Constructing Neural Network Based Models for Simulating Dynamical Systems

CHRISTIAN LEGAARD, Aarhus University
THOMAS SCHRANZ and GERALD SCHWEIGER, TU Graz
JÁN DRGOŇA, Pacific Northwest National Laboratory
BASAK FALAY, AEE–Institute for Sustainable Technologies
CLÁUDIO GOMES, ALEXANDROS IOSIFIDIS, MAHDI ABKAR, and
PETER LARSEN, Aarhus University

Dynamical systems see widespread use in natural sciences like physics, biology, and chemistry, as well as engineering disciplines such as circuit analysis, computational fluid dynamics, and control. For simple systems, the differential equations governing the dynamics can be derived by applying fundamental physical laws. However, for more complex systems, this approach becomes exceedingly difficult. Data-driven modeling is an alternative paradigm that seeks to learn an approximation of the dynamics of a system using observations of the true system. In recent years, there has been an increased interest in applying data-driven modeling techniques to solve a wide range of problems in physics and engineering. This article provides a survey of the different ways to construct models of dynamical systems using neural networks. In addition to the basic overview, we review the related literature and outline the most significant challenges from numerical simulations that this modeling paradigm must overcome. Based on the reviewed literature and identified challenges, we provide a discussion on promising research areas.

CCS Concepts: • **Computing methodologies** → **Neural networks**; **Continuous simulation**; **Continuous models**; *Supervised learning by regression*; • **Applied computing** → *Physics*; *Engineering;*

Additional Key Words and Phrases: Neural ODEs, physics-informed neural networks, physics-based regularization

We acknowledge the Poul Due Jensen Foundation for funding the project Digital Twins for Cyber-Physical Systems (DiT4CPS) and Legaard would also like to acknowledge partial support from the MADE Digital project.
This research was supported by the Data Model Convergence (DMC) initiative via the Laboratory Directed Research and Development (LDRD) investments at Pacific Northwest National Laboratory (PNNL). PNNL is a multi-program national laboratory operated for the U.S. Department of Energy (DoE) by Battelle Memorial Institute under Contract No. DE-AC05-76RL0-1830.
The Digital Twins for Cyber-Physical Systems (DiT4CPS) project was funded by the Poul Due Jensen Foundation. C. Legaard received partial support from the MADE Digital project. This research was supported by the Data Model Convergence (DMC) initiative via the Laboratory Directed Research and Development (LDRD) investments at Pacific Northwest National Laboratory (PNNL). PNNL is a multi-program national laboratory operated for the U.S. Department of Energy (DoE) by Battelle Memorial Institute under Contract No. DE-AC05-76RL0-1830.
Authors' addresses: C. Legaard, C. Gomes, A. Iosifidis, M. Abkar, and P. Larsen, Aarhus University, Aarhus, Denmark; emails: {cml, claudio.gomes, ai}@ece.au.dk, abkar@mpe.au.dk, pgl@ece.au.dk; T. Schranz and G. Schweiger, TU Graz, Graz, Austria; emails: {thomas.schranz, gerald.schweiger}@tugraz.at; J. Drgoňa, Pacific Northwest National Laboratory, Richland, WA; email: jan.drgona@pnnl.gov; B. Falay, AEE–Institute for Sustainable Technologies, Gleisdorf, Austria; email: b.falay@aee.at.

0360-0300/2023/02-ART236 $15.00
https://doi.org/10.1145/3567591

ACM Reference format:
Christian Legaard, Thomas Schranz, Gerald Schweiger, Ján Drgoňa, Basak Falay, Cláudio Gomes, Alexandros Iosifidis, Mahdi Abkar, and Peter Larsen. 2023. Constructing Neural Network Based Models for Simulating Dynamical Systems. *ACM Comput. Surv.* 55, 11, Article 236 (February 2023), 34 pages.
https://doi.org/10.1145/3567591

1 INTRODUCTION

Mathematical models are fundamental tools for building an understanding of the physical phenomena observed in nature [13]. Not only do these models allow us to predict what the future may look like, but they also allow us to develop an understanding of what causes the observed behavior. In engineering, models are used to improve the system design [33, 118], design optimal control policy [23, 25, 35], simulate faults [84, 94], forecast future behavior [122], or assess the desired operational performance [51].

The focus of this survey is on the type of models that allow us to predict how a physical system evolves over time for a given set of conditions. Dynamical systems theory provides an essential set of tools for formalizing and studying the dynamics of this type of model. However, when studying complex physical phenomena, it becomes increasingly difficult to derive models by hand that strike an acceptable balance between accuracy and speed. This has led to the development of fields that are concerned with creating models directly from data such as *system identification* [76, 87], **machine learning (ML)** [9, 85], and, more recently, **deep learning (DL)** [40].

In recent years, the interest in DL has increased rapidly, as is evident from the volume of research being published on the topic [95]. The exact causes behind the success of **neural networks (NNs)** are hard to pinpoint. Some claim that practical factors like the availability of large quantities of data, user-friendly software frameworks [1, 93], and specialized hardware [82] are the main cause for its success, whereas others claim that the success of NNs can be attributed to their structure being well suited to solving a wide variety of problems [95].

The goal of this survey is to provide a practical guide on how to construct models of dynamical systems using NNs as primary building blocks. We do this by walking the reader through the most important classes of models found in the literature, for many of which we provide an example implementation. We put special emphasis on the process for training the models, since it differs significantly from traditional applications of DL that do not consider evolution over time. More specifically, we describe how to split the trajectories used during training, and we introduce optimization criteria suitable for simulation. After training, it is necessary to validate that the model is a good representation of the true system. Like other data-driven models, we determine the validity empirically by using a separate set of trajectories for validation. We introduce some of the most important properties and how they can be verified.

It should be emphasized that the type of model we wish to construct should allow us to obtain a simulation of the system. Rather than providing a formal definition of simulation, we refer to Figure 1, which shows several topics related to simulation that are not covered in this article.

The source code and instructions for running the experiments can be accessed at GitHub.[1]

1.1 Related Surveys

We provide an overview of existing surveys related to our work. Then we compare our work with these surveys and describe the structure of the remainder of the article.

[1]https://github.com/clegaard/deep_learning_for_dynamical_systems.

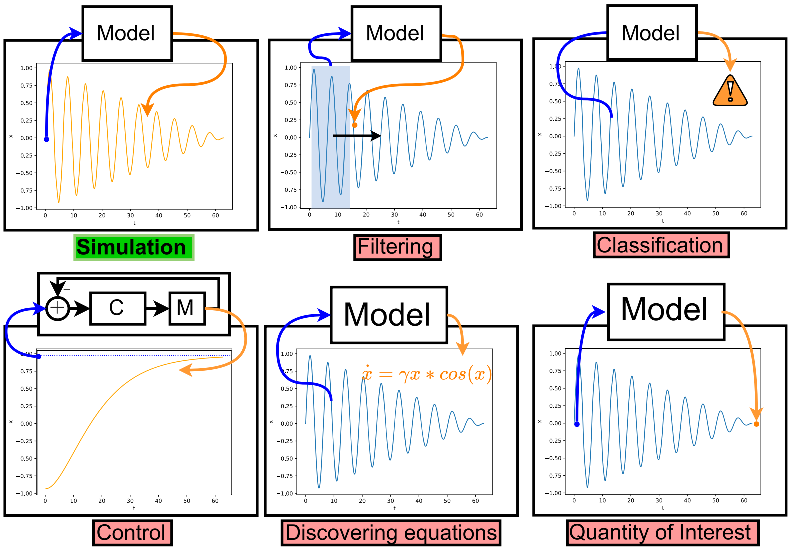


Fig. 1. *Simulation* and related application areas where ML techniques are commonly applied. The focus of the survey is exclusively on techniques that can generate a simulation based on an initial condition, as shown in the top left. Although interesting on their own, topics other than simulation are not covered by the survey. *Filtering* refers to applications where a sliding window over past observations is used to predict the next sample or some other quantity of interest. *Classification* refers to applications where a model takes a sequence of observations and produces a categorical label, for instance, indicating that the system is in an abnormal state. *Control* refers to applications where a NN-based controller is used to drive the system to a desired state. *Discovering Equations* refers to techniques based on ML that aim to discover the underlying equations of the system. *Quantity of Interest* refers to applications where an NN is used to provide a mapping from an initial condition to some quantity of interest, such as the steady-state of the system.

Application Domain. The broader topic of using ML in scientific fields has received widespread attention within several application domains [11, 12, 19, 108]. These review papers commonly focus on providing an overview of the prospective use cases of ML within their domains but put limited emphasis on how to apply the techniques in practice.

Surrogate Modeling. The field of surrogate modeling—that is, the theory and techniques used to produce faster models—is intimately related to the field of simulation with NNs. So it is important that we highlight some surveys in this field. The work of Koziel and Pietrenko-Dabrowska [61] presents a thorough introduction to data-driven surrogate modeling, which encompasses the use of NNs. Viana et al. [127] summarize advanced and yet simple statistical tools commonly used in the design automation community: (i) screening and variable reduction in both the input and the output spaces, (ii) simultaneous use of multiple surrogates, (iii) sequential sampling and optimization, and (iv) conservative estimators. Since optimization is an important use case of surrogate modeling, Forrester and Keane [31] reviewed advances in surrogate modeling in this field. Finally, with a focus on applications to water resources and building simulation, we highlight the work of Razavi et al. [105] and Westermann and Evins [135].

Prior Knowledge. One of the major trends to address some challenges arising in NN-based simulation is to encode prior knowledge such as physical constraints into the network itself or during the training process, ensuring the trained network is physically consistent. The work of Kelly et al. [54] coins this *theory-guided data science* and provides several examples of how knowledge

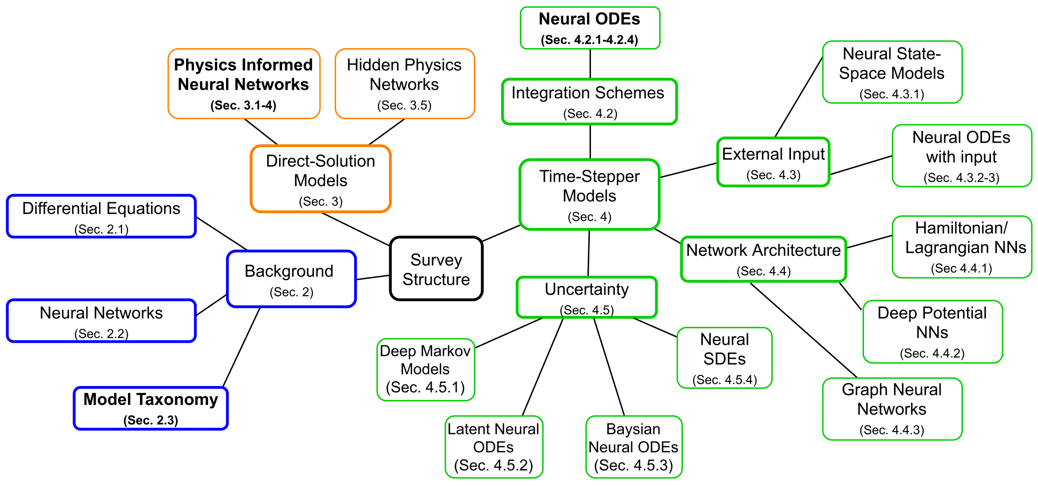

Neural ODEs (Sec. 4.2.1-4.2.4)
Physics Informed Neural Networks (Sec. 3.1-4)
Hidden Physics Networks (Sec. 3.5)
Integration Schemes (Sec. 4.2)
Neural State-Space Models (Sec. 4.3.1)
Direct-Solution Models (Sec. 3)
Time-Stepper Models (Sec. 4)
External Input (Sec. 4.3)
Neural ODEs with input (Sec. 4.3.2-3)
Differential Equations (Sec. 2.1)
Survey Structure
Background (Sec. 2)
Network Architecture (Sec. 4.4)
Hamiltonian/ Lagrangian NNs (Sec 4.4.1)
Neural Networks (Sec. 2.2)
Uncertainty (Sec. 4.5)
Deep Potential NNs (Sec. 4.4.2)
Deep Markov Models (Sec. 4.5.1)
Neural SDEs (Sec. 4.5.4)
Model Taxonomy (Sec. 2.3)
Graph Neural Networks (Sec. 4.4.3)
Latent Neural ODEs (Sec. 4.5.2)
Baysian Neural ODEs (Sec. 4.5.3)

Fig. 2. A mind map of the topics and model types covered in the survey.

may be incorporated in practice. Closely related to this is the work of Rai and Sahu [100] and von Rueden et al. [128, 129], which propose a detailed taxonomy describing the various paths through which knowledge can be incorporated into a NN model.

Comparison with This Survey. Our work complements the preceding surveys by providing an in-depth review focused specifically on NNs rather than ML as a whole. The concrete example helps the reader's understanding and highlights the similarities and inherent deficiencies of each approach.

We also outline the inherent challenges of simulation and establish a relationship between numerical simulation challenges and DL-based simulation challenges. The benefit of our approach is that the reader gets the intuition behind some approaches used to incorporate knowledge into the NNs. For instance, we relate energy-conserving numerical solvers to Hamiltonian NNs, whose goal is to encode energy conservation, and we discuss concepts such as numerical stability and solver convergence, which are crucial in long-term prediction using NNs.

1.2 Survey Structure

The remainder of the article is structured according to the mind map shown in Figure 2. First, Section 2 introduces the central concepts of dynamical systems, numerical solvers, and NNs. In addition, the section proposes a taxonomy describing the fundamental differences of how models can be constructed using NNs. The following two sections are dedicated to describing the two classes of models identified in the taxonomy: *direct-solution models* and *time-stepper models* in Sections 3 and 4, respectively. For each of the two categories, we describe the following:

- The structure of the model and the mechanism used to produce simulations of a system
- How the parameters are tuned to match the behavior of the true system
- Key challenges and extensions of the model designed to address them

Following this, Section 5 discusses the advantages and limitations of the two distinct model types and outlines future research directions. Finally, Section 6 provides a brief summary of the contributions of the article and the outlined research directions.

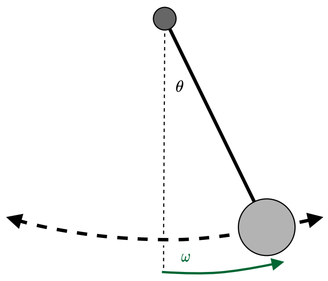


Fig. 3. The ideal pendulum system used as a case study throughout the article. The pendulum is characterized by an angle, θ, and an angular velocity, ω.

2 BACKGROUND

Models are an integral tool in natural sciences and engineering that allow us to deepen our understanding of nature or improve the design of engineered systems. One way to categorize models is by the *modeling* technique used to derive the model: *first principles* models are derived using fundamental physical laws, and *data-driven* models are created based on experimental data.

First, in Section 2.1, a running example is introduced, where we describe how differential equations can be used to model a simple mechanical system and how a solver is used to obtain a simulation. Then, Section 2.2 introduces the different ways NN-based models of the system can be constructed and trained. Finally, Section 2.3 introduces a taxonomy of the different ways NNs can be used to construct models of dynamical systems.

2.1 Differential Equations

An *ideal pendulum*, shown in Figure 3, refers to a mathematical model of a pendulum that, unlike its physical counterpart, neglects the influence of factors such as friction in the pivot or bending of the pendulum arm. The state of this system can be represented by two variables: its angle θ (expressed in radians) and its angular velocity ω. These variables correspond to a mathematical description of the system's state and are referred to as *state variables*. The way that a given point in the state-space evolves over time can be described using *differential equations*. Specifically, for the ideal pendulum, we may use the following **ordinary differential equation (ODE)**:

$$\frac{\partial^2\theta}{\partial t^2} + \frac{g}{l}\sin\theta = 0, \tag{1}$$

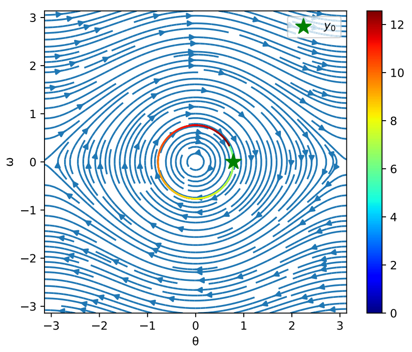
where g is the gravitational acceleration and l is the length of the pendulum arm. The ideal pendulum (Equation (1)) falls into the category of *autonomous* and *time-invariant* systems, since the system is not influenced by external stimulus and the dynamics do not change over time. Although this simplifies the notation and the way in which models can be constructed, it is not the general case. We discuss the implication of these issues in Section 4.3.1.

The equation can be rewritten as two first-order differential equations and expressed compactly using vector notation as follows:

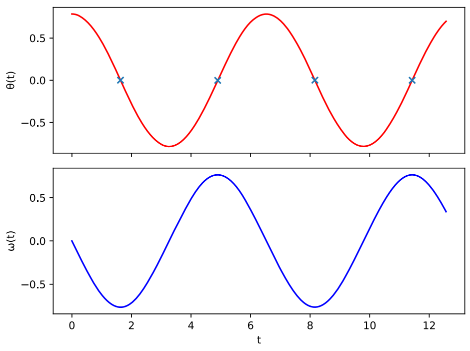
$$f(x) = \begin{bmatrix} \frac{\partial\omega}{\partial t} \\ \frac{\partial\theta}{\partial t} \end{bmatrix} = \begin{bmatrix} -\frac{g}{l}\sin\theta \\ \omega \end{bmatrix}. \tag{2}$$

where x is a vector of the system's state variables. In the context of this article, we refer to $f(x)$ as the *derivative function* or as the derivative of the system.

Although the differential equations describe how each state variable will evolve over the next time instance, they do not provide any way of determining the solution $x(t)$ on their own.



(a) Phase portrait of the ideal pendulum with a single trajectory drawn onto the phase space. The color denotes time.



(b) Solution of equation (3) for the initial condition marked with a star in figure 4(a).

Fig. 4. Diagram of the pendulum system and an example of the trajectory generated when solving the equation using a numerical solver.

Obtaining the solution of an ODE $f(x)$ given some *initial conditions* x_0 is referred to as an **initial value problem (IVP)** and can be formalized as follows:

$$\frac{\partial}{\partial t}x(t) = f(x(t)), \tag{3}$$

$$x(t_0) = x_0 \tag{4}$$

where $x(\cdot)$ is called the *solution*, $x : \mathbb{R} \to \mathbb{R}^n$, and $n \in \mathbb{N}$ is the dimension of the system's state space.

The result of solving the IVP corresponding to the pendulum can be seen in Figure 4(b), which shows how the two state variables θ and ω evolve from their initial state. An alternative view of this can be seen in the *phase portrait* in Figure 4(a).

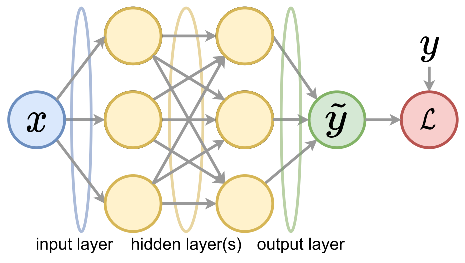
In many cases, it is impossible to find an exact analytical solution to the IVP, and instead numerical methods are used to approximate the solution. Numerical solvers are algorithms that approximate a continuous IVP, as the one in Equation (2), into a discrete-time dynamical system. These systems are often modeled with difference equations:

$$x_{i+1} = F(x_i), \tag{5}$$

where x_i represents the state vector at the i-th time point, x_{i+1} represents the next state vector, and $F : \mathbb{R}^n \to \mathbb{R}^n$ models the system behavior. Just as with ODEs, the initial state can be represented by a constraint on x_0, and the solution to Equation (5) with an initial value defined by such constraint is a function x_i defined for all $i \geq 0$. In Equation (5), time is implicitly defined as a discrete set.

We start by introducing the simplest and most intuitive numerical solver because it highlights the main challenges well. There are many numerical solvers, each presenting unique trade-offs. The reader is referred to the work of Cellier and Kofman [14] for an introduction to this topic, to the work of Hairer and Wanner [44] and Wanner and Hairer [133] for more detailed expositions on the numerical solution of ODEs and differential-algebraic system of equations (DAEs), to the work of LeVeque [69] for the numerical solution to **partial differential equations (PDEs)**, to the work of Marsden and West [78] for an overview of more advanced numerical schemes, and to the work of Kofman and Junco [60] for an introduction to quantized state solvers.

Given an IVP (Equation (3)) and a simulation step size $h > 0$, the **forward Euler (FE)** method computes a sequence in time of points $\tilde{x}_i$, where $\tilde{x}_i$ is the approximation of the solution to the



(a)

```
ỹ = 𝒩(x)
```

(b) Inference.

```
ỹ = 𝒩(x)
loss = ℒ(ỹ, y)
optimizer.step(loss)
```

(c) Training. This step is typically repeated many times for different inputs and desired output values.

Fig. 5. An FC NN is used to perform regression from an input x to y, where $\tilde{y}$ represents the approximation provided by the NN. Each layer of the network is characterized by a set of weights that are tuned during training to produce the desired output for a given input. During training, the loss function $\mathcal{L}$ is used to measure the divergence between the output produced by the network, $\tilde{y}$, and the desired output, y.

IVP at time hj: $\tilde{x}_i \approx x_i = x(hi)$. It starts from the given initial value $\tilde{x}_0 = x(0)$ and then computes iteratively:

$$\tilde{x}_{i+1} = \tilde{x}_i + hf(t_i, \tilde{x}_i), \tag{6}$$

where $f : \mathbb{R} \times \mathbb{R}^n \rightarrow \mathbb{R}^n$ is the ODE right-hand side in Equation (2) and $t_i = hi$.

A graphical representation of the solutions IVP starting from different initial conditions can be seen in Figure 4(a). For a specific point, the solver evaluates the derivative (depicted as curved arrows in the plot) and takes a small step in this direction. Applying this process iteratively results in the full trajectory, which for the pendulum corresponds to the circle in the phase space. The circle in the phase space implies that the solution is repeating itself—that is, corresponds to an oscillation in time as seen in Figure 4(b).

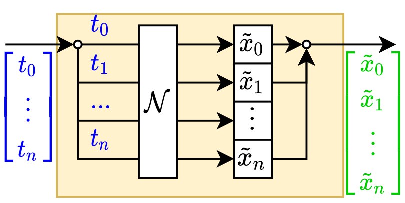
The ideal pendulum is an example of a well-studied dynamical system for which the dynamics can be described using simple ODEs that can be solved using standard solvers. Unfortunately, the simplicity of the idealized model comes at the cost of neglecting several factors that are present in a real pendulum. For example, the arm of the real pendulum may bend and energy may be lost in the pivot due to friction. The idealized model can be extended to account for these factors by incorporating models of friction and bending. However, this is time consuming, leads to a model that is harder to interpret, and does not guarantee that all factors are accounted for.

2.2 Neural Networks

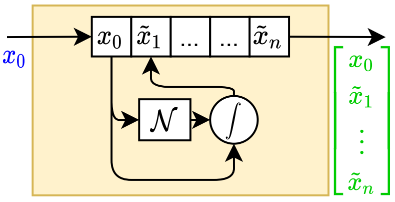
Today, the term *neural network* has come to encompass a whole family of models, which collectively have proven to be effective building blocks for solving a wide range of problems. In this article, we focus on a single class of networks, the **fully connected (FC)** NNs, due to their simplicity and the fact that they will be used to construct the models introduced in later sections. We refer the reader to the work of Goodfellow et al. [40] for a general introduction to the field of DL.

Like other data-driven models, NNs are generic structures that have no behavior specific to the problem they are being applied to before training. For this reason, it is essential to consider not only how the network produces its outputs but also how the network's parameters are tuned to solve the problem. For instance, we may consider using an FC NN to perform regression from a scalar input, x, to a scalar output, y, as shown in Figure 5(a).

We will refer to the process of producing predictions as *inference* and the process of tuning the network's weights to produce the desired results as *training*. There can be quite drastic differences



(a) Direct-solution model. An NN is used to parameterize a mapping from a time instance to the solution corresponding to that time instance.



(b) Time-stepper model. The network, $\mathcal{N}$, provides the derivative of the system at various points in state-space, which is then integrated by a numerical solver, here depicted as $\int$.

Fig. 6. Overview of two distinct model types. Direct-solution models are trained to produce a simulation without performing numerical integration explicitly. Conversely, time-stepper models use the same techniques known from numerical simulation to produce a simulation of the system.

in the complexity of the two phases, the training phase typically being the most complex and computationally intensive. During training, a loss function defines a mapping from the predicted quantity to a scalar that is a measure of how close the prediction is to the true trajectory. Differentiating the loss function with respect to the parameters of the NN allows us to update them in a way such that the loss is minimized.

Batching and Notation. During the training of an NN, we often wish to perform the forward pass individually for multiple inputs grouped in a batch. By convention, many DL frameworks treat any leading dimensions as being batches of samples. We adopt this convention as well to simplify notation. Thus, $\mathcal{N}(x) \in \mathbb{R}^{*\times n}$ when $\mathcal{N} : \mathbb{R}^n \to \mathbb{R}^n$ and $x \in \mathbb{R}^{*\times n}$ ('$*$' indicating any number of leading dimensions).

2.3 Model Taxonomy

A challenge of studying any fast-evolving research field such as DL is that the terminology used to describe important concepts and ideas may not always have converged. This is especially true in the intersection between DL, numerical simulation, and physics, due to the influx of ideas and terminology from the different fields. In the literature, there is also a tendency to focus on the success of a particular technique in a specific application, with little emphasis on explaining the inner workings and limitations of the technique. A consequence of this is that important contributions to the field become lost due to the papers being hard to digest.

In an attempt to alleviate this, we propose a simple taxonomy describing how models can be constructed consisting of two categories: *direct-solution models* and *time-stepper models*, as shown in Figure 6. Direct-solution models, described in Section 3, do not employ integration but rather produce an estimate of the state at a particular time by feeding in the time as an input to the network. Time-stepper models, found in Section 4, can be characterized by using a similar approach to numerical solvers, where the current state is used to calculate the state at some time into the future. The difference between the time-stepper and continuous models has significant implications for how the model deals with varying initial conditions and inputs. Per design, the time-stepper models handle different initial conditions and inputs, whereas direct-solution models have to be retrained. In other words, the time-stepper models learn the dynamics while the direct-solution models learn a solution to an IVP for a given initial state and set of inputs.

Table 1. Comparison of Direct-Solution Models

Name	In_{NN}	Out_{NN}	Out_{AD}	Uses Equations
Vanilla Direct-Solution	t	θ, ω		
Automatic Differentiation Direct-Solution	t	θ	ω	
Physics-Informed Neural Network	t	θ	$\omega, \partial\omega$	✓
Hidden Physics Neural Network	t	θ, l	$\omega, \partial\omega$	✓

3 DIRECT-SOLUTION MODELS

One approach to obtaining the trace of a system is to construct a model that maps a set of time instances $t \in \mathbb{R}^m$ to the solution $\tilde{x} \in \mathbb{R}^{m \times n}$. We refer to this type of model as a *direct-solution* model.

To construct the model, an NN is trained to provide an exact solution for a set of *collocation* points that are sampled from the true system. Another way to view this is that the NN acts as a trainable interpolation engine, which allows the solution to be evaluated at arbitrary points in time, not only those of the collocation points. An important limitation of this approach is that a trained model is fixed for a specific set of initial conditions. To evaluate the solution for different initial conditions, a new model would have to be trained on new data.

In the literature, this type of model is often applied to learn the dynamics of systems governed by PDEs and less frequently for systems governed by ODEs. Several factors are likely to influence this pattern of use. First, PDEs are generally harder and more computationally expensive to solve than ODEs, which provides a stronger motivation for applying NNs as a means to obtain a solution. Second, many practical uses of ODEs require that they can easily be evaluated for different initial conditions, which is not the case for direct-solution models.

Although the motivation for applying direct-solution networks may be strongest for PDEs, they can also be applied to model ODEs. The main difference is that a network to model an ODE takes time as the only input, whereas the network used to model a PDE would take both time and spatial coordinates.

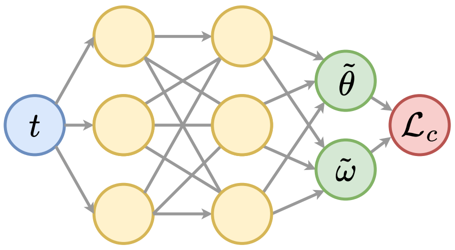
A key challenge in training direct-solution NNs is the amount of data required to reach an acceptable level of accuracy and generalization. A vanilla approach that does not leverage prior knowledge, like the one described in Section 3.2, is likely to fit the collocation points very well but fails to reproduce the underlying trend. A recent trend popularized by **physics-informed neural networks (PINNs)** [101] is to apply **automatic differentiation (AD)** and to use equations encoding prior knowledge to improve the generalization of the model.

The remaining part of this section describes how the different types of direct-solution models, shown in Table 1, can be applied to simulate the ideal pendulum system for a specific initial condition. First, the architecture of the NNs used for the experiments is introduced in Section 3.1. Next, the simplest approach is introduced in Section 3.2, before progressively building up to a model type that incorporates features from all prior models in Section 3.5.

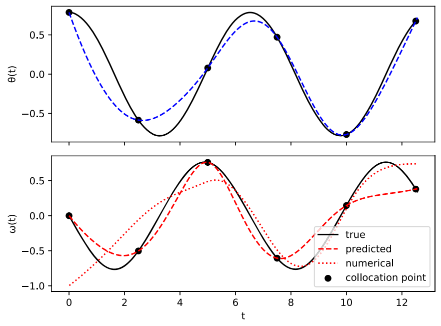
3.1 Methodology

The examples of direct-solution models shown in this section use an FC NN with three hidden layers consisting of 32 neurons each. The output of each hidden layer is followed by a softplus activation function.

Each model is trained on a trajectory corresponding to the simulation for a single initial condition, which is sampled to obtain a set of collocation points as shown in Figure 7(b). The goal is to obtain a model that can predict the solution at any point in time, not only those coinciding with the collocation points.



(a) Network structure.



(b) Predictions.

```
θ̃,ω̃ = 𝒩(t)
loss = ℒc((θ̃, ω̃), (θ, ω))
optimizer.step(loss)
```

(c) Training.

```
θ̃,ω̃ = 𝒩(t)
```

(d) Inference.

Fig. 7. Vanilla direct-solution model. The network $\mathcal{N} : \mathbb{R} \rightarrow \mathbb{R}^2$ maps time instances t to the solution $\tilde{\theta}, \tilde{\omega}$. Black dots indicate the collocation points (i.e., the points in which the loss function is minimized). The network fits all collocation points well but fails to generalize in the interval between points. In addition, $\tilde{\omega}$ is very different from the approximation obtained using numerical differentiation of $\tilde{\theta}$.

3.2 Vanilla Direct-Solution

Direct-solution models produce an estimate of the system's state at a given time, t_i, by introducing it as an input to an NN.

To model the pendulumm we use a feed-forward network with a single input t and two outputs $\tilde{\theta}$ and $\tilde{\omega}$, as depicted in Figure 7(a). To obtain the solution for multiple time instances, the network can simply be evaluated multiple times. There are no dependencies between the estimates of multiple states, allowing us to evaluate all of these in parallel.

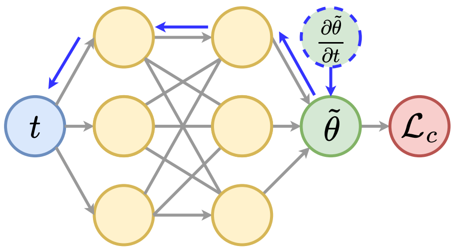
The network is trained by minimizing the difference between the predicted and the true trajectory in the collocation points shown in Figure 7(b) using a distance metric such as MSE defined by Equation (7):

$$\mathcal{L}_c(\tilde{x}, x) = \frac{1}{mn} \sum_{i=0}^{m-1} \sum_{j=0}^{n-1} (\tilde{x}_{ij} - x_{ij})^2, \tag{7}$$

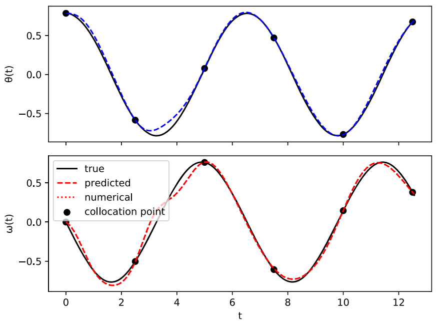
where m is the length of the trajectory, n is the dimension of the system's state-space, and x_{ij} denotes the value of the j-th state at the i-th point of time of the trajectory.

It is important to emphasize that the models learn a sequence of system states characterized by a specific set of initial conditions—that is, the initial conditions are encoded into the trainable parameters of the network during training and cannot be modified during inference.

Direct-solution models are sensitive to the quality of training data. NNs are used to find mappings between sparse sets of input data and the output. Even a simple example in the data-sampling strategy can influence their generalization performance. Consider the trajectory in Figure 7(b); the model has only learned the correct solution in the collocation points and fails to generalize anywhere else.



(a) Network structure.



(b) Predictions

```
θ̃ = 𝒩(t)
ω̃ = gradient(θ̃, t)
loss = ℒ_c((θ̃, ω̃), (θ, ω))
optimizer.step(loss)
```

(c) Training.

```
θ̃ = 𝒩(t)
ω̃ = gradient(θ̃, t)
```

(d) Inference.

Fig. 8. AD in the direct-solution model. The network $\mathcal{N} : \mathbb{R} \to \mathbb{R}$ maps the time instances t to the pendulum's angle $\tilde{\theta}$. The angular velocity $\tilde{\omega}$ is obtained by differentiating $\tilde{\theta}$ with respect to time using AD. This approach ensures that an output, representing the derivative of another output, acts like a true derivative. As a result, the network generalizes significantly better across both state variables.

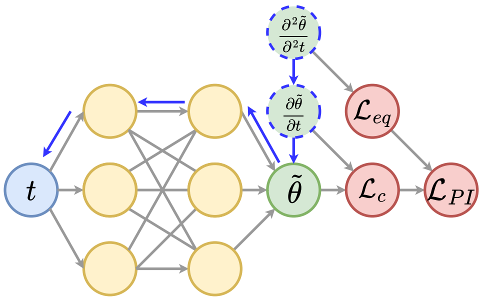
It is worth noting that there are many ways that this can go wrong—that is, given a sufficiently sparse sampling, it is not just one specific choice of training points that makes it impossible for the network to learn the true mapping. The obvious way to mitigate the issue is to obtain more data by sampling at a higher rate. However, there are cases where data acquisition is expensive, impractical, or it is simply impossible to change the sampling frequency.

Consider a system where one state variable is the derivative of the other, a setting that is quite common in systems that can be described by differential equations. A vanilla direct-solution model cannot guarantee that the relationship between the predicted state variables respects this property. Figure 7(b) provides a graphical representation of the issue. Although the model predicts both system state variables correctly in the collocation points, it can clearly be seen that the estimate for ω is neither the derivative of $\tilde{\theta}$ nor does it come close to the true trajectory.

3.3 AD Direct-Solution

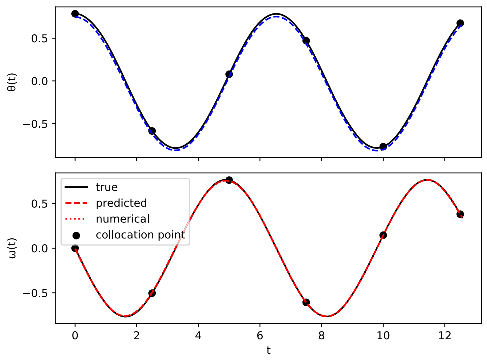
One way to leverage known relations is to calculate derivatives of state variables using automatic differentiation instead of having the network predict them as explicit outputs. In the case of the pendulum, this means using the network to predict $\tilde{\theta}$ and then obtaining $\tilde{\omega}$ by calculating the first-order derivative of $\tilde{\theta}$ with respect to time, as described in Figure 8(c) and (d). Figure 8(b) shows how much closer the predicted trajectories are to the true ones when using this approach.

A drawback of obtaining ω using AD is an increased computation cost and memory consumption depending on which mode of AD is used. Using reverse mode AD (backpropagation) as depicted in Figure 8(a) requires another pass of the computation graph, as indicated by the arrow going from output θ to input t. For training, this is not problematic since the computations carried out during backpropagation are necessary to update the weights of the network as well. However, using backpropagation during inference is not ideal because it introduces unnecessary memory and computation cost. An alternative is to use forward AD where the derivatives are computed



$\frac{\partial^2\tilde{\theta}}{\partial^2 t}$
$\frac{\partial\tilde{\theta}}{\partial t}$
t
$\tilde{\theta}$
$\mathcal{L}_{eq}$
$\mathcal{L}_c$
$\mathcal{L}_{PI}$

(a) Network structure.



(b) Predictions.

```
θ̃ = 𝒩(t)
ω̃ = gradient(θ̃, t)
∂ω̃ = gradient(ω̃, t)
loss = ℒc((θ̃,ω̃),(θ,ω))+ℒeq(θ̃,∂ω̃)
optimizer.step(loss)
```

(c) Training.

```
θ̃ = 𝒩(t)
ω̃ = gradient(θ̃, t)
```

(d) Inference.

Fig. 9. Physics-informed neural network. The network $\mathcal{N} : \mathbb{R} \rightarrow \mathbb{R}$ maps the time instances t to the pendulum's angle $\tilde{\theta}$. The angular velocity $\tilde{\omega}$ and its derivative are obtained using AD. The network is trained by minimizing Equation (8).

during the forward pass, thus dispensing of the separate backward pass. Unfortunately, not all DL frameworks provide support for forward mode AD (see Table 5 in the work of Baydin et al. [5]). A likely explanation is that the typical task of evaluating the derivative of the loss with respect to the network's weights is more efficient using reverse-mode AD (backpropagation).

3.4 Physics-Informed Neural Networks

In modeling scenarios where the equations describing the dynamics of the system are known, we can use them to train the model as another way of addressing the data-sampling issue. In what is known as *physics-informed neural networks* [101], knowledge about the physical laws governing the system is used to impose structure on the NN model. This can be accomplished by extending the loss function with an *equation loss* term that ensures the solution obeys the dynamics described by the governing equations. Although this technique was originally proposed for solving PDEs, it can also be applied to solve ODEs. For instance, to model the ideal pendulum using a PINN, we could integrate the expression of $\frac{\partial\tilde{\omega}}{\partial t}$ from Equation (1) to formulate the loss as

$$\mathcal{L}_{PI}\left(\tilde{\theta}, \frac{\partial\tilde{\theta}}{\partial t}, \frac{\partial\tilde{\omega}}{\partial t}\right) = \mathcal{L}_c\left(\tilde{\theta}, \frac{\partial\tilde{\theta}}{\partial t}\right) + \mathcal{L}_{eq}\left(\tilde{\theta}, \frac{\partial\tilde{\omega}}{\partial t}\right) \tag{8}$$

$$\mathcal{L}_{eq}\left(\tilde{\theta}, \frac{\partial\tilde{\omega}}{\partial t}\right) = \frac{1}{m}\sum_{i=0}^{m-1}\left(\frac{\partial\tilde{\omega}_i}{\partial t} - \frac{g}{l}\sin\tilde{\theta}_i\right)^2.$$

Again, we can use automatic differentiation to obtain $\frac{\partial\tilde{\omega}}{\partial t}$ by differentiating $\tilde{\theta}$ twice, depicted in the computation graph shown in Figure 9(a). As shown in Figure 9(c), this requires only a few lines of code when using AD.

A motivation for incorporating the equation loss term is to constrain the search space of the optimizer to parameters that yield physically consistent solutions. It should be noted that both the loss term penalizing the prediction error and the equation error are necessary to constrain the predictions of the network. On its own, the equation error guarantees that the predicted state satisfies the ODE, but not necessarily that it is the solution at a particular time. Introducing the prediction error ensures that the predictions are not only valid but also the correct solutions for the particular points used to calculate the prediction error. In addition, it should be noted that the collocation and equation loss terms may be evaluated for a different set of times. For instance, the equation-based loss term may be evaluated for an arbitrary number of time instances, since the term does rely on accessing the true solution for particular time instances.

In addition to proposing the introduction of the equation loss, PINNs also apply the idea of using backpropagation to calculate the derivatives of the state variables rather than adding them as outputs to the network, as depicted in Figure 9(a). Being able to obtain the n-th order derivatives is very useful for PINNs, as they often appear in differential equations on which the equation loss is based. For the ideal pendulum, this technique can be used to obtain $\frac{\partial^2\theta}{\partial t^2}(t)$ from a single output of the network θ, which can then be plugged into Equation (2) to check that the prediction is consistent. A benefit of using backpropagation compared to adding state variables as outputs of the network is that this structurally ensures that the derivatives are in fact partial derivatives of the state variables.

Training PINNs using gradient descent requires careful tuning of the learning rate. Specifically, it has been observed that the boundary conditions and the physics regularization terms may converge at different rates. In some cases, this manifests itself as a large misfit specifically at the boundary points. Wang et al. [131, 132] propose a strategy for weighing the different terms of the loss function to ensure consistent minimization across all terms.

3.5 Hidden Physics Networks

Hidden physics neural networks (HNNs) [103] can be seen as an extension of PINNs that use governing equations to extract features of the data that are not present in the original training data. We refer to the unobserved variable of interest as a *hidden variable*. This technique is useful in cases where the hidden variable is difficult to measure compared to the known variables or simply impossible to measure since no sensor exists that can reliably measure it.

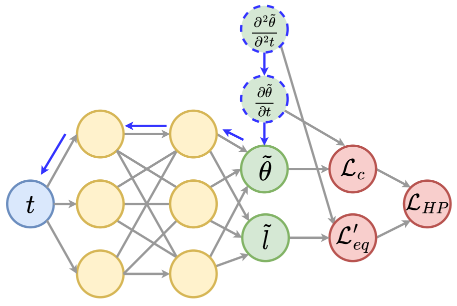
For the sake of demonstration, we may suppose that the length of the pendulum arm is unknown and that it varies with time, as shown in Figure 10(b). For the training, this is problematic since l is required to calculate the equation loss. A solution to this is to add an output $\tilde{l}$ to the network that serves as an approximation of the true length l, as depicted in Figure 10(d). We modify Equation (8) to define a new loss function that takes the estimate of $\tilde{l}$ into account

$$\mathcal{L}_{HP}\left(\tilde{\theta}, \frac{\partial\tilde{\theta}}{\partial t}, \frac{\partial\tilde{\omega}}{\partial t}, \tilde{l}\right) = \mathcal{L}_c\left(\tilde{\theta}, \frac{\partial\tilde{\theta}}{\partial t}\right) + \mathcal{L}'_{eq}\left(\tilde{\theta}, \frac{\partial\tilde{\omega}}{\partial t}, \tilde{l}\right) \tag{9}$$

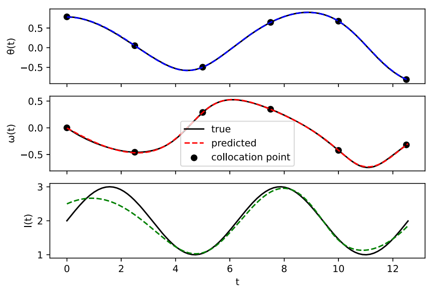
$$\mathcal{L}'_{eq}\left(\tilde{\theta}, \frac{\partial\tilde{\omega}}{\partial t}, \tilde{l}\right) = \frac{1}{m}\sum_{i=0}^{m-1}\left(\frac{\partial\tilde{\omega}_i}{\partial t} - \frac{g}{\tilde{l}_i}\sin\tilde{\theta}_i\right)^2.$$

It should be emphasized that $\tilde{l}$ is not part of the collocation loss term, since the true value l is not known. It is only as a result of the equation loss that the network is constrained to produce estimates of l satisfies the system's dynamics.

Raissi et al. [103] use this technique to extract pressure and velocity fields based on measured dye concentrations. In this particular case, the dye concentration can be measured by a camera, since


$\frac{\partial^2\tilde{\theta}}{\partial^2 t}$
$\frac{\partial\tilde{\theta}}{\partial t}$
t
$\tilde{\theta}$
$\tilde{l}$
$\mathcal{L}_c$
$\mathcal{L}'_{eq}$
$\mathcal{L}_{HP}$

(a) Network structure.



(b) Predictions.

```
θ̃, l̃ = 𝒩(t)
ω̃ = gradient(θ̃, t)
∂ω̃ = gradient(ω̃, t)
loss = ℒc((θ̃,ω̃),(θ,ω))+ℒ'eq(θ̃, ∂ω̃, l̃)
optimizer.step(loss)
```

(c) Training.

```
θ̃, l̃ = 𝒩(t)
ω̃ = gradient(θ̃, t)
```

(d) Inference.

Fig. 10. Hidden physics network. This network $\mathcal{N} : \mathbb{R} \rightarrow \mathbb{R}^2$ is an extension of the PINN and maps the time instances t to the pendulum's angle $\tilde{\theta}$ and the length of the pendulum $\tilde{l}$ that is set to vary in time for the sake of demonstration. Note that $\tilde{l}$ is not part of $\mathcal{L}_c$ since there is no training data for it; instead, it is part of the equation loss $\mathcal{L}'_{eq}$.

the opacity of the fluid is proportional to the dye concentration. They show that this technique also works well even in cases where the dye concentration is sampled at only a few points in time and in space. Like PINNs, HNNs are easily applied to PDEs, but at the cost of the initial conditions being encoded in the network during training.

The difference between PINNs and HNNs is very subtle; both utilize similar network architectures and use loss functions that penalize any incorrect prediction violations of governing equations. A distinguishing factor is that, in HNNs, the hidden variable is inferred based on physical laws that relate the hidden variable to the observed variables. Since the hidden variables are not part of the training data, they can only be enforced through equations.

4 TIME-STEPPER MODELS

Consider the approach used to model an ideal pendulum, described in Section 2. First, a set of differential equations, Equation (2), was used to model the derivative function of the system. Next, using the function, a numerical solver was used to obtain a simulation of the system for a particular initial condition. The challenge of this approach is that identifying the derivative function analytically is difficult for complex systems.

An alternative approach is to train an NN to approximate the derivative function of the system, allowing the network to be used in place of the hand-derived function, as depicted in Figure 11. We refer to this type of model as a *time-stepper model* since it produces a simulation by taking multiple steps in time, like a numerical solver. An advantage of this is that it allows well-studied numerical solvers to be integrated into a model with relative ease.

The main differences between two given models can be attributed to (i) how the derivatives are produced by the network and (ii) what sort of integration scheme is applied. For instance, the difference between the *direct* (Section 4.2.1) and *Euler* time-stepper models (Section 4.2.2) is that the former does not employ any integration scheme, whereas the latter is similar to the FE

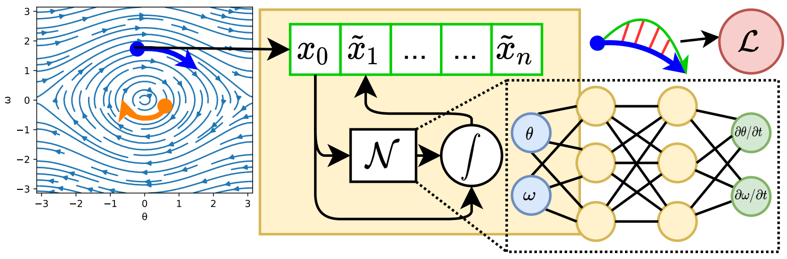


Fig. 11. Time-stepper model. Starting from a given initial condition x_0, the next state of the system, $\tilde{x}_{i+1}$, is obtained by feeding the current state $\tilde{x}_i$ into the derivative network $\mathcal{N}$, producing a derivative that is integrated using an integration scheme $\int$. The loss $\mathcal{L}$ is evaluated by comparing the predicted with the training trajectory. The process can be repeated for multiple trajectories to improve the generalization of the derivative network.

(recall Equation (6)), leading to a significant difference in predictive ability. Other networks, such as the *Lagrangian* time-stepper, Section 4.4.1, distinguish themselves by the way the NN produces the derivatives. Specifically, this approach does not obtain $\partial\theta$ and $\partial\omega$ as outputs from a network but instead uses AD in an approach similar to Section 3.3. Similar to how an ODE can be solved with different numerical solvers, the Lagrangian time-stepper could be modified to use a different integration scheme than FE.

Given the independent relationship between the choice of NN and the numerical solver used, the models introduced in the sections should not be viewed as an exhaustive list of combinations. Rather, the aim is to describe and compare the models commonly encountered in the literature.

4.1 Methodology

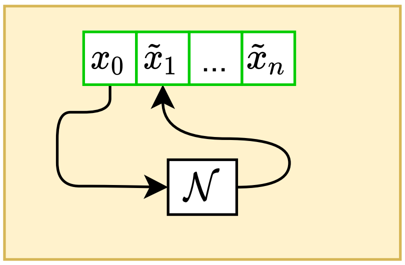
A time-stepper must be able to produce accurate simulations for different initial conditions. It would be possible to train a time-stepper using a single trajectory; however, this is unlikely to generalize well to different initial conditions. Another approach is to use multiple, potentially shorter trajectories as training data. We can extend Equation (7) to take the mean error over l trajectories:

$$\mathcal{L}_{ts}(\tilde{X}, X) = \frac{1}{l}\sum_{i=0}^{l-1}\mathcal{L}_c(\tilde{X}_i, X_i), \tag{10}$$

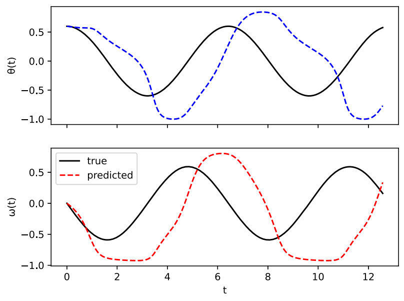
where $X \in R^{l\times m\times n}$ are the training trajectories and $\tilde{X} \in R^{l\times m\times n}$ are the predicted trajectories.

Each time-stepper model is trained on 100 trajectories, each consisting of two samples: the initial state and the state one step into the future. The initial states are sampled in the interval $\theta : (-1, 1)$ and $\omega : (-1, 1)$ using Latin hyper-cube sampling (see Figure 11). Each model uses an FC network consisting of eight hidden layers with 32 neurons each. Each layer of the network applies a softplus activation function. The number of inputs and outputs is determined by the number of states characterizing the system, which is 2 for the ideal pendulum. Exceptions to this are networks such as the Lagrangian network described in Section 4.4.1, for which the derivatives are obtained using AD rather than as outputs of a network.

To validate the performance of each model, 100 new initial conditions are sampled in a grid. For each initial condition in the validation set, the system is simulated for 4π seconds using the original ODE and compared with the corresponding prediction made by the trained model. For simplicity, we show only the trajectory corresponding to a single initial condition, like the one in Figure 12(b).



(a) Network structure.



(b) Predictions.

```
X̃[:,0,:] = $X_{t_0}$
for i in 0...m-1
  X̃[:,i+1,:] = $\mathcal{N}$(X̃[:,i,:])
loss = $\mathcal{L}_{ts}$(X,X̃)
optimizer.step(loss)
```

(c) Training.

```
x̃[0,:] = $x_{t_0}$
for i in 0...m-1
  x̃[i+1,:] = $\mathcal{N}$(x̃[i,:])
```

(d) Inference.

Fig. 12. Direct time-stepper. The output of the network $\mathcal{N} : \mathbb{R}^2 \to \mathbb{R}^2$ is used as the prediction for the next step without any form of numerical integration. An issue of this type of model is that it fails to generalize beyond the exact points in state space that it has been trained for. Over several steps, the error compounds, which leads to an inaccurate simulation.

4.2 Integration Schemes

An important characteristic of a time-stepper model is how the derivatives are evaluated and integrated to obtain a simulation of the system. Again, it should be emphasized that the choice of the numerical solver is independent of the architecture of the NN used to approximate the derivative function. In other words, for a given choice of NN architecture, the performance of the trained model may depend on the choice of solver.

The choice of numerical solver not only determines how the model produces a simulation of the system but also influences how the model must be trained. Specifically, when minimizing any criterion that is a function of the integrated state, the choice of solver determines how the state is produced.

In the following section, we demonstrate how various numerical solvers can be used and evaluate their impacts on the performance of the models.

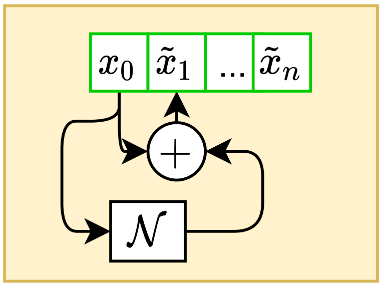
4.2.1 Direct Time-Stepper. The simplest approach to obtaining the next state is to use the prediction produced by the network directly, as summarized in Figure 12(a):

$$\tilde{x}_{i+1} = \mathcal{N}(\tilde{x}_i),$$

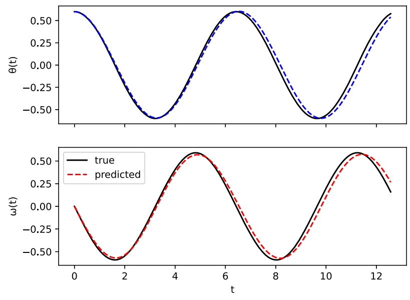
where N represents a generic NN with arbitrary architecture and $\tilde{x}_0 = x_0$.

The network is trained to produce an estimate of the next state, $\tilde{x}_{i+1}$, from the current state, x_i. During training, this operation can be vectorized such that every state at every timestamp, omitting the last, is mapped one step into the future using a single invocation of the network, as shown in Figure 12(c). The reason for leaving out the last sample in when invoking the NN is that this would produce a prediction, x_{N+1}, for which there does not exist a sample in the training set.

At inference time, only the initial state x_0 is known. The full trace of the system is obtained by repeatedly introducing the current state into the network, as depicted in Figure 12(d). Note that the inference phase cannot be parallelized in time, since predictions for time $i + 1$ depend on



(a) Network structure.



(b) Predictions.

```
X̃[:,0,:] = $X_0$
for i in 0...m-1
  $\Delta X$ = $\mathcal{N}$(X̃[:,i,:])
  X̃[:,i+1,:] = X̃[:,i,:] + $\Delta X$
loss = $\mathcal{L}_{ts}$(X,X̃)
optimizer.step(loss)
```

(c) Training.

```
x̃[0,:] = $x_0$
for i in 0...m-1
  $\Delta x$ = $\mathcal{N}$(x̃[i,:])
  x̃[i+1,:] = x̃[i,:] + $\Delta x$
```

(d) Inference.

Fig. 13. Residual time-stepper. The output of the network is added to the current state to form a prediction of the next state. Compared to the direct time-stepper, this method produces simulations that are much closer to the true system.

predictions for time k. However, it is possible to simulate the system for multiple initial states in parallel, as they are independent of each other.

The simulation for a single initial condition can be seen in Figure 12(b). Although the simulation is accurate for the first few steps, it quickly diverges from the true dynamics.

4.2.2 Residual Time-Stepper. A network can be trained to predict a derivative-like quantity that can then be added to the current state to yield the next as shown in Figure 13(a):

$$\tilde{x}_{i+1} = \tilde{x}_i + \mathcal{N}(\tilde{x}_i).$$

DL practitioners may recognize this as a residual block that forms the basis for *residual networks* (ResNets) [45], which are used with great success in applications spanning from image classification to natural language processing. Readers familiar with numerical simulation will likely notice that the previous equation closely resembles the accumulated term in the FE integrator (recall Equation (6)) but without the term that accounts for the step size. If the data is sampled at equidistant timesteps, the network scales the derivative to adapt the step size.

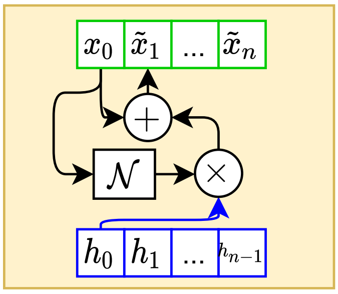
The central motivation for using a residual network is that it may be easier to train a network to predict how the system will change rather than a direct mapping between the current and next state.

4.2.3 Euler Time-Stepper. Alternatively, the step size can be encoded in the model by scaling the contribution of the derivative by the step size h_i as shown in Figure 14(a):

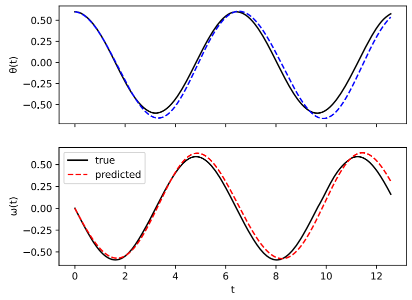
$$\tilde{x}_{i+1} = \tilde{x}_i + h_i * \mathcal{N}(\tilde{x}_i). \quad (11)$$

This resemblance has been noted several times [97] and has resulted in work that interprets residual networks as ODEs allowing classical stability analysis to be used [15, 110, 111].

The FE integrator shown in Equation (11) is simple to implement. However, it accumulates a higher error than more advanced methods, such as the Midpoint, for a given step size. This issue



(a) Network structure.



(b) Predictions.

```
X̃[:,0,:] = $X_0$
for i in 0...m-1
  $\Delta X$ = $\mathcal{N}$(X̃[:,i,:])
  X̃[:,i+1,:] = X̃[:,i,:] + $h$[i]*$\Delta X$
loss = $\mathcal{L}_{ts}$($X$,X̃)
optimizer.step(loss)
```

(c) Training.

```
x̃[0,:] = $x_0$
for i in 0...m-1
  $\Delta x$ = $\mathcal{N}$(x̃[i,:])
  x̃[i+1,:] = x̃[i,:] + $h$[i]*$\Delta x$
```

(d) Inference.

Fig. 14. Euler time-stepper. The output of the network is multiplied by the step size and is added to the current state to form a prediction for the next state. In this case, accounting for the step size leads to minimal improvements, if any, compared to the residual time-stepper. This is likely due to the fact that the step size used during training is the same as the one used to plot the trajectory in Figure 14(b).

has motivated the integration of more sophisticated numerical solvers in time-stepper models. For example, **linear multistep (LMS)** methods are used in the work of Raissi et al. [102]. LMS uses several past states and their derivatives to predict the next state, resulting in a smaller error compared to FE. Like FE, LMS only requires a single function evaluation per step, making it a very efficient method. But if the system is not continuous, this method needs to be reinitialized after a discontinuity occurs [36].

4.2.4 Neural Ordinary Differential Equations. **Neural ordinary differential equations (NODEs)** [18] is a method used to construct models by combining a numerical solver with an NN that approximates the derivative of the system. Unlike the previously introduced models, the term *NODEs* is not used to refer to models using a specific integration scheme but rather to the idea of treating an ML problem as a dynamical system that can be solved using a numerical solver. Part of their contribution was the implementation of differentiable solvers accessible via a simple API, allowing users to implement NODEs in only a few lines of code, as shown in Figure 16.

Some confusion may arise from the fact that NODEs are frequently used for image classification throughout the literature, which may seem completely unrelated to numerical simulations. The underlying idea is that an image can be represented as a point in state-space that moves on a trajectory defined by an ODE, as shown in Figure 15. The goal of this is to find an ODE that results in images of the same class converging to a cluster that is easily separable from that of unrelated classes. For single inference (e.g., in image classification), intermediate predictions have no inherent meaning—that is, they typically do not correspond to any measurable quantity of the system; we are only interested in the final estimate $\hat{x}_m$. Due to the lack of training samples corresponding to intermediate steps, it is impossible to minimize the single-step error.

Chen et al. [18] motivate the use of an adaptive step-size solver by its ability to adjust the step size to match the desired balance between numerical error and performance. An alternative way

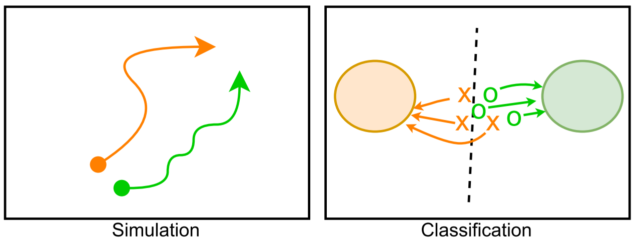
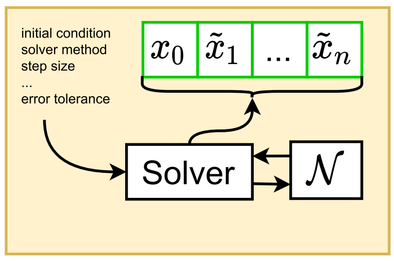
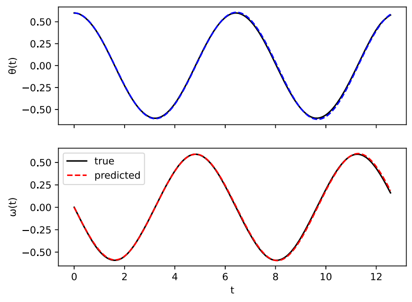


Fig. 15. Different applications of NODEs. NODEs can be used to simulate a dynamical system with the goal of obtaining a trajectory corresponding to an initial condition. In this case, the goal is to train the network to produce a derivative that provides a good estimate of the true state at every step of the trajectory. Another use is for classification by treating each input sample as a point in state-space, which evolves according to the derivative produced by the network. In this case, the goal is to train the network to learn dynamics that leads to samples belonging to each class ending in distinct clusters that are easily separable.



(a) Network Structure.



(b) Predictions.

```
X̃ = odeint(N, X₀, t_start, t_end, "rk4")
loss = L_ts(X, X̃)
optimizer.step(loss)
```

(c) Training.

```
x̃ = odeint(N, x₀, t_start, t_end, "rk4")
```

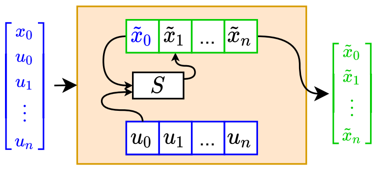
(d) Inference.

Fig. 16. Neural ordinary differential equations. NODEs generally refer to models that are constructed to use a numerical solver to integrate the derivatives through time. Unlike the previously introduced integration schemes that mapped to concrete architectures, NODEs refer to the idea of using well-established numerical solvers inside a model. Part of NODEs' popularity is due to the fact that it mimics the programming APIs of traditional numerical solvers, which makes it easy to switch between different types of solvers.

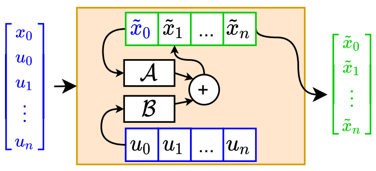
to view NODEs is as a *continuous-depth model* where the number of layers is a result of the step size chosen by the solver.

From this perspective, the stability of NODEs is closely related to the stability of integration schemes of classical ODEs. To address the convergence issues during training, some authors propose NODEs with stability guarantees by exploiting Lyapunov stability theory [79] and spectral projections [99]. Another standing issue of NODEs is their large computational overhead during training compared to classical NNs. Finlay et al. [28] demonstrated that stability regularization may improve convergence and reduce the training times of NODEs. Poli et al. [96] propose graph NODEs resulting in training speedups, as well as improved performance due to incorporation of prior knowledge.

To improve the performance, others have introduced various inductive biases such as Hamiltonian NODE architecture [142] or penalizing higher-order derivatives of the NODEs in the



(a) State and input stacked and fed into the same network.



(b) State and input fed into separate networks.

Fig. 17. Incorporation of inputs in the time-stepping model.

loss function [55]. To account for the noise and uncertainties, some authors proposed stochastic NODEs [42, 48, 70, 74] as generalizations of deterministic NODEs.

A fundamental issue of interpreting trained NODEs as proper ODEs is that they may have trajectory crossings, and their performance can be sensitive to the step size used during inference [92]. Contrary to this, the solutions of ODEs with unique solutions would never have intersecting trajectories, as this would imply that for a given state (the point of intersection), the system could evolve in two different ways. Some authors have noted that there seems to be a critical step size for which the trained network starts behaving like a proper ODE [92]. In other words, if trained with a particular step size, the network will perform equally well or better if used with a smaller step size during inference. Another approach is to use regularization to constrain the parameters of the network to ensure that solutions are unique. For ResNets, this can be achieved by ensuring that the Lipschitz constant of the network is less than 1 for any point in the state-space, which guarantees a unique solution [7].

To deal with external inputs in NODEs, Dupont et al. [27] and Norcliffe et al. [88] propose lifting the state-space via additional augmented variables. A more general way of explicitly modeling the input dynamics via additional NNs is proposed by Massaroli et al. [80].

4.3 External Input

So far, we have only considered how to apply time-stepper models to systems where the derivative function is determined exclusively by the system's state. In practice, many systems encountered are influenced by an external stimulus that is independent of the dynamics, such as external forces acting on the system or actuation signals of a controller. To avoid confusion, we refer to these external influences as *external input* to distinguish them from the general concept of an NN's inputs.

The structure of a time-stepper model lends itself well to introducing external inputs at every evaluation of the derivative function. As a result, it is possible to integrate external inputs in time-stepper models in many ways.

4.3.1 Neural State-Space Models. Inputs can be added to the time-stepper models in a couple of ways. One way is to concatenate the inputs with the states, as illustrated in Figure 17(a):

$$\tilde{x}_{i+1} = \mathcal{N}([x_i, u_i]), \tag{12}$$

where $\tilde{x}_i$ and u_i represent states and inputs at time t_i, respectively. The evolution of the future state x_{i+1} is fully determined by the derivative network $\mathcal{N}$. A possible rationale for lumping system states and inputs are parameter-varying systems, where the inputs influence the system differently depending on the current state. This approach does not impose any structure on how the state and input information are aggregated in the network, since the layers of the network make no distinction between the two.

Alternatively, two separate networks $\mathcal{A}$ and $\mathcal{B}$ can be used to model contributions of the autonomous and forced parts of the dynamics, respectively, as seen in Figure 17(b). This information can then be aggregated by taking the sum of the two terms:

$$\tilde{x}_{i+1} = \mathcal{A}(\tilde{x}_i) + \mathcal{B}(u_i). \tag{13}$$

This approach is suitable for systems where the influence of the inputs is known to be independent of the state of the system since it structurally enforces models that are independent.

In system identification and control theory, both variants (12) and (13) are referred to as **state-space models (SSMs)** [56, 66, 116, 117]. More recently, researchers [43, 64, 104, 123] proposed to model non-linear SSMs by using NNs, which we refer to as neural SSMs.

Some works proposed to combine neural approximations with classical approaches with linear state transition dynamics $\mathcal{A}$, resulting in Hammerstein [91] and Hammerstein-Wiener [47] architectures, or using linear operators representing transfer function as layers in deep NNs [30]. However, others leverage encoder-decoder neural architectures to handle partially observable dynamics [37, 81]. Some authors [26, 120, 121] applied principles of gray-box modeling by imposing physics-informed constraints on a learned neural SSM. Ogunmolu et al. [90] analyzed the effect of different neural architectures on the system identification performance of non-linear systems and concluded that compared to classical non-linear regressive models, deep NNs scale better and are easier to train.

4.3.2 NODEs with External Input. The challenge of introducing external input to NODEs is that the numerical solver may try to evaluate the derivative function at time instances that align with the sampled values of the external input. For instance, an adaptive step-size solver may choose its own internal step size based on how rapidly the derivative function changes in the neighborhood of the current state. The issue can be solved using interpolation to obtain values of external inputs for time instances that do not coincide with the sampling.

External input can also be used to represent static parameter values that remain constant through a simulation. In the context of the ideal pendulum system, we could imagine that the length of the pendulum could be made a parameter of the model, allowing the model to simulate the system under different conditions. Lee and Parish [67] call this approach *parameterized* NODEs and use this mechanism to train models that can solve PDEs for different parameter values.

Another approach is **neural controlled differential equations (NCDEs)** [57]. The term *controlled* should not be confused with the field of *control theory* but rather the mathematical concept of controlled differential equations from the field of rough analysis. The core idea of NCDEs is to treat the progression of time and the external inputs as a signal that *drives* the evolution of the system's state over time. The way that a specific system responds to this signal is approximated using an NN. A benefit of this approach is that it generalizes how a system's autonomous and forced dynamics are modeled. Specifically, it allows NCDEs to be applied to systems where NODEs would be applied, as well as systems where the output is purely driven by the external input to the system.

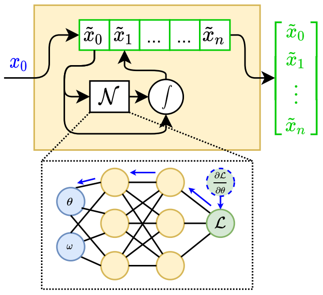
4.4 Network Architecture

Part of the success of NNs can be attributed to the ease of integrating specialized architectures into a model. In this section, we introduce a few examples of how to integrate domain-specific NNs into a time-stepper model.

First, Section 4.4.1 describes how energy-conserving dynamics can be enforced by encoding the problem using Hamiltonian or Lagrangian mechanics. Next, Section 4.4.2 demonstrates another way of enforcing energy conservation, which is often encountered in **molecular dynamics (MD)**.

Finally, Section 4.4.3 describes how graphs can be integrated with a time-stepper to solve problems that can naturally be represented as graphs.

4.4.1 Hamiltonian and Lagrangian Networks. Recall that the movement in some physical systems happens as a result of energy transfers within the system, as opposed to systems where energy is transferred to/from the system. The former is called an *energy conservative system*. For instance, if the pendulum introduced in Figure 3 had no friction and no external forces acting on it, it would oscillate forever, with its kinetic and potential energy oscillating without a change in its total energy. In physics, a special class of closely related functions, called *Hamiltonian and Lagrangian functions*, has been developed for describing the total energy of a system. Both Hamiltonian $\mathcal{H}$ and Lagrangian $\mathcal{L}$ are defined as a sum of total kinetic T and potential energy V of the system. We start with the Hamiltonian defined as



x_0
$\tilde{x}_0$ $\tilde{x}_1$ ... ... $\tilde{x}_n$
$\mathcal{N}$
$\int$
θ
ω
$\frac{\partial \mathcal{L}}{\partial \theta}$
$\mathcal{L}$
$\tilde{x}_0$ $\tilde{x}_1$ $\vdots$ $\tilde{x}_n$

Fig. 18. Lagrangian time-stepper. The Lagrangian, $\mathcal{L}$ (not to be confused with the loss function), is differentiated using AD to obtain the derivative of the state.

$$\mathcal{H}(x) = T(x) - V(x), \tag{14}$$

where $x = [q, p]$ represents the concatenated state vector of generalized coordinates q and generalized momenta p. By taking the gradients of the energy function (14), we can derive a corresponding differential $\dot{x} = f(x)$ equation as

$$\dot{x} = S\nabla\mathcal{H}(x), \tag{15}$$

where S is a symplectic matrix. Please note that the difference between $\mathcal{H}$ and $\mathcal{L}$ is their corresponding coordinate system: for the Lagrangian, instead of $x = [q, p]$, we consider $x = [q, \dot{q}]$, where $\dot{p} = M(q)\dot{q}$, with $M(q)$ being a generalized mass matrix.

Despite their mathematical elegance, deriving analytical Hamiltonian and Lagrangian functions for complex dynamical systems is a grueling task. In recent years, the research community turned its attention to deriving these types of scalar-valued energy functions by means of data-driven methods [41, 77, 142]. Specifically, the goal is to train an NN to approximate the Hamiltonian/Lagrangian of the system, as shown in Figure 18. A key aspect of this approach is that the derivatives of the states are not outputs of the network but are instead obtained by differentiating the output of the network $\mathcal{L}$, with respect to the state variables $[\theta, \omega]$ and plugging the results into Equation (14). The main advantage of Hamiltonian [41, 124] NNs and the closely related Lagrangian [21, 77] NNs is that they naturally incorporate the preservation of energy into the network structure itself. Research into the simulation of energy-preserving systems has yielded a special class of solvers, called *symplectic solvers*. Jin et al. [52] propose a new specialized network architecture, referred to as *symplectic networks*, to ensure that the dynamics of the model are energy conserving. Similarly, Finzi et al. [29] propose extensions for including explicit constraints via Lagrange multipliers for improved training efficiency and accuracy.

4.4.2 Deep Potential Energy Networks. A similar concept to that of Hamiltonian and Lagrangian NNs involves learning neural surrogates for potential energy functions $V(x)$ of a dynamical system, where the primary difference with Hamiltonians and Lagrangians is that the kinetic terms are

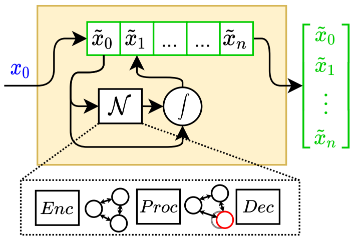

x_0
$\tilde{x}_0$ $\tilde{x}_1$ ... ... $\tilde{x}_n$
$\mathcal{N}$
$\int$
$\tilde{x}_0$ $\tilde{x}_1$ $\vdots$ $\tilde{x}_n$
Enc
Proc
Dec

Fig. 19. A simplified view of a graph time-stepper. During each step of the simulation, the current state is encoded as a graph (Enc) that is then used to compute the change in state variable between the current and next timestep (Proc). Finally, the change in state is decoded to the original state space to update the state of the system (Dec).

encoded explicitly in the time stepper by considering classical Newtonian laws of motion:

$$\tilde{x}_{i+1} = \tilde{x}_i + \tilde{v}_i, \tag{16a}$$

$$\tilde{v}_{i+1} = \tilde{v}_i - \frac{\nabla \mathcal{V}(x)}{m}, \tag{16b}$$

where x_i, and v_i are positional and velocity vectors of the system. The gradients of the potential function are equal to the interaction forces $F = -\nabla \mathcal{V}(x)$, whereas m is a vector of "masses."

This approach is extensively used, mainly in the domain of MD simulations [6, 50, 125, 126, 130, 139]. In modern data-driven MD, the learned neural potentials $V(x)$ replace expensive quantum chemistry calculations based, for example, on density functional theory (DFT). The advantage of this approach for large-scale systems, compared to directly learning high-dimensional maps of the time-steppers, is that the learning of the scalar-valued potential function $V(x) : \mathbf{R}^n \to \mathbf{R}$ represents a much simpler regression problem. Furthermore, this approach allows prior information to be encoded in the architecture of the deep potential functions $V(x)$, such as considering only local interactions between atoms [119], and encoding spatial symmetries [34, 140]. As a result, these methods are allowing researchers in MD to achieve unprecedented scalability, allowing simulation of up to 100M atoms on supercomputers [49]. In contrast, training a single naive time-stepper for such a model would require learning a 300M-dimensional mapping.

4.4.3 Graph Time-Steppers. Many complex real-world systems from social networks and molecules to power grid systems can be represented as graph structures describing the interactions between individual subsystems. Recent research in **graph neural networks (GNNs)** embraces this idea by embedding or learning the underlying graph structure from data. There exists a large body of work on GNNs, but covering this is outside the scope of this survey. We refer the interested reader to overview papers [3, 10, 115, 137, 141, 143]. For the purposes of this section, we focus solely on GNN-based time-stepper models applied to model dynamical systems [58, 71].

The core idea of using GNNs inside time-steppers is to use a GNN-based pipeline to estimate the derivatives of the system, as shown in Figure 19. Generally, the pipeline can be split into three steps; first, the current state of the system is encoded as a graph; next, the graph is processed to produce an update of the system's state; and finally, the update is decoded and used to update the state of the system.

One of the early works includes interaction networks [4] or *neural physics engine* (NPE) [16] demonstrating the ability to learn the dynamics in various physical domains in smaller scale dimensions, such as n-body problems, rigid-body collision, and non-rigid dynamics. Since then, the use of GNNs rapidly expanded, finding its use in NODE time-steppers [112] including control inputs [72, 114], dynamic graphs [109], or considering feature encoders enabling learning dynamics directly from the visual signals [134]. Modern GNNs are trained using message-passing algorithms introduced in the context of quantum chemistry application [39]. In GNNs, each node has associated latent variables representing values of physical quantities such as positions, charges, or velocities, then in the message-passing step, the aggregated values of the latent states are passed through the edges to update the values of the neighboring nodes. This abstraction efficiently encodes local structure-preserving interactions that commonly occur in the natural world. Although early implementations of GNN-based time-steppers suffered from larger computational complexity, more recent works [113] have demonstrated their scalability to ever larger dynamical systems with thousands of state variables over long prediction horizons. Due to their expressiveness and generic nature, GNNs could in principle be applied in all the time-stepper variants summarized in this article, some of which would represent novel architectures up to date.

4.5 Uncertainty

So far, we have considered only the cases of modeling systems where noise-free trajectories were available for training. In reality, it is likely that the data captured from the system does not represent the true state of the system, x, but rather a noisy version of the original signal perturbed by measurement noise. Another source of uncertainty is that the dynamics of the system itself may exhibit some degree of randomness. One cause of this would be unidentified external forces acting on the system. For instance, the dynamics of a physical pendulum may be influenced by vibrations from its environment. The following sections introduce several models that explicitly incorporate uncertainty in their predictions.

4.5.1 Deep Markov Models. A **deep Markov model (DMM)** [2, 32, 65, 73, 86] is a probabilistic model that combines the formalism of Markov chains with the use of an NN for approximating unknown probability density functions. A Markov chain is a latent variable model, which assumes that the values we observe from the system are determined by an underlying latent variable, which cannot be observed directly. This idea is very similar to an SSM, the difference being that a Markov chain assumes that the mapping from the latent to the observed variable is probabilistic and that evolution of the latent variable is not fully deterministic.

The relationship between the observed and latent variables of a DMM can be specified as follows:

$$z_{i+1} \sim \mathcal{Z}(N_t(z_i)), \qquad \text{(Transition)} \tag{17a}$$
$$x_i \sim \mathcal{X}(N_e(z_i)), \qquad \text{(Emission)} \tag{17b}$$

where z_i represents the latent state vector and x_i is the output vector. Here, $\mathcal{Z}$ and $\mathcal{X}$ represent probability distributions, commonly Gaussian distributions, modeled by maps $N_T(\mathbf{z}_i)$ or $N_e(\mathbf{z}_i)$, respectively.

A natural question to ask is how the observed and latent variables are represented, given that they are probability density functions and not numerical values—a solution to pick distributions that can be represented in terms of a few characteristic parameters. For instance, a Gaussian can be represented by its mean and covariance. The process of performing inference using a DMM is shown in Figure 20.

An obstacle to training DMMs using supervised learning is that the training data only contains targets for the observed variables x, not the latent variables z. A popular approach for training DMMs is using **variational inference (VI)**. It should be noted that VI is a general method for fitting the parameters of statistical models to data. In this special case, we happen to be applying it in a case where there is a dependence between samples in time. For a concrete example of a

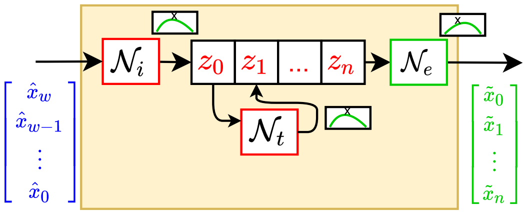


Fig. 20. Deep Markov model with an inference network. The value of z_0 is estimated by an inference network N_i based on several samples of the observed variable. The transmission function, approximated by the network N_t, maps the current value of z to a distribution over z one step ahead in time. The emission function, approximated by N_e, maps each predicted latent variable to a distribution of the corresponding x value in the original observed space. Note that the output of each network is the parameters of a distribution, which is then sampled to obtain a value that can be fed into the next stage of the model.

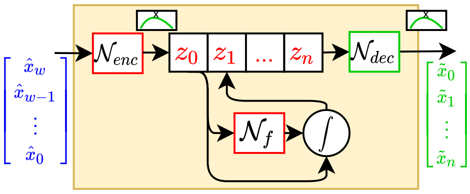


Fig. 21. Latent NODEs. An encoder network is used to obtain a latent representation of the system's initial state, z_0, by aggregating information from several observations of the systems $[\hat{x}_w, \hat{x}_{w-1}, \dots, \hat{x}_0]$. The system is simulated for multiple steps to obtain $[z_0, z_1, \dots, z_n]$. Finally, the latent variables are mapped back to the original state space by a decoder network.

training algorithm based on VI that is suitable for training DMM, we refer to the work of Krishnan et al. [65].

Although probability distributions in classical DMMs are assumed to be Gaussian, recent extensions proposed the use of more expressive but also more computationally expensive deep normalizing flows [38, 106]. Another variant of DMM includes additional graph structure for possible encoding of useful inductive biases [98]. DMMs are typically trained using the stochastic counterpart of the backpropagation algorithm [107], which is part of popular open source libraries such as PyTorch-based Pyro [8] or TensorFlow Probability [24]. Applications in dynamical systems modeling span from climate forecasting [17], MD [136], or generic time series modeling with uncertainty quantification [83].

4.5.2 Latent NODEs. Latent NODEs [18] is an extension of NODEs that introduces an encoder and decoder NN to the model as shown in Figure 21. The core of the idea is that information from multiple observations can be aggregated by the encoder network N_{enc} to obtain a latent state z_0, which characterizes the specific trajectory. A convenient choice of encoder network for time series is an RNN because it can handle a variable number of observations. The system can then be simulated using the same approach as NODEs to produce a solution in the latent space. Finally, a decoder network maps each point of the latent solution to the observable space to obtain the final solution.

Separating the measurement, $\mathbf{x}_i$, from the latent system dynamics, $\mathbf{z}_i$, allows us to exploit the modeling flexibility of wider NNs capable of generating more complex latent trajectories. However, by doing so, it creates an inference problem of estimating unknown initial conditions of the hidden states for both deterministic [68, 121] and stochastic time-steppers [20, 62, 63, 68].

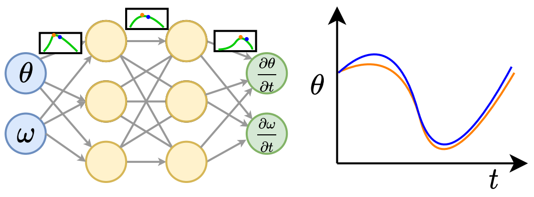

θ
ω
$\frac{\partial \theta}{\partial t}$
$\frac{\partial \omega}{\partial t}$
θ
t

Fig. 22. Bayesian neural ordinary differential equations. The parameters of the network are characterized by a probability distribution. The parameter distributions are sampled multiple times and used to simulate the system, producing multiple trajectories as shown on the right. To get a single prediction, the predictions can be averaged.

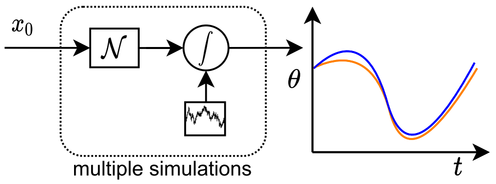


Fig. 23. Neural SDEs. The network $\mathcal{N}$ is used to approximate the deterministic drift term of the SDE and the diffusion term is a Wiener process. Multiple trajectories are produced by solving the SDE multiple times, corresponding to different realizations of the Wiener process.

A difference between latent NODEs and DMMs is that the former treats the state variable as a continuous-time variable and the latter treats it as discrete-time. In addition, latent NODEs assume that the dynamics are deterministic.

4.5.3 Bayesian Neural Ordinary Differential Equations. **Bayesian neural ordinary differential equations (BNODEs)** [22] combine the concept of a NODE with the stochastic nature of **Bayesian neural networks (BNNs)** [53]. In the context of a BNN, the term *Bayesian* refers to the fact that the parameters of the network are characterized by a probability density function rather than an exact value. For instance, the weights of the networks may be assumed to be approximately distributed according to a multivariate Gaussian.

A possible motivation for applying this formalism is that the uncertainty of the model's predictions can be quantified, which would otherwise not be possible. To obtain an estimate of the uncertainty, the model can be simulated several times using different realizations of the model's parameters, resulting in several trajectories as shown in Figure 22. The ensemble of trajectories can then be used to infer confidence bounds and to obtain the mean value of the trajectories.

A drawback of using BNNs and extensions like BNODEs is that they use specialized training algorithms that generally do not scale well to large network architectures. An alternative approach is to introduce sources of stochasticity during the training and inference, such as by using dropout. A categorization of ways to introduce stochasticity that do not require specialized training algorithms is provided in Section 8 of the work of Jospin et al. [53].

4.5.4 Neural Stochastic Differential Equations. Neural **stochastic differential equations (SDEs)** [75] can be viewed as a generalization of an ODE that includes one or more stochastic terms in addition to the deterministic dynamics, as shown in Figure 23. Like the DTMC, an SDE often includes a deterministic drift term and a stochastic diffusion term, such as *Wiener process*:

$$dX = f(x(t))dt + g(x(t))dW_t. \tag{18}$$

Table 2. Comparison of Direct-Solution and Time-Stepper Models

Name	Advantages	Limitations
Direct-solution	+ Easy to apply to PDEs + No discretization of time and spatial coordinates + No accumulation of error during simulation + Parallel evaluation of simulation	- Fixed initial condition - Fixed temporal and spatial domain - Difficult to incorporate inputs
Time-stepper	+ Initial condition not fixed + Easy to incorporate inputs + Leverage knowledge from numerical simulation	- Not trivial to apply to PDEs - Accumulation of error during simulation - No parallel evaluation of simulation

Conventionally, SDEs are expressed in *differential form* unlike the derivative form of an ODE. The reason for this is that many stochastic processes are continuous but cannot be differentiated. The meaning of Equation (18) is per definition the integral equation:

$$x(t) = x_0 + \int_0^t f(x(s))ds + \int_0^t g(x(s))dW_s. \tag{19}$$

As is the case for ODEs, most SDEs must be solved numerically, since only very few SDEs have analytical solutions. Solving SDEs requires the use of algorithms that are different from those used to solve deterministic ODEs. Covering the solvers is outside the scope of this work; instead, we refer to Chapter 9 of the work of Kloeden and Platen [59] for an in-depth coverage. However, in the context of neural SDEs, we can simply think of the solver as a means to simulate systems with stochastic dynamics.

There are several choices for how to incorporate the use of NNs for modeling SDEs. For instance, if the stochastic diffusion term is known, an NN can be trained to approximate the deterministic drift term in Equation (18) as in the case of the work of Liu et al. [74] and Oganesyan et al. [89]. Another approach is to use NNs to parameterize both the drift and diffusion terms [46]. In addition, there are approaches such as those of Xu et al. [138] that incorporate the idea from both neural SDEs and BNNs, by modeling both evolution of the state variables and network parameters as SDEs.

Although neural SDEs provide a strong theoretical framework for modeling uncertainty, they are complex compared to their deterministic counterparts. One way to address this is to examine if simpler and computationally efficient mechanisms like injecting noise or using dropout can achieve some of the same effects as adopting a fully SDE-based framework.

5 DISCUSSION

An important question is how to pick the right type of model for a given application. The two fundamentally different approaches for simulating a system are (i) having an NN approximate the solution of the problem, as described in Section 3, or (ii) having an NN approximate the dynamics of the system, as described in Section 4. Each approach has inherent advantages and limitations, which can be derived by looking at what the NN is used for within the respective type of model. A comparison between the two types of models can be seen in Table 2.

In this survey, we described several variants of direct-solution and time-stepper models. The way that these are presented in the literature often gives the impression that they are fundamentally different. However, applying them to the ideal pendulum system makes it clear that many models are closely related, set apart only by a small extension of the original idea. In the case of the direct-solution models, we observed that the differences between the vanilla direct-solution and the PINN is the application of physics-based regularization and use of AD for obtaining the velocity. In the case of time-stepper models, the main differences boil down to the architecture of the NN and the numerical integration scheme being applied. The ability to pick an NN architecture for a specific application makes it possible to model a wide range of physical phenomena. In addition, the ideas of one model can easily be transferred to another, allowing for the creation of novel architectures. This inherent variability makes it difficult to define concrete guidelines for picking a type of model for a certain application. Instead, we urge the reader to consider what capabilities are needed for the application and how knowledge of the physics incorporated. The topics described by Figure 2 may serve as a starting point for this.

Evaluating the performance of different models on a benchmark dataset consisting of data from various dynamical systems would be quite useful. This dataset should be representative of the systems that are encountered in disciplines such as physics, chemistry, and engineering. This would allow us to identify general trends and heuristics, which would serve as a starting point for new practitioners and future applications. Drawing inspiration from other applications of DL, such as image classification, we see that large image databases have contributed greatly toward developing better NN architectures. A standardized benchmark dataset is an essential step toward gaining more insight into which types of models work well. Not only would it allow for a fair comparison between the NN-based models, but it would also allow us to answer the question of how well these models work compared to traditional models originating from various fields.

Another valuable contribution, would be to define a procedure for evaluating a model's ability to approximate a dynamical system. We are interested in verifying that the model can produce accurate simulations for the initial conditions we would encounter when using the model. Given the diverse nature of these dynamical systems, some may be more difficult for an NN to approximate than others. For instance, a small approximation error in a chaotic system may result in the accumulation of a large error over time. An interesting research topic is determining metrics that allow a fair comparison across multiple dynamical systems.

Another valuable contribution would be to develop concrete guidelines on how to train models of dynamical systems. Finding a rule of thumb for how much training data is necessary to reach a certain degree of accuracy would make it easier to determine if a data-driven approach is feasible for a given application. In addition to determining how much data we need, it would be useful to develop best practices on how to split the data into training and validation sets. For instance, in the context of training time-stepper models, we may examine which length of trajectories result in a good ratio between accuracy and training time. Likewise, it would be useful to determine how to formulate the loss function such that the process of optimizing the model's parameters is fast and robust.

6 SUMMARY

In recent years, there has been an increased interest in applying NNs to solve a diverse set of problems encountered in various branches of engineering and natural sciences. This has resulted in a wealth of papers, each proposing how a particular physical phenomenon can be simulated using NNs. As a consequence, the terminology and notation used in each paper vary greatly, making it difficult to digest for all but experts in the respective field. These papers, often constrained in space, put great emphasis on describing the application and the physics involved, often at a cost of omitting details like how the NN was trained and limitations of proposed methods.

This survey provides an easy-to-follow overview of the techniques for simulating dynamical systems based on NNs. Specifically, we categorized the models encountered in the literature into two distinct types: direct-solution- and time-stepper models. For each type of model, we provided a concrete guide on how to construct, train, and use the model for simulation. Starting from the simplest possible model, we incrementally introduced more advanced variants and established the differences and similarities between the models. In addition, we supply source code for many of the models described in the article, which can be used as a reference for detailed implementation of each model.

An open research question is determining how well these methods work across a broad set of problems that are representative of real-world applications. It is our hope that this survey will support this goal by presenting the most important concepts in a way that is accessible to practitioners coming from DL as well as various branches of physics and engineering.

REFERENCES

[1] Martín Abadi, Ashish Agarwal, Paul Barham, Eugene Brevdo, Zhifeng Chen, Craig Citro, Greg S. Corrado, et al. 2016. TensorFlow: Large-scale machine learning on heterogeneous distributed systems. arXiv:1603.04467 (2016).

[2] Maren Awiszus and Bodo Rosenhahn. 2018. Markov chain neural networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*. 2180–2187.

[3] Peter W. Battaglia, Jessica B. Hamrick, Victor Bapst, Alvaro Sanchez-Gonzalez, Vinícius Flores Zambaldi, Mateusz Malinowski, et al. 2018. Relational inductive biases, deep learning, and graph networks. *CoRR* abs/1806.01261 (2018).

[4] Peter W. Battaglia, Razvan Pascanu, Matthew Lai, Danilo Jimenez Rezende, and Koray Kavukcuoglu. 2016. Interaction networks for learning about objects, relations and physics. *CoRR* abs/1612.00222 (2016).

[5] Atilim Gunes Baydin, Barak A. Pearlmutter, Alexey Andreyevich Radul, and Jeffrey Mark Siskind. 2018. Automatic differentiation in machine learning: A survey. arXiv:1502.05767 [cs, stat] (Feb. 2018).

[6] Jörg Behler. 2015. Constructing high-dimensional neural network potentials: A tutorial review. *International Journal of Quantum Chemistry* 115, 16 (2015), 1032–1050. https://doi.org/10.1002/qua.24890

[7] Jens Behrmann, Will Grathwohl, Ricky T. Q. Chen, David Duvenaud, and Joern-Henrik Jacobsen. 2019. Invertible residual networks. In *Proceedings of the 36th International Conference on Machine Learning*. Proceedings of Machine Learning Research, Vol. 97, Kamalika Chaudhuri and Ruslan Salakhutdinov (Eds.). PMLR, 573–582.

[8] Eli Bingham, Jonathan P. Chen, Martin Jankowiak, Fritz Obermeyer, Neeraj Pradhan, Theofanis Karaletsos, Rohit Singh, Paul Szerlip, Paul Horsfall, and Noah D. Goodman. 2019. Pyro: Deep universal probabilistic programming. *Journal of Machine Learning Research* 20, 1 (2019), 973–978.

[9] Christopher Bishop. 2006. *Pattern Recognition and Machine Learning*. Springer-Verlag, New York, NY.

[10] Michael M. Bronstein, Joan Bruna, Taco Cohen, and Petar Velickovic. 2021. Geometric deep learning: Grids, groups, graphs, geodesics, and gauges. *CoRR* abs/2104.13478 (2021).

[11] Steven L. Brunton, Bernd R. Noack, and Petros Koumoutsakos. 2020. Machine learning for fluid mechanics. *Annual Review of Fluid Mechanics* 52, 1 (2020), 477–508. https://doi.org/10.1146/annurev-fluid-010719-060214

[12] Keith T. Butler, Daniel W. Davies, Hugh Cartwright, Olexandr Isayev, and Aron Walsh. 2018. Machine learning for molecular and materials science. *Nature* 559, 7715 (July 2018), 547–555. https://doi.org/10.1038/s41586-018-0337-2

[13] François Edouard Cellier. 1991. *Continuous System Modeling*. Springer Science & Business Media.

[14] François Edouard Cellier and Ernesto Kofman. 2006. *Continuous System Simulation*. Springer Science & Business Media.

[15] Bo Chang, Lili Meng, Eldad Haber, Frederick Tung, and David Begert. 2018. Multi-level residual networks from dynamical systems view. arXiv:1710.10348 [cs, stat] (Feb. 2018).

[16] Michael B. Chang, Tomer Ullman, Antonio Torralba, and Joshua B. Tenenbaum. 2016. A compositional object-based approach to learning physical dynamics. *CoRR* abs/1612.00341 (2016).

[17] Zhengping Che, Sanjay Purushotham, Guangyu Li, Bo Jiang, and Yan Liu. 2018. Hierarchical deep generative models for multi-rate multivariate time series. In *Proceedings of the 35th International Conference on Machine Learning*. Proceedings of Machine Learning Research, Vol. 80, Jennifer Dy and Andreas Krause (Eds.). PMLR, 784–793.

[18] Ricky T. Q. Chen, Yulia Rubanova, Jesse Bettencourt, and David Duvenaud. 2019. Neural ordinary differential equations. arXiv:1806.07366 [cs, stat] (Dec. 2019).

[19] Travers Ching, Daniel S. Himmelstein, Brett K. Beaulieu-Jones, Alexandr A. Kalinin, Brian T. Do, Gregory P. Way, Enrico Ferrero, et al. 2018. Opportunities and obstacles for deep learning in biology and medicine. *Journal of the Royal Society Interface* 15, 141 (April 2018), 20170387. https://doi.org/10.1098/rsif.2017.0387

[20] Junyoung Chung, Kyle Kastner, Laurent Dinh, Kratarth Goel, Aaron C. Courville, and Yoshua Bengio. 2015. A recurrent latent variable model for sequential data. In *Advances in Neural Information Processing Systems*, C. Cortes, N. Lawrence, D. Lee, M. Sugiyama, and R. Garnett (Eds.), Vol. 28. Curran Associates.

[21] Miles Cranmer, Sam Greydanus, Stephan Hoyer, Peter Battaglia, David Spergel, and Shirley Ho. 2020. Lagrangian neural networks. arXiv:2003.04630 [physics, stat] (July 2020).

[22] Raj Dandekar, Karen Chung, Vaibhav Dixit, Mohamed Tarek, Aslan Garcia-Valadez, Krishna Vishal Vemula, and Chris Rackauckas. 2021. Bayesian neural ordinary differential equations. arXiv:2012.07244 [cs] (March 2021).

[23] Moritz Diehl, H. Georg Bock, Johannes P. Schlöder, Rolf Findeisen, Zoltan Nagy, and Frank Allgöwer. 2002. Real-time optimization and nonlinear model predictive control of processes governed by differential-algebraic equations. *Journal of Process Control* 12, 4 (2002), 577–585. https://doi.org/10.1016/S0959-1524(01)00023-3

[24] Joshua V. Dillon, Ian Langmore, Dustin Tran, Eugene Brevdo, Srinivas Vasudevan, Dave Moore, Brian Patton, Alex Alemi, Matthew D. Hoffman, and Rif A. Saurous. 2017. TensorFlow distributions. *CoRR* abs/1711.10604 (2017).

[25] Ján Drgoňa, Javier Arroyo, Iago Cupeiro Figueroa, David Blum, Krzysztof Arendt, Donghun Kim, Enric Perarnau Ollé, et al. 2020. All you need to know about model predictive control for buildings. *Annual Reviews in Control* 50 (2020), 190–232. https://doi.org/10.1016/j.arcontrol.2020.09.001

[26] Jan Drgona, Aaron R. Tuor, Vikas Chandan, and Draguna L. Vrabie. 2020. Physics-constrained deep learning of multi-zone building thermal dynamics. arXiv:2011.05987 [cs.LG] (2020).

[27] Emilien Dupont, Arnaud Doucet, and Yee Whye Teh. 2019. Augmented neural ODEs. arXiv:1904.01681 [stat.ML] (2019).

[28] Chris Finlay, Jörn-Henrik Jacobsen, Levon Nurbekyan, and Adam M. Oberman. 2020. How to train your neural ODE: The world of Jacobian and kinetic regularization. arXiv:2002.02798 [stat.ML] (2020).

[29] Marc Finzi, Ke Alexander Wang, and Andrew Gordon Wilson. 2020. Simplifying Hamiltonian and Lagrangian neural networks via explicit constraints. *CoRR* abs/2010.13581 (2020).

[30] Marco Forgione and Dario Piga. 2020. dynoNet: A neural network architecture for learning dynamical systems. arXiv:2006.02250 [cs.LG] (2020).

[31] Alexander I. J. Forrester and Andy J. Keane. 2009. Recent advances in surrogate-based optimization. *Progress in Aerospace Sciences* 45, 1-3 (Jan. 2009), 50–79. https://doi.org/10.1016/j.paerosci.2008.11.001

[32] Marco Fraccaro, Søren Kaae Sønderby, Ulrich Paquet, and Ole Winther. 2016. Sequential neural models with stochastic layers. arXiv preprint arXiv:1605.07571 (2016).

[33] Jonathan Friedman and Jason Ghidella. 2006. Using model-based design for automotive systems engineering—Requirements analysis of the power window example. *Journal of Passenger Cars: Electronic and Electrical Systems* 115 (2006), 516–521. https://doi.org/10.4271/2006-01-1217

[34] Xiang Gao, Farhad Ramezanghorbani, Olexandr Isayev, Justin S. Smith, and Adrian E. Roitberg. 2020. TorchANI: A free and open source PyTorch-based deep learning implementation of the ANI neural network potentials. *Journal of Chemical Information and Modeling* 60, 7 (2020), 3408–3415. https://doi.org/10.1021/acs.jcim.0c00451

[35] Carlos E. García, David M. Prett, and Manfred Morari. 1989. Model predictive control: Theory and practice—A survey. *Automatica* 25, 3 (1989), 335–348. https://doi.org/10.1016/0005-1098(89)90002-2

[36] C. W. Gear and O. Osterby. 1984. Solving ordinary differential equations with discontinuities. *ACM Transactions on Mathematical Software* 10, 1 (Jan. 1984), 23–44. https://doi.org/10.1145/356068.356071

[37] Daniel Gedon, Niklas Wahlström, Thomas B. Schön, and Lennart Ljung. 2020. Deep state space models for nonlinear system identification. arXiv:2003.14162 [eess.SY] (2020).

[38] Anubhab Ghosh, Antoine Honoré, Dong Liu, Gustav Eje Henter, and Saikat Chatterjee. 2021. Robust classification using hidden Markov models and mixtures of normalizing flows. *CoRR* abs/2102.07284 (2021).

[39] Justin Gilmer, Samuel S. Schoenholz, Patrick F. Riley, Oriol Vinyals, and George E. Dahl. 2017. Neural message passing for quantum chemistry. *CoRR* abs/1704.01212 (2017).

[40] Ian Goodfellow, Yoshua Bengio, Aaron Courville, and Yoshua Bengio. 2016. *Deep Learning*, Vol. 1. MIT Press, Cambridge, MA.

[41] Samuel Greydanus, Misko Dzamba, and Jason Yosinski. 2019. Hamiltonian neural networks. In *Advances in Neural Information Processing Systems 32*, H. Wallach, H. Larochelle, A. Beygelzimer, F. F. d'Alché-Buc, E. Fox, and R. Garnett (Eds.). Curran Associates, 15379–15389.

[42] Batuhan Güler, Alexis Laignelet, and Panos Parpas. 2019. Towards robust and stable deep learning algorithms for forward backward stochastic differential equations. arXiv:1910.11623 [stat.ML] (2019).

[43] Danijar Hafner, Timothy P. Lillicrap, Ian Fischer, Ruben Villegas, David Ha, Honglak Lee, and James Davidson. 2018. Learning latent dynamics for planning from pixels. *CoRR* abs/1811.04551 (2018).

[44] Ernst Hairer and Gerhard Wanner. 1996. *Solving Ordinary Differential Equations II: Stiff and Differential-Algebraic Problems*. Number 14. Springer-Verlag, Berlin, Germany.

[45] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2015. Deep residual learning for image recognition. arXiv:1512.03385 [cs] (Dec. 2015).

[46] Pashupati Hegde, Markus Heinonen, Harri Lähdesmäki, and Samuel Kaski. 2018. Deep learning with differential Gaussian process flows. arXiv:1810.04066 [cs, stat] (Oct. 2018).

[47] Jeen-Shing Wang and Yi-Chung Chen. 2008. A Hammerstein-W recurrent neural network with universal approximation capability. In *Proceedings of the 2008 IEEE International Conference on Systems, Man, and Cybernetics*. 1832–1837. https://doi.org/10.1109/ICSMC.2008.4811555

[48] Junteng Jia and Austin R. Benson. 2019. Neural jump stochastic differential equations. *CoRR* abs/1905.10403 (2019).

[49] Weile Jia, Han Wang, Mohan Chen, Denghui Lu, Lin Lin, Roberto Car, Weinan E, and Linfeng Zhang. 2020. Pushing the limit of molecular dynamics with ab initio accuracy to 100 million atoms with machine learning. arXiv:2005.00223 [physics.comp-ph] (2020).
[50] Bin Jiang, Jun Li, and Hua Guo. 2016. Potential energy surfaces from high fidelity fitting of ab initio points: The permutation invariant polynomial–neural network approach. *International Reviews in Physical Chemistry* 35, 3 (2016), 479–506. https://doi.org/10.1080/0144235X.2016.1200347
[51] Zhihao Jiang, Miroslav Pajic, Rajeev Alur, and Rahul Mangharam. 2014. Closed-loop verification of medical devices with model abstraction and refinement. *International Journal on Software Tools for Technology Transfer* 16, 2 (April 2014), 191–213. https://doi.org/10.1007/s10009-013-0289-7
[52] Pengzhan Jin, Aiqing Zhu, George Em Karniadakis, and Yifa Tang. 2020. Symplectic networks: Intrinsic structure-preserving networks for identifying Hamiltonian systems. *CoRR* abs/2001.03750 (2020).
[53] Laurent Valentin Jospin, Wray Buntine, Farid Boussaid, Hamid Laga, and Mohammed Bennamoun. 2020. Hands-on Bayesian neural networks—A tutorial for deep learning users. arXiv:2007.06823 [cs, stat] (July 2020).
[54] Anuj Karpatne, Gowtham Atluri, James H. Faghmous, Michael Steinbach, Arindam Banerjee, Auroop Ganguly, Shashi Shekhar, Nagiza Samatova, and Vipin Kumar. 2017. Theory-guided data science: A new paradigm for scientific discovery from data. *IEEE Transactions on Knowledge and Data Engineering* 29, 10 (Oct. 2017), 2318–2331. https://doi.org/10.1109/TKDE.2017.2720168
[55] Jacob Kelly, Jesse Bettencourt, Matthew James Johnson, and David Duvenaud. 2020. Learning differential equations that are easy to solve. arXiv:2007.04504 [cs.LG] (2020).
[56] Gaëtan Kerschen, Keith Worden, Alexander F. Vakakis, and Jean-Claude Golinval. 2006. Past, present and future of nonlinear system identification in structural dynamics. *Mechanical Systems and Signal Processing* 20, 3 (2006), 505–592. https://doi.org/10.1016/j.ymssp.2005.04.008
[57] Patrick Kidger, Ricky T. Q. Chen, and Terry Lyons. 2020. "Hey, that's not an ODE": Faster ODE adjoints with 12 lines of code. arXiv:2009.09457 [cs, math] (Sept. 2020).
[58] Thomas Kipf, Ethan Fetaya, Kuan-Chieh Wang, Max Welling, and Richard Zemel. 2018. Neural relational inference for interacting systems. arXiv:1802.04687 [stat.ML] (2018).
[59] Peter E. Kloeden and Eckhard Platen. 1992. *Numerical Solution of Stochastic Differential Equations*. Springer.
[60] Ernesto Kofman and Sergio Junco. 2001. Quantized-state systems: A DEVS approach for continuous system simulation. *Transactions of the Society for Modeling and Simulation International* 18, 3 (2001), 123–132.
[61] Slawomir Koziel and Anna Pietrenko-Dabrowska. 2020. *Basics of Data-Driven Surrogate Modeling*. Springer International Publishing, Cham, Switzerland, 23–58. https://doi.org/10.1007/978-3-030-38926-0_2
[62] R. Krishnan, U. Shalit, and D. Sontag. 2017. Structured inference networks for nonlinear state space models. In *Proceedings of the 31st AAAI Conference on Artificial Intelligence (AAAI'17)*. 2101–2109.
[63] Rahul G. Krishnan, Uri Shalit, and David Sontag. 2015. Deep Kalman filters. arXiv:1511.05121 [stat.ML] (2015).
[64] Rahul G. Krishnan, Uri Shalit, and David Sontag. 2016. Structured inference networks for nonlinear state space models. arXiv:1609.09869 [stat.ML] (2016).
[65] Rahul G. Krishnan, Uri Shalit, and David Sontag. 2016. Structured inference networks for nonlinear state space models. arXiv:1609.09869 [cs, stat] (Dec. 2016).
[66] Andreas Kroll and Horst Schulte. 2014. Benchmark problems for nonlinear system identification and control using soft computing methods: Need and overview. *Applied Soft Computing* 25 (2014), 496–513. https://doi.org/10.1016/j.asoc.2014.08.034
[67] Kookjin Lee and Eric J. Parish. 2021. Parameterized neural ordinary differential equations: Applications to computational physics problems. *Proceedings of the Royal Society A: Mathematical, Physical and Engineering Sciences* 477, 2253 (Sept. 2021), 20210162. https://doi.org/10.1098/rspa.2021.0162
[68] I. Lenz, Ross A. Knepper, and A. Saxena. 2015. DeepMPC: Learning deep latent features for model predictive control. In *Proceedings of the Conference on Robotics: Science and Systems*.
[69] Randall J. LeVeque. 2007. *Finite Difference Methods for Ordinary and Partial Differential Equations: Steady-State and Time-Dependent Problems*, Vol. 98. SIAM.
[70] Xuechen Li, Ting-Kam Leonard Wong, Ricky T. Q. Chen, and David Duvenaud. 2020. Scalable gradients for stochastic differential equations. arXiv:2001.01328 [cs.LG] (2020).
[71] Yunzhu Li, Jiajun Wu, Russ Tedrake, Joshua B. Tenenbaum, and Antonio Torralba. 2018. Learning particle dynamics for manipulating rigid bodies, deformable objects, and fluids. *CoRR* abs/1810.01566 (2018).
[72] Yunzhu Li, Jiajun Wu, Jun-Yan Zhu, Joshua B. Tenenbaum, Antonio Torralba, and Russ Tedrake. 2018. Propagation networks for model-based control under partial observation. *CoRR* abs/1809.11169 (2018).
[73] Dong Liu, Antoine Honoré, Saikat Chatterjee, and Lars K. Rasmussen. 2019. Powering hidden Markov model by neural network based generative models. arXiv preprint arXiv:1910.05744 (2019).

[74] Xuanqing Liu, Tesi Xiao, Si Si, Qin Cao, Sanjiv Kumar, and Cho-Jui Hsieh. 2019. Neural SDE: Stabilizing neural ODE networks with stochastic noise. arXiv:1906.02355 [cs.LG] (2019).

[75] Xuanqing Liu, Tesi Xiao, Si Si, Qin Cao, Sanjiv Kumar, and Cho-Jui Hsieh. 2019. Neural SDE: Stabilizing neural ODE networks with stochastic noise. arXiv:1906.02355 [cs, stat] (June 2019).

[76] Lennart Ljung. 2006. Some aspects of non linear system identification. *IFAC Proceedings Volumes* 39, 1 (2006), 110–121. https://doi.org/10.3182/20060329-3-AU-2901.00009

[77] Michael Lutter, Christian Ritter, and Jan Peters. 2019. Deep Lagrangian networks: Using physics as model prior for deep learning. arXiv:1907.04490 [cs, eess, stat] (July 2019).

[78] J. E. Marsden and M. West. 2001. Discrete mechanics and variational integrators. *Acta Numerica* 10 (May 2001), 357–514. https://doi.org/10.1017/S096249290100006X

[79] Stefano Massaroli, Michael Poli, Michelangelo Bin, Jinkyoo Park, Atsushi Yamashita, and Hajime Asama. 2020. Stable neural flows. arXiv:2003.08063 [cs.LG] (2020).

[80] Stefano Massaroli, Michael Poli, Jinkyoo Park, Atsushi Yamashita, and Hajime Asama. 2021. Dissecting neural ODEs. arXiv:2002.08071 [cs.LG] (2021).

[81] D. Masti and A. Bemporad. 2018. Learning nonlinear state-space models using deep autoencoders. In *Proceedings of the 2018 IEEE Conference on Decision and Control (CDC'18)*. 3862–3867.

[82] Sparsh Mittal and Shraiysh Vaishay. 2019. A survey of techniques for optimizing deep learning on GPUs. *Journal of Systems Architecture* 99 (Oct. 2019), 101635. https://doi.org/10.1016/j.sysarc.2019.101635

[83] George Montanez, Saeed Amizadeh, and Nikolay Laptev. 2015. Inertial hidden Markov models: Modeling change in multivariate time series. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 29.

[84] Mehrdad Moradi, Cláudio Gomes, Bentley James Oakes, and Joachim Denil. 2019. Optimizing fault injection in FMI co-simulation. In *Proceedings of the 2019 Summer Simulation Conference*. 12. https://doi.org/10.5555/3374138.3374170

[85] Kevin P. Murphy. 2012. *Machine Learning: A Probabilistic Perspective*. MIT Press, Cambridge, MA.

[86] Mohammed Kyari Mustafa, Tony Allen, and Kofi Appiah. 2019. A comparative review of dynamic neural networks and hidden Markov model methods for mobile on-device speech recognition. *Neural Computing and Applications* 31, 2 (2019), 891–899.

[87] Oliver Nelles. 2001. *Nonlinear System Identification: From Classical Approaches to Neural Networks and Fuzzy Models*. Springer-Verlag, Berlin, Germany. https://doi.org/10.1007/978-3-662-04323-3

[88] Alexander Norcliffe, Cristian Bodnar, Ben Day, Nikola Simidjievski, and Pietro Liò. 2020. On second order behaviour in augmented neural ODEs. arXiv:2006.07220 [cs.LG] (2020).

[89] Viktor Oganesyan, Alexandra Volokhova, and Dmitry Vetrov. 2020. Stochasticity in neural ODEs: An empirical study. arXiv:2002.09779 [cs, stat] (June 2020).

[90] Olalekan Ogunmolu, Xuejun Gu, Steve Jiang, and Nicholas Gans. 2016. Nonlinear systems identification using deep dynamic neural networks. arXiv:1610.01439 [cs] (Oct. 2016).

[91] Olalekan P. Ogunmolu, Xuejun Gu, Steve B. Jiang, and Nicholas R. Gans. 2016. Nonlinear systems identification using deep dynamic neural networks. *CoRR* abs/1610.01439 (2016).

[92] Katharina Ott, Prateek Katiyar, Philipp Hennig, and Michael Tiemann. 2020. When are neural ODE solutions proper ODEs? arXiv:2007.15386 [cs, stat] (July 2020).

[93] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, et al. 2019. PyTorch: An imperative style, high-performance deep learning library. In *Advances in Neural Information Processing Systems 32*, H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett (Eds.). Curran Associates, 8026–8037.

[94] Ludovic Pintard, Jean-Charles Fabre, Karama Kanoun, Michel Leeman, and Matthieu Roy. 2013. Fault injection in the automotive standard ISO 26262: An initial approach. In *Dependable Computing*, David Hutchison, Takeo Kanade, Josef Kittler, Jon M. Kleinberg, Friedemann Mattern, John C. Mitchell, Moni Naor, et al. (Eds.), Vol. 7869. Springer, Berlin, Germany, 126–133. https://doi.org/10.1007/978-3-642-38789-0_11

[95] Alessio Plebe and Giorgio Grasso. 2019. The unbearable shallow understanding of deep learning. *Minds and Machines* 29, 4 (Dec. 2019), 515–553. https://doi.org/10.1007/s11023-019-09512-8

[96] Michael Poli, Stefano Massaroli, Junyoung Park, Atsushi Yamashita, Hajime Asama, and Jinkyoo Park. 2020. Graph neural ordinary differential equations. arXiv:1911.07532 [cs.LG] (2020).

[97] Tong Qin, Kailiang Wu, and Dongbin Xiu. 2019. Data driven governing equations approximation using deep neural networks. *Journal of Computational Physics* 395 (Oct. 2019), 620–635. https://doi.org/10.1016/j.jcp.2019.06.042

[98] Meng Qu, Yoshua Bengio, and Jian Tang. 2019. GMNN: Graph Markov neural networks. In *Proceedings of the International Conference on Machine Learning*. 5241–5250.

[99] Alessio Quaglino, Marco Gallieri, Jonathan Masci, and Jan Koutník. 2020. SNODE: Spectral discretization of neural ODEs for system identification. arXiv:1906.07038 [cs.NE] (2020).

[100] R. Rai and C. K. Sahu. 2020. Driven by data or derived through physics? A review of hybrid physics guided machine learning techniques with cyber-physical system (CPS) focus. *IEEE Access* 8 (2020), 71050–71073.

[101] M. Raissi, P. Perdikaris, and G. E. Karniadakis. 2019. Physics-informed neural networks: A deep learning framework for solving forward and inverse problems involving nonlinear partial differential equations. *Journal of Computational Physics* 378 (Feb. 2019), 686–707. https://doi.org/10.1016/j.jcp.2018.10.045

[102] Maziar Raissi, Paris Perdikaris, and George Em Karniadakis. 2018. Multistep neural networks for data-driven discovery of nonlinear dynamical systems. arXiv:1801.01236 [nlin, physics:physics, stat] (Jan. 2018).

[103] Maziar Raissi, Alireza Yazdani, and George Em Karniadakis. 2020. Hidden fluid mechanics: Learning velocity and pressure fields from flow visualizations. *Science* 367, 6481 (Feb. 2020), 1026–1030. https://doi.org/10.1126/science.aaw4741

[104] Syama S. Rangapuram, Matthias W. Seeger, Jan Gasthaus, Lorenzo Stella, Yuyang Wang, and Tim Januschowski. 2018. Deep state space models for time series forecasting. In *Advances in Neural Information Processing Systems 31*, S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett (Eds.). Curran Associates, 7785–7794.

[105] Saman Razavi, Bryan A. Tolson, and Donald H. Burn. 2012. Review of surrogate modeling in water resources. *Water Resources Research* 48, 7 (July 2012), 1–32. https://doi.org/10.1029/2011WR011527

[106] Danilo Jimenez Rezende and Shakir Mohamed. 2016. Variational inference with normalizing flows. arXiv:1505.05770 [stat.ML] (2016).

[107] Danilo Jimenez Rezende, Shakir Mohamed, and Daan Wierstra. 2014. Stochastic backpropagation and approximate inference in deep generative models. In *Proceedings of the International Conference on Machine Learning*. 1278–1286.

[108] David Rolnick, Priya L. Donti, Lynn H. Kaack, Kelly Kochanski, Alexandre Lacoste, Kris Sankaran, Andrew Slavin Ross, et al. 2019. Tackling climate change with machine learning. arXiv:1906.05433 [cs, stat] (Nov. 2019).

[109] Emanuele Rossi, Ben Chamberlain, Fabrizio Frasca, Davide Eynard, Federico Monti, and Michael M. Bronstein. 2020. Temporal graph networks for deep learning on dynamic graphs. *CoRR* abs/2006.10637 (2020).

[110] Lars Ruthotto and Eldad Haber. 2018. Deep neural networks motivated by partial differential equations. arXiv:1804.04272 [cs, math, stat] (Dec. 2018).

[111] Lars Ruthotto and Eldad Haber. 2020. Deep neural networks motivated by partial differential equations. *Journal of Mathematical Imaging and Vision* 62, 3 (April 2020), 352–364. https://doi.org/10.1007/s10851-019-00903-1

[112] Alvaro Sanchez-Gonzalez, Victor Bapst, Kyle Cranmer, and Peter W. Battaglia. 2019. Hamiltonian graph networks with ODE integrators. *CoRR* abs/1909.12790 (2019).

[113] Alvaro Sanchez-Gonzalez, Jonathan Godwin, Tobias Pfaff, Rex Ying, Jure Leskovec, and Peter W. Battaglia. 2020. Learning to simulate complex physics with graph networks. *CoRR* abs/2002.09405 (2020).

[114] Alvaro Sanchez-Gonzalez, Nicolas Heess, Jost Tobias Springenberg, Josh Merel, Martin A. Riedmiller, Raia Hadsell, and Peter W. Battaglia. 2018. Graph networks as learnable physics engines for inference and control. *CoRR* abs/1806.01242 (2018).

[115] Franco Scarselli, Marco Gori, Ah Chung Tsoi, Markus Hagenbuchner, and Gabriele Monfardini. 2009. The graph neural network model. *IEEE Transactions on Neural Networks* 20, 1 (2009), 61–80. https://doi.org/10.1109/TNN.2008.2005605

[116] Johan Schoukens and Lennart Ljung. 2019. Nonlinear system identification: A user-oriented roadmap. *CoRR* abs/1902.00683 (2019).

[117] M. Schoukens and J. P. Noël. 2017. Three benchmarks addressing open challenges in nonlinear system identification. *IFAC-PapersOnLine* 50, 1 (2017), 446–451. https://doi.org/10.1016/j.ifacol.2017.08.071

[118] Dieter Schramm, Wildan Lalo, and Michael Unterreiner. 2010. Application of simulators and simulation tools for the functional design of mechatronic systems. *Solid State Phenomena* 166–167 (Sept. 2010), 1–14. https://doi.org/10.4028/www.scientific.net/SSP.166-167.1

[119] Kristof T. Schütt, Pieter-Jan Kindermans, Huziel E. Sauceda, Stefan Chmiela, Alexandre Tkatchenko, and Klaus-Robert Müller. 2017. SchNet: A continuous-filter convolutional neural network for modeling quantum interactions. arXiv:1706.08566 [stat.ML] (2017).

[120] Elliott Skomski, Jan Drgona, and Aaron Tuor. 2020. Physics-informed neural state space models via learning and evolution. arXiv:2011.13497 [cs.NE] (2020).

[121] Elliott Skomski, Soumya Vasisht, Colby Wight, Aaron Tuor, Jan Drgona, and Draguna Vrabie. 2021. Constrained block nonlinear neural dynamical models. arXiv:2101.01864 [math.DS] (2021).

[122] B. Sohlberg and E. W. Jacobsen. 2008. Grey box modelling—Branches and experiences. *IFAC Proceedings Volumes* 41, 2 (2008), 11415–11420. https://doi.org/10.3182/20080706-5-KR-1001.01934

[123] Heung-Il Suk, Chong-Yaw Wee, Seong-Whan Lee, and Dinggang Shen. 2016. State-space model with deep learning for functional dynamics estimation in resting-state fMRI. *NeuroImage* 129 (2016), 292–307. https://doi.org/10.1016/j.neuroimage.2016.01.005

[124] Peter Toth, Danilo Jimenez Rezende, Andrew Jaegle, Sébastien Racanière, Aleksandar Botev, and Irina Higgins. 2020. Hamiltonian generative networks. arXiv:1909.13789 [cs, stat] (Feb. 2020).

[125] Oliver T. Unke and Markus Meuwly. 2018. A reactive, scalable, and transferable model for molecular energies from a neural network approach based on local information. *Journal of Chemical Physics* 148, 24 (2018), 241708. https://doi.org/10.1063/1.5017898

[126] Oliver T. Unke and Markus Meuwly. 2019. PhysNet: A neural network for predicting energies, forces, dipole moments, and partial charges. *Journal of Chemical Theory and Computation* 15, 6 (2019), 3678–3693. https://doi.org/10.1021/acs.jctc.9b00181

[127] Felipe A. C. Viana, Christian Gogu, and Raphael T. Haftka. 2010. Making the most out of surrogate models: Tricks of the trade. In *Volume 1: 36th Design Automation Conference, Parts A and B*. ASMEDC, Montreal, Quebec, Canada, 587–598. https://doi.org/10.1115/DETC2010-28813

[128] Laura von Rueden, Sebastian Mayer, Katharina Beckh, Bogdan Georgiev, Sven Giesselbach, Raoul Heese, Birgit Kirsch, et al. 2020. Informed machine learning—A taxonomy and survey of integrating knowledge into learning systems. arXiv:1903.12394 [cs, stat] (Feb. 2020).

[129] Laura von Rueden, Sebastian Mayer, Rafet Sifa, Christian Bauckhage, and Jochen Garcke. 2020. Combining machine learning and simulation to a hybrid modelling approach: Current and future directions. *Advances in Intelligent Data Analysis XVIII* 12080 (2020), 548–560. https://doi.org/10.1007/978-3-030-44584-3_43

[130] Jiang Wang, Simon Olsson, Christoph Wehmeyer, Adrià Pérez, Nicholas E. Charron, Gianni de Fabritiis, Frank Noé, and Cecilia Clementi. 2019. Machine learning of coarse-grained molecular dynamics force fields. *ACS Central Science* 5, 5 (2019), 755–767. https://doi.org/10.1021/acscentsci.8b00913

[131] Sifan Wang, Yujun Teng, and Paris Perdikaris. 2020. Understanding and mitigating gradient pathologies in physics-informed neural networks. arXiv:2001.04536 [cs, math, stat] (Jan. 2020).

[132] Sifan Wang, Xinling Yu, and Paris Perdikaris. 2020. When and why PINNs fail to train: A neural tangent kernel perspective. arXiv:2007.14527 [cs, math, stat] (July 2020).

[133] G. Wanner and E. Hairer. 1991. *Solving Ordinary Differential Equations I: Nonstiff Problems*, Vol. 1. Springer-Verlag.

[134] Nicholas Watters, Daniel Zoran, Theophane Weber, Peter Battaglia, Razvan Pascanu, and Andrea Tacchetti. 2017. Visual interaction networks: Learning a physics simulator from video. In *Advances in Neural Information Processing Systems*, I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett (Eds.), Vol. 30. Curran Associates, 4542–4550.

[135] Paul Westermann and Ralph Evins. 2019. Surrogate modelling for sustainable building design—A review. *Energy and Buildings* 198 (Sept. 2019), 170–186. https://doi.org/10.1016/j.enbuild.2019.05.057

[136] Hao Wu, Andreas Mardt, Luca Pasquali, and Frank Noe. 2018. Deep generative Markov state models. arXiv preprint arXiv:1805.07601 (2018).

[137] Zonghan Wu, Shirui Pan, Fengwen Chen, Guodong Long, Chengqi Zhang, and Philip S. Yu. 2019. A comprehensive survey on graph neural networks. *CoRR* abs/1901.00596 (2019).

[138] Winnie Xu, Ricky T. Q. Chen, Xuechen Li, and David Duvenaud. 2021. Infinitely deep Bayesian neural networks with stochastic differential equations. arXiv:2102.06559 [cs, stat] (Aug. 2021).

[139] Linfeng Zhang, Jiequn Han, Han Wang, Roberto Car, and Weinan E. 2018. Deep potential molecular dynamics: A scalable model with the accuracy of quantum mechanics. *Physical Review Letters* 120, 14 (April 2018), 143001. https://doi.org/10.1103/PhysRevLett.120.143001

[140] Linfeng Zhang, Jiequn Han, Han Wang, Wissam A. Saidi, Roberto Car, and Weinan E. 2018. End-to-end symmetry preserving inter-atomic potential energy model for finite and extended systems. arXiv:1805.09003 [physics.comp-ph] (2018).

[141] Ziwei Zhang, Peng Cui, and Wenwu Zhu. 2018. Deep learning on graphs: A survey. *CoRR* abs/1812.04202 (2018).

[142] Yaofeng Desmond Zhong, Biswadip Dey, and Amit Chakraborty. 2019. Symplectic ODE-Net: Learning Hamiltonian dynamics with control. *CoRR* abs/1909.12077 (2019).

[143] Jie Zhou, Ganqu Cui, Shengding Hu, Zhengyan Zhang, Cheng Yang, Zhiyuan Liu, Lifeng Wang, Changcheng Li, and Maosong Sun. 2020. Graph neural networks: A review of methods and applications. *AI Open* 1 (2020), 57–81. https://doi.org/10.1016/j.aiopen.2021.01.001

Received 2 November 2021; revised 23 September 2022; accepted 4 October 2022