Physically-informed machine learning for modelling the dynamics of plant-pathogens molecular interactions

Prabal Ghosh[1][0009−0004−3449−5811]

Universite Cote d'Azur, Sophia Antipolis, France
prabal5ghosh@gmail.com

1 Research Project

I worked under the guidance of Silvia Bottini, Junior Professor Chair INRAe/UniCA, Sophia-Antipolis.

2 Abstract

Horticulture, as a critical branch of agriculture, has significantly contributed to the development of human civilization [13]. Plants, however, face ongoing challenges from biotic threats such as pathogens, which require them to employ sophisticated defense mechanisms. These defense systems involve complex signaling networks responsible for surveillance, perception, and activation of immune responses. The effectiveness of these processes is influenced by various spatial and temporal factors. The molecular dialogue between pathogens and plant hosts unfolds over time, ultimately determining the success or failure of an infection. With the advent of omics technologies, particularly transcriptomics, we now have the ability to study these intricate biological systems at a molecular level. Transcriptomics allows for the quantification of gene expression changes over time, providing valuable insights into plant responses during pathogen attacks. However, traditional methods for analyzing time-course transcriptomics data often treat each time point independently or rely on profile analysis techniques that fail to capture the temporal continuity of the data. Although alternative methods, such as regression and spline models, exist, they often fall short in providing mechanistic interpretations of the data. Furthermore, high-resolution temporal transcriptomic analysis in plant tissues is challenging due to limitations in longitudinal experiments, often involving only a few time points. This creates difficulty in drawing statistically significant conclusions regarding the changes that occur over the course of an infection.

To address these challenges, this research leverages the potential of Physics-Informed Neural Networks (PINNs), particularly Physics-Informed Dynamical Variational Autoencoders (ϕ-DVAE), as an advanced framework for analyzing time-dependent transcriptomics data. PINNs integrate observational data with underlying physical principles, making them especially suitable for handling high-dimensional, noisy, and sparse time-series datasets. Despite their success in solving various scientific problems, PINNs have yet to be fully explored in omics domains, likely due to the generally unknown governing physical systems. This study investigates the application of PINNs, specifically ϕ-DVAE, to analyze longitudinal multi-transcriptomics data related to plant defense responses. These methods enable the integration of data-driven approaches with physical constraints, allowing for a deeper understanding of the temporal dynamics of plant-pathogen interactions. By addressing limitations in traditional analyses, this framework aims to uncover new insights into the molecular mechanisms underlying plant defense, offering a novel perspective on pathogen-related dynamics even in the presence of noisy and incomplete data. This work represents a significant step toward advancing the use of machine learning in systems biology and enhancing our understanding of plant immune responses.

Keywords: Physics Informed Machine Learning, Physics-Informed Dynamical Variational Autoencoder, Partial Differential Equation, Stochastic Differential Equation, Extended Kalman Filter, Discretise Statistical Finite Element Method, Monte Carlo, Latent State-Space Model

3 Introduction

3.1 Transcriptomics and Plant Immune Responses

Omics technologies have significantly advanced the study of biological systems by enabling detailed molecular-level analysis of complex processes. These high-throughput methods encompass genomics, transcriptomics, proteomics, and metabolomics, each offering insights into different biological aspects [12].

Transcriptomics analyzes gene expression changes under various conditions or over time, providing insights into how plants regulate immune responses at the transcriptional level during stress or pathogen attacks [1]. By examining the expression patterns of thousands of genes across time and varying conditions, transcriptomics allows researchers to uncover the complex regulatory networks involved in plant immune responses. This molecular-level understanding is critical for horticulture, where plants frequently encounter biotic threats, as it helps improve disease resistance and crop protection. When a plant encounters a pathogen, it activates a series of defense mechanisms orchestrated by a network of genes. These mechanisms include both general immune responses and specific reactions, tailored to particular pathogens. Transcriptomics identifies genes that are upregulated or downregulated during pathogen attacks, offering valuable insights into the pathways and molecular processes involved in the plant's immune response. By analyzing gene expression changes at various stages of infection, transcriptomics reveals how plants perceive pathogens, activate defense genes, and regulate their immune signaling networks.

In horticulture, the use of transcriptomics is essential for identifying critical genes associated with disease resistance. This knowledge aids in developing crops with enhanced pathogen resistance, thereby improving agricultural productivity and sustainability. Additionally, transcriptomics data can inform breeding programs by pinpointing genes vital for plant defense, enabling the creation of disease-resistant plant varieties. By leveraging transcriptomics data, researchers can contribute to improved plant health, enhanced food security, and the development of sustainable agricultural practices.

3.2 Physics-Informed Neural Networks (PINNs) for Data-Driven Inverse Problems

Physics-Informed Neural Networks (PINNs) represent a powerful tool for solving data-driven inverse problems, especially in situations where the governing Partial Differential Equations (PDEs) are either unknown or only partially understood. By integrating neural networks with fundamental mathematical principles, PINNs are able to infer both the hidden dynamics and the unknown parameters directly from observed data. This is achieved by leveraging automatic differentiation, which allows the network to approximate solutions and their derivatives, thus enabling the discovery of the underlying equations while adhering to physical constraint [6]s.

The fundamental concept behind PINNs is to assume a general form for the PDE and then combine data-driven loss functions with residual terms that ensure consistency with the physical laws governing the system. The residual terms quantify the discrepancy between the model's predictions and the physical constraints imposed by the PDE. During the training process, the PINNs minimize a total loss function, which includes both observed data and the residuals of the unknown governing equations. This approach allows the model to simultaneously learn the solution to the problem and uncover the structure of the PDE, including any unknown terms and parameters [8] [2]. This flexibility makes PINNs especially effective when dealing with noisy or incomplete data and provides a unified framework for addressing challenges such as parameter estimation, missing boundary conditions, and discovering nonlinear operators in PDEs.

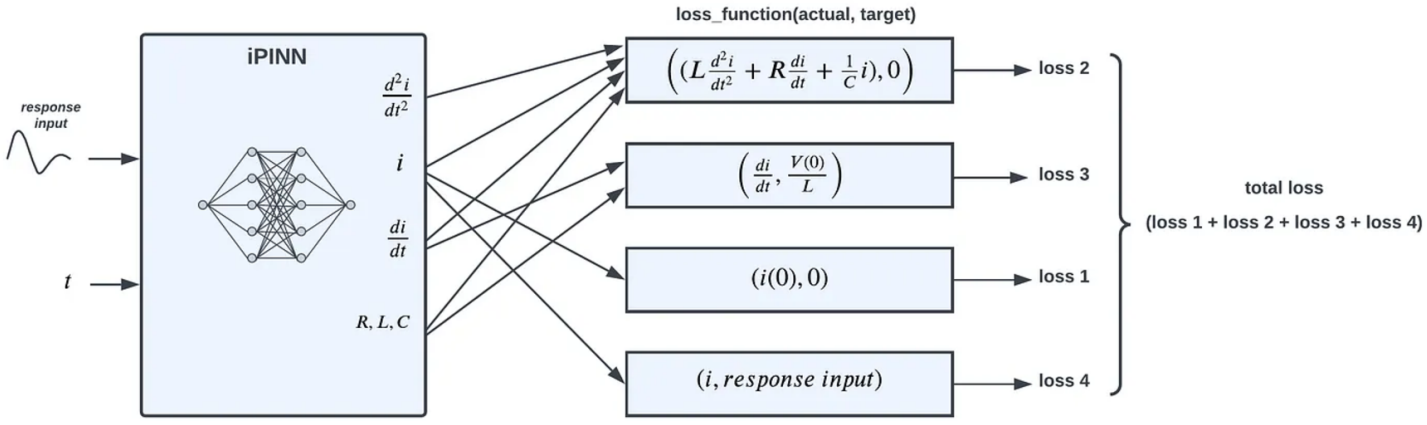

Fig. 1: Physics-Informed Neural Networks (PINNs).[11]

To begin with, the general form of the PDE is assumed without prior knowledge of the specific terms. This equation is typically expressed as:

$$\mathcal{N}[u;\lambda] = 0,$$

where $u(x,t)$ is the solution to the PDE, and $\mathcal{N}[u;\lambda]$ represents a nonlinear operator parameterized by unknown coefficients λ. These coefficients can correspond to various physical properties, such as reaction rates, diffusion coefficients, or other system parameters. The form of $\mathcal{N}$ might include terms involving derivatives, nonlinear reactions, or diffusion processes, but the exact nature of these terms is typically unknown at the start.

In the PINN framework, a neural network, $N(x,t)$, is used to approximate the solution $u(x,t)$. The network is designed to approximate not just the solution itself but also its derivatives. Since automatic differentiation allows for the efficient calculation of derivatives such as u_t, u_x, and u_{xx}, the neural network can compute the necessary quantities to construct the governing equations.

To enforce the residuals of the assumed PDE, a physics-based loss function is defined. The physics loss function can be expressed as:

$$f(x,t) = N_t + \mathcal{N}[N;\lambda].$$

This function ensures that the output of the neural network adheres to the governing PDE, thereby incorporating the underlying physics into the learning process.

During the training of the PINN, two primary loss functions are minimized: the data loss and the physics loss. The data loss quantifies the difference between the predicted values from the neural network and the observed data, and it is defined as:

$$L_{\text{data}} = \frac{1}{N}\sum_{i=1}^{N} \left| N(x_i,t_i) - u_{\text{true}}(x_i,t_i) \right|^2,$$

where $u_{\text{true}}(x,t)$ is the true observed data at the points (x_i,t_i). The data loss function ensures that the model's predictions are as close as possible to the observed data, providing a means of training the model with real-world information.

The physics loss, on the other hand, minimizes the residual of the physics-based loss, and is given by:

$$L_{\text{physics}} = \frac{1}{M}\sum_{j=1}^{M} \left| f(x_j,t_j) \right|^2,$$

where M represents the number of collocation points sampled within the domain of the PDE. This term ensures that the solution produced by the neural network is consistent with the underlying physics of the problem, enforcing the structure of the PDE during the learning process.

The total loss function for the PINN is then the sum of the data loss and the physics loss:

$$L = L_{\text{data}} + L_{\text{physics}}.$$

This combined loss function enables the neural network to learn both the solution to the problem and the structure of the governing PDE, thereby addressing the inverse problem in a unified manner.

As the PINN is trained, it not only learns the solution to the PDE but also infers the unknown parameters λ, which are the coefficients of the governing equations. These parameters may represent physical quantities such as material properties, reaction rates, or other system parameters that are critical for determining the system's behavior. In addition to learning the coefficients, the PINN framework also uncovers the form of the nonlinear operator $\mathcal{N}[u;\lambda]$, providing valuable insights into the physical processes governing the system.

PINNs are particularly well-suited for biological systems, geophysics, fluid dynamics, and material science, where the physics of the system is often complex or partially understood [10]. By bridging the gap between data and scientific modeling, PINNs enable the discovery of interpretable and physically consistent equations, offering a powerful tool for advancing scientific research in domains reliant on sparse, noisy, or incomplete datasets.

4 Related Works

4.1 (ϕ-DVAE): Physics-Informed Dynamical Variational Autoencoders

The ϕ-DVAE framework integrates variational autoencoders (VAEs) with physics-informed modeling to bridge the gap between unstructured data and systems governed by partial differential equations (PDEs) or stochastic

differential equations (SDEs). This innovative approach combines machine learning with physical laws, enabling the assimilation of noisy and sparse data into complex dynamical systems. Unlike traditional models, ϕ-DVAE preserves physical consistency while inferring latent dynamics and estimating unknown parameters, making it a powerful tool for state and parameter estimation.

A key feature of ϕ-DVAE is the embedding of latent state dynamics within a physics-constrained framework. These dynamics are governed by ordinary differential equations (ODEs), PDEs, or SDEs, providing physical realism to the latent space. In contrast to standard VAEs, which operate in an unconstrained latent space, this incorporation of physical laws adds a layer of reliability and interpretability to the generative process.

The latent dynamics in ϕ-DVAE follow stochastic differential equations (SDEs), which capture the continuous evolution of the system, accounting for both deterministic and stochastic influences. To achieve this, the framework employs statistical finite element methods (statFEM) to discretize the SDEs, enabling effective integration of physical models with machine learning.

The generative process in ϕ-DVAE transforms high-dimensional unstructured data, such as videos or images, into a low-dimensional latent space using the VAE encoder. These latent states evolve over time according to the latent dynamics described by SDEs. The decoder then reconstructs observations probabilistically from the latent states, modeling the relationship between noisy high-dimensional data and their corresponding latent representations.

One of the significant strengths of ϕ-DVAE is its ability to perform joint inference of latent states, unknown parameters governing the physical dynamics, and neural network parameters (encoder and decoder). This is achieved through variational inference and Monte Carlo sampling, providing a probabilistic framework for parameter estimation and uncertainty quantification.

Uncertainty quantification is another critical aspect of ϕ-DVAE. By using variational Bayesian methods, the framework estimates the posterior distribution of both latent states and parameters, capturing the uncertainty associated with the underlying physical model. Additionally, ϕ-DVAE employs Extended Kalman Filtering (ExKF) to infer latent states from pseudo-observations, ensuring accurate state inference even in the presence of noisy or incomplete data.

The mathematical foundation of ϕ-DVAE relies on probabilistic relationships between latent dynamics, observations, and pseudo-observations. The latent dynamics are modeled as:

$$u_n | u_{n-1}, \Lambda \sim p(u_n | u_{n-1}, \Lambda),$$

where u_n represents the latent states, Λ denotes unknown parameters in the physical model, and $p(u_n|u_{n-1}, \Lambda)$ describes the transition dynamics.

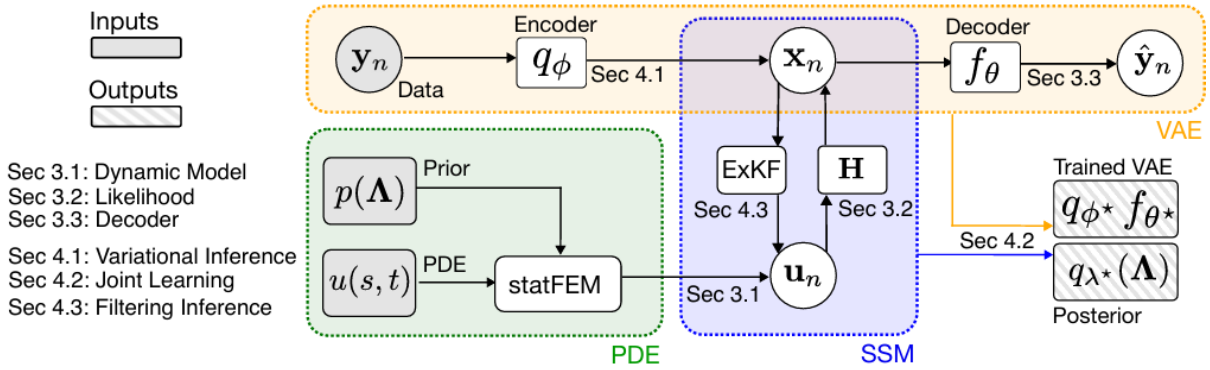

Fig. 2: Flow diagram describing connections between the specified partial differential equation (PDE), latent state-space model (SSM) and variational autoencoder (VAE).[4]

Observations, such as noisy video frames (y_n), are generated from the latent states through a probabilistic decoder:

$$y_n | x_n \sim p_\theta(y_n | x_n),$$

where x_n are the pseudo-observations, and $p_\theta(y_n|x_n)$ represents the likelihood function.

The mapping between latent states and pseudo-observations is given by:

$$x_n = Hu_n + r_n, \quad r_n \sim \mathcal{N}(0, R),$$

where H is the observation matrix, and r_n is Gaussian noise.

To approximate the true posterior distribution of the latent states and parameters, ϕ-DVAE employs a variational posterior:

$$q(u_{1:N}, x_{1:N}, \Lambda | y_{1:N}) = q(u_{1:N} | x_{1:N}, \Lambda) q_\phi(x_{1:N} | y_{1:N}) q_\lambda(\Lambda),$$

where $q_\phi(x_{1:N}|y_{1:N})$ maps data to pseudo-observations, and $q_\lambda(\Lambda)$ provides the variational approximation of the model parameters.

The model is trained by maximizing the Evidence Lower Bound (ELBO):

$$\log p(y_{1:N}) \geq E_{q_\phi} \left[\log p_\theta(y_{1:N}|x_{1:N}) - \log q_\phi(x_{1:N}|y_{1:N}) \right] + E_{q_\lambda} \left[\log p(x_{1:N}|\Lambda) + \log p(\Lambda) - \log q_\lambda(\Lambda) \right].$$

The ϕ-DVAE framework has shown great versatility in modeling both linear and nonlinear dynamical systems. It has been successfully applied to a variety of systems, including the advection equation, the Lorenz-63 system, and the Korteweg–de Vries (KdV) equation. In these applications, ϕ-DVAE has outperformed traditional methods like Kalman VAE (KVAE), accurately estimating unknown parameters and predicting system behavior even in the presence of noisy data. This demonstrates its ability to effectively integrate physical laws with machine learning for robust system identification and prediction.

5 Methodology

The project focuses on the analysis of time-course transcriptomics data from tomato plants infected by three distinct pathogens, with data collected at various time points for both leaves and roots. This dataset is crucial for studying the dynamic mechanisms of disease resistance and plant responses. While the dataset does not have missing values, it presents significant analytical challenges due to its high sparsity, consisting of 12 times points and approximately 30,000 genes, along with a limited number of samples. These characteristics make traditional analytical approaches inadequate for capturing the temporal and biological complexities inherent in the data. To address these challenges, the project employs advanced computational methods, specifically Physics-Informed

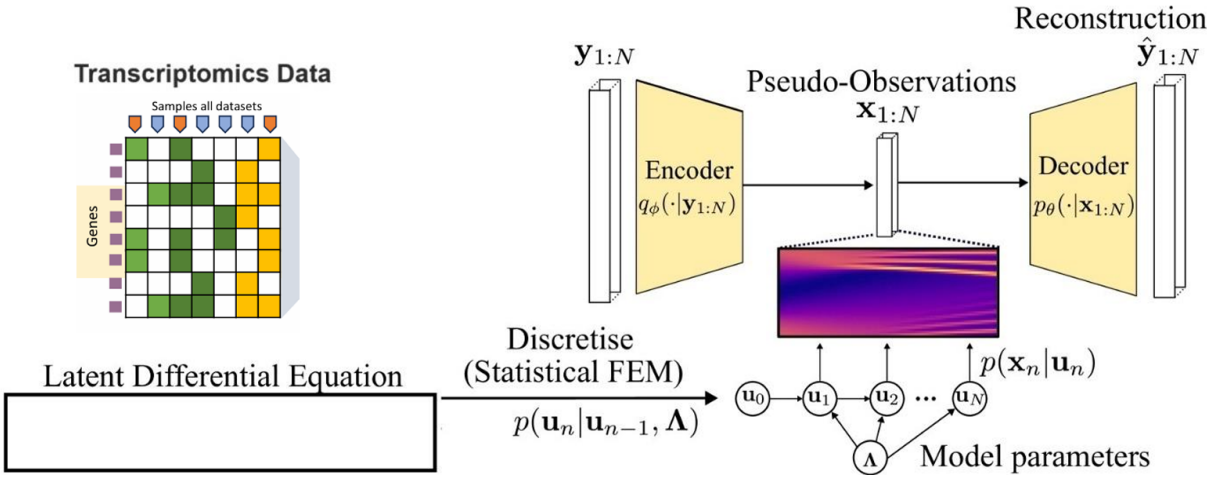

Fig. 3: The architecture of the Physics-Informed Dynamical Variational Autoencoder is being developed for time-series transcriptomics data analysis. Currently, the governing partial differential equations (PDEs) are unknown, and efforts are focused on defining the neural network for the model.

Neural Networks (PINNs) and Physics-Informed Dynamical Variational Autoencoders (ϕ-DVAE). These methods integrate observational data with physical or mathematical models to provide a robust framework for analyzing

biological dynamics, mitigating batch effects, and uncovering interactions within plant defense mechanisms over time.

The physics underlying omics data is largely unknown, which means that mathematical differential equations precisely describing the system are not predefined. This necessitates the development of foundational analytical elements from scratch. The primary focus of the project is to design neural network models capable of handling the time-dependent nature of the data. To begin addressing the challenge of undefined physical systems, simplified partial differential equations (PDEs) can be proposed for a specific part of the data. These equations serve as an approximation of the temporal interactions within specific parts of the dataset, enabling the PINNs to focus on capturing sequential relationships and estimating parameters. By training PINNs on subsets of the time-series transcriptomics data, the project explores how these networks can be adapted to high-dimensional biological datasets and used to uncover temporal dependencies [9] [5].

In addition to the use of PINNs, the project incorporates Dynamical Variational Autoencoders (DVAEs) to process sequential data. DVAEs extend traditional Variational Autoencoders (VAEs) by combining them with temporal models, often incorporating state-space models or recurrent neural networks[3] [7]. This allows for unsupervised representation learning tailored to sequential data, enabling the generation of latent vectors that encode the essential features and temporal interactions of the transcriptomics dataset. These latent representations provide a reduced-dimensional yet comprehensive view of the data, capturing critical elements relevant to plant defense mechanisms and their responses to pathogen infections. We also integrate simplified PDEs from the part of omics data to introduce physics into the DVAE (ϕ-DVAE). These state-of-the-art techniques combine observational data with physical or mathematical models, enabling the capture of complex biological dynamics and interactions. The proposed methodology focuses on removing batch effects and uncovering temporal dependencies in transcriptomic data to reveal the underlying patterns of plant-pathogen interactions.

References

1. Yunpeng Cao, Xiaoxu Li, Hui Song, Muhammad Abdullah, and Muhammad Aamir Manzoor. Multi-omics and computational biology in horticultural plants: from genotype to phenotype, volume ii, 2024.
2. Salvatore Cuomo, Vincenzo Schiano Di Cola, Fabio Giampaolo, Gianluigi Rozza, Maziar Raissi, and Francesco Piccialli. Scientific machine learning through physics–informed neural networks: Where we are and what's next. *Journal of Scientific Computing*, 92(3):88, 2022.
3. Laurent Girin, Simon Leglaive, Xiaoyu Bie, Julien Diard, Thomas Hueber, and Xavier Alameda-Pineda. Dynamical variational autoencoders: A comprehensive review. *arXiv preprint arXiv:2008.12595*, 2020.
4. Alex Glyn-Davies, Connor Duffin, O Deniz Akyildiz, and Mark Girolami. ϕ-dvae: Physics-informed dynamical variational autoencoders for unstructured data assimilation. *Journal of Computational Physics*, 515:113293, 2024.
5. Paguiel Javan Hossie, Béatrice Laroche, Thibault Malou, Lucas Perrin, Thomas Saigre, and Lorenzo Sala. Simulating interactions in microbial communities through physics informed neural networks: towards interaction estimation. 2024.
6. George Em Karniadakis, Ioannis G Kevrekidis, Lu Lu, Paris Perdikaris, Sifan Wang, and Liu Yang. Physics-informed machine learning. *Nature Reviews Physics*, 3(6):422–440, 2021.
7. Diederik P Kingma, Max Welling, et al. An introduction to variational autoencoders. *Foundations and Trends® in Machine Learning*, 12(4):307–392, 2019.
8. Isaac E Lagaris, Aristidis Likas, and Dimitrios I Fotiadis. Artificial neural networks for solving ordinary and partial differential equations. *IEEE transactions on neural networks*, 9(5):987–1000, 1998.
9. Lu Lu, Xuhui Meng, Zhiping Mao, and George Em Karniadakis. Deepxde: A deep learning library for solving differential equations. *SIAM review*, 63(1):208–228, 2021.
10. Chuizheng Meng, Sungyong Seo, Defu Cao, Sam Griesemer, and Yan Liu. When physics meets machine learning: A survey of physics-informed machine learning. *arXiv preprint arXiv:2203.16797*, 2022.
11. Maziar Raissi, Paris Perdikaris, and George E Karniadakis. Physics-informed neural networks: A deep learning framework for solving forward and inverse problems involving nonlinear partial differential equations. *Journal of Computational physics*, 378:686–707, 2019.
12. Jeyachandran Sivakamavalli and Baskaralingam Vaseeharan. An overview of omics approaches: Concept, methods and perspectives. 2020.
13. Xiaoyu Zhang, Jingqing Zhang, Kai Sun, Xian Yang, Chengliang Dai, and Yike Guo. Integrated multi-omics analysis using variational autoencoders: application to pan-cancer classification. In *2019 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, pages 765–769. IEEE, 2019.