ACD/Labs

Data sheet

Impurity Profiling Module

ACD/Percepta

Overview

According to the FDA Guidance for Industry, impurities and degradants in drug products identified below the ICH qualification thresholds may be evaluated for genotoxicity and carcinogenicity based on structural activity relationship (SAR) assessments using computational software.

ACD/Labs has developed an in silico package for such assessment, in collaboration with FDA Center for Food Safety and Nutrition (CFSAN). The ACD/Labs Genotoxicity predictor provides several distinct tools to help you better determine the hazardous potential of genotoxic impurities and other compounds you are working with:

- Impurity profiling module, which aggregates the output of probabilistic predictive models for a variety of endpoints characterizing genotoxic and/or carcinogenic potential of chemicals, and displayed structural alerts produced by the knowledge-based expert system
- A dedicated Ames Test module providing fast and accurate predictions of the mutagenic potential of candidate compounds with the possibility of model training
- A browsable database of experimental Ames Test data

This combination of probabilistic predictors and a knowledge-based expert system provides two different approaches to the determination of toxic potential, ensuring greater confidence in results.

These powerful tools enable researchers to quickly identify genotoxic impurities, and eliminate potentially hazardous substances aiding compound selection and prioritization of genotoxicity testing in risk assessment.

Features

- Calculates the probability for a compound to produce positive results in a battery of assays that detect various mechanisms of hazardous action, including mutagenicity, clastogenicity, carcinogenicity, and reproductive toxicity.
- All predictions are supported by Reliability Index (RI) values that represent a quantitative evaluation of prediction confidence.
- Displays experimental results from a particular assay, for up to 5 similar structures taken from the relevant training set, along with each prediction.
- The knowledge-based ‘Genotoxicity/Carcinogenicity Hazards’ expert system identifies and highlights structural fragments potentially involved in toxic activity, along with their statistical significance, short descriptions of their mechanism of action, and literature references.
- Ames test module visualizes the statistical contributions of different parts of the molecule to the overall predicted result by color-mapping them onto the structure (red – atoms or fragments associated with genotoxicity, green – not involved in genotoxic effect).
- A fully searchable database of over 5500 experimental Ames Test results provides information about individual studies conducted with each compound corresponding to various bacterial strains tested, presence or absence of metabolic activation, as well as other experimental conditions.

- Batch calculation mode allows automatically calculating acute toxicity of hundreds of molecules per minute without user intervention.
- Ames test predictor is trainable – increase prediction accuracy with your experimental data.

Technical Information

Experimental Data

Overall, ACD/Impurity profiling package includes predictive models for 21 distinct endpoints that detect different mechanisms of hazardous activity. These can be divided into three large groups – genetic toxicity endpoints, carcinogenicity studies, and reproductive toxicity studies. The sources of experimental data for each of the groups are listed below, while data set sizes are given in Table 1.

Genetic toxicity: data sets for standard assays reflecting different mechanisms of genetic damage were obtained from the FDA. Gene mutation tests and techniques detecting clastogenic/aneugenic effects are included. The original data sources were EPA GENE-TOX database and scientific literature.

Carcinogenicity: results of chronic (two-year term) carcinogenicity studies in rats and mice were received from the FDA. This data was based on NTP technical reports, IARC monographs, the Carcinogenic Potency DataBase (CPDB) and other publicly available sources. Raw data was converted to a binary classification using a weight of evidence (WOE) approach. Classification using the WOE threshold corresponding to “potent carcinogens” was used to build the models in the current study.

Reproductive toxicity: experimental data characterizing the potential for endocrine system disruption due to Estrogen receptor α binding were acquired from METI database and original publications. Compounds were classified as binders/non-binders on the basis of their relative binding affinities (RBA) compared to reference ligand estradiol. Two cut-offs were used: Log*RBA* > -3 (“general binding”), and Log*RBA* > 0 (“strong binding”).

Table 1. Bioassays considered in the study, the respective dataset sizes, and model performance.

Mechanism/ test system	Endpoint	No. of compounds (% positives)	Sensitivity	Specificity
Genetic toxicity >				
Mutagenicity >	Salmonella	7826 (49.5%)	87.1%	81.7%
Procaryote	Escherichia	1479 (26.1%)	72.5%	87.0%
Eukaryote	Eukaryote composite	2901 (54.9%)	78.1%	64.0%
	Yeast	658 (52.7%)	86.7%	80.0%
	Drosophila	600 (48.8%)	70.6%	81.8%
	ML51	1272 (60.0%)	76.2%	64.7%
	CHO/CHL all loci	1229 (47.6%)	80.0%	67.5%
Clastogenicity >				
Chromosome aberrations	CA in vitro	2034 (46.3%)	74.6%	71.5%
Micronucleus test	MNT in vivo	1299 (31.0%)	62.1%	69.5%
DNA damage	UDS composite	593 (28.0%)	66.7%	76.3%
Carcinogenicity	Rodent composite	2211 (30.5%)	58.8%	81.9%
Reproductive toxicity >				
Endocrine disruption >	Log *RBA* > 0	1464 (51.6%)	82.9%	97.1%
Estrogen receptor binding	Log *RBA* > -3	1464 (23.8%)	92.1%	85.6%

Modeling details & prediction accuracy

The predictive models for all considered genetic and reproductive toxicity endpoints were derived using GALAS (Global, Adjusted Locally According to Similarity) modeling methodology. A GALAS model consists of two parts:

- **Global** (baseline) statistical model based on binomial PLS with multiple bootstrapping, using a predefined set of fragmental descriptors.
- **Local** correction to baseline prediction based on the analysis of model performance for similar compounds from the training set (the so called Self-training Library).

In carcinogenicity modeling, the local correction step was only used to produce RI values, while probability estimation also took into account the output of Ames test and Endocrine disruption models as well as the presence of specific hazardous fragments typical to carcinogens.

Reliability Index (*RI*)

Local part of the model provides the basis for estimating reliability of prediction by the means of calculated Reliability Index (*RI*) values. *RI* is a number ranging from 0 to 1 (0 – unreliable prediction, 1 – idealistic, fully reliable prediction). The following two criteria are applied for reliability estimation:

- Similarity of the analyzed molecule to compounds in the Self-training Library (a reliable prediction cannot be made if no similar compounds have been found in the Library).
- Consistency of model predictions with experimental data for similar compounds (inconsistent experimental outcomes for very similar molecules lead to lower *RI* values).

RI can serve as a valuable tool for interpreting prediction results. If a compound obtains *RI* lower than a certain cut-off value (typically, set at 0.3), it means that this compound falls outside of the Model Applicability Domain, and the respective prediction should be discarded from further analysis regardless of calculated probabilities.

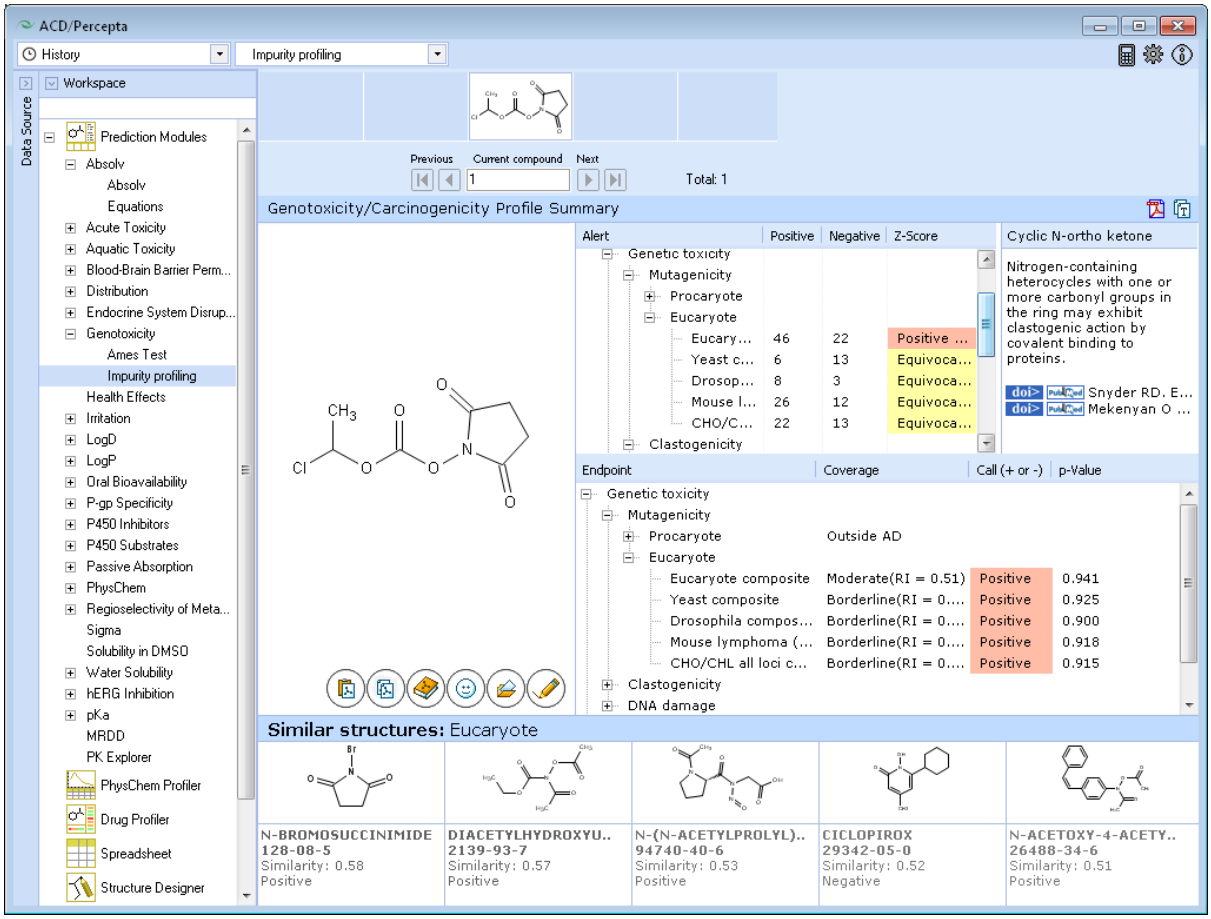


Figure 1. *Screenshot of ACD/Percepta Impurity profiling module*

Model Performance

The predictive performance of Ames test (bacterial composite) mutagenicity model is summarized in Table 2. The accuracy of predictions is very high with almost 90% of the validation set compounds falling within Model Applicability Domain (*RI* ≥ 0.3), and almost 90% of these – classified correctly. A high/moderate reliability index was reported for 65% of compounds in the validation set, and in this case the observed number of misclassifications is even lower.

Table 2. Statistical characteristics of the obtained Ames test classification model.

Accuracy testing		Calculated probability		Statistical parameters		Overall accuracy
		<0.5	≥ 0.5			
Test set (RI ≥ 0.3) 1483 compounds (86.6% covered)	Safe	**392 (26.4%)**	96 (6.5%)	Specificity	80.3%	**89.0%**
	Mutagenic	67 (4.5%)	**928 (62.6%)**	Sensitivity	93.3%	
Test set (RI ≥ 0.5) 1117 compounds (65.2% covered)	Safe	**257 (23.0%)**	51 (4.6%)	Specificity	83.4%	**93.4%**
	Mutagenic	23 (2.0%)	**786 (70.4%)**	Sensitivity	97.2%	

Performance of the derived models for other genotoxicity/carcinogenicity endpoints is given in Table 1 along with the data set sizes.

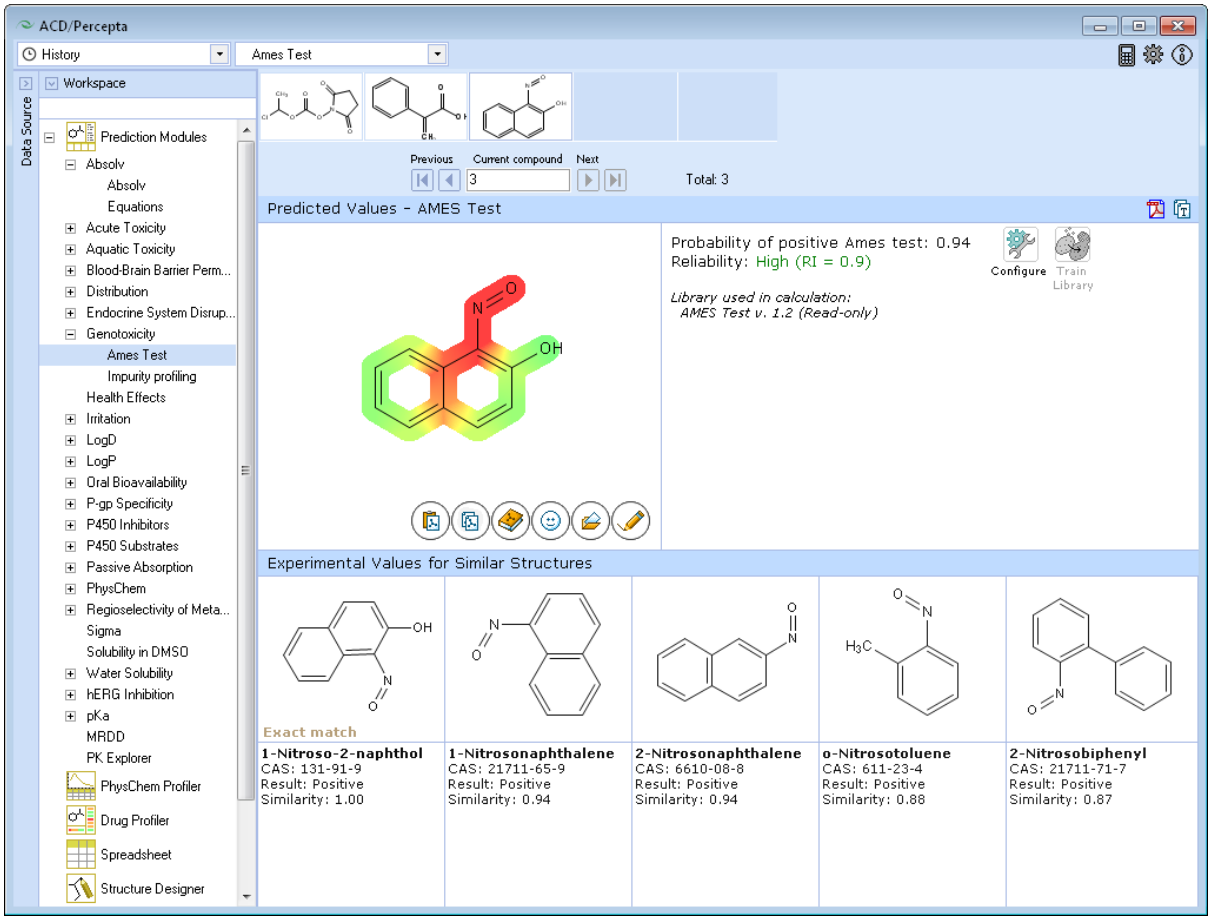


Figure 2. *Screenshot of ACD/Percepta Ames test module*

Improving Prediction Accuracy via Training

The ACD/Genotoxicity Ames Test module also implements the Trainability feature. It addresses the issue of the chemical space of 'in-house' libraries being considerably wider than that of publicly available data which results in limited applicability of most third-party QSARs for analysis of 'in-house' data. The 'Training engine' makes appropriate corrections for systematic deviations produced by the baseline QSAR model based on analysis of similar compounds from the experimental data library.

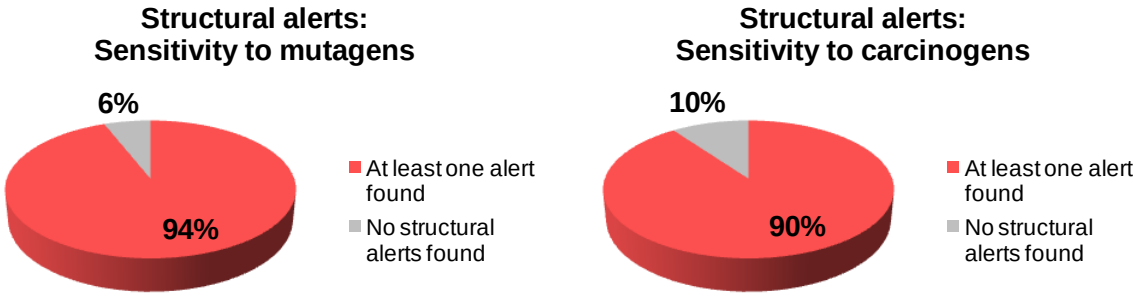
Addition of user-defined experimental data to the model Self-training Library leads to an instant improvement of prediction accuracy for the respective compound classes, therefore avoiding the need for time-consuming rebuilding of the models from scratch when reasonably large amounts of experimental data for new compounds become available.

Genotoxicity/Carcinogenicity Hazards

Hazards is a knowledge-based expert system that identifies and highlights structural moieties that are frequently present in compounds that tested positive in the Ames test, eucaryote gene mutation tests, and chromosomal damage assays, as well as in carcinogens acting by non-genotoxic (epigenetic) mechanisms. Thorough analysis yielded a list of 70 structural alerts, of which 33 represent mutagens, 24–clastogens, and 13–epigenetic carcinogens (androgens, peroxisome proliferators, etc.).

For each hazardous fragment, the software displays additional relevant information, such as its statistical significance, short descriptions of the mechanism of action, and literature references.

Overall, the expert system was able to detect 94% of mutagens in the Ames test DB and 90% of compounds labeled as potent carcinogens by the FDA:



Mutagenicity DB

ACD/Labs Genotoxicity prediction package is supplemented with a fully searchable database that contains over 5500 compounds with experimental Ames Test results, and information about individual studies conducted with each compound. The overall outcome for a compound is reported in the following manner:

- **"+":** positive
- **"–":** negative
- **"(+)":** weakly positive
- **"?"**: inconclusive, discrepant data

Mutagenicity DB includes data corresponding to a variety of bacterial strains tested (*S. typhimurium* TA97, TA98, TA100, TA102, TA104, TA1535, TA1537, TA1538; *E. coli* WP2 uvrA), presence or absence of metabolic activation, as well as other experimental conditions.