Learning from Imperfect Data: Robust Inference of Dynamic Systems Using Simulation-Based Generative Model

Hyunwoo Cho[1*], **Hyeontae Jo**[2,3†], **Hyung Ju Hwang**[1,4†]

[1]Department of Mathematics, Pohang University of Science and Technology, 87 Cheongam-Ro, 37673, Pohang, Republic of Korea

[2]Division of Applied Mathematical Sciences, Korea University, 2511 Sejong-ro, 30019, Sejong City, Republic of Korea

[3]Biomedical Mathematics Group, Pioneer Research Center for Mathematical and Computational Sciences, Institute for Basic Science, 55 Expo-ro, Yuseong-gu, 34126, Daejeon, Republic of Korea

[4]AMSquare Corp., 87 Cheongam-Ro, 37673, Pohang, Republic of Korea

Abstract

System inference for nonlinear dynamic models represented by ordinary differential equations (ODEs) remains a significant challenge in many fields, particularly when the data are noisy, sparse, or partially observable. In this paper, we propose a Simulation-based Generative Model for Imperfect Data (SiGMoID), that enables precise and robust inference for dynamic systems. The proposed approach integrates two key methods: (1) HyperPINN, and (2) W-GAN. We demonstrate that SiGMoID quantifies data noise, estimates system parameters, and infers unobserved system components. Its effectiveness is validated by analyzing examples based on realistic experiments, showcasing its broad applicability in various domains, from scientific research to engineered systems, and enabling the discovery of full system dynamics.

Code — https://github.com/CHWmath/SiGMoID

Introduction

Many scientific fields, such as gene regulation (Hirata et al. 2002; Jo et al. 2024), biological rhythms (Forger 2024), disease transmission (Smith et al. 2018; Jung et al. 2020; Hong et al. 2024), and ecology (Busenberg 2012), require the investigation of complex behaviors of d_y system components $\{y_i(t)|i = 1, \ldots, d_y\}$, over time t. The temporal interactions among $y_i(t)$ can be modeled using ordinary differential equations (ODEs) governed by a system function $\mathbf{f} \in \mathbb{R}^{d_y} \times \mathbb{R}^{d_p} \to \mathbb{R}^{d_y}$, as follows:

$$\frac{d\mathbf{y}(t)}{dt} = \mathbf{f}(\mathbf{y}(t), \mathbf{p}), \quad t \in [0, T], \tag{1}$$

where $\mathbf{y} = (y_1, y_2, \ldots, y_{d_y})$ denotes a vector comprising the system components and $\mathbf{p} \in \mathbb{R}^{d_p}$ denotes the d_p-dimensional vector comprising the model parameters to be estimated based on observed (or experimental) data $\mathbf{y}^o(t) = (y_1^o(t), \ldots, y_{d_y}^o(t))$ at N_o observation time points, $t \in \{t_1, \ldots, t_{N_o}\} \subset [0, T]$.

Recent advances in experimental data acquisition have greatly enhanced the ability to monitor dynamic systems, enabling more accurate fitting of solutions to observed data $\mathbf{y}^o(t)$. However, these observations are typically collected at discrete time points and are subject to measurement noise. Accordingly, we model the observed data as:

$$\mathbf{y}^o(t) = \mathbf{y}(t) + \mathbf{e}(t), \tag{2}$$

where $\mathbf{e}(t)$ represents the measurement error governed by the noise level, σ. Additional challenges persist when some elements of the system are difficult to observe due to limitations in measurement resolution (Hirata et al. 2002; Smith et al. 2018; Jo et al. 2024; Hong et al. 2024). In order to address the aforementioned challenges, we classify imperfect datasets, $\mathbf{y}^o(t)$, into two distinct types to capture their characteristics effectively:

- **Noisy and Sparse (NS):** i^{th} component, y_i^o, is observable, for all $i \in \{1, \ldots, d_y\}$, but observations are noisy and recorded at sparse time points (i.e., $\mathbf{y}^o = \mathbf{y} + \mathbf{e}$).
- **NS with Missing Components (NSMC):** A subset of system components $\{y_i^o(t),$ for some $i \in \{1, \ldots, d_y\}\}$ in the NS data is unobservable (See example in Figure 1(a)). We also denote by S_o the set of observable state indices.

To address the challenges associated with NS and NSMC data, we propose a Simulation-based Generative Model for Imperfect Data (SiGMoID). This method utilizes the following two deep-learning models: 1) physics informed neural networks with hyper-networks (HyperPINN), which directly provides solutions of Equation (1) corresponding to any predetermined set of parameters $\mathbf{p}$ (Figure 1(b), and 2) Wasserstein generative adversarial betworks (W-GAN), which adjust the priors by matching the corresponding solutions to imperfect data (Figure 1(c)). By leveraging these two types of models, SiGMoID quantifies the observation noise $\mathbf{e}$ accurately, enabling precise inference of both the underlying system parameters $\mathbf{p}$ and the missing components.

*First author: H. Cho (chw51@postech.ac.kr)

†Corresponding authors: Email: H. Jo (korea_htj@korea.ac.kr), H. J. Hwang (hjhwang@postech.ac.kr)

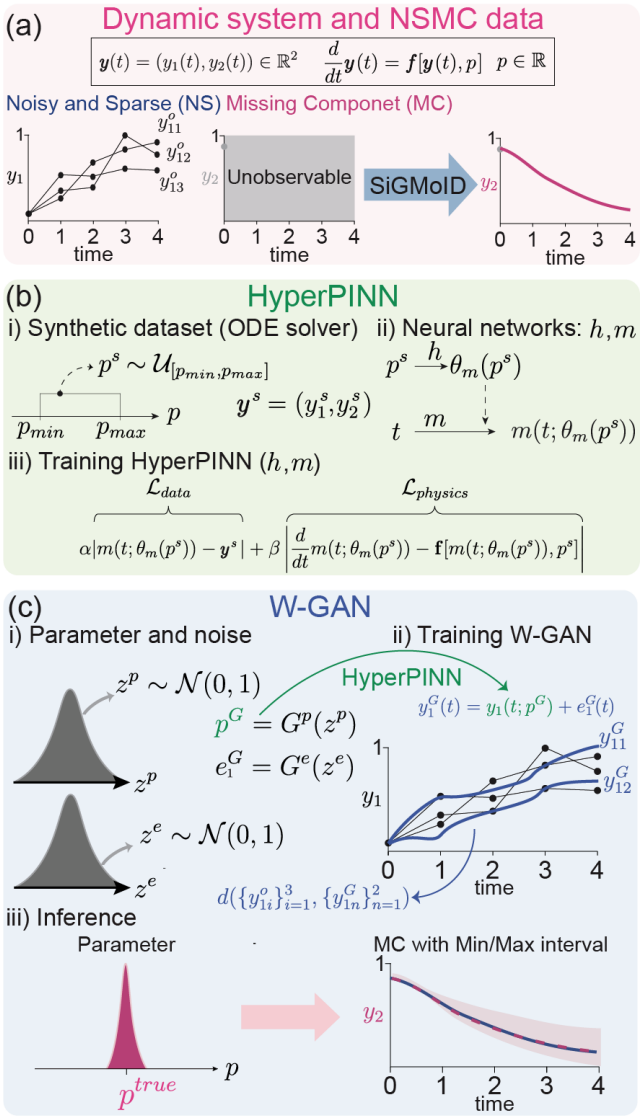

Figure 1: Graphical illustration of the functioning of SiGMoID on NSMC data. (a) ODE system with observable and missing components (b) HyperPINN training from simulated parameter-solution pairs (c) Parameter estimation and missing data recovery using W-GAN

We also evaluate the efficiency of SiGMoID on four distinct types of dynamic systems: 1) FitzHugh–Nagumo (FN) (FitzHugh 1961), 2) Protein transduction (Vyshemirsky and Girolami 2008), 3) Gene regulatory networks (*Hes1*) (Hirata et al. 2002), and 4) Lorenz system (Lorenz 2017; Stepaniants et al. 2024). Our results reveal the following contributions of SiGMoID:

Improved parameter estimation on NS Data: SiGMoID exhibits superior parameter estimation accuracy compared to existing methods on NS data.

Enhanced prediction of full dynamics on NSMC Data: SiGMoID accurately infers unobserved components on NSMC data, overcoming the limitations of current methods that fail to capture these dynamics completely.

Advancing system inference using deep learning-based ODE solvers: The deep learning-based ODE solver utilized in SiGMoID exhibits potential scalability beyond NS and NSMC datasets, making it applicable to a wide range of datasets.

Therefore, we anticipate that SiGMoID will be broadly applicable across domains ranging from scientific research to engineered systems, facilitating comprehensive analyses of system dynamics and offering robust solutions to challenges involving noisy, sparse, or partially observed data.

Related Works

Inference of Dynamic Systems

Inference for dynamic systems has been extensively studied in various fields, and numerous methods have been developed to address the challenges posed by NS data. Among these, the penalized likelihood approach (Ramsay et al. 2007) offers a ODE solution method that bypasses the need for numerical computation. It applies B-spline bases to smooth data, while simultaneously incorporating penalties for deviations from the ODE system. Although effective in many scenarios, this method often requires significant manual tuning, particularly for systems with unobserved components, which diminishes their scalability and ease of implementation.

Another line of research employs Gaussian processes (GPs), which provide a flexible and analytically tractable framework for representing system states, derivatives, and observations, thereby enabling parameter inference (Calderhead, Girolami, and Lawrence 2008; Dondelinger et al. 2013; Wenk et al. 2019; Yang, Wong, and Kou 2021). In this context, (Dondelinger et al. 2013) introduced adaptive gradient matching (AGM), which utilizes GP to approximate gradients and fits them to the gradients defined by the ODE system. Further, (Wenk et al. 2019) developed fast GP-based gradient matching (FGPGM) which effectively reduces computational cost by optimizing the GP framework. More recently, (Yang, Wong, and Kou 2021) introduces manifold-constrained Gaussian process inference (MAGI). MAGI explicitly incorporates the ODE structure into the GP model by conditioning the GP's derivatives on the constraints defined by the ODEs, effectively fitting the nature of the stochastic process to the deterministic dynamics of the system. This approach avoids numerical integration entirely, making it computationally efficient and providing a principled Bayesian framework that ensures consistency between the GP and ODE models. Therefore, to validate the efficiency of the SiGMoID method proposed in this study, we perform a comprehensive comparison with the aforementioned models in terms of their prediction performance on three different examples involving NS and NSMC data.

More recently, deep generative models, such as generative adversarial networks (GANs) (Goodfellow et al. 2014), have been proposed to imitate data based on dynamic systems. For instance, (Kadeethum et al. 2021; Patel, Ray, and Oberai 2022) utilized a GAN to determine the joint probability distribution of parameters and solutions of Equation (1). Sub-

sequently, (Kadeethum et al. 2021) modified the conditional GAN (Mirza and Osindero 2014) to infer parameters based on real data samples. However, GAN-based methods are not explicitly designed to reconstruct unobservable system components, leading to inaccuracies in system inference. Thus, a modification of the GAN architecture and a novel framework that can pre-learn the dynamics systems are required.

Simulation-Based Inference of Dynamic Systems

Similar to the framework proposed in this study, simulation-based inference (SBI) has been developed to infer the joint probability distributions of solutions and corresponding parameters of Equation (1), $\pi(\mathbf{y}, \mathbf{p})$, by incorporating both deep generative models and simulators based on dynamic systems and their underlying parameters, such as numerical DE solvers (Ramesh et al. 2022; Gloeckler et al. 2024), with the aim of inferring the true underlying system parameters based on the observed data. For instance, (Ramesh et al. 2022) successfully applied GANs to estimate the distribution of system parameters. More recently, (Gloeckler et al. 2024) proposed a model that integrates both transformers and denoising diffusion implicit models to predict posterior distributions on sparse datasets.

However, these methods have not yet been applied to NSMC datasets. To address this research desideratum, we propose SiGMoID by modifying the SBI strategy. The proposed framework handles NSMC data by combining an ODE solver capable of learning system dynamics with a W-GAN specifically designed to generate noisy data.

Methods

We develop a Simulation-based Generative Model for Imperfect Data (SiGMoID), which captures system parameters $\mathbf{p}$ in Equation (1) and noise level $\mathbf{e}$ in Equation (2). The detailed procedure is provided in Methodological Details section of the Appendix, along with pseudo-algorithms (Algorithm 1). For better understanding, we present an overview of the key concepts and provide a graphical illustration of SiGMoID on NSMC data in Figure 1, assuming $d_y = 2$, $d_p = 1$, and $N_s = 3$ observed time series in Equation (1) for simplicity. Under this assumption, $\mathbf{y}(t; p^{true}) = \mathbf{y}(t) = (y_1(t), y_2(t)) \in \mathbb{R}^2$ denotes the solution of Equation (1) over time t with a true system parameter $p^{true} \in \mathbb{R}$ (Figure 1(a)). Three NS data corresponding to y_1, $\{y_{1i}^o\}_{i=1}^3$, are given based on observations (NS, blue) at time points $t_j = j - 1$, for $j = 1, 2, 3, 4, 5$. However, the data corresponding to y_2 are not observable (that is, MC). To address this, the SiGMoID framework is utilized to infer the noise distribution of y_1, estimate the true parameter p^{true}, and reconstruct the underlying true solution y_2.

In SiGMoID framework, we first set the parameter boundary, $[p_{\min}, p_{\max}]$ based on the empirically determined feasible range of true parameter values (Figure 1(b)-i)). A random parameter value p^s is then sampled from the uniform distribution $\mathcal{U}_{[p_{\min}, p_{\max}]}$. Using a numerical solver, the solution $\mathbf{y}^s(t) = \mathbf{y}(t; p^s)$ of the system considered in (Figure 1(a)) is computed with p^s. Subsequently, the pair $(p^s, \mathbf{y}^s)$ is used to train the HyperPINN framework, which consists of two neural networks—a hypernetwork h and a main network m (Figure 1(b)-ii)). The hypernetwork h takes the sampled parameter p^s as its input and generates weights and biases $\theta_m(p^s)$ for the main network m. The main network $m^s(t) = m(t; \theta_m(p^s))$ produce the solution $\mathbf{y}^s(t)$, for an arbitrary time $t \in [0, T]$. For given time collocation points $\{t_j^c\}_{j=1}^{T_{col}}$, the networks h and m are trained jointly by minimizing the weighted sum of data loss $\mathcal{L}_{data}$ and physics loss $\mathcal{L}_{physics}$ with distinct weights $\alpha, \beta > 0$ (Figure 1(b)-iii)):

$$\mathcal{L}_{data} = \sum_{j=1}^{T_{col}} \left\| m^s(t_j^c) - \mathbf{y}(t_j^c; p^s) \right\|^2,$$

$$\mathcal{L}_{physics} = \sum_{j=1}^{T_{col}} \left\| \frac{d}{dt} m^s(t_j^c) - \mathbf{f}\left[m^s(t_j^c), p^s\right] \right\|^2,$$

$$\mathcal{L} = \alpha \mathcal{L}_{data} + \beta \mathcal{L}_{physics},$$

where $\|\mathbf{y}\|^2$ denotes L^2 norm $\|\mathbf{y}\|^2 = y_1^2 + \cdots + y_{d_y}^2$.

For W-GAN step (Figure 1(c)), latent variables z^p and z^e are sampled from the standard normal distribution, $\mathcal{N}(0, 1)$ (Figure 1(c)-i)). These serve as inputs to the two generators G^p and G^e, to yield a candidate system parameter p^G and data noise e_1^G for $\{y_{1i}^o\}_{i=1}^3$, respectively. Using the HyperPINN trained in (b)-ii), the solution $y_1^G = m_1^G + e_1^G$ corresponding to y_{1i} is then obtained(Figure 1(c)-ii)). By repeating this procedure N times, each time sampling random parameters and noise values (Figure 1(c)-i) ($N = 2$ for simplicity), we obtain N samples $\{(p_n^G, y_{1n}^G)\}_{n=1}^N$ from the generator. The discrepancy between $\{y_{1i}^o\}_{i=1}^3$ and $\{y_{1n}^G\}_{n=1}^2$ is measured in terms of the Wasserstein distance

$$d(\{y_{1i}^o\}_{i=1}^3, \{y_{1n}^G\}_{n=1}^2)$$
$$= \sup_{\phi \in \mathrm{Lip}_1} \mathbb{E}_{\mathbf{y}^o}[\phi(\{y_{1i}^o\}_{i=1}^3)] - \mathbb{E}_{\mathbf{y}^G}[\phi(\{y_{1n}^G\}_{n=1}^2)],$$

where $\{y_{1i}^o\}_{i=1}^3 \sim \mathbf{y}^o$ and $\{y_{1n}^G\}_{n=1}^2 \sim \mathbf{y}^G$ denote distributions of observed and generated outputs, respectively. The W-GAN is trained by minimizing this objective. Subsequently, the parameter values $\{p_n^G\}_{n=1}^2$ generated by G^p are expected to converge within a narrow region containing the true system parameter p^{true} (Figure 1(c)-iii), (left)). Using the parameter distribution, the numerical solver is reapplied to generate MC data for y_2 (right). A general description of this procedure is provided in Methodological Details in Appendix, and detailed descriptions of HyperPINN and GAN, including stability measures, hyperparameters, and architecture configurations, are presented in Table A1 to Table A2.

Experiments

We apply SiGMoID to four real-world problems, requiring inference of system parameters and reconstruction of MC datasets: 1) the FitzHugh–Nagumo (FN) model (FitzHugh 1961), 2) the protein transduction model (Vyshemirsky and Girolami 2008), 3) the *Hes1* model (Hirata et al. 2002), and 4) Lorenz system (Lorenz 2017). For each example, the NS and NSMC datasets are generated using a numerical solver (explicit Runge-Kutta of order four) based on the simulation configurations obtained from the references. In particular,

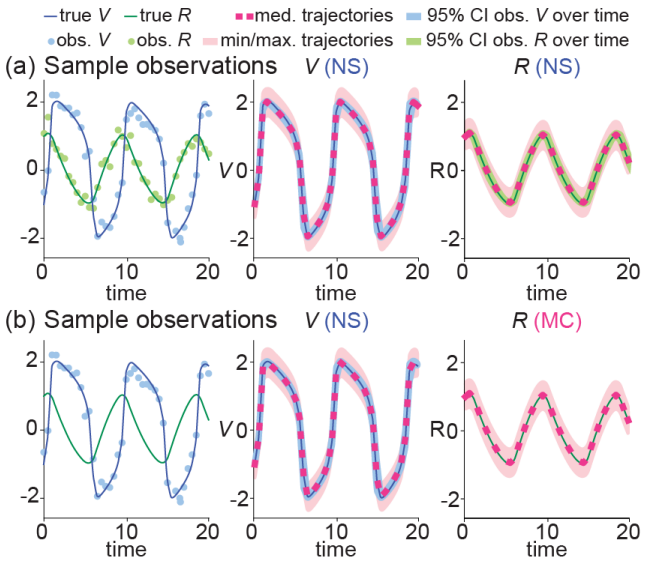

Figure 2: System inference for the FN equation. Demonstration of the capability of SiGMoID to infer true system solutions (V and R) for the FN model. (a) NS datasets for V and R are provided (sample observations - blue and green dots). SiGMoID infers the true system solutions (V (NS), R (NS) - red dashed line) accurately. The red region represents the range of the solutions inferred using SiGMoID, while the blue region represents the 95% confidence interval (CI) of the observed component over time. (b) The experimental dataset for R is missing, while the analogue for V is available (Sample observations - blue dots). Despite the absence of R in the dataset, SiGMoID fits both the true V and true R successfully. The CI for the observed R is omitted owing to the absence of any corresponding dataset.

the simulations adopt the parameter configurations and initial conditions established in previous studies. The results of this evaluation demonstrate that SiGMoID not only outperforms existing methods (Yang, Wong, and Kou 2021; Stepaniants et al. 2024), but also excels at reconstructing unobserved components. The computational costs for all cases are summarized in Table A7 in the Appendix.

FitzHugh–Nagumo

Spike potentials in ion channels can be interpreted as the interaction between the membrane potential voltage of a neuron V and the recovery variable associated with the neuron current, R. This relationship is described by the FN equations for $\mathbf{y} = (V, R)$. The equations for $\mathbf{y} = (V, R)$ are expressed as follows:

$$f(\mathbf{y}, \mathbf{p}) = \begin{pmatrix} c\left(V - \frac{V^3}{3} + R\right) \\ -\frac{1}{c}\left(V - a + bR\right) \end{pmatrix},$$

where the set of parameters $\mathbf{p} = (a, b, c)$ indicates the equilibrium voltage level for the system a, the coupling strength between the recovery variable and the membrane potential b, and the timescale and sensitivity of the voltage dynamics c.

DATA	METHOD	V	R
NS	SIGMOID	**0.028**	**0.012**
	MAGI	0.103	0.070
	FGPGM	0.257	0.094
	AGM	1.177	0.662
NSMC	SIGMOID	0.038	0.016

Table 1: Trajectory RMSEs of each component in the FN system, comparing the average trajectory RMSE values of the four methods over one hundred simulated datasets.

To generate sample observations, we initialize the true parameters values as $\mathbf{p}^{true} = (0.2, 0.2, 3)$ and set the initial condition to be $\mathbf{y}(0) = (-1, 1)$ (FitzHugh 1961). Using these conditions, we compute the true underlying trajectories (Figure 2(a), Sample observations-true V, R). Next, one hundred observed trajectories for V and R are generated corresponding to a noise level of 0.2 using 41 observation points to construct the NS dataset (Figure 2(a), Sample observations-observed V, R) (See also the detailed scenario reported in (Dondelinger et al. 2013; Wenk et al. 2019)). Under this configuration, the parameter estimation problem becomes non-identifiable (See Figure A1 in Appendix for details).

SiGMoID is used to infer both the true underlying trajectories (Figure 2(a), V and R) and the parameters (Table A3, NS). It is observed to exhibit the lowest root mean square error (RMSE), $\frac{1}{N_s}\sum_{j=1}^{N_s}\sqrt{\frac{1}{N_t}\sum_{k=1}^{N_o}|y_{true,i}(t_k) - y_i^G(t_k)|^2}$, between the true $y_{true,i}(t) = y_i(t; p^{true})$ and estimated trajectories $y_{ij}^G(t) = y_i(t; p_j^G)$ for each component i compared to the other three methods—MAGI (Yang, Wong, and Kou 2021), FGPGM (Wenk et al. 2019), and AGM (Dondelinger et al. 2013) (Table 1, NS). Further, the parameters inferred using SiGMoID are observed to be the closest to the true values compared to those obtained using the other methods. These results demonstrate that SiGMoID not only accurately reconstructs the true underlying trajectories but also outperforms existing methods (MAGI, FGPGM, and AGM) in terms of estimation accuracy, establishing its effectiveness and reliability. Although the neuronal potential voltage V is easily observable, R cannot be directly measured experimentally and must be inferred based on the dynamic system (FitzHugh 1961; Samson, Tamborrino, and Tubikanec 2025). To this end, we apply SiGMoID to the NSMC dataset by removing the sample observations for R in Figure 2(a) (Figure 2(b), sample observations). Despite the absence of observed data for R, the parameters inferred using SiGMoID based on the NSMC dataset are observed to be accurate and closely aligned with the true values Table A3, NSMC).

Protein transduction

The dynamics of signaling proteins and their interactions in biochemical pathways can be described using a reaction network with variables $\mathbf{y} = (S, S_d, R, S_R, R_{pp})$. Specifically, the signaling protein S interacts with the receptor protein R to form the complex S_R with a binding rate of k_2. The

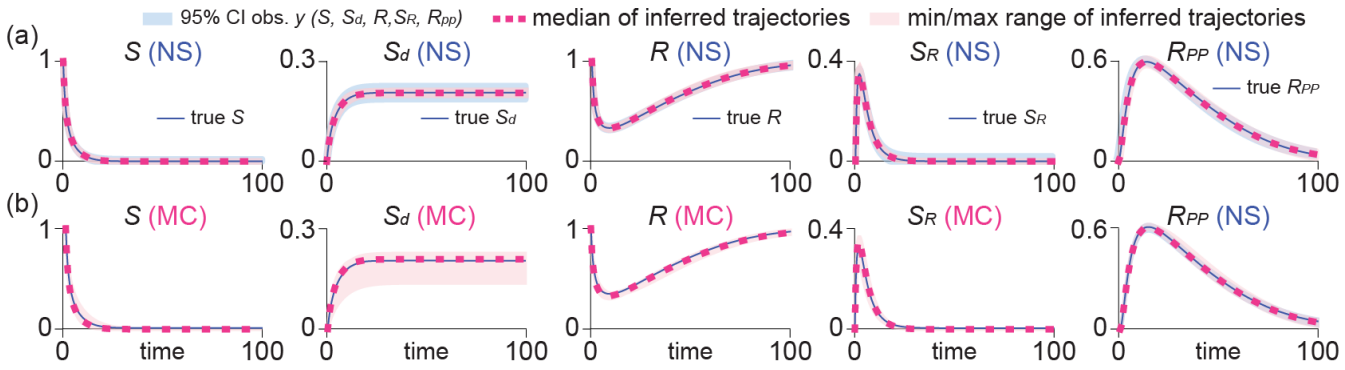

Figure 3: System inference for a protein transduction model. We infer the true system solutions (true S, S_d, R, S_R, and R_{pp}) using SiGMoID. (a) When all the components are given in the NS dataset, the solutions inferred using SiGMoID match the true solutions accurately. (b) Similar to (a), the solutions inferred using SiGMoID match the true solutions accurately, even though the components, S, S_d, R, and S_R are not observable (MC).

complex S_R can dissociate back into S and R with a dissociation rate of k_3. In addition, S_R facilitates the activation of R into its phosphorylated form R_{pp} with an activation rate of k_4. The signaling protein S also degrades into its inactive form S_d with a degradation rate of k_1. Finally, the activated receptor R_{pp} is deactivated via Michaelis-Menten kinetics, where V represents the maximum deactivation rate and K_m denotes the Michaelis constant, which defines the concentration of R_{pp} at which the deactivation rate becomes half of its maximum value. These dynamics are described by the following system of ODEs with $\mathbf{p} = (k_1, k_2, k_3, k_4, V, K_m)$:

$$f(\mathbf{y}, \mathbf{p}) = \begin{pmatrix} -k_1 S - k_2 SR + k_3 S_R \\ k_1 S \\ -k_2 SR - k_3 S_R + \frac{V R_{pp}}{K_m + R_{pp}} \\ k_2 SR - k_3 S_R - k_4 S_R \\ k_4 S_R - \frac{V R_{pp}}{K_m + R_{pp}} \end{pmatrix}.$$

To generate sample observations, we set $\mathbf{p}^{true} = (0.07, 0.6, 0.05, 0.3, 0.017, 0.3)$ and $\mathbf{y}(0) = (1, 0, 1, 0, 0)$. Using these conditions, the true underlying trajectories are computed (Figure 3(a), Sample observations-true S, S_d, R, S_R, R_{pp}). Subsequently, one hundred observed trajectories are generated corresponding to a noise level of 0.01, and 26 observation time points are used to construct the NS dataset (Figure 3(a), Sample observations- 95% CI of obs. S, S_d, R, S_R, R_{pp}). Notably, the selected noise level represents a severely challenging scenario, as described in (Vyshemirsky and Girolami 2008).

Subsequently, both the true underlying trajectories (Figure 3(b), S, S_d, R, S_R and R_{pp}) and the parameters are inferred using SiGMoID. As demonstrated previously, SiGMoID achieves the lowest RMSE values between the true and estimated trajectories compared to the other three methods (MAGI, FGPGM, and AGM) (Table 2, NS). Further, identifiability problems are observed in the parameter estimates for the protein transduction model (Dondelinger et al. 2013; Wenk et al. 2019), where the system exhibits low sensitivity to certain parameters, complicating the unique determination of their values based on the observed data. Despite this phenomenon, the parameter estimates obtained using SiGMoID are observed to match the true values most accurately Table A4, NS). Similar to the application of SiGMoID to the FN model, we obtain the NSMC dataset by removing sample observations corresponding to all components except R_{pp} (NS) (Figure 3(b)), which is consistent with the experimental results reported in (Vyshemirsky and Girolami 2008). Notably, this unobserved component renders the parameter inference problem non-identifiable, (See Figure A2 in Appendix for details). Despite the absence of all components except R_{pp}, SiGMoID exhibits robust performance in inferring both trajectories (Figure 3(b), S, S_d, R, S_R, R_{pp} and Table 2 NSMC) and parameters Table A4, NSMC).

DATA	METHOD	S	S_d	R	S_R	R_{pp}
NS	SIGMOID	**0.3**	**0.8**	**1.2**	**0.4**	**1.3**
	MAGI	12.2	4.3	16.7	13.5	13.6
	FGPGM	12.8	8.9	21.0	13.6	30.9
	AGM	67.1	312.5	413.8	98.0	297.3
NSMC	SIGMOID	11.2	11.2	4.9	3.8	2.9

Table 2: Trajectory RMSEs of each component in the protein transduction system, comparing the average trajectory RMSE of the four methods over one hundred simulated datasets. (scaled by $\times 10^3$)

Hes1 model

METHOD	P	M	H
SIGMOID	**0.38**	**0.12**	**1.96**
MAGI	0.97	0.21	2.57
(RAMSAY ET AL. 2007)	1.30	0.40	59.47

Table 3: Trajectory RMSEs of each component in the hes1 system, comparing the average trajectory RMSE of the three methods over $2,000$ simulated datasets. Note that H denotes the missing component.

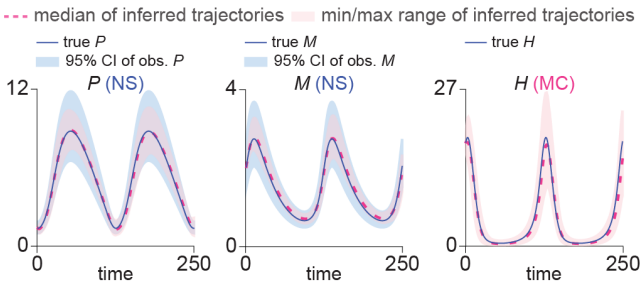

Figure 4: System inference for a *Hes1* model. Datasets for the two components, P and M, in the *Hes1* model are provided as NS data, while the component, H, is unobserved (MC). SiGMoID accurately infers not only the true system solutions for all components but also captures the noise in the data (95% CI of obs. P and M).

The oscillation of *Hes1* mRNA M and *Hes1* protein P levels in cultured cells can be described using a dynamic system (Hirata et al. 2002). This system postulates the involvement of an interacting factor H that contributes to stable oscillations, representing a biological rhythm (Forger 2024). The system of ODEs for the three-component state vector $\mathbf{y} = (P, M, H)$ is expressed as follows:

$$f(\mathbf{y}, \mathbf{p}) = \begin{pmatrix} -aPH + bM - cP \\ -dM + \frac{e}{1+P^2} \\ -aPH + \frac{f}{1+P^2} - gH \end{pmatrix}, \quad (3)$$

where $\mathbf{p} = (a, b, c, d, e, f, g)$ are the associated parameters.

The *Hes1* model is a prime example of the challenges of inference in systems with unobserved components and asynchronous observation times. In (Hirata et al. 2002), the theoretical oscillatory behavior of the system was established using the true parameter values $a = 0.022, b = 0.3, c = 0.031, d = 0.028, e = 0.5, f = 20$, and $g = 0.3$, revealing an oscillation cycle of approximately 2 hours. The initial conditions were defined as the lowest value of P during oscillatory equilibrium, with $P = 1.439, M = 2.037$, and $H = 17.904$. To construct the NSMC dataset from the true solution in this study, a noise level of $\sigma = 0.15$ is adopted based on (Yang, Wong, and Kou 2021). This choice is aligned with the reported standard errors, which account for approximately 15% of P (protein) and M (mRNA) levels in repeated measurements. Accordingly, simulation noise is modeled as multiplicative, using a lognormal distribution with $\sigma = 0.15$. Because of the multiplicative nature of the error in this strictly positive system, a log-transformation is applied to Equation (3) to produce Gaussian error distributions. Additionally, the component H is treated as unobserved and excluded from the simulated observations. The absence of this component leads to a non-identifiable parameter inference problem (See Figure A3 in Appendix).

Using the above setting, $2,000$ simulated datasets are generated. Three methods (SiGMoID, MAGI, and (Ramsay et al. 2007)) are then implemented on each dataset to infer system solutions and estimate system parameters (Figure 4). Among these methods, SiGMoID accurately infers both observable (P, M) and unobservable (H) true solutions. Further, compared to the two other methods, SiGMoID achieves the lowest RMSE values related to the difference between the inferred and true trajectories (Table 3).

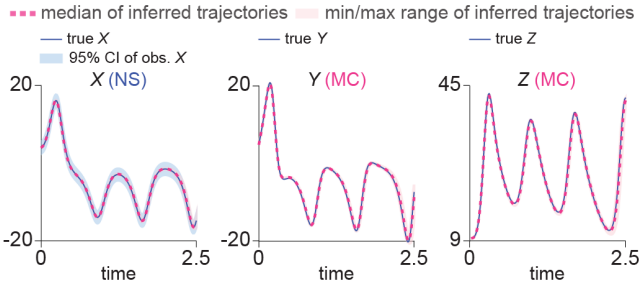

Figure 5: System inference for a Lorenz equation. We infer the true system solutions (true X, Y and Z) using SiGMoID. (a) When all the components are given in the NS dataset, the solutions inferred using SiGMoID match the true solutions accurately. (b) Similar to (a), the solutions inferred using SiGMoID match the true solutions accurately, even though the components, Y and Z are not observable (MC).

We also compare the parameter estimates results in Table A5. Note that while the temporal dynamics of the two observable components, P and M, are influenced by various combinations of four parameters (b, c, d, and e), the unobservable component H is determined exclusively by the other three parameters (a, f, and g), leading to an identifiability problem (Yang, Wong, and Kou 2021). This shows that our estimation approach is well-suited for inferring H, as it can be accurately reconstructed once a, f, and g are correctly determined.

Lorenz system

The Lorenz system was proposed to model simplified atmospheric convection using three key variables, $\mathbf{y} = (X, Y, Z)$ (Lorenz 2017). In this formulation, X denotes the intensity of convective motion, Y represents the temperature difference between rising and sinking air flows within a convection cell, and Z indicates the deviation from thermal equilibrium or, more specifically, the vertical temperature distribution in the system:

$$f(\mathbf{y}, \mathbf{p}) = \begin{pmatrix} \sigma(Y - X) \\ X(\rho - Z) - Y \\ XY - \beta Z \end{pmatrix},$$

where $\mathbf{p} = (\sigma, \rho, \beta)$ represents physical parameters with specific interpretations: the Prandtl number σ controls the ratio of fluid viscosity to thermal diffusivity, the Rayleigh number ρ quantifies the driving force of convection due to temperature differences, and the parameter β is associated with the damping effect on convective motion.

We examine the capability of SiGMoID to estimate the parameters of the Lorenz system under the NSMC setting, where two system components, Y and Z are unobservable (Figure 5). The Lorenz parameters are set to $\sigma = 10$, $\rho =$

28, and $\beta = 8/3$, with the initial condition set to $\mathbf{y}(0) = (4.67, 5.49, 9.06)$ (Hirsch, Smale, and Devaney 2013).

In this setting, we first generate the NS dataset by integrating the Lorenz system using the RK4 method, adding additive noise with a noise level of 0.1. For the NSMC dataset, we omit both Y and Z variables. This setup is motivated by the framework introduced in (Stepaniants et al. 2024), and such missing components still introduce challenges for parameter identifiability (See Figure A4 in Appendix). Next, one hundred observed trajectories for X is generated using 9 observation points.

Although Y and Z are remained unobserved (Figure 5), SiGMoID accurately infers both unobservable states (Y, Z). Compared to the result of (Stepaniants et al. 2024), the estimation result obtained by SiGMoID shows significantly improved accuracy in capturing the difference between the inferred and true trajectories. Furthermore, the inferred parameters closely match the true values, as demonstrated in Table A6 in Appendix.

Disucssion

SiGMoID enables accurate parameter estimation, robust noise quantification, and precise inference of unobserved system components across various synthetic examples by integrating HyperPINN with W-GAN. These experimental results indicate the potential applicability of existing deep learning-based models across diverse fields. In particular, the proposed framework shows strong potential to address real-world NSMC challenges. For example, in virology, only viral load may be observed in target-cell limited datasets (Smith et al. 2018); in epidemiological modeling, the number of individuals in the latent stage may not be recorded over time (Hong et al. 2024); and in cell-signaling studies, intermediate steps in the pathway are often hidden (Jo et al. 2024). In such diverse fields, full system observations are rarely available, highlighting the potential of SiGMoID to contribute to data-driven scientific discovery under incomplete or partially observed conditions.

The main task of SiGMoID is not to develop a new surrogate model, but to reconstruct missing components from incomplete (NSMC) data by combining two modules: (i) a surrogation model (HyperPINN) and (ii) an inference module (W-GAN). While HyperPINN was used as the surrogate model in this study, this choice is not essential. Depending on the characteristics of the underlying dynamical system, other surrogate architectures such as Fourier neural operators (Li et al. 2020) can be employed to further improve computational efficiency and scalability. This modular design underpins the generality of SiGMoID and supports its applicability to a broad class of dynamical systems beyond the four examples presented in this paper.

The use of a HyperPINN-based ODE solver significantly enhanced simulation efficiency by quickly generating solutions of differential equations across a wide range of parameter settings. This addresses the computational inefficiencies inherent in traditional PINN methods, offering scalability for high-dimensional or real-time system analysis. However, there are clear areas for improvement in this approach. Specifically, training HyperPINNs requires a priori knowledge of the distribution of parameters within the dataset (de Avila Belbute-Peres, Chen, and Sha 2021). The number of artificial trajectories required for training grows proportionally with the size of the parameter space, leading to increased computational cost and training time. This limitation highlights an ongoing challenge in the field of deep learning-based operator learning. Consequently, the overall efficiency of SiGMoID can be further enhanced in parallel with advances in ODE solver methodologies (e.g., (Lee, Ko, and Hong 2025)).

Recent generative models such as diffusion and flow-matching methods (Ho, Jain, and Abbeel 2020; Lipman et al. 2022) excel in high-dimensional data generation but are designed for data-to-data settings, unlike our parameter-to-data inference task. These models also require large datasets, which is unrealistic in biological contexts where fewer than one hundred trajectories are typically available. To address this data-scarce setting, we adopt W-GAN, which provides stable distribution matching and shows robust performance in our experiments.

A key future direction is extending the framework to cases where the governing equations are partially or entirely unknown. Existing approaches such as SINDy and neural ODEs require strong prior assumptions or suffer from identifiability issues when multiple dynamics fit the same data (Champion et al. 2019; Chen et al. 2018). Transformer-based structure-discovery models (d'Ascoli et al. 2023) show promise but remain sensitive to noise and limited data. Developing hybrid methods that combine such models with robust inference techniques and domain-specific priors is an important but challenging direction.

Conclusion

In this study, we introduce SiGMoID (Simulation-based Generative Model for Imperfect Data), a framework designed to handle noisy, sparse, and partially observed data in dynamic systems. By integrating HyperPINN with W-GAN, SiGMoID enables simultaneous inference of system parameters and reconstruction of missing dynamics, achieving superior accuracy and robustness compared to conventional methods (Dondelinger et al. 2013; Wenk et al. 2019; Yang, Wong, and Kou 2021; Stepaniants et al. 2024). Therefore, these results show that deep learning–based inference models can be broadly applicable across various domains, even when full system observations are scarce. Consequently, SiGMoID offers a powerful approach for data-driven scientific discovery under incomplete or partially observed conditions.

Acknowledgments

Hyung Ju Hwang was supported by the National Research Foundation of Korea (NRF) grant funded by the Korea government (MSIT) (RS-2023-00219980 and RS-2022-00165268) and by Institute for Information & Communications Technology Promotion (IITP) grant funded by the Korea government (MSIP) (No. RS-2019-II191906, Artificial Intelligence Graduate School Program (POSTECH)).

Hyeontae Jo was supported by the National Research Foundation of Korea (RS-2024-00357912) and a Korea University grant (K2418321, K2425881).

References

Busenberg, S. 2012. *Differential Equations and Applications in Ecology, Epidemics, and Population Problems*. Elsevier.

Calderhead, B.; Girolami, M.; and Lawrence, N. 2008. Accelerating Bayesian inference over nonlinear differential equations with Gaussian processes. *Advances in neural information processing systems*, 21.

Champion, K.; Lusch, B.; Kutz, J. N.; and Brunton, S. L. 2019. Data-driven discovery of coordinates and governing equations. *Proceedings of the National Academy of Sciences*, 116(45): 22445–22451.

Chen, R. T.; Rubanova, Y.; Bettencourt, J.; and Duvenaud, D. K. 2018. Neural ordinary differential equations. *Advances in neural information processing systems*, 31.

d'Ascoli, S.; Becker, S.; Mathis, A.; Schwaller, P.; and Kilbertus, N. 2023. Odeformer: Symbolic regression of dynamical systems with transformers. *arXiv preprint arXiv:2310.05573*.

de Avila Belbute-Peres, F.; Chen, Y.-f.; and Sha, F. 2021. HyperPINN: Learning parameterized differential equations with physics-informed hypernetworks. In *The symbiosis of deep learning and differential equations*.

Dondelinger, F.; Husmeier, D.; Rogers, S.; and Filippone, M. 2013. ODE parameter inference using adaptive gradient matching with Gaussian processes. In *Artificial intelligence and statistics*, 216–228. PMLR.

FitzHugh, R. 1961. Impulses and physiological states in theoretical models of nerve membrane. *Biophysical journal*, 1(6): 445–466.

Forger, D. B. 2024. Biological clocks, rhythms, and oscillations: the theory of biological timekeeping.

Gloeckler, M.; Deistler, M.; Weilbach, C.; Wood, F.; and Macke, J. H. 2024. All-in-one simulation-based inference. *arXiv preprint arXiv:2404.09636*.

Goodfellow, I.; Pouget-Abadie, J.; Mirza, M.; Xu, B.; Warde-Farley, D.; Ozair, S.; Courville, A.; and Bengio, Y. 2014. Generative adversarial nets. *Advances in neural information processing systems*, 27.

Hirata, H.; Yoshiura, S.; Ohtsuka, T.; Bessho, Y.; Harada, T.; Yoshikawa, K.; and Kageyama, R. 2002. Oscillatory expression of the bHLH factor Hes1 regulated by a negative feedback loop. *Science*, 298(5594): 840–843.

Hirsch, M. W.; Smale, S.; and Devaney, R. L. 2013. *Differential equations, dynamical systems, and an introduction to chaos*. Academic press.

Ho, J.; Jain, A.; and Abbeel, P. 2020. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33: 6840–6851.

Hong, H.; Eom, E.; Lee, H.; Choi, S.; Choi, B.; and Kim, J. K. 2024. Overcoming bias in estimating epidemiological parameters with realistic history-dependent disease spread dynamics. *Nature Communications*, 15(1): 8734.

Jo, H.; Hong, H.; Hwang, H. J.; Chang, W.; and Kim, J. K. 2024. Density physics-informed neural networks reveal sources of cell heterogeneity in signal transduction. *Patterns*, 5(2).

Jung, S. Y.; Jo, H.; Son, H.; and Hwang, H. J. 2020. Real-world implications of a rapidly responsive COVID-19 spread model with time-dependent parameters via deep learning: Model development and validation. *Journal of medical Internet research*, 22(9): e19907.

Kadeethum, T.; O'Malley, D.; Fuhg, J. N.; Choi, Y.; Lee, J.; Viswanathan, H. S.; and Bouklas, N. 2021. A framework for data-driven solution and parameter estimation of pdes using conditional generative adversarial networks. *Nature Computational Science*, 1(12): 819–829.

Lee, J. Y.; Ko, S.; and Hong, Y. 2025. Finite Element Operator Network for Solving Elliptic-Type Parametric PDEs. *SIAM Journal on Scientific Computing*, 47(2): C501–C528.

Li, Z.; Kovachki, N.; Azizzadenesheli, K.; Liu, B.; Bhattacharya, K.; Stuart, A.; and Anandkumar, A. 2020. Fourier neural operator for parametric partial differential equations. *arXiv preprint arXiv:2010.08895*.

Lipman, Y.; Chen, R. T.; Ben-Hamu, H.; Nickel, M.; and Le, M. 2022. Flow matching for generative modeling. *arXiv preprint arXiv:2210.02747*.

Lorenz, E. N. 2017. Deterministic Nonperiodic Flow 1. In *Universality in Chaos, 2nd edition*, 367–378. Routledge.

Mirza, M.; and Osindero, S. 2014. Conditional generative adversarial nets. *arXiv preprint arXiv:1411.1784*.

Patel, D. V.; Ray, D.; and Oberai, A. A. 2022. Solution of physics-based Bayesian inverse problems with deep generative priors. *Computer Methods in Applied Mechanics and Engineering*, 400: 115428.

Ramesh, P.; Lueckmann, J.-M.; Boelts, J.; Tejero-Cantero, Á.; Greenberg, D. S.; Gonçalves, P. J.; and Macke, J. H. 2022. GATSBI: Generative adversarial training for simulation-based inference. *arXiv preprint arXiv:2203.06481*.

Ramsay, J. O.; Hooker, G.; Campbell, D.; and Cao, J. 2007. Parameter estimation for differential equations: a generalized smoothing approach. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 69(5): 741–796.

Samson, A.; Tamborrino, M.; and Tubikanec, I. 2025. Inference for the stochastic FitzHugh-Nagumo model from real action potential data via approximate Bayesian computation. *Computational Statistics & Data Analysis*, 204: 108095.

Smith, A. P.; Moquin, D. J.; Bernhauerova, V.; and Smith, A. M. 2018. Influenza virus infection model with density dependence supports biphasic viral decay. *Frontiers in microbiology*, 9: 1554.

Stepaniants, G.; Hastewell, A. D.; Skinner, D. J.; Totz, J. F.; and Dunkel, J. 2024. Discovering dynamics and parameters of nonlinear oscillatory and chaotic systems from partial observations. *Physical Review Research*, 6(4): 043062.

Vyshemirsky, V.; and Girolami, M. A. 2008. Bayesian ranking of biochemical system models. *Bioinformatics*, 24(6): 833–839.

Wenk, P.; Gotovos, A.; Bauer, S.; Gorbach, N. S.; Krause, A.; and Buhmann, J. M. 2019. Fast Gaussian process based gradient matching for parameter identification in systems of nonlinear ODEs. In *The 22nd International Conference on Artificial Intelligence and Statistics*, 1351–1360. PMLR.

Yang, S.; Wong, S. W.; and Kou, S. 2021. Inference of dynamic systems from noisy and sparse data via manifold-constrained Gaussian processes. *Proceedings of the National Academy of Sciences*, 118(15): e2020397118.