EPA
www.epa.gov

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

J. L. Bundy[1], R. Judson[1], A.J. Williams[1], C. Grulke[1], I. Shah[1], L. J. Everett[1]
1) US EPA, Research Triangle Park, NC

Link to BMC BioData Mining manuscript describing this work:

The views expressed in this presentation are those of the author(s) and do not necessarily represent the views or policies of the Agency.

Joseph L. Bundy I bundy.joseph@epa.gov

Introduction

Goal: U.S. EPA is developing new approach methodologies (NAMs) to identify potential toxicity pathways. Some NAMs are using mechanistic data, such as high throughput transcriptomics (HTTr), to connect apical effects with molecular initiating events (MIEs). To meet this challenge, we are developing a machine learning based method that integrates HTTr data and chemical-MIE labels to predict MIEs.

Key points:

- Integrated LINCS L1000 CMAP gene expression compendium [1]
- Used RefChemDB database of chemical-protein target interactions [2]
- Trained binary classifiers on integrated data sets with the following parameters:
  - MCF7-derived gene expression profiles in LINCS L1000 CMAP data
  - 52 MIEs
  - 3 Training Feature Types
  - 6 Classification Algorithms

Classifier Training Overview

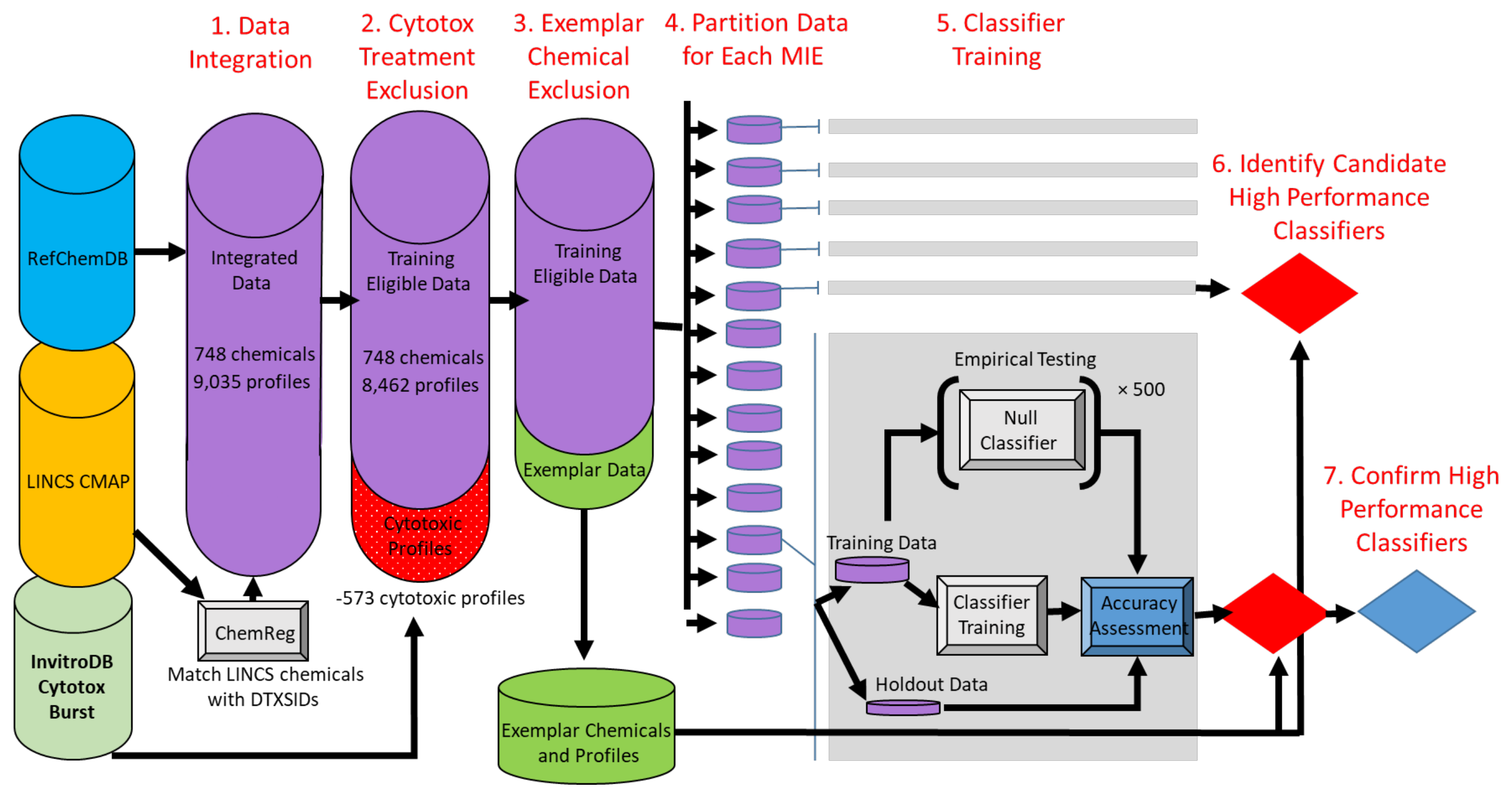

Figure 1. Data processing and classifier training workflow

The prediction of chemical bioactivity at the level of MIEs required the integration of Chemical-MIE labels and a large gene expression compendium (Figure 1).

Method:

1. Chemical treatments associated with LINCS L1000 profiles were matched to EPA substance identifiers (DTXSIDs) using ChemReg [3] and identifiers in LINCS metadata.
2. LINCS profiles corresponding to chemical treatments above the InvitroDB cytotoxic burst value were dropped from the analysis
3. Exemplar chemicals associated with MIEs of interest were excluded from classifier training for downstream validation
4. Remaining data were partitioned into MIE-specific training data sets (see Figure 2.)
5. Binary classifiers were trained independently for each of 52 distinct MIEs using the R package *caret*. 500 Null classifiers were trained for each combination of MIE and classification algorithm
6. Classifiers that passed empirical significance testing were flagged as candidate high performance classifiers
7. Candidate high performance classifiers were validated with training excluded exemplar chemicals

Selection of Training Data

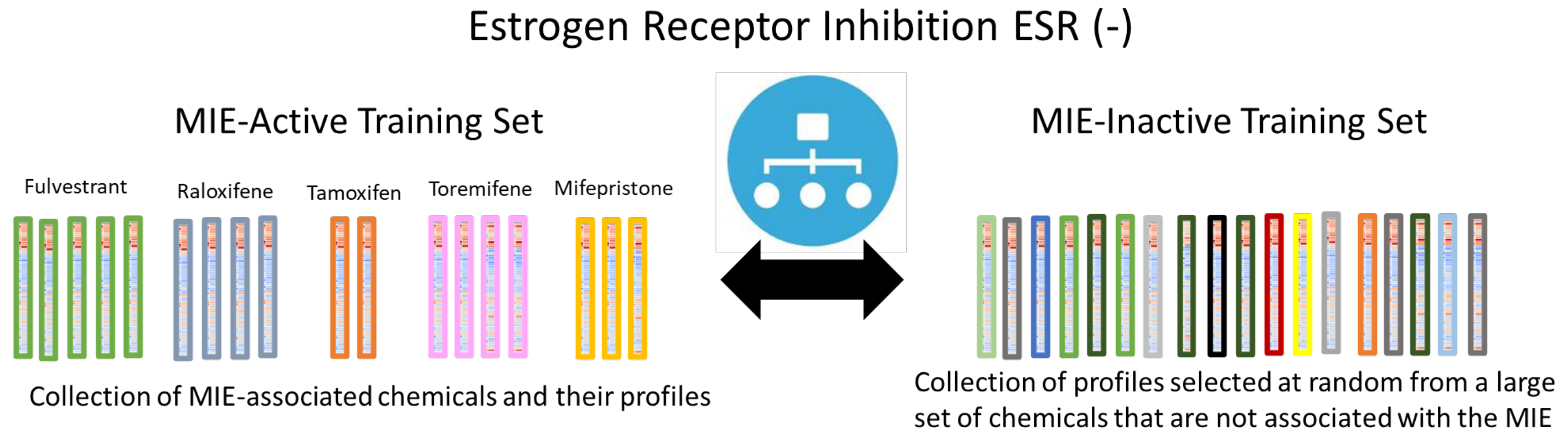
Estrogen Receptor Inhibition ESR (-)
MIE-Active Training Set
Fulvestrant
Raloxifene
Tamoxifen
Toremifene
Mifepristone
Collection of MIE-associated chemicals and their profiles
MIE-Inactive Training Set
Collection of profiles selected at random from a large set of chemicals that are not associated with the MIE

Figure 2. Example of training data structure for Estrogen Receptor inhibition. Binary classifiers were trained for each MIE using size-matched collections of LINCS L1000 gene expression profiles (represented by vertical bars) partitioned into a MIE-Active and MIE-Inactive category. MIE-Active profiles were associated with a chemical treatment that is linked to a given MIE in RefChemDB. MIE-Inactive profiles are selected at random from a collection of chemicals with no association with the given MIE in RefChemDB.

Classifier Optimization

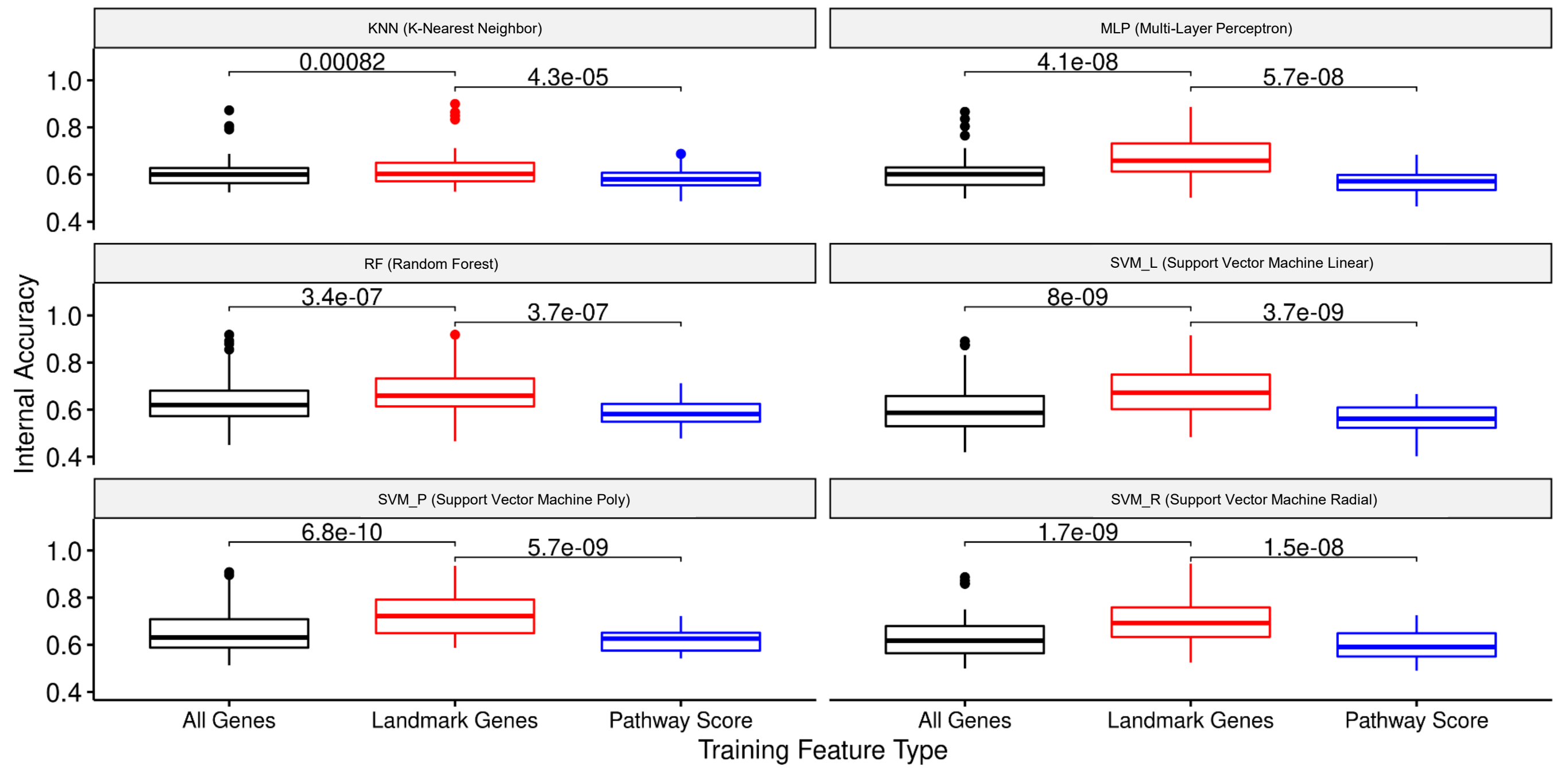

Figure 3. Comparison of internal accuracy distributions for classifiers trained each combination of training algorithm and training feature type. P-values are from a two tailed, paired, Wilcoxon test.

- To optimize classifiers, we evaluated model performance across all the 3 types of gene expression feature sets and 6 classification algorithms (Figure 3).
- Classifiers were trained using 6 different algorithms
- Classifiers were trained using three different sets of features:
  1. Landmark genes
     - ~1,000 transcripts that are directly measured in the L1000 assay
  2. All genes
     - Landmark genes plus ~11,000 genes with inferred expression
  3. Pathway scores
     - Canonical pathways from MSigDB [4] scored with gene set enrichment analysis [5] from calculated from "All Genes" features
- Cross-fold validation accuracies were compared for the 52 MIE classifiers trained on different feature types using a paired Wilcoxon test
- Landmark Gene based classifiers consistently out-performed "All Gene" and "Pathway Score" based classifiers, regardless of algorithm

Empirical Significance Testing

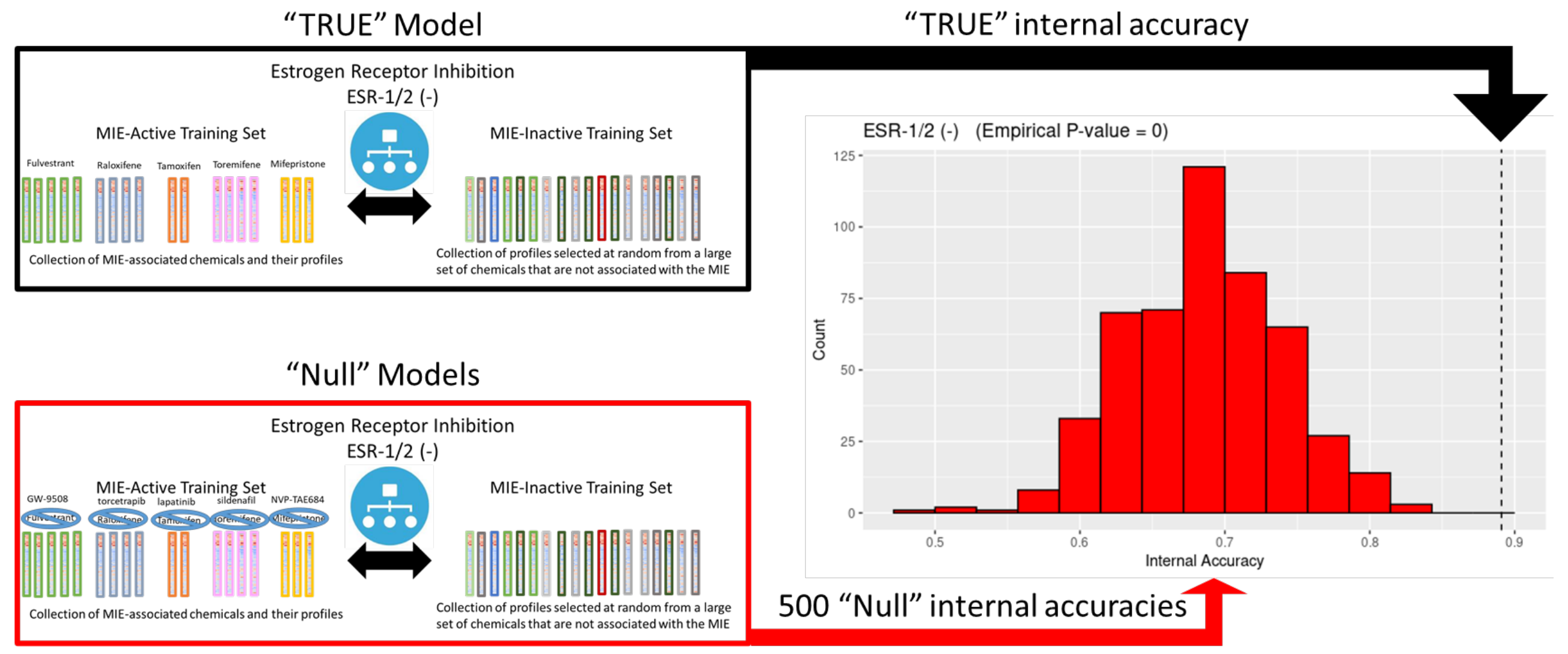

Figure 4. Example of empirical significance calculation using internal accuracy distribution of null models

- Identified candidate high performance classifiers using an empirical significance testing approach
- Classifiers that generated an internal accuracy that is higher than 95% of their "null" counterparts (p-value < 0.05) were retained for further analysis
- 47 candidate high performance classifiers spanning 12 MIEs passed empirical significance testing and were then validated on exemplar chemicals

Generation of Ensemble Classifiers

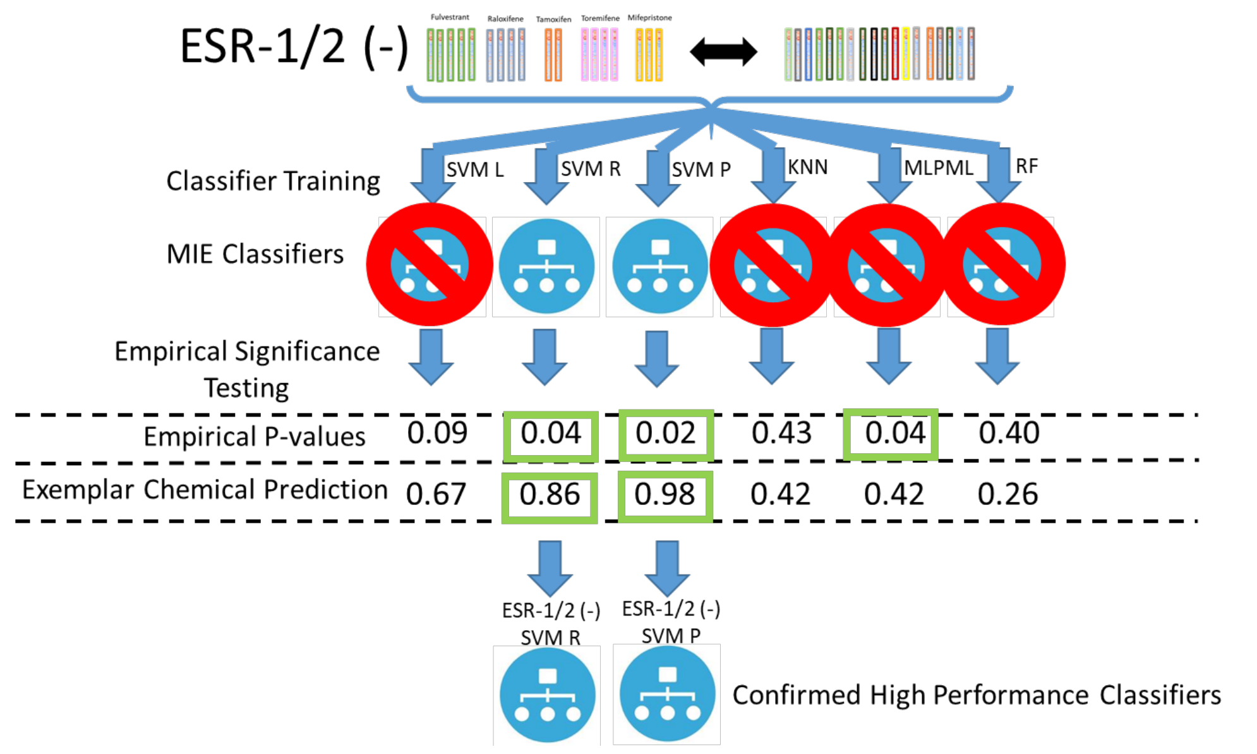

Figure 5. Schematic showing example of classifier validation steps. Classifiers generated for each MIE using 6 training algorithms. Classifiers are filtered based on Empirical Significance testing (pvalue < 0.05) and exemplar chemical prediction (prediction > 0.75). Classifiers that meet both criteria are termed "confirmed high performance classifiers.

- MIE activation predictions were generated for all MCF7-derived gene expression profiles in the LINCS L1000 CMAP data set
- 45 classifiers were validated using training-excluded exemplar chemicals. Retained classifiers must generate a prediction for their training-excluded exemplar chemical that is greater than 75% of the chemicals in the LINCS L1000 CMAP data set
- Predictions for high performance classifiers associated with the same MIE were averaged to generate ensemble predictions
- Confirmed high performance classifiers spanned 11 MIEs

References

1. Subramanian A, Narayan R, Corsello SM, Peck DD, Natoli TE, Lu XD, et al. A Next Generation Connectivity Map: L1000 Platform and the First 1,000,000 Profiles. Cell. 2017;171(6):1437
2. Judson RS, Thomas RS, Baker N, Simha A, Howey XM, Marable C, et al. Workflow for Defining Reference Chemicals for Assessing Performance of In Vitro Assays. Altex-Altern Anim Ex. 2019;36(2):261-76
3. Grulke CM, Williams AJ, Thillanadarajah I, Richard AM. EPA's DSSTox database: History of development of a curated chemistry resource supporting computational toxicology research. Computational Toxicology. 2019;12:100096.
4. Liberzon A, Subramanian A, Pinchback R, Thorvaldsdottir H, Tamayo P, Mesirov JP. Molecular signatures database (MSigDB) 3.0. Bioinformatics. 2011;27(12):1739-40.
5. Barbie DA, Tamayo P, Boehm JS, Kim SY, Moody SE, Dunn IF, et al. Systematic RNA interference reveals that oncogenic KRAS-driven cancers require TBK1. Nature. 2009;462(7269):108-U22.

Ensemble Classifier Predictions

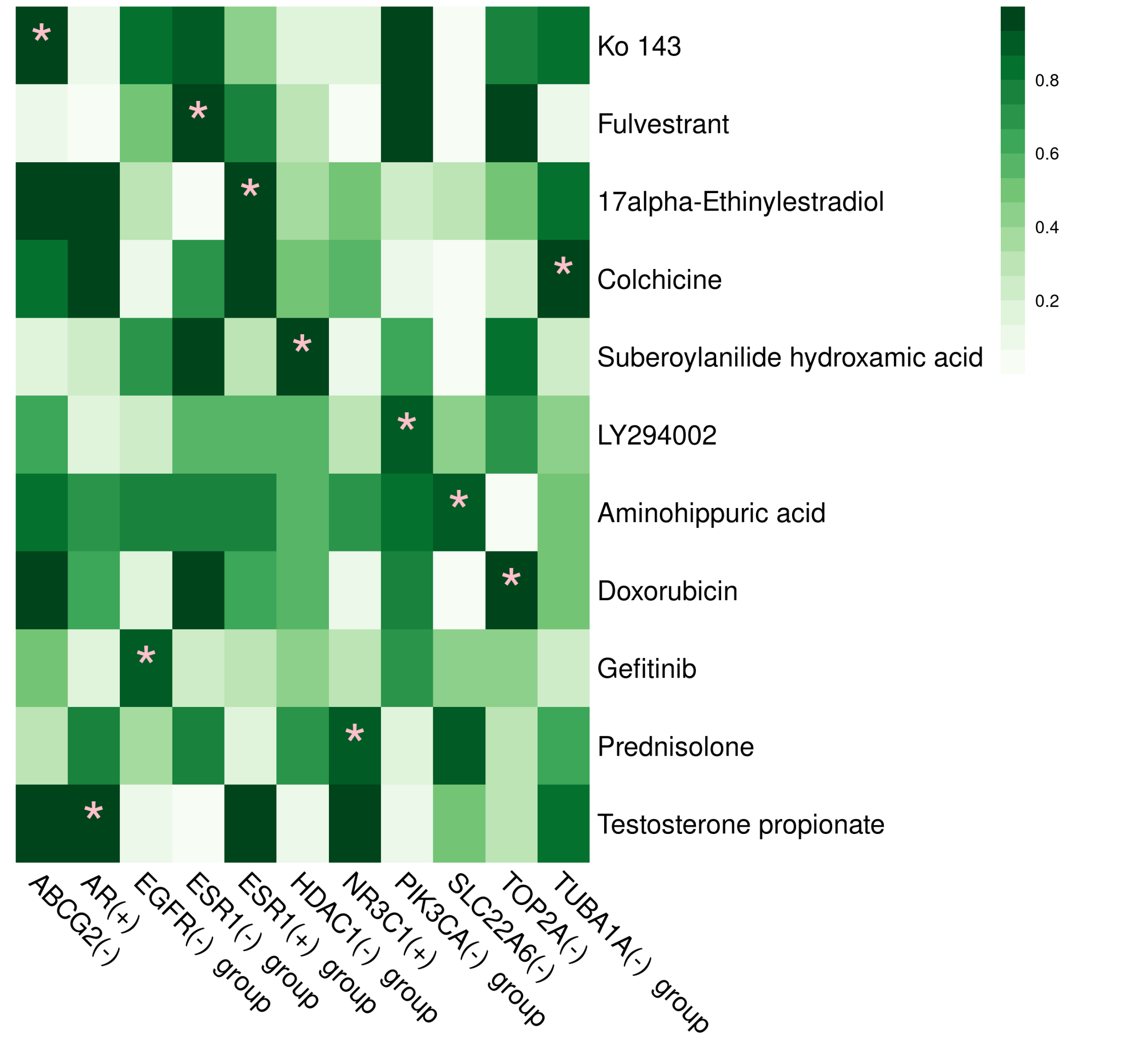

Figure 6. Heatmap showing MIE activation predictions for 11 training-excluded exemplar chemicals for the 11 MIEs modeled with confirmed high-performance classifiers. Shading indicates the "percent rank" of each chemical's prediction relative to all other chemicals screened in the MCF7 cell line in LINCS. Dark green indicates affirmative prediction of MIE activation. * = presence of chemical-MIE linkage in RefChemDB

- Confirmed high performance classifiers correct predicted MIE activation for their corresponding training-excluded exemplar chemicals
- Some classifiers predicted MIE activation for chemicals not annotated for the MIE
  - Possibly false positive predictions
  - Possibly the result of convergence of modeled MIEs onto key events activated by training-excluded exemplar chemicals

Discussion / Conclusions

- Integrated RefChemDB chemical-MIE annotations with LINCS chemical identifiers and gene expression profiles in a machine learning framework
- Trained binary classifiers were trained to predict activation of 52 distinct MIEs
- Classifiers trained on landmark genes yielded the highest internal accuracy
- 47 classifiers that modeled MIEs significantly better than null models, 45 of which were validated with training-excluded exemplar chemicals.
- 11 MIEs modeled with the remaining classifiers showed correctly predicted MIE activation of training-excluded exemplar chemicals
  - Some exemplar chemicals predicted as MIE active in the absence of literature annotations
- Findings suggest that ML-based methods for predicting MIEs may be helpful in prioritizing chemicals for further study based on transcriptomic profiling and may inform decisions on suitable cell-types for further screening.