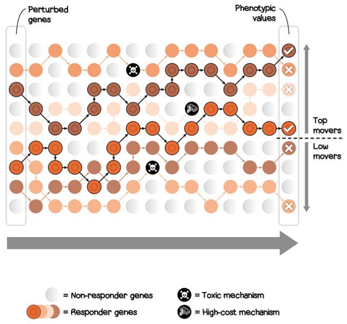
DiscoBAX: Discovery of Optimal Intervention Sets in Genomic Experiment Design

Clare Lyle[1], Arash Mehrjou[2], Pascal Notin[3], Andrew Jesson[3], Stefan Bauer[4, 5], Yarin Gal[3], Patrick Schwab[2]

[1]University of Oxford (now at DeepMind), [2]GSK plc, [3]University of Oxford, [4]Helmholtz AI, [5]TUM

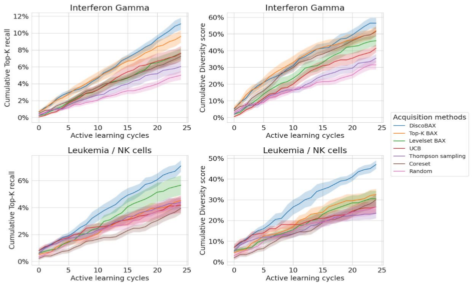
Introduction

We want to identify a set of candidate mechanisms via experiments (leftmost phase) which maximizes the odds of at least one intervention on a mechanism in this set succeeding in successive phases of therapeutic validation (moving right).

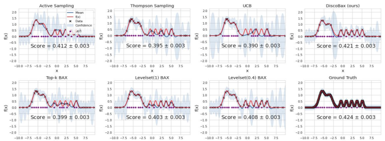


Perturbing different genes affects different cellular mechanisms. Each mechanism consists of various steps (circles). Lower right are not moved enough by the perturbations. Among the top right circles, those that include a toxic step in their mechanism or other cost are not desired. DiscoBAX spreads its budget over diverse mechanisms to increase the chance of finding a mechanism which both moves the phenotype and avoids toxic side effects.

Experiments and results



Top-K recall and Diversity score vs acquisition cycles (x-axis). The two top plots are for the Interferon γ assay[1], and the two bottom plots are based on the Leukemia assay [2].



Existing methods struggle to accurately capture both the high- and low-valued local optima. DiscoBAX finds a decent trade-off between value seeking and mode coverage.

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

Performance comparison on GeneDisco CRISPR assays. We report the aggregated performance of DiscoBAX and other methods on all assays from the GeneDisco benchmark.

Takeaways

- Active learning aims to learn the underlying function as accurate as possible.
- In drug discovery, the goal is often not to fully know the underlying mechanism, but to know **which mechanisms are safe and effective**.
- We showed that this goal can be compactly formulated and approached by DiscoBAX – inspired by Bayesian Algorithm Execution.
- We demonstrate the empirical success of this approach on real-world data from the GeneDisco dataset.

DiscoBAX

$$\max_{S \subseteq \mathcal{X}} \mathbb{E}_\eta \left[\max_{\mathbf{x} \in S} f_{\text{out}} \left(\mathbf{x}; \eta \right) \right]$$

- η : Captures the randomness caused by unknown toxic or costly mechanisms.
- f_{out} : The end-to-end molecular mechanism from a perturbed target to the measured phenotype.
- $\mathcal{S}$: The chosen set of targets (genes)
- $\mathcal{X}$: The set of all available genes to perturb.

Example (Bernoulli) noise model:

$$f_{\text{out}}(x;\eta) = f_{\text{ip}}(x)\eta(x) \mid \eta(x) \in \{0,1\}$$

Where f_{ip} denotes the intermediate phenotype value we can measure in our gene knockout experiments.

Inner loop runs a greedy algorithm to maximize the objective assuming a known intermediate phenotype value for all gene knockouts (sampled from model's posterior)

Outer loop uses Bayesian Algorithm Execution to select knockout experiments which maximize the information gain about the output of the inner loop given a posterior belief over as-yet-unseen intermediate phenotype values.

References

[1] Schmidt et al. Crispr activation and interference screens in primary human t cells decode cytokine regulation

[2] Zhuang et al.. Genome-wide crispr screen reveals cancer cell resistance to nk cells induced by nk-derived ifn-γ. Frontiers in Immunology, 10

[3] Mehrjou et al. GeneDisco: A Benchmark for Experimental Design in Drug Discovery. ICLR 2022

ICML | 2023