Predicting Molecular Initiating Events from Gene Expression 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 Background

A Challenge in Chemical Hazard Identification:

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

A Solution:

Screen thousands of chemicals with high throughput transcriptomics to survey transcriptional perturbation

Train binary classifiers to predict MIE activation using gene expression data from high confidence reference chemicals

Use classifiers to predict MIE activation for data-poor chemicals and flag candidates for screening with targeted tier 2 assays

What Are Molecular Initiating Events (MIEs)?

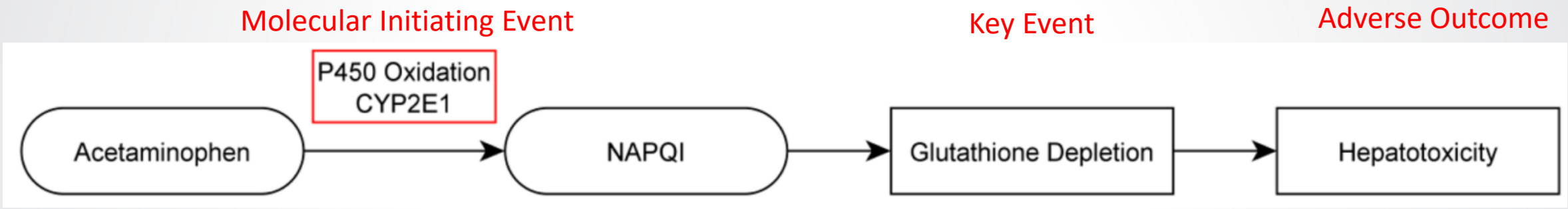
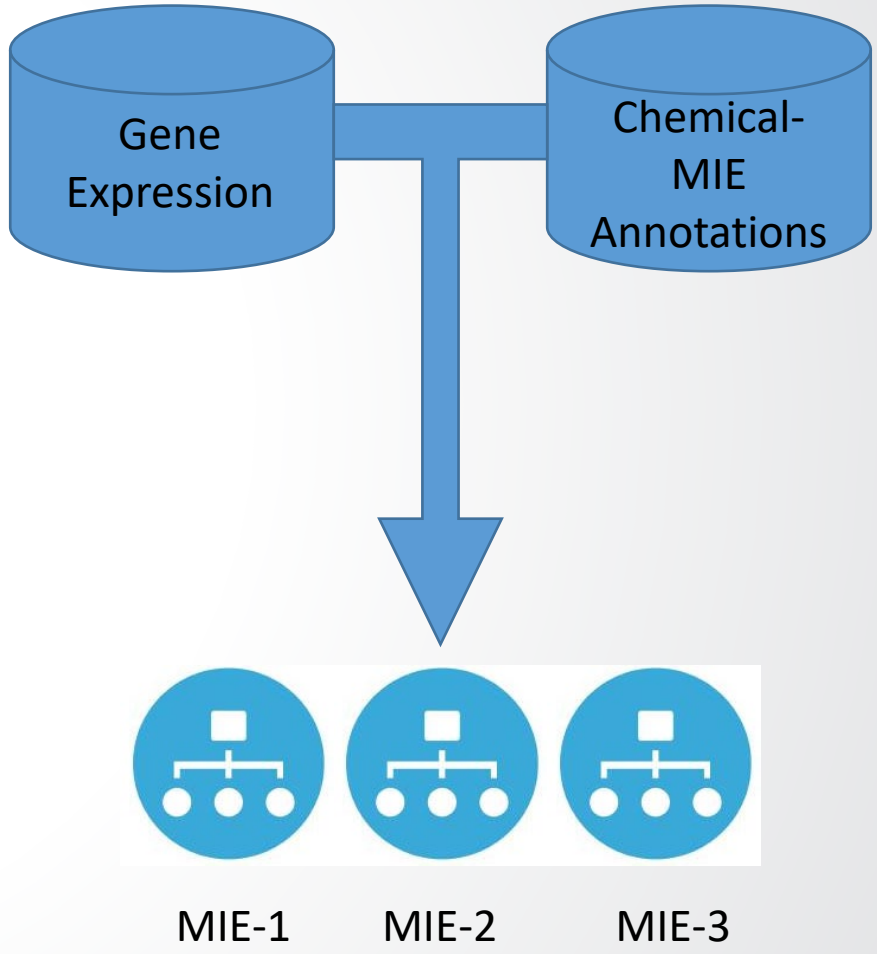


Figure 4, Allen et al. 2014

- MIEs are a concept presented 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
- For these methods, MIEs are defined as a gene (or family of genes that share similar chemical modulators) and a mode of modulation (activation or inhibition)

Predicting MIEs from Gene Expression Data

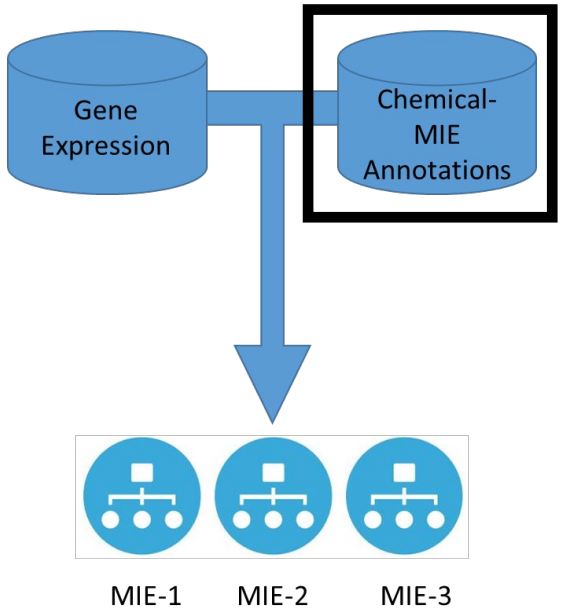
- Integrate publicly available gene expression data with a database that links reference chemicals to molecular targets
- Train separate binary classifiers for each MIE to predict whether chemicals activate the MIE being modeled
- Train a family of MIE classifiers to predict activation of a spectrum of possible MIEs



Data sources

RefChemDB

- Database of chemical-protein interactions compiled from multiple sources
- Contains “support” field indicative of how many different sources evidence the chemical-protein relationship
- Contains “mode” field indicating the nature of the interaction (activator, inhibitor, unspecified interaction)
- ~330K total annotations
- ~31K unique DTXSIDs
- ~9.5K unique MIEs



Research Article

Workflow for Defining Reference Chemicals for Assessing Performance of *In Vitro* Assays

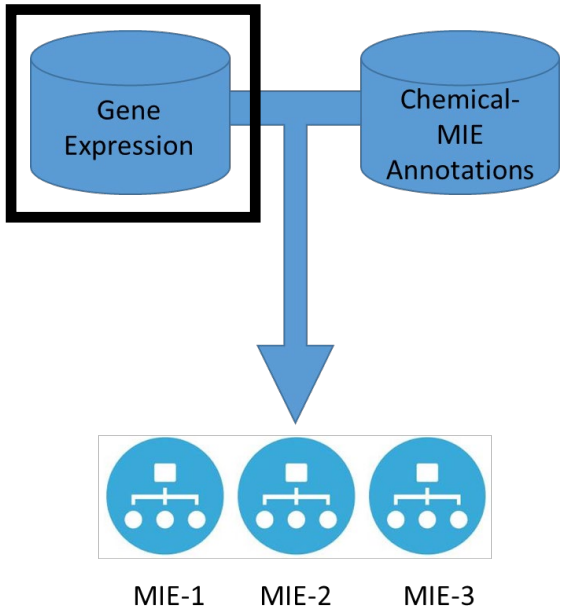
Richard S. Judson[1], Russell S. Thomas[1], Nancy Baker[2], Anita Simha[3], Xia Meng Howey[3], Carmen Marable[3], Nicole C. Kleinstreuer[4] and Keith A. Houck[1]

[1]US EPA, National Center for Computational Toxicology, Research Triangle Park, NC, USA; [2]Leidos, Inc., Research Triangle Park, NC, USA; [3]ORAU, contractor to U.S. Environmental Protection Agency through the National Student Services Contract, Research Triangle Park, NC, USA; [4]National Toxicology Program, Interagency Center for the Evaluation of Alternative Toxicological Methods, Research Triangle Park, NC, USA

Data sources

LINCS L1000 CMAP Data

- “The LINCS L1000 project has collected gene expression profiles for thousands of perturbagens at a variety of time points, doses, and cell lines. ”
- **L**ibrary of **I**ntegrated **N**etwork-based **C**ellular **S**ignatures
- Gene expression for 978 “landmark” transcripts is measured directly. Expression values for an additional 11,350 genes are inferred
- ~300K gene expression profiles
- ~20K unique chemical treatments



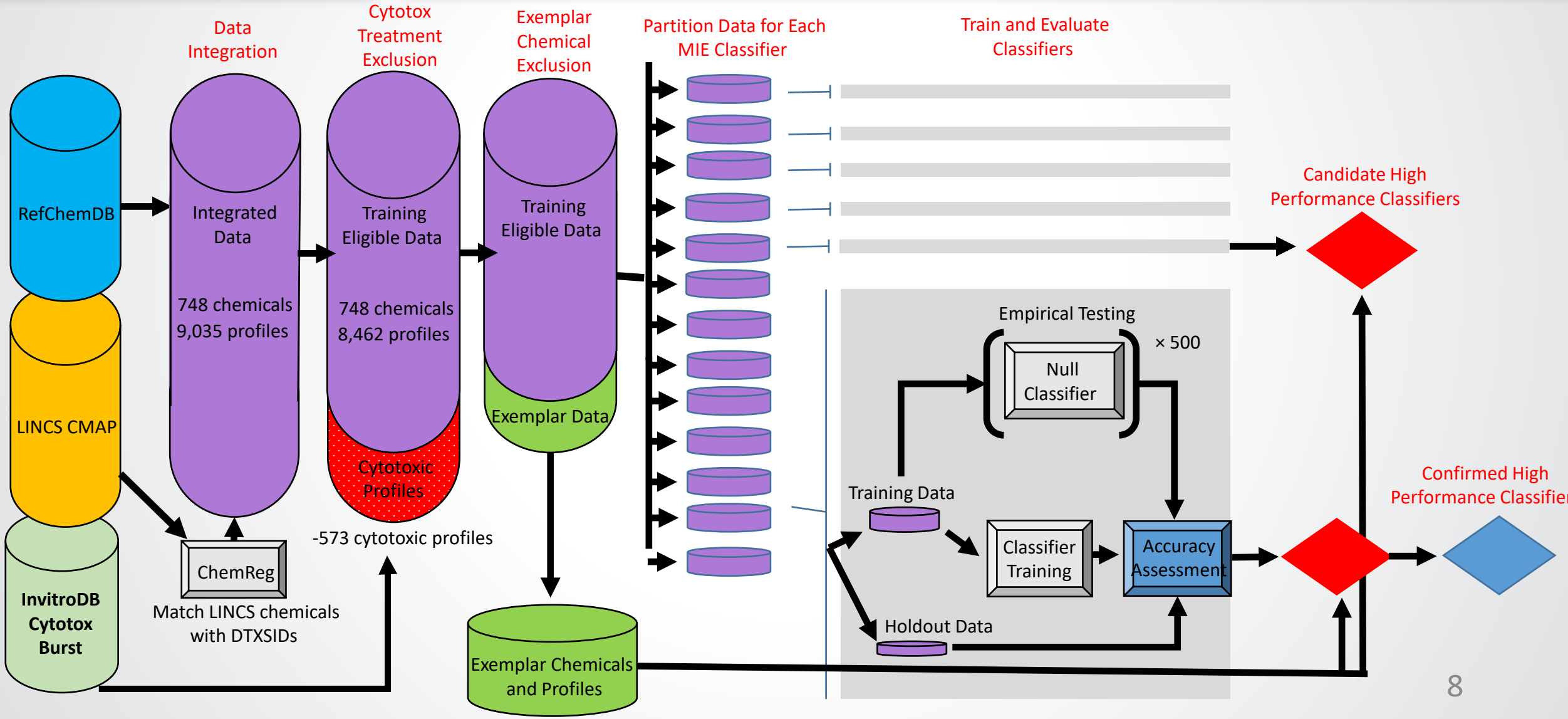
Resource

Cell

A Next Generation Connectivity Map: L1000 Platform and the First 1,000,000 Profiles

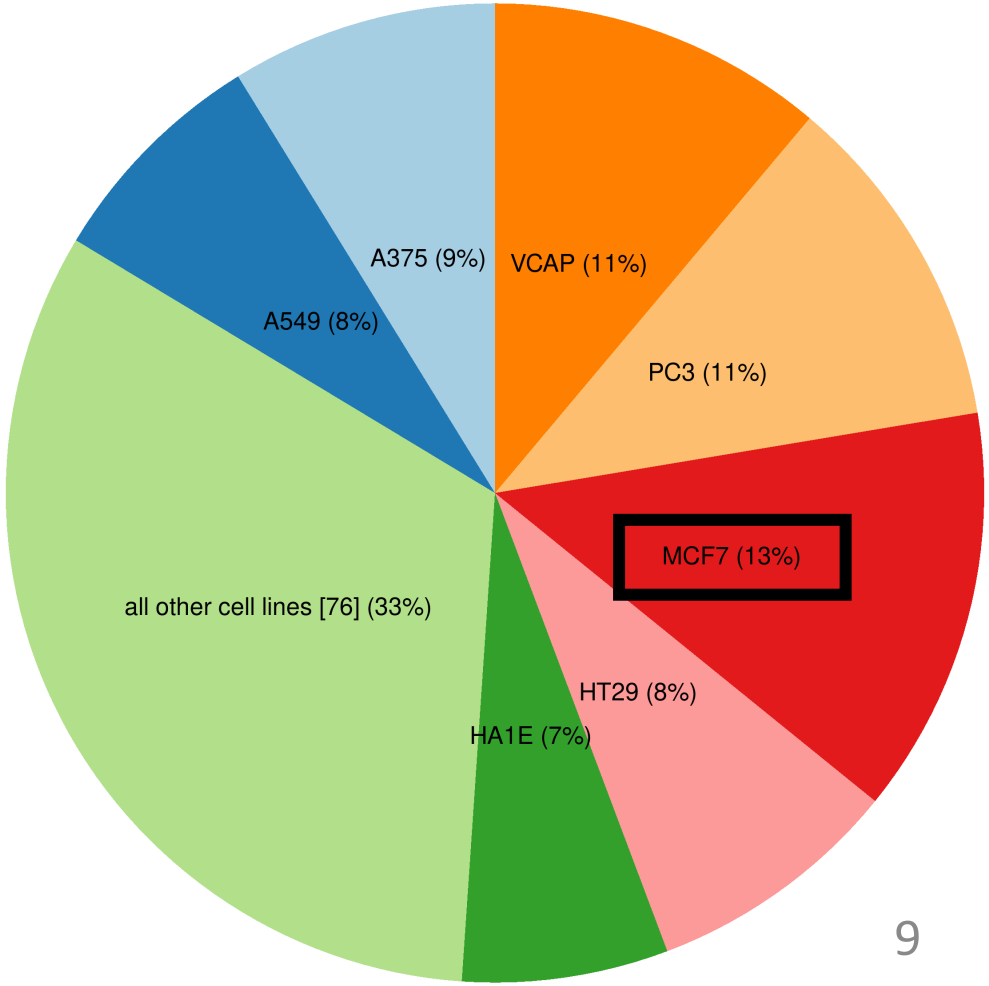
Aravind Subramanian,[1,9] Rajiv Narayan,[1,9] Steven M. Corsello,[1,2,3,9] David D. Peck,[1] Ted E. Natoli,[1] Xiaodong Lu,[1] Joshua Gould,[1] John F. Davis,[1] Andrew A. Tubelli,[1] Jacob K. Asiedu,[1] David L. Lahr,[1] Jodi E. Hirschman,[1] Zihan Liu,[1] Melanie Donahue,[1] Bina Julian,[1] Mariya Khan,[1] David Wadden,[1] Ian C. Smith,[1] Daniel Lam,[1] Arthur Liberzon,[1] Courtney Toder,[1] Mukta Bagul,[1] Marek Orzechowski,[1] Oana M. Enache,[1] Federica Piccioni,[1] Sarah A. Johnson,[1] Nicholas J. Lyons,[1] Alice H. Berger,[1,2,3,10] Alykhan F. Shamji,[1] Angela N. Brooks,[1,2,3,11] Anita Vrcic,[1] Corey Flynn,[1] Jacqueline Rosains,[1,12] David Y. Takeda,[1,2,3] Roger Hu,[1] Desiree Davison,[1] Justin Lamb,[1,13] Kristin Ardlie,[1] Larson Hogstrom,[1] Peyton Greenside,[1,15] Nathanael S. Gray,[1,3,4] Paul A. Clemons,[1] Serena Silver,[1] Xiaoyun Wu,[1] Wen-Ning Zhao,[1,3,5] Willis Read-Button,[1,13] Xiaohua Wu,[1] Stephen J. Haggarty,[1,3,5] Lucienne V. Ronco,[1,14] Jesse S. Boehm,[1] Stuart L. Schreiber,[1,6,7] John G. Doench,[1] Joshua A. Bittker,[1] David E. Root,[1] Bang Wong,[1] and Todd R. Golub[1,3,7,8,16,*]

Data Processing and Classifier Training Workflow



Multiple Cell Lines Present in LINCS L1000 CMAP Data Set

- There are ~83 different cell lines annotated in LINCS metadata
- Comparing models trained different cell lines may identify cell lines more sensitive to specific MIEs
- Trained initial set of classifiers on gene expression profiles from the MCF7 cell line
  - Breast cancer derived
  - Largest number of gene expression profiles in LINCS



MIE Classifier Training Parameters

52 MIEs had sufficient data for training based on 2 criteria:

- Valid MIEs must be linked to at least 5 chemicals
- Valid MIEs must be linked to at least 50 gene expression profiles
- Limited amount of training data

MIE Name	# Chemicals	# Profiles
KCNH2_Negative	37	387
ABCB1_Negative	30	331
PTGS2_Negative_group	29	252
PIK3CA_Negative_group	16	236
HDAC1_Negative_group	12	206
SLC6A4_Negative_group	19	202
EGFR_Negative_group	12	181
CYP3A4_Negative	15	174
ESR1_Positive_group	13	165
KDR_Negative_group	12	158

Model optimization variables:

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

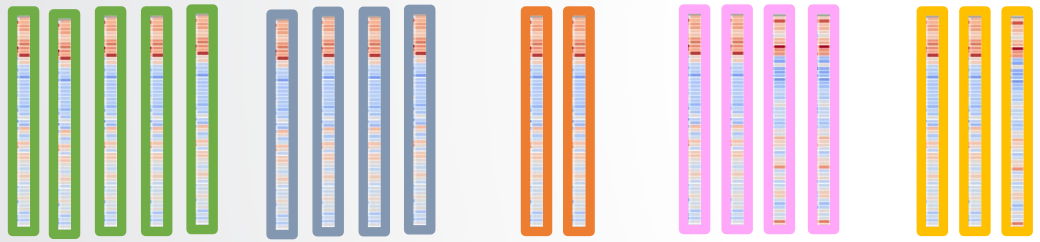
- Classifiers trained with 6 algorithms
  1. Support Vector Machine Linear
  2. Support Vector Machine Polynomial
  3. Support Vector Machine Radial
  4. K-Nearest Neighbor
  5. Multilayer Perceptron
  6. Random Forest

Example of MIE Specific Training Data Set

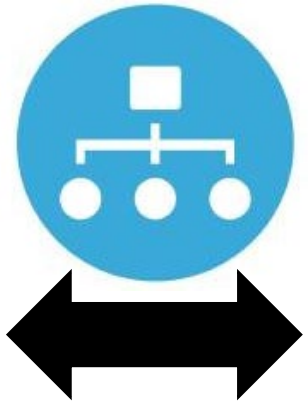
Estrogen Receptor Inhibition ESR (-)

MIE-Active Training Set

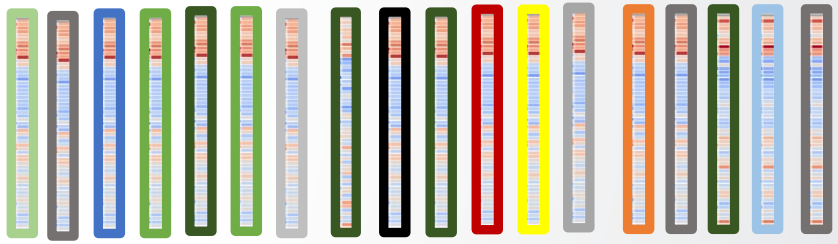
Fulvestrant Raloxifene Tamoxifen Toremifene Mifepristone



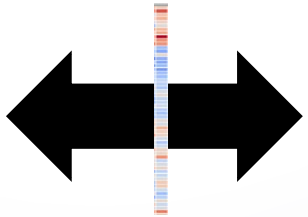
Collection of MIE-associated chemicals and their profiles



MIE-Inactive Training Set



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



MIE-Active or MIE-Inactive?

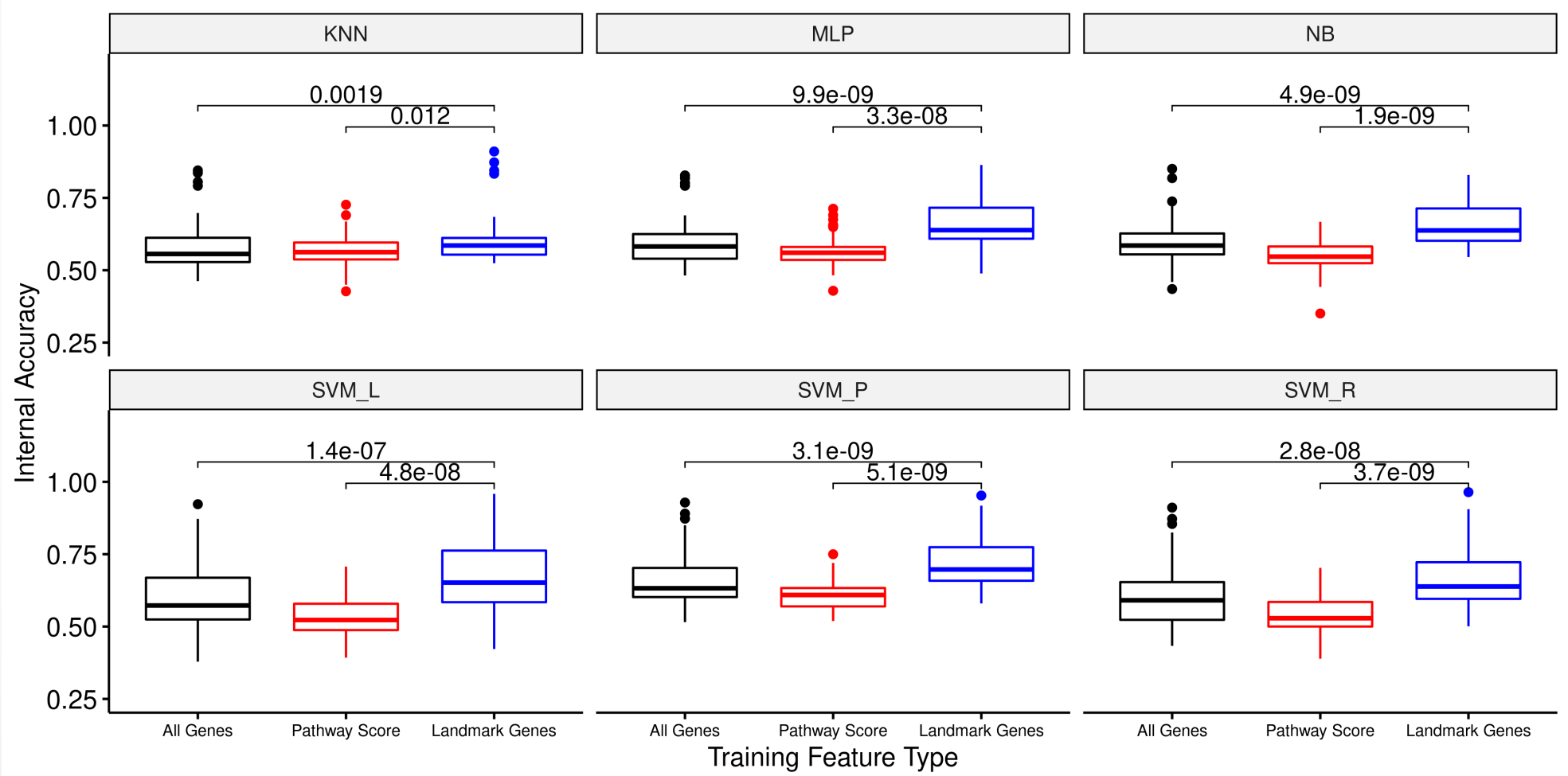
Comparison of Training Feature Types

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

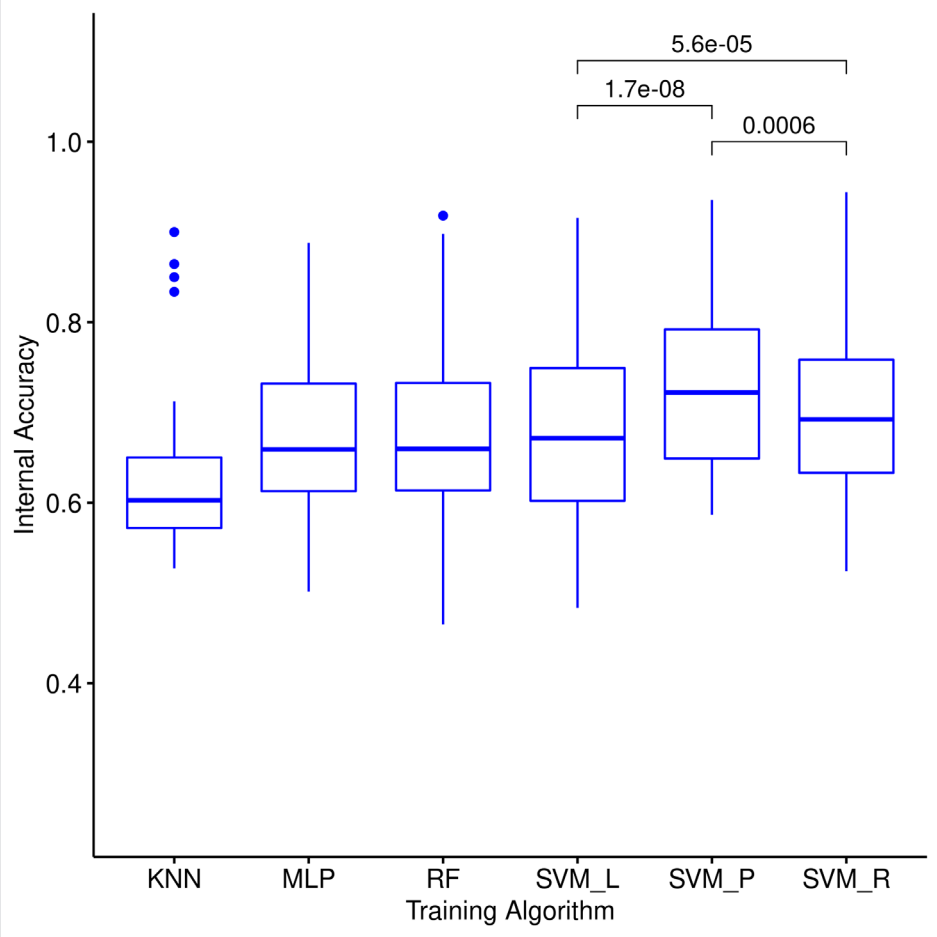
All genes
978 landmark genes + 11,350 inferred genes

Pathway scores
~1,000 Pathway scores

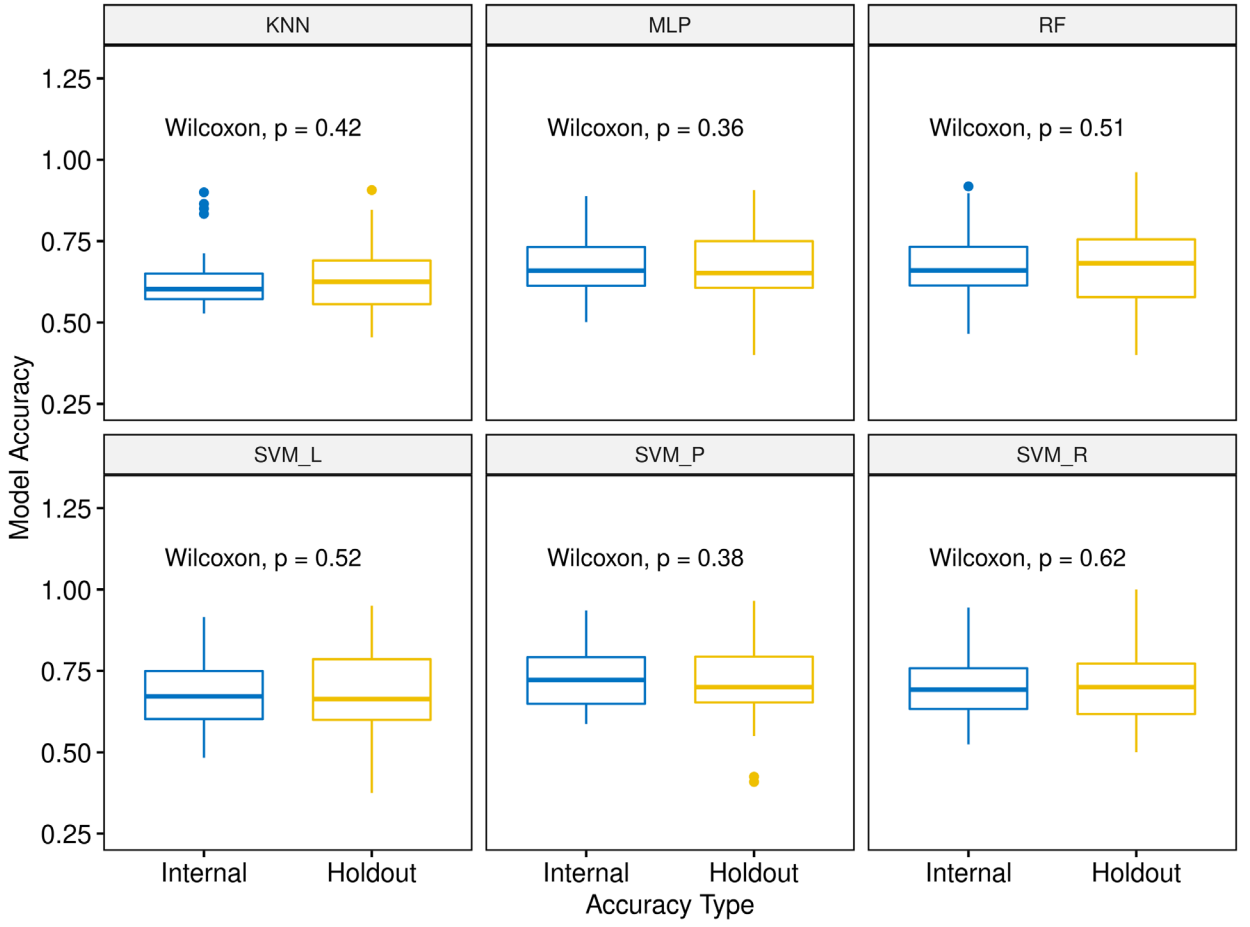
Landmark genes
978 genes measured in L1000 assay



Some Approaches Achieve Higher Accuracy than Others

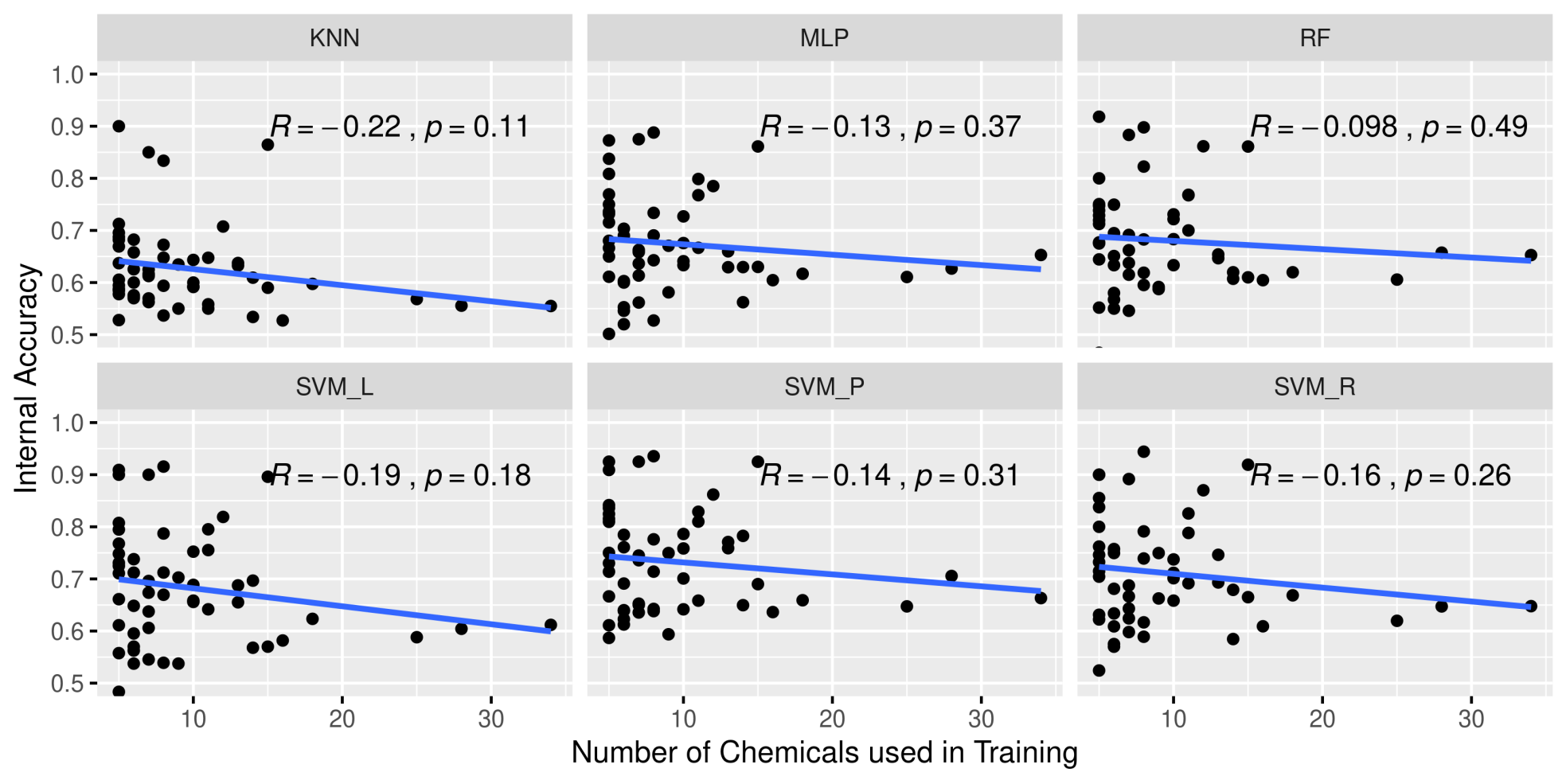


Support Vector Machine algorithm with a polynomial kernel produced the highest internal accuracy

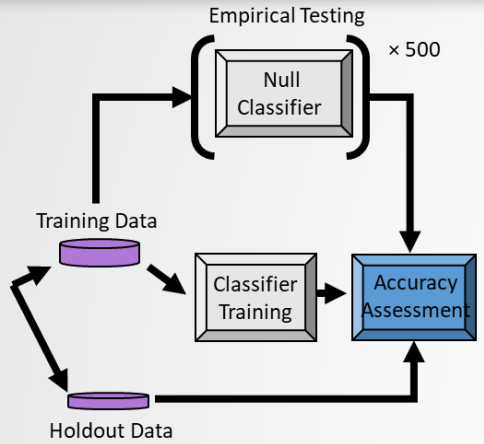


Comparison of internal and holdout accuracies revealed no evidence of systematic overfitting

Models Trained on a Smaller Set of Chemicals can Achieve High Accuracy



Empirical Significance Analysis



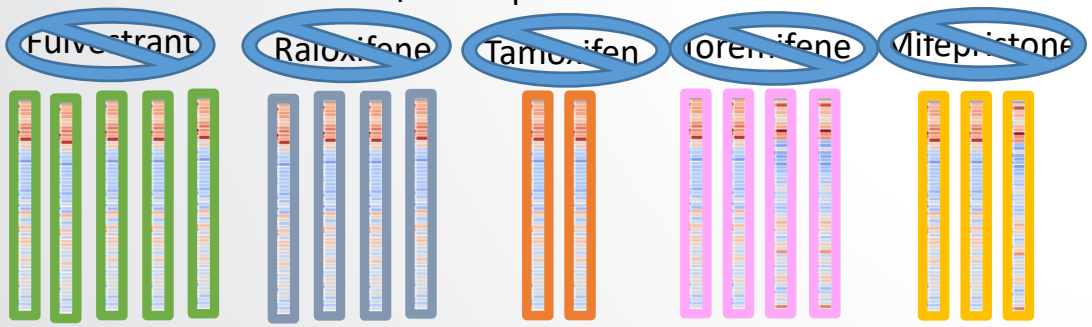
Train multiple "null" classifiers by permuting chemical-MIE associations

Estrogen Receptor Inhibition
ESR (-)

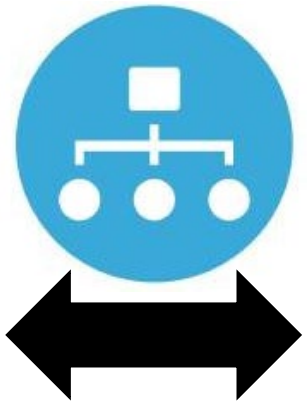
×500

MIE-Active Training Set

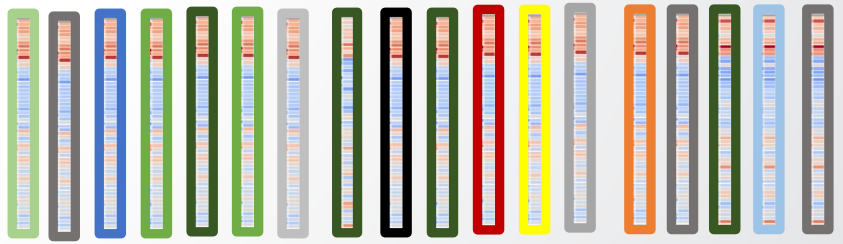
GW-9508 torcetrapib lapatinib sildenafil NVP-TAE684



Collection of MIE-associated chemicals and their profiles

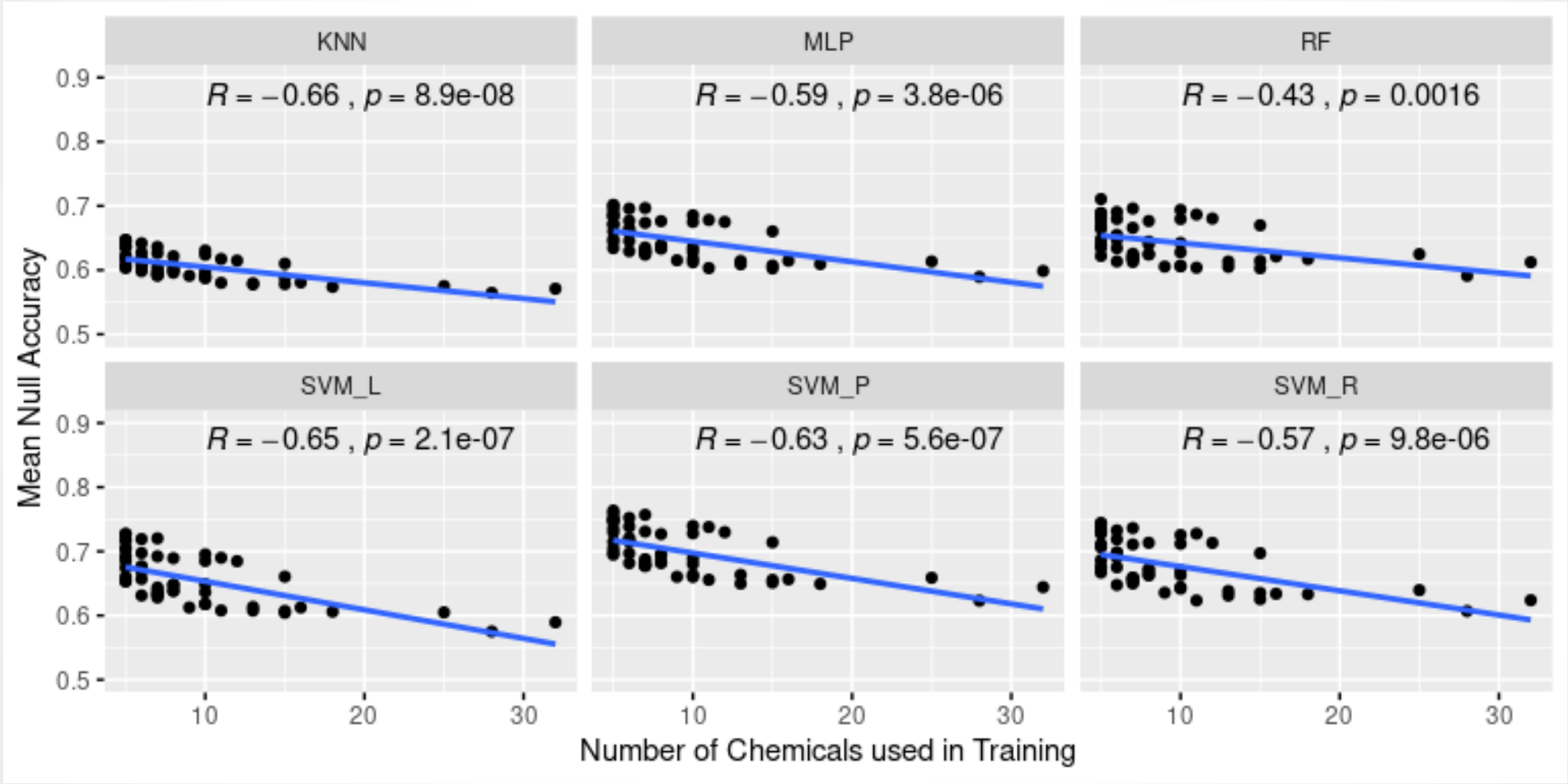


MIE-Inactive Training Set

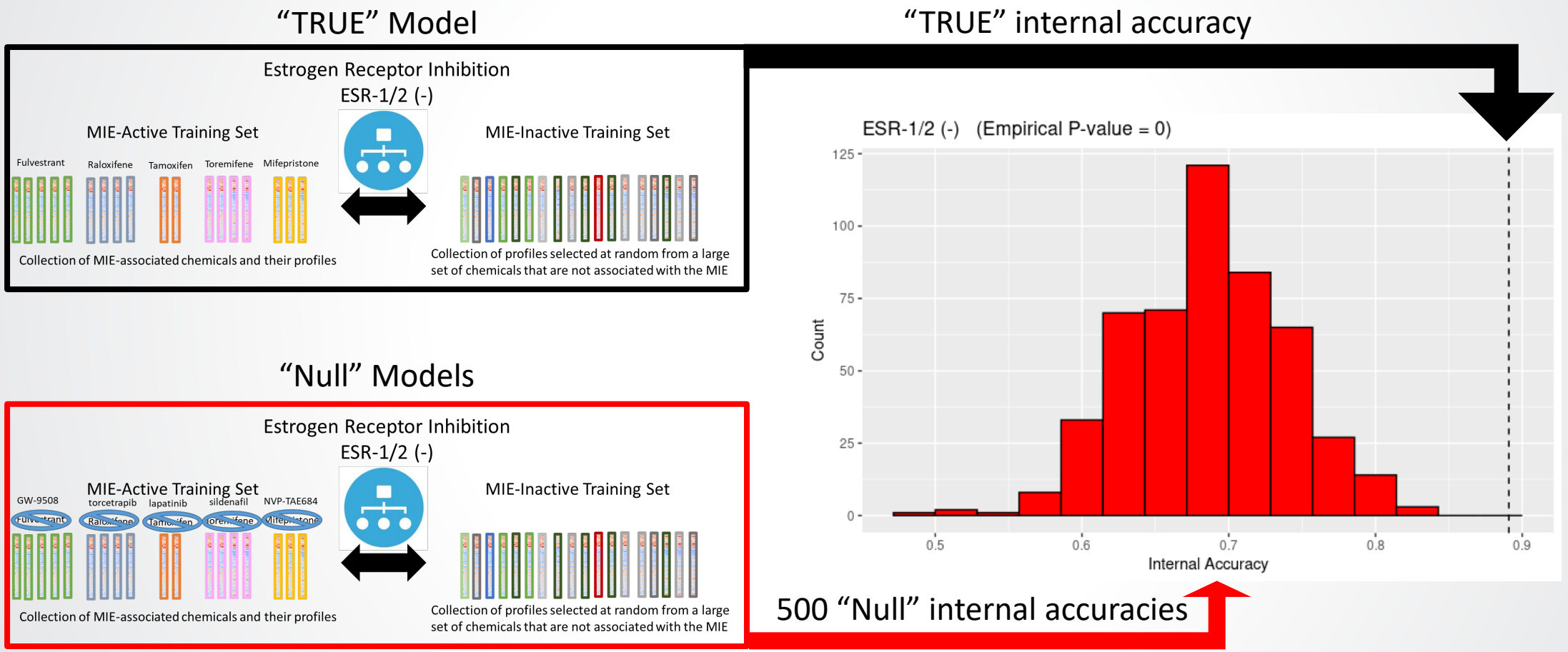


Collection of profiles selected at random from a large set of chemicals that are not associated with the MIE

Null models reveal inverse relationship between training data set size and accuracy

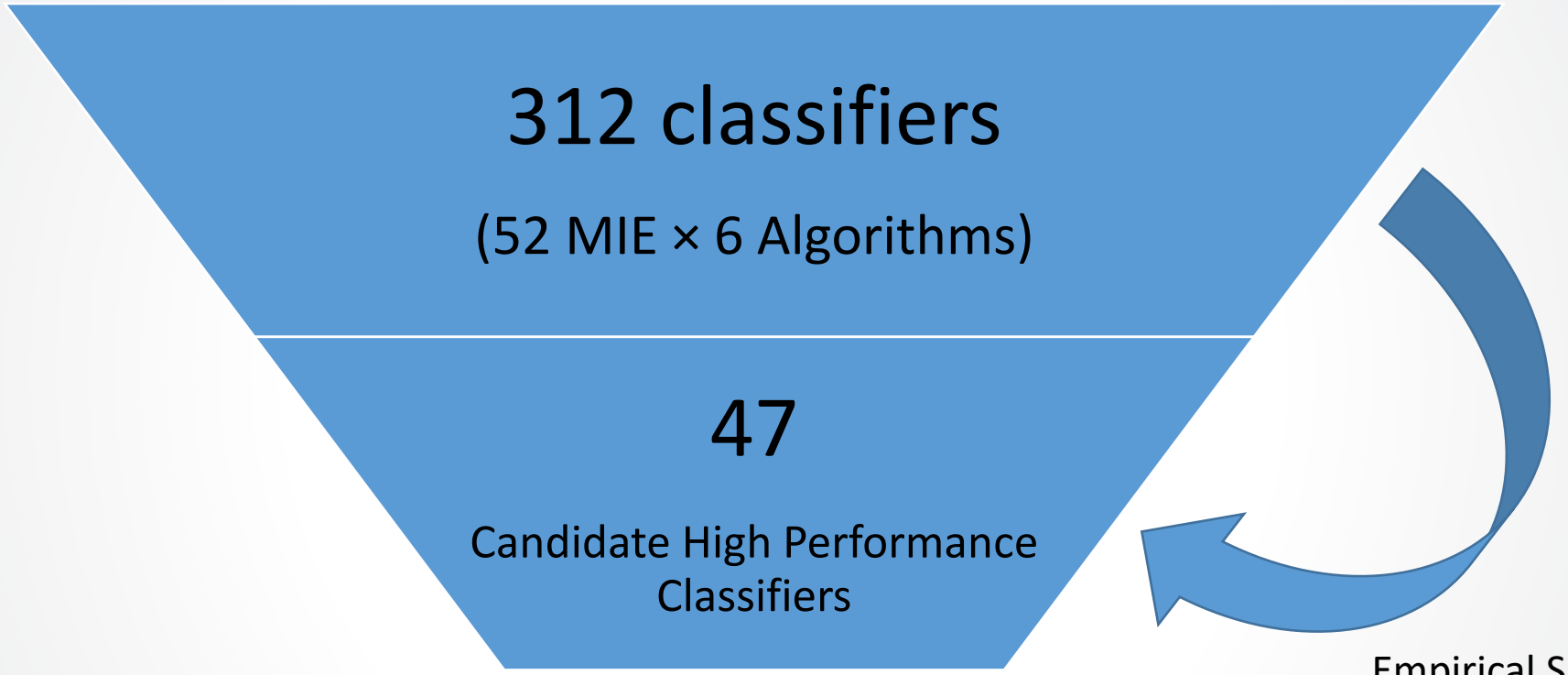


Empirical Significance Analysis

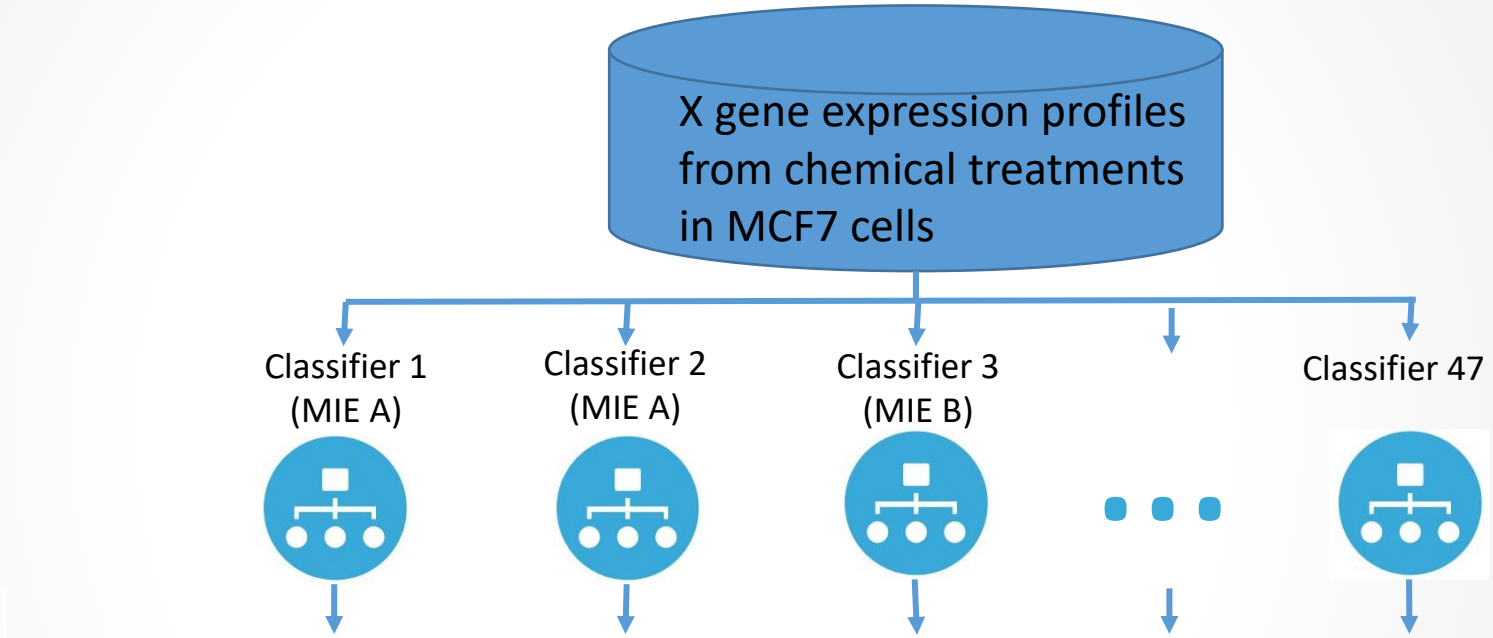


- Calculate percentile rank of "true" un-permuted model accuracy relative to accuracy scores of the 500 permuted models

Selection of High Performing Classifiers



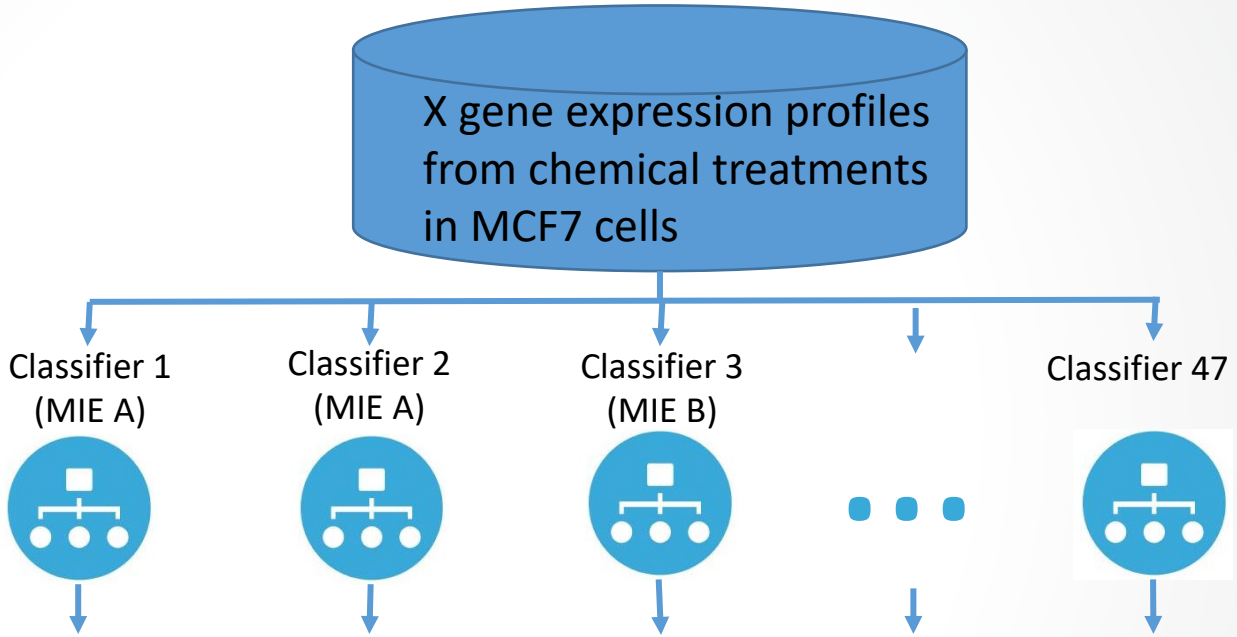
Validation of Candidate High Performance Classifiers with Exemplar Chemical Predictions



Profile Index	MIE A	MIE A	MIE B	…	…
	Algorithm 1	Algorithm 2	Algorithm 1	…	…
Profile 1	0.47	0.49	0.26	…	0.11
Profile 2	0.77	0.17	0.37	…	0.67
Profile 3	0.27	0.58	0.30	…	0.67
Profile 4	0.58	0.92	0.83	…	0.99
Profile 5	0.27	0.28	0.99	…	0.74

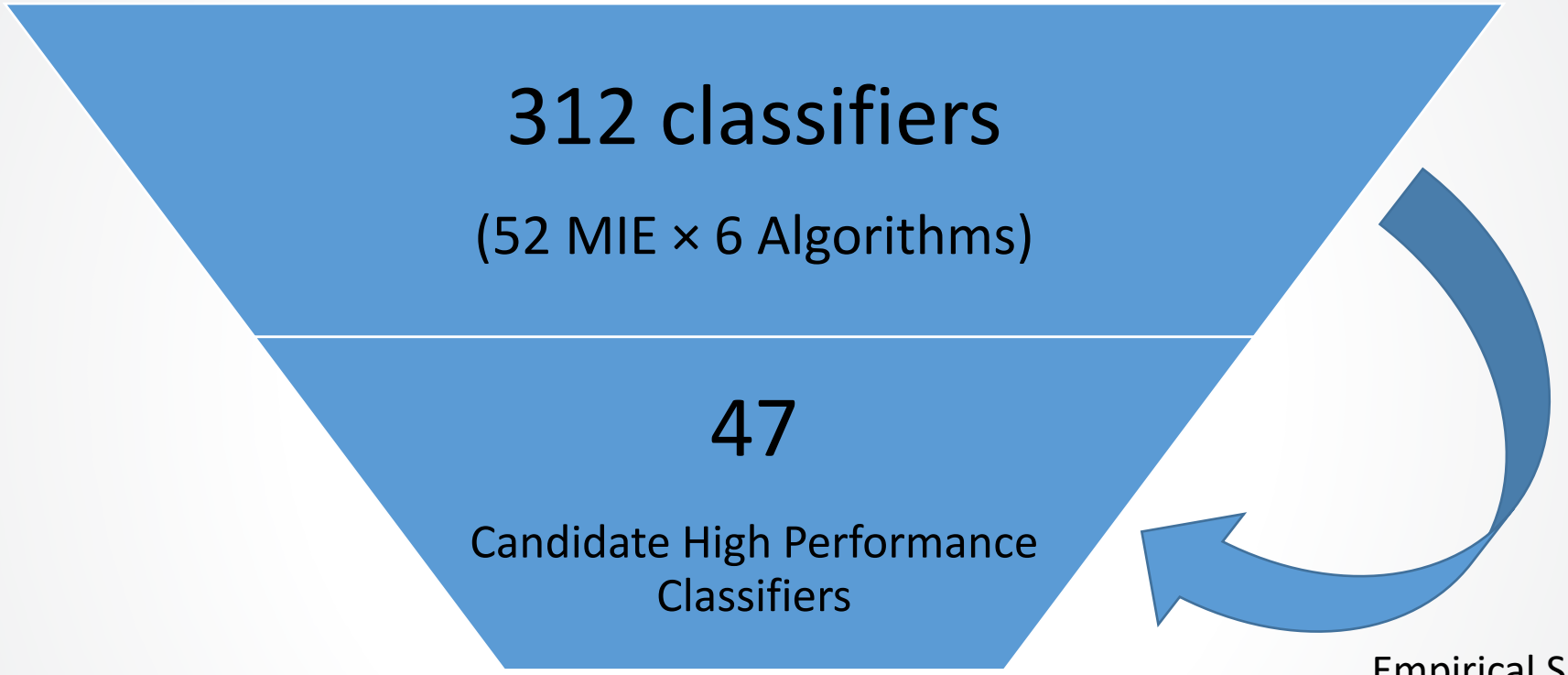
Validation of Candidate High Performance Classifiers with Exemplar Chemical Predictions

Classifiers were retained only if the mean prediction for their corresponding exemplar chemical was ≥ 75% of other chemicals in LINCS



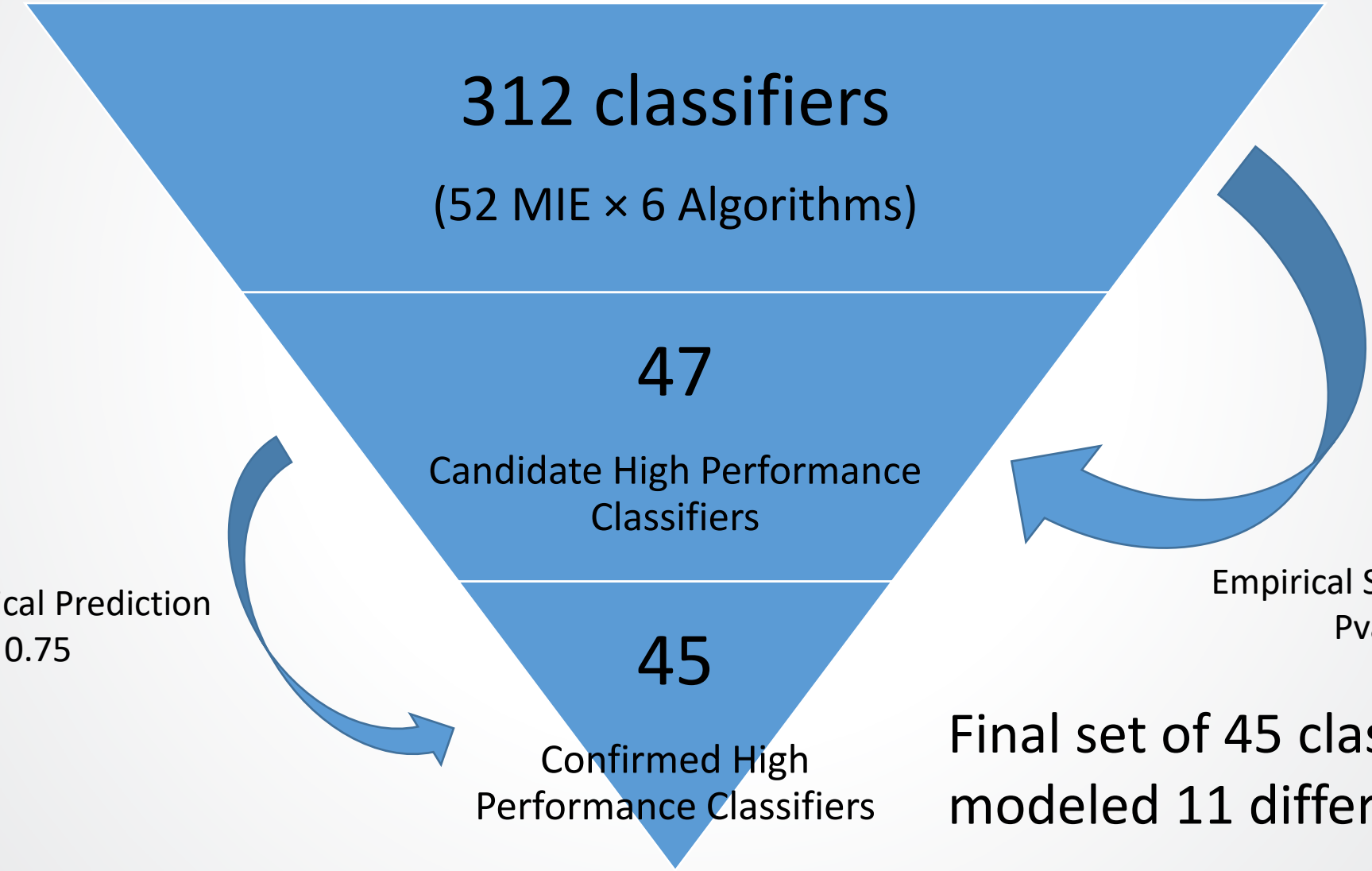
Exemplar Chemical Annotation	Profile Index	MIE A	MIE A	MIE B	…	…
		Algorithm 1	Algorithm 2	Algorithm 1	…	…
MIE A	Profile 1	0.47	0.49	0.26	…	0.11
MIE A	Profile 2	0.77	0.17	0.37	…	0.67
NA	Profile 3	0.27	0.58	0.30	…	0.67
MIE B	Profile 4	0.58	0.92	0.83	…	0.99
MIE B	Profile 5	0.27	0.28	0.99	…	0.74

Selection of High Performing Classifiers



Empirical Significance Testing
Pvalue ≤ 0.05

Selection of High Performing Classifiers



Final set of 45 classifiers modeled 11 different MIEs

Generating Per-MIE Ensemble Predictions

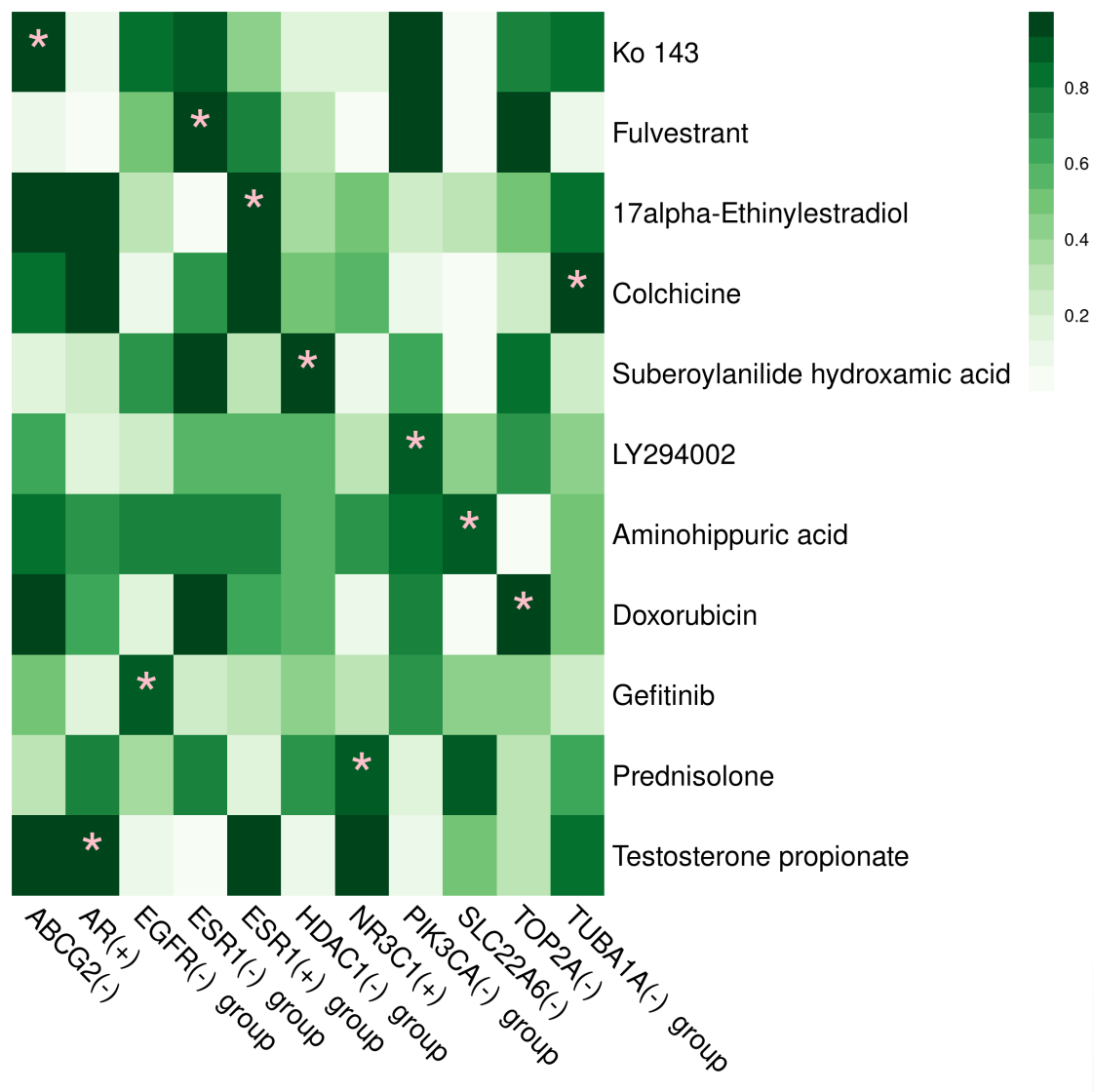
Confirmed High Performance Classifiers

	MIE 1	MIE 1	MIE 2	MIE 2	MIE 2
	Algorithm 1	Algorithm 2	Algorithm 1	Algorithm 2	Algorithm 3
Chemical 1	0.47	0.49	0.26	0.37	0.11
Chemical 2	0.73	0.17	0.37	0.82	0.67
Chemical 3	0.27	0.58	0.30	0.73	0.67

MIE activation predictions from classifiers that model the same MIE but differ in training algorithm were averaged

	MIE 1	MIE 2
Chemical 1	0.48	0.25
Chemical 2	0.45	0.37
Chemical 3	0.43	0.57

Exemplar Predictions for Confirmed High Performance Classifiers



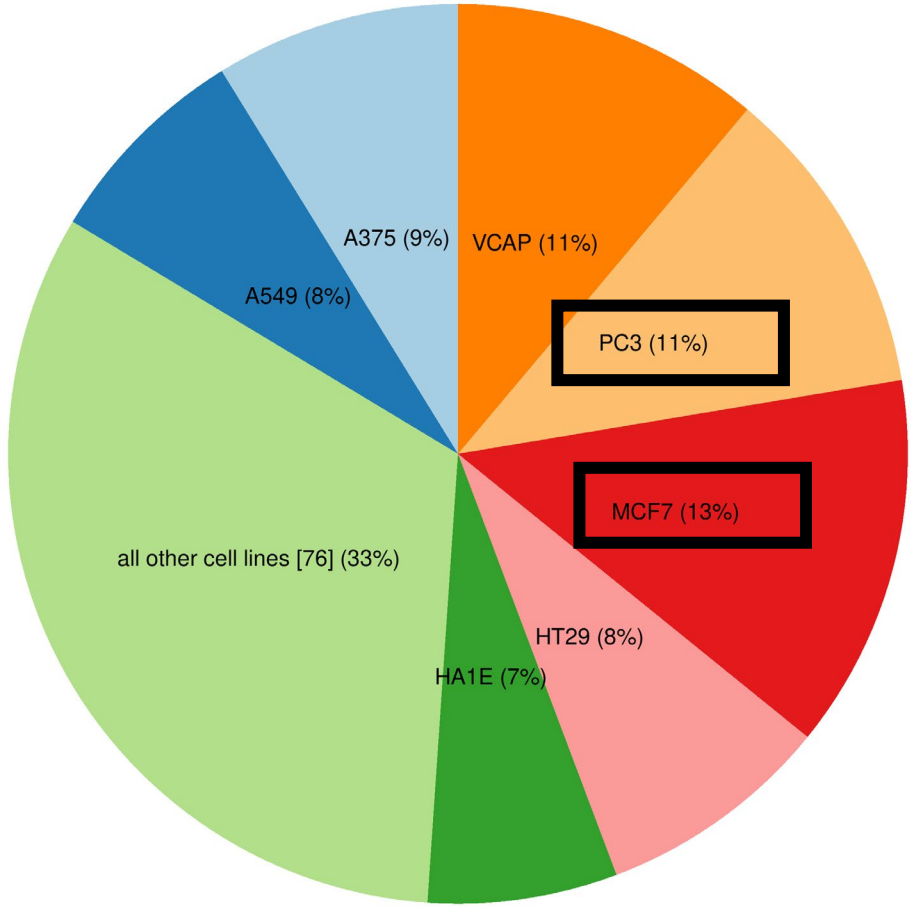
Exemplar chemical predictions for 11 MIEs modeled with confirmed high performance classifiers

Cell shading indicates the percent rank of a training excluded exemplar chemical (rows) for a given MIE (columns)

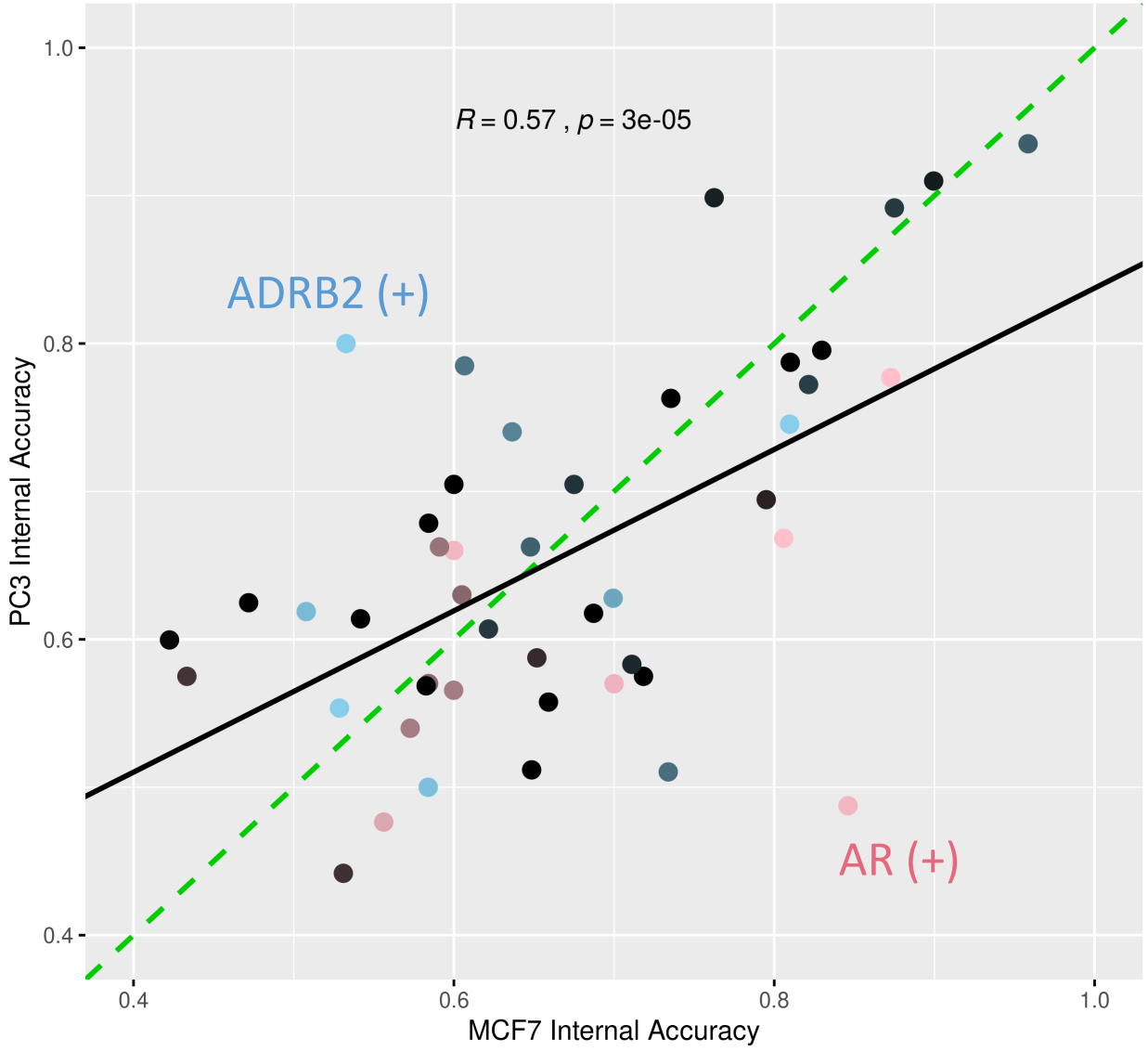
* = MIE-chemical linkage according to RefChemDB

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 47 of the 52 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

LINCS L1000 Project Summary

Key Points

Trained predictive models for 52 MIEs by integrating LINCS L1000 gene expression data with RefChemDB chemical-target labels

- Explored factors that affected model accuracy
- Identified 11 MIEs modeled with high performance classifiers
- Compared classifiers trained on MCF7 and PC3-derived data, revealing that some MIEs are better modeled in one cell type

Ongoing and Future Work

- Improve MIE prediction using alternative gene expression and annotation data sets
  - EPA is currently analyzing TempO-Seq based transcriptomic chemical screens in multiple human cell lines (MCF7, HepaRG, U2OS)
  - Ongoing efforts to improve reference chemical annotations from literature mining approaches
- Explore deep learning approaches
  - Implemented convolutional neural network

Acknowledgements

EPA High-Throughput Transcriptomics Team

CCTE Leadership
Rusty Thomas (CCTE)
Sid Hunter (BCTD)
John Cowden (CTBB)

Project Contributors
Antony Williams
Chris Grulke
Logan Everett
Richard Judson
Imran Shah
Katie Paul-Friedman
Jesse Rogers

47 Candidate High Performance Classifiers Identified

MIE Name	Algorithm	Internal Accuracy	Holdout Accuracy	MIE Active Profiles	Mean Null Accuracy	Empirical Significance
ABCG2(-)	MLP	0.88	0.75	50	0.68	0.00
ABCG2(-)	RF	0.85	0.75	50	0.68	0.02
ABCG2(-)	SVM_L	0.94	0.80	50	0.70	0.00
ABCG2(-)	SVM_P	0.89	0.80	50	0.75	0.01
AR(+)	RF	0.83	0.86	58	0.62	0.00
AR(+)	SVM_P	0.82	0.86	58	0.69	0.00
AR(+)	SVM_R	0.82	0.91	58	0.67	0.00
CYP2D6(-)	SVM_R	0.82	0.85	52	0.67	0.01
EGFR(-) group	SVM_L	0.81	0.82	151	0.68	0.00
EGFR(-) group	SVM_P	0.83	0.87	151	0.73	0.00
EGFR(-) group	SVM_R	0.81	0.83	151	0.71	0.01
ESR1(-) group	KNN	0.91	0.88	68	0.65	0.00
ESR1(-) group	MLP	0.90	0.81	68	0.70	0.00
ESR1(-) group	RF	0.95	0.96	68	0.69	0.00
ESR1(-) group	SVM_L	0.92	0.92	68	0.73	0.01
ESR1(-) group	SVM_P	0.93	0.88	68	0.76	0.00
ESR1(-) group	SVM_R	0.91	0.88	68	0.74	0.01

Validation of Candidate High Performance Classifiers

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	…	…	…	…

Generate predictions for every gene expression profile for every candidate high performance classifier

Distill per-profile 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)	…	…	…

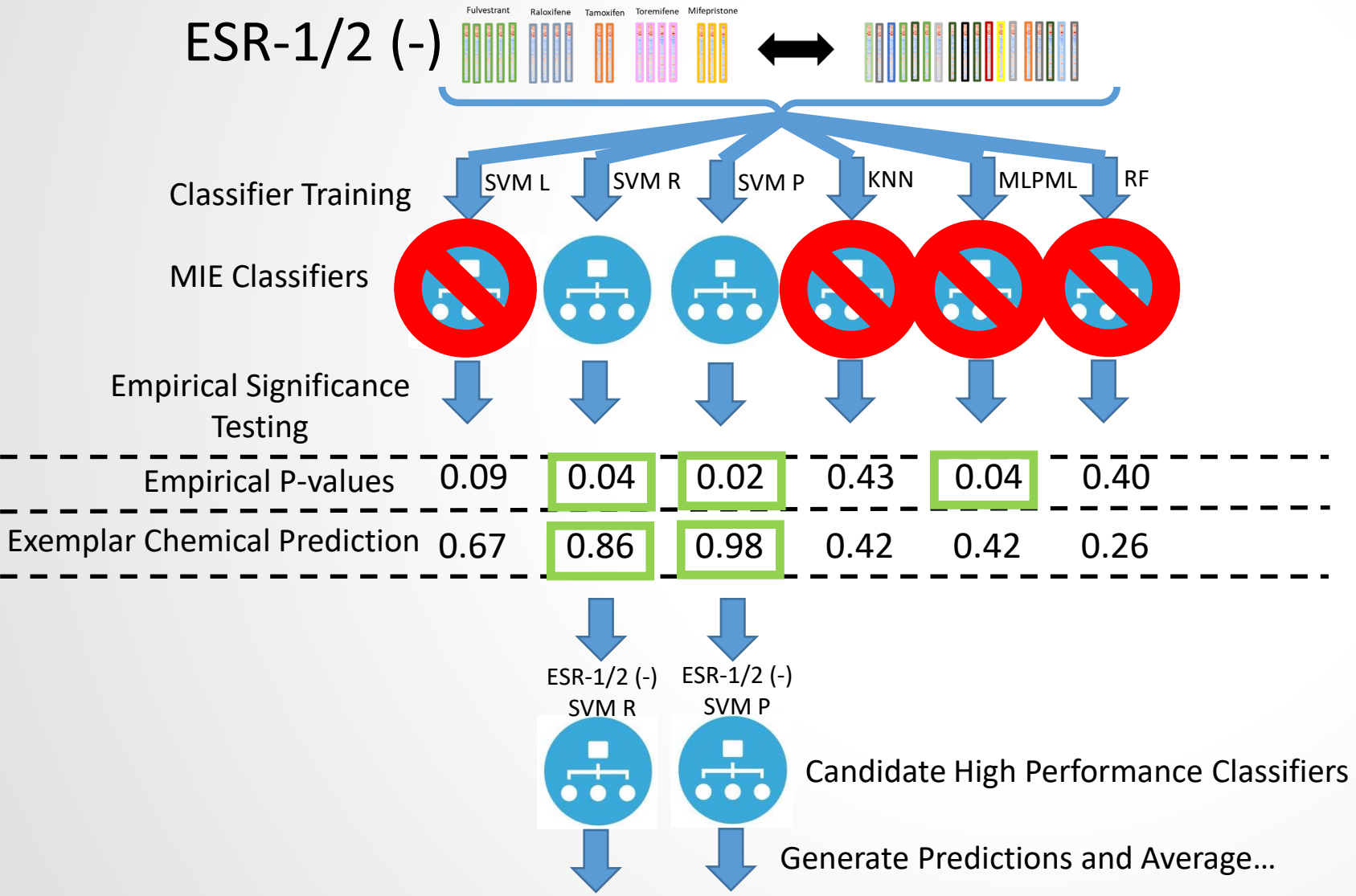
Calculate the percentile rank for each chemical

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)	…	…	…

45 High Performance Classifiers Retained after Exemplar Chemical based Validation

MIE Name	Algorithm	Internal Accuracy	Holdout Accuracy	MIE Active Profiles	Mean Null Accuracy	Empirical Significance	Exemplar Chemical	Exemplar Rank	Exemplar Percent Rank
ABCG2(-)	MLP	0.88	0.75	50	0.68	0.00	Ko 143	3927	0.66
ABCG2(-)	RF	0.85	0.75	50	0.68	0.02	Ko 143	2746.5	0.76
ABCG2(-)	SVM_L	0.94	0.80	50	0.70	0.00	Ko 143	620	0.95
ABCG2(-)	SVM_P	0.89	0.80	50	0.75	0.01	Ko 143	646	0.94
AR(+)	RF	0.83	0.86	58	0.62	0.00	Testosterone propionate	8	1.00
AR(+)	SVM_P	0.82	0.86	58	0.69	0.00	Testosterone propionate	150	0.99
AR(+)	SVM_R	0.82	0.91	58	0.67	0.00	Testosterone propionate	78	0.99
CYP2D6(-)	SVM_R	0.82	0.85	52	0.67	0.01	Quinidine	6555	0.44
EGFR(-) group	SVM_L	0.81	0.82	151	0.68	0.00	Gefitinib	1463	0.87
EGFR(-) group	SVM_P	0.83	0.87	151	0.73	0.00	Gefitinib	1058	0.91
EGFR(-) group	SVM_R	0.81	0.83	151	0.71	0.01	Gefitinib	1005	0.91
ESR1(-) group	KNN	0.91	0.88	68	0.65	0.00	Fulvestrant	1.5	1.00
ESR1(-) group	MLP	0.90	0.81	68	0.70	0.00	Fulvestrant	1	1.00
ESR1(-) group	RF	0.95	0.96	68	0.69	0.00	Fulvestrant	8	1.00
ESR1(-) group	SVM_L	0.92	0.92	68	0.73	0.01	Fulvestrant	3	1.00
ESR1(-) group	SVM_P	0.93	0.88	68	0.76	0.00	Fulvestrant	2	1.00
ESR1(-) group	SVM_R	0.91	0.88	68	0.74	0.01	Fulvestrant	1	1.00

Combining Multiple Models into an Ensemble Classifier



Classifiers must pass both empirical significance testing and exemplar chemical validation to be retained in the analysis

Classifiers that pass these tests that correspond to the same MIE are retained and have their predictions averaged

Alternatives to LINCS L1000 CMAP Training Data

Pros of using LINCS L1000 data for MIE prediction

- Publicly available
- Contains profiles from thousands of chemical treatments
- Spans multiple cell lines

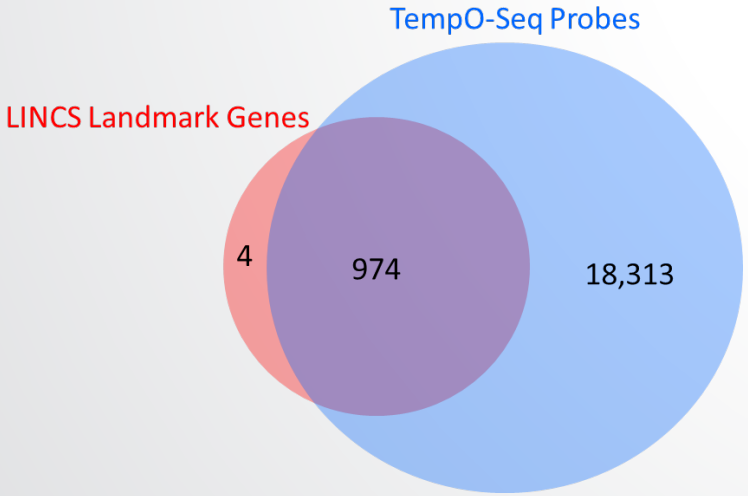
Cons of using LINCS L1000 data for MIE prediction

- Chemicals in LINCS were not selected with a priority on hazard identification
- Modest overlap between LINCS and RefChemDB chemicals
- Treatment concentrations vary between chemicals, some screened in single conc
- 978 transcripts are measured – most gene expression is inferred

Predicting MIEs from HTTr TempO-Seq Data

EPA High Throughput Transcriptomics team is currently developing methods to analyze gene expression data from large TempO-Seq based chemical screens

- HTTr MCF7 screen spans 2,049 unique DTXSIDs
- Chemicals are screened at 8 concentrations, consistent from chemical to chemical



Trained MIE classifiers on gene expression profiles MCF7 TempO-Seq chemical screen

Only 191 (~9.3%) of screened chemicals are annotated in RefChemDB with support level >= 5

MIE	TempO-Seq Profiles	TempO-Seq Chemicals
ESR1(+) group	104	17
AR(-)	41	12
NR3C1(+)	39	6
PTGS2(-) group	33	12
CA2(-) group	31	12
ABCB1(-)	31	7
PPARA(+) group	30	12
ESR1(-) group	29	6
NR1I2(+)	28	6

MCF7-Trained Candidate High Performance Classifiers

Target Name	Algorithm	Internal Accuracy	Holdout Accuracy	Target Members	Mean Null Internal Accuracy	Empirical Pval
AHR_Positive	rf	0.90	0.89	135	0.68	0.00
AR_Positive	rf	0.81	0.90	52	0.62	0.00
AR_Negative	svmRadial	0.76	0.71	270	0.62	0.00
AR_Negative	svmPoly	0.80	0.70	270	0.64	0.00
AR_Negative	svmLinear	0.78	0.71	270	0.63	0.00
NR3C1_Positive	svmRadial	0.86	0.84	156	0.68	0.00
NR3C1_Positive	svmPoly	0.97	0.90	156	0.73	0.00
NR3C1_Positive	svmLinear	0.96	0.90	156	0.73	0.00
NR3C1_Positive	rf	0.93	0.85	156	0.68	0.00
NR3C1_Positive	mlpML	0.80	0.60	156	0.63	0.00

- 10 classifiers passed empirical significance testing
- These classifiers spanned 4 of the original 20 MIEs

Exemplar Chemical Predictions for TempO-Seq trained Candidate High Performance Classifiers