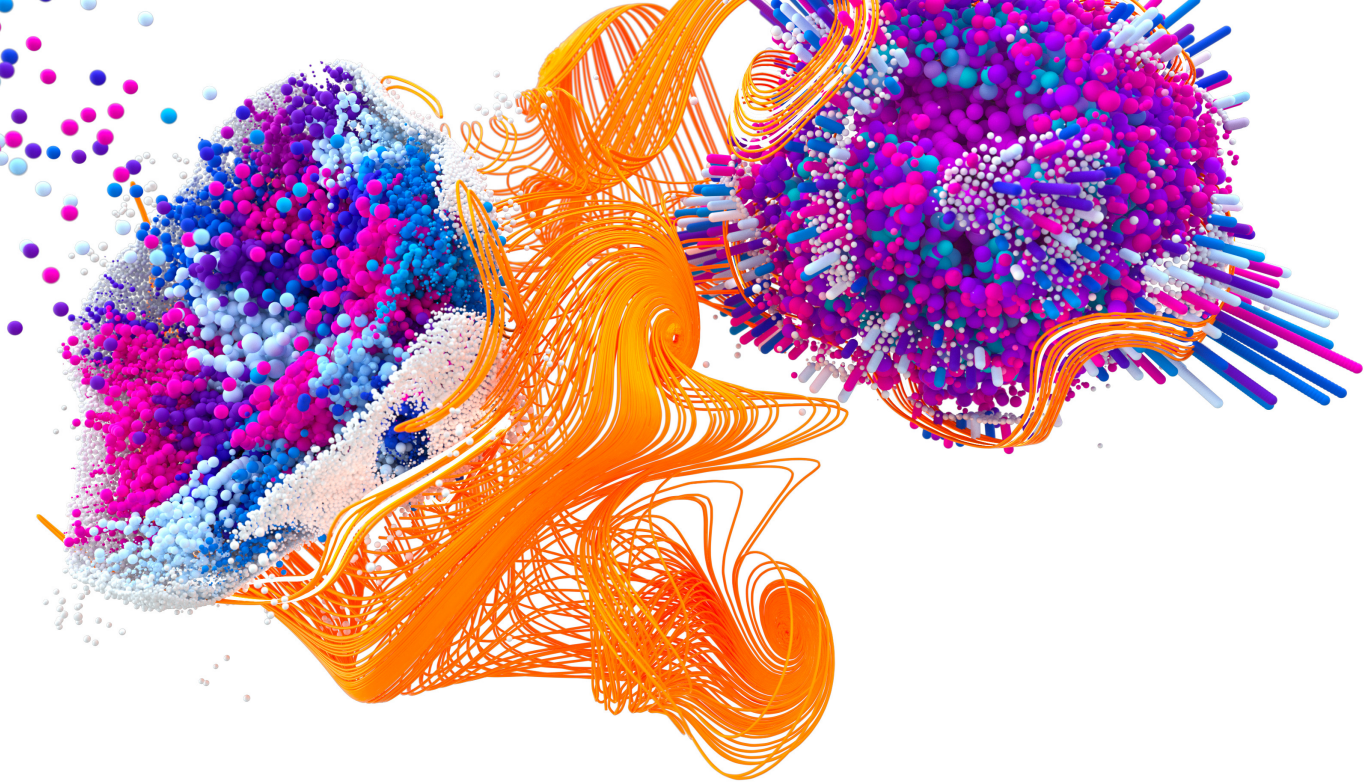
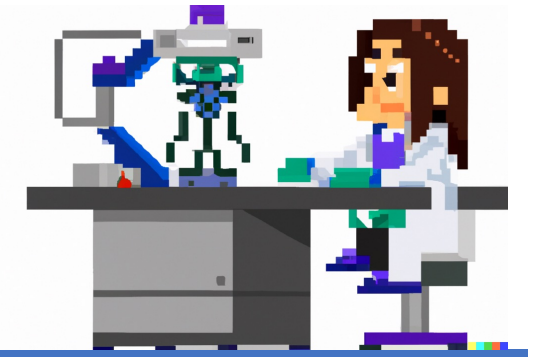


GSK



9 July 2024 / OxML Health x Bio

Working smarter, not harder in drug development: AI assistants for accelerating biological discovery

Patrick Schwab, Head Biomedical AI, GSK.ai, on behalf of the team

GSK.ai Biomedical AI group

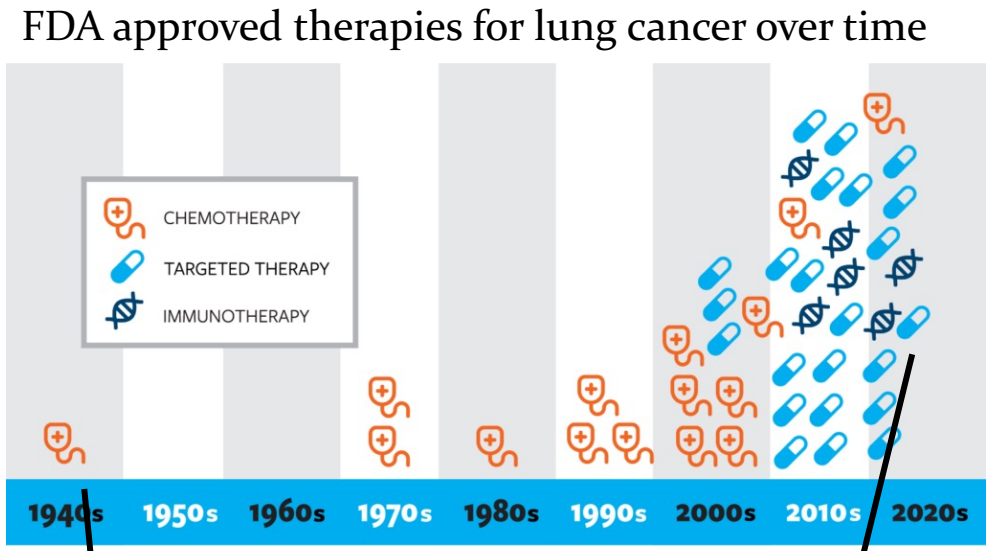
- AI for Health and Biology, Software, and Technology
- Based in Heidelberg/Germany, Zug/Switzerland and London/UK

- We **help identify, monitor, and treat disease** with Clinical AI
- We **create a map of the immune system** using AI-guided experimentation
- We **advance the science of AI for Health** in partnership with leading European/Swiss institutions (e.g., ETH Zurich, Oxford, Cambridge, King's College)

the team

Introduction

The good: drug development works for society

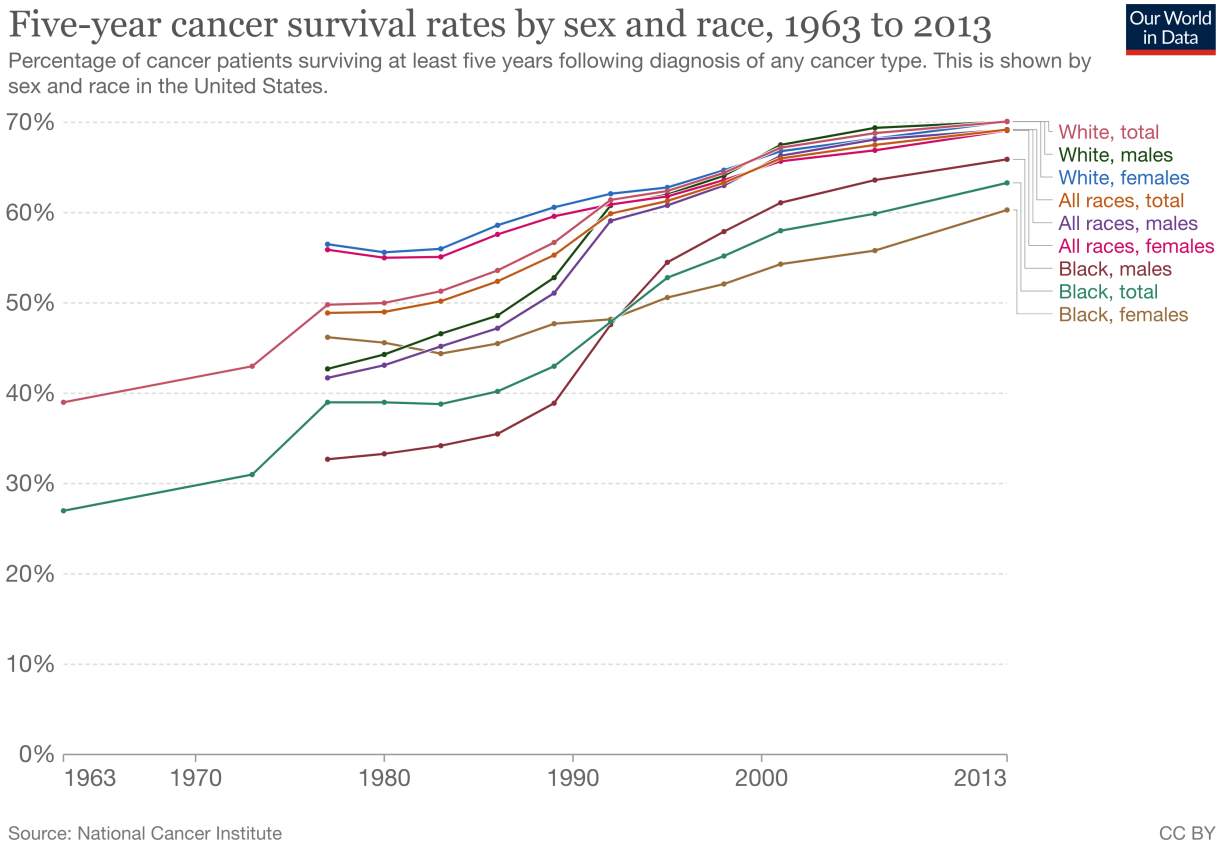


Mechlorethamine Hydrochloride

Candidate chemical warfare agent research

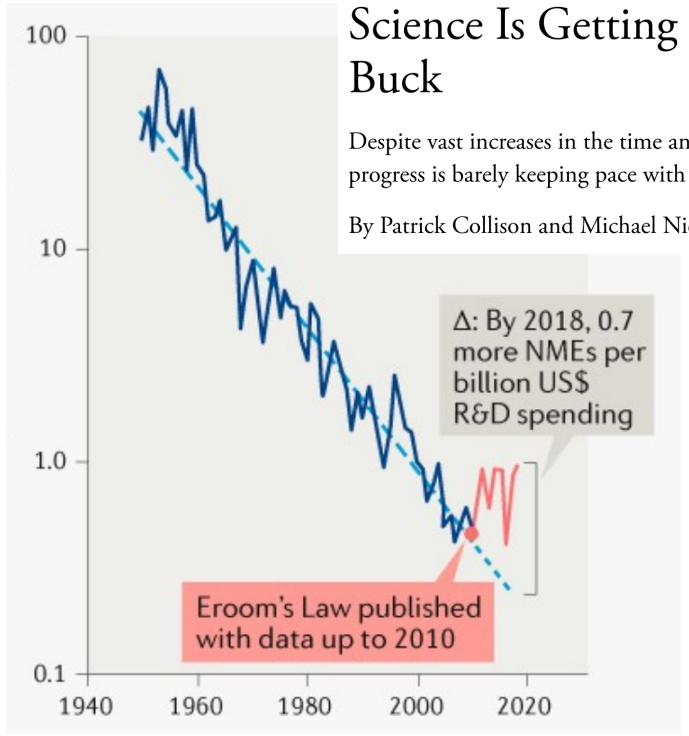
Mobocertinib

NSCLC with EGFR exon 20 insertion w/ progression after platinum therapy



Innovative medicines play a crucial role in enabling humans to live longer, healthier lives.

The bad: we are doing less, with (exponentially) more



"Eroom's law":
Exponential drop in R&D productivity.

- **Failure is the norm**: Probability of success for a new medicine is **~5.5%**
- **Median cost** of a new drug is **$1.1 billion**
- Trend stable for decades with recent reversal (potentially) due to **increasing personalization** and **molecular/genetic evidence**.

Ringel, Michael S., et al. "Breaking Eroom's Law." Nature reviews. Drug discovery (2020).
Wouters et al. "Estimated Research and Development Investment Needed to Bring a New Medicine to Market, 2009-2018". JAMA (2020)
Ringel et al. "Breaking Eroom's law." Nature reviews Drug Discovery (2020)

What do we know works?

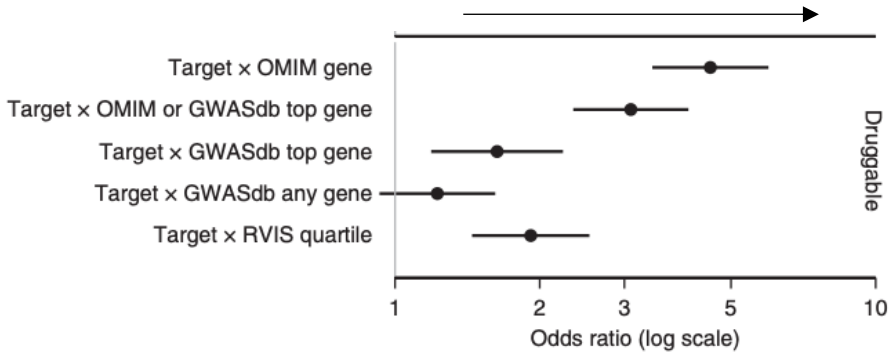
Population-scale Biobanks –
Variation in Genetic Background
(>100’000 people)

FINNGEN

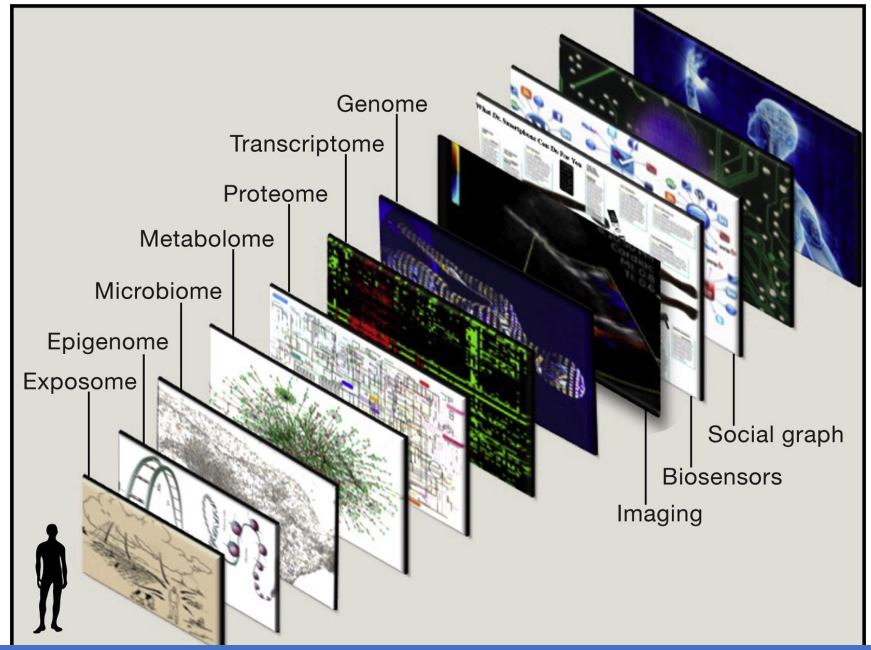
biobank uk

All of Us RESEARCH PROGRAM

Higher probability of approval (EU/US)



Targeted Biomarkers supporting Therapy
2x (immune) to 8x (oncology) higher probability of trial success



A causal & targeted link between disease of interest and a molecular mechanism substantially improves our chances of successful translation.

Wong et al. “Estimation of clinical trial success rates and related parameters “ Biostatistics (2018)

Topol, Eric J. "Individualized medicine from prewomb to tomb." Cell 157.1 (2014): 241-253.

Nelson, Matthew R., et al. "The support of human genetic evidence for approved drug indications." Nature genetics 47.8 (2015): 856-860.

Ringel, Michael S., et al. "Breaking Eroom's Law." Nature reviews. Drug discovery (2020).

How can we utilise machine learning and experimentation to bring targeted medicines with strong causal evidence to patients faster?

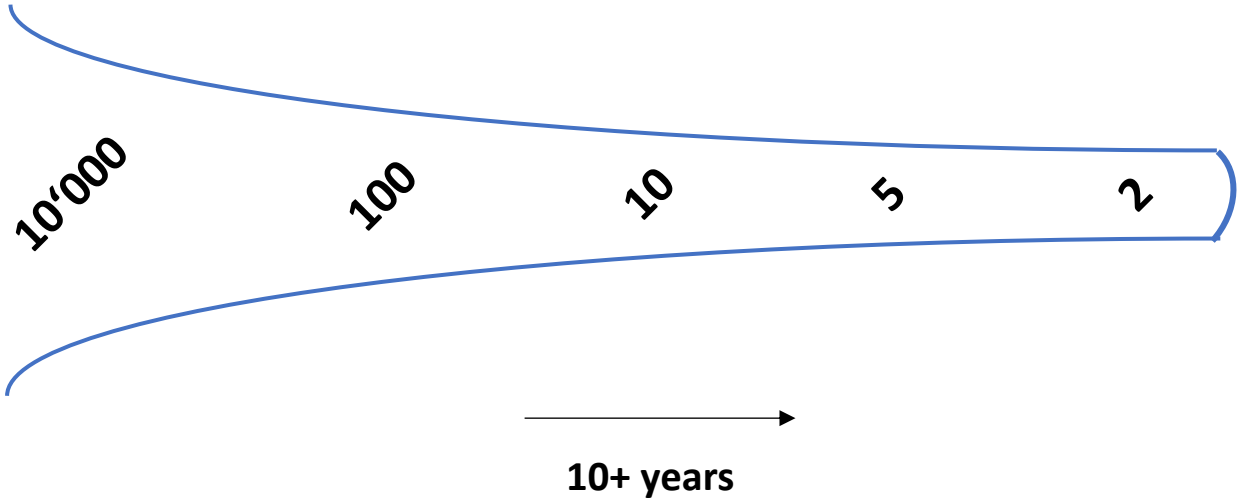
Agenda

1. Introduction
2. In-silico experimentation
3. Optimal experimental design
4. Causal discovery
5. Multimodal learning
6. Conclusion & discussion

} The AI-driven Stack for accelerated Biological Discovery

Drug development process

	Discovery	Preclinical	Phase I	Phase II	Phase III
Probability of success		60%	50%	30%	60%

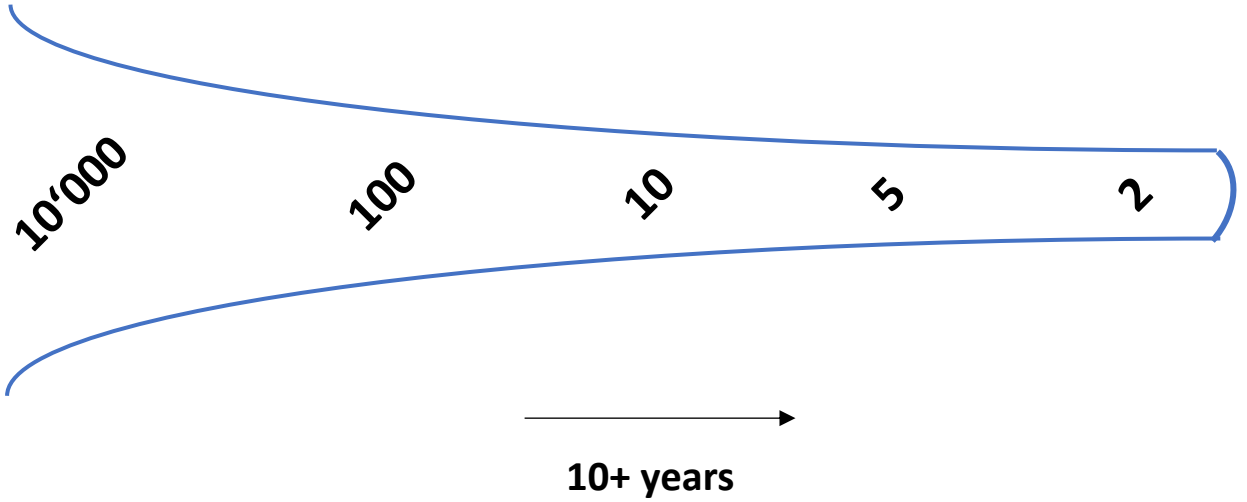


1 approved medicine

Cook, David, et al. "Lessons learned from the fate of AstraZeneca's drug pipeline: a five-dimensional framework." Nature reviews Drug discovery 13.6 (2014): 419-431.

Drug development process

	Discovery	Preclinical	Phase I	Phase II	Phase III
Probability of success		60%	50%	30%	60%



1 approved medicine

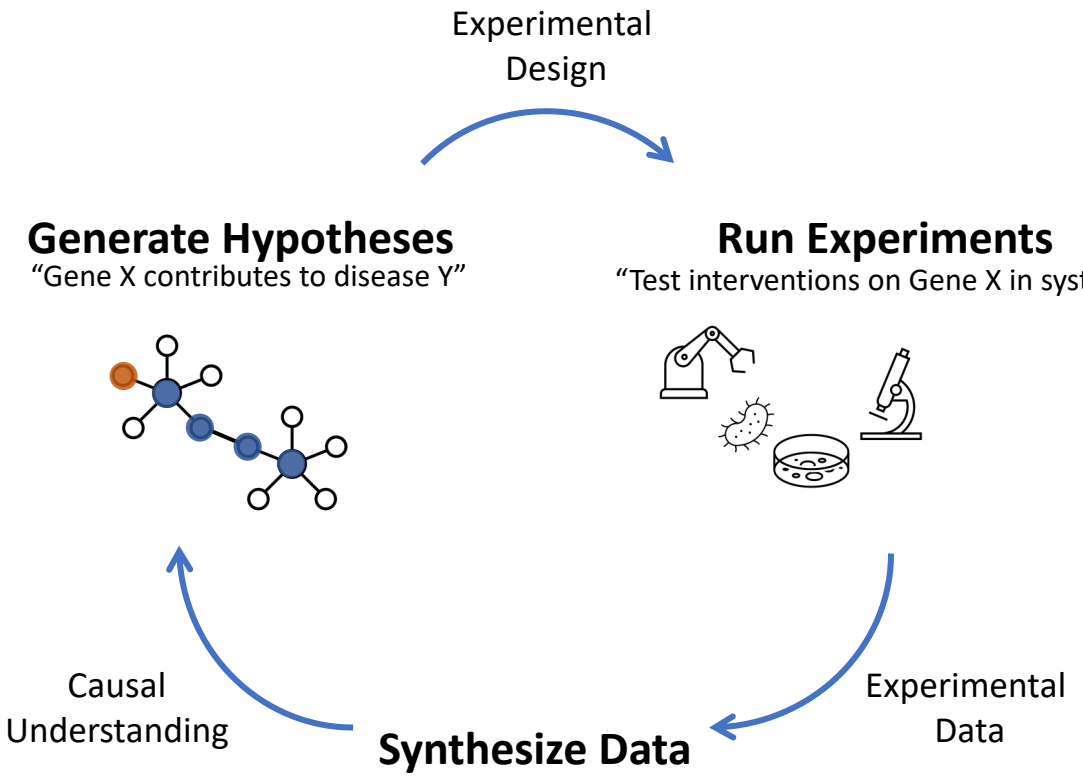
Drug discovery is an information science mainly concerned with making optimal decisions under uncertainty at every step of the development process.

Possibly the most complex human endeavor.

Cook, David, et al. "Lessons learned from the fate of AstraZeneca's drug pipeline: a five-dimensional framework." Nature reviews Drug discovery 13.6 (2014): 419-431.

An AI-driven stack for biological discovery

How can we accelerate biological discovery with modern machine learning tools?



An AI-driven stack for biological discovery

How can we accelerate biological discovery with modern machine learning tools?

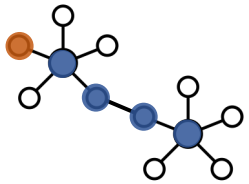
Experimental Design

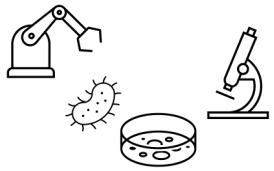
2 **Run experiments that yield more**

Optimal experimental design

Generate Hypotheses
"Gene X contributes to disease Y"

Run Experiments
"Test interventions on Gene X in system Z"





1 **Run faster experimental cycles**

In-silico experimentation

4 **Make more of experiments**

Multimodal learning

Causal Understanding

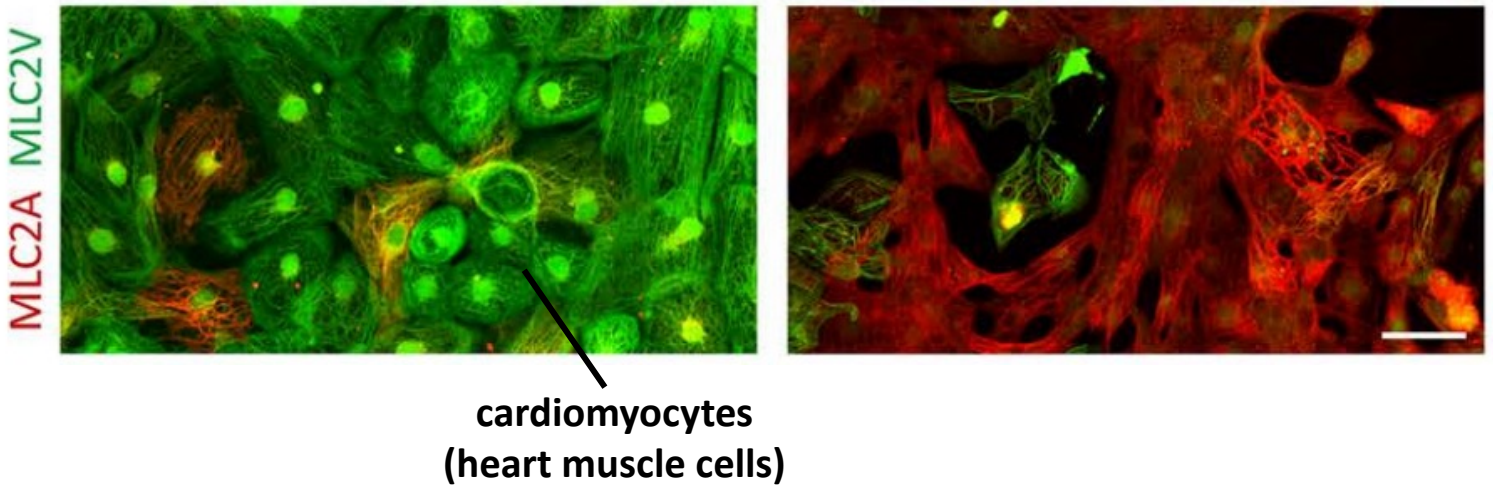
Synthesize Data

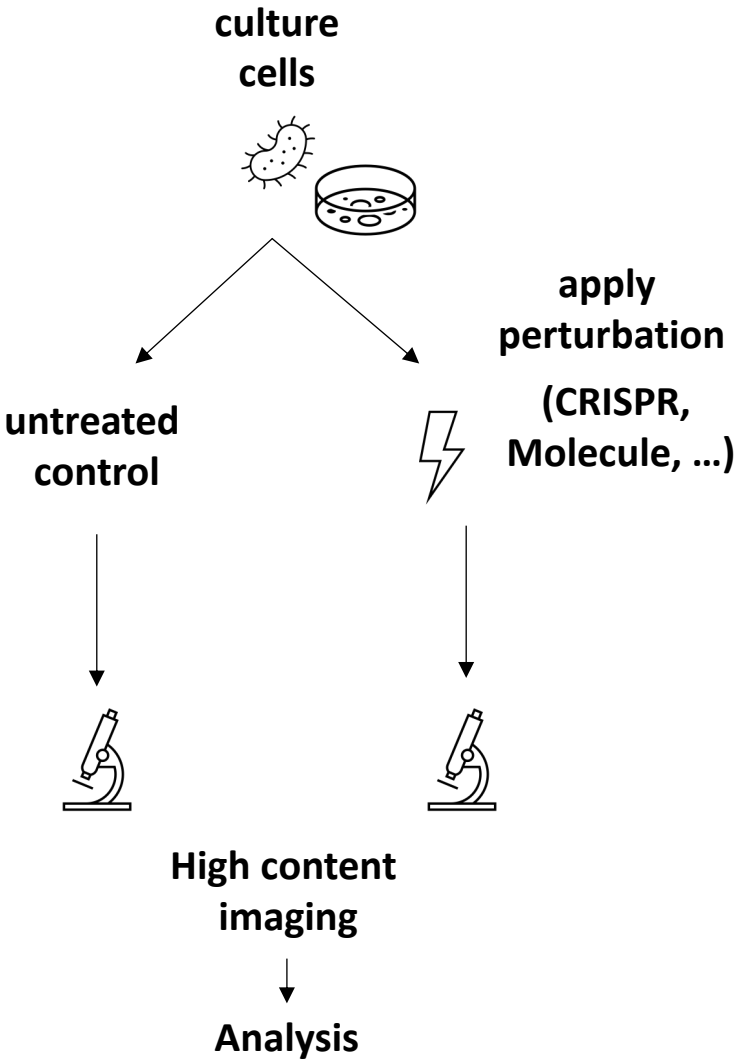
Experimental Data

3 **Distill robust causal relationships from data**

Causal discovery

Decoding the Language of the Cell with Perturbation Experiments





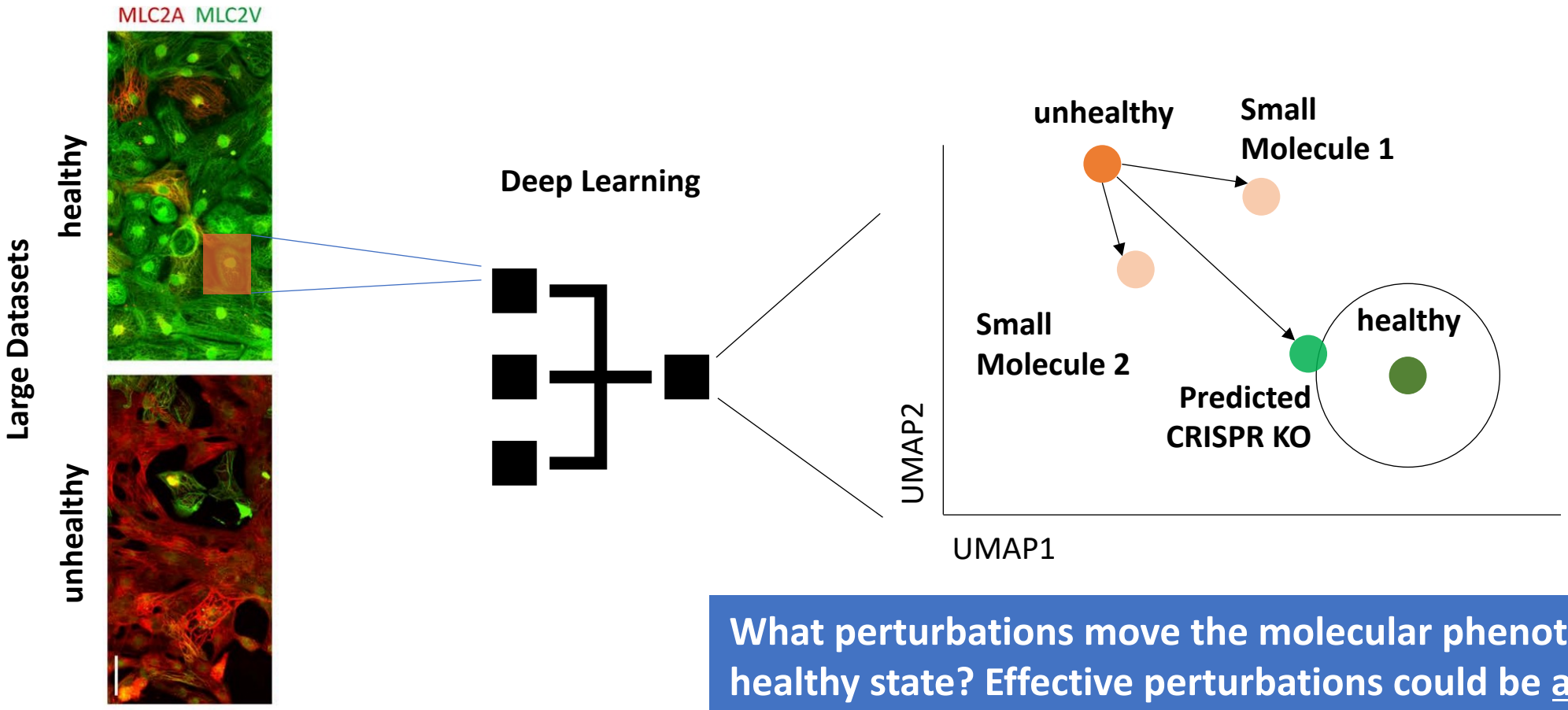
Comparison between experimental readouts in control and intervention settings enables us to understand causal effects of cellular perturbations.

1 Run faster experimental cycles

Cyganek, Lukas, et al. "Deep phenotyping of human induced pluripotent stem cell–derived atrial and ventricular cardiomyocytes." JCI insight 3.12 (2018).

In-silico experimentation

Decoding the Language of the Cell with Perturbation Experiments



What perturbations move the molecular phenotype towards the healthy state? Effective perturbations could be attractive treatments for further development.

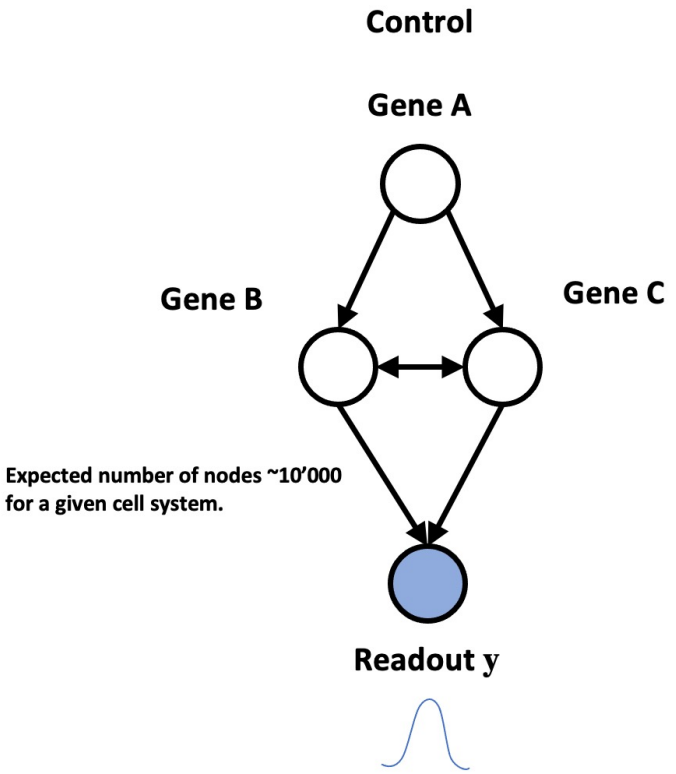
1 Run faster experimental cycles

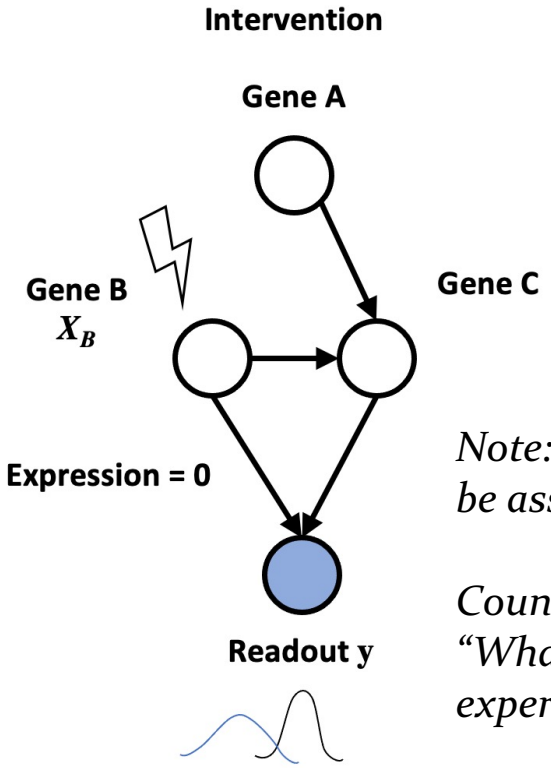
Cyganek, Lukas, et al. "Deep phenotyping of human induced pluripotent stem cell–derived atrial and ventricular cardiomyocytes." JCI insight 3.12 (2018).

In-silico experimentation

In-silico experimental perturbation models

Observed conformation change in causal generative process (CGP)!





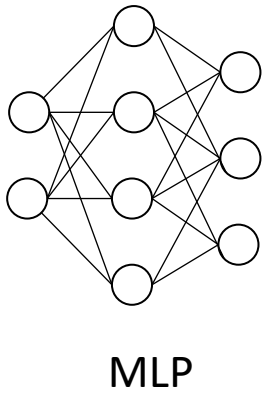
Note: acyclicity can generally not be assumed.

Counterfactual prediction: "What if I conducted experiment…?"

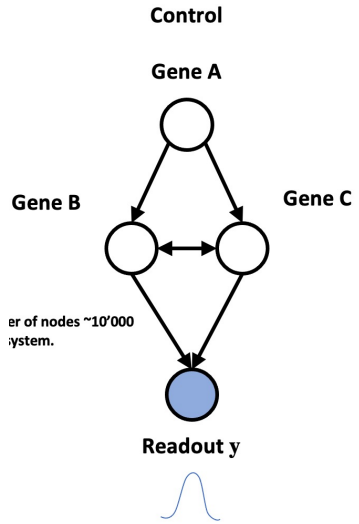
e.g., markers, imaging, omics

The problem: supervised learning methods assume smoothness

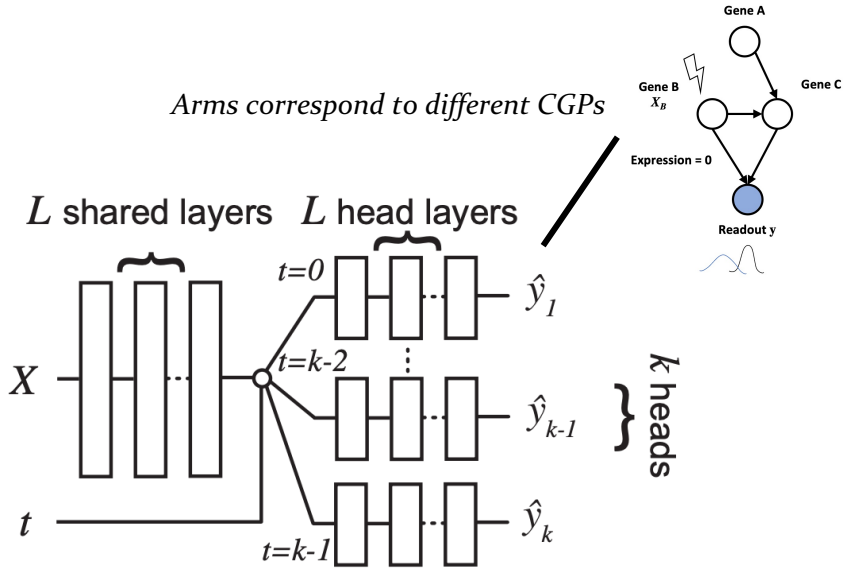
Supervised learning methods assume **smoothly interpolate-able** generative process



MLP



Need **conditional plasticity** to model conformational change in generative process



Schwab et al. "Perfect Match: A Simple Method for Learning Representations For Counterfactual Inference With Neural Networks" arXiv preprint arXiv:1810.00656 (2018)
Parbhoo et al. "NCoRE: Neural Counterfactual Representation Learning for Combinations of Treatments" arXiv preprint arXiv:2103.11175 (2021)
Schwab et al. "Learning Counterfactual Representations for Estimating Individual Dose-Response Curves" AAAI 2020

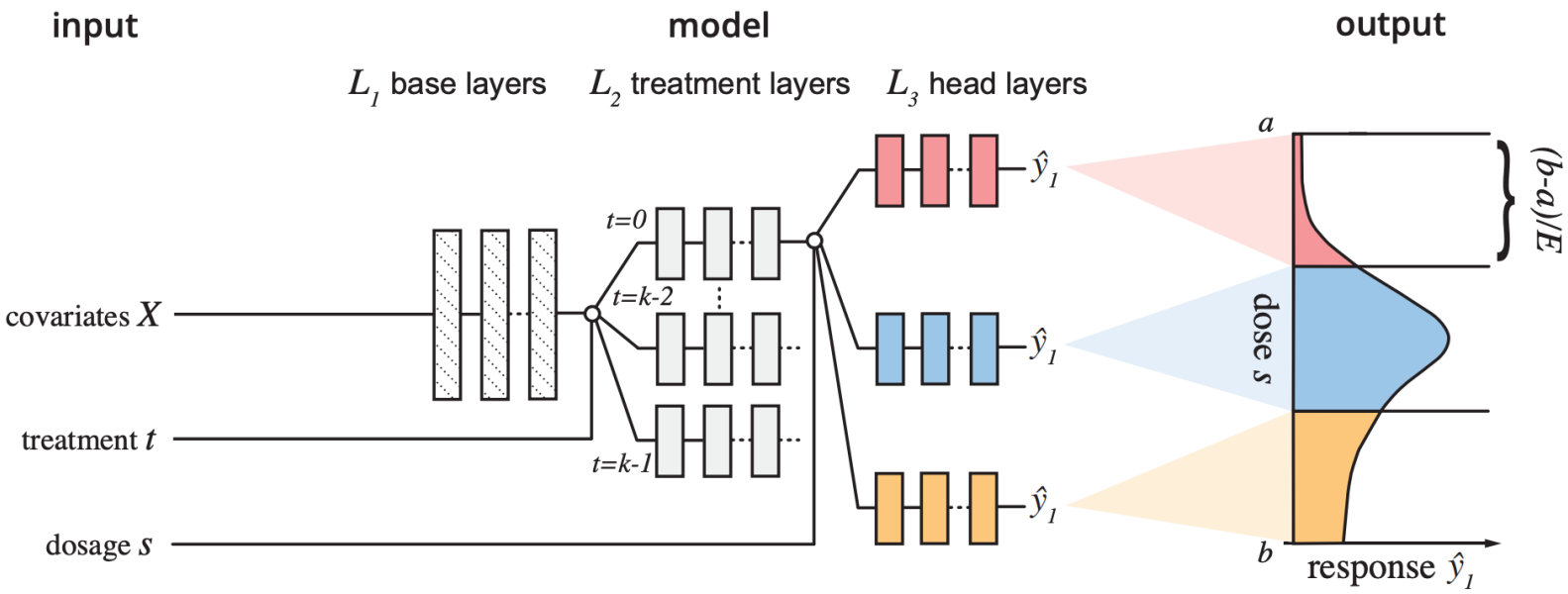
1 Run faster experimental cycles

In-silico experimentation

Different settings of interest require different levels of conditional plasticity

Multi-treatment with dosages

Dose-response Networks (DRNet)

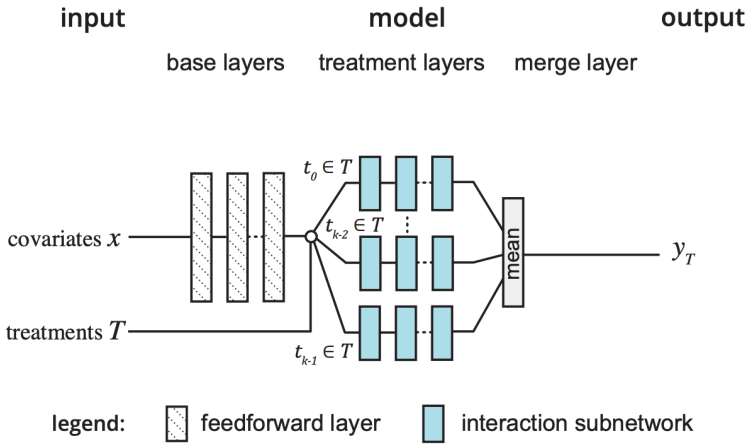


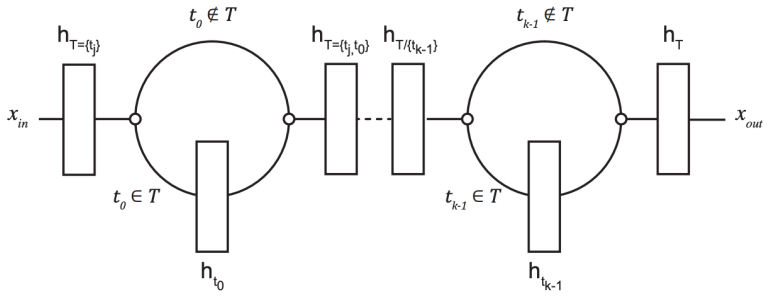
Schwab et al. "Perfect Match: A Simple Method for Learning Representations For Counterfactual Inference With Neural Networks" arXiv preprint arXiv:1810.00656 (2018)
Parbhoo et al. "NCoRE: Neural Counterfactual Representation Learning for Combinations of Treatments" arXiv preprint arXiv:2103.11175 (2021)
Schwab et al. "Learning Counterfactual Representations for Estimating Individual Dose-Response Curves" AAAI 2020

Different settings of interest require different levels of conditional plasticity

Multi treatment & combinations

Neural Counterfactual Relation Estimation (NCoRE)



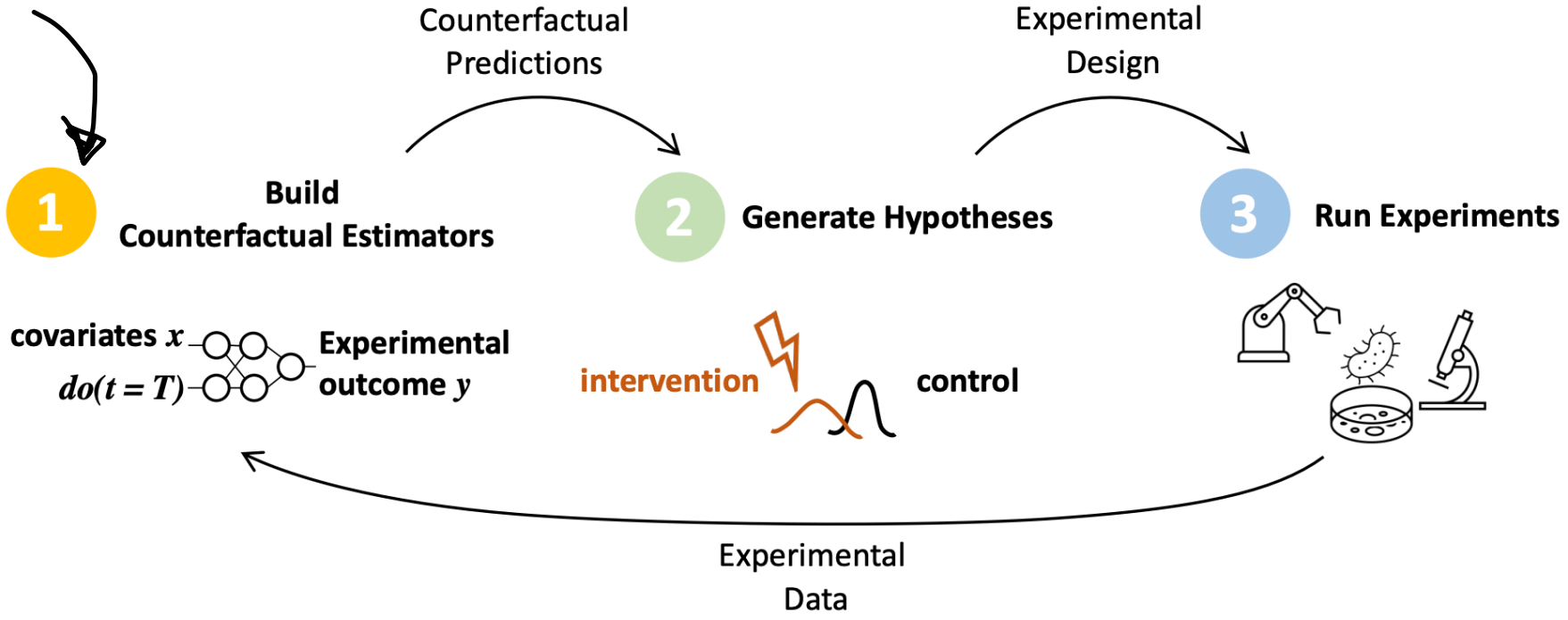


Conditional plasticity in models enables flexibility to accurately reflect conformation changes in the underlying causal generative process.

Schwab et al. "Perfect Match: A Simple Method for Learning Representations For Counterfactual Inference With Neural Networks" arXiv preprint arXiv:1810.00656 (2018)
Parbhoo et al. "NCoRE: Neural Counterfactual Representation Learning for Combinations of Treatments" arXiv preprint arXiv:2103.11175 (2021)
Schwab et al. "Learning Counterfactual Representations for Estimating Individual Dose-Response Curves" AAAI 2020

Integrating *in-silico* experimentation into exploration campaigns

Insert *in-silico* counterfactuals into experimentation campaigns



1 Run faster experimental cycles

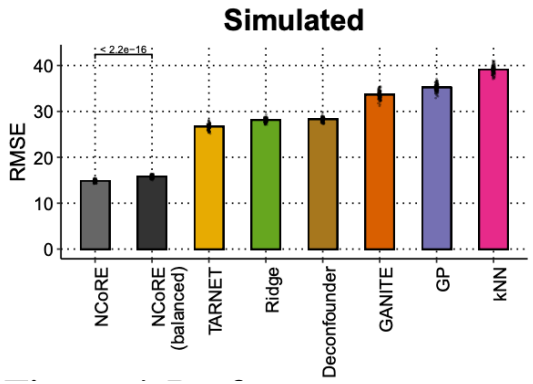
Mehrjou et al GeneDisco: A benchmark for experimental design in drug discovery. ICLR 2022

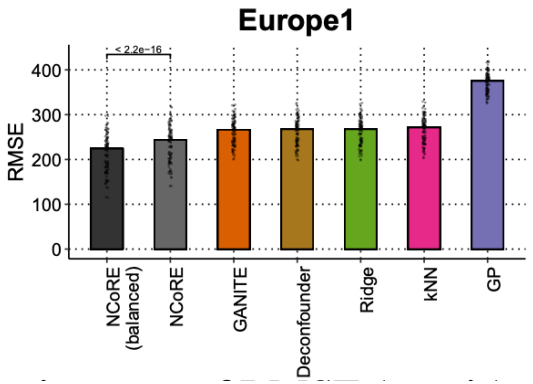
In-silico experimentation

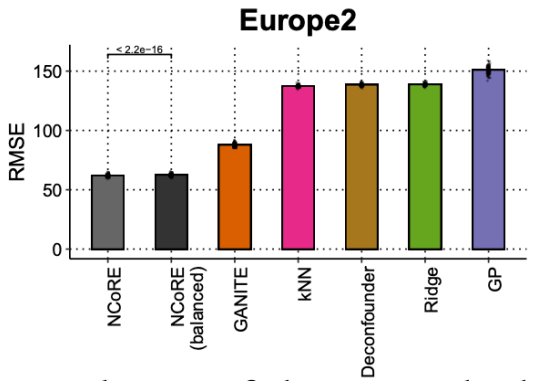
Results on counterfactual prediction

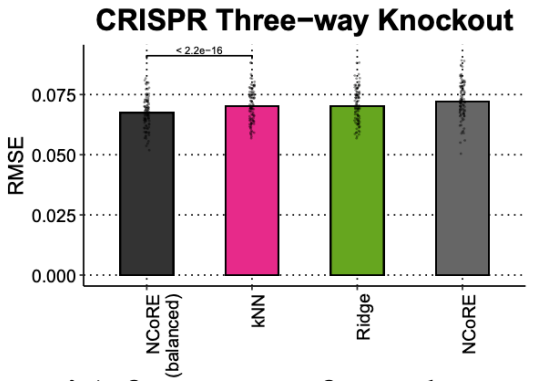
Model inputs: patient covariates, gene embeddings, **output**: counterfactuals under intervention

Known graph	HIV combination therapy	HIV combination therapy	3-way CRISPR KO







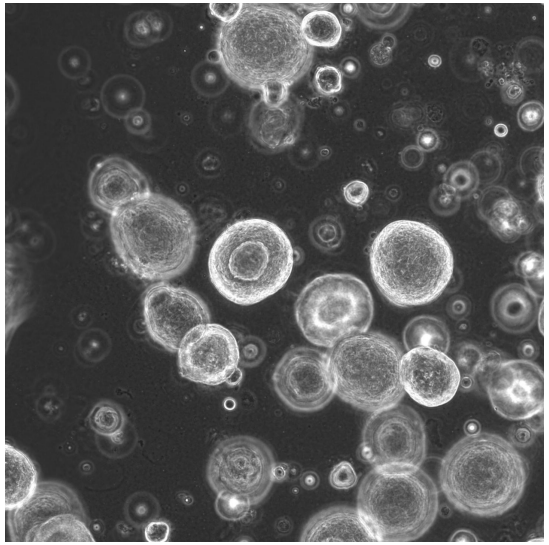


Improving models' ability to address loss of smoothness in the causal generative process leads to better performance in estimating "what if..?" outcomes.

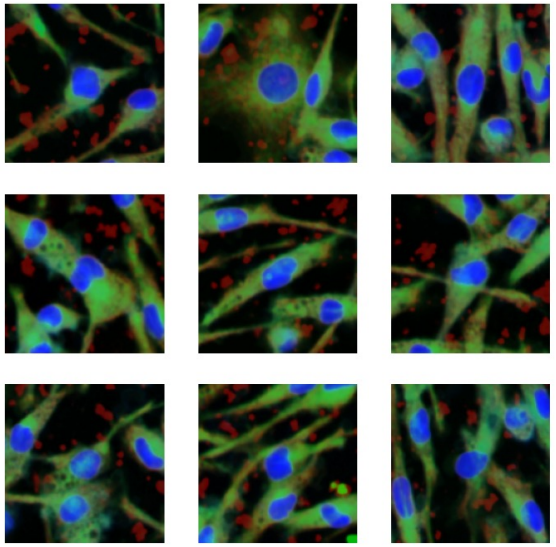
Schwab et al. "Perfect Match: A Simple Method for Learning Representations For Counterfactual Inference With Neural Networks" arXiv preprint arXiv:1810.00656 (2018)
Parbhoo et al. "NCoRE: Neural Counterfactual Representation Learning for Combinations of Treatments" arXiv preprint arXiv:2103.11175 (2021)
Schwab et al. "Learning Counterfactual Representations for Estimating Individual Dose-Response Curves" AAAI 2020

Future challenge: Experimental context matters!

Patient derived organoid



In-vitro cell cultures



Animal model

The same stimulus can elicit very different responses depending on context.
Context matters when simulating experiments.

PATIENT-DERIVED ORGANOIDS CAN PREDICT RESPONSE TO CHEMOTHERAPY IN METASTATIC COLORECTAL CANCER PATIENTS. Science Translational Medicine (2019)

1 Run faster experimental cycles

In-silico experimentation

An AI-driven stack for biological discovery

How can we accelerate biological discovery with modern machine learning tools?

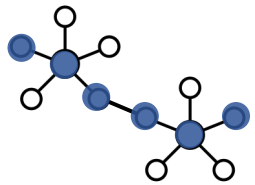
Experimental Design

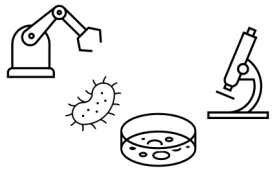
2 **Run experiments that yield more**

Optimal experimental design

Generate Hypotheses
"Gene X contributes to disease Y"

Run Experiments
"Test interventions on Gene X in system Z"





1 **Run faster experimental cycles**

In-silico experimentation

4 **Make more of experiments**

Multimodal learning

Causal Understanding

Synthesize Data

Experimental Data

3 **Distill robust causal relationships from data**

Causal discovery

Now that we have tools to predict what results may be surfaced in genetic experiments, how do we choose the optimal set under a given capacity budget?

DISCOBAX: DISCOVERY OF OPTIMAL INTERVENTION SETS IN GENOMIC EXPERIMENT DESIGN

Clare Lyle[2], **Arash Mehrjou**[1], **Andrew Jesson**[2], **Pascal Notin**[2], **Stefan Bauer**[1,3], **Yarin Gal**[2], **Patrick Schwab**[1]
[1]GlaxoSmithKline, Artificial Intelligence & Machine Learning
[2] Department of Computer Science, University of Oxford, [3] KTH Stockholm

GENEDISCO: A BENCHMARK FOR EXPERIMENTAL DESIGN IN DRUG DISCOVERY

Arash Mehrjou[1], **Ashkan Soleymani**[2], **Andrew Jesson**[3], **Pascal Notin**[3], **Yarin Gal**[3], **Stefan Bauer**[1,4,5], **Patrick Schwab**[1]
[1]GlaxoSmithKline, Artificial Intelligence & Machine Learning
[2] MIT, [3] Department of Computer Science, University of Oxford
[4] CIFAR Azrieli Global Scholar, [5] KTH Stockholm

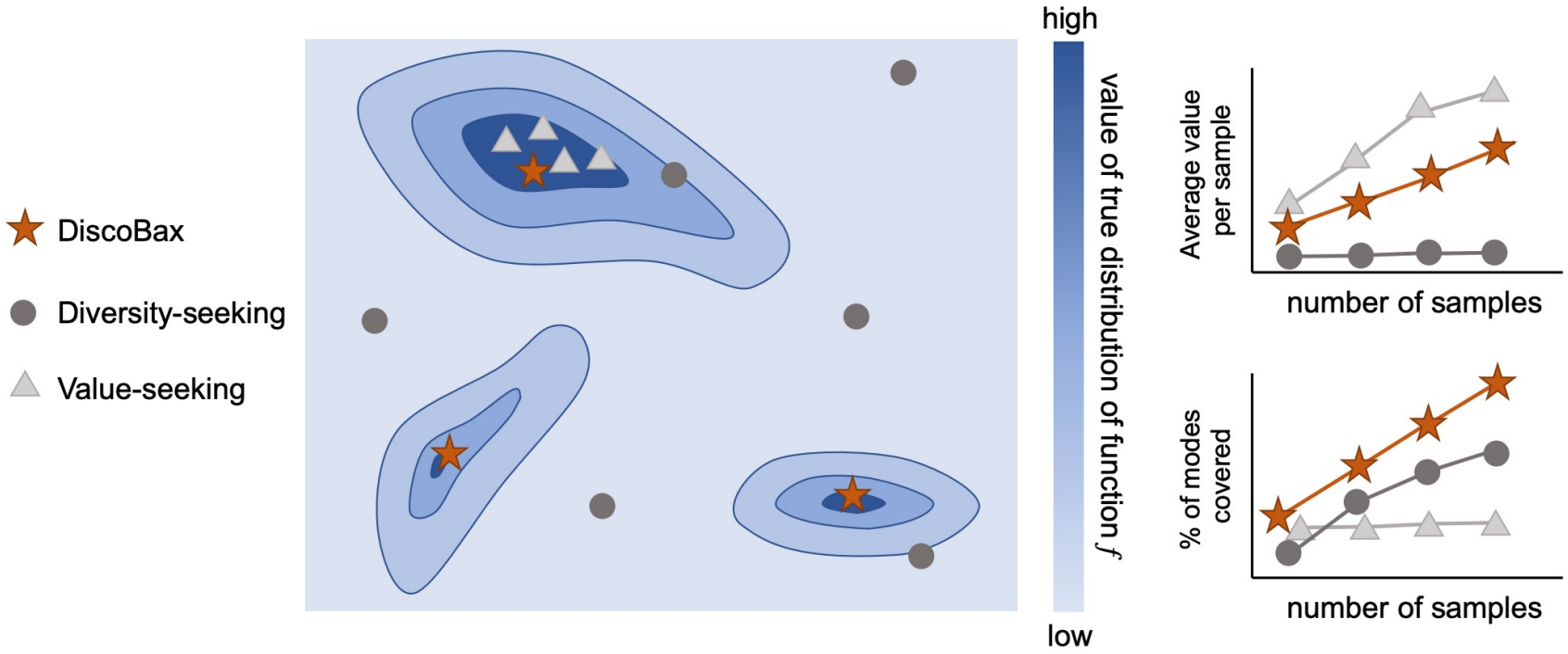
Lyle et al. DiscoBAX - Discovery of optimal intervention sets in genomic experiment design https://openreview.net/forum?id=mBkUeW8rpD6

 Run experiments that yield more

Optimal experimental design

The problem: traditional active learning methods seek data value or diversity

In genetic experiments, we most often want to maximize coverage of all potential mechanisms

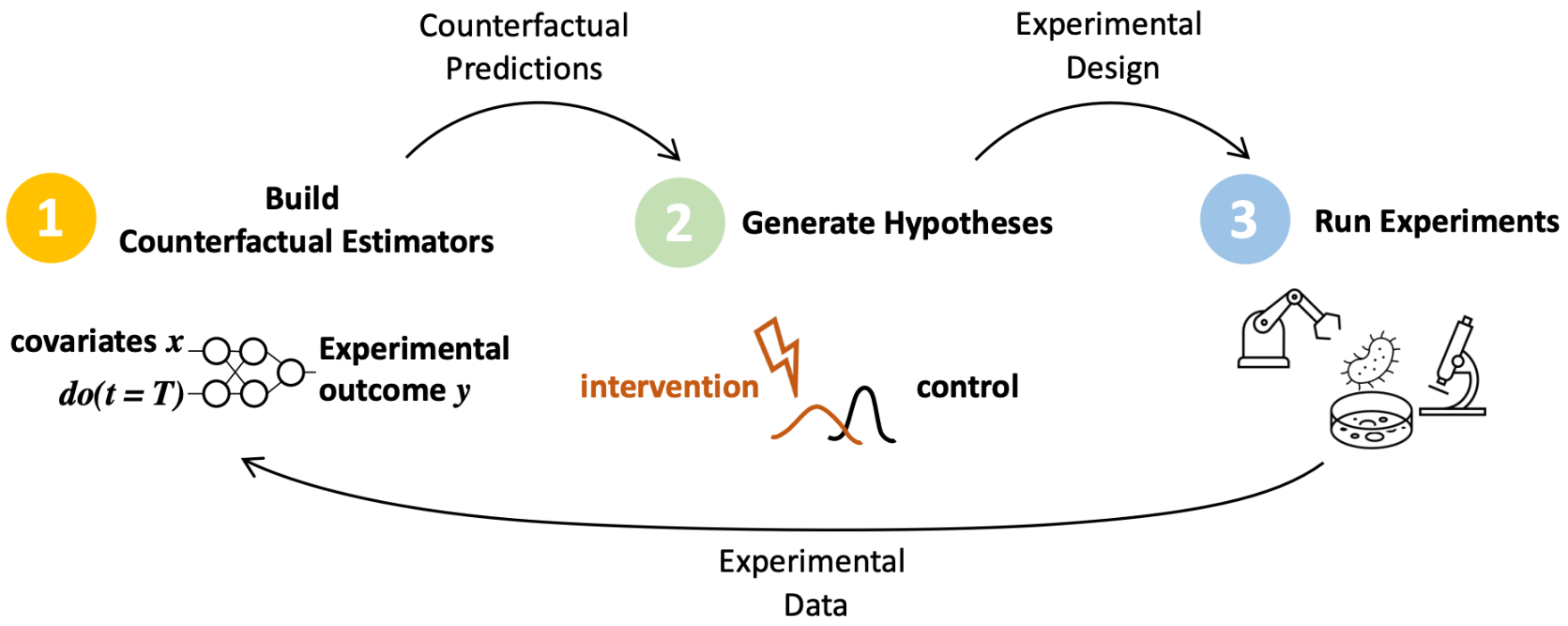


Biological model systems are imperfect models for humans and covering the potential mechanisms could help maximize the coverage of mechanisms that may translate to downstream experiments.

Lyle et al. DiscoBAX - Discovery of optimal intervention sets in genomic experiment design https://openreview.net/forum?id=mBkUeW8rpD6

How do we measure our ability to improve experimental design in genetic experiments?

Setting overview



Lyle et al. DiscoBAX - Discovery of optimal intervention sets in genomic experiment design https://openreview.net/forum?id=mBkUeW8rpD6

Run experiments that yield more

Optimal experimental design

How do we measure our ability to improve experimental design in genetic experiments?

Range of experimental configurations

Optimal experimental design

Batch Sizes	Tasks	Methods	Predictors	Models
16	T-cell Proliferation	Random	CCLE	Random Forest
32	IL-2	Top-uncertain	STRING	Bayesian NN
64	IFNy	Soft-uncertain		
128	NK Cell Cytotoxicity	K-means (data)		
256	Tau Modulation	K-means (embedding)		
	SARS-Cov-2 Endosomal Entry	Margin		
		Coreset		
		BADGE		
		Adversarial BIM		

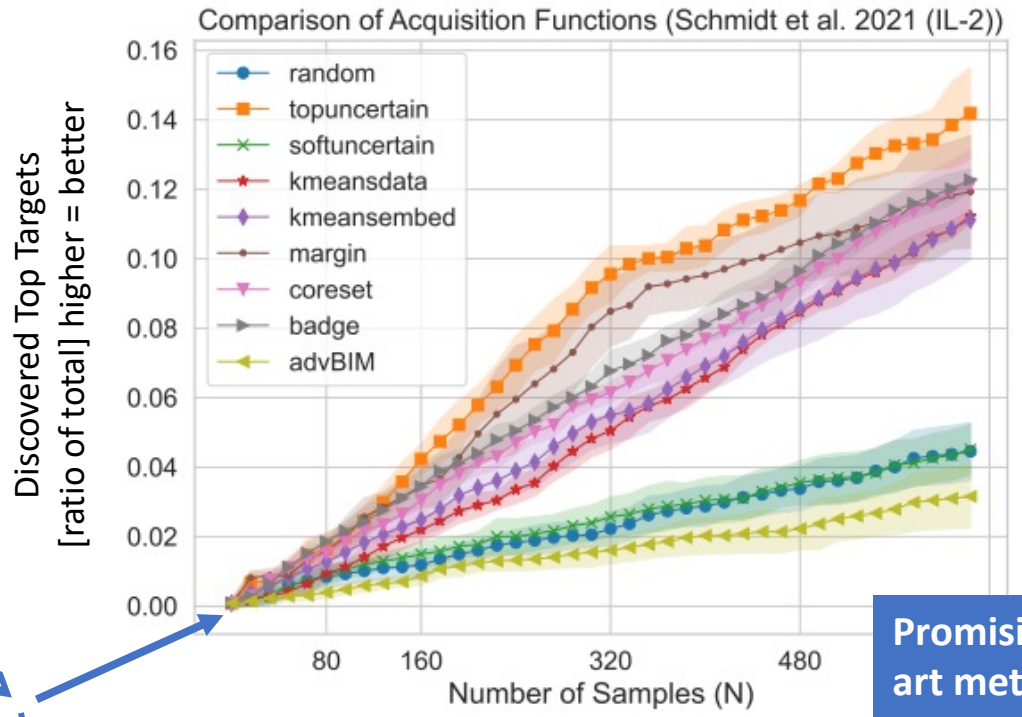
We tested a total of 5x6x9x2x2 = 1080 configurations using a total of >20 000 CPU hours of compute (5 random seeds each).

Lyle et al. DiscoBAX - Discovery of optimal intervention sets in genomic experiment design https://openreview.net/forum?id=mBkUeW8rpD6

GeneDisco: Results from simulated iterative campaigns using state-of-the-art methods

We observe settings in which we find up to 4x more significant movers than baselines with the same capacity (more data across screens available in reference)

CRISPRi T-cell IL-2 production screen
(IL-2 readout, primary T-cells)



Optimized Exploration
~16% discovered

Naïve Exploration
~4% discovered

Every data point = experimental cycle

Promising early results with state-of-the-art methods – but can we do better?

Lyle et al. DiscoBAX - Discovery of optimal intervention sets in genomic experiment design https://openreview.net/forum?id=mBkUeW8rpD6

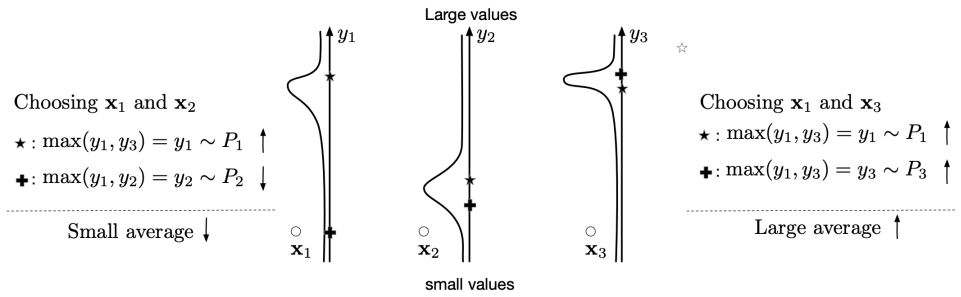
DiscoBAX (Bayesian Algorithm Execution) for optimal intervention set selection

Uncertainty-aware models: Gaussian processes, Bayesian neural networks with Monte Carlo dropout

DiscoBAX consists of two stages:

- **Subset maximization (Alg. 1)**: Identify an uncorrelated subset *S* of interventions that maximizes expected value over the unmeasured downstream phenotype (e.g., ground truth unobserved phenotype) using submodularity (diminishing returns) for an approximate solution.

  Assumption: f_out and f_ip are connected via Bernoulli multiplicative and additive Gaussian noise.
- **Active sampling (Alg. 2):** Optimize set expected information gain (EIG) out of subset *S*.



Algorithm 1 SubsetSelect (Multiplicative Noise)

Require: integer $k > 0$, set $\mathcal{X}$, distribution $P(\eta)$, sampled $\widehat{f}_{\text{ip}} : \mathcal{X} \to \mathbb{R}$

$S \leftarrow \emptyset$

$\widehat{f}_{\text{out}}(\mathbf{x}; \eta) := \widehat{f}_{\text{ip}}(\mathbf{x})\eta(\mathbf{x})$

for $i < k$ **do**

$$S \leftarrow S \cup \{\arg\max_{x \in \mathcal{X} \setminus S} \mathbb{E}_\eta[\max_{y \in S \cup \{x\}} \widehat{f}_{\text{out}}(\mathbf{x}; \eta)]\}$$

end for

return S

Algorithm 2 DiscoBAX

Require: finite sample set $\mathcal{X}$, budget T, Monte Carlo parameter $\ell \in \mathbb{N}$

$\mathcal{D} \leftarrow \emptyset$

for $i < T$ **do**

sample $\{\widehat{f}_{\text{ip}}\}_{j=1}^{\ell} \sim P(f_{\text{ip}}|\mathcal{D})$

$S_j \leftarrow \text{SubsetSelect}(\widehat{f_{\text{ip},j}}), \forall j = 1, \ldots, \ell$

$\mathbf{x}_i \leftarrow \arg\max_{\mathbf{x} \in \mathcal{X}} \text{EIG}^v(\mathbf{x}, S_{j=1}^{\ell})$

query $f_{\text{ip}}(\mathbf{x}_i)$

$\mathcal{D} = \mathcal{D} \cup \{(\mathbf{x}_i, f_{\text{ip}}(\mathbf{x}_i)\}$

end for

return $\mathcal{D}$

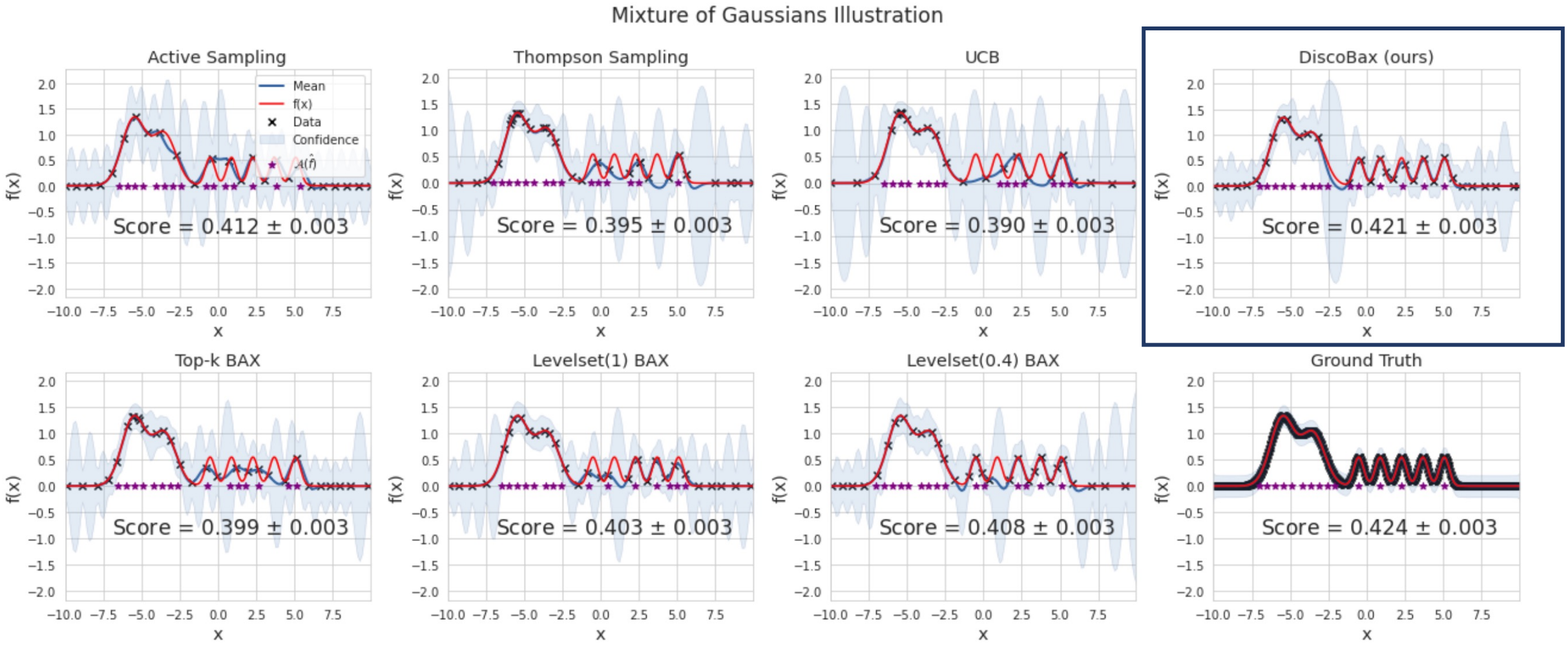
Lyle et al. DiscoBAX - Discovery of optimal intervention sets in genomic experiment design https://openreview.net/forum?id=mBkUeW8rpD6

Run experiments that yield more

Optimal experimental design

Simulation results (DiscoBAX)

Setting: Mixture of Gaussians, 30 iterations, batch size 1



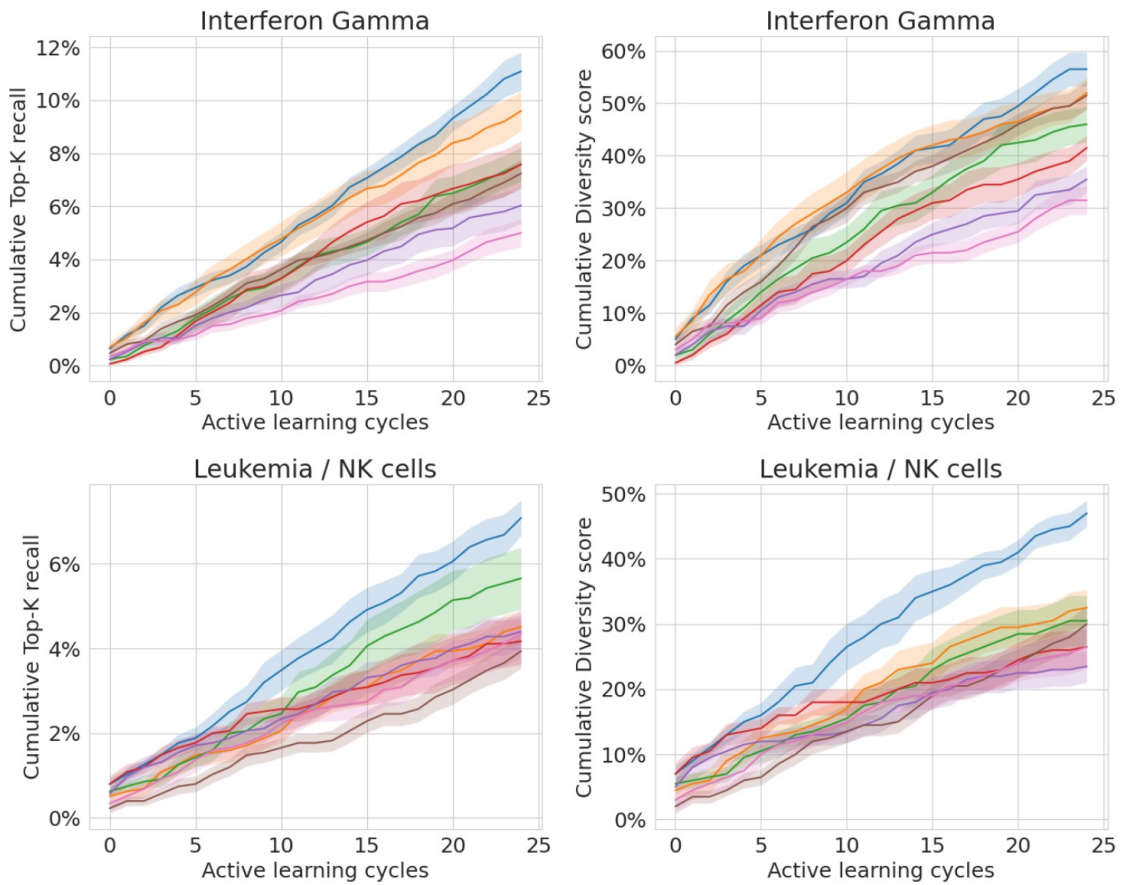
Lyle et al. DiscoBAX - Discovery of optimal intervention sets in genomic experiment design https://openreview.net/forum?id=mBkUeW8rpD6

Run experiments that yield more

Optimal experimental design

Benchmark results (DiscoBAX)

Setting: GeneDisco Leukemia/IFNg assays



(Leukemia/NK cells = cytotoxic sensitivity assay of K562 to primary NK cells; Zhuang et al 2019)

Method	Category	Top-K recall	Diversity score	Overall score
Random	-	29.3% (1.4%)	4.9% (0.3%)	12.0% (0.6%)
Thompson Sampling	Bandits	27.5% (1.5%)	4.8% (0.4%)	11.5% (0.7%)
UCB	Bayesian Optim.	33.5% (2.0%)	5.9% (0.5%)	14.1% (1.0%)
Coreset	Active learning	39.3% (1.9%)	5.5% (0.3%)	14.7% (0.8%)
Levelset Bax	BAX	35.4% (2.2%)	6.3% (0.4%)	15.0% (0.9%)
Top-K Bax	BAX	38.8% (2.3%)	6.8% (0.6%)	16.2% (1.2%)
DiscoBax (ours)	BAX	**44.1% (2.2%)**	**7.8% (0.5%)**	**18.6% (1.1%)**

We observe a ~doubling of the significant findings with phenotypes of interest over random selection with a more diverse coverage of mechanisms than baselines.

Lyle et al. DiscoBAX - Discovery of optimal intervention sets in genomic experiment design https://openreview.net/forum?id=mBkUeW8rpD6

An AI-driven stack for biological discovery

How can we accelerate biological discovery with modern machine learning tools?

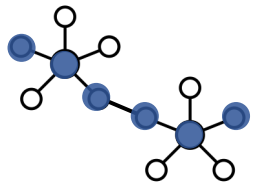
Experimental Design

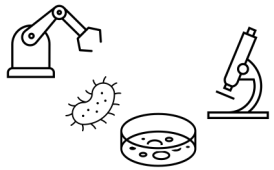
2 Run experiments that yield more

Optimal experimental design

Generate Hypotheses
"Gene X contributes to disease Y"

Run Experiments
"Test interventions on Gene X in system Z"





1 Run faster experimental cycles

In-silico experimentation

4 Make more of experiments

Multimodal learning

Causal Understanding

Synthesize Data

Experimental Data

3 Distill robust causal relationships from data

Causal discovery

Now that we are confronted with more data being generated faster: How do we distill the wealth of information into causal relationships?

CAUSALBENCH: A LARGE-SCALE BENCHMARK FOR NETWORK INFERENCE FROM SINGLE-CELL PERTURBATION DATA

Mathieu Chevalley[1], **Yusuf H Roohani**[1,2], **Arash Mehrjou**[1], **Jure Leskovec**[2], **Patrick Schwab**[1]
[1] GSK plc, Artificial Intelligence & Machine Learning, Zug, Switzerland
[2] Department of Computer Science, Stanford University, Stanford, CA, USA

Deriving Causal Order from Single-Variable Interventions: Guarantees & Algorithm

Mathieu Chevalley[1,2] **Patrick Schwab**[1] **Arash Mehrjou**[1]
[1]GSK.ai [2]ETH Zürich

THE CAUSALBENCH CHALLENGE: A MACHINE LEARNING CONTEST FOR GENE NETWORK INFERENCE FROM SINGLE-CELL PERTURBATION DATA

Organizers:

Mathieu Chevalley[5,7] **Jacob Sackett-Sanders**[7] **Yusuf Roohani**[7,8] **Pascal Notin**[9,10]
Arash Mehrjou[7†] **Patrick Schwab**[7†]

Participants:

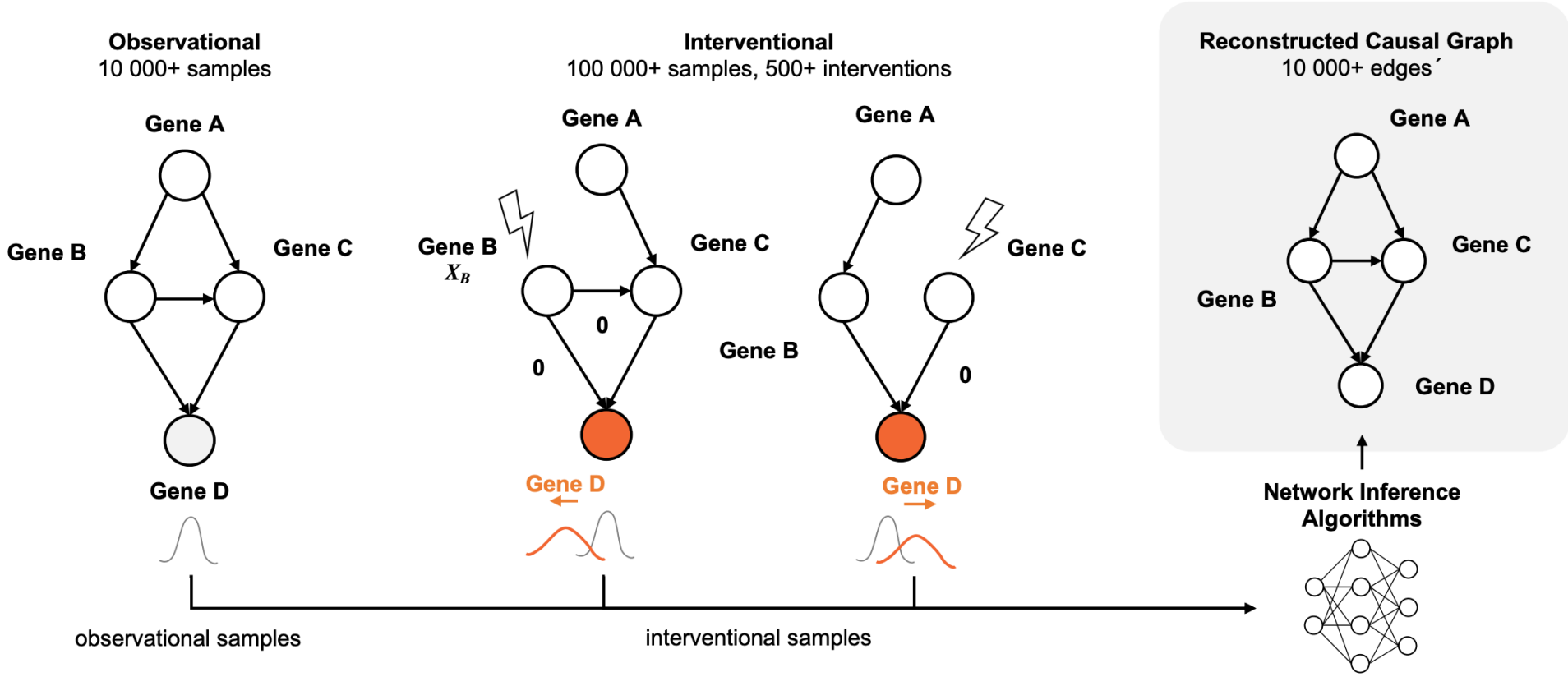
Artemy Bakulin[2] **Dariusz Brzezinski**[1] **Kaiwen Deng**[3] **Yuanfang Guan**[3]
Justin Hong[4] **Michael Ibrahim**[1] **Wojciech Kotlowski**[1] **Marcin Kowiel**[6]
Panagiotis Misiakos[5] **Achille Nazaret**[4] **Markus Püschel**[5] **Chris Wendler**[5]
[1]Poznan University of Technology [2] Lomonosov Moscow State University [3]University of Michigan
[4]Columbia University [5]ETH Zürich [6]Ryvu Therapeutics [7]GSK.ai [8]Stanford University
[9]Harvard Medical School [10]University of Oxford

3 Distill robust causal relationships from data

Causal discovery

The problem: Network inference from large-scale perturbational single-cell data

A new at-scale challenge for causal discovery



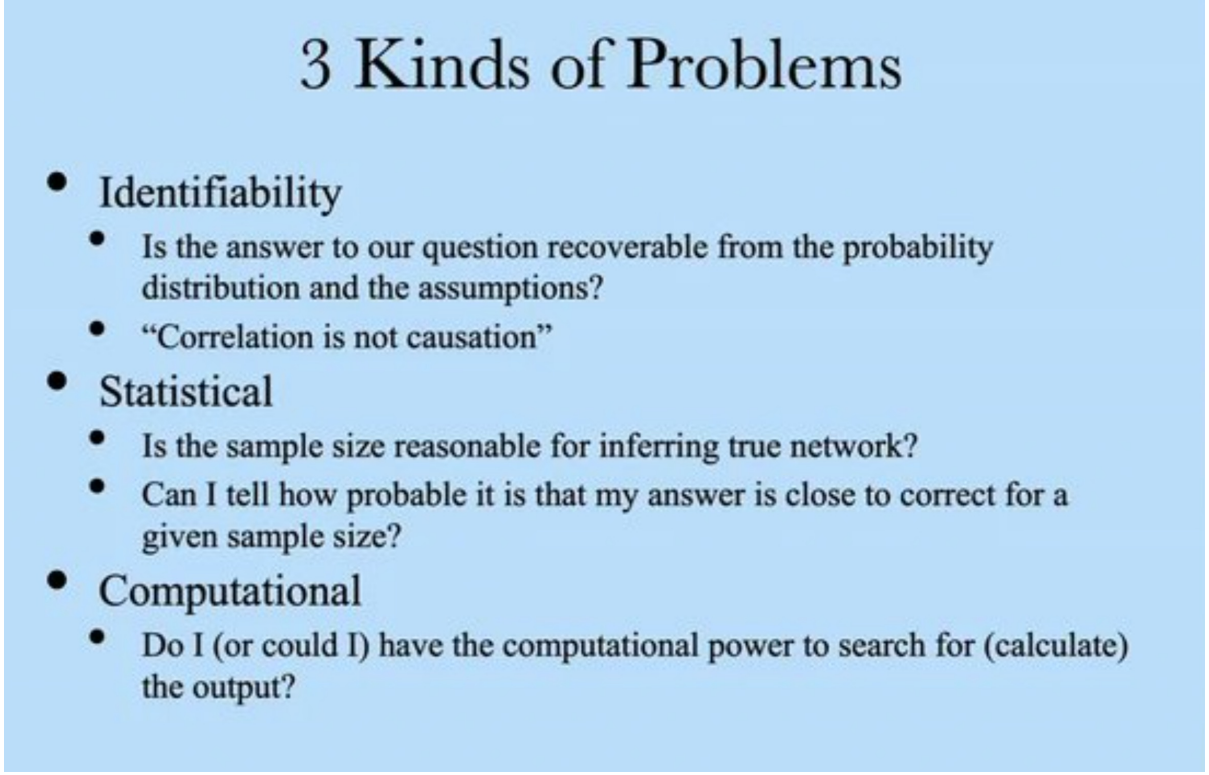
Perturbational single-cell datasets with transcriptome-wide readouts (e.g., PerturbSeq) generate interventional experimental data of unprecedented size also compared to other scientific disciplines.

Lyle et al. DiscoBAX - Discovery of optimal intervention sets in genomic experiment design https://openreview.net/forum?id=mBkUeW8rpD6

Distill robust causal relationships from data

Challenges in causal discovery from observational and interventional data

Core issues known since decades - slides repurposed from Spirtes (NeurIPS 2022)



Lyle et al. DiscoBAX - Discovery of optimal intervention sets in genomic experiment design https://openreview.net/forum?id=mBkUeW8rpD6

Related works

PC algorithm (Spirtes & Glymour 2000) –
Based on conditional independence tests.

Greedy Equivalence Search (Chickering 2002) –
Score-based method for equivalence classes.

Co-expression based methods (observational only, e.g., GRNBoost/SCENIC) -
Best method in observational benchmarks, fundamentally cannot distinguish Equivalence Classes.

Methods combining deep learning / causal discovery (e.g., ENCO, DCDI, NOTEARS)—
Recent developments promising to be more scalable and flexible than alternatives with potential to integrate perturbational data. (DCDI-DSF – normalizing flows, DCDI-G – Gaussian c. distr.)

Common limitations include underlying acyclicity assumptions, SUTVA, time-dependence/cell-cycle not considered, scalability. Results are often overstated and based only on simulation studies (lack of real world data).

Lyle et al. DiscoBAX - Discovery of optimal intervention sets in genomic experiment design https://openreview.net/forum?id=mBkUeW8rpD6

Benchmarking network inference from mixed observational and interventional data at single-cell scale

No ground truth. Approach: Compare networks against (i) biological plausibility and (ii) statistical evidence.

Evaluating gene-gene networks' biological plausibility:

- Physical interactions (STRING physical layer)
- Protein-protein interactions (STRING network layer)
- Protein complexes (CORUM database)
- CHIP-Seq (in same cell line or cell-type)

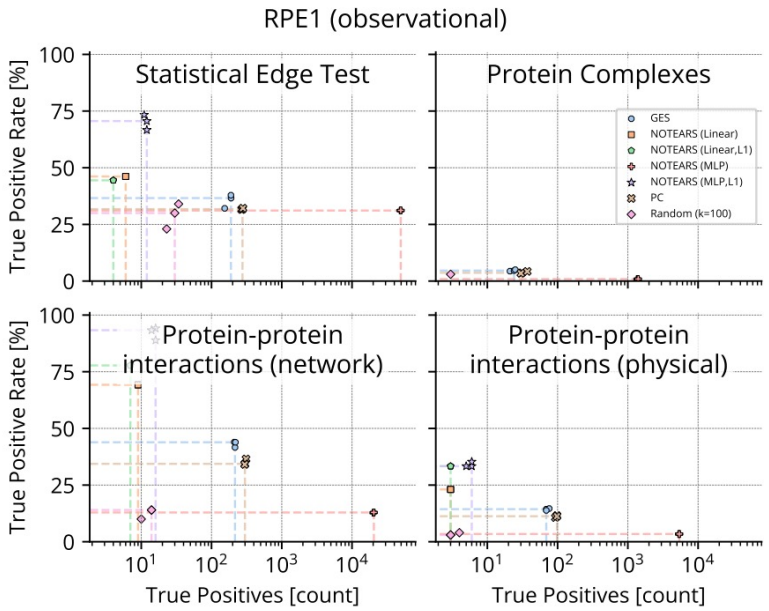
Evaluating against statistical evidence:

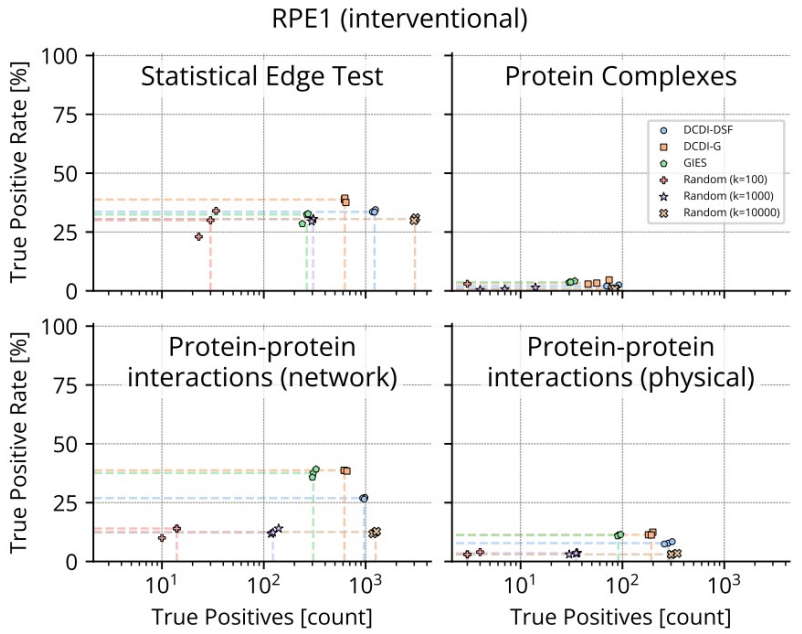
- For edge A->B, compare **observational distribution of B** with **interventional distribution when A is targeted** – test for equal central tendencies (MWW test).
- Threshold free: Continuous distribution distance (e.g., Wasserstein)

Challenges: Cell-type specificity and time-dependence/cell-cycle dependence for biological plausibility. Precision is difficult to estimate given it is possible that links are substantiated by other biological evidence.

Lyle et al. DiscoBAX - Discovery of optimal intervention sets in genomic experiment design https://openreview.net/forum?id=mBkUeW8rpD6

Benchmarking network inference from mixed observational and interventional data at single-cell scale





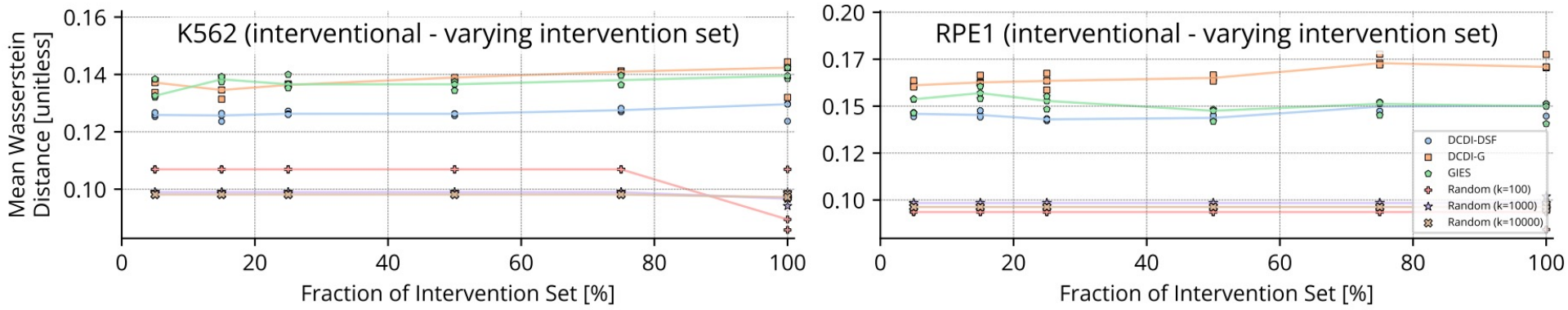
Model	Average rank
Random (k = 100)	10.5
Random (k = 1000)	9.99
Random (k = 10000)	8.66
GIES	8.33
PC	7.66
GES	6.33
NOTEARS (MLP, L1)	5.16
NOTEARS (Linear)	5.0
DCDI-DSF	4.16
NOTEARS (Linear, L1)	3.83
DCDI-G	3.66
NOTEARS (MLP)	2.83
GRNBoost	1.83

Equal weight precision/recall.

Overall, while NOTEARS performs the best out of the network inference methods, no method that utilizes perturbational data consistently outperforms GRNBoost that only uses observational data.

Lyle et al. DiscoBAX - Discovery of optimal intervention sets in genomic experiment design https://openreview.net/forum?id=mBkUeW8rpD6

Performance when varying the number of perturbational data points available

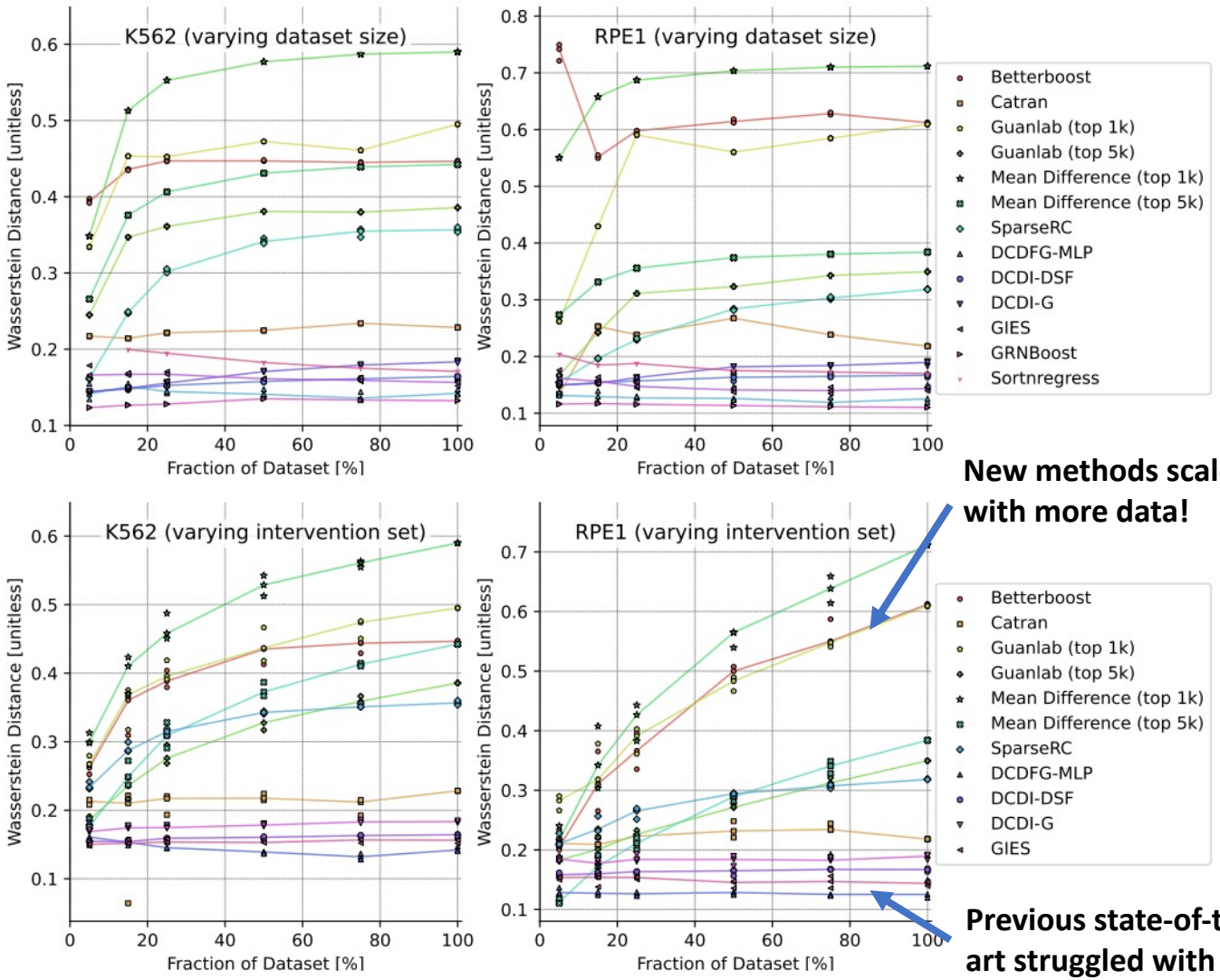


Besides DCDI-DSF, no method is able to translate additional perturbational data into better performance.

Our benchmarking results highlight the need for new methods that better integrate observational and interventional data and the non-representative nature of simulation studies that algorithms are typically tested on.

Lyle et al. DiscoBAX - Discovery of optimal intervention sets in genomic experiment design https://openreview.net/forum?id=mBkUeW8rpD6

In search of causal discovery that scales with more perturbation data



New methods scale with more data!

Previous state-of-the-art struggled with using perturbations.

New causal discovery methods developed in the CausalBench challenge improved our ability to correctly identify causal interactions between genes from perturbation data.

GBP 6000

Achille Nazaret and Justin Hong
Columbia University

GBP 3000

Kaiwen Deng and Yuangfang Guan
University of Michigan

GBP 1500

Panagiotis Misiakos, Chris Wendler, Markus Püschel
ETH Zurich

Chevalley et al. CausalBench: A large-scale Benchmark for Network Inference from Single-Cell Perturbation Data. arXiv 2022

3 Distill robust causal relationships from data

Causal discovery

Successful methods rely on significant distribution shifts for downstream interactors..

e.g. BetterBoost (Nazaret & Hong, 2023)

eps-interventional faithfulness

We here introduce a new concept that we name *interventional faithfulness*. Akin to the common assumption of faithfulness of a joint distribution with respect to the true graph, which assumes that all d-separations in the graph are reflected by conditional independences in the entailed distribution, ϵ-interventional faithfulness assumes that all paths in the graphs are revealed by changes in distribution under intervention above a significance threshold ϵ. This assumption is also similar to how Mooij et al. (2016) define causality as the existence of interventions inducing a change in distribution. We formalize our assumption as follows:

Definition 1. Given the distributions $P_X^{\mathcal{C},(\emptyset)}$ and $P_X^{\mathcal{C},do(X_k:=\tilde{N}_k)}, \forall k \in \mathcal{I}$, we say that the tuple $(\tilde{N}, \mathcal{C})$ is *ϵ-interventionally faithful* to the graph $\mathcal{G}$ associated to $\mathcal{C}$ if for all $i \neq j, i \in \mathcal{I}, j \in V$, $D\left(P_{X_j}^{\mathcal{C},(\emptyset)}, P_{X_j}^{\mathcal{C},do(X_i:=\tilde{N}_i)}\right) > \epsilon$ if and only if there is a directed path from i to j in $\mathcal{G}$.

Successful methods in CausalBench leveraged comparison between interventional distribution vs unperturbed distribution – why does this work?

an optimal score for causal order

Given an observational distribution $P_X^{\mathcal{C},(\emptyset)}$ and a set of interventional distributions $\mathcal{P}_{int} = \{P_X^{\mathcal{C},do(X_k:=\tilde{N}_k)}, k \in \mathcal{I}\}$, $\mathcal{I} \subseteq V$, we define the following score for a permutation π, for some statistical distance $D : \mathcal{P}(M) \times \mathcal{P}(M) \to [0, \infty)$, $\epsilon > 0$:

$$\mathcal{S}(\pi, \epsilon, D, \mathcal{I}, P_X^{\mathcal{C},(\emptyset)}, \mathcal{P}_{int}) = \sum_{\pi(i)<\pi(j), i\in\mathcal{I}, j\in V} \left(D\left(P_{X_j}^{\mathcal{C},(\emptyset)}, P_{X_j}^{\mathcal{C},do(X_i:=\tilde{N}_i)}\right) - \epsilon\right) \quad (1)$$

In fully intervened case, we are guaranteed to recover correct causal order.

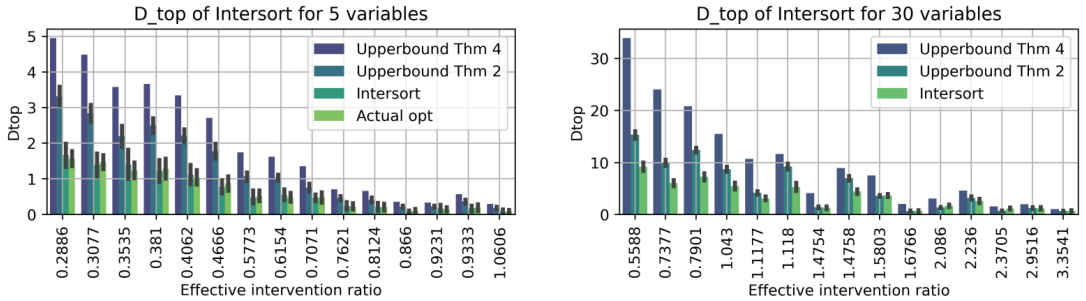
N .. Set of interventional variables
C .. Structural causal model (SCM)
P .. Distribution (interventional or obs)
D_top .. Top order divergence (b/w graph & order)

Chevalley et al. Deriving Causal Order from Single-Variable Interventions: Guarantees & Algorithma. arXiv 2024

3 Distill robust causal relationships from data

Causal discovery

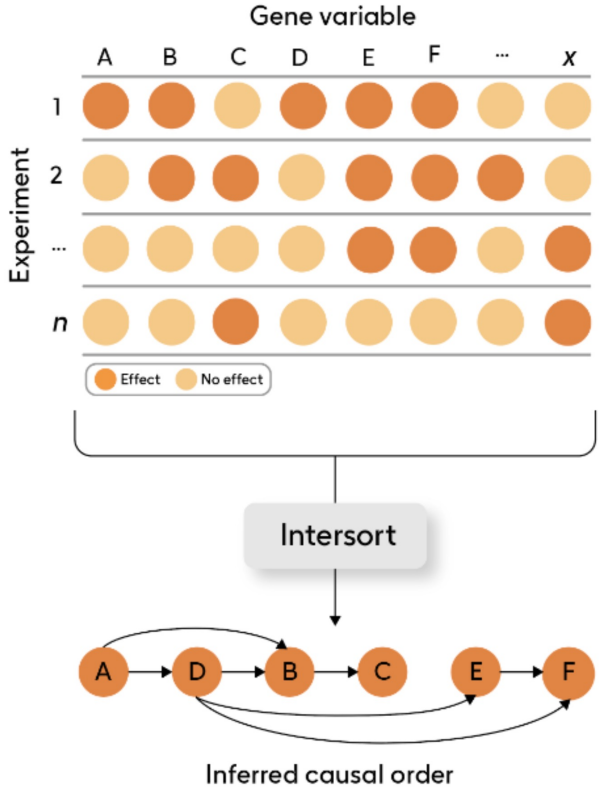
Intersort: an approximate algorithm for deriving causal order from perturbation data



(a) Simulation with 5 variables

(b) Simulation with 30 variables

Figure 1: Simulation and comparison between the two bounds, and between Intersort and the exact π_{opt}. For each setting, we draw 20 graphs per setting, where a setting is the tuple (p_{int}, p_e). Then, for each graph, we run the algorithm on 10 configurations, where each configuration corresponds to a draw of the targeted variables following p_{int}. For both 5 and 30 variables, we have $p_{int} \in \{0.25, 0.33, 0.5, 0.66, 0.75\}$. In the 5 variables setting on the left, we have $p_e \in \{0.5, 0.66, 0.75\}$. In the 30 variables settings on the right, we have $p_e \in \{0.05, 0.1, 0.2\}$. The settings are ordered on the x-axis following what we call the effective intervention ratio $\frac{p_{int}}{\sqrt{p_e}}$. We observe that the error is approximately monotonic when ordered by the effective intervention ratio.



We find an approximate algorithm using eps-interventional faithfulness consistently (& more than existing methods) benefits from more perturbations with guarantees.

Chevalley et al. Deriving Causal Order from Single-Variable Interventions: Guarantees & Algorithma. arXiv 2024

An AI-driven stack for biological discovery

How can we accelerate biological discovery with modern machine learning tools?

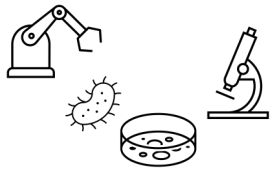
Experimental Design

2 Run experiments that yield more

Optimal experimental design

Generate Hypotheses
"Gene X contributes to disease Y"

Run Experiments
"Test interventions on Gene X in system Z"



1 Run faster experimental cycles

In-silico experimentation

4 Make more of experiments

Multimodal learning

Causal Understanding

Synthesize Data

Experimental Data

3 Distill robust causal relationships from data

Causal discovery

Bonus section: Making more of experimental data

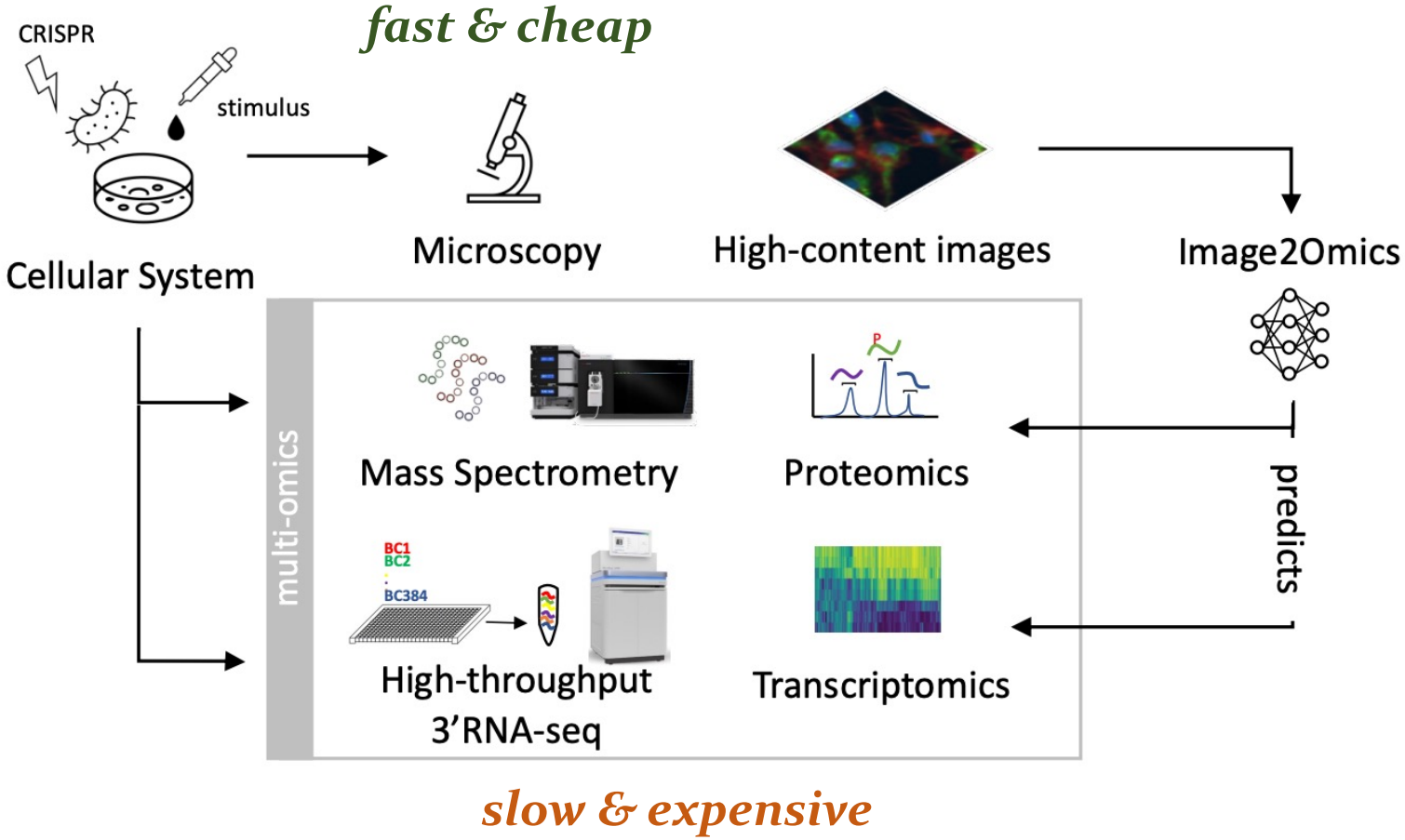
Multi-omics Prediction from High-content Cellular Imaging with Deep Learning

Rahil Mehrizi[1,a], Arash Mehrjou[1,a], Maryana Alegro[1], Yi Zhao[1], Benedetta Carbone[1], Carl Fishwick[1], Johanna Vappiani[1], Jing Bi[1], Siobhan Sanford[1], Hakan Keles[1], Erin Edwards[1], Guillaume Heger[1], Mathias Kalxdorf[1], Marcus Bantscheff[1], Cuong Nguyen[1], and Patrick Schwab[1,*]

[1] *GSK plc, United Kingdom*
[a] *Joint first authors*
[*] *Corresponding author*

Mehrizi & Mehrjou et al. Multi-omics Prediction from High-content Cellular Imaging with Deep Learning. arXiv 2023

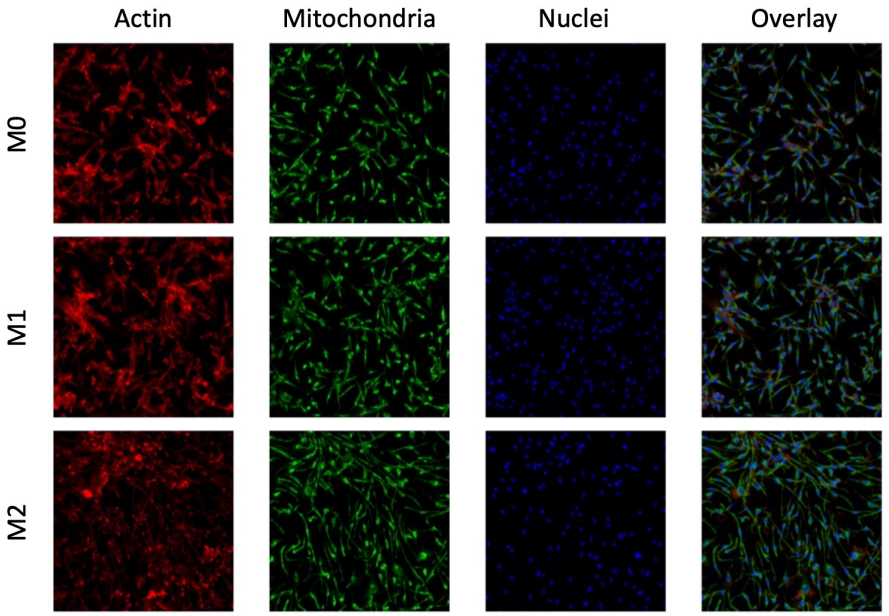
The challenge: Connect high-throughput experiments across biological modalities

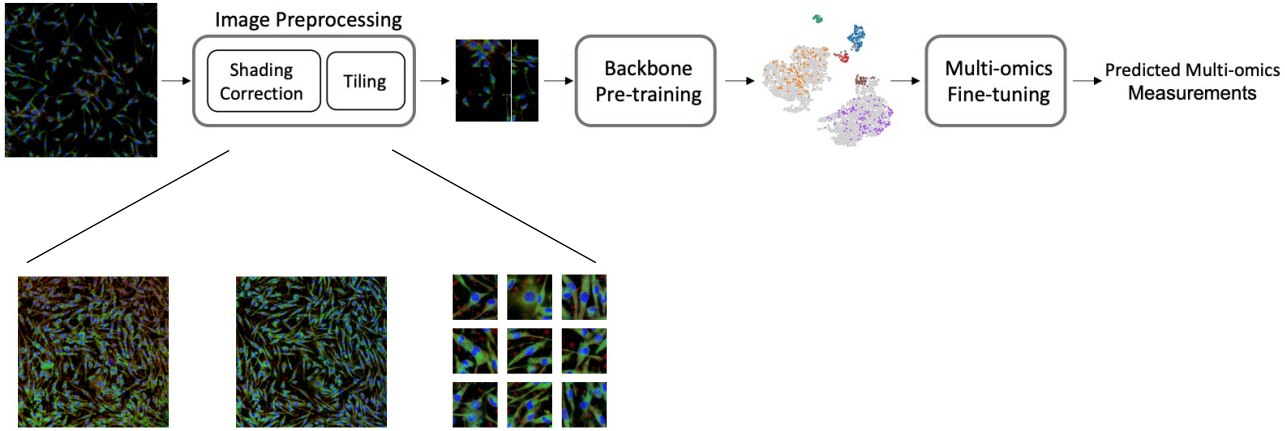


Make more of experimental data

Mehrizi & Mehrjou et al. Multi-omics Prediction from High-content Cellular Imaging with Deep Learning. arXiv 2023

Image2Omics predicts omics-measurements from high content imaging



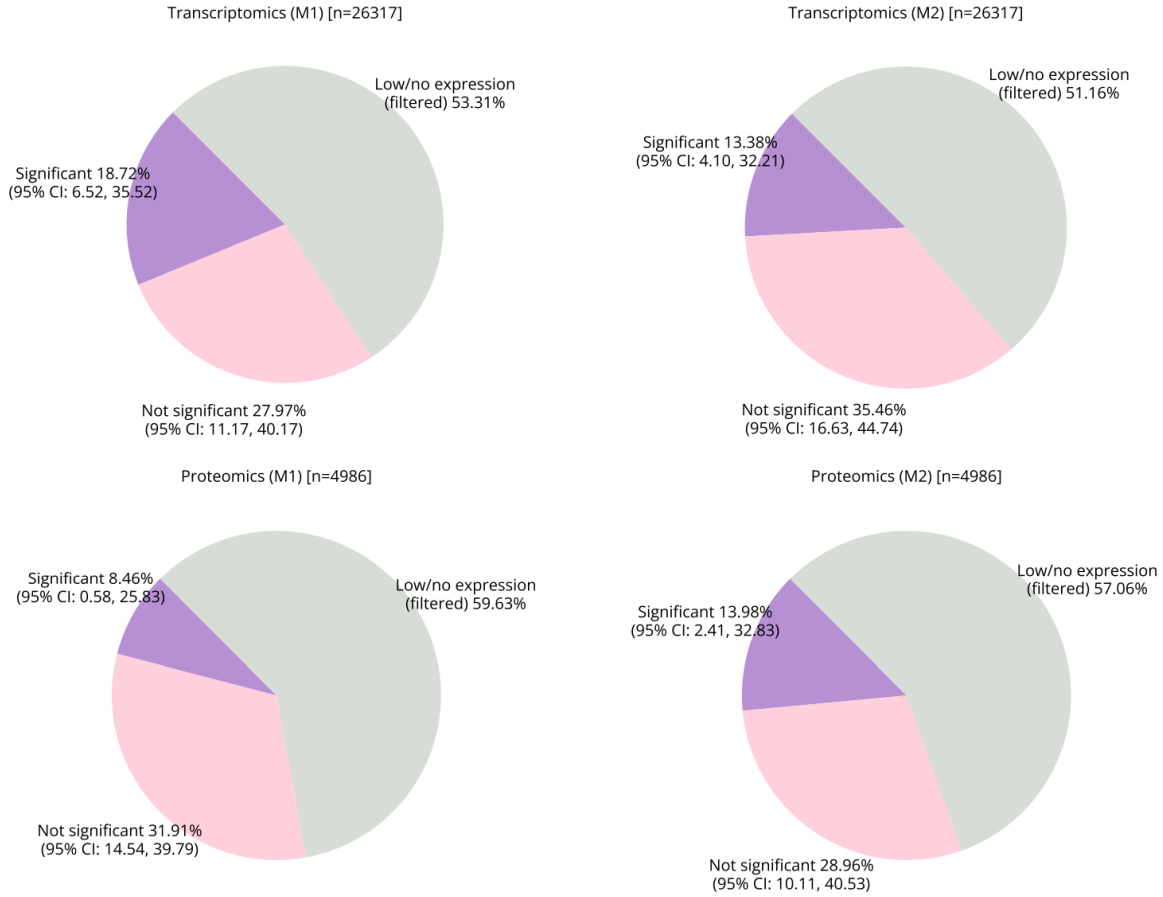


Make more of experimental data

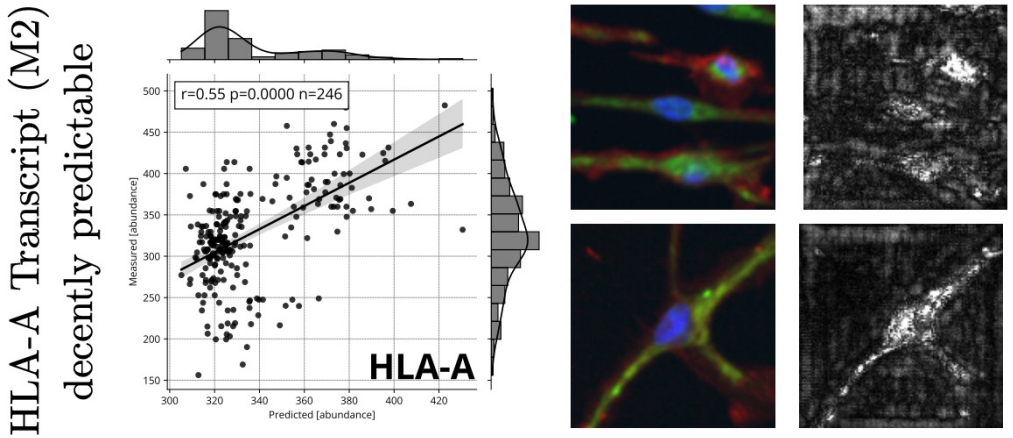
Mehrizi & Mehrjou et al. Multi-omics Prediction from High-content Cellular Imaging with Deep Learning. arXiv 2023

Multimodal learning

8 to 19% of transcriptome / proteome are significantlys predictable from imaging

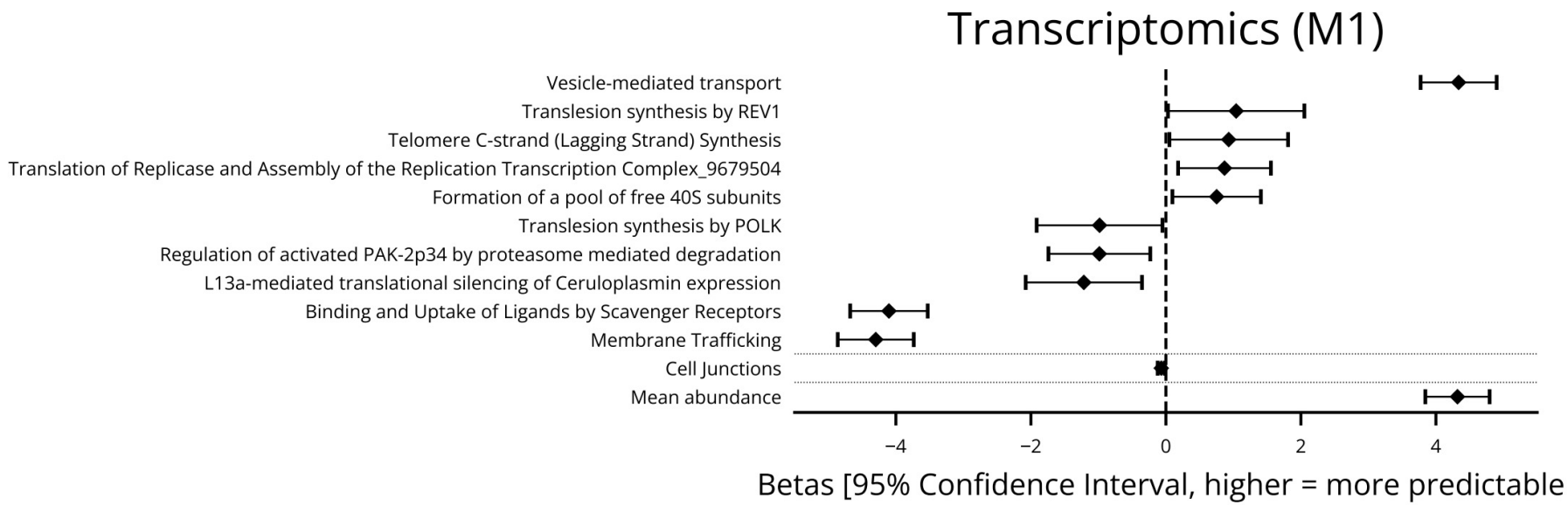


Example (HLA-A Tx)



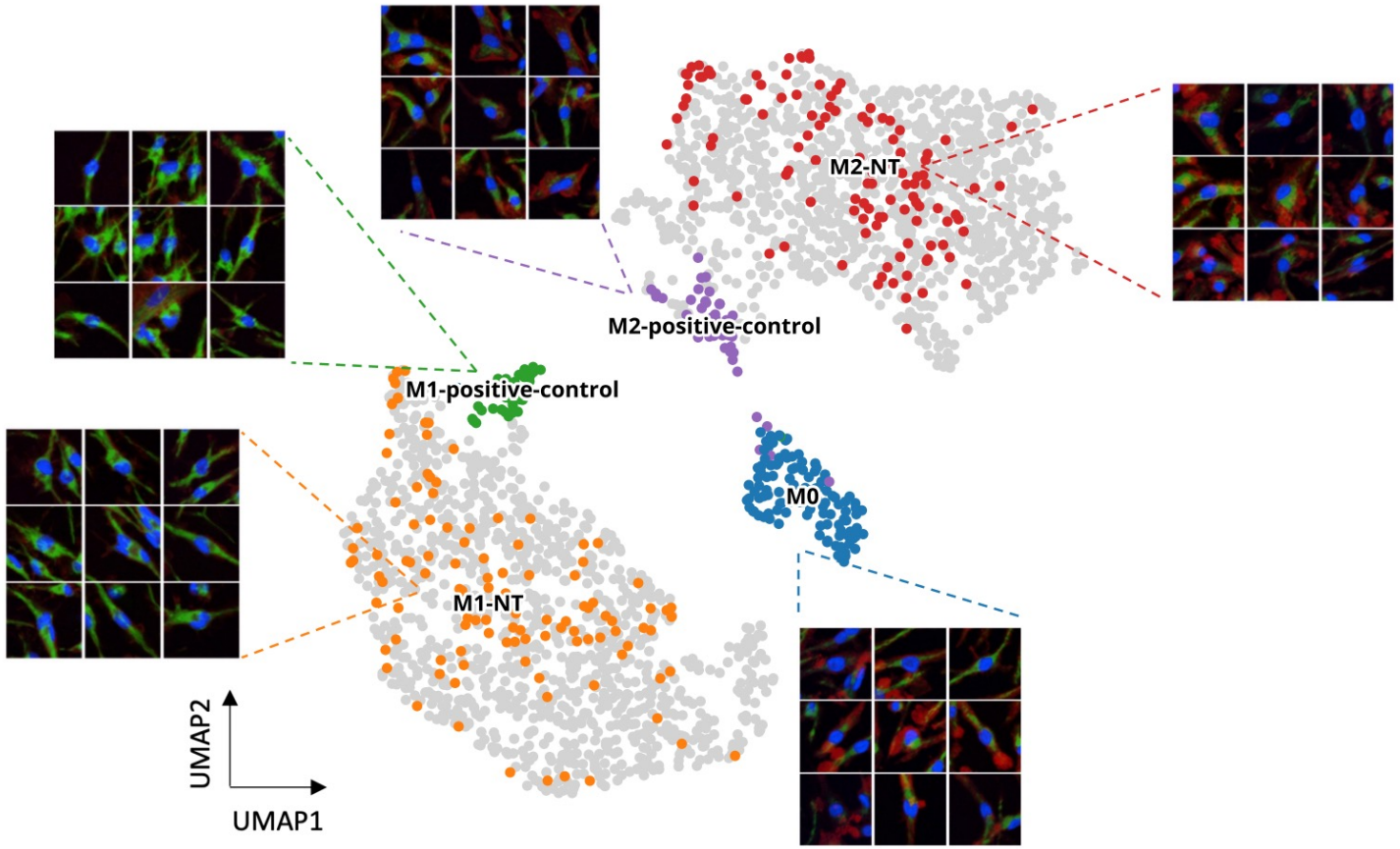
Mehrizi & Mehrjou et al. Multi-omics Prediction from High-content Cellular Imaging with Deep Learning. arXiv 2023

More abundant products and certain pathways are more highly predictable



Mehrizi & Mehrjou et al. Multi-omics Prediction from High-content Cellular Imaging with Deep Learning. arXiv 2023

Image2omics learns map of cellular states to predict omics markers



Mehrizi & Mehrjou et al. Multi-omics Prediction from High-content Cellular Imaging with Deep Learning. arXiv 2023

An AI-driven stack for biological discovery

How can we accelerate biological discovery with modern machine learning tools?

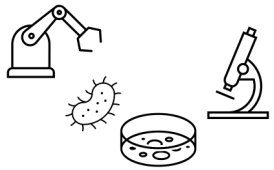
Experimental Design

2 Run experiments that yield more

Optimal experimental design

Generate Hypotheses
"Gene X contributes to disease Y"

Run Experiments
"Test interventions on Gene X in system Z"



1 Run faster experimental cycles

In-silico experimentation

4 Learn across experiments

Multimodal learning

Causal Understanding

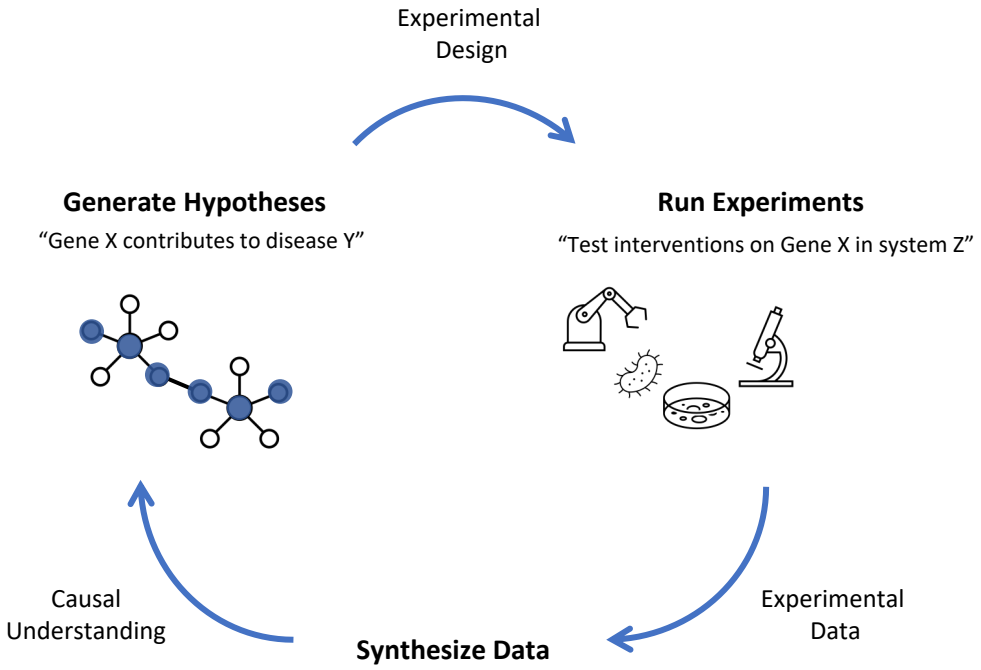
Synthesize Data

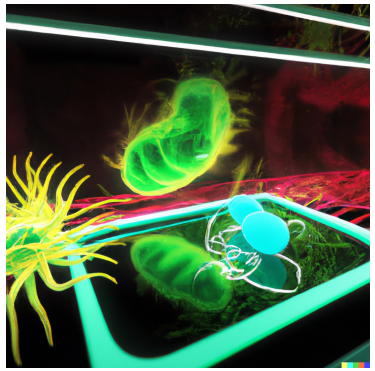
Experimental Data

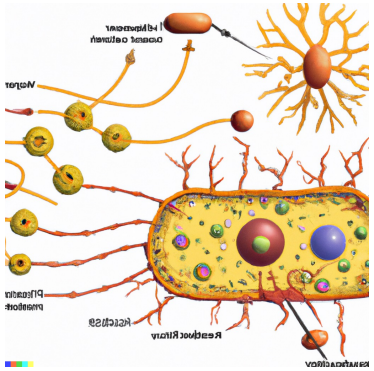
3 Distill robust causal relationships from data

Causal discovery

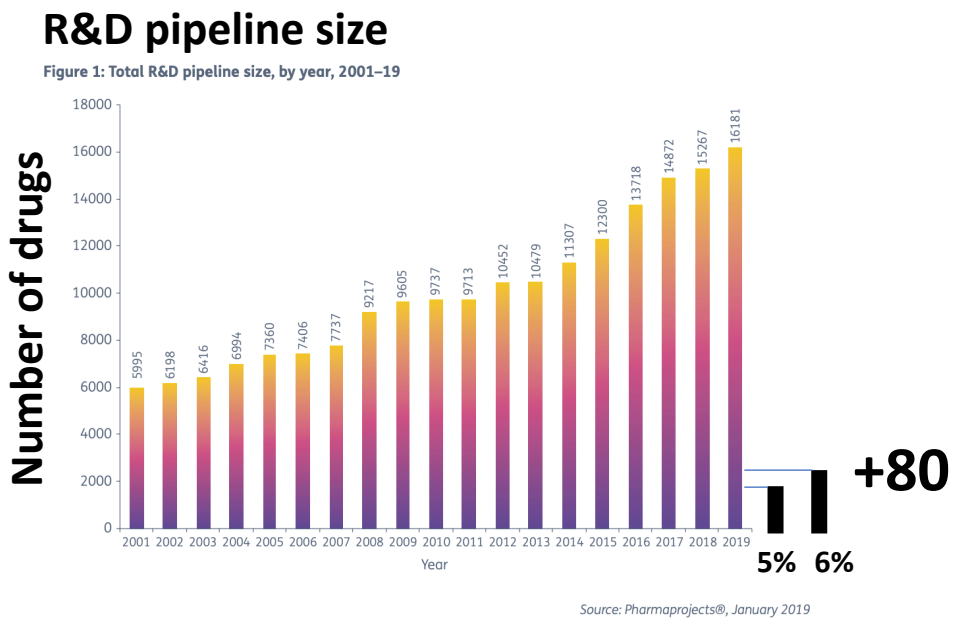
In-silico biological discovery complements wetlab discovery and has the potential to accelerate therapeutic discovery







Summary and conclusion



An increase of the probability of success by just one percent would lead to the discovery of 80 potentially life-changing medicines for patients – an increase of almost 20%.

- 🧬 + 💻 = 🪄
- **Promising early results and new challenges** in applying AI assistants to simulate experimental outcomes completely in-silico.
- AI assistants have the potential to considerably **accelerate biological discovery and medicines development**.
- Advancing the probability of success of drug discovery is expected to accelerate development of **new innovative medicines for patients**.

1 Lloyd, Ian, Alexandra Shimmings, and Pink Sheet Scrip. "Pharma R&D annual review 2019." London: Pharma Intelligence (2019).

Thank you! Questions more than welcome.

GSK

- Lyle et al. DiscoBAX: Discovery of optimal intervention sets in genomic experiment design, ICML (2023)
- Mehrjou et al. GeneDisco: A Benchmark for Experimental Design in Drug Discovery. ICLR (2022)
- Scherrer et al. Learning Neural Causal Models with Active Interventions. arXiv (2021)
- Chevalley et al. CausalBench: A Large-scale Benchmark for Network Inference from Single-cell Perturbation Data. arXiv (2022)
- Chevalley et al. Deriving Causal Order from Single-variable Interventions: Guarantees & Algorithm. arXiv (2024)
- Mehrizi et al. Multi-omics prediction from high-content cellular imaging with deep learning. arXiv (2023)
- Schwab et al. Learning counterfactual representations for estimating individual dose-response curves. AAAI (2020)

Patrick Schwab
patrick.x.schwab@gsk.com
@schwabpa

Interested to get active?
We are looking for
AI/ML, Software Engineers, Internships
see https://gsk.ai

OATML(Yarin Gal)

UNIVERSITY OF OXFORD

CIFAR

Stefan Bauer

Mila

B. Schölkopf

Max Planck Institute for Intelligent Systems

UNIVERSITY OF CAMBRIDGE

the team