



# Modeling via the WebTEST2.0 Platform

*Todd Martin<sup>1\*</sup>, Gabriel Sinclair<sup>2</sup>, Christian Ramsland<sup>2</sup>,  
Nathaniel Charest<sup>2</sup>, and Antony Williams<sup>1</sup>*

*<sup>1</sup> US EPA/ORD/CCTE*

*<sup>2</sup> ORAU*

*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.*

February 9, 2022

# QSAR Methods

## ➤ Python based QSAR methods

- **RF** - Random Forest
- **SVM** – Support Vector Machine
- **DNN** – Deep Neural Network
- **XGBoost** – eXtreme Gradient Boosting
- **Consensus** – average of above methods

➤ Easily implementable as web services for both model building and model prediction

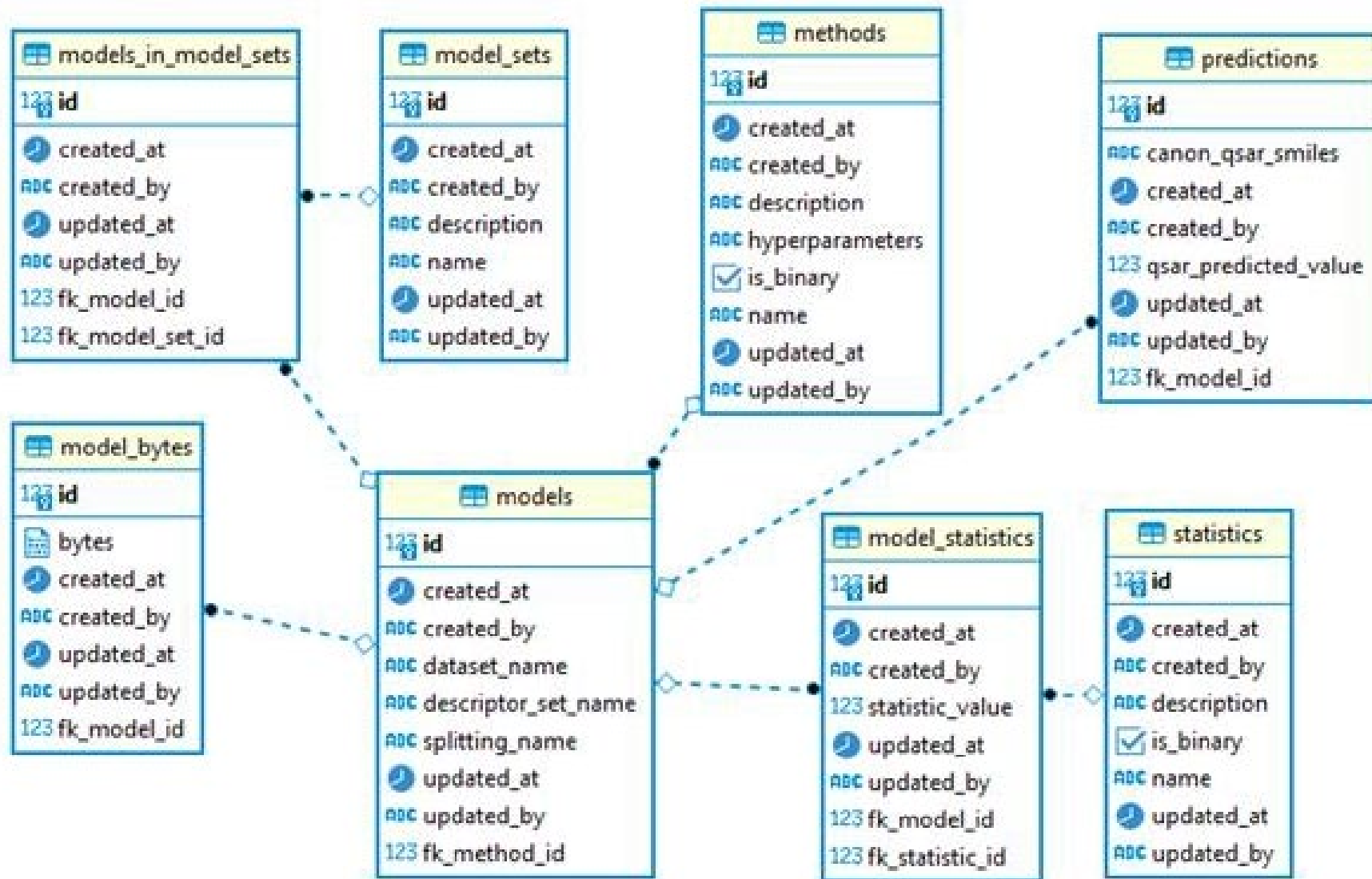
# Splitting into training and test sets

- Representative random forest splitting
  - Data set is split into 5-fold training and test sets
  - Random splitting is chosen using the splitting that gives the performance most similar to the average performance using random forest method
  - This approach was chosen so that we had a single training and test set for each dataset to reduce complexity for users
  - Other splittings can be added to allow comparison of performance with other researchers

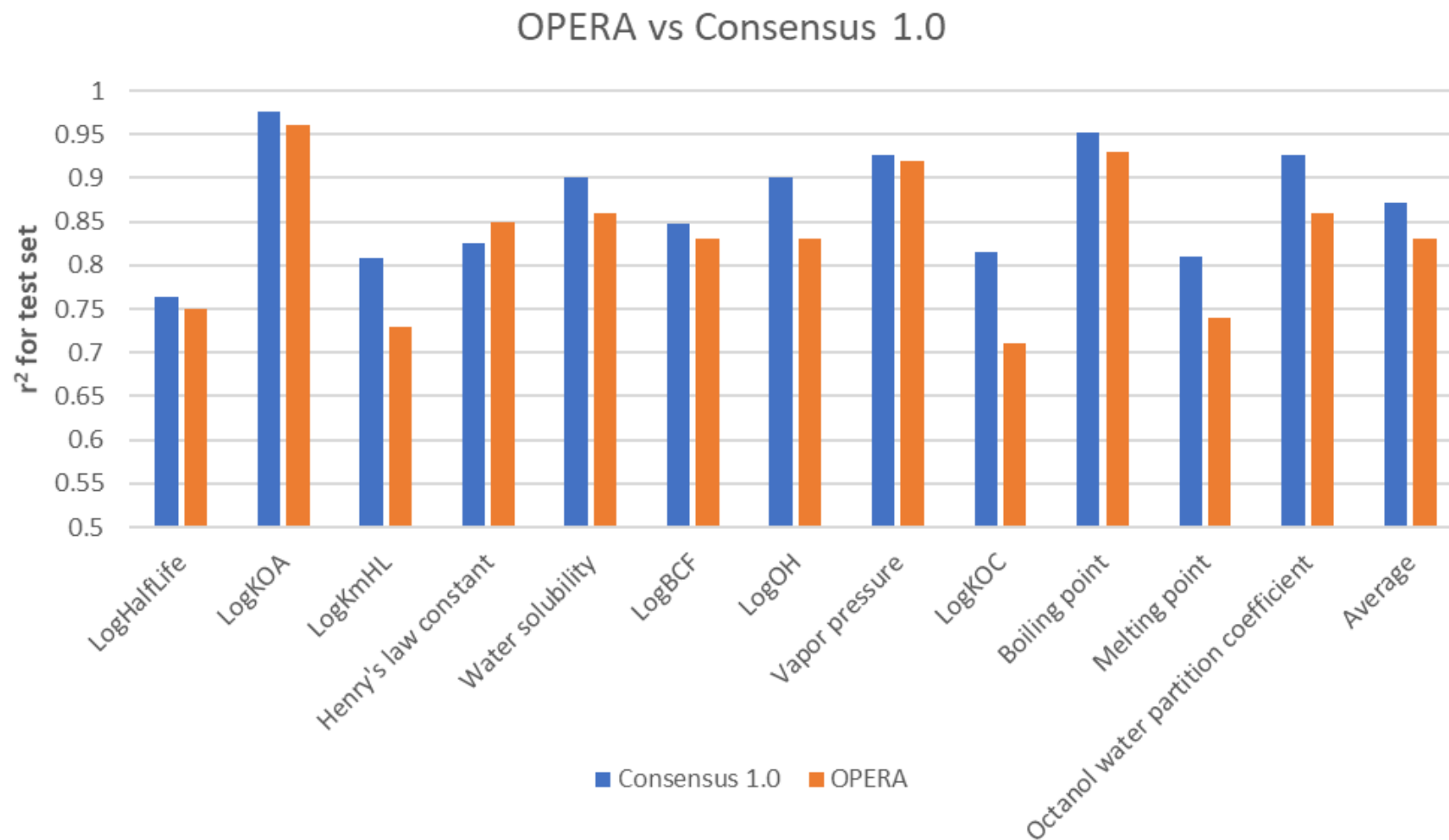
# Molecular descriptors

- Descriptors for model building:
  - T.E.S.T. 5.1
  - Padel
  - RDKit
  - Mordred
  - Others?
- Descriptors available via web API
- Descriptors are filtered prior to model building
  - Remove correlated, constant descriptors

# QSAR models schema



# Comparison to OPERA



# Prediction reports

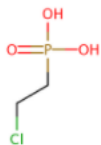
## Predicted LC50 for 16672-87-0 from Consensus method

Prediction results

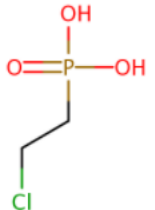
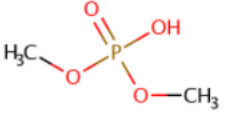
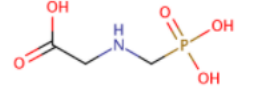
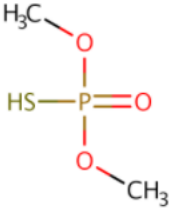
| Endpoint                   | Experimental value (CAS= 16672-87-0) | Predicted value <sup>a</sup> |
|----------------------------|--------------------------------------|------------------------------|
| LC <sub>50</sub> -log10(M) | 3.05                                 | 3.13                         |
| LC <sub>50</sub> mg/L      | 129.97                               | 106.46                       |

<sup>a</sup>Note: the test chemical was present in the training set. The prediction *does not* represent an external prediction.

| Individual Predictions |  |                           |
|------------------------|--|---------------------------|
| Method                 | Method Description   | Predicted value -log10(M) |
| svm_regressor_1.1      | <a href="#">sklearn implementation of SVM using NuSVR for regression</a> | 3.23                      |
| dnn_regressor_1.8      | <a href="#">tensorflow/keras implementation of DNN</a>                   | 3.11                      |
| rf_regressor_1.1       | <a href="#">sklearn implementation of random forest</a>                  | 3.14                      |
| xgb_regressor_1.0      | <a href="#">python implementation of extreme gradient boosting</a>       | 3.05                      |

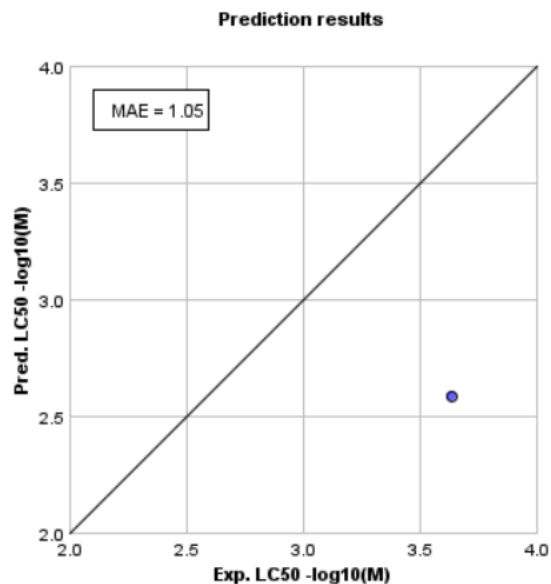

ClCCOP(=O)(O)O

Applicability domain

| Test chemical   | Training Neighbor 1   | Training Neighbor 2  | Training Neighbor 3   | Result  |
|---|---|--|---|---|
|  <p>exp=3.05</p> |  <p>SC=0.79<br/>exp=3.85</p> |  <p>SC=0.78<br/>exp=4.57</p> |  <p>SC=0.71<br/>exp=3.52</p> | <p>SC for 95% training coverage=0.52<br/>Average SC=0.76<br/><b>Result: Inside AD</b></p> |

## Predictions for the test chemical and for the most similar chemicals in the external test set

If the predicted value matches the experimental values for similar chemicals in the test set (and the similar chemicals were predicted well), one has greater confidence in the predicted value.



Results for entire set vs  
results for similar chemicals

| Chemicals                         | MAE* |
|-----------------------------------|------|
| Entire set                        | 0.49 |
| Similarity coefficient $\geq 0.5$ | 1.05 |

\*Mean absolute error in  $-\log_{10}(M)$

Color legend

| Color  | Range*              |
|--------|---------------------|
| Green  | $SC \geq 0.9$       |
| Blue   | $0.8 \leq SC < 0.9$ |
| Yellow | $0.7 \leq SC < 0.8$ |
| Orange | $0.6 \leq SC < 0.7$ |
| Red    | $0.6 < SC$          |

\*SC = similarity coefficient

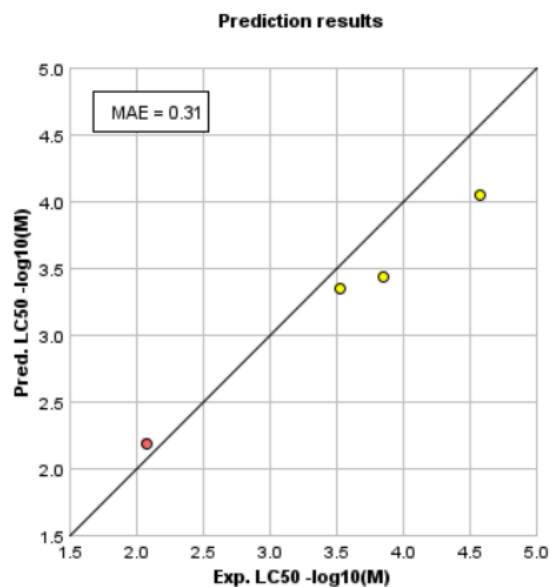
Results for similar chemicals

| CAS                           | Structure | Similarity Coefficient | Experimental value $-\log_{10}(M)$ | Predicted value $-\log_{10}(M)$ |
|-------------------------------|-----------|------------------------|------------------------------------|---------------------------------|
| 16672-87-0<br>(test chemical) |           |                        | 3.05                               | 3.13                            |
| 2074-67-1                     |           | 0.84                   | 3.64                               | 2.58                            |



## Predictions for the test chemical and for the most similar chemicals in the training set

If the predicted value matches the experimental values for similar chemicals in the training set (and the similar chemicals were predicted well), one has greater confidence in the predicted value.



Results for entire set vs  
results for similar chemicals

| Chemicals                         | MAE* |
|-----------------------------------|------|
| Entire set                        | 0.22 |
| Similarity coefficient $\geq 0.5$ | 0.31 |

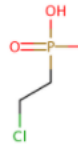
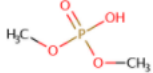
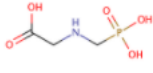
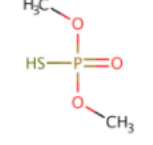

\*Mean absolute error in  $-\log_{10}(M)$

Color legend

| Color  | Range*              |
|--------|---------------------|
| Green  | $SC \geq 0.9$       |
| Blue   | $0.8 \leq SC < 0.9$ |
| Yellow | $0.7 \leq SC < 0.8$ |
| Orange | $0.6 \leq SC < 0.7$ |
| Red    | $0.6 < SC$          |

\*SC = similarity coefficient

Results for similar chemicals

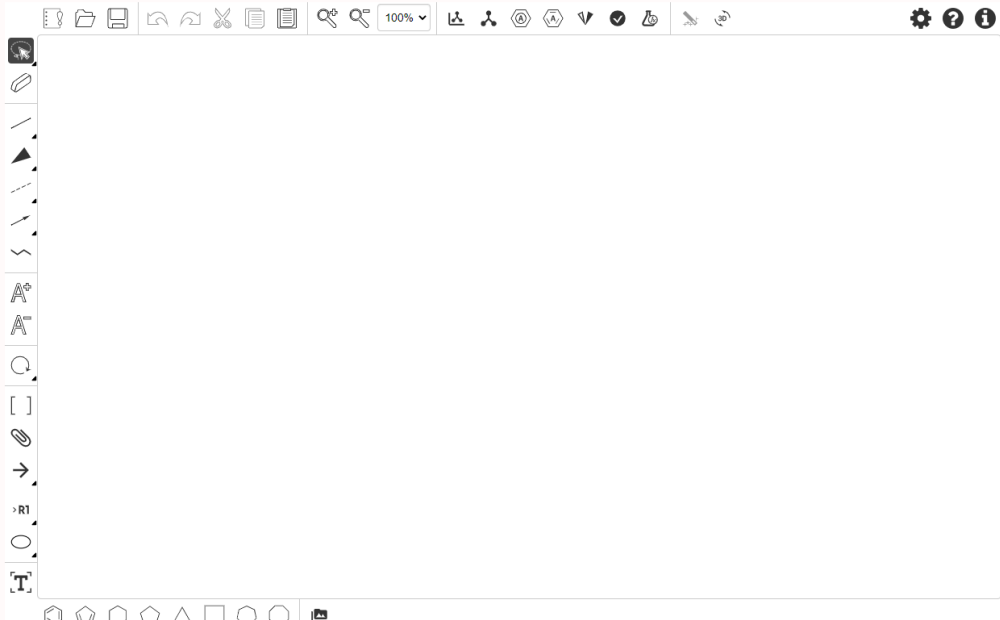
| CAS                           | Structure   | Similarity Coefficient | Experimental value $-\log_{10}(M)$ | Predicted value $-\log_{10}(M)$ |
|-------------------------------|---|------------------------|------------------------------------|---------------------------------|
| 16672-87-0<br>(test chemical) |    |                        | 3.05                               | 3.13                            |
| 813-78-5                      |    | 0.79                   | 3.85                               | 3.44                            |
| 1071-83-6                     |    | 0.78                   | 4.57                               | 4.05                            |
| 1112-38-5                     |   | 0.71                   | 3.52                               | 3.35                            |
| 627-30-5                      |  | 0.51                   | 2.07                               | 2.19                            |

# Single chemical real time predictions

Model set **WebTEST 2.0**

**Predictions**

Search for chemical by systematic name, synonym, CAS number, DTXSID or InChIKey



|                                     | Property                         | Dataset                               | Descriptor Set      | Splitting                 | Methods                        | Type          |
|-------------------------------------|----------------------------------|---------------------------------------|---------------------|---------------------------|--------------------------------|---------------|
| <input checked="" type="checkbox"/> | <u>Skin sensitization</u>        | <u>LLNA 1.0</u>                       | <u>T.E.S.T. 5.1</u> | <u>RND_REPRESENTATIVE</u> | <u>SVM, DNN, XGBoost, RF</u> ⓘ | Toxicological |
| <input checked="" type="checkbox"/> | <u>Carcinogenicity (binary)</u>  | <u>Carcinogenicity (binary) 1.0</u>   | <u>Padelpy</u>      | <u>RND_REPRESENTATIVE</u> | <u>SVM, DNN, XGBoost, RF</u> ⓘ | Toxicological |
| <input checked="" type="checkbox"/> | <u>Carcinogenicity (potency)</u> | <u>Carcinogenicity (potency) 1.0</u>  | <u>T.E.S.T. 5.1</u> | <u>RND_REPRESENTATIVE</u> | <u>SVM, DNN, XGBoost, RF</u> ⓘ | Toxicological |
| <input checked="" type="checkbox"/> | <u>Water solubility</u>          | <u>Water solubility 1.0</u>           | <u>T.E.S.T. 5.1</u> | <u>RND_REPRESENTATIVE</u> | <u>SVM, DNN, XGBoost, RF</u> ⓘ | Physical      |
| <input checked="" type="checkbox"/> | <u>logKow</u>                    | <u>logKow 1.0</u>                     | <u>T.E.S.T. 5.1</u> | <u>RND_REPRESENTATIVE</u> | <u>SVM, DNN, XGBoost, RF</u> ⓘ | Physical      |
| <input checked="" type="checkbox"/> | <u>logKow</u>                    | <u>logKow 1.0</u><br><u>PFAS only</u> | <u>T.E.S.T. 5.1</u> | <u>RND_REPRESENTATIVE</u> | <u>SVM, DNN, XGBoost, RF</u> ⓘ | Physical      |
| <input checked="" type="checkbox"/> | <u>Functional use</u>            | <u>Functional use 1.0</u>             | <u>T.E.S.T. 5.1</u> | <u>RND_REPRESENTATIVE</u> | <u>SVM, DNN, XGBoost, RF</u> ⓘ | Miscellaneous |

## Future work

- Add third party predictions via web services and adding model metadata to the database
  - EPISUITE, OPERA, VEGA, WebTEST1.0

# Questions???

The views expressed in this presentation are those of the author and do not necessarily represent the views or policies of the U.S. Environmental Protection Agency