Critical assessment of QSAR models to predict environmental toxicity against *T. pyriformis*: Applicability domain and overfitting by variable selection

I.V. Tetko,¹ I. Sushko,¹ A.K. Pandey,¹ H. Zhu,² A. Tropsha,² E. Papa,³ T. Öberg,⁴ R. Todeschini,⁵ D. Fourches,⁶ A. Varnek⁶

¹Helmholtz Zentrum München (Germany), ²University of North Carolina (USA), ³University of Insubria (Italy), ⁴University of Kalmar (Sweden), ⁵University of Milano-Bicocca (Italy) and ⁶Louis Pasteur University (France)
REACH

Registration, Evaluation, Authorisation and Restriction of Chemical substances

European Chemicals Agency (ECHA) in Helsinki
REACH and QSAR (Quantitative Structure Activity Relationship) models

> 30,000 chemicals to be registered ... is a lot!

It is expensive to measure all of them ($200,000 per compound), a lot of animal testing

QSAR models can be used to prioritize compounds

- Compound is predicted to be toxic
  - Biological testing will be done to prove/disprove the models
- Compound is predicted to be not toxic
  - tests can be avoided, saving money, animals
  - but ... only if we are confident in the predictions
Requirements of biological testing following QSAR model prediction

<table>
<thead>
<tr>
<th>model prediction</th>
<th>prediction confidence</th>
</tr>
</thead>
<tbody>
<tr>
<td>non-toxic, IC50 &lt; LIMIT</td>
<td>low need</td>
</tr>
<tr>
<td></td>
<td>low need (depends on other properties)</td>
</tr>
<tr>
<td>toxic, IC50 &gt; LIMIT</td>
<td>strong need (depends on other properties)</td>
</tr>
<tr>
<td></td>
<td>moderate need</td>
</tr>
</tbody>
</table>

Acceptance of decisions will be more accurate if confidence intervals (prediction errors) are known and are taken into analysis: concept of applicability domain.
Declining R&D productivity in the pharmaceutical industry

Dwindling R&D Productivity in the Pharmaceutical Industry

Source: PhRMA 2007, FDA
Reasons for failure in drug development

> 60% of drug failures are due to absorption, distribution, metabolism, excretion and **toxicology** (ADME/T) problems
"One can not embrace the unembraceable."

**Possible:** $10^{60} - 10^{100}$ molecules theoretically exist
(> $10^{80}$ atoms in the Universe)

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

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

**Measured:** $10^{2} - 10^{4}$ molecules with measured 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!

Both environmental & health sciences have similar problems!

**Ionic Liquids ca $10^{18}$ (Prof. Jastorff)**

Methods that can estimate the accuracy of predictions are required.
Models can fail due to chemical diversity of training & test sets

Training set data used to develop a model

Our model given the training set

Correct model

New data to be estimated
It is easy to build a QSAR model

but it is much more difficult to estimate its accuracy for new data
Representation of Molecules for Quantitative Structure-Activity Relationship (QSAR)

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

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

One of these sets of descriptors is usually used for determination the applicability domain of models.

Distance to model:
Goals of this study

• Develop new models for prediction of environmental toxicity against *T. pyriformis*

• Benchmark different applicability domains (*distances to models*)

• Is accuracy of predictions limited by the approach or by the data themselves?

• Is there a best (“universal”) AD?
Estimation toxicity of *T. pyriformis*

Initial Dataset\(^1,2\)

\(n=983\) molecules

\(n=644\) training set

\(n=339\) test set 1

Test set 2:

\(n=110\) molecules\(^1,2\)

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


Overview of analyzed QSAR approaches and distances to models

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

# Overview of analyzed distances to models (DMs)

<table>
<thead>
<tr>
<th>EUCLID</th>
<th>$$EU_m = \frac{1}{m} \sum_{i=1}^{k} d_j$$</th>
<th>k is number of nearest neighbors, m index of model</th>
</tr>
</thead>
<tbody>
<tr>
<td>EUCLID</td>
<td>$$EU = \frac{1}{k} \sum_{j=1}^{k} d_j$$</td>
<td></td>
</tr>
<tr>
<td>TANIMOTO</td>
<td>$$\text{Tanimoto}(a,b) = \frac{\sum x_{a,i} x_{b,i} + \sum x_{a,i} - \sum x_{b,i} - \sum x_{a,i} x_{b,i}}{\sum x_{a,i} + \sum x_{b,i} - \sum x_{a,i} x_{b,i}}$$</td>
<td>x_{a,i} and x_{b,i} are fragment counts</td>
</tr>
<tr>
<td>LEVERAGE</td>
<td>$$\text{LEVERAGE} = x^T(X^TX)^{-1}x$$</td>
<td></td>
</tr>
<tr>
<td>PLSEU (DModX)</td>
<td>Error in approximation (restoration) of the vector of input variables from the latent variables and PLS weights.</td>
<td></td>
</tr>
<tr>
<td>STD</td>
<td>$$\text{STD} = \frac{1}{N-1} \sum (y_i - \bar{y})^2$$</td>
<td></td>
</tr>
<tr>
<td>y_i is value calculated with model i and (\bar{y}) is average value</td>
<td></td>
<td></td>
</tr>
<tr>
<td>CORREL</td>
<td>$$\text{CORREL}(a) = \max_j \text{CORREL}(a,j) = R^2(Y_{\text{calc}}^a, Y_{\text{calc}}^j)$$</td>
<td></td>
</tr>
<tr>
<td>Y^a=(y_1,\ldots,y_N) is vector of predictions of molecule i</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Analysis of two simulated datasets

A) Errors do not depend on the distance to model (DM)

B) Errors depend on the DM
Mixture of Gaussian Distributions (MGD)

Idea is to find a MGD, which maximize likelihood (probability)

\[ \Pi N(0, \sigma^2(e_i)) \]

of the observed distribution of errors

\[
N(0, \sigma^2(e)) = \frac{1}{\sigma \sqrt{2\pi}} \exp \left( -\frac{e^2}{2\sigma^2} \right) \\
S(G_g) = \Sigma \log N(0, \sigma^2(e_i))
\]
MGDs for the simulated datasets

A) Non significant MGD was found

B) A MGD composed of 3 Gaussian distributions was found
Analysis of DMs for a linear model

\[
\log(\text{IGC}_{50}^{-1}) = -18(\pm0.7) + 0.065(\pm0.002)\text{AMR} - 0.50(0.04)\text{O56} - 0.30(0.03)\text{O58} - 0.29(0.02)\text{nHAcc} + 0.046(0.005)\text{H-046} + 16(0.7)\text{Me}
\]

The use of various DM provides different discrimination of molecules with low and large errors.

Performances of MGDs calculated with different definitions of Distance to Models (DM)

<table>
<thead>
<tr>
<th>DM</th>
<th>average rank</th>
<th>highest rank</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>LOO</td>
<td>5-CV</td>
</tr>
<tr>
<td>STD-CONS</td>
<td>1</td>
<td>1.8</td>
</tr>
<tr>
<td>STD-ASNN</td>
<td>2</td>
<td>1.2</td>
</tr>
<tr>
<td>STD-kNN-DR</td>
<td>6.6</td>
<td>4.3</td>
</tr>
<tr>
<td>STD-kNN-MZ</td>
<td>9.2</td>
<td>8.3</td>
</tr>
<tr>
<td>EUCLID-kNN-DR</td>
<td>7.1</td>
<td>4.9</td>
</tr>
<tr>
<td>LEVERAGE-PLS</td>
<td>8.4</td>
<td>5</td>
</tr>
<tr>
<td>EUCLID-kNN-MZ</td>
<td>7.5</td>
<td>7.1</td>
</tr>
<tr>
<td>TANIMOTO-kNN-FR</td>
<td>7</td>
<td>6.1</td>
</tr>
<tr>
<td>TANIMOTO-MLR-FR</td>
<td>8.3</td>
<td>8.3</td>
</tr>
<tr>
<td>CORREL-ASNN</td>
<td>10.7</td>
<td>10.8</td>
</tr>
<tr>
<td>LEVERAGE-OLS-DR</td>
<td>12.3</td>
<td>12.6</td>
</tr>
<tr>
<td>EUCLID-MLR-FR</td>
<td>7</td>
<td>9.3</td>
</tr>
<tr>
<td>PLSEU-PLS</td>
<td>11.1</td>
<td>11.8</td>
</tr>
<tr>
<td>EUCLID-kNN-FR</td>
<td>12.1</td>
<td>13.3</td>
</tr>
</tbody>
</table>

*Ordered by performance of the DMs on the validation dataset

# Standard Deviation of Models (STD)

<table>
<thead>
<tr>
<th>Country</th>
<th>Modeling Techniques</th>
<th>Descriptors</th>
<th>Abbreviation</th>
<th>STD</th>
</tr>
</thead>
<tbody>
<tr>
<td>(UNC)</td>
<td>kNN ensemble</td>
<td>MolconnZ</td>
<td>kNN-MZ</td>
<td>1.12</td>
</tr>
<tr>
<td></td>
<td>kNN ensemble</td>
<td>Dragon</td>
<td>kNN-DR</td>
<td>1.02</td>
</tr>
<tr>
<td></td>
<td>SVM</td>
<td>MolconnZ</td>
<td>SVM-MZ</td>
<td>0.97</td>
</tr>
<tr>
<td></td>
<td>SVM</td>
<td>Dragon</td>
<td>SVM-DR</td>
<td>0.91</td>
</tr>
<tr>
<td>(ULP)</td>
<td>SVM</td>
<td>Fragments</td>
<td>SVM-FR</td>
<td>0.88</td>
</tr>
<tr>
<td></td>
<td>kNN</td>
<td>Fragments</td>
<td>kNN-FR</td>
<td>0.95</td>
</tr>
<tr>
<td></td>
<td>MLR</td>
<td>Fragments</td>
<td>MLR-FR</td>
<td>0.99</td>
</tr>
<tr>
<td></td>
<td>MLR</td>
<td>CODESSA-Pro</td>
<td>MLR-COD</td>
<td>1.14</td>
</tr>
<tr>
<td>(UI)</td>
<td>OLS</td>
<td>Dragon</td>
<td>OLS-DR</td>
<td>1.06</td>
</tr>
<tr>
<td>(UK)</td>
<td>PLS</td>
<td>Dragon</td>
<td>PLS-DR</td>
<td>1.08</td>
</tr>
<tr>
<td>(HMGU)</td>
<td>neural networks</td>
<td>E-state indices</td>
<td>ASNN-ESTATE</td>
<td>1.10</td>
</tr>
<tr>
<td>consensus (average)</td>
<td></td>
<td></td>
<td>CONS</td>
<td>1.02</td>
</tr>
<tr>
<td>STD</td>
<td></td>
<td></td>
<td>STD-CONS</td>
<td>0.09</td>
</tr>
</tbody>
</table>
Errors using MGD & STD distance to models
Estimations of errors using STD distance to models

- RMSE

- Set I
- Predicted I
- Set II
- Predicted II
Estimations based on training set errors calculated with incorrect validation protocol

Prediction of data from the training and two external database

Experimental accuracy:

Estimated experimental accuracy:¹

\[ SE = 0.38 \] reactive and

\[ SE = 0.21 \] narcosis mechanism of action

Sustainable or Green Chemistry

Twelve Principles

- Prevent waste
- Design safer chemicals and products
- Use renewable feedstocks
- Use catalysts, not stoichiometric reagents
- Avoid chemical derivatives
- Maximize atom economy
- Use safer solvents and reaction conditions
- Increase energy efficiency
- Design chemical and products to degrade after use
- Analyze in real time to prevent pollution
- Minimize the potential for accidents
QSAR for Sustainable or Green Chemistry

Twelve Principles

• Prevent waste
  ✓ Design safer chemicals and products
• Use renewable feedstocks
• Use catalysts, not stoichiometric reagents
• Avoid chemical derivatives
• Maximize atom economy
• Use safer solvents and reaction conditions
• Increase energy efficiency
  ✓ Design chemical and products to degrade after use
• Analyze in real time to prevent pollution
• Minimize the potential for accidents
Conclusions

• Development of green chemistry (environmental sciences) and discovery of drugs (health sciences) share similar problems

• The use of QSAR approaches can help to identify toxic/non-toxic compounds before start of their commercial exploitation in chemical industry or clinical testing in the drug discovery

• Data (diversity, accuracy) but not the methods dominate in determination of the accuracy of model predictions

• The standard deviation of models provided the best discrimination of molecules with low and high prediction accuracy

• Models are available at http://www.qspr.org (in development)

• Models can reliably predict only small % of molecules from the REACH-like database
Do you need more information?

http://www.vcclab.org

http://www.qspr.eu*

Acknowledgements

All collaborators

• Ester Papa
• Tomas Öberg
• Roberto Todeschini
• Alexander Tropsha & Hao Zhu
• Alexandre Varnek & Denis Fourches

Paola Gramatica
Mark Hewitt
Mark Cronin

Our team

• Iurii Sushko
• Robert Koerner
• Sergii Novatarskyi
• Anil Kumar Pandey

Terry Schultz

This study has been partially supported with BMBF GoBio and FP7 CADASTER projects.

Thank you very much for your attention!