Predicting Five Rat Acute Toxicity Endpoints with ANNE Models using ADMET Predictor™

Pankaj R. Daga, Michael Lawless, Marvin Waldman, Robert Fraczkiewicz, Robert D. Clark, John DiBella, and Michael B. Bolger

Simulations Plus, Inc., 42505 10th Street West, Lancaster, CA 93534, USA (www.simulations-plus.com)

Introduction

• Alternative methods are being explored to predict the toxicity of chemicals to reduce use of animals.
  ✓ Laboratory/Animal tests are costly in time and money
  ✓ Cheminformatics (QSTR) presents a good alternative to animal testing

Why Artificial Neural Network Ensemble (ANNE)?

• Toxicity prediction is a tough problem
  ✓ Multiple underlying mechanisms of action
  ✓ Datasets studied (e.g., rat LD50) are large and chemically diverse
  ✓ Multiple and wide variety of data sources
  ✓ Simple regression methods like MLR may prove insufficient

• Ensemble methods, such as ANNE and Random Forest, have proven to be robust enough to tackle this intensive task

Five endpoints were provided to model

• Rat LD50 and “Very Toxic”, “Non Toxic”, “EPA Cat” & “GHS Cat”
  ✓ The labels in the four endpoints are dependent upon rat LD50

Why Dataset Curation Necessary?

• The “QSAR-ready structures” provided as training set needed careful curation

Uncertainty in structures is not useful

• Matched Molecular Pair Analysis shows a few large activity cliffs
  ✓ The data is questionable and hence excluded

Molecular Descriptors

ADMET Predictor™ generated 341 molecular descriptors
✓ Constitutional Descriptors
✓ Topological Indices
✓ Electropotential State Indices
✓ Charge-based Descriptors
✓ Hydrogen Bonding Descriptors
✓ Moriguchi Descriptors
✓ Functional Groups

Applicability Domain

Models predicted fewer false negatives compared to false positives. Thus, they erred on the side of caution, e.g., fewer toxic compounds were incorrectly predicted. This can be seen in the EPA and GHS category predictions which show fewer incorrect compounds in the lower right-hand corner than the upper left-hand corner.

ADMET Predictor™ generated 341 molecular descriptors

Molecular Descriptors

Constitutional Descriptors
Topological Indices
Electrotopological State Indices
Charge-based Descriptors
Hydrogen Bonding Descriptors
Moriguchi Descriptors
Functional Groups

Model Building Steps

• All the models show comparable performance on both training & test set
• Overall statistics suggests that models are NOT OVERTRAINEd
• Almost all compounds were predicted within applicability domain of models.
• Only ~50 compounds (1.5%) were predicted out of the AD, 48 contained a Si, Se, or heavy metal atom and 2 compounds exceeded the 256 heavy atom limit of ADMET Predictor.

Model Performance & Analysis

• Matched Molecular Pair Analysis shows a few large activity cliffs
• The data is questionable and hence excluded

ADMET Predictor™

Incorrect tautomer updated to Correct tautomer

Correct tautomer assignment is necessary in model building exercise as well as for correct prediction

Model Endpoint Validation Data Size Training Set Test Set Outside AD (%) Performance On Training Performance On Ext Test

EPA Cat_1 EPA class (1-4) 2812 6531 1633 50 (1.8%) 0.698 0.696
EPA Cat_2 EPA class (1-4) 2812 6531 1633 51 (1.8%) 0.693 0.691
GHS Cat_1 GHS class (1-5) 2882 6951 1648 51 (1.8%) 0.708 0.666
GHS Cat_2 GHS class (1-5) 2882 6951 1648 51 (1.8%) 0.689 0.671
LD50_1 LD50 ≥ 2,000 mg/kg 2887 7059 1246 54 (1.9%) 0.765 0.750
NonTox_1 LD50 ≤ 2,000 mg/kg 2887 7059 1246 55 (1.9%) 0.771 0.748
VeryTox_1 LD50 ≤ 50 mg/kg 2891 6699 1675 52 (1.8%) 0.675 0.620
VeryTox_2 LD50 ≤ 50 mg/kg 2891 6699 1675 53 (1.8%) 0.809 0.825

Corrected chemical structures are ready for modeling exercise as well as for correct prediction.

1 BA for EPA, GHS, NT, and VT. TST_RMSE for LD50 2 Existing model from ADMET Predictor was used to predict LD50.