Collaborative modeling efforts for the AcuteTox endpoints

Kamel Mansouri

Predictive Models for Acute Oral Systemic Toxicity Workshop
April 11-12, 2018

Disclaimer: ILS and ScitoVation staff provide technical support for NICEATM, but do not represent NIEHS, NTP, or the official positions of any federal agency.
Outline:

• Preparation for modeling
  • Available data overview
  • Chemical structures standardization and processing
  • Training set/evaluation set split
  • Prediction set preparation

• Modeling efforts
  • Participants
  • Submissions
  • Evaluation
  • Analysis

• Consensus
  • Analysis
  • Combining predictions
  • Analysis of the consensus predictions
  • Implementation of the consensus models to screen new chemicals

• Conclusions
Available data for modeling

15,688 chemicals total
21,200 LD50 values

- Very toxic: 11886
- Nontoxic: 11871
- EPA: 11755
- GHS: 11845
- LD50: 8908
QSAR-ready standardization workflow

Aim of the KNIME workflow:
• Combine different procedures and ideas
• Minimize the differences between the structures used for prediction
• Produce a flexible free and open source workflow to be shared

Mansouri et al. (http://ehp.niehs.nih.gov/15-10267/)
• The same training and test chemicals across all endpoints
• Split into training (75%) and evaluation (25%)
• Similar distributions and variability for values and categories
• Similar distribution of chemical structures sources

Training set: 9888 chemicals  Evaluation set: 2888 chemicals
Similar distribution of values and variability (LD50)

Replicates distribution between training and test set

Stdev distribution between training and test set

LD50 values distribution between training and test set
Similar distribution for true and false (NT, VT)

VT classes distribution between training and test set

NT classes distribution between training and test set
Similar distribution of categories (EPA, GHS)

EPA categories distribution between training and test set

GHS categories distribution between training and test set
Prediction set

Lists:
• ToxCast/Tox21
• EDSP
• TSCA
• Substances on the market (EPA Dashboard list)

After QSAR-ready standardization: 48137 structures to be predicted
Modeling efforts, participants

Previous collaborations:

**CERAPP**
Collaborative Estrogen Receptor Activity Prediction Project

Mansouiri et al. (http://ehp.niehs.nih.gov/15-10267/)

**CoMPARA**
Collaborative Modeling Project for Androgen Receptor Activity

Mansouiri et al. (DOI: 10.13140/RG.2.2.19612.80009)
Participant groups locations

Interactive map: https://batchgeo.com/map/9d3ff810a72d8a84093c74ab0601f01d
Received models

- Very Toxic: 31 models
- Non-toxic: 32 models
- EPA categories: 24 models
- GHS categories: 21 models
- LD50: 24 models

Total: 132 models
Evaluation procedure:

**Qualitative evaluation:**
- Documentation
- Defined endpoint
- Unambiguous algorithm
- Availability of code
- Applicability domain definition
- Availability of data used for modeling
- Mechanistic interpretation

**Quantitative evaluation:**
- Goodness of fit: training statistics
- Evaluation set predictivity: statistics on the evaluation set
- Robustness: balance between (Goodness of fit) & (Test set predictivity)

\[ S = 0.3 \times (\text{Goodness of fit}) + 0.45 \times (\text{Test set predictivity}) + 0.25 \times (\text{Robustness}) \]

**Categorical models (binary and multi-class):**

- **Goodness of fit**: \( BA_{Tr} \) + 0.3 \times (1 - |Sn_{Tr} - Sp_{Tr}|)
- **Test set predictivity**: 0.7 \times (BA_{Tst}) + 0.3 \times (1 - |Sn_{Tst} - Sp_{Tst}|)
- **Robustness**: 1 - |BA_{Tr} - BA_{Tst}|

**Continuous models:**

- **Goodness of fit**: \( R_{Tr}^2 \)
- **Test set predictivity**: \( R_{Tst}^2 \)
- **Robustness**: 1 - |\( R_{Tr}^2 \) - \( R_{Tst}^2 \)|
Evaluation results
Evaluation of the VT and NT models

Evaluation scores

- VT
- NT
Evaluation of the EPA and GHS models

Evaluation scores

[Bar chart with scores for different models, comparing EPA and GHS]
Evaluation of the LD50 models
Coverage of the models
Consensus

- **Categorical models:**
  Weighted majority rule

- **Continuous models:**
  Weighted average

For each chemical of the prediction set, the weights and consensus are calculated based on predictions within the AD only.

The predicted consensus value \( C_i \) of the chemical \( i \) is calculated as:

\[
C_i = \sum_{j=1}^{n} w_j \cdot P_j
\]

where \( n \) is the number of models that provided predictions for the chemical \( i \), and \( P_j \) is the prediction of each one of them. The weight \( w \) of each model \( j \) is calculated as:

\[
w_j = S_j / \sum_{k=1}^{n} S_k
\]

So that the sum of weights is equal to 1.
### Consensus results: Binary and LD50

<table>
<thead>
<tr>
<th></th>
<th>VT Train</th>
<th>VT Test</th>
<th>NT Train</th>
<th>NT Test</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sn</td>
<td>0.87</td>
<td>0.67</td>
<td>0.93</td>
<td>0.70</td>
</tr>
<tr>
<td>Sp</td>
<td>0.94</td>
<td>0.96</td>
<td>0.96</td>
<td>0.88</td>
</tr>
<tr>
<td>BA</td>
<td>0.93</td>
<td>0.81</td>
<td>0.94</td>
<td>0.79</td>
</tr>
</tbody>
</table>

The balanced accuracy of the replicate animal data for predicting VT and NT categories was 81% and 89%, respectively.

<table>
<thead>
<tr>
<th></th>
<th>LD50 Train</th>
<th>LD50 Test</th>
</tr>
</thead>
<tbody>
<tr>
<td>R2</td>
<td>0.84</td>
<td>0.64</td>
</tr>
<tr>
<td>RMSE</td>
<td>0.32</td>
<td>0.51</td>
</tr>
</tbody>
</table>

The reproducibility of the replicate animal data for predicting LD50 had R2 of 0.8 and RMSE of 0.42.
### Consensus results: EPA and GHS

<table>
<thead>
<tr>
<th></th>
<th>EPA Train</th>
<th>EPA Test</th>
<th>GHS Train</th>
<th>GHS Test</th>
</tr>
</thead>
<tbody>
<tr>
<td>Median Sn</td>
<td>0.73</td>
<td>0.5</td>
<td>0.63</td>
<td>0.45</td>
</tr>
<tr>
<td>Median Sp</td>
<td>0.96</td>
<td>0.91</td>
<td>0.91</td>
<td>0.92</td>
</tr>
<tr>
<td>BA</td>
<td>0.83</td>
<td>0.71</td>
<td>0.77</td>
<td>0.68</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>EPA Train Cat 1</th>
<th>EPA Train Cat 2</th>
<th>EPA Train Cat 3</th>
<th>EPA Train Cat 4</th>
<th>EPA Test Cat 1</th>
<th>EPA Test Cat 2</th>
<th>EPA Test Cat 3</th>
<th>EPA Test Cat 4</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sn</td>
<td>0.55</td>
<td>0.83</td>
<td>0.92</td>
<td>0.65</td>
<td>0.45</td>
<td>0.54</td>
<td>0.80</td>
<td>0.38</td>
</tr>
<tr>
<td>Sp</td>
<td>1</td>
<td>0.94</td>
<td>0.75</td>
<td>0.98</td>
<td>0.98</td>
<td>0.86</td>
<td>0.59</td>
<td>0.96</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>GHS Train Cat 1</th>
<th>GHS Train Cat 2</th>
<th>GHS Train Cat 3</th>
<th>GHS Train Cat 4</th>
<th>GHS Train Cat 5</th>
<th>GHS Test Cat 1</th>
<th>GHS Test Cat 2</th>
<th>GHS Test Cat 3</th>
<th>GHS Test Cat 4</th>
<th>GHS Test Cat 5</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sn</td>
<td>0.34</td>
<td>0.48</td>
<td>0.63</td>
<td>0.91</td>
<td>0.69</td>
<td>0.18</td>
<td>0.43</td>
<td>0.44</td>
<td>0.76</td>
<td>0.53</td>
</tr>
<tr>
<td>Sp</td>
<td>1</td>
<td>1</td>
<td>0.95</td>
<td>0.71</td>
<td>0.98</td>
<td>1</td>
<td>0.96</td>
<td>0.91</td>
<td>0.61</td>
<td>0.92</td>
</tr>
</tbody>
</table>

The accuracy of the animal data for predicting EPA and GHS categories was 78% and 74%, respectively.
Model concordance

VT consensus concordance

NT consensus concordance

LD50 consensus concordance
Model concordance

Concordance relaxed: fraction of models predicting the consensus category +/- 1 category.

EPA consensus, concordance

GHS consensus concordance

EPA consensus, concordance relaxed

GHS consensus, concordance relaxed
Discordance analysis

<70% concordance

- VT: 1374
- NT: 12778
- EPA: 27364
- GHS: 24659
- LD50: 21043
Structural similarity to the training set

Most disconcordant (<0.7)
4135 chemicals

Most concordant (>0.7)
7525 chemicals
Models to consensus evaluation

Coverage and BA of VT models Vs the Consensus

Coverage: number of predicted chemicals within the AD
Coverage and BA of NT models Vs the Consensus

Coverage: number of predicted chemicals within the AD
Models to consensus evaluation

Coverage and BA of EPA models Vs the Consensus

Coverage: number of predicted chemicals within the AD
Coverage: number of predicted chemicals within the AD
Coverage and R2 of LD50 models Vs the Consensus

Coverage: number of predicted chemicals within the AD
Consensus implementation

>=85% concordance
- VT: 44523
- NT: 21659

>=75% concordance
- EPA: 16959
- GHS: 20215
- LD50: 22738

Implementation for regulatory use:
- A defined endpoint
- An unambiguous algorithm
- A defined domain of applicability
- Appropriate measures of goodness-of-fit, robustness and predictivity
- Mechanistic interpretation, if possible
OPERAs are a suite of property predictions from the National Center for Computational Toxicology at the US Environmental Protection Agency. OPERA was derived from curated data, an automated curation procedure for addressing chemical errors and inconsistencies in public datasets used in QSAR modeling.

**Mansouri et al. OPERA models**

OPERA prediction report
Desktop and online Predictions:

https://github.com/kmansouri/OPERA

Standalone app:
batch mode for new chemicals

EPA Comptox dashboard:
batch mode download or drawing
Summary

• Generated high quality data and models that can be used prospectively to screen the chemical universe

• Screened tens of thousands of chemicals in a fast accurate and economic way.

• Free & open-source code and workflows

• Consensus models being implemented for future use to help with regulatory process

• Data and predictions will be soon available via the EPA’s CompTox dashboard and the NICEATM Integrated Chemical Environment
Acknowledgments

• All collaborating groups

• EPA/NCCT
  • Grace Patlewicz
  • Jeremy Fitzpatrick
  • Prachi Pradeep

• ILS/NICEATM
  • Dave Allen
  • Shannon Bell
  • Agnes Karmaus
  • Patricia Ceger
  • Judy Strickland

• NTP/NICEATM
  • Nicole Kleinstreuer
  • Warren Casey