Computational Biology Department

Carnegie Mellon University
School of Computer Science

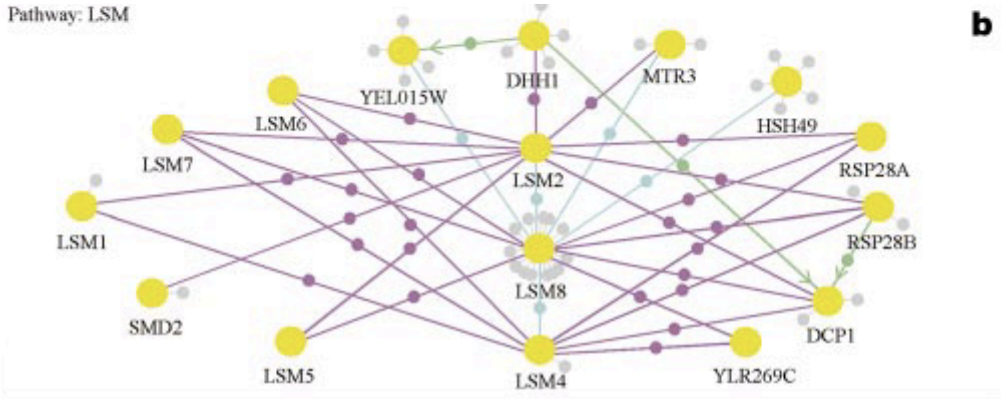
Self-driving instruments: the need for active machine learning in biomedical research

Robert F. Murphy
Ray & Stephanie Lane Professor of Computational Biology and
Professor of Biological Sciences, Biomedical Engineering and Machine Learning
Head, Computational Biology Department, School of Computer Science

February 2018

The failure of Reductionism

- For many decades, biomedical research was based on ***reductionism***, the assumption that biological components could be understood in isolation
- By the 80's it was becoming clear that many, many components interacted



A comprehensive analysis of protein–protein interactions in *Saccharomyces cerevisiae*

Peter Uetz*†, Loic Giot*‡, Gerard Cagney†, Traci A. Mansfield‡, Richard S. Judson‡, James R. Knight‡, Daniel Lockshon†, Vaibhav Narayan‡, Maithreyan Srinivasan‡, Pascale Pochart‡, Alia Qureshi-Emili†§, Ying Li‡, Brian Godwin‡, Diana Conover†§, Theodore Kalbfleisch‡, Govindan Vijayadamodar‡, Meijia Yang‡, Mark Johnston†‖, Stanley Fields†§ & Jonathan M. Rothberg‡

Reductionism gives way to Systems Biology

- Cells, Tissue, Organs and Organisms were recognized to be “complex systems” — systems whose properties as a whole cannot be inferred from their individual properties
- The need for computational methods to produce **predictive models** became recognized
- Need data, so many “big science” projects were started

A big problem…

- Assuming *n* genes, one gene=one function and reductionism, the number of experiments equals the number of genes, about 10,000
  - at one experiment per day, 28 years
- Given *m* average genes per function and *n* genes, the number of experiments is $n^m \sim 10^{4m} \sim 10^{20}$
  - at 10^9 experiments per day, 2 million centuries!

and another one…

- Emphasis in systems biology on “validating” or “proving” of models by doing selective experiments
- But empirical models cannot be proven!

Solution?

- Use ***active*** machine learning
- Choose experiments not to ***prove*** model but to ***improve*** model

NATURE CHEMICAL BIOLOGY | VOL 7 | JUNE 2011

commentary

An active role for machine learning in drug development

Robert F Murphy

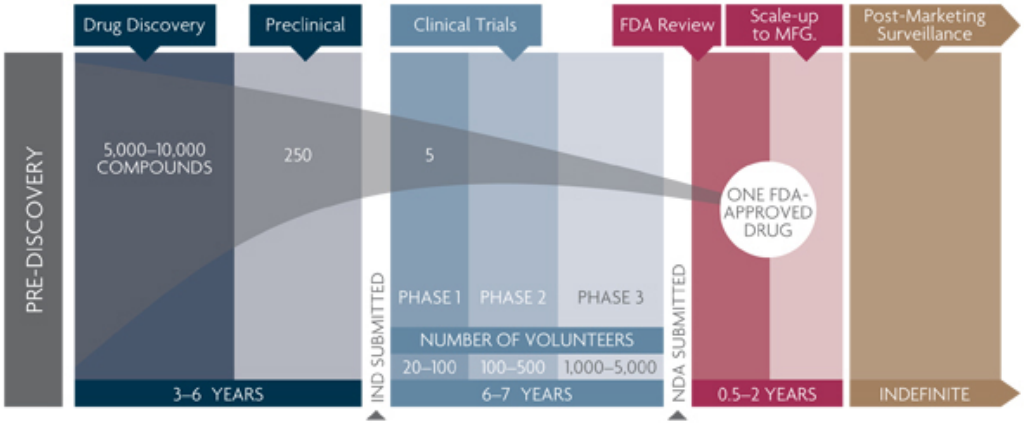
Because of the complexity of biological systems, cutting-edge machine-learning methods will be critical for future drug development. In particular, machine-vision methods to extract detailed information from imaging assays and active-learning methods to guide experimentation will be required to overcome the dimensionality problem in drug development.

Systems Biology, Big Data and 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

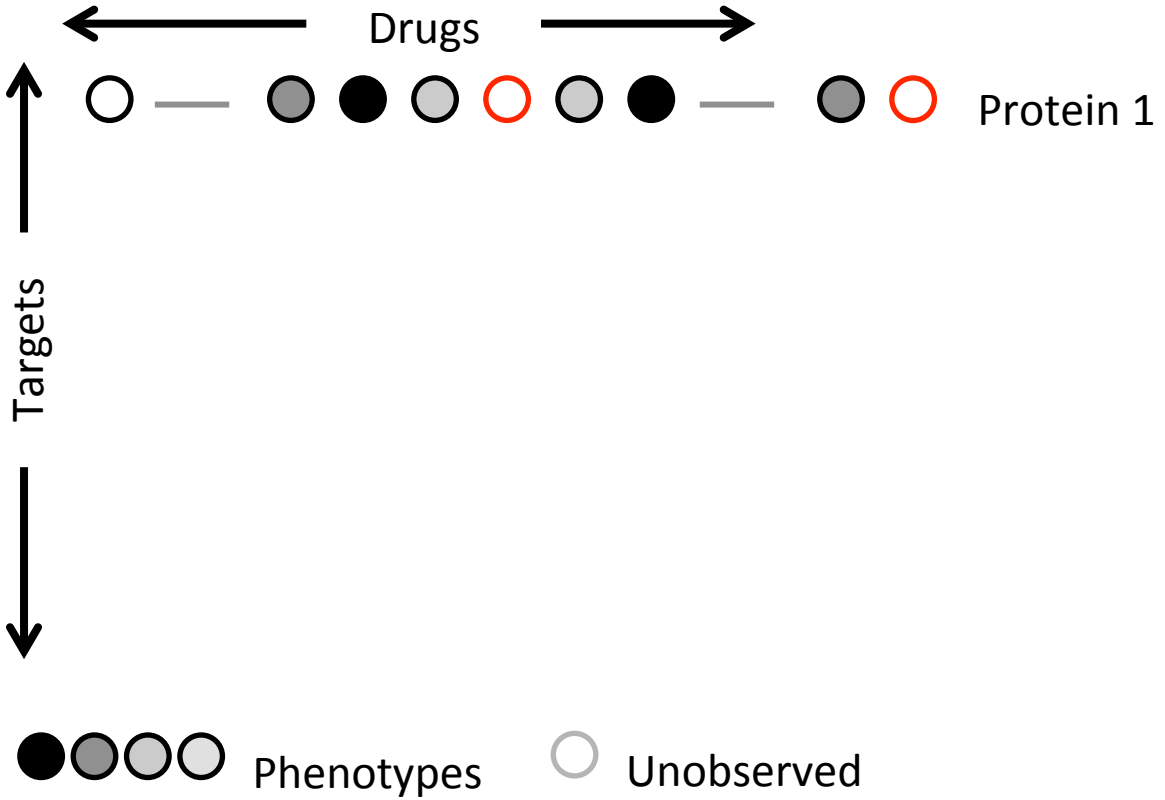
Further…

- 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 not just about finding hits on desired target but also about avoiding others

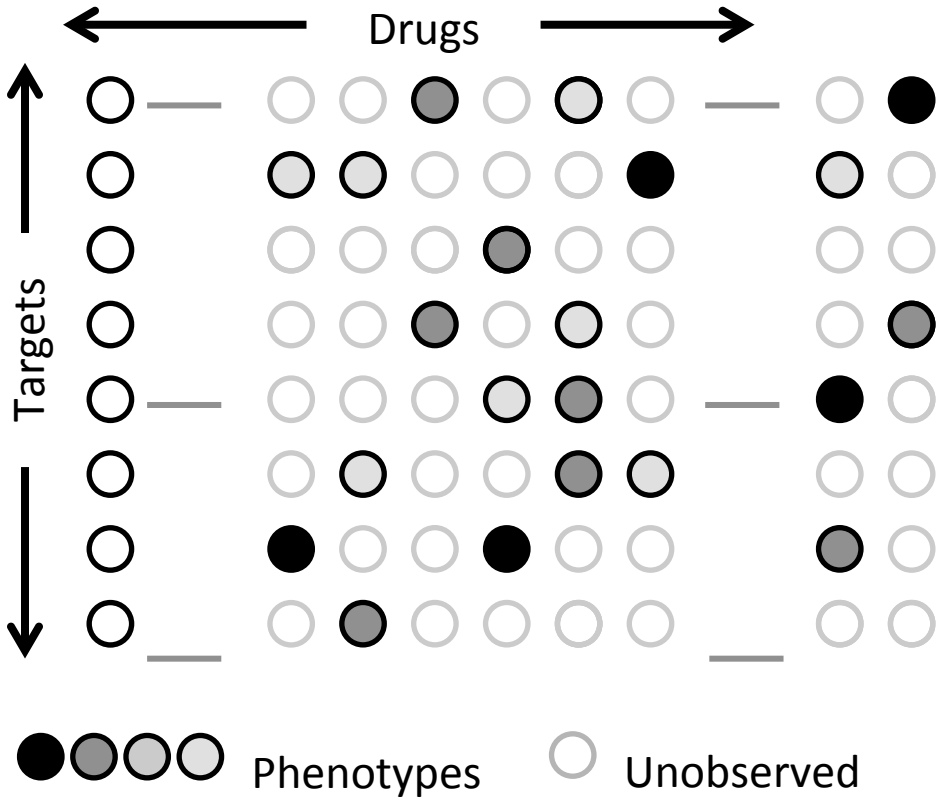


Source: PhRMA

Current practice in drug screening: consider each target separately



Where we'd like to be: consider all drugs and all targets



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

Playing Battleship with Drugs and Cells

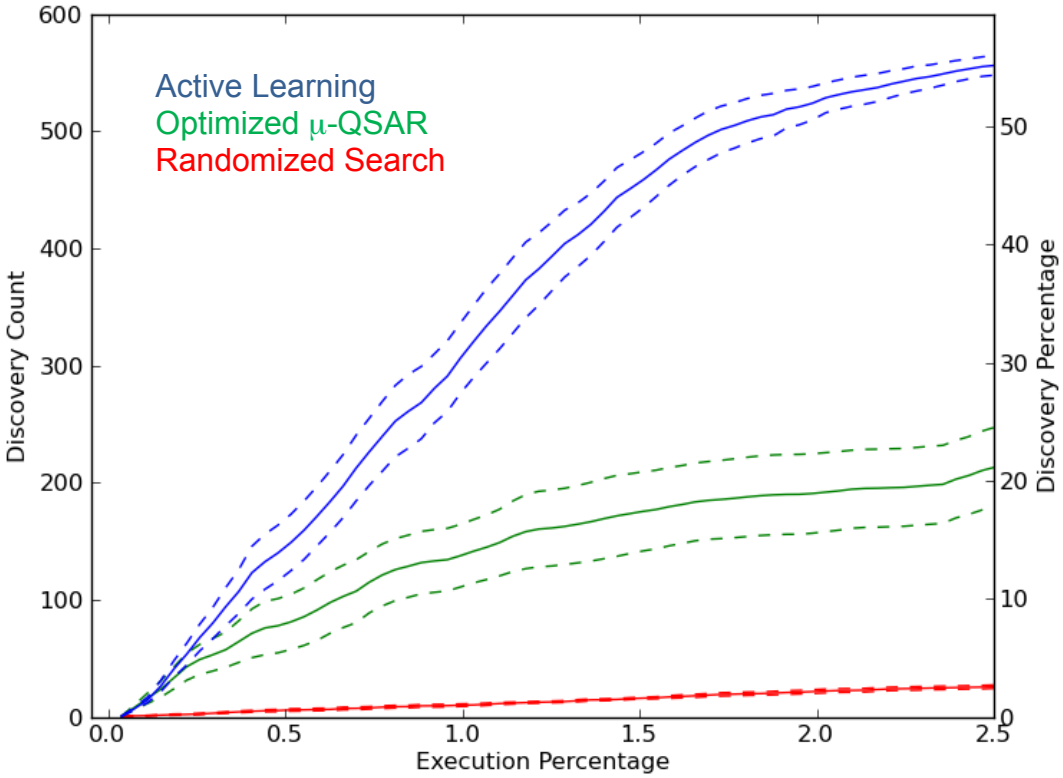
	A	B	C	D	E	F	G	H	I	J
1	■	■		■			■	■	■	■
2				■						
3	■			■		■	■	■		■
4	■		X							■
5	■					X	X			
6	■	X				■		X		X
7				X		■				X
8	X	X						X		■
9										
10					■	■	■	■	■	■

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

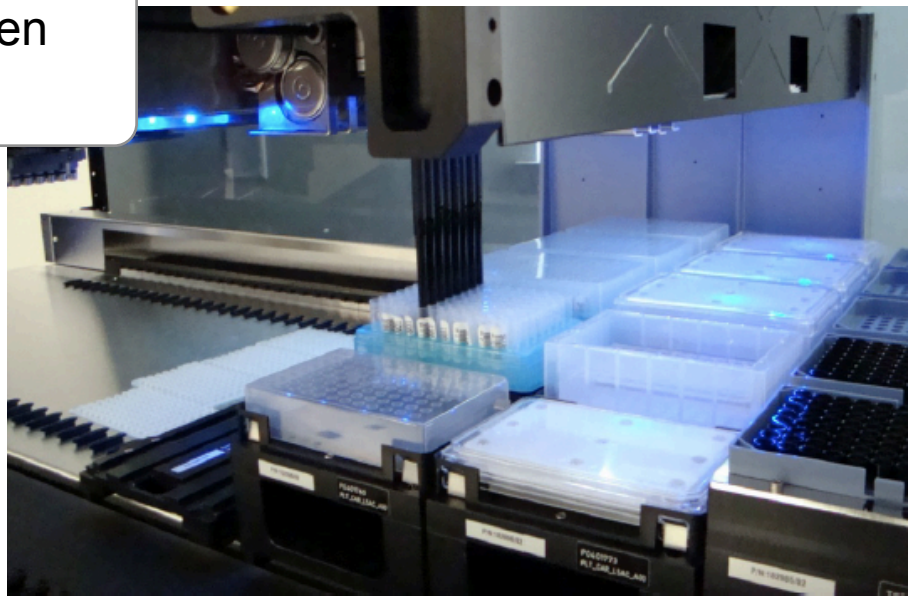
With only **2.5%** of the matrix covered, we can identify **57%** of the active compounds!

Kangas, Naik, Murphy, BMC Bioinformatics 2014

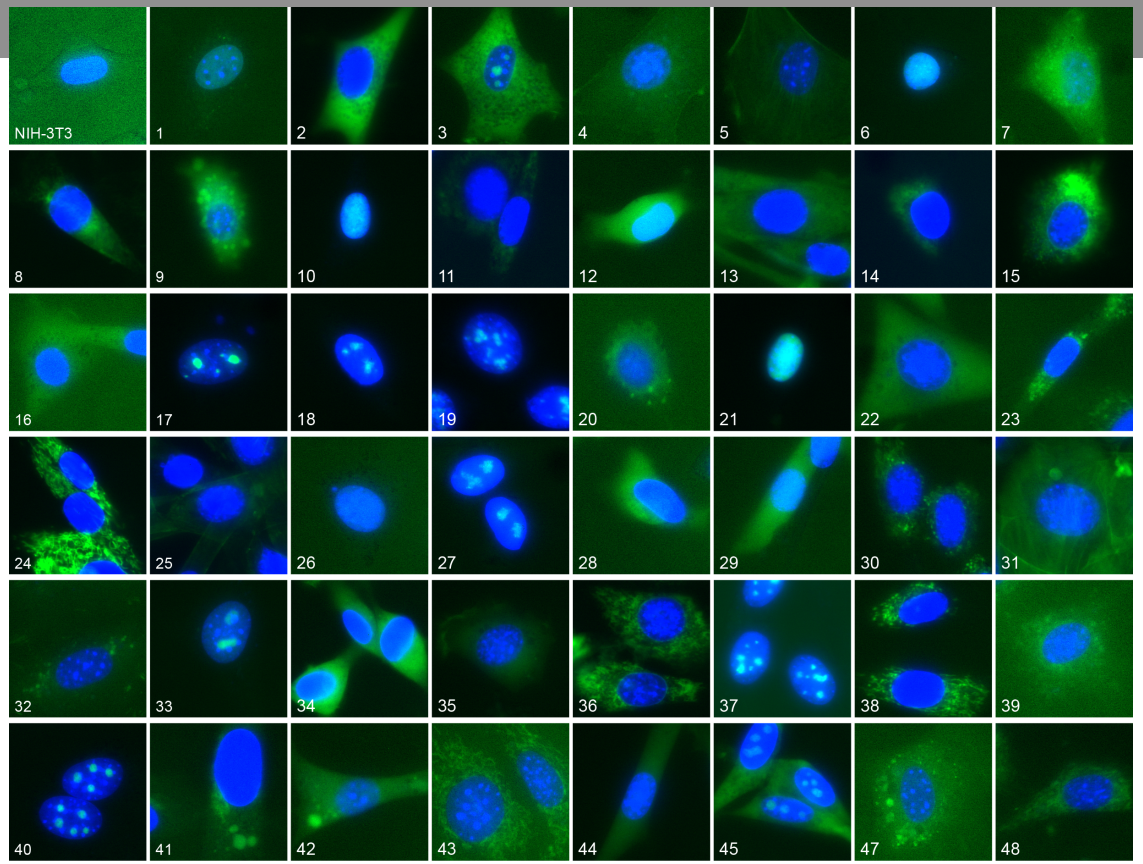


First Prospective Use of Active Learning for Complex System

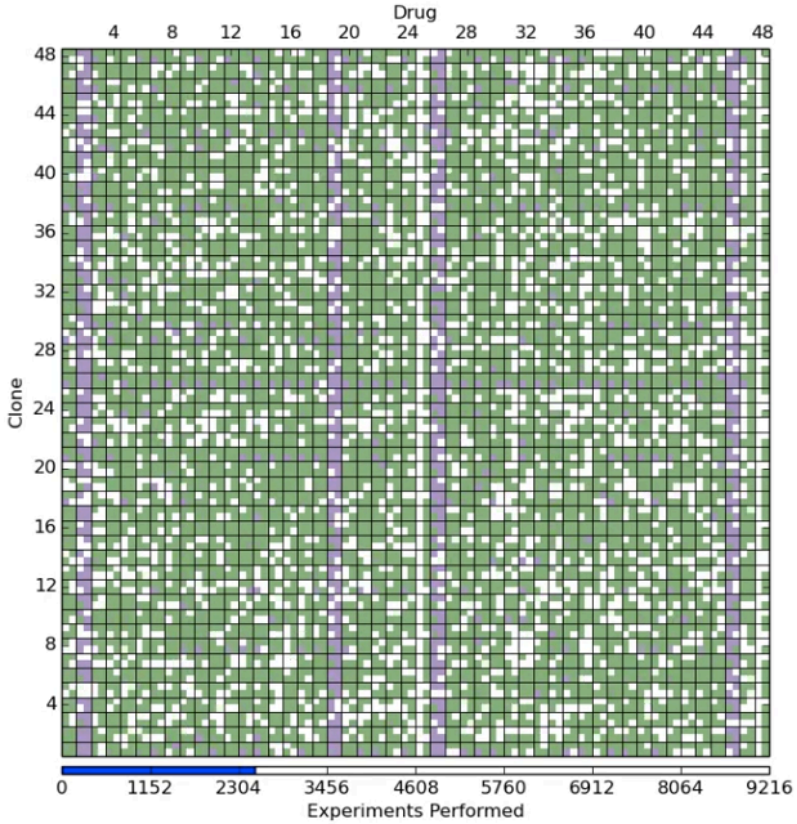
Use liquid handling robots and automated microscope to execute experiments chosen by an active learner



Try to learn the effects of 96 drugs upon 96 GFP-tagged proteins, *without doing experiments for all drugs and proteins*

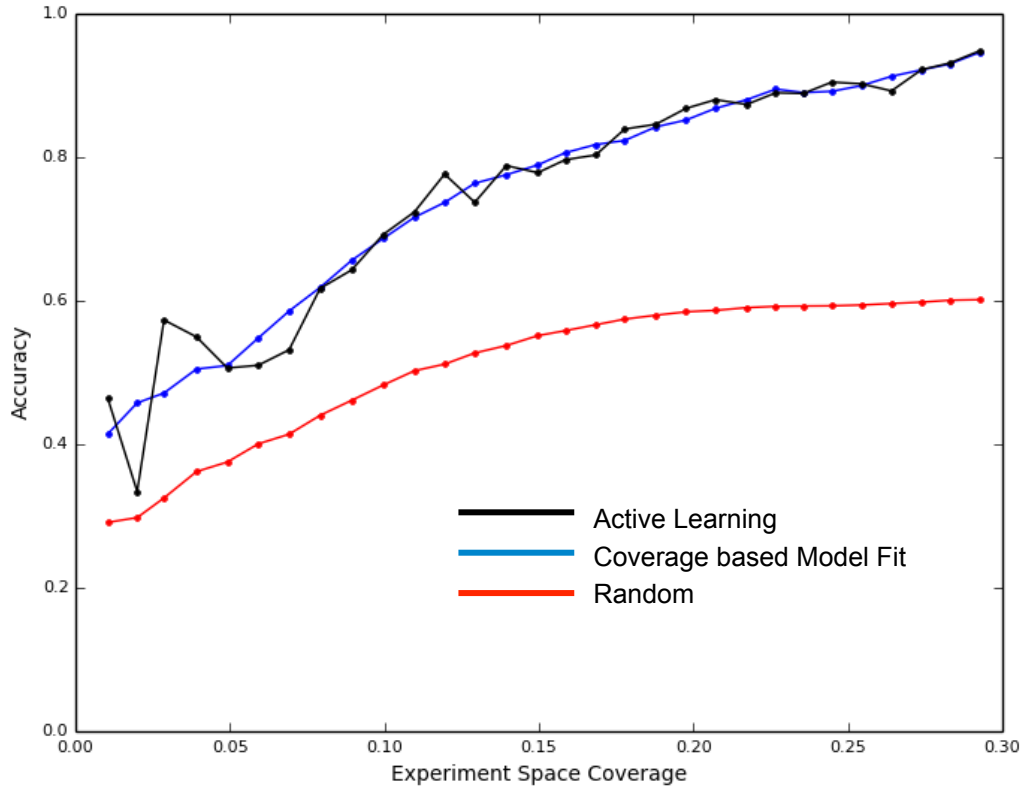


- Each small box is one drug and one target (but due to duplication there are four combinations)
- Green shows accurate prediction, purple is inaccurate, white shows experiments done



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

Naik, Kangas, Sullivan, Murphy, eLife 2016



Conclusions?

- Deep learning may find low-hanging fruit in existing data but fundamentally limited
- Embracing complexity in higher dimensional models combined with active machine learning to guide experimentation needed in many areas of biomedical research

Supported by:

NSF

NIH National Institute of General Medical Sciences