HYPED: A MULTIMODAL HYBRID PERTURBATION GENE EXPRESSION AND IMAGING DATASET

Anonymous authors
Paper under double-blind review

ABSTRACT

Integrating multimodal, high-resolution biological data is a useful way to characterize biological processes, such as how cells respond to perturbations. Cell perturbation prediction is a major experimental challenge and has motivated substantial research in machine learning for biology. In this work, we generated a multimodal benchmark dataset that captures the dynamic response of human fibroblasts to transient transcription factor perturbations. We performed time-series live cell imaging with fluorescent cell cycle reporters over 72 hours and collected long-read single-cell RNA sequencing data from the same population of cells. We release the processed dataset, preprocessing pipelines and benchmarking code along with the evaluation of existing models using our data as ground truth. This work supports the development and evaluation of machine learning methods for modeling dynamical systems from multimodal datasets. HYPED makes the cell perturbation problem accessible to machine learning researchers with state-of-the-art experimental data.

1 INTRODUCTION

Direct cellular reprogramming, the process of converting one cell type into another without passing through an intermediate pluripotent state, holds promise for personalized and regenerative medicine (65; 53). Reprogramming requires the introduction of targeted perturbations capable of altering cell state, most often through the introduction of cell-type specific transcription factors (TFs) (23; 61). Current methods for delivering TFs, such as lentiviral vectors or CRISPR-based systems, permanently alter the cell's DNA. This irreversible modification raises concerns about uncharacterized and potentially harmful effects, limiting their utility for a wide range of research and clinical applications. To address this, recent advancements in transient RNA-based methods for TF delivery have emerged as safer alternatives to lentiviral vector or CRISPR-based methods (57; 24).

Despite experimental progress made over the past decade, direct reprogramming efforts face key challenges: conversion efficiency remains low, and identifying new TF combinations is difficult due to limited knowledge of their combinatorial effects (32; 61). Evaluating the effects of perturbations is essential for optimizing existing cellular reprogramming protocols. This requires capturing molecular and phenotypic dynamics with high temporal and cellular resolution. While several works have captured output measurements from reprogramming experiments, most existing datasets are limited to a single data modality, and few datasets exist for the evaluation of transient TF delivery methods (17; 49; 70; 21; 14; 5). The limitation of data compatible with current experimental perturbation techniques poses a challenge for evaluating and benchmarking machine learning models for cell perturbation prediction. In this work, we collected high-resolution, long-read single-cell transcriptomic data and live-cell fluorescent microscopy data from a state-of-the-art TF reprogramming protocol on human fibroblasts using a combination of forced overexpression (*MYOD1*) and gene suppression (*PRRX1*). Our dataset is a valuable resource for the machine-learning evaluation of cellular perturbations.

Biological Background. The cellular reprogramming problem originated in the 1980s with H. Weintraub's use of the TF MyoD to convert fibroblasts into skeletal muscle (66). In the early 2000s, S. Yamanaka reprogrammed the first induced pluripotent stem cells (iPSC) (53). Both approaches perform cell perturbations with viruses that are inserted into the host cell's DNA permanently changing the cell. Many contemporary perturbation screens use CRISPR, through methods like

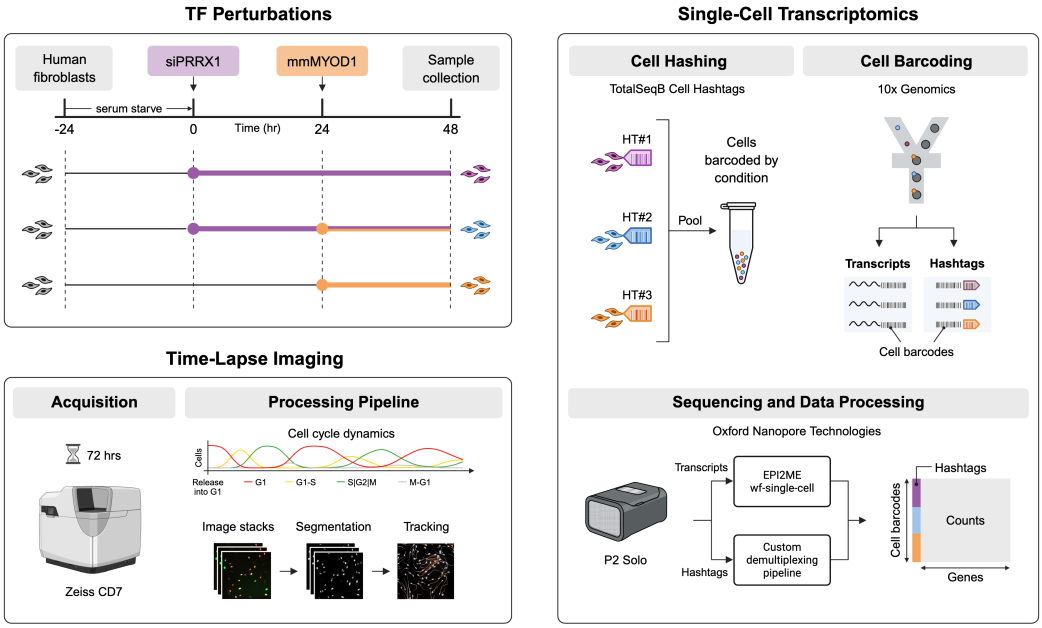

Figure 1: **Experimental Dataset.** The HYPED dataset includes time-series imaging and long-read sequencing to characterize transient cellular perturbations. **Top left:** Human fibroblasts are perturbed using combinations of *MYOD1* overexpression and *PRRX1* attenuation, delivered via modified mRNA and siRNA, respectively. **Bottom left:** Live-cell fluorescent imaging with the Zeiss Celldiscoverer 7 (CD7) captures the temporal dynamics of the cell cycle under TF perturbation. **Right:** Long-read single-cell sequencing captures the final transcriptional state of the perturbed cells.

Perturb-seq, to upregulate or downregulate gene expression, also irreversibly modifying the host cell genome (11). These methods have produced many datasets for perturbation prediction using machine learning; however, the permanent modification of a host cell's DNA is not suitable for many clinical, research, and biomanufacturing applications. The dynamics of the cell cycle, a central factor in cell reprogramming, are not accounted for in the choice to irreversibly modify a cell's DNA and hence are not considered by machine learning perturbation models (40).

To address this challenge from an experimental standpoint, non-integrative approaches using modified mRNA (mmRNA) and small interfering RNA (siRNA) have become favored for direct reprogramming (62; 60; 56). mmRNA enables transient, high-level expression of reprogramming TFs without genomic integration, while siRNA facilitates the attenuation of endogenous mRNAs that maintain the original cell identity. Together, these approaches promote efficient and controlled lineage conversion without permanent genetic alteration (50; 35). To our knowledge, we present the first cellular perturbation dataset based on state-of-the-art experimental methods for machine learning.

Data Modalities. HYPED is a multimodal dataset that combines high-throughput sequencing with live-cell imaging (fig. 1). Recent advancements in long-read omics technologies, such as single-cell RNA sequencing (scRNA-seq), have expanded our ability to investigate complex intracellular mechanisms across multiple regulatory levels (27; 58; 26). In particular, scRNA-seq with Oxford Nanopore Technologies enables the direct sequencing of full-length RNA transcripts, providing insights into transcriptional regulation and cell state dynamics. The long-read capability of Nanopore sequencing allows for the identification of alternative splicing events, quantification of isoform diversity, and detection of RNA modifications, which are often inaccessible features through short-read platforms (25; 27). This enables a more accurate and comprehensive characterization of the transcriptome, facilitating the study of gene expression heterogeneity at the single-cell level.

RNA sequencing remains one of the most widely used data modalities to assess cellular responses to perturbation (43; 55; 45). However, sequencing requires cell lysis, resulting in destruction of

Table 1: **Overview of Related Single Cell Datasets.** Overview of benchmark datasets used in machine learning for cellular imaging, sequencing, and perturbation.

Dataset		Imaging	Sequencing	Perturbation
ML Task				
Only Imaging				
LIVECell	(12)	✓		
NeurIPS Competition 2022	(34)	✓		
Only Sequencing				
CellXGene	(41)		✓	
NeurIPS Competition 2021	(33)		✓	
Multi-Modal Perturbation (Perturbation + Sequencing)				
scPerturb	(19)		✓	✓
PerturbBase	(64)		✓	✓
sc-pert	(28)		✓	✓
NeurIPS Competition 2023	(8)		✓	✓
Multi-Modal (Imaging + Perturbation + Sequencing)				
HYPED (ours)		✓	✓	✓

the cell during sample preparation. In contrast, advances in live-cell imaging technologies have significantly improved spatial resolution and throughput, offering a non destructive, time-resolved method for studying dynamic cellular processes. These systems enable high-resolution monitoring of features like cell morphology, proliferation activity, migration, and apoptosis over time. When combined with fluorescent reporters such as the Fluorescent Ubiquitination-based Cell Cycle Indicator (FUCCI) system (18), live-cell imaging allows real-time tracking of specific protein expression to explore intracellular drivers of phenotypic changes (48). FUCCI distinguishes cell cycle phases by leveraging fluorescently-tagged proteins whose degradation is regulated by the cell cycle. This approach provides a dynamic view of cell cycle progression in live cells. In the HYPED dataset, time-series imaging offers high temporal resolution to capture perturbation dynamics, complementing sequencing-based approaches. This addition introduces a data modality for machine learning studies of cellular perturbations, enabling multimodal and temporal analysis of the perturbation prediction problem (see table 1).

Machine Learning Approaches. Given the significance of the cellular reprogramming problem, a wide range of machine learning approaches have been developed. Early efforts include graph- and network-based methods (42; 13), as well as studies that frame reprogramming through system identification, trajectory optimization, and control theory (43; 7). More recent work leverages transformers and other modern architectures to represent and embed cell states measured with sequencing (55; 39; 10). Although these approaches have achieved successes at the machine learning problem, there remain substantial gaps in the success of experimental perturbations, which motivates the development and application of new models to state-of-the-art experimental datasets.

Contributions. Imaging and transcriptomics each offer distinct yet complementary insights into cellular biology; imaging captures spatial and morphological dynamics, while transcriptomics reveals the molecular and regulatory state of the cell. However, these modalities are typically collected in isolation, providing only fragmented views of genetic perturbations. Integrating these experimental modalities promises a more comprehensive understanding of the cell as a dynamical system, revealing how gene expression and physical behavior co-evolve over time. This integration is technically challenging, as it involves harmonizing data with vastly different structures, resolutions, and temporal characteristics. Emerging machine learning approaches, particularly multimodal representation learning and generative modeling, offer a powerful bridge, enabling alignment, translation, and predictive modeling across these disparate data types. Leveraging these computational frameworks is key to unlocking unified models of cellular behavior that incorporate both molecular and phenotypic dimensions.

To address this challenge, we designed and performed a perturbation experiment that captures live-cell imaging and scRNA-seq of the same cells, providing a state-of-the-art ground truth dataset for a variety of challenges, including improving perturbation prediction outcomes. We overcome several limitations of existing datasets, which often lack synchronized, time-resolved measurements across modalities. Our contributions provide not only a valuable resource for benchmarking multimodal models but also a framework for studying dynamic cellular processes through an integrative lens.

2 Related Work

2.1 Cellular Reprogramming

Direct cellular reprogramming is a complex process that involves activating lineage-specific target networks while simultaneously silencing the original cell identity (37). Most reprogramming strategies have relied on viral vectors or plasmids to deliver lineage-specifying genes. Pioneer factors, such as *MYOD* in muscle (66) or *OCT4* in pluripotency (53), can initiate reprogramming but often require cooperative TFs to fully activate and maintain the target gene regulatory network (20). Without the proper TFs, most cells fail to convert fully, instead stalling in unstable intermediate states (9; 38). Researchers have mapped regulatory networks that govern cell identity and have developed libraries of transcription factors capable of inducing cell transitions (47; 69; 67; 29).

In parallel, RNA-based approaches have gained traction as non-integrative tools for cellular reprogramming (63; 62; 60). Synthetic mmRNA delivery allows for transient and tunable expression of reprogramming factors, minimizing risk of genomic integration and long-term mutagenesis (50; 7; 6). Similarly, siRNA can be used to transiently suppress endogenous factors that stabilize native cell identity (56; 31). While RNA-based systems offer less fine-tuned control compared to viral or plasmid-based delivery methods, they have significantly greater clinical promise (24). Their transient nature is well suited to control the cell cycle and reduces the risk of mutagenesis and tumorigenesis, making them attractive for therapeutic applications (4). As RNA delivery technologies and engineering strategies mature, RNA-based reprogramming offers a safe path to clinical translation without compromising on effectiveness (32).

2.2 Datasets and Models for Gene Perturbations

Gene perturbations are a core laboratory technique for cellular reprogramming and have become a focus of machine learning in biology. RNA-seq is the primary technique for measuring cellular responses to perturbations, with scRNA-seq enabling high-resolution analysis of development, disease, and treatment effects. This has led to the development of large single cell atlases (22; 52; 15; 54) and foundation models (55; 16; 51). Together, these datasets and models provide a rich resource for developing machine learning strategies to infer cell identity and predict perturbation outcomes.

Perturb-seq, which combines CRISPR perturbations with scRNA-seq, enables high-throughput measurement of gene function and has led to the creation of several public datasets and resources (11). The Gene Perturbation Atlas (GPA) compiles single-gene perturbation data across diverse cell types to systematically assess how individual genes influence cell identity (68). The Perturbation Cell and Tissue Atlas (PCTA) extends this effort by integrating genetic and environmental perturbations with molecular and imaging readouts to support causal inference in cell and tissue biology (44).

Early computational frameworks for direct reprogramming combine gene expression, regulatory networks, and epigenetic information to predict optimal TF combinations (30; 43; 42; 13). However, the low efficiency of current reprogramming protocols has positioned cell perturbation prediction as a standard machine learning challenge (45; 55; 28; 72). While many models focus on forward prediction—forecasting the cellular response to a given perturbation—few address how to identify the optimal perturbation strategy. Moreover, existing methods often overlook important biological priors such as cellular dynamics, cell cycle phase, and other factors known to influence reprogramming success (71).

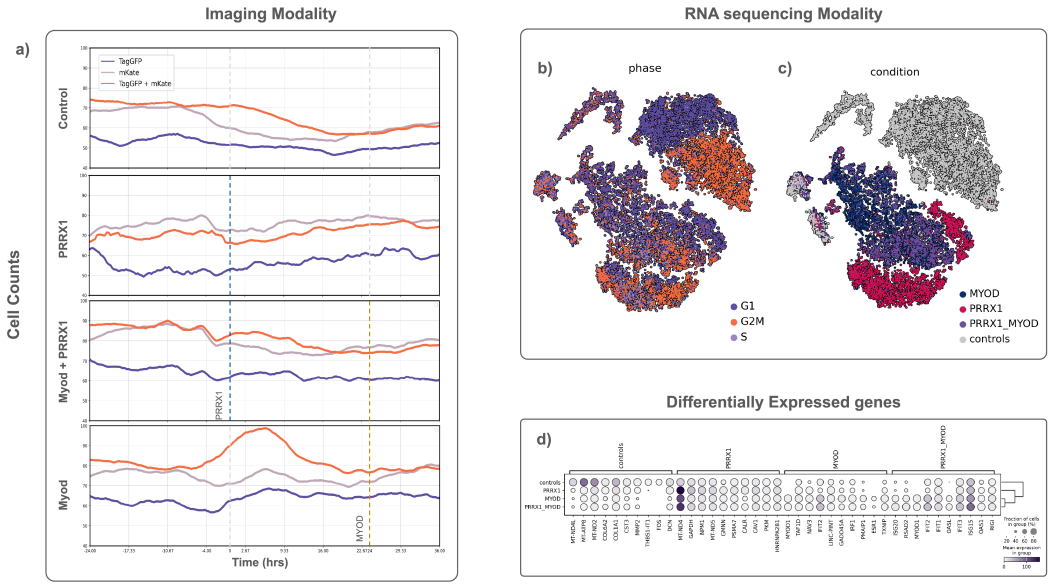

Figure 2: **HYPED Dataset Overview. (a)** Shows the Distribution of cells across different cell cycle phases, quantified from Incucyte imaging data across the different fluorescence channels. The **(b)** t-SNE projection of cells colored by cell cycle phase shows **(c)** t-SNE projection of cells colored by experimental condition. **(d)** Top 10 differentially expressed genes across experimental conditions.

3 Dataset

We present a multimodal dataset that captures the dynamic response of human fibroblasts to transcription factor (TF) perturbations. The dataset includes long-read scRNA-seq and time-lapse imaging data collected under four experimental conditions: *MYOD1* activation via modified mRNA (mmMYOD1), *PRRX1* suppression via siRNA (siPRRX1), sequential dual perturbation (siPRRX1 followed by mmMYOD1), and an unperturbed control. The scRNA-seq dataset comprises approximately 20,000 cells ($\sim$11,000 perturbed, $\sim$9,000 controls), enabling isoform-level transcriptomic profiling. Time-lapse imaging data were collected at 20 minute intervals over 72 hours, capturing high-resolution trajectories of cell cycle dynamics and morphological changes.

3.1 Experimental Setup

Human neonatal foreskin fibroblasts (BJ, ATCC CRL-2522) carrying the Incucyte Cell Cycle Reporter (Sartorius 4779) were used for perturbation experiments. Cells were cultured at 37°C in 5% CO_2 on standard cultureware in full media (FM; Dulbecco's Modified Eagle's Medium (DMEM, Gibco 11965-092) + 10% Fetal Bovine Serum (FBS, Corning 35-015-CV) + 1% Penicillin-Streptomycin (P/S, Gibco 15140-122)). Cells were seeded at 1.03 - 1.33 x 10^4 cells/cm^2 in 6- and 48-well plates, followed by 24 hours of serum starvation (0.2% FBS) for G0/G1 phase synchronization. At t=0h, cells were released from serum starvation, and respective cells were transfected with PRRX1 siRNA (25 nM; Dharmacon) in FM using the TransIT-X2 Dynamic Delivery System (Mirus MIR6000). At t=24h, siRNA was washed out with FM. Respective cells were then transfected with MYOD1 modified mRNA (1 ng/μl; Pseudouridine + Silica purification, Trilink WOTL39876) in FM supplemented with 2 μM 4OH-Tamoxifen (Sigma-Aldrich H7904) using the Lipofectamine MessengerMAX Transfection Reagent (Invitrogen 100026485). At t=48h, cells were harvested with TrypLE Express (Gibco 12604-013) for scRNA-seq sample preparation. For imaging plates, cells were treated the same as described above with two differences: (1) Cells were cultured in imaging media (FluoroBrite DMEM (Gibco A18967-01) + 10% FBS + 1% P/S) instead of standard FM; (2) Cells were stained with 250 nM SiR-DNA (Cytoskeleton, Inc. CY-SC007) at t=-24h to enable tracking of cell nuclei.

Table 2: **Imaging Data Structure.** These dimensions are used to operate on the imaging files, named according to CZI specification.

Symbol	Description
H	Phase
S	Scene (plate/well)
T	Time
C	Channel
Z	Depth dimension
M	Mosaic (determines sub-tile)
Y	Vertical axis
X	Horizontal axis

3.2 Single Cell RNA Sequencing with Long Reads

Cells from siPRRX1, siPRRX1/mmMYOD1, and mmMYOD1 were individually labeled with TotalSeq-B human cell hashing antibodies (BioLegend, Cat# 394631, #394633, and #394635, respectively). 20,000 cells from each condition were mixed together (viability >98%) and single cell barcoded on the 10x Genomics Chromium Controller (X) using the Next GEM Single Cell 3' Kit V4. Barcoded cDNA amplicons from transcript and hashing libraries were prepared according to the Oxford Nanopore Technologies (ONT) 3' cDNA protocol (SQK-LSK114, SST_9198_v114_revJ_13Nov2024). Libraries were assessed for quality following ONT recommendations. Prepared libraries were sequenced on the ONT PromethION Solo 2 (P2) sequencer. Raw reads were base-called with Dorado v0.9.1 (3) using the High Accuracy basecalling model and stored as fastq files.

For unperturbed controls, fibroblasts were sorted by fluorescence activated cell sorting (FACS) into their respective cell cycle phases after staining with 16 μM Hoechst 33342 for 50 minutes. Cells in G1, S, and G2/M were individually labeled with TotalSeq-B hashing antibodies (BioLegend, Cat# 394631, #394603, #394605, respectively), mixed, and single cell barcoded (viability >97%). Transcript and hashing libraries were processed and sequenced as described above.

Raw sequencing files obtained from scRNA-seq of the perturbed cells and controls were processed using the EPI2ME wf-single-cell pipeline v3.1.0 (2) with the human GRCh38 reference genome. The processed gene expression matrices were combined into a single AnnData object containing cells as rows (obs) and genes as columns (var) (59).

Respective perturbation conditions (siPRRX1, siPRRX1/mmMYOD1, mmMYOD1) were assigned to individual cell barcodes using a custom demultiplexing pipeline (1). We add the $assigned\_condition$ column to AnnData.obs.

3.3 Time Series Microscopy

The Zeiss Celldiscoverer 7 (CD7) live-cell imaging system was used to capture time-lapse images over the course of perturbation experiments. Oblique contrast and fluorescence microscopy was performed with a Plan-Apochromat 20x/0.7 objective and 0.5x tube lens. Images were taken using an Axiocam 506 with 14 bit resolution. Cells were imaged at 37°C in 5% CO_2 in imaging media. Images were captured every 20 minutes over 72 hours. Raw CZI files were exported from the Zen Blue 3.0 software for image processing and downstream analyses.

3.4 Data Storage

The HYPED dataset is made available on kaggle along with the software used to process the raw experimental data from the sequencers. Each modalitiy of the dataset is structured separately. The raw CZI files exported from the microscope contain data with dimensions HSTCZMYX (table 2). There are 4 channels that correspond to markers that may be used to delinieate the cell cycle phase:

- Cy5 - cell nuclei
- MKate - G2/M phase
- tagGFP - G1 phase
- Oblique - oblique contrast

Each scene is associated with one of the four experimental conditions. We extract only the relevant scenes across time frames where each frame is split into mosaics of 6×5. These mosaic tiles are stitched together to get the full frame. Each frame was then converted into a four-mode numpy array (T, C, X, Y) with dimensions (202, 4, 2826, 3245). To make the dataset standardized and accessible, tensor data of each frame was embedded into a HDF5 (Hierarchical Data Format). Each HDF5 file consists of a single frame with the four channels at a particular image capture point of the experiment. The dataset is designed and to load frames across any range of time points into compute memory

The long read sequencing gene expression data is stored in the HDF5 AnnData format *.h5ad*, a commonly-used HDF5-based format with extensive support in Python and R. This format stores the measured gene expression in each single cell in a cell by gene matrix, with the rows and columns annotated to correspond to the genes and experimental conditions of the cells.

4 BENCHMARKING

Sequencing. We compare the ground truth results obtained from the experiment and the outputs from predictive models. We use two models to benchmark: the foundation model Geneformer (55) and the perturbation prediction model GEARS (46) and report on Cosine similarity, Mean Squared Error (MSE), Mean Absolute Error (MAE) and Pearson correlation coefficient between the predicted and observed values (see table 3).

Geneformer is a Transformer-based Deep Learning model which utilizes unsupervised self-attention mechanism to generate Transcriptomic representations. This generates embeddings in a lower-dimensional space that can be fine-tuned for a variety of tasks.

We tokenize the transcriptome of each cell in the data set with rank value encoding. For the in-silico perturbation, we set the expression level of *MYOD1* to a value greater than the highest expressed gene, which places it at the top of the ranked value encoding, simulating the activation of *MYOD1* within the cell. Similarly, we simulate suppression of *PRRX1* by setting the expression level to zero before rank value encoding. The tokenized data is then run through Geneformer generating embeddings. This embedded data is compared with those of experimentally perturbed cells, and the metrics are reported.

GEARS is a Deep Learning perturbation prediction model which can predict gene expression vectors under certain perturbation conditions. We first split our experimental data into Test, Train and validation splits and created a GEARS graph dataset. We then performed in silico prediction for all three conditions using GEARS. This generates the predicted gene expression vector for the top 500 genes which we compare with the average expression of our held out conditional cells, and calculate different metrics.

Imaging. To evaluate the performance of noise removal techniques, we performed a benchmarking task on our dataset, first generating regions of interest from the original data by employing image processing techniques and segmenting the cells. These cells are then labeled as foreground, and everything else is considered noise and the Signal-to-Noise Ratio is calculated. We then run the images through a machine learning algorithm and various quantitative measures including PSNR and SSIM were calculated. FM2S (36) is a Deep Neural Network based Fluorescent Microscopy Imaging denoiser. We Run this model on different conditions of our dataset and report the average metrics (see table 4).

5 DISCUSSION

Here, we introduced the HYPED benchmark dataset for cellular reprogramming, designed using state-of-the-art perturbation techniques. The use of transient techniques with modified mmRNA and siRNA provides the first multimodal perturbation dataset suitable for settings that prevent gene editing.

Algorithm 1 Geneformer Validation Experiment

Require: perturbed_group, control_group, perturbation_genes
Ensure: cos_sim

// Initialize lists to store control and perturbed group embeddings
$E_c,\ E_p \leftarrow [\,], [\,]$

// Compute Geneformer embeddings of perturbed data
for each cell x in perturbed_group **do**
  $e \leftarrow \texttt{Geneformer}(x)$
  $E_p \leftarrow [E_p\ e]$
end for

// Compute Geneformer embeddings with in silico perturbations
for each cell x in control_group **do**
  if MYOD $\in$ perturbation_genes **then**
    $x[\text{MYOD}] \leftarrow \texttt{max}(x)$
  end if
  if PRRX1 $\in$ perturbation_genes **then**
    $x[\text{PRRX1}] \leftarrow 0$
  end if
  $e \leftarrow \texttt{Geneformer}(x)$
  $E_c \leftarrow [E_c\ e]$
end for

// Determine the average embedding of each experimental group
$e_p \leftarrow \texttt{mean}(E_p)$
$e_c \leftarrow \texttt{mean}(E_c)$

Return $\texttt{CosineSimilarity}(u, v)$ =0

Table 3: **Benchmarking Perturbation Models Against Experimental Data.** Single-cell control data were computationally perturbed using Geneformer and GEARS, and the resulting profiles were compared to experimentally perturbed counterparts.

Model	Perturbation	Cell Cycle	Cosine Sim.	MSE	MAE	Pearson
Geneformer	+ MYOD1	G1	0.9808	0.0148	0.0902	0.9807
		S	0.9776	0.0172	0.0988	0.9775
		G2/M	0.9715	0.0221	0.1119	0.9714
	− PRRX1	G1	0.9913	0.0068	0.0605	0.9913
		S	0.9900	0.0078	0.0660	0.9900
		G2/M	0.9900	0.0078	0.0680	0.9900
	− PRRX1 + MYOD1	G1	0.9902	0.0076	0.0645	0.9902
		S	0.9887	0.0086	0.0697	0.9887
		G2/M	0.9851	0.0115	0.0810	0.9851
GEARS	− PRRX1	All Phases	0.9518	0.0502	0.1079	0.9358
	+ MYOD1	All Phases	0.9741	0.0129	0.0622	0.9625
	− PRRX1 + MYOD1	All Phases	0.9772	0.0116	0.0598	0.9676

Unlike widely used lentiviral, CRISPR, and Perturb-seq datasets, which though commonplace are not suitable for many biological applications, HYPED provides a biologically relevant alternative better suited for learning and predicting perturbation dynamics based on established biological constraints.

Table 4: **Image Benchmarking.** Image Quality Metrics Averaged Across Timepoints.

Model	MSE	PSNR (dB)	SSIM	Calculated SNR	Denoised SNR
FM2S	0.000146	86.480553	0.877678	14.082471	15.051044

Moreover, HYPED's live cell imaging provides one of the first perturbation datasets to capture cell cycle dynamics with perturbations. It has long been recognized by cellular reprogramming and biological researchers that cell perturbations are transient processes, yet many current models including GEARS, Geneformer, scGPT, and others do not account for or model the dynamics. This dataset provides an improved opportunity to train models that account for perturbation and biological dynamics at a higher temporal resolution and under current experimental conditions than previously possible.

To advance cellular reprogramming and related biological problems, it is essential that machine learning models are aligned with the capabilities and limitations of contemporary experiments by training on modern experimental data modalities. Many existing approaches are trained on idealized or outdated datasets that fail to reflect the transient, dynamic nature of real biological systems. By employing modified mmRNA and siRNA and including live-cell imaging with long-read single-cell transcriptomics, the HYPED dataset serves as a resource for researchers to develop models that are not only more predictive, but also more actionable in laboratory and clinical settings.

6 Limitations

We identified two main limitations in our study. First, the experiment conducted includes the effects of three unique transcription factor perturbations captured on skin fibroblasts. Although this describes the activity of a small subset, the vast combinatorial space of transcription factors and cell types remains largely unexplored. Second, we considered a limited number of tasks for each modality, but there are a variety of challenges that can be explored with our dataset.

7 Ethical Considerations

All resources provided as part of this cell reprogramming study are strictly for research purposes only and should not be used in clinical settings and diagnostic procedures. No sensitive information is included in the dataset. With the aforementioned restrictions, we have not identified any potential adverse impacts from the HYPED dataset.

Acknowledgments

We would like to thank all Rajapakse lab members for helpful and inspiring discussions. This work was supported by the Defense Advanced Research Projects Agency (DARPA) award number [HR00112490472 to I.R.], the Air Force Office of Scientific Research (AFOSR) award number [FA9550-22-1-0215 to I.R.], support from NVIDIA [to I.R.], and support from the National Institute of General Medical Sciences (NIGMS) award number [GM150581 to J.P.].

References

[1] CooperStansbury/ont_10x_feature_barcodes.

[2] epi2me-labs/wf-single-cell. original-date: 2022-08-19T09:19:03Z.

[3] nanoporetech/dorado. original-date: 2022-05-17T23:12:13Z.

[4] Aurelio Balsalobre and Jacques Drouin. Pioneer factors as master regulators of the epigenome and cell fate. *Nature Reviews Molecular Cell Biology*, 23(7):449–464, 2022.

[5] Brent A. Biddy, Wenjun Kong, Kenji Kamimoto, Chuner Guo, Sarah E. Waye, Tao Sun, and Samantha A. Morris. Single-cell mapping of lineage and identity in direct reprogramming. *Nature*, 564(7735):219–224, December 2018. Publisher: Nature Publishing Group.

[6] Anna K Blakney, Paul F McKay, and Robin J Shattock. Delivery of mrna-based vaccines and therapeutics with lipid nanoparticles. *Biochemical Society Transactions*, 47(4):1209–1218, 2019.

[7] Simone Bruno, Thorsten M Schlaeger, and Domitilla Del Vecchio. Epigenetic oct4 regulatory network: stochastic analysis of cellular reprogramming. *npj Systems Biology and Applications*, 10(1):3, 2024.

[8] Daniel Burkhardt, Andrew Benz, Robrecht Cannoodt, Mauricio Cortes, Scott Gigante, Christopher Lance, Richard Lieberman, Malte Luecken, and Angela Pisco. Single-cell perturbation prediction: generalizing experimental interventions to unseen contexts. `https://neurips.cc/virtual/2023/competition/66586`, 2023. NeurIPS 2023 Competition Track.

[9] Patrick Cahan, Hu Li, Samantha A. Morris, Edroaldo Lummertz da Rocha, George Q. Daley, and James J. Collins. CellNet: Network Biology Applied to Stem Cell Engineering. *Cell*, 158(4):903–915, August 2014.

[10] Haotian Cui, Chloe Wang, Hassaan Maan, Kuan Pang, Fengning Luo, Nan Duan, and Bo Wang. scgpt: toward building a foundation model for single-cell multi-omics using generative ai. *Nature Methods*, 21(8):1470–1480, 2024.

[11] Atray Dixit, Oren Parnas, Biyu Li, Jenny Chen, Charles P Fulco, Livnat Jerby-Arnon, Nemanja D Marjanovic, Danielle Dionne, Tyler Burks, Raktima Raychowdhury, et al. Perturb-seq: dissecting molecular circuits with scalable single-cell rna profiling of pooled genetic screens. *cell*, 167(7):1853–1866, 2016.

[12] Christoffer Edlund, Timothy R Jackson, Nabeel Khalid, Nicola Bevan, Timothy Dale, Andreas Dengel, Sheraz Ahmed, Johan Trygg, and Rickard Sjögren. Livecell—a large-scale dataset for label-free live cell segmentation. *Nature methods*, 18(9):1038–1045, 2021.

[13] Ryohei Eguchi, Momoko Hamano, Michio Iwata, Toru Nakamura, Shinya Oki, and Yoshihiro Yamanishi. Transdire: data-driven direct reprogramming by a pioneer factor-guided trans-omics approach. *Bioinformatics*, 38(10):2839–2846, 2022.

[14] Mirko Francesconi, Bruno Di Stefano, Clara Berenguer, Luisa de Andrés-Aguayo, Marcos Plana-Carmona, Maria Mendez-Lago, Amy Guillaumet-Adkins, Gustavo Rodriguez-Esteban, Marta Gut, Ivo G Gut, Holger Heyn, Ben Lehner, and Thomas Graf. Single cell RNA-seq identifies the origins of heterogeneity in efficient cell transdifferentiation and reprogramming. *eLife*, 8:e41627, March 2019. Publisher: eLife Sciences Publications, Ltd.

[15] Oscar Franzén, Li-Ming Gan, and Johan L M Björkegren. PanglaoDB: a web server for exploration of mouse and human single-cell RNA sequencing data. *Database*, 2019:baz046, January 2019.

[16] Xi Fu, Shentong Mo, Alejandro Buendia, Anouchka P. Laurent, Anqi Shao, Maria del Mar Alvarez-Torres, Tianji Yu, Jimin Tan, Jiayu Su, Romella Sagatelian, Adolfo A. Ferrando, Alberto Ciccia, Yanyan Lan, David M. Owens, Teresa Palomero, Eric P. Xing, and Raul Rabadan. A foundation model of transcription across human cell types. *Nature*, 637(8047):965–973, January 2025. Publisher: Nature Publishing Group.

[17] Andreia M. Gomes, Ilia Kurochkin, Betty Chang, Michael Daniel, Kenneth Law, Namita Satija, Alexander Lachmann, Zichen Wang, Lino Ferreira, Avi Ma'ayan, Benjamin K. Chen, Dmitri Papatsenko, Ihor R. Lemischka, Kateri A. Moore, and Carlos-Filipe Pereira. Cooperative Transcription Factor Induction Mediates Hemogenic Reprogramming. *Cell Reports*, 25(10):2821–2835.e7, December 2018.

[18] Gavin D Grant, Katarzyna M Kedziora, Juanita C Limas, Jeanette Gowen Cook, and Jeremy E Purvis. Accurate delineation of cell cycle phase transitions in living cells with pip-fucci. *Cell cycle*, 17(21-22):2496–2516, 2018.

[19] Tessa Durakis Green, Stefan Peidli, Ciyue Shen, Torsten Gross, Joseph Min, Samuele Garda, Jake P Taylor-King, Debora Susan Marks, Augustin Luna, Nils Blüthgen, et al. scperturb: Information resource for harmonized single-cell perturbation data. In *NeurIPS 2022 Workshop on Learning Meaningful Representations of Life*, 2022.

[20] Chuner Guo and Samantha A Morris. Engineering cell identity: establishing new gene regulatory and chromatin landscapes. *Curr. Opin. Genet. Dev.*, 46:50–57, October 2017.

[21] Lin Guo, Lihui Lin, Xiaoshan Wang, Mingwei Gao, Shangtao Cao, Yuanbang Mai, Fang Wu, Junqi Kuang, He Liu, Jiaqi Yang, Shilong Chu, Hong Song, Dongwei Li, Yujian Liu, Kaixin Wu, Jiadong Liu, Jinyong Wang, Guangjin Pan, Andrew P. Hutchins, Jing Liu, Duanqing Pei, and Jiekai Chen. Resolving Cell Fate Decisions during Somatic Cell Reprogramming by Single-Cell RNA-Seq. *Molecular Cell*, 73(4):815–829.e7, February 2019. Publisher: Elsevier.

[22] Zhisong He, Leander Dony, Jonas Simon Fleck, Artur Szałata, Katelyn X. Li, Irena Slišković, Hsiu-Chuan Lin, Malgorzata Santel, Alexander Atamian, Giorgia Quadrato, Jieran Sun, Sergiu P. Pașca, J. Gray Camp, Fabian J. Theis, and Barbara Treutlein. An integrated transcriptomic cell atlas of human neural organoids. *Nature*, 635(8039):690–698, November 2024. Publisher: Nature Publishing Group.

[23] Kenichi Horisawa and Atsushi Suzuki. Direct cell-fate conversion of somatic cells: Toward regenerative medicine and industries. *Proceedings of the Japan Academy, Series B*, 96(4):131–158, 2020.

[24] Masahito Inagaki. Cell reprogramming and differentiation utilizing messenger rna for regenerative medicine. *Journal of Developmental Biology*, 12(1):1, 2023.

[25] Miten Jain, Robin Abu-Shumays, Hugh E Olsen, and Mark Akeson. Advances in nanopore direct rna sequencing. *Nature methods*, 19(10):1160–1164, 2022.

[26] Miten Jain, Sergey Koren, Karen H Miga, Josh Quick, Arthur C Rand, Thomas A Sasani, John R Tyson, Andrew D Beggs, Alexander T Dilthey, Ian T Fiddes, et al. Nanopore sequencing and assembly of a human genome with ultra-long reads. *Nature biotechnology*, 36(4):338–345, 2018.

[27] Miten Jain, Hugh E Olsen, Benedict Paten, and Mark Akeson. The oxford nanopore minion: delivery of nanopore sequencing to the genomics community. *Genome biology*, 17:1–11, 2016.

[28] Yuge Ji, Mohammad Lotfollahi, F Alexander Wolf, and Fabian J Theis. Machine learning for perturbational single-cell omics. *Cell Systems*, 12(6):522–537, 2021.

[29] Julia Joung, Sai Ma, Tristan Tay, Kathryn R. Geiger-Schuller, Paul C. Kirchgatterer, Vanessa K. Verdine, Baolin Guo, Mario A. Arias-Garcia, William E. Allen, Ankita Singh, Olena Kuksenko, Omar O. Abudayyeh, Jonathan S. Gootenberg, Zhanyan Fu, Rhiannon K. Macrae, Jason D. Buenrostro, Aviv Regev, and Feng Zhang. A transcription factor atlas of directed differentiation. *Cell*, 186(1):209–229.e26, January 2023.

[30] Uma S Kamaraj et al. Computational methods for direct cell conversion. *Cell Cycle*, 15(24):3343–3354, 2016.

[31] Rosemary Kanasty, J. Robert Dorkin, Arturo Vegas, and Daniel Anderson. Delivery materials for sirna therapeutics. *Nature Materials*, 12(11):967–977, 2013.

[32] Riya Keshri, Damien Detraux, Ashish Phal, Clara McCurdy, Samriddhi Jhajharia, Tung Ching Chan, Julie Mathieu, and Hannele Ruohola-Baker. Next-generation direct reprogramming. *Frontiers in Cell and Developmental Biology*, 12:1343106, 2024.

[33] Christopher Lance, Malte D Luecken, Daniel B Burkhardt, Robrecht Cannoodt, Pia Rautenstrauch, Anna Laddach, Aidyn Ubingazhibov, Zhi-Jie Cao, Kaiwen Deng, Sumeer Khan, et al. Multimodal single cell data integration challenge: results and lessons learned. *BioRxiv*, pages 2022–04, 2022.

[34] Kwanyoung Lee, Hyungjo Byun, and Hyunjung Shim. Cell segmentation in multi-modality high-resolution microscopy images with cellpose. In *Competitions in Neural Information Processing Systems*, pages 1–11. PMLR, 2023.

[35] Hanqin Li, Houbo Jiang, Xinzhen Yin, Jonathan E Bard, Baorong Zhang, and Jian Feng. Attenuation of prrx2 and hey2 enables efficient conversion of adult human skin fibroblasts to neurons. *Biochemical and biophysical research communications*, 516(3):765–769, 2019.

[36] Jizhihui Liu, Qixun Teng, Qing Ma, and Junjun Jiang. Fm2s: Towards spatially-correlated noise modeling in zero-shot fluorescence microscopy image denoising, 2025.

[37] Samantha A Morris. Direct lineage reprogramming via pioneer factors; a detour through developmental gene regulatory networks. *Development*, 143(15):2696–2705, August 2016.

[38] Samantha A. Morris, Patrick Cahan, Hu Li, Anna M. Zhao, Adrianna K. San Roman, Ramesh A. Shivdasani, James J. Collins, and George Q. Daley. Dissecting Engineered Cell Types and Enhancing Cell Fate Conversion via CellNet. *Cell*, 158(4):889–902, August 2014.

[39] Eric Nguyen, Michael Poli, Marjan Faizi, Armin Thomas, Michael Wornow, Callum Birch-Sykes, Stefano Massaroli, Aman Patel, Clayton Rabideau, Yoshua Bengio, et al. Hyenadna: Long-range genomic sequence modeling at single nucleotide resolution. *Advances in neural information processing systems*, 36:43177–43201, 2023.

[40] Paul Nurse, Yoshio Masui, and Leland Hartwell. Understanding the cell cycle. *Nature medicine*, 4(10):1103–1106, 1998.

[41] CZI Cell Science Program, Shibla Abdulla, Brian Aevermann, Pedro Assis, Seve Badajoz, Sidney M Bell, Emanuele Bezzi, Batuhan Cakir, Jim Chaffer, Signe Chambers, et al. Cz cellxgene discover: a single-cell data platform for scalable exploration, analysis and modeling of aggregated data. *Nucleic Acids Research*, 53(D1):D886–D900, 2025.

[42] Owen JL Rackham, Jaber Firas, Hai Fang, Matt E Oates, Melissa L Holmes, Anja S Knaupp, FANTOM Consortium, Harukazu Suzuki, Christian M Nefzger, Carsten O Daub, et al. A predictive computational framework for direct reprogramming between human cell types. *Nature genetics*, 48(3):331–335, 2016.

[43] Scott Ronquist, Geoff Patterson, Lindsey A Muir, Stephen Lindsly, Haiming Chen, Markus Brown, Max S Wicha, Anthony Bloch, Roger Brockett, and Indika Rajapakse. Algorithm for cellular reprogramming. *Proceedings of the National Academy of Sciences*, 114(45):11832–11837, 2017.

[44] Jennifer E Rood, Anna Hupalowska, and Aviv Regev. Toward a foundation model of causal cell and tissue biology with a perturbation cell and tissue atlas. *Cell*, 187(17):4520–4545, 2024.

[45] Yusuf Roohani, Kexin Huang, and Jure Leskovec. Predicting transcriptional outcomes of novel multigene perturbations with gears. *Nature Biotechnology*, 42(6):927–935, 2024.

[46] Yusuf Roohani, Kexin Huang, and Jure Leskovec. Predicting transcriptional outcomes of novel multigene perturbations with GEARS. *Nature Biotechnology*, 42(6):927–935, June 2024. Publisher: Nature Publishing Group.

[47] Rita Silvério-Alves, Ilia Kurochkin, Anna Rydström, Camila Vazquez Echegaray, Jakob Haider, Matthew Nicholls, Christina Rode, Louise Thelaus, Aida Yifter Lindgren, Alexandra Gabriela Ferreira, Rafael Brandão, Jonas Larsson, Marella F. T. R. de Bruijn, Javier Martin-Gonzalez, and Carlos-Filipe Pereira. GATA2 mitotic bookmarking is required for definitive haematopoiesis. *Nature Communications*, 14(1):4645, August 2023. Publisher: Nature Publishing Group.

[48] Amar M Singh, Robert Trost, Benjamin Boward, and Stephen Dalton. Utilizing fucci reporters to understand pluripotent stem cell biology. *Methods*, 101:4–10, 2016.

[49] Cooper Stansbury, Jillian Cwycyshyn, Joshua Pickard, Walter Meixner, Indika Rajapakse, and Lindsey A. Muir. Data-guided direct reprogramming of human fibroblasts into the hematopoietic lineage, August 2024. tex.ids= stansburyDataguidedDirectReprogramming2024a pages: 2024.08.26.609589 section: New Results.

[50] Clara Steichen, Eléanor Luce, Jérôme Maluenda, Lucie Tosca, Inmaculada Moreno-Gimeno, Christophe Desterke, Noushin Dianat, Sylvie Goulinet-Mainot, Sarah Awan-Toor, Deborah Burks, et al. Messenger rna-versus retrovirus-based induced pluripotent stem cell reprogramming strategies: analysis of genomic integrity. *Stem Cells Translational Medicine*, 3(6):686–691, 2014.

[51] Artur Szałata, Karin Hrovatin, Sören Becker, Alejandro Tejada-Lapuerta, Haotian Cui, Bo Wang, and Fabian J. Theis. Transformers in single-cell omics: a review and new perspectives. *Nature Methods*, 21(8):1430–1443, August 2024. Publisher: Nature Publishing Group.

[52] John A. Tadross, Lukas Steuernagel, Georgina K. C. Dowsett, Katherine A. Kentistou, Sofia Lundh, Marta Porniece-Kumar, Paul Klemm, Kara Rainbow, Henning Hvid, Katarzyna Kania, Joseph Polex-Wolf, Lotte Bjerre-Knudsen, Charles Pyke, John R. B. Perry, Brian Y. H. Lam, Jens C. Brüning, and Giles S. H. Yeo. Human HYPOMAP: A comprehensive spatio-cellular map of the human hypothalamus, September 2023. Pages: 2023.09.15.557967 Section: New Results.

[53] Kazutoshi Takahashi and Shinya Yamanaka. Induction of pluripotent stem cells from mouse embryonic and adult fibroblast cultures by defined factors. *cell*, 126(4):663–676, 2006.

[54] THE TABULA SAPIENS CONSORTIUM. The Tabula Sapiens: A multiple-organ, single-cell transcriptomic atlas of humans. *Science*, 376(6594):eabl4896, May 2022. Publisher: American Association for the Advancement of Science.

[55] Christina V Theodoris, Ling Xiao, Anant Chopra, Mark D Chaffin, Zeina R Al Sayed, Matthew C Hill, Helene Mantineo, Elizabeth M Brydon, Zexian Zeng, X Shirley Liu, et al. Transfer learning enables predictions in network biology. *Nature*, 618(7965):616–624, 2023.

[56] Yasuhiro Tomaru, Ryota Hasegawa, Takahiro Suzuki, Taiji Sato, Atsutaka Kubosaki, Masanori Suzuki, Hideya Kawaji, Alistair RR Forrest, Yoshihide Hayashizaki, FANTOM Consortium, et al. A transient disruption of fibroblastic transcriptional regulatory network facilitates trans-differentiation. *Nucleic acids research*, 42(14):8905–8913, 2014.

[57] Connor A Tsuchida, Kevin M Wasko, Jennifer R Hamilton, and Jennifer A Doudna. Targeted nonviral delivery of genome editors in vivo. *Proceedings of the National Academy of Sciences*, 121(11):e2307796121, 2024.

[58] Erwin L Van Dijk, Yan Jaszczyszyn, Delphine Naquin, and Claude Thermes. The third revolution in sequencing technology. *Trends in Genetics*, 34(9):666–681, 2018.

[59] Isaac Virshup, Sergei Rybakov, Fabian J. Theis, Philipp Angerer, and F. Alexander Wolf. anndata: Access and store annotated data matrices. *Journal of Open Source Software*, 9(101):4371, 2024.

[60] Aline Yen Ling Wang. Application of modified mrna in somatic reprogramming to pluripotency and directed conversion of cell fate. *International journal of molecular sciences*, 22(15):8148, 2021.

[61] Haofei Wang, Yuchen Yang, Jiandong Liu, and Li Qian. Direct cell reprogramming: approaches, mechanisms and progress. *Nature Reviews Molecular Cell Biology*, 22(6):410–424, June 2021. Number: 6 Publisher: Nature Publishing Group.

[62] Luigi Warren and Cory Lin. mrna-based genetic reprogramming. *Molecular Therapy*, 27(4):729–734, 2019.

[63] Luigi Warren, Philip D Manos, Tim Ahfeldt, Yuin-Han Loh, Hu Li, Frank Lau, Wataru Ebina, Pankaj K Mandal, Zachary D Smith, Alexander Meissner, et al. Highly efficient reprogramming to pluripotency and directed differentiation of human cells with synthetic modified mrna. *Cell stem cell*, 7(5):618–630, 2010.

[64] Zhiting Wei, Duanmiao Si, Bin Duan, Yicheng Gao, Qian Yu, Zhenbo Zhang, Ling Guo, and Qi Liu. Perturbase: a comprehensive database for single-cell perturbation data analysis and visualization. *Nucleic Acids Research*, 53(D1):D1099–D1111, 2025.

[65] Harold Weintraub, Robert Davis, Stephen Tapscott, Matthew Thayer, Michael Krause, Robert Benezra, T Keith Blackwell, David Turner, Ralph Rupp, Stanley Hollenberg, et al. The myod gene family: nodal point during specification of the muscle cell lineage. *Science*, 251(4995):761–766, 1991.

[66] Harold Weintraub, Stephen J Tapscott, Robert L Davis, Mathew J Thayer, Mohammed A Adam, Andrew B Lassar, and A Dusty Miller. Activation of muscle-specific genes in pigment, nerve, fat, liver, and fibroblast cell lines by forced expression of myod. *Proceedings of the National Academy of Sciences*, 86(14):5434–5438, 1989.

[67] Thomas P. Wytock and Adilson E. Motter. Cell reprogramming design by transfer learning of functional transcriptional networks. *Proceedings of the National Academy of Sciences*, 121(11):e2312942121, March 2024. tex.ids= wytockCellReprogrammingDesign2024a publisher: Proceedings of the National Academy of Sciences.

[68] Yun Xiao, Yonghui Gong, Yanling Lv, Yujia Lan, Jing Hu, Feng Li, Jinyuan Xu, Jing Bai, Yulan Deng, Ling Liu, et al. Gene perturbation atlas (gpa): a single-gene perturbation repository for characterizing functional mechanisms of coding and non-coding genes. *Scientific reports*, 5(1):10889, 2015.

[69] Zhiyuan Xie, Ilya Sokolov, Maria Osmala, Xue Yue, Grace Bower, J. Patrick Pett, Yinan Chen, Kai Wang, Ayse Derya Cavga, Alexander Popov, Sarah A. Teichmann, Ekaterina Morgunova, Evgeny Z. Kvon, Yimeng Yin, and Jussi Taipale. DNA-guided transcription factor interactions extend human gene regulatory code. *Nature*, pages 1–10, April 2025. Publisher: Nature Publishing Group.

[70] Qiao Rui Xing, Chadi El Farran, Pradeep Gautam, Yu Song Chuah, Tushar Warrier, Cheng-Xu Delon Toh, Nam-Young Kang, Shigeki Sugii, Young-Tae Chang, Jian Xu, James J. Collins, George Q. Daley, Hu Li, Li-Feng Zhang, and Yuin-Han Loh. Diversification of reprogramming trajectories revealed by parallel single-cell transcriptome and chromatin accessibility sequencing. *Science Advances*, 6(37):eaba1190, September 2020. Publisher: American Association for the Advancement of Science.

[71] Shinya Yamanaka. Elite and stochastic models for induced pluripotent stem cell generation. *Nature*, 460(7251):49–52, 2009.

[72] Bo Yuan, Ciyue Shen, Augustin Luna, Anil Korkut, Debora S Marks, John Ingraham, and Chris Sander. Cellbox: interpretable machine learning for perturbation biology with application to the design of cancer combination therapy. *Cell systems*, 12(2):128–140, 2021.