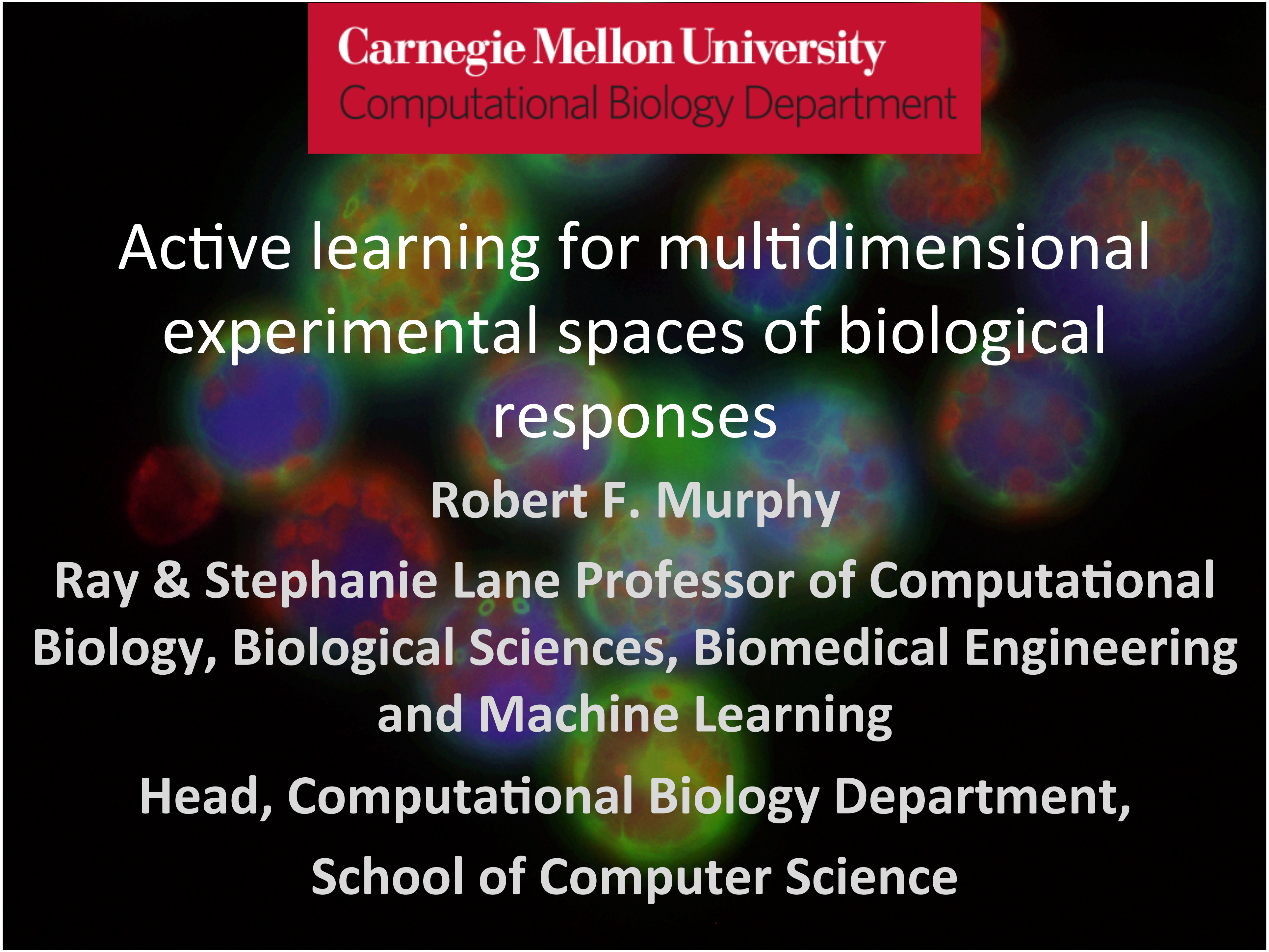


Carnegie Mellon University
Computational Biology Department

Active learning for multidimensional experimental spaces of biological responses

Robert F. Murphy

Ray & Stephanie Lane Professor of Computational Biology, Biological Sciences, Biomedical Engineering and Machine Learning

Head, Computational Biology Department,

School of Computer Science

Big problems, Little Data: Drug Development

- Diseases can be extremely heterogeneous and based on many factors (e.g., diabetes)
- Drug effects can be very different depending on the patient and disease
- Ideally, need to know how all drugs will affect all diseases in all patients
- Too many combinations to measure everything

Further…

- Leading cause of drug failures in early development is not lack of effectiveness but safety concerns (and in late development, discovery of undesirable side effects)
- Drug development is not just about finding compounds that hit a desired target-also about finding compounds that miss all other targets

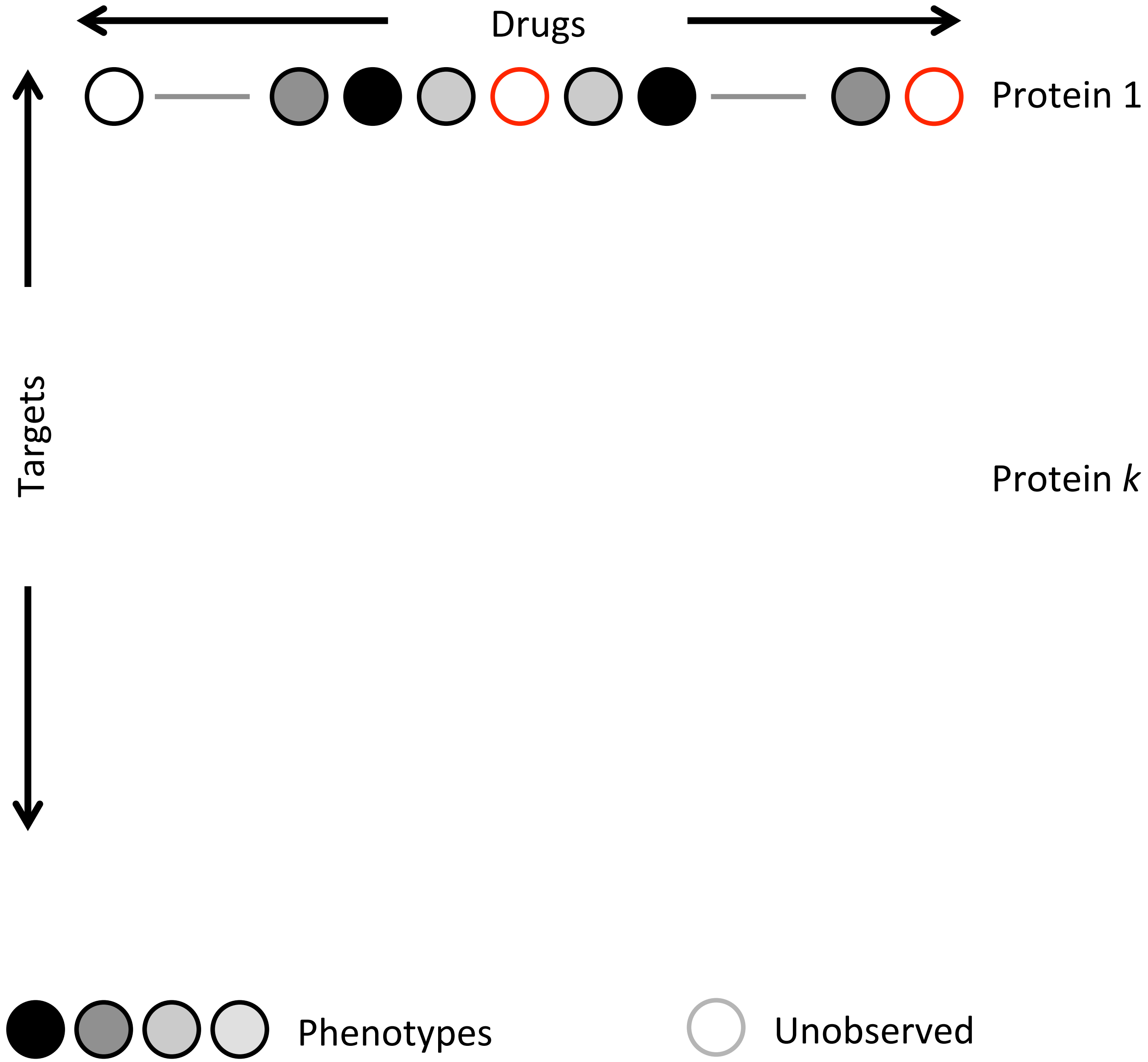
Big problems, Little Data: Basic Biological Research

- Cells/Tissues/Organisms are complex systems without rules/laws
- Every process/cell type/organelle/protein may be affected by drugs, gene variation, environment
- Need to learn all of these changes
- Millions of potential perturbations/gene variations, tens of thousands of proteins, hundreds of cell types

Predictive modeling

- Need to learn a complete matrix/tensor to show whether a particular drug affects a particular target in a particular genotype
- Same for which genes affect which metabolic processes, etc.
- Try to learn the matrix without doing all experiments
- Measure some and build a predictive model for the rest
- But which measurements should be done?

Current practice: consider each target separately



Drugs

Protein 1

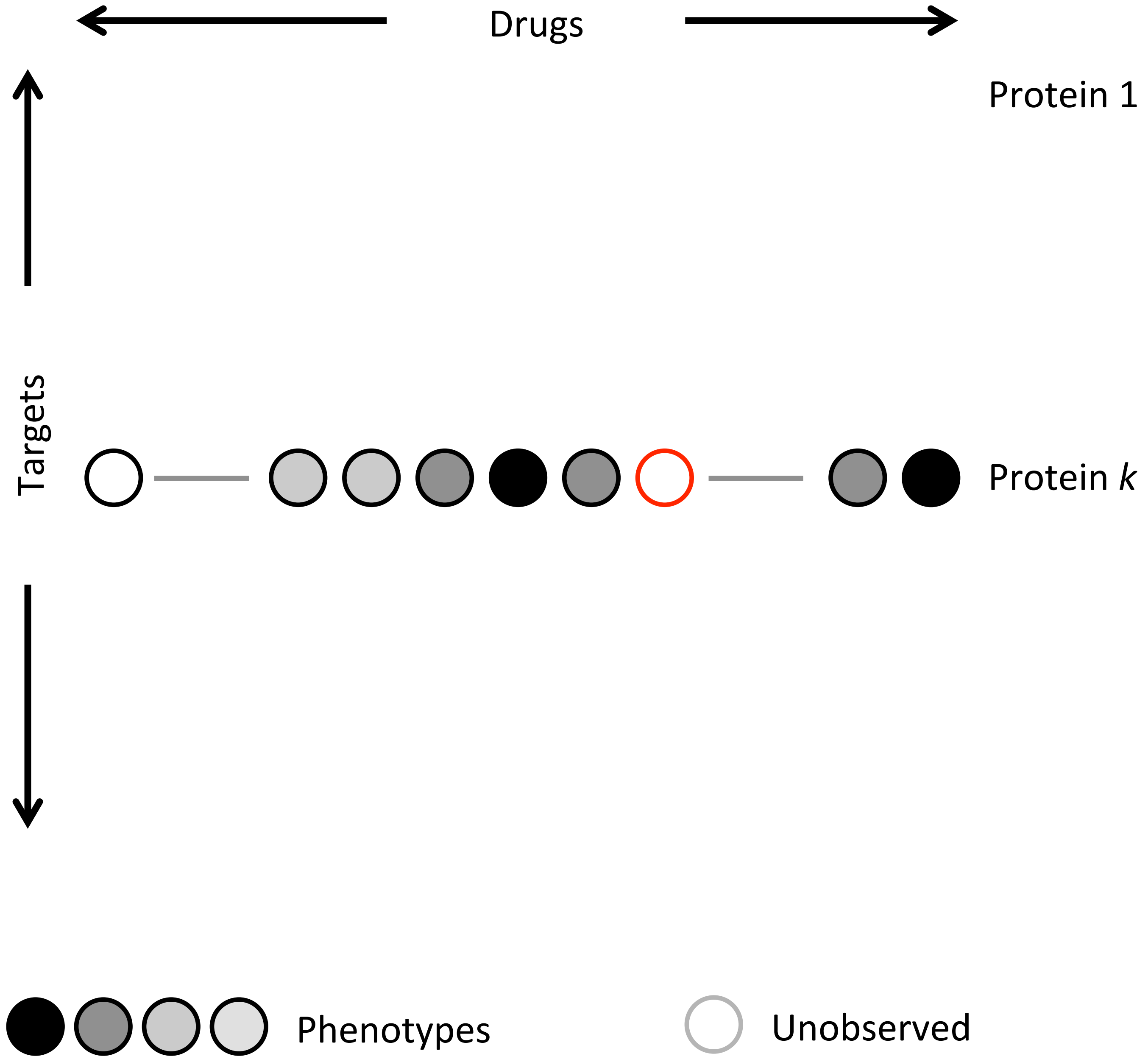
Targets

Protein *k*

Phenotypes

Unobserved

Current practice: consider each target separately



Drugs

Protein 1

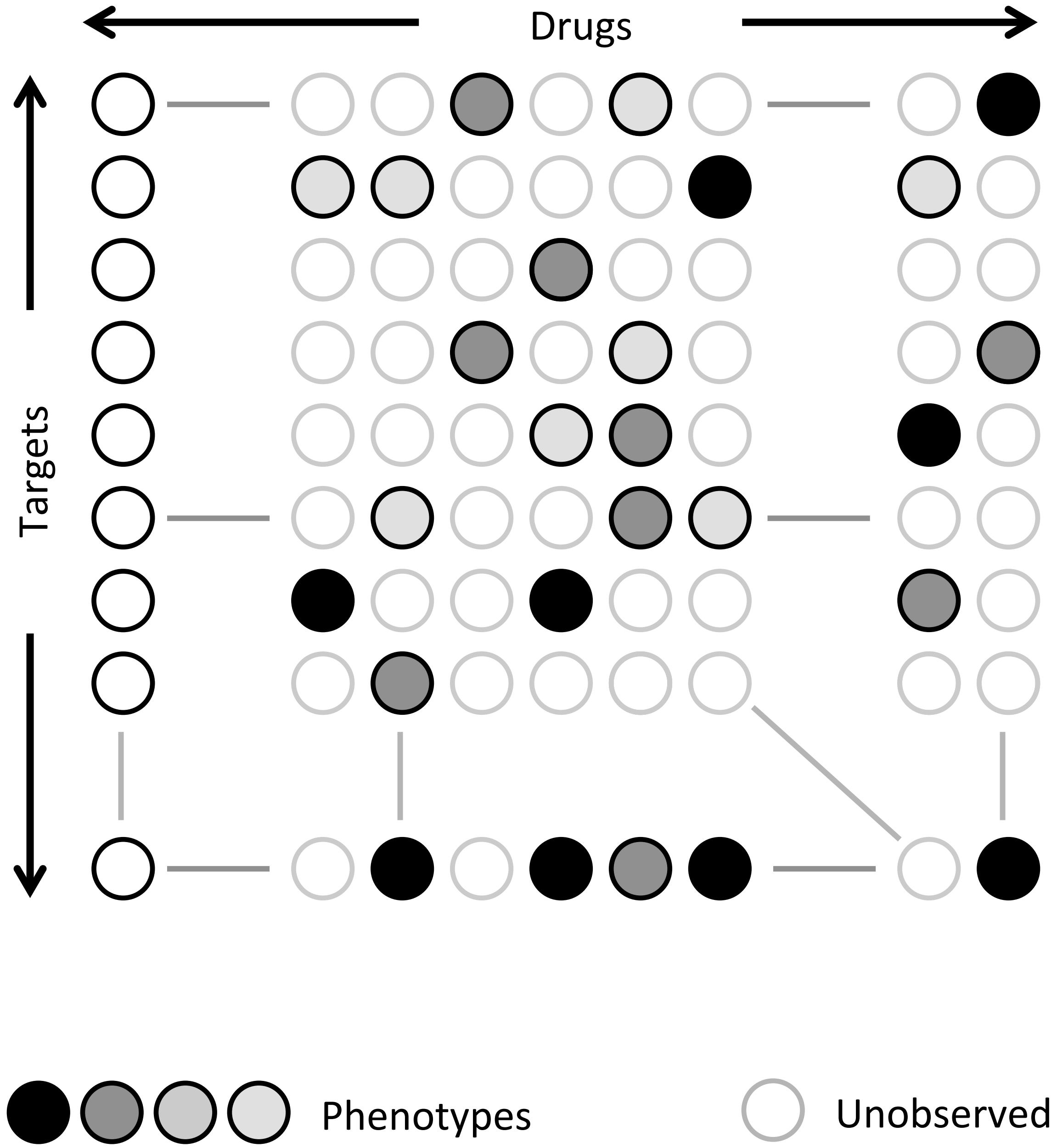
Targets

Protein *k*

Phenotypes

Unobserved

Can set up a Sparse, Matrix Factorization/Completion Problem



Dempster et al (1977)
Hill et al. (1995);
Lee & Seung (1999);
Buchanan & Fitzgibbon (2005);
Salakhutdinov & Mnih (2008);
Mitra (2010);
Gönen (2012); …

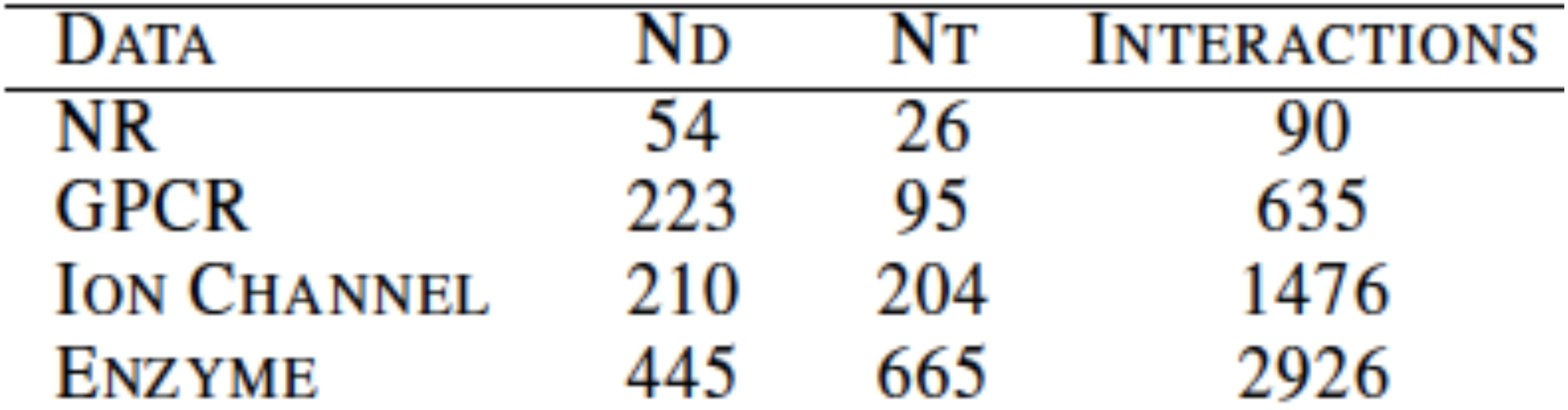
Three considerations

- How much/what data is missing?
  - Little: **matrix completion** (passive learning)
  - None for some, all for others: **matrix factorization**
  - Most/all: **<u>not addressed</u>** (need active learning)
- Any basis for *ab initio* predictions?
  - **Yes**, features for drugs/targets
    - e.g., chemical fingerprints
  - **<u>No</u>**
- What are we predicting?
  - **Real values or binary values** (all prior work)
  - **<u>Classes</u>**

Retrospective Studies

- Widely used for demonstrating "real world-applicability" of methods
- Always concern about generalizability of results due to possibility of making model choices using testing data
- In drug effects space, mostly done with small datasets (50-500 drugs, 20-700 targets) for which complete data was available

Use Curated Drug Interaction Datasets

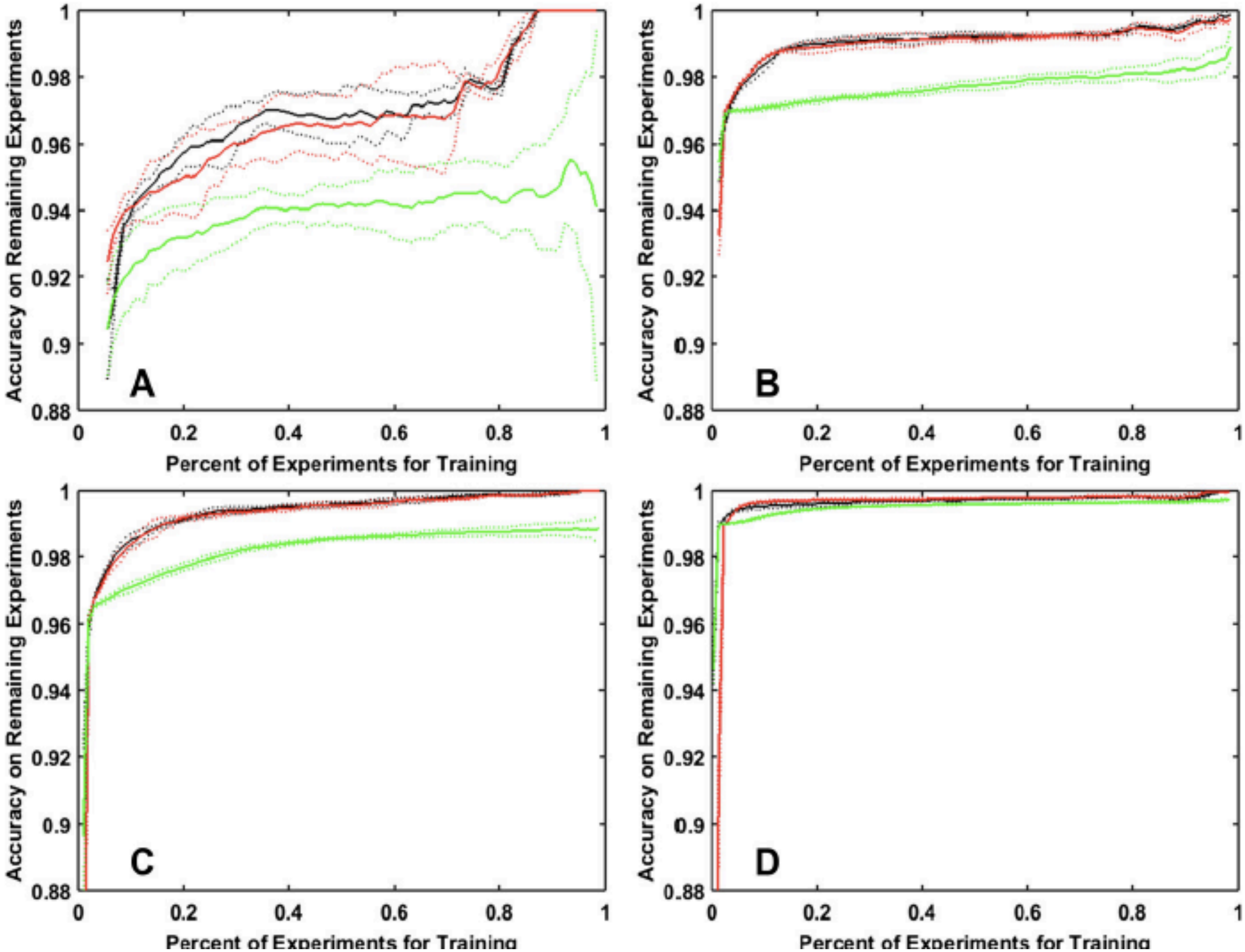


DATA	ND	NT	INTERACTIONS
NR	54	26	90
GPCR	223	95	635
ION CHANNEL	210	204	1476
ENZYME	445	665	2926

Previous studies (e.g., Gönen 2012)
tested ability to predict for 20% of
drugs using training with 80%

Temerinac-Ott, Naik & Murphy, BMC Bioinformatics 2015

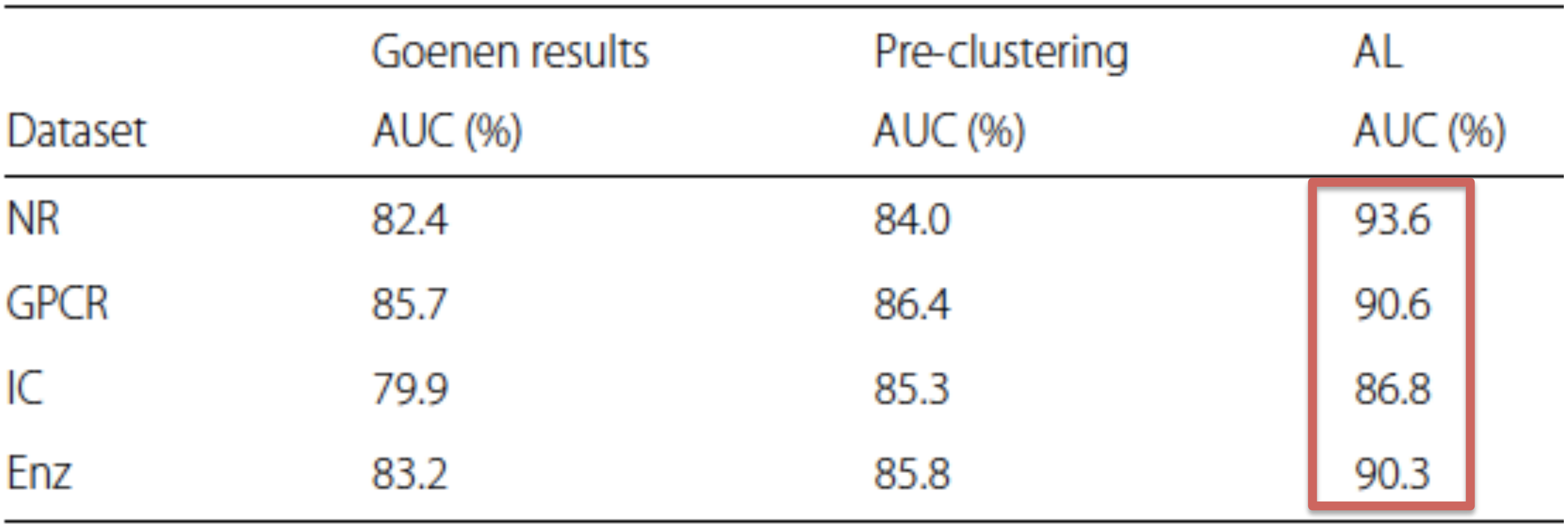
Active learning of factored models



Green: random

Black,Red: active

Active Learning of 80% compared to random or clustering by features



	Goenen results	Pre-clustering	AL
Dataset	AUC (%)	AUC (%)	AUC (%)
NR	82.4	84.0	93.6
GPCR	85.7	86.4	90.6
IC	79.9	85.3	86.8
Enz	83.2	85.8	90.3

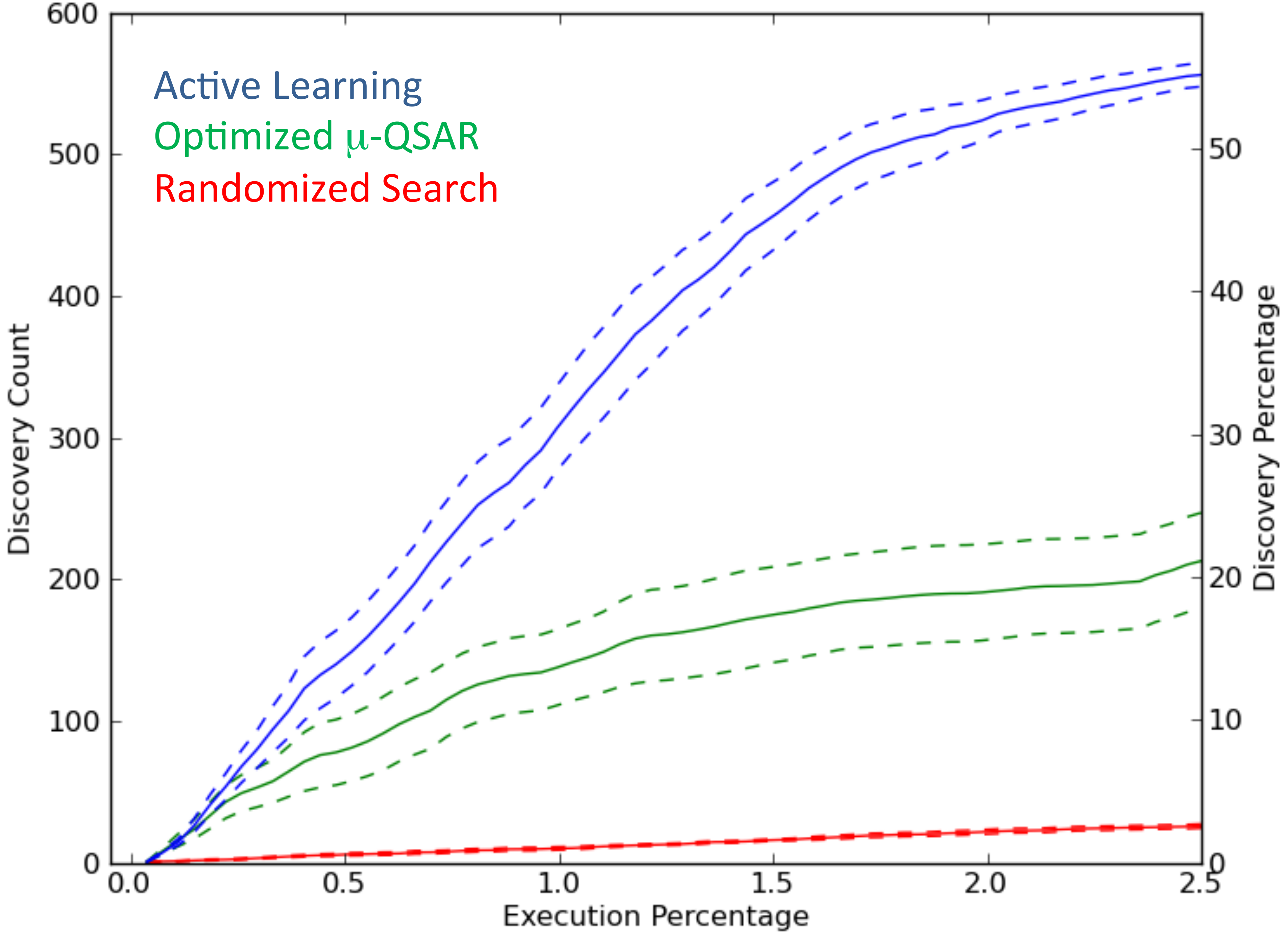
Kangas, Naik, Murphy, BMC Bioinformatics 2014

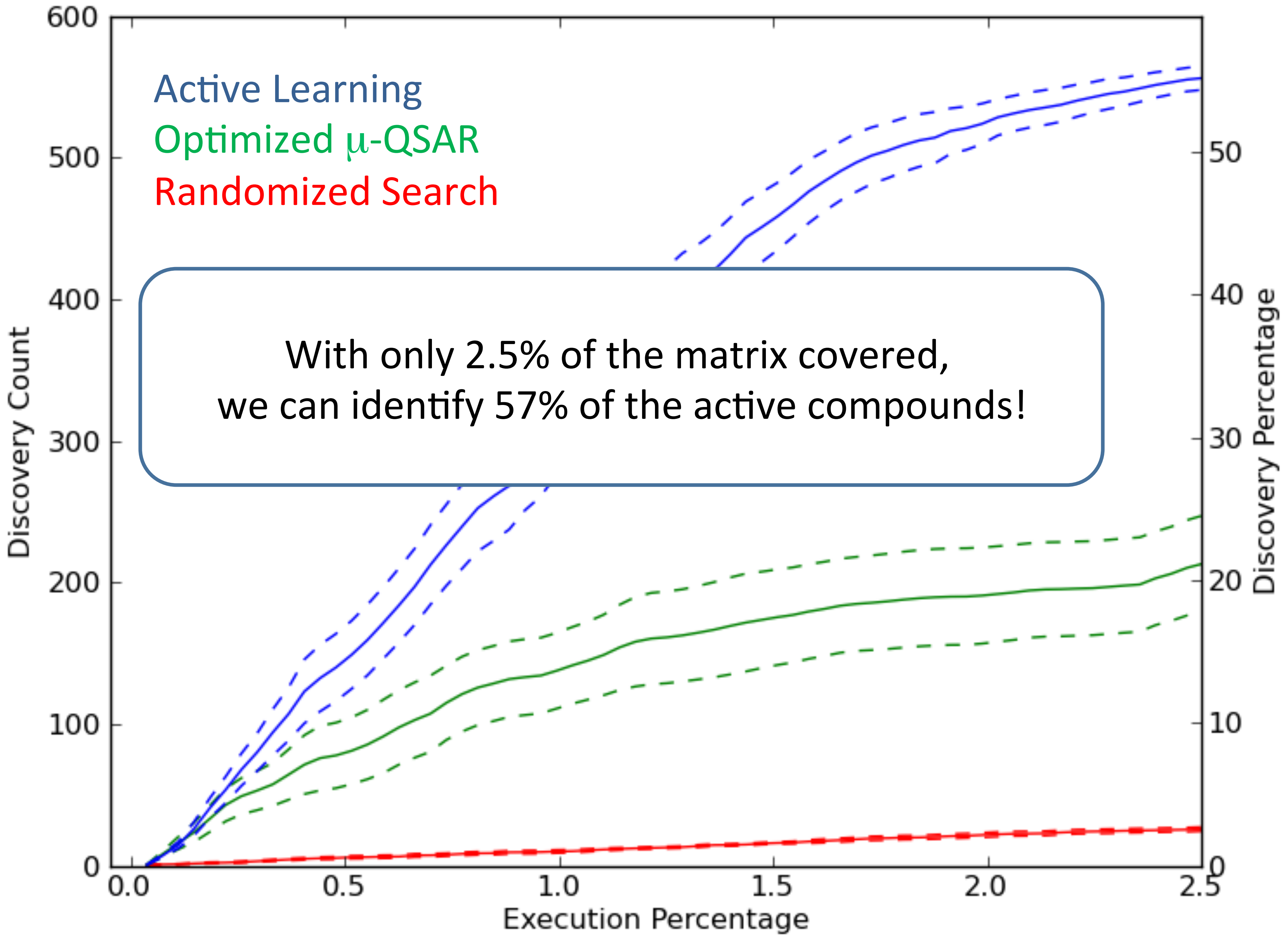
Use subset of PubChem Data

- Assays: 177
- Unique Protein Targets: 133
- Compounds: 20,000
- Experiments: ~1,000,000 (30% coverage)
- Use features to measure similarity between drugs and between targets
- Compare discovery rate across different methods
  - Discovery: a drug-protein pair whose |rank score| > 80

Sparse model

- Need model that can be built from very limited data during initial acquisition
- Used LASSO models for each target and for each drug, average predictions from each
- Used 50% greedy/50% uncertainty hybrid
- Used “memory limitation” to focus learning models from recently acquired data


Active Learning
Optimized μ-QSAR
Randomized Search
Discovery Count
600
500
400
300
200
100
0
0.0
0.5
1.0
1.5
2.0
2.5
Execution Percentage
50
40
30
20
10
0
Discovery Percentage



With only 2.5% of the matrix covered,
we can identify 57% of the active compounds!

- These methods are based on
  - having estimates of similarity among drugs and/or targets (normally both), typically from descriptive features
    - permits predictions to be made about drugs or targets for which few or no experiments have been done
  - having binary or real-valued experimental outputs
- What do we do when
  - features are not reliable, or not possible
  - outputs are multidimensional?

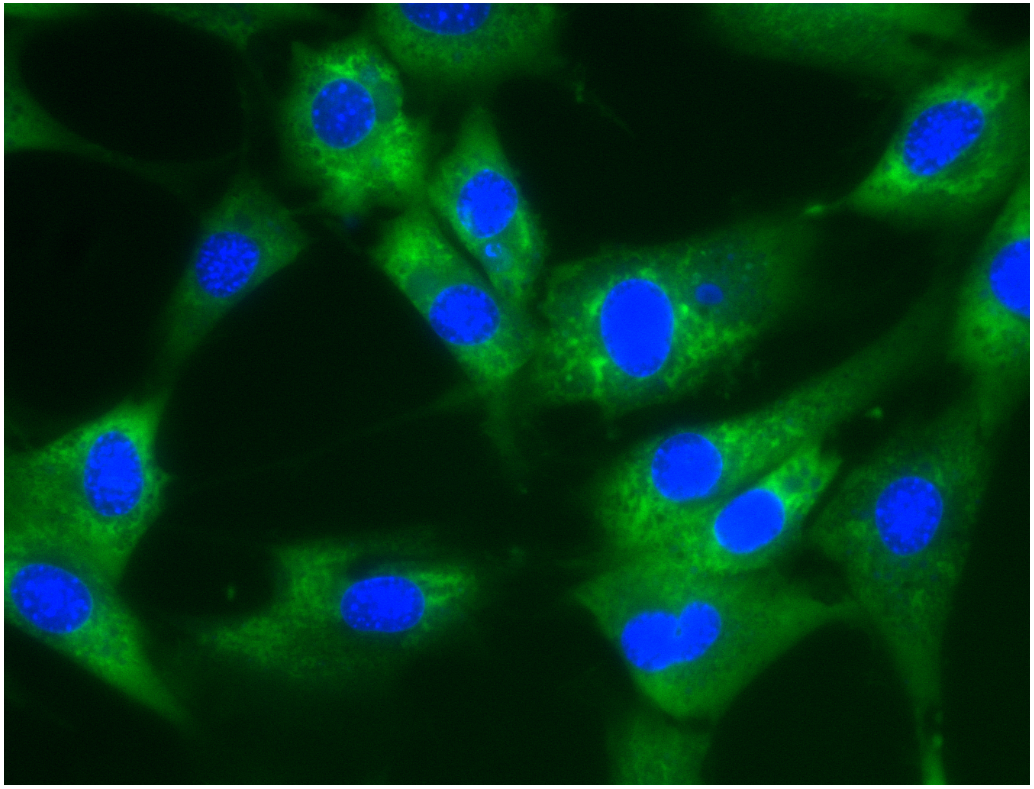
Example: image-based screening

Drugs

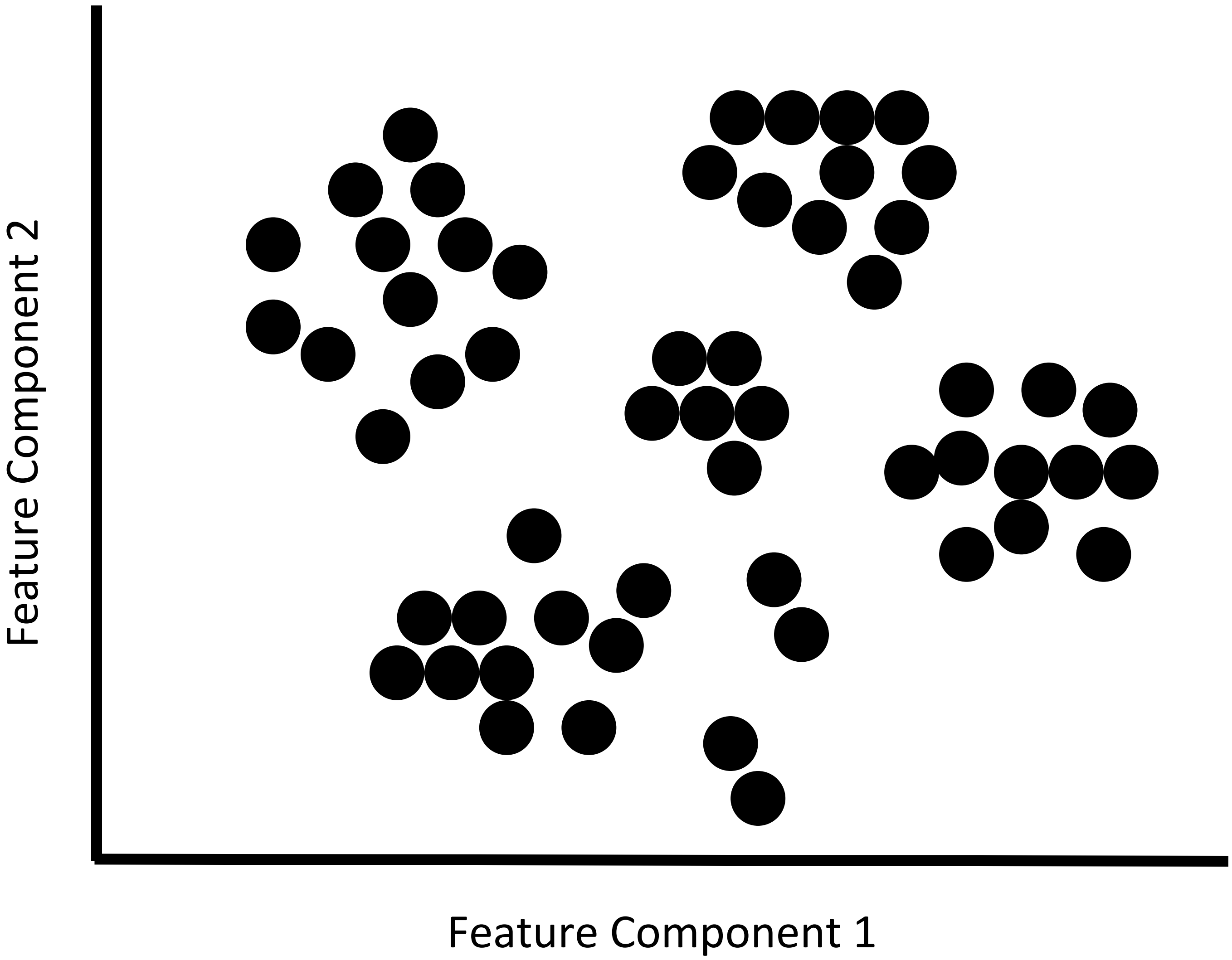
Proteins

Drug *j*

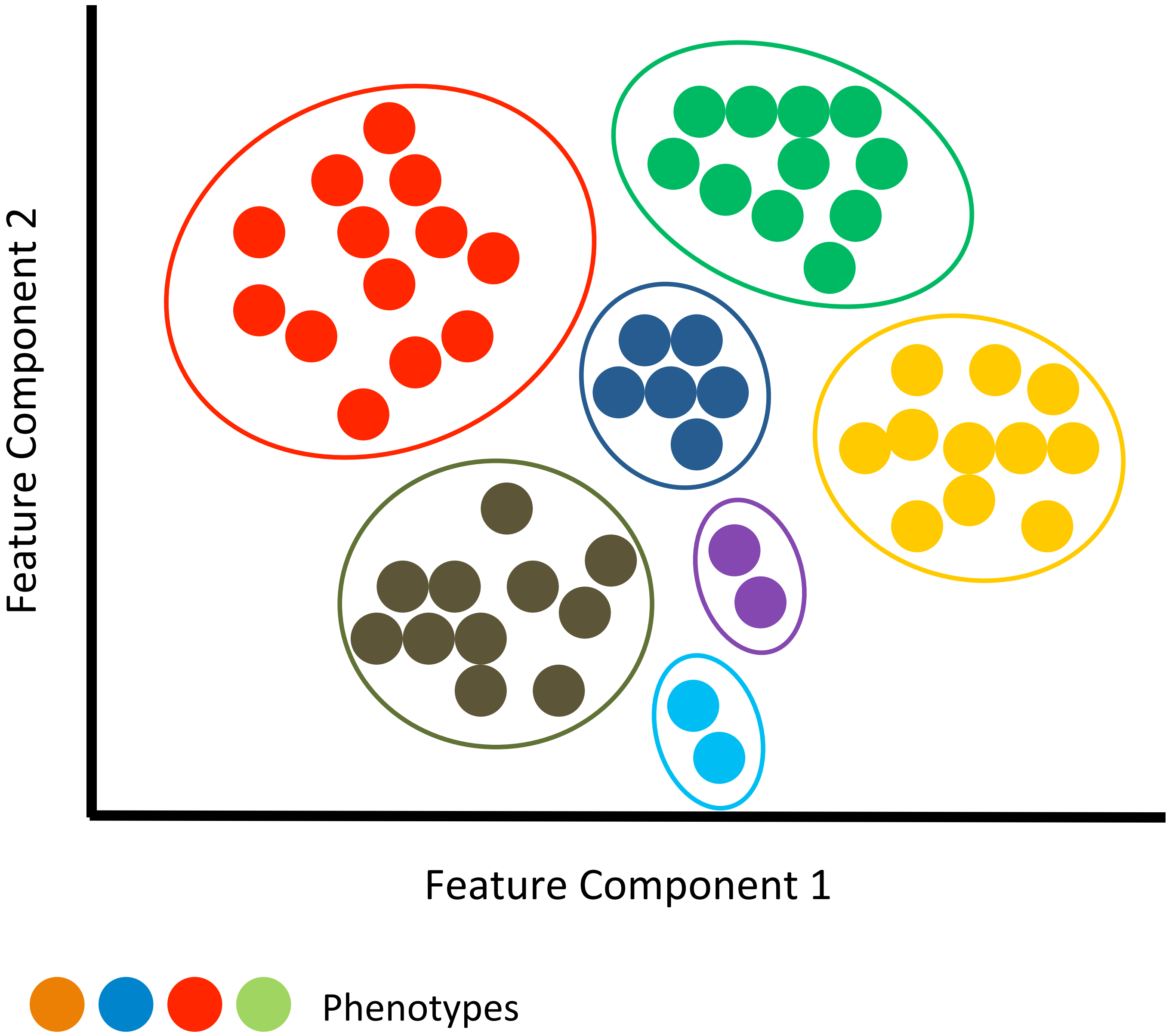
Protein *k* ●



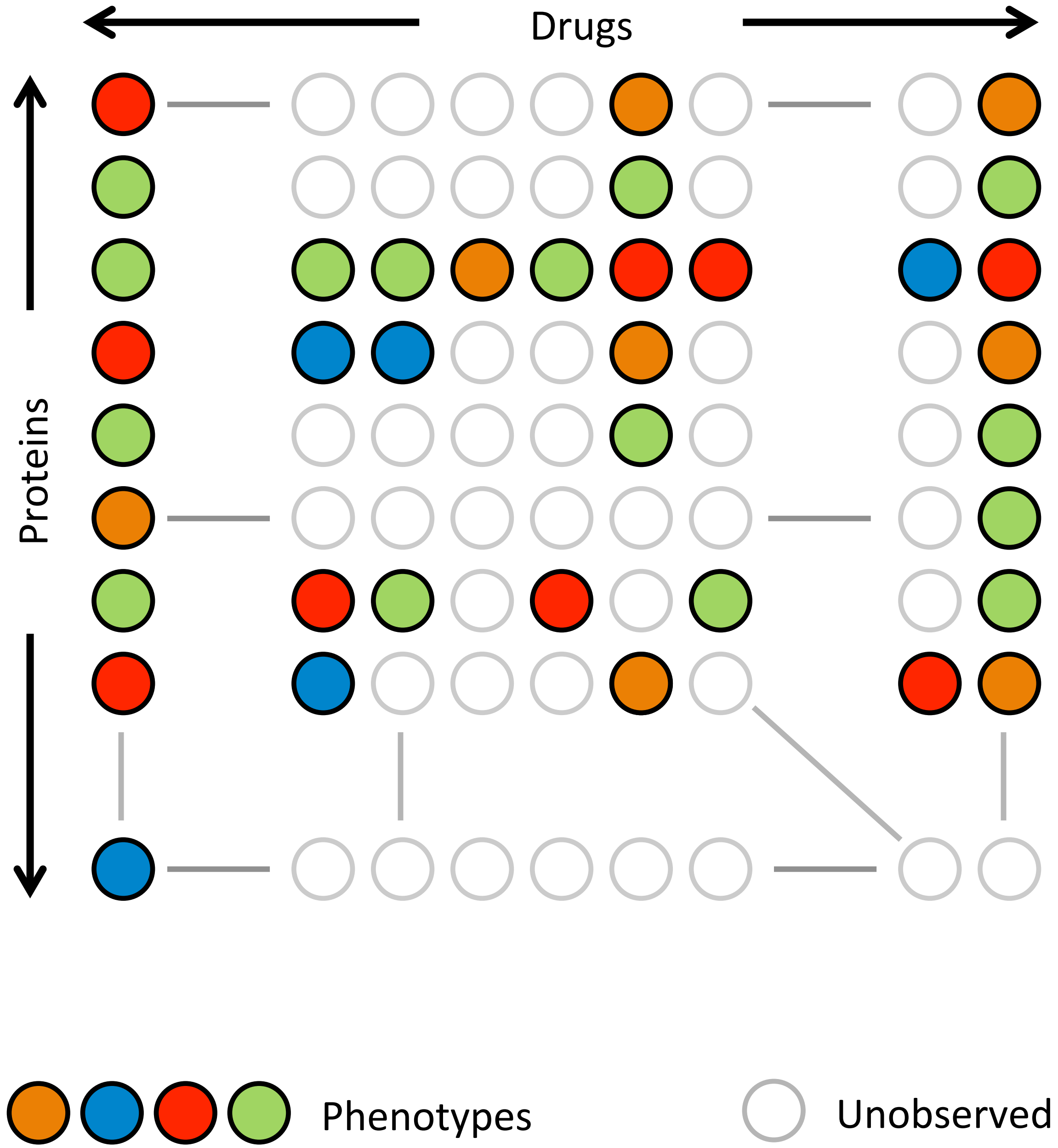
Consider each experiment in a Feature Space



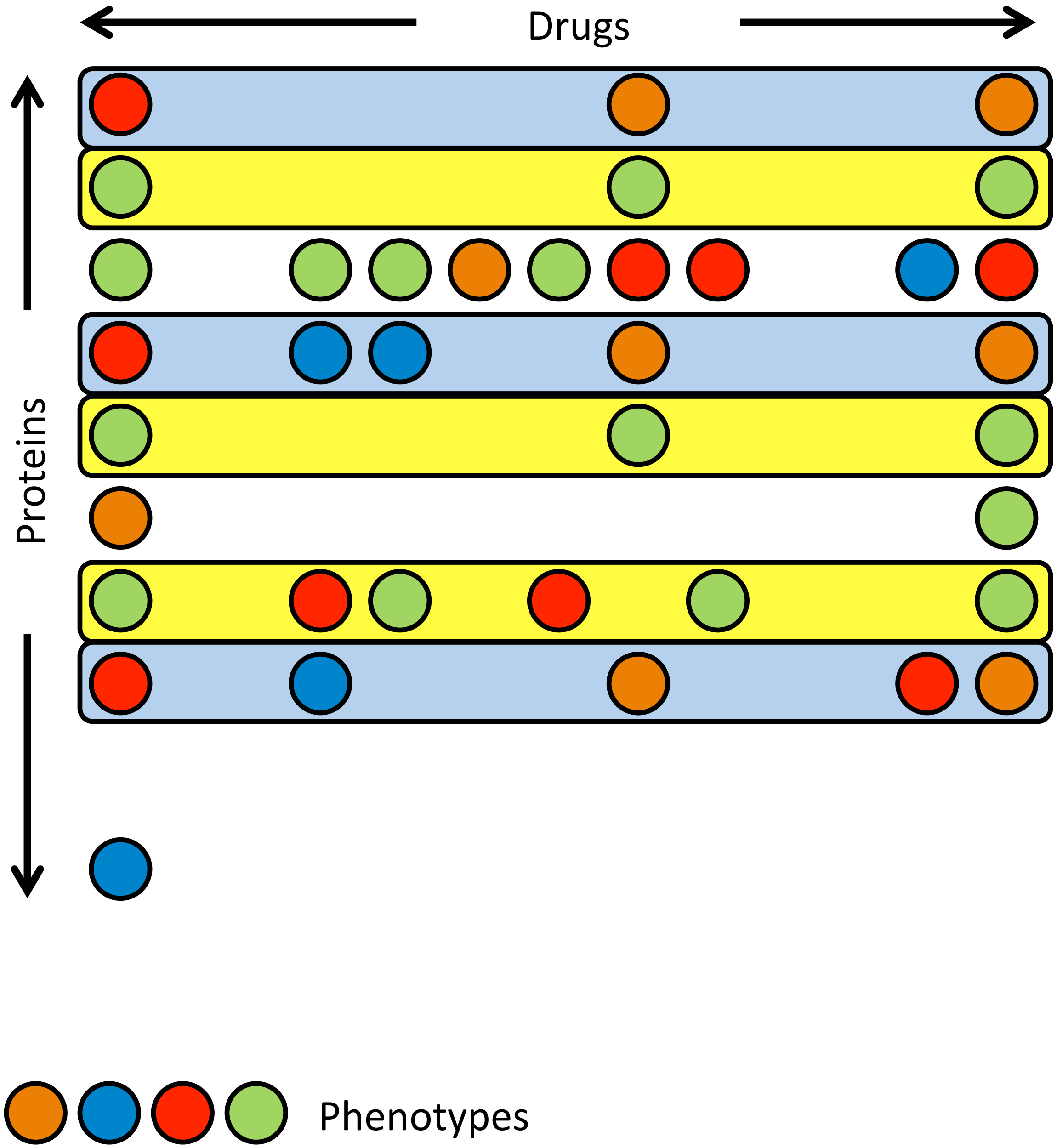
Cluster to form Phenotypes



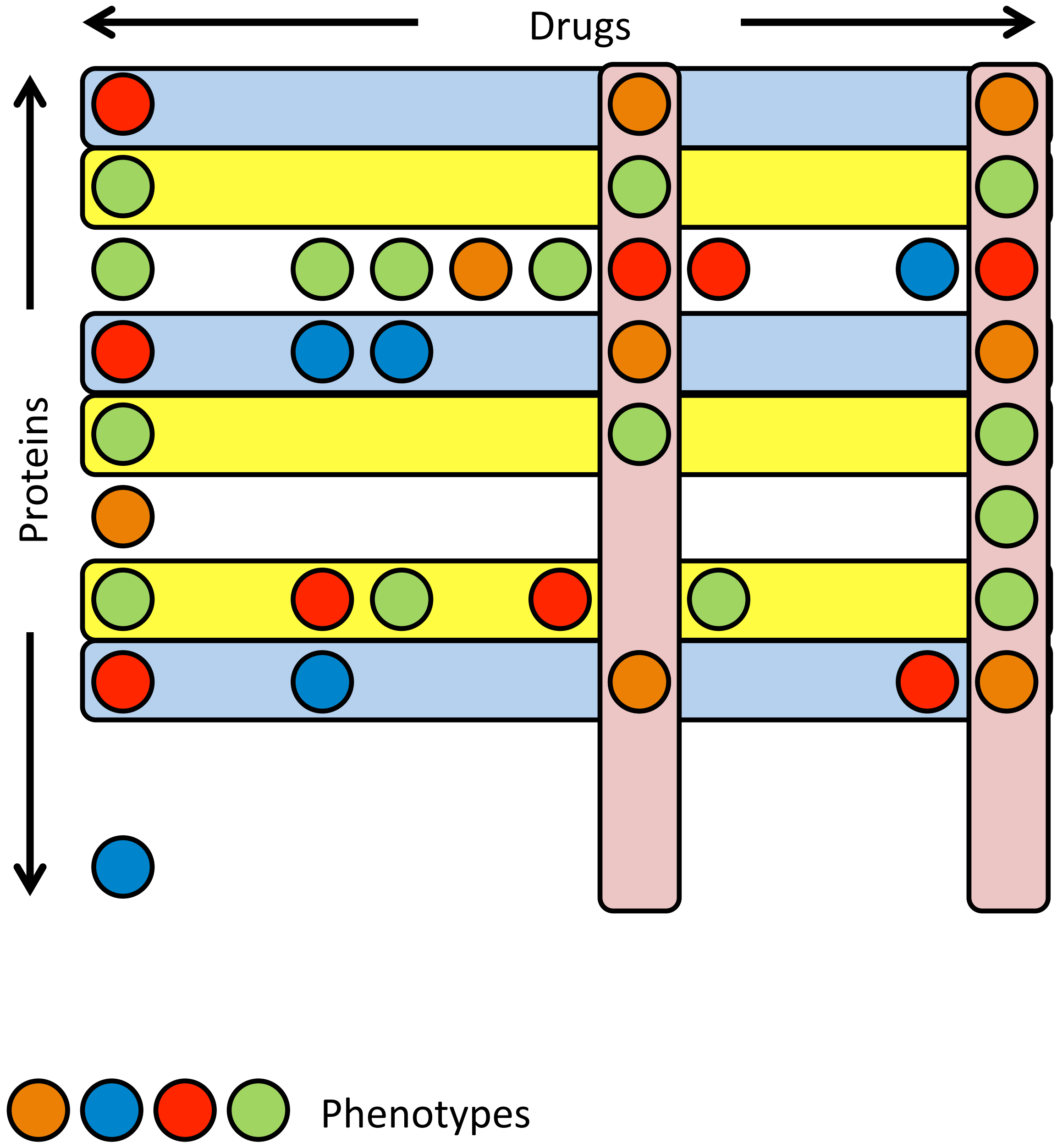
How do we form Predictions for Unobserved Experiments?



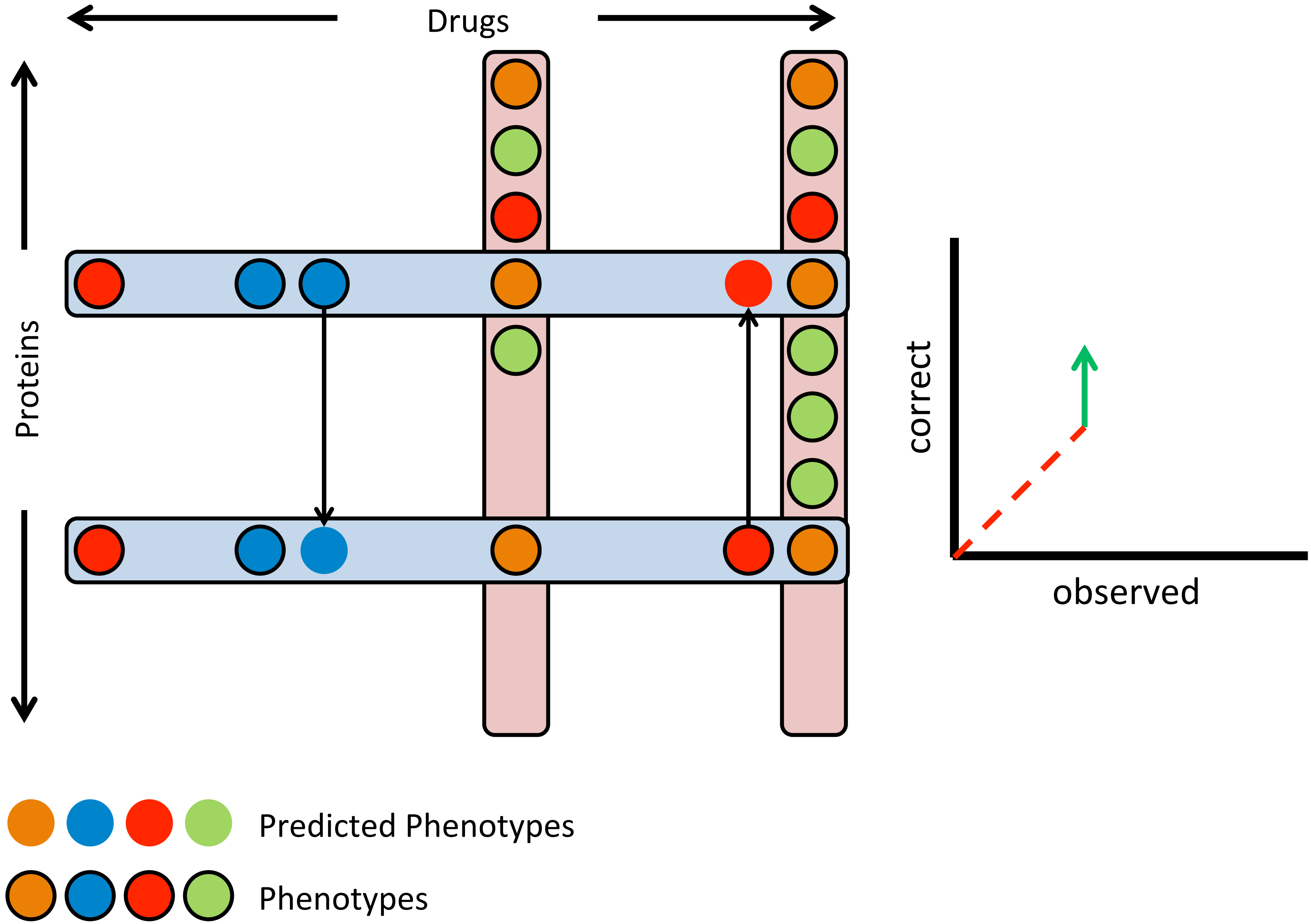
Identify Proteins with Similar Responses to Drugs



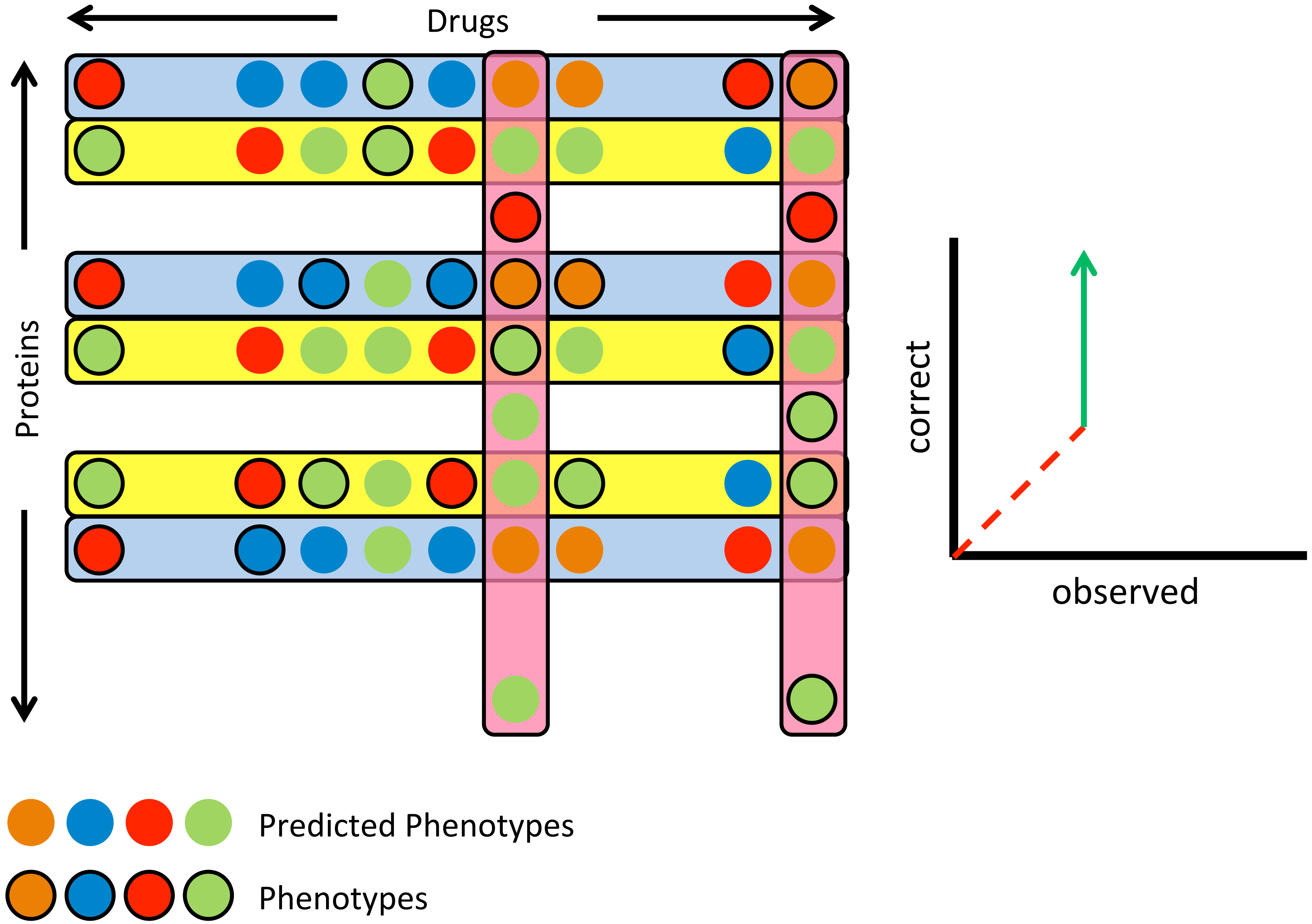
Identify Drugs with Similar Effects on Proteins



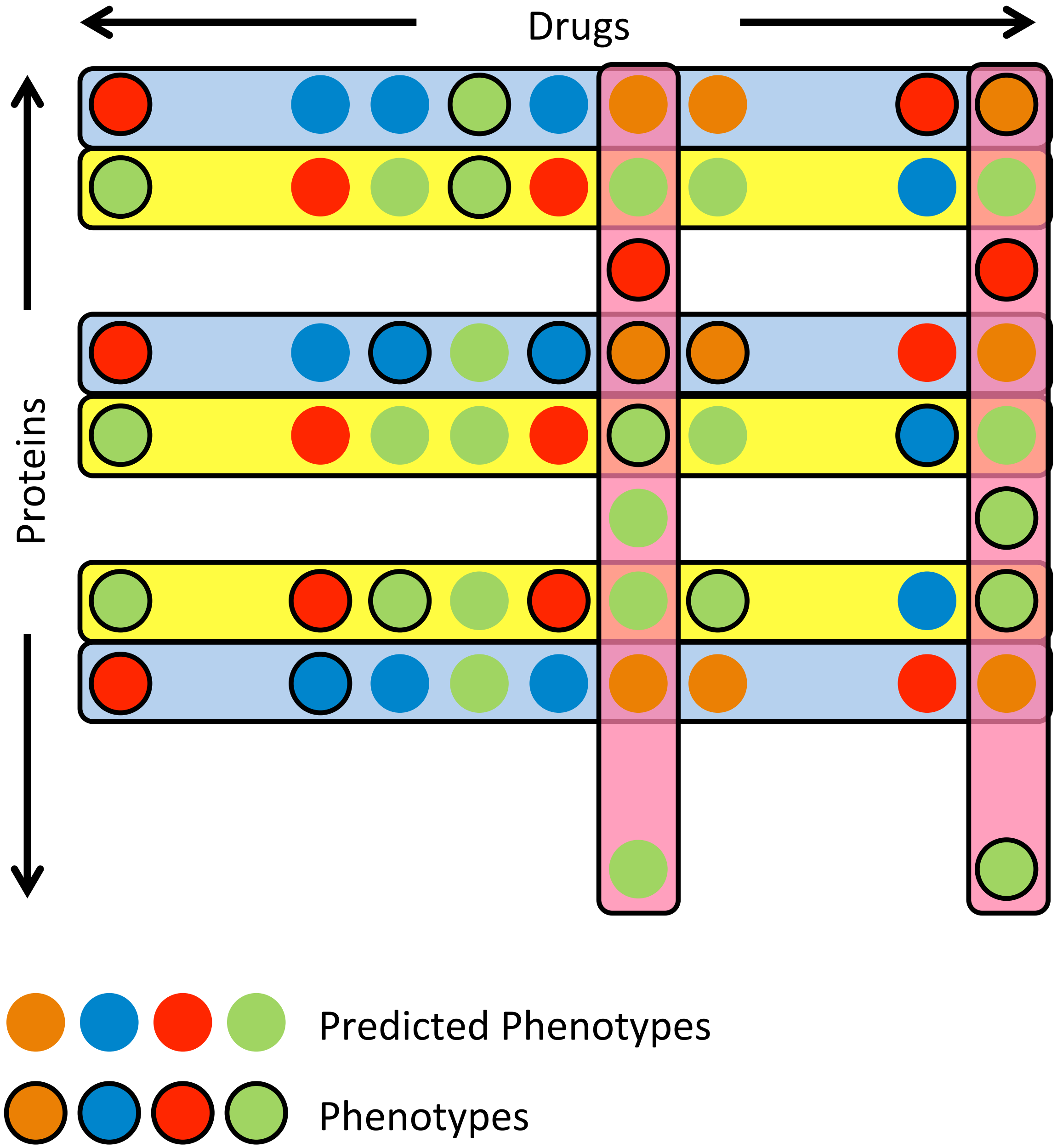
Use Similarities to Predict (matrix factorization without prior kernels)



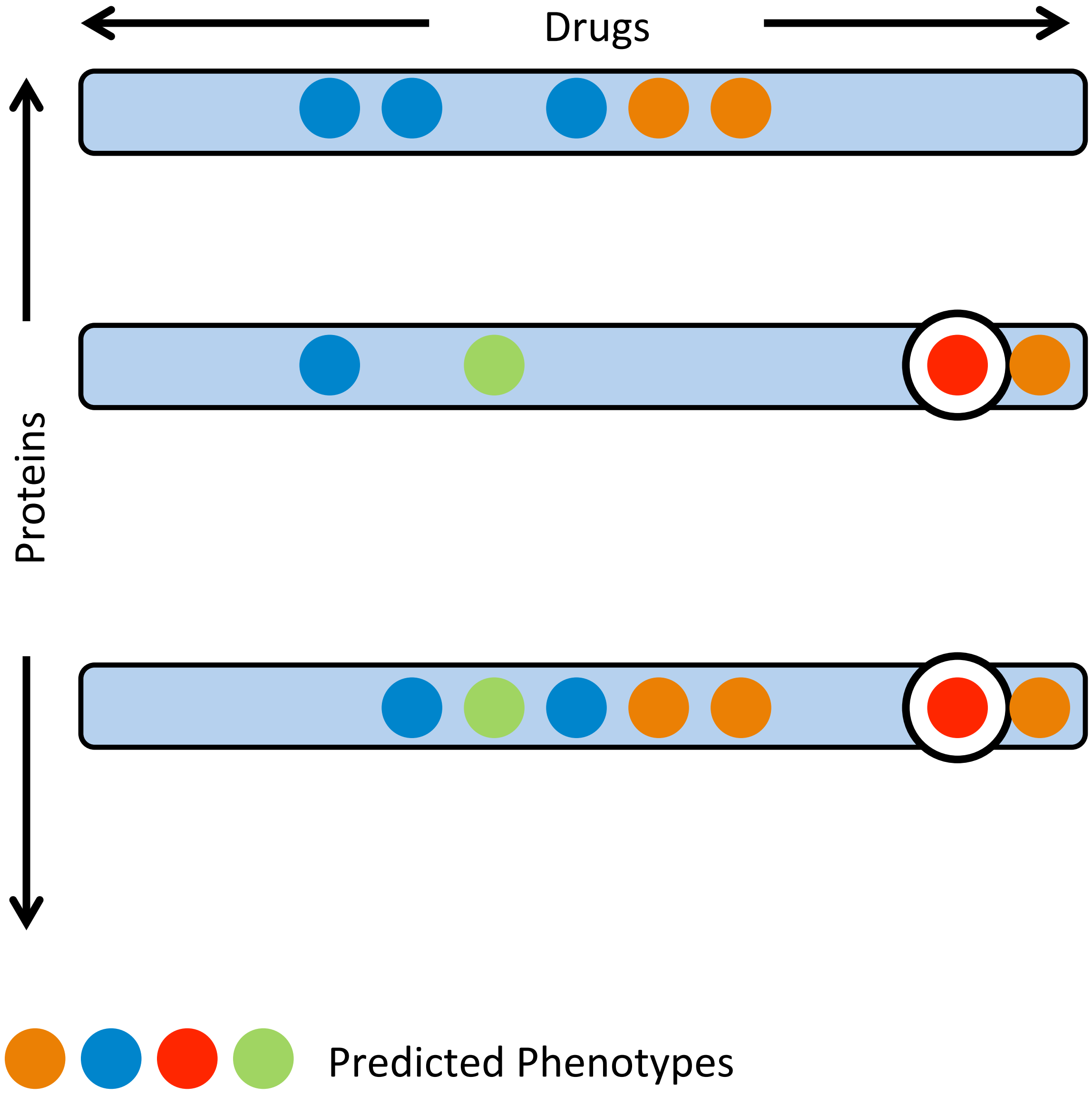
These considerations lead to a predictive model



How do we choose the next experiments?



Which experiments test equivalences?

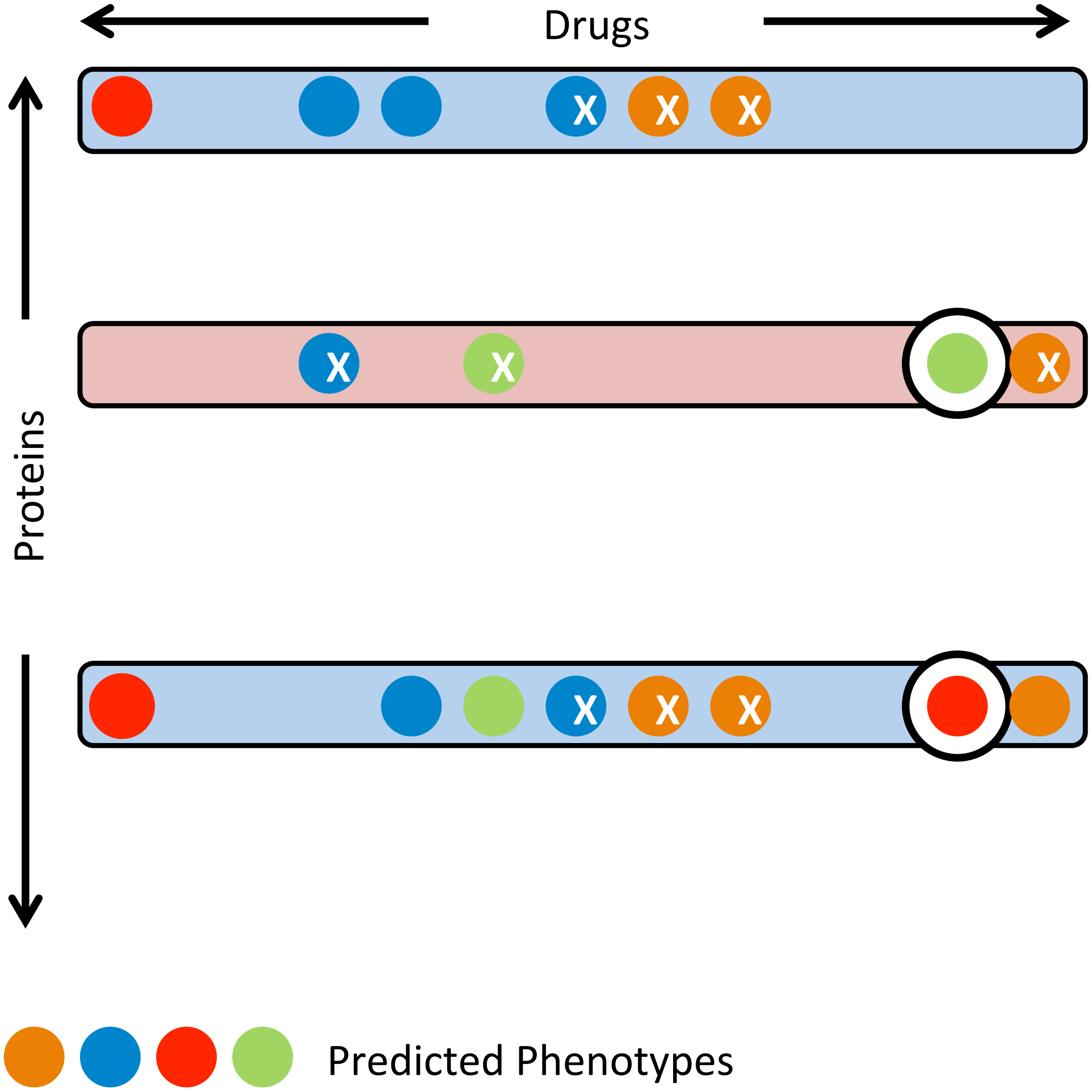


Drugs

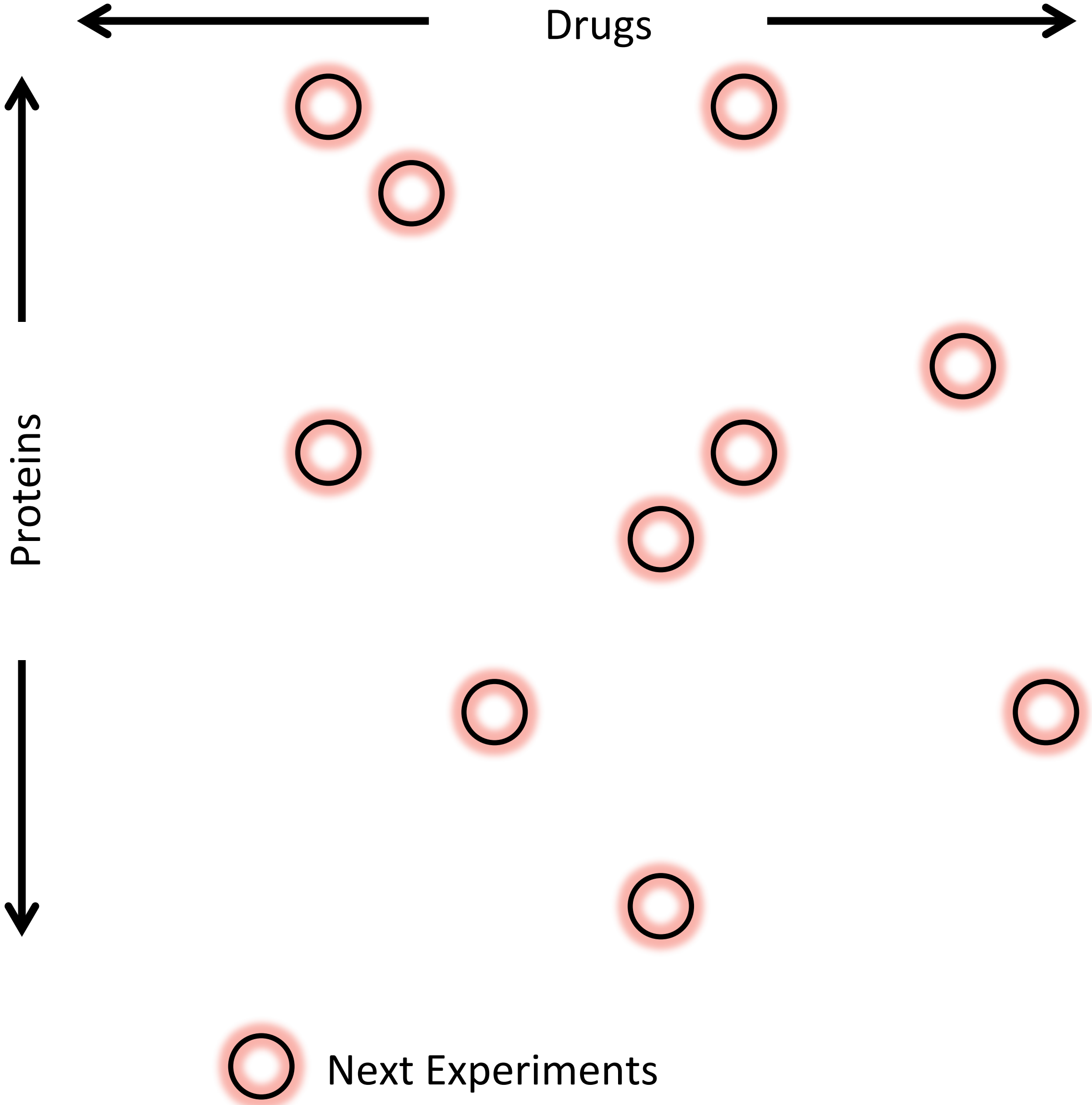
Proteins

Predicted Phenotypes

Impact of falsification of equivalence



Identify an Informative Batch of Experiments to Perform Next



Drugs

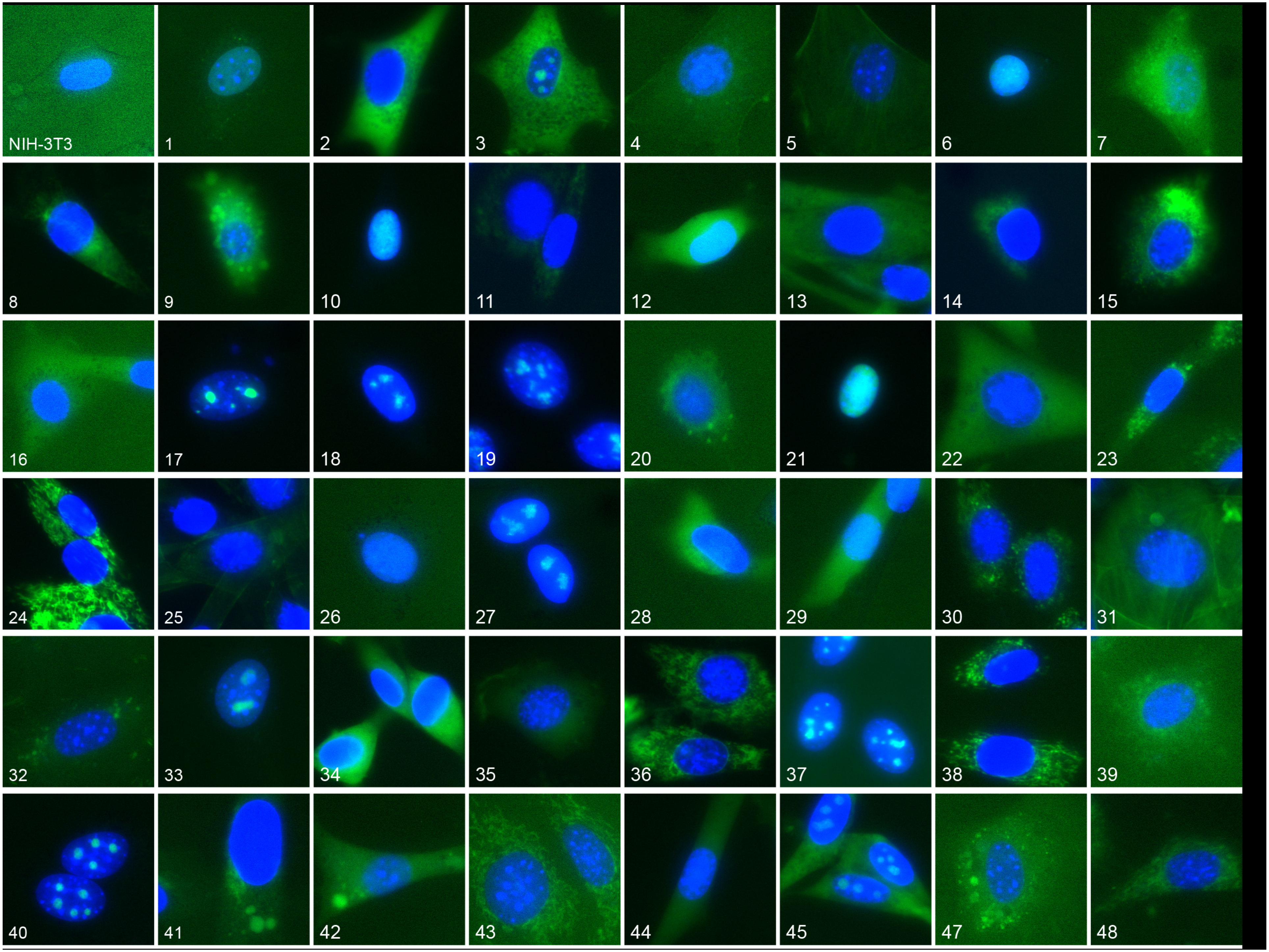
Proteins

○ Next Experiments

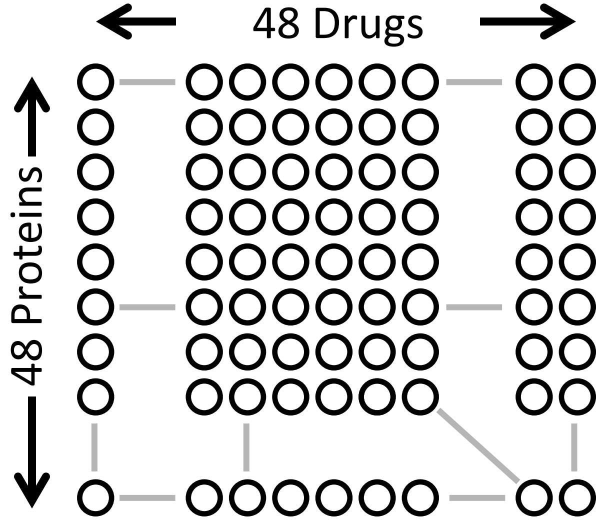
Naik, Kangas, Sullivan, Murphy, eLife 2016

Testing Prospectively

- Learning the effects of many compounds (drugs) on the subcellular localization of many proteins

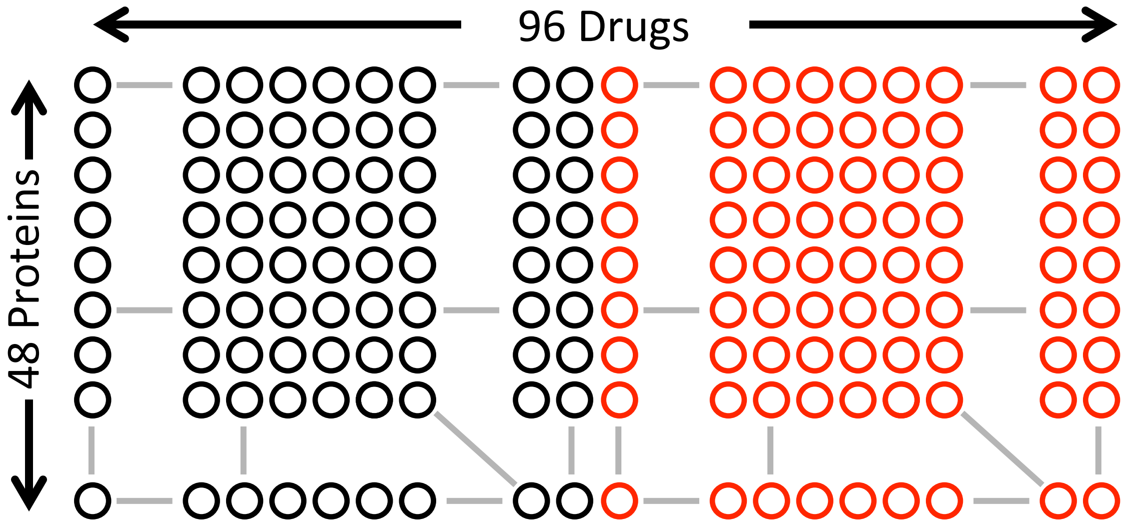

NIH-3T3
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23
24
25
26
27
28
29
30
31
32
33
34
35
36
37
38
39
40
41
42
43
44
45
47
48

Underlying Experiment Space: 48 Proteins x 48 Drugs

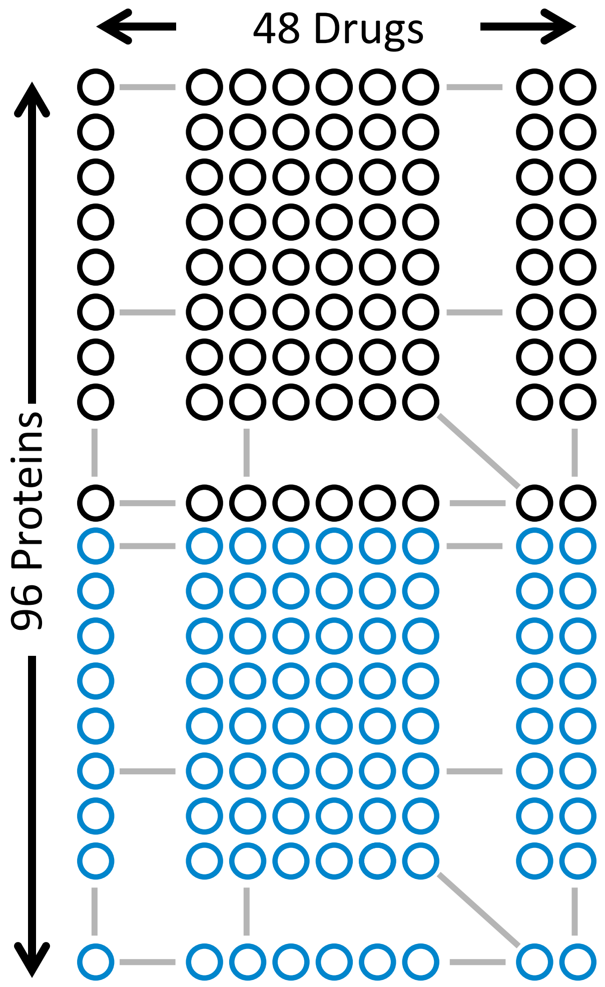


Since no information available on what effects to expect, need some way to evaluate effectiveness of active learning.
→ Use “hidden” duplication of drugs and proteins

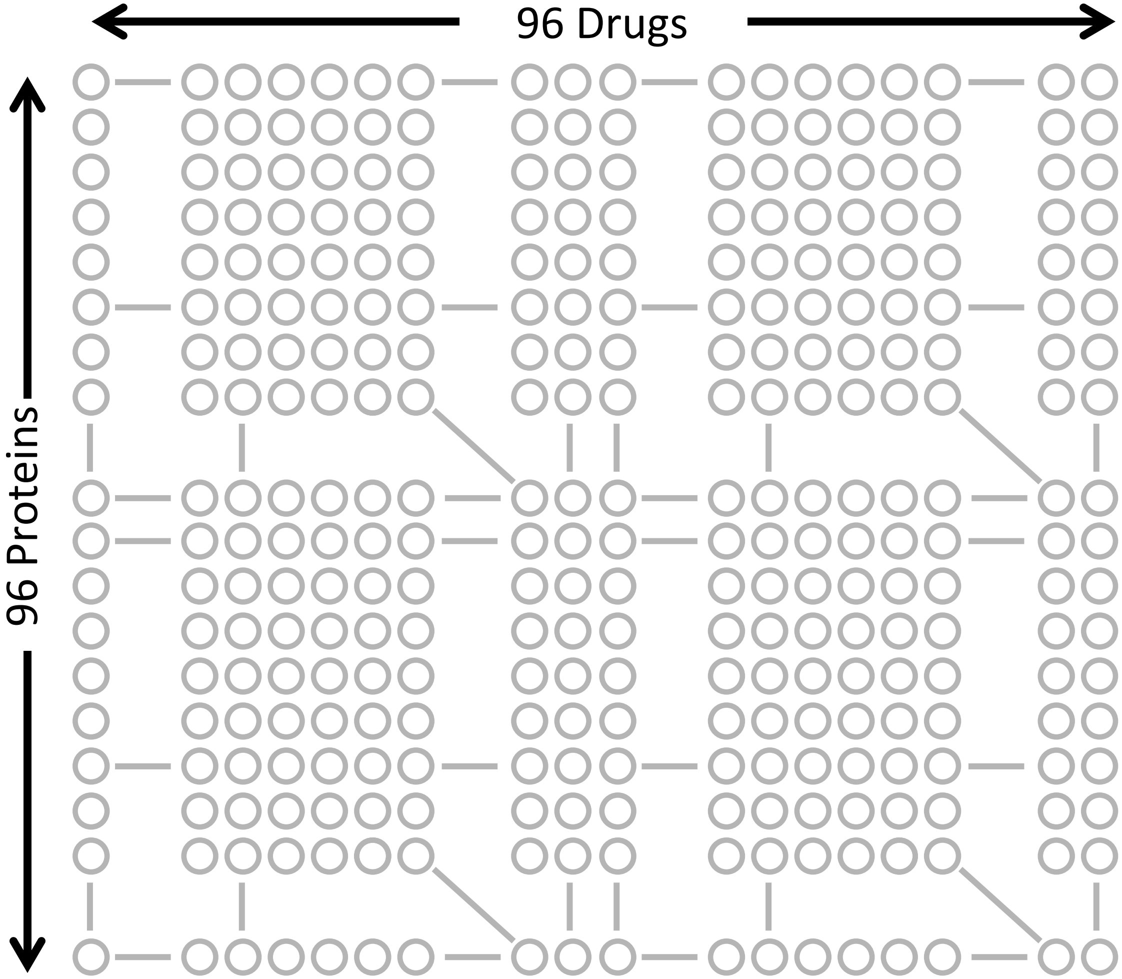
Silently Duplicate Drugs



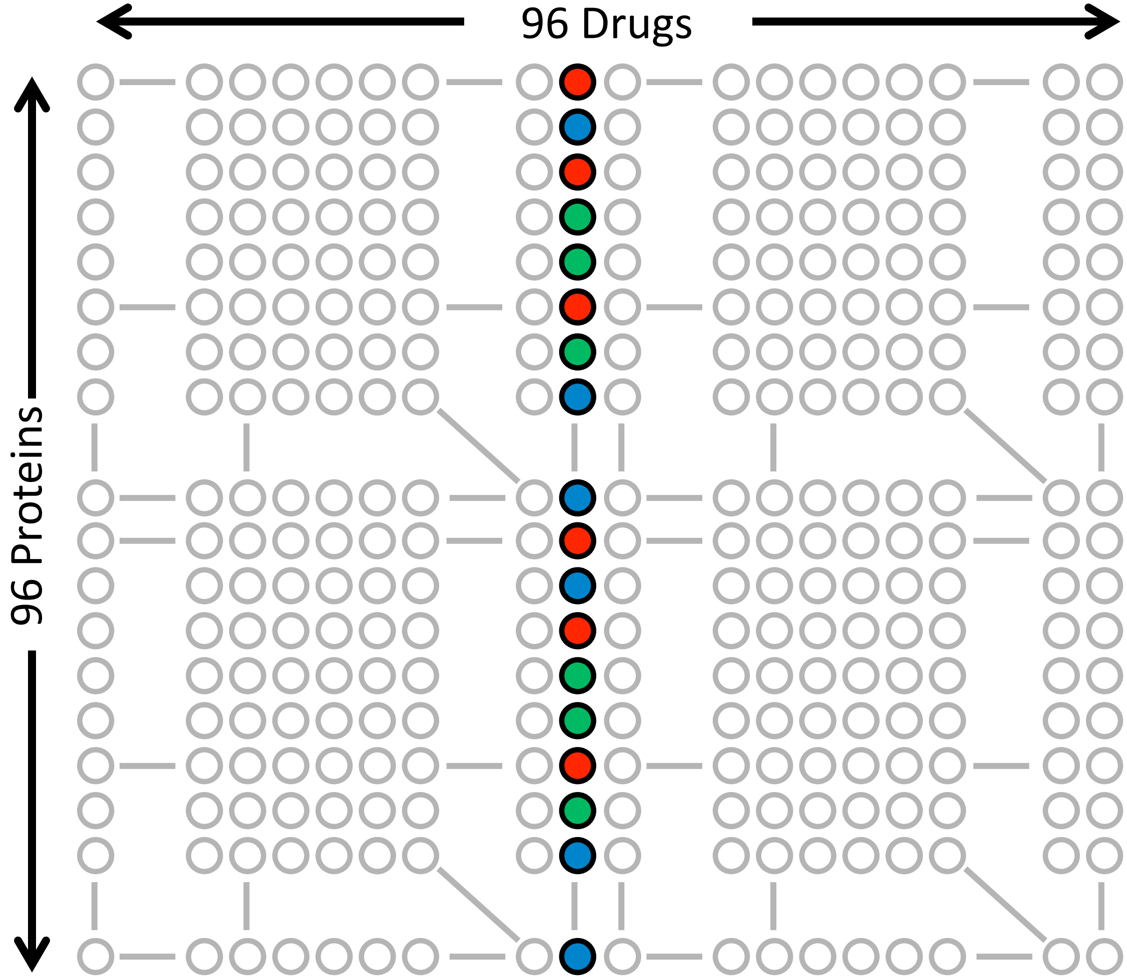
Silently Duplicate Proteins



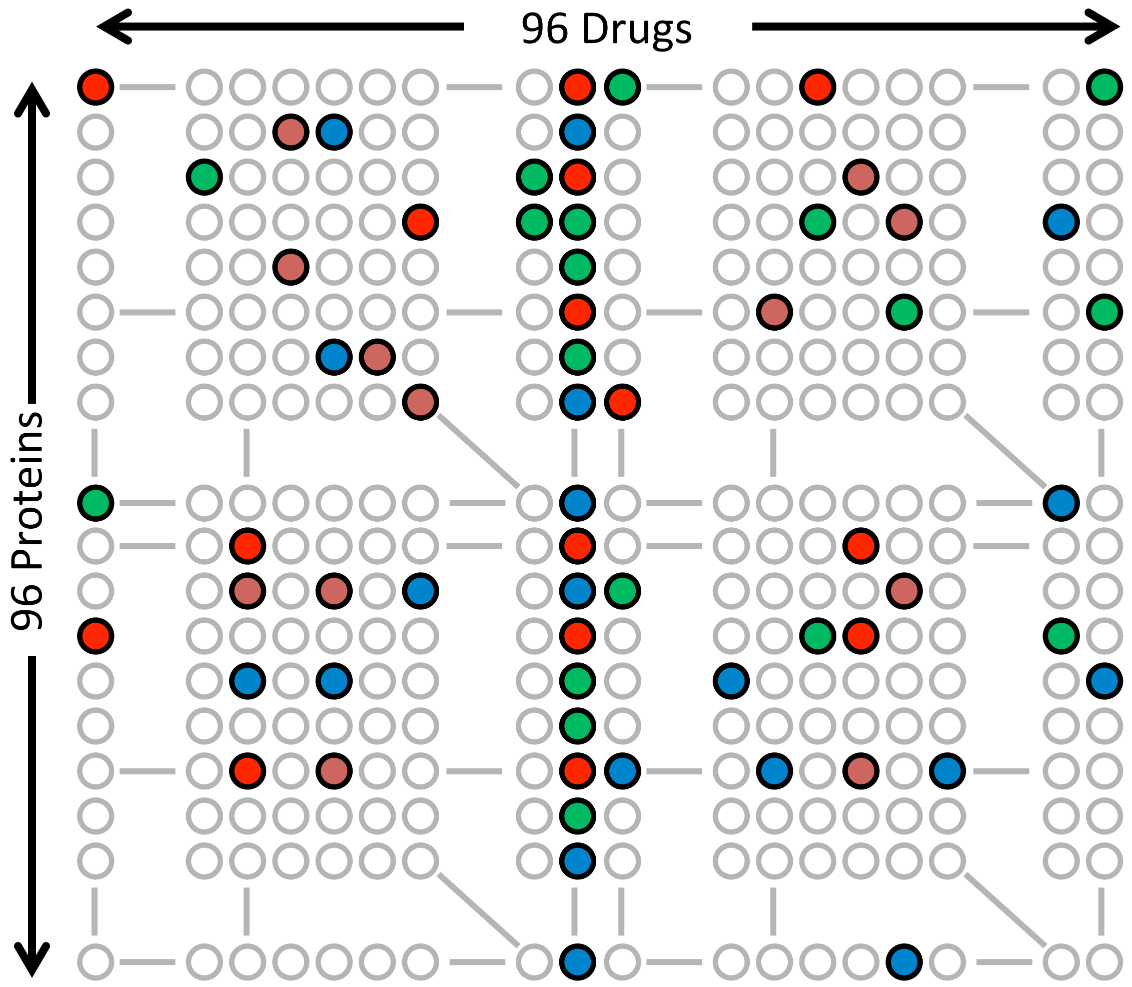
Silently Duplicate Proteins and Drugs to 96x96



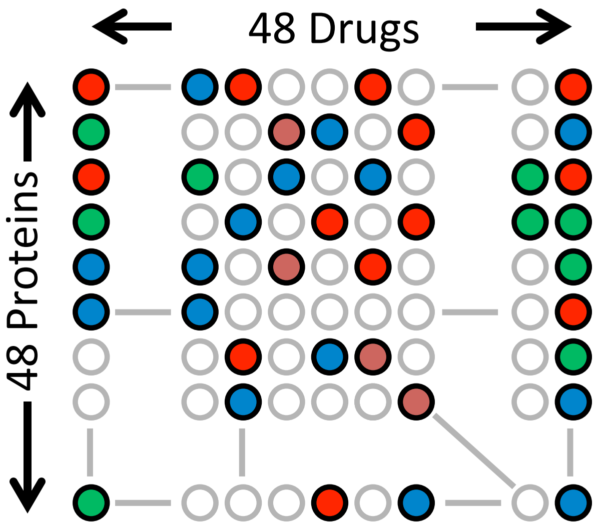
Starting data: All 96 Proteins with No Drug



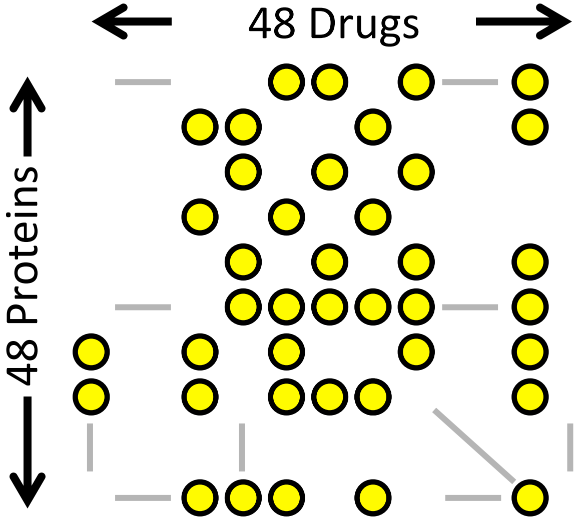
Actively Sampled 30 Batches (=28% of the 96x96 experiment space)



The 30 Batches covered 72% of the 48x48 space

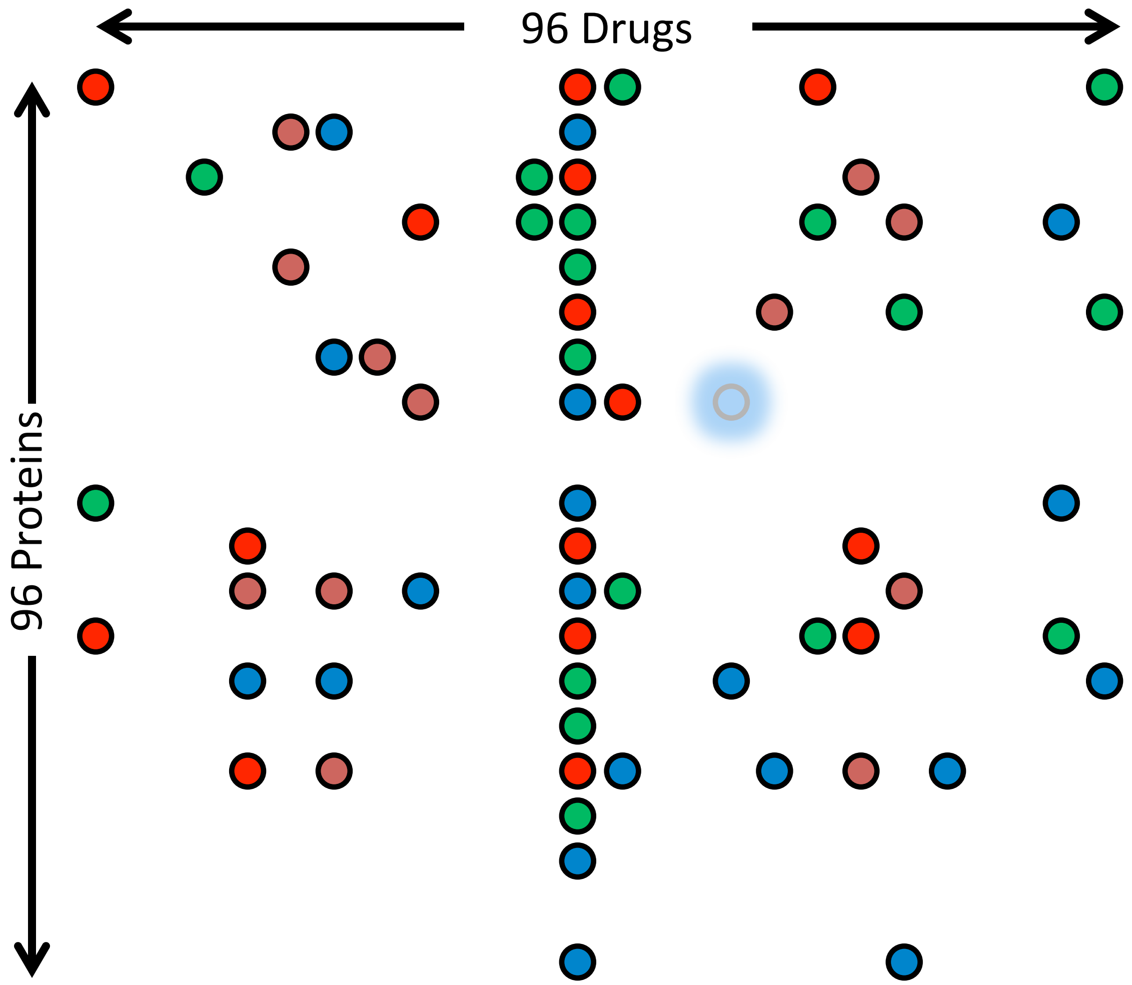


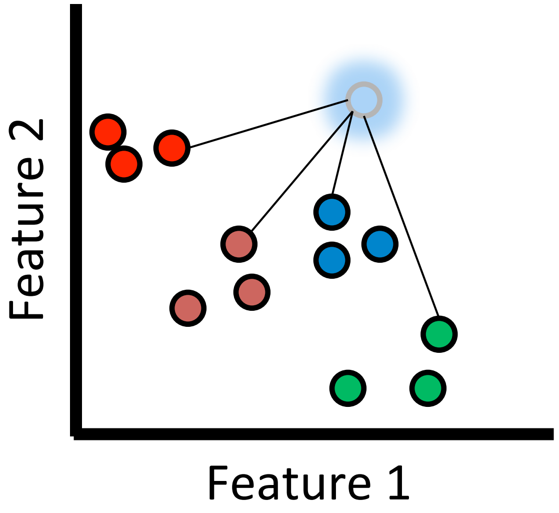
Performed remaining unique (protein, drug) combinations

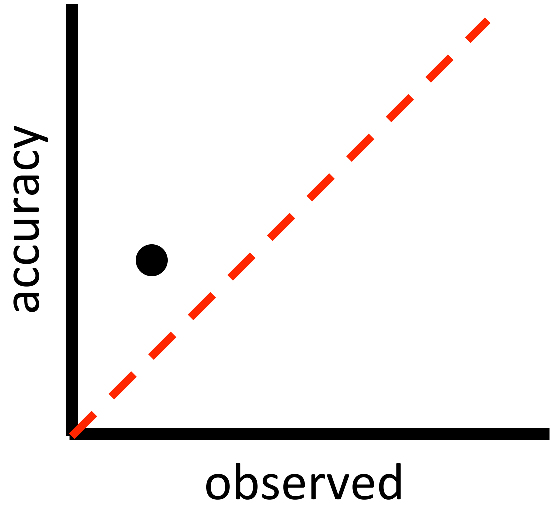


● 48x48 space filled in data

How well did it learn? Measure generalization performance

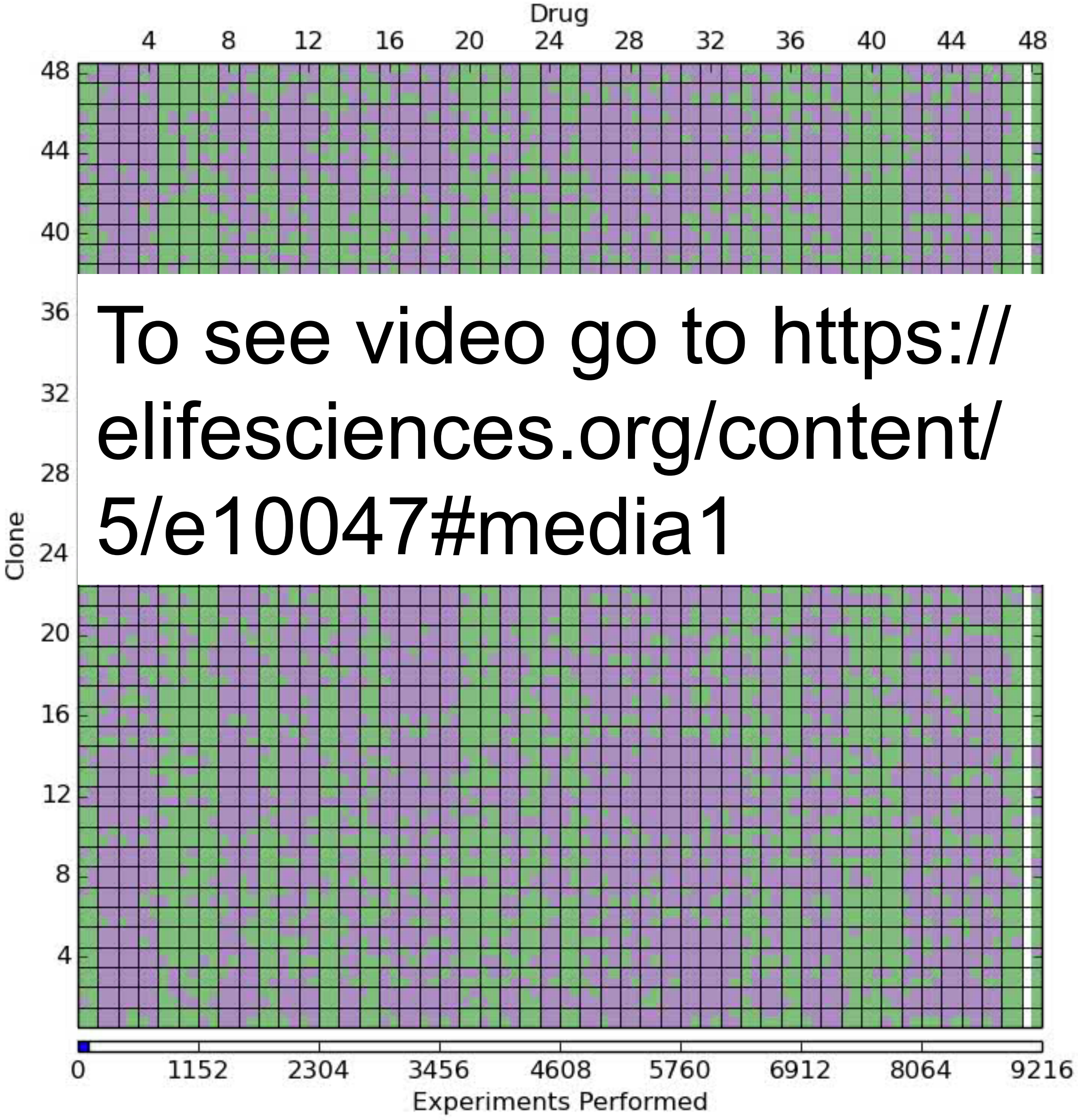




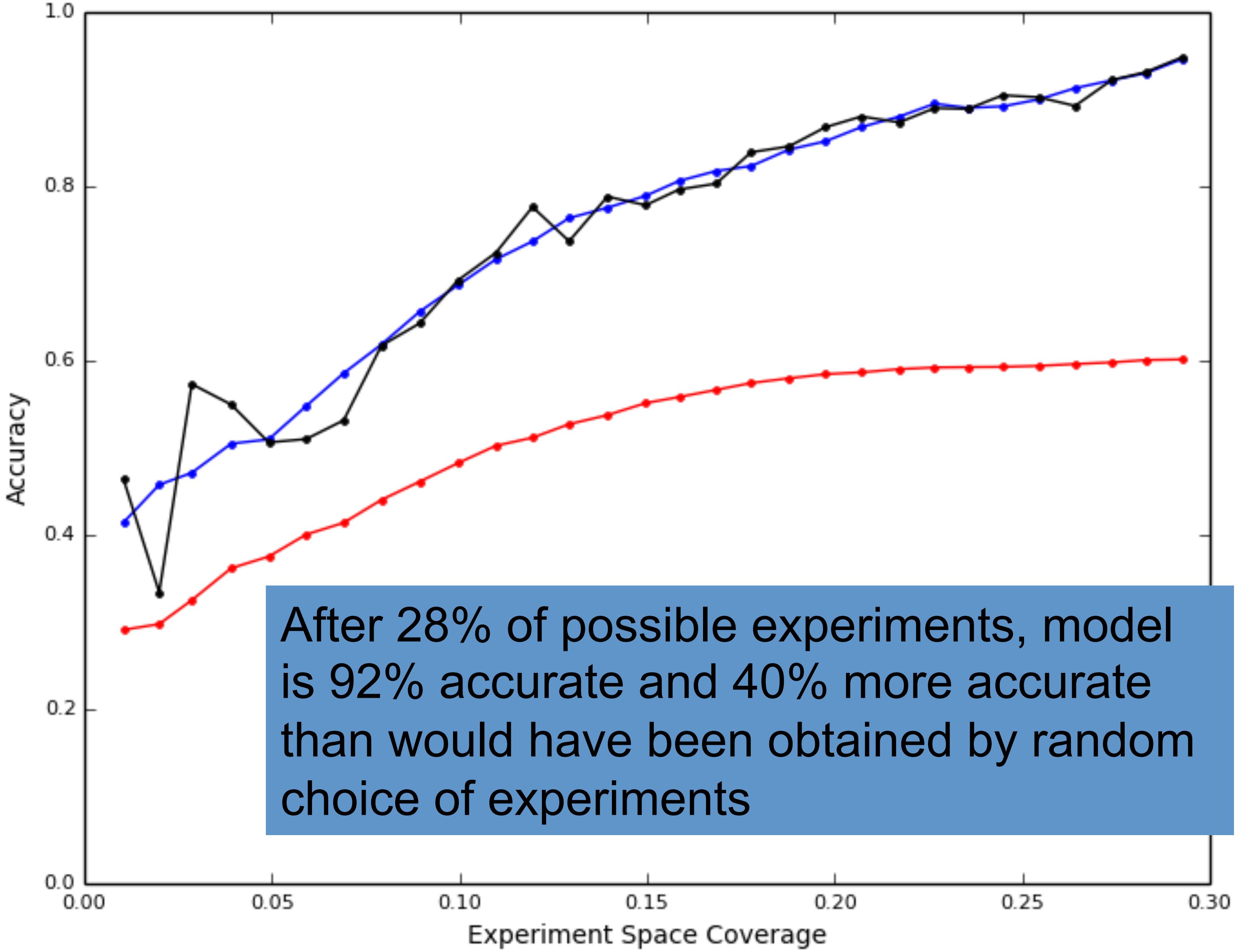


Automated, Prospective Active Learning

- Each small box is one drug and one target (but due to duplication there are four combinations)
- Green shows accurate prediction, purple is inaccurate, white shows experiments done



To see video go to https://elifesciences.org/content/5/e10047#media1



After 28% of possible experiments, model is 92% accurate and 40% more accurate than would have been obtained by random choice of experiments

Knowing when to stop AL

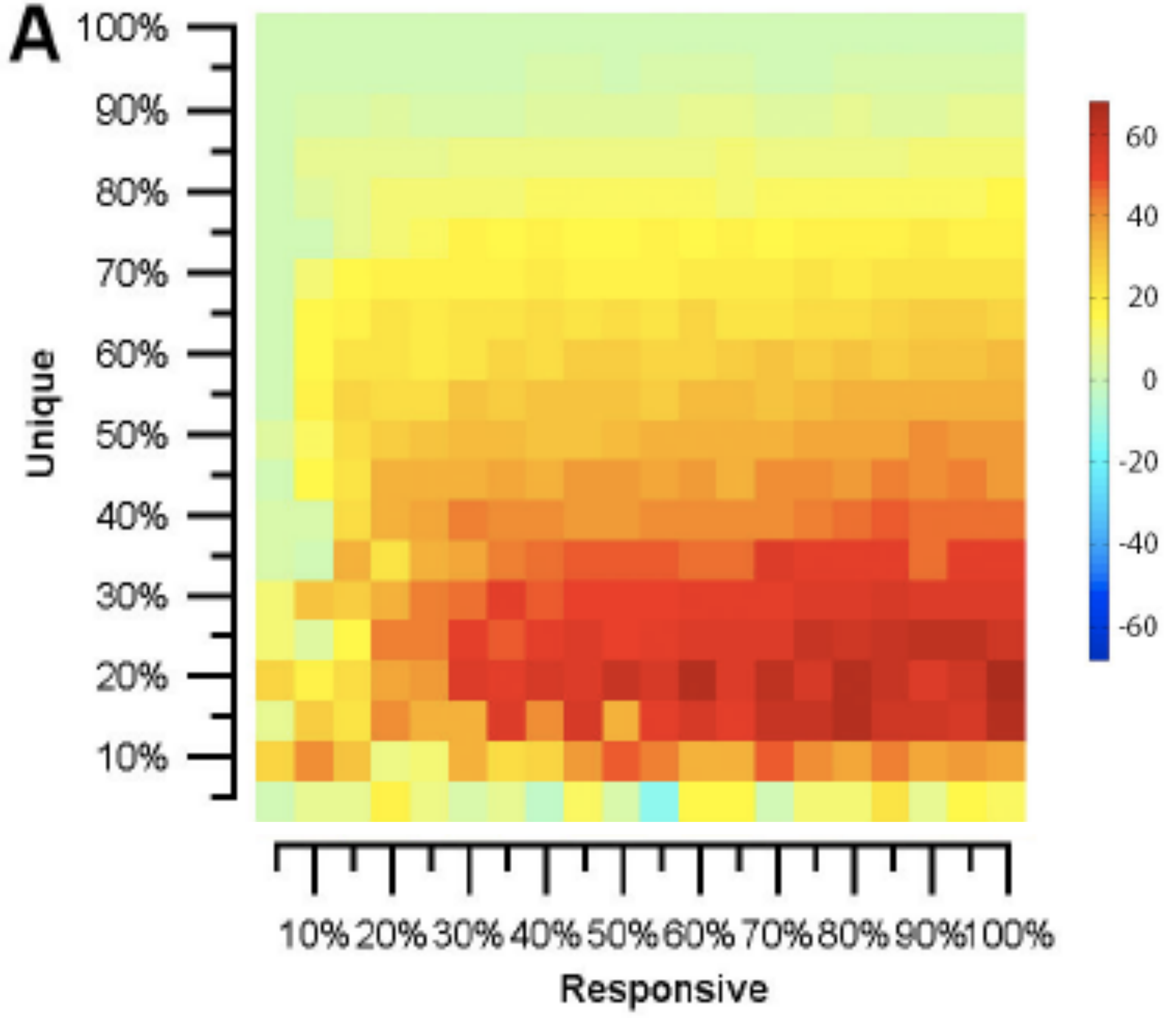
- When evaluating retrospectively, can calculate accuracy of any model using full data to decide how well we are doing
- In any prospective application, can't do that
- Need stopping criterion
- Past proposals of single criterion, typically based upon consistency/confidence of predictions
- We propose a machine learned criterion based on active learning trajectory

Naik, Kangas & Murphy, PLoS ONE 2013

Characterizing experimental spaces

- Basis of both matrix factorization and active learning is presence of correlations (low rank)
- Sparseness of interactions influences ability to learn correlations
- Define uniqueness as probability that all drugs and targets have different responses (100% = full rank)
- Define responsiveness as probability that any drug will affect any target (low%=sparse)

Active learning simulations for different experimental spaces

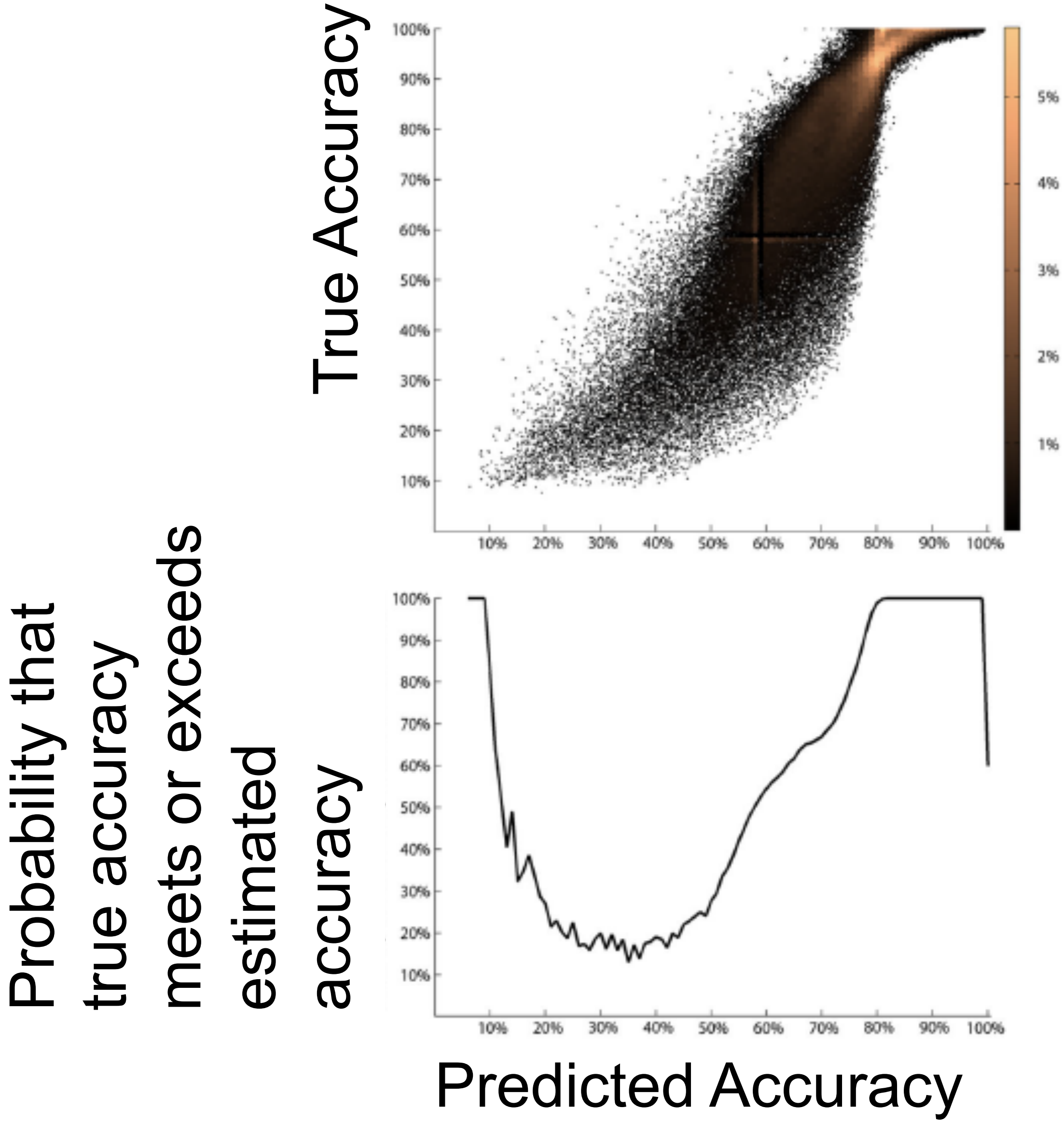


Savings from AL over random

Learning a stopping criterion

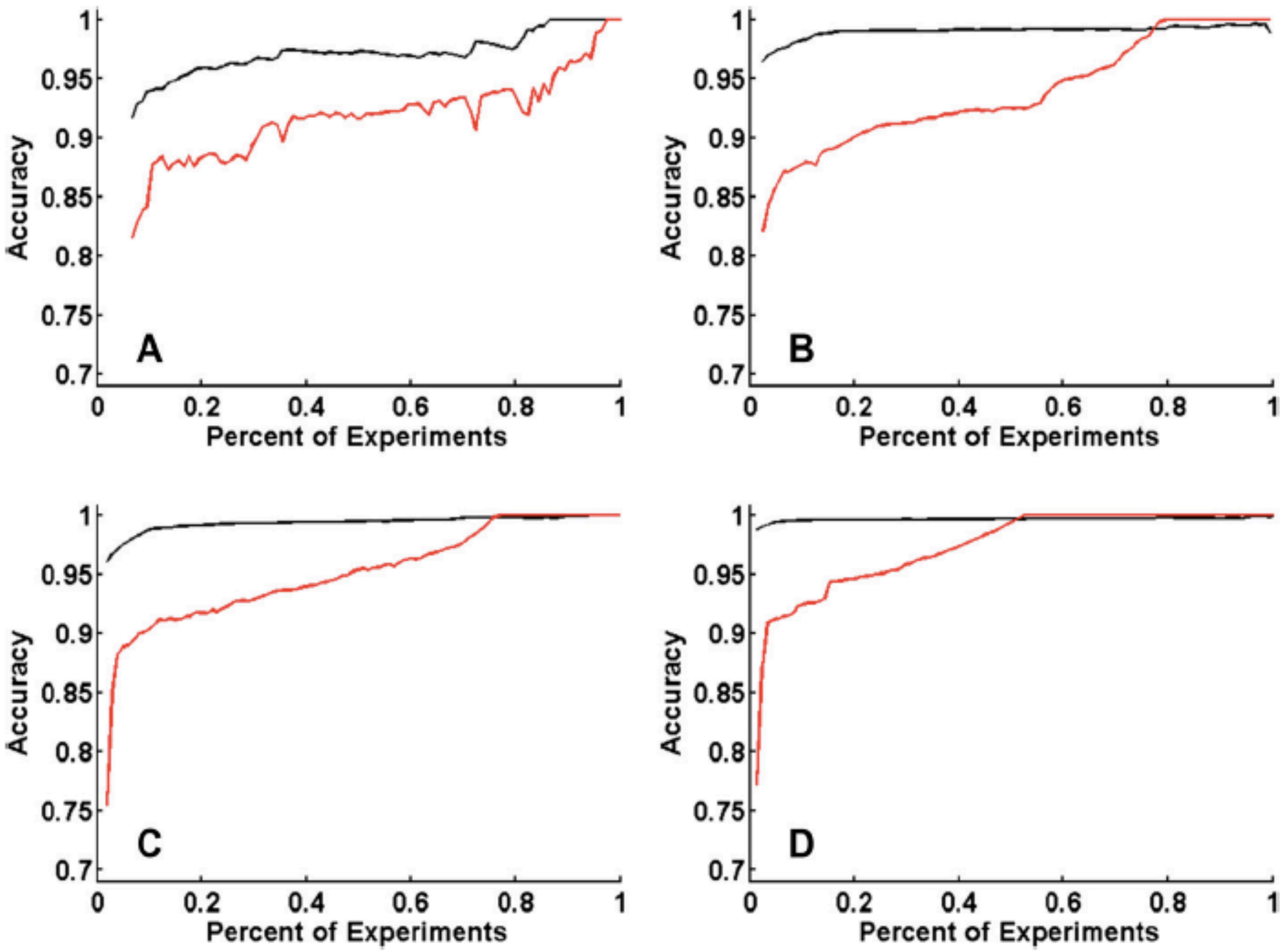
- Assuming a parameterization of an experimental space (such as uniqueness and responsiveness), perform many simulations over that space and record features for each active learning run (e.g., number of phenotypes observed, consistency of new experiments with predictions, number of conditions that differ within a target)
- Learn a regression function over all simulations to predict accuracy of model from these features

Learning the stopping criterion



Temerinac-Ott, Naik & Murphy, BMC Bioinformatics 2015

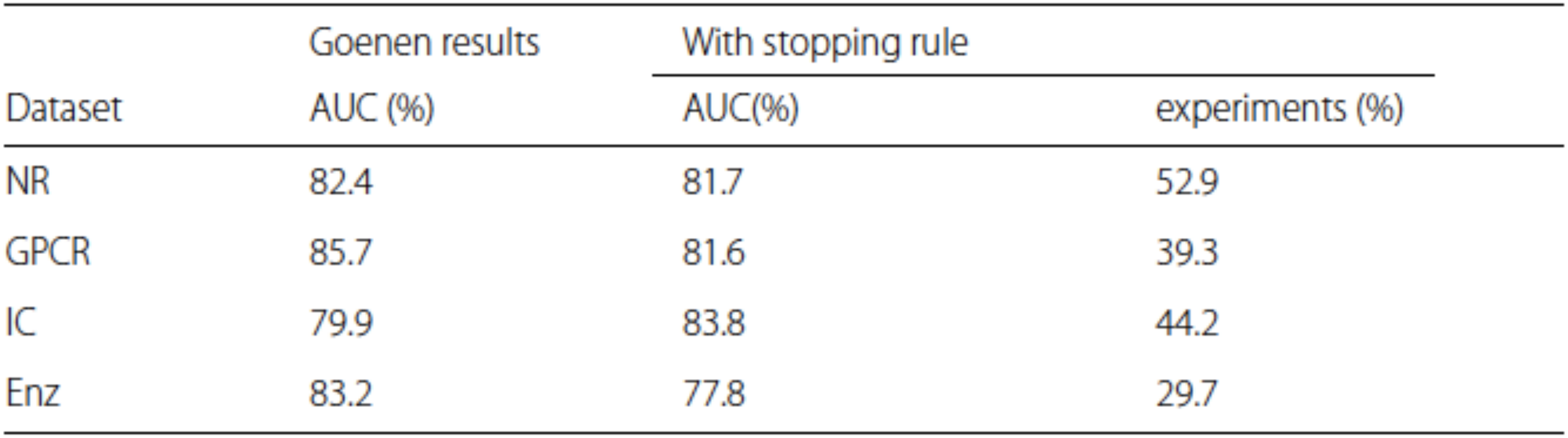
Estimating accuracy during active learning



Black: Actual

Red: estimated

Reduced number of experiments chosen by stopping criterion



	Goenen results	With stopping rule	
Dataset	AUC (%)	AUC(%)	experiments (%)
NR	82.4	81.7	52.9
GPCR	85.7	81.6	39.3
IC	79.9	83.8	44.2
Enz	83.2	77.8	29.7

Stopping when estimated
accuracy = 90%

Summary

- Empirical results for value of active learning for “large” heterogeneous experimental spaces starting with little data
- First prospective demonstration of active learning driven experimentation for unknown phenotypes
- Machine learning approach for learning stopping criteria

Acknowledgments

Josh Kangas

Armaghan Naik

Maja Temerinac-Ott

Devin Sullivan

Chris Langmead

Support:
NIH grants GM075205, GM090033, GM103712
German Ministry for Education and Research FKZ0316185