

Multi-Label-Classification to predict repeated dose toxicity in the context of REACH

Batke M.¹, Bitsch A.¹, Gundert-Remy U.², **Gütlein M.**³, Helma C.⁴, Kramer S.⁵, Maunz A.⁶, Partosch F.², Seeland M.⁷, Stahlmann R.²

¹ Fraunhofer Institut für Toxikologie und Experimentelle Medizin, Chemikalienbewertung - Hannover, Deutschland

² Charité - Universitätsmedizin Berlin, Institut für Klinische Pharmakologie und Toxikologie - Berlin, Deutschland

³ Albert-Ludwigs-Universität Freiburg, Deutschland, guetlein@informatik.uni-freiburg.de

⁴ in silico toxicology gmbh - Basel, Schweiz

⁵ Johannes Gutenberg - Universität Mainz, Institut für Informatik - Mainz, Deutschland

⁶ Oncotest GmbH - Freiburg, Deutschland

⁷ Technische Universität München, Institut für Informatik - Garching, Deutschland

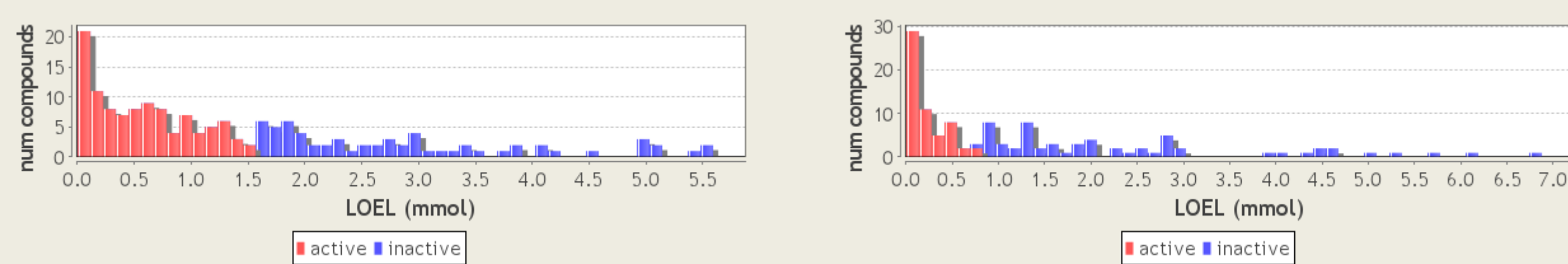
Introduction

- Freely available (Q)SAR models are created for repeated dose toxicity
- Multi-Label-Classification (MLC) is applied to simultaneously predict 28 toxic effects of chemical compounds
- Training data set:
 - 1022 compounds that are tested in over 2000 studies in rats
 - Joined from the RepDose data base and the ELINCS data base
 - Sub-acute and sub-chronic study duration
 - Oral and inhalation application
 - 82% of the endpoint values are missing

Discretization of numeric endpoints

- Modeling the data is challenging due to high in vivo error rate and aggregation of numerous studies
 - Numeric data (LOELs) is converted into nominal data with classes *active* (low values) and *inactive* (high values)
- k-Means clustering is employed to discretize the data:
 - μmol value for sub-chronic data is divided by half
 - Iterative approach with increasing k until a cluster threshold is found between 1.5 and 2.0 μmol
- 49% mean rate of active compounds (minimum: 34%, maximum: 70%)

Discretization for endpoint "red blood cell" into 160 active and 155 inactive compounds sub-acute experiments



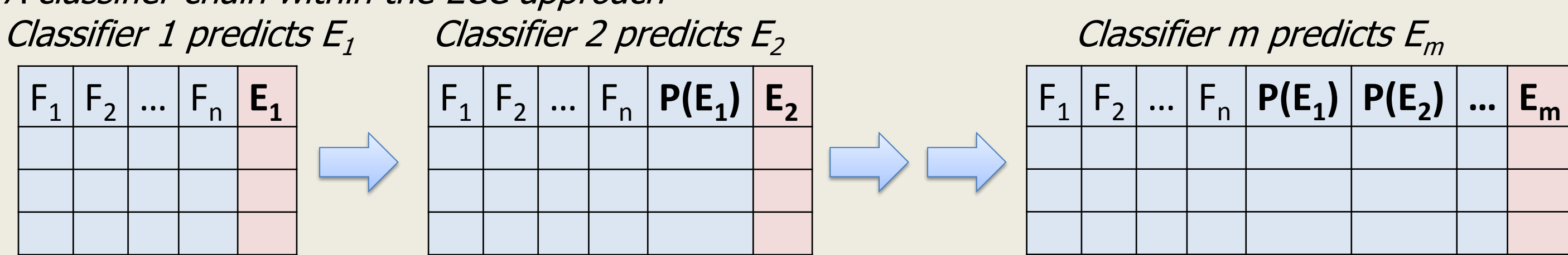
Compound descriptor calculation

- A combined feature set is created with physico-chemical descriptors and structural fragments (622 features in total)
- 243 physico-chemical descriptors computed with freely available libraries (Open Babel and The Chemistry Development Kit (CDK))
- 379 structural fragments found by matching pre-defined lists with SMARTS fragments (included in Open Babel):
 - MACCS keys
 - Inte:Ligand functional groups (Open Babel fingerprint FP4)
 - Checkmol functional groups (Open Babel fingerprint FP3)

MLC with Ensemble of Classifier Chains (ECC)^[1]

- Multi-Label-Classification (MLC) exploits the inter-correlation of classes (i.e. endpoints) and simultaneously predicts multiple classes for one compound
- ECC is a MLC approach that sequentially combines single class classifiers (here: random forest classifier)
- ECC creates multiple chains with randomly ordered endpoints
- Input features for each classifier in the chain: standard descriptors F_1 + predicted endpoints $P(E_i)$ of the preceding classifiers
- Predictions of chains are aggregated via weighted majority vote

A classifier chain within the ECC approach

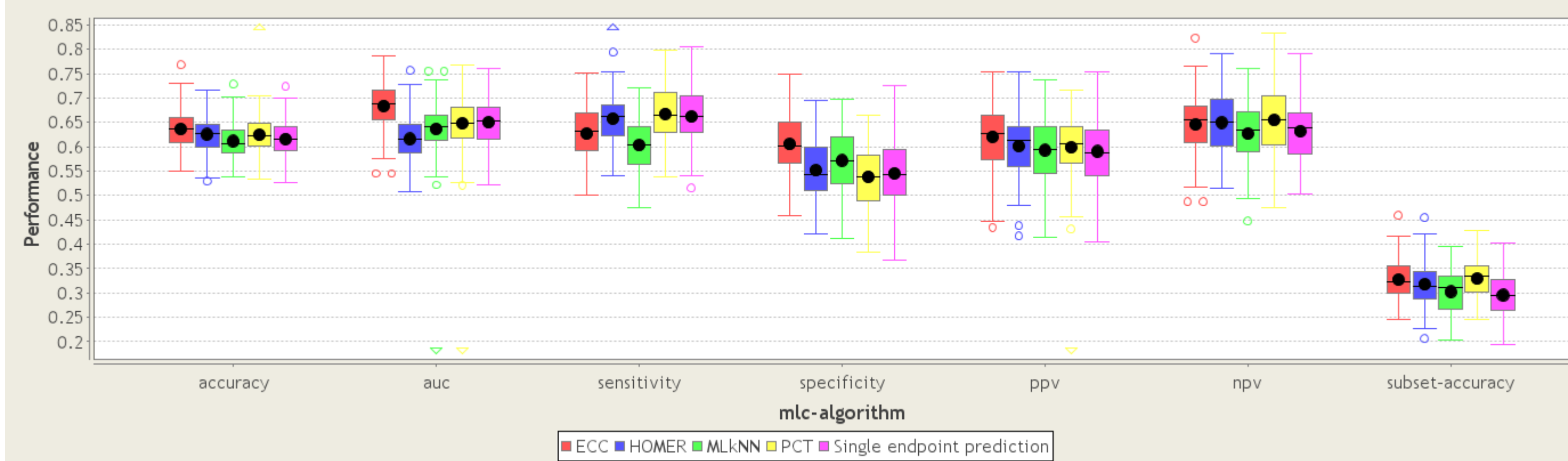


Imputation of missing values

- 82% of endpoint values are missing
- Mean number of compounds per endpoint: 167 (minimum: 44, maximum: 517)
- Important to fill data gaps in order to make efficient use of the endpoint correlations
- Imputation: method to substitute missing values in the training data with predicted values
- ECC has inherent imputation mechanism (predicted endpoint of preceding classifiers in the chain is used)
- Explicit imputation is applied to improve the predictivity of other MLC methods (see validation)

Validation of MLC methods

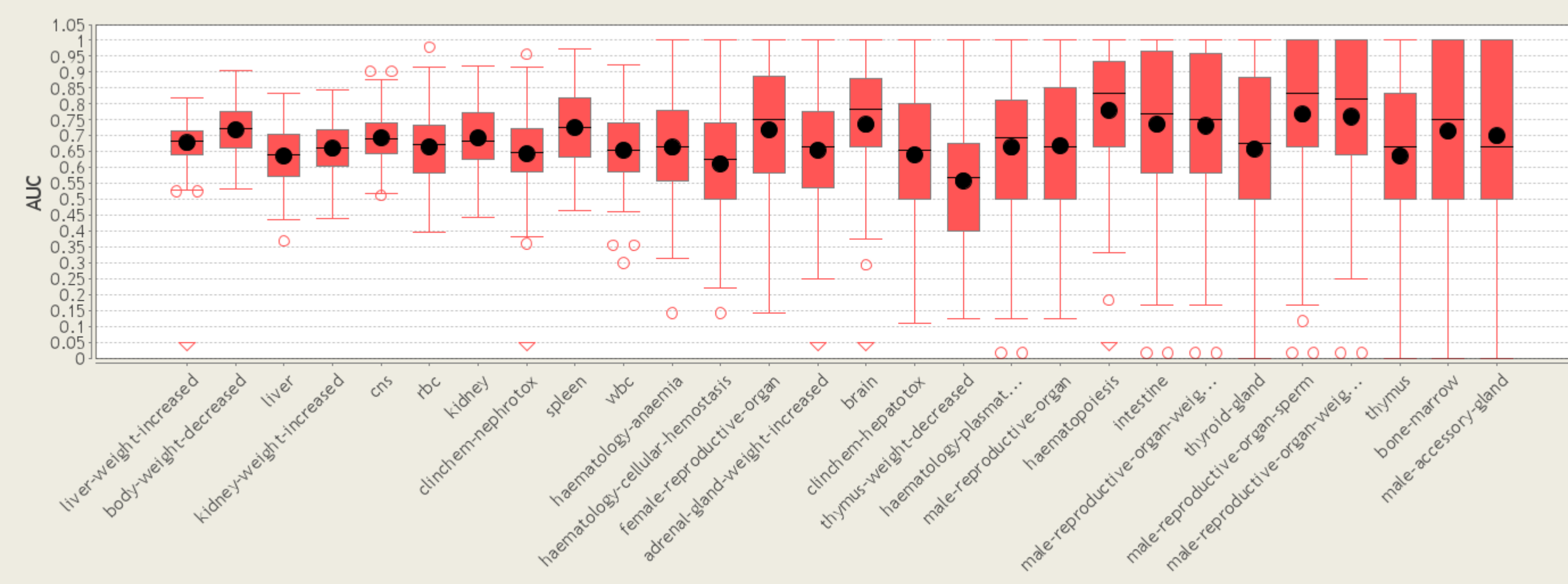
- 10 times repeated 10-fold cross-validation
- Performance of the ECC approach is 68% Area-under-ROC (AUC) (63% sensitivity, 62% selectivity)
- ECC outperforms the following MLC methods (with imputation):
 - Hierarchy Of Multilabel classifiers (HOMER)
 - ML-kNN (Multi-label k-nearest neighbor)
 - Predictive Clustering Trees (PCT)
- Single prediction of each endpoint uses random forests
- Implementation is based on the *mulan* library^[2] (for MLC, extended with imputation and missing value handling) and the *clus* library (for PCT)



Endpoint predictivity

Predictability of endpoints varies from 56% to 78% AUC (using ECC)

Endpoints are sorted according to number of measured compounds



Prediction service

MLC models are published and can be applied to unseen compounds at:

<http://mlc-reach.informatik.uni-mainz.de>

Model: RepDose-Neustoff-ECC

The model predicts various rat toxicity endpoints. The training data set contains repeated dose toxicity data and oral or inhalation application. The RepDose Database contains publicly available repeated dose toxicity data of industrial chemicals. The Neustoff Database comprises new confidential industrial chemicals registered in Europe between 1985 and 2013. The database also contains data of Notified Chemical Substances (ELINCS) which have been tested in subacute or subchronic studies. The Competent Authorities.

Model documentation:

- Technical model description
- Predicted endpoints
- Training dataset compounds
- Validation results

Make compound prediction (Insert one SMILES or InChI per line):

[C]#N1C#CC23C4=C-C-C-C(C3=[O+])C3=C2C=C(C=C14)C#CC3=O][O-]

Predict

Recent performed predictions:

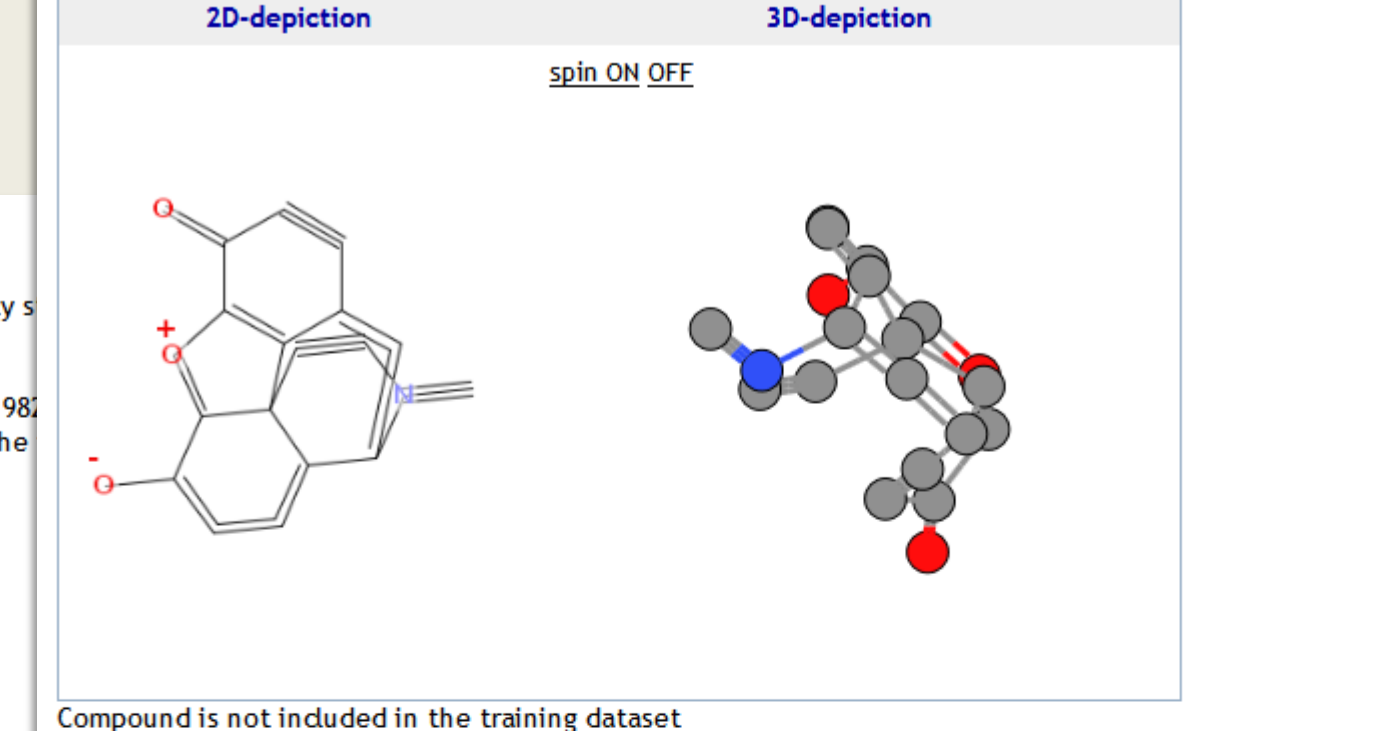
- 2014-02-07 14:17 - [C]#N1C#CC23C4=C-C-C-C(C3=[O+])C3=C2C=C(C=C14)C#CC3=O][O-]
- 2014-01-28 11:31 - CCCCC(c1cc0)c0(c0)c1=O
- 2013-12-06 17:00 - O=C(C=C)O

Compound prediction

Prediction with model RepDose-Neustoff-ECC

[C]#N1C#CC23C4=C-C-C-C(C3=[O+])C3=C2C=C(C=C14)C#CC3=O][O-]

InChI: InChI=1S/C17H19NO3/c:18-7-6-17-10-3-5-13(20)16(17)21-15-12(19)4-2-9(14)15(17)8-11(10)18



Compound is not included in the training dataset

endpoint	predicted activity (with confidence)	prediction correct
liver-weight-increased	inactive (low confidence: 1.7%)	52.5%
body-weight-decreased	active (low confidence: 14.3%)	56%
liver	active (low confidence: 22.6%)	60.7%
kidney-weight-increased	active (low confidence: 1.2%)	57.6%
cms	active (low confidence: 11.7%)	60.6%
rbc	active (low confidence: 2.9%)	55.1%
kidney	active (low confidence: 18.9%)	54.6%
clinchem-nephrotox	active (low confidence: 16.6%)	60.7%
spleen	inactive (low confidence: 29.3%)	62.3%
wbc	active (low confidence: 5%)	58.7%
haematology-anaemia	active (low confidence: 9.7%)	59.2%
haematology-cellular-hemostasis	active (low confidence: 2.9%)	51.7%
female-reproductive-organ	active (low confidence: 7.7%)	58.7%
adrenal-gland-weight-increased	inactive (low confidence: 18%)	49.2%
brain	inactive (low confidence: 12.8%)	67.8%
clinchem-hepatotox	active (low confidence: 1.9%)	61.3%
thymus-weight-decreased	inactive (low confidence: 3.5%)	47.4%
haematology-plasmatic-hemostasis	active (low confidence: 19.3%)	68.8%
male-reproductive-organ	inactive (low confidence: 5.5%)	39%
haematopoiesis	inactive (low confidence: 17.9%)	66.7%
interstine	active (low confidence: 0.2%)	57.5%
male-reproductive-organ-weight-increased	inactive (low confidence: 20.8%)	59.3%
thyroid-gland	active (medium confidence: 39.5%)	71.1%
male-reproductive-organ-sperm	inactive (low confidence: 2%)	66.6%
male-reproductive-organ-weight-decreased	active (low confidence: 1.1%)	66%
thymus	inactive (medium confidence: 34.7%)	67%
bone-marrow	inactive (low confidence: 2.4%)	55.6%
male-accessory-gland	inactive (low confidence: 7.1%)	50.4%

References

- Read, J., Pfahringer, B., Holmes, G., & Frank, E. (2011). Classifier chains for multi-label classification. *Machine learning*, 85(3), 333-359.
- <http://mulan.sourceforge.net>