

Beyond Predictions in Neural ODEs: Identification and Interventions

Hananeh Aliee
Helmholtz Munich

hananeh.aliee@helmholtz-muenchen.de

Fabian Theis
Helmholtz Munich

fabian.theis@helmholtz-muenchen.de

Niki Kilbertus
Technical University of Munich & Helmholtz Munich

niki.kilbertus@helmholtz-muenchen.de

Abstract

Spurred by tremendous success in pattern matching and prediction tasks, researchers increasingly resort to machine learning to aid original scientific discovery. Given large amounts of observational data about a system, can we uncover the rules that govern its evolution? Solving this task holds the great promise of fully understanding the causal interactions and being able to make reliable predictions about the system’s behavior under interventions. We take a step towards such system identification for time-series data generated from systems of ordinary differential equations (ODEs) using flexible neural ODEs. Neural ODEs have proven successful in learning dynamical systems in terms of recovering observed trajectories. However, their efficacy in learning ground truth dynamics and making predictions under unseen interventions are still underexplored. We develop a simple regularization scheme for neural ODEs that helps in recovering the dynamics and causal structure from time-series data. Our results on a variety of (non)-linear first and second order systems as well as real data validate our method. We conclude by showing that we can also make accurate predictions under interventions on variables or the system itself.

1 Introduction

Many research areas increasingly embrace data-driven machine learning techniques not only for prediction, but also with the hope of leveraging data for original scientific discoveries. We may formulate the core task of an “automated scientist” as follows: *Given observational data about a system, identify the underlying rules governing it!* A key part of this quest is to determine how variables depend on each other. Putting this question at its center, causality is a natural candidate to surface scientific insights from data.

Great attention has been given to “static” settings, where each observable under consideration is a random variable whose distribution is given as a deterministic function of other variables—its causal parents. The structure of “which variables listen to what other variables” is typically encoded as a directed acyclic graph (DAG), giving rise to graphical structural causal models (SCM) (Pearl, 2009). SCMs have been successfully deployed both for inferring causal structure as well as estimating the strength of causal effects. However, in numerous scientific fields, we are interested in systems that jointly evolve over time with dynamics governed by differential equations, which cannot be easily captured in a standard SCM. In such interacting systems, the instantaneous derivative of a variable is a function of other variables (and their derivatives). We can thus interpret the variables (or derivatives) that enter this function as causal parents (Mooij et al., 2013). Unlike static SCMs, this accommodates cyclic dependencies and temporal co-evolution (Bongers et al., 2016).

The hope is that machine learning may sometimes be able to deduce true laws of nature purely from observational data, promising reliable predictions not only within the observed setting, but also under

interventions.¹ In this work we take a step towards system identification and causal structure inference from time-series data of a *fully observed* system that is assumed to *jointly evolve according to an ODE*. We use scalable and flexible neural ODE estimators (Chen et al., 2018) that allow for *a-priori unknown non-linear interactions*. We start with noise-free observations in continuous time, but also provide results for additive observation noise, irregularly sampled time points, and real gene expression data. Generally, recovering the ODE from a single observed solutions is an ill-posed problem. However, existing theory suggests (informally) that for certain classes of ODEs “most systems” will be identifiable. Motivated by these findings, we make the following contributions:

- We discuss potential regularizers to enforce sparsity in the number of causal interactions such that variables depend on few other variables, a common assumption in modern causal modeling Schölkopf et al. (2021).
- We develop a causal structure inference technique from time-series data called *causal neural ODE (C-NODE)* combining flexible neural ODE estimators with suitable regularization techniques. C-NODE works for non-linear ODEs with cyclic causal structure and identifies the full ODE in certain cases. Our results suggest that the proposed regularizer improves identifiability of the governing dynamics.
- We demonstrate the efficacy of C-NODE on a variety of low-dimensional (non-)linear first and second order systems, where it also makes accurate predictions under *unseen interventions*. On simulated autonomous, linear, homogeneous systems we show that C-NODE scales to tens of variables.
- Finally, C-NODE yields promising results for gene regulatory network inference on real, noisy, irregularly sampled single-cell RNA-seq data.

Related work. There is a large body of work on the discovery of causal DAGs within the static SCM framework (Heinze-Deml et al., 2018; Glymour et al., 2019; Vowels et al., 2021). One key idea for causal discovery on time-series data is based on *Granger causality*, where we attempt to forecast one time-series based on past values of others (Granger, 1988). We review the basic ideas and contemporary methods in Section 2.3. Typically, these methods also return an acyclic directed interaction model, though feedback of the forms $X_t \rightarrow X_{t+1}$ or $X_t \rightarrow Y_{t+1}$ and $Y_t \rightarrow X_{t+1}$ is allowed. Inferring Granger causality often relies on conditional independence tests (Malinsky & Spirtes, 2018) or score-based methods (Pamfil et al., 2020). Certain extensions of SCMs to cyclic dependence structures that retain large parts of the causal interpretation (Bongers et al., 2016) also allow for causal discovery of cyclic models (Lacerda et al., 2012; Hyttinen et al., 2012; Mooij et al., 2011). The framing of our work differs from the above in that they cannot model evolutions of instantaneously interacting systems and only aim at the causal structure instead of the system specifics.

Another line of research has explored connections between (asymptotic) equilibria of differential equations and (extended) SCMs that preserve behavior under interventions (Mooij et al., 2013; Bongers & Mooij, 2018; Rubenstein et al., 2018; Blom et al., 2020). Pfister et al. (2019) focus on recovering the causal dependence of a single target variable in a system of differential equations by leveraging data from multiple heterogeneous environments. Following earlier work (Dondelinger et al., 2013; Raue et al., 2015; Benson, 1979; Ballnus, 2019; Champion et al., 2019), they consider mass-action kinetics only taking into account linear combinations of up to degree one interactions of the target variable’s parents. They enforce sparsity by only allowing a fixed number of such terms to be non-zero. More broadly, parameter identifiability and estimation has been thoroughly investigated for discrete dynamical systems (McGoff et al., 2015), when the entire solution space is available (Grewal & Glover, 1976), or time-lag is required with no instantaneous interactions (Runge, 2018; Ye et al., 2015). Building on Takens (1981), the “convergent cross mapping method” (Sugihara et al., 2012) overcomes separability assumptions of Granger causality and has successfully been extended to identify causal structure (and sometimes systems) for chaotic, strongly non-linear, or time-lagged systems among others (Ye et al., 2015; Runge et al., 2019; De Brouwer et al., 2020). Current methods for ODE parameter estimation (with known parametric form) often deal with structural (Cobelli & Distefano, 1980) and practical (e.g., partial observability (Raue et al., 2009)) non-identifiability via empirical uncertainty analysis (Raue et al., 2015). For example, Sindy (Brunton et al., 2016), a popular sparse regression method for identification of nonlinear systems, explicitly assumes sparsity in a set of candidate basis functions. This poses a limitation on the scalability of the method as the dictionary size grows combinatorially in the number of variables that are

¹An automated scientist would undoubtedly be more powerful if it can interact with the system and perform experiments. By focusing on the purely observational setting, we avoid an ad-hoc specification of which experiments can be conducted and adhere to current settings, where algorithms are not granted direct access to real-world interventions.

allowed to interact and nested non-linearities have to be added explicitly to the dictionary. Hence, Brunton et al. (2016) acknowledge that Sindy fails to recover the dynamics already for a system of seven variables and that allowing for “a broader function search space is an important area of current and future work”.

In contrast to the works above, we assume neither a semantically meaningful pre-specified parametric form of the ODEs with a small set of parameters, nor the existence of equilibria. We consider fully observed, non-delay ODEs (no time-lag, only instantaneous interactions) with observations from a single environment. Our focus is on scalability to many variables with unknown non-linearities and leverage neural networks as flexible, yet efficiently learnable, function approximators.² Finally, we aim at fully identifying the ODE system, not only the causal structure, to be able to make predictions under interventions with a special focus on the multivariate case, i.e., systems of ODEs with sparse dependency structures. Note that the causal structure is implied by the ODE system, hence ODE identification is strictly harder than causal structure identification in our setting. In most applications, we consider the causal structure as an informative byproduct of our attempt at ODE system identification.

Vorbach et al. (2021); Massaroli et al. (2021); Bellot et al. (2021) aim at interpreting the inner workings of NODEs. Our main difference to Vorbach et al. (2021); Massaroli et al. (2021) is that they do not go beyond predictive performance and disregard whether the true system has been learned. Vorbach et al. (2021); Bellot et al. (2021) do not discuss the behavior of the system under interventions.

This research has potential applications in diverse fields including biology (Pfister et al., 2019) (e.g., gene-regulatory network inference (Matsumoto et al., 2017; Qiu et al., 2020)), robotics (Murray et al., 1994; Kipf et al., 2018), and economics (Zhang, 2005).

2 Setup and Background

Assume we observe the temporal evolution of n real-valued variables $X^* : [a, b] \rightarrow \mathbb{R}^n$ on a continuous time interval $a < b$, such that X^* solves the system of ODEs³

$$\begin{aligned} \dot{X} &= f^*(X, t), \text{ for } f^* \in \mathcal{F} \text{ where} \\ \mathcal{F} &:= \{f : \mathbb{R}^n \times [a, b] \rightarrow \mathbb{R}^n \mid f \text{ uniformly Lipschitz-continuous in } X \text{ and continuous in } t\}. \end{aligned} \quad (1)$$

That is, we observe a single solution trajectory of some ODE (determined by) $f^* \in \mathcal{F}$.

Our **main goal** is the following identification task: Given X^* , identify $f^* \in \mathcal{F}$. (2)

This is the inverse problem of “solving an ODE”, for which the celebrated Picard-Lindelöf theorem guarantees the existence and uniqueness of a solution of the *initial value problem* (IVP) $\dot{X} = f(X, t)$, $X(a) = x_0$ for all $f \in \mathcal{F}$ and $x_0 \in \mathbb{R}^n$ on an interval $t \in (a - \epsilon, a + \epsilon)$ for some $\epsilon > 0$. We remark that higher-order ODEs, in particular second-order ODEs $\ddot{X} = f(X, \dot{X}, t)$, can be reduced to first-order systems via $U := (X, \dot{X}) \in \mathbb{R}^{2n}$, $\dot{U} = (\dot{X}, \ddot{X}) = (U_2, f(U, t))$.⁴ Hence, it suffices to continue our analysis for first-order systems.

2.1 Causal Interpretation

In SCMs causal relationships are typically described by directed parent-child relationships in a DAG, where the causes (parents) of a variable X_i are denoted by $pa(X_i) \subset X$. For ODEs an analogous relationship can be described by which variables “enter into f_i ”. Formally, we define the **causal parents of X_i in system f** , denoted by $pa_f(X_i)$, as follows: $X_j \in pa_f(X_i)$ if and only if there exist $x_1, \dots, x_{j-1}, x_{j+1}, \dots, x_n \in \mathbb{R}$ such that $f_i(x_1, \dots, x_{j-1}, \bullet, x_{j+1}, \dots, x_n) : \mathbb{R} \rightarrow \mathbb{R}$ is not constant. This notion analogously extends to second

²Local minima are always a concern in high-dimensional non-convex optimization (training), but orthogonal to our work.

³We use X^* for the real observed function and X for a generic function. We refer to components X_i as the observed variables of interest. Similarly, f^* is the ground truth system and f a generic one. Lower case letters denote observations at a fixed time, e.g., $x_0 = X(t = 0)$.

⁴Second order systems require $X(a)$ and $\dot{X}(a)$ as initial values for a unique solution. In practice, when only X^* is observed, we assume that we can infer $\dot{X}^*(a)$, either from forward finite differences or during NODE training, see Section 2.2. Any higher-order ODE can iteratively be reduced to a first order system.

and higher order equations by defining $pa_f(X_i)$ as the variables X_j for which any (higher order) derivative of X_j enters f_i . Thereby, identifying f^* in eq. (2) also yields the causal structure—it is an immediate yet informative byproduct.

One of the key advantages of causal models compared to merely predictive ones is that they enable us to make predictions about hypothetical interventions not in the training data. In the ODE setting, different types of interventions can be conceived of. We will focus on the following types of interventions.

- **Variable interventions:** For one or multiple $i \in \{1, \dots, n\}$, we fix $X_i := c_i, f_i := 0$ and replace every occurrence of X_i in the remaining f_j with c_i (for some constant(s) $c_i \in \mathbb{R}$). We interpret these interventions as externally clamping certain variables to a fixed value.
- **System interventions:** We replace one or multiple f_i with \tilde{f}_i . Here we can further distinguish between *causality preserving* system interventions in which the causal parents remain unchanged, that is $pa_f(X_i) = pa_{\tilde{f}}(X_i)$ for all i , and others.

As an illustration, consider an ODE describing the positions of masses in a spring-mass system. A variable intervention could amount to externally keeping one mass at a fixed point in space. A system intervention could describe changing the stiffness of some of the springs. We will analyze variable interventions in (non-linear) ODEs and system interventions primarily in linear settings with interpretable system parameters like in the chemical reaction example.

2.2 (Neural) Ordinary Differential Equations

In neural ODEs (NODE) a machine learning model (often a neural network with parameters θ) is used to learn the function $f_\theta \approx f$ from data (Chen et al., 2018). Starting from the initial observation $X^*(a)$, an explicit iterative ODE solver is applied to predict $X^*(t)$ for $t \in (a, b]$ using the current derivative estimates from f_θ . The parameters θ are then updated via backpropagation on the mean squared error between predictions and observations. We mostly build on an augmented variant called SONODE that also works for second order systems and estimates the initial values $X^*(a)$ in an end-to-end fashion (Norcliffe et al., 2020).

Recently, NODEs have been extended to irregularly-sampled timesteps (Rubanova et al., 2019), stochastic DEs (Li et al., 2020; Oganessian et al., 2020), partial DEs (Sun et al., 2019), Bayesian NODEs (Dandekar et al., 2020), and delay ODEs (Zhu et al., 2021). While these extensions could benefit our method, we use vanilla SONODE to disentangle the performance of our method from tuning the underlying NODE method. As discussed extensively in the literature, NODEs can outperform traditional ODE parameter inference techniques in terms of reconstruction error, especially for non-linear f (Chen et al., 2018; Dupont et al., 2019). A subsequent advantage over previous methods is that we need not pre-suppose a parameterization of f in terms of a small set of semantically meaningful parameters.

Recently, a number of regularization techniques has been proposed for NODEs, where NODEs are viewed as infinite depth limits of residual neural networks (typically for classification) and there is no single true underlying dynamic law (Finlay et al., 2020; Kelly et al., 2020; Ghosh et al., 2020; Pal et al., 2021; Grathwohl et al., 2019). We emphasize that these existing techniques regularize the number of model evaluations to improve efficiency in learning one out of many possible dynamics yielding good performance on a downstream predictive task. They are unrelated to our regularizer, which aims at identifying a single true underlying dynamical system from time series data. Our regularizer targets neither the number of function evaluations, nor sparsity in the weights of the neural network directly (like pruning), but the number of causal dependencies. Finally, another alternative to train neural networks that depend on few inputs, regularizing input gradients (Ross et al., 2017b; Ross & Doshi-Velez, 2017; Ross et al., 2017a) does not scale to high-dimensional regression tasks.

2.3 Granger Causality

Granger causality is a classic method for causal discovery in time series data that primarily exploits the directionality of time (Granger, 1988). Informally, a time series X_i *Granger causes* another time series X_j if predicting X_j becomes harder when excluding the values of X_i from a universe of all time series. Assuming

that X is stationary, multivariate Granger causality analysis usually fits a vector autoregressive model

$$X(t) = \sum_{\tau=0}^k W^{(\tau)} X(t - \tau) + E(t), \quad (3)$$

where $E(t) \in \mathbb{R}^n$ is a Gaussian random vector and k is a pre-selected maximum time lag. We seek to infer the $W^{(\tau)} \in \mathbb{R}^{n \times n}$ from $X(t)$. In this setting, we call X_i a Granger cause of X_j if $|W_{i,j}^{(\tau)}| > 0$ for some $\tau \in \{0, \dots, k\}$. We need to ensure that $W^{(0)}$ encodes an acyclic dependence structure to avoid circular dependencies at the current time. Pamfil et al. (2020) then estimate the parameters in eq. (3) via

$$\min_W \left\| X(t) - \sum_{\tau=0}^k W^{(\tau)} X(t - \tau) \right\|_2 + \xi \|W^{(0)}\|_{1,1} + \rho \|W^{\setminus 0}\|_{1,1}, \quad (4)$$

where $W = (W^{(\tau)})_{\tau=0}^k$, $W^{\setminus 0} = (W^{(\tau)})_{\tau=1}^k$, and $\|\cdot\|_{1,1}$ is the element-wise ℓ_1 norm, which is used to encourage sparsity in the system. In addition, to ensure that the graph corresponding to $W^{(0)}$ interpreted as an adjacency matrix is acyclic, a smooth score encoding ‘‘DAG-ness’’ proposed by Zheng et al. (2018) is added with a separate regularization parameter. While extensions to nonlinear cases exist (Diks & Wolski, 2016), we primarily compare to `Dynotears` by Pamfil et al. (2020)—a choice motivated further in Appendix A.

3 Theoretical Considerations

First, we note that our main goal in eq. (2) is ill-posed. A solution exists by assumption, but it may not be unique.⁵ We provide a simple example of two different autonomous, linear, homogeneous systems that have at least one solution in common in Appendix B (where we provide all proofs).⁶ This means that the underlying system is *unidentifiable* from observational data.

Autonomous, linear, homogeneous systems are worthy a closer look:

$$\mathcal{F}_{\text{lin}} := \{f(X) = AX \mid A \in \mathbb{R}^{n \times n}\} \subset \mathcal{F}. \quad (5)$$

First, they are common models for chemical reactions or oscillating physical systems. Second, unlike for larger classes of ODEs, identifiability is reasonably well understood. Within \mathcal{F}_{lin} we can use A and f interchangeably. For such systems, Stanhope et al. (2014) developed beautiful graphical criteria for the identifiability of the system A given X^* , namely that the trajectory X^* is not confined to a proper subspace of \mathbb{R}^n . Qiu et al. (2022) recently showed that an equivalent characterization is that A has n distinct eigenvalues, which implies that almost all $A \in \mathbb{R}^{n \times n}$ are uniquely identifiable from X^* for almost all initial values $x_0 \in \mathbb{R}^n$.⁷ In this sense, all unidentifiable systems are non-generic and likely require some ‘‘fine-tuning’’ like our example in Appendix B, indicating that non-identifiability may not be a prevalent issue for the average case in practice. While these results largely extend to affine linear systems (Duan et al., 2020), little is known about identifiability in general non-linear systems (Miao et al., 2011). One may suspect non-identifiability to be a greater issue there, but highly non-linear or even chaotic systems are sometimes known to be identifiable (Takens, 1981; Sugihara et al., 2012; De Brouwer et al., 2020).

While these results are encouraging, we are particularly interested in ‘‘simple’’ interactions, which we capture by sparsity. That is, we assume that in natural systems each variable depends on only few other variables as causal parents (Schölkopf et al., 2021). Counting the number of parent-child relationships in a system f as $\|f\|_{\text{causal}} := \sum_{i=1}^n |pa_f(X_i)|$, we are thus interested only in possible ground truths f^* with small $\|f^*\|_{\text{causal}}$ (compared to n^2). For \mathcal{F}_{lin} , this amounts to matrices $A \in \mathbb{R}^{n \times n}$ with at most k non-zero entries. To the best

⁵This violates one of the three Hadamard properties for well-posed problems. We use the word ‘solution’ somewhat ambiguously and care must be taken not to confuse solutions to a given ODE or IVP (find X given $f \in \mathcal{F}$) and a solution of our main goal (finding $f^* \in \mathcal{F}$ given X^*).

⁶*Autonomy* means that f does not explicitly depend on time $f(X, t) = f(X)$. *Linear* systems are ones where f is linear in X , i.e., $f = A(t)X + b(t)$. *Homogeneous* systems are linear systems in which $b(t) = 0$.

⁷Specifically, this holds with respect to the Lebesgue measure on $\mathbb{R}^{n \times n}$ and \mathbb{R}^n . The result still holds for probability measures from most common random matrix ensembles such as the Gaussian orthogonal, the Wishart, or the Ginibre ensembles.

of our knowledge, it remains an open problem whether among such sparse matrices still almost all of them are identifiable from X^* (for almost all initial conditions x_0). Since sparse matrices are more likely to have repeated eigenvalues, the existing theory for dense matrices does not carry over to our setting. Hence, there is a gap between predictive performance (reconstruction error of X) and *identifying the governing system*, which is required to make *predictions under interventions*. NODEs perform well in terms of predictive performance, but theoretically they may do so by learning the “false” system. This is a key motivation for our empirical analysis in this work. In the following, we develop a method to identify such sparse systems from a single observed solution trajectory via regularization and assess identifiability empirically not only in the linear, but also non-linear case.

We first formulate our **regularized goal**: Given X^* , find $f \in \mathcal{F}$ such that X^* solves $\dot{X} = f(X, t)$ and $f \in \arg \min_{g \in \mathcal{F}} \|g\|$ for some measure of complexity $\|\cdot\| : \mathcal{F} \rightarrow \mathbb{R}_{\geq 0}$. Without knowing f^* a priori, $\|\cdot\|_{\text{causal}}$ is difficult to enforce as a complexity measure in practice. Instead, we approximate the requirement of “not being constant w.r.t. to an argument” in our definition of causal parents via the following regularizer

$$\|f\|_{\epsilon} := \sum_{i,j=1}^n \mathbf{1}\{\|\partial_j f_i\|_2 > \epsilon\} \quad \text{for some } \epsilon \geq 0, \quad (6)$$

where $\|\cdot\|_2$ is the L^2 norm on the Hilbert space of square-integrable real functions (with respect to the Lebesgue measure). For $A \in \mathcal{F}_{\text{lin}}$ this captures sparsity in the common sense $\|A\|_{\epsilon=0} = \sum_{i,j=1}^n \mathbf{1}\{A_{ij} \neq 0\} = \|A\|_{\text{causal}}$.⁸ There are still two hurdles to implementing $\|f\|_{\epsilon}$ in practice. (a) For the full non-linear case $f \in \mathcal{F}$ the L^2 -norms of partial derivatives are difficult to evaluate efficiently and accurately. (b) Even for $A \in \mathcal{F}_{\text{lin}}$, $\|A\|_{\epsilon}$ is not differentiable. For the linear case, non-differentiability of $\|A\|_{\epsilon}$ is typically overcome by enforcing sparsity via an entry-wise ℓ_1 norm $\|A\|_{1,1} = \sum_{i,j=1}^n |A_{ij}|$ as a penalty term, like in eq. (4). While this covers the linear case, in section 4 we develop a technique to partially overcome problem (a) in the non-linear case, by reformulating $\|f\|_{\epsilon}$ as an entry-wise ℓ_1 norm for a matrix derived from the parameters θ of the neural network f_{θ} approximating f .

Remarks. In the realm of neural ODEs one may be tempted to enforce sparsity in the neural network parameters θ directly. While this can be a sensible regularization scheme to improve generalization (Liebenwein et al., 2021), it does not directly translate into interpretable properties of the ODE f_{θ} . For fully connected neural networks even a sparse θ typically leads to dense input-output connections such that $\|f\|_{\epsilon} = n^2$.

Alternatively, one may train a separate neural network $f_{i,\theta_i} : \mathbb{R} \times \mathbb{R}^n \rightarrow \mathbb{R}$ with parameters θ_i for each component f_i . Stacking the outputs of all f_{i,θ_i} we can then train each network separately from the same training signal. For such a parallel setup we can enforce sparsity via $\|f_i\|_{\epsilon}^{\text{single}} := \sum_{j=1}^n \mathbf{1}\{\|\partial_j f_i\|_2 > \epsilon\}$ in the first layer parameters of each component neural network separately to “zero out” certain inputs entirely (Bellot et al., 2021). However, this differs from our sparsity measure $\|f\|_{\epsilon}$ in eq. (6). The latter allows large $\|f_i\|_{\epsilon}^{\text{single}}$ for *some* components f_i as long as the entire system is sparse. The parallel approach did not perform better in empirical experiments, but is computationally more expensive.

4 Method

Practical regularization. Let us write out f_{θ} explicitly as a fully connected neural network with L hidden layers parameterized by $\theta := (W^l, b^l)_{l=1}^{L+1}$, with l -th layer weights W^l and biases b^l

$$f_{\theta}(Y) = W^{L+1} \sigma(\dots \sigma(W^2 \sigma(W^1 Y + b^1) + b^2) \dots), \quad (7)$$

with element-wise non-linear activations σ . With this parameterization we now approximate our desired regularization $\|f_{\theta}\|_{\text{causal}}$ in terms of θ . A major drawback of the natural candidate $\|f_{\theta}\|_{\epsilon}$ from eq. (6) is that it is piece-wise constant in θ —an obstacle to gradient-based optimization. Instead, we aim at replacing $\|f\|_{\epsilon}$ with a differentiable surrogate. Ideally, we would like to use ℓ_1 regularization on the strengths of all input to output connections $j \rightarrow i$ in f_{θ} .

⁸ $\|A\|_{\epsilon=0}$ and therefore also $\|A\|_{\text{causal}}$ is not a norm; it violates the triangle-inequality.

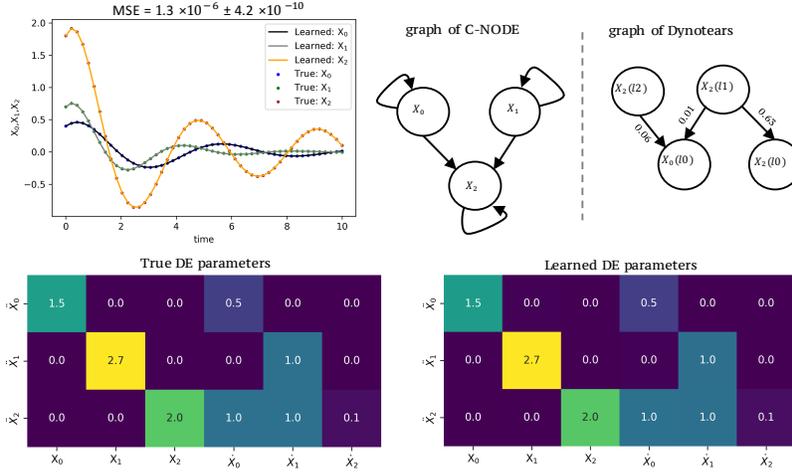


Figure 1: Exemplary second-order linear system.

- **The linear case.** Recall that for \mathcal{F}_{lin} it suffices to choose linear activation functions (specifically, $\sigma(x) = x$) and $b^l = 0$, such that $f_\theta(Y) = AY$ for some $A = W^{L+1} \cdot \dots \cdot W^1$.⁹ Hence, for $f_\theta \in \mathcal{F}_{\text{lin}}$, we can directly implement a continuous ℓ_1 surrogate of the desired regularizer in terms of θ

$$\|\theta\|_{\text{simple}}^{\text{lin}} := \|A\|_{1,1} = \|W^{L+1} \cdot \dots \cdot W^1\|_{1,1}. \quad (8)$$

We then have $X_j \in pa_{f_\theta}(X_i)$ if and only if $A_{ij} \neq 0$. When restricting ourselves to \mathcal{F}_{lin} , using $\sigma(x) = x$, $b^l = 0$ together with the above sparsity constraint is thus a viable and theoretically sound method. In this case we infer A directly from θ , i.e., we identify f , from which the causal structure follows.

- **The non-linear case.** In practice, we do not know whether $f^* \in \mathcal{F}_{\text{lin}}$ a priori. Thus we remain open to the possibility of non-linear f^* via non-linear activation functions σ . Then $\|\theta\|_{\text{simple}}^{\text{lin}}$ is not equivalent to $\|f_\theta\|_{\text{causal}}$ anymore, because we may have $X_j \notin pa_{f_\theta}(X_i)$ despite $A_{ij} \neq 0$. In this case, we use the absolute weights for the regularizer:

$$\|\theta\|_{\text{simple}}^{\text{non-lin}} := \| |W^{L+1}| \cdot \dots \cdot |W^1| \|_{1,1}. \quad (9)$$

While we can still have $\| |W^{L+1}| \cdot \dots \cdot |W^1| \|_{i,j} \neq 0$ even though $X_j \notin pa_{f_\theta}(X_i)$, $\| |W^{L+1}| \cdot \dots \cdot |W^1| \|_{i,j} = 0$ always implies $X_j \notin pa_{f_\theta}(X_i)$ for $b^l = 0$. Hence, using $\|\theta\|_{\text{simple}}^{\text{non-lin}}$ as a regularizer aims at minimizing an upper bound of $\|f_\theta\|_{\text{causal}}$. We show empirically that enforcing this upper bound of the desired regularizer serves as an effective inductive bias to enforce sparsity in the causal connections.

From now on, we will drop the superscript of $\|\theta\|_{\text{simple}}$ when it is clear from context.

Causal structure inference. While we could read off the causal structure directly from θ via A in the linear case, for non-linear f_θ we can validate our results via partial derivatives $\partial_j(f_\theta)_i$ over time, showing how each \dot{X}_i depends on each X_j . Following the reasoning of the regularizer $\|f_\theta\|_\epsilon$ in eq. (6), we then reconstruct the causal relationships via $X_j \in pa_{f_\theta}(X_i)$ if and only if $\sum_{k=1}^N |\partial_j f_{\theta,i}(t_k)| > \epsilon$ for the N observations at times $a = t_1 < \dots < t_N = b$ and some threshold $\epsilon > 0$.¹⁰ Thus we can still infer the causal structure in non-linear cases where evaluating whether our method identified the correct f^* is challenging (as we cannot compare parameters directly, but would have to compare a neural network to an analytically known function in symbolic form). The choice of ϵ is sensitive to the scale of the data, which we account for by normalizing data before training. Empirically we did not observe strong dependence of the inferred causal structure on the choice of ϵ for normalized data. A simpler method is available to determine the *absence* of causal dependencies in the non-linear setting: if the entry $\| |W^{L+1}| \cdot \dots \cdot |W^1| \|_{i,j}$ is (close to) zero, then $X_j \notin pa_{f_\theta}(X_i)$.

Summary. C-NODE adds a practical, differentiable regularizer $\lambda \|\theta\|_{\text{simple}}$ with a tuneable regularization parameter λ to the NODE loss to recover sparse dynamics corresponding to our regularized goal. Our regularizer captures $\|f\|_{\text{causal}}$ perfectly in the linear case and enforces minimization of an upper bound in the

⁹When biases are non-zero, we are in the realm of inhomogeneous, linear, autonomous systems.

¹⁰This is but one simple approach to estimate practically whether $\|\partial_j f_{\theta,i}\|_2 \neq 0$, i.e., whether the i -th output of the neural network f_θ depends on the j -th input. While a host of more sophisticated methods may be used here, we found this simple approach sufficient in our experiments.

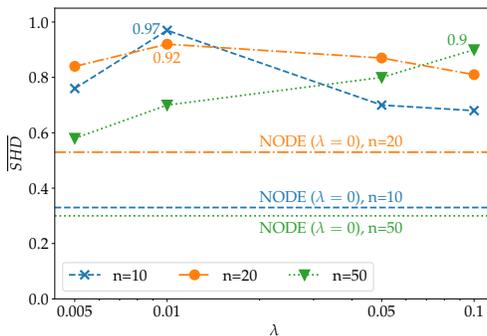


Figure 2: Dependence on the sparsity parameter λ . Dashed lines labeled $\lambda = 0$ refer to vanilla NODE.

Table 1: Experimental results using synthetic datasets with $n \in \{10, 20, 50\}$, varying noise level σ , and sampling irregularity (irr). SHD is the structural Hamming distance and $\overline{\text{SHD}} = 1 - \text{SHD}$.

σ	dim	irr = 0.0			irr = 0.2			irr = 0.5			irr = 0.7		
		$\overline{\text{SHD}}$	TPR	TNR									
0.0	10	0.97	0.95	0.98	0.98	0.97	0.98	0.93	0.97	0.91	0.92	0.93	0.91
	20	0.92	0.82	0.97	0.84	0.72	0.88	0.85	0.74	0.88	0.86	0.75	0.89
	50	0.90	0.71	0.92	0.85	0.67	0.87	0.92	0.69	0.96	0.93	0.70	0.96
0.05	10	0.88	0.81	0.92	0.87	0.80	0.92	0.86	0.84	0.88	0.67	0.48	0.87
	20	0.86	0.78	0.89	0.84	0.72	0.88	0.72	0.59	0.75	0.69	0.53	0.73
	50	0.91	0.64	0.94	0.90	0.69	0.93	0.90	0.67	0.93	0.89	0.65	0.93
0.1	10	0.78	0.68	0.74	0.71	0.68	0.71	0.65	0.77	0.58	0.50	0.60	0.51
	20	0.82	0.79	0.82	0.79	0.74	0.80	0.68	0.53	0.74	0.61	0.48	0.65
	50	0.86	0.65	0.90	0.89	0.59	0.93	0.87	0.61	0.91	0.86	0.68	0.89

non-linear case. We also devised a method to recover the causal structure from linear and non-linear f_θ . In Section 5 we show empirically that potentially remaining theoretical unidentifiability does not practically impede C-NODE from recovering f^* , allowing for accurate predictions under variable and system interventions. In Appendix D we discuss extensions for latent dynamics and data from heterogeneous environments.

5 Experiments

We now illustrate the robustness of our method in several case studies. The general principles readily extend to more complex model classes. Among several methods developed for causal inference from time-series data based on Granger causality (Tank et al., 2021; Hyvärinen et al., 2010; Runge et al., 2019; Amornbunchornvej et al., 2019), we compare to *Dynotears* (Pamfil et al., 2020), because it outperforms most competing methods in their evaluation. For system identification, we also compare to GroupLasso (Bellot et al., 2021) and PySINDy (Brunton et al., 2016; de Silva et al., 2020). Details on all parameter settings, evaluation, and implementation choices are provided in Appendix C.

Linear ODEs. We first study second-order, homogeneous, autonomous, linear ODEs

$$\ddot{X} = W_1 \dot{X} + W_2 X. \quad (10)$$

We begin with $n = 3$ and randomly chosen true weight matrices W_1^*, W_2^* from which we generate X^* using a standard ODE solver, see Appendix C. Figure 1 shows that our method not only accurately predicts X^* , but it also identifies W_1^*, W_2^* within a maximum absolute difference of 0.018. Thus the causal graph is also inferred correctly. The poor performance of *Dynotears* may be due to cyclic dependencies in W_1^*, W_2^* .

We extend these results to study the *scalability* of our method and its performance when the observations are *irregularly* sampled (a fixed fraction of observations is dropped uniformly at random) with *measurement noise* (additive zero-mean Gaussian noise with standard deviation σ). The data generation specifics for three synthetic datasets with 10, 20, and 50 variables are described in Appendix C. We evaluate the inferred causal graph using the structural hamming distance (SHD) (Lachapelle et al., 2019) for the fraction of wrongly predicted edges, the true positive rate (TPR), and the true negative rate (TNR). The results in

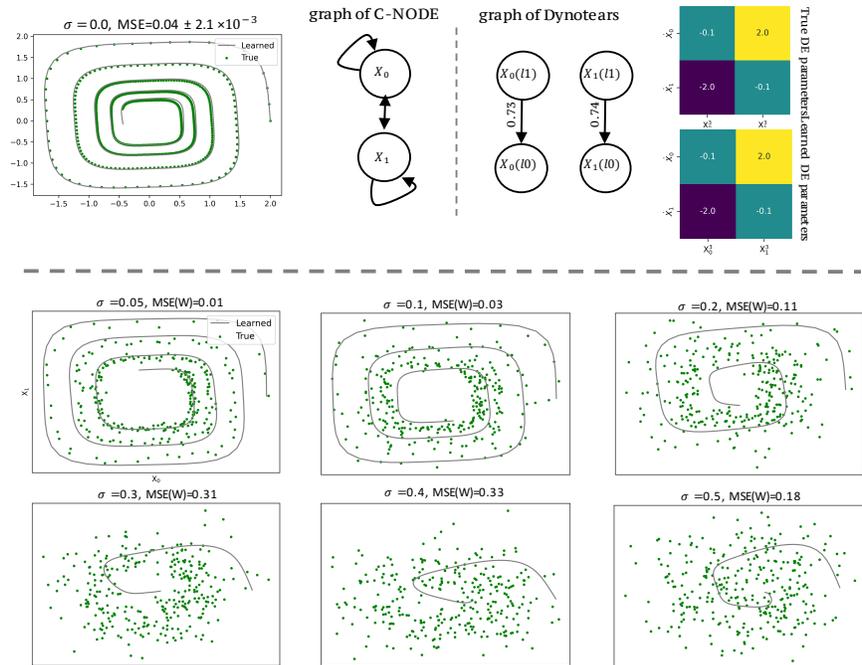


Figure 3: Results for the spiral ODE.

Table 1 show that C-NODE performs well for non-noisy data ($\sigma = 0$) and is robust to randomly removing samples from the observation. Accuracy drops with increasing noise levels, which is further exacerbated by sampling irregularities, suggesting improved robustness to observation noise as an interesting direction for future work. In Figure 2 we show the dependence of $\overline{\text{SHD}} = 1 - \text{SHD}$ on the regularization parameter λ for high-dimensional systems following eq. (10). C-NODE clearly outperforms vanilla NODE for causal structure inference with only moderate sensitivity to λ even though they achieve similar predictive accuracy. This indicates that for sparse systems unidentifiability may indeed be a problem (as hypothesized in section 3) in that the solution trajectory can be perfectly reconstructed also by “false” dense systems. This provides evidence that regularization is indeed crucial for recovering the correct sparse ground truth dynamics. In Figures 8 and 9, we provide further results and show that C-NODE also outperforms both vanilla NODE Chen et al. (2018), GroupLasso (Bellot et al., 2021) and PySINDy (de Silva et al., 2020).

A concrete application example for a common (synthetic) chemical reaction network of transcriptional gene dynamics is also provided in Appendix E.2.

Spiral ODEs. The spiral ODE model is given by

$$\dot{X}_0 = -\alpha X_0^3 + \beta X_1^3, \quad \dot{X}_1 = -\beta X_0^3 + \alpha X_1^3 \quad (11)$$

and features cyclic dependencies and self-loops. We follow the parameterization in Chen et al. (2018). While Dynotears fails to estimate the cyclic causal graph, Figure 3 shows that C-NODE infers the actual ODE parameters α, β and thus the causal structure correctly. Again, Figure 3 illustrates that predictive performance slowly degrades as observation noise levels increase raising the mean-squared error (MSE) of the inferred adjacency matrix substantially for higher variance noise. However, the deduced causal structure remains correct.

Lotka-Volterra ODEs. The Lotka-Volterra predator-prey model is given by the non-linear system

$$\dot{X}_0 = -\alpha X_0 - \beta X_0 X_1, \quad \dot{X}_1 = -\delta X_1 + \gamma X_0 X_1. \quad (12)$$

We use the same parameters as Dandekar et al. (2020) and ReLU activations for non-linearity. Figure 4 shows the excellent predictive performance of C-NODE (left). Because of the non-linearity, we resort to our non-linear causal structure inference method and show the partial derivatives $\partial_j f_{\theta,i}$ for the learned f_{θ} in Figure 4 (right). For example, from eq. (12) we know that $\partial \dot{X}_0 / \partial X_1 = -\beta X_0$ and indeed $\partial_1 f_{\theta,0}$ in

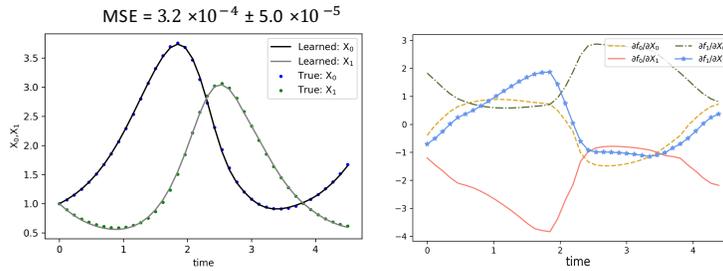


Figure 4: Results for the Lotka-Volterra example.

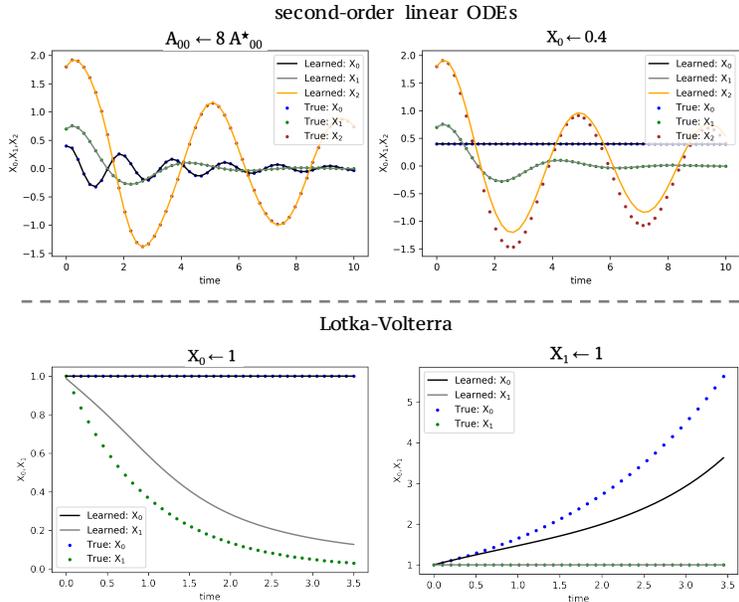


Figure 5: Predictions under interventions.

Figure 4 (right) resembles X_1 in Figure 4 (left) up to rescaling and a constant offset. Similarly, the remaining dependencies estimated from f_θ strongly correlate with the true dependencies encoded in eq. (12), giving us confidence that f_θ has indeed correctly identified f^* .

Interventions. To back up this claim, we assess whether we can predict the behavior of systems under interventions. We consider C-NODEs trained on observational data (without interventions) from Figures 1 (simple linear system) and 4 (Lotka-Volterra). We apply two types of interventions: (1) a system intervention replacing one entry of A^* via $\tilde{A}_{00} := 8A_{00}^*$ (for example, a temperature change that increases some reaction rate eightfold) in the linear setting, and (2) variable interventions $X_0 := 0.4$ in the linear as well as $X_0 := 1$ and $X_1 := 1$ in the non-linear Lotka-Volterra setting (for example, keeping the number of predators fixed via culling quotas and reintroduction). Figure 5 (top left) shows that C-NODE successfully predicts the linear system’s evolution under the system intervention. For the variable intervention in the linear setting (top right), X_1 correctly remains unaffected, while the new behavior of X_2 is predicted accurately.

In the Lotka-Volterra example, both variable interventions impact the other variable. Fixing either the predator or prey population should lead to an exponential increase or decay of the other, depending on whether the fixed levels can support higher reproduction than mortality. Figure 5 (bottom row) shows that our method correctly predicts an exponential decay (increase) of X_0 (X_1) for fixed $X_1 := 1$ ($X_0 := 1$) respectively. The quantitative differences between predicted and true values stem from small inaccuracies in the predicted parameters which amplify exponentially to seemingly large quantitative differences.

Real single-cell RNA-seq data. Finally, we apply C-NODE to learn gene-gene interactions. Gene (feature) interactions, also known as causal dependencies between genes, are often represented as a gene regulatory network (GRN) where nodes correspond to genes and directed edges indicate regulatory (or

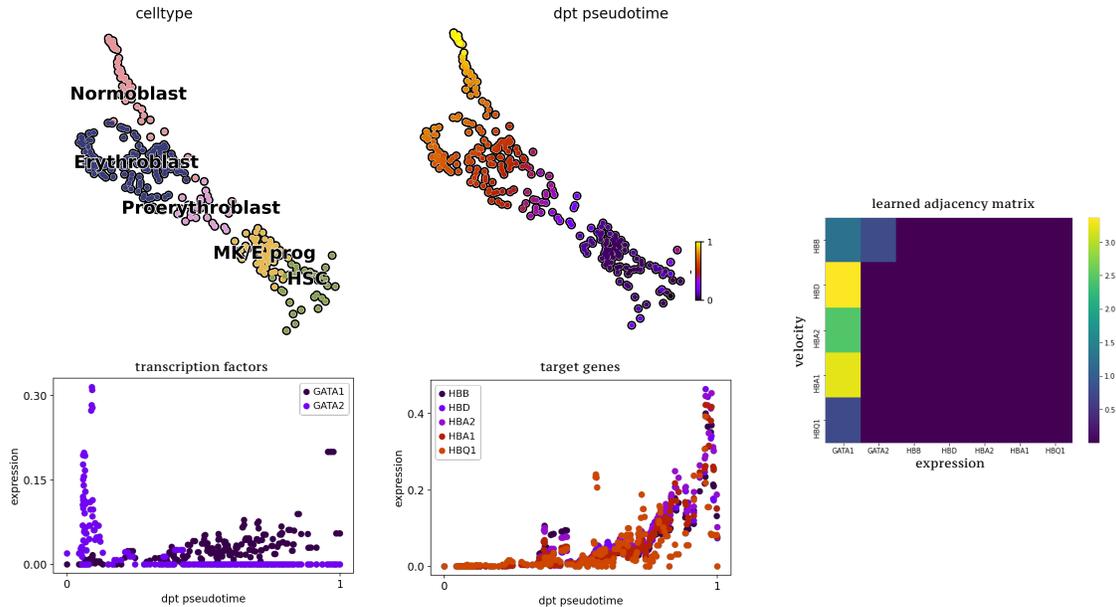


Figure 6: Gene regulatory network inference results using human bone marrow data with 7 genes.

causal) interactions between genes. GRN inference from observations is known to be an exceptionally difficult task Perkel (2022). The inferred GRN is expected to be sparse as regulatory genes known as transcription factors do not individually target all genes. Therefore, sparsity in the number of interactions is essential.

In this experiment, we first explore how pruning improves GRN inference from human hematopoiesis single-cell multiomics data Luecken et al. (2021) (GEO accession code: GSE194122). We select a branch of data in which hematopoiesis stem cells (HSCs) differentiate into Erythroid cells with a total of 280 cells (or samples). The count matrix is normalized to one million counts per cell. Figure 6 (top row) shows a UMAP representation (McInnes et al., 2018) of the data where each point corresponds to a cell colored by cell type (left) and an inferred continuous pseudotime (right). This pseudotime aims at identifying how far a cell has advanced in the differentiation process and is inferred via a diffusion map based manifold learning technique called *dpt* (Haghverdi et al., 2016) on 2,000 highly variable genes (Wolf et al., 2018; Bergen et al., 2020). We take measured gene expression levels over pseudotime t as our observations $X^*(t)$.

In this setting, domain knowledge asserts that GATA genes ($GATA1$, $GATA2$) regulate the expression of hemoglobin subunits (HBB , $HBA1$, $HBA2$, HBD , $HBQ1$) (Ding et al., 2010; Johnson et al., 2002; Suzuki et al., 2013; Shearstone et al., 2016). The normalized expression of genes related to these subunits over pseudotime is presented in Figure 6 (middle row). The expressions are scaled between 0 and 1 for each gene before training. We first apply C-NODE to these 7 genes with known ground truth. The bottom row of Figure 6 shows the row-wise normalized absolute values of the adjacency matrix inferred by C-NODE. Our approach properly assigns hemoglobin subunit changes to GATA genes, even though visually the hemoglobin target genes appear to be more correlated among themselves than with the GATA drivers.

We expect the regulatory elements and their target genes to be similar across species. To show that the previous results are stable across species, we next apply C-NODE to mouse single-cell RNA-seq data from (Pijuan-Sala et al., 2019) (GEO accession number: GSE87038) where blood progenitors similarly differentiate into Erythroid cells. Consistent with previous results, we also observe in Figure 11 that hemoglobin genes depend on GATA genes in Erythroid lineage (more details are discussed in Appendix E.3).

Finally to study the scalability of the proposed method and the effectiveness of the sparsity regularizer, we select 529 highly-variable genes from the whole human hematopoiesis data Luecken et al. (2021) and apply C-NODE on cells from the Erythroid lineage. Many of those genes are spurious for the Erythroid lineage as their expression does not change along the lineage. After training, the model selects 141 key genes as important with at least one interaction to other genes. Our first observation is that the important genes are ranked as highly variable only for the Erythroid lineage (Figure 14) which shows that C-NODE avoids using spurious features for prediction (Figure 7). Using the chromatin accessibility features available in the

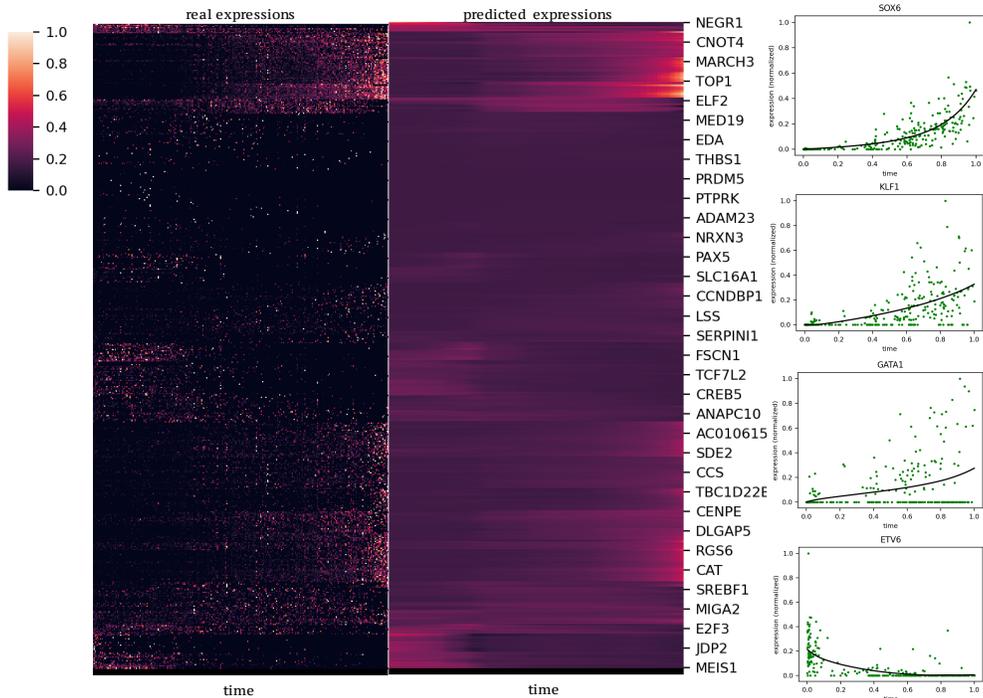


Figure 7: Prediction performance using human bone marrow data with 529 genes. Shown are the predictions for 141 important features (on left), and selected transcription factors (on right).

human immune cells datasets as reference, we also observe 22 regulatory genes known as transcription factors among those important genes. Assessed by the literature, all the 22 transcription factors are important for the differentiation of the Erythroid lineage. The expression and the prediction of four transcription factors are shown in Figure 7, left. Other transcription factors are listed in Table 2.

Since ground truth dynamics are not known for GRNs, validating the inferred interactions is challenging. In order to assess the biological relevance of the learned interactions, we perform Gene Set Enrichment Analysis (GSEA) via the enrichr method Chen et al. (2013) (discussed in Appendix E.3). The enrichment test also captures many relevant processes including hematopoietic stem cell differentiation, Erythrocyte differentiation, and some immune related signaling pathways. This provides evidence that indeed key regulatory processes for the differentiation have been captured by our model.

6 Conclusion

We proposed C-NODE, an approach to identification and causal structure learning of (sparse) ODE-based dynamical systems using Neural ODEs. First, we observe that even if a method achieves perfect predictive accuracy, it may not be able to predict the system’s behavior under interventions, as ODE identification is generally ill-posed. Therefore, a key focus of our work lies on the predominantly neglected issue of restricting the search space in meaningful ways to recover the underlying sparse system and causal structure.

We devised a simple and practical method to extract causal dependencies (including cyclic relationships) from a learned neural ODE derivative network. We then demonstrated that C-NODE performs well on causal structure identification for a wide variety of settings and further corroborated our findings by correctly predicting the effect of different forms of interventions targeting both the evolving variables as well as parameters of the governing ODE itself.

In our experiments we analyze C-NODE on synthetic and real-world gene regulatory data with varying numbers of variables, noise levels, and irregular sampling intervals. While unidentifiability indeed affects vanilla NODE, C-NODE still reliably infers sparse causal structures. Going forward, our results suggest an in-depth analysis of the conditions under which unidentifiability manifests itself in practice as a fruitful direction for future work. At the same time, extending C-NODE for successful hypotheses generation in high-

dimensional real datasets with stochasticity, delay, unobserved confounding, or heterogeneous environments is an exciting challenge for further research.

Given these limitations, we highlight that caution must be taken when informing consequential decisions, e.g., in healthcare, based on causal structures learned purely from observational data. At the same time, we hope that causal modelling of dynamical systems broadly and C-NODE in particular can be valuable tools for hypothesis-generation in various scientific domains to suggest promising experiments for in-depth follow up.

References

- Chainarong Amornbunchornvej, Elena Zheleva, and Tanya Y. Berger-Wolf. Variable-lag granger causality for time series analysis. *2019 IEEE International Conference on Data Science and Advanced Analytics (DSAA)*, Oct 2019. doi: 10.1109/dsaa.2019.00016. URL <http://dx.doi.org/10.1109/DSAA.2019.00016>.
- Benjamin Ballnus. *Development and Evaluation of Sampling-based Parameter Estimation Methods for Dynamic Biological Processes*. Dissertation, Technische Universität München, München, 2019.
- Alexis Bellot, Kim Branson, and Mihaela van der Schaar. Graphical modelling in continuous-time: consistency guarantees and algorithms using neural odes, 2021.
- M. Benson. Parameter fitting in dynamic models. *Ecological Modelling*, 6(2):97–115, 1979. ISSN 0304-3800. doi: [https://doi.org/10.1016/0304-3800\(79\)90029-2](https://doi.org/10.1016/0304-3800(79)90029-2). URL <https://www.sciencedirect.com/science/article/pii/0304380079900292>.
- Volker Bergen, Marius Lange, Stefan Peidli, F. Alexander Wolf, and Fabian J. Theis. Generalizing RNA velocity to transient cell states through dynamical modeling. *Nature Biotechnology*, 38(12), 2020. ISSN 15461696. doi: 10.1038/s41587-020-0591-3.
- Tineke Blom, Stephan Bongers, and Joris M Mooij. Beyond structural causal models: Causal constraints models. In *Uncertainty in Artificial Intelligence*, pp. 585–594. PMLR, 2020.
- Stephan Bongers and Joris M Mooij. From random differential equations to structural causal models: The stochastic case. *arXiv preprint arXiv:1803.08784*, 2018.
- Stephan Bongers, Jonas Peters, Bernhard Schölkopf, and Joris M Mooij. Theoretical aspects of cyclic structural causal models. *arXiv preprint arXiv:1611.06221*, 2016.
- Steven L Brunton, Joshua L Proctor, and J Nathan Kutz. Discovering governing equations from data by sparse identification of nonlinear dynamical systems. *Proceedings of the national academy of sciences*, 113(15):3932–3937, 2016.
- Kathleen Champion, Bethany Lusch, J. Nathan Kutz, and Steven L. Brunton. Data-driven discovery of coordinates and governing equations. *Proceedings of the National Academy of Sciences*, 116(45):22445–22451, 2019. ISSN 0027-8424. doi: 10.1073/pnas.1906995116. URL <https://www.pnas.org/content/116/45/22445>.
- Edward Y Chen, Christopher M Tan, Yan Kou, Qiaonan Duan, Zichen Wang, Gabriela Vaz Meirelles, Neil R Clark, and Avi Ma’ayan. Enrichr: interactive and collaborative HTML5 gene list enrichment analysis tool. *BMC Bioinformatics*, 14(1):128, 2013. ISSN 1471-2105. doi: 10.1186/1471-2105-14-128. URL <https://doi.org/10.1186/1471-2105-14-128>.
- Ricky T. Q. Chen, Yulia Rubanova, Jesse Bettencourt, and David K Duvenaud. Neural ordinary differential equations. In S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett (eds.), *Advances in Neural Information Processing Systems*, volume 31. Curran Associates, Inc., 2018. URL <https://proceedings.neurips.cc/paper/2018/file/69386f6bb1dfed68692a24c8686939b9-Paper.pdf>.
- Claudio Cobelli and Joseph J Distefano. Parameter and structural identifiability concepts and ambiguities: a critical review and analysis. *American Journal of Physiology-Regulatory, Integrative and Comparative Physiology*, 239(1):R7–R24, 1980.

-
- Raj Dandekar, Vaibhav Dixit, Mohamed Tarek, Aslan García-Valadez, and Chris Rackauckas. Bayesian neural ordinary differential equations, 2020. ISSN 23318422.
- Edward De Brouwer, Adam Arany, Jaak Simm, and Yves Moreau. Latent convergent cross mapping. In *International Conference on Learning Representations*, 2020.
- Brian de Silva, Kathleen Champion, Markus Quade, Jean-Christophe Loiseau, J. Kutz, and Steven Brunton. Pysindy: A python package for the sparse identification of nonlinear dynamical systems from data. *Journal of Open Source Software*, 5(49):2104, 2020. doi: 10.21105/joss.02104. URL <https://doi.org/10.21105/joss.02104>.
- Cees Diks and Marcin Wolski. Nonlinear Granger Causality: Guidelines for Multivariate Analysis. *Journal of Applied Econometrics*, 31(7), 2016. ISSN 10991255. doi: 10.1002/jae.2495.
- Ya Li Ding, Cheng Wang Xu, Zhi Dong Wang, Yi Qun Zhan, Wei Li, Wang Xiang Xu, Miao Yu, Chang Hui Ge, Chang Yan Li, and Xiao Ming Yang. Over-expression of EDAG in the myeloid cell line 32D: Induction of GATA-1 expression and erythroid/megakaryocytic phenotype. *Journal of Cellular Biochemistry*, 110(4), 2010. ISSN 10974644. doi: 10.1002/jcb.22597.
- Frank Dondelinger, Dirk Husmeier, Simon Rogers, and Maurizio Filippone. Ode parameter inference using adaptive gradient matching with gaussian processes. In Carlos M. Carvalho and Pradeep Ravikumar (eds.), *Proceedings of the Sixteenth International Conference on Artificial Intelligence and Statistics*, volume 31 of *Proceedings of Machine Learning Research*, pp. 216–228, Scottsdale, Arizona, USA, 29 Apr–01 May 2013. PMLR. URL <http://proceedings.mlr.press/v31/dondelinger13a.html>.
- X Duan, JE Rubin, and D Swigon. Identification of affine dynamical systems from a single trajectory. *Inverse Problems*, 36(8):085004, 2020.
- Emilien Dupont, Arnaud Doucet, and Yee Whye Teh. Augmented neural ODEs, 2019. ISSN 23318422.
- Chris Finlay, Jörn-Henrik Jacobsen, Levon Nurbekyan, and Adam M Oberman. How to train your neural ode: the world of jacobian and kinetic regularization, 2020.
- Arnab Ghosh, Harkirat Singh Behl, Emilien Dupont, Philip H. S. Torr, and Vinay Namboodiri. Steer: Simple temporal regularization for neural odes, 2020.
- Clark Glymour, Kun Zhang, and Peter Spirtes. Review of causal discovery methods based on graphical models. *Frontiers in genetics*, 10:524, 2019.
- C. W.J. Granger. Some recent development in a concept of causality. *Journal of Econometrics*, 39(1-2), 1988. ISSN 03044076. doi: 10.1016/0304-4076(88)90045-0.
- Will Grathwohl, Ricky T. Q. Chen, Jesse Bettencourt, Ilya Sutskever, and David Duvenaud. Ffjord: Free-form continuous dynamics for scalable reversible generative models. *International Conference on Learning Representations*, 2019.
- MS Grewal and Keith Glover. Identifiability of linear and nonlinear dynamical systems. *IEEE Transactions on automatic control*, 21(6):833–837, 1976.
- Laleh Haghverdi, Maren Büttner, F. Alexander Wolf, Florian Buettner, and Fabian J. Theis. Diffusion pseudotime robustly reconstructs lineage branching. *Nature Methods*, 13(10), 2016. ISSN 15487105. doi: 10.1038/nmeth.3971.
- Christina Heinze-Deml, Marloes H Maathuis, and Nicolai Meinshausen. Causal structure learning. *Annual Review of Statistics and Its Application*, 5:371–391, 2018.
- Morris Hirsch and Stephen Smale. Differential equations, dynamical systems, and linear algebra. *Pure and Applied Mathematics, Vol. 60*, 1974.

-
- Antti Hyttinen, Frederick Eberhardt, and Patrik O Hoyer. Learning linear cyclic causal models with latent variables. *The Journal of Machine Learning Research*, 13(1):3387–3439, 2012.
- Aapo Hyvärinen, Kun Zhang, Shohei Shimizu, and Patrik O. Hoyer. Estimation of a structural vector autoregression model using non-gaussianity. *Journal of Machine Learning Research*, 11(56):1709–1731, 2010. URL <http://jmlr.org/papers/v11/hyvarinen10a.html>.
- Kirby D. Johnson, Jeffrey A. Grass, Meghan E. Boyer, Carol M. Kiekhäfer, Gerd A. Blobel, Mitchell J. Weiss, and Emery H. Bresnick. Cooperative activities of hematopoietic regulators recruit RNA polymerase II to a tissue-specific chromatin domain. *Proceedings of the National Academy of Sciences*, 99(18):11760–11765, sep 2002. ISSN 0027-8424. doi: 10.1073/PNAS.192285999. URL <https://www.pnas.org/content/99/18/11760>.
- Eric Jones, Travis Oliphant, Pearu Peterson, et al. SciPy: Open source scientific tools for Python, 2001. URL <http://www.scipy.org/>.
- Jacob Kelly, Jesse Bettencourt, Matthew James Johnson, and David Duvenaud. Learning differential equations that are easy to solve. In *Neural Information Processing Systems*, 2020. URL <https://arxiv.org/abs/2007.04504>.
- Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization, 2017.
- Thomas Kipf, Ethan Fetaya, Kuan-Chieh Wang, Max Welling, and Richard Zemel. Neural relational inference for interacting systems, 2018.
- Gustavo Lacerda, Peter L Spirtes, Joseph Ramsey, and Patrik O Hoyer. Discovering cyclic causal models by independent components analysis. *arXiv preprint arXiv:1206.3273*, 2012.
- Sébastien Lachapelle, Philippe Brouillard, Tristan Deleu, and Simon Lacoste-Julien. Gradient-based neural dag learning. *arXiv preprint arXiv:1906.02226*, 2019.
- Xuechen Li, Ting-Kam Leonard Wong, Ricky TQ Chen, and David Duvenaud. Scalable gradients for stochastic differential equations. In *International Conference on Artificial Intelligence and Statistics*, pp. 3870–3882. PMLR, 2020.
- Lucas Liebenwein, Ramin Hasani, Alexander Amini, and Daniela Rus. Sparse flows: Pruning continuous-depth models. *arXiv preprint arXiv:2106.12718*, 2021.
- Malte D Luecken, Daniel Bernard Burkhardt, Robrecht Cannoodt, Christopher Lance, Aditi Agrawal, Hananeh Aliee, Ann T Chen, Louise Deconinck, Angela M Detweiler, Alejandro A Granados, Shelly Huynh, Laura Isacco, Yang Joon Kim, Dominik Klein, BONY DE KUMAR, Sunil Kuppasani, Heiko Lickert, Aaron McGeever, Honey Mekonen, Joaquin Caceres Melgarejo, Maurizio Morri, Michaela Müller, Norma Neff, Sheryl Paul, Bastian Rieck, Kaylie Schneider, Scott Steelman, Michael Sterr, Daniel J. Treacy, Alexander Tong, Alexandra-Chloe Villani, Guilin Wang, Jia Yan, Ce Zhang, Angela Oliveira Pisco, Smita Krishnaswamy, Fabian J Theis, and Jonathan M. Bloom. A sandbox for prediction and integration of DNA, RNA, and proteins in single cells. In *Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 2)*, 2021. URL <https://openreview.net/forum?id=gN35BGa1Rt>.
- Daniel Malinsky and Peter Spirtes. Causal Structure Learning from Multivariate Time Series in Settings with Unmeasured Confounding. *Proceedings of 2018 ACM SIGKDD Workshop on Causal Discovery*, 2018.
- Stefano Massaroli, Michael Poli, Jinkyoo Park, Atsushi Yamashita, and Hajime Asama. Dissecting neural odes, 2021.
- Hirotaaka Matsumoto, Hisanori Kiryu, Chikara Furusawa, Minoru S H Ko, Shigeru B H Ko, Norio Gouda, Tetsutaro Hayashi, and Itoshi Nikaido. SCODE: an efficient regulatory network inference algorithm from single-cell RNA-Seq during differentiation. *Bioinformatics*, 33(15):2314–2321, 04 2017. ISSN 1367-4803. doi: 10.1093/bioinformatics/btx194. URL <https://doi.org/10.1093/bioinformatics/btx194>.

-
- Kevin McGoff, Sayan Mukherjee, and Natesh Pillai. Statistical inference for dynamical systems: A review. *Statistics Surveys*, 9:209–252, 2015.
- Leland McInnes, John Healy, and James Melville. Umap: Uniform manifold approximation and projection for dimension reduction. *arXiv preprint arXiv:1802.03426*, 2018.
- Hongyu Miao, Xiaohua Xia, Alan S Perelson, and Hulin Wu. On identifiability of nonlinear ode models and applications in viral dynamics. *SIAM review*, 53(1):3–39, 2011.
- Joris M Mooij, Dominik Janzing, Tom Heskes, and Bernhard Schölkopf. On causal discovery with cyclic additive noise models. In J. Shawe-Taylor, R. Zemel, P. Bartlett, F. Pereira, and K. Q. Weinberger (eds.), *Advances in Neural Information Processing Systems*, volume 24. Curran Associates, Inc., 2011. URL <https://proceedings.neurips.cc/paper/2011/file/d61e4bbd6393c9111e6526ea173a7c8b-Paper.pdf>.
- Joris M Mooij, Dominik Janzing, and Bernhard Schölkopf. From ordinary differential equations to structural causal models: the deterministic case. *arXiv preprint arXiv:1304.7920*, 2013.
- Richard M. Murray, S. Shankar Sastry, and Li Zexiang. *A Mathematical Introduction to Robotic Manipulation*. CRC Press, Inc., USA, 1st edition, 1994. ISBN 0849379814.
- Alexander Norcliffe, Cristian Bodnar, Ben Day, Nikola Simidjievski, and Pietro Liò. On second order behaviour in augmented neural odes. *Advances in Neural Information Processing Systems*, 33:5911–5921, 2020.
- Alexander Norcliffe, Cristian Bodnar, Ben Day, Jacob Moss, and Pietro Liò. Neural ode processes, 2021.
- Viktor Oganessian, Alexandra Volokhova, and Dmitry Vetrov. Stochasticity in Neural ODEs: An Empirical Study, 2020. ISSN 23318422.
- Avik Pal, Yingbo Ma, Viral Shah, and Christopher Rackauckas. Opening the blackbox: Accelerating neural differential equations by regularizing internal solver heuristics, 2021.
- Roxana Pamfil, Nisara Sriwattanaworachai, Shaan Desai, Philip Pilgerstorfer, Konstantinos Georgatzis, Paul Beaumont, and Bryon Aragam. Dynotears: Structure learning from time-series data. In *International Conference on Artificial Intelligence and Statistics*, pp. 1595–1605. PMLR, 2020.
- Judea Pearl. *Causality*. Cambridge university press, 2009.
- J. M. Perkel. Smart software untangles gene regulation in cells. *Nature*, 609(7926), 2022. doi: 10.1038/d41586-022-02826-1.
- Niklas Pfister, Stefan Bauer, and Jonas Peters. Learning stable and predictive structures in kinetic systems. *Proceedings of the National Academy of Sciences*, 116(51):25405–25411, 2019.
- Blanca Pijuan-Sala, Jonathan A. Griffiths, Carolina Guibentif, Tom W. Hiscock, Wajid Jawaid, Fernando J. Calero-Nieto, Carla Mulas, Ximena Ibarra-Soria, Richard C.V. Tyser, Debbie Lee Lian Ho, Wolf Reik, Shankar Srinivas, Benjamin D. Simons, Jennifer Nichols, John C. Marioni, and Berthold Göttgens. A single-cell molecular map of mouse gastrulation and early organogenesis. *Nature*, 566(7745), 2019. ISSN 14764687. doi: 10.1038/s41586-019-0933-9.
- Xiaojie Qiu, Arman Rahimzamani, Li Wang, Bingcheng Ren, Qi Mao, Timothy Durham, José L. McFaline-Figueroa, Lauren Saunders, Cole Trapnell, and Sreeram Kannan. Inferring causal gene regulatory networks from coupled single-cell expression dynamics using scribe. *Cell Systems*, 10(3):265–274.e11, 2020. ISSN 2405-4712. doi: <https://doi.org/10.1016/j.cels.2020.02.003>. URL <https://www.sciencedirect.com/science/article/pii/S2405471220300363>.
- Xing Qiu, Tao Xu, Babak Soltanalizadeh, and Hulin Wu. Identifiability analysis of linear ordinary differential equation systems with a single trajectory. *Applied Mathematics and Computation*, 430:127260, 2022.

-
- Andreas Raue, Clemens Kreutz, Thomas Maiwald, Julie Bachmann, Marcel Schilling, Ursula Klingmüller, and Jens Timmer. Structural and practical identifiability analysis of partially observed dynamical models by exploiting the profile likelihood. *Bioinformatics*, 25(15):1923–1929, 2009.
- Andreas Raue, Bernhard Steiert, Max Schelker, Clemens Kreutz, Tim Maiwald, Helge Hass, Joep Vanlier, Christian Tönsing, Lorenz Adlung, Raphael Engesser, et al. Data2dynamics: a modeling environment tailored to parameter estimation in dynamical systems. *Bioinformatics*, 31(21):3558–3560, 2015.
- Andrew Ross, Isaac Lage, and Finale Doshi-Velez. The neural lasso: Local linear sparsity for interpretable explanations. In *Workshop on Transparent and Interpretable Machine Learning in Safety Critical Environments, 31st Conference on Neural Information Processing Systems*, volume 4, 2017a.
- Andrew Slavin Ross and Finale Doshi-Velez. Improving the adversarial robustness and interpretability of deep neural networks by regularizing their input gradients, 2017.
- Andrew Slavin Ross, Michael C. Hughes, and Finale Doshi-Velez. Right for the right reasons: Training differentiable models by constraining their explanations, 2017b.
- Yulia Rubanova, Ricky T.Q. Chen, and David Duvenaud. Latent ODEs for irregularly-sampled time series, 2019. ISSN 23318422.
- Paul K. Rubenstein, Stephan Bongers, Bernhard Schölkopf, and Joris M. Mooij. From deterministic ODEs to dynamic structural causal models. In Amir Globerson and Ricardo Silva (eds.), *Proceedings of the 34th Conference on Uncertainty in Artificial Intelligence (UAI-18)*. AUAI Press, August 2018. URL <http://auai.org/uai2018/proceedings/papers/43.pdf>.
- Jakob Runge. Causal network reconstruction from time series: From theoretical assumptions to practical estimation. *Chaos: An Interdisciplinary Journal of Nonlinear Science*, 28(7):075310, 2018.
- Jakob Runge, Peer Nowack, Marlene Kretschmer, Seth Flaxman, and Dino Sejdinovic. Detecting and quantifying causal associations in large nonlinear time series datasets. *Science advances*, 5(11):eaau4996, 2019.
- Bernhard Schölkopf, Francesco Locatello, Stefan Bauer, Nan Rosemary Ke, Nal Kalchbrenner, Anirudh Goyal, and Yoshua Bengio. Toward causal representation learning. *Proceedings of the IEEE*, 109(5):612–634, 2021.
- Jeffrey R. Shearstone, Olga Golonzhka, Apurva Chonkar, David Tamang, John H. Van Duzer, Simon S. Jones, and Matthew B. Jarpe. Chemical inhibition of histone deacetylases 1 and 2 induces fetal hemoglobin through activation of GATA2. *PLoS ONE*, 11(4), 2016. ISSN 19326203. doi: 10.1371/journal.pone.0153767.
- S Stanhope, Jonathan E Rubin, and David Swigon. Identifiability of linear and linear-in-parameters dynamical systems from a single trajectory. *SIAM Journal on Applied Dynamical Systems*, 13(4):1792–1815, 2014.
- George Sugihara, Robert May, Hao Ye, Chih-hao Hsieh, Ethan Deyle, Michael Fogarty, and Stephan Munch. Detecting causality in complex ecosystems. *science*, 338(6106):496–500, 2012.
- Yifan Sun, Linan Zhang, and Hayden Schaeffer. NeuPDE: Neural network based ordinary and partial differential equations for modeling time-dependent data, 2019. ISSN 23318422.
- Mikiko Suzuki, Maki Kobayashi-Osaki, Shuichi Tsutsumi, Xiaoqing Pan, Shin’ya Ohmori, Jun Takai, Takashi Moriguchi, Osamu Ohneda, Kinuko Ohneda, Ritsuko Shimizu, et al. Gata factor switching from gata 2 to gata 1 contributes to erythroid differentiation. *Genes to Cells*, 18(11):921–933, 2013.
- Floris Takens. Detecting strange attractors in turbulence. In *Dynamical systems and turbulence, Warwick 1980*, pp. 366–381. Springer, 1981.
- Alex Tank, Ian Covert, Nicholas Foti, Ali Shojaie, and Emily B Fox. Neural granger causality. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pp. 1–1, 2021. ISSN 1939-3539. doi: 10.1109/tpami.2021.3065601. URL <http://dx.doi.org/10.1109/TPAMI.2021.3065601>.

-
- Charles Vorbach, Ramin Hasani, Alexander Amini, Mathias Lechner, and Daniela Rus. Causal navigation by continuous-time neural networks, 2021.
- Matthew J Vowels, Necati Cihan Camgoz, and Richard Bowden. D'ya like dags? a survey on structure learning and causal discovery. *arXiv preprint arXiv:2103.02582*, 2021.
- F. Alexander Wolf, Philipp Angerer, and Fabian J. Theis. SCANPY: Large-scale single-cell gene expression data analysis. *Genome Biology*, 2018. ISSN 1474760X. doi: 10.1186/s13059-017-1382-0.
- Hao Ye, Ethan R Deyle, Luis J Gilarranz, and George Sugihara. Distinguishing time-delayed causal interactions using convergent cross mapping. *Scientific reports*, 5(1):1–9, 2015.
- W.-B. Zhang. *Differential equations, bifurcations, and chaos in economics*. World Scientific, 2005.
- Xun Zheng, Bryon Aragam, Pradeep K Ravikumar, and Eric P Xing. Dags with no tears: Continuous optimization for structure learning. *Advances in Neural Information Processing Systems*, 31, 2018.
- Qunxi Zhu, Yao Guo, and Wei Lin. Neural delay differential equations. *arXiv preprint arXiv:2102.10801*, 2021.

A Discretization of ODE Systems

The choice to compare primarily to `Dynotears` is motivated by the fact that the autoregressive model in eq. (3) can be viewed as a finite difference approximation to linear systems of ODEs that also incorporates sparsity. Since this recent method by Pamfil et al. (2020) demonstrates superior performance to most other competing methods for inferring causality from time series data, we chose it as a strong competitor to our method. For numerical treatment of derivatives in ODEs we often employ *finite differences*, where the derivative at time step t is approximated via differences of function values at slightly different time steps. For example, the backward finite difference for the first derivative is given by $(f(t) - f(t - h))/h$ for some $h > 0$. For time series data, we often use (without loss of generality) $h = 1$ and can thus approximate the derivative via $f(t) - f(t - 1)$. Similar approximations of higher order derivatives require more terms. Generally, to approximate the k -th derivative, information from $k + 1$ different time points is needed. Hence, finite combinations of the form

$$\sum_{\tau=0}^k W_{\tau} f(t - \tau), \quad (13)$$

which is also used by `Dynotears`, can in principle encode (linear combinations of) derivatives of f at time t up to order $k - 1$. Hence, time-series based methods such as `Dynotears` could in principle be expected to be able to model ODE systems correctly.

B Proofs

Here, we provide the example of unidentifiability mentioned in the main text. We start with some notation and known results. The space of general solutions of $\dot{X} = AX$ forms an n -dimensional sub vector space \mathcal{L} of all continuously differentiable functions from \mathbb{R} to \mathbb{R}^n .¹¹ For a basis ϕ_1, \dots, ϕ_n of \mathcal{L} , we call $\Phi_A = (\phi_1, \dots, \phi_n) : \mathbb{R} \rightarrow \mathbb{R}^{n \times n}$ a *fundamental system* of the differential equation, which in this setting is simply given by $\Phi_A(t) = e^{At}$. The solution to the IVP with $X(t_0) = x_0 \in \mathbb{R}^n$ is then $X(t) = \Phi_A(t)h$ for an h such that $x_0 = \Phi_A(t_0)h$, which exists since $\Phi_A(t)$ has full rank for all t . Hence, our main goal restricted to \mathcal{F}_{lin} reads: *Assuming $f^*(X, t) = A^*X$ and w.l.o.g. $a = 0$, can we uniquely identify A^* given X^* ?* In other words, does $e^{At}x_0 = e^{A't}x_0$ imply $A = A'$? This is not the case generally as we show with the following example. Non-uniqueness of A amounts to the existence of $B \in \mathbb{R}^{n \times n}$ different from A such that $\Phi_A h = \Phi_B g$ for some $g, h \in \mathbb{R}^n$ with $\Phi_A(0)h = \Phi_B(0)g = X(0)$. For $n = 2$, $X(0) := (1, 1)$ and

$$A = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}, \quad B = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix} \quad (14)$$

we have $X(t) = e^{At}h = e^{Bt}g = (e^t, e^t)$ for all $t \in \mathbb{R}$.

Remarks. First, we note that this result appears intuitive when writing out the two systems, showing that we can choose any constant initial value $X(0) = (\alpha, \alpha)$. Another way to understand unidentifiability intuitively for autonomous systems is to consider the flow vector field $f(X)$ of the ODE in which a particular solution is a single one-dimensional trajectory $X(t) \in \mathbb{R}^n$ following the vector field $f(X)$ at every point. An alternative system f' also solved by $X(t)$ only needs to preserve the vector field $f(X)$ along its trajectory ($f'|_{X(t)} = f|_{X(t)}$ for all $t \in [a, b]$), but could arbitrarily differ (without violating continuity assumptions) from $f(X)$ outside of $\{X(t) \mid t \in [a, b]\}$.

Finally, most results discussed here also hold for inhomogeneous, autonomous, linear systems ($f(X) = AX + b$ for $b \in \mathbb{R}^n$). In this case, the n -dimensional solution vector space is an affine subspace $\mathcal{L} + \tilde{X}$, where \mathcal{L} is the solution vector space of the homogeneous system and \tilde{X} is any specific solution of the inhomogeneous one.

Ill-posedness despite regularization. Here we show that our main goal is still ill-posed even under sensible sparsity regularizations. We begin again with the example from eq. (14), where the matrices A and B have equal values for $\|\cdot\|_{\text{causal}}$, $\|A\|_1$, and $\|A\|_{1,1}$. Let $C \in \mathbb{R}^{2 \times 2}$ be any system that has $X(t) = (e^t, e^t)$ as a solution for the initial value $X(0) = (1, 1)$ on all of \mathbb{R} . Since $X_1 = X_2 = \tilde{X}_1 = \tilde{X}_2$ on \mathbb{R} the coefficients in

¹¹The statements in this paragraph are proven in most textbooks on ODEs, e.g., see (Hirsch & Smale, 1974).

each row of C must sum up to 1. Hence, the minimum achievable value for $\|C\|_{1,1} = \sum_{i,j=1}^2 |C_{ij}|$ is 2, which is achieved by both A and B . Similarly, the minimum achievable value for $\|C\|_1 = \max_{j=1,2} \sum_{i=1}^2 |C_{ij}| = \max\{|C_{11}| + |C_{21}|, |C_{12}| + |C_{22}|\}$ under the row-unit-sum constraint is 1, which is also achieved by both A and B . Finally, in the linear case the minimum achievable $\|C\|_{\text{causal}} = \|C\|_{\epsilon=0} = \sum_{i,j=1}^2 \mathbf{1}\{C_{ij} \neq 0\}$ is also 1, again achieved by both A and B . Therefore, the systems A and B in eq. (14) are indeed among the minimum complexity solutions under all considered regularization schemes.

While other types of regularization may yield unique solutions (Tikhonov regularization for ill-posed problems), these typically clash with the explicit demand for sparsity in the system, with many relationships not just being weak, but desired to be non-existent. Our focus lies on sparsity enforcing regularization motivated by $\|\cdot\|_{\text{causal}}$ throughout. We also remark the close resemblance of our arguments to the fact that Ridge regularization yields unique solutions in linear regression whereas Lasso may not. In our example, consider $\|\cdot\|_2$ or $\|\cdot\|_{2,2}$ instead. In this case the minimum complexity system is *uniquely* given by $C_{ij} = 0.5$ for $i, j \in \{1, 2\}$. It is important to recognize though that this argument does *not* suffice to prove that our statement does not hold for these regularizers. It merely illustrates that our examples and proof techniques fail for these regularizers. Refining these unidentifiability results is an interesting direction for future work.

C Implementation Details

C.1 Synthetic Data Generation

We generate random datasets with 10, 20, and 50 variables and at most 200 observations. For each dataset, we first start with a random ground truth adjacency matrix and generate the discrete observations using off-the-shelf explicit numerical Runge-Kutta style ODE solvers (Jones et al., 2001). We then add zero-mean, fixed-variance Gaussian noise to each variable and observation independently. Finally, a percentage of samples (denoted by ‘irr’) is randomly dropped from the dataset to simulate irregular observation times.

C.2 Architecture and Training Procedure

All neural networks used in this work are fully connected, feed-forward neural networks. The initial velocities are predicted using a neural network with two hidden layers with 20 neurons each and tanh activations. The main architecture to infer velocities (or accelerations) also contains two hidden layers of sizes 20, 50, or 100 depending on the size of the input and ELU (or for some experiments linear) activation function. ELU and tanh were used because they allow for negative values in the ODE (Norcliffe et al., 2020). As an ODE solver, we use an explicit 5-th order Dormand-Prince solver commonly denoted by `dopri5`.

All models are optimized using Adam (Kingma & Ba, 2017) with an initial learning rate of 0.01. We use PyTorch’s default weight initialization scheme for the weights and set the regularization parameter λ for the L1 penalty to 0.01 in the 10, and 20-dimensional examples and to 0.1 for the 50-dimensional example. All models can be trained entirely on CPUs on consumer grade Laptop machines within minutes or hours. To compute the MSE uncertainty, the experiments in Figures 1,3 and 4 are run 10 times with random seeds. Finally, the presence of edges in the weight matrices was determined by thresholding absolute values at 0.05 for the synthetic datasets and 0.001 for the real dataset.

D Extensions

D.1 Measurement noise

Considering a deterministic version of dynamical systems with measurement noise, we have:

$$\tilde{X}(t) = X(t) + E(t). \tag{15}$$

Where, $X(t)$ is assumed to be governed by the ODE system and the noises E_i are assumed to be jointly independent with zero mean. We show that our causal inference model based on NODEs can be relatively resistant to measurement noise (see the example in Figure 3). However, we can still extend our approach

to latent-variable models for more complicated systems. In a general form, consider a (generative) model e which encodes the initial position \tilde{X}_0 into the latent variable Z_0 as follows:

$$\mathbf{Z}_0 = e(\tilde{\mathbf{X}}, \theta_e). \quad (16)$$

This is then used by the ODE function f and the ODE solver consequently:

$$\begin{aligned} \dot{Z} &= f(Z, t, \theta) \\ Z_0, \dots, Z_N &= \text{ODESolver}(f, \mathbf{Z}_0, (t_0, \dots, t_N)) \end{aligned} \quad (17)$$

The latent variables are then decoded as follows:

$$\hat{X}_0, \dots, \hat{X}_N = d(Z, \theta_d). \quad (18)$$

Such a latent-variable model is causally interpretable and can be integrated into our method, if we can learn \hat{X} as a function of \hat{Z} naturally from the model. While most of the extensions to NODEs for noisy and irregularly-sampled observations perform well with respect to the reconstruction accuracy, they are hardly interpretable due to their model complexity (Rubanova et al., 2019; Norcliffe et al., 2021).

D.2 Data from Heterogeneous Systems

Our algorithm could potentially also be extended to (noisy) observations generated from heterogeneous experiments. Following Pfister et al. (2019), in this case we treat the observations from each experiments as a time-series sample. Training the model with heterogeneous samples, we may hope to identify the causal model that is invariant across the experiments (Pfister et al., 2019). It is an interesting direction for future work to determine whether heterogeneous experiments allow us to overcome the unidentifiability results in Section 3.

E Extended Results

E.1 Synthetic Linear ODEs

The learned adjacency matrices using vanilla NODE Chen et al. (2018), GroupLasso (Bellot et al., 2021) and PySINDy (de Silva et al., 2020) are illustrated in Figure 8.

The SHD matrices using C-NODE are presented in Figure 9. S_{ij} in an SHD matrix represents the presence or absence of an edge between X_i and X_j including the sign of the effect (e.g., $S_{ij} = -1$ means $X_j \rightarrow X_i$ with a negative coefficient).

E.2 Chemical reaction networks

Here, we study a model of transcriptional dynamics which captures transcriptional induction and repression of unspliced precursor mRNAs $u(t)$ splicing into mature mRNAs $s(t)$ at rate β . The mature mRNAs eventually degrade with rate γ .

$$u = \alpha - \beta u, \quad s = \beta u - \gamma s, \quad (19)$$

where α is the reaction rate of the transcription. We assume β and γ to be constant and the transcription rate α to vary over time. The results in Figure 10 show that our method can successfully learn the structural graph as well as the ODE parameters while `Dynotears` fails on the same task. Note that in this case we add α as a variable in the system with a fixed, pre-specified time-dependence in such a way that it satisfies an ODE separately from u, s . Our method successfully identifies this structure where the evolution of α does not depend on u and s , but conversely, the derivatives of u and s depend on α . Encouraged by these synthetic results, we also tested our method on real single-cell gene expression data, described in Section 5.

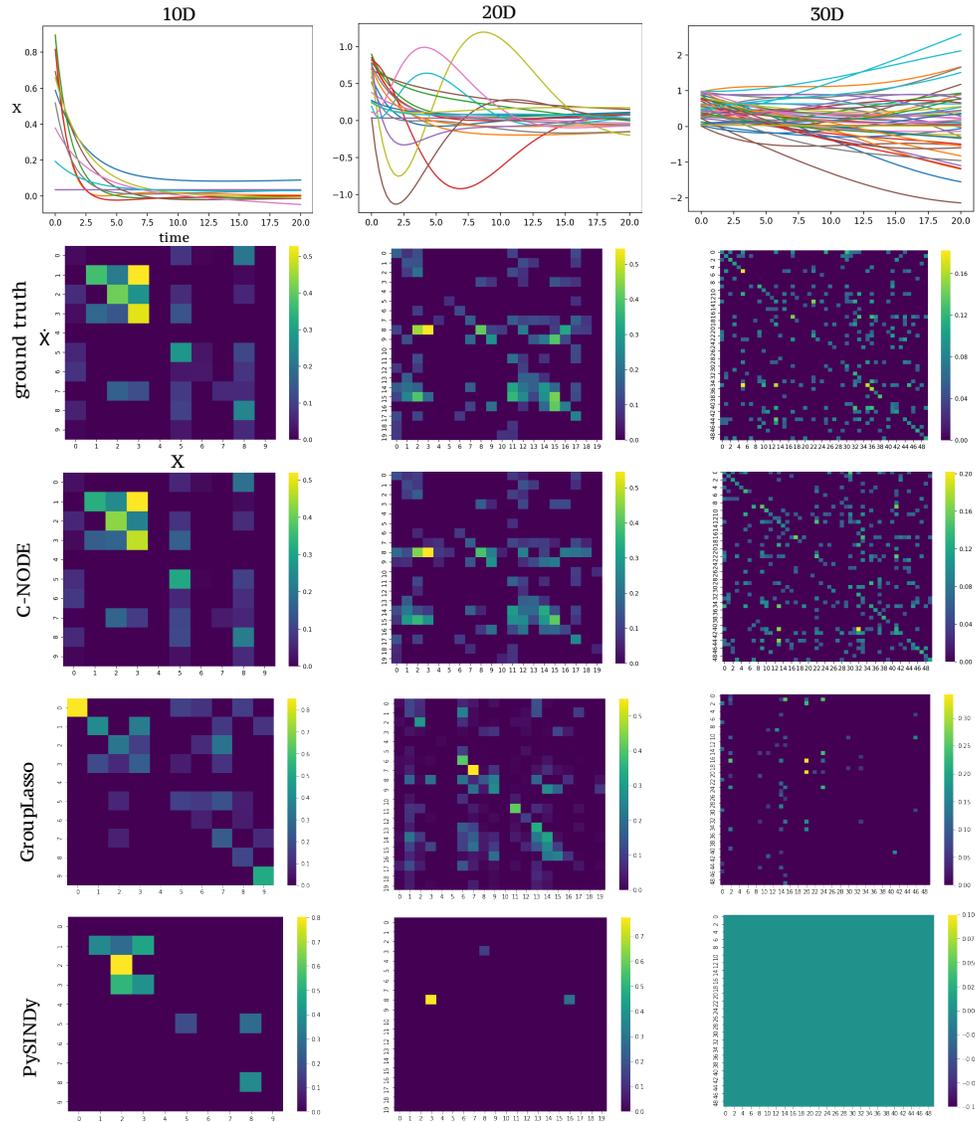


Figure 8: The synthetic datasets and the inferred adjacency matrices using C-NODE, GroupLasso (Bellot et al., 2021) and PySINDy (de Silva et al., 2020). Shown are the absolute values for the adjacency matrices referred to the regularly sampled ($\text{irr} = 0.0$) and non-noisy ($\sigma = 0$) setting.

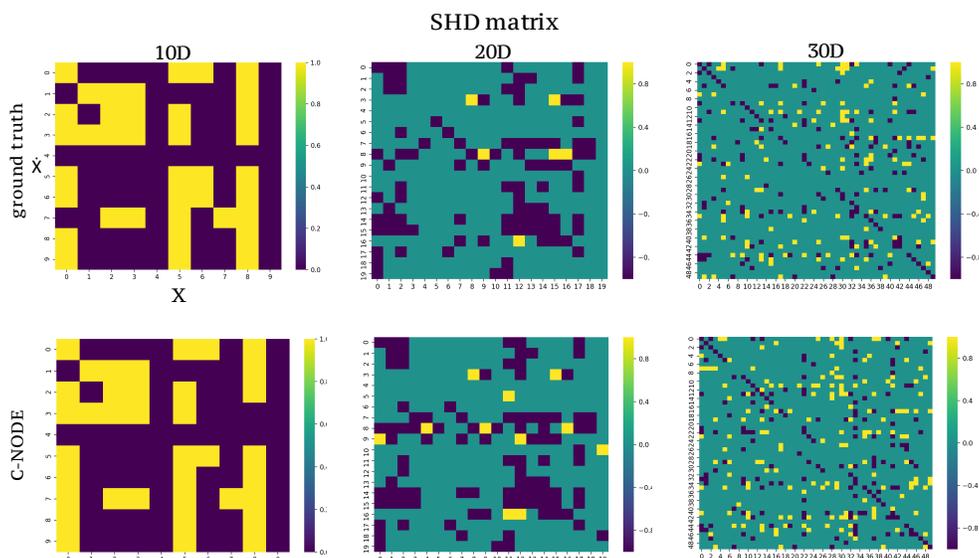


Figure 9: The true and the inferred $\overline{\text{SHD}}$ matrices for the synthetic datasets using C-NODE, related to Table 1.

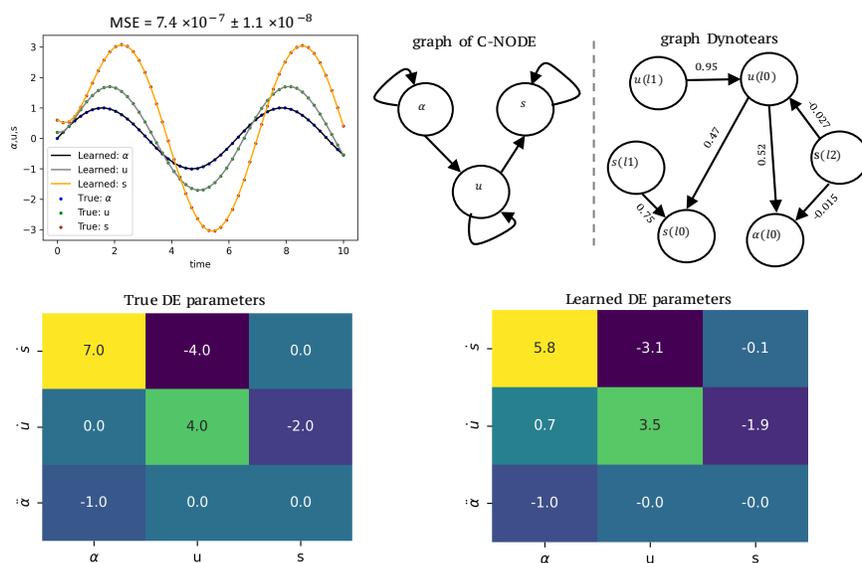


Figure 10: Example of a chemical reaction network modeling the transcriptional dynamics of a gene.

E.3 Gene Regulatory Network Inference

We apply C-NODE to real mouse single-cell RNA-seq data from (Pijuan-Sala et al., 2019) (GEO accession number: GSE87038). We select a branch of data in which blood progenitors differentiate into Erythroid cells with a total of 9,192 cells (or samples). The count matrix is normalized to one million counts per cell. We randomly subset 300 cells from all 9,192 as training data for C-NODE. We observe in Figure 11 that the inferred GRN using GATA and hemoglobin genes resembles the one in Figure 6.

In Figure 12 and Figure 13, we show additional results for GRN inference of single-cell mouse data. Figure 12 shows the predictions of C-NODE. The partial derivatives in Figure 13 indicate that despite non-linear

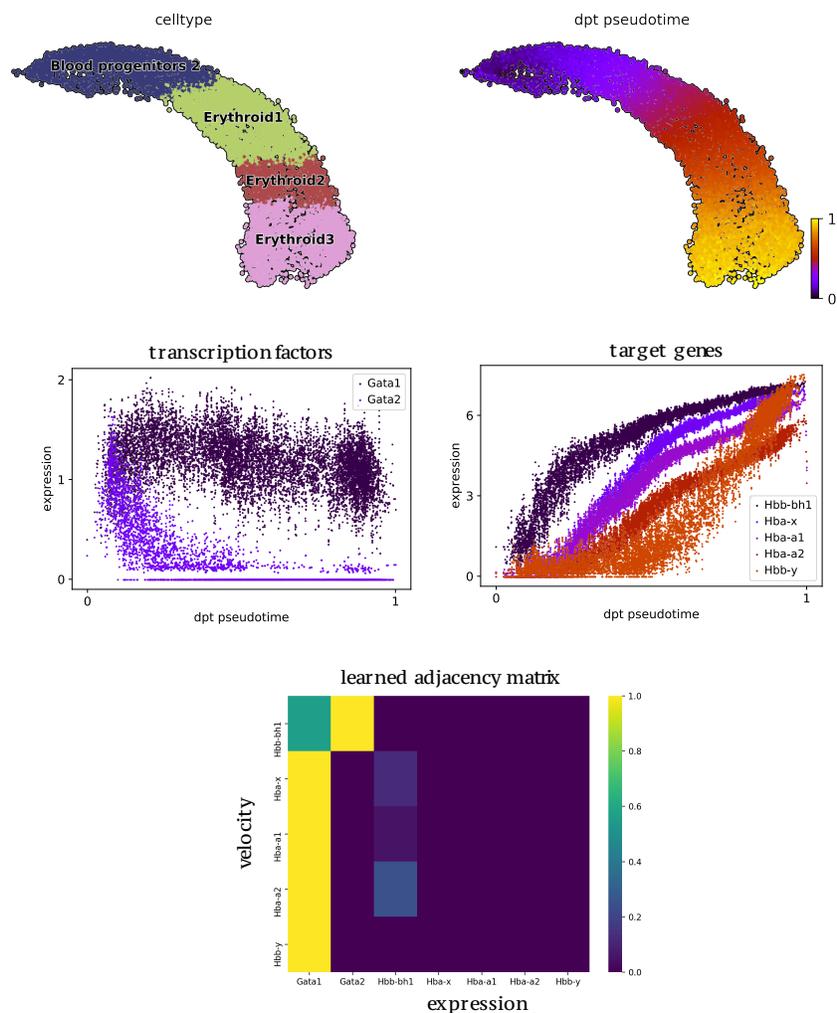


Figure 11: Gene regulatory network inference results using mouse dataset with 7 genes.

Table 2: The inferred TFs for Erythroids using C-NODE.

CTCF, E2F1, E2F3, E2F8, ETV6, GATA1, GFI1B, KLF1, KLF13, MAFG, MAZ, MXI1, MYBL2, NFE2L2, NFIA, SP4, RREB1, RUNX1, SOX4, SOX6, SREBF1, TAL1

activations, the associations are mostly linear for the target genes except for *Hbb-y*. This suggests that linear $f^* \in \mathcal{F}_{lin}$ may indeed be a decent approximation for certain gene regulatory networks.

For the large dataset shown in Figure 12, the gene set enrichment analysis uses a priori gene sets that involve in known biological pathways. For each list of dependent genes inferred using C-NODE, we then analyze whether the majority of genes in each pathway fall in the extremes of this list.

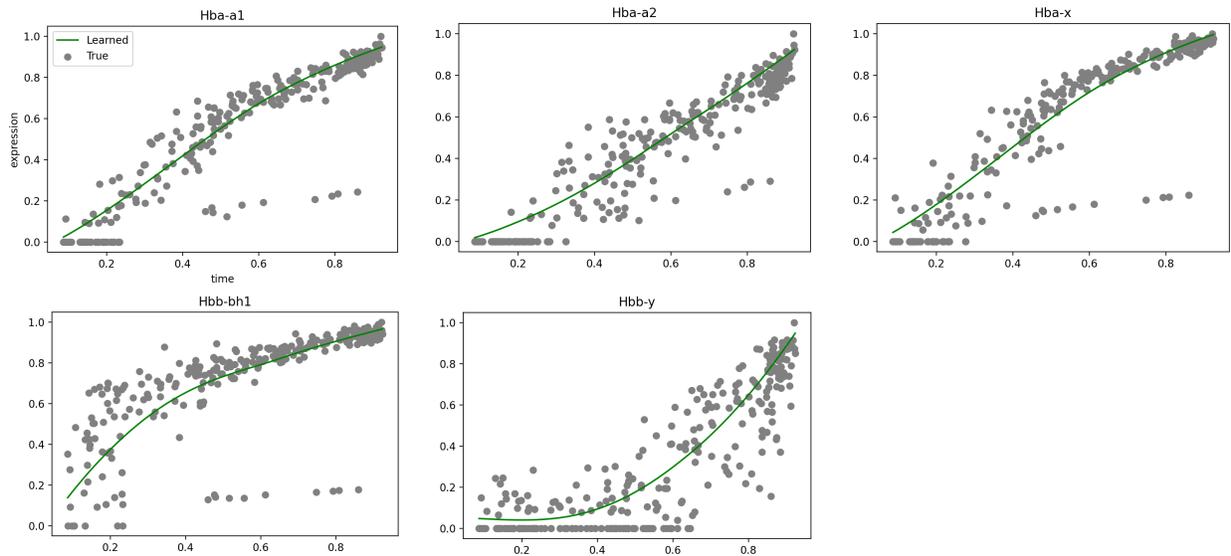


Figure 12: Predictions of the expression of the target genes together with the 300 samples used for training, related to Figure 11.

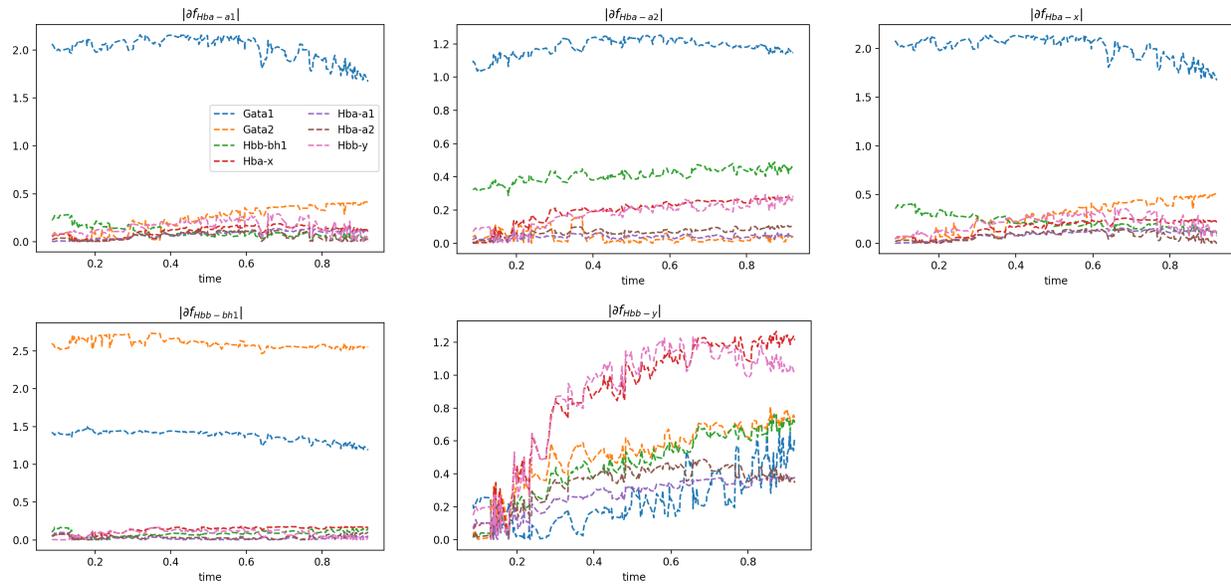


Figure 13: Gradients for the target genes, related to Figure 11.

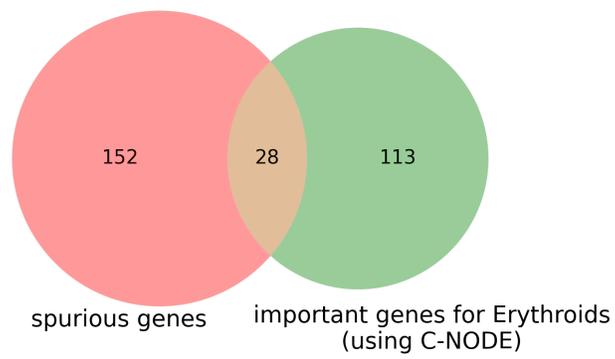


Figure 14: The important genes extracted using C-NODE are Erythroid specific and have negligible overlap with spurious genes.