>Fragments</td>
<td>MLR-FR</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>MLR</td>
<td>Molec. properties (CODESSA-Pro)</td>
<td>MLR-COD</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>OLS</td>
<td>Dragon</td>
<td>OLS-DR</td>
<td>LEVERAGE</td>
<td></td>
<td></td>
</tr>
<tr>
<td>PLS</td>
<td>Dragon</td>
<td>PLS-DR</td>
<td>LEVERAGE</td>
<td>PLSEU</td>
<td></td>
</tr>
<tr>
<td>ensemble of 100 neural networks</td>
<td></td>
<td>E-state indices</td>
<td>ASNN-ESTATE</td>
<td>CORREL, STD</td>
<td></td>
</tr>
<tr>
<td>All consensus model</td>
<td></td>
<td>-</td>
<td>CONS</td>
<td>STD</td>
<td></td>
</tr>
</tbody>
</table>

Descriptor space: DM **does not work**

Property-based space: DM does work!

STD

# Ranking of Distance to Models (DM)

<table>
<thead>
<tr>
<th>DM</th>
<th>average rank</th>
<th>highest rank$^1$</th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>LOO</td>
<td>5-CV</td>
<td>Valid.*</td>
<td>LOO</td>
</tr>
<tr>
<td>STD-CONS</td>
<td>1</td>
<td>1.8</td>
<td>1.1</td>
<td>12</td>
</tr>
<tr>
<td>STD-ASNN</td>
<td>2</td>
<td>1.2</td>
<td>2.5</td>
<td>10</td>
</tr>
<tr>
<td>STD-kNN-DR</td>
<td>6.6</td>
<td>4.3</td>
<td>4.1</td>
<td></td>
</tr>
<tr>
<td>STD-kNN-MZ</td>
<td>9.2</td>
<td>8.3</td>
<td>5.3</td>
<td></td>
</tr>
<tr>
<td>EUCLID-kNN-DR</td>
<td>7.1</td>
<td>4.9</td>
<td>5.4</td>
<td></td>
</tr>
<tr>
<td>LEVERAGE-PLS</td>
<td>8.4</td>
<td>5</td>
<td>6.3</td>
<td></td>
</tr>
<tr>
<td>EUCLID-kNN-MZ</td>
<td>7.5</td>
<td>7.1</td>
<td>6.4</td>
<td></td>
</tr>
<tr>
<td>TANIMOTO-kNN-FR</td>
<td>7</td>
<td>6.1</td>
<td>6.8</td>
<td></td>
</tr>
<tr>
<td>TANIMOTO-MLR-FR</td>
<td>8.3</td>
<td>8.3</td>
<td>9</td>
<td></td>
</tr>
<tr>
<td>CORREL-ASNN</td>
<td>10.7</td>
<td>10.8</td>
<td>9.4</td>
<td></td>
</tr>
<tr>
<td>LEVERAGE-OLS-DR</td>
<td>12.3</td>
<td>12.6</td>
<td>11.1</td>
<td></td>
</tr>
<tr>
<td>EUCLID-MLR-FR</td>
<td>7</td>
<td>9.3</td>
<td>11.5</td>
<td></td>
</tr>
<tr>
<td>PLSEU-PLS</td>
<td>11.1</td>
<td>11.8</td>
<td>11.5</td>
<td></td>
</tr>
<tr>
<td>EUCLID-kNN-FR</td>
<td>12.1</td>
<td>13.3</td>
<td>12.1</td>
<td></td>
</tr>
</tbody>
</table>

*Ordered by performance of the DMs on the validation dataset

Accuracy of toxicity prediction against *T. pyriformis* for training and two industrial sets

**Left side:** error of ASNN estimation, individual points

**Right side:** % of molecules predicted with the error estimated by the black line

High Production Volume - HPV (USA-EPA): 3182

EINECS (REACH): 48774

**Warning:** using the available data one can reliably predict only few % molecules from the industry related datasets!

2: Benchmarking of logP calculators

Existing Dogma:

- Prediction of physico-chemical properties, in particular log P, is simple
- There is no need to measure them
- We have enough number of good computational methods

Is this true?
Data & background models

18 methods (major commercial providers and public software)

_in house_ data:
95809 molecules from Prizer
889 molecules from Nycomed

**Arithmetic Average Model (AAM):**
mean logP was used as a prediction (one value for all molecules)

**Rank III:** models with errors ($RMSE \geq AAM$), i.e. non-predictive
**Rank I:** models with $RMSE$ identical or close to the best method
**Rank II:** remaining models
## Benchmarking of logP methods for in-house data of Pfizer & Nycomed

<table>
<thead>
<tr>
<th>Method</th>
<th>Pfizer set (N = 95 809)</th>
<th>Nycomed set (N = 882)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>RMSE</td>
<td>% in error range</td>
</tr>
<tr>
<td></td>
<td></td>
<td>&lt;0.5</td>
</tr>
<tr>
<td>Consensus logP</td>
<td>0.95</td>
<td>48</td>
</tr>
<tr>
<td>ALOGPS</td>
<td>1.02</td>
<td>41</td>
</tr>
<tr>
<td>S+logP</td>
<td>1.02</td>
<td>44</td>
</tr>
<tr>
<td>NC+NHET</td>
<td>1.04</td>
<td>38</td>
</tr>
<tr>
<td>MLOGP(S+)</td>
<td>1.05</td>
<td>40</td>
</tr>
<tr>
<td>XLOGP3</td>
<td>1.07</td>
<td>43</td>
</tr>
<tr>
<td>MiLogP</td>
<td>1.10</td>
<td>41</td>
</tr>
<tr>
<td>AB/LogP</td>
<td>1.12</td>
<td>39</td>
</tr>
<tr>
<td>ALOGP</td>
<td>1.12</td>
<td>39</td>
</tr>
<tr>
<td>ALOGP98</td>
<td>1.12</td>
<td>40</td>
</tr>
<tr>
<td>OsirisP</td>
<td>1.13</td>
<td>39</td>
</tr>
<tr>
<td>AAM</td>
<td>1.16</td>
<td>33</td>
</tr>
<tr>
<td>CLOGP</td>
<td>1.23</td>
<td>37</td>
</tr>
<tr>
<td>ACD/logP</td>
<td>1.28</td>
<td>35</td>
</tr>
<tr>
<td>CSlogP</td>
<td>1.29</td>
<td>37</td>
</tr>
<tr>
<td>COSMOFrag</td>
<td>1.30</td>
<td>32</td>
</tr>
<tr>
<td>QikProp</td>
<td>1.32</td>
<td>31</td>
</tr>
<tr>
<td>KowWIN</td>
<td>1.32</td>
<td>33</td>
</tr>
<tr>
<td>QLogP</td>
<td>1.33</td>
<td>34</td>
</tr>
<tr>
<td>XLOGP2</td>
<td>1.80</td>
<td>15</td>
</tr>
<tr>
<td>MLOGP(Dragon)</td>
<td>2.03</td>
<td>34</td>
</tr>
</tbody>
</table>

Large number of methods could not perform better than the **AAM** model!

Catastrophe!?
ALOGPS 2.1

- LogP: 75 variables, 12908 molecules, RMSE=0.35, MAE=0.26

- LogS: 33 variables, 1291 molecules, RMSE=0.49, MAE=0.35


ALOGPS self-learns new data to cover new scaffolds

\[ N = 95809 \] (in house Pfizer data)

ALOGPS Blind prediction

![Graph showing ALOGPS Blind prediction](image)

RMSE = 1.02

ca 30 minutes of calculations on a notebook!

ALOGPS LIBRARY

![Graph showing ALOGPS LIBRARY](image)

RMSE = 0.59

ALOGPS distinguishes reliable vs. non-reliable predictions in property-based space (CORREL)

CORREL identifies 60% of molecules predicted with average accuracy of 0.3 log units
ALOGPS dramatically improves accuracy

![Bar chart showing calculated accuracy vs. estimated accuracy for blind predictions.](image)

Only reliable predictions (and we can distinguish them!) have much higher accuracy.
3: Prediction of Ames Mutagenicity set

http://ml.cs.tu-berlin.de/toxbenchmark
Toxicity against *Salmonella typhimurium*

Data set: 4361 molecules
67% with mutagenic effect (background model)

Large international collaboration effort of 13 labs from USA, Canada, EU, Russia, Ukraine & China

Prof. Bruce N. Ames
Inventor of the test (1975)
Associative Neural Network analysis of Ames set

Only reliable predictions (15% of all data points) are 22%/5% = 4 times more accurate!

*Coverage of model
4: Prediction of CYP450 1A2 inhibitors


Bioassay AID 410

One of the test performed within NIH Roadmap

4177 active molecules
3680 inactive molecules

53% were inhibitors of CYP (background accuracy)

Dr. Elias Zerhouni
Former NIH director (2002-2008)

The most reliable predictions (30% of all molecules) are 21%/5% = 4 times more accurate!
5: *In vivo* rodent toxicity (ZEBET database\(^1\))

- **361 compounds**
  - Cytotoxicity IC50 and both rat and/or mouse LD50

- **291 compounds**
  - Inorganics, mixtures and heavy metal salts are removed

- **253 compounds**
  - Both in vitro IC50 values and rat LD50 results

**Random split**

- **230 compounds**
  - Modeling set
- **23 compounds**
  - Validation set

ZEBET - The national center for documentation and evaluation of alternative methods to animal experiments

Poor in vitro-in vivo Correlation Between IC50 and Rat LD50 Values

Two steps model: first classify and then predict!
Use of applicability domain increased accuracy of prediction for the new compounds

Table 3. Comparison between TOPKAT and the two-step model prediction of the external compounds.

<table>
<thead>
<tr>
<th>Measure</th>
<th>No applicability domain</th>
<th>With applicability domain</th>
<th>No applicability domain</th>
<th>With applicability domain</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>Two-step model</td>
<td></td>
<td>TOPKAT</td>
</tr>
<tr>
<td>Prediction of 27 new ZEBET compounds</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$R^2$</td>
<td>0.64</td>
<td>0.85</td>
<td>0.16</td>
<td>0.60</td>
</tr>
<tr>
<td>MAE</td>
<td>0.38</td>
<td>0.29</td>
<td>0.78</td>
<td>0.50</td>
</tr>
<tr>
<td>Coverage (%)</td>
<td>100</td>
<td>67</td>
<td>100</td>
<td>67</td>
</tr>
<tr>
<td>Prediction of 1,562 RTECS compounds with 70% confidence level</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$R^2$</td>
<td>0.26</td>
<td>0.33</td>
<td>0.19</td>
<td>0.22</td>
</tr>
<tr>
<td>MAE</td>
<td>0.65</td>
<td>0.54</td>
<td>0.76</td>
<td>0.65</td>
</tr>
<tr>
<td>Coverage (%)</td>
<td>100</td>
<td>62</td>
<td>100</td>
<td>62</td>
</tr>
<tr>
<td>Prediction of 1,562 RTECS compounds with 90% confidence level</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$R^2$</td>
<td>0.42</td>
<td>0.62</td>
<td>0.19</td>
<td>0.26</td>
</tr>
<tr>
<td>MAE</td>
<td>0.60</td>
<td>0.42</td>
<td>0.84</td>
<td>0.68</td>
</tr>
<tr>
<td>Coverage (%)</td>
<td>12</td>
<td>6</td>
<td>12</td>
<td>6</td>
</tr>
</tbody>
</table>

Registry of Toxic Effects of Chemical Compounds (RTECS)
Developed methodology allows navigation in space of molecules with a confidence and:

✓ to develop targeted (local) models to cover specific series.
✓ to reliably estimate which compounds can/can’t be reliably predicted.
✓ to provide experimental design and to minimize costs of new measurements.

☐ This is our expertise and “know-how” that we are applying to new data.
Acknowledgements

My group
Mr I. Sushko
Mr S. Novotarskyi
Mr A.K. Pandey
Mr R. Körner
Mr S. Brandmaier
Dr. M. Rupp

Visiting Scientists
Dr V. Kovalishyn
Dr V. Prokopenko

Collaborators:
Dr. G. Poda (Pfizer)
Dr. C. Ostermann (Nycomed)
Dr. C. Höfer (DMPKore)
Prof. A. Tropsha (NC, USA)
Prof. T. Oprea (New Mexico, USA)
Prof. A. Varnek (Strasbourg, France)
Prof. R. Mannhold (Düsseldorf, Germany)
+ 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