Universiteit
Leiden
The Netherlands

Bachelor Computer Science & Economics

Evaluating an Ornstein–Uhlenbeck Model for Perturbation Prediction from Steady-State Gene Expression Data

Maciej Wrona

First Supervisor:
Saber Salehkaleybar

Second Supervisor:
Ana Letícia Garcez Vicente

BACHELOR THESIS

Leiden Institute of Advanced Computer Science (LIACS)
www.liacs.leidenuniv.nl

25/08/2026

Contents

Abstract

Predicting cellular responses to genetic perturbations requires models that incorporate causal information while operating on the steady-state measurements commonly produced by large-scale single-cell experiments. This thesis evaluates the Ornstein–Uhlenbeck-based least-squares framework proposed by (Salehkaleybar, 2026) for learning linear stochastic dynamics from observational and interventional steady-state data. The method is evaluated on synthetic systems with known parameters and on two single-cell perturbation datasets, CrunchDAO Broad Institute Obesity and Perturb-CITE-seq. In the synthetic experiments, complete intervention contexts are held out during fitting and used only to evaluate prediction of unseen perturbations with Maximum Mean Discrepancy (MMD) and Pearson Delta. The biological experiments use the same metrics and compare predictions with a perturbed-centroid baseline. For the real datasets, covariance terms are omitted because their computation is impractical at the evaluated dimensionality and, for Perturb-CITE-seq, strongly affected optimisation through noisy covariance estimates. On synthetic data, the OU model outperforms the baseline on both metrics for every evaluated intervention setup, although performance does not improve monotonically as intervention coverage increases. On CrunchDAO, the model improves both metrics relative to the baseline under the best-overall configuration. On Perturb-CITE-seq, it achieves higher Pearson Delta but higher MMD. These results show that, in the evaluated datasets, the OU-based model can achieve useful predictive performance on held-out perturbations, although its performance relative to the baseline depends on the dataset and evaluation metric.

1 Introduction

Predicting and understanding how biological systems respond to interventions is a fundamental goal in systems biology (Šimić et al., 2025). Disruptions to regulatory circuits contribute to a large proportion of human disease, such as obesity and related disorders (Ramirez et al., 2020; Di Rocco et al., 2024), among others (Bhatia and Kleinjan, 2014; Kyono et al., 2019; Cheng et al., 2024). Identifying disease mechanisms and candidate drug targets depends on predicting how gene expression changes in response to specific perturbations. Gene expression is governed by gene regulatory networks (GRNs): systems of interactions in which genes, primarily through the transcription factors they encode, activate or repress one another. These networks frequently contain feedback loops (Paige et al., 2015; Dunn et al., 2014; Freimer et al., 2022), meaning that two genes can regulate each other directly or indirectly. As a result, gene regulatory networks are generally cyclic rather than acyclic. This property poses a specific computational challenge, because many established causal-inference methods assume that the underlying structure can be represented as a directed acyclic graph under strong assumptions (Brouillard et al., 2020; Lopez et al., 2022; Zheng et al., 2018; Lippe et al., 2021). Understanding how genes regulate one another is therefore not only a matter of biological interest. It is also a methodological one.

The underlying dynamics of a GRN can be represented by a stochastic differential equation (SDE), with its steady-state distribution capturing both the interaction structure and the associated kinetic parameters (Gardiner, 2009; Villaverde et al., 2016). By recovering these parameters, one can infer the causal relationships governing the variables and predict how the system will respond to unseen perturbations (Pearl, 2009). In many biological experiments and related research settings,

observations are collected at isolated time points, providing only static snapshots of the underlying system rather than continuous temporal information (Cao et al., 2019; Schiebinger et al., 2019). This snapshot format is considerably more scalable to collect than time-resolved measurements, but it also means that any method aiming to recover the underlying regulatory dynamics must do so from equilibrium statistics alone. However, many SDE estimation methods require time-series trajectories (Zhang et al., 2024; Oh et al., 2024). To help mitigate the lack of temporal data, experimental interventions such as gene knockdown (Datlinger et al., 2017) can provide complementary insights by deliberately perturbing specific components of the system and revealing their effects. This may resolve non-identifiability issues resulting from being reliant on solely observational data (Peters et al., 2017). For an OU process, observational steady-state data constrain the model through its stationary mean and covariance, which satisfy $\Lambda\mu = b, \Lambda\Sigma + \Sigma\Lambda^\top = D$. These equilibrium equations do not, in general, determine a unique parameterisation of the underlying dynamics: different parameter values can yield the same stationary distribution. Interventional steady-state measurements provide additional moment equations under modified dynamics and therefore restrict the set of parameterisations consistent with the observed data Salehkaleybar, 2026.
Furthermore, advances in single-cell sequencing, combined with CRISPR-based perturbation techniques, have enabled researchers to probe GRNs systematically (Replogle et al., 2022; Datlinger et al., 2021; Norman et al., 2019). Thousands of genes in millions of cells can now be perturbed, after which the resulting change in expression of mRNA can be measured across all genes in each cell. These experiments produce large-scale perturbation datasets that support the study of causal relationships between genes.

There is a rich theoretical literature on causal inference using linear stochastical systems. Early works apply the Lyapunov equation to recover drift structures from steady-state data (Varando and Richard Hansen, 2020). In this setting, (Dettling et al., 2024) proposed a Lasso-based estimator. However, identifiability issues arise as they do not employ interventional data and rely solely on observational data. (Lorch et al., 2024) instead concentrates on modeling the observational and interventional distributions, but their approach does not establish guarantees for recovering the underlying model parameters. (Rohbeck et al., 2024) attains good predictive results, while integrating both observational and interventional measurements. However, the applicability of their approach depends on having intervention data for a large proportion, or potentially the entirety, of genes in the system.

Another direction of study does not attempt to recover interpretable parameters at all. Instead, it trains flexible predictive models, including large neural network architectures, to predict post-perturbation gene expression directly from data (Roohani et al., 2024). Powerful transistor models such as State (Adduri et al., 2025) and XPert (Guo et al., 2026) have also been introduced. These models can achieve high predictive accuracy, but their internal representations cannot be interpreted as explicit regulatory interactions. Moreover, recent benchmarking studies have shown that simple, non-causal baseline predictors can match or exceed the performance of several such models (Viñas Torné et al., 2025), which raises questions about the extent to which their predictive accuracy reflects genuine causal understanding of the underlying system.

A practical constraint in experimental perturbation studies is that intervention coverage is limited. Systematically perturbing a large number of genes increases the experimental burden, while

combinatorial perturbations grow rapidly with the number of candidate genes (He et al., 2025). Consequently, causal inference methods that depend on broad intervention coverage - such as Bycicle (Roohani et al., 2024) - may not match the intervention patterns available in large perturbation datasets. The question is therefore whether interpretable stochastic dynamics can be recovered from steady-state data when interventions are available for only a subset of the variables.

Recent theoretical work by (Salehkaleybar, 2026) addresses this question for multivariate Ornstein–Uhlenbeck processes. The analysis shows that, under structural and spectral assumptions, interventional steady-state measurements can provide sufficient information to recover the underlying parameters up to a global scaling, even when interventions are not available for every variable (Salehkaleybar, 2026). The result relates the required intervention coverage to the strongly connected components of the drift graph rather than to the total number of variables. This provides a theoretical basis for learning interpretable linear stochastic dynamics from incomplete intervention data.

This thesis evaluates the practical application of the least-squares estimation framework introduced by (Salehkaleybar, 2026) to observational and interventional steady-state gene expression data. The study first uses synthetic OU systems with known parameters to examine parameter recovery as the number of available interventions varies. It then applies the method to two real perturbation datasets, the CrunchDAO Broad Institute Obesity dataset[1] and Perturb-CITE-seq (Frangieh et al., 2021), and evaluates its ability to predict unseen perturbation responses. Predictive performance is compared with a perturbed centroid baseline using Maximum Mean Discrepancy and Pearson Delta. The objective is not to establish new identifiability theory, but to assess whether an interpretable stochastic model can be estimated under partial intervention coverage and whether the resulting model provides useful predictions on real perturbation data.

2 Background

2.1 Gene Regulatory Networks

Gene regulatory networks (GRNs) describe the regulatory interactions that determine gene expression within a cell. These interactions are primarily mediated by transcription factors, which activate or repress target genes by binding to specific DNA sequences. Through this mechanism, GRNs control essential cellular processes, such as metabolic regulation by preadipocyte differentiation (Ramirez et al., 2020).

From a computational perspective, GRNs can be represented as directed graphs in which nodes correspond to genes and edges represent regulatory influences. GRNs frequently contain feedback loops, meaning that genes can regulate each other directly or indirectly. Consequently, the underlying regulatory structure is generally cyclic due to feedback loops (Paige et al., 2015; Dunn et al., 2014; Freimer et al., 2022), which makes it unsuitable for methods that assume directed acyclic graphs (DAGs). Modeling these cyclic dependencies accurately is therefore an important challenge in causal inference and perturbation prediction.

[1] https://hub.crunchdao.com/competitions/broad-obesity-1/resources/datasets

2.2 Single-Cell Perturbation Data

Advances in single-cell sequencing have enabled gene expression to be measured at the resolution of individual cells, revealing patterns of cellular heterogeneity that bulk measurements cannot capture. When combined with CRISPR-based perturbation technologies, such as Perturb-seq, these methods reveal how targeted interventions affect the expression of other genes, producing large-scale datasets for studying causal relationships between genes.

Most perturbation datasets consist of steady-state, or snapshot, measurements. Rather than observing the temporal evolution of gene expression, each cell is measured only after the system has reached equilibrium following a perturbation. Collecting steady-state data is considerably more scalable than performing time-series experiments. However, the absence of temporal information makes recovering the underlying regulatory dynamics challenging, since a single equilibrium snapshot does not reveal how that state was reached.

2.3 Causal Learning

Traditional machine learning models identify statistical associations between variables, whereas causal learning aims to infer how interventions alter the underlying system. This distinction is central to perturbation prediction, where the objective extends beyond predicting gene expression values to estimating how perturbing one gene influences the expression of others.

By actively modifying specific genes, perturbation experiments provide information that cannot generally be recovered from observational data alone. Interventions of this kind help separate direct causal effects from indirect correlations, thereby reducing ambiguity during model estimation.

A closely related concept is identifiability, which concerns whether the parameters of a model can be uniquely recovered from the available data under a given set of assumptions. If a model is not identifiable, multiple parameterizations may produce identical observations, making it impossible to determine the true underlying system. Establishing identifiability is therefore a prerequisite for developing interpretable causal models.

2.4 Stochastic Dynamical Systems

Many computational approaches model gene regulation as a continuous-time stochastic dynamical system. Stochastic differential equations (SDEs) provide a suitable framework for this purpose, as they combine deterministic system dynamics with stochastic fluctuations that capture random variation in gene expression.

This thesis considers the Ornstein–Uhlenbeck (OU) process, a linear SDE that describes systems evolving toward a stable equilibrium. Within this framework, the drift matrix represents the interactions that govern the system dynamics, while the diffusion term models stochastic noise. The mathematical properties of the OU process make it particularly suitable for learning from steady-state data, as its stationary distribution is fully characterized by its first two statistical moments: the mean and the covariance.

2.5 Learning from Steady-State Data

Since steady-state datasets do not contain temporal trajectories, parameter estimation relies on the statistical properties of the equilibrium distribution rather than on dynamic observations. For the OU process, the stationary mean and covariance are directly related to the model parameters, which allows parameter estimation to be formulated as a moment-matching problem. In particular, the stationary covariance satisfies the continuous Lyapunov equation, which links the observed equilibrium statistics to the underlying system dynamics. The corresponding mathematical formulation is presented in Section 4.4.

Equilibrium statistics alone are insufficient to uniquely recover the parameters of a stochastic dynamical system, since observational data cannot distinguish between multiple networks that produce the same equilibrium distribution. Incorporating interventional steady-state measurements introduces additional constraints, which improve parameter identifiability and enable more reliable estimation of the underlying dynamics. Combining observational and interventional data therefore forms the foundation of the method proposed in (Salehkaleybar, 2026), which will be the basis for this thesis.

3 Related work

Research on perturbation prediction has evolved from stochastic models that estimate causal system dynamics to deep learning models that directly predict perturbation responses. This section reviews the methods most relevant to this thesis and discusses the limitations that motivate the proposed method.

3.1 Causal dynamical models

Initial work on perturbation prediction using stochastic dynamical systems focused on learning from observational steady-state data. (Varando and Richard Hansen, 2020) introduced the Graphical Continuous Lyapunov Model (GCLM), which estimates the drift matrix of an OU process by solving a sparsity-regularized optimization problem constrained by the continuous Lyapunov equation. Within this setting, (Dettling et al., 2023) provided a complete characterization of identifiability. Specifically, they proved that the stationary covariance uniquely determines the drift matrix exactly when the diffusion matrix is known and the drift graph is simple, i.e., it contains no directed two-cycles. These assumptions, however, restrict the applicability of their results to observational data and exclude models with bidirectional dependencies or unknown diffusion parameters.

(Dettling et al., 2024) extended this framework by introducing a Lasso-based estimator and deriving statistical conditions under which the underlying drift matrix can be recovered consistently. Their method improves sparse parameter estimation under specific assumptions but remains limited to observational data. Thus, identifiability remains dependent on restrictive assumptions that are unlikely to hold in large-scale gene regulatory networks.

On the other hand, assuming a low-rank drift matrix, (Zweig et al., 2026) investigated whether linear SDEs can be uniquely identified from mean-shift intervention data. While (Lorch et al., 2024)

introduced a kernel-based method for learning stationary SDEs via distribution matching. Their work emphasizes density estimation and predictive generalization rather than parameter recovery.

Meanwhile, the availability of large-scale perturbation datasets motivated methods that explicitly incorporate interventions. For example, Bicycle estimates model parameters by jointly fitting observational and interventional steady-state distributions (Rohbeck et al., 2024). Including perturbation data reduces the ambiguity present in observational learning and improves the prediction of unseen interventions. Its theoretical analysis, however, assumes interventions on nearly every variable in the system, which limits its applicability when perturbation experiments are available for only a subset of genes.

Expanding on the above, (Salehkaleybar, 2026) studied the identifiability of linear stochastic dynamics from steady-state observational and interventional data. For multivariate Ornstein–Uhlenbeck processes, the authors show that, under suitable structural and spectral conditions, a single intervention per strongly connected component (SCC) of the drift graph is generically sufficient to recover the underlying drift, external input, and diffusion parameters, up to a global scaling factor. In particular, their result requires the condensation graph of SCCs to be connected with a single root, together with a set of spectral nondegeneracy conditions. This provides a substantially weaker intervention requirement than approaches that assume interventions on nearly every variable, and is therefore particularly relevant to large-scale perturbation settings where experimental coverage is limited. The author further proposes a recursive identification procedure that processes SCCs in topological order, as well as a regularized least-squares estimator based on the steady-state mean and covariance equations.

(Sethuraman, Lopez, Mohan, Fekri, Biancalani, and Hütter, 2023) proposed NODAGS-Flow, which models nonlinear cyclic causal systems using residual normalizing flows. Instead of matching equilibrium moments, the method learns the equilibrium distribution through maximum likelihood. This enables the model to represent nonlinear dependencies and feedback loops while avoiding acyclicity assumptions. However, the method does not provide theoretical guarantees that the recovered model parameters uniquely correspond to the underlying dynamical system. Additionally, the authors evaluated NODAGS-Flow on a Perturb-CITE-seq dataset containing 218,331 melanoma cells across control, co-culture, and IFN-γ conditions and restricted the analysis to 61 genes because of computational limitations. This 61-gene IFN-γ setting is also used as the Perturb-CITE-seq benchmark in this thesis.

3.2 Deep learning approaches

Deep learning approaches formulate perturbation prediction as a supervised learning problem. Rather than estimating the parameters of a causal model, these methods predict post-perturbation gene expression directly from training data.

GEARS uses a graph neural network that combines perturbation data with prior biological knowledge to predict unseen perturbations (Roohani et al., 2024). Incorporating known gene relationships into the model architecture improves generalization while reducing the amount of training data required for new perturbation combinations.

More recent methods employ transformer architectures to learn representations of cellular responses from large perturbation datasets. State (Adduri et al., 2025) and XPert (Guo et al., 2026) use self-attention to model dependencies between genes and report improved prediction accuracy on benchmark datasets compared with earlier neural-network-based approaches. Because these models optimize predictive performance rather than parameter estimation, their learned representations cannot be interpreted as explicit causal interactions. Additionally, a notable limitation of such models is their substantial data requirements for effective training. For instance, State (Adduri et al., 2025) was trained on more than 100 million perturbed cells.

Moreover, recent benchmarking studies have shown that simple baseline methods achieve predictive performance comparable to, and in some cases exceeding, that of state-of-the-art models (Viñas Torné et al., 2025).

4 Methodology

This chapter describes the methodology used to estimate the parameters of a linear stochastic dynamical system from observational and interventional steady-state gene expression data. Our implementation follows the framework proposed by (Salehkaleybar, 2026), where a simple modification has been introduced for large-scale datasets, where the covariance-based objective is computationally infeasible. The proposed implementation is evaluated on publicly available perturbation datasets and compared against a perturbed centroid-based baseline using standardised perturbation prediction metrics.

4.1 Research Design

This research follows a quantitative experimental design in which the proposed learning algorithm is evaluated through comparative benchmarking on publicly available single-cell perturbation datasets. Rather than proposing a new mathematical framework, this thesis implements the recently proposed least-squares estimator for OU processes and investigates its applicability to large-scale perturbation prediction.

The study consists of two experimental phases. In the first phase, the proposed method is evaluated on synthetically generated OU systems, for which the true model parameters are known. These experiments assess whether the optimisation procedure can recover the underlying parameters under controlled conditions, and evaluate how parameter recovery depends on the number of interventions. In the second phase, the method is evaluated on real-world perturbation datasets, where the learned parameters are used to predict unseen perturbation responses and compared against a centroid-based baseline.

The implementation closely follows the methodology introduced by (Salehkaleybar, 2026). Their work establishes theoretical identifiability guarantees, showing that, under suitable assumptions, the parameters of the underlying stochastic dynamical system can be recovered up to a global scaling factor from observational and interventional steady-state moments. As the theoretical proofs

fall outside the scope of this thesis, only the assumptions necessary for the implementation are summarised in Section **??**.

The complete experimental workflow is summarised below.

1. Generate synthetic OU systems and simulate observational and interventional data.
2. Acquire real observational and interventional steady-state datasets from CrunchDAO and Perturb-CITE-seq.
3. Compute empirical stationary moments.
4. Estimate the model parameters through least-squares optimisation.
5. Predict stationary responses for unseen perturbations.
6. Evaluate parameter recovery on synthetic data.
7. Evaluate predictive performance using Maximum Mean Discrepancy and Pearson Delta.

4.2 Datasets

This study evaluates the proposed method using both synthetic and real-world datasets. The synthetic dataset is used to assess parameter recovery under controlled conditions, since the true model parameters are known. The real-world datasets consist of two CRISPR-mediated single-cell perturbation experiments and are used to evaluate the ability of the learned model to predict unseen perturbation responses across different biological datasets.

4.2.1 Synthetic Dataset

In addition to the biological datasets, synthetic datasets were generated to evaluate the method under controlled conditions where the true OU parameters are known. Each dataset was generated from a randomly sampled stable ten-dimensional OU process using the same intervention mechanism assumed by the estimator. Off-diagonal entries of the drift matrix were sampled independently with probability 0.3 from a Gaussian distribution with mean 0.2 and standard deviation 0.8. The diagonal entries were subsequently chosen to make each row strictly diagonally dominant, ensuring positive stability of the generated drift matrix. The constant vector b was sampled uniformly from $[0.2, 1.5]$, while the diagonal entries of the diffusion matrix were sampled uniformly from $[0.2, 0.4]$.

For each system, the observational stationary mean and covariance were computed from the linear mean equation and continuous Lyapunov equation. Interventions were generated by zeroing the off-diagonal entries of the targeted row of Λ while retaining its diagonal entry and leaving b and D unchanged. Cells were then sampled from the corresponding multivariate Gaussian stationary distribution. Each system contains 40,000 observational cells and 20,000 cells for every generated intervention.

The experiments use six intervention-availability settings, $2, 4, 6, 8, 10$. Each setting contains 40 independently generated systems with 10 genes, giving 240 generated systems in total. In setup S, interventions are generated for nodes $0, \ldots, S-1$. For predictive evaluation, available intervention contexts are divided into fixed training and held-out sets, as detailed in 4.7.1.

4.2.2 Perturb-CITE-seq Dataset

One of the two real-world benchmarks used in this thesis is the Perturb-CITE-seq dataset introduced by (Frangieh et al., 2021) and subsequently used by (Sethuraman, Lopez, Mohan, Fekri, Biancalani, and Hütter, 2023) in their evaluation of NODAGS-Flow. The original experiment investigated genetic drivers of resistance to immune checkpoint inhibitors in melanoma cells. The dataset contains 218,331 cells across three conditions: control, co-culture, and interferon-γ (IFN-γ), containing 57,627, 73,114, and 87,590 cells, respectively.

Following Sethuraman et al. (2023), this thesis restricts the analysis to 61 genes from the IFN-γ condition. The resulting dataset contains single-gene perturbations targeting the selected genes, together with the corresponding gene expression measurements. All 61 selected genes are represented by single-gene interventions in the dataset used in this thesis.

The Perturb-CITE-seq dataset is used as a second real-world benchmark for evaluating whether the proposed method generalises beyond the CrunchDAO dataset. Unlike the synthetic experiments, the underlying regulatory parameters are unknown. The learned model is therefore evaluated through its ability to predict the expression distribution following unseen perturbations.

4.2.3 CrunchDAO Dataset

The second real-world dataset is the Broad Institute Obesity Challenge dataset, provided through the CrunchDAO platform[2] . The dataset consists of observational and interventional single-cell gene expression measurements collected after CRISPR perturbations. The preprocessing pipeline required to construct the released dataset was performed by the competition organisers prior to publication and is described in Section 4.3.3.

The dataset is provided by CrunchDAO as two separate files.

- `obesity_challenge_1.h5ad`: the training set, containing single-cell RNA-seq profiles for a subset of gene perturbations. This file contains 88,202 cells across 21,592 genes, covering 122 unique perturbed genes plus non-targeting control (NC) cells, drawn from the full set of 157 targeted genes used across the challenge, of which 150 are transcription factors.
- `obesity_challenge_1_local_gtruth.h5ad`: a local test set containing a distinct set of single-gene perturbations, provided to allow offline evaluation prior to submission to the official leaderboard. This file contains 5 held-out test perturbations.

Each perturbation is profiled via single-cell RNA sequencing, which captures gene expression at the individual-cell level. For modelling, the gene space was restricted to a working subset of 10,238

[2] https://hub.crunchdao.com/competitions/broad-obesity-1/resources/datasets

genes. Evaluation metrics were computed on a fixed, randomly seeded panel of 1,000 genes sampled from this working space, ensuring a consistent comparison between model predictions and the baseline across all perturbations. The file `obesity_challenge_1_local_gtruth.h5ad` was used as the evaluation set throughout this thesis, as the complete evaluation set was not shared publicly; this local ground truth was the only test set available, since the non-local ground truth was not released.

4.3 Data Preprocessing

4.3.1 Synthetic Data

As the synthetic datasets were generated directly from the stationary distribution of the OU process, no preprocessing was required. Empirical stationary means and covariance matrices were estimated directly from the simulated observational and interventional samples and used directly in the optimisation procedure.

4.3.2 Perturb-CITE-seq

The Perturb-CITE-seq data were restricted to the IFN-γ condition before model estimation. From the approximately 20,000 genes measured in the original dataset, 61 genes were selected for the analysis, following the experimental setup of (Sethuraman, Lopez, Mohan, Fekri, Biancalani, and Huetter, 2023). The corresponding single-gene perturbations were retained as interventional conditions.

For each retained condition, the gene-expression measurements were used to estimate the empirical stationary mean. The control measurements provide the observational condition, while cells associated with the selected single-gene perturbations provide the interventional conditions. Only the 61 selected genes were included in the resulting expression vectors.

No protein measurements were used in the proposed model. The analysis therefore uses the transcriptomic component of the Perturb-CITE-seq experiment only.

4.3.3 CrunchDAO

For the CrunchDAO dataset, preprocessing was performed by the competition organisers prior to release. The implementation developed in this thesis therefore uses the processed gene expression matrices supplied by the competition directly. The preprocessing pipeline applied by the organisers is summarised below.

- **Quality control:** low-quality cells and doublets were removed based on sequencing library complexity, gene detection rate, and mitochondrial gene content.
- **Guide filtering:** cells were restricted to those with a single, confidently assigned guide corresponding to one perturbation; guides represented by fewer than 10 cells were excluded.
- **Gene filtering:** genes detected in fewer than 10 cells were removed; known signature genes (from `signature_genes.csv`) were subsequently re-introduced regardless of this threshold.

- **Normalisation:** raw counts were normalised to a target sum of 10^5 per cell, followed by a standard single-cell RNA-seq log-transformation. Normalised values are stored in `adata.X`, while raw counts are preserved in `adata.layers['counts']` for reproducibility.
- **Perturbation annotation:** each cell is labelled in `adata.obs['gene']` as either a non-targeting control (`NC`) or the name of its targeted gene.
- **Cell state annotation:** cells were further annotated for enrichment of *pre-adipocyte*, *adipocyte*, and *lipogenic* transcriptional programmes (or *other*, if none applied), based on expert-curated canonical signature genes; per-perturbation programme proportions are provided in `program_proportion.csv` (not used in our thesis, still include?).

4.4 Proposed Method

4.4.1 Ornstein–Uhlenbeck Formulation

Gene expression dynamics are modelled using the linear OU process

$$dx = (-\Lambda x + b)\,dt + \sigma\,dW_t, \tag{1}$$

where

- $x \in \mathbb{R}^n$ denotes the vector containing the expression levels of all genes,
- $\Lambda \in \mathbb{R}^{n\times n}$ is the drift matrix describing regulatory interactions,
- $b \in \mathbb{R}^n$ vector that represents external input,
- $\sigma \in \mathbb{R}^{n\times n}$ is the diffusion matrix,
- W_t denotes an n-dimensional Wiener process.

The drift matrix determines how perturbations propagate through the regulatory network. Diagonal entries describe self-regulation, whereas off-diagonal entries represent directed regulatory influences between genes. Positive stability of Λ guarantees convergence towards a unique stationary distribution.
The diffusion term models stochastic fluctuations in gene expression, allowing the stationary mean and covariance to be computed directly from the model parameters. As both non-synthetic datasets considered in this thesis consist exclusively of steady-state observations, parameter estimation is based on the stationary distribution rather than on temporal trajectories.

4.4.2 Stationary Mean

For a stable OU process, the stationary mean satisfies

$$\mu = \mathbb{E}[x_\infty] = \Lambda^{-1} b, \tag{2}$$

where x_∞ denotes a random vector distributed according to the stationary distribution.

Equation (2) provides a direct linear relationship between the drift matrix and the observed equilibrium mean. Every observational or interventional dataset therefore contributes information about the unknown parameters through its empirical stationary mean.
Rearranging Equation (2) yields

$$\Lambda\mu = b, \tag{3}$$

which forms the first component of the optimisation objective.

4.4.3 Stationary Covariance

The stationary covariance matrix

$$\Sigma = \mathbb{E}\left[(x_\infty - \mu)(x_\infty - \mu)^\top\right]$$

satisfies the continuous Lyapunov equation

$$\Lambda\Sigma + \Sigma\Lambda^\top = D, \tag{4}$$

where

$$D = \sigma\sigma^\top.$$

Throughout this work, the diffusion matrix is assumed to be diagonal,

$$D = \mathrm{diag}(d_1, \ldots, d_n),$$

with strictly positive diagonal entries.

The Lyapunov equation establishes the relationship between the covariance of the stationary distribution and the underlying dynamical system. Observational covariance matrices therefore provide additional constraints during parameter estimation, beyond those obtained from the stationary mean alone.

4.4.4 Modelling Genetic Perturbations

CRISPR knockout experiments are modelled by modifying the drift matrix, rather than by changing the observed data directly. For an intervention targeting gene i, the corresponding row of the drift matrix is replaced by

$$\tilde{\Lambda}^{(i)}_{kj} = \begin{cases} \Lambda_{kj}, & k \neq i, \\ 0, & k = i,\ j \neq i, \\ \Lambda_{ii}, & k = i,\ j = i. \end{cases}$$

This intervention removes all incoming regulatory effects on the perturbed gene, while leaving the regulatory interactions among all other genes unchanged. The intervened dynamics become

$$dx = (-\tilde{\Lambda}^{(i)}x + b)\,dt + \sigma\,dW_t.$$

The stationary mean of the intervened system satisfies

$$\mu^{(i)} = (\tilde{\Lambda}^{(i)})^{-1} b, \tag{5}$$

or, equivalently,

$$\tilde{\Lambda}^{(i)} \mu^{(i)} = b. \tag{6}$$

Likewise, the stationary covariance satisfies

$$\tilde{\Lambda}^{(i)} \Sigma^{(i)} + \Sigma^{(i)} (\tilde{\Lambda}^{(i)})^{\top} = D. \tag{7}$$

Each intervention therefore contributes an additional set of linear constraints, which reduce the ambiguity present when learning solely from observational data.

4.4.5 Identifiability

The optimisation procedure used in this thesis follows the identifiability framework of (Salehkaleybar, 2026). Observational steady-state moments constrain the OU parameters through

$$\Lambda \mu = b, \qquad \Lambda \Sigma + \Sigma \Lambda^{\top} = D,$$

but these constraints do not generally determine a unique parameter triple (Λ, b, D). Interventional stationary moments provide additional equations through the corresponding modified drift matrices. For this model, identifiability up to a global scale means that any parameter triple $(\widehat{\Lambda}, \widehat{b}, \widehat{D})$ producing the same observational and interventional stationary moments must satisfy

$$\widehat{\Lambda} = c\Lambda, \qquad \widehat{b} = cb, \qquad \widehat{D} = cD, \qquad c > 0.$$

This remaining ambiguity is intrinsic to stationary data because the first- and second-moment equations are invariant under common positive scaling of Λ, b, and D.

The drift graph is not assumed to be known during estimation. It is defined by the unknown drift matrix, with an edge $j \to i$ whenever $\Lambda_{ij} \neq 0$. The estimator receives observational and interventional data and the corresponding intervention targets, while Λ, b, and D are estimated. Under the spectral conditions specified by (Salehkaleybar, 2026), these parameters are generically identifiable up to a global scale if the SCC condensation graph of the underlying drift matrix is connected with a single root and at least one intervention is available within every SCC. The graph conditions therefore characterize when parameter recovery is possible.

These results provide the theoretical justification for the least-squares optimisation procedure described in the following section.

4.4.6 Least-Squares Optimisation

For the synthetic experiments, the estimated parameters can be compared directly with the known ground-truth parameters, which allows parameter recovery to be evaluated explicitly. For the

biological datasets, where the underlying parameters are unknown, the estimated model is instead evaluated through its ability to predict the stationary response to unseen perturbations.
In practice, no ground truth is available; only empirical estimates of the stationary moments are available, $\hat{\mu}, \hat{\Sigma}, \hat{\mu}^{(i)}, \hat{\Sigma}^{(i)}$.

Replacing the exact moments with empirical estimates introduces sampling noise, which renders the system of equations generally inconsistent. Rather than solving the equations exactly, parameter estimation is therefore formulated as a least-squares optimisation problem.
For the parameter vector

$$\Theta = (\mathrm{vec}(\Lambda), b, d),$$

the objective function becomes

$$\begin{aligned} L(\Theta) = \quad & \alpha_O \Big(\Big\| \Lambda\hat{\mu} - b \Big\|_2^2 + \Big\| \Lambda\hat{\Sigma} + \hat{\Sigma}\Lambda^\top - D \Big\|_F^2 \Big) \\ & + \alpha_I \Big(\sum_{i \in I} \Big\| \tilde{\Lambda}^{(i)}\hat{\mu}^{(i)} - b \Big\|_2^2 + \sum_{i \in I} \Big\| \tilde{\Lambda}^{(i)}\hat{\Sigma}^{(i)} + \hat{\Sigma}^{(i)}(\tilde{\Lambda}^{(i)})^\top - D \Big\|_F^2 \Big), \end{aligned} \tag{8}$$

where $\| \cdot \|_F$ denotes the Frobenius norm. The first term measures agreement between the observational stationary mean and the estimated parameters. The second term minimises the residual of the observational Lyapunov equation. The remaining two terms enforce consistency between the estimated parameters and every available interventional steady-state distribution.

Following the original method, the observational and interventional losses are weighted separately using the coefficients α_O and α_I, which allows observational data to receive a larger weight when more observational samples are available.

4.4.7 ℓ_1 Regularisation

To promote sparsity in the estimated drift graph, following (Salehkaleybar, 2026), an ℓ_1 penalty is imposed on the off-diagonal entries of Λ,

$$R(\Lambda) = \gamma \sum_{i \neq j} |\Lambda_{ij}|,$$

where $\gamma 0$ and denotes the regularisation coefficient. The complete optimisation problem becomes

$$\min_{\Theta} L(\Theta) + R(\Lambda), \tag{9}$$

subject to positivity constraints, such as $\Lambda_{kk} > 0, d_k > 0$. These positivity constraints are imposed following (Salehkaleybar, 2026).

The value of γ was selected through a hyperparameter sweep using Weights & Biases (W&B).

4.4.8 Adaptation for Large-Scale Real-World Datasets

The original optimisation procedure incorporates both the stationary mean equations and the Lyapunov equations. For the real-world datasets considered in this thesis, however, the covariance-based terms were not included in the optimisation. Computing and repeatedly evaluating the covariance residuals for the high-dimensional gene expression data of the CrunchDAO dataset was computationally impractical given the available hardware. The choice to not include the covariance-based terms in the Perturb-Cite-seq dataset was because it would catch too much noise, resulting in an enormous starting loss of $1e-11$.
The optimisation for both the CrunchDAO and Perturb-CITE-seq datasets therefore uses only the observational and interventional mean equations,

$$L_{\text{CrunchDAO}} = \alpha_O \|\Lambda\hat{\mu} - b\|_2^2 + \alpha_I \sum_{i \in I} \|\tilde{\Lambda}^{(i)}\hat{\mu}^{(i)} - b\|_2^2. \tag{10}$$

The same ℓ_1 regularisation term is retained for both datasets,

$$\min_{\Theta} L_{\text{CrunchDAO}} + R(\Lambda).$$

Consequently, the real-world experiments evaluate whether the drift matrix can provide useful perturbation predictions when estimated from steady-state means alone. The covariance-based terms are retained only in the synthetic experiments, where the smaller dimensionality makes their computation feasible.

4.4.9 Prediction of Unseen Perturbations

After optimisation, the estimated parameters were used to predict the stationary response of unseen perturbations. For a previously unseen intervention, the corresponding drift matrix was modified according to the intervention model described earlier. The predicted stationary mean was then obtained by solving

$$\hat{\mu}^{(i)} = (\tilde{\Lambda}^{(i)})^{-1}\hat{b}.$$

The estimated parameters are identifiable only up to a positive global scaling factor. This ambiguity does not affect perturbation prediction, however, because the scaling cancels in the stationary mean,

$$(c\Lambda)^{-1}(cb) = \Lambda^{-1}b.$$

All parameterisations that differ only by a global scaling factor therefore produce the same predicted steady-state gene expression.

4.5 Baseline Method

The proposed method is compared against the centroid baseline. The baseline summarises each perturbation in the training set by its centroid, that is, the mean gene expression profile across all cells associated with that perturbation. Predictions for unseen perturbations are then obtained

using these centroids.

In contrast to the proposed method, the centroid baseline does not estimate an explicit gene regulatory network or dynamical system. Instead, it predicts gene expression directly from the reference perturbation profiles contained in the training data. The baseline therefore serves as a reference against which the proposed causal modelling approach can be evaluated.

Both methods are evaluated using the same training and test splits, and compared using the evaluation metrics described in the following section.

4.6 Evaluation Metrics

4.6.1 Maximum Mean Discrepancy

Maximum Mean Discrepancy measures the distance between two probability distributions from their samples. Given two distributions P and Q, the squared MMD is defined as

$$\mathrm{MMD}^2(P, Q) = \left\| \mathbb{E}_{x \sim P}[\phi(x)] - \mathbb{E}_{y \sim Q}[\phi(y)] \right\|^2, \tag{11}$$

where $\phi(\cdot)$ denotes the feature mapping induced by a kernel function.

In this work, MMD quantifies the discrepancy between the predicted and observed gene expression distributions following each perturbation. Smaller values indicate that the predicted distribution more closely matches the experimental observations.

4.6.2 Pearson Delta

Pearson Delta evaluates, in the case, the agreement between the predicted and observed perturbation effects by computing the Pearson correlation coefficient between their differential expression profiles. For two vectors x and y, the Pearson correlation coefficient is given by

$$r = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}\sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2}}, \tag{12}$$

where $\bar{x}$ and $\bar{y}$ denote the sample means.

Pearson Delta measures how closely the predicted perturbation effect follows the observed change in gene expression, irrespective of differences in absolute magnitude. Larger positive values indicate stronger agreement between the predicted and observed perturbation responses, while negative values indicate a negative correlation.

4.7 Experimental Setup

The model was optimised using the Adam optimiser. Hyperparameter selection was performed using W&B sweeps, where the value was chosen randomly and uniformly per sweep within its search space as depicted in the tables below. This does not hold for a search space with discrete values.

Furthermore, the Bayesian method was chosen for the sweeping agent during hyperparameter optimisation.

4.7.1 Synthetic Setup

Each W&B sweep used a fixed synthetic instance and static perturbation split, keeping data, test contexts, and baseline constant. Train/test counts were 1/1, 3/1, 4/2, 6/2, and 8/2, for each setup 2, 4, 6, 8, 10, respectively. The model was trained for 5,000 epochs for each intervention count, while 20 sweeps were performed each time.

Table 2 lists the best hyperparameter configuration found for each intervention count, selected via sweeping.

Table 1: Hyperparameter search space for synthetic experiments. The search space consists of a range denoted by the minimum and maximum values.

Hyperparameter	**Search space**
Learning rate (lr)	10^{-5} – 10^{-1}
a_0	0.1 – 10.0
a_1	10^{-3} – 1.0
L1 coefficient	10^{-6} – 0.1

Table 2: Best hyperparameters for synthetic experiments by intervention count.

Interventions	**lr**	a_0	a_1	**L1 coef.**	**Loss**
2	0.0994	5.678	0.5295	0.018	0.0345
4	0.0254	2.867	0.0586	0.0004	0.0024
6	0.0152	5.232	0.3138	0.0003	0.0043
8	0.0966	5.042	0.9339	0.0038	0.0184
10	0.0490	9.664	0.3264	0.0057	0.0458

4.7.2 Perturb-CITE-seq Setup

The Perturb-CITE-seq experiments used the hyperparameter search space shown in Table 3. Hyperparameter configurations were evaluated using the same prediction metrics as the other experiments. The best configurations according to MMD, Pearson Delta, and an overall selection criterion are reported in Table 4. For this dataset, the model was trained for 20,000 epochs, and a total of 50 sweeps were performed.

4.7.3 CrunchDAO Setup

Table 6 lists the best hyperparameter configuration found for each selection criterion on the CrunchDAO dataset. The learning rate was swept over 2 values, and the L1 coefficient had a constant value. This is due to manual searching prior to sweeping, as full sweeping would otherwise have been too costly in terms of time, given the size of the dataset. In total, 30 sweeps were performed, while the model trained for 2,000 epochs each time.

Table 3: Hyperparameter search space for Perturb-CITE-Seq experiments.

Hyperparameter	**Search space**
Learning rate (lr)	1e-5 – 1e-1
a_0	0.2 – 10.0
a_1	0.1 – 1.0
L1 coefficient	1e-5 – 1e-1

Table 4: Best hyperparameters for Perturb-CITE-Seq experiments by selection criterion.

Configuration	**lr**	a_0	a_1	**L1 coef.**	**Loss**
Best MMD	0.0067	1.2173	0.1245	0.0335	4.8470
Best Pearson	0.0576	9.3346	0.1801	0.0001	3.9025
Best overall	0.0998	9.4785	0.1495	0.0351	5.3733

5 Results

This section reports the empirical results obtained for the proposed OU-based method. All experiments are evaluated using the same two metrics: Maximum Mean Discrepancy (MMD) and Pearson Delta. The only baseline considered is the perturbed centroid baseline. Results are reported for synthetic systems and for two real-world datasets: Perturb-CITE-seq and CrunchDAO.

5.1 Synthetic Experiments

Synthetic experiments were conducted with 2, 4, 6, 8, and 10 interventions, in order to assess how the proposed method behaves as the amount of interventional information increases.

As mentioned in 4.7.1, for every setup, the available interventions were divided using a fixed perturbation-level split. With 2, 4, 6, 8, and 10 available interventions, respectively 1, 3, 4, 6, and 8 interventions were used for fitting, leaving 1, 1, 2, 2, and 2 complete intervention contexts for testing. No cells, means, or covariances from these held-out contexts entered the least-squares objective. The split was kept fixed across W&B sweep runs so that the baseline, training data, and evaluation perturbations did not change between hyperparameter configurations. This makes the predictive metrics out-of-sample with respect to intervention identity.

Table **??** reports predictive performance on the unseen perturbations. The proposed model achieves a lower MMD than the perturbed-centroid baseline in every evaluated setup. Model MMD is 0.0250 with two available interventions, decreases to 0.0074 with four, and then increases to 0.0088, 0.0255, and 0.0611 for six, eight, and ten interventions, respectively. The corresponding baseline values are 0.5392, 0.5297, 0.3912, 0.3963, and 0.4540. The model therefore maintains substantially better distributional agreement than the baseline, but MMD does not decrease as intervention availability increases.
For the Pearson Delta, the model achieves values of 0.9509, 0.9745, 0.9445, 0.8522, and 0.8052 for the 2, 4, 6, 8, and 10 intervention settings, respectively. The corresponding baseline values are -0.2901, 0.0151, 0.1392, 0.0043, and -0.4742. The highest model Pearson Delta is observed for the four-intervention setting, while it keeps decreasing until the last, ten-intervention, setting. Across

Table 5: Hyperparameter search space for CrunchDAO experiments.

Hyperparameter	**Search space**
Learning rate (lr)	1e-3 & 1e-2
a_0	0.2 – 2.0
a_1	0.1 – 1.0
L1 coefficient	1e-4

Table 6: Best hyperparameters for CrunchDAO experiments by selection criterion.

Configuration	**lr**	a_0	a_1	**L1 coef.**	**Loss**
Best MMD	0.01	0.4426	0.2878	0.001	0.0046
Best Pearson	0.001	0.3333	0.1546	0.0001	0.0696
Best overall	0.01	1.9481	0.1114	0.001	0.0388

all five evaluated settings, the model Pearson Delta is higher than that of the perturbed-centroid baseline.

5.2 Perturb-CITE-seq Experiments

Table 8 reports the results obtained on the Perturb-CITE-seq dataset for three model configurations, selected according to MMD, Pearson Delta, and the overall selection criterion, alongside the perturbed centroid baseline.

The perturbed centroid baseline achieves the lowest MMD, with a value of 0.0105, whereas the best MMD obtained by the proposed method is 0.06027. The proposed method therefore does not improve distributional agreement according to MMD on Perturb-CITE-seq. Pearson Delta shows the opposite result. The baseline achieves a value of 0.0649, while the three model configurations achieve values of 0.1716, 0.2960, and 0.2428, respectively. The best-Pearson configuration therefore achieves the highest agreement with the observed perturbation effects. The results show that the two evaluation metrics favour different predictions on this dataset. The centroid baseline more closely matches the observed distribution according to MMD, whereas the proposed model achieves stronger correlation with the observed perturbation effects according to Pearson Delta.

5.3 CrunchDAO Experiments

Table 9 reports results on the CrunchDAO dataset for three model configurations, selected respectively for the lowest MMD, the highest Pearson Delta, and the best overall trade-off between the two metrics, alongside the perturbed centroid baseline.

The best-MMD configuration reduces MMD to 0.0805, approximately a ninefold improvement over the baseline value of 0.7484, but achieves a Pearson Delta of only 0.0036, close to the baseline value of 0.0008. The best-Pearson configuration achieves the highest Pearson Delta of all configurations, at 0.0229, nearly a thirtyfold improvement over the baseline, but its MMD of 0.7281 is close to the baseline and more than nine times higher than that of the best-MMD configuration. The best-overall configuration achieves a MMD of 0.0954, comparable to the best-MMD configuration, together with a Pearson Delta of 0.0201, close to the highest value obtained by the best-Pearson configuration.

Table 7: Synthetic experiment results across intervention counts.

Interventions	Baseline MMD	Model MMD	Baseline Pearson	Model Pearson
2	0.5392	0.0250	-0.2901	0.9509
4	0.5297	0.0074	0.0151	0.9745
6	0.3912	0.0088	0.1392	0.9445
8	0.3963	0.0255	0.0043	0.8522
10	0.454	0.0611	-0.4742	0.8052

Table 8: Perturb-CITE-seq results: baseline and best model configurations by selection criterion.

Configuration	MMD	Pearson Delta
Baseline (perturbed centroid)	0.0105	0.0649
Best MMD	0.06027	0.1716
Best Pearson	0.09802	0.296
Best overall	0.06392	0.2428

These results indicate a clear trade-off between the two metrics on the CrunchDAO dataset: the hyperparameter configuration that minimises MMD does not coincide with the configuration that maximises Pearson Delta. The best-overall configuration demonstrates that a substantial improvement in MMD can be retained without sacrificing most of the gain in Pearson Delta, suggesting that the two objectives are not entirely in conflict and that a suitable choice of hyperparameters can approximate both simultaneously.

5.4 Comparison with the Baseline

The proposed method shows different relative performance across the two real-world datasets. On CrunchDAO, the proposed method improves both MMD and Pearson Delta relative to the perturbed centroid baseline. On Perturb-CITE-seq, the baseline achieves a lower MMD, whereas the proposed method achieves a higher Pearson Delta under all three evaluated configurations. On the synthetic data, the proposed method outperforms the baseline on both metrics whenever interventional data are available. Pearson Delta increases from 0.9801 with two interventions to 0.9993 with ten interventions, while MMD remains substantially below the baseline at every intervention count. The real-world results therefore do not show uniform superiority of the proposed method across metrics. Instead, they show that the model can improve agreement with perturbation-specific effects while distributional agreement depends on the dataset. This distinction is particularly visible in Perturb-CITE-seq, where the centroid baseline obtains the lower MMD but the proposed model obtains the higher Pearson Delta.

6 Discussion

The results in Section 5 show that the proposed least-squares estimator can predict perturbation responses, but its behaviour differs between synthetic and biological datasets. In the synthetic experiments, the model outperforms the centroid baseline for every evaluated intervention setup,

Table 9: CrunchDAO results: baseline and best model configurations by selection criterion.

Configuration	MMD	Pearson Delta
Baseline (perturbed centroid)	0.7484	0.0008
Best MMD	0.0805	0.0036
Best Pearson	0.7281	0.0229
Best overall	0.0954	0.0201

although performance does not improve monotonically with increasing intervention availability. On the real datasets, relative performance depends on both the dataset and metric. This section interprets these findings in relation to the research question, discusses limitations of the synthetic and real-data designs, and considers their implications for the trade-off between causal interpretability and predictive performance identified in the Related Work 3 section.

6.1 Interpretation of the Synthetic Results

The synthetic experiments evaluate prediction on entire intervention contexts that are excluded from parameter estimation. Across all evaluated setups, the OU model performs substantially better than the perturbed-centroid baseline on both MMD and Pearson Delta. Model MMD ranges from 0.0074 to 0.0611, compared with baseline values between 0.3912 and 0.5392. Pearson Delta ranges from 0.8052 to 0.9745, whereas the baseline remains between -0.4742 and 0.1392. The results therefore show that parameters estimated from observational data and a subset of interventions contain information that generalises to unseen intervention contexts.

Performance does not improve monotonically with the number of available interventions. The strongest results occur in the four-intervention setup, where MMD is 0.0074 and Pearson Delta is 0.9745, while both metrics are weaker in the eight- and ten-intervention setups. This pattern should not be interpreted as evidence that providing more interventions worsens prediction. The intervention-count settings are based on separately generated OU systems, so the underlying model parameters differ between settings. In addition, different perturbations are held out for testing. The observed differences therefore reflect both the amount of training intervention data and differences between the generated systems and test perturbations. The experiment demonstrates that the estimated model can predict unseen perturbations well, but it does not isolate the effect of intervention count on predictive performance.

6.2 Interpretation of the Perturb-CITE-seq Results

The Perturb-CITE-seq results show a discrepancy between distributional agreement and perturbation-effect agreement. The perturbed centroid baseline achieves an MMD of 0.0105, which is lower than the value of 0.06027 obtained by the best-MMD model configuration. According to MMD, the baseline therefore provides a closer match to the observed expression distribution. Pearson Delta produces the opposite ordering. The baseline achieves 0.0649, whereas the best-Pearson configuration reaches 0.2960. The proposed model therefore provides stronger agreement with the observed perturbation effects according to this metric. The best-overall configuration achieves a Pearson Delta of 0.2428 while maintaining an MMD of 0.06392. These results indicate that the two

metrics measure different aspects of prediction quality. MMD evaluates similarity between predicted and observed distributions, whereas Pearson Delta evaluates the correspondence between predicted and observed perturbation effects. The Perturb-CITE-seq results therefore do not support a claim of uniform predictive superiority, but they provide evidence that the estimated model captures information about perturbation effects that is not captured by the centroid baseline.

6.3 Interpretation of the CrunchDAO Results

On the CrunchDAO dataset, the proposed method outperforms the perturbed centroid baseline on MMD by close to an order of magnitude under the best-overall configuration, while also improving Pearson Delta relative to the baseline. This indicates that the estimated model can be used to predict held-out perturbations despite the fact that the training data contain interventions for only 122 of the 157 targeted genes in the challenge.

At the same time, the absolute values of Pearson Delta obtained on CrunchDAO remain modest, reaching at most 0.0229 under the best-Pearson configuration. This is considerably lower than the values obtained on synthetic data, where Pearson Delta consistently exceeded 0.98 once interventions were available. This gap suggests that gene-level rank agreement is more difficult to recover in a real, high-dimensional biological system than in the lower-dimensional synthetic systems used for validation, where the number of genes, the sparsity structure of the drift matrix, and the noise properties are all controlled by construction. Real gene regulatory networks are likely to violate at least some of the simplifying assumptions of the model, including linearity of the drift term and a diagonal diffusion matrix, which may limit the extent to which fine-grained, per-gene predictions can be recovered even when the overall distributional fit, as measured by MMD, is strong.

The trade-off observed between the best-MMD and best-Pearson configurations further suggests that the two metrics emphasise different aspects of predictive quality. MMD is sensitive primarily to the overall shape of the predicted distribution, whereas Pearson Delta depends on the precise ordering and magnitude of individual gene responses. The best-overall configuration demonstrates that a hyperparameter setting exists which retains most of the MMD improvement while approaching the highest observed Pearson Delta, indicating that the two objectives are not entirely in conflict, even though no single configuration in this evaluation was optimal for both metrics simultaneously.

6.4 Implications for the Research Gap

The theoretical work of (Salehkaleybar, 2026) establishes conditions under which the parameters of the OU model are identifiable up to a global scale from observational and partially interventional steady-state data. The contribution of this thesis is therefore not to establish whether partial intervention coverage is theoretically sufficient, but to evaluate whether the resulting estimated model also provides useful predictions of unseen perturbations according to MMD and Pearson Delta.

The experiments show that this predictive performance depends on the dataset and evaluation metric. In the synthetic experiments, the OU model achieves lower MMD and higher Pearson Delta than the perturbed-centroid baseline for every evaluated held-out intervention setting. On CrunchDAO, the proposed method also improves both metrics under the best-overall configuration.

On Perturb-CITE-seq, however, it improves Pearson Delta while the centroid baseline achieves lower MMD. These results show that an explicitly parameterised stochastic model can generalise to unseen perturbations, although the extent of this advantage depends on the dataset and on whether prediction quality is evaluated through distributional similarity or perturbation-effect correlation.

6.5 Limitations

Several limitations affect the interpretation of these results. First, the covariance-based terms of the loss function were omitted for both real-world datasets, owing to the computational cost of evaluating covariance residuals for high-dimensional gene expression data and abundant catching of noise. The resulting objective relies exclusively on the observational and interventional mean equations. This differs from the full formulation used in the synthetic experiments. Since the covariance contains additional information about the stochastic dynamics and regulatory structure, excluding it may reduce parameter identifiability and predictive accuracy on real-world data.

Second, the evaluation on CrunchDAO relies on a local ground-truth test set containing five held-out perturbations, since the complete evaluation set used by the official leaderboard was not made publicly available. Conclusions drawn from five perturbations are necessarily limited in statistical power, and the reported metrics may not generalise to the full, undisclosed evaluation set.

Third, the Perturb-CITE-seq evaluation is restricted to 61 genes from the IFN-γ condition. This restriction was adopted for computational feasibility and follows the experimental setup used by Sethuraman et al. (2023). Consequently, the results do not establish whether the method performs similarly when applied to the full transcriptome or to the other experimental conditions in the original dataset.

Fourth, the synthetic experiments assume a data-generating process that matches the model exactly: a linear OU process with a diagonal diffusion matrix and interventions that remove incoming regulatory edges exactly as specified in Section 4.4. This alignment between the assumed and true generative model is unlikely to hold for real gene regulatory networks, which may exhibit nonlinear regulatory relationships, non-diagonal noise structure, or intervention effects that are not fully captured by row deletion in the drift matrix. The synthetic results should therefore be interpreted as an upper bound on achievable performance under favourable, well-specified conditions, rather than as a direct predictor of performance on biological data.

7 Conclusion

This thesis investigated whether a linear stochastic model of gene regulation can be estimated from partial intervention data and used to predict unseen perturbations. A least-squares estimator for Ornstein–Uhlenbeck dynamics was implemented following the framework of (Salehkaleybar, 2026). The Lyapunov objective was retained for synthetic systems, while the covariance terms were omitted for the high-dimensional biological datasets where their computation or estimation was infeasible.

In the synthetic experiments, intervention contexts were divided into fixed training and held-out sets. The OU model outperformed the perturbed-centroid baseline on both MMD and Pearson Delta for every evaluated intervention setup. Its best performance occurred with four available interventions, with an MMD of 0.0074 and Pearson Delta of 0.9745, while performance was weaker at larger intervention counts. Because the intervention-count setups were generated from different OU systems and used different held-out contexts, this pattern does not demonstrate that additional interventions reduce predictive performance. It instead shows that the estimated dynamics can generalise to unseen interventions under the controlled synthetic model.

On CrunchDAO, the model outperformed the baseline on both metrics under the best-overall configuration. On Perturb-CITE-seq, it improved Pearson Delta but not MMD. These findings indicate that an interpretable OU model can support unseen perturbation prediction under partial intervention coverage, while its empirical advantage depends on the dataset, evaluation metric, and degree to which the OU assumptions describe the biological system. Overall, the results support the practical relevance of causal stochastic modelling, but do not establish uniform superiority.

7.1 Future work

Future work could reintroduce the covariance-based loss terms for large-scale datasets through a computationally efficient approximation and evaluate whether the additional covariance information improves prediction on datasets such as CrunchDAO. Extending the model to larger gene sets would also allow the effect of dimensionality on parameter estimation and prediction to be investigated.

Future work should also include comparisons with a broader range of perturbation-prediction models. Currently, the method is compared only with the perturbed centroid baseline. Evaluating methods from different categories, such as deep learning models like GEARS (Roohani et al., 2024) and transformer-based models such as State (Adduri et al., 2025) and Xpert (Guo et al., 2026), would provide a more comprehensive assessment of its performance.

References

Adduri, A., Gautam, D., Bevilacqua, B., Imran, A., Shah, R., Naghipourfar, M., Teyssier, N., Ilango, R., Nagaraj, S., Dong, M., Ricci-Tam, C., Carpenter, C., Subramanyam, V., Winters, A., Tirukkovular, S., Sullivan, J., Plosky, B., Eraslan, B., Youngblut, N., ... Roohani, Y. (2025). *Predicting cellular responses to perturbation across diverse contexts with state*, bioRxiv.

Bhatia, S., & Kleinjan, D. A. (2014). Disruption of long-range gene regulation in human genetic disease: A kaleidoscope of general principles, diverse mechanisms and unique phenotypic consequences. *Hum. Genet.*, *133*(7), 815–845.

Brouillard, P., Lachapelle, S., Lacoste, A., Lacoste-Julien, S., & Drouin, A. (2020). Differentiable causal discovery from interventional data. https://arxiv.org/abs/2007.01754

Cao, J., Spielmann, M., Qiu, X., Huang, X., Ibrahim, D. M., Hill, A. J., Zhang, F., Mundlos, S., Christiansen, L., Steemers, F. J., Trapnell, C., & Shendure, J. (2019). The single-cell transcriptional landscape of mammalian organogenesis. *Nature*, *566*(7745), 496–502.

Cheng, Y. H. H., Bohaczuk, S. C., & Stergachis, A. B. (2024). Functional categorization of gene regulatory variants that cause mendelian conditions. *Hum. Genet.*, *143*(4), 559–605.

Datlinger, P., Rendeiro, A. F., Boenke, T., Senekowitsch, M., Krausgruber, T., Barreca, D., & Bock, C. (2021). Ultra-high-throughput single-cell RNA sequencing and perturbation screening with combinatorial fluidic indexing. *Nat. Methods*, *18*(6), 635–642.

Datlinger, P., Rendeiro, A. F., Schmidl, C., Krausgruber, T., Traxler, P., Klughammer, J., Schuster, L. C., Kuchler, A., Alpar, D., & Bock, C. (2017). Pooled CRISPR screening with single-cell transcriptome readout. *Nat. Methods*, *14*(3), 297–301.

Dettling, P., Drton, M., & Kolar, M. (2024, January). On the lasso for graphical continuous lyapunov models. In F. Locatello & V. Didelez (Eds.), *Proceedings of the third conference on causal learning and reasoning* (pp. 514–550, Vol. 236). PMLR. https://proceedings.mlr.press/v236/dettling24a.html

Dettling, P., Homs, R., Améndola, C., Drton, M., & Hansen, N. R. (2023). Identifiability in continuous lyapunov models. *SIAM Journal on Matrix Analysis and Applications*, *44*(4), 1799–1821. https://doi.org/10.1137/22M1520311

Di Rocco, G., Trivisonno, A., Trivisonno, G., & Toietta, G. (2024). Dissecting human adipose tissue heterogeneity using single-cell omics technologies. *Stem Cell Res. Ther.*, *15*(1), 322.

Dunn, S.-J., Martello, G., Yordanov, B., Emmott, S., & Smith, A. G. (2014). Defining an essential transcription factor program for naıve pluripotency. *Science*, *344*(6188), 1156–1160.

Frangieh, C. J., Melms, J. C., Thakore, P. I., Geiger-Schuller, K. R., Ho, P., Luoma, A. M., Cleary, B., Jerby-Arnon, L., Malu, S., Cuoco, M. S., Zhao, M., Ager, C. R., Rogava, M., Hovey, L., Rotem, A., Bernatchez, C., Wucherpfennig, K. W., Johnson, B. E., Rozenblatt-Rosen, O., ... Izar, B. (2021). Multimodal pooled Perturb-CITE-seq screens in patient models define mechanisms of cancer immune evasion. *Nat. Genet.*, *53*(3), 332–341.

Freimer, J. W., Shaked, O., Naqvi, S., Sinnott-Armstrong, N., Kathiria, A., Garrido, C. M., Chen, A. F., Cortez, J. T., Greenleaf, W. J., Pritchard, J. K., & Marson, A. (2022). Systematic discovery and perturbation of regulatory genes in human T cells reveals the architecture of immune networks. *Nat. Genet.*, *54*(8), 1133–1144.

Gardiner, C. (2009). *Stochastic methods* (Vol. 4). Springer Berlin Heidelberg.

Guo, Y., Zhang, H., Hu, H., Wu, J., Cao, J., Hsieh, C.-Y., & Yang, B. (2026). Modelling drug-induced cellular perturbation responses with a biologically informed dual-branch transformer. *Nat. Mach. Intell.*

He, C., Zhang, J., Dahleh, M., & Uhler, C. (2025, July). *MORPH predicts the single-cell outcome of genetic perturbations across conditions and data modalities*, bioRxiv.

Kyono, Y., Kitzman, J. O., & Parker, S. C. J. (2019). Genomic annotation of disease-associated variants reveals shared functional contexts. *Diabetologia*, *62*(5), 735–743.

Lippe, P., Cohen, T., & Gavves, E. (2021). Efficient neural causal discovery without acyclicity constraints.

Lopez, R., Hütter, J.-C., Pritchard, J. K., & Regev, A. (2022). Large-scale differentiable causal discovery of factor graphs. https://arxiv.org/abs/2206.07824

Lorch, L., Krause, A., & Schölkopf, B. (2024, February). Causal modeling with stationary diffusions. In S. Dasgupta, S. Mandt, & Y. Li (Eds.), *Proceedings of the 27th international conference on artificial intelligence and statistics* (pp. 1927–1935, Vol. 238). PMLR. https://proceedings.mlr.press/v238/lorch24a.html

Norman, T. M., Horlbeck, M. A., Replogle, J. M., Ge, A. Y., Xu, A., Jost, M., Gilbert, L. A., & Weissman, J. S. (2019). Exploring genetic interaction manifolds constructed from rich single-cell phenotypes. *Science*, *365*(6455), 786–793.

Oh, Y., Lim, D., & Kim, S. (2024). Stable neural stochastic differential equations in analyzing irregular time series data. In B. Kim, Y. Yue, S. Chaudhuri, K. Fragkiadaki, M. Khan, & Y. Sun (Eds.), *International conference on learning representations* (pp. 38231–38262, Vol. 2024). https://proceedings.iclr.cc/paper_files/paper/2024/file/a61023ce36d21010f1423304f8ec49af-Paper-Conference.pdf

Paige, S. L., Plonowska, K., Xu, A., & Wu, S. M. (2015). Molecular regulation of cardiomyocyte differentiation. *Circ. Res.*, *116*(2), 341–353.

Pearl, J. (2009). *Causality* (2nd ed.). Cambridge University Press.

Peters, J., Janzing, D., & Scholkopf, B. (2017). *Elements of causal inference: Foundations and learning algorithms*. The MIT Press.

Ramirez, A. K., Dankel, S. N., Rastegarpanah, B., Cai, W., Xue, R., Crovella, M., Tseng, Y.-H., Kahn, C. R., & Kasif, S. (2020). Single-cell transcriptional networks in differentiating preadipocytes suggest drivers associated with tissue heterogeneity. *Nat. Commun.*, *11*(1), 2117.

Replogle, J. M., Saunders, R. A., Pogson, A. N., Hussmann, J. A., Lenail, A., Guna, A., Mascibroda, L., Wagner, E. J., Adelman, K., Lithwick-Yanai, G., Iremadze, N., Oberstrass, F., Lipson, D., Bonnar, J. L., Jost, M., Norman, T. M., & Weissman, J. S. (2022). Mapping information-rich genotype-phenotype landscapes with genome-scale perturb-seq. *Cell*, *185*(14), 2559–2575.e28.

Rohbeck, M., Clarke, B., Mikulik, K., Pettet, A., Stegle, O., & Ueltzhöffer, K. (2024, January). Bicycle: Intervention-based causal discovery with cycles. In F. Locatello & V. Didelez (Eds.), *Proceedings of the third conference on causal learning and reasoning* (pp. 209–242, Vol. 236). PMLR. https://proceedings.mlr.press/v236/rohbeck24a.html

Roohani, Y., Huang, K., & Leskovec, J. (2024). Predicting transcriptional outcomes of novel multigene perturbations with GEARS. *Nat. Biotechnol.*, *42*(6), 927–935.

Salehkaleybar, S. (2026). One intervention per component is enough: Towards identifiability in linear stochastic dynamics from steady state. https://openreview.net/forum?id=vjI9tsebtP

Schiebinger, G., Shu, J., Tabaka, M., Cleary, B., Subramanian, V., Solomon, A., Gould, J., Liu, S., Lin, S., Berube, P., Lee, L., Chen, J., Brumbaugh, J., Rigollet, P., Hochedlinger, K., Jaenisch, R., Regev, A., & Lander, E. S. (2019). Optimal-transport analysis of single-cell gene expression identifies developmental trajectories in reprogramming. *Cell*, *176*(4), 928–943.e22.

Sethuraman, M. G., Lopez, R., Mohan, R., Fekri, F., Biancalani, T., & Huetter, J.-C. (2023, 25–27 Apr). Nodags-flow: Nonlinear cyclic causal structure learning. In F. Ruiz, J. Dy, & J.-W. van de Meent (Eds.), *Proceedings of the 26th international conference on artificial intelligence and statistics* (pp. 6371–6387, Vol. 206). PMLR. https://proceedings.mlr.press/v206/sethuraman23a.html

Sethuraman, M. G., Lopez, R., Mohan, R., Fekri, F., Biancalani, T., & Hütter, J.-C. (2023). Nodags-flow: Nonlinear cyclic causal structure learning. https://arxiv.org/abs/2301.01849

Šimić, I., Veas, E., & Sabol, V. (2025). A comprehensive analysis of perturbation methods in explainable AI feature attribution validation for neural time series classifiers. *Sci. Rep.*, *15*(1), 26607.

Varando, G., & Richard Hansen, N. (2020, March). Graphical continuous lyapunov models. In J. Peters & D. Sontag (Eds.), *Proceedings of the 36th conference on uncertainty in artificial intelligence (uai)* (pp. 989–998, Vol. 124). PMLR. https://proceedings.mlr.press/v124/varando20a.html

Villaverde, A. F., Barreiro, A., & Papachristodoulou, A. (2016). Structural identifiability of dynamic systems biology models. *PLoS Comput. Biol.*, *12*(10), e1005153.

Viñas Torné, R., Wiatrak, M., Piran, Z., Fan, S., Jiang, L., Teichmann, S. A., Nitzan, M., & Brbić, M. (2025). Systema: A framework for evaluating genetic perturbation response prediction beyond systematic variation. *Nat. Biotechnol.*, 1–10.

Zhang, X., Pu, Y., Kawamura, Y., Loza, A., Bengio, Y., Shung, D. L., & Tong, A. (2024). Trajectory flow matching with applications to clinical time series modelling. In A. Globerson, L. Mackey, D. Belgrave, A. Fan, U. Paquet, J. Tomczak, & C. Zhang (Eds.), *Advances in neural information processing systems* (pp. 107198–107224, Vol. 37). Curran Associates, Inc. https://doi.org/10.52202/079017-3404

Zheng, X., Aragam, B., Ravikumar, P., & Xing, E. P. (2018). DAGs with NO TEARS: Continuous optimization for structure learning.

Zweig, A., Lin, Z., Azizi, E., & Knowles, D. A. (2026, 17–21 Aug). Towards identifiability of interventional stochastic differential equations. In E. Perković & D. Malinsky (Eds.), *Proceedings of the 42nd conference on uncertainty in artificial intelligence* (pp. 8349–8374, Vol. 337). PMLR. https://proceedings.mlr.press/v337/zweig26a.html