Predicting Molecular Initiating Events from High Throughput Transcriptomic Screening using Machine Learning

Joseph Bundy
US EPA, Research Triangle Park, NC

Disclaimer

The views expressed in this presentation are those of the authors and do not necessarily represent the views or policies of the US EPA.

Project Context

A Current Challenge in Chemical Hazard Identification:

There are approximately 883,000 chemicals registered on the CompTox Chemicals Dashboard. Many chemicals have limited associated chemical safety information.

Solution:

New Approach Methodologies (NAMs) such as High-Throughput Transcriptomics (HTTr) combined with machine learning methods can help identify Molecular Initiating Events (MIEs) induced by chemical treatment for hundreds / thousands of chemicals at a time.

What are Molecular Initiating Events?

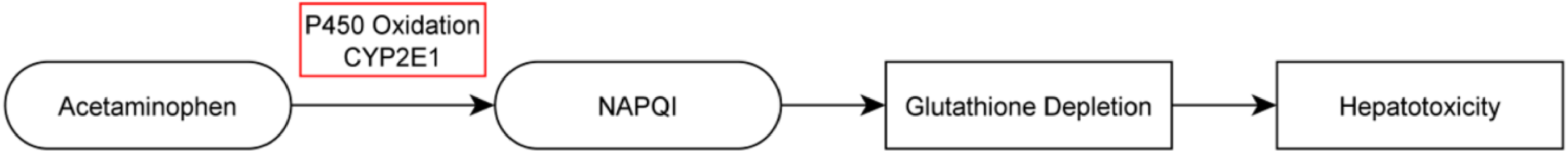


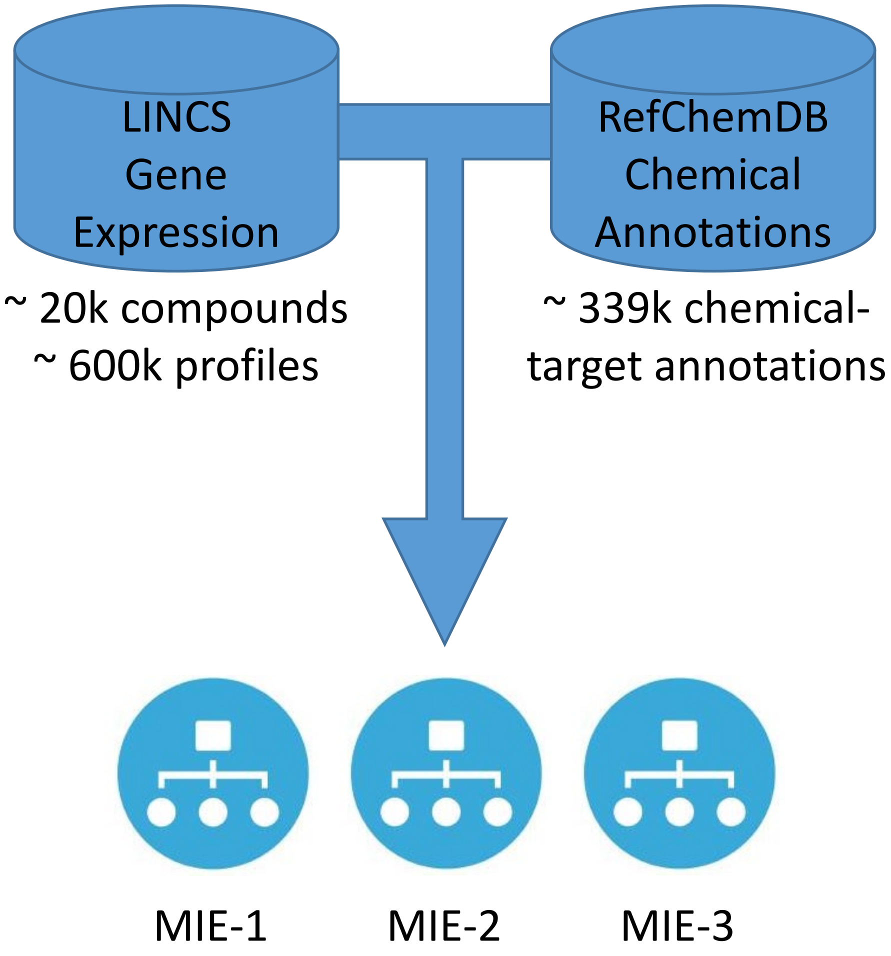
Figure 4, Allen et al. 2014

- Molecular Initiating Events (MIEs) are a concept in the Adverse Outcome Pathway (AOP) paradigm

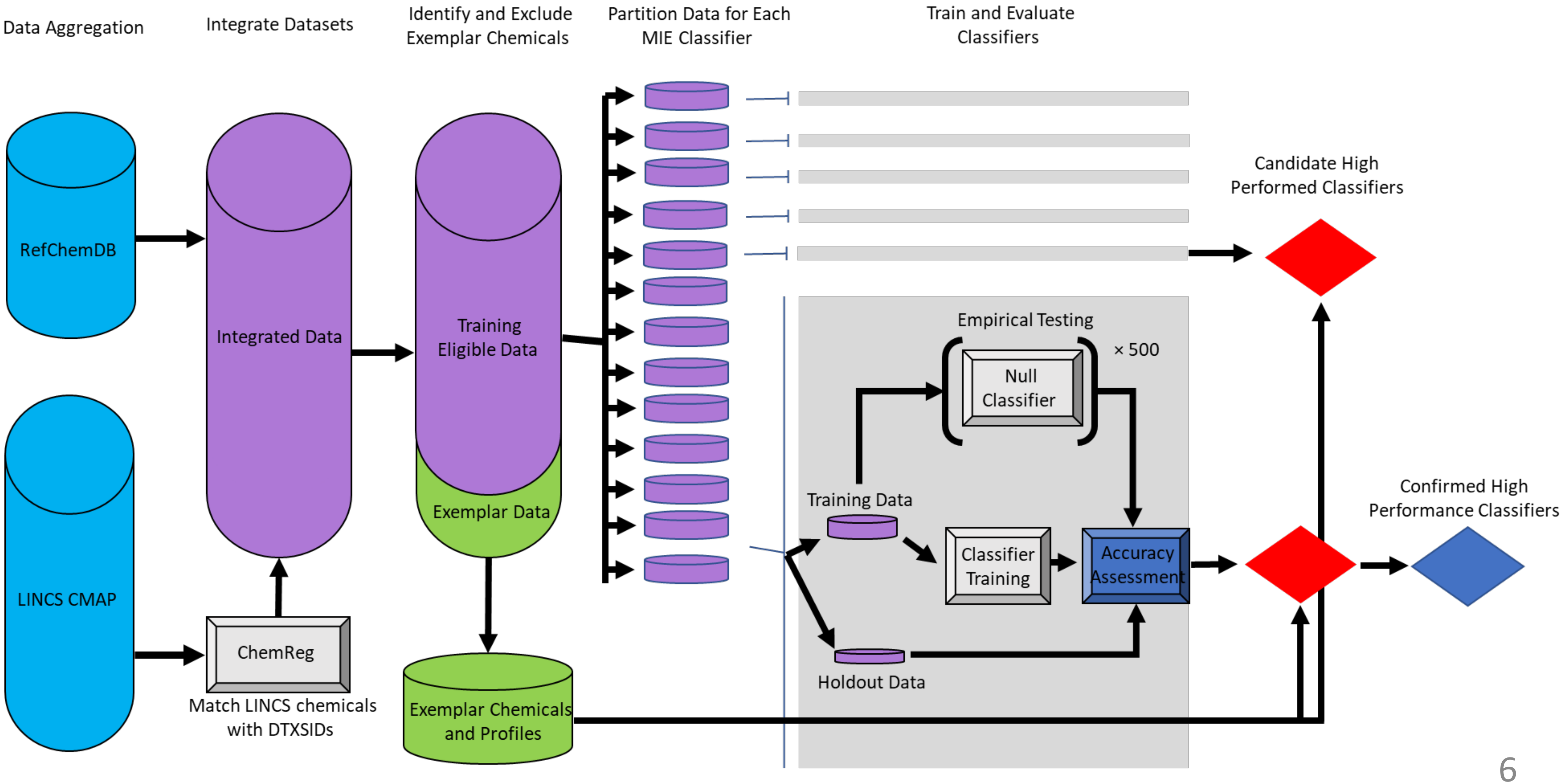
- MIEs are the initial molecular interactions between a chemical and a biological system that trigger downstream key events, culminating in an adverse outcome

Predicting MIEs from Gene Expression Data

- Integrate publicly available gene expression data with a database that links reference chemicals to molecular targets
- Train binary classifiers to predict activation of MIEs by chemical treatment
- Train a separate classifier for each MIE using machine learning



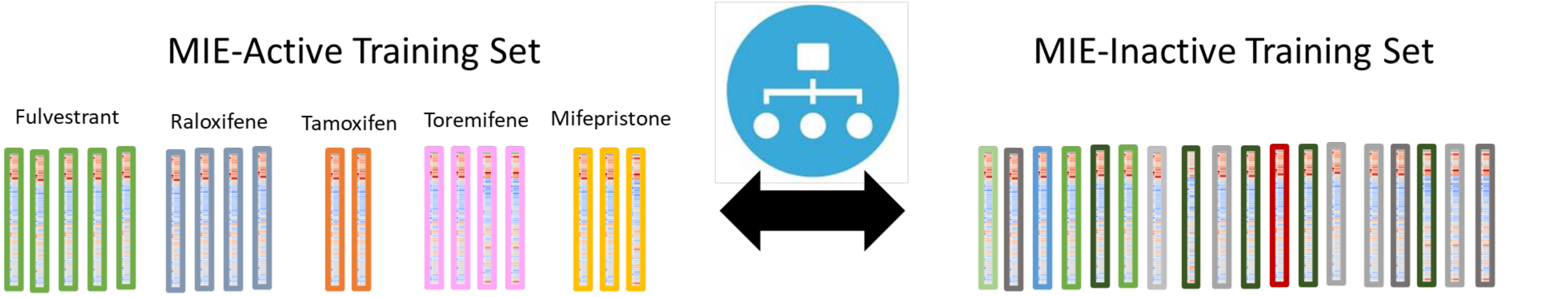
Data Processing and Classifier Training Workflow



Data Aggregation
Integrate Datasets
Identify and Exclude Exemplar Chemicals
Partition Data for Each MIE Classifier
Train and Evaluate Classifiers
RefChemDB
LINCS CMAP
Integrated Data
ChemReg
Match LINCS chemicals with DTXSIDs
Training Eligible Data
Exemplar Data
Exemplar Chemicals and Profiles
Candidate High Performed Classifiers
Empirical Testing
× 500
Null Classifier
Training Data
Classifier Training
Accuracy Assessment
Holdout Data
Confirmed High Performance Classifiers

Example of Classifier Training Data Set

Estrogen Receptor Inhibition
ESR-1/2 (-)



MIE-Active Training Set

Collection of MIE-associated chemicals and their profiles

MIE-Inactive Training Set

Collection of profiles selected at random from a large set of chemicals that are not associated with the MIE

MIE Classifier Training Parameters

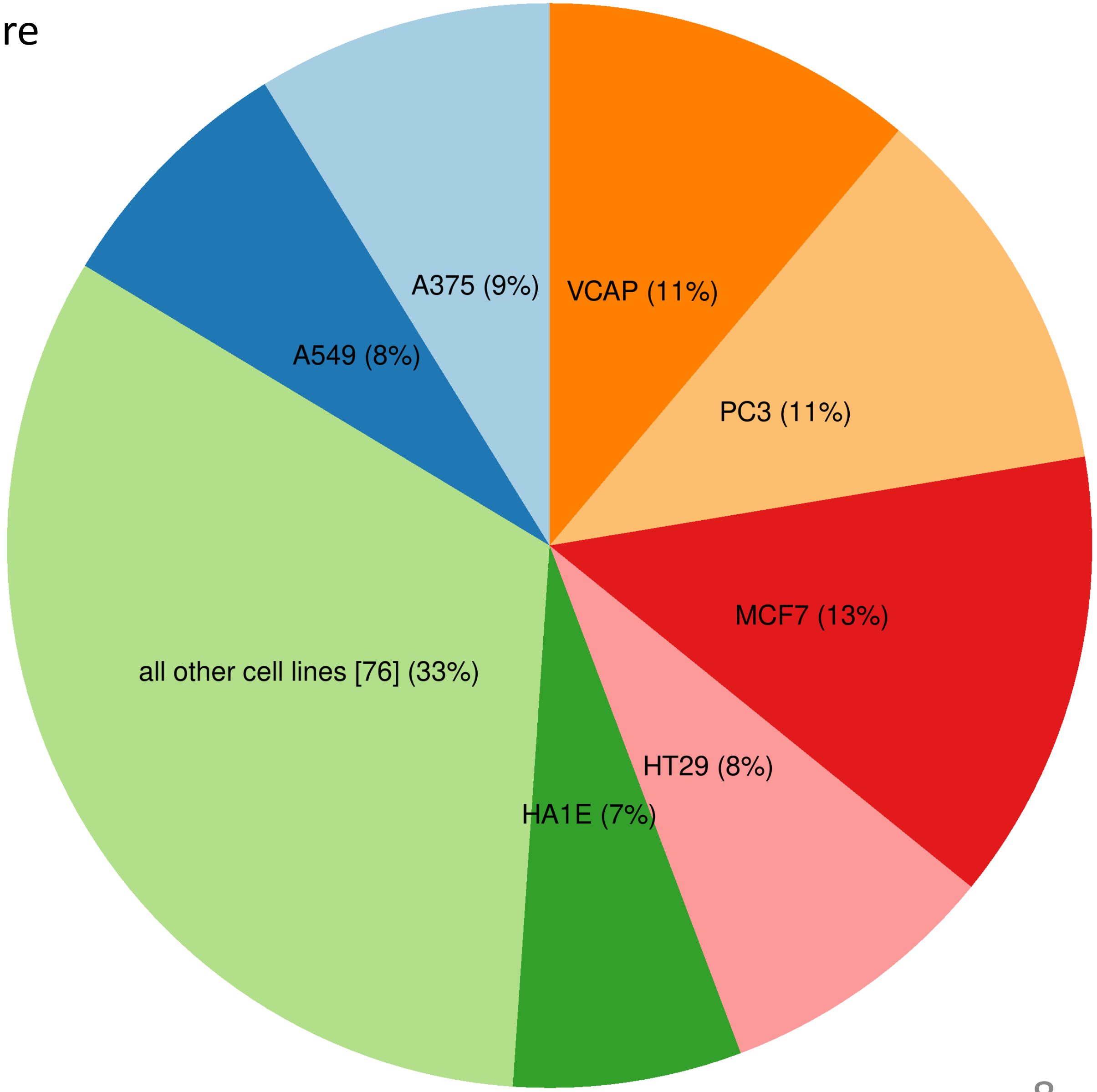
In MCF7 data, 51 MIEs with sufficient training data were identified

- Valid MIEs must be linked to at least 5 Chemicals
- Valid MIEs must be linked to at least 50 Gene Expression Profiles

Model optimization variables:

- Training Feature Type
  1. Landmark Genes
  2. All Genes
  3. Pathway Scores

- Classifiers trained with 6 algorithms
  1. Support Vector Machine Linear
  2. Support Vector Machine Polynomial
  3. Support Vector Machine Radial
  4. K-Nearest Neighbor
  5. Multilayer Perceptron
  6. Naïve Bayes



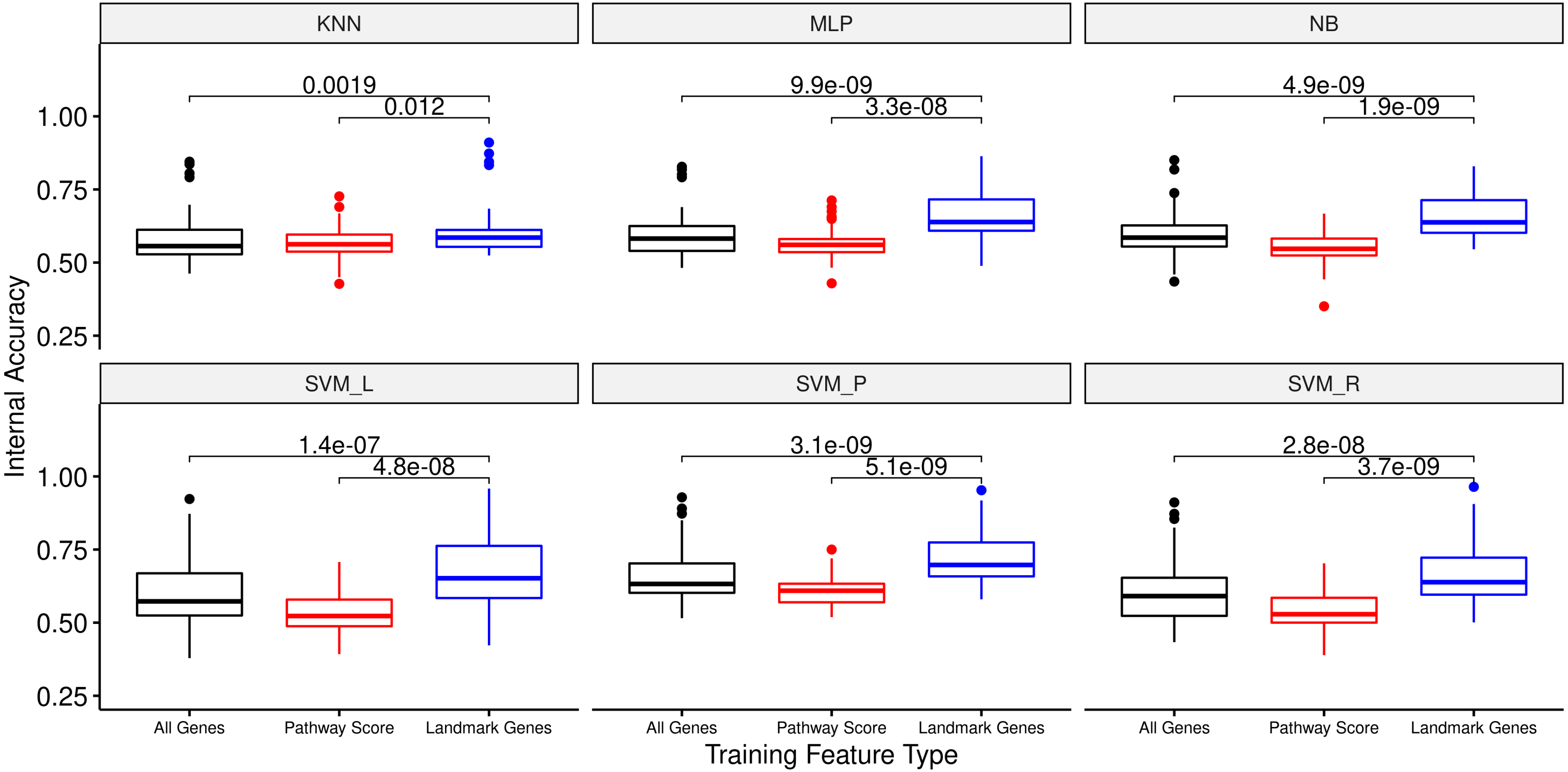
Comparison of Training Feature Types

- Classifiers trained on landmark genes perform better than classifiers trained on pathway score or landmark + inferred genes (all genes)

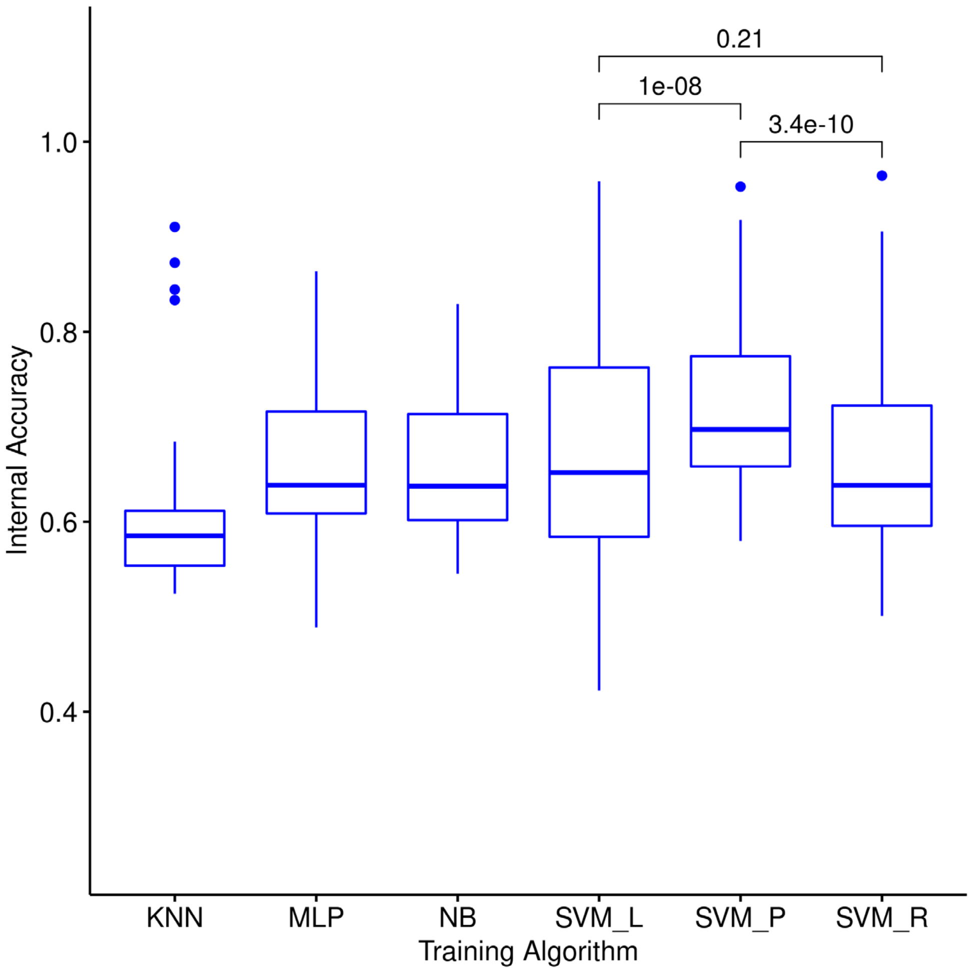
All genes
978 landmark genes + 11,350 inferred genes

Pathway scores
~900 Pathway scores

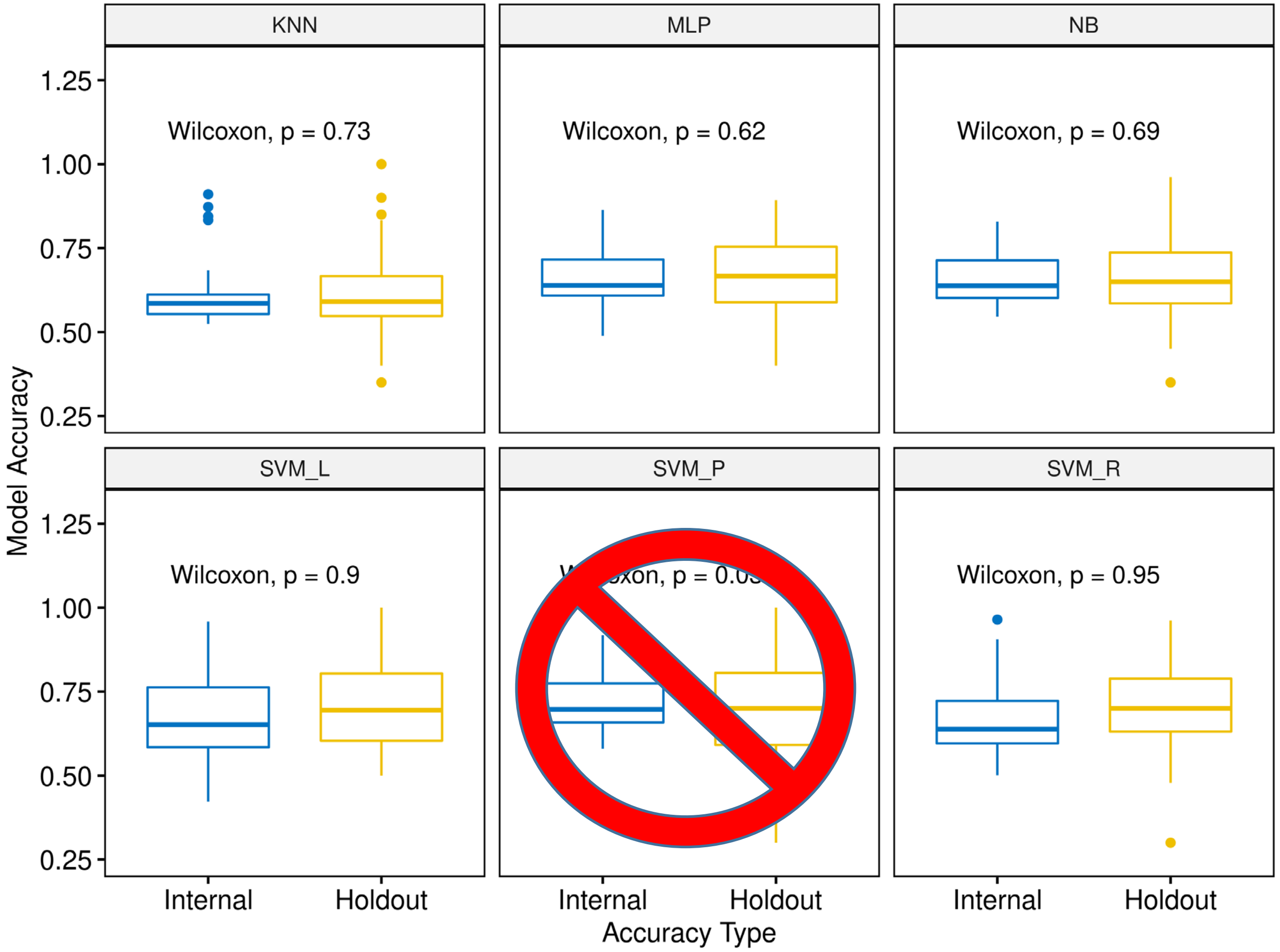
Landmark genes
978 genes measured in L1000 assay



Comparison of Classifier Algorithm Performance

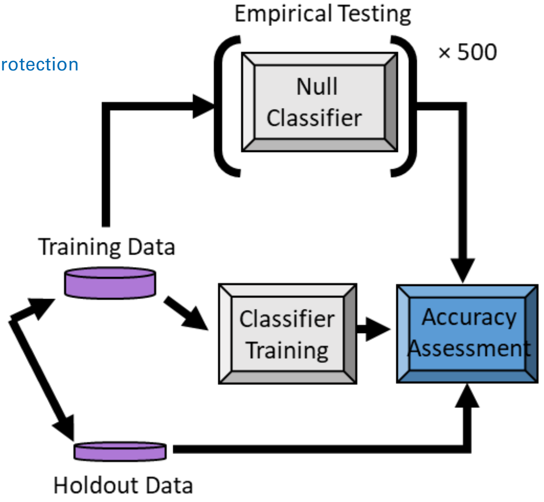


Support Vector Machine algorithm with a polynomial kernel produced the highest internal accuracy



Comparison of internal and Holdout accuracies revealed that SVM_P based classifiers were likely overfit

Empirical Significance Analysis


Empirical Testing

× 500

Null Classifier

Training Data

Classifier Training

Accuracy Assessment

Holdout Data

Train multiple “null” classifiers by permuting chemical-MIE associations

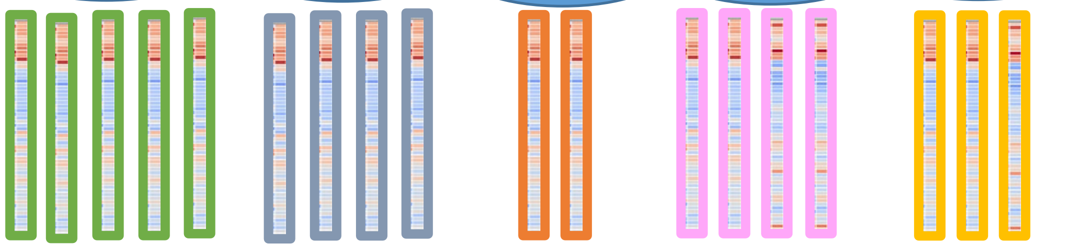
Estrogen Receptor Inhibition
ESR-1/2 (-)

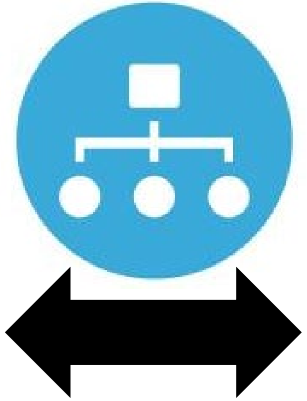
×500

MIE-Active Training Set

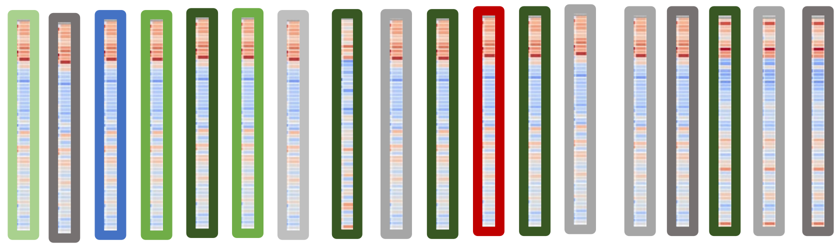
GW-9508 torcetrapib lapatinib sildenafil NVP-TAE684

~~Fulvestrant~~ ~~Raloxifene~~ ~~Tamoxifen~~ ~~Toremifene~~ ~~Mifepristone~~





MIE-Inactive Training Set

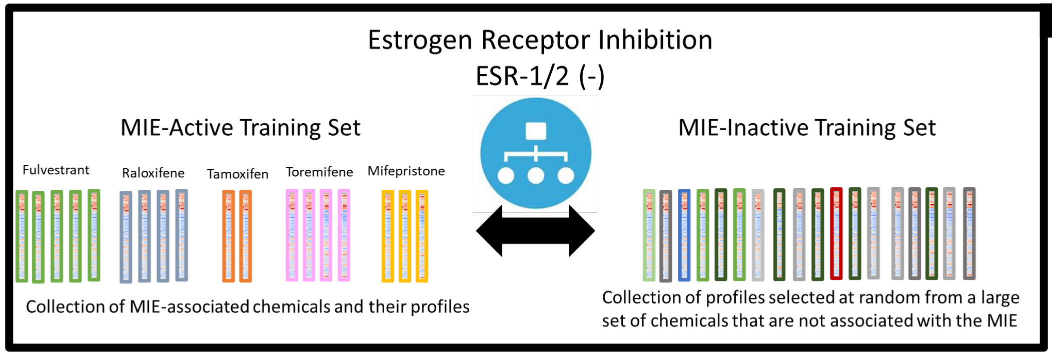


Collection of MIE-associated chemicals and their profiles

Collection of profiles selected at random from a large set of chemicals that are not associated with the MIE

Empirical Significance Analysis

“TRUE” Model



Estrogen Receptor Inhibition
ESR-1/2 (-)

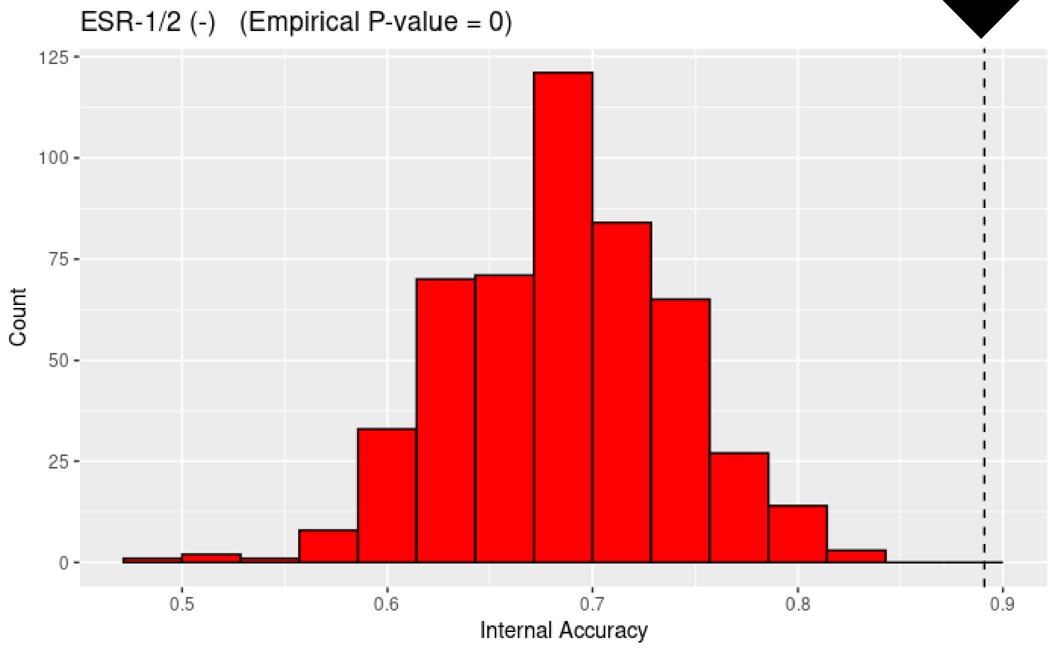
MIE-Active Training Set

Collection of MIE-associated chemicals and their profiles

MIE-Inactive Training Set

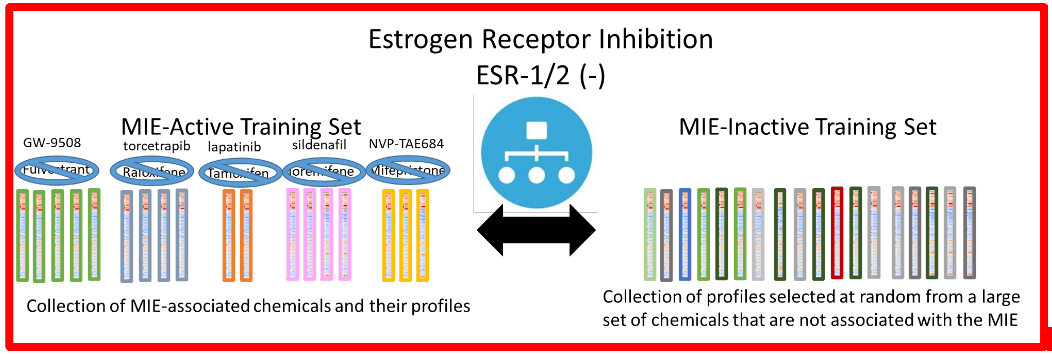
Collection of profiles selected at random from a large set of chemicals that are not associated with the MIE

“TRUE” internal accuracy



ESR-1/2 (-) (Empirical P-value = 0)

“Null” Models



Estrogen Receptor Inhibition
ESR-1/2 (-)

MIE-Active Training Set

Collection of MIE-associated chemicals and their profiles

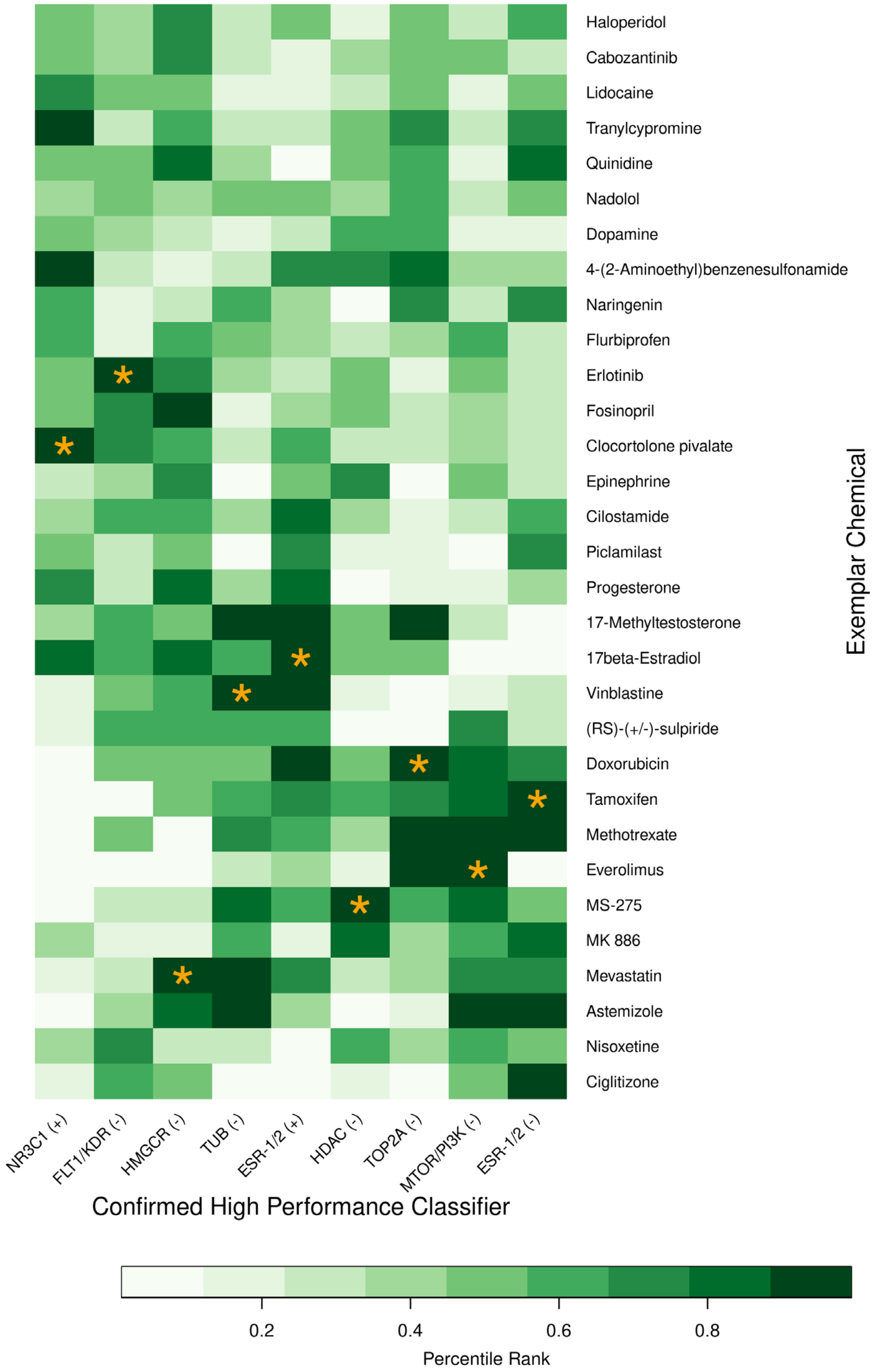
MIE-Inactive Training Set

Collection of profiles selected at random from a large set of chemicals that are not associated with the MIE

500 “Null” internal accuracies

- Calculate percentile rank of “true” un-permuted model accuracy relative to accuracy scores of the 500 permuted models

MIE Name	Classification Algorithm	Internal Accuracy	Holdout Accuracy	MIE Active Profiles	MIE Active Chemicals	Mean Null Accuracy	Empirical Pvalue	Exemplar Chemical	Exemplar Percent Rank
ADRA2A (+)	SVM_R	0.72	0.86	58	7	0.60	0.03	Epinephrine	0.76
ALOX5 (-)	NB	0.73	0.55	51	5	0.61	0.01	MK 886	0.63
AR (+)	NB	0.71	0.60	52	8	0.61	0.03	17-Methyltestosterone	0.12
DRD2 (-)	SVM_R	0.68	0.54	118	14	0.58	0.03	Haloperidol	0.74
ESR-1/2 (-)	MLP	0.89	0.92	68	5	0.69	0.00	**Tamoxifen**	**1.00**
ESR-1/2 (+)	SVM_L	0.85	0.79	145	12	0.64	0.00	**17beta-Estradiol**	**0.96**
FLT1/KDR (-)	MLP	0.75	0.69	122	10	0.66	0.02	**Erlotinib**	**0.90**
HDAC (-)	SVM_L	0.82	0.78	174	10	0.67	0.00	**MS-275**	**0.97**
HMGCR (-)	MLP	0.79	0.85	50	4	0.66	0.03	**Mevastatin**	**0.92**
HRH1 (-)	MLP	0.71	0.61	110	14	0.61	0.01	Astemizole	0.24
JAK2 (-)	SVM_L	0.88	0.85	54	5	0.71	0.01	NA	NA
KCNH2 (-)	SVM_R	0.66	0.64	369	34	0.58	0.00	Haloperidol	0.70
MAPK14 (-)	SVM_L	0.86	0.93	78	5	0.73	0.03	NA	NA
MET (-)	SVM_L	0.83	0.70	114	7	0.70	0.01	Cabozantinib	0.54
MTOR/PIK3 (-)	SVM_R	0.90	0.88	204	12	0.70	0.00	**Everolimus**	**1.00**
NR3C1 (+)	SVM_R	0.73	0.68	100	10	0.60	0.01	**Clocortolone pivalate**	**0.97**
PTGS-1/2 (-)	SVM_R	0.65	0.65	247	28	0.58	0.00	Flurbiprofen	0.59
SLC22A6 (-)	KNN	0.70	0.64	55	6	0.58	0.02	Methotrexate	0.26
TOP2A (-)	SVM_L	0.88	0.87	75	7	0.67	0.00	**Doxorubicin**	**1.00**
TUB (-)	SVM_L	0.94	0.90	104	8	0.59	0.00	**Vinblastine**	**1.00**



Exemplar Chemical Predictions for 9 High Performance Classifiers

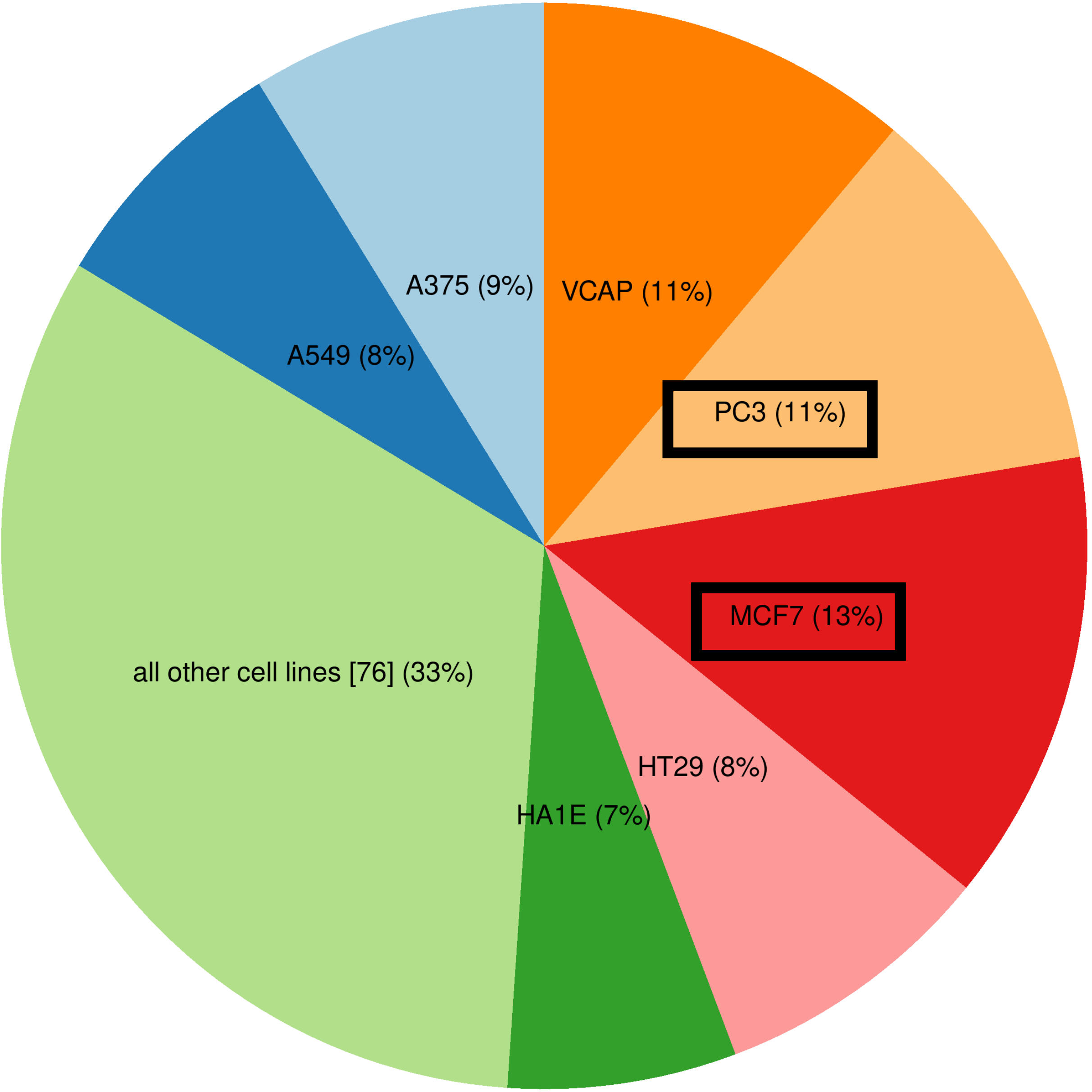
- High performance classifiers generated high ranking predictions for their respective training-excluded exemplar reference chemicals

- A subset of exemplar chemicals returned high ranking predictions for MIEs for which they are not annotated (Methotrexate and ESR-1/2 (-), MTOR/PI3K (-))
  - Likely the result of molecular cross-talk and the convergence of signaling pathways shared between MIEs

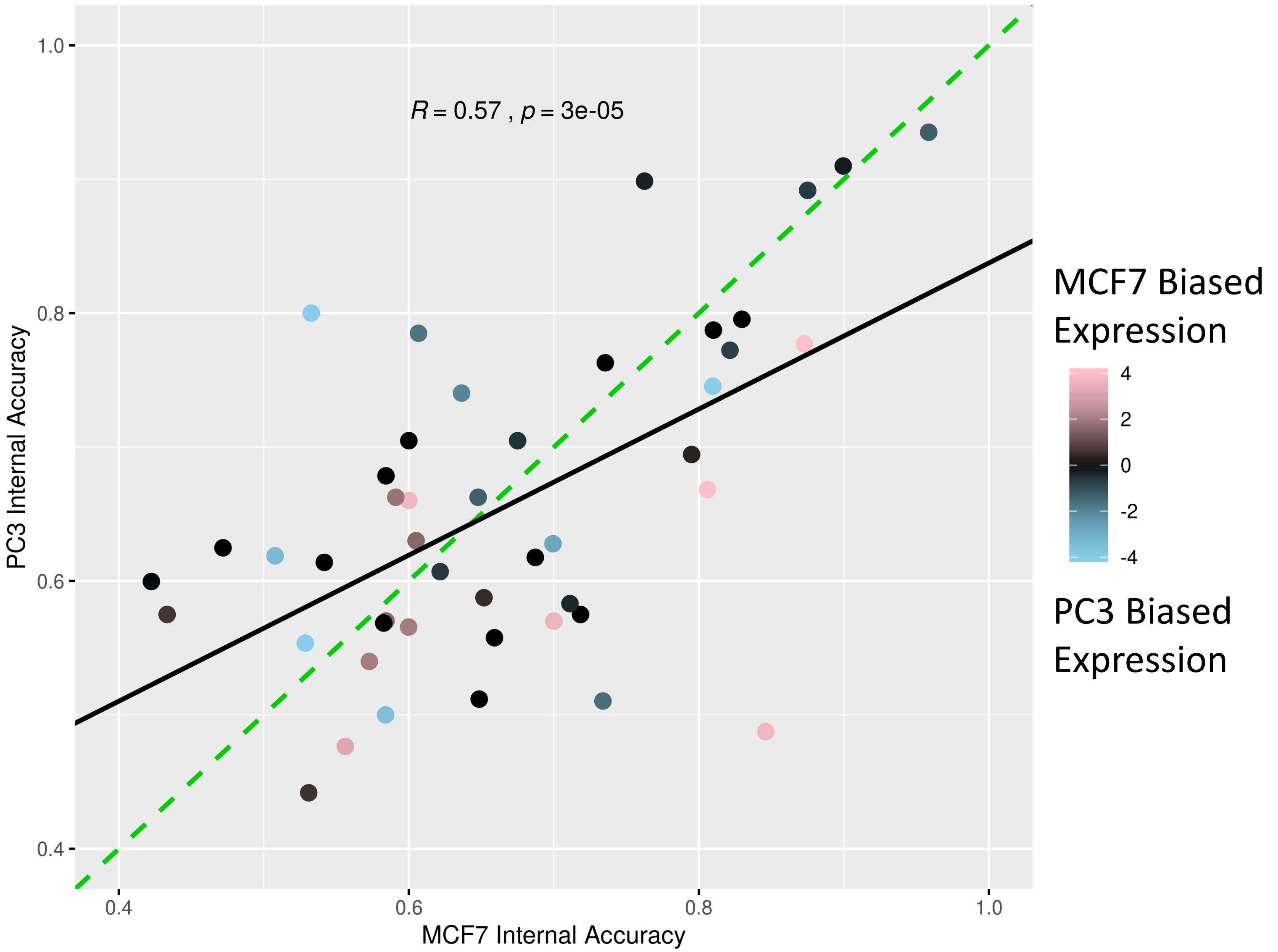
How does MIE Classifier Performance Vary Across Cell Lines?

- Trained a second set of MIE classifiers on PC3-derived data (prostate cancer cell line)
  - PC3 cell line has the second most gene expression profiles in LINCS L1000 CMAP dataset

- PC3 classifiers were trained for 46 of the 51 MIEs modeled in the MCF7 cell line



Comparison of Internal Accuracies for MCF7 and PC3-trained Classifiers



- Modest correlation between internal accuracies of MCF7 and PC3 trained classifiers
- Some variation in internal accuracy likely attributable to differences in baseline expression of MIE gene targets
  - Gene expression values derived from human protein atlas
  - MIEs may be more readily triggered (and better modeled) in cell types where the associated target protein is highly expressed

Conclusions

- Trained predictive models for 51 distinct MIEs by integrating gene expression data with chemical-target labels
  - Identified 9 MIEs modeled with high performance classifiers
- Explored factors that affected model accuracy
  - Feature type
  - Classification algorithm
- Trained classifiers using profiles from different cell types (MCF7 and PC3)
- Identified several MIEs that are well-modeled in both cell types
  - A subset of classifiers showed a disparity in performance as a function of cell type and shed light on MIEs that may be better screened in one cell type over another

Acknowledgements

US Environmental Protection Agency

Office of Research and Development

Center for Computational Biology and Exposure

Biomolecular and Computation Toxicology Division

Computation Toxicology and Bioinformatic Branch

Antony Williams

Chris Grulke

Logan Everett

Richard Judson

Imran Shah

Signature Index	Chemical Treatment	MIE 1 Prediction	MIE 2 Prediction	MIE 3 Prediction
1	Haloperidol	0.05	0.77	0.42
2	Haloperidol	0.25	0.62	0.23
3	Haloperidol	0.13	0.55	0.26
4	Everolimus	0.88	0.33	0.42
5	Everolimus	0.74	0.18	0.23
6	Everolimus	0.90	0.44	0.32
7	Dopamine	0.23	0.43	0.98
8	Dopamine	0.27	0.21	0.76
… 42,049	…	…	…	…

Distill per-signature predictions into per-chemical predictions by taking the median

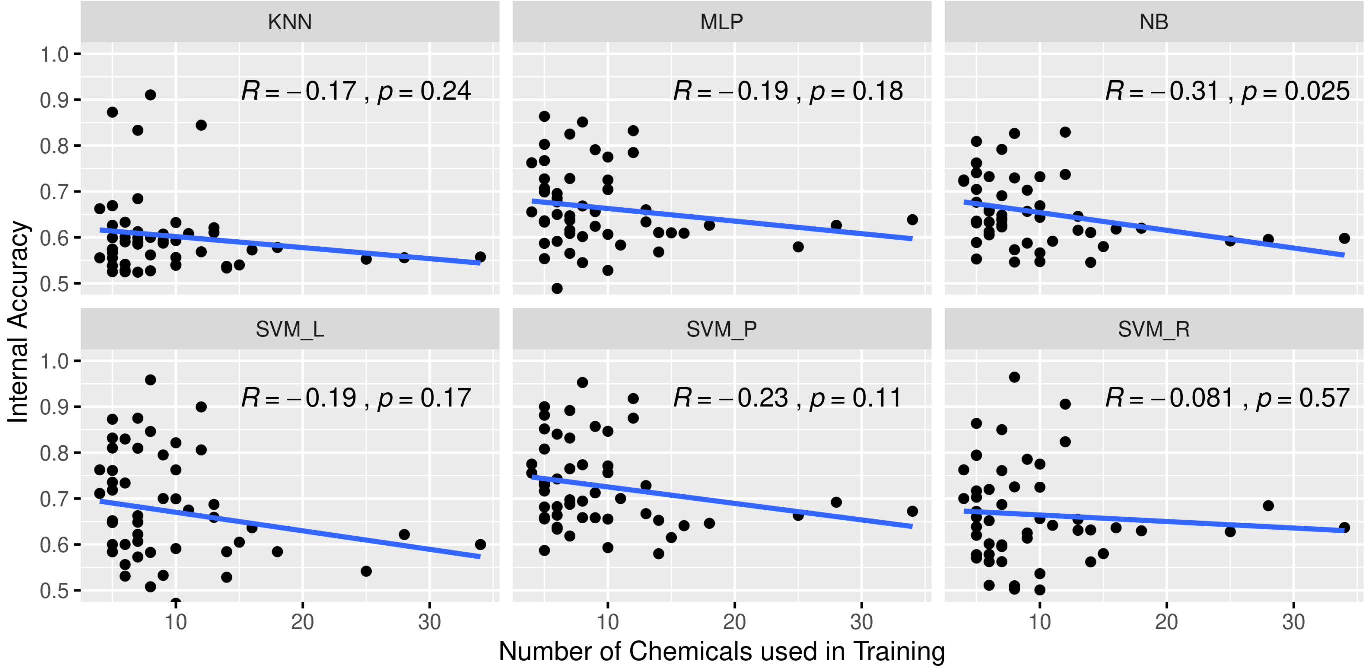
Chemical Treatment	MIE 1 Prediction	MIE 2 Prediction	MIE 3 Prediction
Haloperidol	0.13	0.62	0.26
Everolimus	0.74	0.33	0.32
Dopamine	0.25	0.32	0.87
… (11,712)	…	…	…

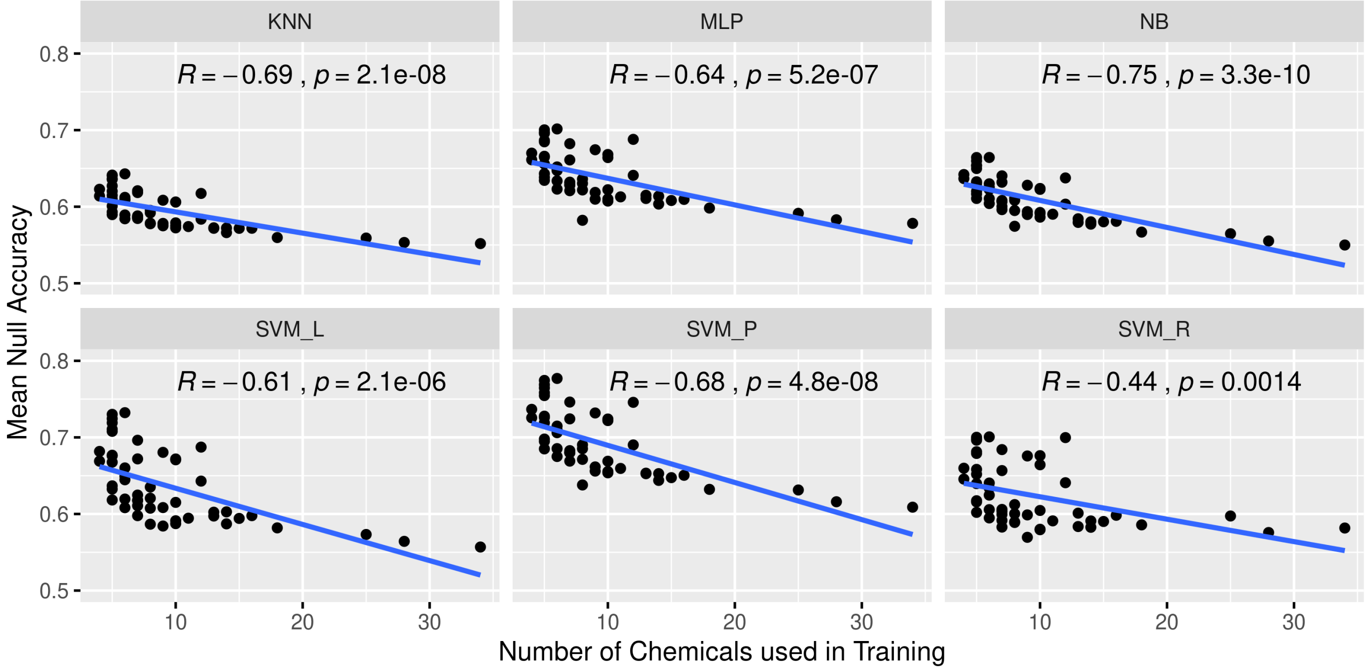
Calculate the MIE-wise rank for each chemical

Chemical Treatment	MIE 1 Prediction	MIE 2 Prediction	MIE 3 Prediction
Haloperidol	6,239/11,712	963/11,712	9,842/11,712
Everolimus	354/11,712	9,426/11,712	9,436/11,712
Dopamine	1453/11,712	9,448/11,712	173/11,712
… (11,712)	…	…	…

Calculate the percentile rank for each chemical

Chemical Treatment	MIE 1 Prediction	MIE 2 Prediction	MIE 3 Prediction
Haloperidol	0.47	0.92	0.16
Everolimus	0.97	0.20	0.19
Dopamine	0.88	0.19	0.99
… (11,712)	…	…	…


KNN
MLP
NB
SVM_L
SVM_P
SVM_R
$R = -0.17$, $p = 0.24$
$R = -0.19$, $p = 0.18$
$R = -0.31$, $p = 0.025$
$R = -0.19$, $p = 0.17$
$R = -0.23$, $p = 0.11$
$R = -0.081$, $p = 0.57$
Internal Accuracy
1.0
0.9
0.8
0.7
0.6
0.5
10
20
30
Number of Chemicals used in Training


KNN
MLP
NB
SVM_L
SVM_P
SVM_R
$R = -0.69$, $p = 2.1e-08$
$R = -0.64$, $p = 5.2e-07$
$R = -0.75$, $p = 3.3e-10$
$R = -0.61$, $p = 2.1e-06$
$R = -0.68$, $p = 4.8e-08$
$R = -0.44$, $p = 0.0014$
Mean Null Accuracy
0.8
0.7
0.6
0.5
10
20
30
Number of Chemicals used in Training