Modeling quantitative acute oral systemic toxicity based on a k-Nearest Neighbor (k-NN) algorithm

Alessandra Roncaglioni, Cosimo Toma, Giovanna Lavado, Kristijan Vukovic, Domenico Gadaleta, and Emilio Benfenati

Istituto di Ricerche Farmacologiche “Mario Negri”
via G. La Masa 19
Milan - Italy
Introduction

Laboratory of Environmental Chemistry and Toxicology

- Modeling exercise based on the dataset released by the NICEATM team
- 2 months, 5 people involved using different tools and algorithms (more results on posters)
- Common internal procedure for the dataset curation
Data preparation: \( LD_{50} \) values

### Aggregation of entries
- De-salting
- Duplicated structures

### Data variability analysis
- Removed entries:
  - with SD ≥ 0.5 for data point values
  - Conflicting classes

### Final dataset: 8476 unique structures
- Median 1st quantile \( LD_{50} \) (mmol/kg)
- Univocal label for classification

<table>
<thead>
<tr>
<th>Count (single point)</th>
<th>Canonical_QSARr</th>
<th>Salt_Solvent</th>
<th>Mean</th>
<th>1st Quantile median</th>
<th>Standard Deviation</th>
</tr>
</thead>
<tbody>
<tr>
<td>4</td>
<td>COP(=S)(=OC)SCN1N=NC2C=CC=CC=2C1=O</td>
<td>Cl(1), C(C)(=O)O(1)</td>
<td>-0.71</td>
<td>-1.84</td>
<td>1.90</td>
</tr>
<tr>
<td>4</td>
<td>NC1CCCCCC1</td>
<td></td>
<td>0.25</td>
<td>-0.95</td>
<td>0.87</td>
</tr>
<tr>
<td>2</td>
<td>CC[P+]C1C=CC=CC=1)(C1C=CC=CC=1)C1C=CC=CC=1</td>
<td><a href="1">Br-</a>, <a href="1">I-</a></td>
<td>-0.50</td>
<td>-0.72</td>
<td>0.32</td>
</tr>
<tr>
<td>5</td>
<td>NCCO</td>
<td></td>
<td>1.72</td>
<td>1.45</td>
<td>0.30</td>
</tr>
</tbody>
</table>
Dataset preparation

- Training and validation set splitting (80%/20%)
- Indigo fingerprints calculated in KNIME and compared using the Tanimoto score
- Stratified sampling based on clusters obtained with the k-Means clustering algorithm
- Nearly-common dataset to all the 5 modeled endpoints
Data distribution

<table>
<thead>
<tr>
<th></th>
<th>n</th>
<th>Min logLD50 (mmol/kg)</th>
<th>Max logLD50 (mmol/kg)</th>
<th>Mean</th>
<th>St.dev</th>
<th>Skewness</th>
</tr>
</thead>
<tbody>
<tr>
<td>TS</td>
<td>5029</td>
<td>-4.66</td>
<td>2.71</td>
<td>0.44</td>
<td>0.90</td>
<td>-0.98</td>
</tr>
<tr>
<td>VS</td>
<td>1251</td>
<td>-3.70</td>
<td>2.38</td>
<td>0.44</td>
<td>0.90</td>
<td>-1.05</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th></th>
<th>%</th>
<th></th>
<th></th>
<th>%</th>
<th></th>
<th></th>
<th>%</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>GHS 1</td>
<td>0.02</td>
<td></td>
<td>EPA 1</td>
<td>0.09</td>
<td></td>
<td>nT</td>
<td>0.43</td>
<td></td>
<td></td>
</tr>
<tr>
<td>GHS 2</td>
<td>0.06</td>
<td></td>
<td>EPA 2</td>
<td>0.22</td>
<td></td>
<td>T</td>
<td>0.57</td>
<td></td>
<td></td>
</tr>
<tr>
<td>GHS 3</td>
<td>0.13</td>
<td></td>
<td>EPA 3</td>
<td>0.50</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>GHS 4</td>
<td>0.35</td>
<td></td>
<td>EPA 4</td>
<td>0.19</td>
<td></td>
<td>not vT</td>
<td>0.08</td>
<td></td>
<td></td>
</tr>
<tr>
<td>GHS 5</td>
<td>0.44</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

BALANCING
Modeling approach

- K-nn algorithm as implemented into in-house software (istKNN by A. Manganaro, Kode srl)

- Simple method, internal benchmark for other more sophisticated modeling techniques (see posters)

- Allows a read-across like approach
Similarity Index (SI)

- accounting for different chemical and structural features of the molecules
- considering different relevant structural aspects, including the size of the molecules

\[ SI = S(FP)^{0.4} \times S(CD)^{0.35} \times S(HE)^{0.1} \times S(FG)^{0.15} \]

FP $\rightarrow$ Fingerprints

CD $\rightarrow$ Structural key with 35 Constitutional descriptors (MW, nr of skeleton atoms, etc.)

HE $\rightarrow$ Structural key with 11 Hetero-atoms descriptors

FG $\rightarrow$ Structural key with 154 Functional groups (specific chemical moieties)

- FP similarity $\rightarrow$ Maxwell-Pilliner index
- CD, HE, FG similarities $\rightarrow$ Bray-Curtis index

Software setting

Batch model development

> 300 models optimizing 5 parameters

Batch models on training set
Select best setting with $R^2_{\text{LOO}}$ and check VS with best model
Use this setting to run a new model on the entire dataset
## Results

### Model parameters

<table>
<thead>
<tr>
<th>Model No.</th>
<th>K</th>
<th>Min Similarity</th>
<th>Min Similarity for single molecule</th>
<th>Enhance factor</th>
<th>Exper. range</th>
<th>R² (LOO)</th>
<th>RMSE (LOO)</th>
<th>Unpred. rate</th>
</tr>
</thead>
<tbody>
<tr>
<td>135</td>
<td>3</td>
<td>0.8</td>
<td>0.85</td>
<td>3</td>
<td>2</td>
<td>0.56</td>
<td>0.59</td>
<td>13.2%</td>
</tr>
</tbody>
</table>

### Training Set (TS)

<table>
<thead>
<tr>
<th>Model No.</th>
<th>K</th>
<th>Min Similarity</th>
<th>Min Similarity for single molecule</th>
<th>Enhance factor</th>
<th>Exper. range</th>
<th>R² (LOO)</th>
<th>RMSE (LOO)</th>
<th>Unpred. rate</th>
</tr>
</thead>
<tbody>
<tr>
<td>135</td>
<td>3</td>
<td>0.8</td>
<td>0.85</td>
<td>3</td>
<td>2</td>
<td>0.56</td>
<td>0.59</td>
<td>13.2%</td>
</tr>
</tbody>
</table>

### Validation Set (VS)

<table>
<thead>
<tr>
<th>Model No.</th>
<th>K</th>
<th>Min Similarity</th>
<th>Min Similarity for single molecule</th>
<th>Enhance factor</th>
<th>Exper. range</th>
<th>R² (LOO)</th>
<th>RMSE (LOO)</th>
<th>Unpred. rate</th>
</tr>
</thead>
<tbody>
<tr>
<td>135</td>
<td>3</td>
<td>0.8</td>
<td>0.85</td>
<td>3</td>
<td>2</td>
<td>0.56</td>
<td>0.59</td>
<td>13.2%</td>
</tr>
</tbody>
</table>

### TS + VS

<table>
<thead>
<tr>
<th>R² (LOO)</th>
<th>RMSE (LOO)</th>
<th>Unpred. rate</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.583</td>
<td>0.575</td>
<td>11.9%</td>
</tr>
</tbody>
</table>

### Graphs

- **R² (LOO) vs. K**
- **R² (LOO) vs. Min Similarity**
- **Unpredicted rate on training set (%) vs. Unpred. rate**
Results on the evaluation set

<table>
<thead>
<tr>
<th>Evaluation set (n = 1865)</th>
</tr>
</thead>
<tbody>
<tr>
<td>R2</td>
</tr>
<tr>
<td>0.58</td>
</tr>
</tbody>
</table>

Applicability domain

*Chemical domain*: Inorganic chemicals and chemicals with unusual elements (i.e., B and Se) are excluded during the data curation. Silicates are included.

*Model domain*: No prediction if 1) too high experimental range of similar molecules or 2) no similar molecules present in the dataset.
Examples

Sim. = 0.881
Exp. LD<sub>50</sub> = 350 mg/kg

Exp. = 390 mg/kg
Pred. = 439 mg/kg

Sim. = 0.83
Exp. LD<sub>50</sub> = 1220 mg/kg

Exp. = 400 mg/kg
No Prediction
Examples

Sim. = 0.941
Exp. LD$_{50}$ = 2900 mg/kg

Sim. = 0.94
Exp. LD$_{50}$ = 3300 mg/kg

Sim. = 0.931
Exp. LD$_{50}$ = 4000 mg/kg

Exp. = 3700 mg/kg
Pred. = 3497 mg/kg
Examples

Sim. = 0.964
Exp. LD$\text{}_{50}$ = 5.97 mg/kg

Sim. = 0.958
Exp. LD$\text{}_{50}$ = 2.91 mg/kg

Sim. = 0.946
Exp. LD$\text{}_{50}$ = 7070 mg/kg

Exp. = 0.812 mg/kg
No Prediction
Conclusions & perspectives

- k-NN demonstrated to be a simple method capable to obtain acceptable results
- Easily adapted for two prediction strategies:
  - automatic way for screening large inventories
  - read-across setting for allowing expert reasoning around available information for hazard assessment
- Easily implementable (freely available) in VEGA (www.vegahub.eu)
Thank you!

Questions?

alessandra.roncaglioni@marionegri.it
Examples

Sim. = 0.913
Exp. LD$_{50}$ = 10 mg/kg

Sim. = 0.912
Exp. LD$_{50}$ = 29 mg/kg

Sim. = 0.931
Exp. LD$_{50}$ = 2 mg/kg

Exp. = 7 mg/kg
Pred. = 7.27 mg/kg
Examples

Sim. = 0.898  
Exp. LD$_{50}$ = 0.55 mg/kg

Sim. = 0.877  
Exp. LD$_{50}$ = 826 mg/kg

Sim. = 0.842  
Exp. LD$_{50}$ = 1700 mg/kg

Exp. = 0.4 mg/kg  
No Prediction