Using Machine Learning to Improve Drug Development

Annamarie Bair*, Matthew B.A. McDermott*, Jennifer Wang†, Wen-Ning Zhao‡, Steven D. Sheridan†, Peter Szolovits*, Isaac Kohane⋆, Stephen J. Haggarty‡, Roy H Perlis†

*CSAIL, MIT; †CEDD & ‡CGM, MGH; ⋆DBMI, Harvard

annabair@mit.edu

Abstract

Drug development is an important endeavor, yet the process is often expensive and time-consuming. The ability to predict properties about a drug, such as biological mechanism of action (MOA) or side effect, have the potential to expedite the process of drug development and provide researchers with insight into the biological processes underlying the effect of a drug. In this work, we utilize the power of large datasets and machine learning methods to offer computational improvements to the drug development pipeline. We successfully predict the side effects and mechanism of action of drugs based on data from gene expression profiles and the chemical structure of the drugs. Additionally, we show that incorporating the chemical structural data can significantly improve performance on our prediction tasks. In evaluating our models, we both use standard accuracy measures and qualitatively compare our predictions to information about the same drugs in the medical and biological literature.

Contributions

- Improve predictive power on side effect and MOA tasks.
- Demonstrate value of combining gene expression and structural data.

Data

Gene Expression Data

Gene expression profiles were obtained from public L1000 datasets.

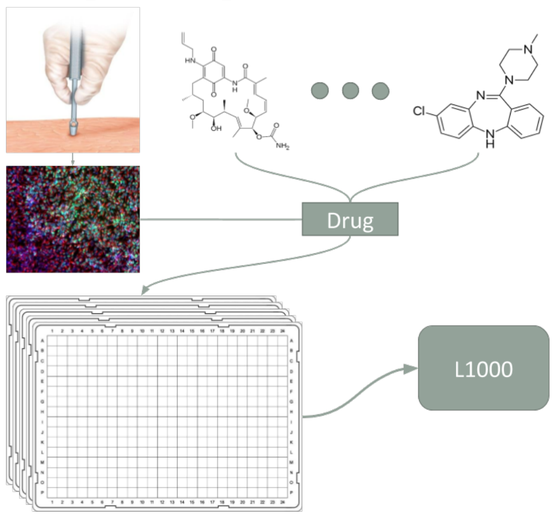


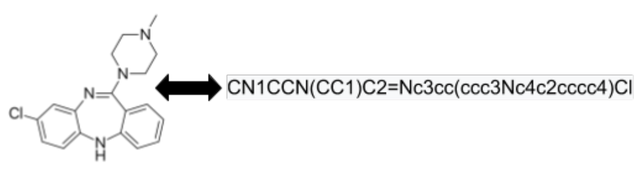
Figure 1: L1000 technique was used to analyze amount of mRNA produced by transcription of each gene in patients' Neural Progenitor Cells.

Task	# Samples	# Cell Lines	# Classes	# Unique drugs
Side Effect	1363	3	6	145
Indication	2629	3	4	290
MOA	2074	3	15	390
All Cells MOA	312,023	82	90	1785

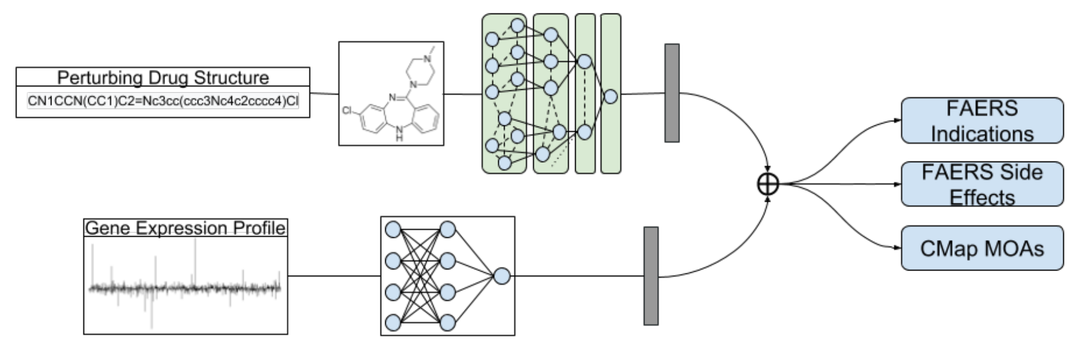
Table 1: Dataset Statistics. All but the last row operate only on NPC cells—the last row summarizes all cell types and is only used when we incorporate structural features.

Chemical Structural Data

Chemical structure represented as a SMILES (Simplified Molecular-Input Line Entry System) string.



Methods



Convolutional Neural Network (CNN) Embedding

- Parse the SMILES string using a CNN in order to create a vector representation of the string.

Graph Convolutional Network (GCN) Embedding

- Represent molecule as a graph: vertices are atoms or bonds and edges indicate a neighboring relationship.
- Process the graph representation to create a vector that captures relationships between atoms and bonds in a molecule.

Results

Drug and gene expression data outperforms gene expression alone on predicting MOA: $40.74\% \pm 6.03\%$ vs. $9.21\% \pm 6.44\%$.

Task	Targets (all $\geq$ 15 drugs)	Chance Acc.	FF-ANN Acc.
Side Effect	Diarrhoea, Pyrexia, Drug Interaction, Dyspnoea, Nausea, Headache	$14.92 \pm 13.50\%$	$35.56 \pm 21.92\%$
Ind.	General, Cardiac Arrhythmias, Epidermal/Dermal, Neurological	$65.86 \pm 20.87\%$	$61.43 \pm 15.87\%$
MOA	Adrenergic receptor agonist, HDAC inhibitor, Dopamine receptor antagonist, Glutamate receptor antagonist, Acetylcholine receptor antagonist, Cyclooxygenase inhibitor, Serotonin receptor antagonist, Histamine receptor antagonist, Serotonin receptor agonist, Adrenergic receptor antagonist, Glucocorticoid receptor agonist, Calcium channel blocker, Phosphodiesterase inhibitor, Dopamine receptor agonist, EGFR inhibitor	$17.09 \pm 18.90\%$	$25.56 \pm 18.27\%$

Table 2: FF-ANN prediction of side effect and MOA outperforms baseline majority classifier

Future Work

- Include more prediction tasks such as toxicity and blood brain barrier penetrance (BBBP).
- Continue refining the GCN embedder.

This research was funded in part by grants from the National Institutes of Health (NIH): National Institute of Mental Health (NIMH) grant P50-MH106933, and National Human Genome Research Institute (NHGRI) grant U54-HG007963.