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 Introduction

Horticulture, as a critical branch of agriculture, has significantly contributed to the development of human civilization [11]. 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 aims to leverage Physics-Informed Neural Networks (PINNs), specifically Physics-Informed Dynamical Variational Autoencoders (ϕ-DVAE), for the analysis of transcriptomics data. These advanced techniques integrate both observational data and physical principles, allowing for a more comprehensive understanding of pathogen-related dynamics at the molecular level. By applying ϕ-DVAE to plant-pathogen interactions, we aim to gain new insights into the temporal dynamics of plant defense mechanisms, despite the challenges posed by sparse and noisy data.

3 Aim

A promising new approach for analyzing time-dependent data is the use of physics-informed neural networks (PINNs). These deep learning algorithms can integrate observational data with physical or mathematical understanding, making them particularly suitable for high-dimensional, noisy data and time-series analysis [9]. However, PINNs have not yet been applied to omics data, likely due to the generally unknown physical systems in this domain.

This research project aims to investigate the potential of PINNs, particularly by using Physics-Informed Dynamical Variational Autoencoders (ϕ-DVAE), and their suitability for analyzing information contained in longitudinal multi-transcriptomics data related to plant defense responses. By exploring these novel methods, we may gain new insights into the complex temporal dynamics of plant-pathogen interactions and improve our understanding of plant defense mechanisms.

4 RELATED WORKS

4.1 Omics data

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 [10].

Types of Omics Data

Genomics : Provides a comprehensive map of an organism's genetic material, aiding in the identification of key genes involved in pathways such as disease resistance and growth regulation.

Transcriptomics : Analyzes gene expression changes under various conditions or over time, revealing how plants regulate immune responses at the transcriptional level during stress or pathogen attacks [1].

Proteomics : Studies proteins, their modifications, and roles in cellular processes, explaining the functional role of proteins like R proteins in defense mechanisms against pathogens.

Metabolomics : Focuses on small molecules and metabolites, identifying compounds like phytoalexins that contribute to chemical defense against pathogens.

Transcriptomics data provides valuable insights into gene expression changes at the molecular level, which are critical for understanding how plants respond to stress, including pathogen attacks. By analyzing the expression patterns of thousands of genes over time and under varying conditions, transcriptomics allows researchers to explore the complex regulatory networks involved in plant immune responses. In horticulture, where plants face constant biotic threats, understanding these responses is essential for improving disease resistance and crop protection.

When a plant encounters a pathogen, it activates a series of defense mechanisms that are orchestrated by a network of genes. These mechanisms involve both general immune responses and specific reactions tailored to particular pathogens, such as effector-triggered immunity (ETI) and pattern-triggered immunity (PTI). Transcriptomics enables the identification of genes that are upregulated or downregulated during pathogen attacks, shedding light on the pathways and molecular processes involved in the plant's immune response. By analyzing changes in gene expression at different stages of the infection, transcriptomics helps us understand how plants perceive and respond to pathogens, how they activate defense genes, and how they regulate their immune signaling networks.

Using transcriptomics data in horticulture is particularly important because it allows for the identification of critical genes involved in disease resistance. This information can be used to develop crops with enhanced resistance to pathogens, improving agricultural productivity and sustainability. Moreover, it can guide breeding programs by pinpointing genes that are crucial for plant defense, enabling the development of disease-resistant varieties. Integrating transcriptomics with other omics technologies, such as genomics, proteomics, and metabolomics, provides a more comprehensive understanding of plant-pathogen interactions and the complex networks that govern plant immunity.

By leveraging transcriptomics data, we can improve plant health, ensure food security, and create sustainable agricultural practices.

4.2 Physics-Informed Neural Networks (PINNs)

Physics-Informed Neural Networks (PINNs) are a transformative tool for solving data-driven inverse problems, particularly when the governing Partial Differential Equations (PDEs) are unknown or only partially understood. In these cases, PINNs integrate neural networks with general mathematical principles to infer both the hidden dynamics and the unknown parameters directly from observed data. By leveraging automatic differentiation, PINNs approximate solutions and their derivatives, enabling the discovery of the underlying equations while adhering to physical constraints [5].

The approach involves defining a general PDE form and combining data-driven loss functions with residual terms that enforce physics consistency [6] [2]. During training, PINNs minimize a total loss that incorporates observed data and the residuals of the unknown governing equations. This allows the model to simultaneously learn the solution to the problem and uncover the PDE structure, including unknown terms and parameters. The flexibility of this method makes it robust to noisy or incomplete data, offering a unified framework for parameter estimation, missing boundary conditions, and discovering nonlinear operators in PDEs.

How PINNs Work for Data-Driven Inverse Problems

General PDE Form Assume the general form of a PDE without prior knowledge of specific terms:

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

where $u(x, t)$ is the solution, and $\mathcal{N}[u; \lambda]$ is a nonlinear operator parameterized by unknown coefficients λ. The structure of $\mathcal{N}$ could include terms like derivatives, reactions, or diffusion.

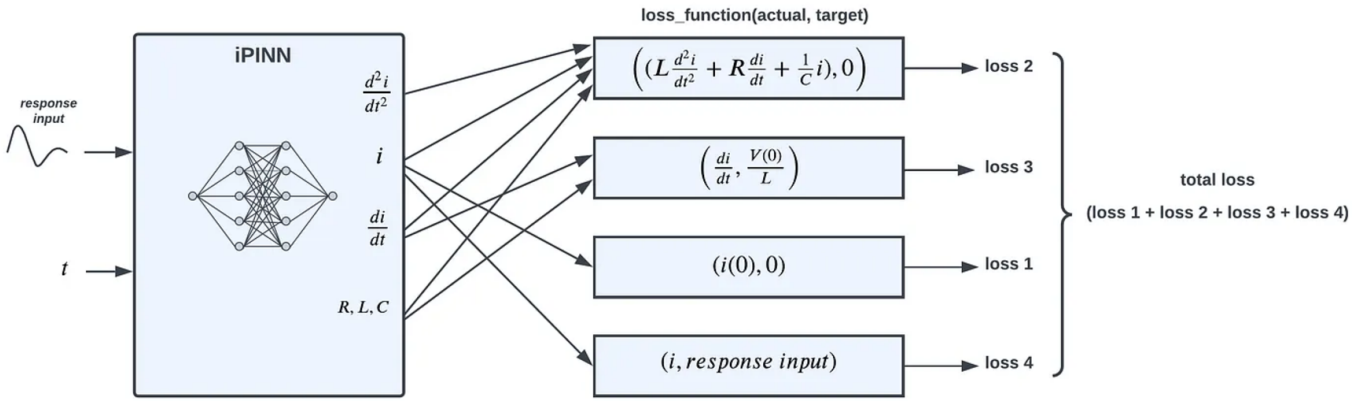

Fig. 1: Physics-Informed Neural Networks (PINNs).[8]

PINN Framework

Neural Network Approximation: A neural network $N(x,t)$ is used to approximate the solution $u(x,t)$. By leveraging automatic differentiation, the network can compute derivatives such as u_t, u_x, and u_{xx}, which are required to construct the governing equations.

Physics-Loss Definition: Define a loss function $f(x,t)$ that enforces the residuals of the assumed PDE:

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

Training Goals During training, the PINN minimizes two main losses:

- **Data Loss:** Minimize the difference between the predicted and observed data:

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

  where $u_{\text{true}}(x,t)$ represents the measured or observed data.
- **Physics Loss:** Minimize the residual of the physics-based loss:

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

  where M represents the collocation points sampled in the domain.

  The total loss is defined as:

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

Inferring Parameters During training, PINNs learn the following:

- The unknown coefficients λ, which are the parameters of the PDE.
- The form of the nonlinear operator $\mathcal{N}[u;\lambda]$.

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 [7]. 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.

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

ϕ-DVAE is a framework that integrates variational autoencoders (VAEs) with physics-informed modeling to assimilate unstructured data (e.g., videos, images) into systems governed by partial differential equations (PDEs) or stochastic differential equations (SDEs). This method blends machine learning with known physical laws, allowing for state and parameter estimation in complex dynamical systems. Unlike traditional models, ϕ-DVAE can infer latent dynamics and unknown parameters from noisy and sparse data while preserving physical consistency.

Key Features of ϕ-DVAE

- **Physics-Informed Latent Dynamics**: ϕ-DVAE embeds the evolution of latent states within a framework constrained by known physical laws, such as those described by ordinary differential equations (ODEs) or partial differential equations (PDEs). This contrasts with traditional VAEs that operate in an unconstrained latent space, thereby incorporating a crucial layer of physical realism into the generative process.
- **Stochastic Differential Equations (SDEs)**: The latent dynamics in ϕ-DVAE are governed by stochastic differential equations (SDEs), which model the continuous evolution of the system while accounting for both deterministic and stochastic influences. The framework applies **statistical finite element methods (statFEM)** to discretize these SDEs, allowing for the effective integration of physical models with machine learning techniques.
- **Generative Model for Data Assimilation**: The ϕ-DVAE framework defines a generative process where high-dimensional unstructured data (e.g., video frames) is embedded into a low-dimensional latent space via the VAE encoder. These latent states evolve over time according to the latent dynamics described by the SDEs. The generative process is probabilistic, with latent states modeled by physical dynamics and observations (such as noisy video frames) being generated from these latent representations.
- **Joint Inference of Latent States and Parameters**: One of the key strengths of ϕ-DVAE is its ability to simultaneously estimate the latent states, the unknown parameters governing the physical dynamics, and the neural network parameters (such as those in the encoder and decoder). This joint inference process is performed using **variational inference** and **Monte Carlo sampling**, providing a probabilistic framework for parameter estimation and uncertainty quantification.
- **Uncertainty Quantification**: The framework allows for robust uncertainty quantification, especially for unknown parameters in the underlying physical model. By using **variational Bayesian methods**, ϕ-DVAE estimates the posterior distribution over parameters, incorporating the uncertainty associated with both the latent states and the parameters governing the physical dynamics.
- **Extended Kalman Filter (ExKF) for Latent State Inference**: To estimate the latent states from pseudo-observations (the low-dimensional representations of the observed data), ϕ-DVAE employs **Extended Kalman Filtering (ExKF)**. This filtering technique is used to propagate uncertainty through the latent state evolution, enabling accurate state inference even in the presence of noisy or incomplete data.

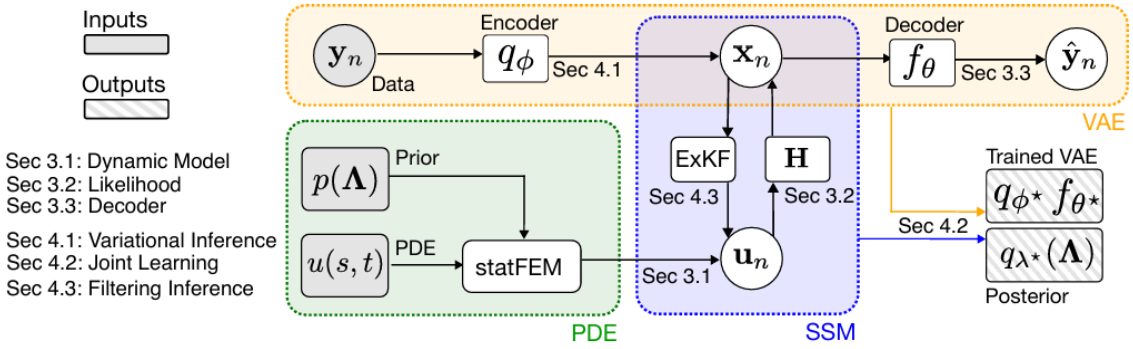

Fig. 2: Flow diagram describing connections between the specified partial differential equation (PDE), latent state-space model (SSM) and variational autoencoder (VAE).[3]

Mathematical Framework The core of the ϕ-DVAE model involves the following probabilistic relationships:

- **Latent Dynamics**: The latent states u_n evolve according to a stochastic process:

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

  where Λ represents the unknown parameters in the physical model, and $p(u_n|u_{n-1}, \Lambda)$ is a transition model describing the latent dynamics.
- **Observation Model**: The observations y_n (e.g., video frames) are generated from the latent states via a probabilistic decoder:

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

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

– **Latent State to Pseudo-Observation Mapping**: The relationship between the 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 represents Gaussian noise.

– **Variational Inference**: The variational posterior is used to approximate the true posterior distribution of the latent states and parameters:

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

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

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

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

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 collecting time-course transcriptomics data from tomato plants infected by three different pathogens, affecting both the leaf and root at various time points. This approach is crucial for understanding the dynamic mechanisms of disease resistance and plant responses. However, a common challenge in such high-throughput omics studies is the presence of missing or incomplete data, which can impede comprehensive analysis. To address this, the project employs advanced methods like physics-informed neural networks (PINNs) and physics-informed dynamic variational autoencoders (ϕ-DVAE). These techniques integrate observational data with underlying physical or mathematical models, enhancing the ability to capture biological dynamics and interactions within plant defense mechanisms. This innovative approach aims to provide robust analysis and insights into how plants respond to pathogen infections over time, despite the challenges posed by incomplete data.

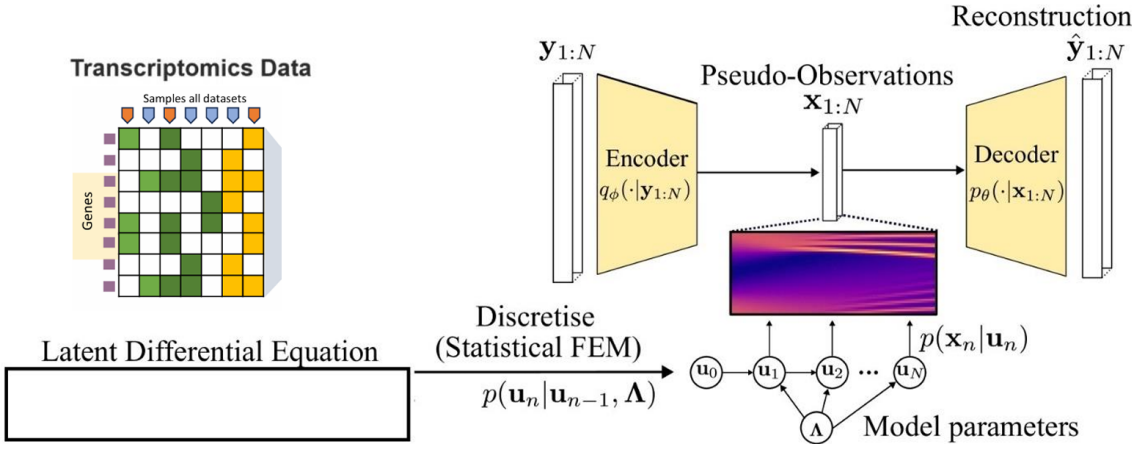

Fig. 3: (ϕ-DVAE) for Transcriptomics data

6 Conclusion

[4]

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. 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.
4. 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.
5. 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.
6. 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.
7. 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.
8. 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.
9. Min Shi and Shamim Mollah. Netoif: A network-based approach for time-series omics data imputation and forecasting. *bioRxiv*, pages 2021–06, 2021.
10. Jeyachandran Sivakamavalli and Baskaralingam Vaseeharan. An overview of omics approaches: Concept, methods and perspectives. 2020.
11. 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.