Genotox-iS™

iS Tox Platform™
Machine-learning models for a new toxicology

Report

Client:
Altox

Username:
Tiago

Study Number:
Genotox-iS_Compound1_

Date:
2019/06/19 - 15:21:02

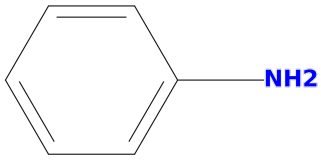
Program Version: 1.8

Molecular Query

Name:
Compound 1

CAS:
NA

SMILES:
Nc1ccccc1



Model Summary

Genotox-iS™ is a computational tool for prediction of genotoxicity by alerts, statistical and machine learning-based models, which were validated based on OECD (Organisation for Economic Co-operation and Development) Principles for the Validation for Regulatory Purposes of (Q)SAR Models. These OECD principles are discussed in each section of this report.

These different models use 102 Structural Alerts (included in 46 categories) and an *in vitro* genotoxicity dataset containing 6,931 rigorously curated structures for the *in vitro* genotoxicity individual predictions (defined endpoint - OECD principle 1).

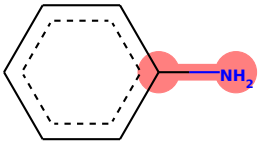
A final conclusion is provided by the Consensus Prediction Path® plot, a strong model with an accuracy rating of 99%, combining the most predictive individual models in a visual decision tree. Results can be positive (mutagen), negative (non-mutagen) or inconclusive (for the low confidence level case).

Alert Analysis

The result below is based on an analysis of fragments assigned to be mutagenic (*in vitro* mutagenicity alerts). When possible, the alerts can provide mechanistic basis of the predicted mutagenicity of the molecule or insights for interpretation of the mechanism based on theories and knowledge of toxicity mechanisms (OECD Principle 5).

Result: (+) Positive

Alerts were found in the molecule. The results are in the table below and a description is provided at the end of the report.

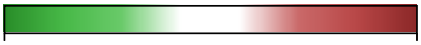
Category	Alert	Alert ID	References
in vitro mutagenicity (Ames alert) alerts by ISS		Primary aromatic amine,hydroxyl amine and its derived esters	Benigni, R., Giuliani, A., Franke, R., and Gruska, A. (2000). Quantitative structure-activity relationships of mutagenic and carcinogenic aromatic amines. Chem.Revs. 100, 3697-3714. Woo, Y. T. and Lai, D. Y. (2001). Aromatic amino and nitro-amino compounds and their halogenated derivatives. In 'Patty's Toxicology. Vol. 4.' (Eds E. Bingham, B. Cohrssen, and C. H. Powell.) pp. 969-1105. (John Wiley and Sons, Inc: New York.)

Machine Learning Models

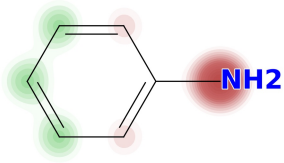
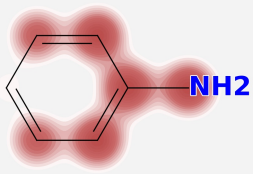
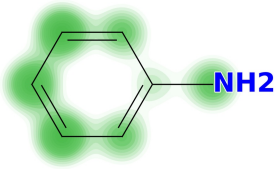
The individual results in the table below were obtained from three statistical and artificial intelligence algorithms: random forest model (RF), k-nearest neighbors algorithm (k-NN) and Deep Learning models.

To ensure some transparency in the description of the model algorithms (an unambiguous algorithm - OECD Principle 2), more detailed information about each model is presented below. The Probability Map indicates the fragments more related to the absence (green) or presence (red) of toxicity, useful for hypotheses for mechanistic interpretations (OECD Principle 5).

Probability Map



Non-Mutagen (-) Mutagen (+)

Method	Prediction (Confidence)	Probability Mapping (SAR)
Random Forest Machine learning decision model implemented with the 2D MACCS fingerprint	Non-Mutagen (-) (93.0%)	
kNN k-nearest neighbors decision model implemented with the 2D Extended Connectivity Fingerprint	Mutagen (+) (71.4%)	
Deep Learning Deep Learning categorical model implemented with hybrid descriptors (ECFP6 fingerprint and physicochemical properties: MW, TPSA, $logK_{ow}$, logD)	Non-Mutagen (-) (98.4%)	

Detailed data about the dataset of chemicals; the end-point and descriptor values; the derivation of the descriptors; the test and training sets; the outliers removed; the statistical parameters and others are in the QMRF (QSAR Model Reporting Format) report, available to queries under a confidentiality agreement. Altox Ltda assures scientific integrity of the data.

Random forest

Our random forest model (RF) was built with 750 individual decision trees models, and the final prediction is based on the ensemble (average) of each individual decision tree model prediction. Also, we employed the MACCS fingerprint bit vector as the independent variable for the RF model. The MACCS fingerprint is considered a global fingerprint based on the 166 SMARTS patterns commonly found in a broad range of molecules.

k-NN model

The k-nearest neighbors algorithm (k-NN) is a type of instance-based learning, or lazy learning, very similar to the “read-across” method. Our k-NN model assigns a prediction for a compound based on majority vote of its seven neighbors, in conjunction with ECFP4 circular fingerprints with 2048 bits and an atom radius of 2 (Morgan2). The Extended-Connectivity Fingerprints (ECFPs) are circular topological fingerprints designed for molecular characterization, similarity searching, and structure-activity modeling.

Deep Learning 3D

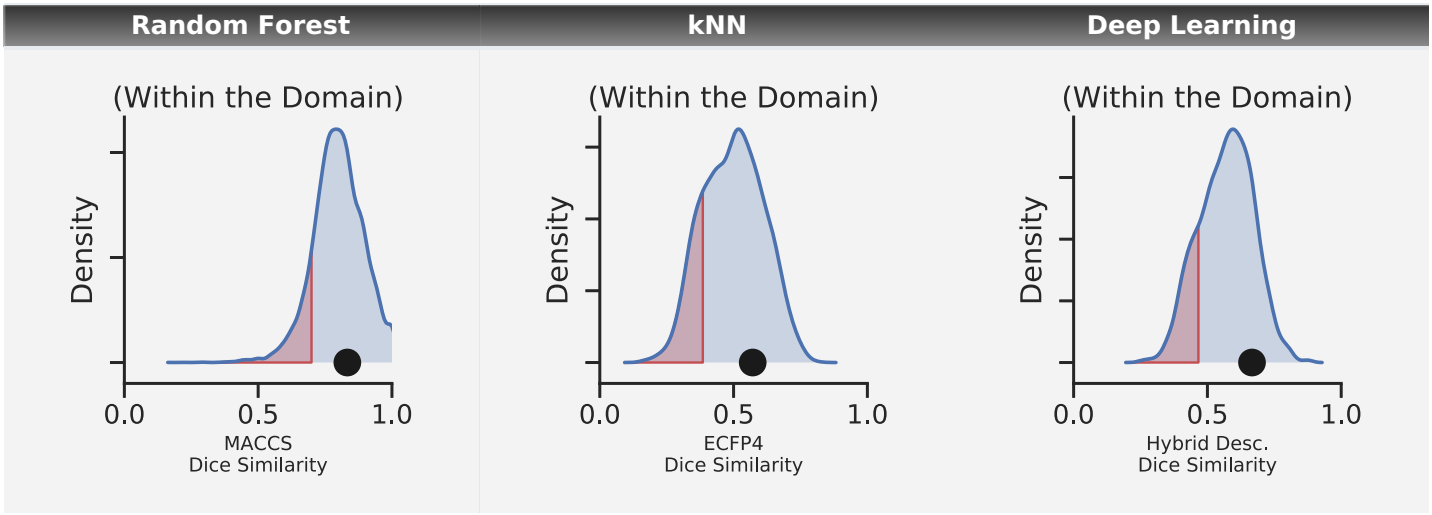
The Deep learning model is a machine learning algorithm based on neural networks or artificial intelligence components with multiple hidden layers that can learn increasingly abstract representations of the chemical fingerprint. In this model, the first hidden layers might only learn local edge patterns. Then, every 14 subsequent layers learn more complex representations. Finally, the last layer can evaluate the mutagenicity of the given compound. Also, the 3D toxicophoric fingerprint was employed. The “E3FP” is an algorithm to calculate 3D conformer fingerprint-like Feature Connectivity Fingerprint (FCFP).

Visual AD Inspection®

The applicability domain (AD) is defined by the chemical structure space and the toxicological response encoded by the developed model, to make new predictions with a given reliability (a defined domain of applicability - OECD Principle 3). Our visual AS Inspection® is used to establish the scope and limitations of the models. Basically, new chemicals must be reasonably similar to training set compounds or a valid prediction cannot be accepted.

Our visual AD inspection is represented by a density plot of the average fingerprint-dice similarity for the k-nearest neighbors of each compound during the 5-Fold external model's validation. The chemical structure is represented by three different types of fingerprints: MACCS, ECFP4 and 3D toxicophoric. At the visual AD inspection, the black circle represents the evaluated compound, the highlighted red area means the forbidden similarity region, and the blue region is the allowed similarity chemical space to predict new compounds.

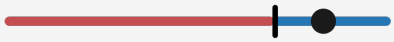
Even though a well-designed AD helps the user to assess the reliability of predictions made by the model, it should not automatically be assumed that all predictions within the defined AD are necessarily reliable.



Final Result

The results are presented below; positive (red) and/or negative (green) predictions are presented with applicability domain (AD) and confidence level (robustness, OECD principle 4) for the Rule-based Expert and Advanced Statistical systems.

For the conclusion, two (Q)SAR prediction methodologies that complement each other should be applied. One methodology is an expert rule-based and the second methodology is composed of statistical-based models. The absence of structural alerts from two complementary (Q)SAR methodologies (expert rule-based and statistical) is sufficient to conclude that the impurity is of no mutagenic concern, and no further testing is recommended (ICH M7).

Method	Prediction (Confidence)	Applicability Domain
Rule-based Expert System		
Structural Alerts	**Mutagen (+)**	-
Advanced Statistical System		
Random Forest Machine learning decision model implemented with the 2D MACCS fingerprint	**Non-Mutagen (-) (93.0%)**	Within 
kNN k-nearest neighbors decision model implemented with the 2D Extended Connectivity Fingerprint	**Mutagen (+) (71.4%)**	Within 
Deep Learning Deep Learning categorical model implemented with hybrid descriptors (ECFP6 fingerprint and physicochemical properties: MW, TPSA, $logK_{ow}$, logD)	**Non-Mutagen (-) (98.4%)**	Within 

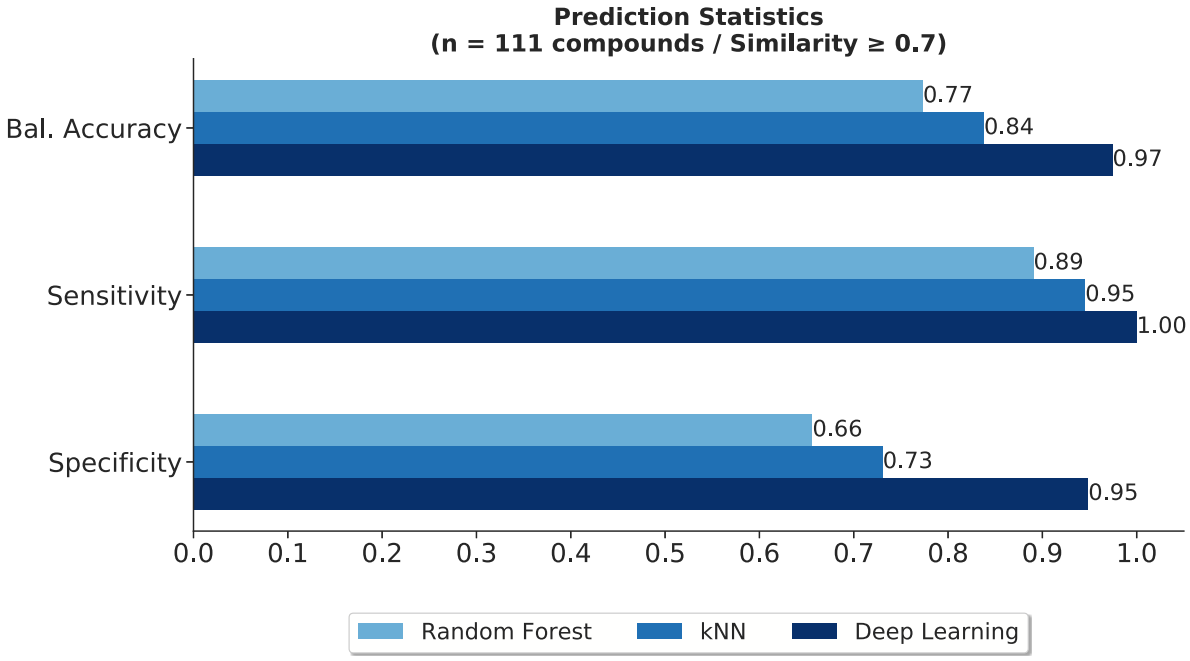
Additional Information

Prediction Confidence

Based on the most similar molecules

With appropriate measures of goodness-of-fit, robustness, and predictivity (OECD Principle 4), our model used different strategies to establish the performance of the model, which consisted of internal model performance (goodness-of-fit and robustness) and external model performance (predictivity).

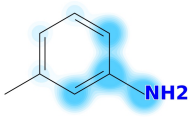
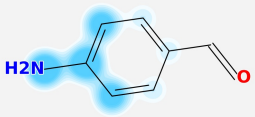
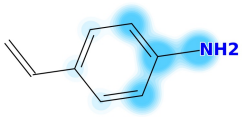
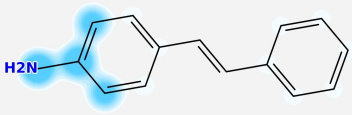
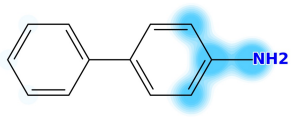
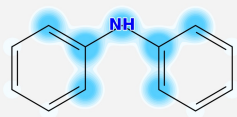
To assess the confidence of the Genotox-iS™ predictions, after of to take into account the applicability domain of the model (Visual applicability domain (AD) Inspection®), additionally, the statistics of the ability to detect known mutagenic compounds (sensitivity), non-mutagenic compounds (specificity), and all molecules in general (concordance) based in the most similar substances are provided below. A map with similarity level for the 10 most similar molecules is provided with the confidence level for each statistical and artificial intelligence model.

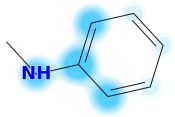
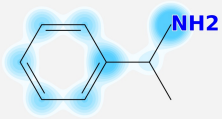
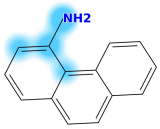
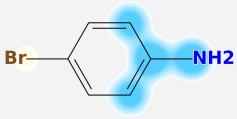


Sorensen-Dice MACCS similarity was used to improve the deep learning confidence by interpolating the confidence equalized by the compound similarity criteria obtained from the dataset chemical space. This helps to improve the *in silico* toxicological model to reduce the false positive and negative error.

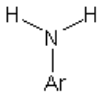
Performance for the 10-most similar molecules

Similarity Map

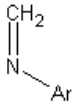
Molecule (Similarity)	Experimental Data	Random Forest Prediction (Confidence)	kNN Prediction (Confidence)	Deep Learning Prediction (Confidence)
 (0.95)	Non-Mutagen	Mutagen (53.0%)	Non-Mutagen (71.4%)	Non-Mutagen (66.2%)
 (0.91)	Non-Mutagen	Non-Mutagen (75.7%)	Non-Mutagen (71.4%)	Non-Mutagen (71.8%)
 (0.91)	Mutagen	Non-Mutagen (68.6%)	Non-Mutagen (57.1%)	Mutagen (67.6%)
 (0.87)	Mutagen	Mutagen (98.1%)	Mutagen (71.4%)	Mutagen (73.2%)
 (0.87)	Mutagen	Mutagen (94.4%)	Mutagen (85.7%)	Mutagen (88.5%)
 (0.86)	Non-Mutagen	Mutagen (53.2%)	Non-Mutagen (57.1%)	Non-Mutagen (73.3%)

Molecule (Similarity)	Experimental Data	Random Forest Prediction (Confidence)	kNN Prediction (Confidence)	Deep Learning Prediction (Confidence)
 (0.86)	Non-Mutagen	Non-Mutagen (79.5%)	Non-Mutagen (57.1%)	Non-Mutagen (68.0%)
 (0.84)	Mutagen	Non-Mutagen (96.3%)	Non-Mutagen (85.7%)	Mutagen (92.3%)
 (0.83)	Non-Mutagen	Mutagen (100.0%)	Mutagen (85.7%)	Non-Mutagen (77.4%)
 (0.83)	Non-Mutagen	Non-Mutagen (63.2%)	Non-Mutagen (71.4%)	Mutagen (73.2%)

Primary aromatic amines, hydroxyl amines and derived esters (genotox)



or amine generating group:



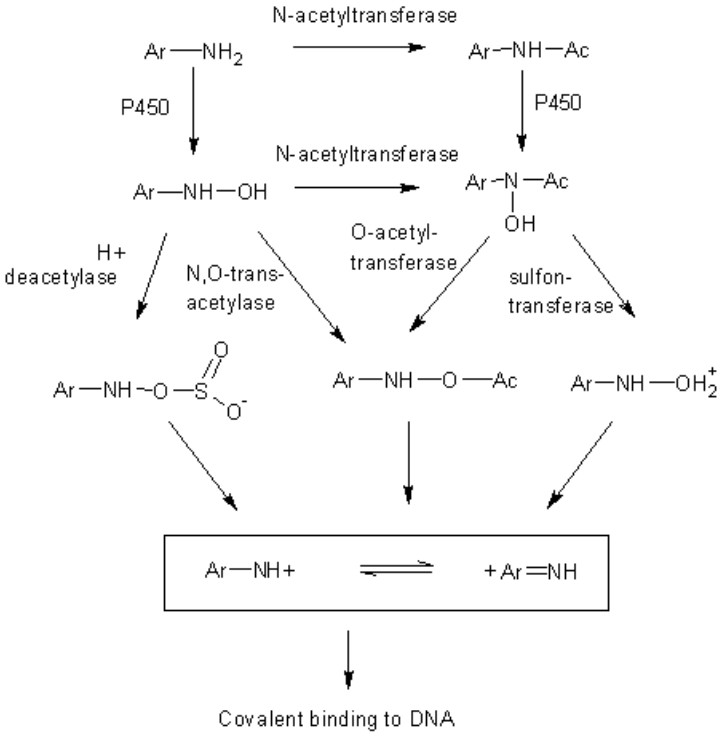
Ar = Any aromatic/heteroaromatic ring

R= Any atom/group

· Chemicals with ortho-disubstitution, or with an ortho carboxylic acid substituent are excluded.

· Chemicals with a sulfonic acid group (-SO3H) on the same ring of the nitro group are excluded.

Aromatic amines have to be metabolized to reactive electrophiles in order to exert their carcinogenic potential. For aromatic amines and amides, this typically involves an initial N-oxidation to N-hydroxyarylamines and N-hydroxyarylamides, which is mediated by cytochrome P-450. Upon further activation by enzymatic esterification, nitrenium ions are formed. These highly reactive intermediates bind covalently to biomolecules, generating aminoaryl derivatives (Benigni et al. 2000); (Woo and Lai 2001).



In addition to the reactions of nitrogen (main activation pathway), certain aromatic amines are converted into electrophilic derivatives through ring oxidation pathways. Ring hydroxylation and subsequent enzymatic or spontaneous dehydrogenation, can result in the formation of iminoquinones, which are directly electrophilic metabolites (Romano Zito, personal communication).

References Cited

Benigni, R., Giuliani, A., Franke, R., and Gruska, A. (2000). Quantitative structure-activity relationships of mutagenic and carcinogenic aromatic amines. Chem.Revs. 100, 3697-3714.

Woo, Y. T. and Lai, D. Y. (2001). Aromatic amino and nitro-amino compounds and their halogenated derivatives. In 'Patty's Toxicology. Vol. 4.' (Eds E. Bingham, B. Cohrssen, and C. H. Powell.) pp. 969-1105. (John Wiley and Sons, Inc: New York.)