Aligning AI-driven Discovery with Human Intuition

Kevin Zhang [1] **Judah Goldfeder** [2] **Hod Lipson** [2]

Abstract

As data-driven modeling of physical dynamical systems becomes more prevalent, a new challenge is emerging: making these models more compatible and aligned with existing human knowledge. AI-driven scientific modeling processes typically begin with identifying hidden state variables, then deriving governing equations, followed by predicting and analyzing future behaviors. The critical initial step of identification of an appropriate set of state variables remains challenging for two reasons. First, finding a compact set of meaningfully predictive variables is mathematically difficult and under-defined. A second reason is that variables found often lack physical significance, and are therefore difficult for human scientists to interpret. We propose a new general principle for distilling representations that are naturally more aligned with human intuition, without relying on prior physical knowledge. We demonstrate our approach on a number of experimental and simulated system where the variables generated by the AI closely resemble those chosen independently by human scientists. We suggest that this principle can help make human-AI collaboration more fruitful, as well as shed light on how humans make scientific modeling choices.

1. Introduction

Recent advances in data-driven methodologies have revolutionized scientific discovery across a wide range of fields (Evans & Rzhetsky, 2010; Fortunato et al., 2018). With increasing computational power and the rise of artificial intelligence (AI), the ability to rapidly analyze vast amounts of data has significantly improved (LeCun et al., 2015; Han et al., 2018; Akiyama et al., 2019). AI algorithms now outperform humans in various tasks, including constructing predictive models (Wagner, 2021), designing experimental protocols (Coley et al., 2019), solving inverse problems (Guo et al., 2024), and testing causal hypotheses (King et al., 1996; Waltz & Buchanan, 2009; King et al., 2004).

Despite the significant progress in data-driven approaches, discovering state variables remains a challenging task. Automated scientific discovery generally involves two key steps: identifying state variables and determining the underlying dynamics governing them (Kramer et al., 2023; Cranmer et al., 2020). This process is illustrated in Figure 1. By following this framework, models can extract meaningful quantities from complex systems and predict future behavior. This work focuses on the initial step of state variable discovery *without prior knowledge of the system*, while simultaneously modeling the governing equations. Identifying these variables is critical for modeling new scientific systems and formulating physical laws (Thompson & Steward, 2002; Hirsch et al., 2012). Automating this process not only deepens our understanding of complex systems that elude human intuition, but also enhances predictive modeling techniques (Kramer et al., 2023). Such advancements have broad implications across various fields, including decision-making (Guo et al., 2021; Barfoot et al., 2019), biology (Keil, 2010; Jumper et al., 2021), economics (Harvey et al., 1994), chemistry (Zhang et al., 2018), engineering (Lin et al., 2025), geology (Houtekamer & Zhang, 2016; Carrassi et al., 2018), and beyond.

Significant progress has been made in modeling dynamics using methods such as sparse regression, genetic programming, and deep learning (Mangan et al., 2017; Kronberger et al., 2020; Raissi & Karniadakis, 2017; Wu & Xiu, 2019; Wu et al., 2019), however selecting an appropriate set of state variables still remains a fundamental challenge (Chen et al., 2022). The difficulty stems from the non-uniqueness of state variables, even in simple systems. For example, a single pendulum can be described using either the angle and angular velocity of its arm or its kinetic and potential energies. Although numerous other combinations of variables could model a pendulum, many lack intuitive physical significance. Consequently, existing methods often rely on assumptions in the data or latent representations to identify

[1]Massachusetts Institute of Technology, Cambridge, MA, USA [2]Columbia University, New York, NY, USA. Correspondence to: Kevin Zhang <kzhang02@mit.edu>, Judah Goldfeder <jag2396@columbia.edu>, Hod Lipson <hod.lipson@columbia.edu>.

Proceedings of the AI for Math Workshop at the 43rd International Conference on Machine Learning (AIforMath@ICML 2026), Seoul, South Korea. 2026. Copyright 2026 by the author(s).

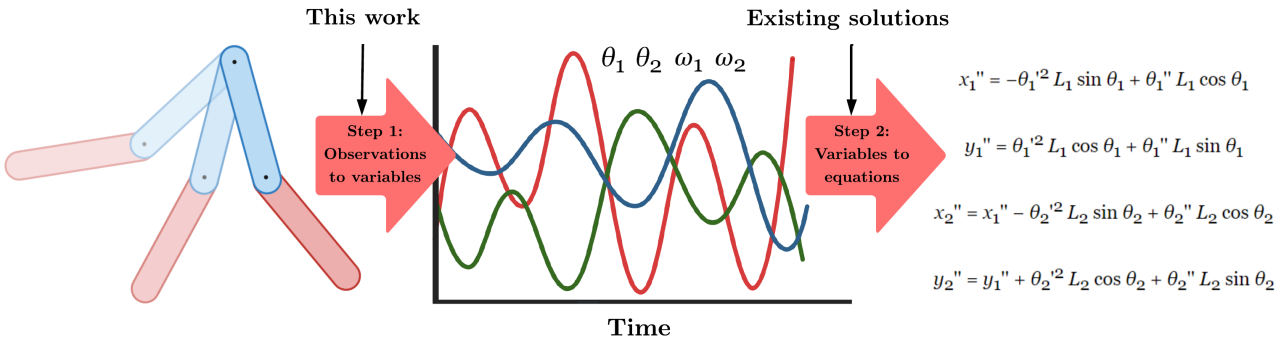


Figure 1. Overview of the automated scientific discovery process. The model first extracts relevant variables from observations of the system then derives the corresponding governing equations for the dynamics. Our work focuses on the initial state variable discovery step.

meaningful state variables (Brunton et al., 2016; Mrowca et al., 2018; Champion et al., 2019; Raissi et al., 2017; Garcon et al., 2022; Kevrekidis et al., 2002; Chen & Bollt, 2007; Bongard & Lipson, 2007; Schmidt & Lipson, 2009). In complex or novel environments where prior assumptions are absent, the problem becomes even more pressing and difficult (Davies et al., 2021). This work proposes a principle to address this challenge.

Various techniques, such as dynamic mode decomposition and singular value decomposition, have been developed for discovering state variables, but these methods typically require prior knowledge of the system or underlying physical laws (Kevrekidis et al., 2002; Chen & Bollt, 2007; Bongard & Lipson, 2007; Schmidt & Lipson, 2009; Kutz et al., 2016). Similarly, several machine learning approaches have been proposed to infer state variables, yet they often rely on specific modeling assumptions. For example, Champion (Champion et al., 2019) constrained the training of an autoencoder by imposing a set of basis functions for dynamic equations, while Garcon (Garcon et al., 2022) integrated an autoencoder with a regression module requiring partial knowledge of variables or the data-generating process. Raissi (Raissi et al., 2017) developed a physics-informed neural network to infer dynamics, assuming certain physical laws or conditions. While these methods are effective in scenarios with known simplifying assumptions, they become impractical in novel environments. The ability to identify state variables and derive governing equations without relying on prior modeling assumptions is crucial for understanding unfamiliar systems and advancing scientific discovery. This capability enables researchers to gain insights into unknown phenomena, particularly when theoretical predictions and empirical observations diverge.

To address the limitation of prior assumptions, Chen et al. (Chen et al., 2022) introduced a two-stage autoencoder architecture designed to extract state variables from videos of physical systems. Their method utilizes manifold learning algorithms to estimate the intrinsic dimension (ID) of the system by employing a high-dimensional latent representation from the first autoencoder. The ID, which represents the minimum number of variables required to fully describe the system, then defines the latent dimension of a second autoencoder tasked with learning the state variables. While this approach operates without prior knowledge of the system, it struggles to produce variables with physical significance or explicit analytical expressions. This limitation stems from the non-uniqueness of the representations. Without interpretability criteria to guide the discovery process, the model tends to favor overly complex representations. Machine learning models often function as black boxes with limited transparency (Lipton, 2017; Ennab & Mcheick, 2022), and can learn representations that do not align with human categories (Lomasov et al., 2025), but for state variable discovery, interpretable representations are essential for facilitating a deeper understanding of the system (Kaur et al., 2020).

In this work, we present a data-driven approach for learning *interpretable* state variables *without prior knowledge of the system*. To accomplish this, we introduce the Temporally-Informed Dynamics Encoder (TIDE), which incorporates key mathematical properties typical of physical variables into its training objective, aligning AI-generated representations with human understanding. We demonstrate that TIDE successfully extracts meaningful variables across various datasets, showing a higher correlation with human-interpretable quantities. Our main contributions are as follows: (1) we propose a method for jointly learning the dynamics, incorporating time-derivative regularization to tackle the state variable discovery problem, and (2) we learn variables that not only align more closely with human-defined quantities but also offer explicit analytical expressions and interpretable latent spaces.

2. Problem formulation and motivation

Problem statement

We propose a general framework for modeling dynamical systems. Following the approach in (Chen et al., 2022), we focus on analyzing videos of physical phenomena, though our methods can be extended to other data modalities. While images offer meaningful visual representations for humans, their high dimensionality and inherent noise make them less suitable for quantitative analysis, particularly in unknown or complex systems. Our objective is to infer the intrinsic dimension (ID) and derive a set of state variables that not only accurately capture the system's dynamics but also provide interpretative value to humans. This enables us to effectively represent, understand, and analyze the underlying dynamics.

Let $\mathcal{X}$ denote the input representation space, and consider a dataset comprising N sequences $\mathcal{D} = \{\mathbf{x}_1, \ldots, \mathbf{x}_N\}$. Each sequence $\mathbf{x}_i$ consists of M observations $\mathbf{x}_{i,1}, \ldots, \mathbf{x}_{i,M} \in \mathcal{X}$ sampled at a discrete time interval Δt. For video data, we adopt the methodology from (Chen et al., 2022) and concatenate pairs of consecutive frames for each observation, as certain variables require information about first-order time derivatives. Thus, $\mathcal{X} \subseteq \mathbb{R}^{2\times H\times W\times C}$, where H, W, and C represent the height, width, and number of channels of each frame, respectively. We make the following assumptions about the data:

(1) There exists a finite dimension state variable representation space $\mathcal{S}$ and a bijective function $f : \mathcal{S} \to \mathcal{X}$ such that for all $\mathbf{x} \in \mathcal{X}$, there exist unique state variables $\mathbf{z} \in \mathcal{S}$ that generate the input $\mathbf{x} = f(\mathbf{z})$.

(2) There exists a deterministic time evolution operator $g : \mathcal{S} \to \mathcal{S}$ such that $\mathbf{z}_{i,j+1} = g(\mathbf{z}_{i,j})$.

(3) Δt is sufficiently small such that consecutive observations over time do not have large changes in the state variable representation.

Assumptions 1 and 2 assert that there exists a finite set of state variables that evolves deterministically, while assumption 3 states that state variable representation is stable. This assumption is generally weak and is typically satisfied by efficient data measurement or high video frame rate.

Our goal is to fully characterize the system and predict its dynamics. To achieve this, it is sufficient to determine the intrinsic dimension (ID) and then specify f and g. Note that assumptions 1 and 2 imply the existence of an evolution operator in the input space, $g' : \mathcal{X} \to \mathcal{X}$, defined as $g' = f \circ g \circ f^{-1}$. However, modeling dynamics directly in terms of pixels is both overly complex and uninformative, so we choose to model f and g separately. We employ a variational autoencoder (VAE) to learn f, while simultaneously optimizing a feed-forward network (referred to as the dynamics module) to learn g. This approach avoids assuming prior knowledge of the system or imposing functional constraints, making it suitable even for unknown environments. Furthermore, we do not enforce statistical dependencies between state variables or restrict the governing equations to parametric forms via basis functions, as such assumptions are not universally applicable. By avoiding these limitations, our model offers a more general framework compared to previous methods.

Aligning state variables

While identifying state variables is crucial, it is equally important to prioritize those that offer interpretive value. However, assessing the alignment between human-understandable and model-learned latent variables without prior knowledge of the system poses a challenge. To address this, consistent with the principle that strong domain-specific priors are necessary for specialized scientific tasks(Goldfeder et al., 2026), we propose three desirable qualities for typical human-interpretable state variables: they should accurately represent the system, have simple governing equations, and be infinitely differentiable over time almost everywhere. We incorporate these properties by training an autoencoder to learn the representation, jointly optimizing a dynamics module, and applying time-derivative regularization, respectively.

3. Methods

Model architecture

We present the components of the Temporally-Informed Dynamics Encoder (TIDE), which includes a variational autoencoder (VAE) backbone, dynamics module, and time-derivative penalty. Our approach uses TIDE networks to compress high-dimensional pixel data into a low-dimensional set of state variables, following a two-step process similar to (Chen et al., 2022).

We use a variational autoencoder (VAE) to encode the input data into a latent representation to model the function g. VAEs are a class of generative models comprising an encoder and decoder parameterized by θ. The encoder compresses the input $\mathbf{x}$ into a low-dimensional latent representation $\mathbf{z}$, while the decoder reconstructs the data by optimizing a maximum likelihood objective. Due to the intractability of the posterior $p_\theta(\mathbf{z})$, VAEs approximate it by variational inference. Using a Gaussian variational family, we learn an encoder h_ϕ, parameterized by ϕ, to approximate the posterior

$$q_\phi(\mathbf{z} \mid \mathbf{x}) = \mathcal{N}(h_\phi^{\boldsymbol{\mu}}, \mathrm{diag}(h_\phi^{\boldsymbol{\sigma}^2})) \approx p_\theta(\mathbf{z} \mid \mathbf{x})$$

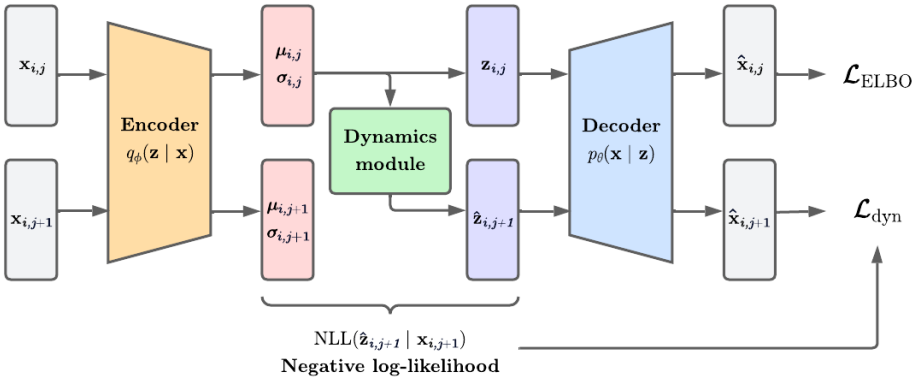


Figure 2. Schematic of the TIDE architecture for interpretable state variable discovery. The VAE backbone consists of an encoder q_ϕ which maps the input $\mathbf{x}_{i,j}$ to $\mathcal{N}(\boldsymbol{\mu}_{i,j}, \boldsymbol{\sigma}^2_{i,j})$ and a decoder p_θ that reconstructs the input from $\mathbf{z}_{i,j} \sim q\phi(\mathbf{z} \mid \mathbf{x}i, j)$. A dynamics module, h_{dyn}, is integrated to predict the next latent representation $\hat{\mathbf{z}}_{i,j+1} = h_{\text{dyn}}(\boldsymbol{\mu}_{i,j})$ by maximizing the likelihood. The overall training objective is a weighted combination of the ELBO $\mathcal{L}_{\text{ELBO}}$, dynamics loss $\mathcal{L}_{\text{dyn}}$, and time-derivative regularization $\mathcal{L}_{\text{reg}}$.

by maximizing the evidence lower bound (ELBO):

$$\mathcal{L}_{\text{ELBO}}(\beta) = \mathbb{E}_{\mathbf{x}}[\mathbb{E}_{q_\phi}[\log p_\theta(\mathbf{x} \mid \mathbf{z})] - \beta D_{KL}(q_\phi(\mathbf{z} \mid \mathbf{x}) \parallel p_\theta(\mathbf{z}))],$$

where β controls the latent capacity as introduced in (Higgins et al., 2017).

We jointly optimize a dynamics module h_{dyn} integrated within the VAE to model the time evolution operator g. This module is trained to maximize the likelihood of the next observation. For input $\mathbf{x}_{i,j}$, let $\boldsymbol{\mu}_{i,j}$ and $\boldsymbol{\sigma}_{i,j}$ denote the mean and standard deviation of $q_\phi(\mathbf{z}_{i,j} \mid \mathbf{x}_{i,j})$, respectively. Let $\hat{\mathbf{z}}_{i,j+1} = h_{\text{dyn}}(\mathbf{x}_{i,j})$ represent the predicted latent state at the next time step. The dynamics objective is to maximize:

$$\mathcal{L}_{\text{dyn}}(\lambda_1) = \mathbb{E}_{\mathbf{x}}\left[\log p_\theta(\mathbf{x}_{i,j+1} \mid \hat{\mathbf{z}}_{i,j}) + \lambda_1 \log q_\phi(\hat{\mathbf{z}}_{i,j} \mid \mathbf{x}_{i,j+1})\right]$$

We introduce a hyperparameter λ_1 to balance the relative weight between maximizing the data likelihood and latent likelihood. The dynamics module is implemented as a fixed-width, three-layer feed-forward neural network (FFN) to maintain simplicity and limit the complexity of the governing dynamics. Alternatively, the dynamics module could be trained using symbolic or sparse regression methods (e.g., SINDy (Brunton et al., 2016)). However, such approaches inherently constrain the dynamics by relying on predefined basis or candidate functions. By using a neural network, we achieve greater generality. The architecture is illustrated in Figure 2.

Time-derivative regularization

We introduce time-derivative regularization to encourage smooth state variables by applying a penalty term based on the magnitudes of discrete derivatives. To prevent global contraction of the latent space, we normalize the latent vectors along each dimension independently using min-max normalization. Given an input $\mathbf{x}$, let $\overline{\boldsymbol{\mu}}$ represent the normalized mean. The discrete n-th derivative at $\overline{\boldsymbol{\mu}}_{i,j}$ is defined as $\overline{\boldsymbol{\mu}}^{(n)}_{i,j} = \overline{\boldsymbol{\mu}}^{(n-1)}_{i,j+1} - \overline{\boldsymbol{\mu}}^{(n-1)}_{i,j}$, where $\overline{\boldsymbol{\mu}}^{(0)}_{i,j} = \overline{\boldsymbol{\mu}}_{i,j}$. The time-derivative regularization $\mathcal{L}_{\text{reg}}$ aggregates the first n discrete derivatives. Since higher-order derivatives tend to vanish due to the lack of renormalization, each term is geometrically scaled by a factor ω. The regularization penalty is expressed as follows:

$$\mathcal{L}_{\text{reg}} = \mathbb{E}_{\mathbf{x}_i \in \mathcal{D}}\left[\sum_{d=1}^{n} \omega^d \, \overline{\boldsymbol{\mu}}^{(d)}_i\right]$$

We use $n = 4$ and $\omega = 5$ in all experiments, as these values yield meaningful state variables. The final objective is a weighted combination of the ELBO, dynamics loss, and the regularization term.

$$\mathcal{L}_{\text{TIDE}} = -\mathcal{L}_{\text{ELBO}}(\beta) - \mathcal{L}_{\text{dyn}}(\lambda_1) + \lambda_2 \mathcal{L}_{\text{reg}}.$$

We introduce a hyperparameter λ_2 to control the weight of the regularization term and refer to our model architecture, comprising the VAE backbone, dynamics module, and time-derivative regularization, as a TIDE network.

Intrinsic dimension estimation

Accurate estimation of the intrinsic dimension (ID) is essential for efficient and precise modeling. In (Chen et al., 2022), the authors observed that directly reducing the bottleneck dimension of an autoencoder does not reliably recover the ID, as reconstructions begin to deteriorate before reaching the true value. Instead, they use a manifold learning algorithm to estimate the ID. Building on their methodology, we train a TIDE network to generate high-dimensional latent representations $\mathbf{y}_{i,j} \in \mathbb{R}^{64}$ and then apply the Dimensionality from Angle and Norm Concentration (DANCo) algorithm (Ceruti et al., 2014) to the latent space. This provides a maximum likelihood estimate of the ID based on the distance and angle between pairs of normalized k-nearest neighbors. The

estimated ID serves as the latent dimension for a second TIDE network, which is used to predict the final state variables. The second network is trained on top of the first (which is kept frozen) to learn state variables $\mathbf{z}_{i,j} \in \mathbb{R}^{\text{ID}}$ by reconstructing both the data $\mathbf{x}_{i,j}$ and the intermediate latent representation $\mathbf{y}_{i,j}$, where the additional likelihood term for $\mathbf{y}_{i,j}$ is weighted by a hyperparameter λ_3.

Symbolic regression

For systems with known standard human state variables, we derive an expression for the predicted state variable representation using PySR (Cranmer, 2023), a symbolic regression library built on multi-population genetic algorithms. We restrict the search space to sine, addition, subtraction, and multiplication operations. These functions prove sufficient for accurate analytical reconstruction, as adding other elementary functions does not significantly improve the final fit.

Dataset

We evaluate our model on nine benchmark datasets from (Chen et al., 2022), consisting of five simulated systems (reaction-diffusion, circular motion, single pendulum, double pendulum, elastic pendulum) and four real-world systems (swing stick, air dancer, lava lamp, fire flame). Figure 3 illustrates these nine evaluation settings.

4. Results

To assess our model, we establish specific evaluation metrics then compare our approach and the methodology in (Chen et al., 2022). Our model accurately learns the dynamics and an interpretable latent representation having higher correlation with human variables. Additionally, we derive analytical expressions for the learned state variables, demonstrating they are meaningful and recoverable.

Evaluation metrics

To assess variable interpretability, we use two metrics: mutual information (MI) and analytical mean squared error (AMSE). MI serves as an information-theoretic measure of correlation between model and human state variables. For two random variables X and Y, the MI quantifies the information gain about one variable by observing the other:

$$\text{MI}(X,Y) = \iint P(x,y) \log \left(\frac{P(x,y)}{P(x)P(y)} \right) dx\, dy.$$

We estimate the probability densities using Gaussian kernel density estimation.

In addition to MI, we calculate the mean squared error between the model's state variables and their analytical fit. State variables that yield accurate and simple analytical expressions are typically more interpretable and aligned with human understanding. This metric, referred to as AMSE, quantifies interpretability by measuring the accuracy of the symbolic expression.

Intrinsic dimension estimation

Accurately estimating the intrinsic dimension is a crucial step in the state variable discovery process, as it determines the latent dimension for the network that will subsequently learn the state variables. In Table 1, we compare the ID estimates from our model (TIDE) with the baseline Neural State Variables (NSV) approach from (Chen et al., 2022). Ground truth ID values, where available, are also included. For the baseline, we follow their outlined methodology to calculate the estimate. To ensure a fair comparison, we use $\mathbb{R}^{64}$ as the intermediate latent space for all ID estimation algorithms. Each method is trained on three random training, validation, and test splits, and we report the mean and standard deviation.

For datasets where the ID is known, our model accurately identifies the necessary number of state variables. TIDE estimates tend to align more closely with the ground truth, likely due to the time-derivative regularization, which encourages a smooth and continuous latent representation, resulting in more precise estimates. Further exploration of the latent space is provided in the subsequent analysis. For datasets with unknown ID, both models generally produce consistent estimates, although TIDE predicts a slightly smaller value for the fire flame dataset. To determine the latent dimension for the subsequent model, we use the ground truth value when available, otherwise rounding the predicted ID to the nearest integer.

State variable visualization

We present the learned latent space and state variables, and qualitatively compare them to the baseline approach. Figure 4 shows the state variables over a single video for both the single and double pendulum. Our learned variables closely resemble human-interpretable counterparts by maintaining continuity over time, while the baseline approach exhibits discontinuous jumps throughout the video. Such behavior diverges from typical human variables, like angle and angular velocity, which do not experience abrupt changes between consecutive observations. Additional visualizations of state variables for other datasets can be found in the Supplemental Information.

Our state variables not only exhibit qualitative similarity to their human equivalents but also capture meaningful information about observable physical phenomena. For example, in the case of the single pendulum, the angle of the pendulum arm is inversely correlated with the value of z_0: the

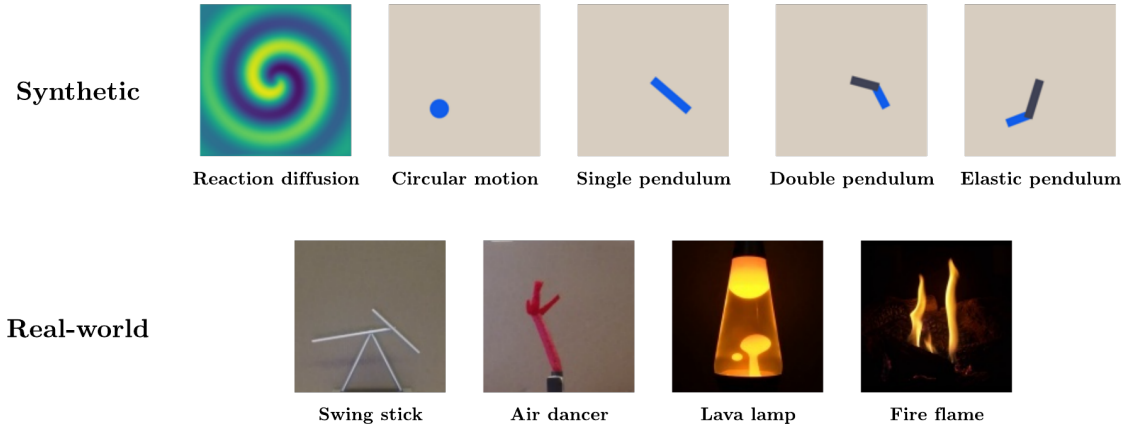


Figure 3. Grid showing the nine evaluation video datasets.

Table 1. Intrinsic dimension (ID) estimates for the nine benchmark datasets from both our model (TIDE) and the baseline (NSV), including the ground truth value where available. The mean and standard deviation of the ID are reported across three random training splits. TIDE consistently provides more precise estimates, which we attribute to the incorporation of time-derivative regularization.

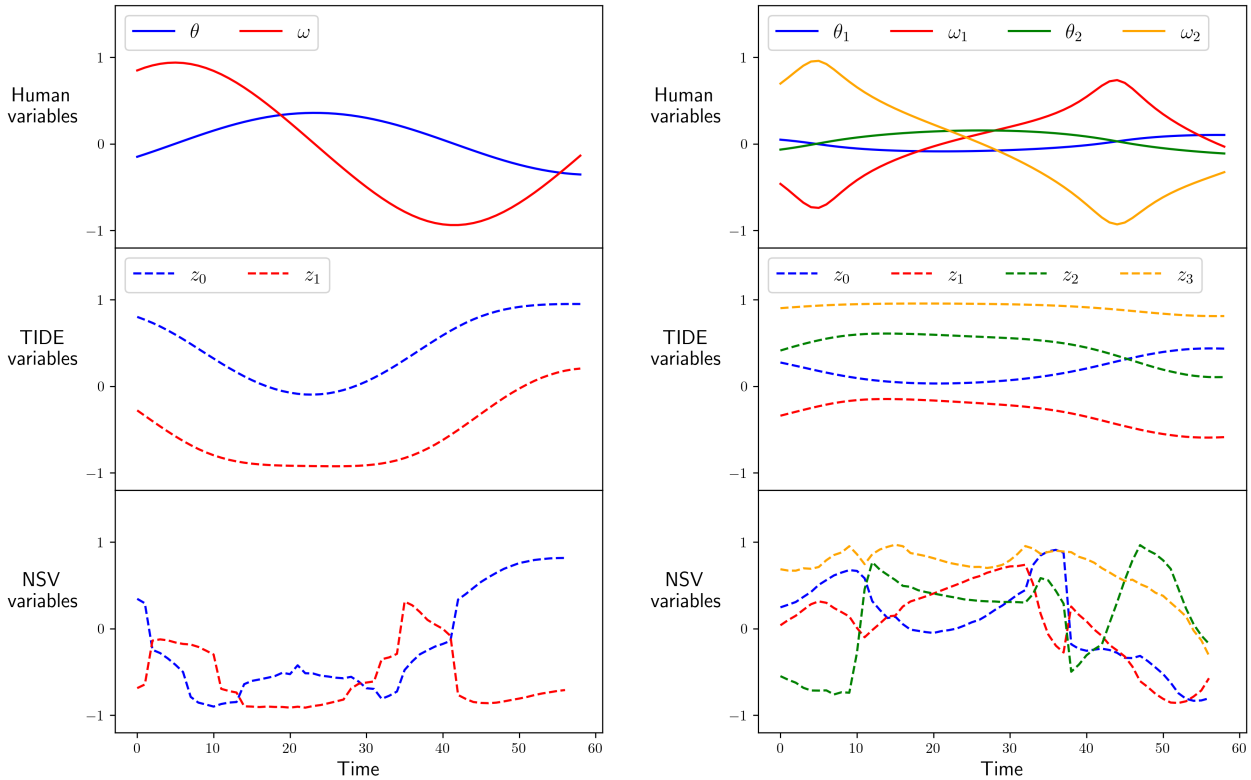
Dataset	TIDE ID (ours)	NSV ID (baseline)	Ground truth ID
Reaction diffusion	2.17 ± 0.07	2.17 ± 0.16	2
Circular motion	2.13 ± 0.02	2.10 ± 0.03	2
Single pendulum	2.17 ± 0.01	2.15 ± 0.03	2
Double pendulum	4.02 ± 0.03	3.52 ± 0.08	4
Elastic pendulum	6.00 ± 0.04	4.46 ± 0.04	6
Swing stick	4.26 ± 0.59	3.86 ± 0.09	4
Air dancer	3.60 ± 0.33	4.29 ± 0.24	Unknown
Lava lamp	4.99 ± 0.35	5.05 ± 0.32	Unknown
Fire flame	8.01 ± 0.01	10.39 ± 0.75	Unknown

angle reaches its maximum when z_0 is at its minimum, and vice versa. It is important to note that our representation does not directly correspond to standard human quantities, as state variables are not unique. Rather than explicitly predicting human variables, our goal is to align the human and AI representations as closely as possible, without relying on prior knowledge of the system.

For the circular motion and single pendulum datasets, we analyze the phase space of the state variables in Figure 5. Each phase space is visualized using two colored point sets, where the colors represent the values of the human variables. Our model generates a continuous latent space, with angular and radial directions corresponding to the angle and angular velocity, respectively. In contrast, the baseline model produces a fragmented phase space. This difference is primarily due to the regularization penalty in our approach, which reduces discontinuities in the phase space by minimizing discrete derivatives. To demonstrate this effect, we show state variable transitions over four observations within a 10-frame segment. Our variables evolve smoothly over time, whereas the baseline model exhibits abrupt changes across different regions in the latent space.

Interpretability metrics

TIDE learns a continuous latent space with smoother state variables that more closely resemble human quantities. We conduct further quantitative analysis to demonstrate a stronger correlation between human and AI variables, based on smoothness, mutual information (MI), and analytical mean squared error (AMSE).



(a) *(b)*

Figure 4. Visual depiction of state variables for the single pendulum 4a and double pendulum 4b over a single video. Human variables are represented with solid lines, while model variables are shown with dashed lines. TIDE learns smoother variables that more closely resemble the angle and angular velocity compared to the baseline. We also depict the symbolic regression fit for the first latent variable in a solid black line. The analytical expressions are provided in Equations 1 and 2.

Table 2. Smoothness and correlation metrics between human and model variables across nine benchmark datasets. Higher values for MI, and lower values for smoothness and AMSE are preferred, as indicated by the arrows. Bold entries highlight the best-performing model for each metric and dataset, while missing values indicate cases where standard human variables are unavailable. TIDE learns smoother state variables that exhibit a stronger correlation with the angle and angular velocity compared to the baseline.

Dataset	**Smoothness (↓)**		**MI (↑)**		**AMSE (↓)**	
	TIDE	NSV	TIDE	NSV	TIDE	NSV
Reaction diffusion	$\mathbf{1.16 \pm 0.06}$	94.85 ± 9.86				
Circular motion	$\mathbf{2.52 \pm 0.08}$	438.09 ± 41.76	$\mathbf{0.44 \pm 0.09}$	0.30 ± 0.01	$\mathbf{0.005 \pm 0.002}$	0.34 ± 0.018
Single pendulum	$\mathbf{1.13 \pm 0.06}$	177.78 ± 29.32	$\mathbf{0.68 \pm 0.04}$	0.48 ± 0.01	$\mathbf{0.006 \pm 0.001}$	0.34 ± 0.033
Double pendulum	$\mathbf{1.84 \pm 0.18}$	52.03 ± 12.93	$\mathbf{1.39 \pm 0.01}$	1.35 ± 0.10	$\mathbf{0.007 \pm 0.002}$	0.26 ± 0.019
Elastic pendulum	$\mathbf{4.89 \pm 0.46}$	35.62 ± 0.77	$\mathbf{2.08 \pm 0.18}$	2.05 ± 0.08	$\mathbf{0.012 \pm 0.007}$	0.20 ± 0.002
Swing stick	$\mathbf{6.26 \pm 0.75}$	29.7 ± 12.89				
Air dancer	$\mathbf{3.79 \pm 0.23}$	83.27 ± 41.0				
Lava lamp	$\mathbf{26.95 \pm 1.17}$	158.63 ± 8.17				
Fire flame	$\mathbf{54.74 \pm 4.27}$	415.45 ± 35.45				

Table 2 reports smoothness penalty ($\mathcal{L}_{\text{reg}}$), MI, and AMSE values for the circular motion, single pendulum, and double pendulum datasets. Across all systems, our variables are consistently smoother compared to the baseline. Ad-

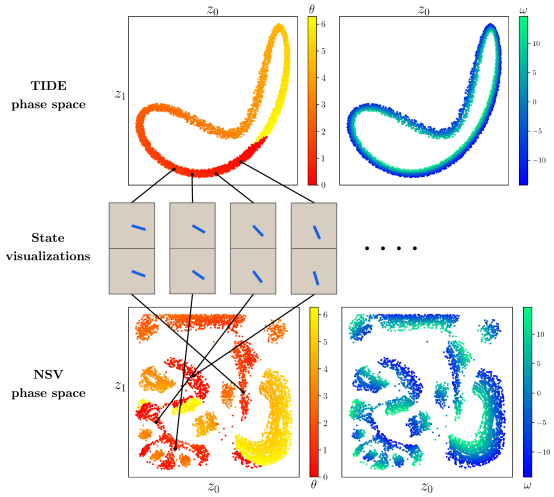


Figure 5. Phase space of the model variables on the single pendulum dataset, colored by the angle θ and angular velocity ω. TIDE (top) generates a continuous latent space, a result of the time-derivative regularization. In contrast, the baseline (bottom) shows discontinuous transitions between consecutive states, which is less aligned with physical intuition.

ditionally, the MI between TIDE and human variables is higher, indicating a stronger correlation. By employing symbolic regression to derive an analytical expression for the latent representation, we achieve significantly lower AMSE, further validating the use of smoothness as a proxy for interpretability when prior knowledge of the system or its dynamics is unavailable.

In Figure 4, we present the state variables and their analytical fits over a single video for the single and double pendulum datasets. We accurately derive an expression for the TIDE variables using symbolic regression, where the first latent component for the single and double pendulum is respectively described as:

$$z_0 = (0.27 - 0.15\sin(\theta + 1.48))\sin(2\theta - 0.9) - 0.74\sin(\theta) - 0.26, \quad (1)$$

$$z_0 = (0.2\sin(\theta_1 - 0.34) + 0.47)\sin(\theta_1 - 0.17) + 0.047(\sin(\theta_1 + 0.6) + 1)\sin(\theta_2 - 0.28) - 0.67. \quad (2)$$

In contrast, no symbolic combination of elementary functions, angle, and angular velocity succeeds in capturing information about the baseline variables. Additional analytical fits and equations are included in the Supplemental Information.

5. Conclusion

Automated scientific discovery is becoming increasingly feasible with the rise of data-driven methodologies. This process involves extracting descriptive variables from observations and formulating new laws and equations based on these variables. Traditionally, identifying state variables has relied on human expertise, but recent advancements aim to automatically discover variables from raw data without prior knowledge of the system. However, previous approaches have struggled to produce human-interpretable physics due to the non-uniqueness of state representations. While identifying state variables is crucial, it is equally important to prioritize those that offer interpretive value for scientific discovery.

In this work, we introduce a novel Temporally-Informed Dynamics Encoder (TIDE) designed to derive meaningful state variables from observational data without requiring prior knowledge of the system, statistical relationships between variables, or parametric assumptions for the dynamics. Our approach emphasizes aligning AI representations with human-interpretable state variables based on three key desired characteristics. These properties are incorporated into an autoencoder framework by jointly optimizing a dynamics module and applying time-derivative regularization. TIDE significantly outperforms the existing neural state variable baseline by learning variables with meaningful physical significance and stronger correlation to human quantities. Furthermore, we derive analytical expressions for the predicted variables using symbolic regression. To the best of our knowledge, this is the first result to derive explicit formulas for state variables learned *without prior knowledge of the system*.

A current limitation of our method is its reliance on the stable state variable representation assumption, which may not apply in scenarios where data collection is inefficient. One potential direction for enhancing the interpretive value of learned variables in broader settings is to incorporate semi-supervised or curriculum learning. In this approach, knowledge of intuitive human variables, such as angle or speed, could be transferred to novel environments where similar underlying variables may be hidden in the data. Although our method aligns model and human variables by encouraging specific properties in the latent space, the resulting representation may not perfectly match human quantities, given the non-uniqueness of state variable configurations. Another promising avenue is to theoretically explore sufficient conditions on the data-generating process under which state variables become identifiable.

6. Acknowledgements

This research was supported in part by NSF AI Institute for Dynamical Systems #2112085, DARPA MTO Lifelong Learning Machines (L2M) Program HR0011-18-2-0020, NSF NRI #1925157, NSF DMS #1937254, NSF DMS #2012562, NSF CCF #1704833, and DE #SC0022317. The code and implementation for our work is available at https://github.com/kzhangm02/tide.

References

Akiyama, K., Alberdi, A., Alef, W., Asada, K., Azulay, R., and et al. First m87 event horizon telescope results. iv. imaging the central supermassive black hole. *The Astrophysical Journal Letters*, 875:L4, 2019. doi: 10.3847/2041-8213/ab0e85.

Barfoot, T. D., Forbes, J. R., and Yoon, D. J. Exactly sparse gaussian variational inference with application to derivative-free batch nonlinear state estimation. *CoRR*, abs/1911.08333, 2019.

Bongard, J. and Lipson, H. Automated reverse engineering of nonlinear dynamical systems. *Proceedings of the National Academy of Sciences*, 104:9943–9948, 2007. doi: 10.1073/pnas.0609476104.

Brunton, S. L., Proctor, J. L., and Kutz, J. N. Discovering governing equations from data by sparse identification of nonlinear dynamical systems. *Proceedings of the National Academy of Sciences*, 113:3932–3937, 2016. doi: 10.1073/pnas.1517384113.

Carrassi, A., Bocquet, M., Bertino, L., and Evensen, G. Data assimilation in the geosciences: an overview of methods, issues, and perspectives. *WIREs Climate Change*, 9, 2018. doi: 10.1002/wcc.535.

Ceruti, C., Bassis, S., Rozza, A., Lombardi, G., Casiraghi, E., and Campadelli, P. Danco: an intrinsic dimensionality estimator exploiting angle and norm concentration. *Pattern Recognition*, 47:2569–2581, 2014. doi: 10.1016/j.patcog.2014.02.013.

Champion, K., Lusch, B., Kutz, J. N., and Brunton, S. L. Data-driven discovery of coordinates and governing equations. *Proceedings of the National Academy of Sciences*, 116:22445–22451, 2019. doi: 10.1073/pnas.1906995116.

Chen, B., Huang, K., Raghupathi, S., Chandratreya, I., Du, Q., and Lipson, H. Automated discovery of fundamental variables hidden in experimental data. *Nature Computational Science*, 2:433–442, 2022. doi: 10.1038/s43588-022-00281-6.

Chen, Y. and Bollt, E. M. Modeling and nonlinear parameter estimation with kronecker product representation for coupled oscillators and spatiotemporal systems. *Physica D: Nonlinear Phenomena*, 227:78–99, 2007. doi: 10.1016/j.physd.2006.12.006.

Coley, C. W., Thomas, D. A., Lummiss, J. A. M., Jaworski, J. N., Breen, C., Schultz, V., Hart, T., Fishman, J., Rogers, L., Gao, H., Hicklin, R. W., Plehiers, P., Byington, J., Piotti, J. S., Green, W., Hart, A. J., Jamison, T. F., and Jensen, K. F. A robotic platform for flow synthesis of organic compounds informed by ai planning. *Science*, 365, 2019. doi: 10.1126/science.aax1566.

Cranmer, M. Interpretable machine learning for science with pysr and symbolicregression.jl, 2023.

Cranmer, M., Sanchez-Gonzalez, A., Battaglia, P., Xu, R., Cranmer, K., Spergel, D., and Ho, S. Discovering symbolic models from deep learning with inductive biases. 2020. doi: 10.48550/arxiv.2006.11287.

Davies, A., Veličković, P., Buesing, L., Blackwell, S., Zheng, D., Tomašev, N., Tanburn, R., Battaglia, P., Blundell, C., Juhász, A., and et al. Advancing mathematics by guiding human intuition with ai. *Nature*, 600(7887): 70–74, 2021. doi: 10.1038/s41586-021-04086-x.

Ennab, M. and Mcheick, H. Designing an interpretability-based model to explain the artificial intelligence algorithms in healthcare. *Diagnostics*, 12:1557, 2022. doi: 10.3390/diagnostics12071557.

Evans, J. A. and Rzhetsky, A. Machine science. *Science*, 329:399–400, 2010. doi: 10.1126/science.1189416.

Fortunato, S., Bergstrom, C. T., Börner, K., Evans, J. A., Helbing, D., Milojević, S., Petersen, A. M., Radicchi, F., Sinatra, R., Uzzi, B., Vespignani, A., Waltman, L., Wang, D., and Barabási, A. L. Science of science. *Science*, 359, 2018. doi: 10.1126/science.aao0185.

Garcon, A., Vexler, J., Budker, D., and Krämer, S. Deep neural networks to recover unknown physical parameters from oscillating time series. *Plos One*, 17:e0268439, 2022. doi: 10.1371/journal.pone.0268439.

Goldfeder, J., Wyder, P., LeCun, Y., and Ziv, R. S. Ai must embrace specialization via superhuman adaptable intelligence. *arXiv preprint arXiv:2602.23643*, 2026.

Guo, G., Goldfeder, J., Lan, L., Ray, A., Yang, A. H., Chen, B., Billinge, S. J., and Lipson, H. Towards end-to-end structure determination from x-ray diffraction data using deep learning. *npj Computational Materials*, 10(1):209, 2024.

Guo, Z. C., Korotkine, V., Forbes, J. R., and Barfoot, T. D. Koopman linearization for data-driven batch state estimation of control-affine systems. *CoRR*, abs/2109.07000, 2021.

Han, J., Zhang, L., Car, R., and Weinan, E. Deep potential: a general representation of a many-body potential energy surface. *Communications in Computational Physics*, 23, 2018. doi: 10.4208/cicp.oa-2017-0213.

Harvey, A., Ruiz, E., and Shephard, N. Multivariate stochastic variance models. *The Review of Economic Studies*, 61 (2):247–264, 1994.

Higgins, I., Matthey, L., Pal, A., Burgess, C., Glorot, X., Botvinick, M., Mohamed, S., and Lerchner, A. beta-vae: Learning basic visual concepts with a constrained variational framework. In *International Conference on Learning Representations*, 2017.

Hirsch, M., Smale, S., and Devaney, R. *Differential Equations, Dynamical Systems, and an Introduction to Chaos*. Academic Press, 2012.

Houtekamer, P. L. and Zhang, F. Review of the ensemble kalman filter for atmospheric data assimilation. *Monthly Weather Review*, 144:4489–4532, 2016. doi: 10.1175/mwr-d-15-0440.1.

Jumper, J., Evans, R., Pritzel, A., Green, T., Figurnov, M., Ronneberger, O., Tunyasuvunakool, K., Bates, R., Žídek, A., Potapenko, A., Bridgland, A., Meyer, C., Kohl, S. A. A., Ballard, A. J., Cowie, A., Romera-Paredes, B., Nikolov, S., Jain, R., Adler, J., Back, T., Petersen, S., Reiman, D., Clancy, E., Zieliński, M., Steinegger, M., Pacholska, M., Berghammer, T., Bodenstein, S., Silver, D., Vinyals, O., Senior, A. W., Kavukcuoglu, K., Kohli, P., and Hassabis, D. Highly accurate protein structure prediction with alphafold. *Nature*, 596:583–589, 2021. doi: 10.1038/s41586-021-03819-2.

Kaur, H., Nori, H., Jenkins, S., Caruana, R., Wallach, H., and Vaughan, J. Interpreting interpretability: understanding data scientists' use of interpretability tools for machine learning. *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, 2020. doi: 10.1145/3313831.3376219.

Keil, A. *Dynamic Variational Level Sets for Cardiac 4D Reconstruction*. PhD thesis, Technische Univ. München, 2010.

Kevrekidis, I. G., Gear, C. W., Hyman, J. M., Kevrekidis, P. G., Runborg, O., and Theodoropoulos, C. Equation-free multiscale computation: enabling microscopic simulators to perform system-level tasks, 2002.

King, R. D., Muggleton, S., Srinivasan, A., and Sternberg, M. J. Structure-activity relationships derived by machine learning: the use of atoms and their bond connectivities to predict mutagenicity by inductive logic programming. *Proceedings of the National Academy of Sciences*, 93: 438–442, 1996. doi: 10.1073/pnas.93.1.438.

King, R. D., Whelan, K. E., Jones, F., Reiser, P. G. K., Bryant, C. H., Muggleton, S., Kell, D. B., and Oliver, S. G. Functional genomic hypothesis generation and experimentation by a robot scientist. *Nature*, 427:247–252, 2004. doi: 10.1038/nature02236.

Kramer, S., Cerrato, M., Džeroski, S., and King, R. Automated scientific discovery: From equation discovery to autonomous discovery systems, 2023.

Kronberger, G., Kammerer, L., and Kommenda, M. Identification of dynamical systems using symbolic regression. *Computer Aided Systems Theory – EUROCAST 2019*, pp. 370–377, 2020. doi: 10.1007/978-3-030-45093-9_45.

Kutz, N. J., Brunton, L. S., Bingni, B. W., and Proctor, L. J. *Dynamic Mode Decomposition: Data-Driven Modeling of Complex Systems*. SIAM, 2016.

LeCun, Y., Bengio, Y., and Hinton, G. E. Deep learning. *Nature*, 521:436–444, 2015. doi: 10.1038/nature14539.

Lin, J., Ning, J., Goldfeder, J., and Lipson, H. Creative synthesis of kinematic mechanisms. *arXiv preprint arXiv:2510.03308*, 2025.

Lipton, Z. C. The mythos of model interpretability, 2017.

Lomasov, S., Goldfeder, J., Erol, M. H., So, M., Yan, Y., Howard, A., Kutz, N., and Ziv, R. S. Exploring human-ai conceptual alignment through the prism of chess. *arXiv preprint arXiv:2510.26025*, 2025.

Mangan, N. M., Kutz, J. N., Brunton, S. L., and Proctor, J. L. Model selection for dynamical systems via sparse regression and information criteria. *Proceedings of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 473:20170009, 2017. doi: 10.1098/rspa.2017.0009.

Mrowca, D., Zhuang, C., Wang, E., Haber, N., Fei-Fei, L., Tenenbaum, J. B., and Yamins, D. L. K. Flexible neural representation for physics prediction, 2018.

Raissi, M. and Karniadakis, G. E. Machine learning of linear differential equations using gaussian processes. 2017. doi: 10.48550/arxiv.1701.02440.

Raissi, M., Perdikaris, P., and Karniadakis, G. E. Physics informed deep learning (part ii): data-driven discovery of nonlinear partial differential equations. 2017. doi: 10.48550/arxiv.1711.10566.

Schmidt, M. D. and Lipson, H. Distilling free-form natural laws from experimental data. *Science*, 324:81–85, 2009. doi: 10.1126/science.1165893.

Thompson, J. M. T. and Steward, H. B. *Nonlinear Dynamics and Chaos*. Wiley, 2002.

Wagner, A. Z. Constructions in combinatorics via neural networks, 2021.

Waltz, D. L. and Buchanan, B. G. Automating science. *Science*, 324:43–44, 2009. doi: 10.1126/science.1172781.

Wu, K. and Xiu, D. Numerical aspects for approximating governing equations using data. *Journal of Computational Physics*, 384:200–221, 2019. doi: 10.1016/j.jcp.2019.01.030.

Wu, K., Qin, T., and Xiu, D. Structure-preserving method for reconstructing unknown hamiltonian systems from trajectory data. 2019. doi: 10.48550/arxiv.1905.10396.

Zhang, L., Han, J., Wang, H., Car, R., and Weinan, E. Deep potential molecular dynamics: a scalable model with the accuracy of quantum mechanics. *Physical Review Letters*, 120, 2018. doi: 10.1103/physrevlett.120.143001.

A. Qualitative reconstructions

To assess reconstruction quality using TIDE, we compare the predicted images from our model and a baseline approach in Figure 6. Specifically, we include visualizations from the double pendulum and fire flame datasets, as these exhibit the largest reconstruction discrepancy. In the double pendulum case, the additional error largely results from distortions in the pendulum arm, although key physical attributes such as angle and angular velocity are still accurately captured. Similarly, for the fire flame dataset, our model generates images with some added blur but successfully reconstructs the shape and orientation of the flames. These artifacts may be reduced through improved hyperparameter selection. Overall, TIDE networks are capable of learning system dynamics while simultaneously producing a more interpretable state variable representation.

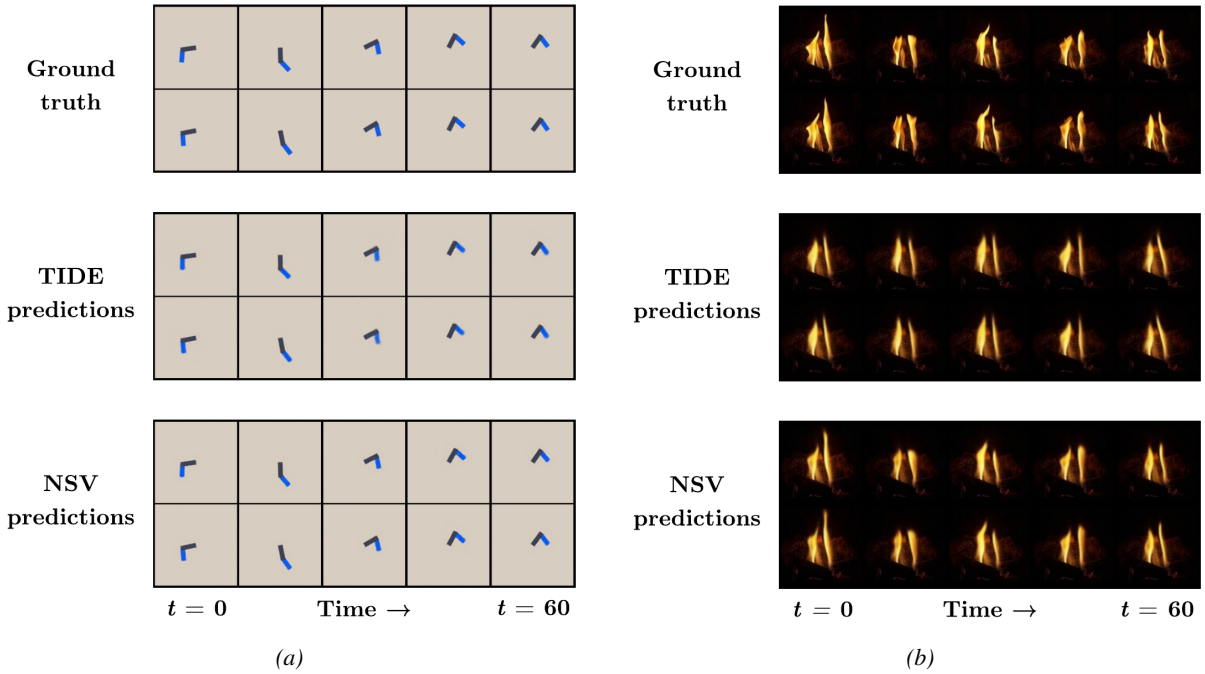


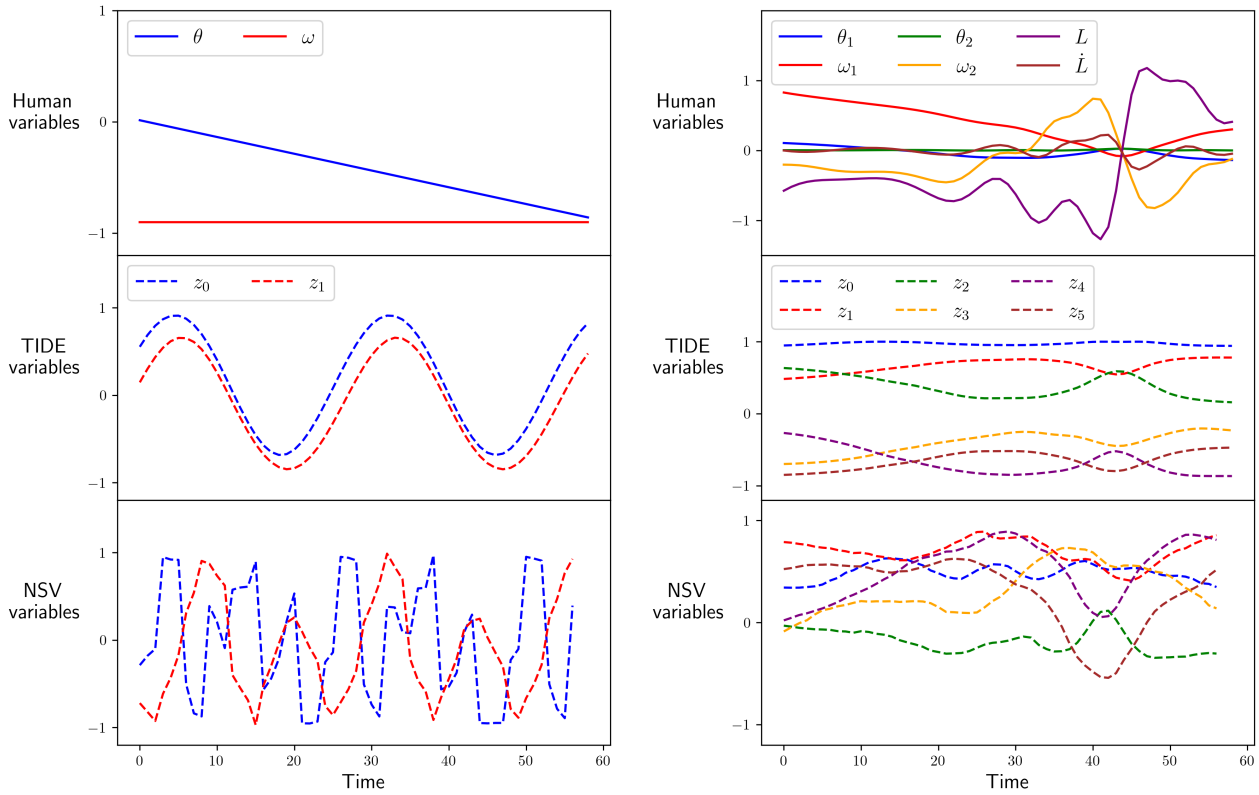
Figure 6. Comparison of reconstructed states for the double pendulum (a) and fire flame (b) datasets. TIDE successfully captures the key physical elements and dynamics of the systems.

B. Additional state variable visualizations

We illustrate the state variables for the circular motion and elastic pendulum datasets in Figure 7. Similar to the single and double pendulum scenarios, TIDE learns variables that are smoother and more closely aligned with human-interpretable quantities. Additionally, Figure 8 shows the phase space of the circular motion dataset, consisting of two distinct ellipses that correspond to positive and negative angular velocity. This observation further supports TIDE's ability to capture physically meaningful quantities. To demonstrate the effect of time-derivative regularization, we include visualizations of four states across a 10-frame segment, yielding results consistent with those observed for the single pendulum. Since the angular velocity remains constant throughout each video, the state variables evolve smoothly over time.

C. Additional state variable analytical fits

We present the analytical fit of all TIDE state variables for the circular motion, single pendulum, double pendulum, and elastic pendulum datasets in Figure 9. Our model accurately predicts state variables expressed in terms of angle and angular velocity, as the learned representation aligns closely with human-interpretable quantities. For conciseness, we omit the analytical fit from the baseline model, as its behavior mirrors the NSV results shown in Figure 2.



(a) *(b)*

Figure 7. Visual depiction of state variables for the circular motion (a) and elastic pendulum (b) across a single video. Standard human variables are represented by solid lines, while the model-predicted variables are shown with dashed lines. In the case of the elastic pendulum, L denotes the length.

D. State variable equations

We present the analytical expressions for TIDE variables obtained through symbolic regression in Table 3. To the best of our knowledge, this is the first instance of deriving explicit formulas for state variables learned *without prior knowledge of the system*.

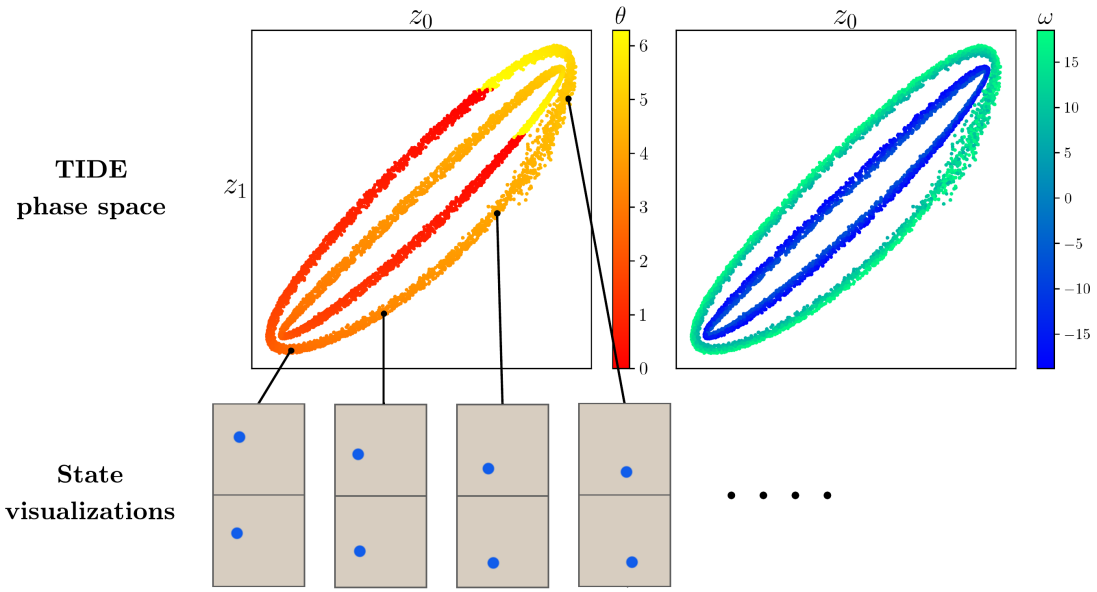


Figure 8. Phase space of TIDE state variables for the circular motion dataset, colored by angle θ and angular velocity ω. Four states from a 10-frame segment are visualized to illustrate the effect of regularization.

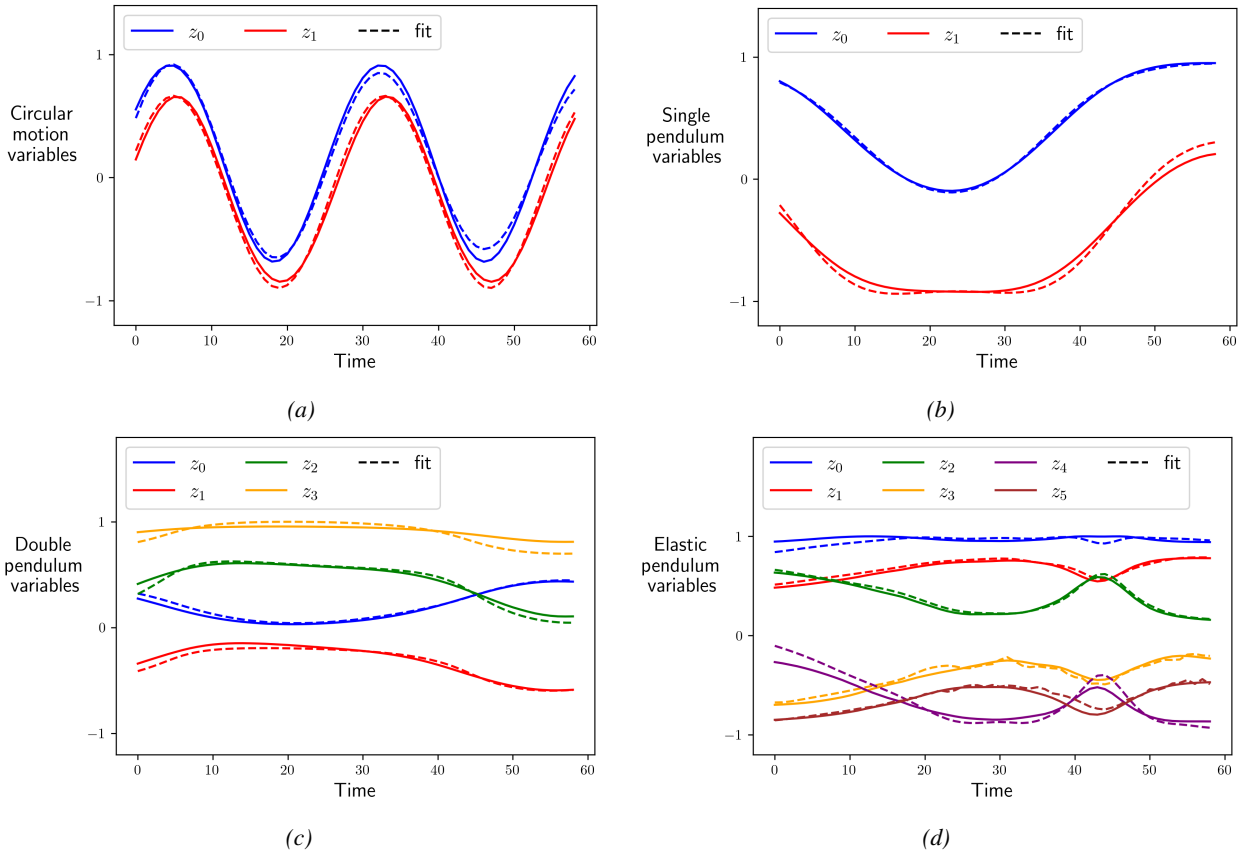


Figure 9. Analytical fits for all TIDE state variables for the circular motion (a), single pendulum (b), double pendulum (c), and elastic pendulum (d) datasets. The model-predicted variables are shown as solid lines, while their corresponding analytical fits are represented by dashed lines.

Dataset	Equations
Circular motion	$z_1 = -0.64\sin(\theta) + 0.5\cos(1.03\theta) - 0.5\cos(0.018\omega) + 0.6$
	$z_2 = 0.0068\omega\sin(\theta + 0.96) - 0.79\sin(\theta - 0.75) - 0.12$
Single pendulum	$z_1 = 0.26(0.0085\omega - 0.52\sin(\theta) - \cos(\theta) - 0.58)^2 - \sin(\theta)$
	$z_2 = -0.6\sin(\theta + 0.27) - (\sin(\theta) + \sin(0.0012\omega) - \sin(\theta - 0.72))^2$
Double pendulum	$z_1 = 0.38\sin(\theta_1 - 0.85)^2 + 0.059\sin(\theta_1 + 0.66)^2 + 0.55\sin(\theta_1)$
	$z_2 = -0.12\sin(\theta_2) - \sin(0.52\theta_1 + 0.51)^2 - 0.057$
	$z_3 = -0.00021\sin(\omega_2(\omega_1 - \theta_1 + 1.17))^2 - 0.66\sin(\theta_1 - 0.58) - 0.18\sin(\theta_2 + 5.3)$
	$z_4 = 0.00063(\omega_2 - \omega_1) - 0.0013\theta_2 - 0.25\sin(\theta_1 - 0.48) + 0.77$
Elastic pendulum	$z_1 = -(\theta_2 + 0.21)\sin(\theta_1)^2 - (\sin(\theta_1) - 0.0076)(\sin(\theta_2) + 0.12) + 0.97$
	$z_2 = -(\sin(1.05\theta_1)\sin(\theta_2 + 0.29) + 0.49)^2 + 0.073\sin(0.075\omega_1^2)$ $- 0.011\sin(\omega_2)^2\sin(L)^2\sin(\theta_2^2\dot{L}^2) + 0.87$
	$z_3 = 0.93(\sin(\theta_2) + 0.45)\sin(\theta_1 + \theta_2 + 0.062) + 0.41$
	$z_4 = -0.0036\omega_2 + (\theta_2 + 0.23)(\theta_2 - 1.8\sin(\theta_1))$ $- 0.0037(\sin(\omega_1) + \sin(L) + 3.03)^2 - 0.43$
	$z_5 = 0.47\theta_2 + 0.0012\omega_2 - 0.00035\dot{L}\sin(\omega_1 + L^2) - 0.0012(\dot{L} - 0.48)^2$ $+ 0.47(\sin(\theta_1 + 0.25) + 0.64)^2 - 0.94$
	$z_6 = -(\sin(\theta_2) + 0.32)\sin(\theta_1) + 0.07\sin(\omega_1\omega_2(L - \dot{L}))^2\sin(\theta_1 - 0.4)^2 - 0.69$

Table 3. Analytical expressions for all TIDE state variables in terms of standard human-interpretable variables for the circular motion, single pendulum, double pendulum, and elastic pendulum datasets. For compactness, all constants and coefficients smaller than 10^{-4} are omitted.