

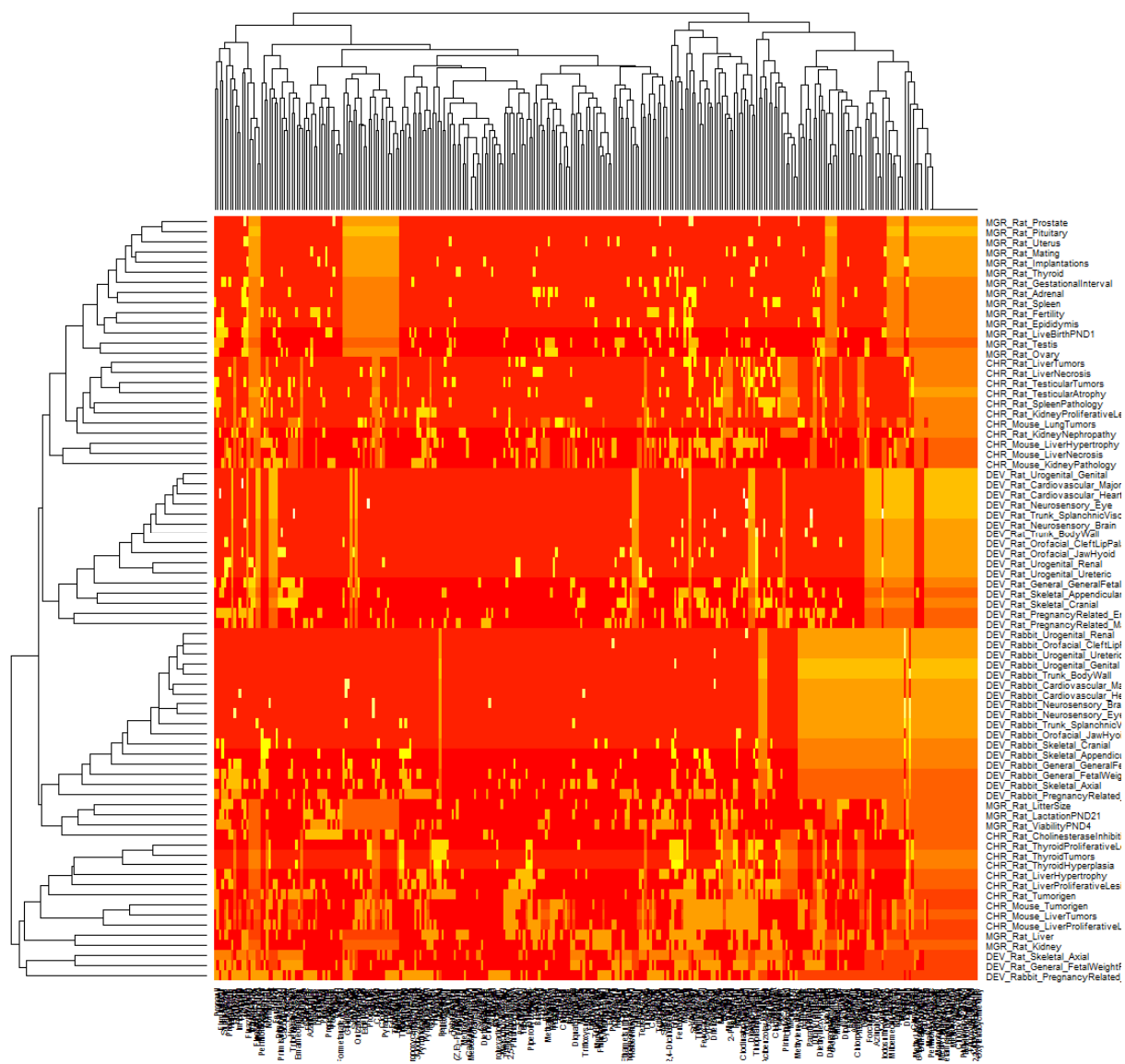
Prediction of *in vivo* toxicity endpoints from ToxCast™ Phase I data using a variety of machine learning approaches

Fred A. Wright^{1,2}, Zhen Li^{1,2}, Hanwen Huang¹, Arpita Ghosh^{1,2}, Wenjun Bao³, Lili Li³, Tzu-Ming Chu³, Russ Wolfinger³, Wei Sun^{1,2}, Fei Zou^{1,2}, Ivan Rusyn^{2,4}

¹Department of Biostatistics, ²The Carolina Centers for Environmental Bioinformatics and Computational Toxicology, University of North Carolina, Chapel Hill, NC, ³SAS Institute, Cary, NC, ⁴Department of Environmental Sciences and Engineering, University of North Carolina.



- The ToxCast Phase I data provides an important testing ground for the development of chemical toxicity signatures based on *in vitro* data and chemical descriptors.
- It is also an interesting testing ground for various data mining/prediction procedures
- The methods described here do not use prior information or beliefs about the predictive ability of certain assays or the amenability of certain endpoints to prediction.
- Initial discussions and analyses focused on endpoints CHR liver tumors and tumorigenicity, for both rat and mouse (4 endpoints), using dichotomized ToxCast assays and endpoints (active/inactive)
- There is considerable structure in the data



Plots of the ToxCast assays are similarly highly structured, although some assays have only a small proportion of actives

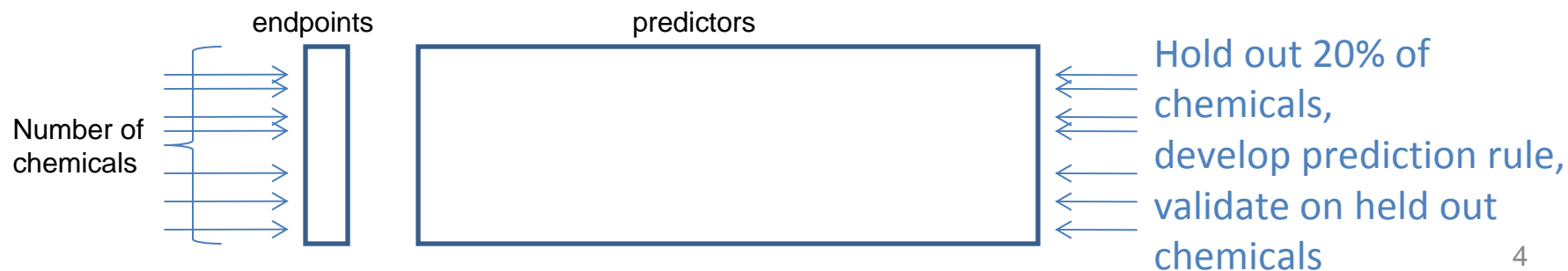
- Several groups/people split off to do independent analyses, deciding to use the following common means of comparison:
 - **Performance** measured using **area under the ROC curve** for each endpoint



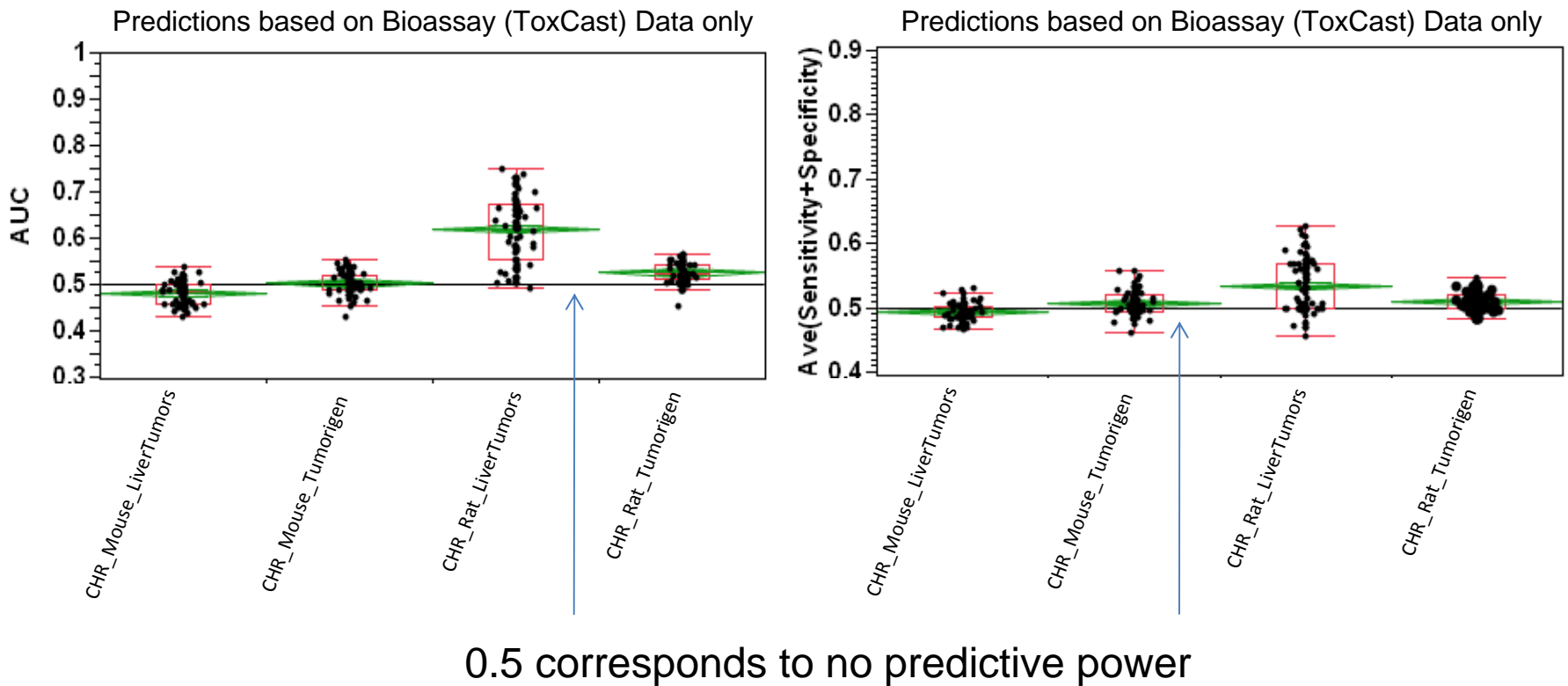
<http://www.anaesthetist.com/mnm/stats/roc/Findex.htm>

as well as the **average of sensitivity and specificity** for a thresholding rule

- **Cross-validation** performed using 5-fold cross-validation, with a large number of independent splits (if possible)



- The SAS investigators ran a large number of models (84, described below), essentially covering all the models also run by UNC
- Thus one might think of the space of models attempted as a kind of population
- The results were not very promising



- In retrospect, perhaps this should not be surprising
- For instance, in Fisher's exact tests of each endpoint vs. each ToxCast predictor, only 4% of p-values are less than 0.05
- We decided to move forward in two directions:
 - 1) Examine more endpoints
 - 2) Include 1224 Dragon chemical descriptors (as predictors of toxicity endpoints), kindly provided by Drs. Hao Zhu and Alex Tropsha of UNC.

For this dataset compared against all in vivo endpoints, 17% of the p-values are less than 0.05.

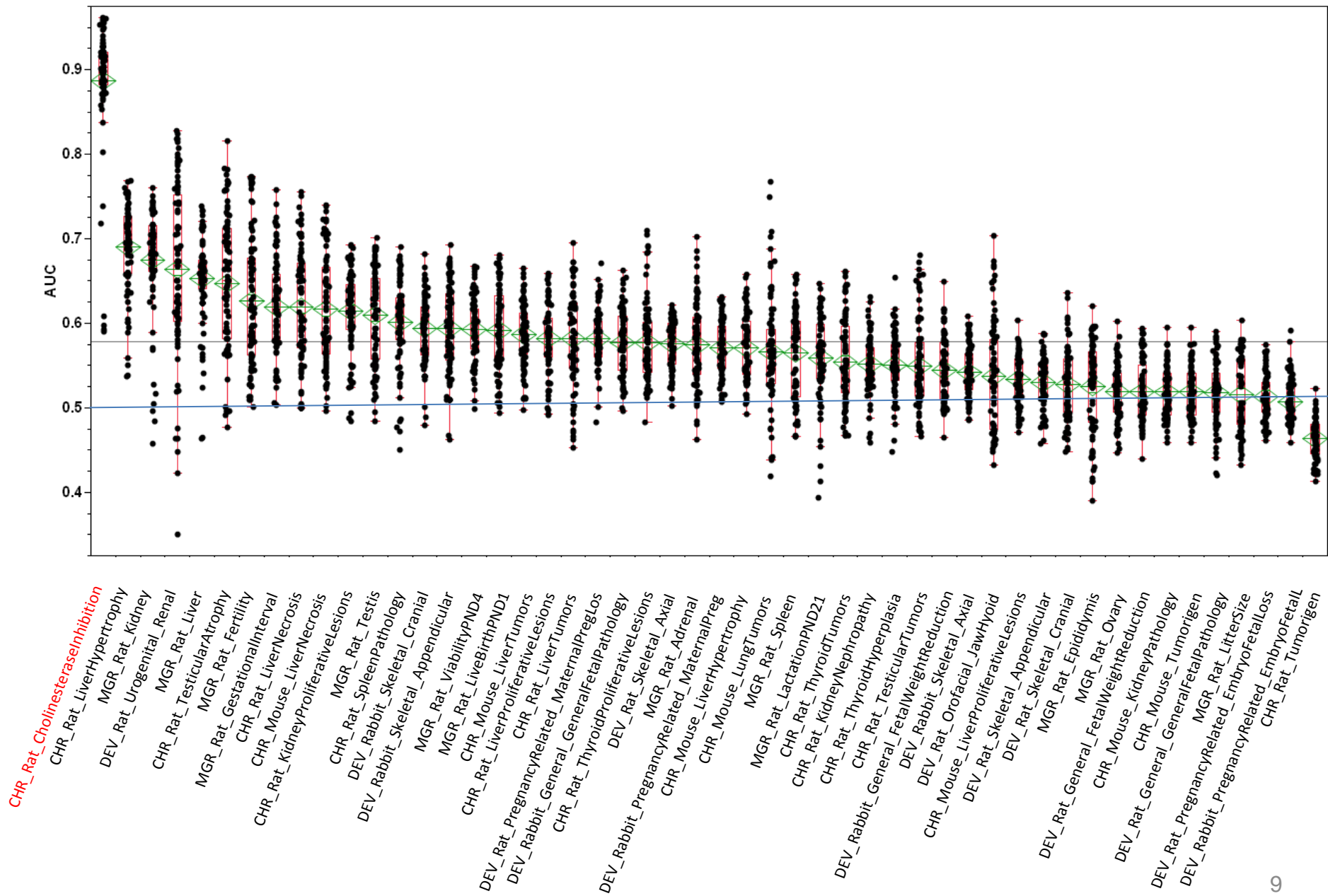
SAS Toxcast Data Prediction

- Fifty Endpoints were selected based on the rank of their frequencies
- Total of 1899 predictors from various tables
- Eight classification methods for prediction, 84 total models
 - Discriminant analysis (DA)
 - Distance Scoring (DS)
 - K-Nearest Neighbor (KNN)
 - General linear model selection (GLM)
 - Logistic regression (LR)
 - Partial least square (PLS)
 - Partition tree (PT)
 - Radial basis machine (RBM)
- 5-fold cross-validation, 10 iterations, performance metrics:
 - AUC
 - Accuracy
 - RMSE

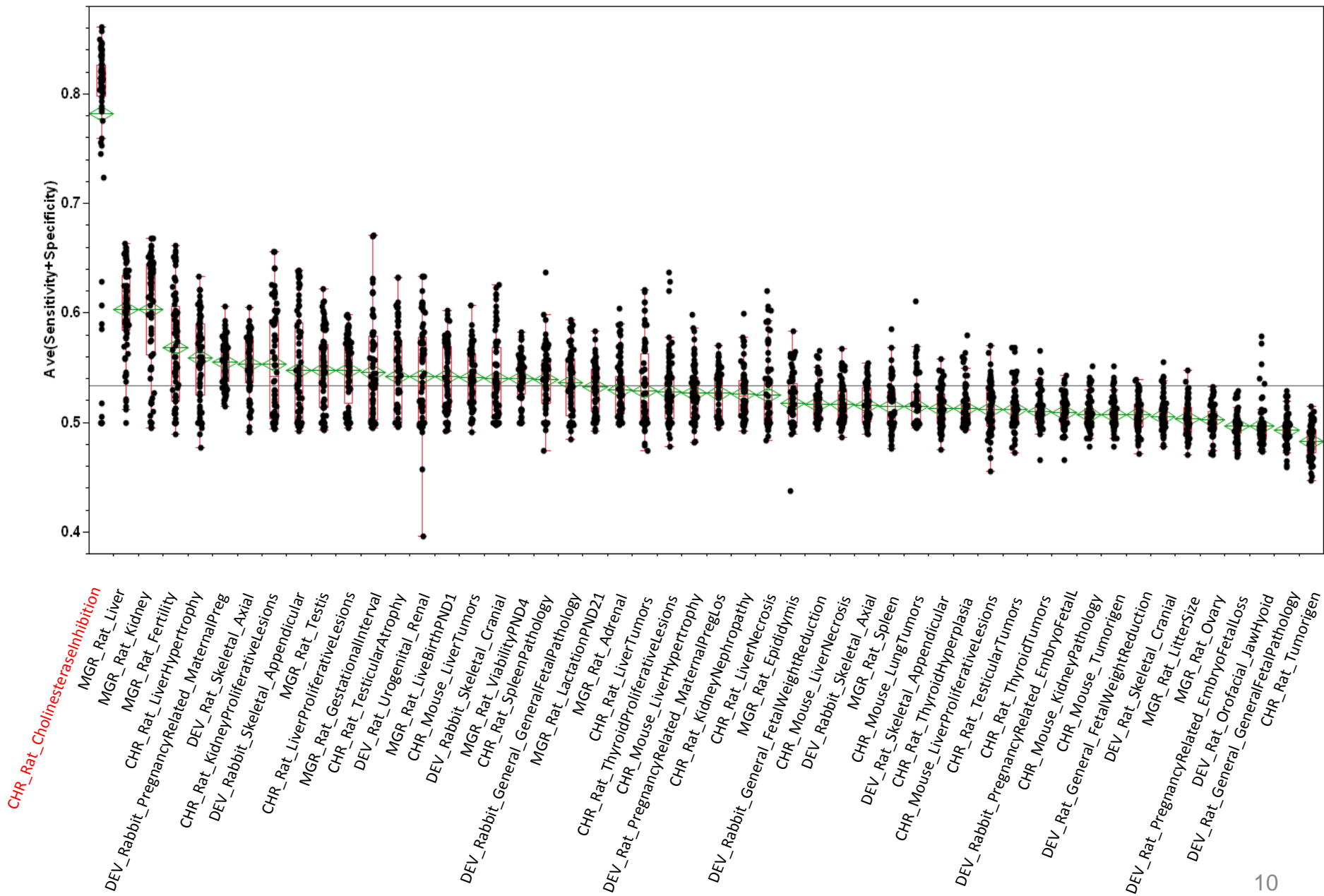
50 endpoints selected based on the rank of the frequency of “actives”

Endpoint	Frequency
DEV_Rat_Skeletal_Axial	111
DEV_Rabbit_PregnancyRelated_Mate	109
MGR_Rat_Liver	104
CHR_Rat_Tumorigen	97
CHR_Mouse_LiverProliferativeLesi	93
CHR_Mouse_Tumorigen	92
DEV_Rat_General_FetalWeightReduc	87
MGR_Rat_Kidney	74
CHR_Mouse_LiverTumors	72
DEV_Rabbit_PregnancyRelated_Embr	70
MGR_Rat_ViabilityPND4	68
CHR_Mouse_LiverHypertrophy	66
CHR_Rat_LiverHypertrophy	65
CHR_Rat_LiverProliferativeLesion	65
DEV_Rabbit_Skeletal_Axial	55
DEV_Rat_PregnancyRelated_EmbryoF	55
DEV_Rabbit_General_FetalWeightRe	49
DEV_Rat_PregnancyRelated_Materna	49
DEV_Rat_Skeletal_Appendicular	47
CHR_Mouse_KidneyPathology	45
CHR_Rat_CholinesteraseInhibition	45
MGR_Rat_LitterSize	43
.....
DEV_Rat_Orofacial_JawHyoid	12

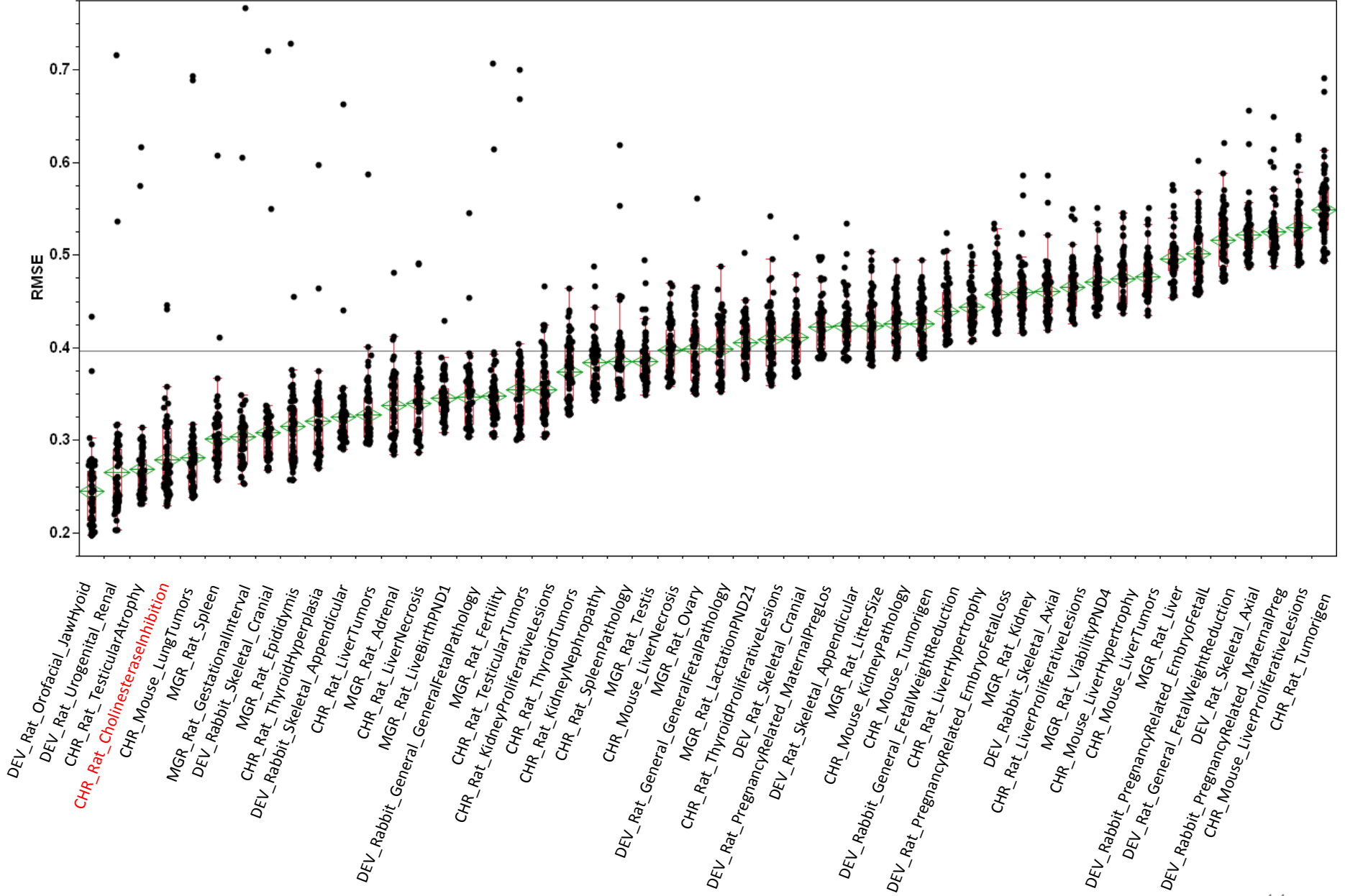
AUC: 84 models



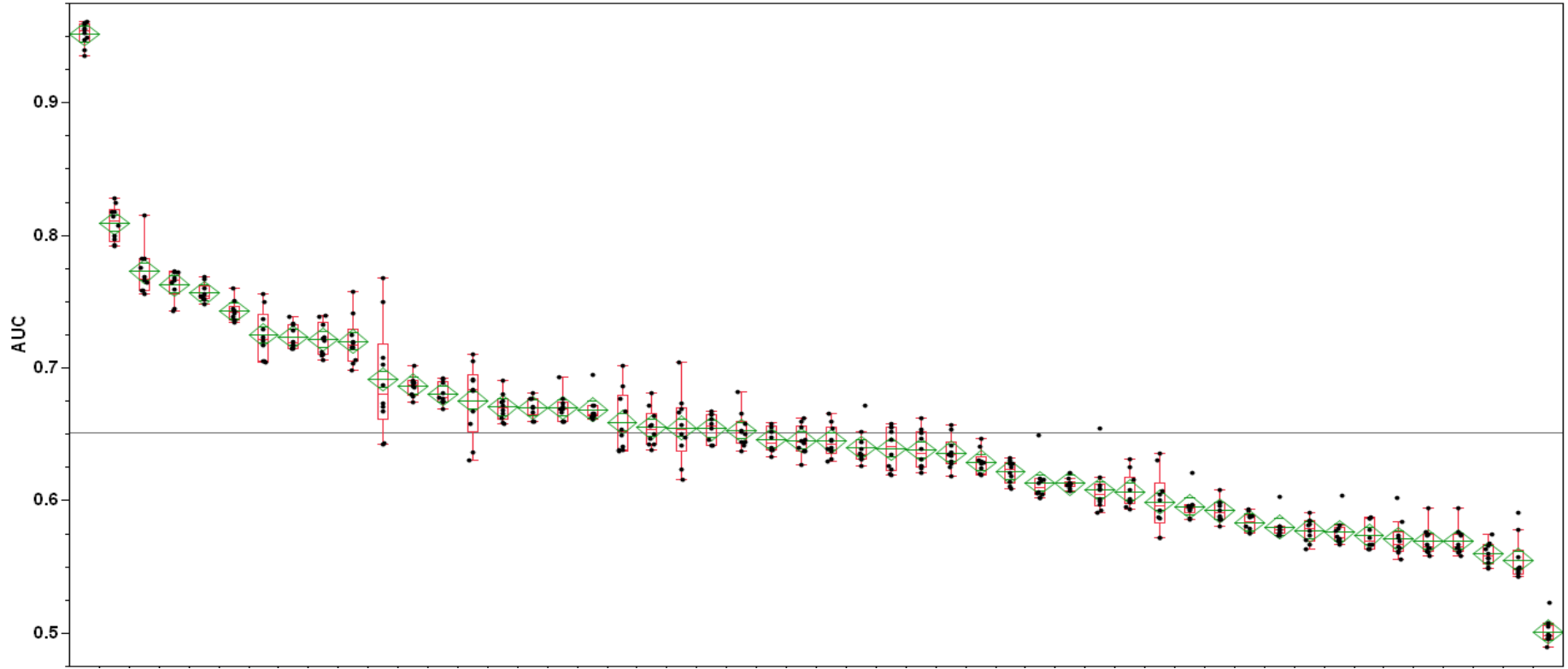
Ave (Sensitivity + Specificity): 84 models



RMSE: 84 models



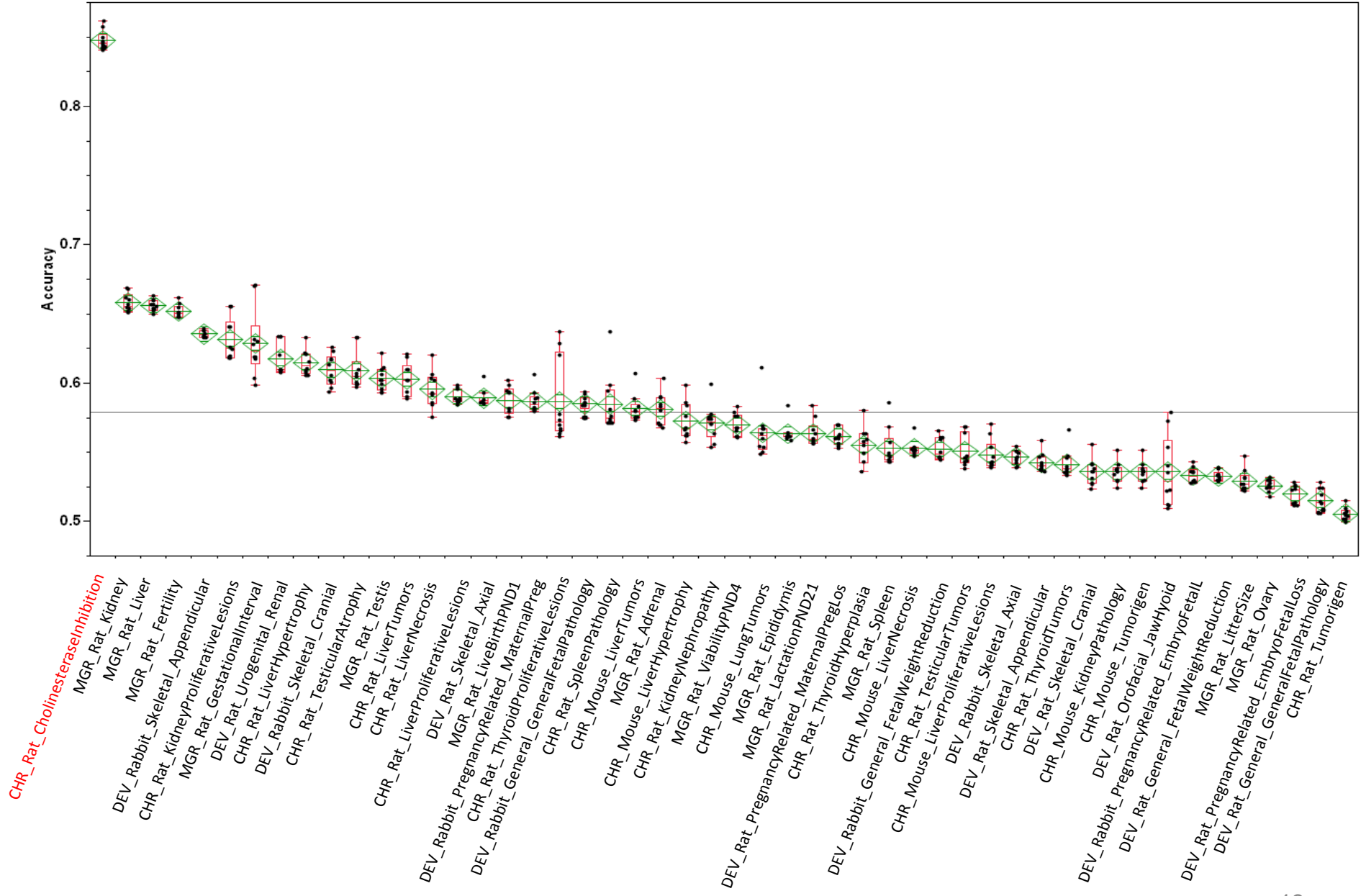
AUC: Top 10 models



- CHR_Rat_CholinesteraseInhibition
- DEV_Rat_Urogenital_Renal
- CHR_Rat_TesticularAtrophy
- MGR_Rat_Fertility
- CHR_Rat_LiverHypertrophy
- MGR_Rat_Kidney
- CHR_Rat_LiverNecrosis
- MGR_Rat_Liver
- CHR_Mouse_GestationalInterval
- MGR_Mouse_LungTumors
- CHR_Rat_KidneyProliferativeLesions
- CHR_Rat_ThyroidProliferativeLesions
- CHR_Rat_SpleenPathology
- MGR_Rat_LiveBirthPND1
- CHR_Rat_Skeletal_Appendicular
- MGR_Rat_LiverTumors
- CHR_Rat_Adrenal
- DEV_Rat_TesticularTumors
- MGR_Rat_Orofacial_JawHyoid
- DEV_Rat_ViabilityPND4
- CHR_Rat_LiverSkeletal_Cranial
- CHR_Rat_ThyroidLesions
- CHR_Mouse_ThyroidTumors
- DEV_Rat_PregnancyRelated_MaternalPregLos
- MGR_Rat_Spleen
- CHR_Mouse_LiverPathology
- MGR_Rat_LiverHypertrophy
- DEV_Rabbit_PregnancyRelated_MaternalPND21
- DEV_Rat_WeightReduction
- CHR_Rat_ThyroidSkeletal_Axial
- CHR_Rat_KidneyHyperplasia
- DEV_Rat_Nephropathy
- MGR_Rat_Skeletal_Cranial
- DEV_Rabbit_Epididymis
- DEV_Rat_General_FetalSkeletal_Axial
- MGR_General_WeightReduction
- CHR_Mouse_LitterSize
- DEV_Mouse_LiverFetalPathology
- DEV_Rat_Skeletal_ProliferativeLesions
- CHR_Mouse_Appendicular
- MGR_Rat_Ovary
- DEV_Rat_PregnancyRelated_EmbryoPathology
- DEV_Rabbit_PregnancyRelated_EmbryoFetalLoss
- CHR_Rat_Tumorigenidial



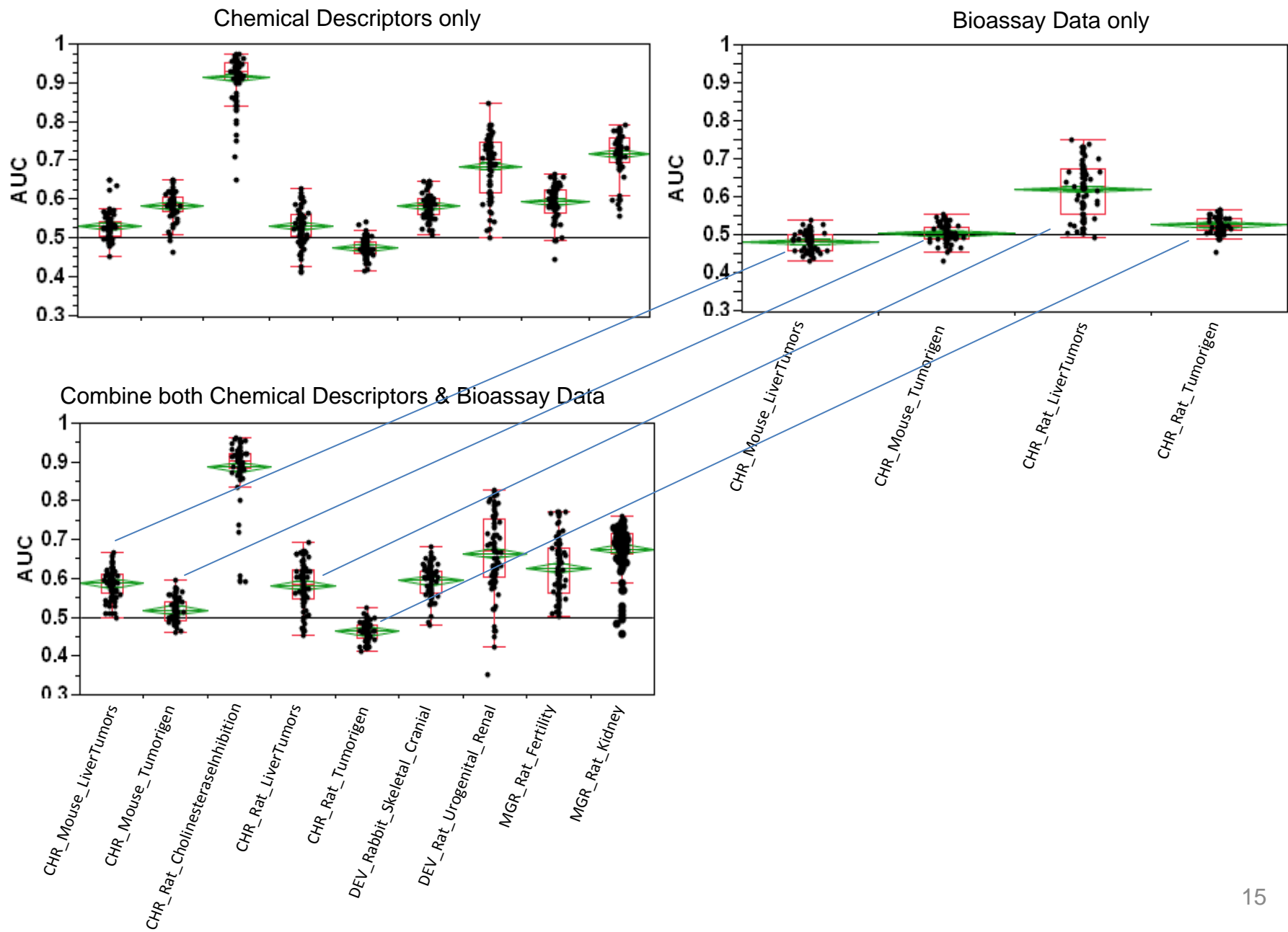
Ave (Sensitivity + Specificity): Top 10 models



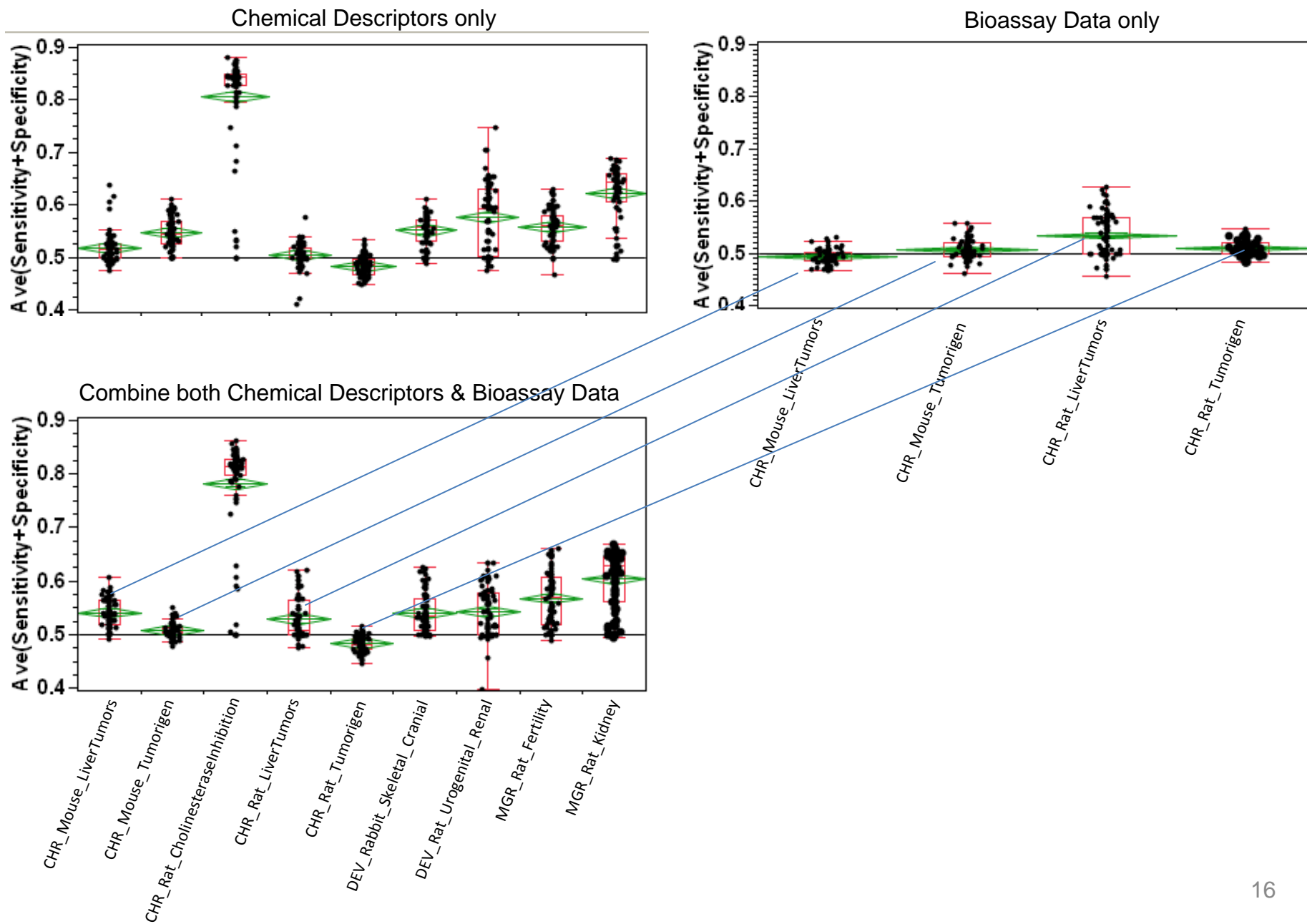
- Can we make general conclusions about performance of various models?

	Discrim Analysis	Distance Scoring	GLM Select	kNN	Logistic	PLS	Partition Trees	Radial Basis Machine
Proportion models examined	.095	.095	.142	.095	.155	.178	.190	.047
Proportion models included among "top 10"	.10	.056	.138	.02	.088	.198	.364	.036

Prediction Comparison Based on AUC

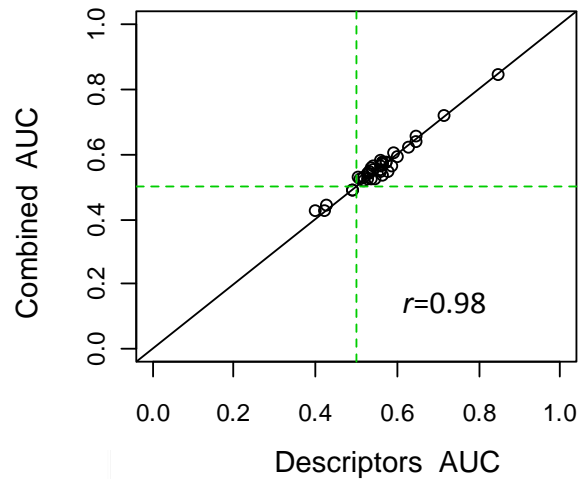
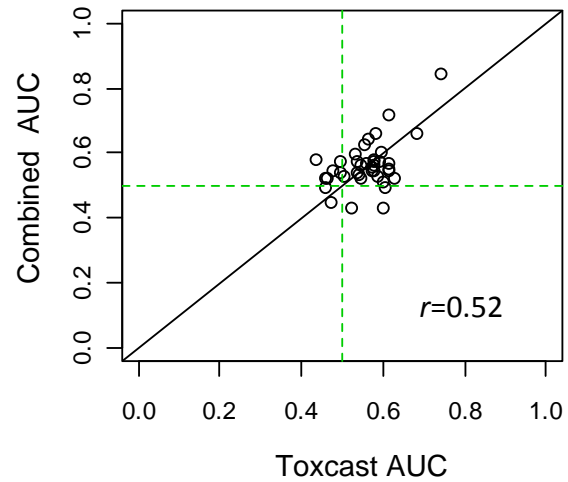
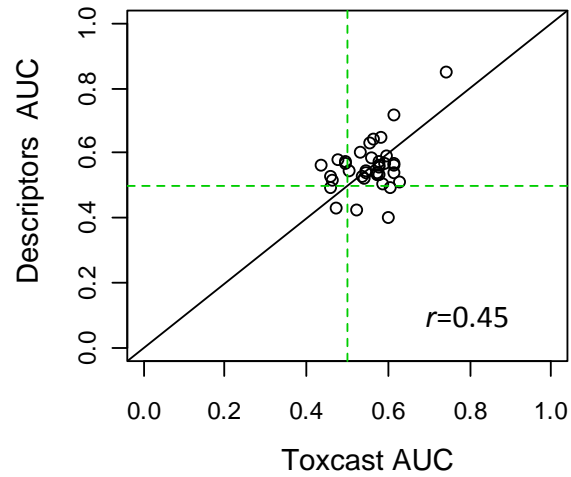


Prediction Comparison Based on Ave (Sensitivity + Specificity)

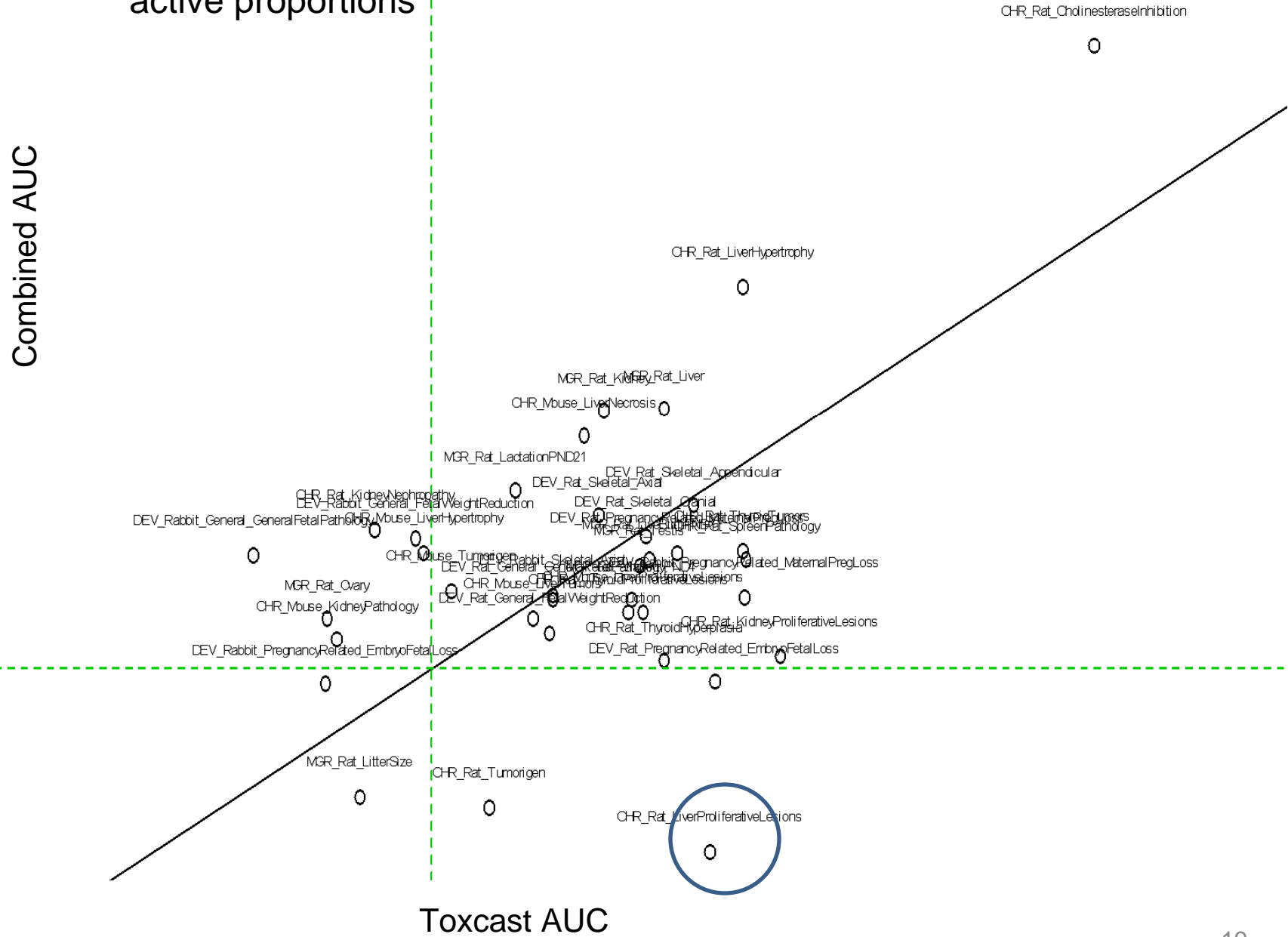


- We were unsure of the implications of these results, so we have investigated further
- Need to more systematically run comparisons of performance using ToxCast data only, the chemical descriptors only, and the combination
- We have lots of results, but can show only a portion due to time constraints
- We ran many additional models, including PAM (shrunken centroid approach), standard LASSO (a penalized regression approach), and an adaptive LASSO procedure under development at UNC.

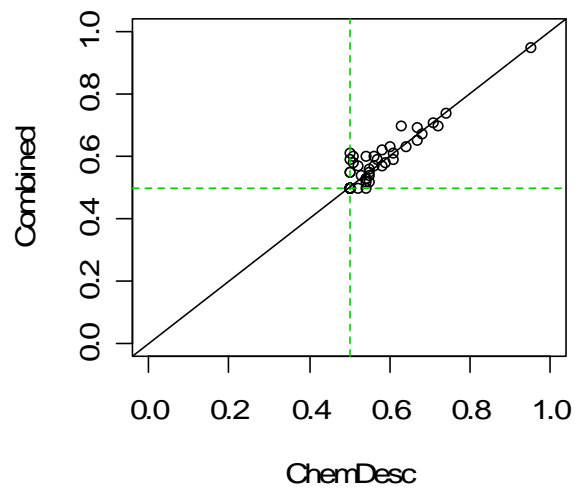
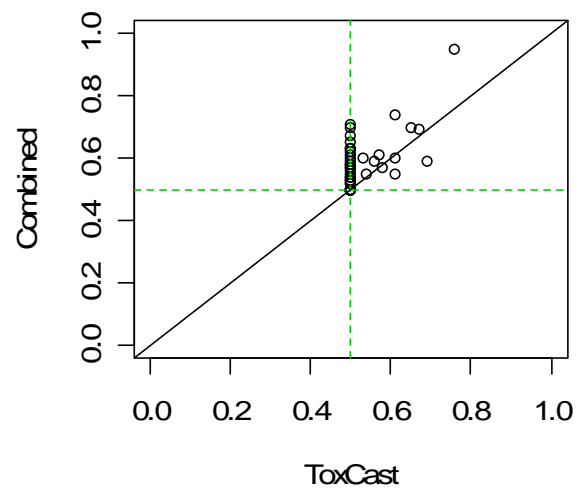
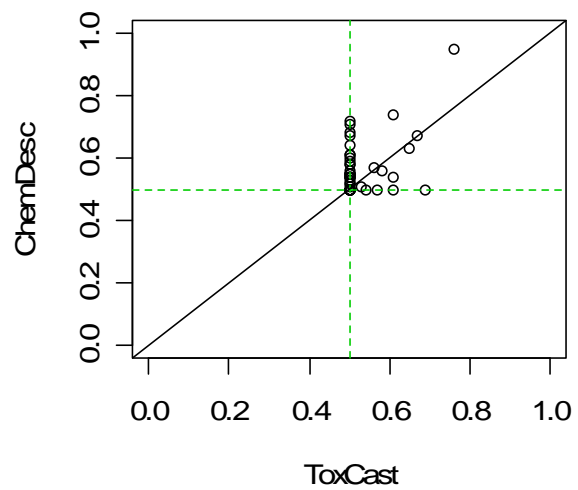
PAM comparative AUC results on 35 endpoints with highest active proportions



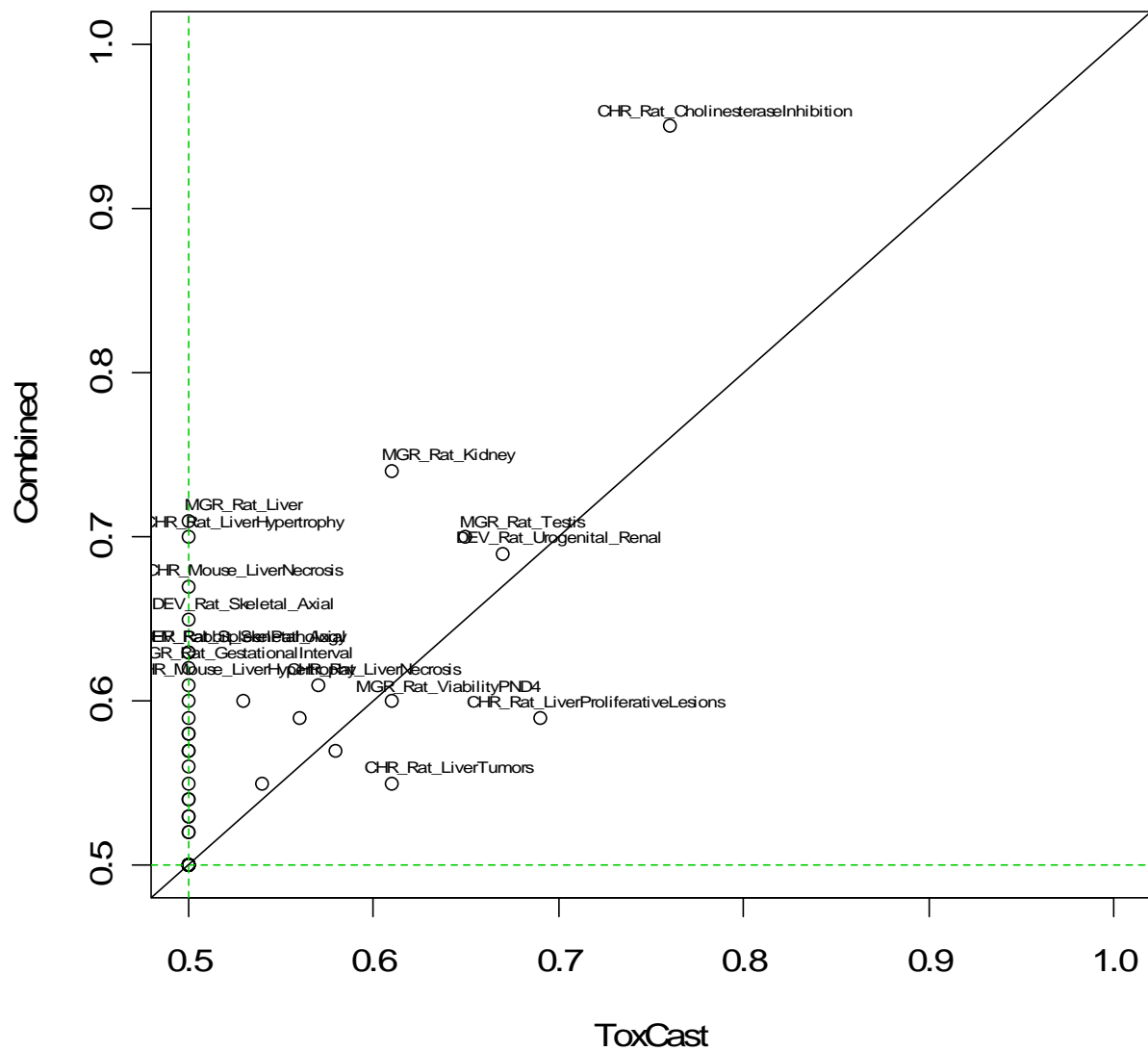
PAM comparative AUC results on 35 endpoints with highest active proportions



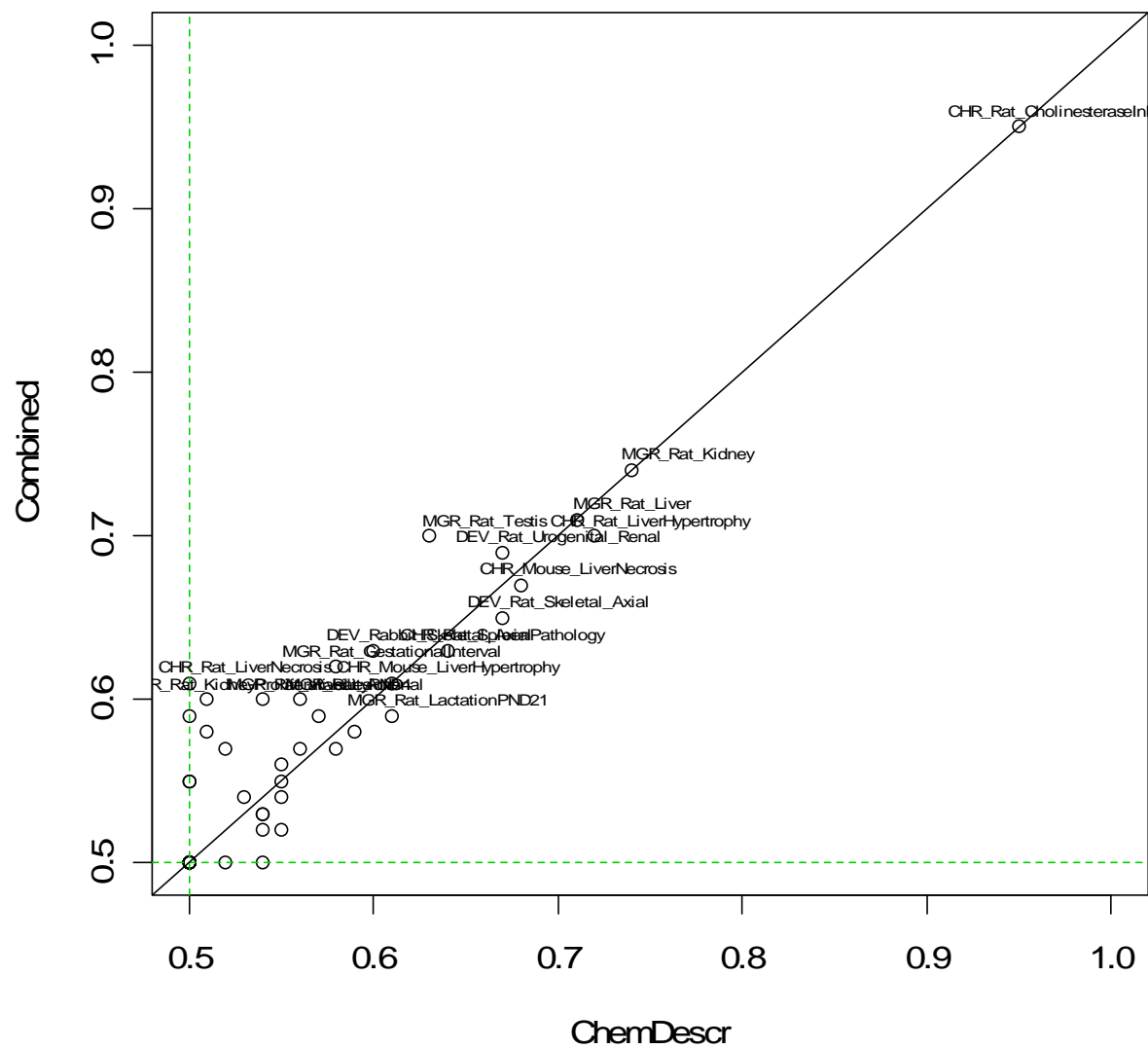
results with LASSO (AUC on each axis)



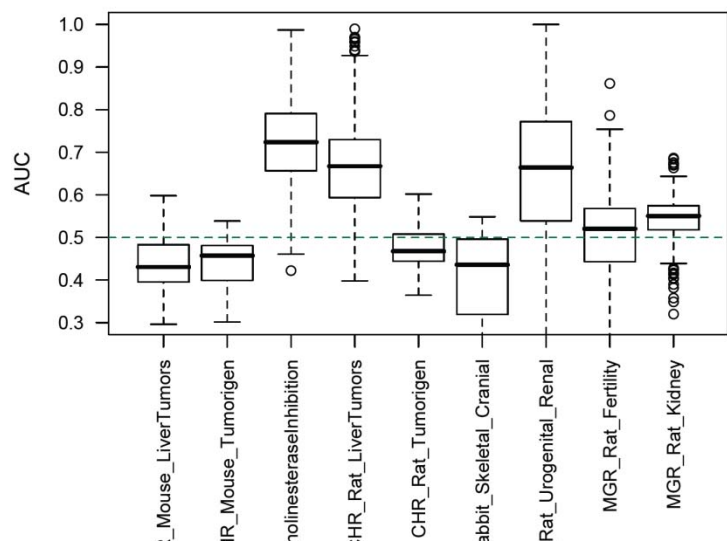
results with LASSO (AUC on each axis)



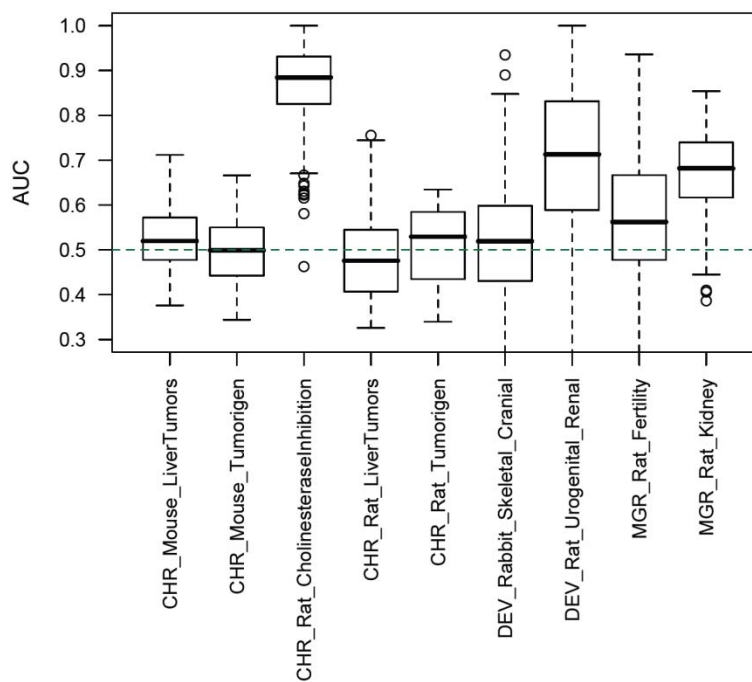
results with LASSO (AUC on each axis)



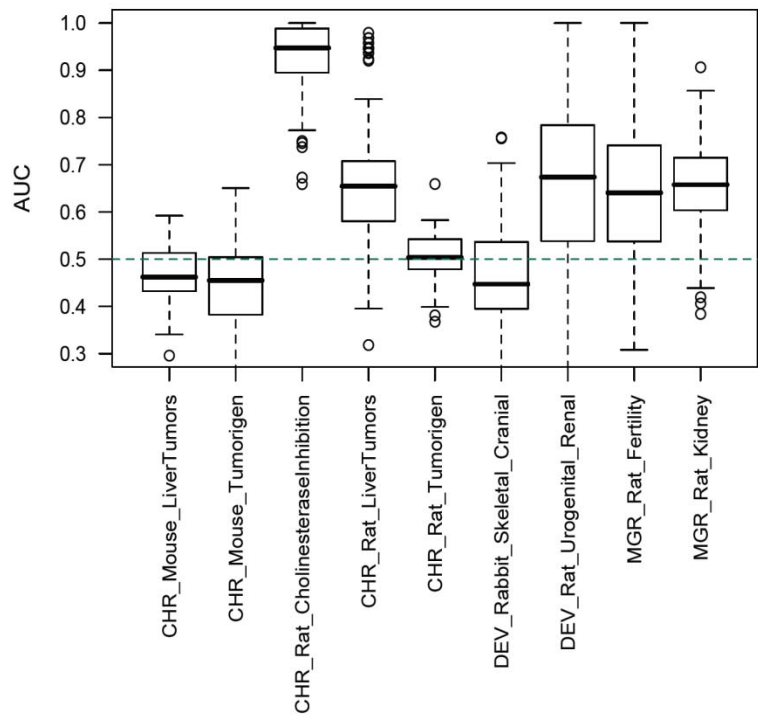
Bio-assay Only



Chemical identifier Only

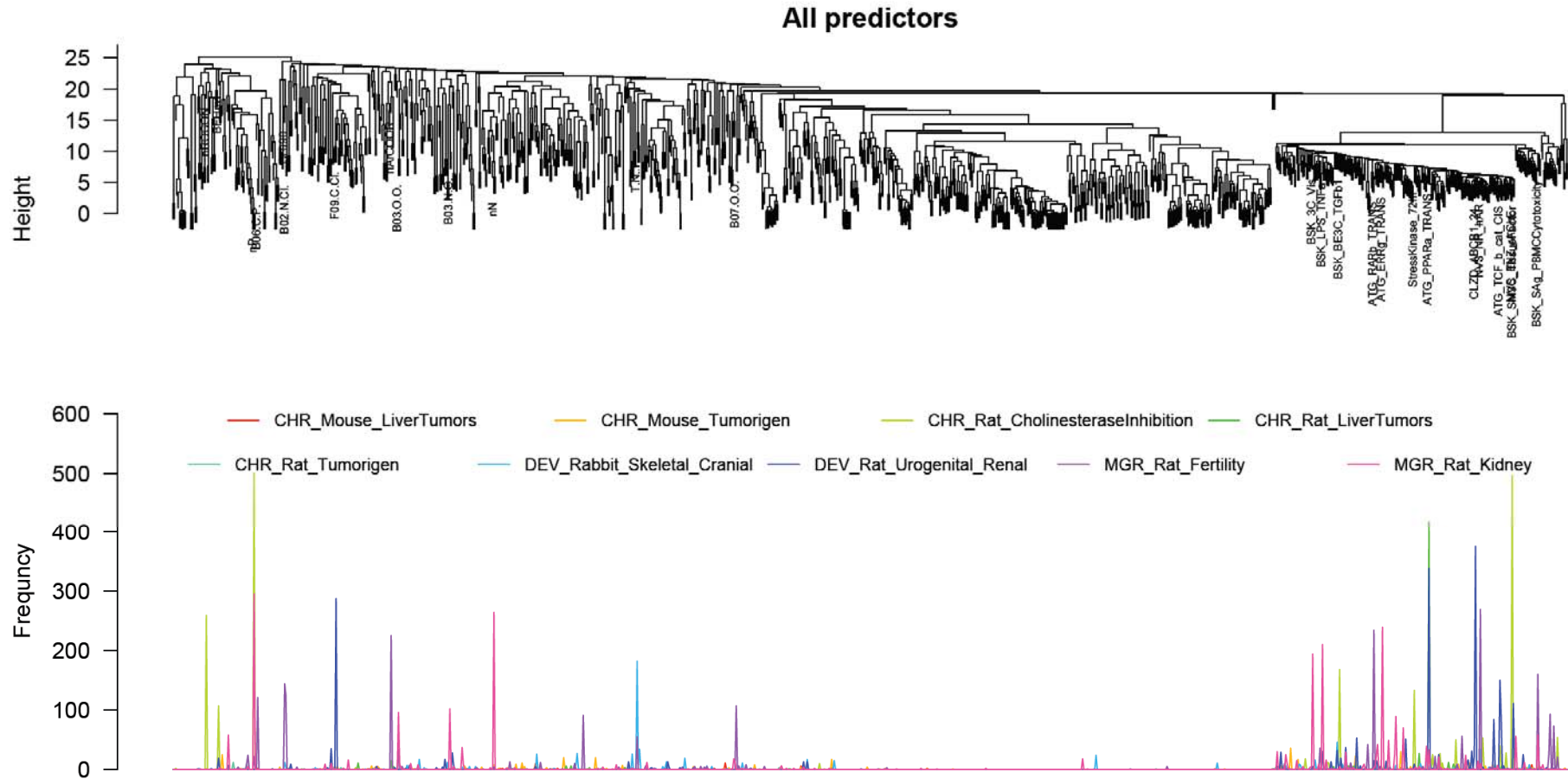


All predictors



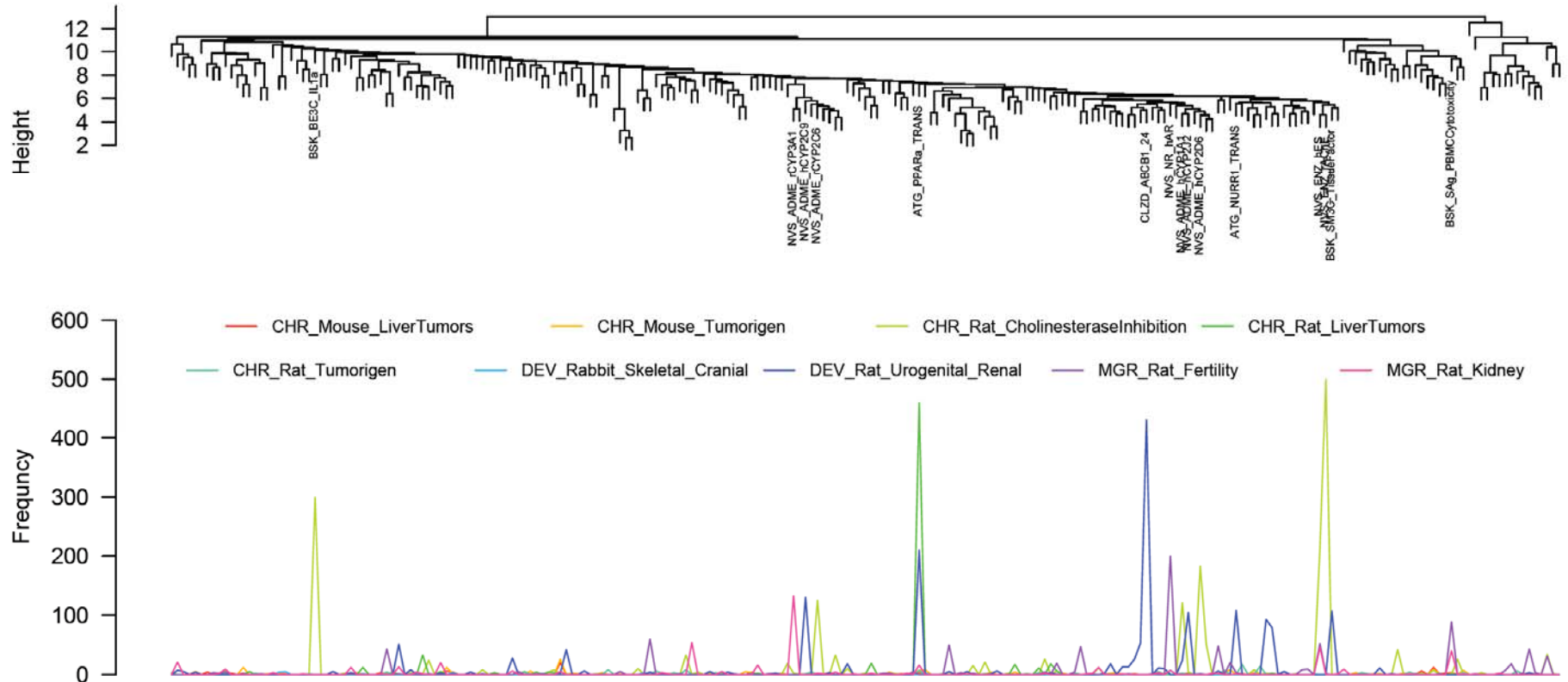
AUC results with adaptive LASSO, original 4 endpoints, plus those with “good” AUC

adaptive LASSO, frequency of predictors selected across 100 5-fold CV iterations



adaptive LASSO, frequency of predictors selected across 100 5-fold CV iterations

Bio-assay Only



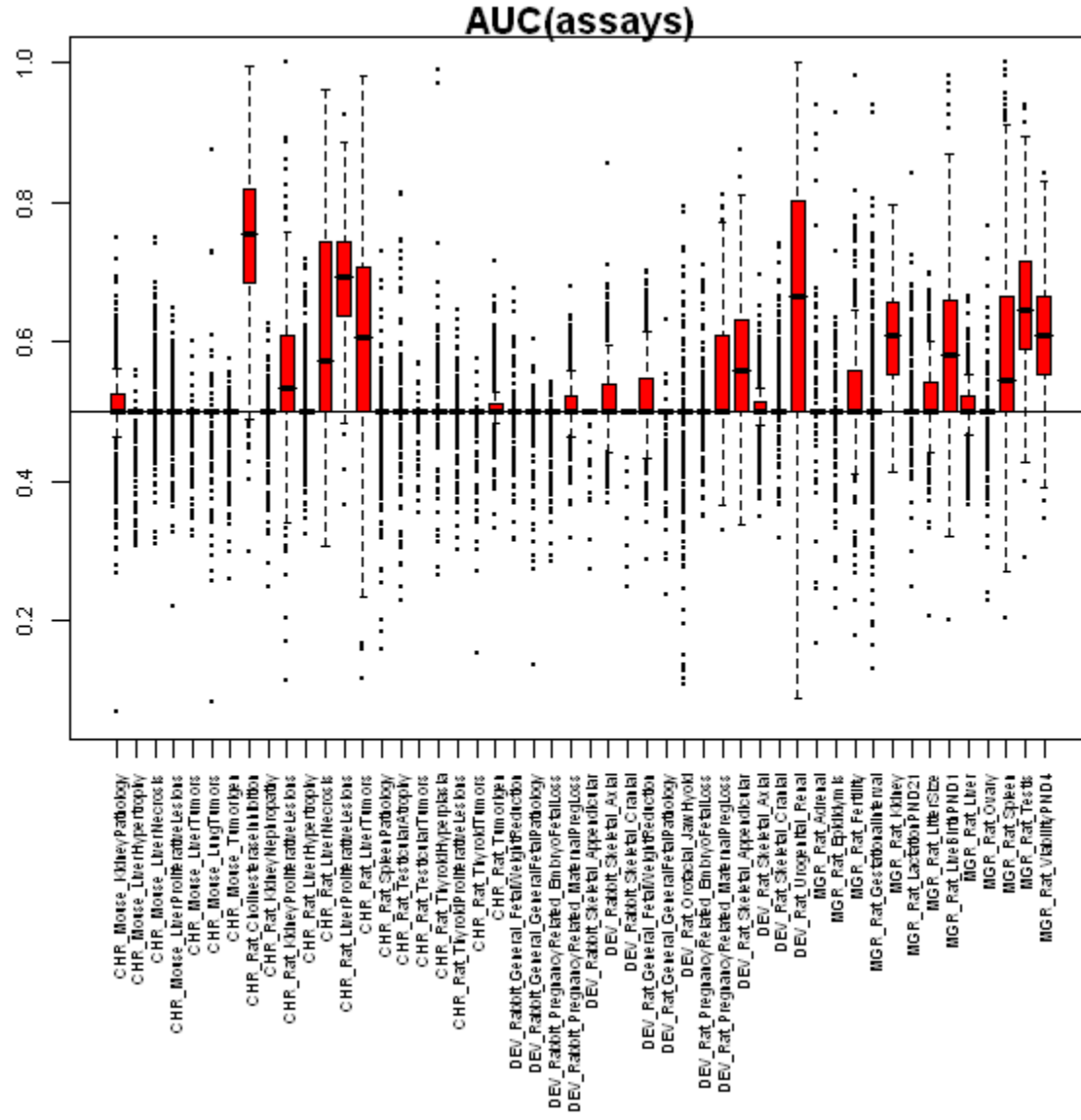
Summary

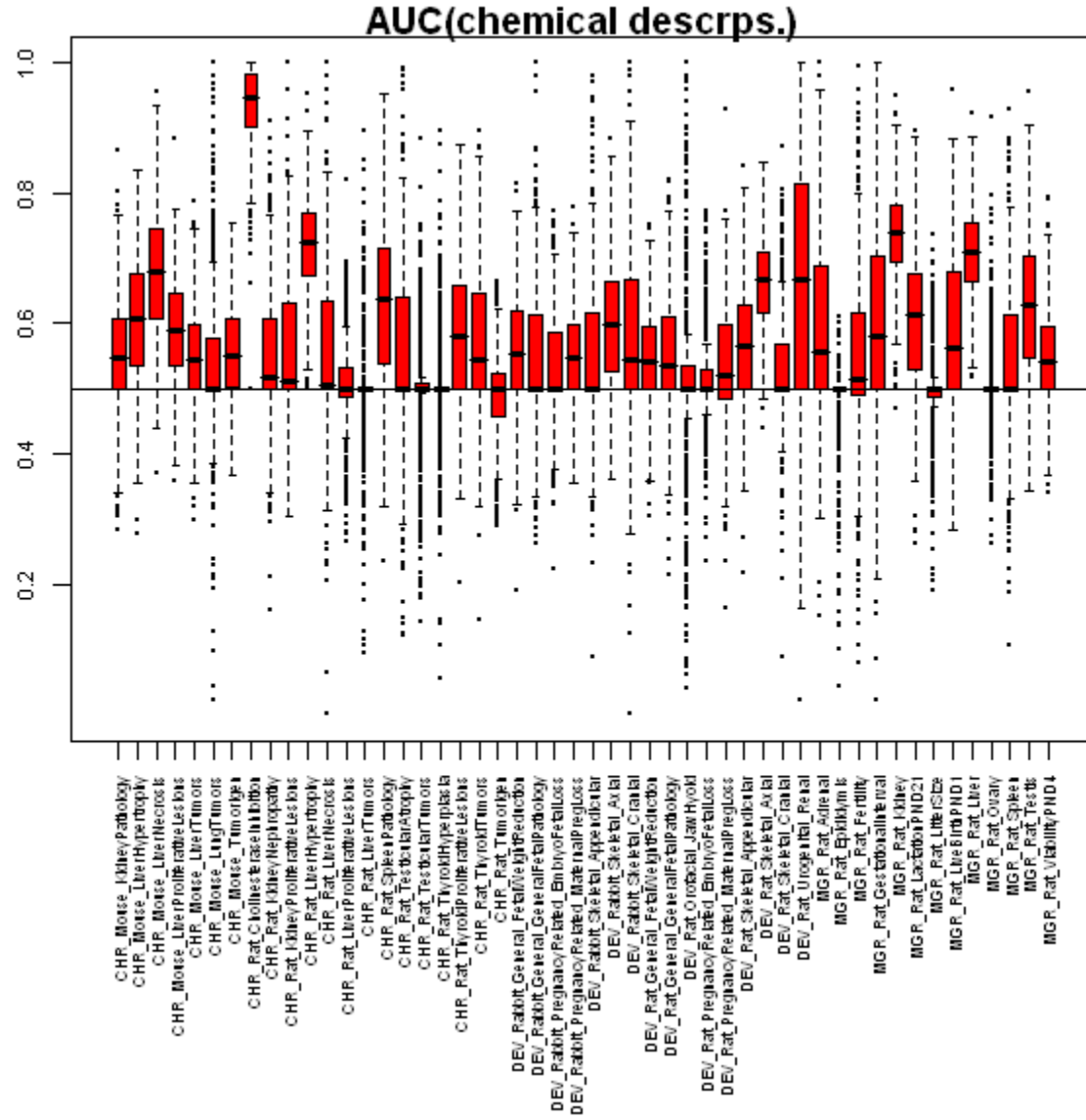
- Our efforts remain a work in progress, as we are still digesting the variables selected and their importance
- Reasonable pre-selection of variables seems to be a good thing
- The chemical descriptors seem to boost predictive accuracy
- However, we suspect that the chemical descriptors might be dominating our models partly due to internal correlation. This can distort interpretation

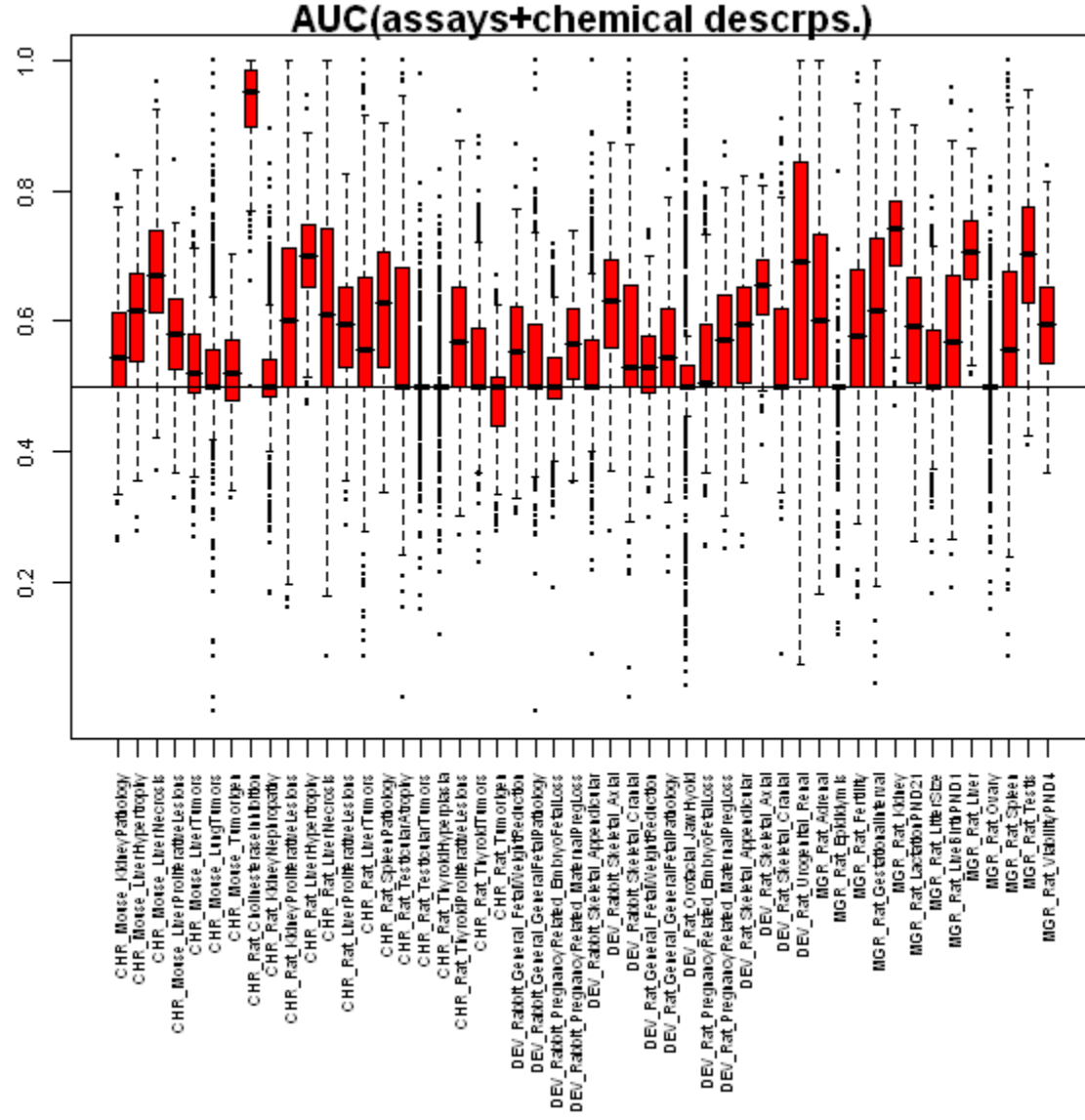
Summary

- We may soon approach a limit of simplistic approaches that do not consider potentially complicated interactions.
- One type of “interaction” might result from different prediction rules holding for different types of chemicals. For example, simple clustering by all toxicity endpoints clearly groups chemicals with differing average activity values.
- Thus some of our next steps will involve attempts at predicting such clusters with ToxCast predictors in order to further boost prediction performance.

EXTRA SLIDES







adaptive LASSO, frequency of predictors selected across 100 5-fold CV iterations

Chemical identifiers

