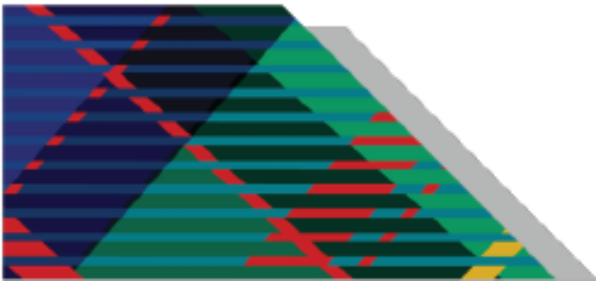
Computational Biology Department

Carnegie Mellon University
School of Computer Science

An Improved Matrix Completion Algorithm For Categorical Variables: Application to Active Learning of Drug Responses

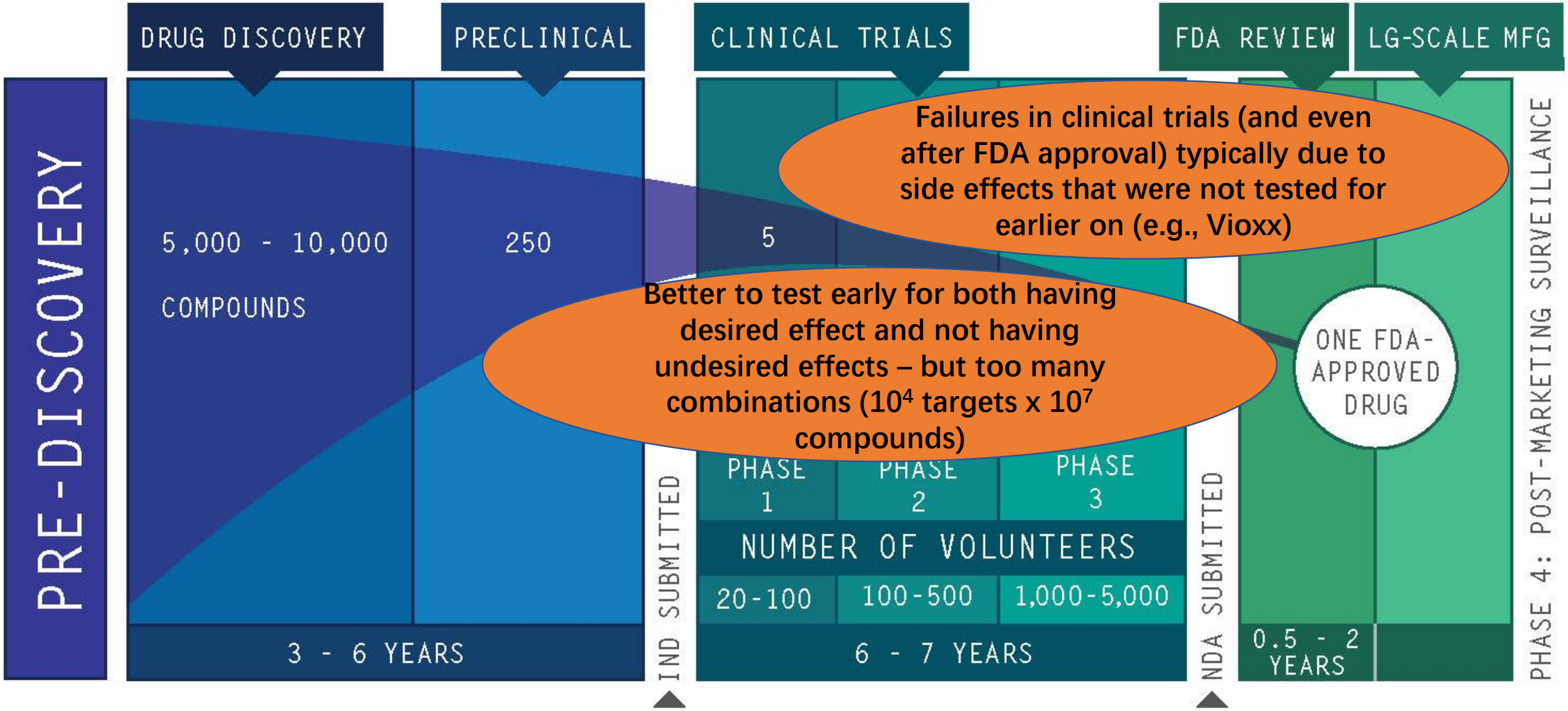
Huangqingbo Sun, Robert F. Murphy

Workshop on Real World Experiment Design and Active Learning at ICML 2020



ICML | 2020
Thirty-seventh International
Conference on Machine Learning

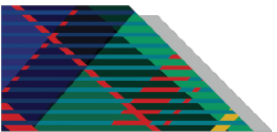
Drug Discovery Funnel



Source: PhRMA

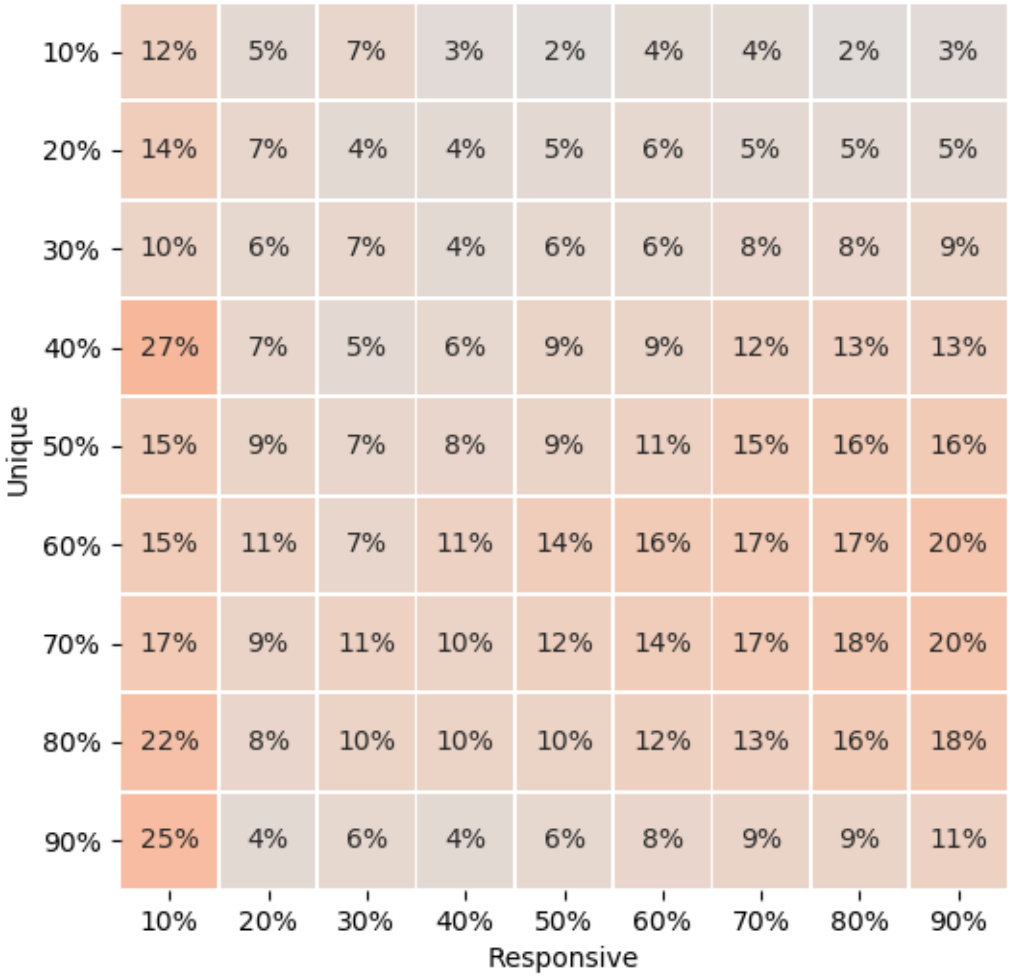
Active Learning - Multiple Phenotypes

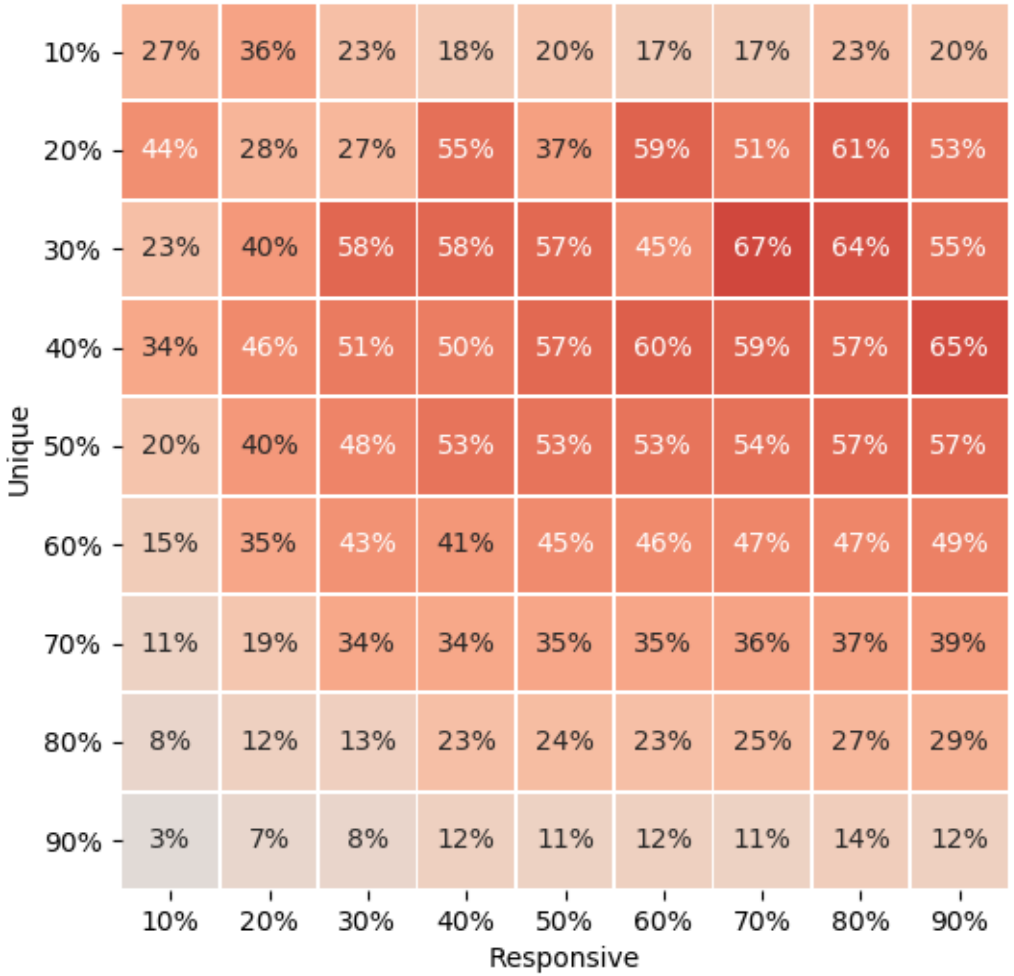
- Solution is active learning of a predictive model of all compound effects on all targets
- But there are also many possible effects that compounds could have on a given target – thus effects are categorical variables
- Assume that there are some similarities in effects among compounds and targets
- Predictive model: completion (imputation) of a very sparse (only a few observed entries) categorical matrix
- For active learning, uncertainty sampling is adopted, with 3 query strategies.



Experiment on Synthetic Data

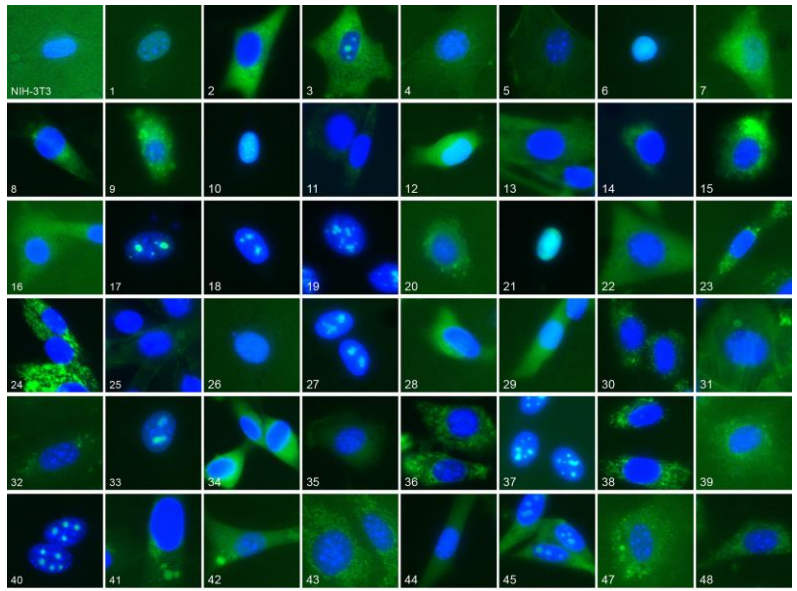
- How fast does Active Learning comparing to random selection?





Performance was measured as the difference in the number of batches to achieve 100% (right) or 90% (left) accuracy between active and random selection.

Experiment Using Microscope Images for Many Drugs and Targets



Learn the effect of 92 drugs on 94 GFP-tagged proteins without doing experiments for all drugs and proteins with the help of Active Learning.

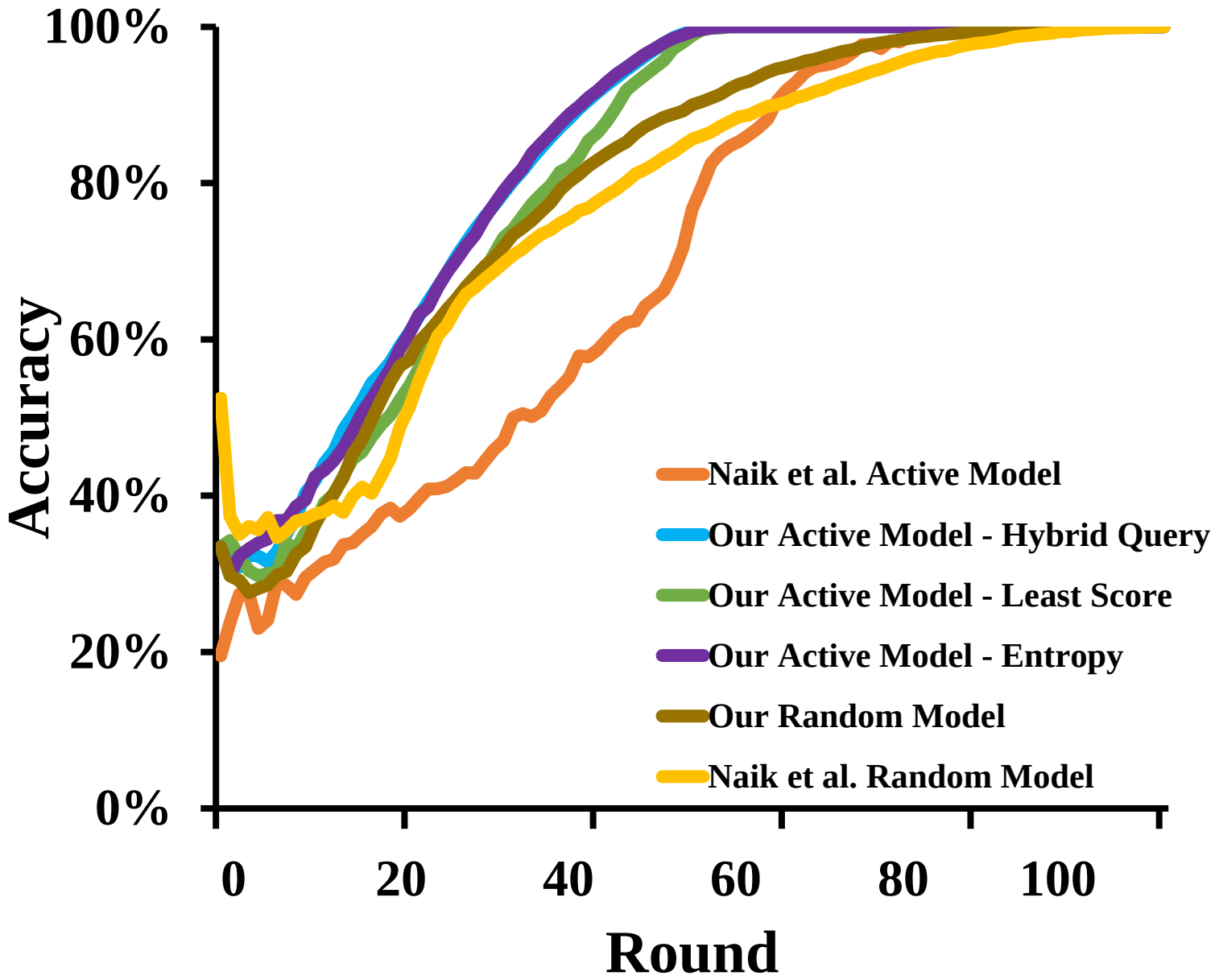


Image source: Naik et al.

Conclusions

- Improved clustering-based, “lazy learning” matrix completion algorithm for categorical matrices.
- Results in improved active learning performance over previous methods.