Prediction of Transcriptomic Response to Perturbagens in Unobserved Cell Lines and Cell States

Tejaswini Ganapathi[1], Edward Moler[1], Rohit Jadhav[1], Munir Yousef[1], Thach Mai[1], Hanadie Yousef[1], Jeremy O'Connell[1]

[1]Juvena Therepeutics, Inc., Redwood City, CA 94063. juvenatherapeutics.com

Introduction

- We developed a framework to predict perturbagen responses on cell lines that are not observed in the training data.
- The framework details methodology to represent small molecule perturbagens as protein ligands that generalize to unobserved interaction combinations across multiple cell lines.
- We identified gene readouts that are good measures of perturbagen effects.

Goals

Problem Statement: Protein perturbagen screening is expensive, but large data sets are required to train predictive models to predict a transcriptional response.

- Can we combine data across assays into single corpus to leverage large models?
- Can we train a model on screening data across cell lines and perturbagens and predict biological effects on a cell line unobserved in training data?
- Can we rank perturbagens based on predictions of biological effects to propose screening order that is more efficient than random/ heuristics ?

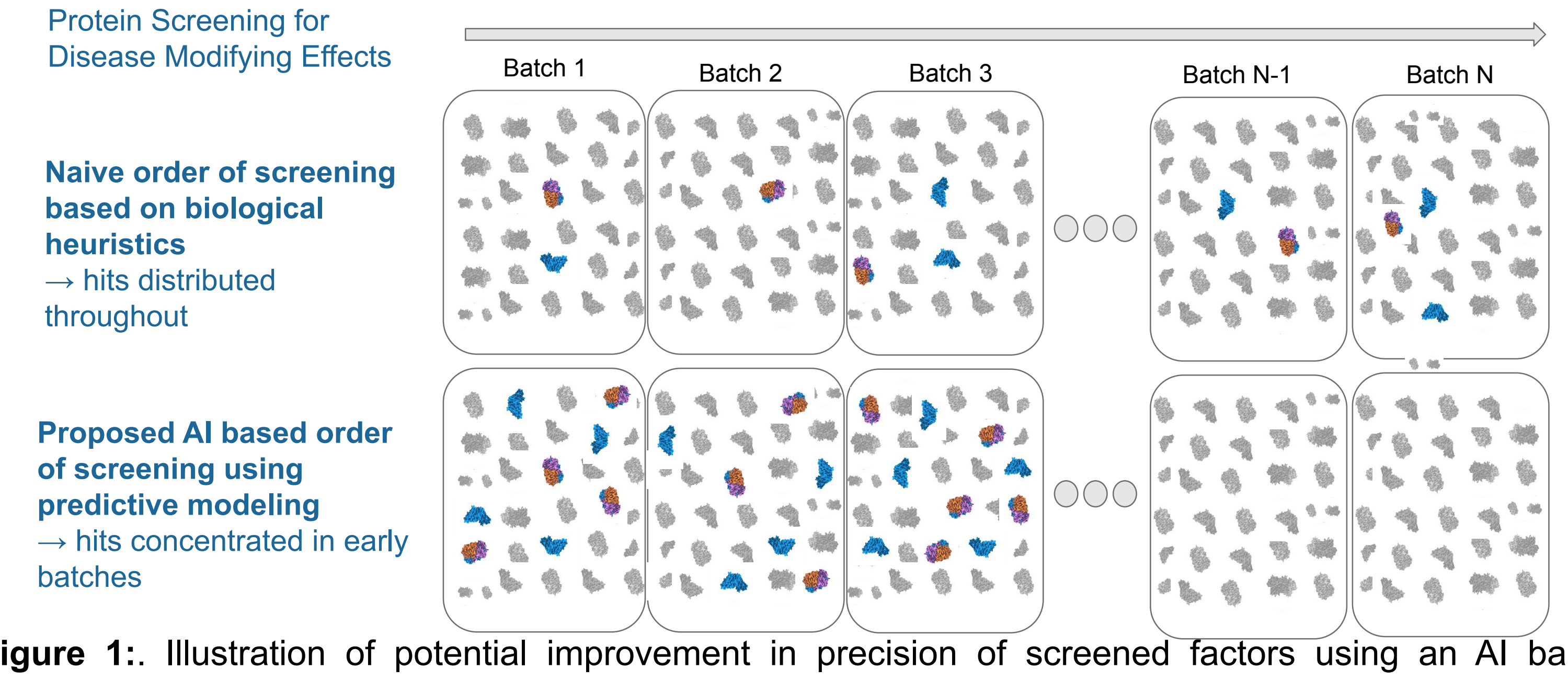

Protein Screening for Disease Modifying Effects
Batch 1
Batch 2
Batch 3
Batch N-1
Batch N
Naive order of screening based on biological heuristics
→ hits distributed throughout
Proposed AI based order of screening using predictive modeling
→ hits concentrated in early batches

Figure 1:. Illustration of potential improvement in precision of screened factors using an AI based approach versus a heuristic approach based on random selection from pre-identified protein factors

Methodology

- LINCS L-1000: Large public dataset with screening data for 58 cell lines, 19,000+ small molecule perturbagens
- Our strategy:
  - Develop a representation for cell lines and small molecule perturbagens
  - Train a predictive model of gene expression changes, that is generalizable to new cells

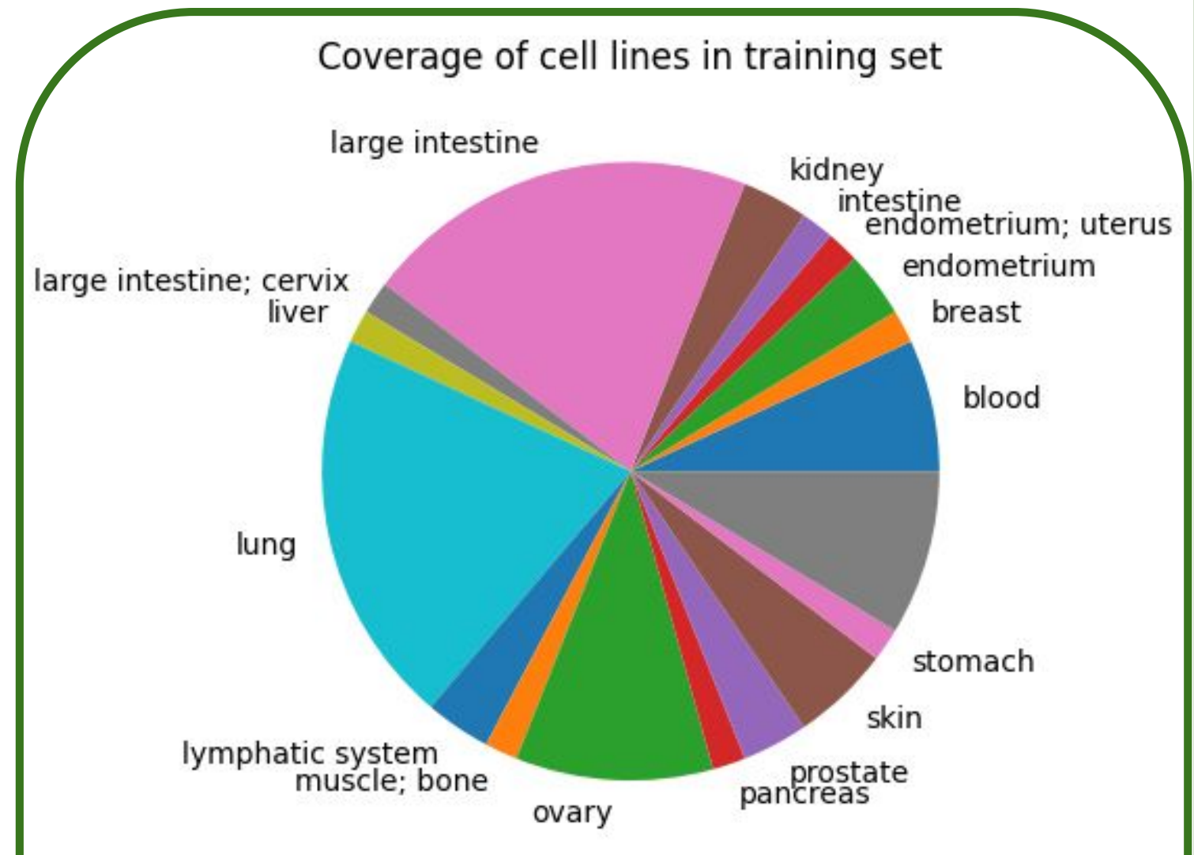


Figure 2:. Cell Lines in LINCS L-1000 cover a large spectrum of cell lines across diseased and normal cell states

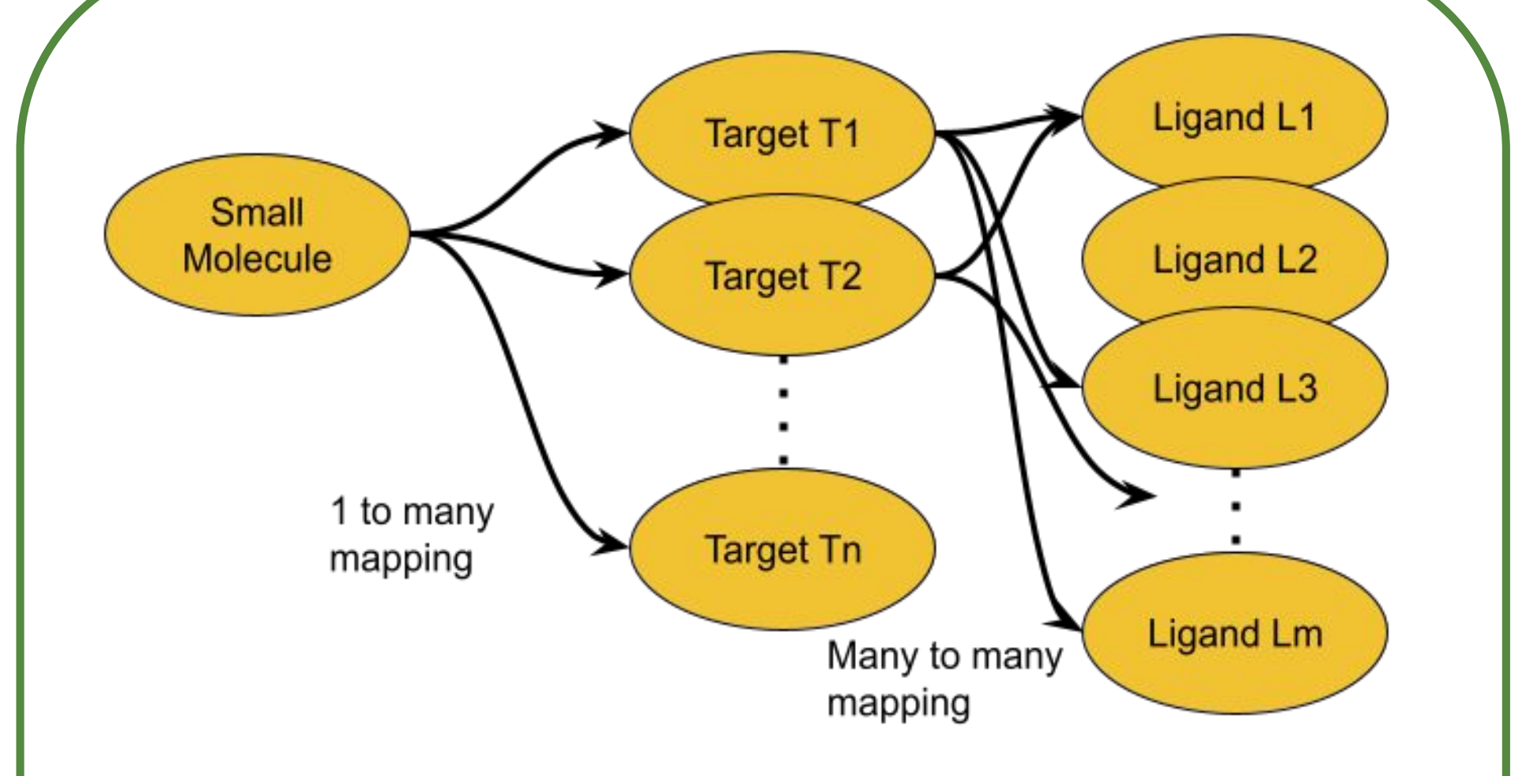
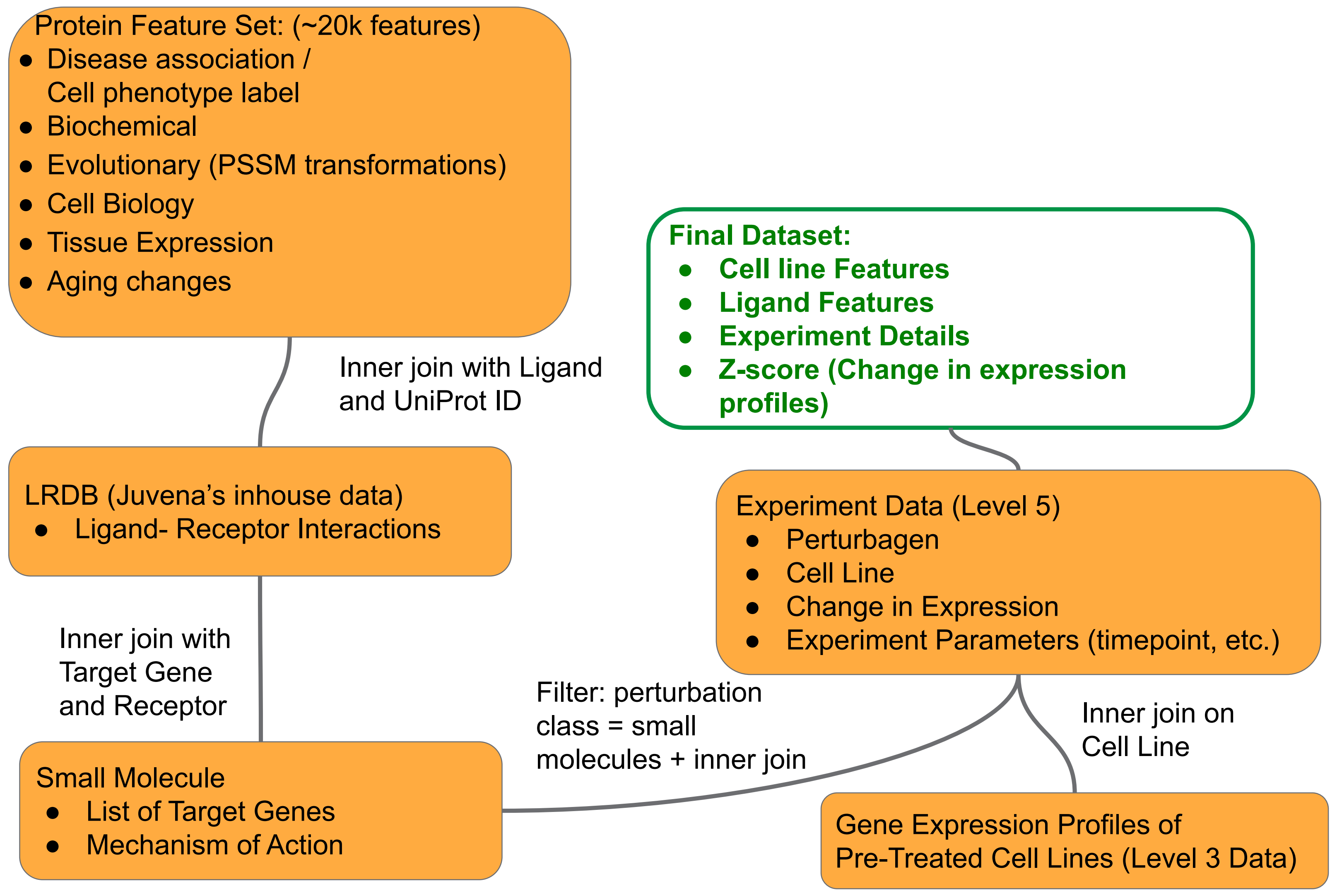


Figure 3:. Small Molecules with known gene targets are represented with Ligands that interact with the gene targets. One small molecule experiment is mapped to "m" Ligand combinations

- Out of ~19,000 small molecules, we filter small molecules with known targets and known mechanism of action.
- We restricted our analysis to inhibitor and agonist MOAs, to arrive at 256 unique perturbagens.
- Each small molecule is represented with the set of Ligands that interact with it's gene target

Methodology

- We augmented LINCS L-1000 with our in-house datasets to produce a dataset for predictive modeling
- **Potential to pre-train; fine tune on lab data**



Protein Feature Set: (~20k features)
- Disease association / Cell phenotype label
- Biochemical
- Evolutionary (PSSM transformations)
- Cell Biology
- Tissue Expression
- Aging changes

Inner join with Ligand and UniProt ID

LRDB (Juvena's inhouse data)
- Ligand- Receptor Interactions

Inner join with Target Gene and Receptor

Small Molecule
- List of Target Genes
- Mechanism of Action

Filter: perturbation class = small molecules + inner join

Final Dataset:
- **Cell line Features**
- **Ligand Features**
- **Experiment Details**
- **Z-score (Change in expression profiles)**

Experiment Data (Level 5)
- Perturbagen
- Cell Line
- Change in Expression
- Experiment Parameters (timepoint, etc.)

Inner join on Cell Line

Gene Expression Profiles of Pre-Treated Cell Lines (Level 3 Data)

Figure 4:. Dataset development - combining LINCS L1000, Juvena's inhouse database of gene/ protein interactions and Juvena's high dimensional protein feature base curated from different sources

- Dataset was segmented by gene readouts and mechanism of action of the small molecule target. For each segment, one cell line was held out in inference, and the model was trained on the remaining.
- Models used were combination of Random Forests, Gradient Boosted Decision Trees, Fully connected neural networks, where the best model/ hyperparameters were chosen.
- Performance Metric: $\frac{||w^T.(y_{true} - y_{best\_model})^2||_2^2}{||w^T.(y_{true} - y_{heuristic})^2||_2^2}$

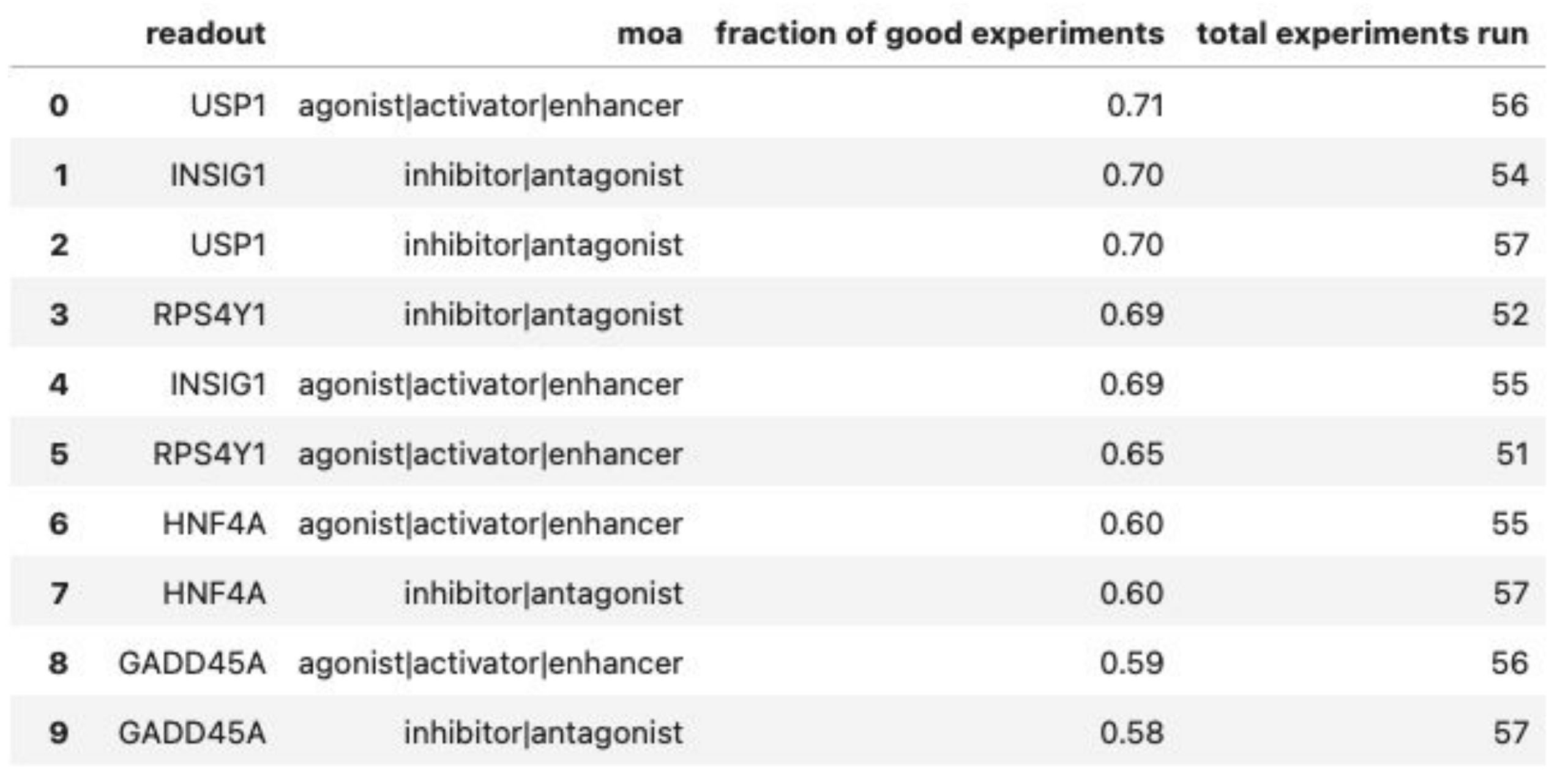
where

$$y_{heuristic} = \mathbb{E}_{training\_set}(zscore)$$

  - The weights are set to the squared value of true gene expression change thereby driving sensitivity towards hits.

Results

- We ran predictions across 10,466 combinations of gene expression readouts, target MOA and inferenced cell line.
- 36% of the combinations showed an > 20% improvement over the heuristic model



	readout	moa	fraction of good experiments	total experiments run
0	USP1	agonist|activator|enhancer	0.71	56
1	INSIG1	inhibitor|antagonist	0.70	54
2	USP1	inhibitor|antagonist	0.70	57
3	RPS4Y1	inhibitor|antagonist	0.69	52
4	INSIG1	agonist|activator|enhancer	0.69	55
5	RPS4Y1	agonist|activator|enhancer	0.65	51
6	HNF4A	agonist|activator|enhancer	0.60	55
7	HNF4A	inhibitor|antagonist	0.60	57
8	GADD45A	agonist|activator|enhancer	0.59	56
9	GADD45A	inhibitor|antagonist	0.58	57

Table 1: (Gene readout, MOA) combinations with maximum number of inferenced cell lines showing > 20% improvement over heuristic models. Similar performance in inference-ability of unseen cell lines across MOAs

Results

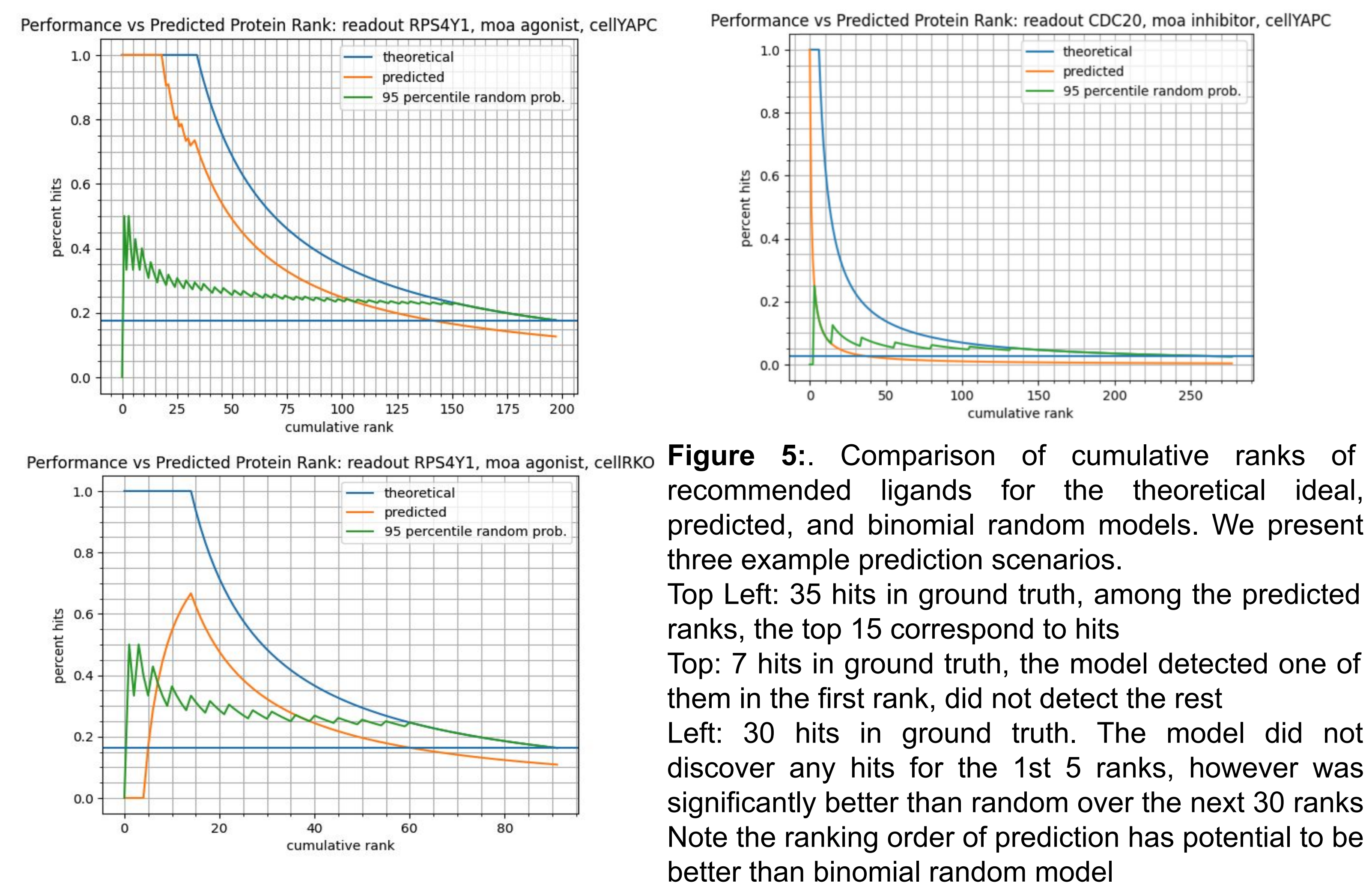


Figure 5:. Comparison of cumulative ranks of recommended ligands for the theoretical ideal, predicted, and binomial random models. We present three example prediction scenarios.
Top Left: 35 hits in ground truth, among the predicted ranks, the top 15 correspond to hits
Top: 7 hits in ground truth, the model detected one of them in the first rank, did not detect the rest
Left: 30 hits in ground truth. The model did not discover any hits for the 1st 5 ranks, however was significantly better than random over the next 30 ranks
Note the ranking order of prediction has potential to be better than binomial random model

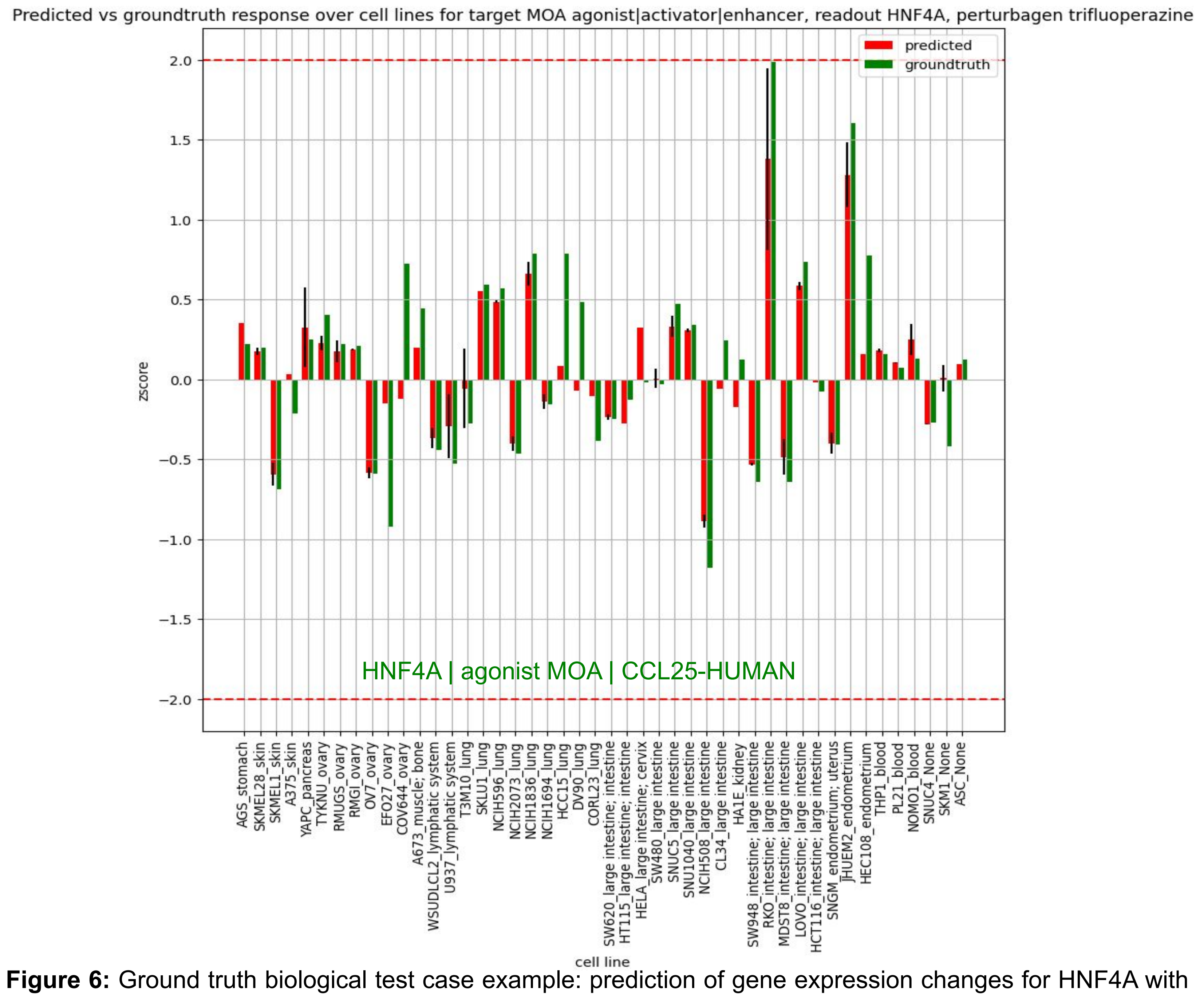


Figure 6: Ground truth biological test case example: prediction of gene expression changes for HNF4A with treatment of trifluoperazine (DRD2 target). Large effects are observed in digestive system cell types consistent with published literature (drug is used as antiemetic, also approved as anti-psychotic). Model predictions capture both direction and magnitude of gene expression change

Conclusions and Next Steps

- Transcriptomic features capture features of cell lines allowing discriminative ability and generalizability to unknown cell lines. We have identified key gene readouts for measuring perturbagen effects.
- Next steps: Pre-train a model on developed dataset; fine-tune on in vitro data from lab.

References

Lv, Duo-Duo, Ling-Yun Zhou, and Hong Tang. "Hepatocyte nuclear factor 4α and cancer-related cell signaling pathways: a promising insight into cancer treatment." *Experimental & molecular medicine* 53.1 (2021)

Keinan, Gal, et al. "Learning Perturbation-specific Cell Representations for Prediction of Transcriptional Response across Cellular Contexts." *bioRxiv* (2023): 2023-03.

tganpathi@juvenatherapeutics.com

https://www.juvenatherapeutics.com

joconnell@juvenatherapeutics.com