Hierarchical multi-label classification of ToxCast datasets

Nina Jeliazkova nina@acad.bg
Vedrin Jeliazkov vedrin@acad.bg
Ideaconsult Ltd., Sofia, Bulgaria

ToxCast Data Analysis Summit
May 14-15, 2009
US EPA, Research Triangle Park, NC

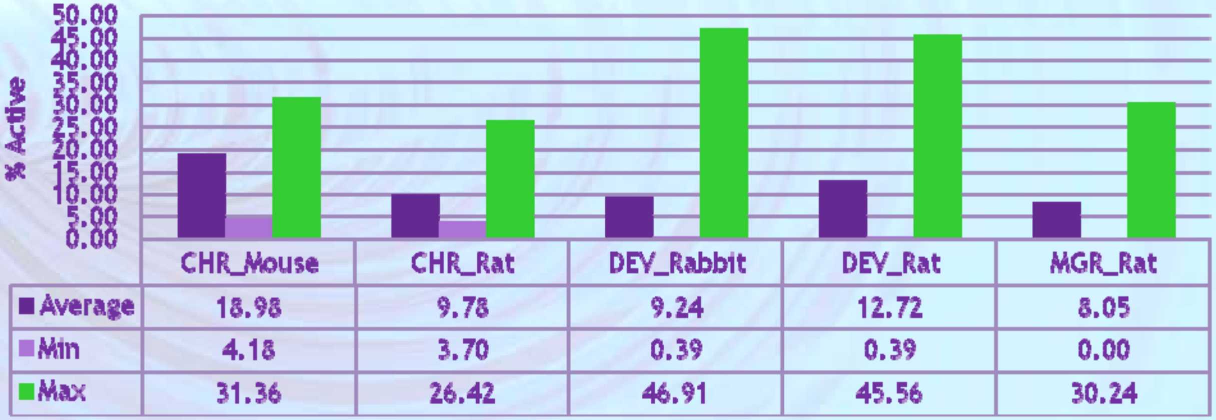
Outline

- Objective
  - Study the possibility to correlate *in-vitro* data with ToxRefDB *in-vivo* test results, using the maximum amount of information available
  - Derive prediction models for *in-vivo* toxicity endpoints
- Data Analysis
- Approach rationale
- Methods and experiments
- Conclusions

Data Analysis (*in-vivo* studies)

- Mixed data type – numerical and nominal
    - Values in mg/kg/day
    - Inactive chemical-assay combinations are (indicated by a value of 1000000)
    - *Preprocessing step : transform the data into nominal (label type) – Active and Inactive*
- Missing data
    - Chemical-assay combinations not tested (indicated by NA)
    - *Preprocessing step : remove missing data (for the sake of simplicity)*

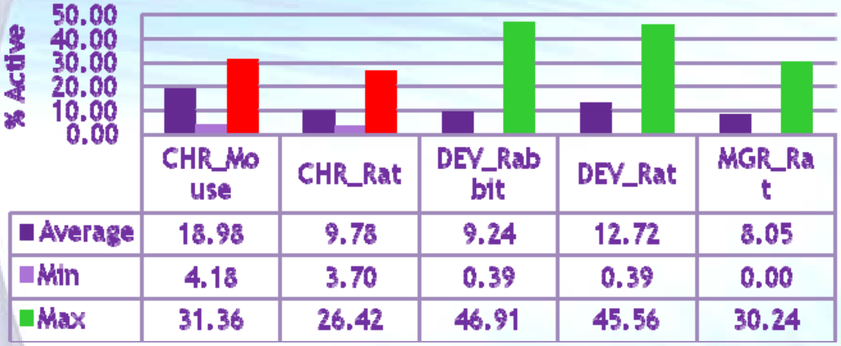
Distribution of Active chemicals - a summary



	CHR_Mouse	CHR_Rat	DEV_Rabbit	DEV_Rat	MGR_Rat
🟪 Average	18.98	9.78	9.24	12.72	8.05
🟪 Min	4.18	3.70	0.39	0.39	0.00
🟩 Max	31.36	26.42	46.91	45.56	30.24

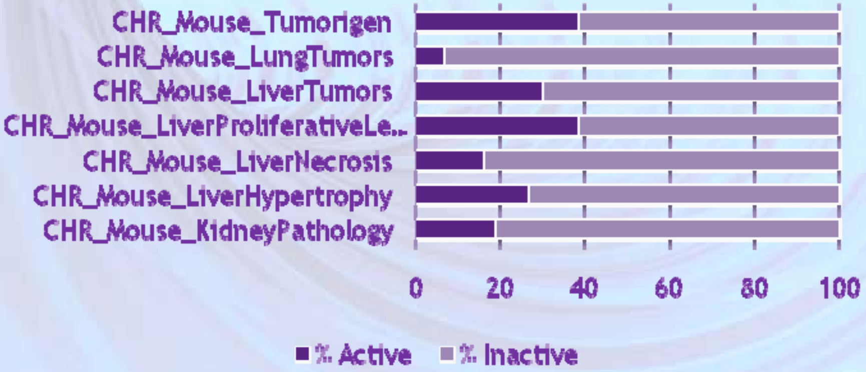
Distribution of *in-vivo* toxicity endpoints

Distribution of Active chemicals - a summary

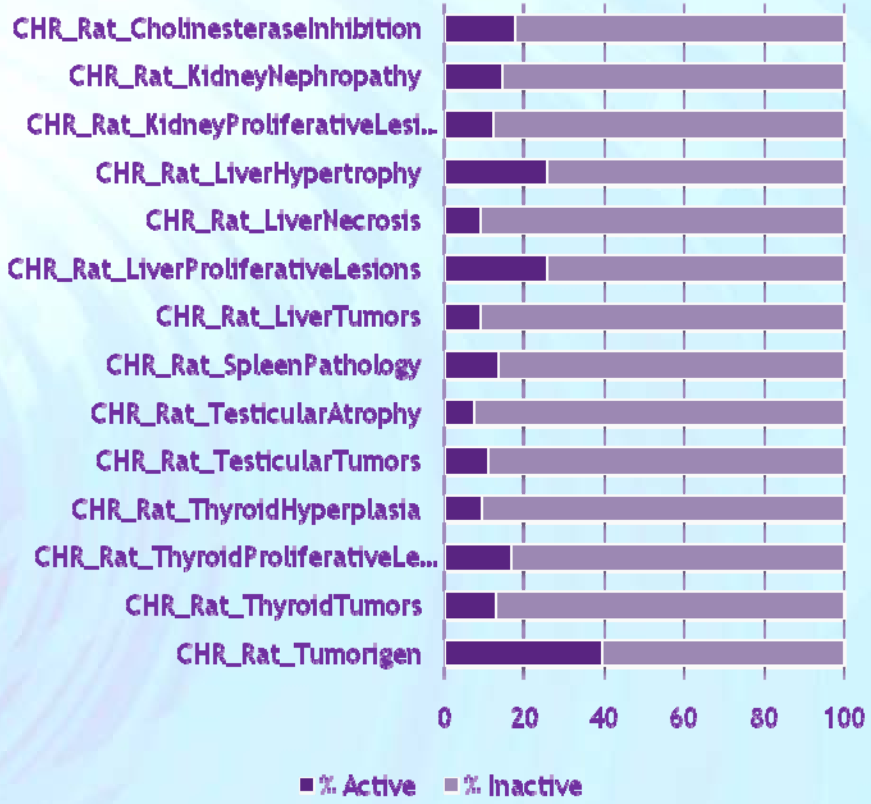


	CHR_Mouse	CHR_Rat	DEV_Rabbit	DEV_Rat	MGR_Rat
■ Average	18.98	9.78	9.24	12.72	8.05
■ Min	4.18	3.70	0.39	0.39	0.00
🟩 Max	31.36	26.42	46.91	45.56	30.24

Chronic toxicity, Mouse

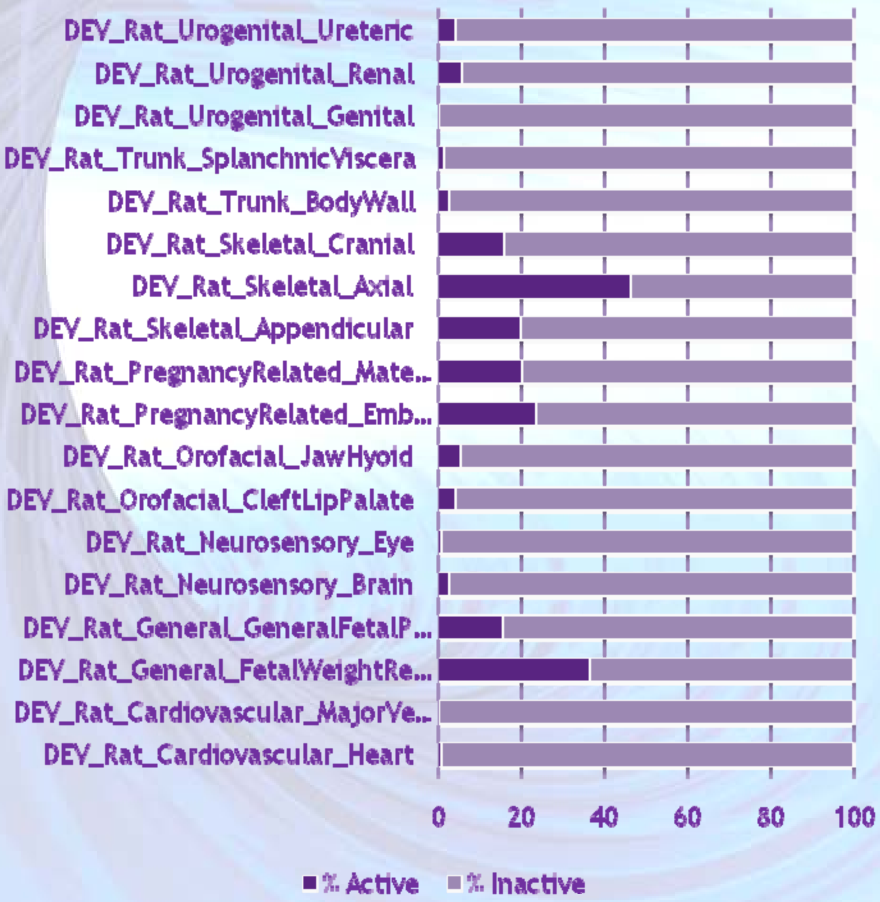


Chronic toxicity, Rat

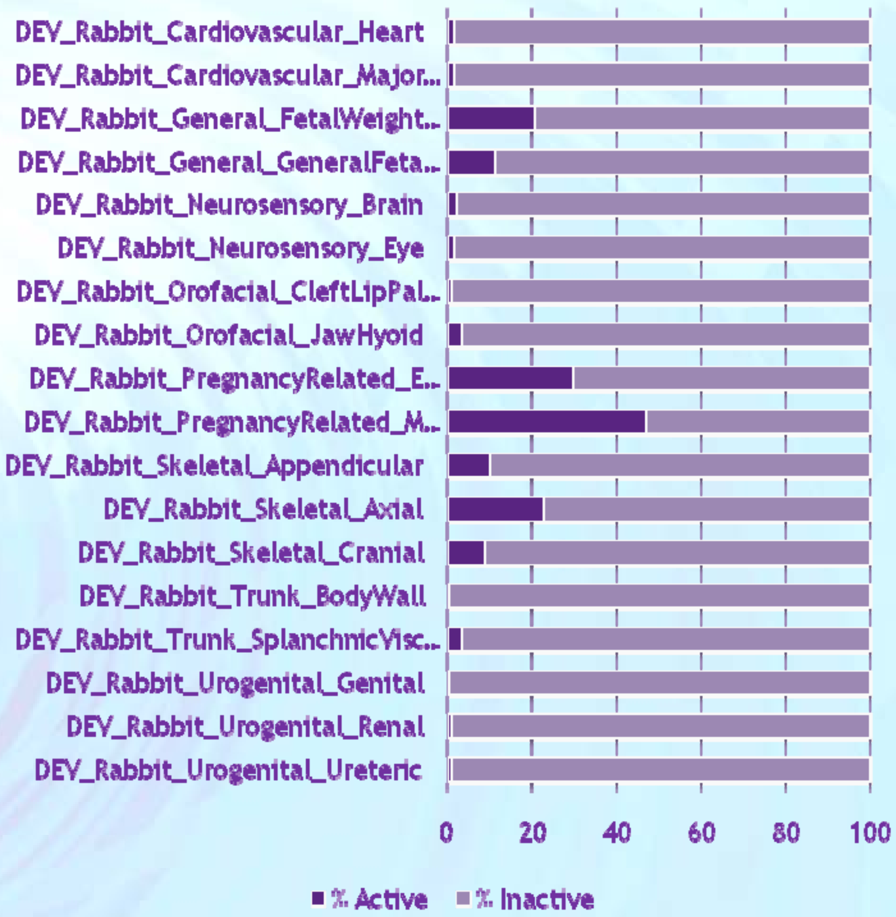


Distribution of *in-vivo* toxicity endpoints

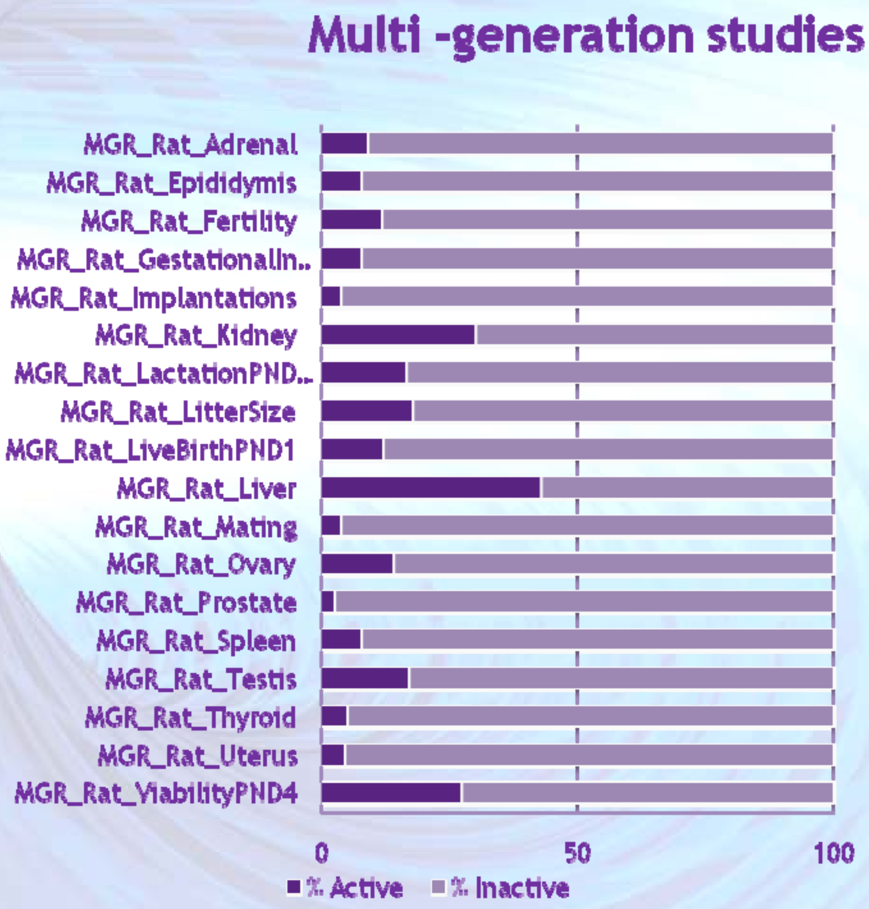
Developmental endpoints, Rat

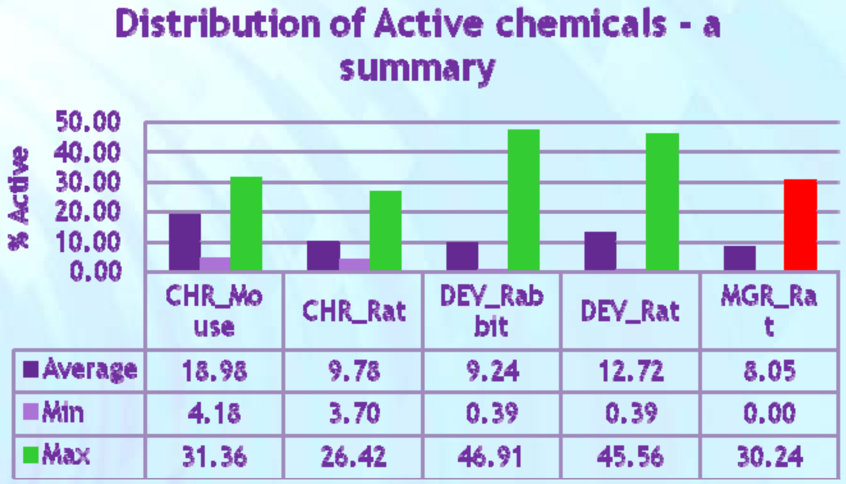


Developmental endpoints, Rabbit



Distribution of *in-vivo* toxicity endpoints


Multi -generation studies


Distribution of Active chemicals - a summary

	CHR_Mouse	CHR_Rat	DEV_Rabbit	DEV_Rat	MGR_Rat
■ Average	18.98	9.78	9.24	12.72	8.05
■ Min	4.18	3.70	0.39	0.39	0.00
■ Max	31.36	26.42	46.91	45.56	30.24

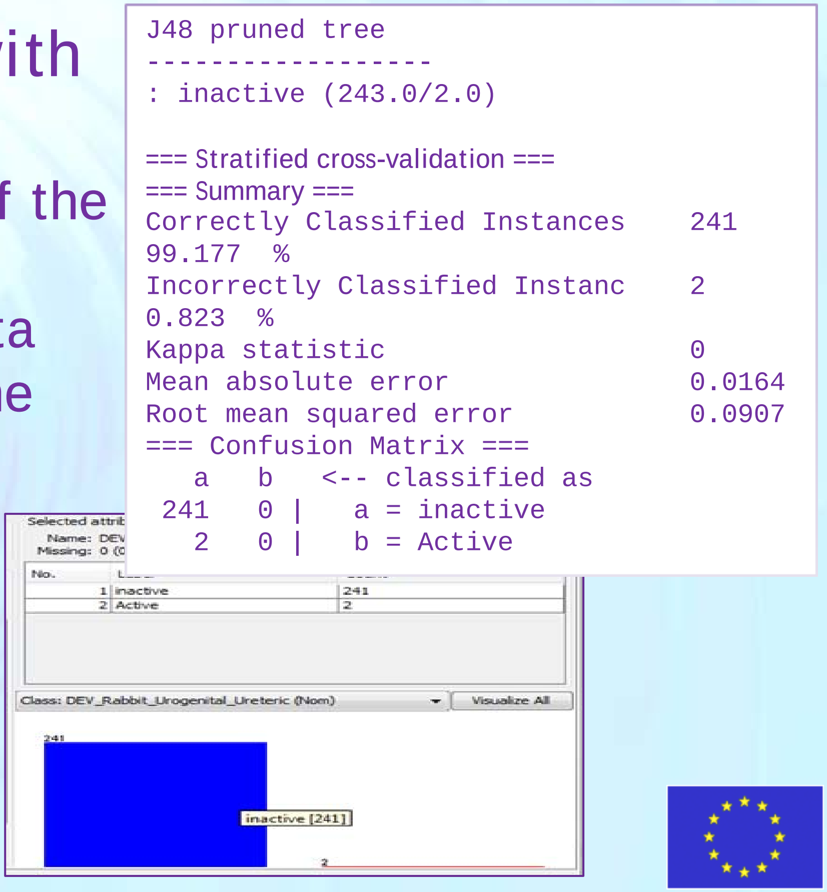
OpenTox

Data Analysis findings (*in-vivo* studies)

- The Active/Inactive classes in ToxCast *in-vivo* data are highly **unbalanced**
    - **This is a potential problem for almost all learning algorithms**

- **Any classification algorithm**, with the
    - Objective to maximize accuracy of the prediction
    - Under the assumption that the data distribution of the training set is the same as the future data

...will hardly be able to improve the predictions over the trivial classifier <u>"all data is from the majority class"</u>



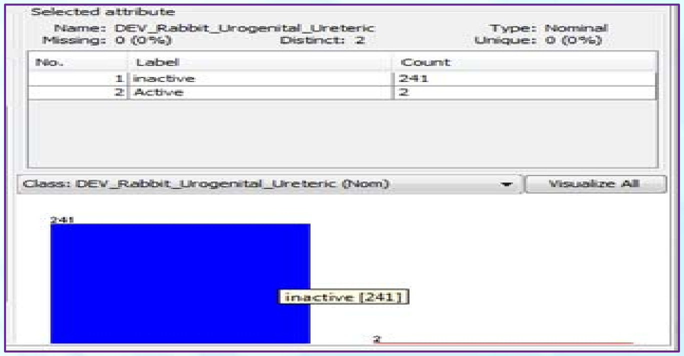
```
J48 pruned tree
------------------
: inactive (243.0/2.0)

=== Stratified cross-validation ===
=== Summary ===
Correctly Classified Instances    241
99.177  %
Incorrectly Classified Instanc    2
0.823  %
Kappa statistic                   0
Mean absolute error               0.0164
Root mean squared error           0.0907
=== Confusion Matrix ===
   a    b   <-- classified as
 241    0 |    a = inactive
   2    0 |    b = Active
```

No.	Label	Count
1	inactive	241
2	Active	2

Unbalanced classes – existing approaches

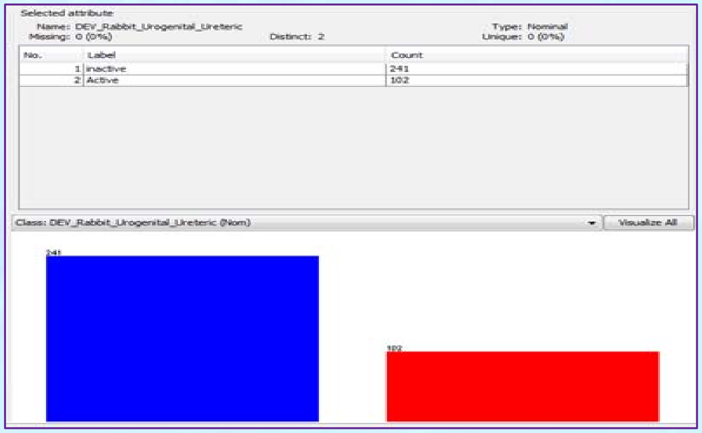
- Modify the balance:
  - Down sampling
    - Throw away data from the majority class
  - Over sampling
    - Add new points to the minority class (copy of the existing or new artificial ones)
- Modify the learning algorithms to treat classes differently
  - E.g. Cost sensitive classification



Selected attribute

Name: DEV_Rabbit_Urogenital_Ureteric Type: Nominal
Missing: 0 (0%) Distinct: 2 Unique: 0 (0%)

No.	Label	Count
1	inactive	241
2	Active	2



Selected attribute

Name: DEV_Rabbit_Urogenital_Ureteric Type: Nominal
Missing: 0 (0%) Distinct: 2 Unique: 0 (0%)

No.	Label	Count
1	inactive	241
2	Active	102

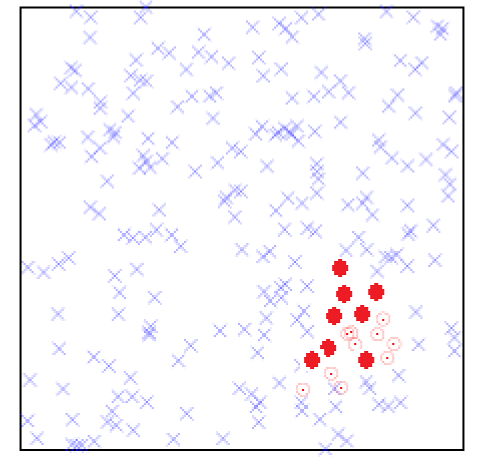
SMOTE - Synthetic Minority Oversampling Technique

Chawla, N.V., Bowyer, K.W., Hall, L.O., Kegelmeyer, W.P. (2002) "SMOTE: Synthetic Minority Over-sampling Technique",

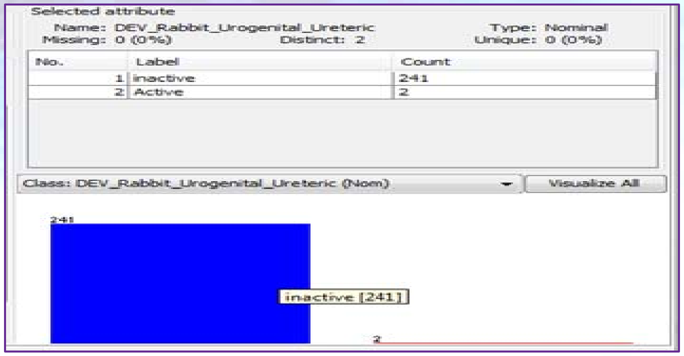
Journal of Artificial Intelligence Research, Volume 16, pages 321-357.

- Generalizes the decision region for the minority class
- Generates new random instances of a given class, based on nearest neighbours
- Recognised as one of the best techniques
- Several extensions available
- Open source WEKA implementation available
- Disadvantages
  - Danger of over-generalization
  - Number of artificial examples fixed

- The algorithm:
  - For each minority data point A find 5 nearest minority class points
  - Randomly choose an example B out of the 5 closest points
  - Calculate the distance D between randomly chosen point and the current point
  - Randomly generate a number less than D
  - Generate a new data point C, such that it lies on the line between A and B and the distance between A and C is D.



The classification example revisited

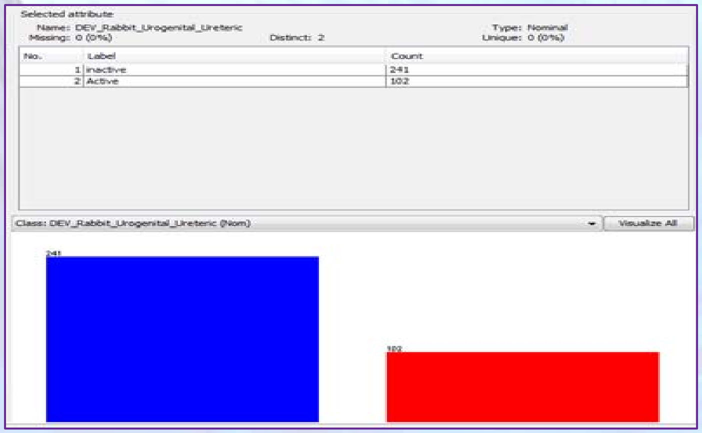


Selected attribute

Name: DEV_Rabbit_Urogenital_Ureteric
Missing: 0 (0%)
Distinct: 2
Type: Nominal
Unique: 0 (0%)

No.	Label	Count
1	inactive	241
2	Active	2

Class: DEV_Rabbit_Urogenital_Ureteric (Nom)



Selected attribute

Name: DEV_Rabbit_Urogenital_Ureteric
Missing: 0 (0%)
Distinct: 2
Type: Nominal
Unique: 0 (0%)

No.	Label	Count
1	inactive	241
2	Active	102

Class: DEV_Rabbit_Urogenital_Ureteric (Nom)

```
J48 pruned tree
------------------
NVS_ENZ_rCNOS <= 945971.200291
|   ATG_DR5_CIS <= 32.321051: Active (97.0)
|   ATG_DR5_CIS > 32.321051
|   |   BSK_BE3C_hLADR <= 40: inactive (2.0)
|   |   BSK_BE3C_hLADR > 40: Active (4.0)
NVS_ENZ_rCNOS > 945971.200291: inactive (240.0/1.0)
=== Stratified cross-validation ===
=== Summary ===
Correctly Classified Instances        339              98.8338 %
Incorrectly Classified Instances        4               1.1662 %
Kappa statistic                      0.9721
Mean absolute error                      0.0152
=== Confusion Matrix ===
  a   b   <-- classified as
 239   2 |   a = inactive
   2 100 |   b = Active
```

Building a prediction model

- Selection of study and endpoint.
  - Chronic toxicity
    - Mouse, Rat
  - Developmental toxicity
    - Rabbit, Rat
  - Multi-generation toxicity
- Selection of a single endpoint in a classic setup effectively ignores the information about other endpoints within the same study.
  - Is it possible to use all study information available?

Multi-label classification

- Classic (single-label)
  - Classes are mutually exlcusive (e.g. Active vs. Inactive)
- Fuzzy classification
  - An instance can be member of several classes, with some probability or degree of uncertainty
- Multi-label classification
  - An instance can be a full member of multiple classes
  - Typical for many domains:
    - Text documents classification
    - Scene recognition
    - Medical diagnosis
    - Toxicology ???

- Single label
  - CHR_Mouse_Tumorigen
    - Yes / No
- Fuzzy
  - CHR_Rat_CholinesteraseInhibition
    - P = 0.5
  - CHR_Rat_LiverTumors
    - P= 0.2
  - CHR_Rat_KidneyProliferativeLesions
    - P = 0.3
- Multi-label

Label/ Chemical	Cholin-esterase Inhibition	Liver Tumors	Kidney Proliferative Lesions
Chemical 1	Yes	No	Yes
Chemical 2	Yes	Yes	Yes
Chemical 3	No	No	No

Building a model

- Data analysis - predictors (*in-vitro* data)
  - High dimensional (>500 columns)
  - Mixed data type (numeric for active chemicals, nominal for inactive)
  - No missing data values
- Approach
  - Unsupervised
    - Clustering
  - Supervised
    - Classification or Regression ?

- Why not Classification by Clustering
  - Predictive Clustering Trees

  Blockeel et al. Top-down induction of clustering trees. In Proc. of the 15th ICML, p.55-63, 1998

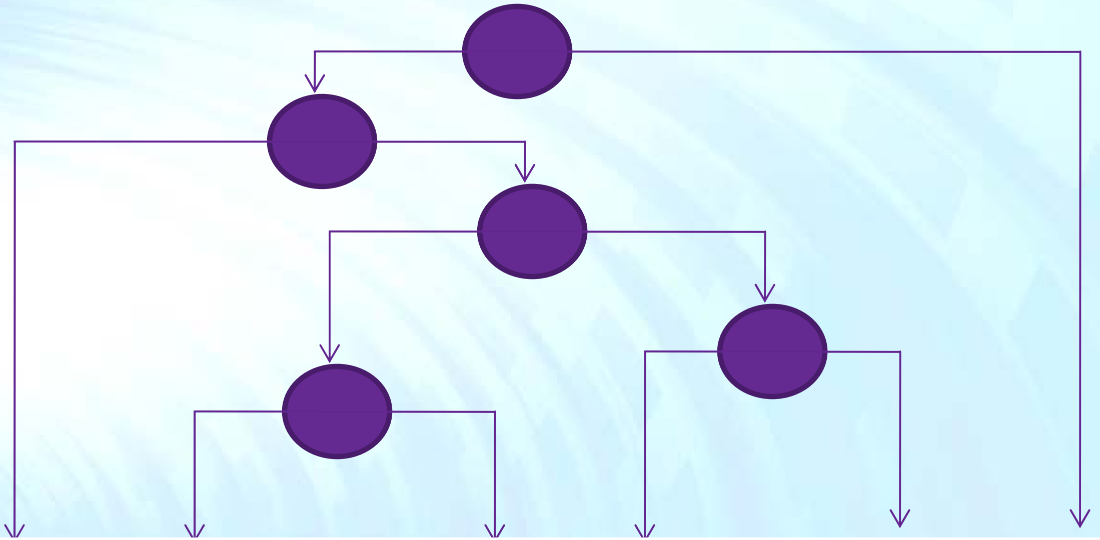
- Decision tree approach advantage
  - When unclear where to start use a Decision Tree ☺
  - Fast
  - Interpretable
  - Relevant features identified during the build process – we can skip the feature selection step

Predictive Clustering Trees

Blockeel et al. Top-down induction of clustering trees. In Proc. of the 15th ICML, p.55-63, 1998

- The decision tree is hierarchy of clusters
- The top node corresponds to the cluster, containing all data, which is recursively partitioned into smaller clusters down the tree.
- The split at each node is selected with the objective to maximize reduction of intra-cluster variance, thus maximizing cluster homogeneity and improving predictive performance
- The variance is treated as a parameter (can be defined in different ways), resulting in (multi-label) classification trees or regression trees as special cases.
- If no test significantly reduce the variance, a leaf is created and is labeled with a prototype (representative instance)
- The prototype is also treated as a parameter and can have various definitions.

Predictive Clustering Trees



Chemical	Cluster1	Cluster2	Cluster3	Cluster4	Cluster5	Cluster6
CHR_Rat_CholinesteraseInhibition	204	11	14	100	0	0
CHR_Rat_KidneyNephropathy	189	14	17	1	50	0
CHR_Rat_KidneyNephropathy	195	16	15	1	100	30

Hierarchical Predictive Clustering Trees

- Predictive Clustering Tree with a special definition of variance

(mean squared distance between each instance label to the set mean label)

Class weights w(c) decrease with the hierarchy

- Advantages vs. separate trees for prediction of multiple classes
  - Identify features with high relevance for multiple classes
  - Hierarchy constraints
  - More efficient
  - Simpler models, if the classes are not independent
  - Learning from skewed distribution

$$\mathrm{Var}(S) = \frac{\sum_i d(v_i, \overline{v})^2}{|S|}$$

$$d(v_1, v_2) = \sqrt{\sum_i w(c_i) \cdot (v_{1,i} - v_{2,i})^2},$$

$$w(c) = w_0^{\mathrm{depth}(c)}$$

$$0 < w_0 < 1$$

<u>Example hierarchy:</u>

1. Top class

1.1. Subclass1

1.2. Subclass 2

Class vector

[1., 1.1., 1.2.]

Class membership

[1, 0, 1]

Why hierarchical classification

- The classes form a hierarchy, i.e. A partial order needs to be defined, such that class C1< C when class C1 is a super-class of C
- ToxCast *in-vivo* toxicity endpoints are obviously related
- Different domain specific relationships can be defined

1.Target
1.1.Liver
1.1.1. Proliferative lesions
1.1.1.1 Neoplasms
1.2. Kidney
1.2.1. Proliferative lesions
1.2.1.1 Neoplasms
1.2.2. Non-proliferative lesions

1.Pathology
1.1. Proliferative
1.1.1. Neoplasms
1.1.1.1.Rat
1.1.1.2. Mouse
1.1.2. Non-neoplastic
1.1.2.1.Rat
1.1.2.2. Mouse
1.2. Non-Proliferative
1.2.1. Rat
1.2.2. Mouse

Open source Implementation - Clus

- Clus is a decision tree and rule induction system that implements the predictive clustering framework. This framework unifies unsupervised clustering and predictive modeling and allows for a natural extension to more complex prediction settings such as multi-task learning and multi-label classification.
- Clus is co-developed by the Declarative Languages and Artificial Intelligence group of the Katholieke Universiteit Leuven, Belgium, and the Department of Knowledge Technologies of the Jožef Stefan Institute, Ljubljana, Slovenia.
- Clus is free software (licensed under the GPL) and can be obtained from
  - http://www.cs.kuleuven.be/~dtai/clus/

Experiments (1)

- Generate single-label prediction models
  - For all *in-vivo* endpoints, available in ToxCast,
  - Build a Predictive Clustering Tree, using *in-vivo* data as predictors
- Generate multi-label prediction models
  - For each study/species combination
    - Generate combinations of 2 and 3 endpoints
  - Build a Predictive Clustering Tree, using *in-vitro* data as predictors

Experiments (2 – balancing via SMOTE)

- Generate single-label prediction models
  - For all *in-vivo* endpoints, available in ToxCast,
  - Apply SMOTE as a pre-processing step
  - Build a Predictive Clustering Tree, using *in-vivo* data as predictors
- Generate multi-label prediction models
  - For all *in-vivo* endpoints, available in ToxCast
    - Generate combinations of 2 and 3 endpoints
  - Apply SMOTE as a pre-processing step
  - Build a Predictive Clustering Tree, using *in-vitro* data as predictors

Performance assessment

- Accuracy is not a good metric!
- Data mining terms
  - Precision = TP / (TP + FP)
  - Recall = TP / (TP + FN) = Sensitivity
- Toxicology terms
  - Sensitivity = TP / (TP + FN) = Recall
  - Specificity = TN/ (TN + FP)
- Receiver Operating Characteristic (ROC)
- Precision – Recall curve (PRC)

Experiments (3 - hierarchical)

- Generate hierarchical multi-label prediction models
  - Specify an hierarchy
  - Match the hierarchy to a ToxRefDB assay
  - Build a decision tree, using *in-vitro* data as predictors
- Performance
  - Precision – Recall curve (PRC)
  - Hierarchical accuracy

An example 4-label tree

```
NVS_NR_hAR > 5.51
+--yes: [inactive,inactive,inactive,inactive] [227.0,202.0,217.0,182.0]
+--no:  ATG_SREBP_CIS > 10.89
        +--yes: [inactive,inactive,inactive,Active] [4.0,4.0,2.0,4.0]
        +--no:  [Active,Active,Active,Active] [5.0,5.0,4.0,5.0]
```

[MGR_Rat_GestationalInterval, MGR_Rat_LitterSize, MGR_Rat_LiveBirthPND1, MGR_Rat_ViabilityPND4]

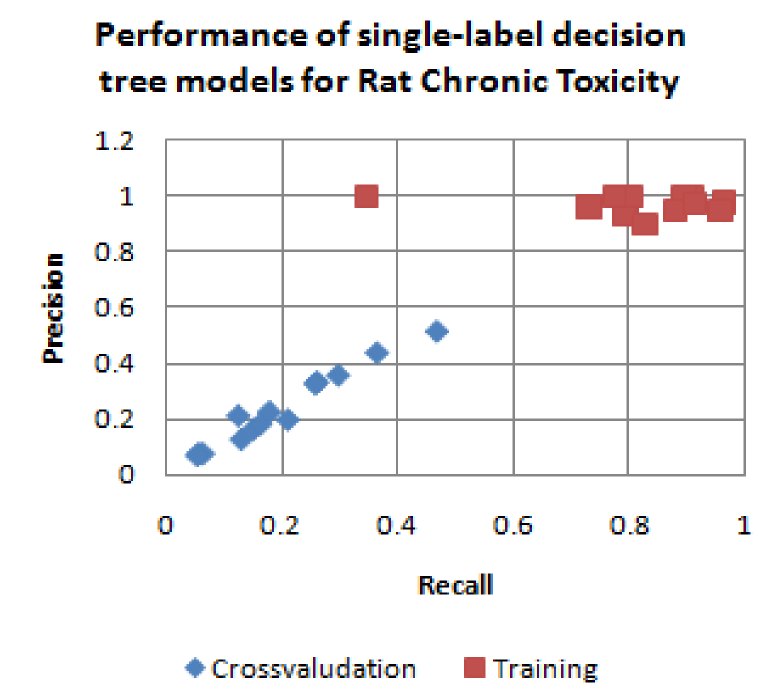
```
R_Rat_GestationalInterval
 REAL\PRED | Active | inactive |
 -----------------------------------
    Active |      5 |       15 |  20
  inactive |      0 |      231 | 231
 -----------------------------------
           |      5 |      246 | 251
 Accuracy: 0.940239
 Cramer's coefficient: 0.484516
```

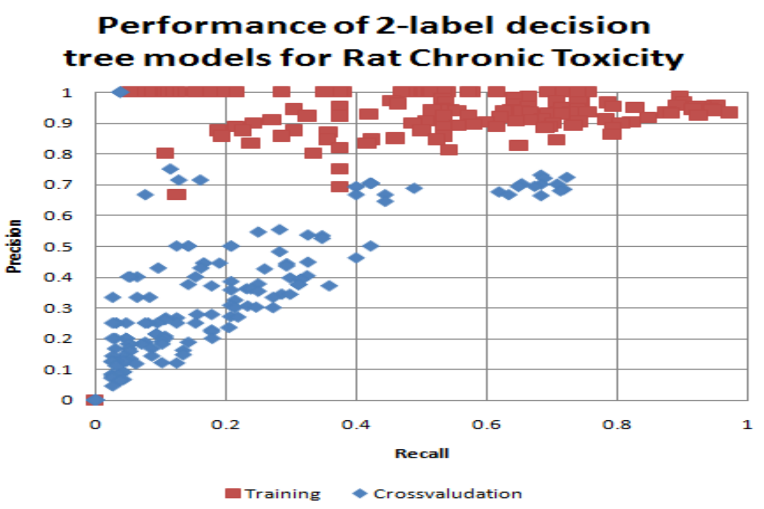
```
: MGR_Rat_LitterSize
PRED | Active | inactive |
----------------------------
ve |      5 |       40 |  45
ive |      0 |      206 | 206
----------------------------
   |      5 |      246 | 251
acy: 0.840637
Cramer's coefficient: 0.305032
```

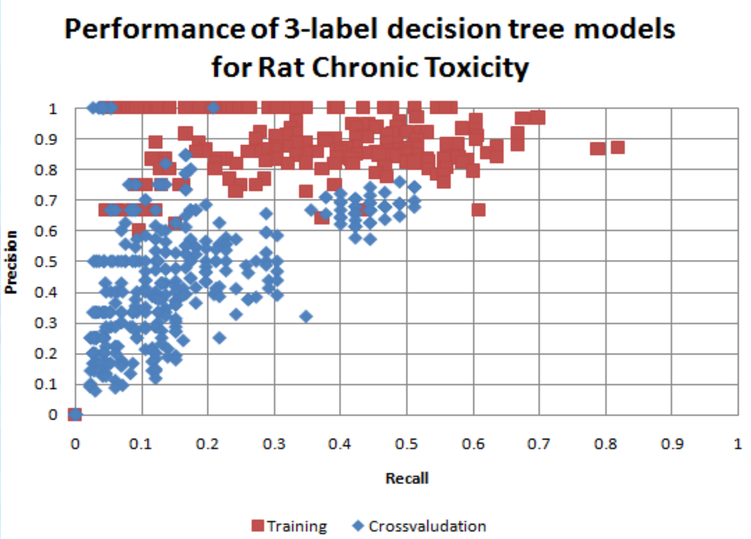
```
MGR_Rat_LiveBirthPND1
    REAL\PRED | inactive | Active |
    -------------------------------
tive |      219 |      1 | 220
tive |       27 |      4 |  31
    -------------------------------
     |      246 |      5 | 251
racy: 0.888446
ner's coefficient: 0.293132
```

```
MGR_Rat_ViabilityPND4
    REAL\PRED | Active | inactive
|
    -----------------------------------
       Active |      9 |       60 |  69
     inactive |      0 |      182 |
182
    -----------------------------------
              |      9 |      242 | 251
    Accuracy: 0.760956
    Cramer's coefficient:
0.313202
```

Rat Chronic/Cancer Toxicity models performance

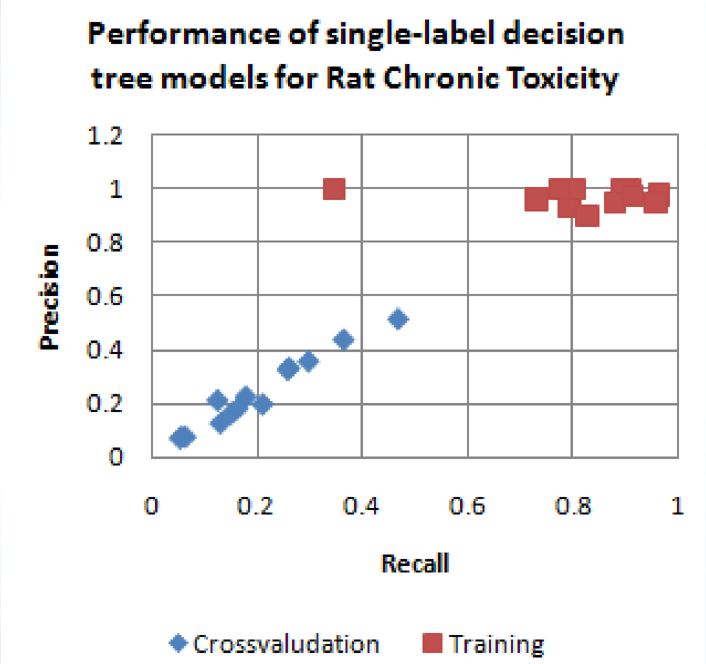


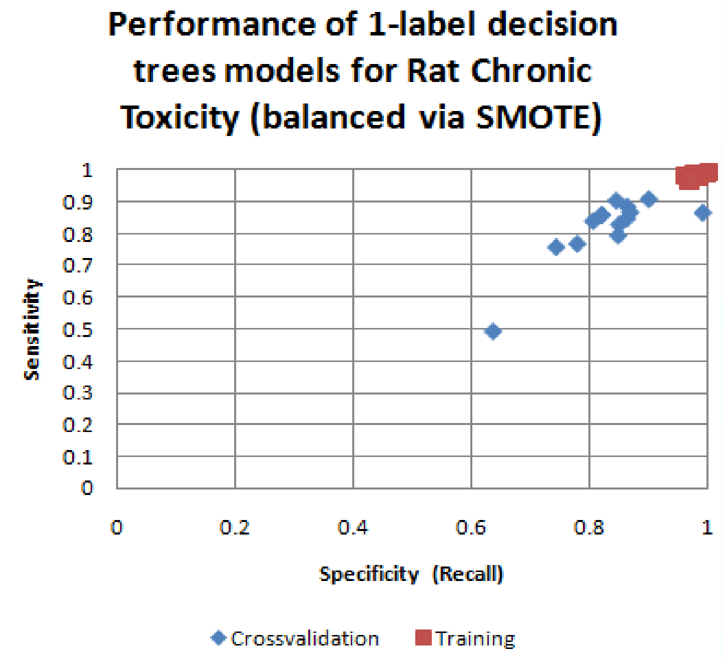




Multi-label trees perform better on average, compared to the single-label tree

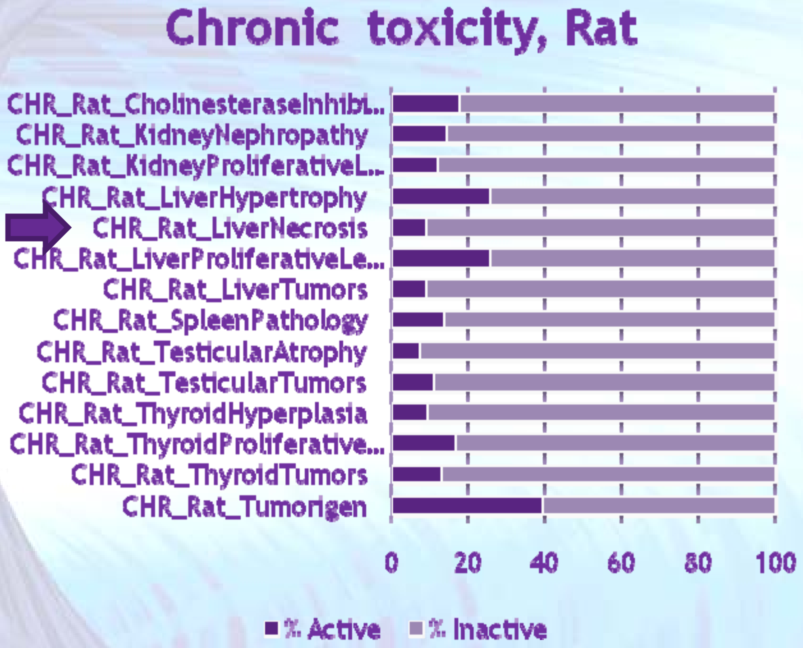
Rat Chronic/Cancer Toxicity models performance (balancing via SMOTE)





Excellent performance with cross-validation!

Example: CHR_Rat_LiverNecrosis model (balancing via SMOTE)



10 fold crossvalidation performance performance:

Real\Predicted	Active	Inactive
Active	92.27%	7.75%
Inactive	9.83%	90.17%

Similar results for other endpoints

```
BSK_3C_MCP1 > 991125.999331
+--yes: ATG_PPARd_TRANS > 947562.256956
|       +--yes: ATG_RARa_TRANS > 6.24888
|       |       +--yes: ATG_M_06_CIS > 100.0
|       |       |       +--yes: [inactive] [153.0]
|       |       |       +--no:  CLZD_CYP2B6_24 > 22.136744
|       |       |               +--yes: [Active] [2.0]
|       |       |               +--no:  [inactive] [3.0]
|       |       +--no:  CLM_StressKinase_1hr > 164.7
|       |               +--yes: [inactive] [5.0]
|       |               +--no:  [Active] [4.0]
|       +--no:  ATG_SREBP_CIS > 33.0
|               +--yes: BSK_LPS_TNFa > 13.333333
|               |       +--yes: [Active] [37.0]
|               |       +--no:  [inactive] [4.0]
|               +--no:  [inactive] [15.0]
+--no:  ATG_Ahr_CIS > 100.0
        +--yes: BSK_3C_IL8 > 40.0
        |       +--yes: BSK_BE3C_IL1a > 40.0
        |       |       +--yes: BSK_KF3CT_TGFb1 > 40.0
        |       |       |       +--yes: BSK_hDFCGF_EGFR > 40.0
        |       |       |       |       +--yes: [Active] [149.0]: 149
        |       |       |       |       +--no:  BSK_hDFCGF_MMP1 > 40.0
        |       |       |       |               +--yes: [inactive] [3.0]
        |       |       |       |               +--no:  [Active] [6.0]
        |       |       |       +--no:  [inactive] [3.0]
        |       |       +--no:  ACEA_LOC3 > 134474.22692
        |       |               +--yes: [inactive] [7.0]
        |       |               +--no:  [Active] [6.0]
        |       +--no:  ACEA_LOC2 > 33.113112
        |               +--yes: [Active] [3.0]
        |               +--no:  [inactive] [11.0]
        +--no:  [inactive] [29.0]
```

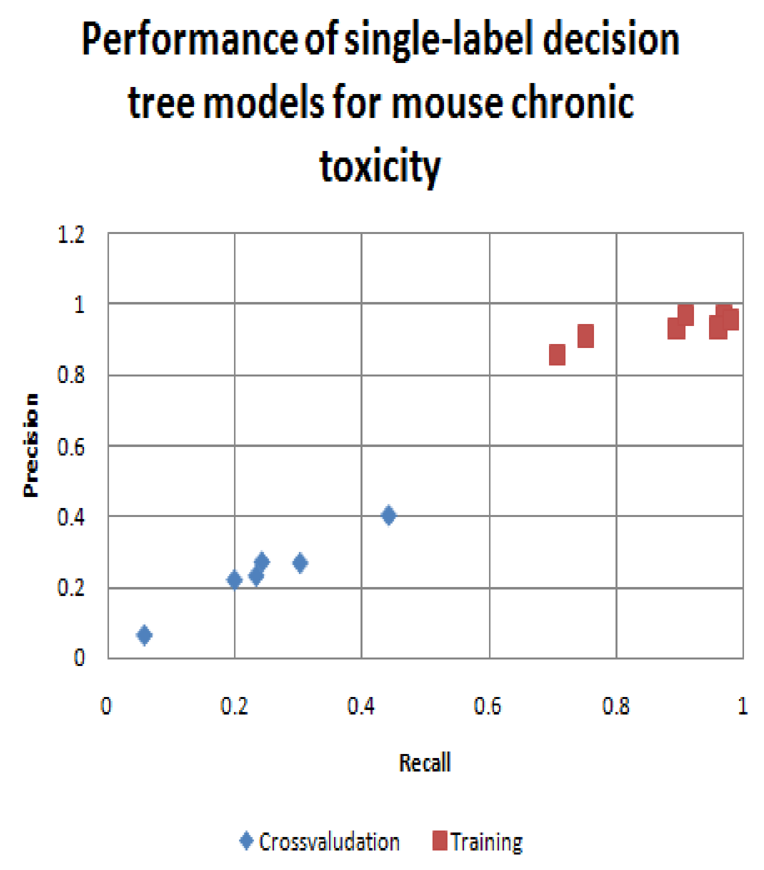
Rat Chronic/Cancer Toxicity models performance (balancing via SMOTE)

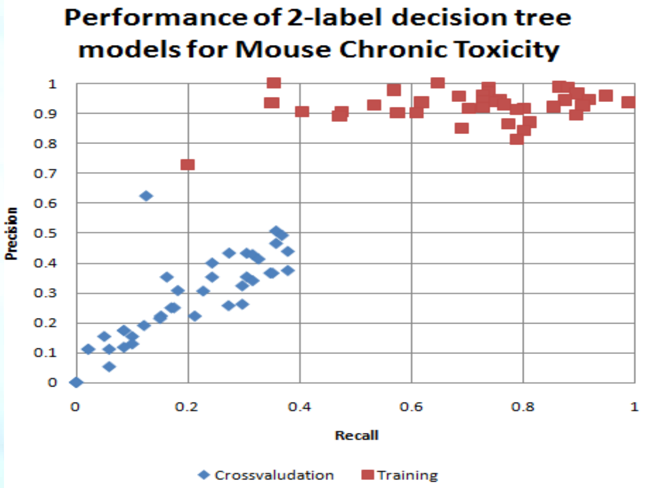
Assay	Number of endpoints selected as relevant tests in all decision trees
ACEA	24
ATG	151
BSK	106
CLM	19
CLZD'	54
NCGC	4
NVS	24
Solidus	2

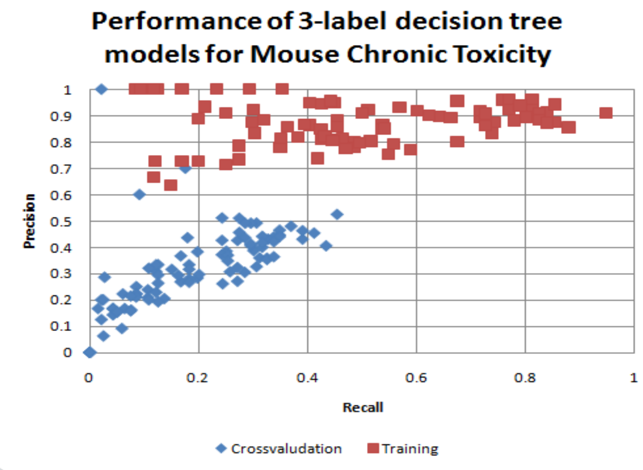
- ACEA_IC50 ACEA_LOC2
ACEA_LOC3 ACEA_LOC4
ACEA_LOC5 ACEA_LOCdec
ACEA_LOCinc ATG_Ahr_CIS
ATG_AP_1_CIS ATG_AR_TRANS
ATG_BRE_CIS ATG_C_EBP_CIS
ATG_CAR_TRANS ATG_CMV_CIS
ATG_CRE_CIS ATG_DR4_LXR_CIS
ATG_DR5_CIS ATG_EGR_CIS
ATG_ERa_TRANS ATG_ERE_CIS
ATG_ERRa_TRANS ATG_FoxA2_CIS
ATG_FXR_TRANS ATG_HIF1a_CIS
ATG_HNF4a_TRANS
ATG_Hpa5_TRANS ATG_IR1_CIS
ATG_ISRE_CIS ATG_LXRa_TRANS
ATG_LXRb_TRANS ATG_M_06_CIS
ATG_M_06_TRANS
ATG_M_19_TRANS ATG_MRE_CIS
ATG_NF_kB_CIS ATG_NFI_CIS
ATG_NRF1_CIS ATG_NRF2_ARE_CIS
ATG_NURR1_TRANS
ATG_Oct_MLP_CIS ATG_PBREM_CIS
ATG_PPARa_TRANS
ATG_PPARd_TRANS
ATG_PPARg_TRANS ATG_PPRE_CIS
ATG_PXRE_CIS ATG_RARa_TRANS
ATG_RARb_TRANS
ATG_RARg_TRANS ATG_RORE_CIS
ATG_RXRb_TRANS ATG_Sp1_CIS
ATG_SREBP_CIS ATG_STAT3_CIS
ATG_TCF_b_cat_CIS ATG_TGFb_CIS
ATG_VDRE_CIS ATG_Xbp1_CIS
BSK_3C_hLADR BSK_3C_ICAM1
BSK_3C_IL8 BSK_3C_MCP1
BSK_3C_Proliferation
BSK_3C_Thrombomodulin
BSK_3C_uPAR BSK_3C_VCAM1
BSK_3C_Vis BSK_4H_MCP1
BSK_4H_Pselectin BSK_4H_VCAM1
BSK_BE3C_hLADR BSK_BE3C_IL1a
BSK_BE3C_IP10 BSK_BE3C_MIG
BSK_BE3C_PAI1 BSK_BE3C_TGFb1
BSK_BE3C_tPA BSK_BE3C_uPA
BSK_BE3C_uPAR BSK_hDFCGF_EGFR
BSK_hDFCGF_IL8 BSK_hDFCGF_IP10
BSK_hDFCGF_MIG
BSK_hDFCGF_MMP1

BSK_hDFCGF_PAI1
BSK_hDFCGF_Proliferation
BSK_hDFCGF_VCAM1 BSK_KF3CT_IL1a
BSK_KF3CT_MCP1 BSK_KF3CT_MMP9
BSK_KF3CT_TGFb1 BSK_KF3CT_TIMP2
BSK_KF3CT_uPA BSK_LPS_Eselectin
BSK_LPS_MCSF BSK_LPS_PGE2
BSK_LPS_TissueFactor BSK_LPS_TNFa
BSK_LPS_VCAM1 BSK_SAg_CD38
BSK_SAg_CD69 BSK_SAg_Eselectin
BSK_SAg_MCP1 BSK_SM3C_MCP1
BSK_SM3C_Proliferation
BSK_SM3C_TissueFactor CLM_CellLoss_72hr
CLM_MicrotubuleCSK_72hr
CLM_MicrotubuleCSK_Destabilizer_72hr
CLM_MitoMass_24hr CLM_MitoMass_72hr
CLM_MitoMembPot_1hr
CLM_MitoMembPot_24hr
CLM_MitoMembPot_72hr
CLM_MitoticArrest_24hr
CLM_MitoticArrest_72hr
CLM_OxidativeStress_24hr
CLM_StressKinase_1hr
CLM_StressKinase_72hr CLZD_ABCB11_24
CLZD_ABCG2_6 CLZD_CYP1A1_24
CLZD_CYP1A1_48 CLZD_CYP1A1_6
CLZD_CYP1A2_24 CLZD_CYP1A2_48
CLZD_CYP1A2_6 CLZD_CYP2B6_24
CLZD_CYP2B6_48 CLZD_CYP3A4_24
CLZD_CYP3A4_48 CLZD_CYP3A4_6
CLZD_HMGCS2_48 CLZD_SULT2A1_24
CLZD_SULT2A1_48 CLZD_SULT2A1_6
CLZD_UGT1A1_24 NCGC_LXR_Agonist
NCGC_PPARg_Agonist
NCGC_PXR_Agonist_human
NCGC_PXR_Agonist_rat NVS_ADME_hCYP1A1
NVS_ADME_hCYP1A2 NVS_ADME_hCYP2A6
NVS_ADME_hCYP2B6 NVS_ADME_hCYP2C9
NVS_ADME_hCYP2J2 NVS_ADME_hCYP3A4
NVS_ADME_hCYP3A5 NVS_ADME_rCYP2C11
NVS_ENZ_hBACE NVS_ENZ_hGSK3b
NVS_ENZ_hPTPMEG2 NVS_ENZ_rabI2C
NVS_ENZ_rAChE NVS_ENZ_rCNOS
NVS_GPCR_hAdnRA2a NVS_GPCR_rAdnRa2B
NVS_GPCR_rSST NVS_IC_hNNR_NBungSens
NVS_NR_hAR NVS_NR_hPXR NVS_TR_hNET
NVS_TR_rSERT Solidus_P450

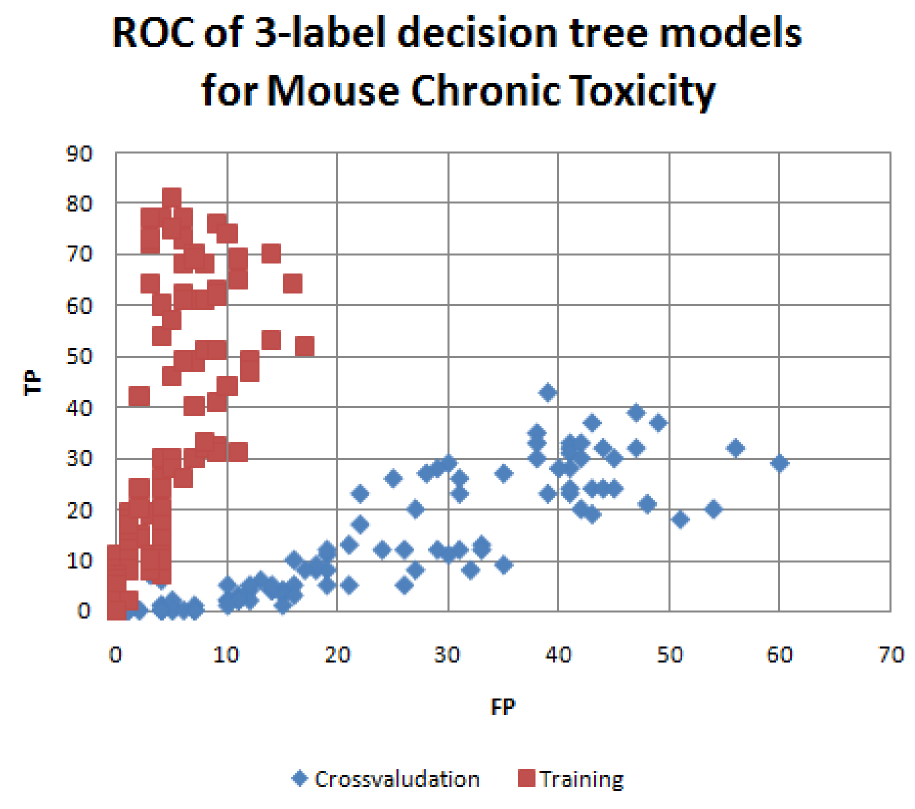
Mouse Chronic/Cancer Toxicity models performance

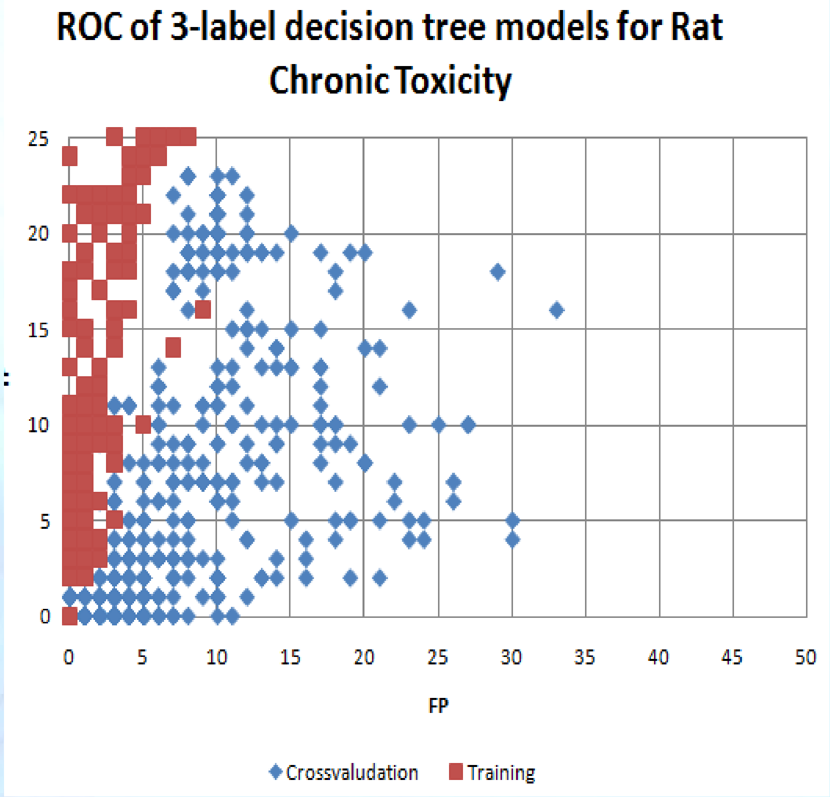

Performance of single-label decision tree models for mouse chronic toxicity


Performance of 2-label decision tree models for Mouse Chronic Toxicity

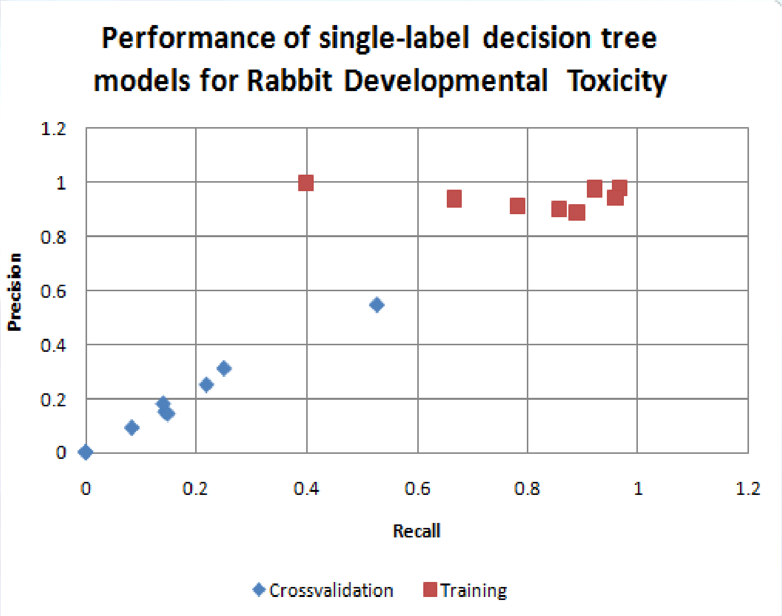

Performance of 3-label decision tree models for Mouse Chronic Toxicity

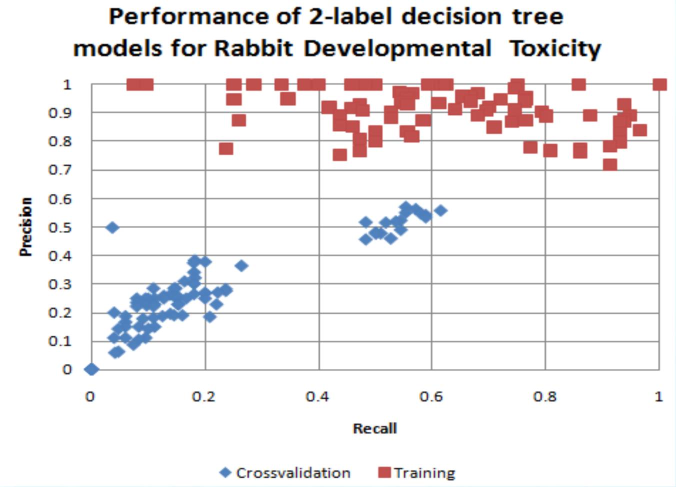
Experiments (3-label)

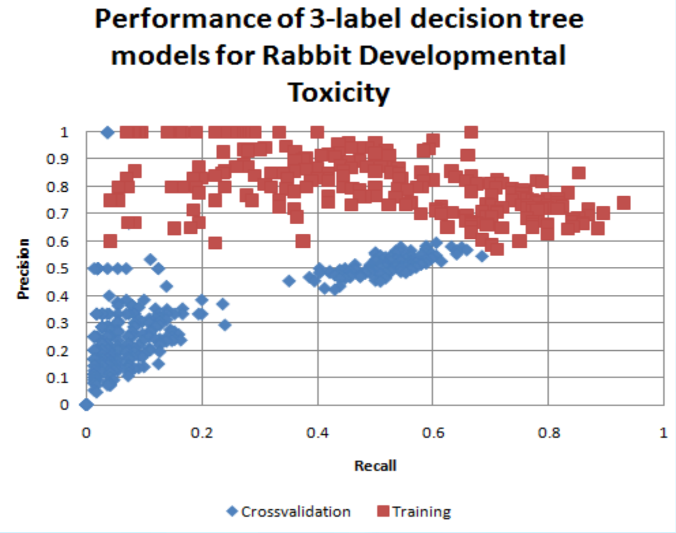




Developmental Toxicity Models performance



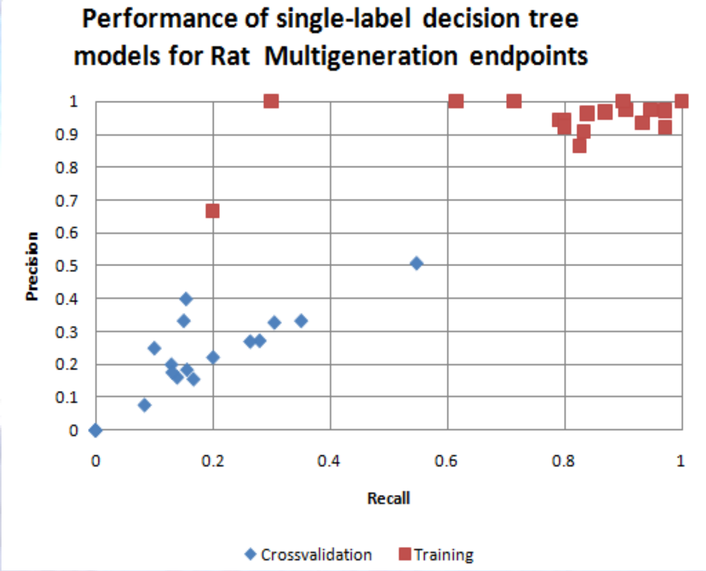


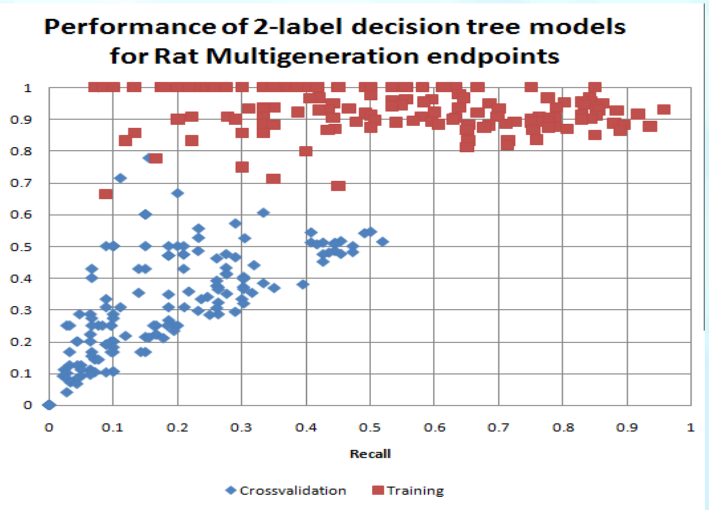


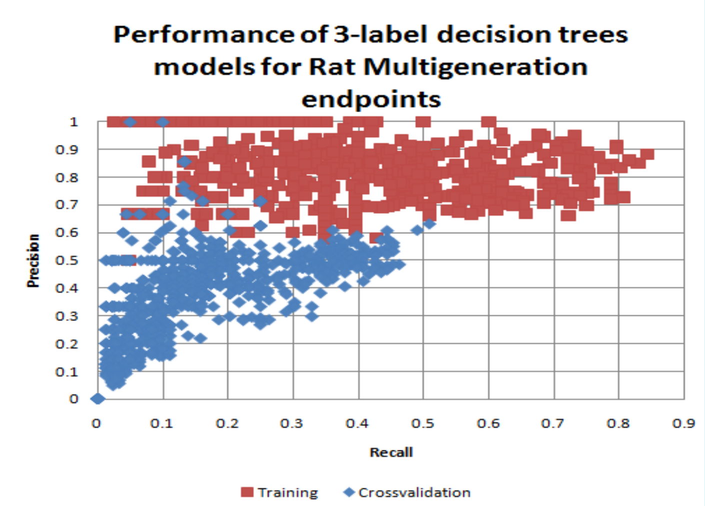
Multi-label trees perform better on average, compared to the single-label tree

OpenTox

Multigeneration Toxicity Models performance





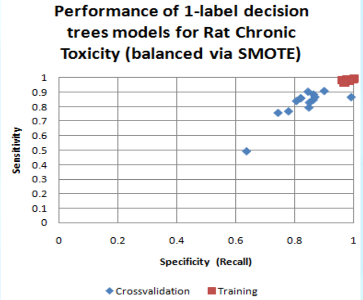


Multi-label trees perform better on average, compared to the single-label tree

OpenTox

Conclusions (single/multi-label trees)

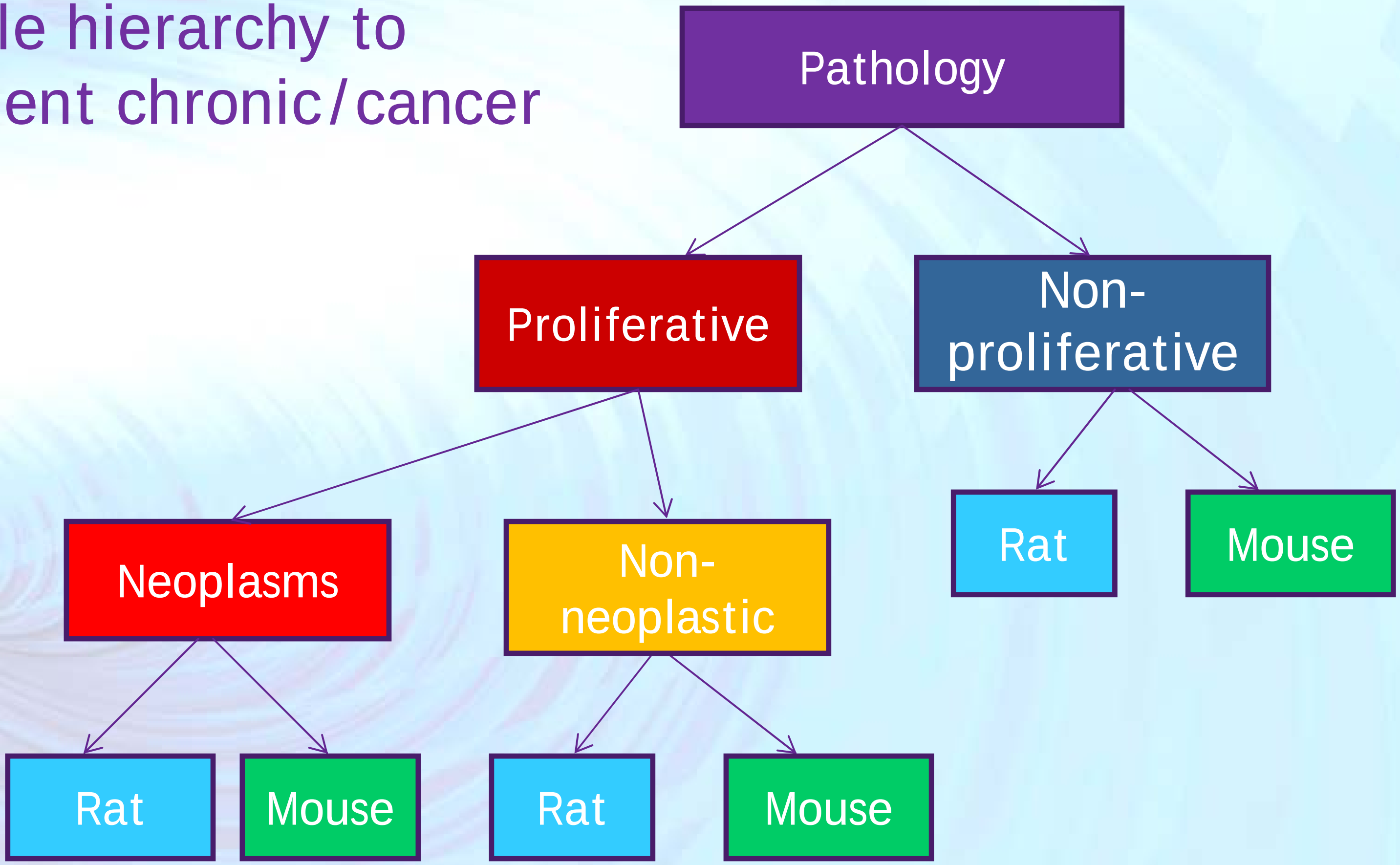
- Original dataset (unbalanced)
  - No successful models!
  - Performance drops significantly with cross validation
- Balanced dataset via SMOTE
  - Excellent results for one-label trees
  - *Unclear how to apply SMOTE for multi-label models – have to balance all classes instead of a single one!*



- The performance of the multi-label classification is better when the classes are related
  - Simpler trees, features relevant for all classes

Hierarchical classification

An example hierarchy to model rodent chronic/cancer endpoints



Pathology

Proliferative

Non-proliferative

Neoplasms

Non-neoplastic

Rat

Mouse

Rat

Mouse

Rat

Mouse

Matching the endpoints to the hierarchy

CHR_Rat_LiverTumors = pathology.proliferative.neoplastic.rat

CHR_Rat_LiverProliferativeLesions = pathology.proliferative.neoplastic.rat , pathology.proliferative.nonneoplastic.rat

CHR_Rat_LiverNecrosis = pathology.non-proliferative.rat

CHR_Rat_LiverHypertrophy = pathology.non-proliferative.rat

CHR_Rat_KidneyNephropathy = pathology.non-proliferative.rat

CHR_Rat_KidneyProliferativeLesions = pathology.proliferative.neoplastic.rat , pathology.proliferative.nonneoplastic.rat

CHR_Rat_ThyroidProliferativeLesions = pathology.proliferative.neoplastic.rat ,pathology.proliferative.nonneoplastic.rat

CHR_Rat_ThyroidTumors = pathology.proliferative.neoplastic.rat

CHR_Rat_ThyroidHyperplasia = pathology.proliferative.nonneoplastic.rat

CHR_Rat_TesticularTumors = pathology.proliferative.neoplastic.rat

CHR_Rat_TesticularAtrophy = pathology.non-proliferative.rat

CHR_Rat_SpleenPathology = pathology.proliferative.neoplastic.rat, pathology.proliferative.nonneoplastic.rat

CHR_Rat_Tumorigen = pathology.proliferative.neoplastic.rat

CHR_Mouse_LiverTumors = pathology.proliferative.neoplastic.mouse

CHR_Mouse_LiverProliferativeLesions = pathology.proliferative.neoplastic.mouse, pathology.proliferative.nonneoplastic.mouse

CHR_Mouse_LiverNecrosis = pathology.non-proliferative.mouse

CHR_Mouse_LiverHypertrophy = pathology.non-proliferative.mouse

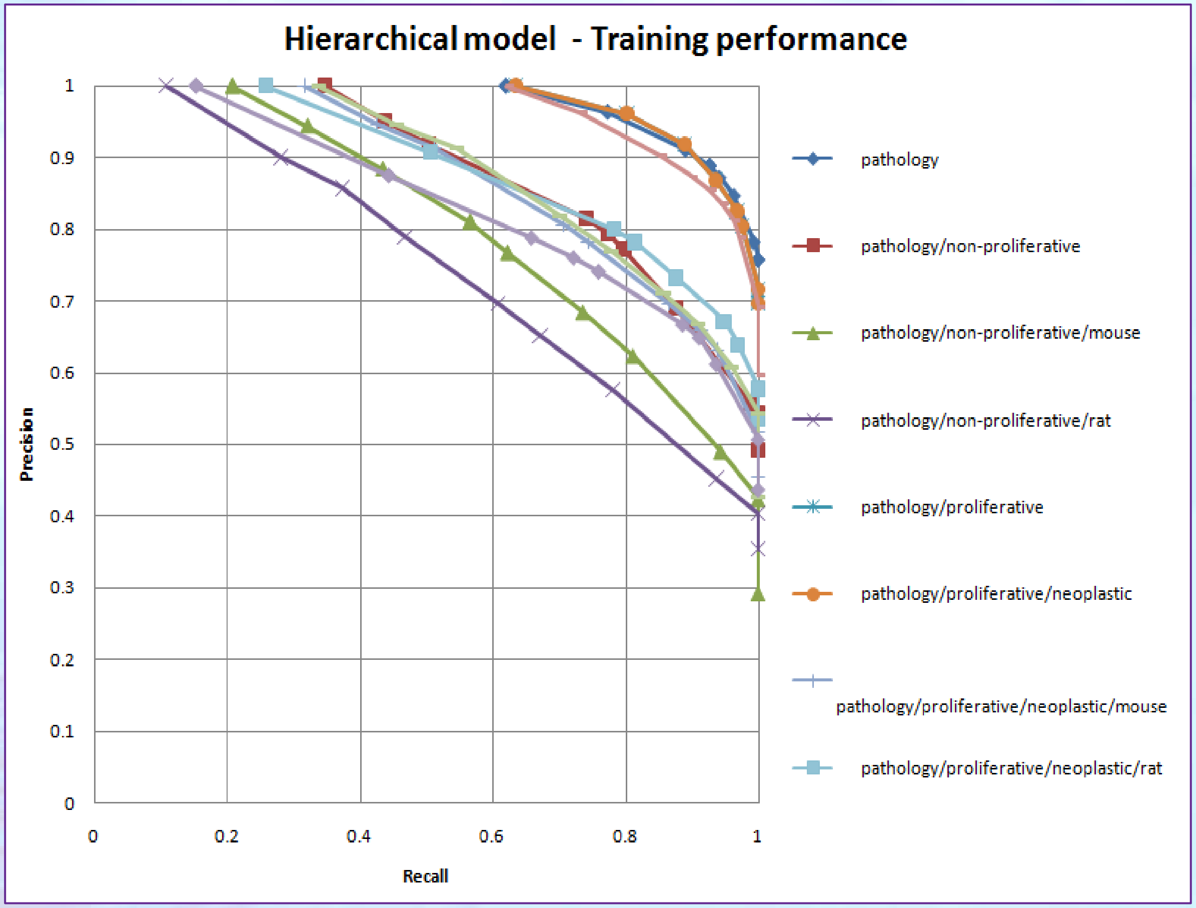
CHR_Mouse_KidneyPathology = pathology.proliferative.neoplastic.mouse, pathology.proliferative.nonneoplastic.mouse

CHR_Mouse_LungTumors = pathology.proliferative.neoplastic.mouse

CHR_Mouse_Tumorigen = pathology.proliferative.neoplastic.mouse

Experiments (4) Chronic/Cancer rodent toxicity (unbalanced dataset)

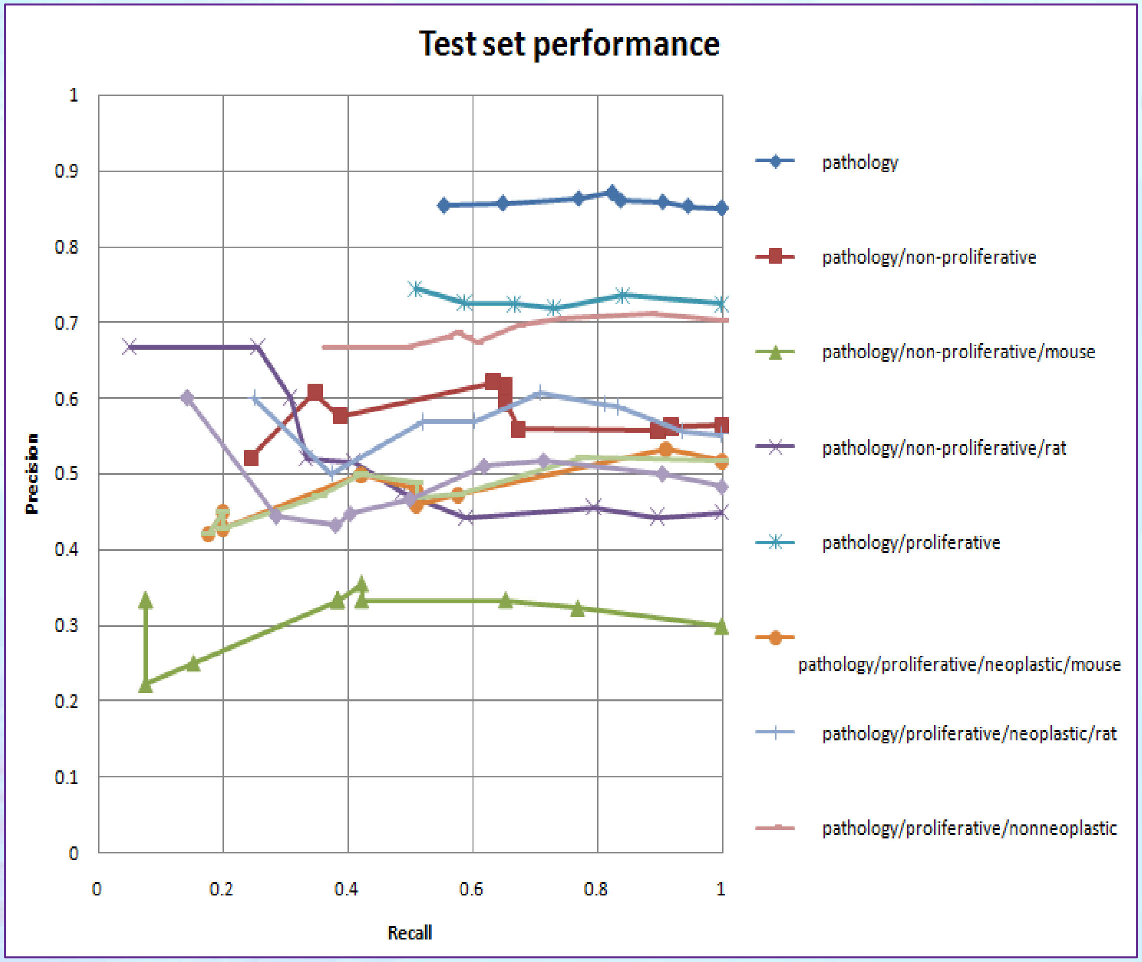
The data set is split into training and test sets 2:1



Precision = TP / (TP + FP)
Recall = TP / (TP + FN) = Sensitivity

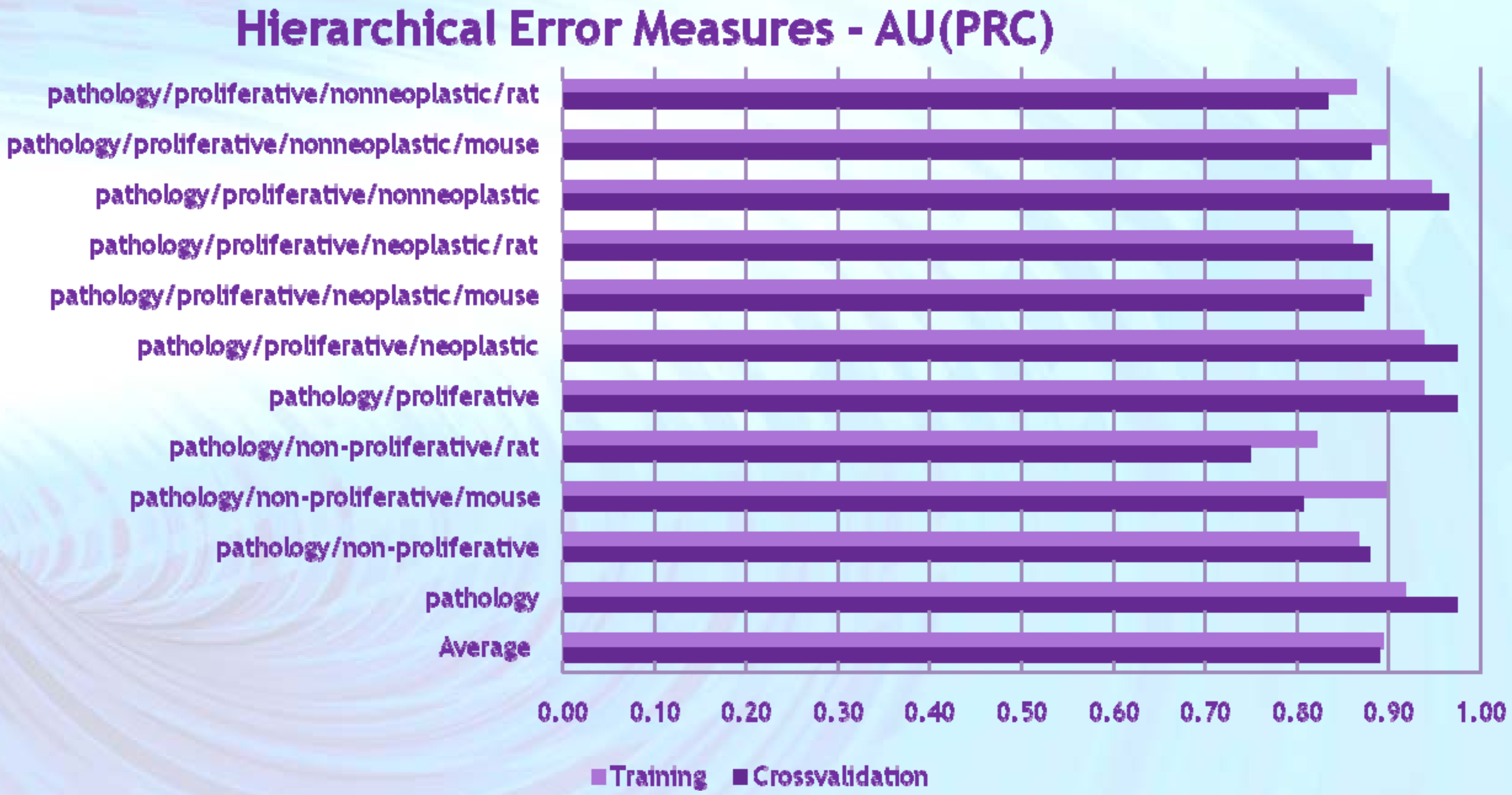
Experiments (4) Chronic/Cancer rodent toxicity (unbalanced dataset)

The data set is split into training and test sets 2:1

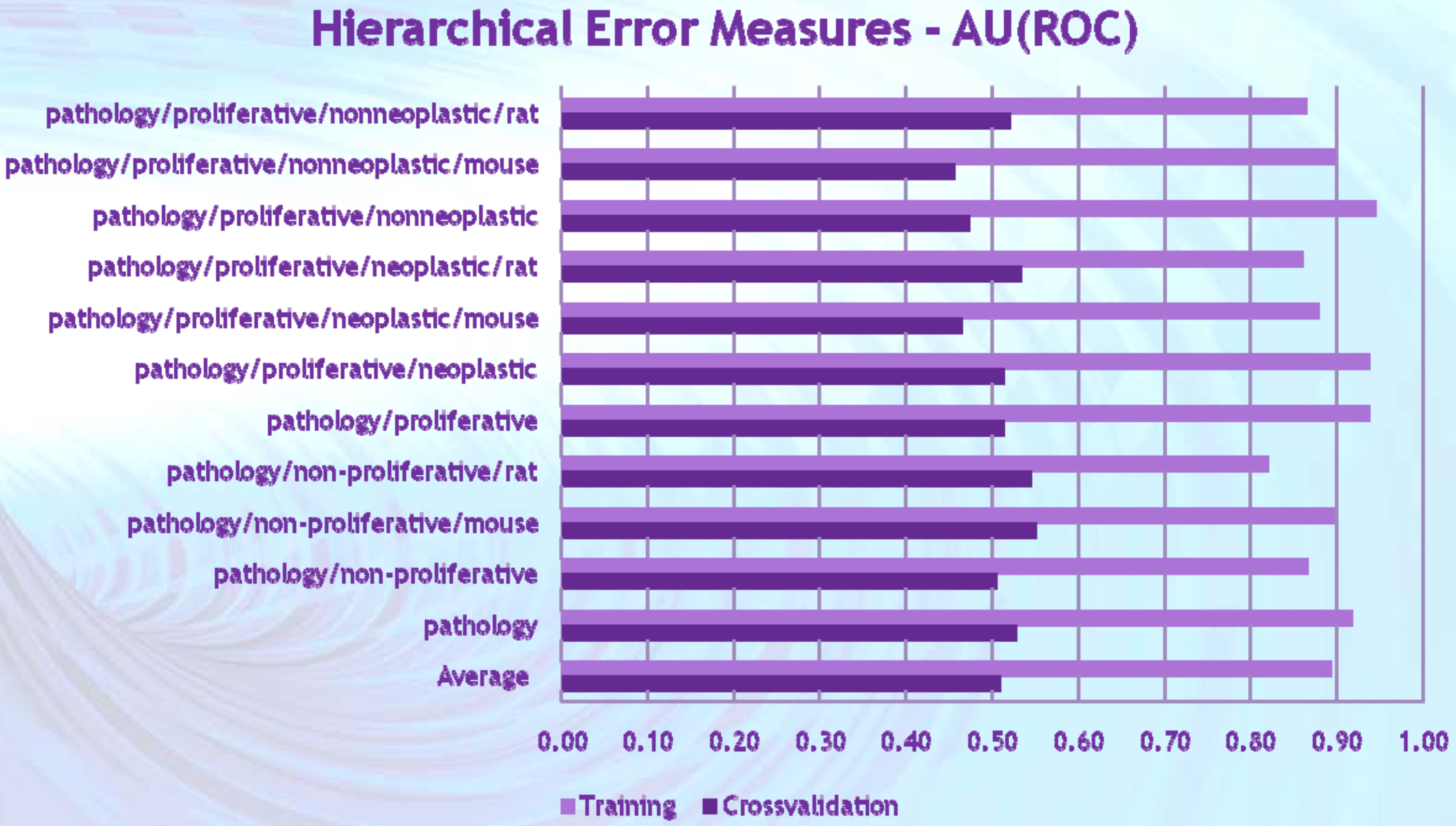


Precision = TP / (TP + FP)
Recall = TP / (TP + FN) = Sensitivity

Hierarchical model : Chronic/Cancer rodent toxicity (unbalanced dataset)



Hierarchical model : Chronic/Cancer rodent toxicity (unbalanced dataset)



OpenTox

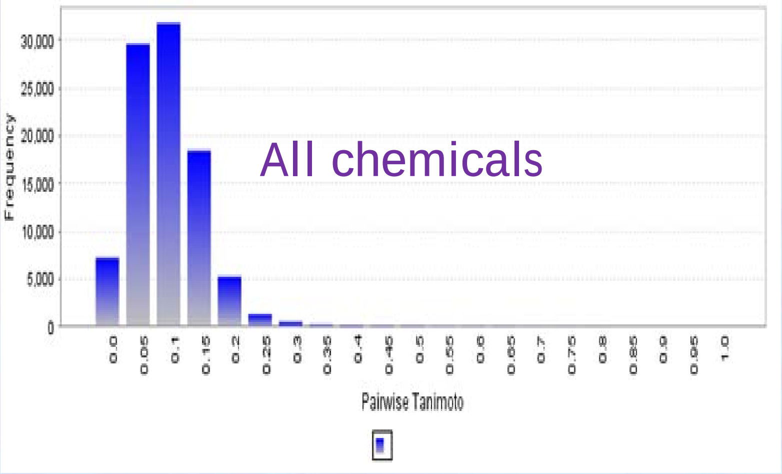
Conclusions (hierarchical model)

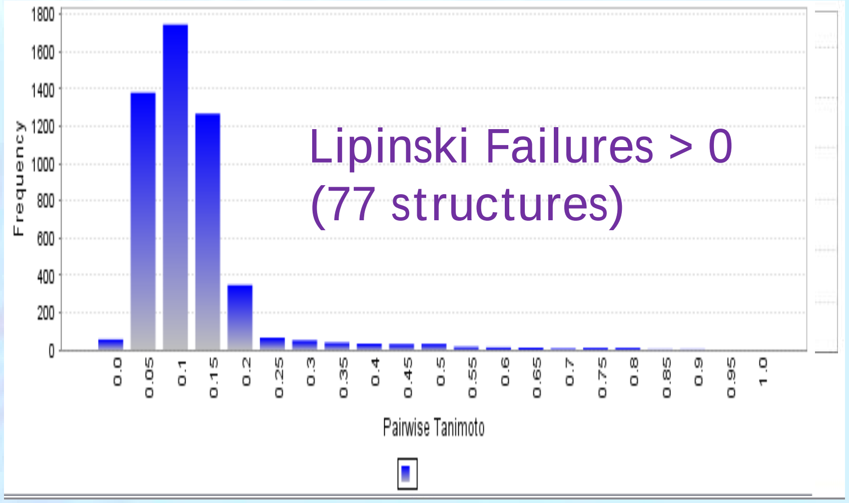
- The hierarchical model performs reasonably well on top level
- The unbalanced dataset most probably is the reason for the worse performance on the lower levels
- SMOTE balancing was not performed; need additional research how to balance multiple classes in a flat or hierarchical setting

Structural diversity

- Addition of a set of structural fragments (from ToxCast ChemicalInfo files) to the *in-vitro* data doesn't make any difference;
- The decision tree didn't select any of the structural alerts as relevant!

- Pairwise similarity matrix of Tanimoto coefficient between every two chemicals calculated by AmbitXT (http://ambit.sourceforge.net)


All chemicals


Lipinski Failures > 0
(77 structures)

Summary

- Continuous *in-vitro* data and binary *in-vivo* data are used to derive predictive clustering trees of 3 types – single label, multi-label and hierarchical
- Multi-label trees on average perform better and are of smaller size, compared to single-label trees
- Modifying class balance is necessary in order to model ToxCast *in-vitro* vs. *in-vivo* data
- Balancing via SMOTE performs very well

Summary

- Data sparsity might be another factor for classification performance over the unbalanced datasets.
- The problem of sparse data, where small number of instances are responsible for a high error rate is known in Machine Learning as "the problem with small disjuncts"
- Thus, ignoring the sparse data areas is not a recommended approach.
- There is no a single remedy for this problem. Recommended approaches are instance-based (lazy) learning, oversampling towards the class with small disjuncts, combining decision trees and lazy learning, etc.
- The combination of noise and small disjuncts in a dataset is prohibitive for the performance.

Future work

- apply cost-sensitive classification instead of balancing for multi-label and hierarchical trees;
- explore hierarchical methods beyond decision trees;
- apply similar approaches to other datasets, e.g. in the framework of the recently launched EU-FP7 funded project Cadaster (http://www.cadaster.eu/)

Final words...

- Modelling ToxCast dataset is challenging, but interesting and definitely promising!

- Acknowledgements:
  - OpenTox project & Barry Hardy
  - Ivelina Nikolova
  - Prof. Boris Aleksiev
  - Jan Struyf (Katholieke Universiteit Leuven, Belgium)

Thank you!

Questions?