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

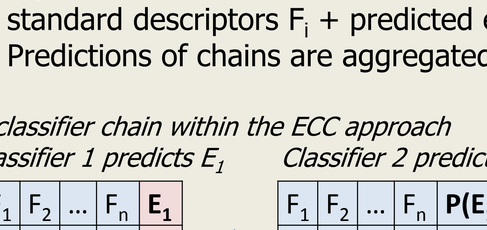
- Freely available (Q)SAR models are created for repeated dose toxicity
- Multi-Label-Classification (MLC) is applied to simultaneously predict 28 toxic effects of chemical compounds
- Training data set:
  - 1022 compounds that are tested in over 2000 studies in rats
  - Joined from the RepDose data 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

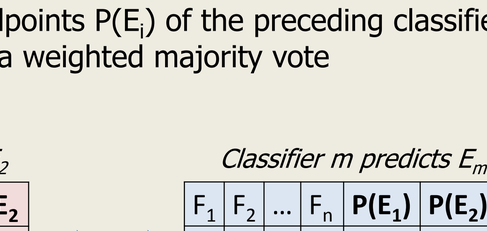
- Modeling the data is challenging due to high in vivo error rate and aggregation of numerous studies
  → Numeric data (LOELs) is converted into nominal data with classes *active* (low values) and *inactive* (high values)
- *k*-Means clustering is employed to discretize the data:
  - µmol value for sub-chronic 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



sub-chronic 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_i + 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

Classifier 1 predicts E_1

F_1	F_2	…	F_n	E_1

Classifier 2 predicts E_2

F_1	F_2	…	F_n	$P(E_1)$	E_2

Classifier m predicts E_m

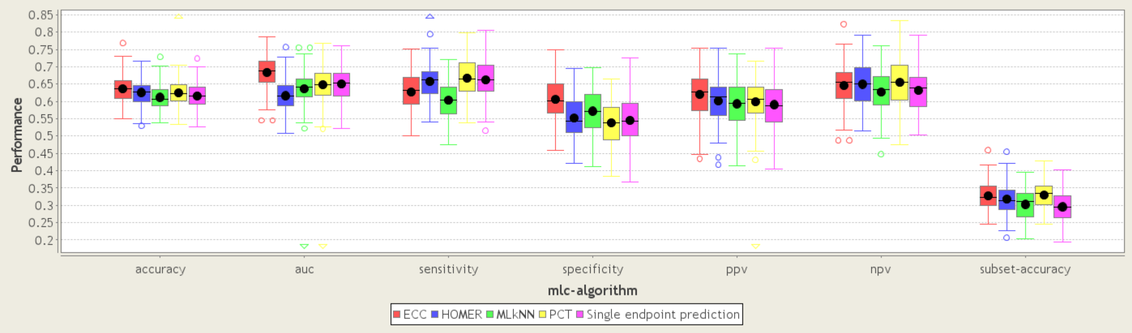
F_1	F_2	…	F_n	$P(E_1)$	$P(E_2)$	…	E_m

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

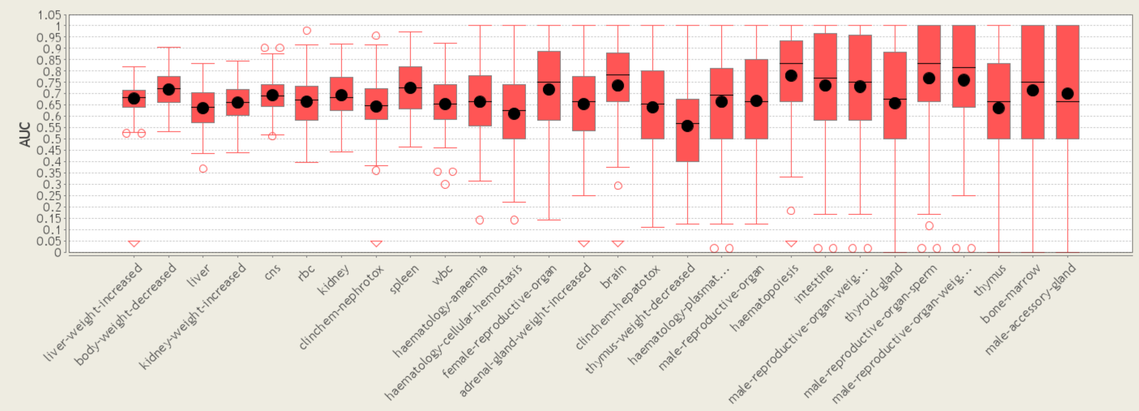
- 10 times repeated 10-fold cross-validation
- Performance of the ECC approach is 68% Area-under-ROC (AUC) (63% sensitivity, 62% selectivity)
- ECC outperforms the following MLC methods (with imputation):
  - Hierarchy Of Multilabel classifiERs (HOMER)
  - ML-kNN (Multi-label k-nearest neighbor)
  - 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

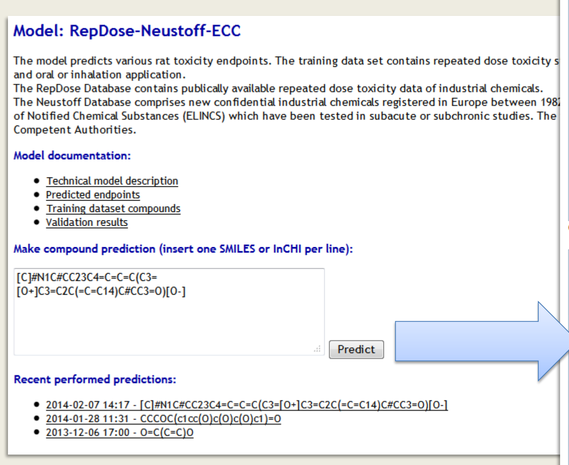
Predictability of endpoints varies from 56% to 78% AUC (using ECC)

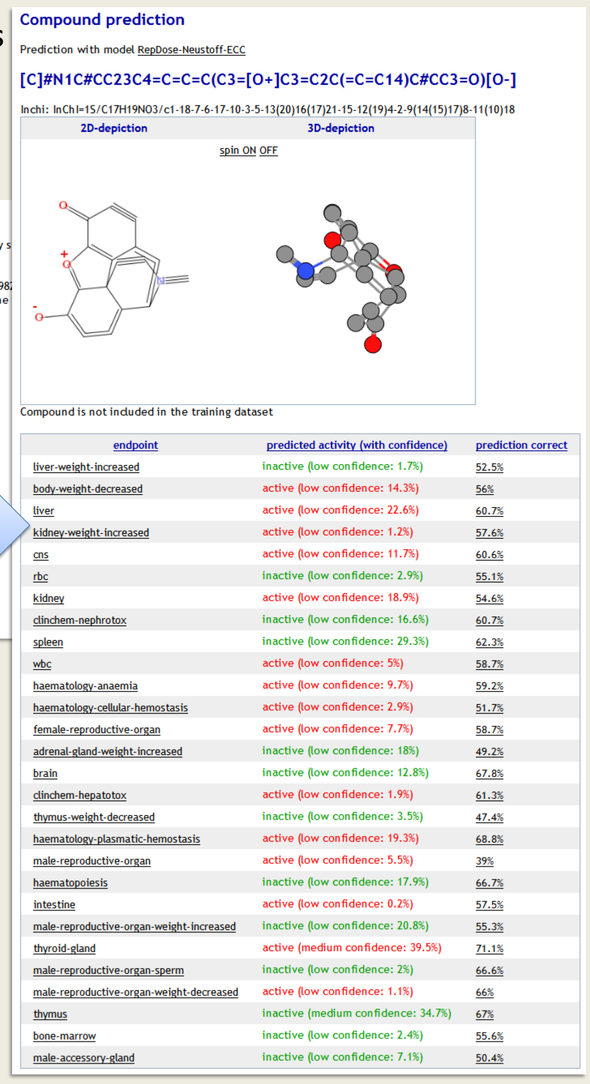
Endpoints are sorted according to number of measured compounds



Prediction service

MLC models are published and can be applied to unseen compounds at:
http://mlc-reach.informatik.uni-mainz.de





References

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