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

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:
  - 51 MIEs
  - 3 Feature Sets
  - 6 Classification Algorithms
  - 2 Cell Types
    - MCF7 profiles
    - PC3 profiles

Classifier Training Overview

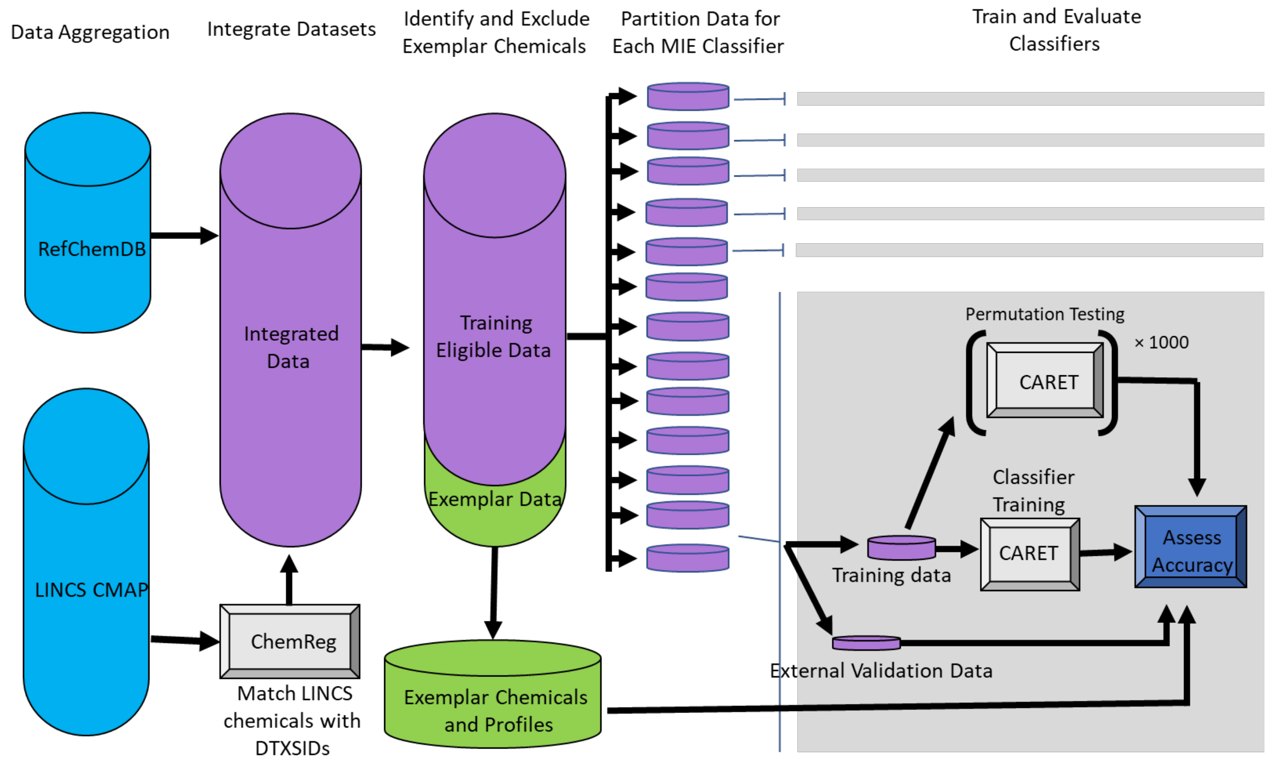

Figure 1. Data processing and classifier training workflow

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. Chemical-MIE linkages in RefChemDB were integrated with LINCS L1000 data to map gene expression profiles to specific MIEs.
3. "Exemplar" chemicals were selected for exclusion from training data to be used later as positive controls.
4. Binary classifiers were trained independently for each of 51 distinct MIEs using the R package *caret*.
5. To measure model performance independent of training data, holdout accuracy was assessed using a hold-out data set consisting of 20% of available gene expression profiles.

Selection of Training Data

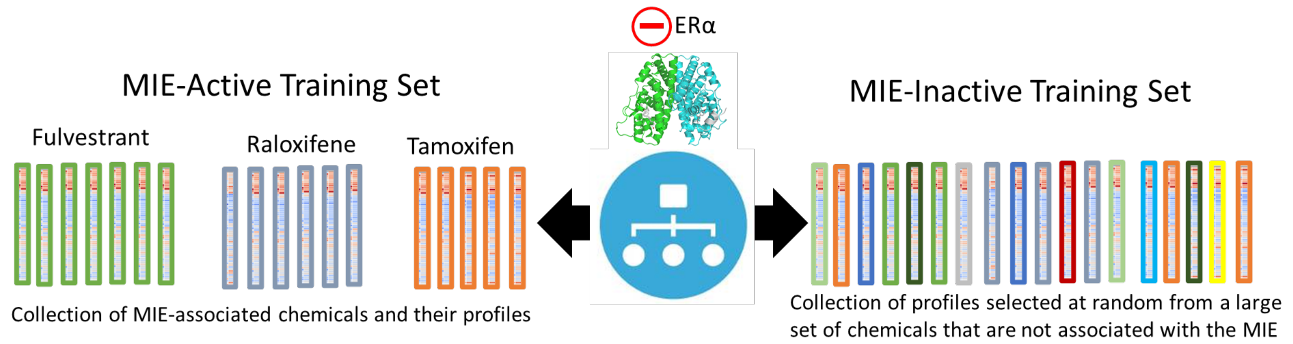
ERα
MIE-Active Training Set
Fulvestrant
Raloxifene
Tamoxifen
Collection of MIE-associated chemicals and their profiles
MIE-Inactive Training Set
Collection of profiles selected at random from a large set of chemicals that are not associated with the MIE

Figure 2. Example of training data structure for Estrogen Receptor-α 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 large collection of chemicals with no association with the given MIE in RefChemDB.

Classifier Optimization

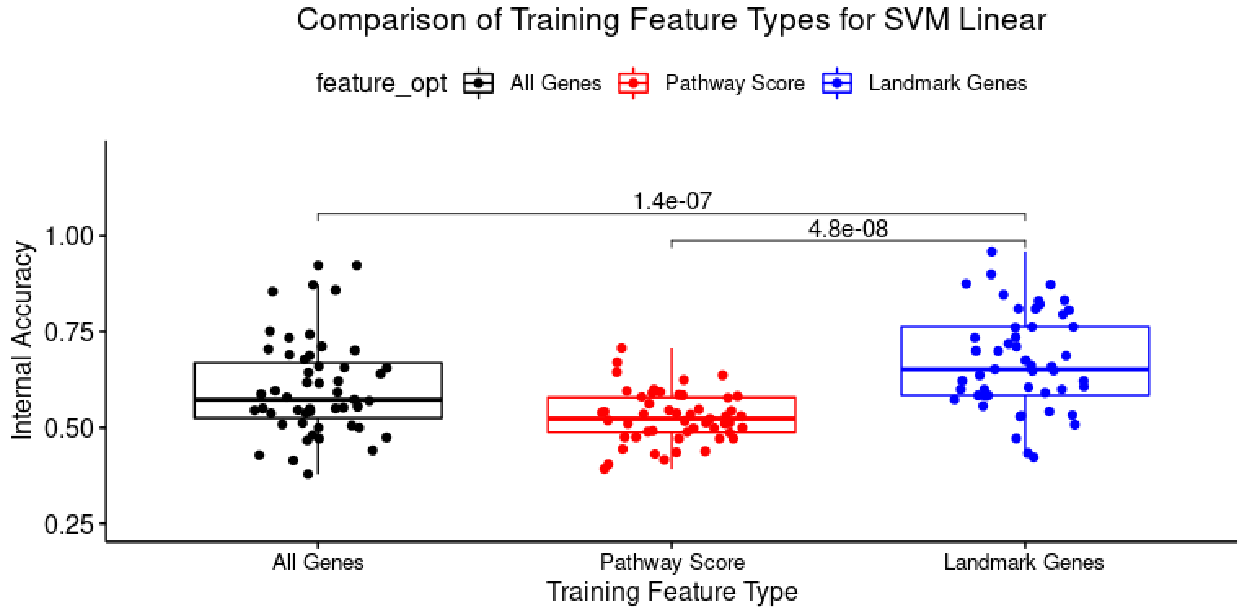
Comparison of Training Feature Types for SVM Linear

Figure 3. Comparison of internal accuracy for Support Vector Machine (SVM) Linear based classifiers trained on different feature types. P-values are from a two tailed, paired, Wilcoxon test.

- To optimize MIE models, we evaluated model performance across all the 3 types of gene expression feature sets (Figure 3) and 6 classification algorithms (Figure 4).
- 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 expression estimates of an additional ~11,000 genes inferred through linear combination of landmark genes
  3. Pathway scores
     - Used the canonical pathways gene set from MSigDB [4]
     - Gene set enrichment scores were calculated from "All genes" features using ssGSEA [5], and the resulting scores used as training features
- Cross-fold validation accuracies were compared for the 51 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**

Classifier Optimization Continued

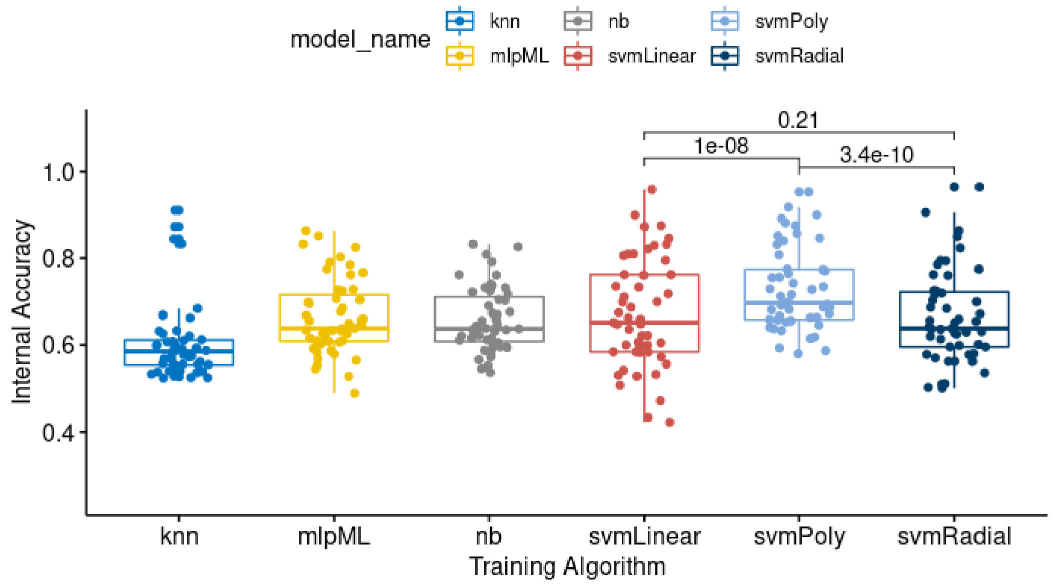

Figure 4. Comparison of internal accuracy across different training algorithms. P-values are from a two-tailed, paired, Wilcoxon test.

- Explored differences in classifier performance as a function of classification algorithm
- Internal accuracies were compared across algorithms using a paired Wilcoxon test
- **svmPoly classifiers achieved significantly higher internal accuracies than svmLinear and svmRadial classifiers**

Test for Overfitting

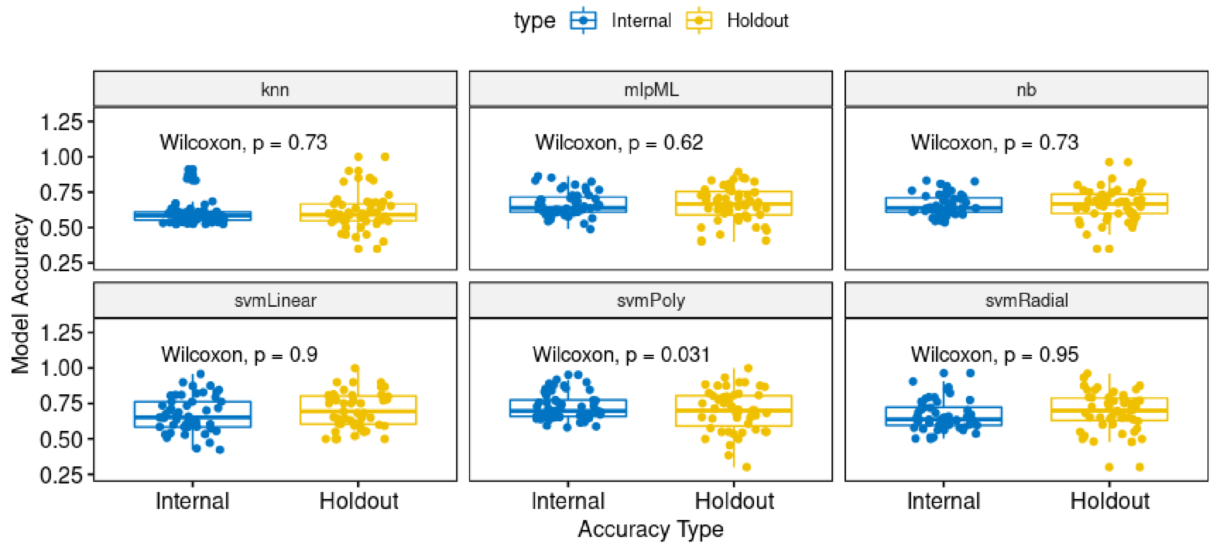

Figure 5. Comparison of internal and hold-out accuracies across classification algorithms. P-values are from a one-tailed, paired, Wilcoxon test.

- Comparison of internal accuracy and hold-out accuracies from svmPoly-based models indicated significantly lower holdout accuracies
  - **Suggests that svmPoly classifiers were systematically overfit**
  - **Restricted further analysis to the runner up svmLinear based classifier**

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.

Cell Type Analysis

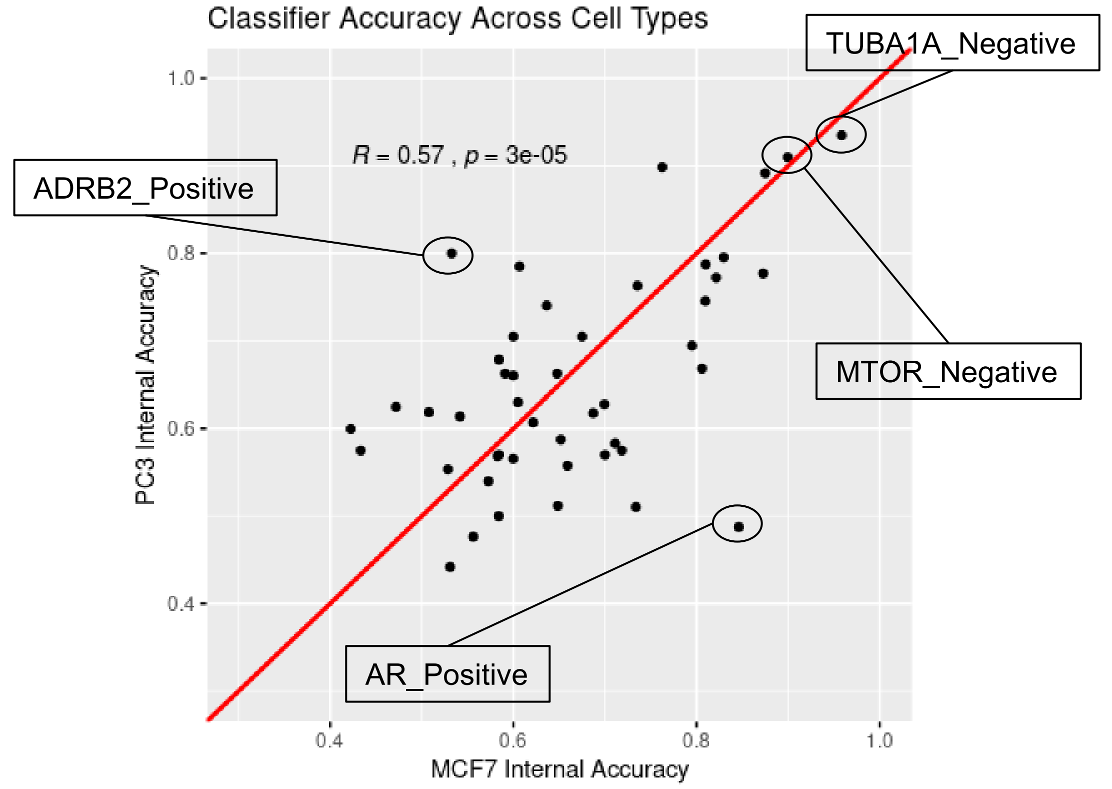
Classifier Accuracy Across Cell Types

Figure 6. Comparison of internal accuracies for MIE models trained on MCF7-derived gene expression profiles vs PC3 derived profiles

- To identify MIEs that were predicting with variable accuracy across cell types, a separate set of classifiers on gene expression profiles derived from the PC3 prostate cancer derived cell line
- Of the 51 classifiers trained on MCF7-derived data, 46 classifiers had sufficient gene expression profiles for training from PC3-derived data
- Accuracies from each cell line were compared with linear regression (Figure 6).
  - **Internal accuracies were generally well correlated**
  - **The two top-performing classifiers as measured by internal accuracy, were the same in both cell types**
    - TUBA1A_Negative
    - MTOR_Negative
  - **Some classifiers showed cell-type specific differences in accuracy.**
    - ADRB2_Positive classifier achieved internal accuracies of 0.80 and 0.53 in PC3 and MCF7 cells
    - AR_Positive achieved an accuracy of 0.85 in MCF7 cells, but only 0.49 in PC3 cells

Discussion / Conclusions

In this study we integrated RefChemDB chemical-MIE annotations with LINCS chemical identifiers and gene expression profiles for the purpose of predicting MIE induction from gene expression profiles. We trained binary classifiers to predict 51 distinct MIEs and explored factors that affected model accuracy such as feature type and classification algorithm. Finally, we trained classifiers on both MCF7 and PC3 derived data and compared accuracies, identifying 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 (AR_Positive in MCF7 cells over PC3 cells). One possible explanation for this disparity is differences in baseline expression of MIE-associated proteins. These 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.

U.S. Environmental Protection Agency
Office of Research and Development