The Development and Deployment of Predictive Toxicology Models

Rajarshi Guha
CICC, Indiana University

and

Stephan Schürer
Scripps, FL

Broad Goals

- Develop methods and tools that facilitate probe development
  - MLSCN centers and to broader user groups
- 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

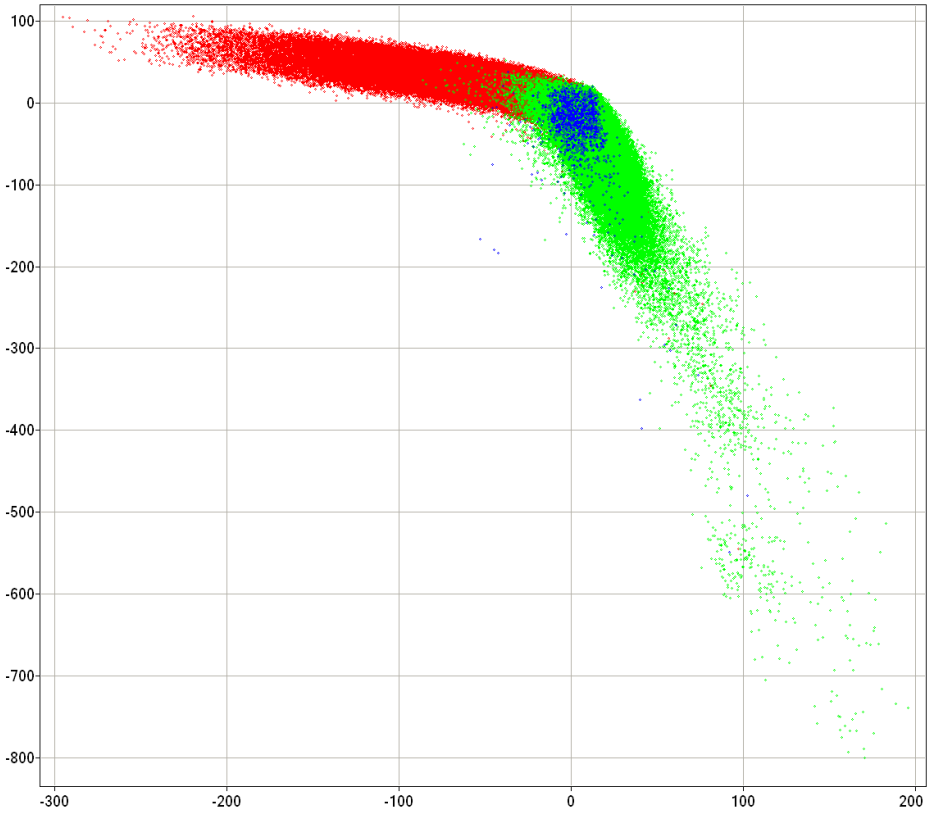
Cytotoxicity and animal toxicity

- Characterize toxicity datasets
  - Structurally and by activity (active chemical classes)
- Are cytotoxicity and animal toxicity related:
  - For which structural classes and mechanism of action does and does not cytotoxicity relate to animal toxicity?
- Model cytotoxicity and animal toxicity
  - Can we identify structural features correlated to toxicity?
  - How do we evaluate model applicability?
  - How do we deploy our final models?

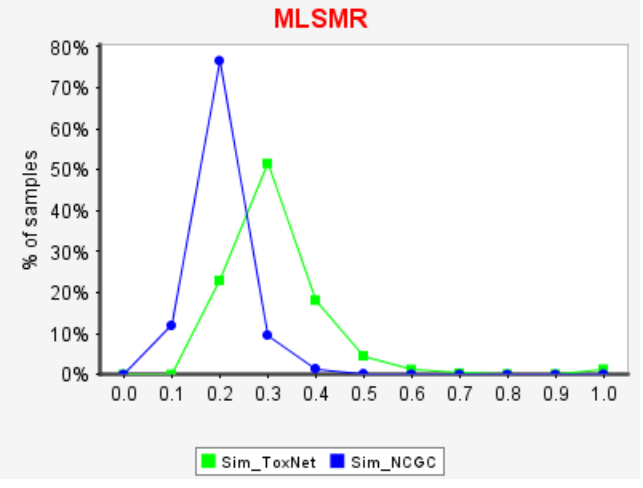
Datasets

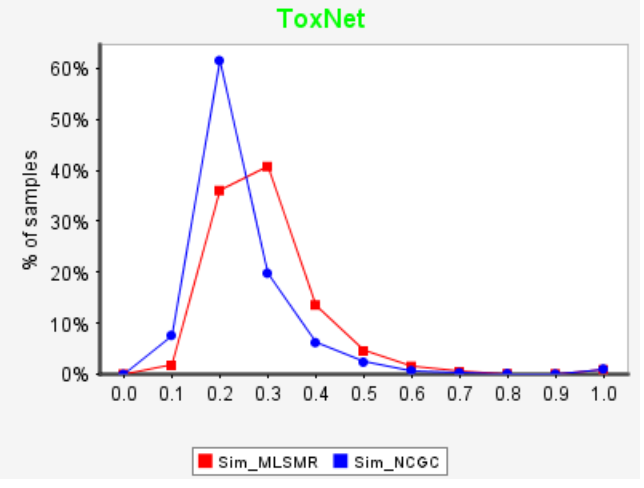
- Animal Acute Toxicity Data was extracted from the ToxNet database (available from MDL)
  - Selected only LD50 data for mouse and rat and three routes of administration
  - Summarized LD50 data by structure, species and route (140,808 LD50 data points, 103,040 structures)
  - Classified into Toxic/Nontoxic 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 IC50 data: 1,334 structure, 13 x 1,334 IC50 values for different cell lines

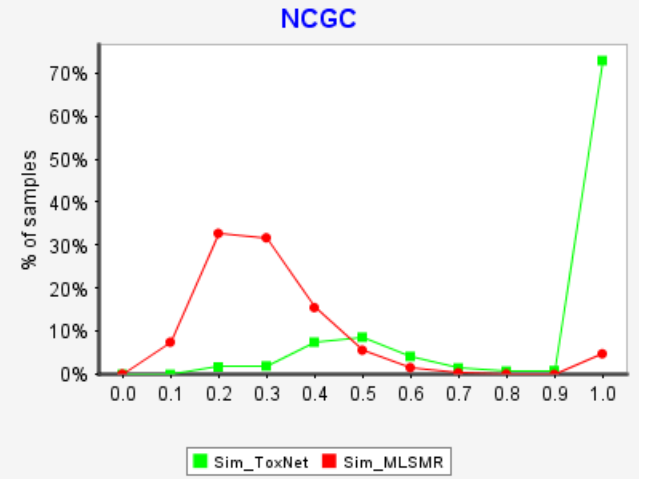
Structure Sets: Fingerprint Similarity



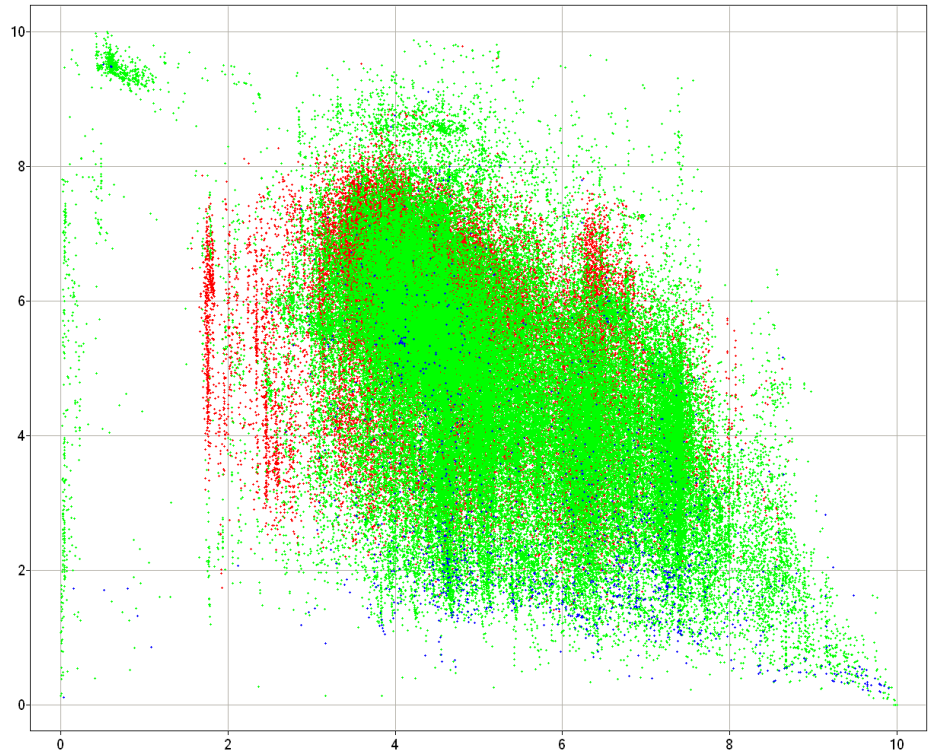
- Only a small fraction of MLSMR structures are similar to ToxNet structures; and vice versa; 4 to 5 % of MLSMR and ToxNet have at least one >50 % similar structure to each other
- NCGC structures are much more similar to ToxNet (86% >50 % max similar) than MLSMR (9% >50 % max similar)



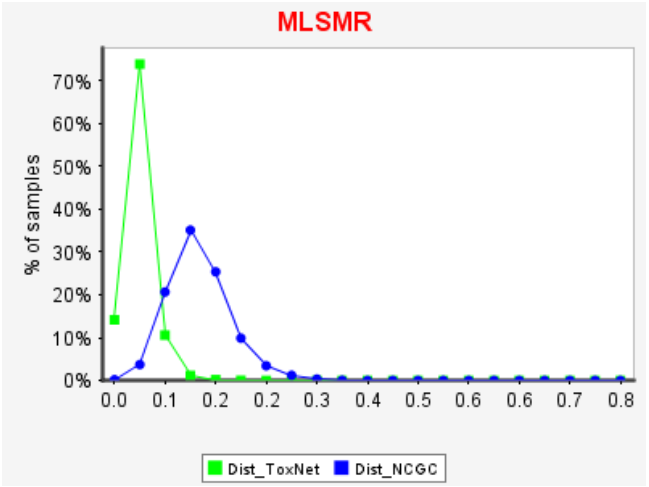


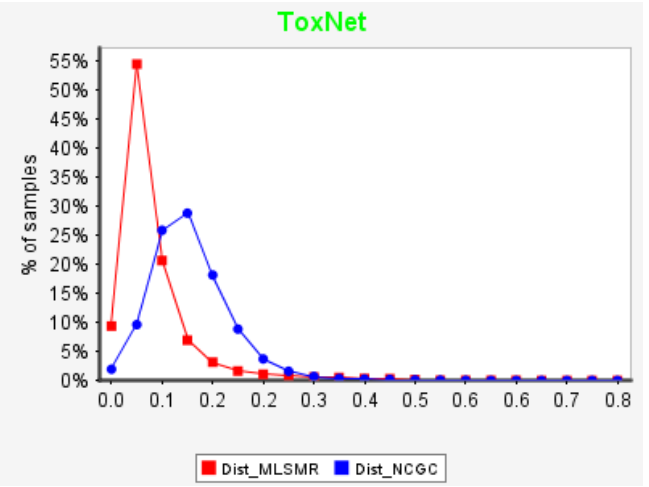


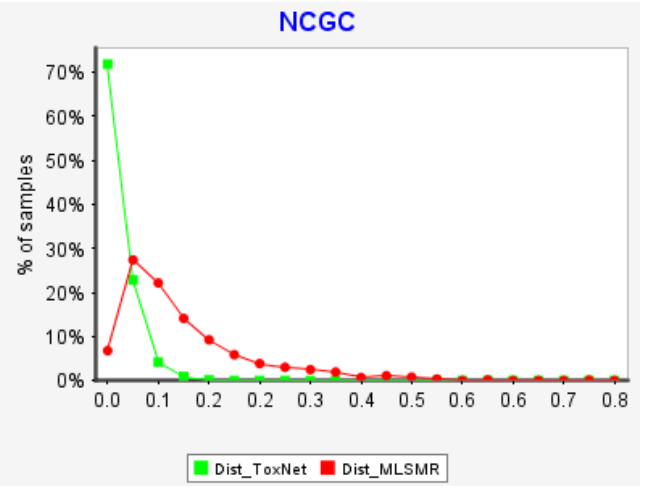
Structure Space - BCUTS



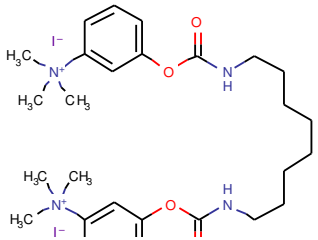
- Similar information to fingerprint similarity
- But BCUTS descriptors are more relevant to diversity than similarity

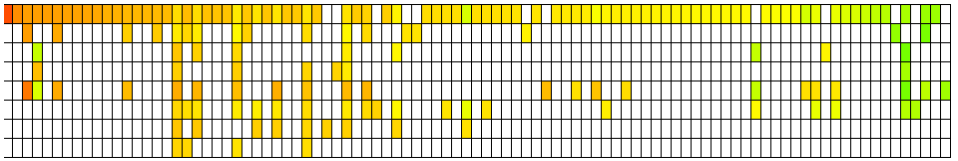


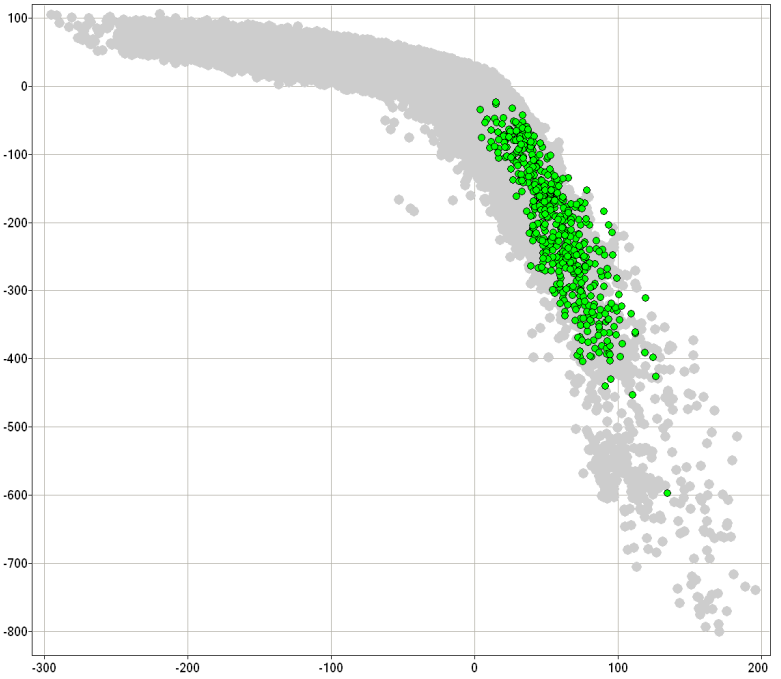


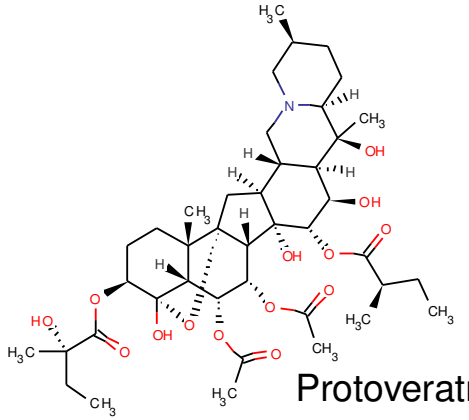


Examples of LD50 toxic classes

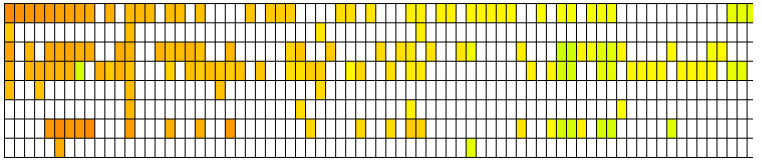


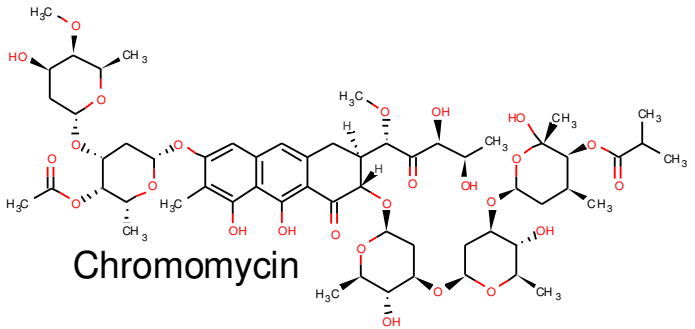




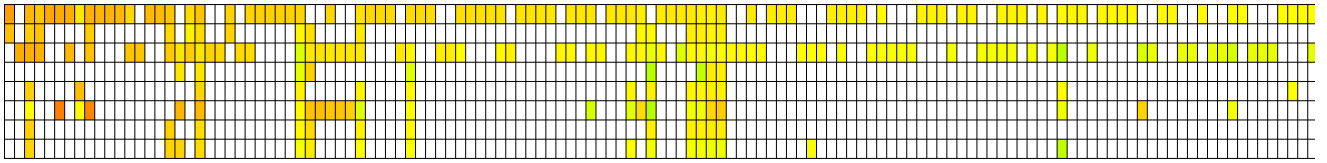


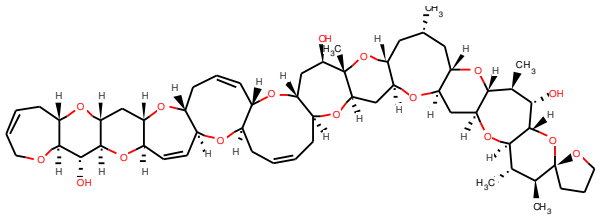
Protoveratrine



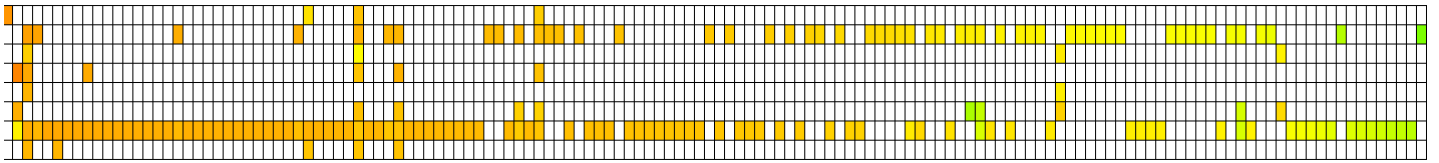


Chromomycin



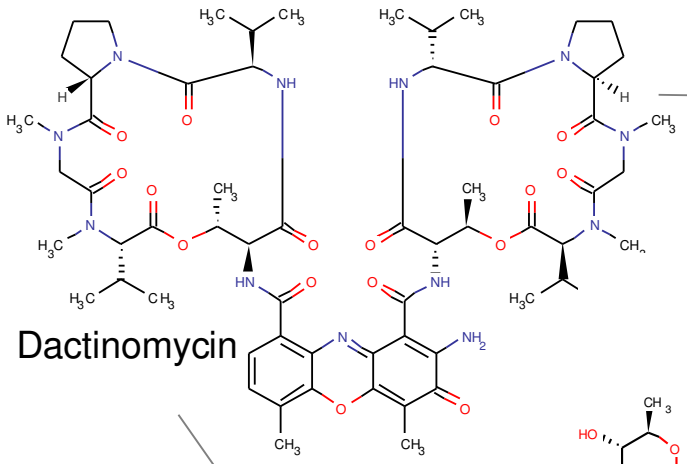


Ciguatoxin CTX 3C

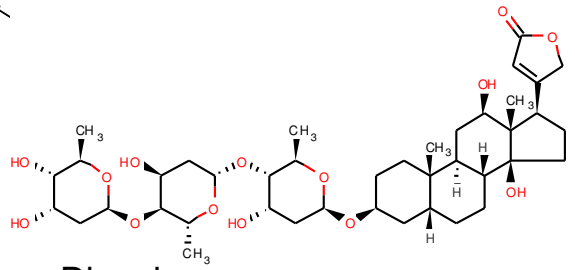


Interactive exploration of structural and activity data

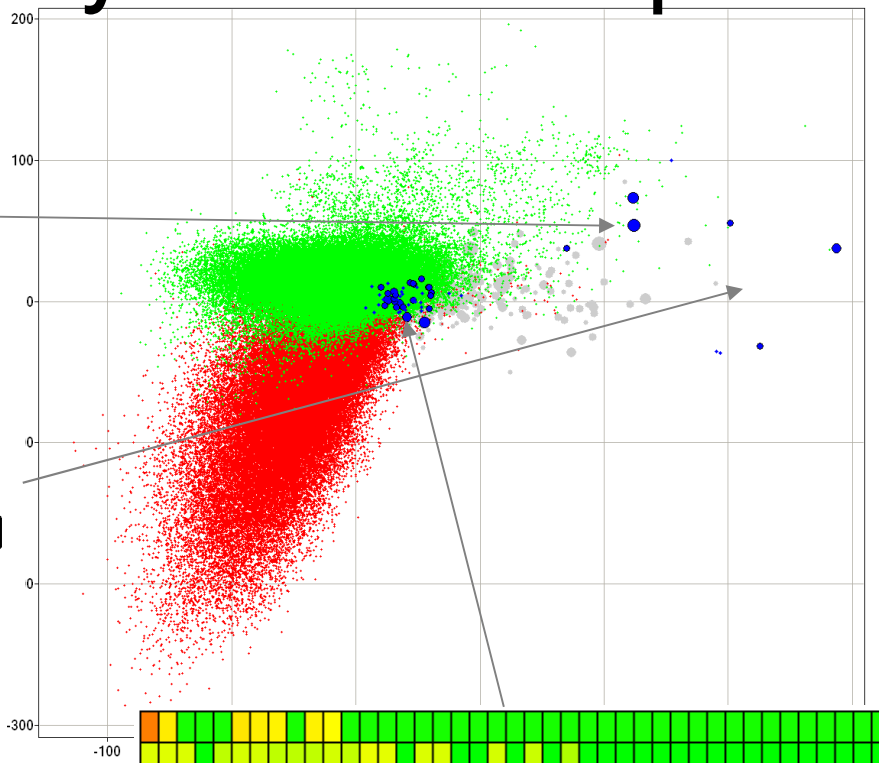
NCGC Cell toxicity data examples

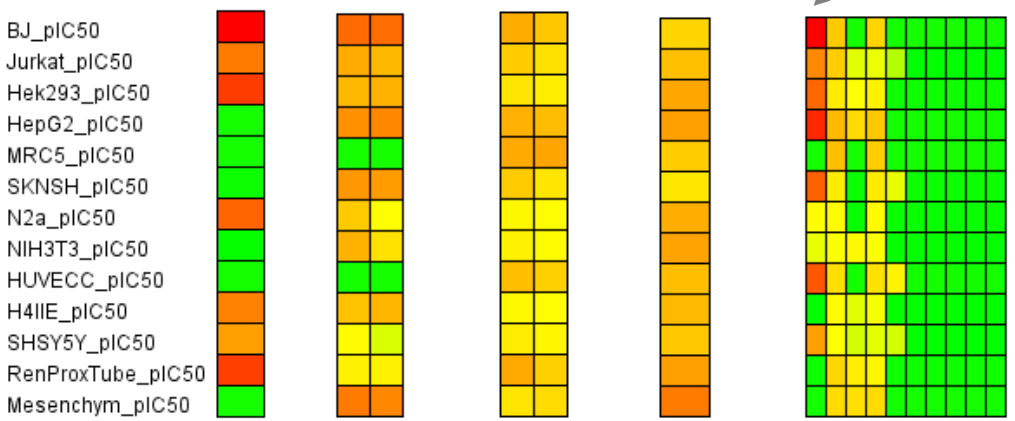


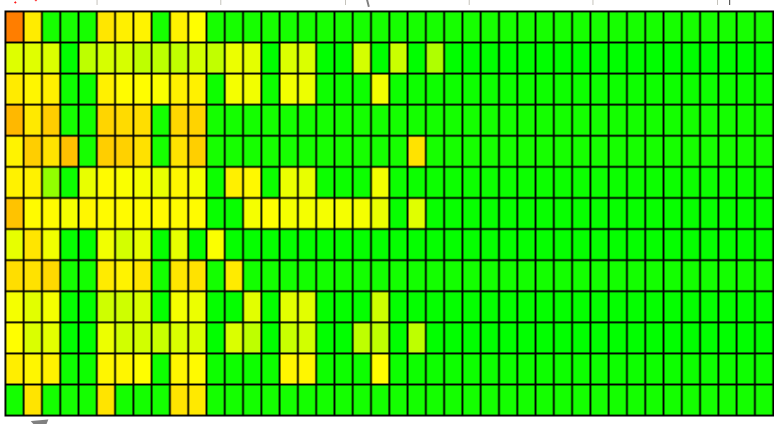
Dactinomycin

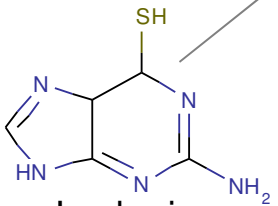


Digoxin

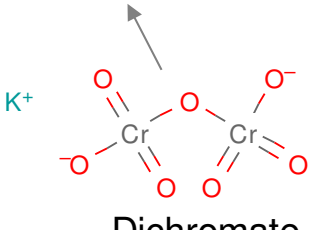




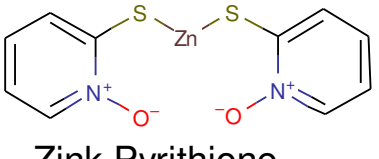




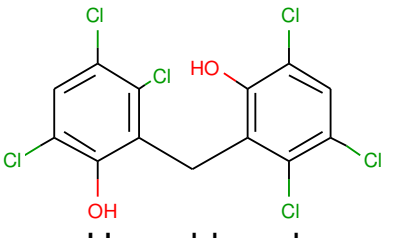
Leukerine



Dichromate

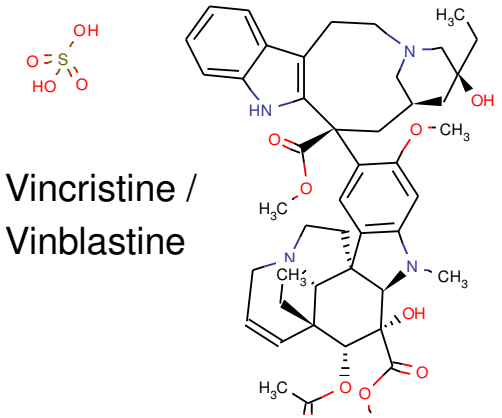


Zink Pyrithione

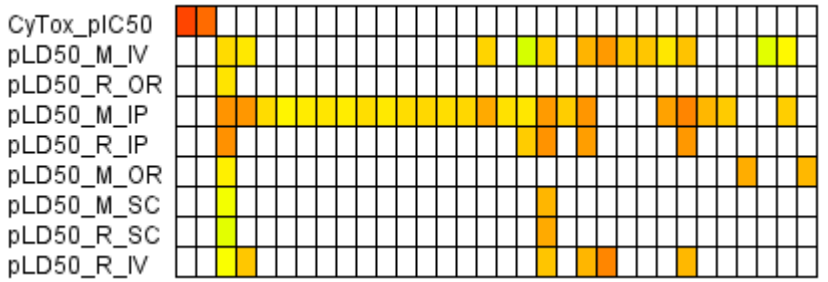


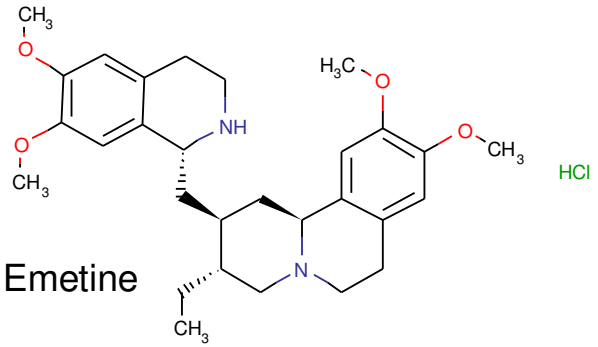
Hexachlorophene

Cytotoxicity and acute LD50 toxicity
Similar activity

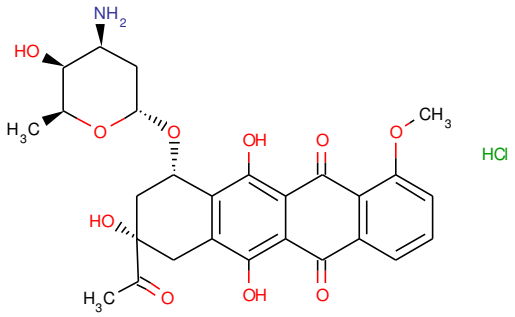


Vincristine /
Vinblastine

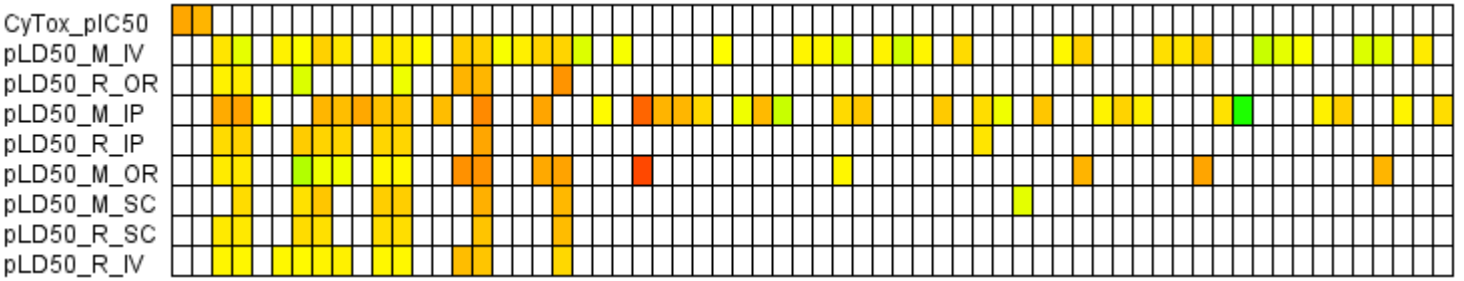




Emetine

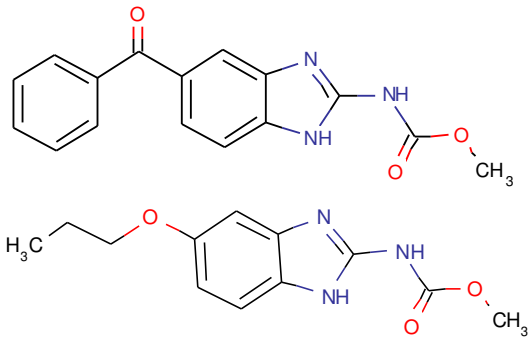


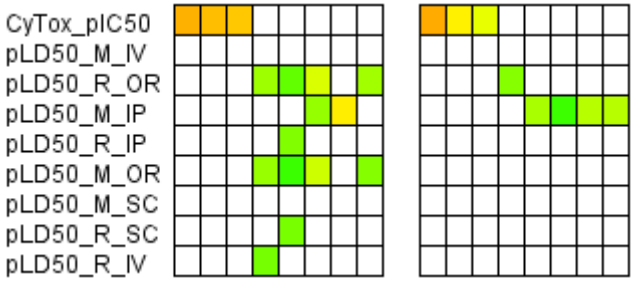
Daunorubicin / Doxorubicin

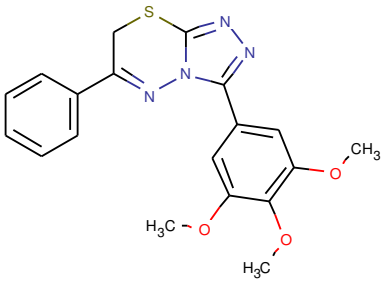


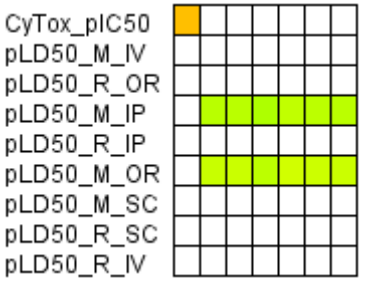
Cytotoxicity and acute LD50 toxicity

cytotoxic but less animal toxic

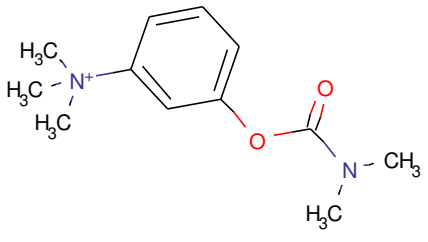




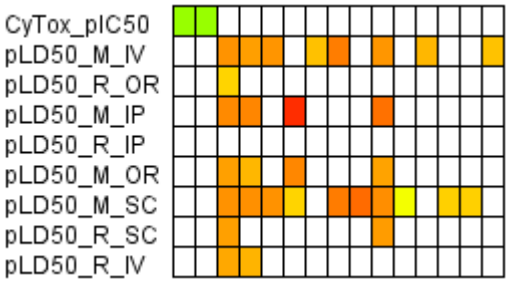


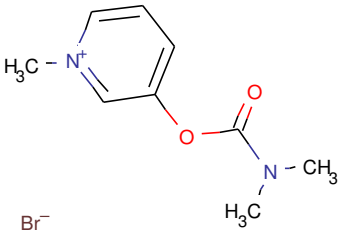


Less cytotoxic but greater animal toxicity

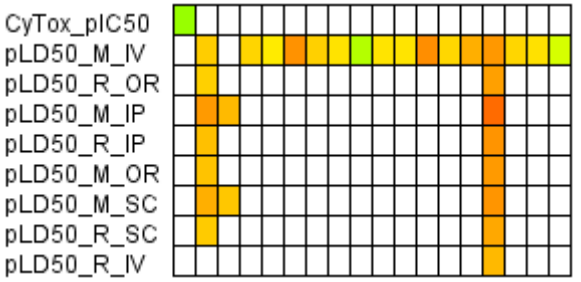


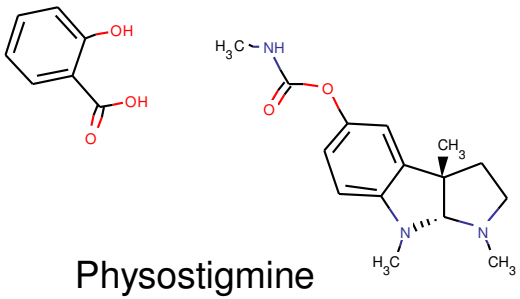
Neostigmine



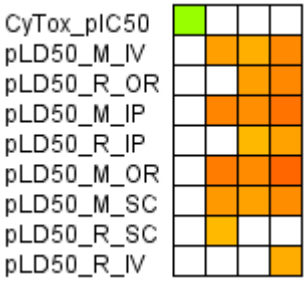


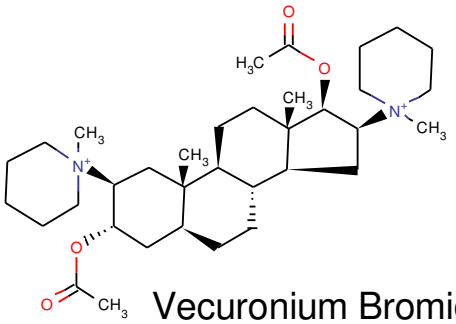
Pyridostimine



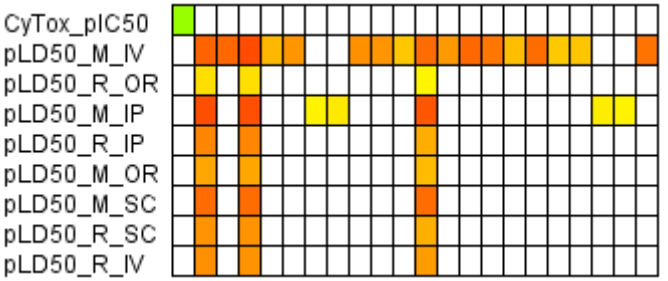


Physostigmine

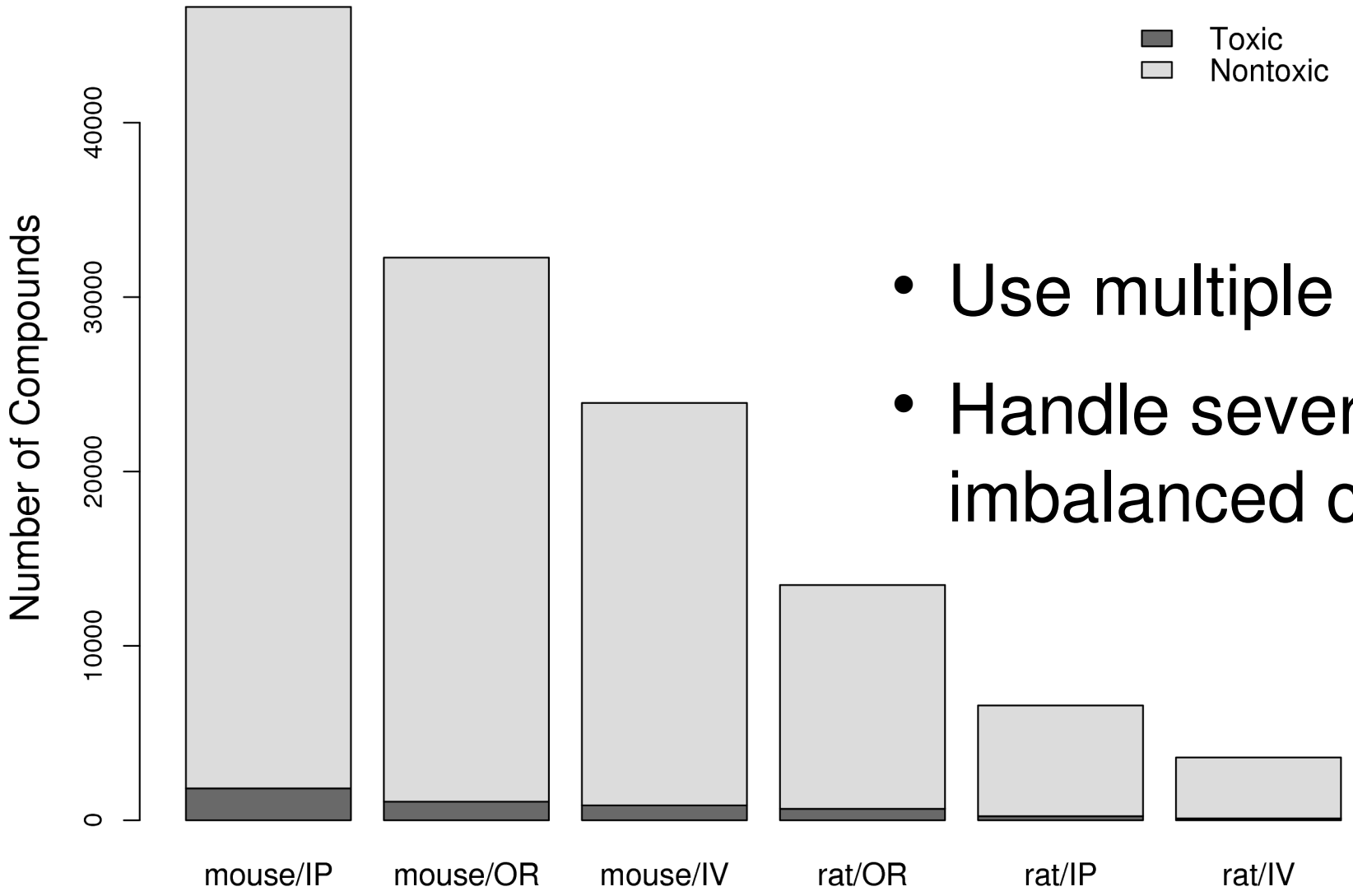




Vecuronium Bromide

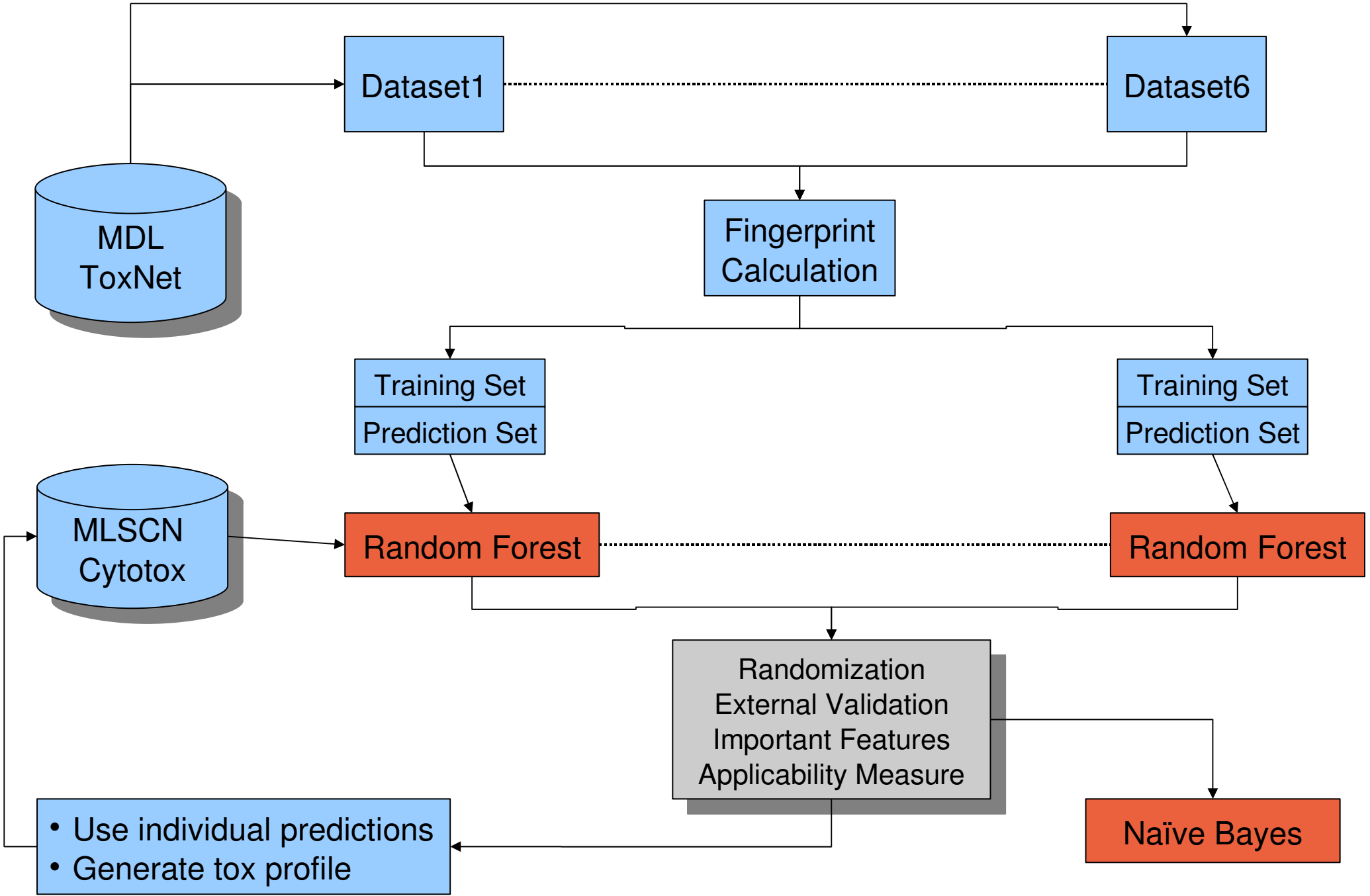


Animal Data



- Use multiple datasets
- Handle severely imbalanced classes

Modeling Protocol



Modeling Protocol

- Descriptors
  - 1052 bit BCI fingerprints
  - Our interest is in fragments indicative of toxicity
  - We don't know the exact mechanisms and thus cannot effectively select mechanism-specific descriptors

von Korff, M. and Sander, T., *J. Chem. Inf. Model.*, **2006**, *46*(2), 536-544
Casalegno, M. et al., *Chem. Res. Tox.*, **2005**, *18*(4), 740-746
von der Ohe, P.C. et al., *Chem. Res. Tox.*, **2005**, *18*(3), 536-555

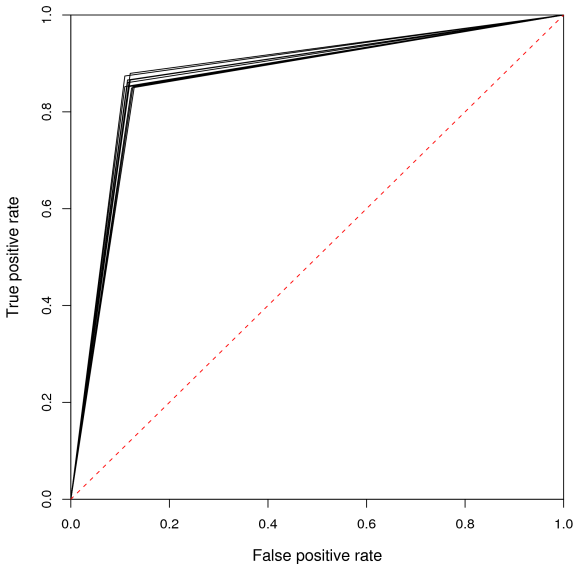
Modeling Protocol

- Random forest models for each dataset
  - Avoids feature selection
  - Identify important features
- We use a sampling procedure for the individual models, to avoid imbalanced classification
  - We take all toxics and an equal number of non-toxics
  - Repeat this 10 times, always keeping the toxics constant
- Each model is built on a training set
- All models tested on a fixed prediction set

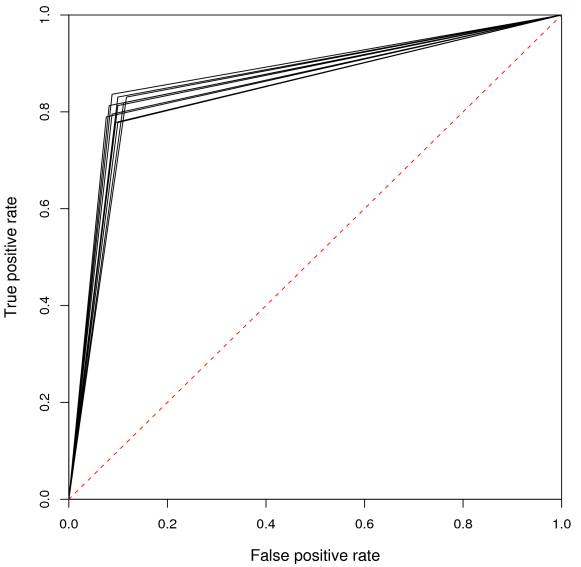
Modeling Protocol

- We end up with 6 random forest ensembles
- Within a given ensemble we find the 100 most important features over all 10 models
- Use these features to develop a Naïve Bayes model ensemble
  - Mainly used to see whether the important subset is better than using all 1052 features

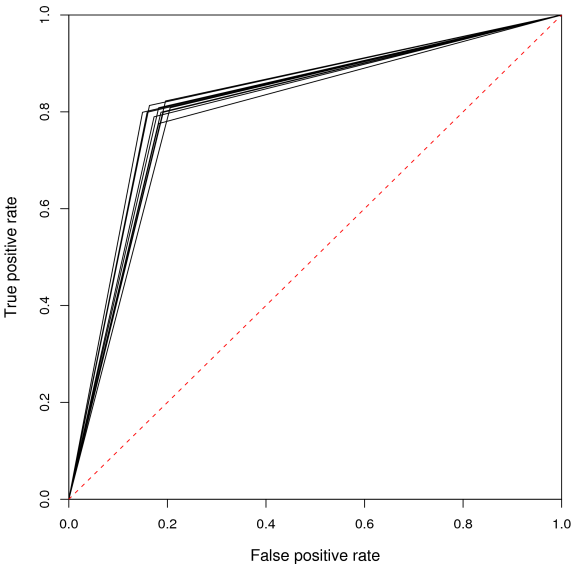
ROC Curves (Random Forests)



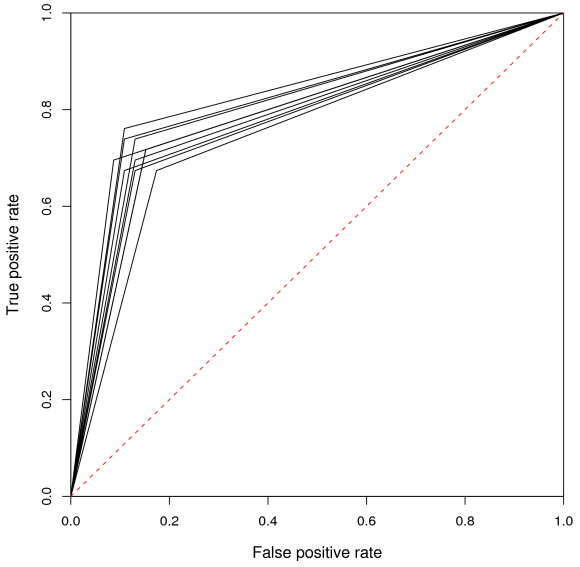
Mouse / IP



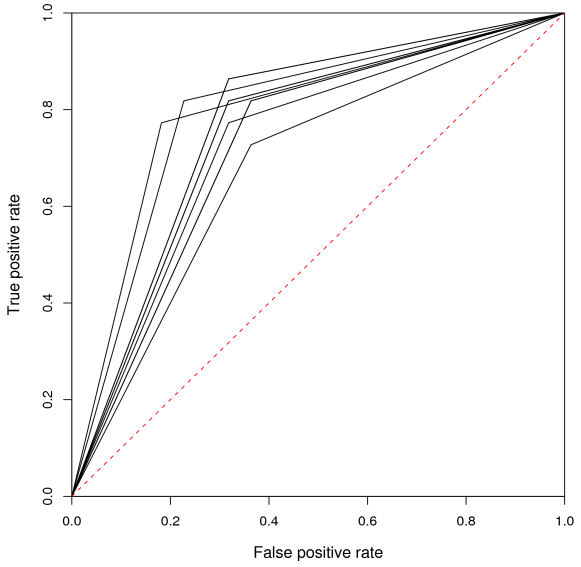
Mouse / IV



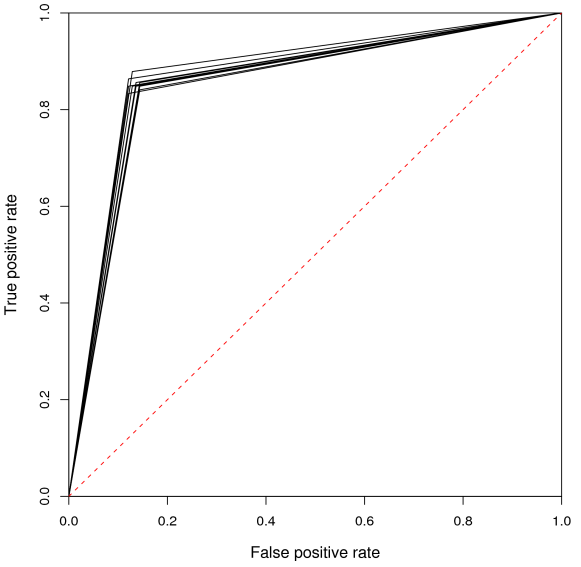
Mouse / OR



Rat / IP

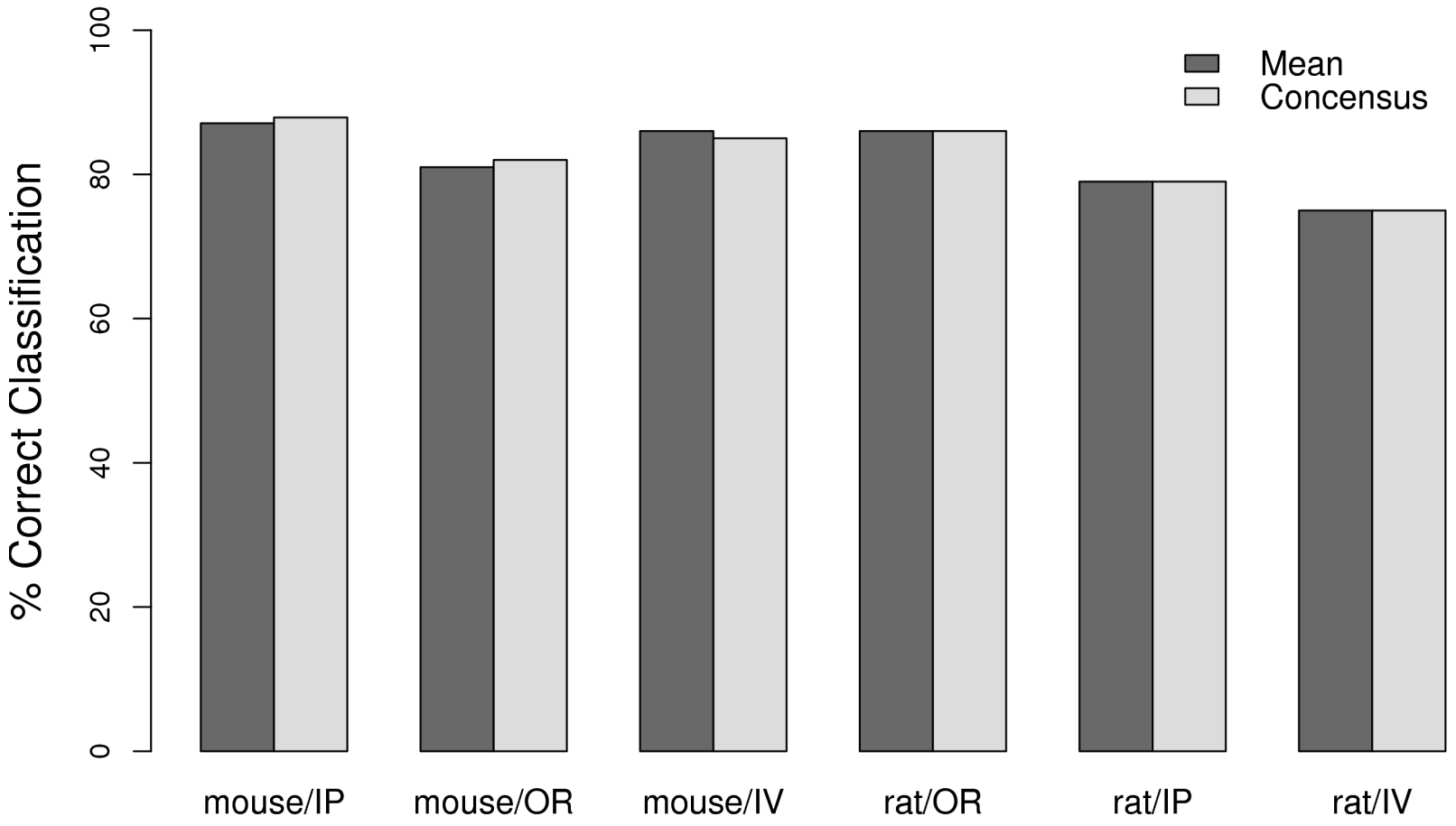


Rat / IV



Rat / OR

Predictive Performance (Random Forest)



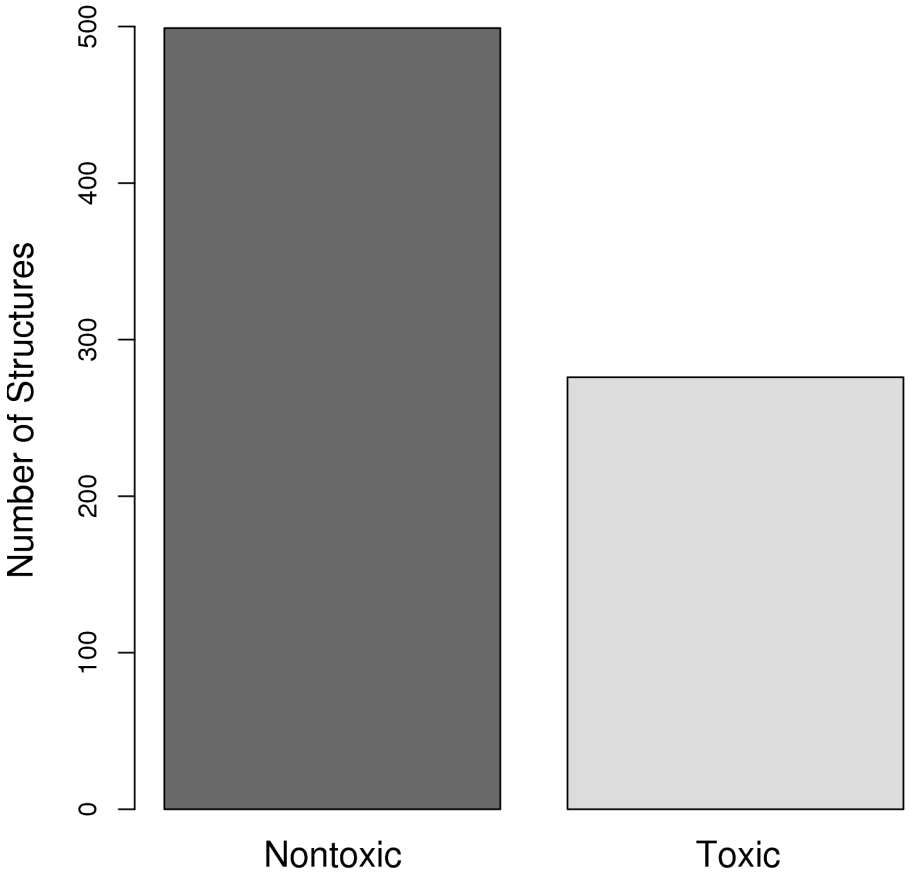
Percent correct classification of the prediction set for each species/mode

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 the 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

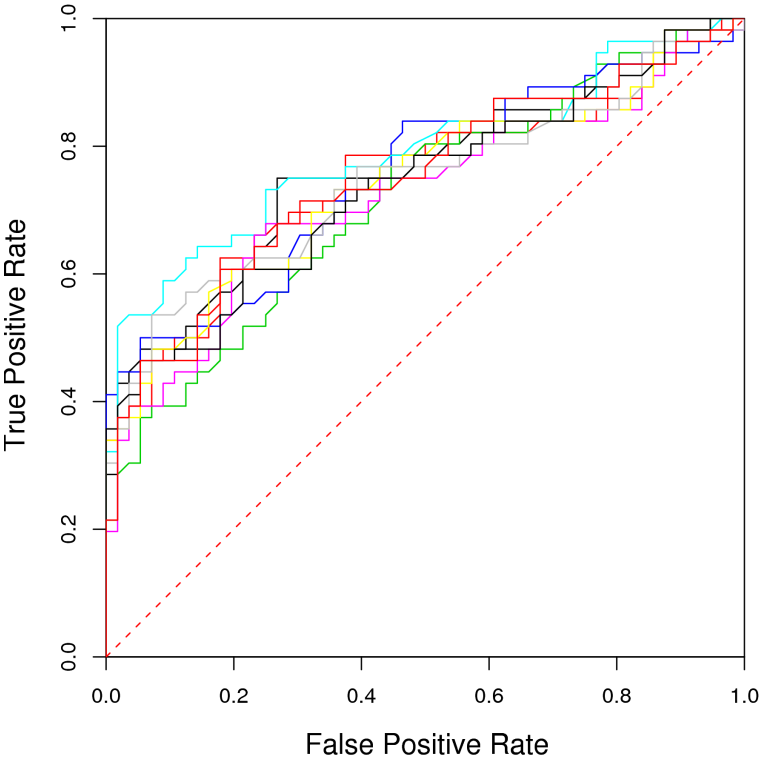


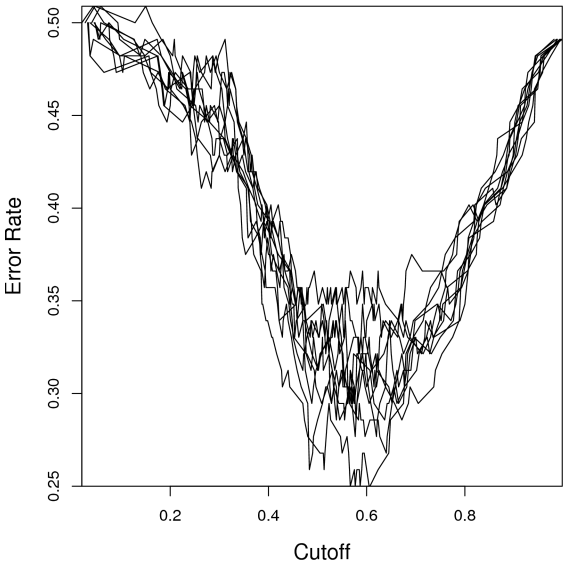
Scripps Cytotoxicity Models

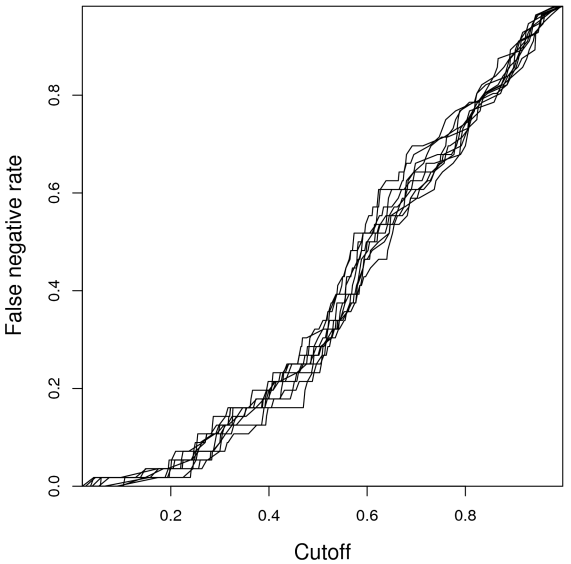
- % correct (ensemble average) = 69%
- % correct (consensus) = 71%

	Nontoxic	Toxic
Nontoxic	39	17
Toxic	15	41

- *Not very good performance*

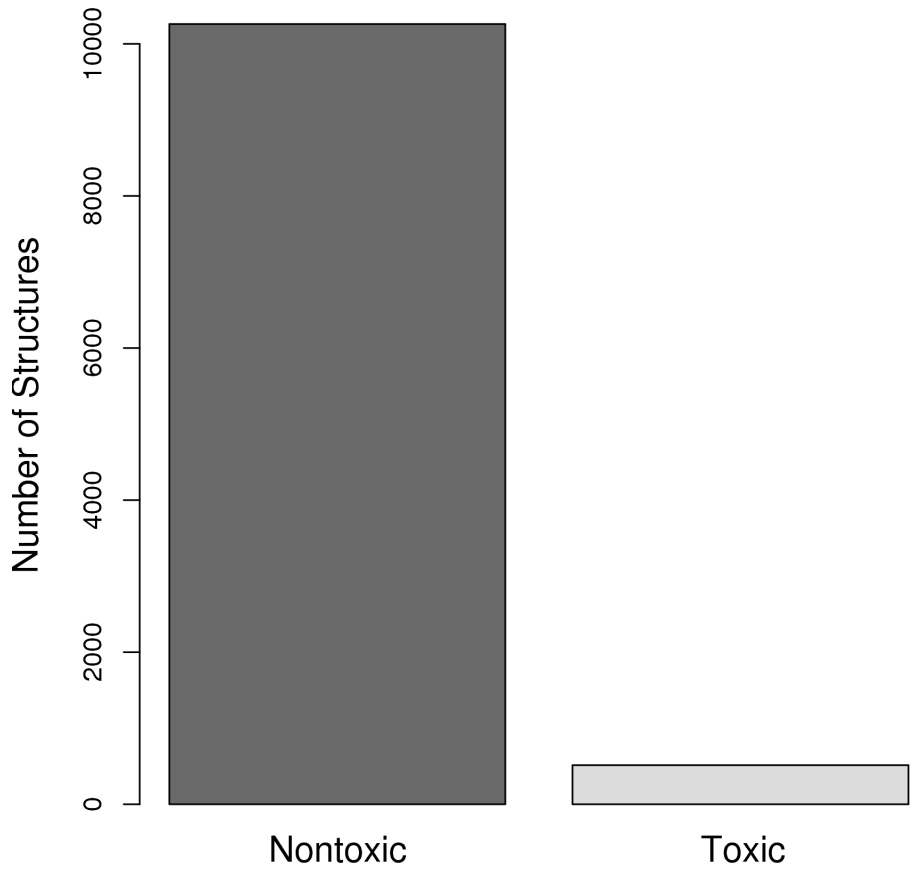






Do More Negatives Help?

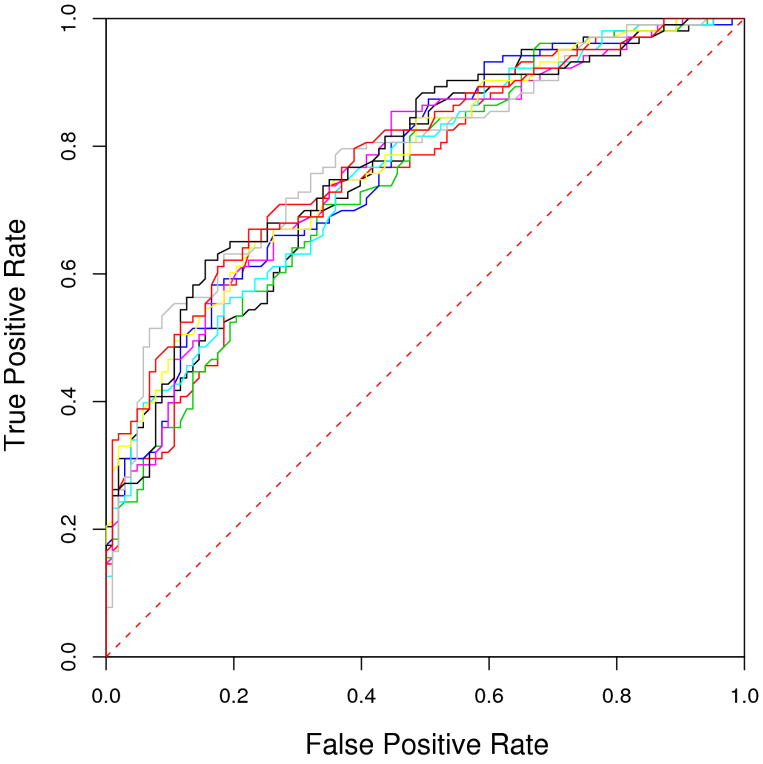
- Include 10,000 structures, randomly selected
  - Primary data, assumed to be nontoxic
- Selected a cutoff pIC50
  - >= 5.0 - toxic
  - < 5.0 – nontoxic
- Used sampling to create 10-member ensemble



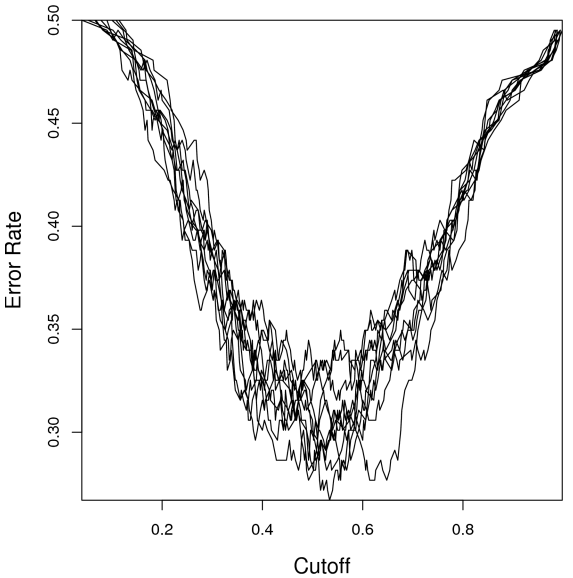
Expanded Cytotoxicity Dataset

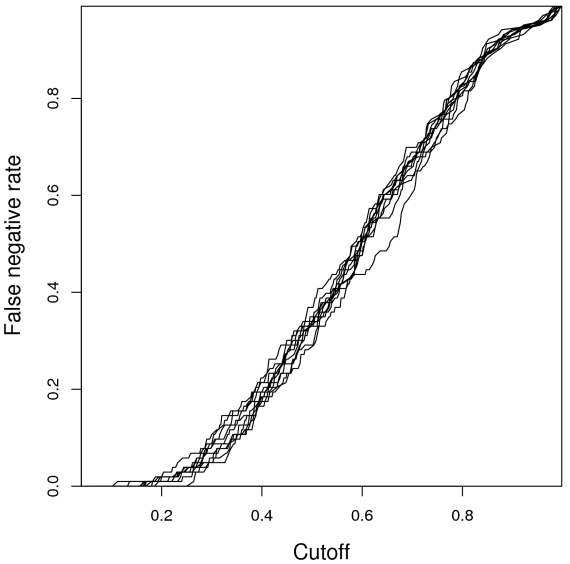
- % correct (averaged over the ensemble) = 69%
- % correct (consensus prediction) = 71%

	Nontoxic	Toxic
Nontoxic	79	24
Toxic	35	68



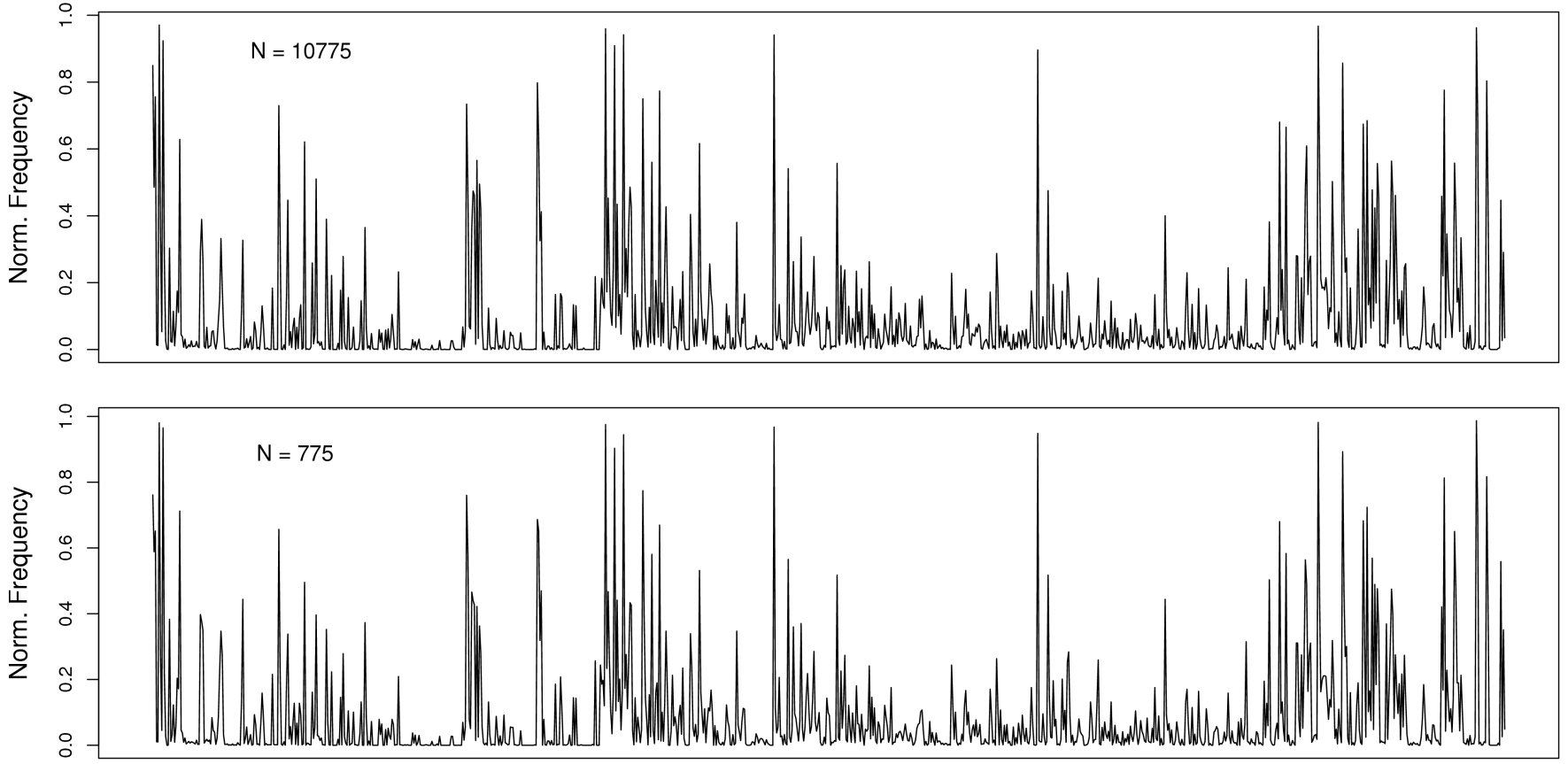
- *Not much improvement*
- *Insufficient sampling of the nontoxics*





We Need More Positives

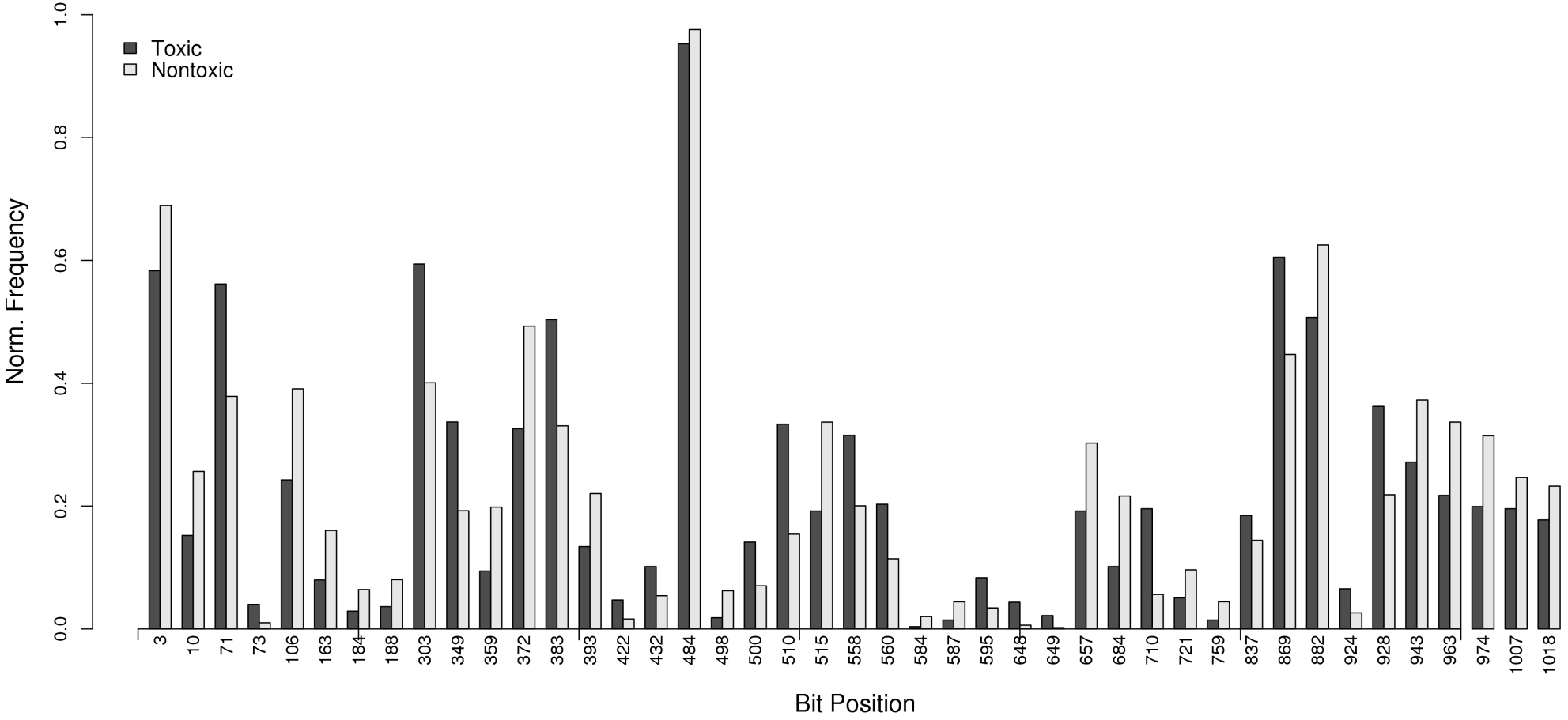
- The two datasets (775 vs 10,775 compounds) are quite similar in terms of *bit spectrum*
- *Normalized Manhattan distance = 0.016*



Important Structural Features

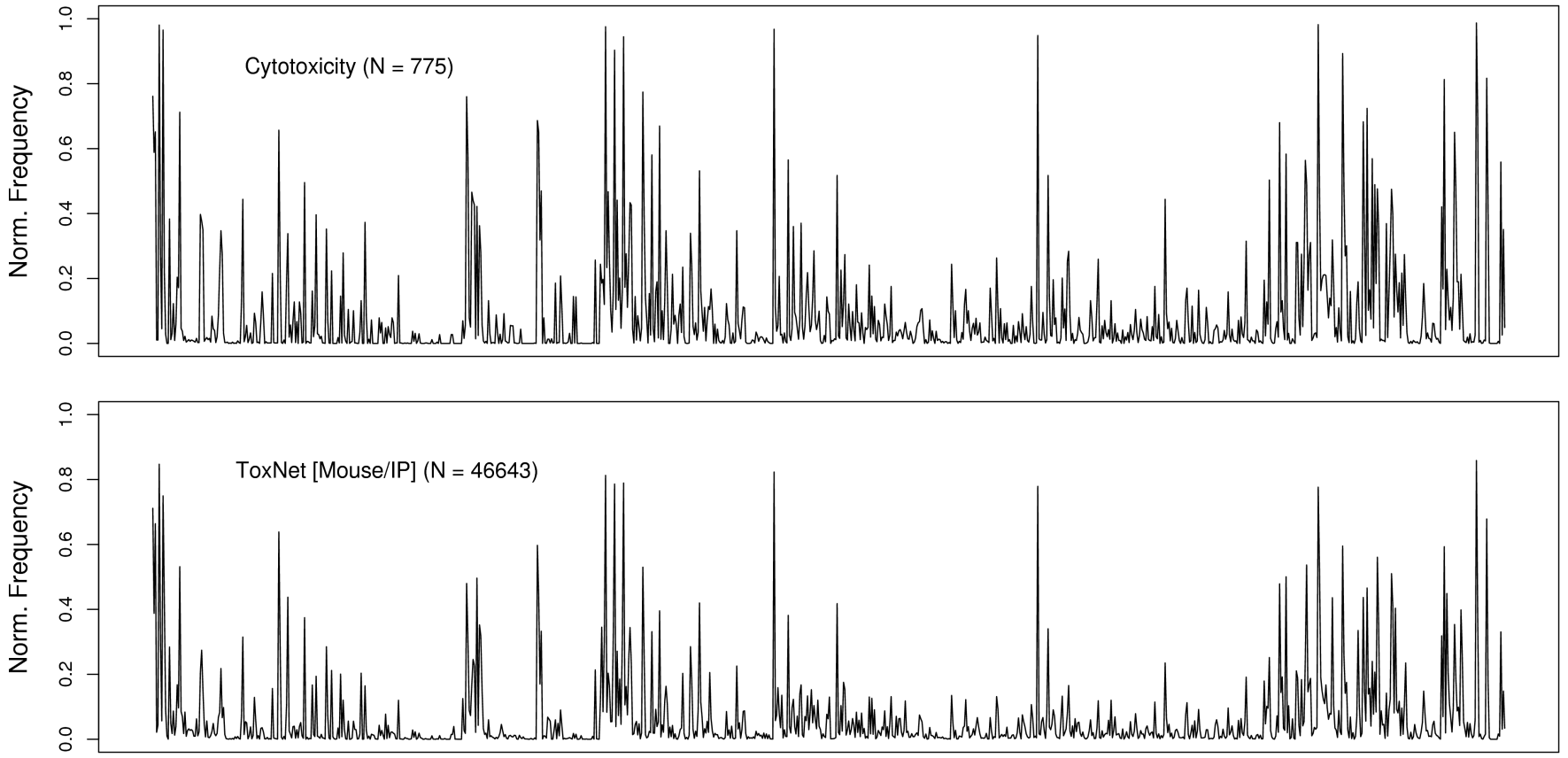
- The 10 most important features for predictive ability across the ensemble leads to 43 unique important bits
- This is a total of 66 structural features
  - The toxic compounds are characterized by having a *slightly* larger number of these features, on average

Important Structural Features



Predicting Animal Toxicity

- We should not 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

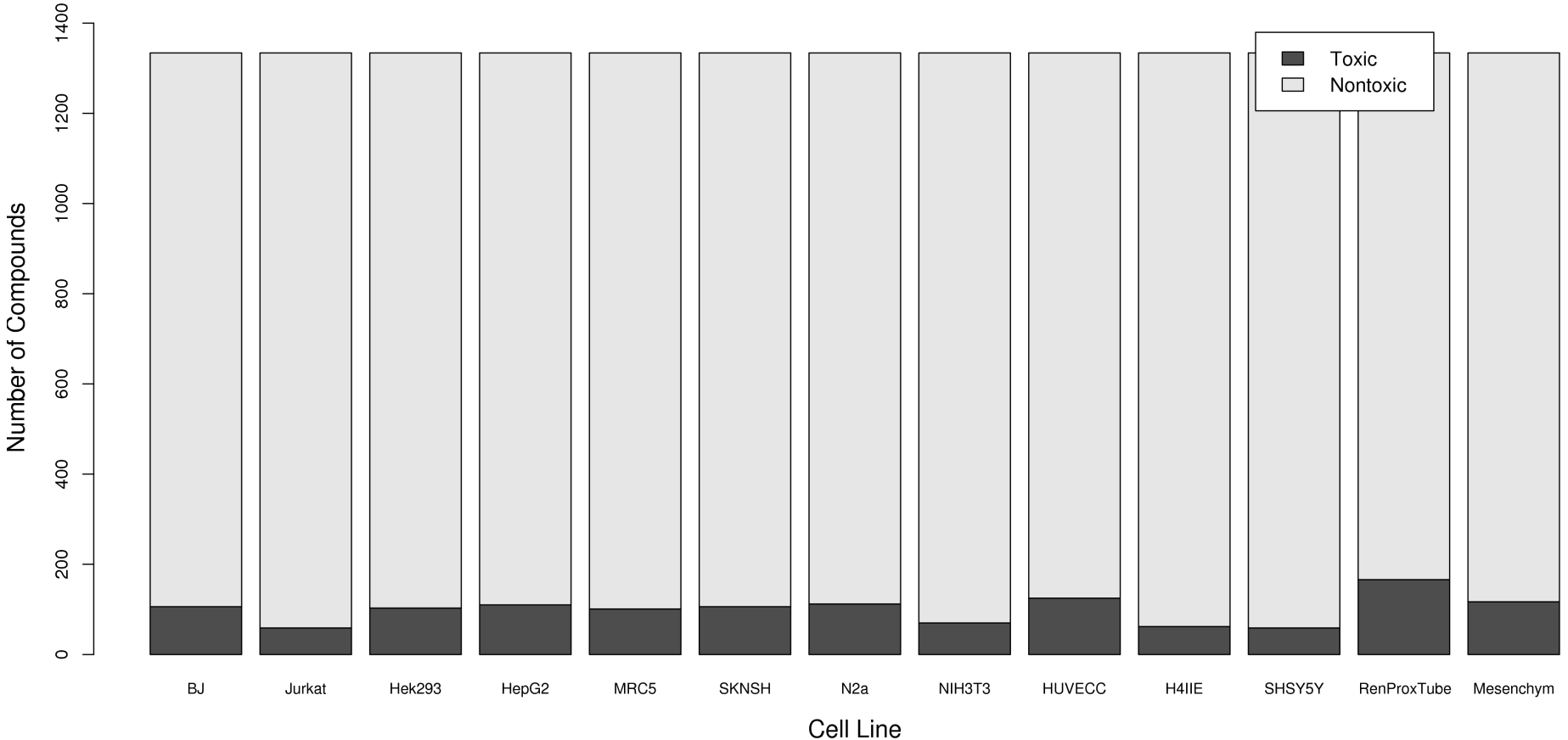
Cutoff = 0.4, 46% correct

NCGC Toxicity Dataset

- 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

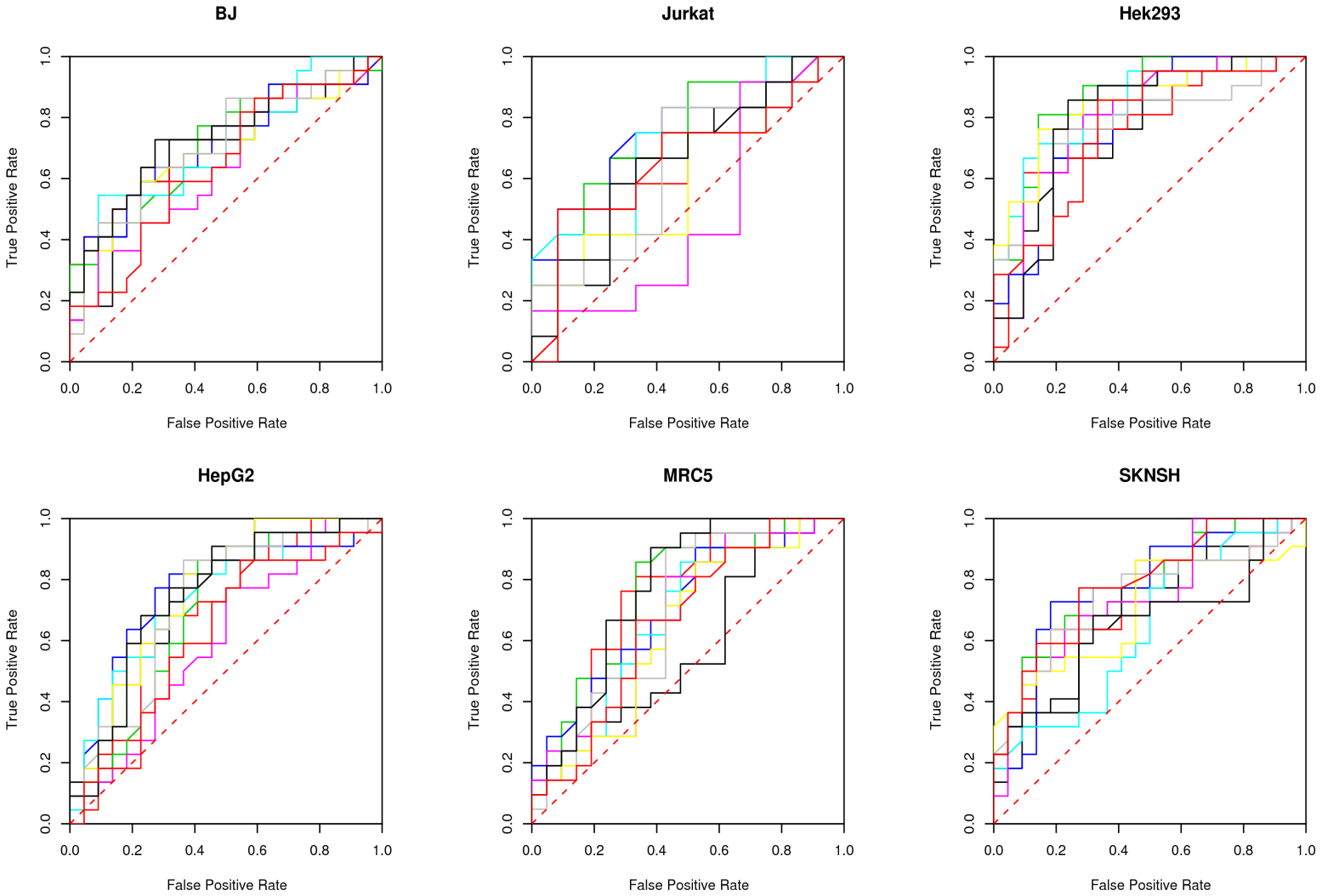
NCGC – Class Distributions

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



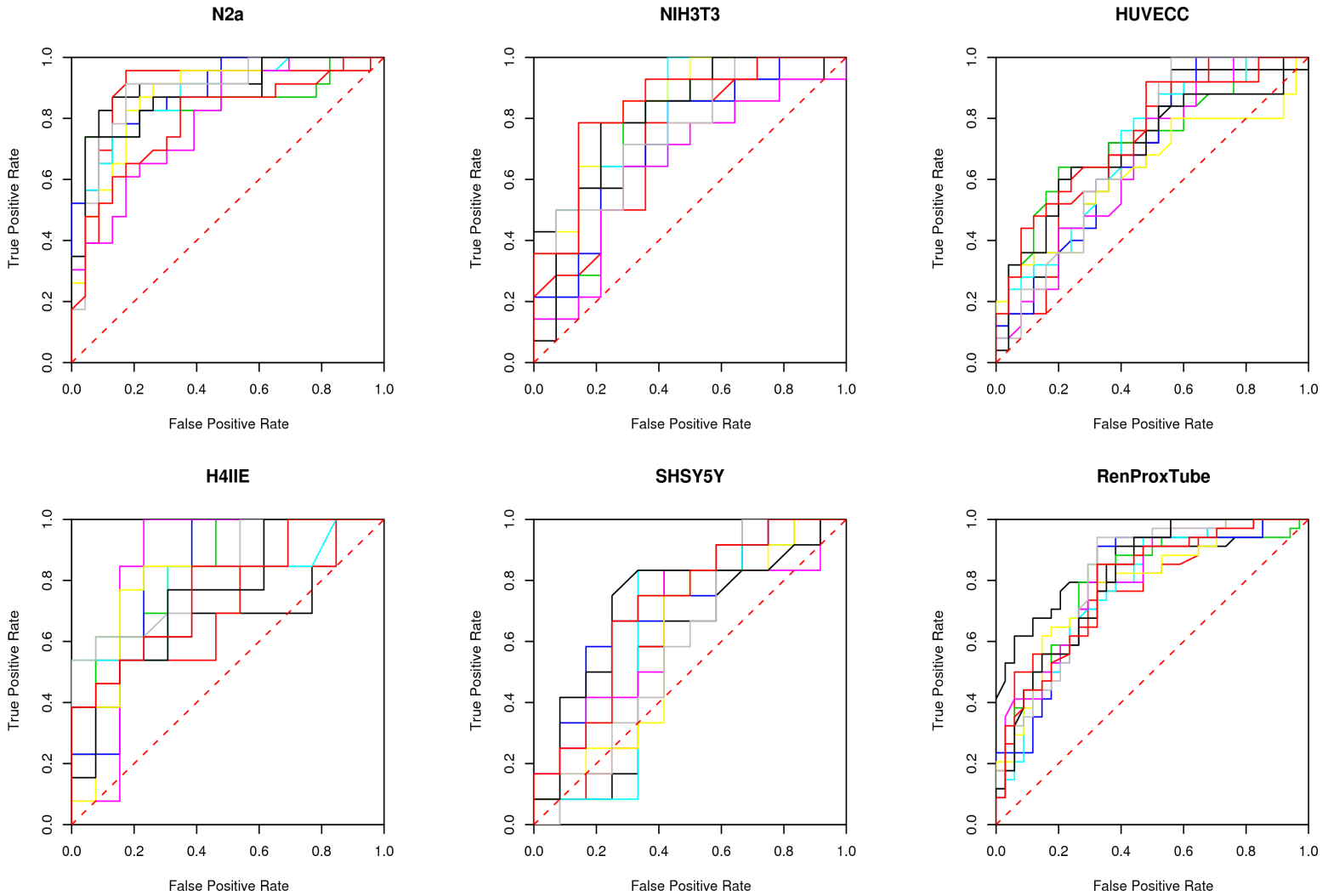
NCGC – ROC Curves

- Sampling the nontoxic class is an issue

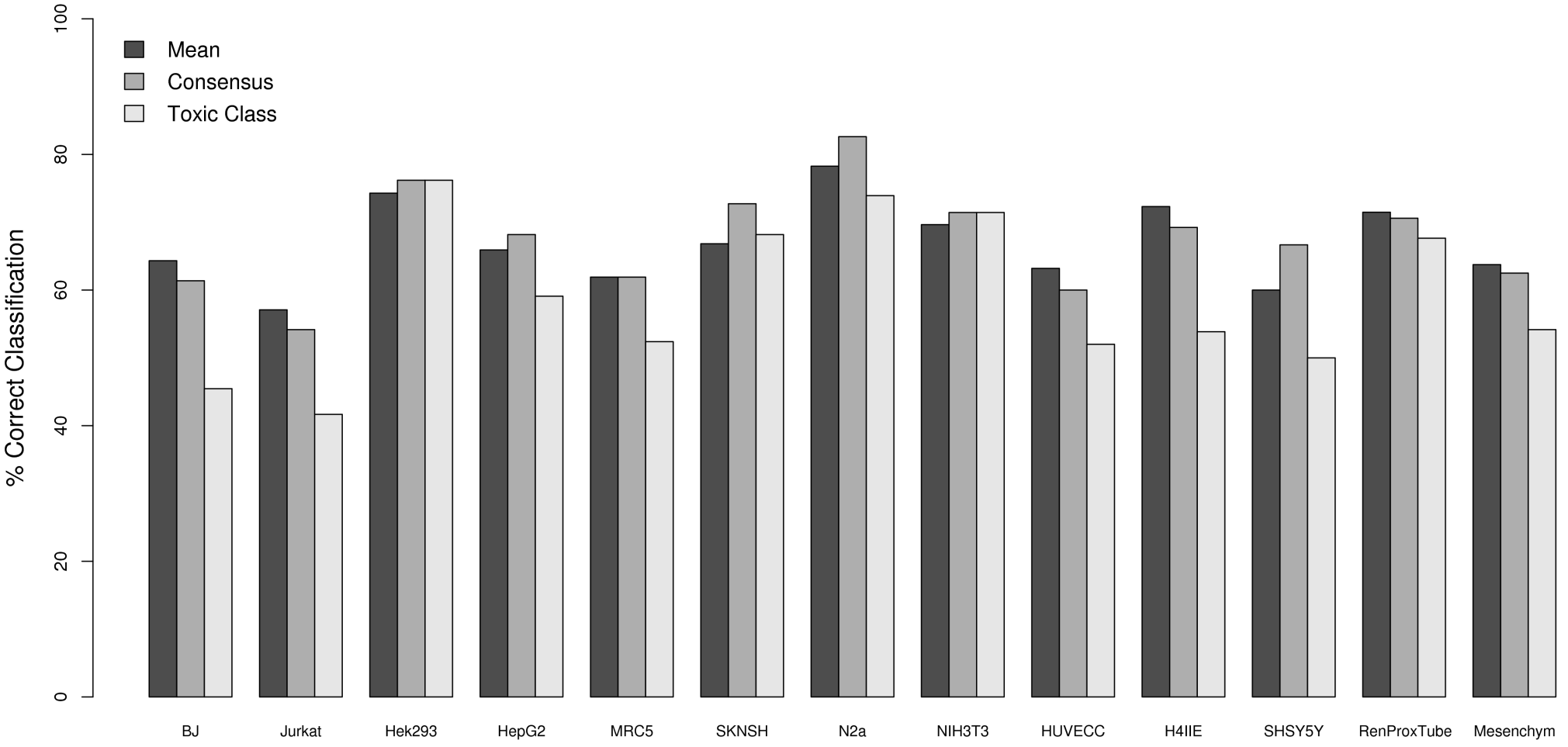


NCGC – ROC Curves

- Sampling the nontoxic class is an issue



NCGC – Model Performance (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

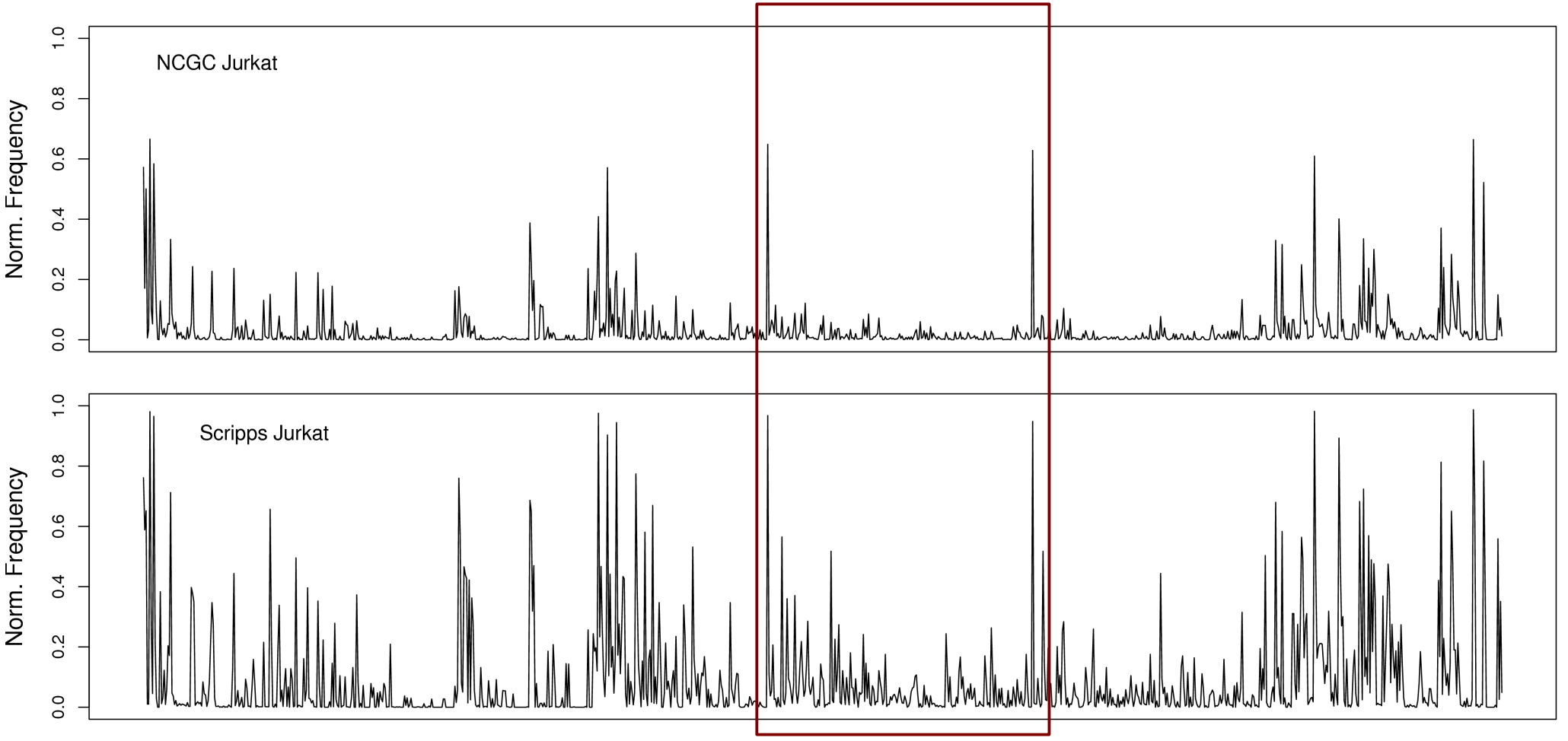
Predictions for the Scripps Cytotox dataset, using the original cutoffs (32% correct)

	Nontoxic	Toxic
Nontoxic	26	90
Toxic	109	550

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

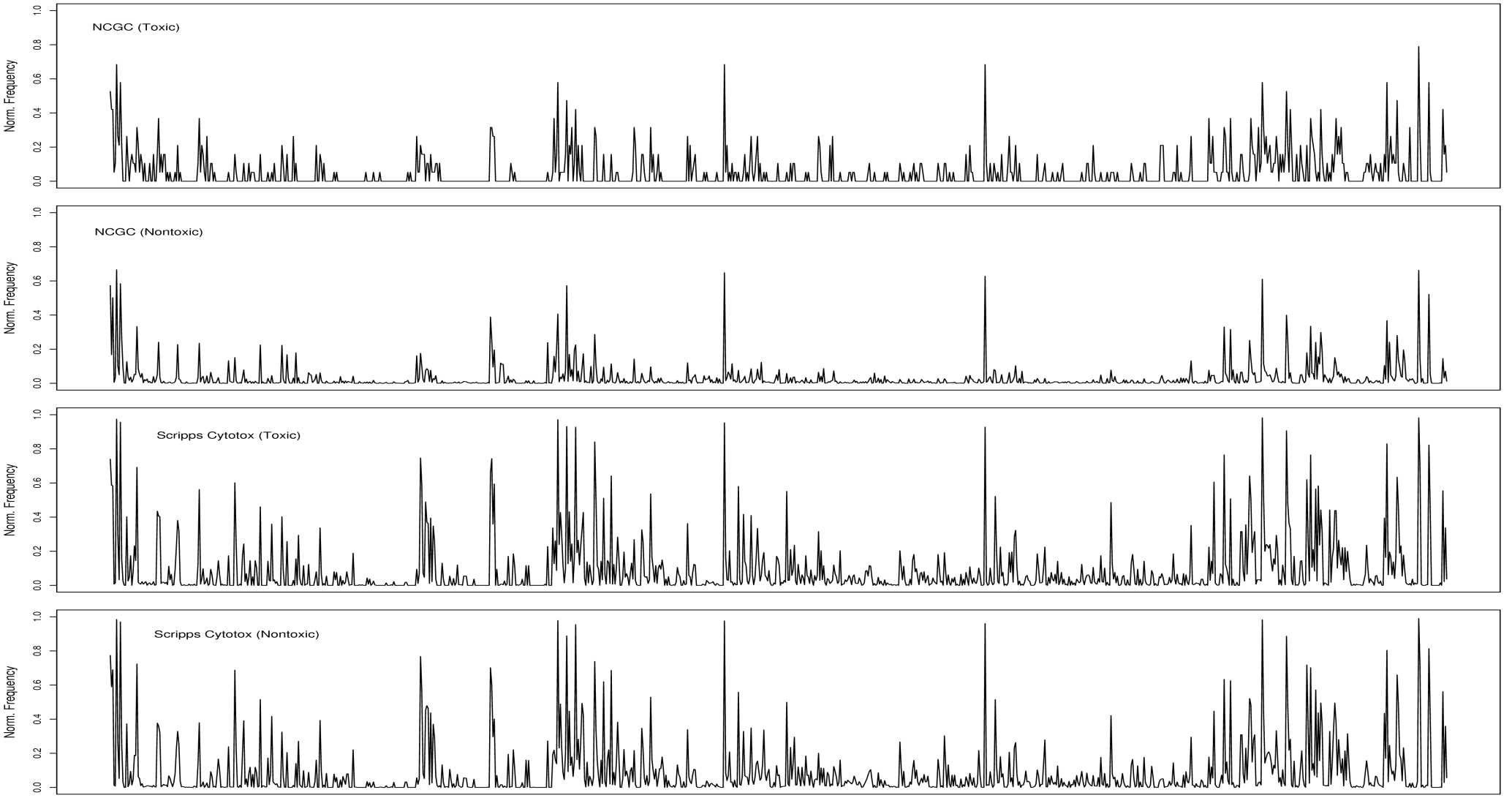
Comparing NCGC & Scripps Datasets

- Comparing the datasets as a whole



Comparing NCGC & Scripps Datasets

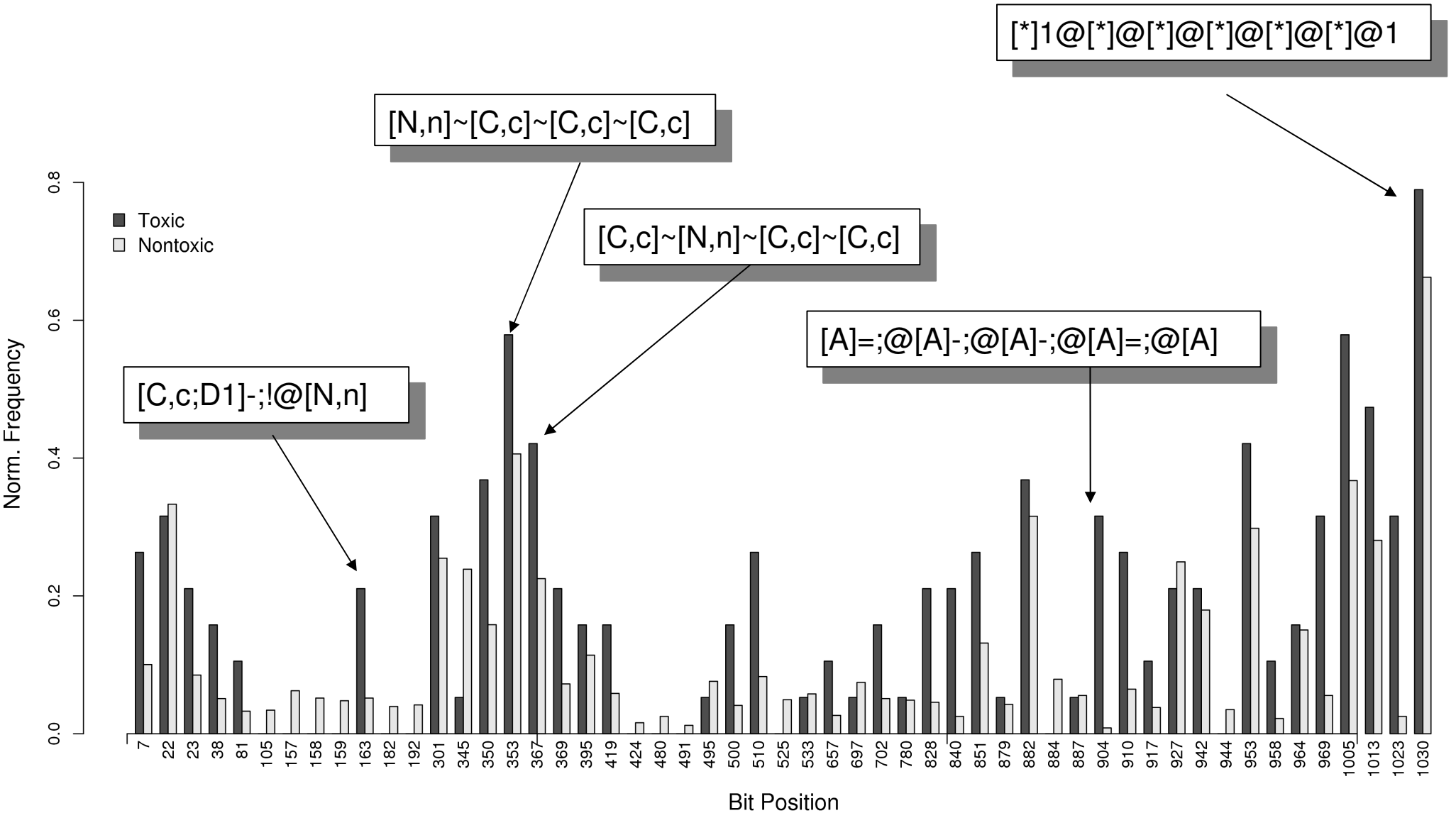
- Comparing datasets class-wise



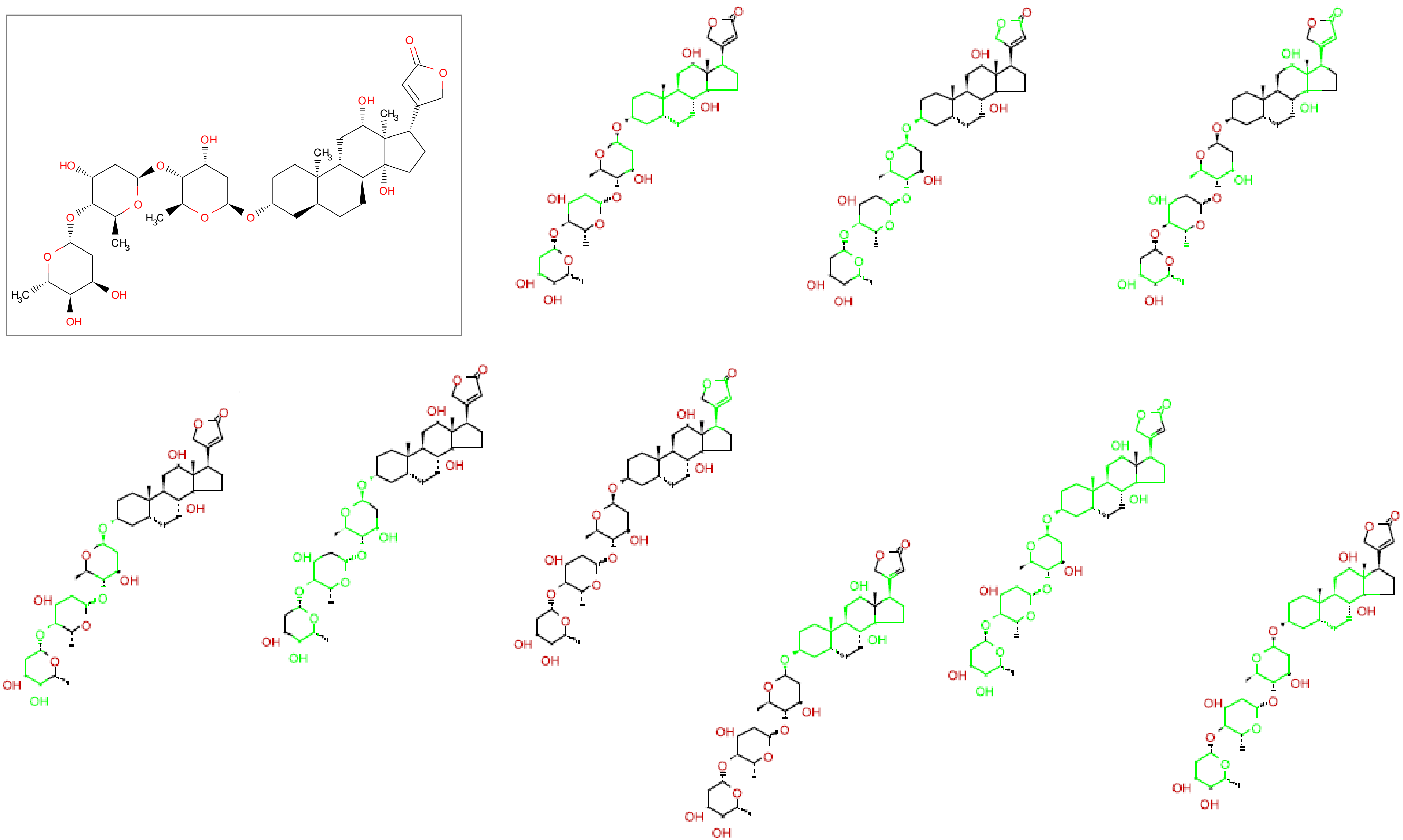
Important Features

- 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 a larger number of these features, on average

Important Features - Distributions



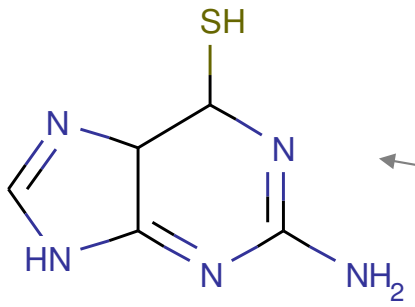
Feature matches for example structure



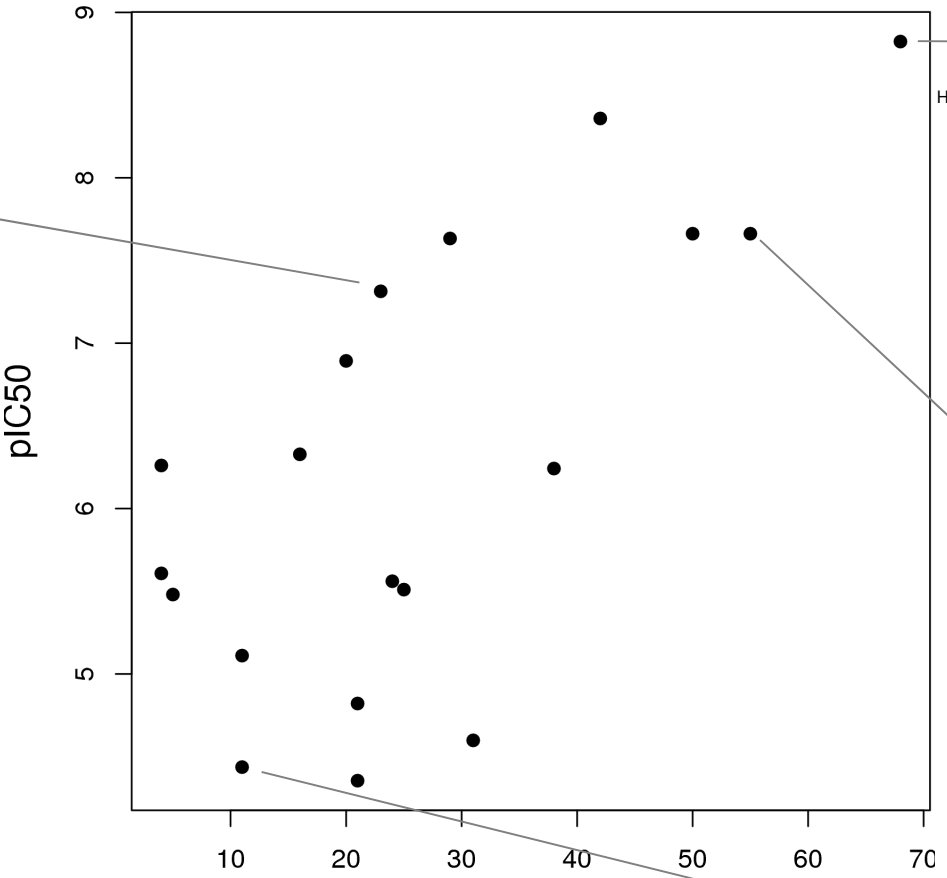
Important Features
Animal Toxicity vs. Cytotoxicity

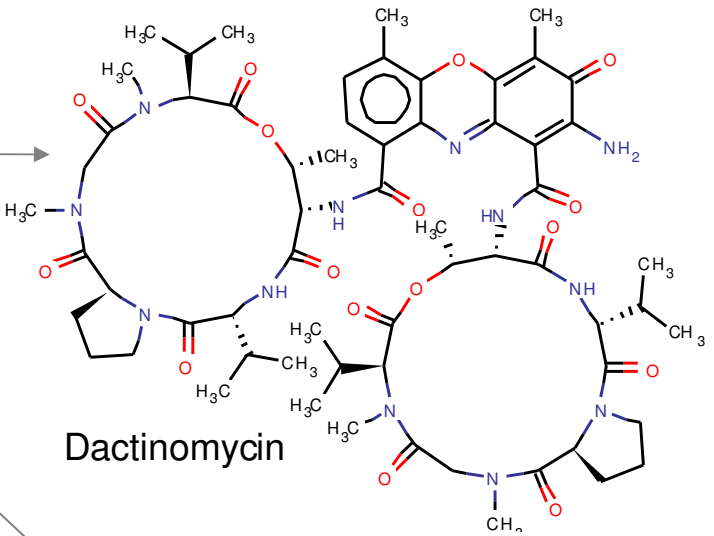
- The ToxNet (Mouse/IP) and NCGC Jurkat models have 130 *important features in common*
- *These features are more common in the NCGC toxic compounds than in the NCGC nontoxic compounds*
- *The average number of these features present in the NCGC dataset, overall, is 18.8*
  - *Very low, might indicate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

Toxicity vs No. of Features - NCGC Dataset

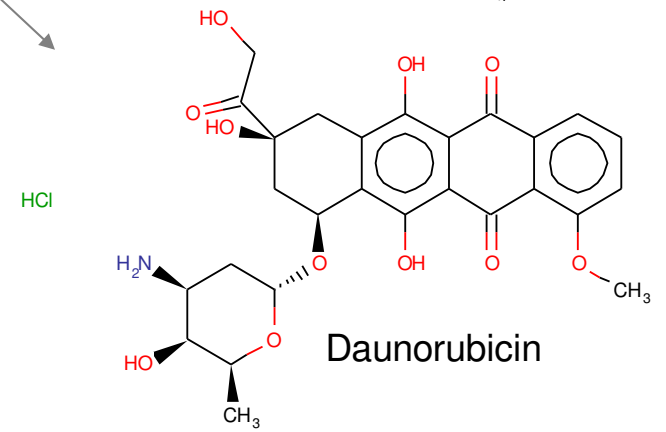


Leukerin deriv.

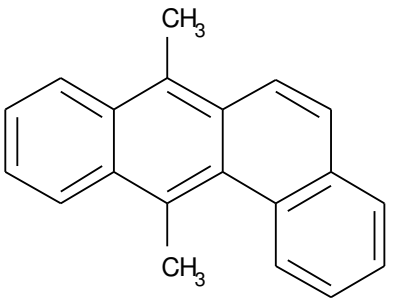




Dactinomycin

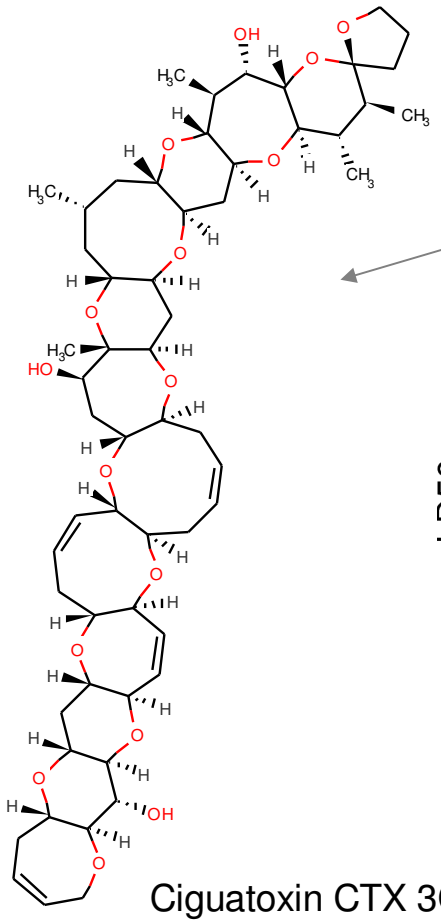


Daunorubicin

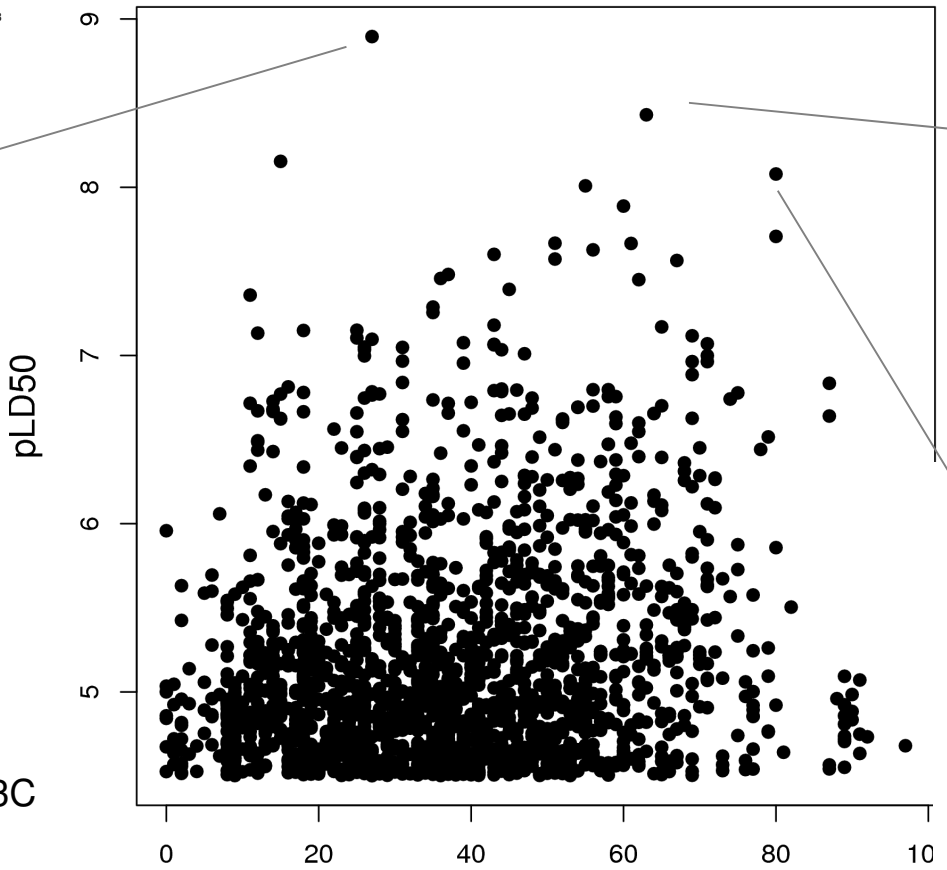


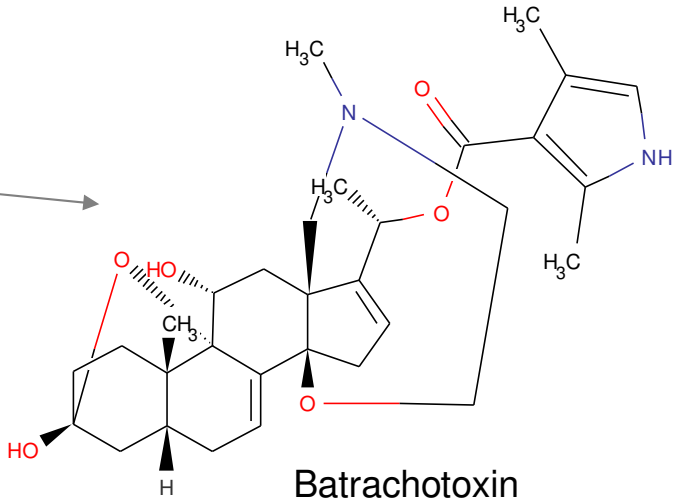
- The important features focus on the toxic class
- No correlation between number of features and pIC50 for nontoxic class

Toxicity vs No. of Features - Mouse/IP

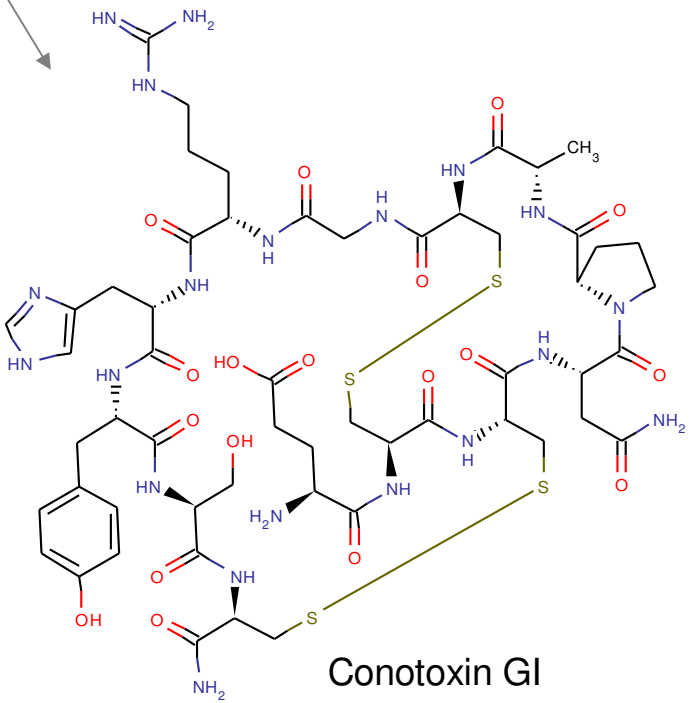


Ciguatoxin CTX 3C





Batrachotoxin



Conotoxin GI

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

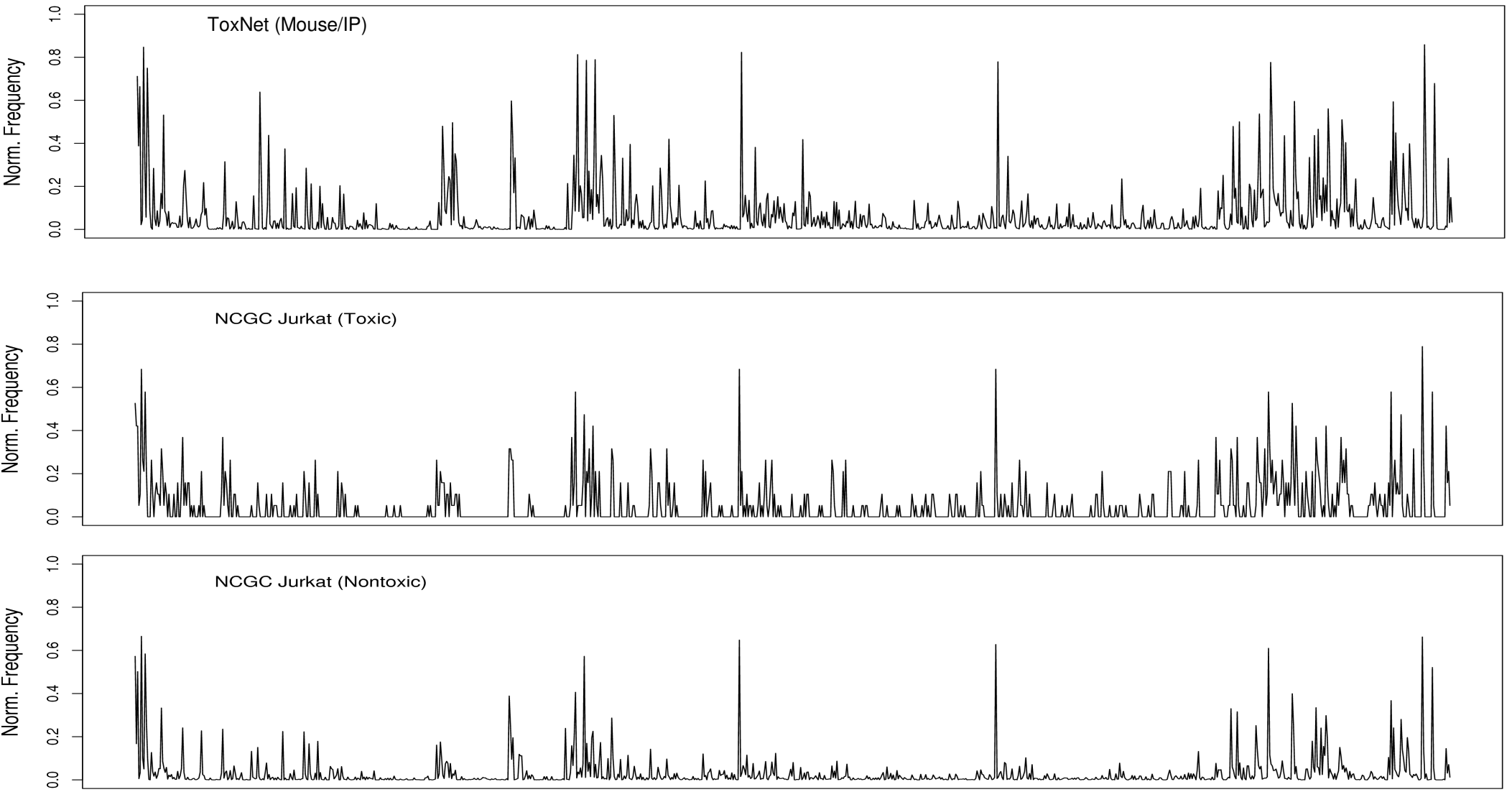
Predicting Animal Toxicity

	Nontoxic	Toxic
Nontoxic	12182	558
Toxic	32638	1265

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

- 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



Model Deployment

- 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

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 experiment
  - 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 data than access large standardized data sets

Standardization Issues - Models

- We've built lots of models and selected ones we think are good
- Why should other people take our word?
  - They shouldn't!
- Users should be able to easily benchmark models against a certain dataset(s)
  - Modelers should also do this, but users may have their own internal benchmark datasets

Standardization Issues - Models

- As a community can we standardize on datasets?
  - NTP
- We need to decide what data characteristics are we trying to test
  - Structural features, cell types, mechanisms
- Models must be easily accessible
  - Models should be downloadable
  - Alternative methods for access should also be provided

Whats Next?

- Further investigate differences in cell lines (cytotoxicity vs. animal toxicity)
- 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 count to prioritize compounds, which may be toxic
  - Improve assessment of model applicability

Summary

- Lots of data available for model development
  - Predictive ability ranges from poor to decent
- Applying models to predict other datasets is dependent on several factors
  - 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 best 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 of 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

An Example

- Identifying frequent hitters
  - Find CID's that are active in multiple assays
- Not (easily?) doable via PubChem
- Downloaded assay data
  - Extracted CID, AID, Activity Outcome and Activity Score
  - Loaded into PostgreSQL database
- Web page allows you to paste CID's/SID's and get a list of the assays they are active in

An Example

- Activity scores are not very rigorous
  - So use what was measured
- But different assays use different column names
  - Difficult to automatically extract IC50 etc.
- There is no specific update schedule for bioassay data
  - Our compound mirror is updated monthly
  - PubChem assay data is not
  - Can't fully sync our data

NCGC Jurkat - Important Features Distributions

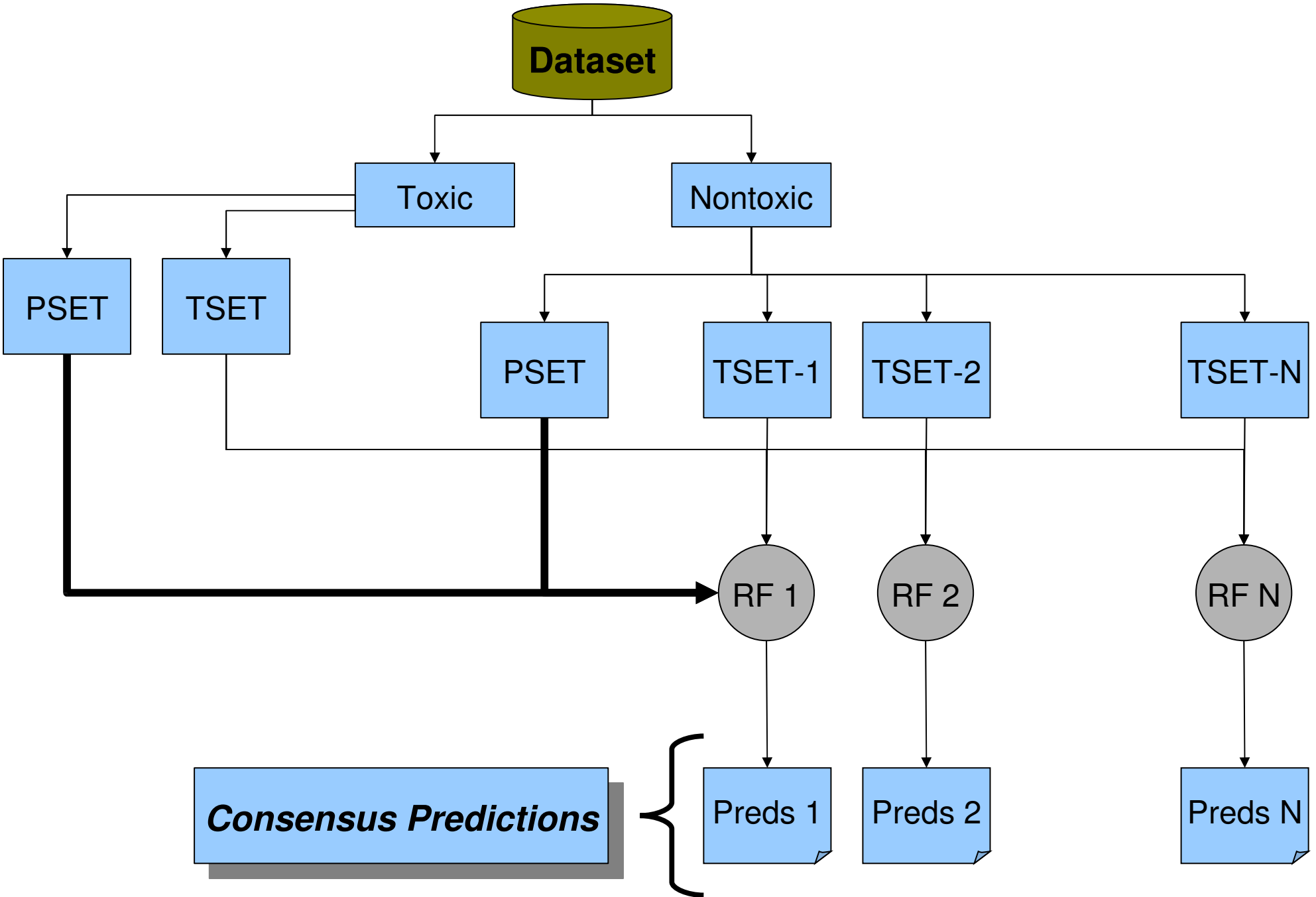
- Top 10 features

	Median	Mean
Toxic	9	9.8
Nontoxic	5	6.2

- Top 100 features

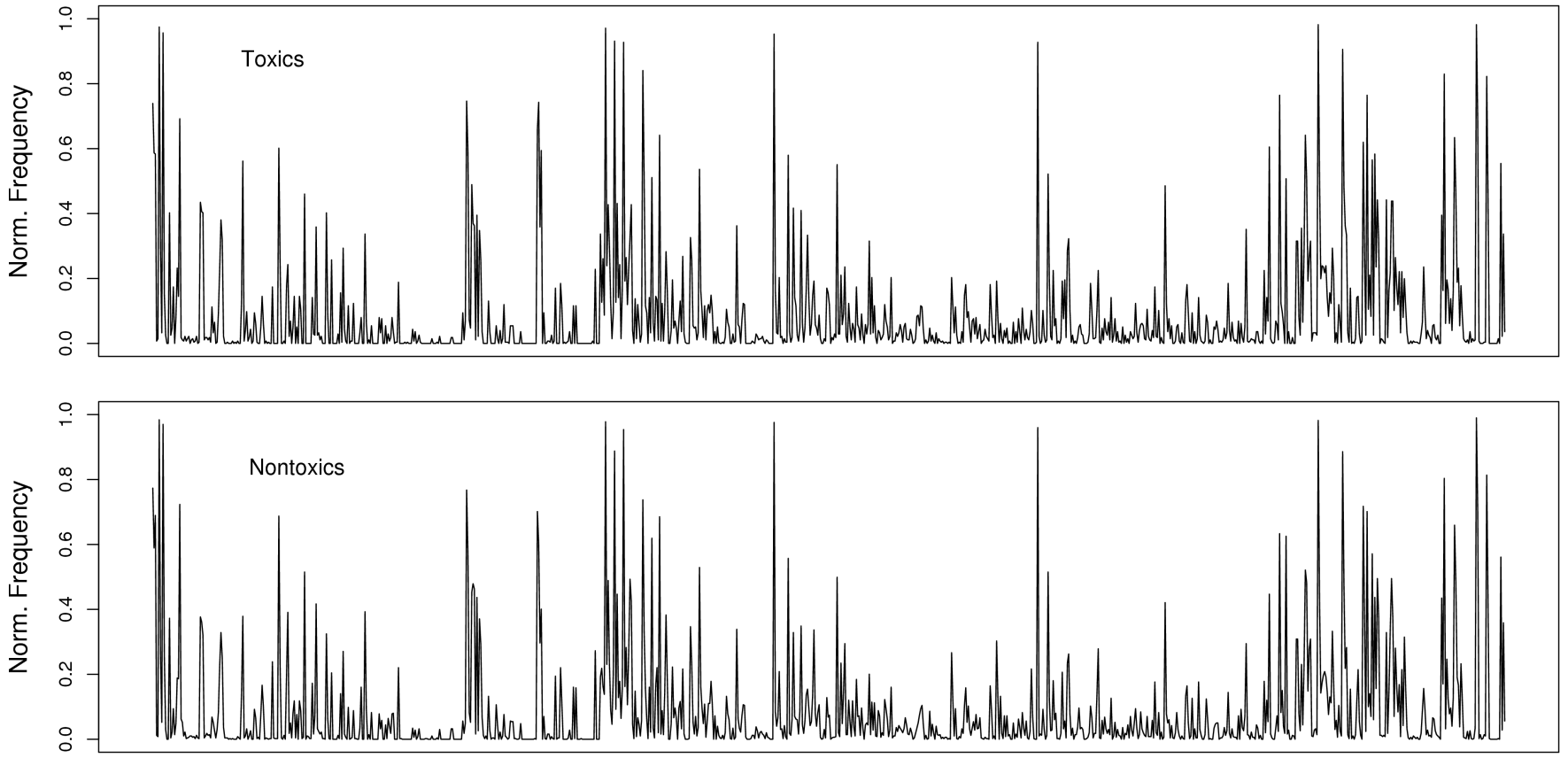
	Median	Mean
Toxic	32	41.53
Nontoxic	21	24.63

Handling Imbalanced Classes



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*