

Continuous–Depth Neural Models for Dynamic Graph Prediction

Michael Poli^{1,*}, Stefano Massaroli^{2,*}, Clayton M. Rabideau³
Junyoung Park¹, Atsushi Yamashita², Hajime Asama², Jinkyoo Park¹

Abstract

We introduce the framework of continuous–depth *graph neural networks* (GNNs). *Neural graph differential equations* (Neural GDEs) are formalized as the counterpart to GNNs where the input–output relationship is determined by a *continuum* of GNN layers, blending discrete topological structures and differential equations. The proposed framework is shown to be compatible with static GNN models and is extended to dynamic and stochastic settings through hybrid dynamical system theory. Here, Neural GDEs improve performance by exploiting of the underlying dynamics geometry, further introducing the ability to accommodate irregularly sampled data. Results prove the effectiveness of the proposed models across applications, such as traffic forecasting or prediction in genetic regulatory networks.

1 Introduction

Introducing appropriate inductive biases on deep learning models is a well-known approach to improving sample efficiency and generalization performance (Battaglia et al., 2018). Graph neural networks (GNNs) represent a general computational framework for imposing such inductive biases when the problem structure can be encoded as a graph or in settings where prior knowledge about entities composing a target system can itself be described as a graph (Li et al., 2018b; Gasse et al., 2019; Sanchez-Gonzalez et al., 2018; You et al., 2019). GNNs have shown remarkable results in various application areas such as node classification (Zhuang and Ma, 2018; Gallicchio and Micheli, 2019), graph classification (Yan et al., 2018) and forecasting (Li et al., 2017; Wu et al., 2019b) as well as generative tasks (Li et al., 2018a; You et al., 2018). A different but equally important class of inductive biases is concerned with the type of temporal behavior of the systems from which the data is collected i.e., discrete or continuous dynamics. Although deep learning has traditionally been a field dominated by discrete models, recent advances propose a treatment of neural net-

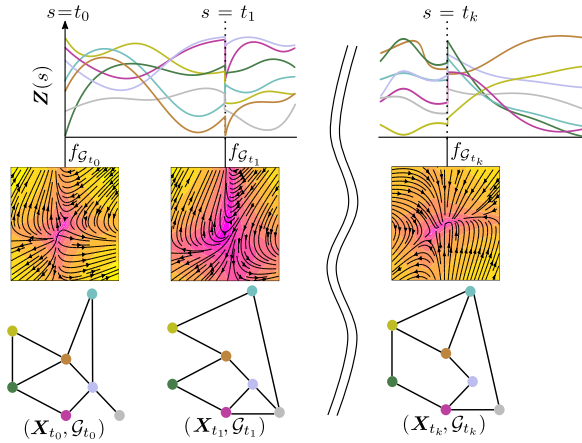


Figure 1: *Neural graph differential equations* (Neural GDEs) model vector fields defined on graphs, both in cases when the structure is fixed or changes in time, via a continuum of *graph neural network* (GNN) layers.

*Equal contribution. Author order was decided via coin flip. ¹KAIST. ²The University of Tokyo ³Syntensor. Corresponding author: Michael Poli, email: poli.m@kaist.ac.kr

works equipped with a continuum of layers (Weinan, 2017; Chen et al., 2018; Massaroli et al., 2020). This view allows a reformulation of the forward and backward pass as the solution of the initial value problem of an *ordinary differential equation* (ODE). The resulting *continuous–depth* paradigm has successfully guided the discovery of novel deep learning models, with applications in prediction (Rubanova et al., 2019; Greydanus et al., 2019), control (Du et al., 2020), density estimation (Grathwohl et al., 2018; Lou et al., 2020; Mathieu and Nickel, 2020), time series classification (Kidger et al., 2020), among others. In this work we develop and experimentally validate a framework for the systematic blending of differential equations and graph neural networks, unlocking recent advances in continuous–depth learning for non–trivial topologies.

● **Blending graphs and differential equations** We introduce the system–theoretic model class of *neural graph differential equations* (Neural GDEs), defined as ODEs parametrized by GNNs. Neural GDEs are designed to inherit the ability to impose relational inductive biases of GNNs while retaining the dynamical system perspective of continuous–depth models. A complete model taxonomy is carefully laid out with the primary objective of ensuring compatibility with modern GNN variants. Neural GDEs offer a grounded approach for the embedding of numerical schemes inside the forward pass of GNNs, in both the deterministic as well as the stochastic case.

● **Dynamic graphs** Additional Neural GDE variants are developed to tackle the spatio–temporal setting of dynamic graphs. In particular, we formalize general Neural Hybrid GDE models as *hybrid dynamical systems* (Van Der Schaft and Schumacher, 2000; Goebel et al., 2009). Here, the structure–dependent vector field learned by Neural GDEs offers a data–driven approach to the modeling of dynamical networked systems (Lu and Chen, 2005; Andreasson et al., 2014), particularly when the governing equations are highly nonlinear and therefore challenging to approach with analytical methods. Neural GDEs can adapt the prediction horizon by adjusting the integration interval of the differential equation, allowing the model to track evolution of the underlying system from irregular observations. The evaluation protocol for Neural GDEs spans several application domains, including traffic forecasting and prediction in biological networks.

2 Neural GDEs ●

We begin by introducing the general formulation. We then provide a taxonomy for Neural GDE models, distinguishing them into *static* and *spatio–temporal* variants.

2.1 General Framework

Without any loss of generality, the inter–layer dynamics of a residual graph neural network (GNN) may be represented in the form:

$$\begin{cases} Z_{k+1} = Z_k + f_{\mathcal{G}}^k(k, Z_k, \theta_k) \\ Z_0 = \ell_{\mathcal{G}}^x(X) \\ \hat{Y}_k = \ell_{\mathcal{G}}^y(Z_k) \end{cases}, \quad k \in \mathbb{N}, \quad (1)$$

with hidden state $Z_k \in \mathbb{R}^{n \times n_z}$, node features $X \in \mathbb{R}^{n \times n_x}$ and output $\hat{Y} \in \mathbb{R}^{n \times n_y}$. $f_{\mathcal{G}}^k$ are generally matrix–valued nonlinear functions conditioned on graph \mathcal{G} , $\theta_k \in \mathbb{R}^{n_{\theta}}$ is the tensor of trainable parameters of the k -th layer and $\ell_{\mathcal{G}}^x$, $\ell_{\mathcal{G}}^y$ represent feature embedding and output layers, respectively. Note that the explicit dependence on k of the dynamics is justified in some graph architectures, such as diffusion graph convolutions (Atwood and Towsley, 2016).

A *neural graph differential equation* (Neural GDE) is constructed as the continuous–depth limit of (1), defined as the nonlinear affine dynamical system:

$$\begin{cases} \dot{Z}_t = f_{\mathcal{G}}(t, Z_t, \theta_t) \\ Z_0 = \ell_{\mathcal{G}}^x(X) \\ \hat{Y}_t = \ell_{\mathcal{G}}^y(Z_t) \end{cases}, \quad s \in \mathcal{S} \subset \mathbb{R}, \quad (2)$$

where $f_{\mathcal{G}} : \mathcal{T} \times \mathbb{R}^{n \times n_z} \times \mathbb{R}^{n_{\theta}} \rightarrow \mathbb{R}^{n \times n_z}$ is a depth–varying vector field defined on graph \mathcal{G} and $\ell_{\mathcal{G}}^x : \mathbb{R}^{n \times n_x} \rightarrow \mathbb{R}^{n \times n_z}$, $\ell_{\mathcal{G}}^y : \mathbb{R}^{n \times n_z} \rightarrow \mathbb{R}^{n \times n_y}$ are two affine linear mappings. Depending on the choice of input transformation $\ell_{\mathcal{G}}^x$, different node feature augmentation techniques can be introduced (Dupont et al., 2019; Massaroli et al., 2020) to reduce stiffness of learned vector fields.

Well-posedness Let $\mathcal{T} := [0, 1]$. Under mild conditions on f_G , namely Lipschitz continuity with respect to Z and uniform continuity with respect to t , for each initial condition (GDE embedded input) $Z_0 = \ell_G^x(X)$, the matrix-valued ODE in (2) admits a unique solution Z_t defined in the whole \mathcal{S} . Thus there is a mapping Φ from $\mathbb{R}^{n \times n_z}$ to the space of absolutely continuous functions $\mathcal{T} \rightarrow \mathbb{R}^{n \times n_z}$ such that $Z_t := \Phi_t(X)$ satisfies the ODE in (2). Symbolically, the output of the Neural GDE is obtained by the following

$$\hat{Y}_t = \ell_G^y \left(\ell_G^x(X) + \int_0^t f_G(s, Z_s, \theta) ds \right).$$

2.2 Neural GDEs on Static Graphs

Based on graph spectral theory (Shuman et al., 2013; Sandryhaila and Moura, 2013), the residual version of *graph convolution network* (GCN) (Kipf and Welling, 2016) layers is obtained by setting:

$$f_G^k(Z_k, \theta_k) := L_G Z_k W_k \quad (3)$$

in (1), where $L_G \in \mathbb{R}^{n \times n}$ is the graph *Laplacian*, $\theta_k := \text{vec}(W_k)$ and σ . The general formulation of the continuous GCNs counterpart, *neural graph convolution differential equations* (Neural GCDEs) is similarly defined by letting the vector field f_G be a multilayer convolution, i.e.

$$\dot{Z}_t = f_G(Z_t, \theta) := f_G^N \circ \sigma \circ f_G^{N-1} \circ \sigma \circ \dots \circ f_G^1 \quad (4)$$

with σ being a nonlinear activation function though to be acting element-wise and $\theta := \text{vec}(W_1, \dots, W_N) \in \mathbb{R}^{n^2}$. Note that the Laplacian L_G can be computed in different ways, see e.g. (Bruna et al., 2013; Defferrard et al., 2016; Levie et al., 2018; Zhuang and Ma, 2018). Alternatively, diffusion-type convolution layers (Li et al., 2017) can be introduced. We note that expressivity of the model is improved via letting the parameters be *time-varying* i.e. $f_G(Z_t, \theta_t) := L_G Z_t W_t$, $\theta_t := \text{vec}(W_t)$ where θ_t is parametrized by spectral or time discretizations (Massaroli et al., 2020).

Additional continuous-time variants We include additional derivations of continuous-time counterparts of common static GNN models such as *graph attention networks* (GATs) (Veličković et al., 2017) and general message passing GNNs in the Appendix. We note that due to the purely algebraic nature of common operations in geometric models such as attention operators (Vaswani et al., 2017), Neural GDEs are compatible with the vast majority of GNN architectures.

3 Neural GDEs on Dynamic Graphs •

Common use cases for GNNs involve prediction in dynamic graphs, which introduce additional challenges. We discuss how Neural GDE models can be extended to address these scenarios, leveraging tools from *hybrid dynamical system* theory (Van Der Schaft and Schumacher, 2000) to derive a Neural Hybrid GDE formulation.

Here, Neural GDEs represent a natural model class for autoregressive modeling of sequences of graphs $\{\mathcal{G}_t\}$ and seamlessly link to dynamical network theory. This line of reasoning naturally leads to an extension of classical spatio-temporal architectures in the form of *hybrid dynamical systems* (Van Der Schaft and Schumacher, 2000; Goebel et al., 2009), i.e., systems characterized by interacting continuous and discrete-time dynamics.

Notation: Let $(\mathcal{K}, >)$, $(\mathcal{T}_e, >)$ be linearly ordered sets; namely, $\mathcal{K} \subset \mathbb{N} \setminus \{0\}$ and \mathcal{T}_e is a set of time instants, $\mathcal{T}_e := \{t_k\}_{k \in \mathcal{K}}$. We suppose to be given a *state-graph data stream* which is a sequence in the form $\{(X_t, \mathcal{G}_t)\}_{t \in \mathcal{T}_e}$. Let us also define a *hybrid time domain* as the set $\mathcal{I} := \bigcup_{k \in \mathcal{K}} ([t_k, t_{k+1}], k)$ and a *hybrid arc* on \mathcal{I} as a function Φ such that for each $k \in \mathcal{K}$, $t \mapsto \Phi(t, k)$ is absolutely continuous in $\{t : (t, j) \in \text{dom } \Phi\}$. Our aim is to build a continuous model predicting, at each $t_k \in \mathcal{T}_e$, the value of $X_{t_{k+1}}$, given (X_t, \mathcal{G}_t) .

3.1 Neural Hybrid GDEs

The core idea is to have a Neural GDE smoothly steering the latent node features between two time instants and then apply some discrete transition operator, resulting in a “jump” of state Z which is then processed by an output layer. Solutions of the proposed continuous spatio-temporal model are therefore hybrid arcs.

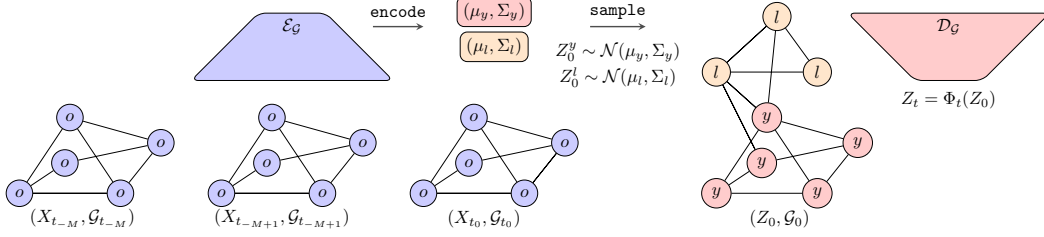


Figure 2: *Latent Neural GDE* model. Data observations are encoded into a compact representation, required to sample from approximate posteriors on observable and latent node features, forming an *augmented* graph. Following this procedure, the dynamics are unrolled by the Neural GDE decoder \mathcal{D}_G . We consider deterministic as well as stochastic decoders in the form of Neural GSDEs.

The general formulation of a Neural Hybrid GDE model can be symbolically represented by:

$$\begin{cases} \dot{Z}_t &= f_{\mathcal{G}_{t_k}}(Z_t, \theta) & t \in [t_{k-1}, t_k] \\ Z_t^+ &= \ell_{\mathcal{G}_{t_k}}^j(Z_t, X_t) & t = t_k \\ \hat{Y}_t &= \ell_{\mathcal{G}_{t_k}}^y(Z_t) \end{cases} \quad k \in \mathcal{K}, \quad (5)$$

where $f_{\mathcal{G}}, \ell_{\mathcal{G}}^j, \ell_{\mathcal{G}}^y$ are GNN-like operators or general neural network layers and Z_t^+ represents the value of Z_t after the discrete transition. The evolution of system (5) is indeed a sequence of hybrid arcs defined on a hybrid time domain. Compared to standard recurrent models which are only equipped with discrete jumps, system (5) incorporates a continuous flow of latent node features Z_t between jumps. This feature of Hybrid Neural GDEs allows them to track the evolution of dynamical systems from observations with irregular time steps. In the experiments we consider $\ell_{\mathcal{G}}^j$ to be a GRU cell (Cho et al., 2014), obtaining *neural graph convolution differential equation-GRU* (GCDE-GRU).

Sequential adjoint for Hybrid GDEs Continuous-depth models, including Neural ODEs and SDEs, can be trained using adjoint sensitivity methods (Pontryagin et al., 1962; Chen et al., 2018; Li et al., 2020). Care must be taken in the case of sequence models such as Neural Hybrid GDEs which often admit losses dependent on solution values at various timestamps, rather than considering terminal states exclusively. In particular, we may consider loss functions of the form

$$L_{\theta}(X_{t_0:N}, Z_{t_0:N}) = \sum_{k \in \mathcal{K}} c_{\theta}(X_{t_k}, Z_{t_k}).$$

In such a case, back-propagated gradient can be computed with an extension of classic adjoint techniques (Pontryagin et al., 1962)

$$\frac{dL_{\theta}}{d\theta} = \frac{\partial L_{\theta}}{\partial \theta} - \int_{\mathcal{T}} \left\langle \lambda, \frac{\partial f_{\mathcal{G}_t}(t, \Phi_t(Z_0), \theta)}{\partial \theta} \right\rangle dt$$

where the Lagrange multiplier $\lambda_t : \mathcal{I}^- \rightarrow \mathbb{R}^{n \times n_x}$ is, however, a hybrid arc on the reversed (backward) hybrid time domain \mathcal{I}^- satisfying the hybrid inclusion

$$\begin{aligned} \dot{\lambda}_t &\in F(t, Z_t, \lambda_t) & t \in \mathcal{T} \\ \lambda_t^+ &\in G(X_t, Z_t, \lambda_t) & t \in \mathcal{T}_e \\ \lambda_{t_K} &= \frac{\partial c_{\theta}}{\partial Z_{t_K}} \end{aligned}$$

with $t_K = \sup \mathcal{T}_e$ and F, G are set-valued mappings $F : \mathcal{T} \times \mathbb{R}^{n \times n_z} \times \mathbb{R}^{n \times n_z} \rightrightarrows \mathbb{R}^{n \times n_z}$, $G : \mathbb{R}^{n \times n_x} \times \mathbb{R}^{n \times n_z} \times \mathbb{R}^{n \times n_z} \rightrightarrows \mathbb{R}^{n \times n_z}$, defined as

$$\begin{aligned} F(t, Z_t, \lambda_t) &:= \left\{ F_k : t \in [t_k, t_{k+1}] \Rightarrow F_k = -\frac{\partial f_{\mathcal{G}_{t_k}}}{\partial Z} \lambda_t \right\} \\ G(X_t, Z_t, \lambda_t) &:= \left\{ G_k : t = t_k \Rightarrow G_k = \lambda_t + \frac{\partial c_{\theta}}{\partial Z_t} \right\}. \end{aligned}$$

3.2 Latent Neural GDEs

The Neural GDE variants introduced so far can be combined into a latent variable model for dynamic graphs. In particular, we might be interested in predicting $(X_{t_1}, \mathcal{G}_{t_1}), \dots, (X_{t_N}, \mathcal{G}_{t_N})$ given past observations $(X_{t_{-M}}, \mathcal{G}_{t_{-M}}), \dots, (X_{t_0}, \mathcal{G}_{t_0})$. To this end, we introduce *Latent Neural GDEs* as **encoder–decoder** models in the form

$$\begin{aligned} E &= \mathcal{E}_{\mathcal{G}}(\mathcal{G}_{t_{-M:0}}, X_{t_{-M:0}}) && \text{Graph Encoder} \\ Z_0 &\sim q_{\mathcal{G}}(Z_0|E) && \text{Reparametrization} \\ Z_t &= \Phi_t(Z_0) && \text{Graph Decoder} \\ \hat{Y}_t &= Z_t^y \end{aligned}$$

where the decoded output \hat{Y}_t in a time-domain \mathcal{T} is obtained via the solution of Neural GDE. Latent Neural GDEs are designed for reconstruction or extrapolation tasks involving dynamic graphs, where the underlying data-generating process is known to be a differential equation.

Rather than constraining the evolution of decoded node features to a latent space, requiring readout layers to map back to data-space, we construct an *augmented* graph with latent nodes Z^l and output-space nodes Z^y . This formulation retains the flexibility of a full latent model while allowing for the embedding of stricter inductive biases on the nature of the latent variables and their effects, as shown experimentally on genetic regulatory networks. Figure 2 depicts an example instance of Latent Neural GDEs, where the approximate posterior on Z^l and Z^y is defined as a multivariate Gaussian.

Latent Neural GDEs are trained via maximum likelihood. The optimization problem can be cast as the maximization of an evidence lower bound (ELBO):

$$\text{ELBO} := \mathbb{E}_{Z_0 \sim q_{\mathcal{G}}} \left[\sum_{k=1}^N \log p(\hat{Y}_{t_k}) - \text{KL}(q_{\mathcal{G}} || \mathcal{N}(0, \mathbb{I})) \right]$$

with an observation-space density defined as $p(\hat{Y}_{t_k}) = \mathcal{N}(Y_{t_k}, \Sigma_{t_k})$ and a covariance Σ_{t_k} hyperparameter.

Embedding stochasticity into Neural GDEs Recent work (Li et al., 2020; Peluchetti and Favaro, 2020; Massaroli et al., 2021) develops extensions of continuous models to stochastic differential equations (Kunita, 1997; Øksendal, 2003) for static tasks or optimal control. These results carry over to the Neural GDEs framework; an example application is to consider *stochastic* Latent Neural GDE decoders in order to capture inherent stochasticity in the samples. Here, given a multi-dimensional Brownian motion B_t , we define and train *Neural graph stochastic differential equations* (Neural GSDEs) of the form

$$dZ_t = f_{\mathcal{G}}(t, Z_t, \theta)dt + g_{\mathcal{G}}(t, Z_t, \theta) \circ dB_t$$

where we adopt a *Stratonovich* SDE formulation.

4 Experiments ●

We evaluate Neural GDEs on a suite of different tasks. The experiments and their primary objectives are summarized below:

- Trajectory extrapolation task on a synthetic multi-agent dynamical system. We compare Neural ODEs and Neural GDEs, providing in addition to the comparison a motivating example for the introduction of additional biases inside GDEs in the form of second-order models (Yildiz et al., 2019; Massaroli et al., 2020; Norcliffe et al., 2020).
- Traffic forecasting on an undersampled version of PeMS (Yu et al., 2018) dataset. We measure the performance improvement obtained by a correct inductive bias on continuous dynamics and robustness to irregular timestamps.
- Flux prediction in genetic regulatory networks such as Elowitz-Leibler repressilator circuits (Elowitz and Leibler, 2000). We investigate Latent Neural GDEs for prediction in biological networks with stochastic dynamics, where prior knowledge on graph structure linking latent and observable nodes plays a key role.

Model	MAPE _{30%}	RMSE _{30%}	MAPE _{70%}	RMSE _{70%}	MAPE _{100%}	RMSE _{100%}
GRU	27.14 ± 0.45	13.25 ± 0.11	27.24 ± 0.19	13.28 ± 0.05	27.20 ± 0.00	13.29 ± 0.00
GCGRU	23.60 ± 0.38	11.97 ± 0.06	21.33 ± 0.16	11.20 ± 0.04	20.92 ± 0.00	10.87 ± 0.00
GCDE-GRU	22.95 ± 0.37	11.67 ± 0.10	20.94 ± 0.14	10.95 ± 0.04	20.46 ± 0.00	10.76 ± 0.00

Table 1: Forecasting test results across 20 runs (mean and standard dev.). MAPE_i indicates an *i*% test sampling undersampling strategy, i.e *i*% of the time series measurements are randomly selected and kept for training.

4.1 • Multi-Agent Trajectory Extrapolation

Experimental setup We evaluate GDEs and a collection of deep learning baselines on the task of extrapolating the dynamical behavior of a synthetic, non-conservative mechanical multi-particle system. Particles interact within a certain radius with a viscoelastic force. Outside the mutual interactions, captured by a time-varying adjacency matrix A_t , the particles would follow a periodic motion, gradually losing energy due to viscous friction. The adjacency matrix A_t is computed along the trajectory as:

$$A_t^{(ij)} = \begin{cases} 1 & 2\|x_i(t) - x_j(t)\| \leq r \\ 0 & \text{otherwise} \end{cases},$$

where $\mathbf{x}_i(t)$ is the position of node i at time t . Therefore, A_t results to be symmetric, $A_t = A_t^\top$ and yields an undirected graph. The dataset is collected by integrating the system for $T = 5s$ with a fixed step-size of $dt = 1.95 \cdot 10^{-3}$ and is split evenly into a training and test set. We consider 10 particle systems. An example trajectory is shown in Appendix C. All models are optimized to minimize mean-squared-error (MSE) of 1-step predictions using Adam (Kingma and Ba, 2014) with constant learning rate 0.01. We measure test *mean average percentage error* (MAPE) of model predictions in different extrapolation regimes. *Extrapolation steps* denotes the number of predictions each model Φ has to perform without access to the nominal trajectory. This is achieved by recursively letting inputs at time t be model predictions at time $t - \Delta t$ i.e $\hat{Y}_{t+\Delta t} = \phi(\hat{Y}_t)$ for a certain number of extrapolation steps, after which the model is fed the actual nominal state X and the cycle is repeated until the end of the test trajectory. For a robust comparison, we report mean and standard deviation across 10 seeded training and evaluation runs. Additional experimental details, including the analytical formulation of the dynamical system, are provided as supplementary material.

Models As the vector field depends only on the state of the system, available in full during training, the baselines do not include recurrent modules. We consider the following models:

- A 3-layer fully-connected neural network, referred to as *Static*. No assumption on the dynamics
- A vanilla Neural ODE with the vector field parametrized by the same architecture as *Static*. ODE assumption on the dynamics.
- A 3-layer convolution Neural GDE, Neural GCDE. Dynamics assumed to be determined by a blend of graphs and ODEs
- A 3-layer, second-order (Yıldız et al., 2019; Massaroli et al., 2020; Norcliffe et al., 2020) Neural GCDE and referred to as *Neural GCDE-II*.

A grid hyperparameter search on number of layers, ODE solver tolerances and learning rate is performed to optimize *Static* and Neural ODEs. We use the same hyperparameters for Neural GDEs. We used the torchdyn Poli et al. (2020) library for Neural ODE baselines.

Results Figure 3 shows the growth rate of test MAPE error as the number of extrapolation steps is increased. *Static* fails to extrapolate beyond the 1-step setting seen during training. Neural ODEs overfit spurious particle interaction terms and their error rapidly grows as the number of extrapolation steps is increased. Neural GCDEs, on the other hand, are able to effectively leverage relational information to track the system: we provide complete visualization of extrapolation trajectory comparisons in the Appendix. Lastly, Neural GCDE-IIs outperform first-order Neural GCDEs as their structure inherently possesses crucial information about the relative relationship of positions and velocities, accurate with respect to the observed dynamical system.

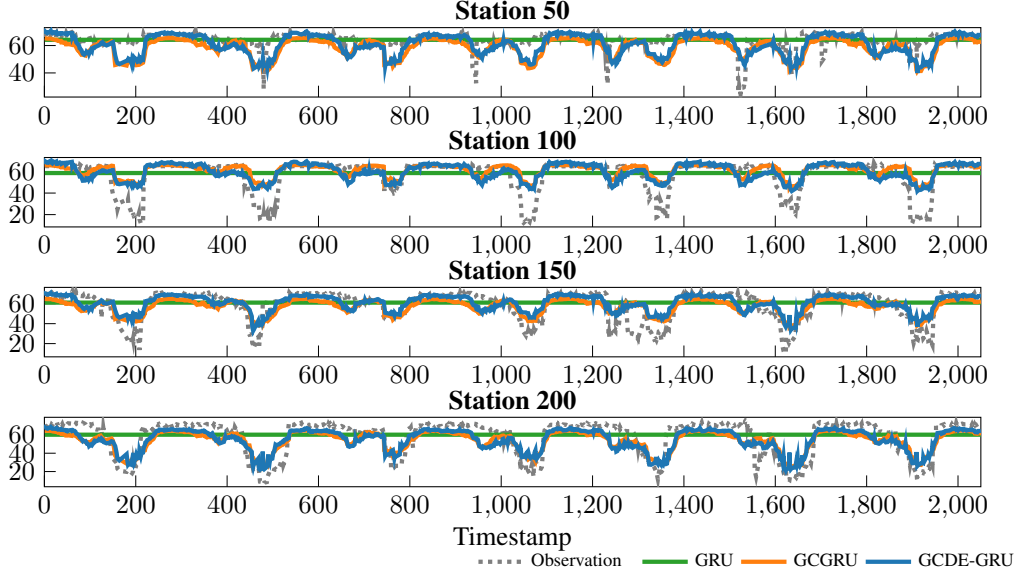


Figure 4: Traffic data prediction results of 50% undersampling. GCDE-GRUs are able to evolve the latents between timestamps and provide a more accurate fit.

4.2 Traffic Forecasting

Experimental setup We evaluate the effectiveness of Neural Hybrid GDE models on forecasting tasks by performing a series of experiments on the established PeMS traffic dataset. We follow the setup of (Yu et al., 2018) in which a subsampled version of PeMS, PeMS7(M), is obtained via selection of 228 sensor stations and aggregation of their historical speed data into regular 5 minute frequency time series. We construct the adjacency matrix A by thresholding the Euclidean distance between observation stations i.e. when two stations are closer than the threshold distance, an edge between them is included. The threshold is set to the 40th percentile of the station distances distribution. To simulate a challenging environment with missing data and irregular timestamps, we undersample the time series by performing independent Bernoulli trials on each data point. Results for 3 increasingly challenging experimental setups are provided: undersampling with 0%, 30%, 50% and 70% of removal. In order to provide a robust evaluation of performance in regimes with irregular data, the testing is repeated 20 times per model, each with a different undersampled version of the test dataset. We collect *root mean square error* (RMSE) and MAPE. More details about the chosen metrics and data are included as supplementary material.

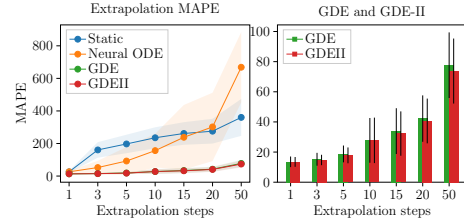


Figure 3: Test extrapolation MAPE averaged across 10 experiments. Shaded area and error bars indicate 1-standard deviation intervals.

Models In order to measure performance gains obtained by Neural GDEs in settings with data generated by continuous time systems, we employ a GCDE-GRU-dpr5 as well as its discrete counterpart GCGRU (Zhao et al., 2018). To contextualize the effectiveness of introducing graph representations, we include the performance of GRUs since they do not directly utilize structural information of the system in predicting outputs. Apart from GCDE-GRU, both baselines have no innate mechanism for handling timestamp information. For a fair comparison, we include timestamp differences between consecutive samples and *sine-encoded* (Petneházi, 2019) absolute time information as additional features. All models receive an input sequence of 5 graphs to perform the prediction.

Results Non-constant differences between timestamps result in a challenging forecasting task for a single model since the average prediction horizon changes drastically over the course of training and

testing. Traffic systems are intrinsically dynamic and continuous in nature and, therefore, a model able to track continuous underlying dynamics is expected to offer improved performance. Since GCDE-GRUs and GCGRUs are designed to match in structure we can measure this performance increase from the results shown in Table 1. GCDE-GRUs outperform GCGRUs and GRUs in all undersampling regimes. Additional details and visualizations are included in Appendix C.

4.3 • Repressilator Reconstruction

Experimental setup We investigate prediction of genetic regulatory network dynamics as a showcase application of Latent Neural GDEs. In particular, we consider biological metabolic fluxes of an Elowitz-Leibler repressilator circuit (Elowitz and Leibler, 2000). Repressilator circuits are a common example of feedback mechanism used to maintain homeostasis in biological systems. The circuit is modeled as a bipartite graph where one set of nodes is comprised of reactions and a second one of biochemical species (mRNA and protein), as shown in Figure 5. The repressilator feedback cycle is structured such that each protein suppresses the expression of mRNA for the next protein in the cycle, yielding oscillatory behaviour. Accurate genome-scale models of dynamic metabolic flux are currently intractable, largely due to scaling limitations or insufficient prediction accuracy when compared with *in vitro* and *in vivo* data. By example, state-of-the-art genome-scale dynamic flux simulations currently require model reduction to core metabolism (Masid et al., 2020). In order to demonstrate the effectiveness of the Neural GDE framework in fitting stochastic systems and allowing for interpretability of underlying mechanisms, we generate a training dataset of ten trajectories by symbolic integration via the τ -leaping method (Gillespie, 2007; Padgett and Ilie, 2016) over a time span of 300 seconds. During training, we split each trajectory into halves and task the model with reconstruction of the last 150 seconds during the decoding phase, conditioned on the first half.

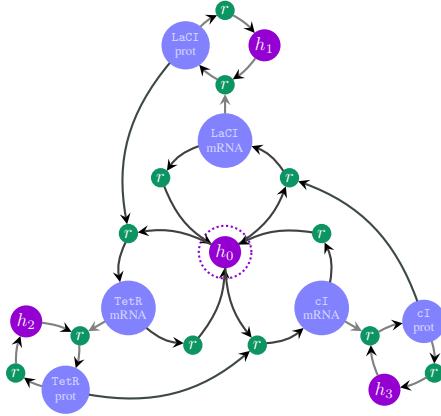


Figure 5: A bipartite graph representation of a metabolic flux schematic for the Elowitz-Leibler repressilator, where h is degradation, and r is reaction.

Models We assess modeling capabilities of Latent Neural GDEs applied to biological networks, with a focus on interpretability. The particular graph structure of the system shown in Figure 5 lends itself to a formulation where protein and mRNA dynamics can be grouped in Z^y , whereas reaction nodes constitute the set of latent nodes Z^l . This in turn allows the decoder to utilize prior knowledge on the edges connecting the two sets of nodes. We report the full adjacency matrix in Appendix B.

The Latent Neural GDE is equipped with an encoder comprised of 2-layers of *temporal convolutions* (TCNs). To model stochasticity and provide uncertainty estimates in predictions, the architecture leverages *neural graph stochastic differential equation* (Neural GSDE) during decoding steps. The drift f_G and diffusion g_G networks of the Neural GSDEs follow an equivalent 3-layer GNN design: [GCN, GAT (Veličković et al., 2017), GCN] with hidden dimension 3 and hyperbolic-tangent activations.

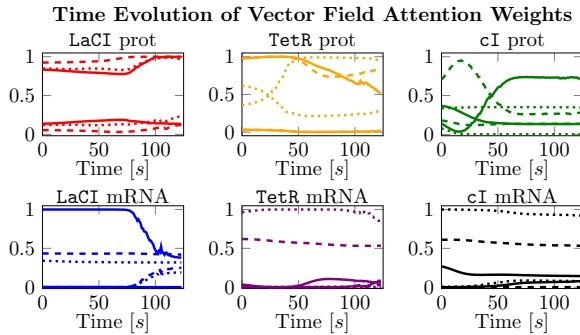


Figure 6: Evolution of GAT Veličković et al. (2017) attention weights of drift networks f_G in the Neural GSDE decoder. The linestyles indicate different edge weights (*incoming* and *outgoing*) for the three edges connected to each protein and mRNA.

Results Figure 7 provides a visual inspection of protein and mRNA trajectory predictions produced by Latent Neural GDEs during testing. The model is able to reconstruct species concentration evolution, while calibrating and balancing decoder diffusion and drift to match the different characteristics of protein and mRNA dynamics. In black, we further highlight model extrapolations beyond the 300 seconds regime, which shows an increase in model uncertainty. The attention coefficients of edges linking reaction nodes with the respective protein and mRNA species is shown in Figure 6. As the dynamics are unrolled by the decoder, the attention weights of the GAT Veličković et al. (2017) layer

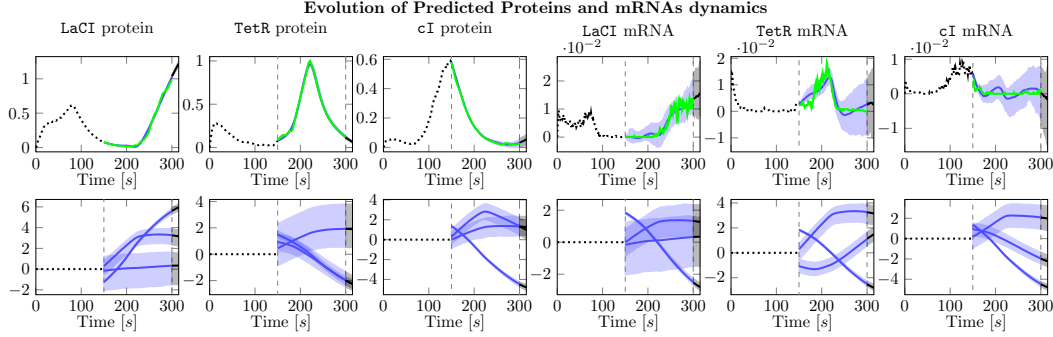


Figure 7: [Above] Reconstruction and extrapolation of protein and mRNA dynamics produced by Latent Neural SGDEs. Min-max intervals produced from multiple samples of the SDE decoder are indicated as shaded areas. [Below] Time-evolution of latent node features corresponding to reaction nodes connected to each of the protein and mRNA species in the repressilator graph. The node reaction latent values evolve over time, modulating the reaction between protein species.

present in the drift f_G evolve over time, modulating reactions between species. Each protein and mRNA node is shown to correspond to six edge attention weights: three incoming and three outgoing.

5 Related work

Since the first seminal paper on Neural ODEs (Chen et al., 2018), several attempts have been made at developing continuous variants of specific GNN models. Here, we provide a detailed comparison of Neural GDEs with other continuous formulations, highlighting in the process the need for a general framework systematically blending graphs and differential equations. Sanchez-Gonzalez et al. (2019) proposes using Graph Networks (GNs) (Battaglia et al., 2018) and ODEs to track Hamiltonian functions, obtaining *Hamiltonian Graph Networks* (HGNs). The resulting model is evaluated on prediction tasks for conservative systems; an HGN can be expressed as a special case of a Neural GDE, and their application domain is fairly limited, being constrained to overly-restrictive Hamiltonian vector fields. In example, we evaluate Neural GDEs on multi-particle non-conservative systems subject to viscous friction, which are theoretically incompatible with HGNs. (Deng et al., 2019) introduces a GNN version of continuous normalizing flows (Chen et al., 2018; Grathwohl et al., 2018), extending (Liu et al., 2019), and deriving a continuous message passing scheme. However, the model is limited to the specific application of generative modeling, and no attempts at generalizing the formulation are made. Finally, Xhonneux et al. (2019) propose *Continuous Graph Neural Networks* (CGNNs), a linear ODE formulation for message passing on graphs. CGNNs rely on rather strong assumptions, such as linearity or depth invariance, and are evaluated exclusively on datasets that can notably be tackled with linear GNNs (Wu et al., 2019a). Fashioning linear flows offers a closed-form solution of the model’s output, though this is achieved at the cost of expressivity and generality. Indeed, CGNNs can also be regarded as a simple form of Neural GDEs.

In contrast, our goal is to develop a system-theoretic framework for continuous-depth GNNs, validated through extensive experiments. To the best of our knowledge, no existing work offers a continuous-depth solution for dynamic graph prediction including latent and stochastic dynamics.

6 Conclusion

In this work we introduce *neural graph differential equations* (Neural GDE), the continuous-depth counterpart to *graph neural networks* (GNN) where the inputs are propagated through a continuum of GNN layers. Neural GDEs are designed to offer a data-driven modeling approach for *dynamical networks*, whose dynamics are defined by a blend of discrete topological structures and differential equations. In sequential forecasting problems, Neural GDEs can accommodate irregular timestamps and track underlying continuous dynamics. Neural GDEs, including latent and stochastic variants, have been evaluated across applications, including traffic forecasting and prediction in genetic regulatory networks.

References

- M. Andreasson, D. V. Dimarogonas, H. Sandberg, and K. H. Johansson. Distributed control of networked dynamical systems: Static feedback, integral action and consensus. *IEEE Transactions on Automatic Control*, 59(7):1750–1764, 2014.
- J. Atwood and D. Towsley. Diffusion-convolutional neural networks. In *Advances in Neural Information Processing Systems*, pages 1993–2001, 2016.
- P. W. Battaglia, J. B. Hamrick, V. Bapst, A. Sanchez-Gonzalez, V. Zambaldi, M. Malinowski, A. Tacchetti, D. Raposo, A. Santoro, R. Faulkner, et al. Relational inductive biases, deep learning, and graph networks. *arXiv preprint arXiv:1806.01261*, 2018.
- J. Bruna, W. Zaremba, A. Szlam, and Y. LeCun. Spectral networks and locally connected networks on graphs. *arXiv preprint arXiv:1312.6203*, 2013.
- D. Chen, Y. Lin, W. Li, P. Li, J. Zhou, and X. Sun. Measuring and relieving the over-smoothing problem for graph neural networks from the topological view. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 3438–3445, 2020.
- T. Q. Chen, Y. Rubanova, J. Bettencourt, and D. K. Duvenaud. Neural ordinary differential equations. In *Advances in neural information processing systems*, pages 6571–6583, 2018.
- K. Cho, B. Van Merriënboer, C. Gulcehre, D. Bahdanau, F. Bougares, H. Schwenk, and Y. Bengio. Learning phrase representations using rnn encoder-decoder for statistical machine translation. *arXiv preprint arXiv:1406.1078*, 2014.
- M. Defferrard, X. Bresson, and P. Vandergheynst. Convolutional neural networks on graphs with fast localized spectral filtering. In *Advances in neural information processing systems*, pages 3844–3852, 2016.
- Z. Deng, M. Nawhal, L. Meng, and G. Mori. Continuous graph flow, 2019.
- J. R. Dormand and P. J. Prince. A family of embedded runge-kutta formulae. *Journal of computational and applied mathematics*, 6(1):19–26, 1980.
- J. Du, J. Futoma, and F. Doshi-Velez. Model-based reinforcement learning for semi-markov decision processes with neural odes. *arXiv preprint arXiv:2006.16210*, 2020.
- E. Dupont, A. Doucet, and Y. W. Teh. Augmented neural odes. *arXiv preprint arXiv:1904.01681*, 2019.
- M. B. Elowitz and S. Leibler. A synthetic oscillatory network of transcriptional regulators. *Nature*, 403(6767):335–338, Jan 2000. ISSN 1476-4687. doi: 10.1038/35002125. URL <https://doi.org/10.1038/35002125>.
- C. Gallicchio and A. Micheli. Fast and deep graph neural networks, 2019.
- M. Gasse, D. Chételat, N. Ferroni, L. Charlin, and A. Lodi. Exact combinatorial optimization with graph convolutional neural networks. *arXiv preprint arXiv:1906.01629*, 2019.
- D. T. Gillespie. Stochastic simulation of chemical kinetics. *Annual Review of Physical Chemistry*, 58(1):35–55, 2007. doi: 10.1146/annurev.physchem.58.032806.104637. PMID: 17037977.
- R. Goebel, R. G. Sanfelice, and A. R. Teel. Hybrid dynamical systems. *IEEE Control Systems Magazine*, 29(2):28–93, 2009.
- W. Grathwohl, R. T. Chen, J. Bettencourt, I. Sutskever, and D. Duvenaud. Ffjord: Free-form continuous dynamics for scalable reversible generative models. *arXiv preprint arXiv:1810.01367*, 2018.
- S. Greydanus, M. Dzamba, and J. Yosinski. Hamiltonian neural networks. *arXiv preprint arXiv:1906.01563*, 2019.
- J. Jia and A. R. Benson. Neural jump stochastic differential equations. *arXiv preprint arXiv:1905.10403*, 2019.
- P. Kidger, J. Morrill, J. Foster, and T. Lyons. Neural controlled differential equations for irregular time series. *arXiv preprint arXiv:2005.08926*, 2020.
- D. P. Kingma and J. Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

- T. N. Kipf and M. Welling. Semi-supervised classification with graph convolutional networks. *arXiv preprint arXiv:1609.02907*, 2016.
- P. E. Kloeden, E. Platen, and H. Schurz. *Numerical solution of SDE through computer experiments*. Springer Science & Business Media, 2012.
- H. Kunita. *Stochastic flows and stochastic differential equations*, volume 24. Cambridge university press, 1997.
- R. Levie, F. Monti, X. Bresson, and M. M. Bronstein. Cayleynets: Graph convolutional neural networks with complex rational spectral filters. *IEEE Transactions on Signal Processing*, 67(1): 97–109, 2018.
- X. Li, T.-K. L. Wong, R. T. Chen, and D. Duvenaud. Scalable gradients for stochastic differential equations. *arXiv preprint arXiv:2001.01328*, 2020.
- Y. Li, R. Yu, C. Shahabi, and Y. Liu. Diffusion convolutional recurrent neural network: Data-driven traffic forecasting. *arXiv preprint arXiv:1707.01926*, 2017.
- Y. Li, O. Vinyals, C. Dyer, R. Pascanu, and P. Battaglia. Learning deep generative models of graphs. *arXiv preprint arXiv:1803.03324*, 2018a.
- Z. Li, Q. Chen, and V. Koltun. Combinatorial optimization with graph convolutional networks and guided tree search. In *Advances in Neural Information Processing Systems*, pages 539–548, 2018b.
- J. Liu, A. Kumar, J. Ba, J. Kiros, and K. Swersky. Graph normalizing flows. In *Advances in Neural Information Processing Systems*, pages 13556–13566, 2019.
- I. Loshchilov and F. Hutter. Sgdr: Stochastic gradient descent with warm restarts. *arXiv preprint arXiv:1608.03983*, 2016.
- A. Lou, D. Lim, I. Katsman, L. Huang, Q. Jiang, S.-N. Lim, and C. De Sa. Neural manifold ordinary differential equations. *arXiv preprint arXiv:2006.10254*, 2020.
- J. Lu and G. Chen. A time-varying complex dynamical network model and its controlled synchronization criteria. *IEEE Transactions on Automatic Control*, 50(6):841–846, 2005.
- M. Masid, M. Ataman, and V. Hatzimanikatis. Analysis of human metabolism by reducing the complexity of the genome-scale models using redHUMAN. *Nature Communications*, 11(1):2821, 2020. ISSN 2041-1723. doi: 10.1038/s41467-020-16549-2. URL <https://doi.org/10.1038/s41467-020-16549-2>.
- S. Massaroli, M. Poli, J. Park, A. Yamashita, and H. Asama. Dissecting neural odes. *arXiv preprint arXiv:2002.08071*, 2020.
- S. Massaroli, M. Poli, S. Peluchetti, J. Park, A. Yamashita, and H. Asama. Learning stochastic optimal policies via gradient descent. *IEEE Control Systems Letters*, 2021.
- E. Mathieu and M. Nickel. Riemannian continuous normalizing flows. *arXiv preprint arXiv:2006.10605*, 2020.
- A. Norcliffe, C. Bodnar, B. Day, N. Simidjievski, and P. Liò. On second order behaviour in augmented neural odes. *arXiv preprint arXiv:2006.07220*, 2020.
- B. Øksendal. Stochastic differential equations. In *Stochastic differential equations*, pages 65–84. Springer, 2003.
- K. Oono and T. Suzuki. Graph neural networks exponentially lose expressive power for node classification, 2019.
- J. M. A. Padgett and S. Ilie. An adaptive tau-leaping method for stochastic simulations of reaction-diffusion systems. *AIP Advances*, 6(3):035217, 2016. doi: 10.1063/1.4944952. URL <https://doi.org/10.1063/1.4944952>.
- S. Peluchetti and S. Favaro. Infinitely deep neural networks as diffusion processes. In *International Conference on Artificial Intelligence and Statistics*, pages 1126–1136. PMLR, 2020.
- G. Petneházi. Recurrent neural networks for time series forecasting. *arXiv preprint arXiv:1901.00069*, 2019.
- M. Poli, S. Massaroli, A. Yamashita, H. Asama, and J. Park. Torchdyn: A neural differential equations library, 2020.

- L. S. Pontryagin, E. Mishchenko, V. Boltyanskii, and R. Gamkrelidze. The mathematical theory of optimal processes. 1962.
- Y. Rubanova, R. T. Chen, and D. Duvenaud. Latent odes for irregularly-sampled time series. *arXiv preprint arXiv:1907.03907*, 2019.
- C. Runge. Über die numerische auflösung von differentialgleichungen. *Mathematische Annalen*, 46(2):167–178, 1895.
- A. Sanchez-Gonzalez, N. Heess, J. T. Springenberg, J. Merel, M. Riedmiller, R. Hadsell, and P. Battaglia. Graph networks as learnable physics engines for inference and control. *arXiv preprint arXiv:1806.01242*, 2018.
- A. Sanchez-Gonzalez, V. Bapst, K. Cranmer, and P. Battaglia. Hamiltonian graph networks with ode integrators. *arXiv preprint arXiv:1909.12790*, 2019.
- A. Sandryhaila and J. M. Moura. Discrete signal processing on graphs. *IEEE transactions on signal processing*, 61(7):1644–1656, 2013.
- D. I. Shuman, S. K. Narang, P. Frossard, A. Ortega, and P. Vandergheynst. The emerging field of signal processing on graphs: Extending high-dimensional data analysis to networks and other irregular domains. *IEEE signal processing magazine*, 30(3):83–98, 2013.
- L. N. Smith and N. Topin. Super-convergence: Very fast training of neural networks using large learning rates. In *Artificial Intelligence and Machine Learning for Multi-Domain Operations Applications*, volume 11006, page 1100612. International Society for Optics and Photonics, 2019.
- A. J. Van Der Schaft and J. M. Schumacher. *An introduction to hybrid dynamical systems*, volume 251. Springer London, 2000.
- A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin. Attention is all you need. In *Advances in neural information processing systems*, pages 5998–6008, 2017.
- P. Veličković, G. Cucurull, A. Casanova, A. Romero, P. Lio, and Y. Bengio. Graph attention networks. *arXiv preprint arXiv:1710.10903*, 2017.
- E. Weinan. A proposal on machine learning via dynamical systems. *Communications in Mathematics and Statistics*, 5(1):1–11, 2017.
- F. Wu, T. Zhang, A. H. d. Souza Jr, C. Fifty, T. Yu, and K. Q. Weinberger. Simplifying graph convolutional networks. *arXiv preprint arXiv:1902.07153*, 2019a.
- Z. Wu, S. Pan, G. Long, J. Jiang, and C. Zhang. Graph wavenet for deep spatial-temporal graph modeling. *arXiv preprint arXiv:1906.00121*, 2019b.
- L.-P. A. Xhonneux, M. Qu, and J. Tang. Continuous graph neural networks. *arXiv preprint arXiv:1912.00967*, 2019.
- S. Yan, Y. Xiong, and D. Lin. Spatial temporal graph convolutional networks for skeleton-based action recognition. In *Thirty-Second AAAI Conference on Artificial Intelligence*, 2018.
- Ç. Yıldız, M. Heinonen, and H. Lähdesmäki. Ode²vae: Deep generative second order odes with bayesian neural networks. *arXiv preprint arXiv:1905.10994*, 2019.
- J. You, R. Ying, X. Ren, W. L. Hamilton, and J. Leskovec. Graphrnn: Generating realistic graphs with deep auto-regressive models. *arXiv preprint arXiv:1802.08773*, 2018.
- J. You, H. Wu, C. Barrett, R. Ramanujan, and J. Leskovec. G2sat: Learning to generate sat formulas. In *Advances in neural information processing systems*, pages 10553–10564, 2019.
- B. Yu, H. Yin, and Z. Zhu. Spatio-temporal graph convolutional networks: A deep learning framework for traffic forecasting. In *Proceedings of the 27th International Joint Conference on Artificial Intelligence (IJCAI)*, 2018.
- X. Zhao, F. Chen, and J.-H. Cho. Deep learning for predicting dynamic uncertain opinions in network data. In *2018 IEEE International Conference on Big Data (Big Data)*, pages 1150–1155. IEEE, 2018.
- C. Zhuang and Q. Ma. Dual graph convolutional networks for graph-based semi-supervised classification. In *Proceedings of the 2018 World Wide Web Conference*, pages 499–508. International World Wide Web Conferences Steering Committee, 2018.

Continuous–Depth Neural Models for Dynamic Graph Prediction

Supplementary Material

Table of Contents

A Neural Graph Differential Equations	13
A.1 Standalone Neural GDE formulation	13
A.2 Computational Overhead	14
A.3 Oversmoothing in Neural GDEs	14
A.4 Additional Neural GDEs	14
A.5 Hybrid Adjoints	15
B Spatio–Temporal Neural GDEs	15
B.1 GCGRU Cell	16
C Additional experimental details	16
C.1 Multi–Agent System Dynamics	16
C.2 Traffic Forecasting	18
C.3 Repressilator Reconstruction	19

A Neural Graph Differential Equations

Notation Let \mathbb{N} be the set of natural numbers and \mathbb{R} the set of reals. Indices of arrays and matrices are reported as superscripts in round brackets.

Let \mathcal{V} be a finite set with $|\mathcal{V}| = n$ whose element are called *nodes* and let \mathcal{E} be a finite set of tuples of \mathcal{V} elements. Its elements are called *edges* and are such that $\forall e_{ij} \in \mathcal{E}, e_{ij} = (v_i, v_j)$ and $v_i, v_j \in \mathcal{V}$. A graph \mathcal{G} is defined as the collection of nodes and edges, i.e. $\mathcal{G} := (\mathcal{V}, \mathcal{E})$. The *adjacency* matrix $A \in \mathbb{R}^{n \times n}$ of a graph is defined as

$$A^{(ij)} = \begin{cases} 1 & e_{ij} \in \mathcal{E} \\ 0 & e_{ij} \notin \mathcal{E} \end{cases}.$$

If \mathcal{G} is an *attributed graph*, the *feature vector* of each $v \in \mathcal{V}$ is $x_v \in \mathbb{R}^{n_x}$. All the feature vectors are collected in a matrix $X \in \mathbb{R}^{n \times n_x}$. Note that often, the features of graphs exhibit temporal dependency, i.e. $X := X_t$.

A.1 Standalone Neural GDE formulation

For clarity and as an easily accessible reference, we include below a general formulation table for Neural GDEs

Neural GDEs

$$\begin{cases} \dot{Z}_t = f_{\mathcal{G}}(t, Z_t, \theta_t) \\ Z_0 = \ell_{\mathcal{G}}^x(X) \\ \hat{Y}_t = \ell_{\mathcal{G}}^y(Z_t) \end{cases} \quad t \in \mathcal{T}$$

Input node features	X	$\mathbb{R}^{n \times n_x}$
Output	Y	$\mathbb{R}^{n \times n_y}$
Graph	\mathcal{G}	n nodes
Latent node features	Z	$\mathbb{R}^{n \times n_z}$
Parameters	θ	$\mathbb{R}^{n_{\theta}}$
Vector Field	$f_{\mathcal{G}}$	$\mathbb{R} \times \mathbb{R}^{n \times n_z} \times \mathbb{R}^{n_{\theta}} \rightarrow \mathbb{R}^{n \times n_z}$
Input Network	$\ell_{\mathcal{G}}^x$	$\mathbb{R}^{n \times n_x} \rightarrow \mathbb{R}^{n \times n_z}$
Output Network	$\ell_{\mathcal{G}}^y$	$\mathbb{R}^{n \times n_z} \rightarrow \mathbb{R}^{n \times n_y}$

Note that the system provided in (5) can serve as a similar reference for the spatio-temporal case.

A.2 Computational Overhead

As is the case for other models sharing the continuous-depth formulation (Chen et al., 2018), the computational overhead required by Neural GDEs depends mainly by the numerical methods utilized to solve the differential equations. We can define two general cases for *fixed-step* and *adaptive-step* solvers.

Fixed-step In the case of fixed-step solvers of k -th order e.g *Runge-Kutta-k* (Runge, 1895), the time complexity is $O(nk)$ where $n := S/\epsilon$ defines the number of steps necessary to cover $[0, S]$ in fixed-steps of ϵ .

Adaptive-step For general adaptive-step solvers, computational overhead ultimately depends on the error tolerances. While worst-case computation is not bounded (Dormand and Prince, 1980), a maximum number of steps can usually be set algorithmically.

A.3 Oversmoothing in Neural GDEs

For each upper limit of integration in $[1, 5, 10]$, we train 100 *Neural Graph Convolutional Differential Equations* (Neural GCDEs) models on standard static dataset Cora and report average metrics, along with 1 standard deviation confidence intervals in Figure 8. Neural GCDEs are shown to be resilient to these changes; however, with longer integration they require more training epochs to achieve comparable accuracy. This result suggests that Neural GDEs are immune to node oversmoothing (Oono and Suzuki, 2019), as their differential equation is not, unless by design, stable, and thus does not reach an equilibrium state for node representations. Repeated iteration of discrete GNN layers, on the other hand, has been empirically observed to converge to fixed points which make downstream tasks less performant (Chen et al., 2020).

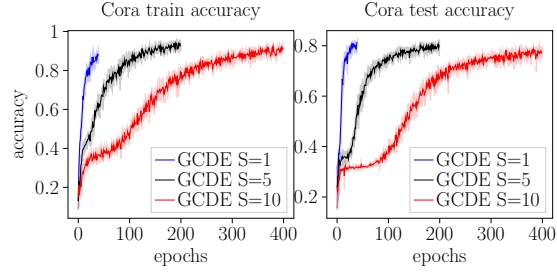


Figure 8: Cora accuracy of Neural GCDE models with different integration intervals S . Longer integration does not result in loss of performance, only an increase in number of training iterations required for convergence.

A.4 Additional Neural GDEs

Message passing Neural GDEs Let us consider a single node $v \in \mathcal{V}$ and define the set of neighbors of v as $\mathcal{N}(v) := \{u \in \mathcal{V} : (v, u) \in \mathcal{E} \vee (u, v) \in \mathcal{E}\}$. Message passing neural networks (MPNNs) perform a spatial-based convolution on the node v as

$$z^{(v)}(s+1) = \phi \left[z^{(v)}(s), \sum_{u \in \mathcal{N}(v)} m \left(z^{(v)}(s), z^{(u)}(s) \right) \right], \quad (6)$$

where, in general, $z^{(v)}(0) = x_v$ while ϕ and m are functions with trainable parameters. For clarity of exposition, let $\phi(x, y) := x + g(y)$ where g is the actual parametrized function. The previous system

(6) becomes

$$z^{(v)}(s+1) = z^{(v)}(s) + g \left[\sum_{u \in \mathcal{N}(v)} m \left(z^{(v)}(s), z^{(u)}(s) \right) \right], \quad (7)$$

and its continuous-depth counterpart, *graph message passing differential equation* (GMDE) is:

$$\dot{z}^{(v)}(s) = f_{\text{MPNN}}^{(v)}(Z, \theta) := g \left[\sum_{u \in \mathcal{N}(v)} m \left(z^{(v)}(s), z^{(u)}(s) \right) \right].$$

Attention Neural GDEs Graph attention networks (GATs) (Veličković et al., 2017) perform convolution on the node v as

$$z^{(v)}(s+1) = \sigma \left(\sum_{u \in \mathcal{N}(v) \cup v} \alpha_{vu} W(s) z^{(u)}(s) \right). \quad (8)$$

Similarly, to GCNs, a *virtual* skip connection can be introduced allowing us to define the *neural graph attention differential equation* (Neural GADE):

$$\dot{z}^{(v)}(s) = f_{\text{GAT}}^{(v)}(Z(s), \theta) := \sigma \left(\sum_{u \in \mathcal{N}(v) \cup v} \alpha_{vu} W z^{(u)}(s) \right), \theta := \text{vec}(W)$$

where α_{vu} are attention coefficient which can be computed following (Veličković et al., 2017).

The attention operator is introduced within drift functions of Latent Neural GDE. Here, visualizing the time-evolution of attention coefficients allows an inspection of neighbouring nodes importance in driving particular node dynamics.

A.5 Hybrid Adjoints

Let us recall the back-propagated adjoint gradients for Hybrid Neural GDEs defined in Section 3.1

$$\begin{aligned} \frac{dL_\theta}{d\theta} &= \frac{\partial L_\theta}{\partial \theta} - \int_{\mathcal{T}} \left\langle \lambda_t, \frac{\partial f_{\mathcal{G}_t}(t, \Phi_t(Z_0), \theta)}{\partial \theta} \right\rangle dt \\ \dot{\lambda}_t &\in F(t, Z_t, \lambda_t) \quad t \in \mathcal{T} \\ \lambda_t^+ &\in G(X_t, Z_t, \lambda_t) \quad t \in \mathcal{T}_e \\ \lambda_{t_K} &= \frac{\partial c_\theta}{\partial Z_{t_K}} \end{aligned}$$

where

$$\begin{aligned} F(t, Z_t, \lambda_t) &:= \left\{ F_k : t \in [t_k, t_{k+1}] \Rightarrow F_k = -\frac{\partial f_{\mathcal{G}_{t_k}}}{\partial Z} \lambda_t \right\} \\ G(X_t, Z_t, \lambda_t) &:= \left\{ G_k : t = t_k \Rightarrow G_k = \lambda_t + \frac{\partial \ell_{\mathcal{G}_{t_k}}^y}{\partial Z_t} + \frac{\partial c_\theta}{\partial Z_t} \right\}. \end{aligned}$$

The above formulation can be derived using the *hybrid inclusions* formalism (Goebel et al., 2009) by extending the results of (Jia and Benson, 2019). In fact, in presence of discontinuities (jumps) in the state the during the forward integration of the (hybrid) differential equation, the adjoint state dynamics becomes itself an hybrid dynamical system with jumps

$$\lambda_t^+ = \lambda_t + \frac{\partial \ell_{\mathcal{G}_{t_k}}^y}{\partial Z_t} + \frac{\partial c_\theta}{\partial Z_t} \text{ if } t = t_k \in \mathcal{T}_e.$$

Due to the time-varying nature of the graph \mathcal{G}_t and, consequently, of the *flow* and *jump* maps $f_{\mathcal{G}_{t_k}}, \ell_{\mathcal{G}_{t_k}}^y$ in the Hybrid Neural GDE, then also the flow and jump maps of the adjoint systems will be different in each interval $[t_k, t_{k+1}]$. The hybrid inclusion representation formalizes this discrete-continuous time-varying nature of the dynamics.

B Spatio-Temporal Neural GDEs

We include a complete description of GCGRUs to clarify the model used in our experiments.

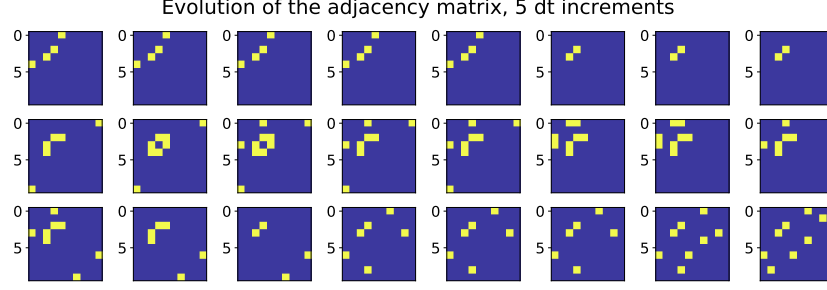


Figure 9: Snapshots of the evolution of adjacency matrix A_t throughout the dynamics of the multi-particle system. Yellow indicates the presence of an edge and therefore a reciprocal force acting on the two bodies

B.1 GCGRU Cell

Following GCGRUs (Zhao et al., 2018), we perform an instantaneous jump of Z at each time t_k using the next input features X_{t_k} . Let $L_{\mathcal{G}_{t_k}}$ be the graph Laplacian of graph \mathcal{G}_{t_k} , which can be computed in several ways (Bruna et al., 2013; Defferrard et al., 2016; Levie et al., 2018; Zhuang and Ma, 2018). Then, let

$$\begin{aligned} H &:= \sigma \left(L_{\mathcal{G}_{t_k}} X_{t_k} W_{xz} + L_{\mathcal{G}_{t_k}} Z W_{hz} \right), \\ R &:= \sigma \left(L_{\mathcal{G}_{t_k}} X_{t_k} W_{xr} + L_{\mathcal{G}_{t_k}} Z W_{hr} \right), \\ \tilde{Z} &:= \tanh \left(L_{\mathcal{G}_{t_k}} X_{t_k} W_{xh} + L_{\mathcal{G}_{t_k}} (R \odot Z) W_{hh} \right). \end{aligned} \quad (9)$$

Finally, the *post-jump* node features are obtained as

$$Z^+ = \text{GCGRU}(Z, X_t) := H \odot Z + (\mathbb{1} - H) \odot \tilde{Z} \quad (10)$$

where W_{xz} , W_{hz} , W_{xr} , W_{hr} , W_{xh} , W_{hh} are matrices of trainable parameters and σ is the standard sigmoid activation and $\mathbb{1}$ is all-ones matrix of suitable dimensions.

C Additional experimental details

Computational resources We carried out all experiments on a cluster of 2x24GB NVIDIA[®] RTX 3090 and CUDA 11.2. The models were trained on GPU. All experiments can be run on a single GPU, as memory requirements never exceeded memory capacity ($\leq 10\text{GB}$).

C.1 Multi-Agent System Dynamics

Dataset Let us consider a planar multi agent system with states x_i ($i = 1, \dots, n$) and second-order dynamics:

$$\ddot{x}_i = -x_i - \sum_{j \in \mathcal{N}_i} f_{ij}(x_i, x_j, \dot{x}_i, \dot{x}_j),$$

where

$$\begin{aligned} f_{ij} &= - \left[\alpha (\|x_i - x_j\| - r) + \beta \frac{\langle \dot{x}_i - \dot{x}_j, x_i - x_j \rangle}{\|x_i - x_j\|} \right] n_{ij}, \\ n_{ij} &= \frac{x_i - x_j}{\|x_i - x_j\|}, \quad \alpha, \beta, r > 0, \end{aligned}$$

and

$$\mathcal{N}_i := \{j : 2\|x_i - x_j\| \leq r \wedge j \neq i\}.$$

The force f_{ij} resembles the one of a spatial spring with drag interconnecting the two agents. The term $-x_i$, is used instead to stabilize the trajectory and avoid the "explosion" of the phase-space. Note that $f_{ij} = -f_{ji}$. The adjacency matrix A_t is computed along a trajectory

Model	MAPE ₁	MAPE ₃	MAPE ₅	MAPE ₁₀	MAPE ₁₅	MAPE ₂₀	MAPE ₅₀
Static	26.12	160.56	197.20	235.21	261.56	275.60	360.39
Neural ODE	26.12	52.26	92.31	156.26	238.14	301.85	668.47
Neural GDE	13.53	15.22	18.76	27.76	33.90	42.22	77.64
Neural GDE-II	13.46	14.75	17.81	27.77	32.28	40.64	73.75

Table 2: Mean MAPE results across the 10 multi-particle dynamical system experiments. MAPE_i indicates results for i extrapolation steps on the full test trajectory.

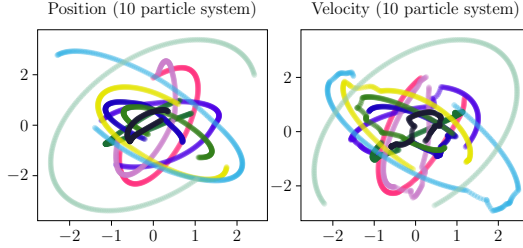


Figure 11: Example position and velocity trajectories of the multi-particle system.

$$A_t^{(ij)} = \begin{cases} 1 & 2\|x_i(t) - x_j(t)\| \leq r \\ 0 & \text{otherwise} \end{cases},$$

which indeed results to be symmetric, $A_t = A_t^\top$ and thus yields an undirected graph. Figure 9 visualizes an example trajectory of A_t , and Figure 11 trajectories of the system. We collect a single rollout with $T = 5$, $dt = 1.95 \cdot 10^{-3}$ and $n = 10$. The particle radius is set to $r = 1$.

Architectural details Node feature vectors are 4 dimensional, corresponding to the dimension of the state, i.e. position and velocity. Neural ODEs and *Static* share an architecture made up of 3 fully-connected layers: $4n, 8n, 8n, 4n$ where $n = 10$ is the number of nodes. The last layer is linear. We evaluated different hidden layer dimensions: $8n, 16n, 32n$ and found $8n$ to be the most effective. Similarly, the architecture of first order Neural GCDEs is composed of 3 GCN layers: 4, 16, 16, 4. Second-order Neural GCDEs, on the other hand, are augmented by 4 dimensions: 8, 32, 32, 8. We experimented with different ways of encoding the adjacency matrix A information into Neural ODEs and *Static* but found that in all cases it lead to worse performance.

Additional results We report in Figure 10 test extrapolation predictions of 5 steps for Neural GDEs and the various baselines. Neural ODEs fail to track the system, particularly in regions of the state space where interaction forces strongly affect the dynamics. Neural GDEs, on the other hand, closely track both positions and velocities of the particles.

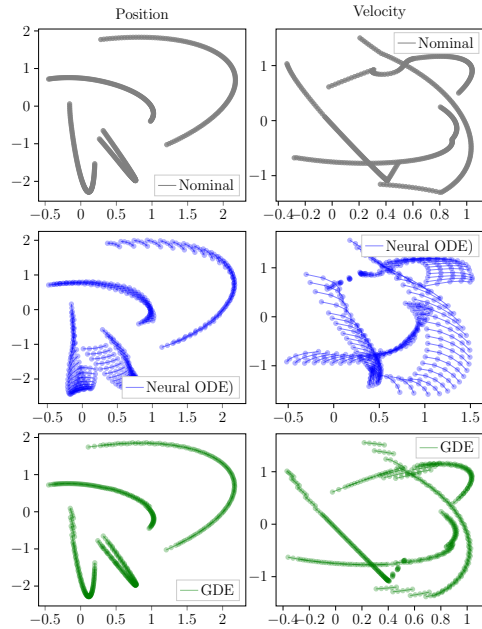


Figure 10: Test extrapolation, 5 steps. Trajectory predictions of Neural ODEs and Neural GDEs. Extrapolation is terminated after 5 steps and the nominal state is fed to the model.

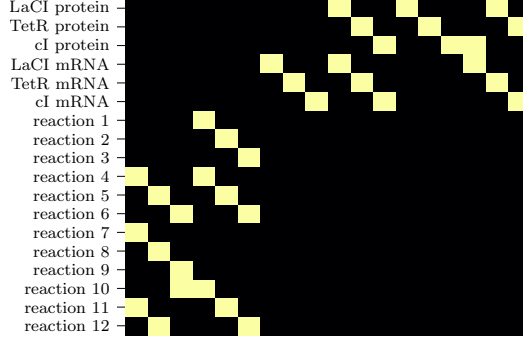


Figure 12: Adjacency matrix of Latent Neural GDE decoders on the repressilator prediction task.

C.2 Traffic Forecasting

Dataset and metrics The timestamp differences between consecutive graphs in the sequence varies due to undersampling. The distribution of timestamp deltas (5 minute units) for the three different experiment setups (30%, 50%, 70% undersampling) is shown in Figure 14.

As a result, GRU takes 230 dimensional vector inputs (228 sensor observations + 2 additional features) at each sequence step. Both GCGRU and GCDE-GRU graph inputs with and 3 dimensional node features (observation + 2 additional feature). The additional time features are excluded for the loss computations. We include MAPE and RMSE test measurements, defined as follows:

$$\text{MAPE}(\mathbf{y}, \hat{\mathbf{y}}) = \frac{100\%}{pT} \left\| \sum_{t=1}^T (y_t - \hat{y}_t) \oslash y_t \right\|_1, \quad (11)$$

where \mathbf{y} , and $\hat{\mathbf{y}} \in \mathbb{R}^p$ is the set of vectorized target and prediction of models respectively. \oslash and $\|\cdot\|_1$ denotes Hadamard division and the 1-norm of vector.

$$\text{RMSE}(\mathbf{y}, \hat{\mathbf{y}}) = \frac{1}{p} \left\| \sqrt{\frac{1}{T} \sum_{t=1}^T (y_t - \hat{y}_t)^2} \right\|_1,$$

where $(\cdot)^2$ and $\sqrt{\cdot}$ denotes the element-wise square and square root of the input vector, respectively. \mathbf{y}_t and $\hat{\mathbf{y}}_t$ denote the target and prediction vector.

Architectural details We employed two baseline models for contextualizing the importance of key components of GCDE-GRU. *GRUs* architectures are equipped with 1 GRU layer with hidden dimension 50 and a 2 layer fully-connected head to map latents to predictions. *GCGRUs* employ a GCGRU layer with 46 hidden dimension and a 2 layer fully-connected head. Lastly, GCDE-GRU shares the same architecture GCGRU with the addition of the flow f_G tasked with evolving the hidden features between arrival times. f_G is parametrized by 2 GCN layers, one with tanh activation and the second without activation. ReLU is used as the general activation function.

Training hyperparameters All models are trained for 40 epochs using Adam (Kingma and Ba, 2014) with $lr = 10^{-2}$. We schedule lr by using cosine annealing method (Loshchilov and Hutter, 2016) with $T_0 = 10$. The optimization is carried out by minimizing the *mean square error* (MSE) loss between predictions and corresponding targets.

Additional results Training curves of the models are presented in the Fig 13. All of models achieved nearly 13 in RMSE during training and fit the dataset. However, due to the lack of dedicated spatial modeling modules, GRUs were unable to generalize to the test set and resulted in a mean value prediction.

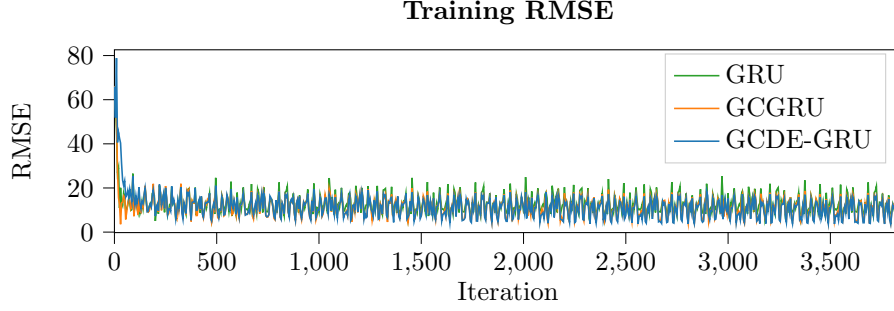


Figure 13: Traffic data *unsmoothed* RMSE during a 50% undersampling training experiment.

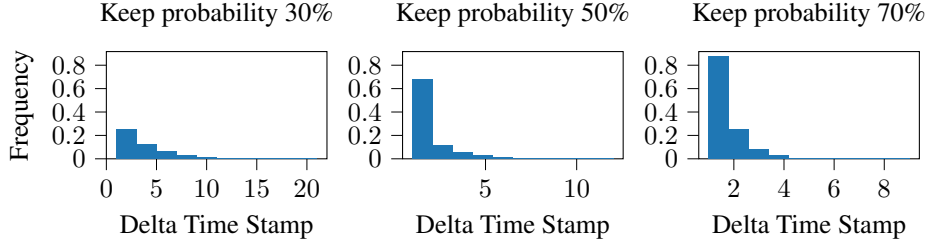


Figure 14: Distribution of deltas between timestamps $t_{k+1} - t_k$ in the undersampled dataset. The time scale of required predictions varies greatly during the task.

C.3 Reprissilator Reconstruction

Dataset We construct a dataset of 10 training and 10 test trajectories of a (Elowitz and Leibler, 2000) Elowitz–Leibler reprissilator system. The data is collected by running stochastic simulations using the τ -leaping method (Gillespie, 2007; Padgett and Ilie, 2016) from an initial condition of $[0, 0, 0, 0, 20, 0]$, where the first three components are concentrations of LaCI, TetR and cI proteins respectively, whereas the last three are concentrations of mRNAs. In the main text, we qualitatively inspect reconstruction capability of the Latent Neural GDE across test trajectories. The samples conditioned on the first 150 seconds of each simulation closely follow the solutions obtained via τ -leaping. The adjacency matrix is constructed as shown in Fig. 12, where the three species of proteins and are disconnected, able to interact only through reaction and mRNA nodes.

Architectural details We employ a Latent Neural GDE model for the reprissilator prediction task. The decoder is comprised of a *neural stochastic graph differential equation* (Neural GSDE) with drift and diagonal diffusion functions constructed as detailed in Table 3. The encoder is constructed with a two layers of *temporal convolutions* (TCNs) and ReLU activations, applied to the 6 protein and mRNA trajectories.

Training hyperparameters Latent Neural GDEs are trained for 1000 epochs using Adam (Kingma and Ba, 2014). We use a one cycle scheduling policy (Smith and Topin, 2019) for the learning rate with maximum at 10^{-2} at 300 epochs and a minimum of $4 \cdot 10^{-4}$. The Neural GSDE decoder is

Layer	Input dim.	Output dim.	Activation
GCN-1	1	3	Tanh
GAT	3	3	Tanh
GCN-2	3	1	None

Table 3: General architecture for Neural GSDEs drift f_G and diffusion g_G functions in decoders of Latent Neural GDEs. The drift function g_G is equipped with an additional sigmoid activation at the end to enforce non-negativity.

solved using an adaptive Euler–Heun ([Kloeden et al., 2012](#)) scheme with tolerances set to 10^{-3} . All data is min–max normalized before being fed into the model.