

# Predictive Modeling of Toxicity Outcomes with Grammatical Evolution Neural Networks

*Nicholas Hardison*

*Alison Motsinger-Reif*

Bioinformatics Research Center

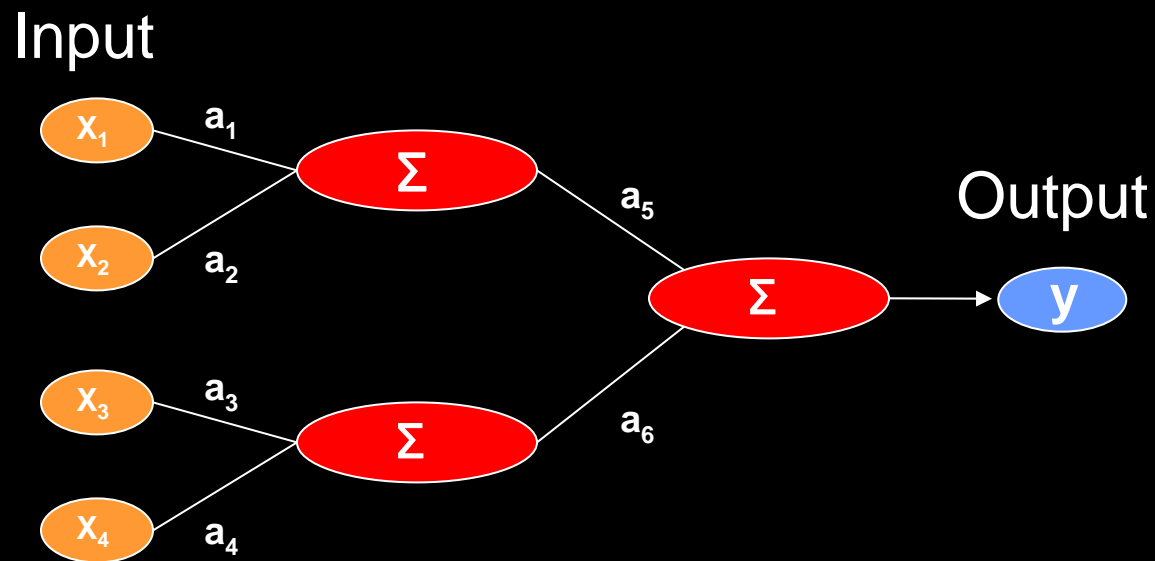
Department of Statistics

North Carolina State University

# Motivation

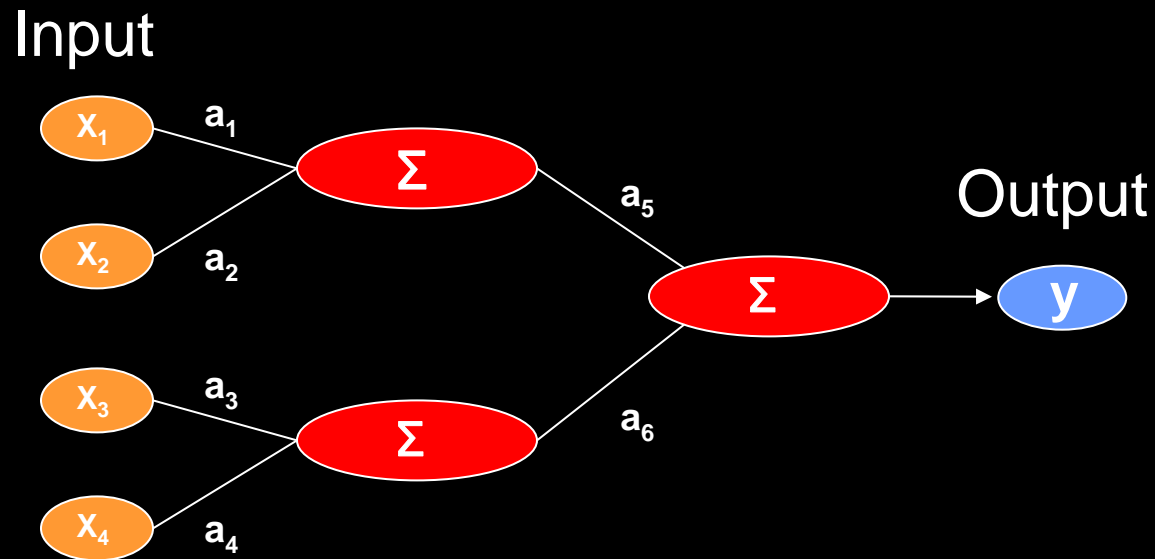
- Identification of complex predictive models
  - Nonlinearity is common in biology
  - Need to identify multivariate, interactive models
- Explosion of high throughput technologies
- Traditional statistical approaches are limited
  - Variable selection
  - Limitations with high dimensionality
- Goals for novel methods
  - Variable Selection
  - Statistical Modeling

# Neural Networks



- Developed 60 years ago
- Originally developed to model/mimic the human brain
- More recently, uses theory of neurons to do computation
- Applications
  - Association, classification, categorization

# Neural Networks



- NNs multiply each input node (i.e. variable, genotype, etc.) by a weight ( $a$ ), the result of which is processed by a function ( $\Sigma$ ), and then compared to a threshold to yield an output (0 or 1).
- Weights are applied to each connection and optimized to minimize the error in the data.

# Neural Networks

- Advantages

- Can handle large quantities of data
- Universal function approximators
- Model-free

- Limitations

- Must fix architecture prior to analysis
- Only the weights are optimized
- Weights are optimized using hill-climbing algorithms

# Neural Networks

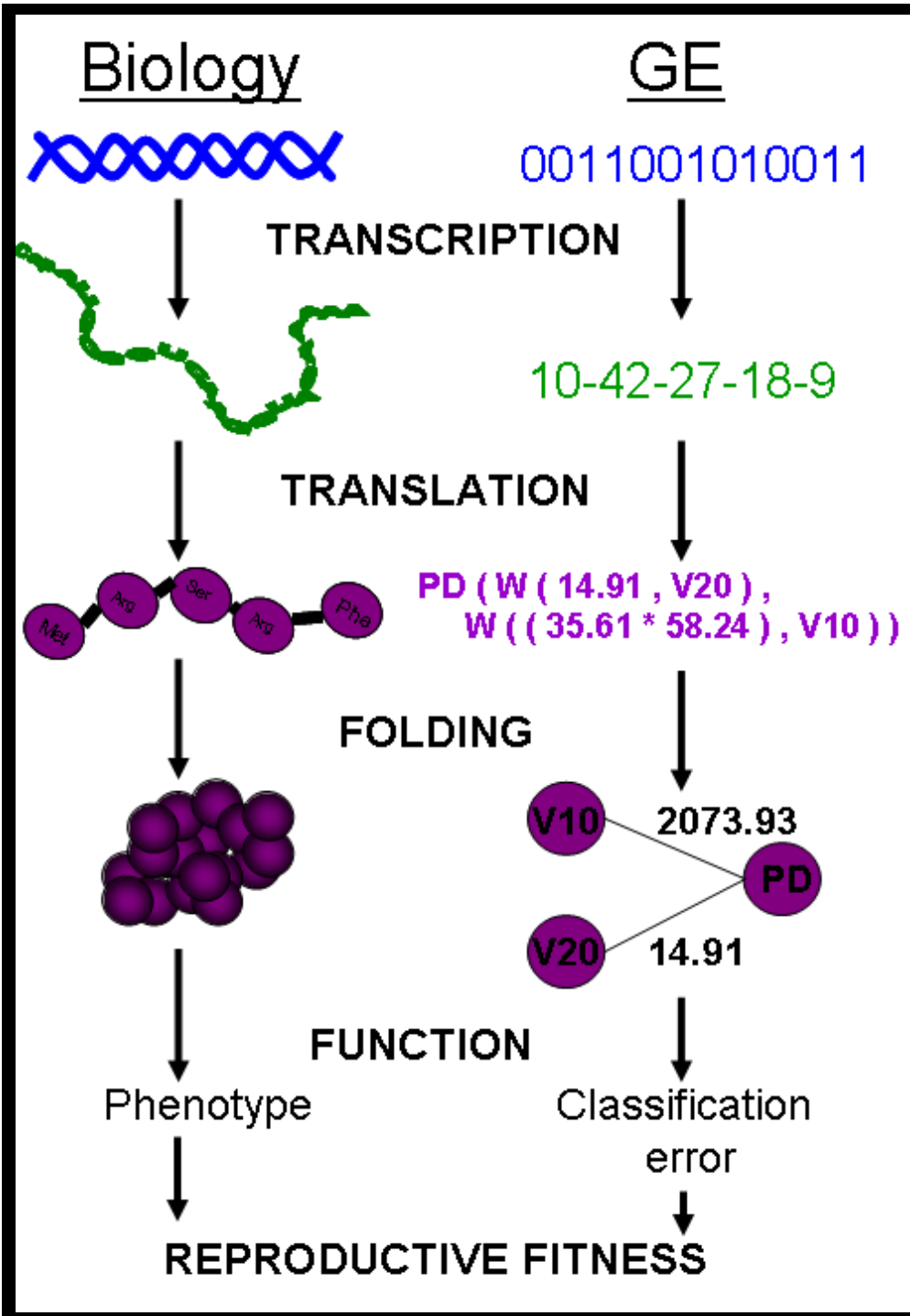
- Advantages

- Can handle large quantities of data
- Universal function approximators
- Model-free

- Limitations

- Must fix architecture prior to analysis
- Only the weights are optimized
- Weights are optimized using hill-climbing algorithms

- Evolutionary computation algorithms can be used for the optimization of the *inputs*, *architecture*, and *weights* of a NN to improve the power to identify gene-gene interactions.



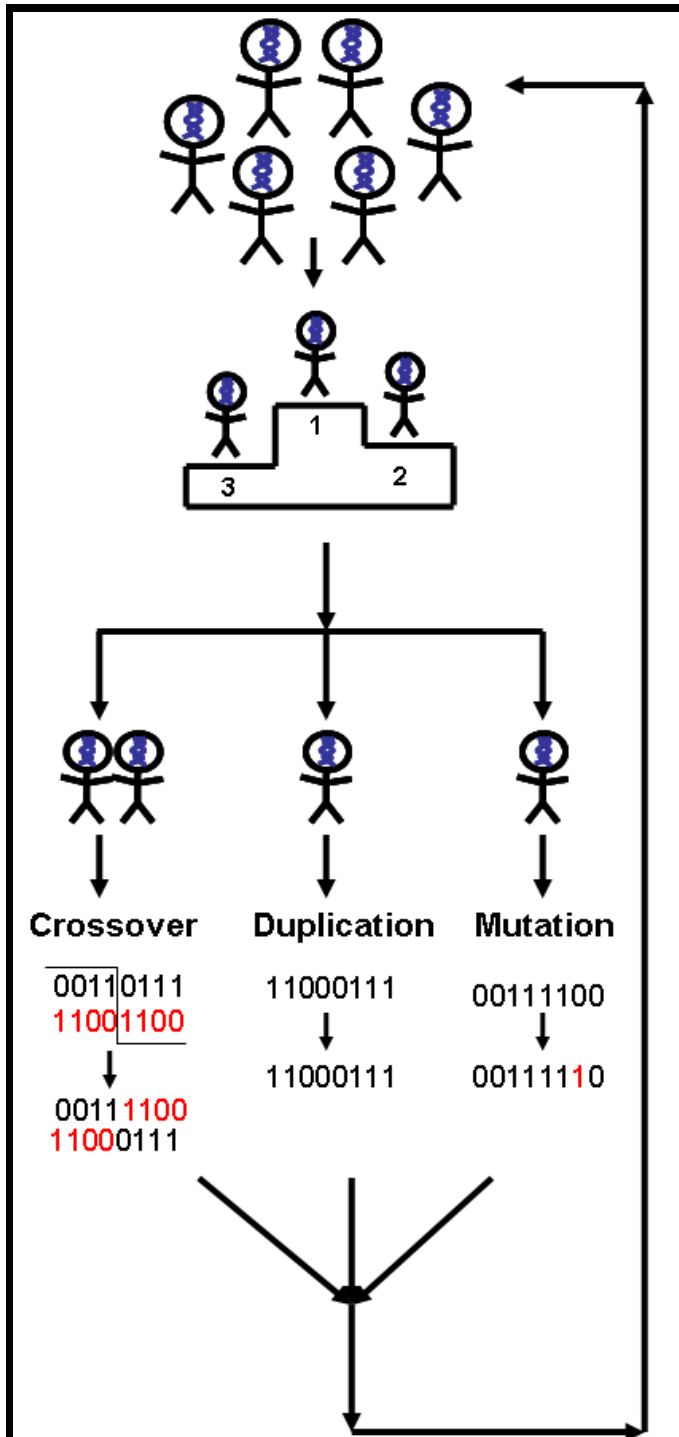
**DNA:** The heritable material in GE is the binary string chromosome. The GE chromosome is divided into codons, undergoes crossover and mutation, and can contain non-coding sequence just as biological DNA.

**RNA:** In GE, the binary chromosome string is transcribed into an integer string. This integer string is a linear copy message of the original heritable material that can then be processed further.

**Polypeptide String:** The integer string is translated using the grammar provided into the code for a functional NN.

**Protein Folding:** The grammar encoding is then interpreted as a multi-dimensional NN. This NN produces a classification error, just as a protein produces a phenotype within an organism.

**Function:** In GE a lower classification error indicates higher fitness. Natural selection will work at the level of reproductive fitness, forcing changes in the heritable material of both biological organisms or GE individuals.



Step 1: A population of individuals is randomly generated, where each individual is a binary string chromosome (genetic material). The number of individuals is user-specified.

Step 2: Individuals are randomly chosen for tournaments – where they compete with other individuals for the highest fitness, and the tournament winners get to pass on their genetic material.

Step 3: Of the winners, user-specified proportions participate in crossover, mutation, or duplication of their genomes to produce offspring.

Step 4: When pooled together, these offspring will become the initial population for the next generation of evolution.

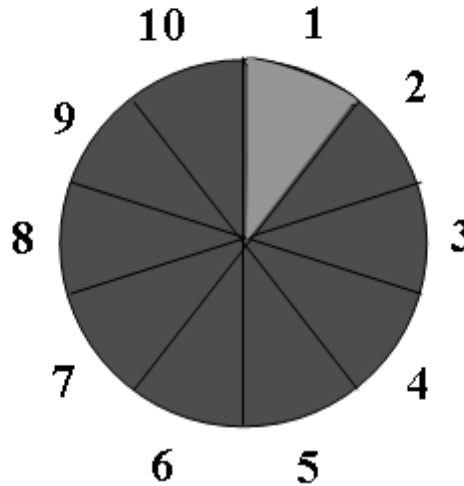
Steps 1-4 are repeated for a user-specified number of generations, to produce offspring with the highest possible fitness.

# GE Neural Networks

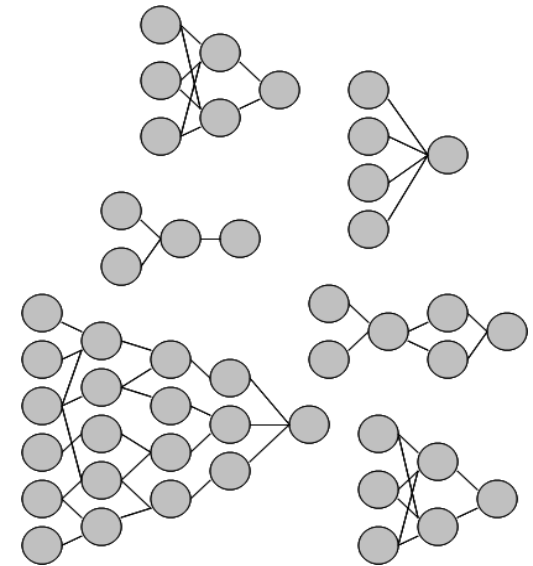
STEP 1

population_size	200
max_generations	50
pvm_exchange_generations	25
random_seed	7
crossover_rate	0.9
mutation_rate	0.01
codon_size	8
wrapper_count	2
min_chrom_size	50
max_chrom_size	1000

STEP 2



STEP 3



STEP 6

STEP 5

STEP 4

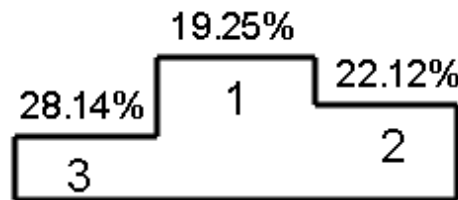
**GENN Model**

Classification Error    Prediction Error  
**19.25**                      **21.55**



**GENN Models**  
Classification Error

**19.25**  
**22.12**  
**24.33**  
**28.14**  
**⋮**



Tournament

# Previous Applications of GENN

- High power to detect a wide range of main effect and interactive models
  - Motsinger-Reif AA, Dudek SM, Hahn LW, and Ritchie MD. Comparison of approaches for machine-learning optimization of neural networks for detecting gene-gene interactions in genetic epidemiology. Genetic Epidemiology 2008 Feb 8 [Epub ahead of print]
- Robust to changes in the evolutionary process
  - Motsinger AA, Hahn LW, Dudek SM, Ryckman KK, Ritchie MD. Alternative cross-over strategies and selection techniques for Grammatical Evolution Optimized Neural Networks. In: Maarten Keijzer et al, eds. Proceeding of Genetic and Evolutionary Computation Conference 2006 Association for Computing Machinery Press, New York, pp. 947-949.
- Higher power than traditional BPNN, GPNN, or random search NN
  - Motsinger AA, Dudek SM, Hahn LW, and Ritchie MD. Comparison of neural network optimization approaches for studies of human genetics. Lecture Notes in Computer Science, 3907: 103-114 (2006).
  - Motsinger-Reif AA, Dudek SM, Hahn LW, and Ritchie MD. Comparison of approaches for machine-learning optimization of neural networks for detecting gene-gene interactions in genetic epidemiology. Genetic Epidemiology 2008 Feb 8 [Epub ahead of print]

# Previous Applications of GENN

- Robust to class imbalance

- Hardison NE, Fanelli TJ, Dudek SM, Ritchie MD, Reif DM, Motsinger-Reif AA. Balanced accuracy as a fitness function in Grammatical Evolution Neural Networks is robust to imbalanced data. Genetic and Evolutionary Algorithm Conference. 2008 353-354.

- Scales linearly in regards to computation with the number of variables

- Motsinger AA, Reif DM, Dudek SM, and Ritchie MD. Dissecting the evolutionary process of Grammatical Evolution Optimized Neural Networks. Proceedings of the IEEE Symposium on Computational Intelligence in Bioinformatics and Computational Biology 2006 pp. 1-8.

- Robust to common types of error in datasets

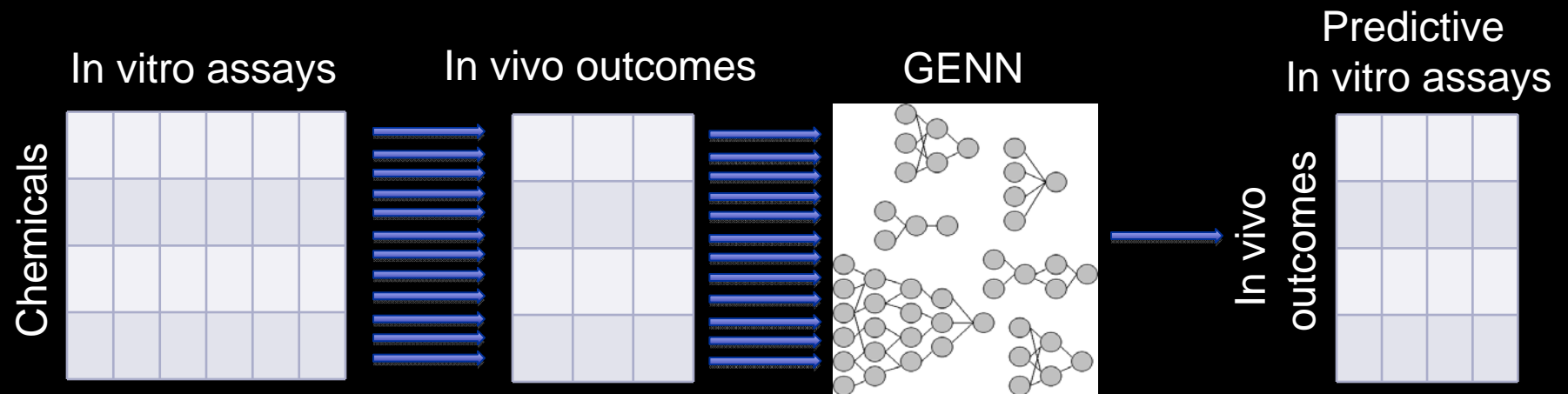
- Motsinger AA, Fanelli TJ, Ritchie MD. Power of Grammatical Evolution Neural Networks to detect gene-gene interactions in the presence of error common to genetic epidemiological studies. BMC Research Notes 2008 Aug 13;1:65.

# Previous Applications of GENN

- Has relatively high power in the presence of heterogeneity
  - Motsinger AA, Fanelli TJ, Ritchie MD. Power of Grammatical Evolution Neural Networks to detect gene-gene interactions in the presence of error common to genetic epidemiological studies. BMC Research Notes 2008 Aug 13;1:65.
- The presence of collinearity between variables increases the power of GENN
  - Motsinger AA, Reif DM, Fanelli TJ, Davis AC, Ritchie MD. Linkage disequilibrium in genetic association studies improves the power of Grammatical Evolution Neural Networks. Proceedings of the IEEE Symposium on Computational Intelligence in Bioinformatics and Computational Biology 2007 pp. 1-8.
- Has been favorably compared to other methods
  - Random Forests, Focused Interaction Testing Framework, Multifactor Dimensionality Reduction, Logistic Regression
  - Motsinger-Reif AA, Reif DM, Fanelli TJ, Ritchie MD. Comparison of computational approaches for genetic association studies. Genetic Epidemiology 2008 Jun 16. [Epub ahead of print]

# Application to ToxCast™


- Goal: Apply GENN for each in vivo outcome to find (combinations of) assays that predict each in vivo response



# ToxCast Data

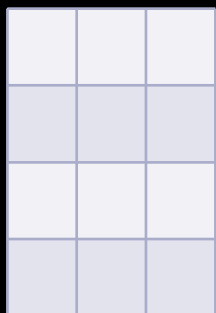
In vitro assays

Chemicals



- Input variables: 524 in vitro assays
  - Discretized to facilitate cross-platform comparability
  - Three categories: {hit, no response, missing}

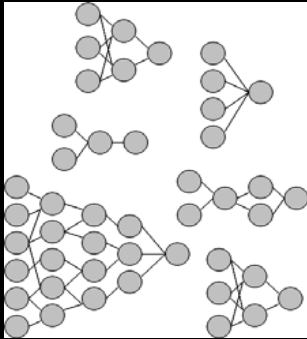
In vivo outcomes



- Instances: chemicals
- Outcomes: 76 in vivo endpoints (binary)
  - Binary: {endpoint observed; not observed}

# Analysis

GENN



- Iterated GENN over all 76 outcomes
  - Parameters used: 2000 generations, crossover rate of .9, mutation rate of .1, 4 demes
- For each outcome:
  - Evaluated training and testing accuracies
  - Selected assays associated with outcomes based on a minimum cross validation consistency (CVC) of 5/10

Predictive

In vitro assays

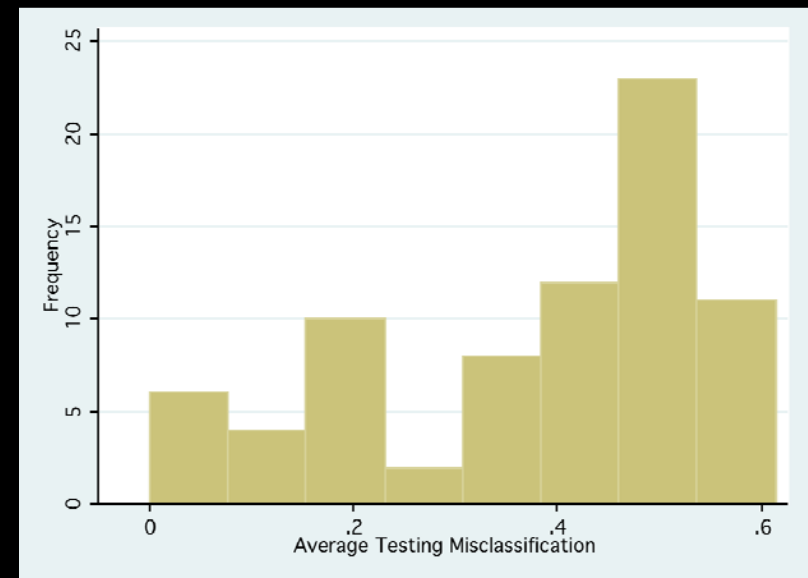
In vivo  
outcomes


# Simplifying Assumptions

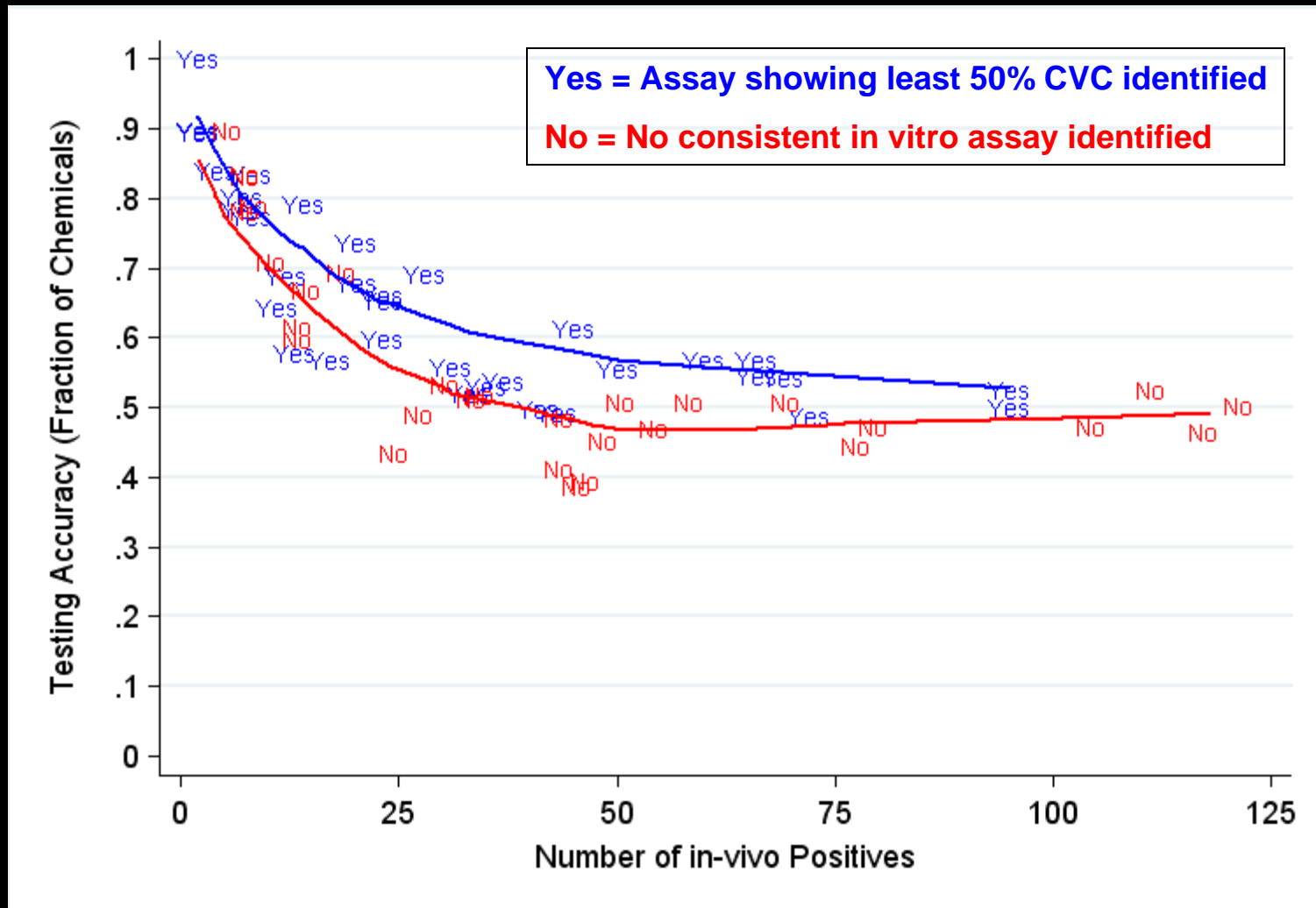
- Chemicals are treated as equally weighted instances
  - Does not weight based on subgroups, exposure, etc
- Assays are uncorrelated
  - Important is assessing CVC
- In vivo endpoints are treated as equally weighted outcomes
  - Does not weight on overall risk or severity, etc

# Results Across All Endpoints

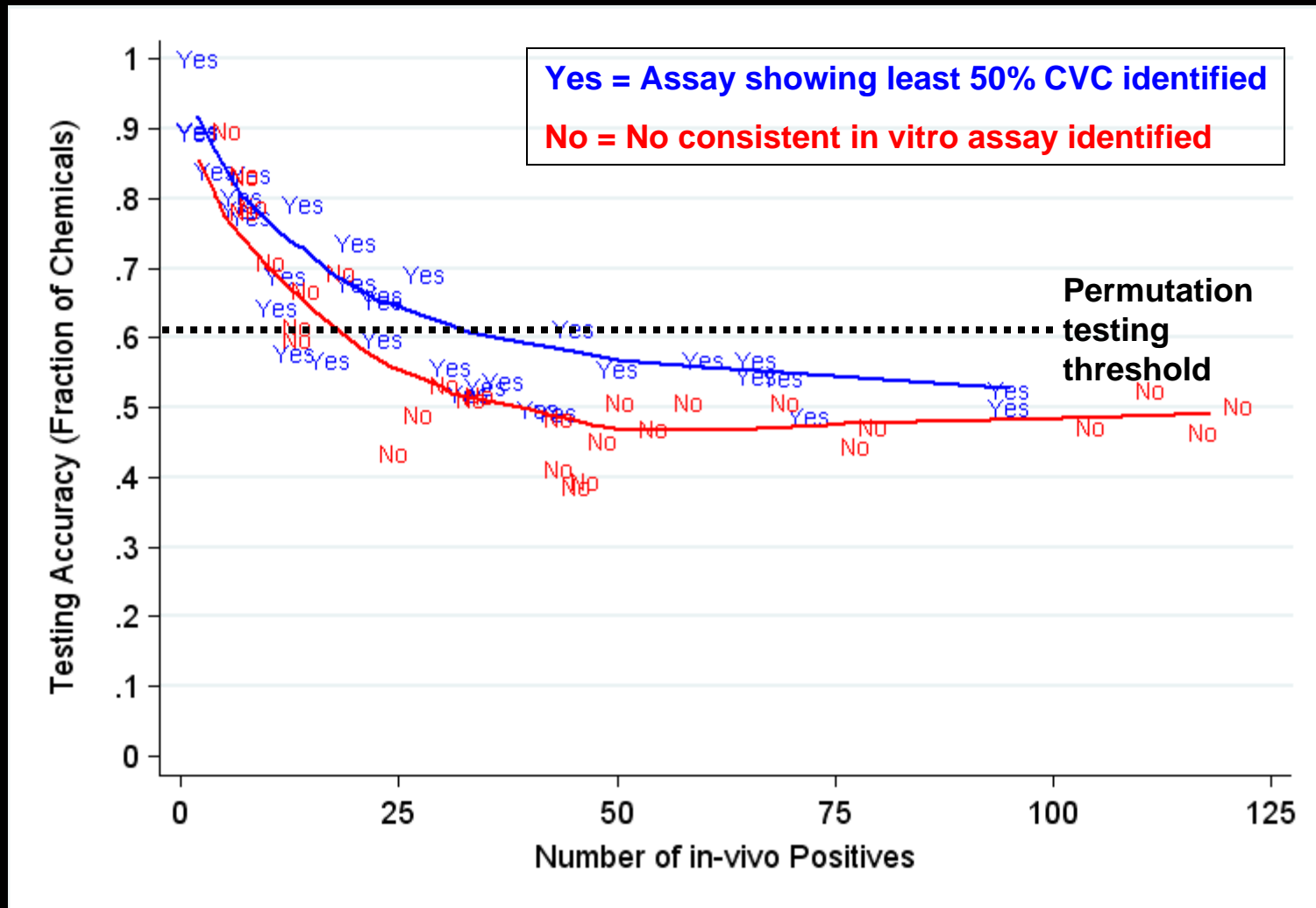
- Prediction errors:  
37.6(.18)
- Training errors:  
24.5 (.113)
- Number of outcomes with assay min 50% CVC:  
42



# Predictive ability (testing accuracy) alone can be misleading



# Predictive ability (testing accuracy) alone can be misleading



# Statistically Significant GENN Results: Chronic Rat Endpoints

Endpoints	Predictive Assays
CHR_Rat_CholinesteraseInhibition	Novascreen_NVS_ENZ_rAChE
CHR_Rat_KidneyNephropathy	BioSeek_BSK_LPS_PGE2
CHR_Rat_LiverNecrosis	Attagene_ATG_M_61_CIS Attagene_ATG_PPARa_TRANS BioSeek_BSK_SM3C_Thrombomodulin
CHR_Rat_LiverTumors	Attagene_ATG_PPARa_TRANS
CHR_Rat_TesticularTumors	Cellumen_CLM_MitoticArrest_72hr Attagene_ATG_NFI_CIS
CHR_Rat_SpleenPathology	BioSeek_BSK_3C_VCAM1 Attagene_ATG_NFI_CIS

# Examine Unique Assay Overlap

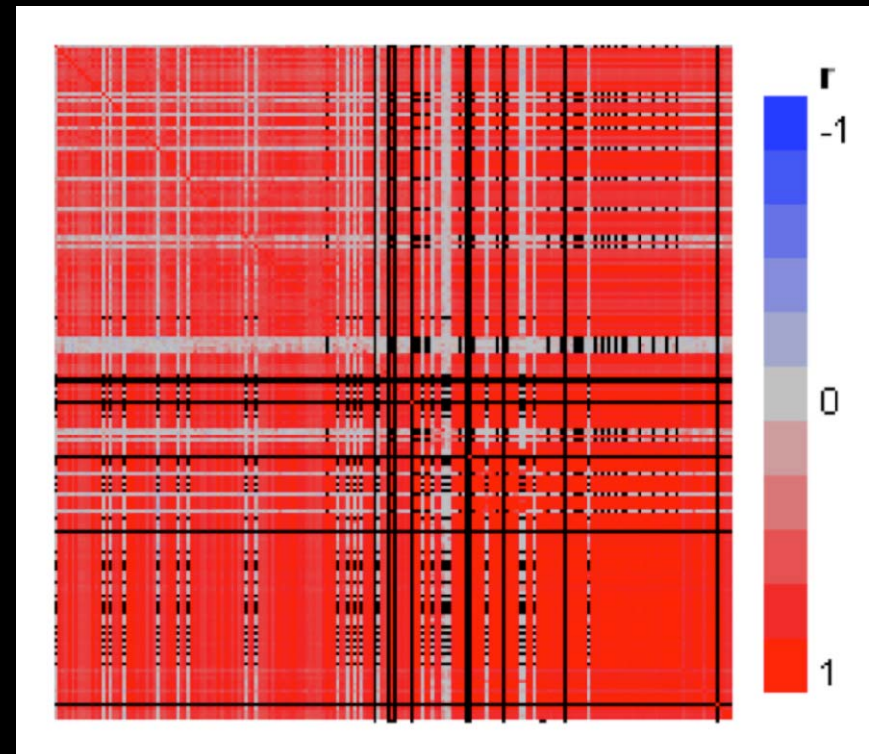
Endpoints	Predictive Assays
CHR_Rat_CholinesteraseInhibition	Novascreen_NVS_ENZ_rAChE
CHR_Rat_KidneyNephropathy	BioSeek_BSK_LPS_PGE2
CHR_Rat_LiverNecrosis	Attagene_ATG_M_61_CIS Attagene_ATG_PPARa_TRANS BioSeek_BSK_SM3C_Thrombomodulin
CHR_Rat_LiverTumors	Attagene_ATG_PPARa_TRANS
CHR_Rat_TesticularTumors	Cellumen_CLM_MitoticArrest_72hr Attagene_ATG_NFI_CIS
CHR_Rat_SpleenPathology	BioSeek_BSK_3C_VCAM1 Attagene_ATG_NFI_CIS

# Examine Unique Assay Overlap

Endpoints	Predictive Assays
CHR_Rat_CholinesteraseInhibition	Novascreen_NVS_ENZ_rAChE
CHR_Rat_KidneyNephropathy	BioSeek_BSK_LPS_PGE2
CHR_Rat_LiverNecrosis	Attagene_ATG_M_61_CIS Attagene_ATG_PPARa_TRANS BioSeek_BSK_SM3C_Thrombomodulin
CHR_Rat_LiverTumors	Attagene_ATG_PPARa_TRANS
CHR_Rat_TesticularTumors	Cellumen_CLM_MitoticArrest_72hr Attagene_ATG_NFI_CIS
CHR_Rat_SpleenPathology	BioSeek_BSK_3C_VCAM1 Attagene_ATG_NFI_CIS
CHR_Rat_LiverProliferativeLesions	Attagene_ATG_Pax6_CIS Attagene_ATG_PPARa_TRANS

# Conclusions/Inconclusions

- Internal model validation can provide an important measure of predictive ability of models
- Cross validation consistency is a related, but not redundant measure
  - Correlation among input variables is a concern



Correlation matrix of assays

# Conclusions/Inconclusions

- How do we properly model prediction accuracies in rare endpoints?
- How do we incorporate/model heterogeneity in the prediction algorithms?
- How do we incorporate pathway knowledge/reconstruction into analysis?
- How do we prepare for increased scale of screening technology?

Questions?