Random Forest Ensembles Applied to MLSCN Screening Data for Prediction and Feature Selection

Rajarshi Guha
School of Informatics
Indiana University
and
Stephan Schürer
Scripps, FL

23rd August, 2007

Broad Goals

- Understand and possibly predict cytotoxicity
  - Utilizing MLSCN screening data and external data
  - Characterize and visualize various screening results
  - Relate screening data to known information
- Model and predict acute toxicity in animals
  - Relate large cytotoxicty data sets to animal toxicity(?)
- Modelling protocols to handle the characteristics of HTS data
  - Large datasets, imbalanced classes, applicability
- Make models publicly available
  - For use in multiple scenarios and accessible by a variety of methods

Datasets

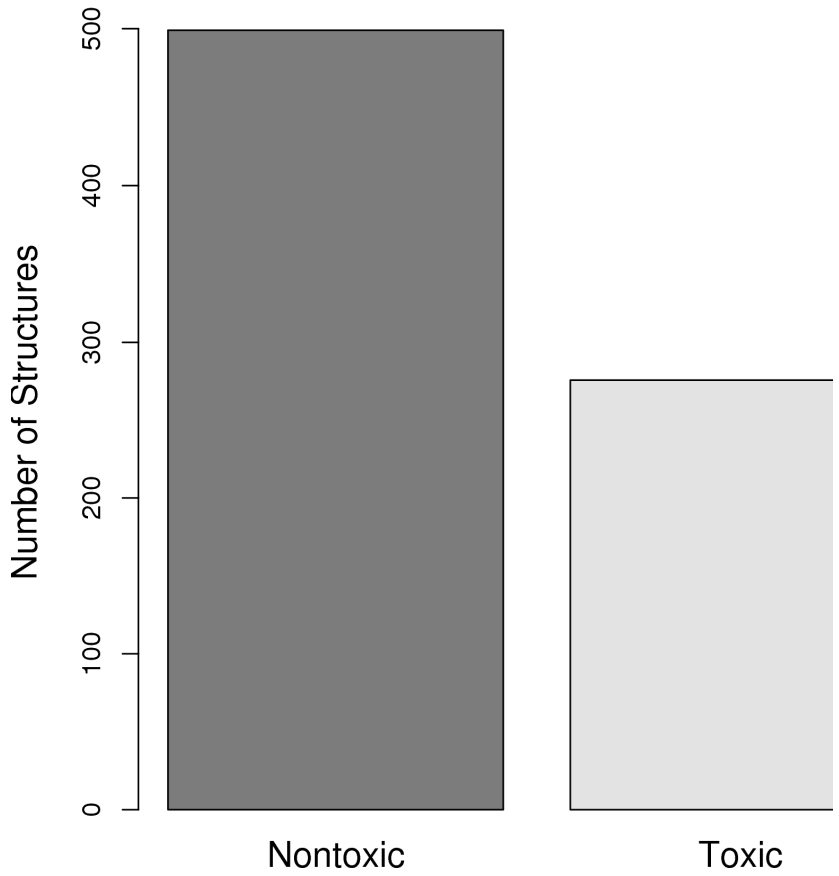
- Animal Acute Toxicity Data was extracted from the ToxNet database (available from MDL)
  - Selected only LD50 data for mouse and rat and three routes of administration
  - Summarized LD50 data by structure, species and route (140,808 LD50 data points, 103,040 structures)
  - Classified into Toxic/Nontoxic using a cutoff
- Cytotoxicity Data was taken as published in PubChem from Scripps and NCGC
  - Scripps Jurkat cytotoxicity assay (59,805 structures with %Inhib, 801 IC50 values)
  - NCGC data from PubChem for 13 cell lines (non-MLSMR structures): summarized multiple sample data by unique structures and extracted IC data: 1,334 structure, 13 x 1,334 IC50 values for different cell lines

Scripps Cytotoxicity Models

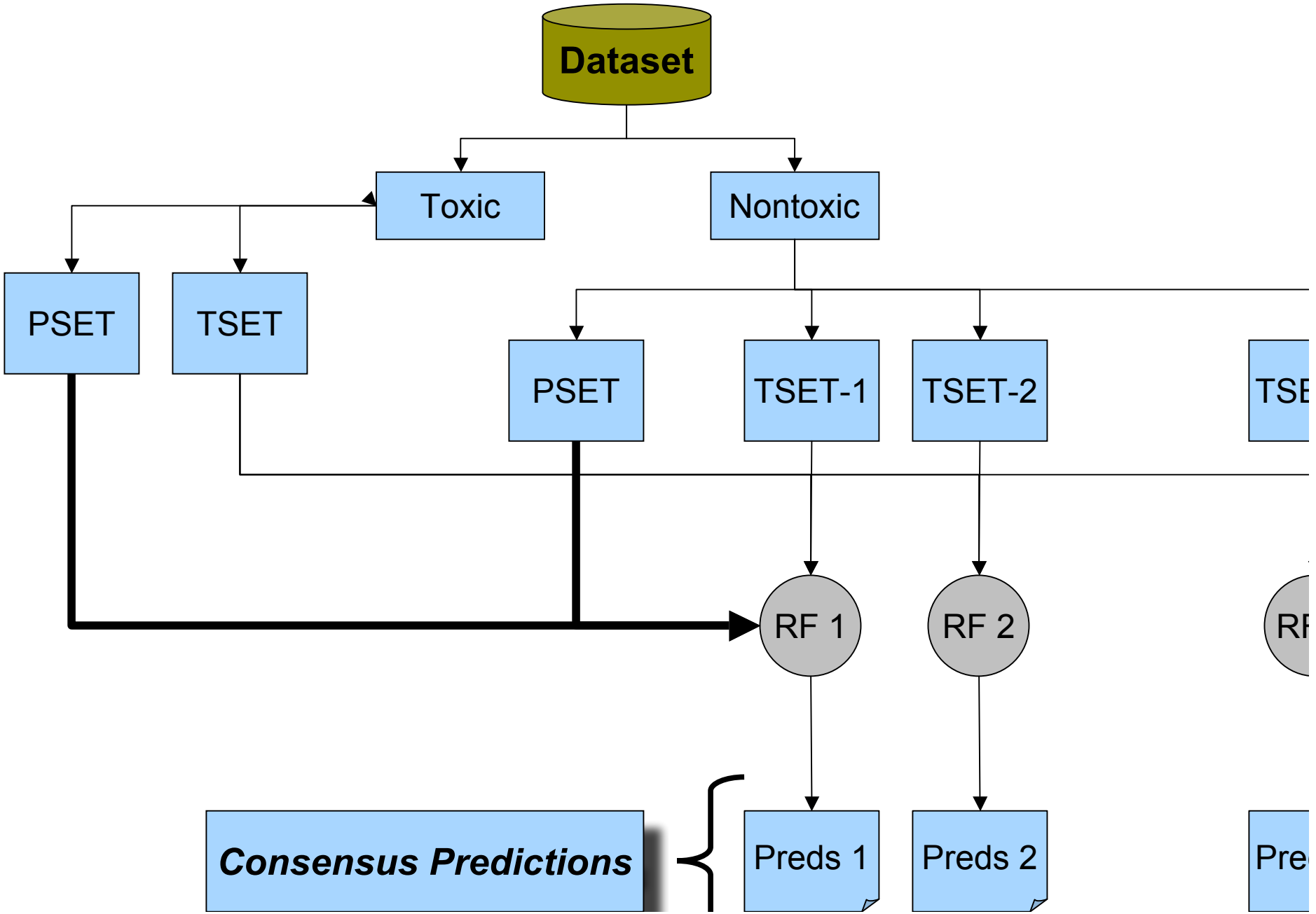
- 57,469 valid structures
- 775 structures with measured IC50
  - Skipped 26 structures that BCI could not parse
- How do we model this dataset?
  - Use all data. Very poor results
  - Use a sampling procedure to get an ensemble of models
  - Consider just the 775 structures

Scripps Cytotoxicity Models

- First considered the 775 structures
- Evaluated 1052 bit BCI fingerprints
- Selected a cutoff pIC50
  - >= 5.5 - toxic
  - < 5.5 – nontoxic
- Used sampling to create 10-member ensemble



Handling Imbalanced Classe

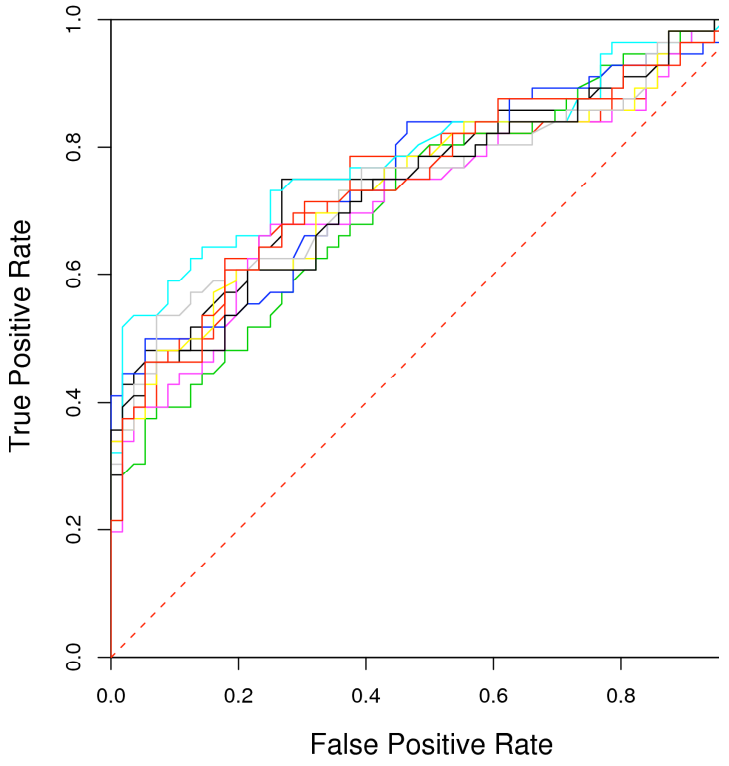


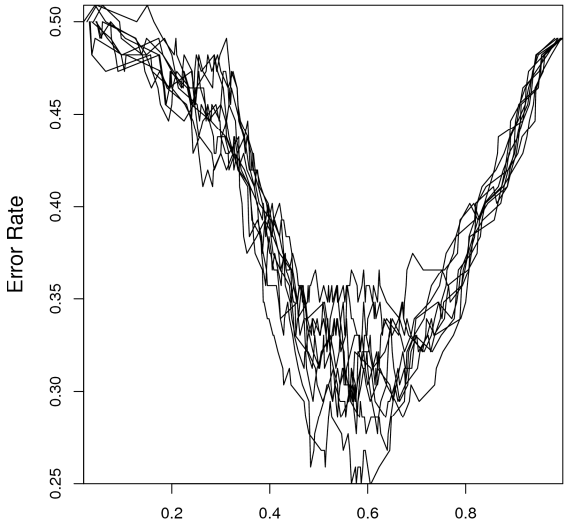
Scripps Cytotoxicity Model

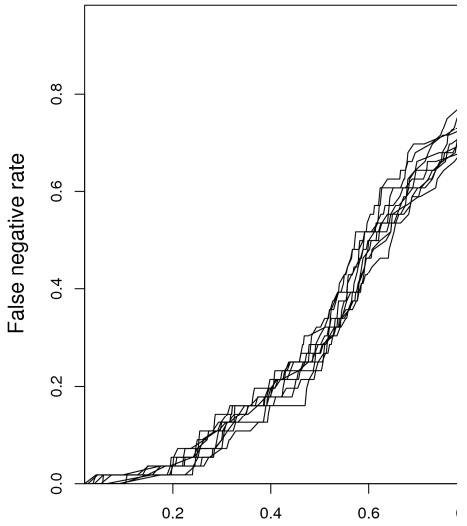
- % correct (ensemble average) = 69%
- % correct (consensus) = 71%

	Nontoxic	Toxic
Nontoxic	39	17
Toxic	15	41

- *Not very good performance*

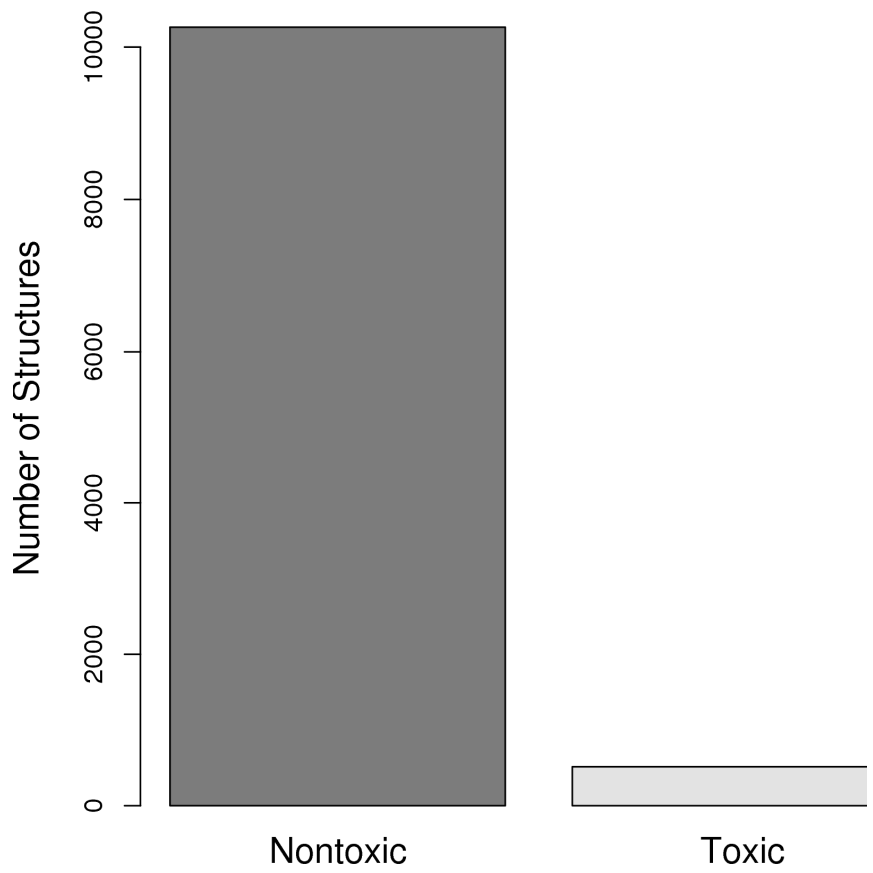






Do More Negatives Help?

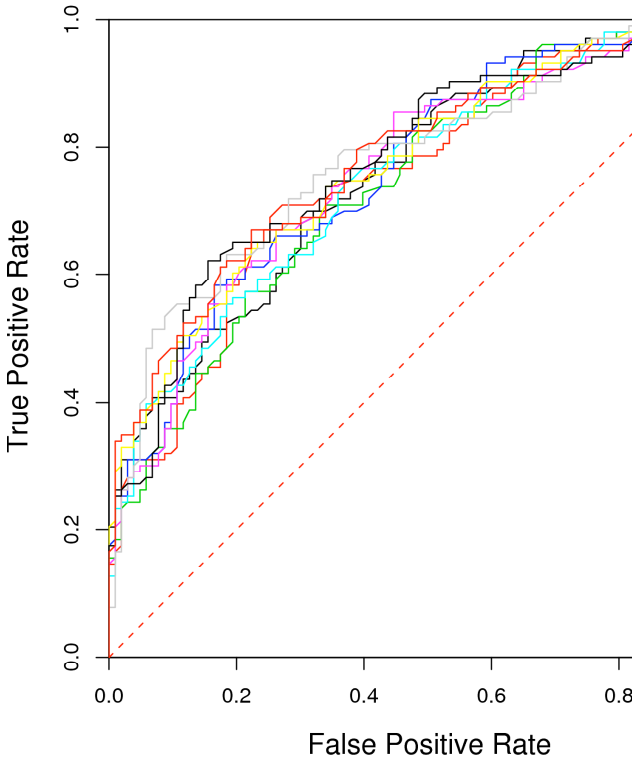
- Include 10,000 structures, randomly selected
  - Primary data, assumed to be nontoxic
- Selected a cutoff pIC50
  - >= 5.0 - toxic
  - < 5.0 – nontoxic
- Used sampling to create 10-member ensemble



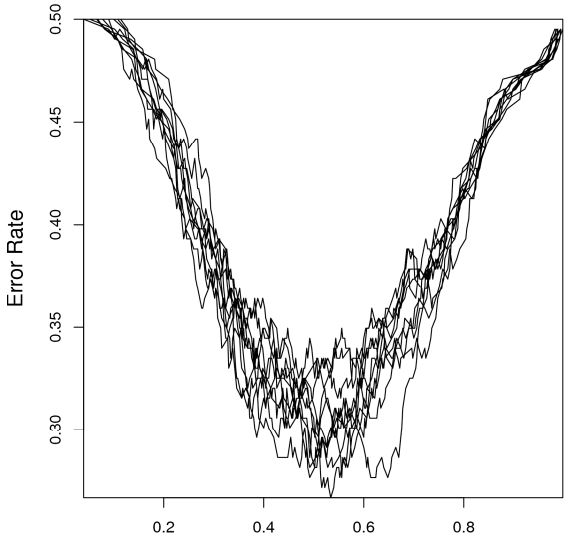
Expanded Cytotoxicity Datase

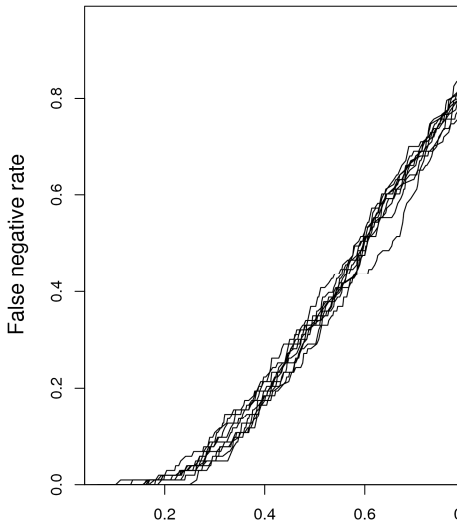
- % correct (averaged over the ensemble) = 69%
- % correct (consensus prediction) = 71%

	Nontoxic	Toxic
Nontoxic	79	24
Toxic	35	68



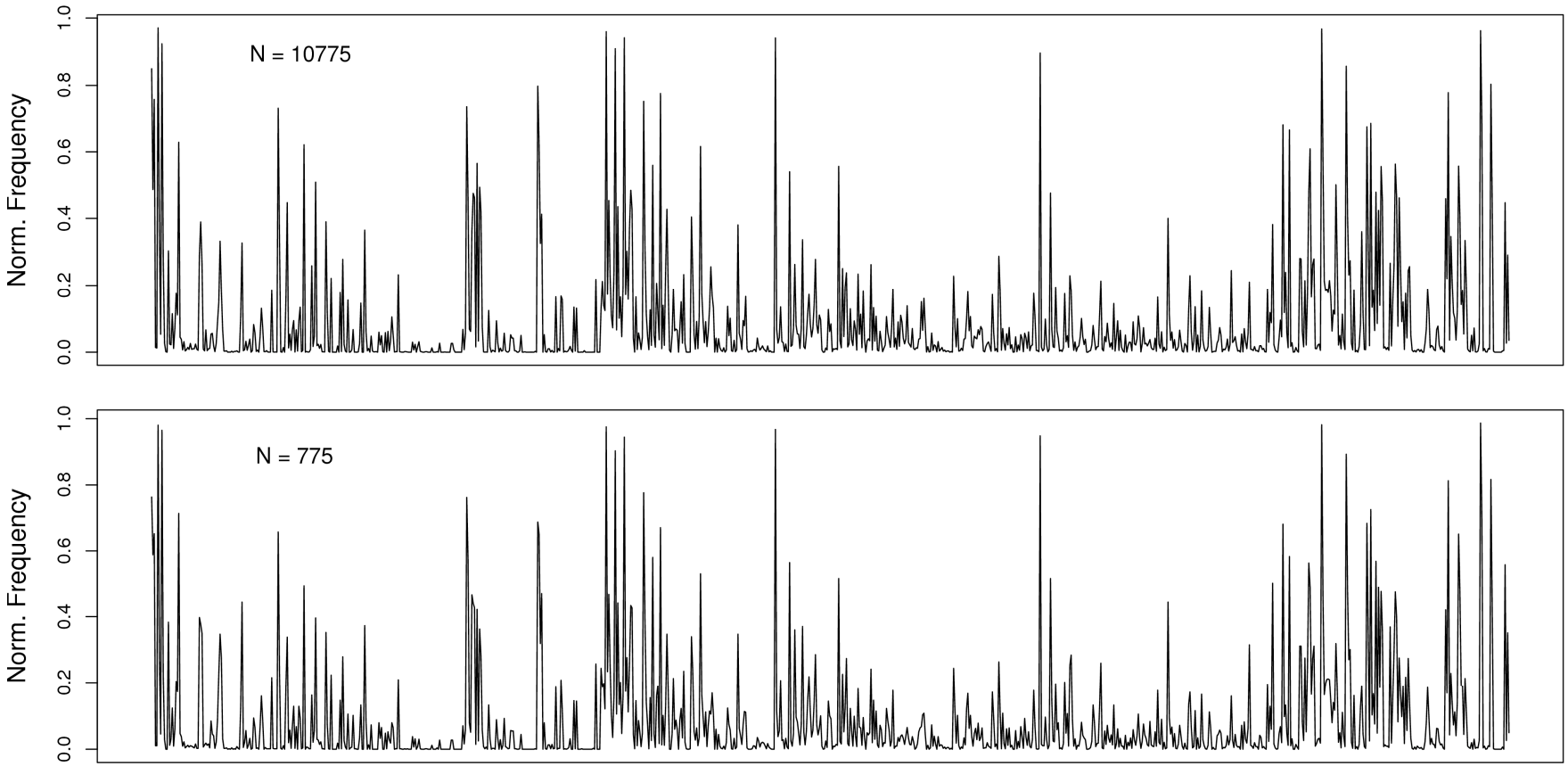
- *Not much improvement*
- *Insufficient sampling of the nontoxics*





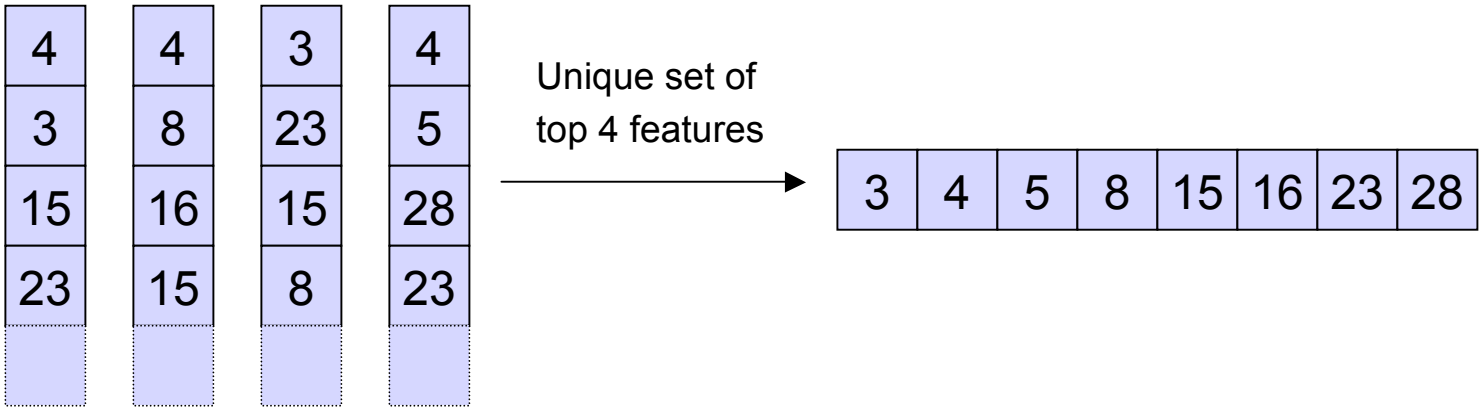
We Need More Positives

- The two datasets (775 vs 10,775 compounds) are quite similar in terms of *bit spectrum*
- *Normalized Manhattan distance = 0.016*



Selecting Important Features

- Identifying the important features in an ensemble
  - Each RF model can rank the input features
  - A consistent ensemble should have similar, but not necessarily identical, features highly ranked
  - Consider the unique set of top N
  - The size of the unique set of top N features provides information about the robustness of the ensemble

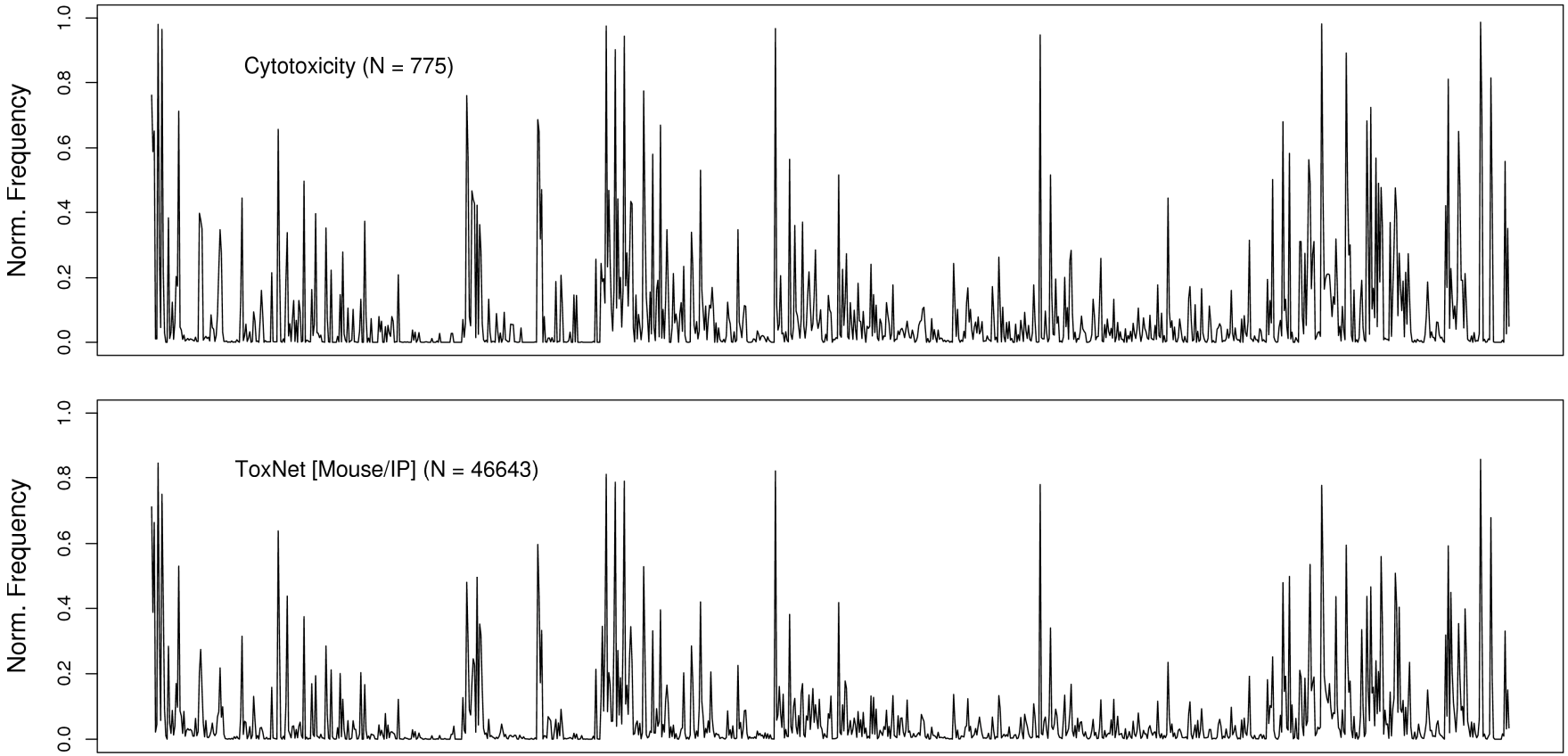


Important Structural Features

- The 10 most important features for predictive ability across the ensemble leads to 43 unique important bits
- This is a total of 66 structural features
  - The toxic compounds are characterized by having *slightly* larger number of these features, on averag

Predicting Animal Toxicity

- Should we use cytotoxicity model to predict animal toxicity?
- *Normalized Manhattan distance = 0.037*



Predicting Animal Toxicity

- Performance really depends on the model cutoff and our goals

	Nontoxic	Toxic
Nontoxic	43072	1683
Toxic	1748	140

Cutoff = 0.6, 93% correct

	Nontoxic	Toxic
Nontoxic	34674	1158
Toxic	10146	665

Cutoff = 0.5, 75% correct

	Nontoxic	Toxic
Nontoxic	20369	587
Toxic	24451	1236

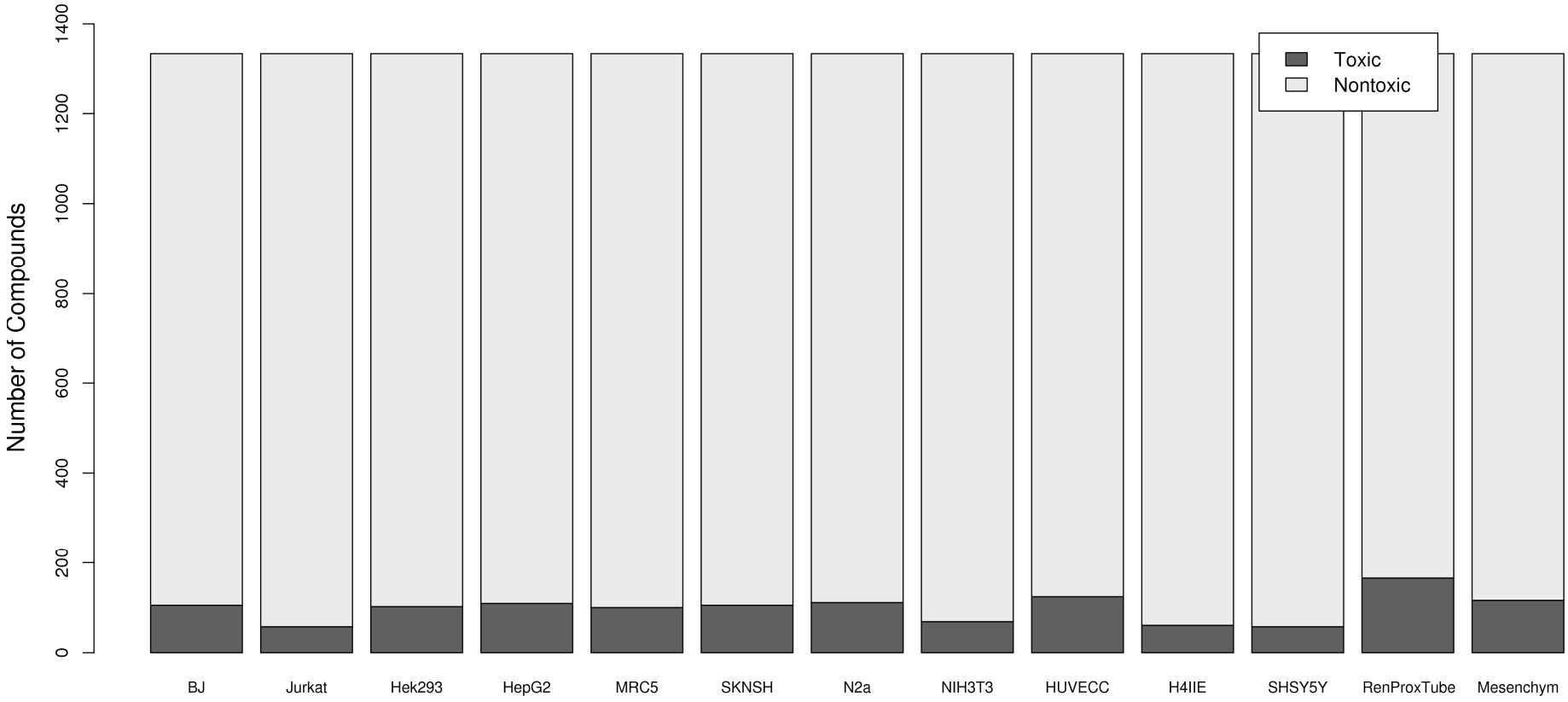
Cutoff = 0.4, 46% correct

NCGC Toxicity Datase

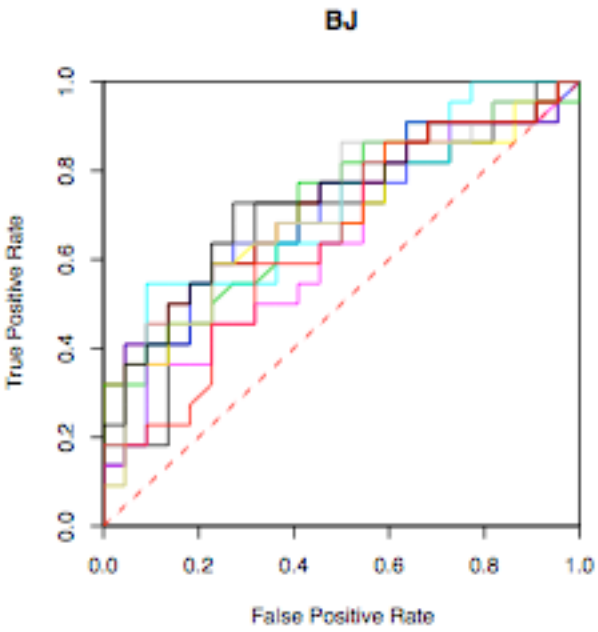
- Considered 13 cell lines, pIC50's
- 1334 compounds, including
  - metals
  - inorganics
- Classified into toxic / nontoxic using a cutoff
  - mean + 2 * SD
- Built models for each cell line

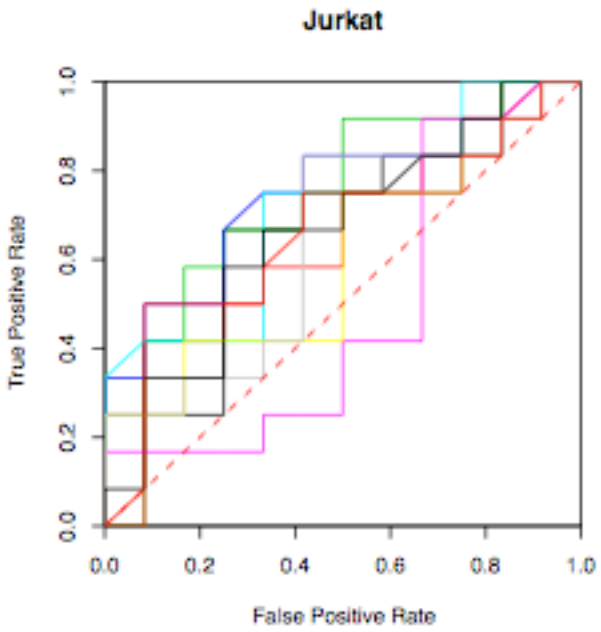
NCGC – Class Distributions

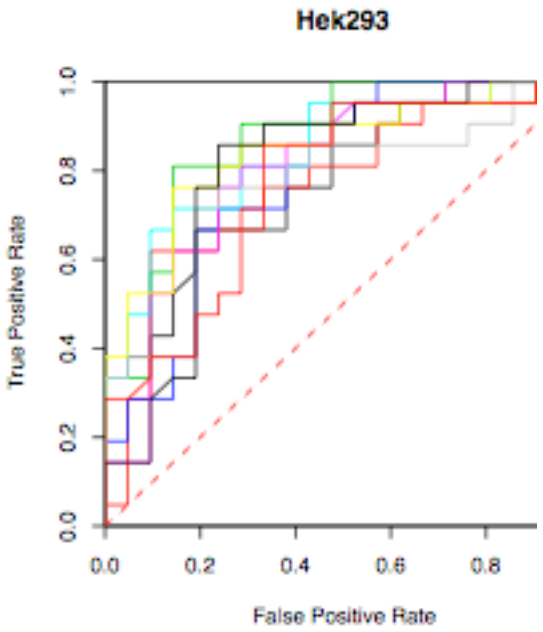
- Cutoff values ranged from 3.56 to 4.72
- Classes are severely imbalanced
- Developed ensembles of RF models

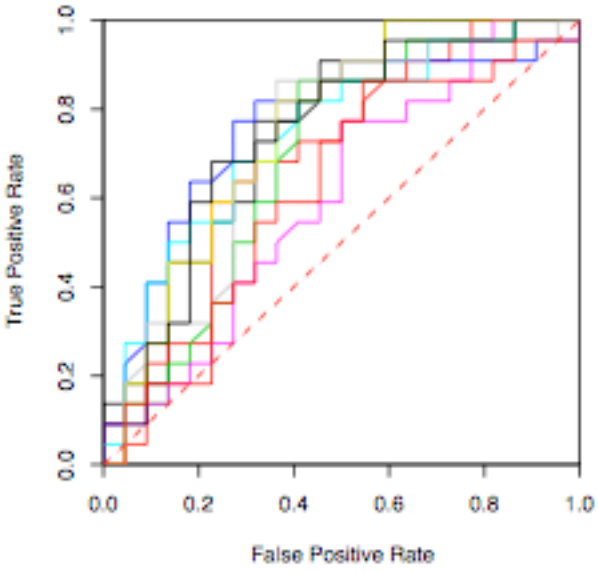


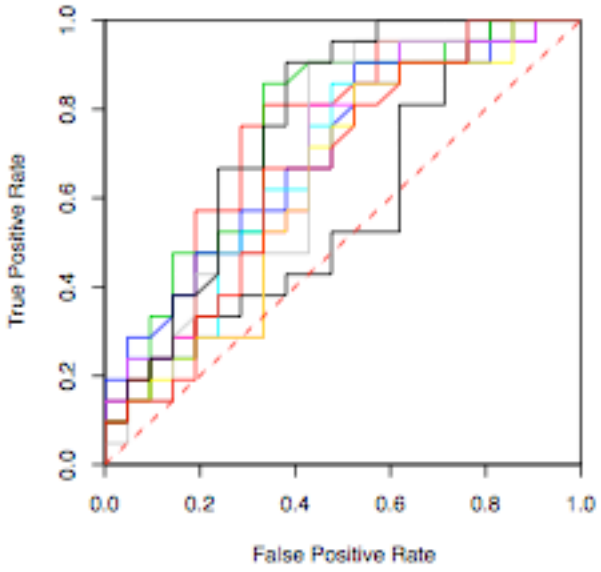
NCGC Model ROC Curves

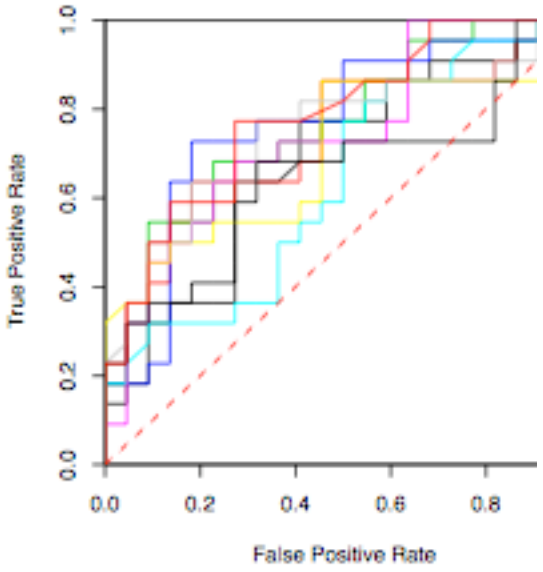




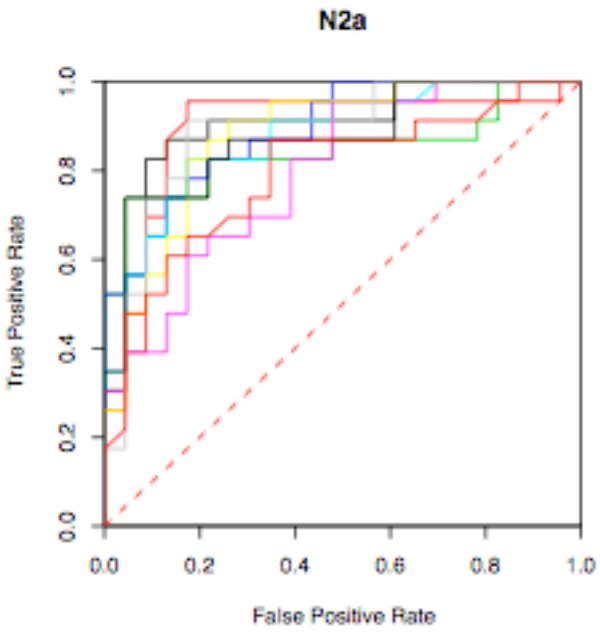


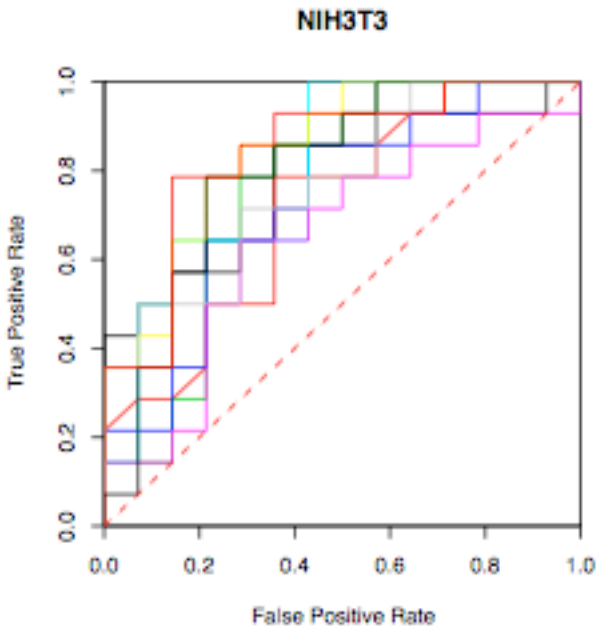


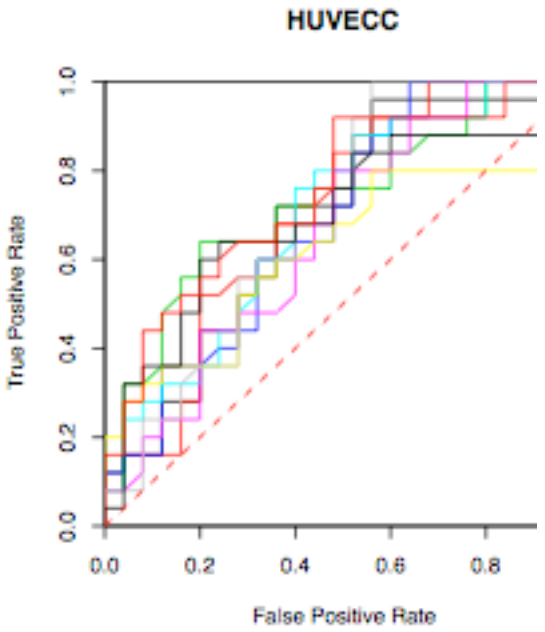


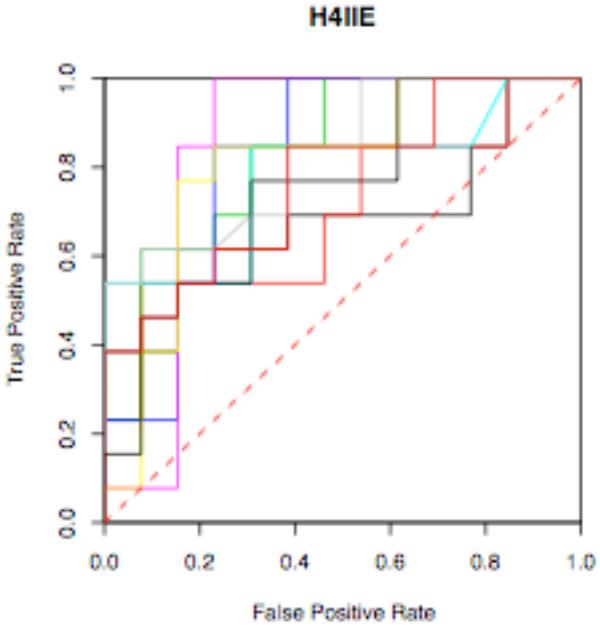


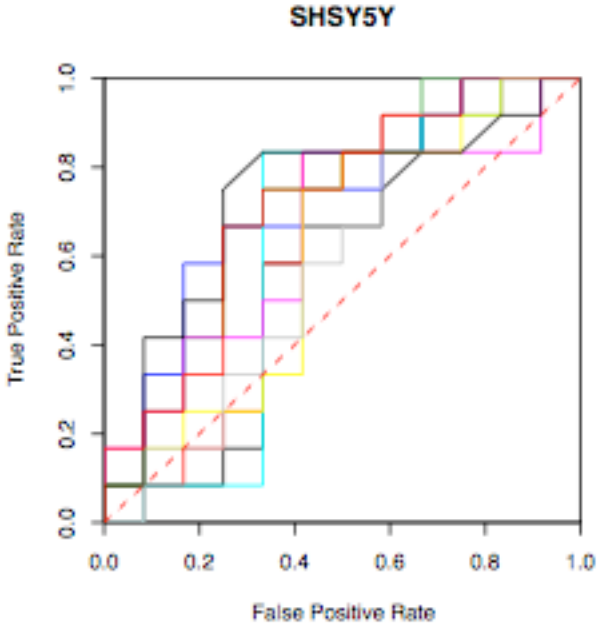
NCGC Model ROC Curves

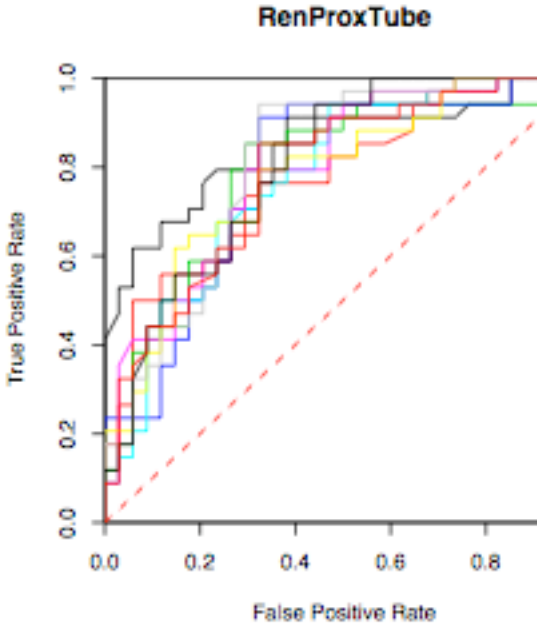




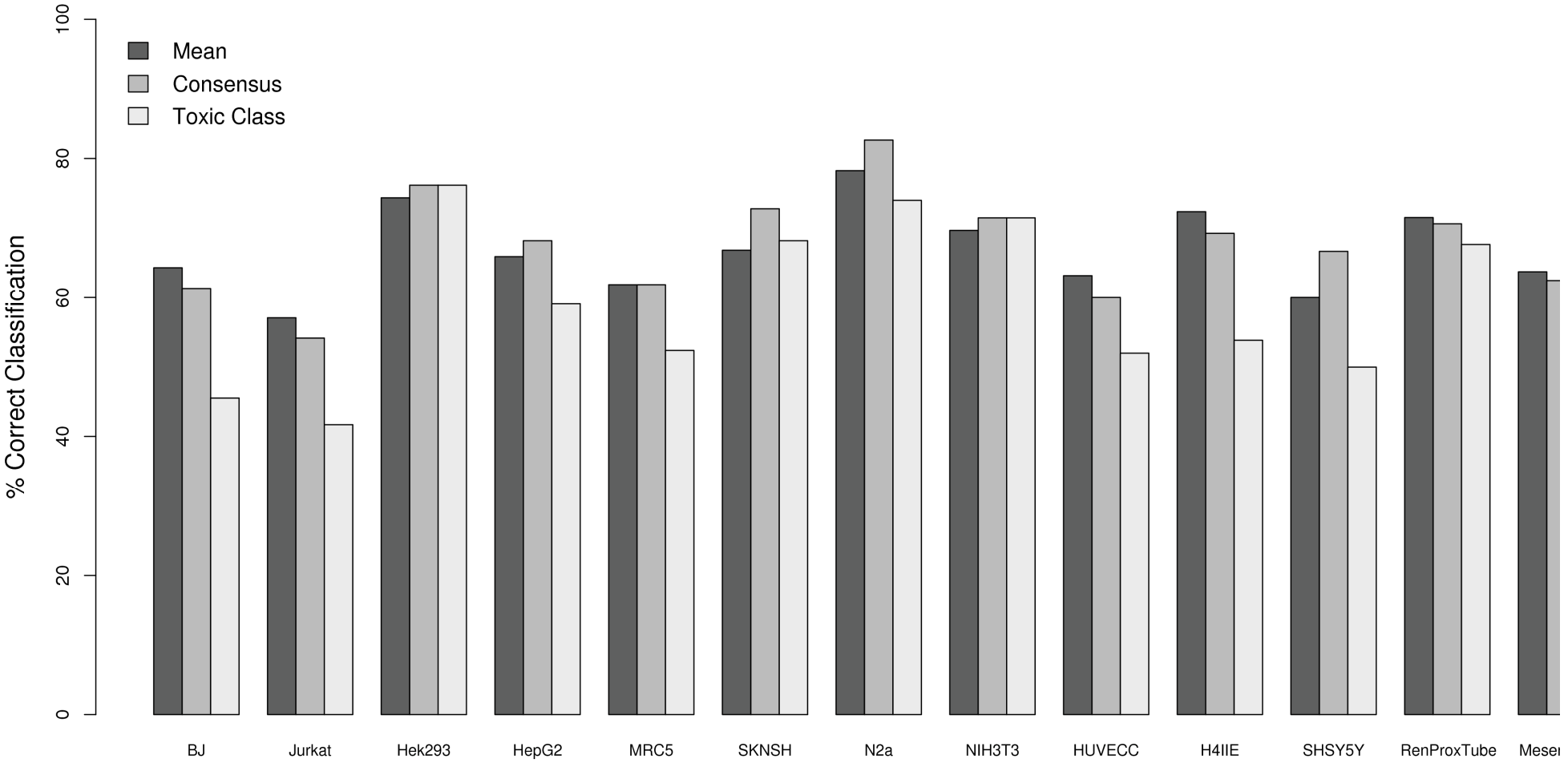








NCGC – Model Performanc
(Prediction Set



NCGC – Using the Models

- Predicted toxicity class for the Scripps Cytotoxicity dataset (775 compounds) using model built for NCGC Jurkat cell line

	Nontoxic	Toxic
Nontoxic	67	49
Toxic	432	227

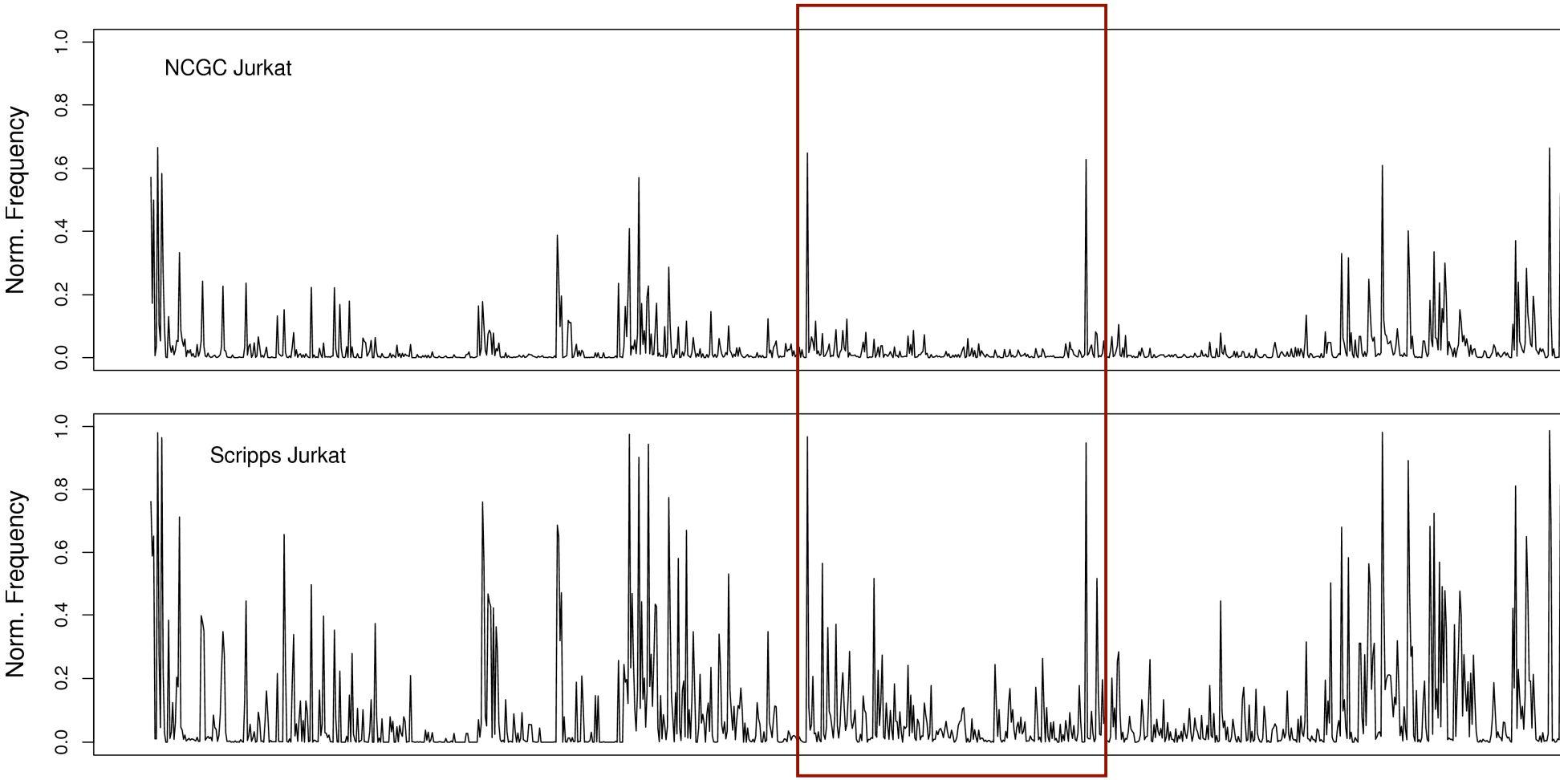
Predictions for the Scripps Cytotox dataset, using the original cutoffs (32% correct

	Nontoxic	Toxic
Nontoxic	26	90
Toxic	109	550

Predictions for the Scripps Cytotox dataset, using the NCGC cutoff (75% correct)

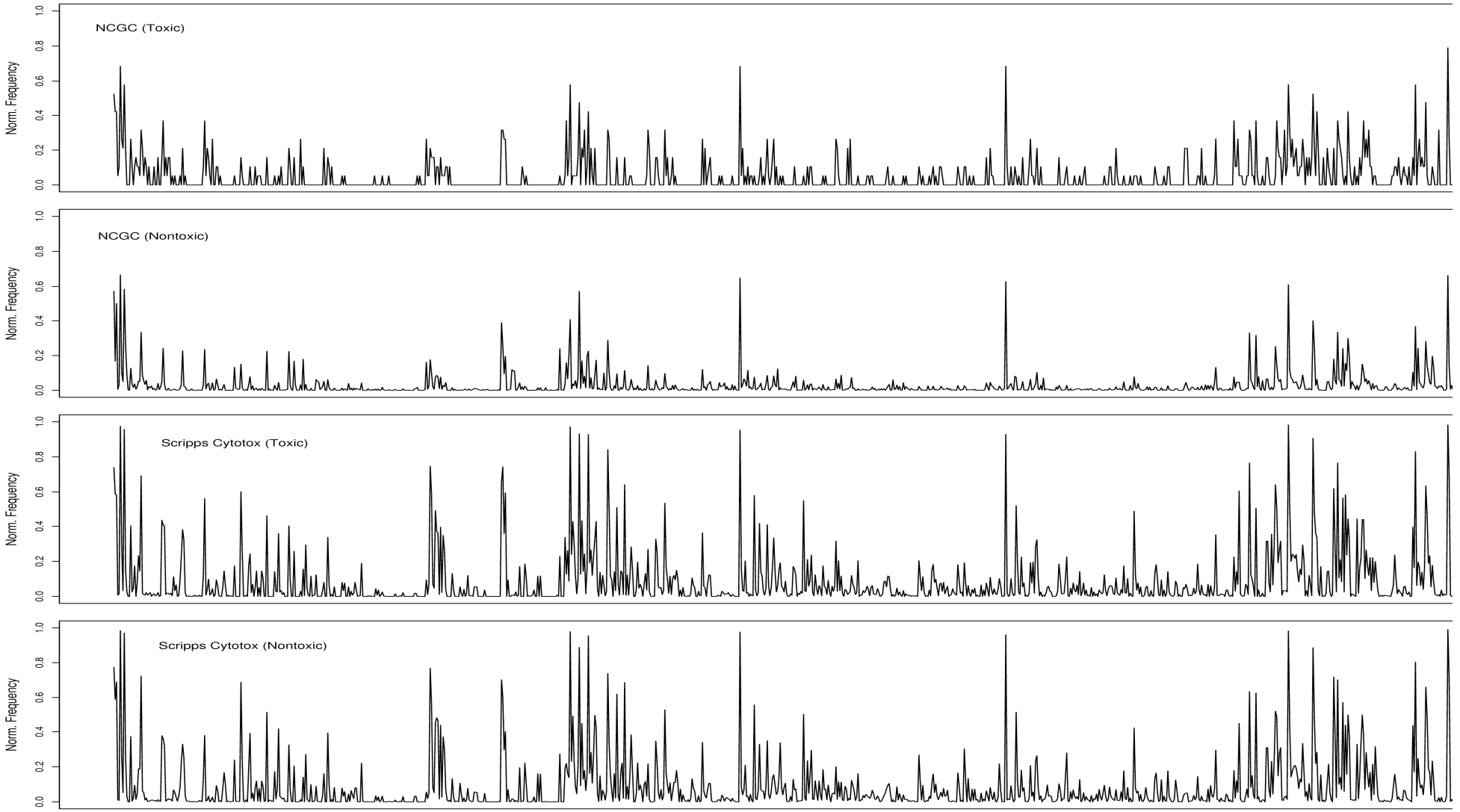
Comparing NCGC & Scripps Datasets

- Comparing the datasets as a whole



Comparing NCGC & Scripps Datase

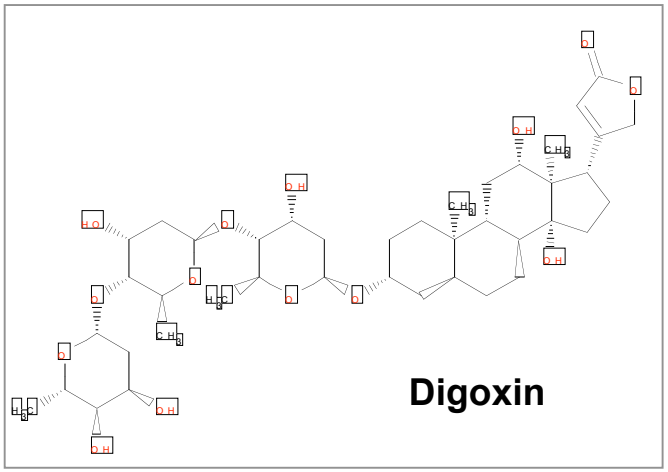
- Comparing datasets class-wise

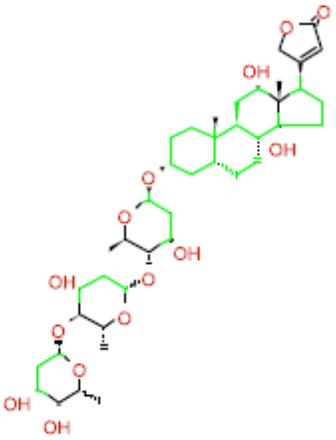


Important Features

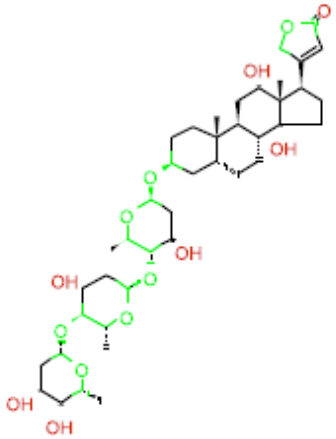
- We consider the NCGC Jurkat cell line
- The 10 most important features for predictive ability across the ensemble leads to 53 unique important bits
- This is a total of 72 structural features
  - The toxic compounds are characterized by having larger number of these features, on average

Feature matches for example structure

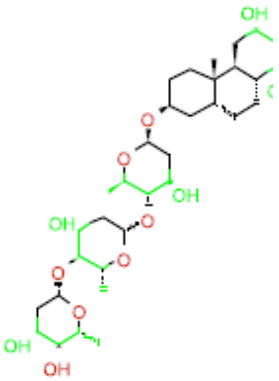




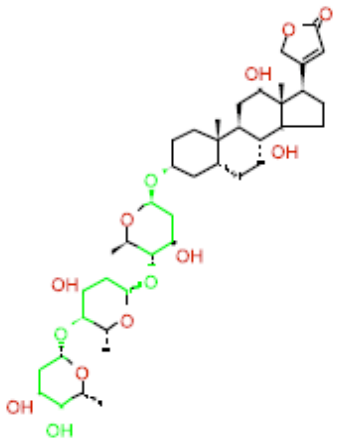
[C;D2](-;@[C,c])-;@[C,c]

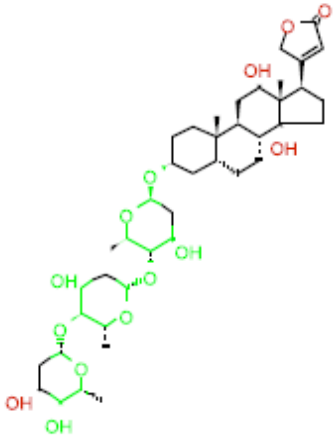


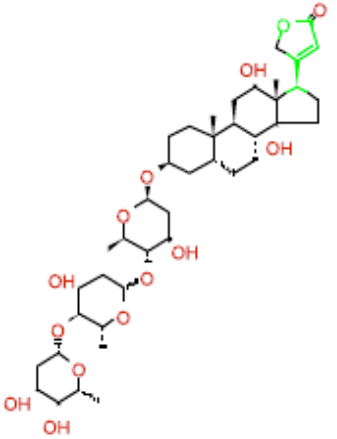
[O,o;D2](~[C,c])~[C,c]



[O,o]!@[C,c]@[C,c]@[C,c]!@



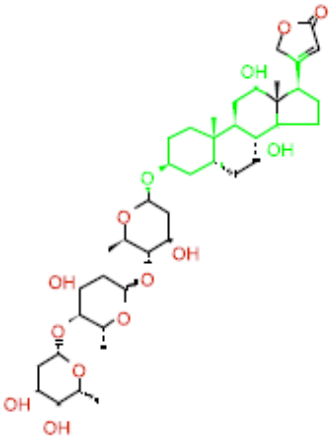


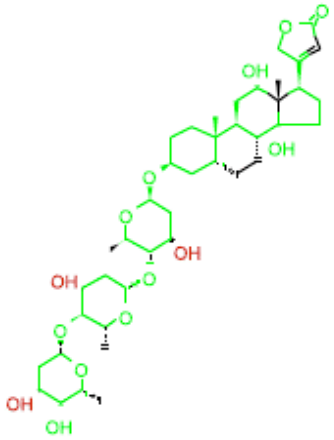


[*]-;@[A]-;@[A]=;@[A]-;!@|

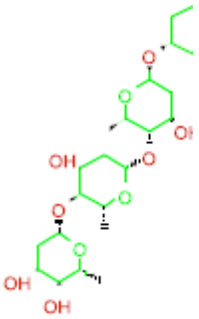
[O,o]~[C,c]~[O,o]~[C,c]~[C,c]~[O,o]

[O,o]!@[C,c]@[C,c]@[C,c]@[C,c]!@[O,o]





]@[*]@[*]@[*]!@[*]



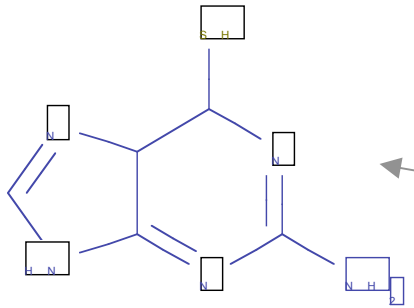
[O,o,S,s,Se,Te,Po]!@[C,c,Si,si,Ge,Sn,Pb]@[C,c,Si,si,Ge,Sn,Pb]@[C,c,Si,si,Ge,Sn,Pb]

[*]1@[*]@[*]@[*]@[*]@

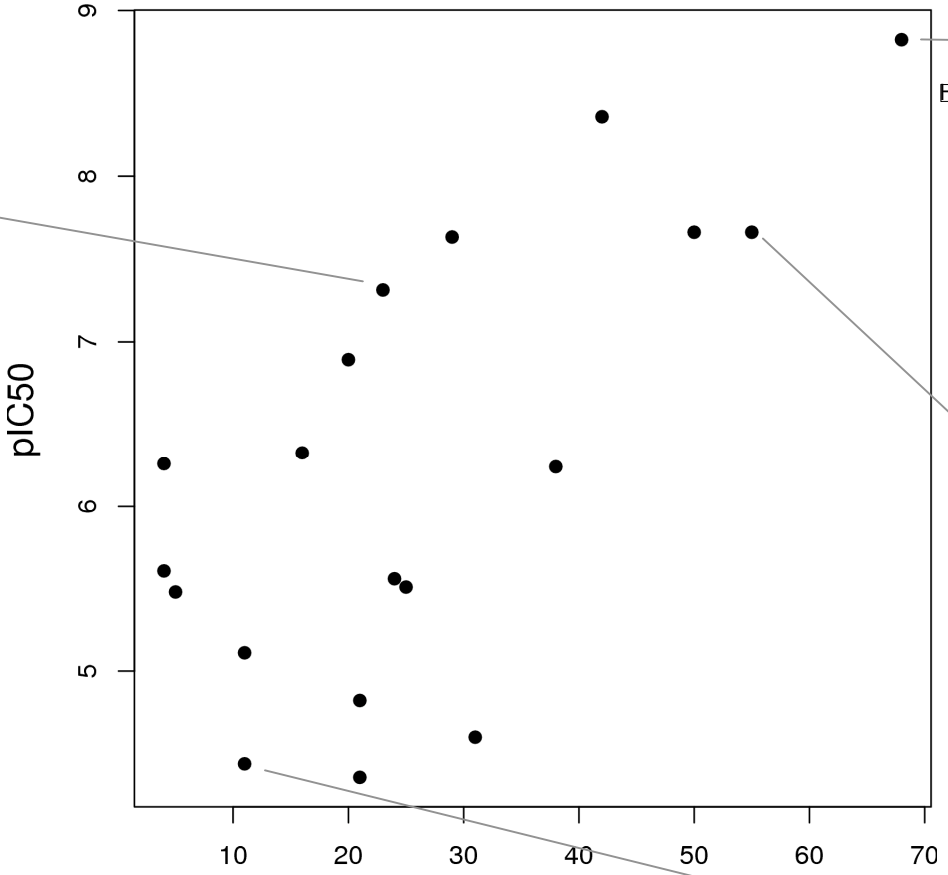
Important Feature
Animal Toxicity vs. Cytotoxicit

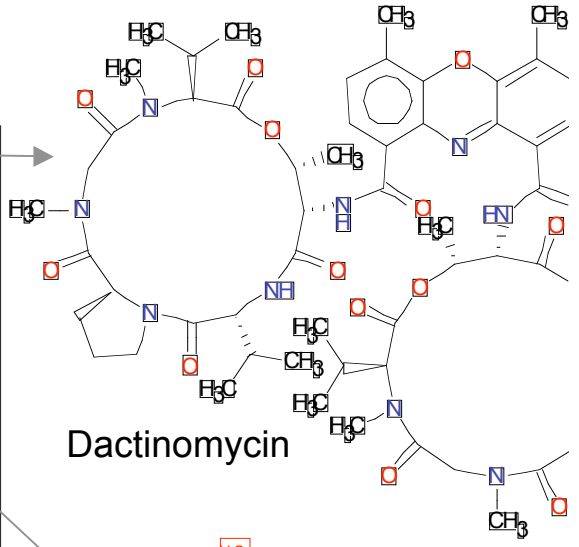
- The ToxNet (Mouse/IP) and NCGC Jurkat models have 130 important features in common
- These features are more common in the NCGC toxic compounds than in the NCGC nontoxic compounds
- The average number of these features present in the NCGC dataset, overall, is 18.8
  - Very low, might indicate that the NCGC model is not going be applicable to the ToxNet data

Toxicity vs No. of Features - NCGC Datas

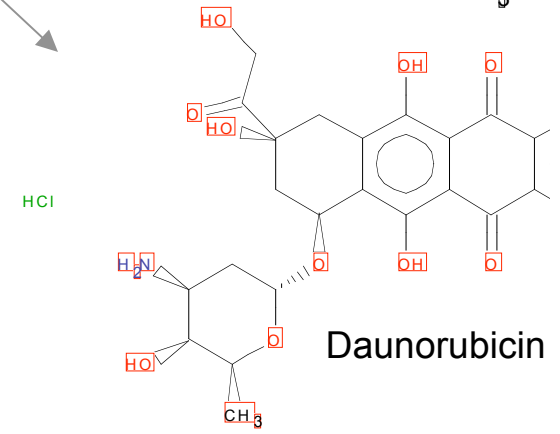


Leukerin deriv.

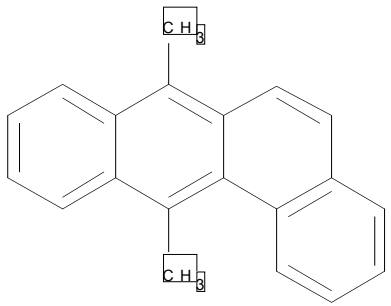




Dactinomycin



Daunorubicin



- The important features focus on the toxic class
- No correlation between number of features and pIC50 for nontoxic class

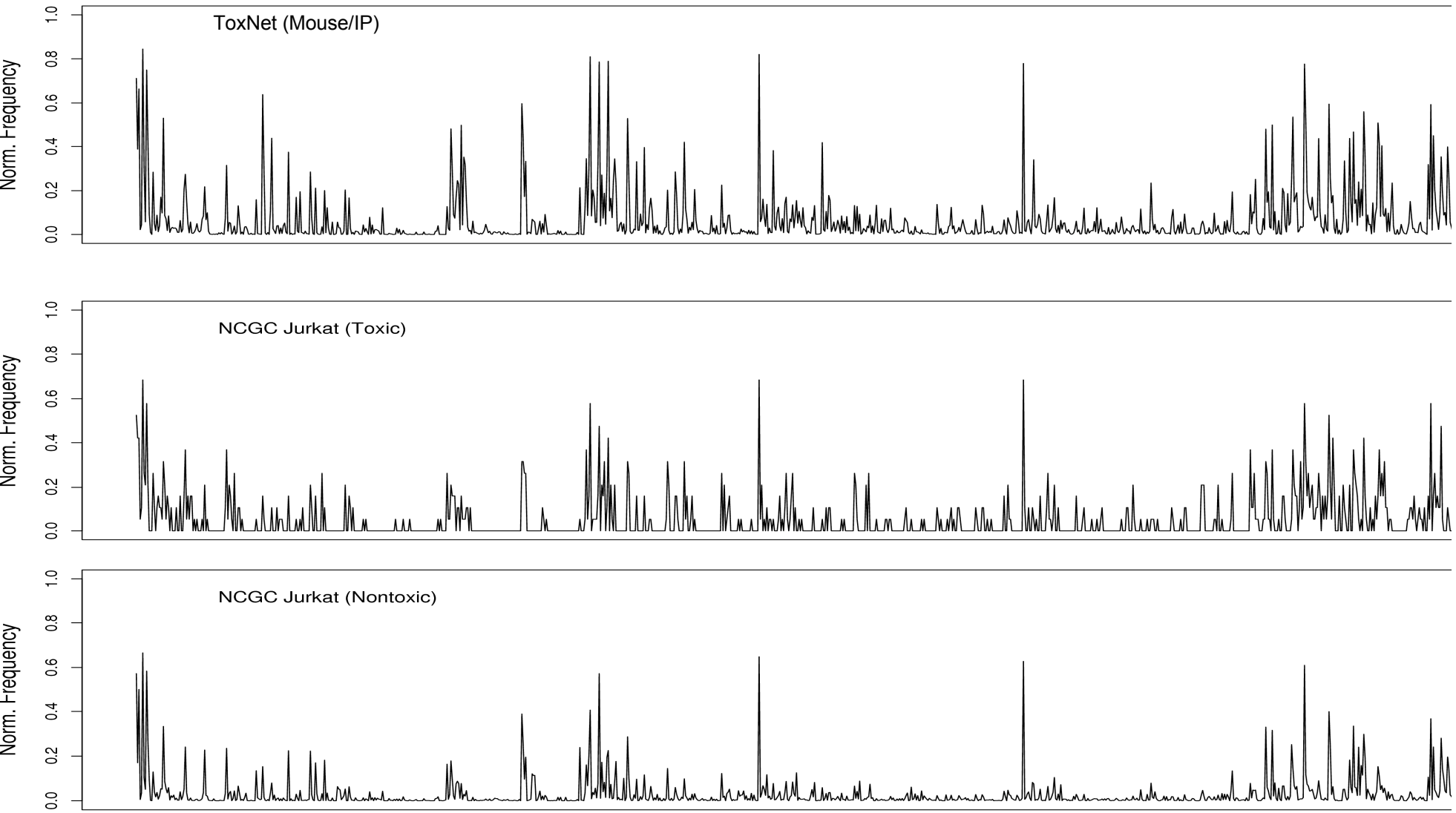
Predicting Animal Toxicity

	Nontoxic	Toxic
Nontoxic	12182	558
Toxic	32638	1265

Predictions for the ToxNet Mouse/IP dataset.
29% correct overall.
70% correct on the toxic clas

- Overall predictive performance is poor
- Possible causes
  - Poor sampling of the nontoxics during training
  - Feature distributions between the two datasets

Feature Distributions – ToxNet vs NCGC



Whats Next?

- Relate structural features to mechanisms of toxicity
- Incorporate these into models / build class models
  - Different cell-lines vs. animal toxicity
  - Structural features vs. mechanisms?
- Based on prediction confidence and model applicability, can we suggest alternative assays?
- Use the vote fraction & common bit counts to prioritize compounds, which may be toxic
  - Improve assessment of model applicability

Summary

- Applying models to predict other datasets is a tricky affair
  - Are the features distributed in a similar manner between training data & the new data?
  - Do toxic/nontoxic labels transfer between datasets
- More secondary data required
  - But this is not the final solution since the NCGC dataset is small but leads to (some) good models

Summary

- Fingerprints may not be the optimal way to get the be predictive ability
  - They do let us look at structural features easily
- We have investigated Molconn-Z descriptors
  - Preliminary results don't indicate significant improvements
- We cannot globally model animal toxicity based on cytotoxicity
  - Animal data sets are biased to toxic compounds
  - Different structural classes behave differently (mechanism action, metabolic effects

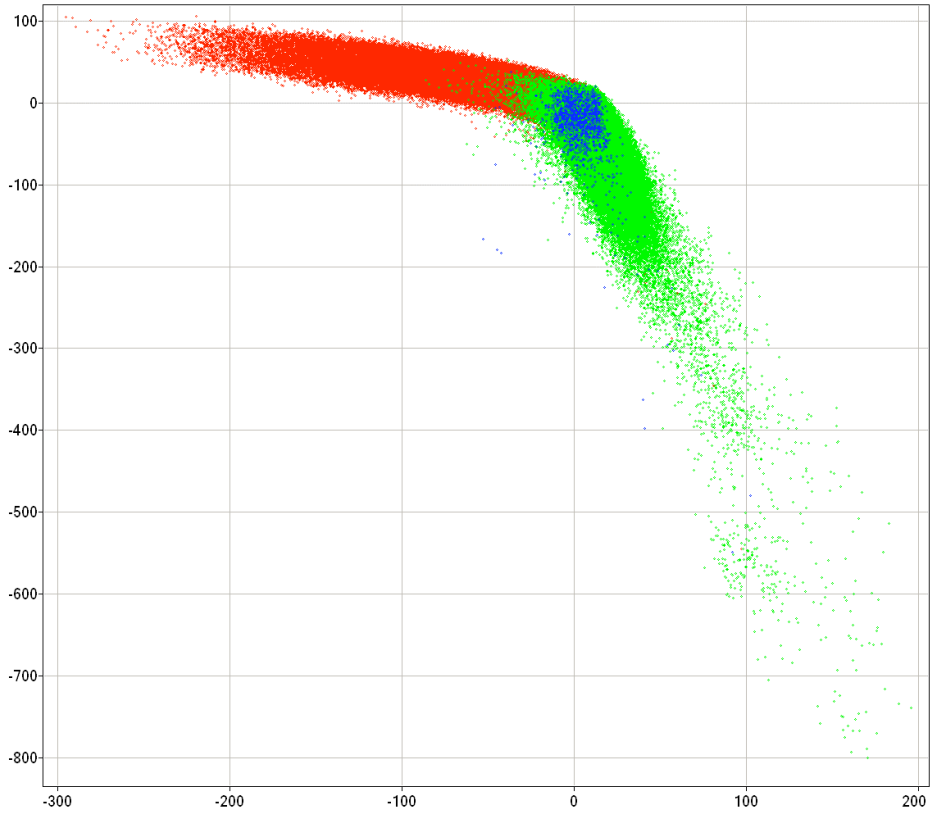
Acknowledgements

- MLSCN data sets / PubChem
- NCGC
- Scripps
  - Screening (Peter Hodder)
  - Informatics (Nick Tsinoremas, Chris Mader)
  - Hugh Rosen
- Alex Tropsha, UNC
- Digital Chemistry
- Tudor Oprea, UNM

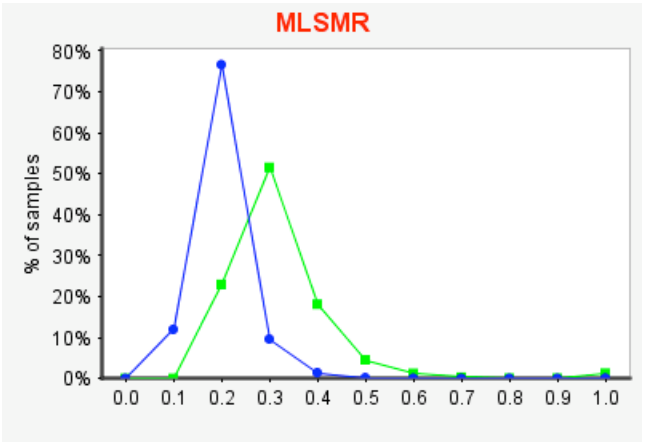
- NIH

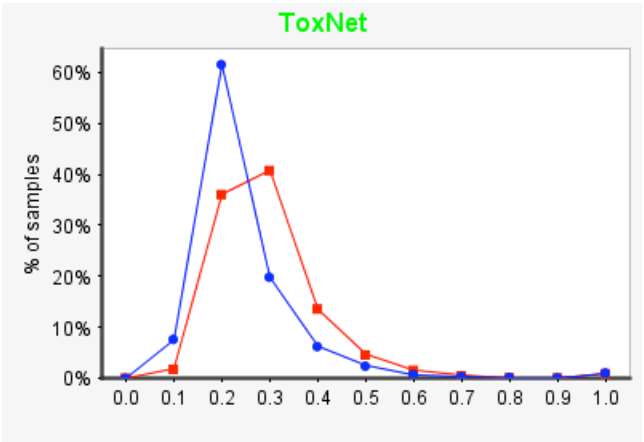
Extras

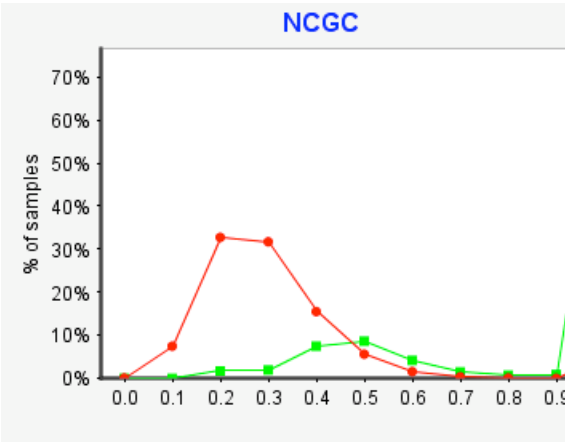
Structure Sets: Fingerprint Similarity



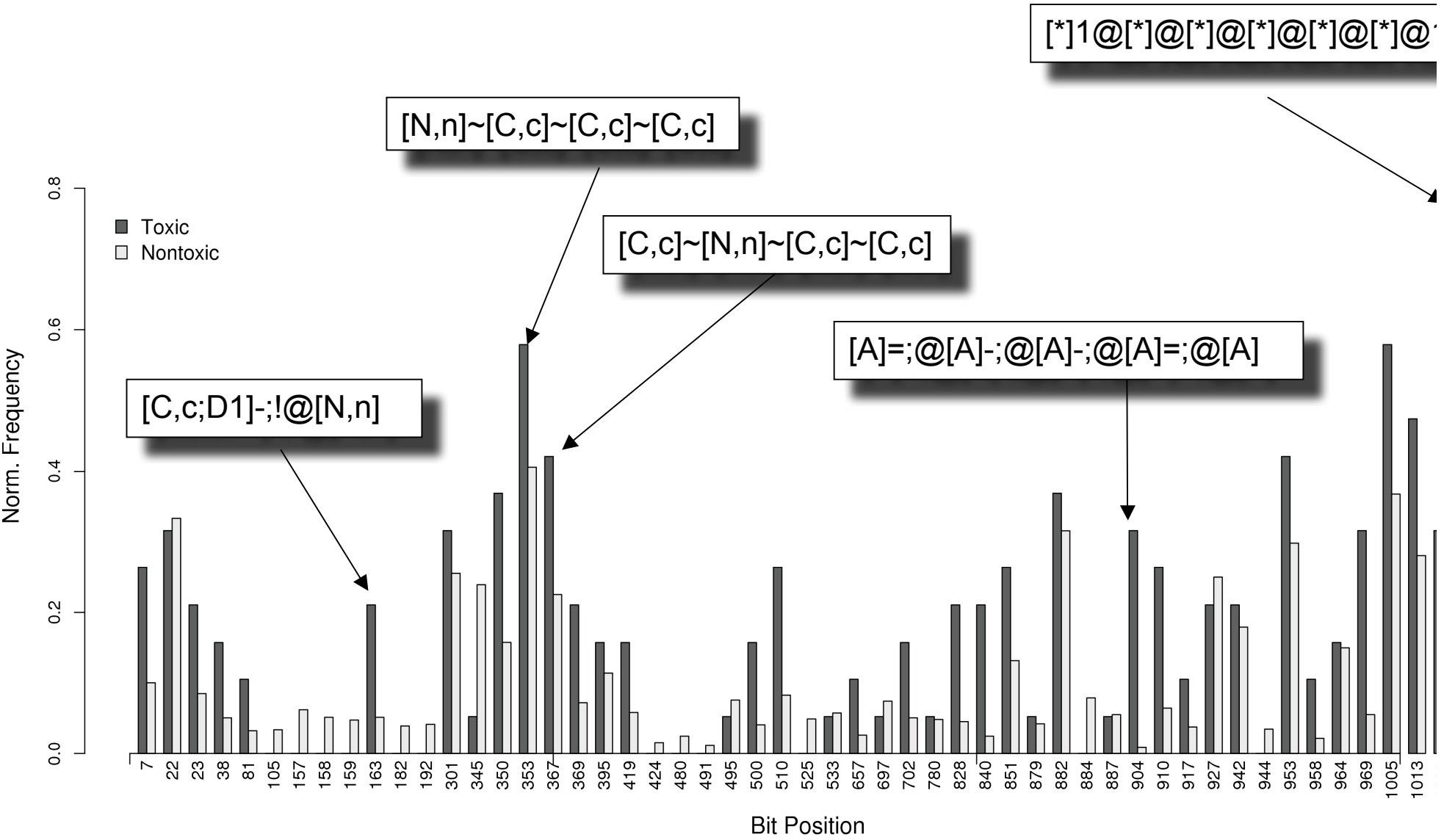
- Only a small fraction of MLSMR structu are similar to ToxNet structures; and vi versa; 4 to 5 % of MLSMR and ToxNet at least one >50 % similar structure to other
- NCGC structures are much more simila ToxNet (86% >50 % max similar) than MLSMR (9% >50 % max similar)





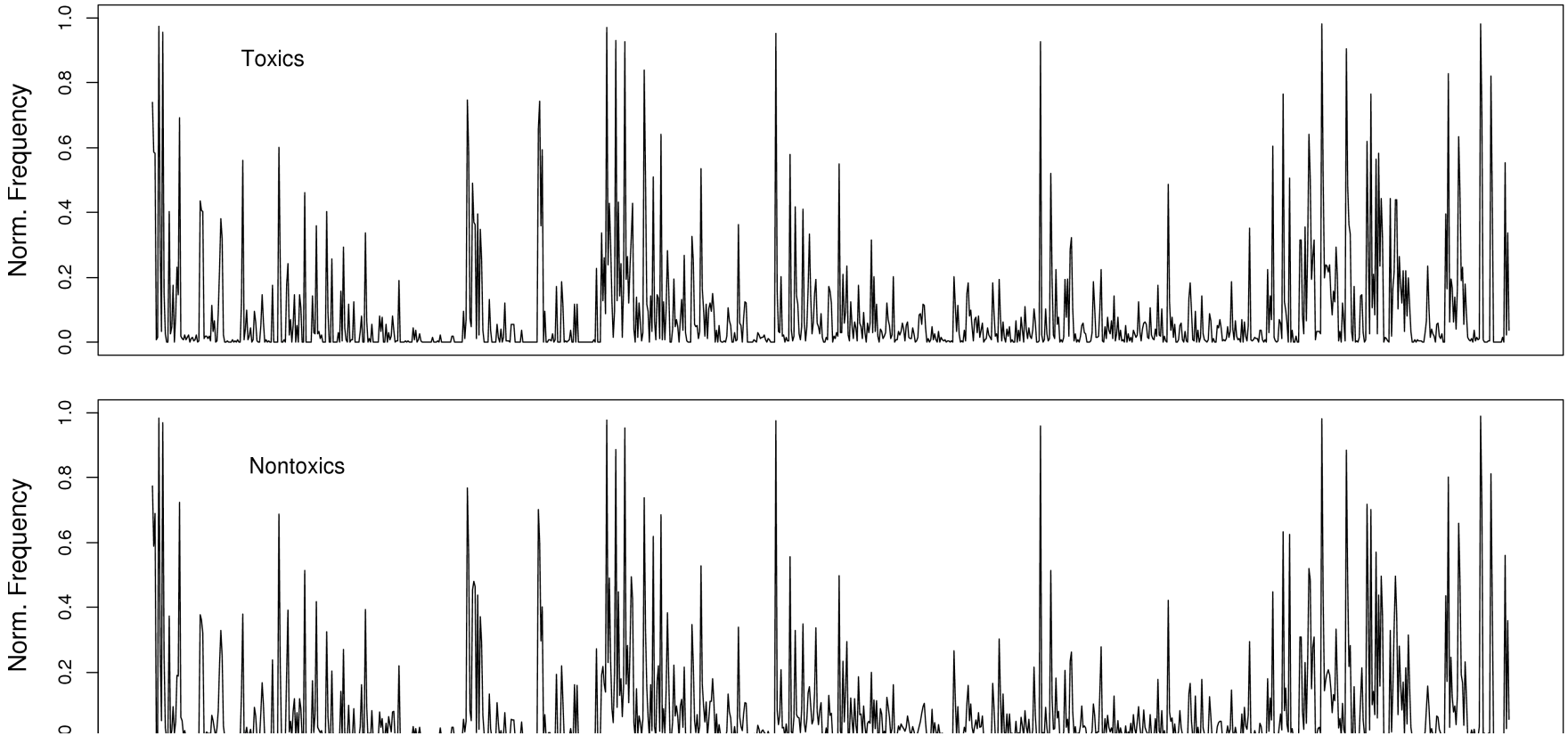


Important Features - Distributions



Are The Cytotox Classes Distinct?

- Poor predictive ability may be explained by the lack of separation between toxic & nontoxic
- *Normalized Manhattan distance = 0.017*



Are The Cytotox Classes Distinct?

- But the situation is a little better if we just look at the *important bits*
- *Normalized Manhattan distance = 0.06*

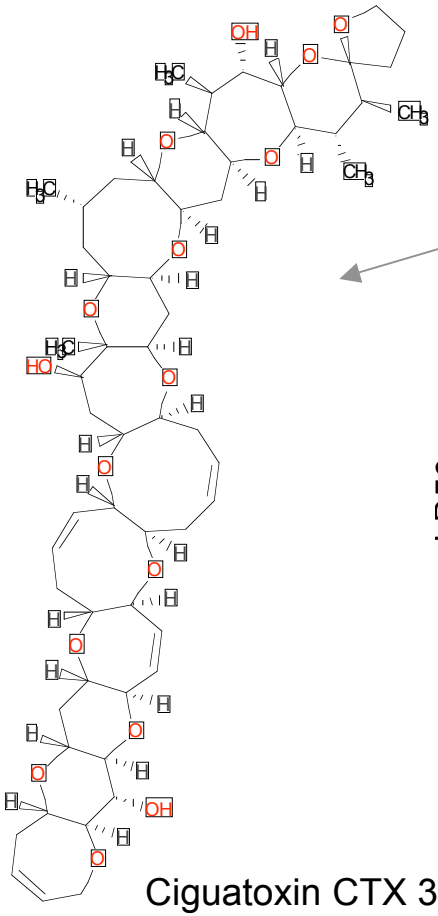
Standardization Issues - Data

- Extracting data sets out of PubChem requires manual curation and post-processing and aggregation of data
  - No standard measures or column definitions
  - Activity score and outcome only valid within one experimen
  - Assay results are not globally comparable
  - No standardization of assay format (e.g. type, readout, etc.)
  - Limited ability to query PubChem for specific data sets
    - `rpubchem` package for R is one option
  - Need better way to access specific bulk data sets
  - No aggregation of assay (sample) data by compound
- PubChem seems better suited to browse individual da than access large standardized data sets

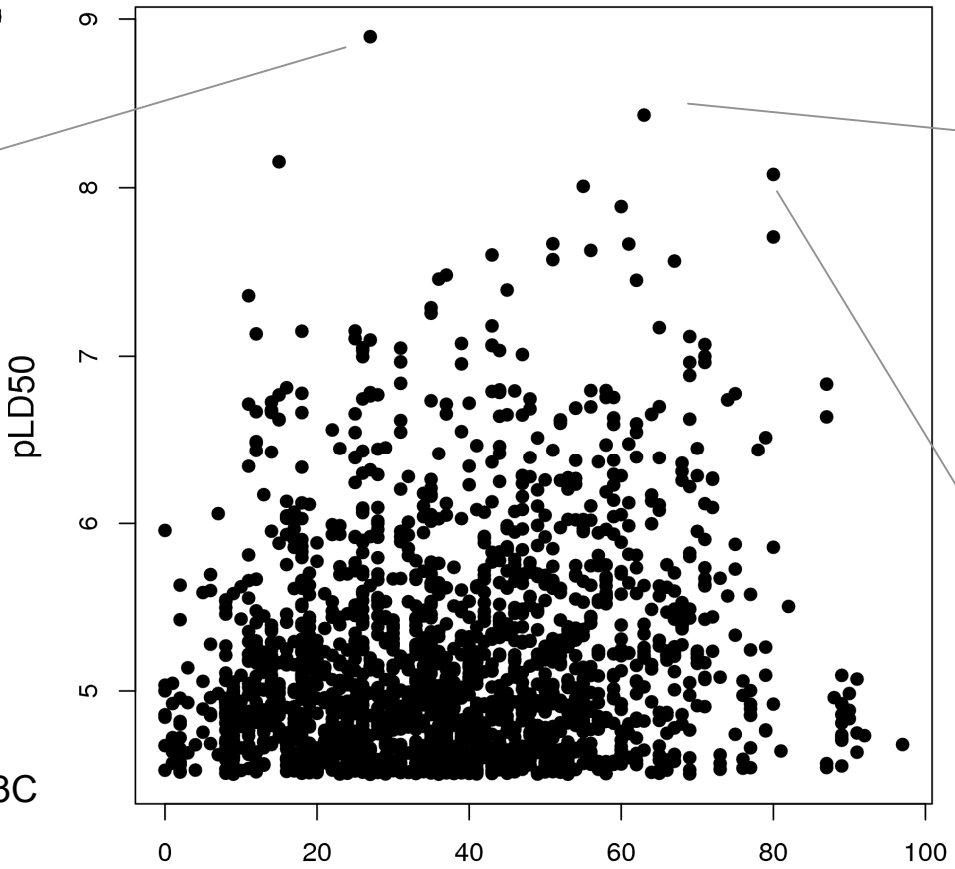
Model Deploymen

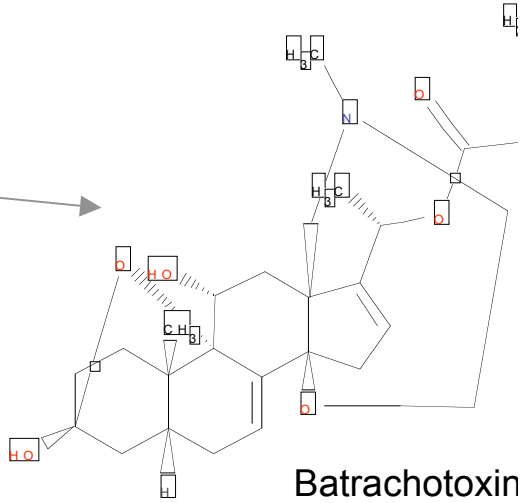
- Final models are deployed in our R WS infrastructure
  - Currently the Scripps Jurkat model is available
- Model file can be downloaded
  - http://www.chembiogrid.org/cheminfo/rws/mlist
- A web page client is available at
  - http://www.chembiogrid.org/cheminfo/rws/scripps
- Incorporated the model into a Pipeline Pilot workflow

Toxicity vs No. of Features - Mouse

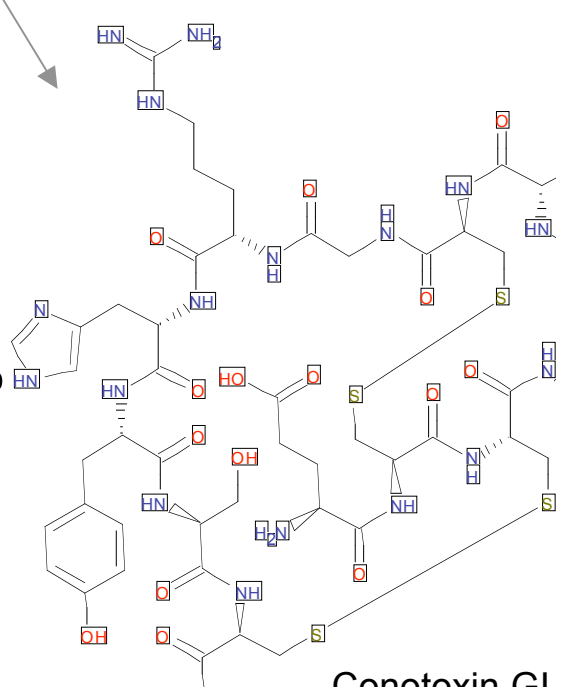


Ciguatoxin CTX 3C





Batrachotoxin



Conotoxin GI

- No correlation between the number of important features and the pLD50 for the toxic class