Obesity Machine Learning Competition: Tackling Metabolic Diseases

Organized by:
Eric and Wendy Schmidt Center at the Broad Institute,
Diabetes Initiative at the Broad Institute of MIT and Harvard,
…

October 2025

Introduction

Adipocytes are specialized cells that play a key role in regulating body metabolism, which is the process of converting food into energy. When nutrients are abundant, adipocytes synthesize and store fatty acids. In energy-demanding conditions, such as fasting, exercise, or cold exposure, they release fatty acids into the bloodstream so that other tissues can use them as an energy source. Most adipose tissue in the body consists of white adipose tissue, which is largely responsible for storing and mobilizing fatty acids according to the body's needs. We also have smaller amounts of brown adipose tissue (making up around 1-2% of body fat), which burns fats and sugar to generate heat. This process, called thermogenesis, is largely mediated by a protein only found in brown adipocytes called Uncoupling Protein 1, or UCP1, along with other enzymes linked to futile energy cycles. While white and brown adipose tissue arise from distinct progenitor cells (aka preadipocytes), under conditions of prolonged cold exposure or sympathetic stimulation, some white adipose tissue may begin to display thermogenic features through a process known as browning.

When there is a chronic imbalance between energy intake and expenditure, the excessive storage of fatty acids in white adipocytes manifests as obesity, a major global health challenge, affecting over 890 million people worldwide. Obesity increases the risk of type 2 diabetes, cardiovascular diseases, and certain cancers. Recently, powerful drugs have been developed which reduce the drive to eat, thus helping to restore energy balance. Unfortunately, these drugs do not work for everyone, and in some cases cause unacceptable side effects. It would be useful, therefore, to develop therapies for obesity that affect the energy expenditure side of the equation, by activating thermogenic adipocytes or shifting the balance between energy-storing and energy-dissipating cell states.

Moreover, deficient adipose tissue formation can also lead to health problems. In conditions like lipodystrophy, there is insufficient adipose development, causing lipids to accumulate in unusual places, such as the liver (liver steatosis), contributing to systemic insulin resistance and other metabolic disorders. Thus, there is a need to understand the pathways regulating normal adipose tissue mass and function.

Recent advances in single-cell transcriptomics have provided insights into the heterogeneity of adipocyte states (PMID: 40360756) revealing subpopulations of preadipocytes, contributing to variable differentiation capacity, and mature adipocytes, contributing to variable lipogenic and thermogenic functions, which may be further modified in the context of metabolic diseases. However, the transcriptional regulators – specialized proteins that regulate gene expression, acting alone or in combination – driving this cell state diversity remain largely unresolved. To study the pathways that drive adipocytes to different functional states, we have adopted an *in vitro* model in which we harvest progenitor cells from human adipose tissue and then attempt to induce them to differentiate into different types of adipocytes in a dish. Thanks to this *in vitro* model, we can investigate factors that affect adipocyte differentiation. By depleting target genes via CRISPR-Cas9 perturbation experiments on progenitor cells exposed to adipocyte differentiation media, our goal is to explore:

- Which transcription factors, or TFs, enhance or prevent differentiation into adipocytes?
- Which genes, or combinations of genes, could potentially convert white adipocytes into brown or thermogenic cells?

Testing all possible TF genes and gene combinations is experimentally unfeasible. Here is where you come in.

Overview of the Challenge

As part of our ongoing Cell Perturbation Prediction Challenge project, this year's goal is to develop algorithms to:

1) Predict the effect of held-out single-gene perturbations.
2) Predict the effects of held-out combinatorial (two genes) perturbations.
3) Propose combinatorial perturbations expected to significantly move cell state from baseline toward a white or brown-like adipocyte state, which we will then test experimentally.

Crunch 1: Predicting the effect of held-out single-gene perturbations.

In this Crunch, we will explore how well we can predict the single-cell transcriptomic response to single-gene perturbations that were not measured and provided in the training dataset.

Dataset

The dataset contains perturbations targeting 157 genes, of which 150 are transcription factors (TFs). For each perturbation, we provide single-cell gene expression (RNA-seq) profiles measured at day 14 of adipocyte differentiation, annotated with gene perturbation identity, quality control (QC) metrics, and cell metadata. The training dataset contains a subset of these perturbations, while a distinct set of single-gene perturbations is held out for validation and test.

- The dataset is provided in AnnData format (`.h5ad`) as `obesity_challenge_1.h5ad`.
- Normalized gene expression values are stored in `.X`. Raw counts were normalized to a target sum of 100,000 per cell, followed by a $\log_2(1+x)$ transformation (standard single-cell RNA-seq normalization; see lecture 2 of the crash course).
- Raw gene expression counts prior to normalization are stored in `.layers['counts']` for reproducibility and alternative preprocessing.
- The perturbation target gene information is provided in `.obs['gene']`, with values corresponding to either "NC" for control cells or to the target gene name if the cell is perturbed. Control cells receive a perturbation that has no effect on the cell's RNA-Seq profile.
- Cell state/program enrichment information is provided in `.obs`, with columns "`pre_adipo`", "`adipo`", "`lipo`", and "`other`" indicating whether each cell was enriched for pre-adipocyte, adipocyte, or lipogenic programs. "Other" was defined as cells that were not enriched for either pre-adipocyte or adipocyte programs. Program enrichment assignments were based on expert-curated canonical signature genes, and the list of signature genes is provided in `signature_genes.csv`. The full analysis workflow used to determine program enrichment is provided in the accompanying notebook, `program_analysis`, which can be consulted for additional methodological details. We provide the cell state proportion for each of the perturbations in a separate file `program_proportion.csv`.
- During preprocessing, standard single-cell quality control (QC) was applied to remove low-quality cells and cell doublets based on sequencing library complexity, gene detection rate, and mitochondrial gene content. The dataset was then restricted to cells with a single confident guide assignment to a perturbation, and guides represented by fewer than 10 cells were excluded. Genes detected in fewer than 10 cells were removed, and known signature genes from `signature_genes.csv` were subsequently re-introduced.
- Top 1000 highly variable genes: In our evaluation, we only compute the evaluation metrics over the highly variable genes (wherever applicable).

Participant output

Participants must submit three outputs:

- `prediction.h5ad` — an AnnData file containing predicted gene expression profiles post-perturbation for 2,863 gene perturbations indicated in `predict_perturbations.txt`, across 10,238 genes listed in `predict_genes.txt`. Predictions should be stored in `adata.X` matrix with the corresponding perturbation identity recorded in `adata.obs['gene']`. For each gene perturbation, we ask you to predict the gene expression profiles for 100 cells to quantify the distribution of each perturbation prediction. The file with predictions should have dimensions [286,300 × 10,238] (cells × genes).
- `predict_program_proportion.csv` — a CSV file reporting the predicted proportion of cells with enriched programs for each gene perturbation listed in `predict_perturbations.txt`. The file should contain one row per perturbation with the following columns: "gene", "pre_adipo", "adipo", "lipo", "other" and "lipo_adipo". The "gene" column should contain the perturbation name, and the "pre_adipo", "adipo", "lipo", and "other" columns should specify the predicted proportion of cells in each corresponding state for that perturbation. The "lipo_adipo" column is defined as the ratio of "lipo" to "adipo" (representing the proportion of adipocytes with enriched lipogenic programs). This file should thus have 2,863 rows and 6 columns.
- `Method description` — a document outlining the approaches used to generate both the predictions and the estimated proportions of cells enriched for each program. This should include sufficient details of the computational models employed and the procedures used to derive cell proportions. The document should be organized into three sections: (1) Method Description; (2) Rationale; and (3) Data and Resources Used. We ask you to please write at least five sentences for each section.

Evaluation

First, we define some terms that will be used for evaluation:

1. **N:** Sample size in the dataset.
2. **P:** set of all perturbation targets.
3. **Mean Expression for perturbation** "p" (X_p): The mean gene expression profile of all the cells with the given perturbation $p \in P$.
4. **Perturbed Mean** (X_{PM}): Mean gene expression of all the perturbed cells (those receiving a single gene perturbation) in the training dataset.

Model performance will be assessed using the following metrics.

Transcriptome-wide metrics. These metrics will be computed independently for each perturbation, restricting inputs to the top 1000 highly variable genes (HVG). The HVG genes are a subset of all genes measured in the training dataset.

- Pearson Delta (ρ) between predicted ($\hat{X}$) and observed (X) perturbation effects relative to perturbed mean (X_p). Formally, $\rho(X, \hat{X}) = \frac{1}{|P|} \sum_{p \in P} cor(\hat{X}_p - X_{PM}, X_p - X_{PM})$ where $|P|$ is the size of the set P, "$cor(a, b)$" is the correlation between vectors a and b.
- Maximum mean discrepancy (MMD) between predicted and observed distributions of single-cell profiles. Let X_p^i be the true expression profile and $\hat{X}_p^i$ be the predicted expression profile for cell "i" with perturbation "p" and. Then we calculate the MMD distance for a particular perturbation using the following formulae:

$$MMD^2(X_p, \hat{X}_p) = \frac{1}{N^2}\sum_{i,j} k(X_p^i, X_p^j) + \frac{1}{N^2}\sum_{k,l} k(\hat{X}_p^k, \hat{X}_p^l) - \frac{2}{N^2}\sum_{m,n} k(X_p^m, \hat{X}_p^n)$$

where $k(a, b)$ is the Gaussian kernel with the bandwidth from the following list: [581.5, 1163.0, 2326.0, 4652.0, 9304.0]. Then we average the MMD score across all the perturbations using the following formulae: $MMD^2(X, \hat{X}) = \frac{1}{|P|} \sum_{p \in P} MMD^2(X_p, \hat{X}_p)$ where $|P|$ is the size of the set P.

Program-level metrics. These metrics evaluate whether models capture meaningful biological outcomes. Two sets of metrics will be used:

- There are four cell state proportions for each perturbation, i.e, pre-adipogenic, adipogenic, lipogenic, and other. For each perturbation "p", we have the ground truth cell-state proportion $S_p = [preadipo, adipo, lipo, other]$. Let S_p^R be the vector that has the proportion of the cell states $[preadipo, adipo, other]$. Let S_p^L denote the condition probability of a cell being in the lipogenic state given the adipogenic state, i.e, $S_p^L = lipo/adipo$. Then we define the program level loss as

  Formally, $L1(\hat{S}, S) = \frac{1}{|P|}\sum_{p \in P} 0.75 * |\hat{S}_p^R - S_p^R|_1 + 0.25 * |\hat{S}_p^L - S_p^L|$ where $|.|_1$ is the L$_1$-distance, |P| is the number of perturbations, and $\hat{S}_p$ is the predicted cell-state proportions.

  Participants are welcome (but not required) to use the provided `program_analysis.ipynb` script to derive the proportion of cells enriched for each program. However, they are also free to apply their own methods if they prefer.

Winners: For challenge 1, we will have 10 winners – one for each of the metrics described above.

Baselines: For comparisons, we will include two baseline models. The first is a *linear model*, which predicts average gene expression profiles given perturbation identity (PMID: 40759747). The second is a *perturbation mean predictor*, which assigns to each unseen perturbation the mean profile of perturbations in the training set as a baseline (PMID: 40854979).

Validation and Test

The held-out genes will be split into a validation set and a test set. 20% of the data will be left out for validation and testing. Participants will have multiple opportunities to submit predictions for the validation set, with the average loss on this set displayed on the public leaderboard. The public leaderboard will be refreshed on a regular weekly schedule to reflect recent submissions. Only one final submission will be allowed for the test set, and three final rankings will be determined by the average losses on the test set.

External resources

The application of external resources (e.g., outside transcriptional datasets, gene ontology, pre-trained embeddings, etc.) is allowed; however, all external resources must be published or in the public domain and properly credited.

Dataset hosting and access

All the datasets are hosted on CrunchDAO. Please see the challenge forum for instructions on how to access the data. A starter Jupyter notebook demonstrating how to load and work with the data is available in the forum.

Crunch 2: Predicting the effect of held-out double-gene perturbations.

In this Crunch, we will explore how well we can predict the single-cell transcriptomic response to double-gene perturbations that were not provided in the training dataset.

Dataset

This dataset contains single-gene and pairwise perturbations targeting a curated set of 18 genes. This includes a set of 153 heterotypic perturbations (i.e., Gene A+Gene B); 18 homotypic perturbations (i.e., Gene A+Gene A); and 18 monogenic perturbations (i.e., Gene A+NC). Each cell receives either 1 guide RNA (resulting in a single genetic perturbation, similar to Crunch 1) or two guide RNAs (leading to heterotypic, homotypic, or monogenic perturbations). Importantly, the number of guides received by a cell has an effect on its underlying transcriptomic distribution: two non-targeting guides (NC+NC) may have a different effect from a single non-targeting guide (NC); similarly, Gene A+NC may have a different effect from Gene A alone. Although the dataset includes some cells that received three guides, the evaluation in this Crunch focuses exclusively on cells that received a single guide or two guides. For each cell, we provide its single-cell gene expression (RNA-seq) profile measured at day 14 of adipocyte differentiation, annotated with the identity of the perturbed genes, quality control (QC) metrics, and cell metadata. The training dataset contains a subset of these perturbations, while a distinct set of perturbations is held out for validation and test.

- The dataset is provided in AnnData format (`.h5ad`) as `obesity_challenge_2.h5ad`.
- Normalized gene expression values are stored in `.X`. Raw counts were normalized to a target sum of 100,000 per cell, followed by a $\log_2(1+x)$ transformation (standard single-cell RNA-seq normalization; see lecture 2 of the crash course).
- Raw gene expression counts prior to normalization are stored in `.layers['counts']` for reproducibility and alternative preprocessing.
- We provide data for cells that received single-guide, double-guide or three-guide perturbation. The perturbation target information for each cell is provided in `.obs['gene']`. Control cells, which receive perturbations with minimal transcriptomic effect, are labeled as “NC” (single-guide) or “NC+NC” (double-guide). For perturbed cells, single-guide perturbations are indicated simply by the target gene name (e.g., 'Gene A'). All double-guide perturbations are denoted using a '+' format, which includes heterotypic gene pairs ('Gene A+Gene B'), homotypic pairs ('Gene A+Gene A'), and monogenic double-guide perturbations ('Gene A+NC'). Similarly, three-guide perturbations extend this format by linking three targets separated by a '+' (e.g., 'Gene A+Gene B+Gene C').
- Cell state/program enrichment information is provided in `.obs`, with columns “`pre_adipo`”, “`adipo`”, “`lipo`”, and “`other`” indicating whether each cell was enriched for pre-adipocyte, adipocyte, or lipogenic programs. “Other” was defined as

cells that were not enriched for either pre-adipocyte or adipocyte programs. Program enrichment assignments were based on expert-curated canonical signature genes, and the list of signature genes is provided in `signature_genes.csv`. The full analysis workflow used to determine program enrichment is provided in the accompanying notebook, `program_analysis`, which can be consulted for additional methodological details. We provide the cell state proportion for each of the perturbations in a separate file `program_proportion.csv`.

- During preprocessing, standard single-cell quality control (QC) was applied to remove low-quality cells and cell doublets based on sequencing library complexity, gene detection rate, and mitochondrial gene content.
- Top 500 highly variable genes: In our evaluation, we only compute the evaluation metrics over the highly variable genes (wherever applicable).

Participant output

Participants must submit three outputs:

- `prediction.h5ad` — an AnnData file containing predicted gene expression profiles post-perturbation for 62 gene perturbations indicated in `predict_perturbations_2.txt`, across 36,601 genes listed in `predict_genes_2.txt`. Predictions should be stored in `adata.X` matrix with the corresponding perturbation identity recorded in `adata.obs['gene']`. For each gene perturbation, we ask you to predict the gene expression profiles for 100 cells to quantify the distribution of each perturbation prediction. The file with predictions should have dimensions [6,200 × 36,601] (cells × genes).
- `predict_program_proportion.csv` — a CSV file reporting the predicted proportion of cells with enriched programs for each gene perturbation listed in `predict_perturbations.txt`. The file should contain one row per perturbation with the following columns: "gene", "pre_adipo", "adipo", "lipo", "other" and "lipo_adipo". The "gene" column should contain the perturbation name, and the "pre_adipo", "adipo", "lipo", and "other" columns should specify the predicted proportion of cells in each corresponding state for that perturbation. The "lipo_adipo" column is defined as the ratio of "lipo" to "adipo" (representing the proportion of adipocytes with enriched lipogenic programs). This file should thus have 62 rows and 6 columns.
- `Method description` — a document outlining the approaches used to generate both the predictions and the estimated proportions of cells enriched for each program. This should include sufficient details of the computational models employed and the procedures used to derive cell proportions. The document should be organized into three sections: (1) Method Description; (2) Rationale; and (3) Data and Resources Used. We ask you to please write at least five sentences for each section.

Evaluation

First, we define some terms that will be used for evaluation:

5. **N:** Sample size in the dataset.
6. **P:** set of all perturbation targets.
7. **Mean Expression for perturbation** "p" (X_p): The mean gene expression profile of all the cells with the given perturbation $p \in P$.
8. **Single Perturbed Mean** (X_{PM1}): Mean gene expression of all the single-perturbed cells (those receiving a single gene perturbation except for "NC" perturbation target) in the training dataset.
9. **Double Perturbed Mean** (X_{PM2}): Mean gene expression of all the double-perturbed cells (those receiving a double gene perturbation, except for "NC+NC" perturbation target) in the training dataset; this in particular includes double-perturbations of the form Gene A+NC.

Model performance will be assessed using the following metrics.

Transcriptome-wide metrics. These metrics will be computed independently for each perturbation, restricting inputs to the top 500 highly variable genes (HVG). The HVG genes are a subset of all genes measured in the training dataset.

- Pearson Delta (ρ) between predicted ($\hat{X}$) and observed (X) perturbation effects relative to perturbed mean (X_p). Formally, $\rho(X, \hat{X}) = \frac{1}{|P|} \sum_{p \in P} cor(\hat{X}_p - X(p)_{PM}, X_p - X(p)_{PM})$ where $X(p)_{PM}$=X_{PM1} if p is a single gene perturbation and $X(p)_{PM}$=X_{PM2} if p is a double gene perturbation, $|P|$ is the size of the set P, "$cor(a, b)$" is the correlation between vectors a and b.
- Maximum mean discrepancy (MMD) between predicted and observed distributions of single-cell profiles. Let X_p^i be the true expression profile and $\hat{X}_p^i$ be the predicted expression profile for cell "i" with perturbation "p" and. Then we calculate the MMD distance for a particular perturbation using the following formulae:

$$MMD^2(X_p, \hat{X}_p) = \frac{1}{N^2} \sum_{i,j} k(X_p^i, X_p^j) + \frac{1}{N^2} \sum_{k,l} k(\hat{X}_p^k, \hat{X}_p^l) - \frac{2}{N^2} \sum_{m,n} k(X_p^m, \hat{X}_p^n)$$

where $k(a, b)$ is the Gaussian kernel with bandwidth from the following list: [212.25, 424.5, 849.0, 1698.0, 3396.0]. Then we average the MMD score across all perturbations using the following formula: $MMD^2(X, \hat{X}) = \frac{1}{|P|} \sum_{p \in P} MMD^2(X_p, \hat{X}_p)$, where $|P|$ is the size of the set P.

Program-level metrics. These metrics evaluate whether models capture meaningful biological outcomes. Two sets of metrics will be used:

- There are four cell state proportions for each perturbation, i.e, pre-adipogenic, adipogenic, lipogenic, and other. For each perturbation “p”, we have the ground truth cell-state proportion $S_p = [preadipo,\ adipo,\ lipo,\ other]$. Let S_p^R be the vector that has the proportion of the cell states $[preadipo,\ adipo,\ other]$. Let S_p^L denote the condition probability of a cell being in the lipogenic state given the adipogenic state, i.e, $S_p^L = lipo/adipo$. Then we define the program level loss as

  $L1(\hat{S}, S) = \frac{1}{|P|}\sum_{p\in P} 0.75 * |\hat{S}_p^R - S_p^R|_1 + 0.25 * |\hat{S}_p^L - S_p^L|$, where $|.|_1$ is the L_1-distance, |P| is the number of perturbations, and $\hat{S}_p$ is the predicted cell-state proportions.

  Participants are welcome (but not required) to use the provided `program_analysis.ipynb` script to derive the proportion of cells enriched for each program. However, they are also free to apply their own method if they prefer.

Winners: For Crunch 2, we will have 10 winners – one for each of the metrics described above.

Baselines: For comparison, we will include the following baseline models.

1. **Single Perturbation Target:**
   a. The first approach is a *linear model* designed to predict average gene expression profiles. This is achieved by fitting a linear model that maps the perturbation target embedding to the resulting gene expression state after perturbation. To derive gene embeddings for all targets, Principal Component Analysis (PCA) is applied to the transpose of the cell-by-gene matrix. We then learn a linear model using the target embedding from the training set to predict the gene-expression. This methodology facilitates the prediction of gene expression for novel, unseen perturbations, as it only requires swapping the gene embedding of targets from the train set to the new perturbation target (PMID: 40759747).
   b. The second is a *perturbation mean predictor*, which assigns to each unseen perturbation the mean profile of all the single perturbations in the training set as a baseline (PMID: 40854979). These baselines are the same as the ones provided in Crunch 1.
2. **Double Perturbation Target:**
   a. Similar to the single-perturbation baseline, we include a perturbation-mean predictor that assigns each unseen perturbation the mean profile of all double perturbations in the training set (PMID: 40854979).

b. An additional baseline, designated as SALT, utilizes the following algorithm, inspired by the paper "Season combinatorial intervention predictions with Salt & Peper" (arxiv: 2404.16907), to compute the effect of a double-perturbation (Gene1+Gene2):
   i. delta1 = X_{G1}- X_{PM1}
   ii. delta2 = X_{G2}- X_{PM1}
   iii. Predicted Expression = X_{PM2}+ delta1 + delta 2

Validation and Test

The held-out perturbations will be split into a validation set and a test set. 30% of the data will be held out for validation and testing. Participants will have multiple opportunities to submit predictions for the validation set, with the average loss on this set displayed on the public leaderboard. The public leaderboard will be refreshed on a regular weekly schedule to reflect recent submissions. Only one final submission will be allowed for the test set to compute the final evaluation. The final evaluation will be based on three distinct metrics, resulting in three separate final rankings (one dedicated leaderboard per metric).

External resources

The use of external resources (e.g., outside transcriptional datasets, gene ontology, pre-trained embeddings, etc.) is allowed; however, all external resources must be published or in the public domain and properly credited.

Dataset hosting and access

All the datasets are hosted on CrunchDAO. Please see the Crunch forum for instructions on how to access the data. A starter Jupyter notebook demonstrating how to load and work with the data is available in the forum.

References

Below are a few references meant to provide more background and some of the approaches researchers are applying in fields relevant to these Crunches. This is not meant to be an exhaustive list and many important works are not listed here. We may provide additional references in response to your questions over the coming months. **Reading these articles is not necessary to complete the Crunches, but we believe these can be a helpful resource.**

Single cell transcriptomic adipocytes datasets

- Dissecting the impact of transcription factor dose on cell reprogramming heterogeneity using scTF-seq
- A single-cell atlas of human and mouse white adipose tissue
- Human subcutaneous and visceral adipocyte atlases uncover classical and nonclassical adipocytes and depot-specific patterns
- Adipose tissue retains an epigenetic memory of obesity after weight loss.
- Unveiling adipose populations linked to metabolic health in obesity
- Spatial mapping reveals human adipocyte subpopulations with distinct sensitivities to insulin
- snRNA-seq reveals a subpopulation of adipocytes that regulates thermogenesis
- Wnt signaling preserves progenitor cell multipotency during adipose tissue development
- Adipogenic and SWAT cells separate from a common progenitor in human brown and white adipose depots
- Single-Nucleus Analysis of Human White Adipose Tissue Reveals Adipocyte Subsets with Distinct Metabolic Profiles
- Mapping the transcriptional landscape of human white and brown adipogenesis using single-nuclei RNA-seq

CRISPR/Cas9 genetic perturbation screens

- Mapping information-rich genotype-phenotype landscapes with genome-scale Perturb-seq
- Exploring genetic interaction manifolds constructed from rich single-cell phenotypes
- X-Atlas/Orion: Genome-wide Perturb-seq Datasets via a Scalable Fix-Cryopreserve Platform for Training Dose-Dependent Biological Foundation Models

Modeling perturbations and causality

- Machine learning for perturbational single-cell omics
- Elements of Causal Inference: Foundations and Learning Algorithms
- Causal Structure and Representation Learning with Biomedical Applications
- scPerturb: Information Resource for Harmonized Single-Cell Perturbation Data
- GeneDisco: A Benchmark for Experimental Design in Drug Discovery

- Systema: a framework for evaluating genetic perturbation response prediction beyond systematic variation
- Deep-learning-based gene perturbation effect prediction does not yet outperform simple linear baselines
- MORPH Predicts the Single-Cell Outcome of Genetic Perturbations Across Conditions and Data Modalities
- Learning Genetic Perturbation Effects with Variational Causal Inference
- Squidiff: predicting cellular development and responses to perturbations using a diffusion model
- Predicting cellular responses to perturbation across diverse contexts with State
- TxPert: Leveraging Biochemical Relationships for Out-of-Distribution Transcriptomic Perturbation Prediction
- GEARS: Predicting transcriptional outcomes of novel multi-gene perturbations
- Learning Causal Representations of Single Cells via Sparse Mechanism Shift Modeling
- Predicting cellular responses to complex perturbations in high-throughput screens
- PerturbNet predicts single-cell responses to unseen chemical and genetic perturbations
- Predicting Cellular Responses to Novel Drug Perturbations at a Single-Cell Resolution
- Active Learning for Optimal Intervention Design in Causal Models
- Control of cell state transitions