

# LQResNet: A Deep Neural Network Architecture for Learning Dynamic Processes

Pawan Goyal\*    Peter Benner†

\*Max Planck Institute for Dynamics of Complex Technical Systems, 39106 Magdeburg, Germany.

Email: [goyalp@mpi-magdeburg.mpg.de](mailto:goyalp@mpi-magdeburg.mpg.de),

†Max Planck Institute for Dynamics of Complex Technical Systems, 39106 Magdeburg, Germany.

Email: [benner@mpi-magdeburg.mpg.de](mailto:benner@mpi-magdeburg.mpg.de),

Mathematical modeling is an essential step, for example, to analyze the transient behavior of a dynamical process and to perform engineering studies such as optimization and control. With the help of first-principles and expert knowledge, a dynamic model can be built, but for complex dynamic processes, appearing, e.g., in biology, chemical plants, neuroscience, financial markets, this often remains an onerous task. Hence, data-driven modeling of the dynamics process becomes an attractive choice and is supported by the rapid advancement in sensor and measurement technology. A data-driven approach, namely operator inference framework, models a dynamic process, where a particular structure of the nonlinear term is assumed. In this work, we suggest combining the operator inference with certain deep neural network approaches to infer the unknown nonlinear dynamics of the system. The approach uses recent advancements in deep learning and possible prior knowledge of the process if possible. We also briefly discuss several extensions and advantages of the proposed methodology. We demonstrate that the proposed methodology accomplishes the desired tasks for dynamics processes encountered in neural dynamics and the glycolytic oscillator.

**Keywords:** Artificial intelligence, machine learning, deep learning, dynamical systems, scientific machine learning

## Introduction

With the rapid development in sensor and measurement technology, time-series data of processes have become available in large amounts with high accuracy. Machine learning and data science play an important role in analyzing and perceiving information of the underlying process dynamics from these data. Building a model describing the dynamics is vital in designing and optimizing various processes, as well as predicting their long-term transient behavior. Inferring a dynamic process model from data, often called system identification, has a rich history; see, e.g., [30, 45]. While linear system identification is well established, nonlinear system identification is still far from being as good understood as for linear systems, despite having a similarly long research history, see, e.g., [25, 43]. Nonlinear system identification often relies on a good hypothesis of the model; thus, it is not entirely a black-box technology. Fortunately, there are several scenarios where one can hypothesize a model structure based on a good understanding of the underlying dynamic behavior using expert knowledge or experience. Towards nonlinear system identification, a promising approach based on a symbolic regression was

proposed [4] to determine the potential structure of a nonlinear system. The method discovers a dynamical system solely from data. However, the approach is computationally expensive; thus, it is not suitable for a large-scale dynamical system. Also, one needs to take care of the overfitting by using the Pareto front that balances model complexity, and data fitting [4]. If nonlinearities are only known to belong to a certain class of mathematical functions (a *dictionary* of functions), sparse regression (see, e.g., [44]) and compressive sensing (see, e.g., [7, 13]) based approaches have emerged as a kind of dictionary learning methods for nonlinear models, including dynamical systems. Here, it is observed that given a high-dimensional nonlinear function space, the process dynamics can often be accurately described by only a few terms from the dictionary [6, 33, 35, 38, 40, 46]. Consequently, one obtains an interpretable process model. However, the success of these approaches highly depends on the quality of the constructed dictionary, i.e., the candidate nonlinear function space [6]. Thus, dictionaries need to be generated carefully using expert knowledge. We mention that there exist other techniques to infer a model of a dynamic process, for instance from time-series data [10, 21], equation-free modeling [22, 48], dynamical models inference [11, 12, 42]. In the category of equation-free modeling, dynamical mode decomposition [37, 41, 47] has also shown promising results. Moreover, deep learning (DL) approaches have been developed successfully for this task. They have the potential to build a model using only the time-history of dependent variables. For this, a deep neural network (DNN) is typically used which can be understood as a composition of functions since the output of a layer is the input of the next layer. DNNs have shown phenomenal performance across various disciplines [18, 24, 27] (e.g., image classification, speech recognition, medical image analysis, to name a few). A primary reason for DNNs' success is that each layer learns a particular feature representation of the input, thus describing a complex function representation. Among several existing DNN architectures, *Residual Neural Networks* (ResNet) [19] closely resemble an Euler-type integration scheme for ordinary differential equations (ODEs). Inspired by this, there are methodological advances for inferring ODEs from data [8]. In recent years, scientific machine learning has emerged as a discipline that combines classical mechanistic modeling and numerical simulation with machine learning techniques. The expectation here is that incorporating existing scientific knowledge allows the (physical) interpretability of models while requiring less training data. Here, it is important to note that engineering processes often do not yield the same amount of data as typical socio-economic big data applications, so overfitting is often encountered. This can be attenuated by including physical constraints during learning, thereby reducing the network parameters, or providing a better-suited network design. Operator inference approaches fall into this category, see, e.g., [2, 34]. Furthermore, physics-informed neural networks (PINN) [36] also employ the existing physical knowledge to improve the learning process that, as an output, yields the solution of a partial differential equation. However, as of yet, PINNs assume the structure of the physical model to be fully known, which is not available in a complex dynamic process.

In the following section, we will propose a DNN architecture dedicated to learning dynamical systems from data while incorporating certain structural assumptions implied by the underlying physics.

## LQResNet: A deep network architecture for learning nonlinear dynamics

In this work, we focus on identifying a dynamical model by using not only collected data in simulations or experiments but also available expert knowledge and physical laws about the process. In general, we aim at inferring a nonlinear dynamical system of the form:

$$\dot{\mathbf{x}}(t) = \mathbf{g}(\mathbf{x}(t)), \quad (1)$$

where the vector  $\mathbf{x}(t) \in \mathbb{R}^n$  denotes the state at time  $t$ , and the function  $\mathbf{g}(\cdot) : \mathbb{R}^n \rightarrow \mathbb{R}^n$  is a continuous function, describing the dynamics of the system. We shall later on extend the dynamical system in Eq. 1 with control and parameters, and its discrete version. In principle, we are interested in determining the function  $\mathbf{g}$  from data. To that aim, we observe time-evaluation of the state vector  $\mathbf{x}(t)$  and the

derivative  $\dot{\mathbf{x}}(t)$  (if possible); otherwise, we approximate the derivative information using the state vector  $\mathbf{x}(t)$  by employing, for example, a five-point stencil method. DL approaches have been used to learn dynamic models for decades [17, 32]. Precisely, they are used to learn the mapping  $\mathbf{x}(t) \rightarrow \dot{\mathbf{x}}(t)$ . To that aim, typically, fully connected DNNs can be utilized as shown in Fig. 1(a); however, they share a common drawback – that is, they are hard to train as the network goes deeper. This is primarily due to gradient vanishing issues that commonly occur in such networks. Consequently, there is little or no updates in the network parameters, yet the network being far from representing the input-output mapping. The gradient vanishing problem can be solved to some extent by a ResNet-type architecture. Fundamentally, in ResNets, the output of a layer is passed not only to the next layer but also directly to deeper layers by using skip connections. This is illustrated in Fig. 1(b). Furthermore, in the direction of scientific machine learning, it is also important to make use of the additional knowledge which may be available from experts or underlying physical laws. Towards this, operator inference (OpInf) approaches have emerged as potential ones [1, 2, 34]. A key observation is that the rate of change of the dependent variable  $\mathbf{x}$  strongly depends linearly or quadratically on the state  $\mathbf{x}$ . This observation is often encountered in engineering and biology problems, e.g., flow problems, Fisher’s equation for gene propagation [14], or FitzHugh-Nagumo [15] describing neural dynamics. Moreover, very recently, the authors in [31] have shown that considering only linear and quadratic dependencies between variables, the dynamics of a single-injector combustion process can be described quite accurately. In this case, one learns dynamical systems of the form:

$$\dot{\mathbf{x}}(t) = \mathbf{A}\mathbf{x}(t) + \mathbf{Q}(\mathbf{x}(t) \otimes \mathbf{x}(t)) + \mathbf{b}, \quad (2)$$

which is typically considered in OpInf [34], and ‘ $\otimes$ ’ denotes the Kronecker product [20], i.e.,

$$\mathbf{x}(t) \otimes \mathbf{x}(t) = [\mathbf{x}_1^2(t), \mathbf{x}_1(t)\mathbf{x}_2(t), \dots, \mathbf{x}_1(t)\mathbf{x}_n(t), \dots, \mathbf{x}_n^2(t)],$$

where  $\mathbf{x}_i(t)$  denotes the  $i$ th component of the vector  $\mathbf{x}(t)$ . Furthermore, the system in Eq. 2 can also be obtained by considering only terms up-to second order of the Taylor series expansion of  $\mathbf{g}(\mathbf{x}(t))$  in Eq. 1, i.e.,

$$\mathbf{g}(\mathbf{x}(t)) = \mathbf{A}\mathbf{x}(t) + \mathbf{Q}(\mathbf{x}(t) \otimes \mathbf{x}(t)) + \mathbf{b} + \mathcal{O}(\mathbf{x}(t)^3). \quad (3)$$

If  $\mathbf{x}(t)$  is small, then the higher-order terms are very small, thus can be neglected. However, if this is not the case, then the higher-order terms would contribute significantly. Hence, in this paper, we consider a model of a dynamical system of the form:

$$\dot{\mathbf{x}}(t) = \mathbf{A}\mathbf{x}(t) + \mathbf{Q}(\mathbf{x}(t) \otimes \mathbf{x}(t)) + \mathbf{f}(\mathbf{x}), \quad \mathbf{x}(0) = \mathbf{x}_0. \quad (4)$$

To learn a model, having the form as in Eq. 4, we propose the architecture shown in Fig. 1(c). The proposed architecture resembles a ResNet, where linear and quadratic connections are directly passed to the output layer, and  $\mathbf{f}(\mathbf{x})$  is also a ResNet in itself. We refer to the architecture as *Linear-Quadratic-Residual Network* (LQResNet). Naturally, we can feed directly  $\mathbf{x}$ , containing all involved independent variables, to learn a function  $\mathbf{g}(\mathbf{x}) := \mathbf{A}\mathbf{x}(t) + \mathbf{Q}(\mathbf{x}(t) \otimes \mathbf{x}(t)) + \mathbf{f}(\mathbf{x})$  in the proposed architecture; however, we empirically make important observations – these are:

- For  $\mathbf{x}_i(t)$  representing different quantities (e.g., concentrations of species in biology/chemical processes), it is efficient to build a network for each component, precisely the mapping  $\mathbf{x}(t) \rightarrow \dot{\mathbf{x}}_i(t)$ , instead of a bigger network, learning the mapping  $\mathbf{x}(t) \rightarrow \dot{\mathbf{x}}(t)$ . A reason behind this is that a deep network aims at learning the features describing the mapping. Typically, we may require very different features to learn the mapping  $\mathbf{x}(t) \rightarrow \dot{\mathbf{x}}_i(t)$ ,  $i \in \{1, \dots, n\}$ . On the other hand, if we learn the mapping  $\mathbf{x}(t) \rightarrow \dot{\mathbf{x}}(t)$  at once, then the network needs to learn many complex features at once, which can require a larger network, and we may require a big training data set and longer training time.
- Moreover, the mappings  $\mathbf{x}(t) \rightarrow \dot{\mathbf{x}}_i(t)$  can be of different complexities; it means that the number of features that need to be learned by a network to describe the mapping can be different. As a result, some of the networks can be smaller/larger than others.

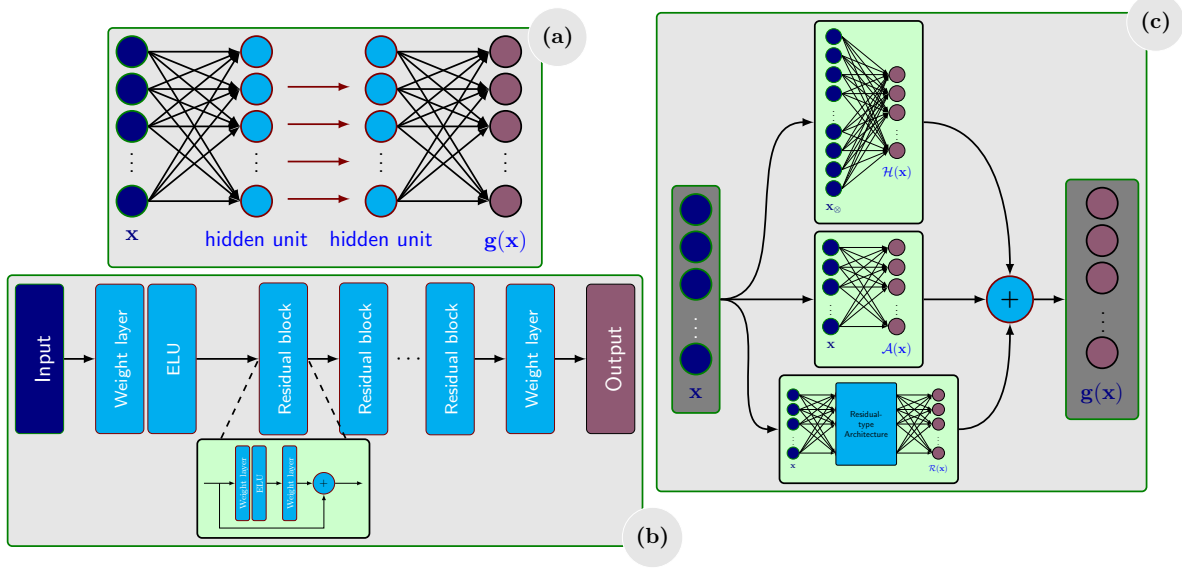


Figure 1: The figure explains different architecture designs of neural networks that essentially aim at mapping the input  $\mathbf{x}$  to  $\mathbf{g}(\mathbf{x})(= \dot{\mathbf{x}})$ . Sub-figure (a) shows a fully connected deep neural network with  $l$  hidden layers; sub-figure (b) illustrates a residual neural network with skip connections that also aims at mapping  $\mathbf{x}$  to  $\mathbf{g}(\mathbf{x})$ ; sub-figure (c) describes the proposed new architecture that, in addition to a residual neural network, has linear and quadratic mappings via skip connections.

- Moreover, if several small networks are built for each mapping, then we can easily parallelize the forward propagation to predict the output  $\dot{\mathbf{x}}(t)$  for a given input  $\mathbf{x}(t)$ .

Thus, we train a network, describing the mapping  $\mathbf{x}(t) \rightarrow \dot{\mathbf{x}}(t)$  by incorporating the linear and quadratic terms. Hence, we aim at solving the following optimization problem:

$$\min_{\mathbf{A}_j, \mathbf{Q}_j, \Theta_i} \sum_{j=1}^{\mathcal{N}} \|\dot{\mathbf{x}}_i^{(j)} - \mathbf{A}_j \mathbf{x}^{(j)} - \mathbf{Q}_j (\mathbf{x}^{(j)} \otimes \mathbf{x}^{(j)}) - \mathcal{R}_{\Theta_i}(\mathbf{x}^{(j)})\|, \quad (5)$$

where  $\mathbf{x}^{(j)}$  is a state vector at a time instance, and  $\dot{\mathbf{x}}^{(j)}$  denotes its derivative with respect to time.  $\mathcal{N}$  is the total number of training data, and  $\mathcal{R}_{\Theta_i}$  is a ResNet (learning  $\mathbf{f}_i(\mathbf{x})$  in Eq. 4 and  $\mathbf{f}_i$  denoting its  $i$ th entry), parameterized by  $\Theta_i$ .

Note that the optimization problem in Eq. 5 is high-dimensional and non-convex; hence, it is a hard problem to solve in theory. Moreover, it may have many local minima. However, as illustrated in [28], neural networks with skip connections such as the proposed one tend to find a global minimum (or get close to it) even with simple stochastic gradient methods. Moreover, such a network does not suffer from the gradient vanishing problem as well in training.

Furthermore, in the following, we state our empirical finding. To solve Eq. 5, one may also think of decoupling the optimization problem as follows. In the first step, one can aim at identifying the linear and quadratic matrices by solving the following optimization problem:

$$\min_{\mathbf{A}_j, \mathbf{Q}_j} \sum_{j=1}^{\mathcal{N}} \|\dot{\mathbf{x}}_i^{(j)} - \mathbf{A}_j \mathbf{x}^{(j)} - \mathbf{Q}_j (\mathbf{x}^{(j)} \otimes \mathbf{x}^{(j)})\|, \quad (6)$$

Once we obtain  $\mathbf{A}_i$  and  $\mathbf{Q}_i$ , we can define the residual as follows:

$$\mathbf{r}_i^{(j)}(\mathbf{x}) = \dot{\mathbf{x}}_i^{(j)} - \mathbf{A}_i \mathbf{x}^{(j)} - \mathbf{Q}_i (\mathbf{x}^{(j)} \otimes \mathbf{x}^{(j)}), \quad (7)$$

which can then be learned using a ResNet. But in our numerical experimental studies, we observe that this yields a poor model, and it seems to be harder for the network to learn the underlying dynamics.

Therefore, we suggest to optimize simultaneously the matrices  $\mathbf{A}_j$ ,  $\mathbf{Q}_j$ , and the parameters for the ResNet  $\mathcal{R}_{\Theta_i}$  that solves Eq. 5. However, the quantity  $\mathbf{r}_i^{(j)}(\mathbf{x})$  can indicate how well the dynamics can be captured only by linear and quadratic terms. Hence, if  $\mathbf{r}_i^{(j)}(\mathbf{x})$  is smaller than a threshold, then we do not need to train a network, and  $\mathbf{A}_j$  and  $\mathbf{Q}_j$  can be given analytically.

## Interpretation of the LQResNet

In what follows, we discuss an interpretation of the LQResNet. For this, we first note that the residual block can be seen as composite functions, i.e.,

$$\mathbf{W}(\mathbf{h}_l(\mathbf{h}_{l-1} \cdots \mathbf{h}_2(\mathbf{h}_1(\mathbf{x})))) + \mathbf{b},$$

where  $\mathbf{h}_i(\mathbf{x}) := \mathbf{x} + \mathbf{W}_i^{(2)}\sigma(\mathbf{W}_i^{(1)}\mathbf{x} + \mathbf{b}^{(1)}) + \mathbf{b}^{(2)}$ , and  $\mathbf{W}$ ,  $\mathbf{W}_i^{(1)}$ ,  $\mathbf{W}_i^{(2)}$  are weights,  $\mathbf{b}$ ,  $\mathbf{b}_i^{(1)}$ ,  $\mathbf{b}_i^{(2)}$  are the biases, and  $\sigma(\cdot)$  is a nonlinear activation function. If we closely look at the network, then we notice that we essentially learn the mapping  $\mathbf{x}(t) \rightarrow \hat{\mathbf{x}}_i(t)$ , and the mapping takes the form:

$$\mathbf{A}_i(\mathbf{x} + \Xi(\mathbf{x})) + \mathbf{Q}_i(\mathbf{x} \otimes \mathbf{x}). \quad (8)$$

Hence, we can think of learning a perturbation to the linear mapping such that it can describe the mapping  $\mathbf{x}(t) \rightarrow \hat{\mathbf{x}}_i(t)$ . If  $\Xi(i)$  is only a small perturbation, then it can be learned using a small network; otherwise, we need a bigger one. However, we shall later on discuss that the network can adaptively be enlarged in the course of the training process if the mapping is not sufficiently accurate. This is possible due to the residual blocks in the network that increasingly learns complex features.

In many scenarios, the dynamics of a physical system are governed by partial differential equations. As a result, if data are collected using a black-box simulator or in an experimental set-up on a spatial grid, then the state vector  $\mathbf{x}$  lies in high-dimensional space. For example, consider the Cahn-Hilliard equation that describes the dynamics of the phase separation in binary alloys. To capture the dynamics accurately even in a 2-dimensional space, several thousands of variables may be required. As the state vector dimension increases, the optimization problem (Eq. 5) becomes computationally expensive as the number of parameters in the optimization problem is of  $\mathcal{O}(n^2)$ , where  $n$  is the dimension of the state vector. However, to ease the problem, we can use the fact that many high-dimensional dynamical systems evolve in a low-dimensional manifold. Thus, the state vector  $\mathbf{x}$  can be well-approximated using  $\hat{n}$  basis functions which can be determined using dimensional reduction techniques. A widely used technique to determine such a basis is proper orthogonal decomposition [3, 26]. In principle, the high-dimensional state vector  $\mathbf{x}(t)$  can be approximated as

$$\mathbf{x}(t) \approx \mathbf{V}\hat{\mathbf{x}}(t), \quad (9)$$

where  $\mathbf{V} \in \mathbb{R}^{n \times \hat{n}}$ ,  $\mathbf{x}(t) \in \mathbb{R}^n$ , and  $\mathbf{V}$  is constructed using left singular vectors of the collected measurements of  $\mathbf{x}(t)$ , corresponding to  $\hat{n}$  leading singular values. As a result, we can rather learn a dynamical system for the low-dimensional variable  $\hat{\mathbf{x}}(t)$ :

$$\dot{\hat{\mathbf{x}}}(t) = \hat{\mathbf{A}}\hat{\mathbf{x}}(t) + \hat{\mathbf{Q}}(\hat{\mathbf{x}}(t) \otimes \hat{\mathbf{x}}(t)) + \hat{\mathbf{f}}(\hat{\mathbf{x}}(t)) \quad (10)$$

where  $\hat{\mathbf{x}}(t)$ . For this, we would require data  $\hat{\mathbf{x}}(t_i)$  and  $\dot{\hat{\mathbf{x}}}(t_i)$  which can be obtained by projecting the high-dimensional  $\mathbf{x}(t_i)$  and  $\dot{\mathbf{x}}(t_i)$  using the projection matrix  $\mathbf{V}$ , i.e.,  $\hat{\mathbf{x}}(t_i) = \mathbf{V}^\top \mathbf{x}(t_i)$  and  $\dot{\hat{\mathbf{x}}}(t_i) = \mathbf{V}^\top \dot{\mathbf{x}}(t_i)$ . Furthermore, when  $\mathbf{x}(t)$  denotes variables, obtained from a discretization of a partial differential equation, all these variables are alike, meaning they are of a similar complexity. Hence, we can directly learn the mapping  $\hat{\mathbf{x}}(t) \rightarrow \dot{\hat{\mathbf{x}}}(t)$  using a single network with linear and quadratic skip connections to the last layer (as depicted in Fig. 1(c)). Moreover, if the data represents different types of variables – it happens when dynamics is governed by a couple of partial differential equations (e.g., shallow water equations [5]) – then we train a network to learn the mapping for each class of variables.

## Possible extensions

Here, we briefly discuss some possible extensions of the above methodology to learn dynamical models.

### Additional prior knowledge

There are various engineering applications where we have more insight into the process that may be known by experts or from first principles, for example, an interconnected topological network in biological species, chemical kinetics reactions, flow problems, and mechanical instruments. We illustrate this with two examples. In the first example, consider the data describing the motion of an inverted pendulum is given. In this case, the first-principle information of a motion of an inverted pendulum can be used – these are, e.g., that the dynamics are of second-order in time and typically contain a sine function, i.e.,

$$\ddot{\mathbf{x}}(t) = \mathbf{a}\dot{\mathbf{x}}(t) + \mathbf{b}\sin(\mathbf{x}(t)), \quad (11)$$

where  $\mathbf{a}, \mathbf{b}$  are constants. However, in practice, there is some additional mechanical friction that is hard to describe analytically. Thus, one may aim at learning the additional friction factor using a residual network and directly add a sine connection to the output layer to make the learning more efficient. Precisely, we can define an optimization problem as follows:

$$\min_{\mathbf{a}, \mathbf{b}, \Theta} \sum_{j=1}^{\mathcal{N}} \|\ddot{\mathbf{x}}^{(j)} - \mathbf{a}\dot{\mathbf{x}}^{(j)} - \mathbf{b}\sin(\mathbf{x}^{(j)}) - \mathcal{R}_{\Theta}(\dot{\mathbf{x}}^{(j)}, x^{(j)})\|, \quad (12)$$

where  $\mathcal{R}_{\Theta}$  is a ResNet, parameterized by  $\theta$ . Having set up this, we can expect a far better model with a smaller amount of data for an inverted pendulum that describes the dynamics as close to reality as possible. In the second example, let us assume the data represents incompressible flow dynamics and denote the velocity field  $\mathbf{v}(t)$  at time  $t$ . If we aim at learning a model describing the dynamics of the flow, we need to ensure that the model enforces the incompressibility condition. For this, we project the velocity field  $\mathbf{v}(t)$  on a manifold, where the condition is met. We can find such a manifold using the singular value decomposition of the data matrix containing the velocity field, i.e.,  $\hat{\mathbf{v}}(t) = \mathbf{V}^{\top} \mathbf{v}(t)$ . Then, we build an LQResNet, describing the dynamics of  $\hat{\mathbf{v}}(t)$  and the velocity field  $\mathbf{v}(t) = \mathbf{V}\hat{\mathbf{v}}(t)$  which will inherently satisfy the mass-conservation law.

### Discrete models

One may also be interested in learning a discrete model which takes the form:

$$\mathbf{x}(k+1) = \mathbf{A}\mathbf{x}(k) + \mathbf{Q}(\mathbf{x}(k) \otimes \mathbf{x}(k)) + \mathbf{G}(\mathbf{x}(k)), \quad (13)$$

where  $\mathbf{x}(k) \in \mathbb{R}^n$  is the state  $\mathbf{x}$  at the  $k$ th time step;  $\mathbf{A} \in \mathbb{R}^{n \times n}$ ,  $\mathbf{Q} \in \mathbb{R}^{n \times n^2}$ , and the nonlinear function  $\mathbf{G}(\cdot) : \mathbb{R}^n \rightarrow \mathbb{R}^n$ . In this case, we can readily apply the approach discussed above that takes the state  $\mathbf{x}$  at time step  $k$  as an input and yields the state at the next time step as an output. An advantage of the approach is that we do not require the derivative information of the state  $\mathbf{x}$ . The derivative information can be hard to estimate when the noise in the state is above a threshold level. However, we note that a technique described in [39] can be adapted that can handle higher-level noises to learn dynamics.

### Controlled and parametric case

Assume that the dynamic process is subject to an external input and involves parameter involved. In this case, inspired by our earlier discussion, we aim at learning a dynamical system of the form:

$$\begin{aligned} \dot{\mathbf{x}}(t, \mu) = & (\mathbf{A} + \mu\mathbf{A}_{\mu}) \mathbf{x}(t, \mu) + (\mathbf{Q} + \mu\mathbf{Q}_{\mu}) (\mathbf{x}(t, \mu) \otimes \mathbf{x}(t, \mu)) \\ & + \mathbf{B}\mathbf{u}(t) + \mathbf{R}(\mathbf{x}(t, \mu), \mathbf{u}(t)), \end{aligned} \quad (14)$$



Problem	Epochs	Learning rate	Batch size	Weight decay
FHN	500	$5 \cdot 10^{-4}$	512	$10^{-4}$
GO	2 000	$10^{-3}$	512	$10^{-4}$

Table 1: The parameters used to train a neural network are listed.

where  $\mu \in \mathbb{R}$ ,  $\mathbf{u}(t) \in \mathbb{R}^m$  denote parameters and control input, respectively;  $\mathbf{x}(t, \mu) \in \mathbb{R}^n$  is the parameter dependent state, and  $\mathbf{R}(\mathbf{x}(t, \mu), \mathbf{u}(t)) : \mathbb{R}^{n+p} \rightarrow \mathbb{R}^n$ . In this case, we can adapt a similar architecture (shown in Fig. 1). Precisely, the adaption is as follows. We have a direct input connection to the output layer, and a residual network can have  $\mu$  and  $\mathbf{u}(t)$  along with  $\mathbf{x}(t, \mu)$  as inputs and the whole network is training simultaneously to predict  $\dot{\mathbf{x}}(t, \mu)$ .

## Demonstration of the Approach

We demonstrate the efficiency of the proposed approach to learn dynamical systems using two examples, namely the FitzHugh-Nagumo model and the Glycolytic oscillator. We collect data by simulating the governing equations using the Python routine `odeint` from `scipy.integrate`. Moreover, given data  $\mathbf{x}(t)$  at time instance  $\{t_0, \dots, t_n\}$ , we approximate  $\dot{\mathbf{x}}(t)$  using a five-point stencil. We have used an exponential linear unit as an activation function [9] that does not suffer from dying neuron problem as well as the activation function is  $C^1$  continuous. The weights of a neural network are optimized using a variant of the Adam method [23], namely rectified Adam [29] that utilizes a warm-up strategy for better convergence. Also, to avoid over-fitting, we regularize the optimization by the 2-norm of the weights involved in the network (often referred to as weight decay). We split the data set into the training and validation data sets in 80:20 ratio. Other important hyper-parameters used to train a network for both problems are listed in Table 1. All experiments were performed using PyTorch in Python running on a Macbook Pro with 2,3 GHz 8-Core Intel Core i9 CPU, 16GB of RAM, and Mac OS X v10.15.6.

### FitzHugh-Nagumo Model

As a first example, we consider the FitzHugh-Nagumo model [16] that describes spiking of a neuron, neuronal dynamical in a simplistic way. The governing equations of such a dynamical behavior are shown in Fig. 2. The variables  $v(t)$  and  $w(t)$  describe activation and de-activation of a neuron. Specifically, the model exhibits a periodic oscillatory behavior when  $I_{\text{ext}}$  exceeds a threshold value. We collect data by simulating the model using 10 randomly chosen initial conditions in the range  $[-1, 1] \times [-1, 1]$ , and for each initial condition, we collect 5 000 equidistant points in the time interval  $[0, 200]$ s. Having these data, we build two models that map  $\{v(t), w(t)\} \mapsto \dot{v}(t)$  and  $\{v(t), w(t)\} \mapsto \dot{w}(t)$ . Before building a deep neural network, we determine the residual as shown in Eq. 7. As a result, we notice that  $\dot{w}(t)$  can accurately be written in the linear and quadratic form of  $\{v(t), w(t)\}$ ; on the other hand, the linear and quadratic forms of  $\{v(t), w(t)\}$  are not sufficient to represent  $\dot{v}(t)$ . Hence, we make use of the residual type neural network to learn the mapping  $\{v(t), w(t)\} \mapsto \dot{v}(t)$  using two residual blocks and each hidden layer having ten neurons. Having trained the model, we test the learned model with the true model for an initial condition – it is neither a part of the training nor the validation set – to examine the predictive capacities of the model. We observe that the learned model replicates the dynamics of the true model very well, even for the unseen initial conditions (see Fig. 2).

### Glycolytic Oscillator

Our next example represents biochemical dynamics; precisely, it describes the dynamical behavior in yeast glycolysis, see, e.g., [12]. The dynamics are given by the following governing equations that

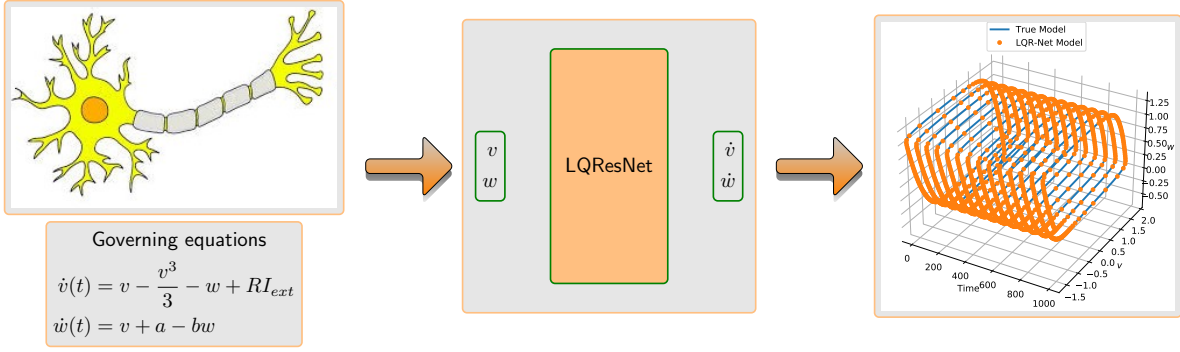


Figure 2: FitzHugh-Nagumo Model: governing equations, learning model, and comparison of the learned and true models. The left-most figure describing neurons is due to courtesy of [49].

predict the concentrations of 7 different biochemical species:

$$\dot{\mathbf{S}}_1 = \mathbf{J}_0 - \frac{k_1 \mathbf{S}_1 \mathbf{S}_6}{1 + (\mathbf{S}_6/k_1)^q} \quad (15a)$$

$$\dot{\mathbf{S}}_2 = 2 \frac{k_1 \mathbf{S}_1 \mathbf{S}_6}{1 + (\mathbf{S}_6/k_1)^q} - k_2 \mathbf{S}_2 (\mathbf{N} - \mathbf{S}_5) - k_6 \mathbf{S}_2 \mathbf{S}_5 \quad (15b)$$

$$\dot{\mathbf{S}}_3 = k_2 \mathbf{S}_2 (\mathbf{N} - \mathbf{S}_5) - k_3 \mathbf{S}_3 (\mathbf{A} - \mathbf{S}_6) \quad (15c)$$

$$\dot{\mathbf{S}}_4 = k_3 \mathbf{S}_3 (\mathbf{A} - \mathbf{S}_6) - k_4 \mathbf{S}_4 \mathbf{S}_5 - \kappa (\mathbf{S}_4 - \mathbf{S}_7) \quad (15d)$$

$$\dot{\mathbf{S}}_5 = k_2 \mathbf{S}_2 (\mathbf{N} - \mathbf{S}_5) - k_4 \mathbf{S}_4 \mathbf{S}_5 - k_6 \mathbf{S}_2 \mathbf{S}_5 \quad (15e)$$

$$\dot{\mathbf{S}}_6 = -2 \frac{k_1 \mathbf{S}_1 \mathbf{S}_6}{1 + (\mathbf{S}_6/k_1)^q} + 2k_3 \mathbf{S}_3 (\mathbf{A} - \mathbf{S}_6) - k_5 \mathbf{S}_6, \quad (15f)$$

$$\dot{\mathbf{S}}_7 = \psi \kappa (\mathbf{S}_4 - \mathbf{S}_7) - \kappa \mathbf{S}_7. \quad (15g)$$

We consider the same model parameters as given in [12, Tab. 1], and the range of initial conditions for all concentrations are taken the same as given [12, Tab. 2]. In order to collect data, we randomly take 30 different initial conditions in the given range and for each initial condition, we take 4 000 equidistant points in the time interval  $[0, 10]$ .

Clearly, the model shows complicated nonlinear dynamics; thus, it often cannot be known or guessed if only data are provided. Albeit complex dynamics, one may notice that the rates of change of concentrations heavily depend on linear and quadratic forms of the concentrations at a given time. Precisely, linear and quadratic forms of the concentrations at time  $t$  with appropriate coefficients can accurately express  $\{\dot{\mathbf{S}}_3(t), \dot{\mathbf{S}}_4(t), \dot{\mathbf{S}}_5(t), \dot{\mathbf{S}}_7(t)\}$ . Since  $\{\dot{\mathbf{S}}_1, \dot{\mathbf{S}}_2, \dot{\mathbf{S}}_6\}$  cannot be accurately given in the linear-quadratic form of the concentrations, we build a LQResNet type deep neural network to learn the mapping  $\{\mathbf{S}_1, \dots, \mathbf{S}_7\} \mapsto \dot{\mathbf{S}}_i, i \in \{1, 2, 6\}$ . For this, the network consists of 5 residual blocks and each hidden layer has 35 neurons. Having built a model, we test the accuracy of the learned model with the true models for an initial condition that has not been considered in the training or validation phase. It is compared in Fig. 3, indicating that the learned model can successfully replicate the dynamics without any prior knowledge of the biochemical process or model. Moreover, we note that the learning process can be even more improved if a topological network describing the interconnection of species is known. To make it clearer, from Eq. 15, we have that the dynamics of  $\mathbf{S}_1$  only dependent on  $\mathbf{S}_1, \mathbf{S}_6$ ; hence, we need to feed only these variables into a network to learn the dynamics of  $\mathbf{S}_1$ .



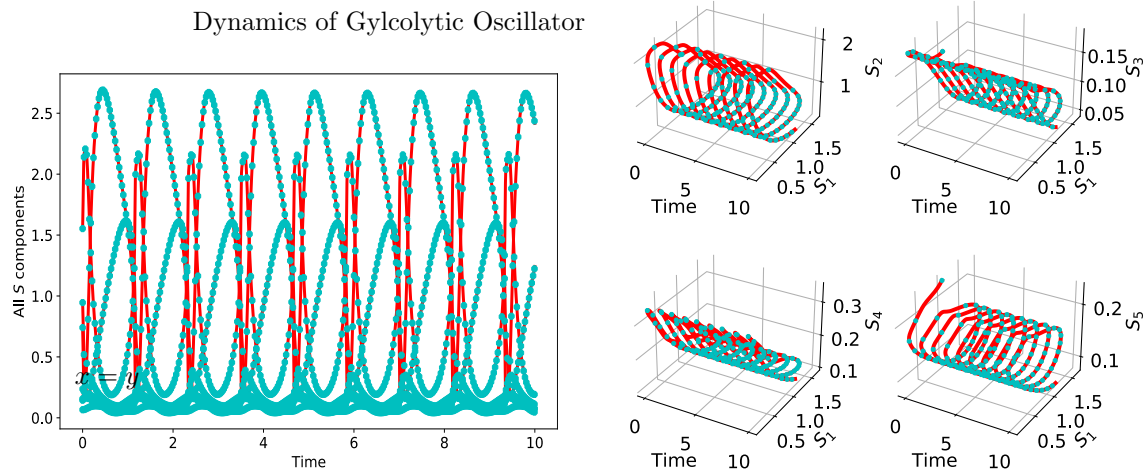


Figure 3: Gylcolytic Oscillator: A comparison of true and learned model for an initial condition that is not used for training. In red color (true model), and cyan dotted (learned model).

## Discussion

In essence, we have discussed a compelling approach to learning complex nonlinear dynamical systems using data that can incorporate any prior knowledge of a process. The approach, in particular, makes use of the observation, which is often found in dynamical systems' modeling – that is, the rate of change of a variable strongly depends on linear and quadratic forms of the variable. Thus, we have proposed an efficient deep learning architecture, LQResNet. We have demonstrated the efficiency of the proposed methodology to learning models using two examples, arising in biology and biochemistry. The approach can be employed in various fields, where abundant data can be obtained. However, underlying physics or models describing dynamics are not completely known, for instance, in biological process modeling, climate science, epidemiology, and financial markets. Moreover, there are various instances where we have a fairly good understanding of the process, which may be derived from physical laws or expert knowledge, but it fails to explain the data collection, for instance, in an experimental setup, due to some hidden dynamics or forces. Hence, the proposed methodology can be applied to learn a correction term using LQResNet, by combining the knowledge of the process as well as data.

## References

- [1] Peter Benner, Pawan Goyal, Jan Heiland, and Igor Pontes Duff. Operator inference and physics-informed learning of low-dimensional models for incompressible flows. *arXiv:2010.06701*, 2020.
- [2] Peter Benner, Pawan Goyal, Boris Kramer, Benjamin Peherstorfer, and Karen Willcox. Operator inference for non-intrusive model reduction of systems with non-polynomial nonlinear terms. *Comp. Meth. Appl. Mech. Eng.*, 372:113433, 2020.
- [3] Gal Berkooz, Philip Holmes, and John L Lumley. The proper orthogonal decomposition in the analysis of turbulent flows. *Annual Rev. Fluid Mech.*, 25(1):539–575, 1993.

- [4] Josh Bongard and Hod Lipson. Automated reverse engineering of nonlinear dynamical systems. *Proc. Nat. Acad. Sci. U.S.A.*, 104(24):9943–9948, 2007.
- [5] Didier Bresch. Shallow-water equations and related topics. In *Handbook of Differential Equations: Evolutionary Equations*, volume 5, pages 1–104. Elsevier, 2009.
- [6] Steven L Brunton, Joshua L Proctor, and J Nathan Kutz. Discovering governing equations from data by sparse identification of nonlinear dynamical systems. *Proc. Nat. Acad. Sci. U.S.A.*, 113(15):3932–3937, 2016.
- [7] Emmanuel J Candès, Justin Romberg, and Terence Tao. Robust uncertainty principles: Exact signal reconstruction from highly incomplete frequency information. *IEEE Trans. Inform. Theory*, 52(2):489–509, 2006.
- [8] Ricky TQ Chen, Yulia Rubanova, Jesse Bettencourt, and David K Duvenaud. Neural ordinary differential equations. In *Advances Neural Inform. Processing Sys.*, pages 6571–6583, 2018.
- [9] Djork-Arné Clevert, Thomas Unterthiner, and Sepp Hochreiter. Fast and accurate deep network learning by exponential linear units (ELUs). *arXiv preprint arXiv:1511.07289*, 2015.
- [10] James P Crutchfield and Bruce S McNamara. Equations of motion from a data series. *Complex Sys.*, 1(417-452):121, 1987.
- [11] Bryan C Daniels and Ilya Nemenman. Automated adaptive inference of phenomenological dynamical models. *Nature Comm.*, 6(1):1–8, 2015.
- [12] Bryan C Daniels and Ilya Nemenman. Efficient inference of parsimonious phenomenological models of cellular dynamics using S-systems and alternating regression. *PLoS One*, 10(3):e0119821, 2015.
- [13] David L Donoho. Compressed sensing. *IEEE Trans. Inform. Theory*, 52(4):1289–1306, 2006.
- [14] Ronald Aylmer Fisher. The wave of advance of advantageous genes. *Annals of Eugenics*, 7(4):355–369, 1937.
- [15] Richard FitzHugh. Mathematical models of threshold phenomena in the nerve membrane. *The Bulletin Math. Biophys.*, 17(4):257–278, 1955.
- [16] Richard FitzHugh. Impulses and physiological states in theoretical models of nerve membrane. *Biophysical J.*, 1(6):445–466, 1961.
- [17] R Gonzalez-Garcia, R Rico-Martinez, and IG Kevrekidis. Identification of distributed parameter systems: A neural net based approach. *Computers & Chemical Engrg.*, 22:S965–S968, 1998.
- [18] Ian Goodfellow, Yoshua Bengio, and Aaron Courville. *Deep learning*. MIT press, 2016.
- [19] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proc. IEEE Conf. Comp. Vision Patt. Recog.*, pages 770–778, 2016.
- [20] Roger A Horn, Roger A Horn, and Charles R Johnson. *Topics in matrix analysis*. Cambridge university press, 1994.
- [21] Holger Kantz and Thomas Schreiber. *Nonlinear Time Series Analysis*, volume 7. Cambridge University Press, 2004.
- [22] Ioannis G Kevrekidis, C William Gear, James M Hyman, Panagiotis G Kevrekidis, Olof Runborg, Constantinos Theodoropoulos, et al. Equation-free, coarse-grained multiscale computation: Enabling microscopic simulators to perform system-level analysis. *Comm. Math. Sci.*, 1(4):715–762, 2003.

- [23] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.
- [24] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. In *Adv. Neural Inform. Processing Sys.*, pages 1097–1105, 2012.
- [25] S Narendra Kumpati and Parthasarathy Kannan. Identification and control of dynamical systems using neural networks. *IEEE Trans. Neural Networks*, 1(1):4–27, 1990.
- [26] Karl Kunisch and Stefan Volkwein. Galerkin proper orthogonal decomposition methods for a general equation in fluid dynamics. *SIAM J. Numer. Anal.*, 40(2):492–515, 2002.
- [27] Yann LeCun, Yoshua Bengio, and Geoffrey Hinton. Deep learning. *Nature*, 521(7553):436–444, 2015.
- [28] Hao Li, Zheng Xu, Gavin Taylor, Christoph Studer, and Tom Goldstein. Visualizing the loss landscape of neural nets. In *Neural Inform. Processing Syst.*, 2018.
- [29] Liyuan Liu, Haoming Jiang, Pengcheng He, Weizhu Chen, Xiaodong Liu, Jianfeng Gao, and Jiawei Han. On the variance of the adaptive learning rate and beyond. *arXiv preprint arXiv:1908.03265*, 2019.
- [30] Lennart Ljung. *System Identification: Theory for the User*. Prentice Hall, NJ, 1999.
- [31] Shane A McQuarrie, Cheng Huang, and Karen Willcox. Data-driven reduced-order models via regularized operator inference for a single-injector combustion process. *arXiv preprint arXiv:2008.02862*, 2020.
- [32] Michele Milano and Petros Koumoutsakos. Neural network modeling for near wall turbulent flow. *J. Comput. Phys.*, 182(1):1–26, 2002.
- [33] Vidvuds Ozoliņš, Rongjie Lai, Russel Caffisch, and Stanley Osher. Compressed modes for variational problems in mathematics and physics. *Proc. Nat. Acad. Sci. U.S.A.*, 110(46):18368–18373, 2013.
- [34] Benjamin Peherstorfer and Karen Willcox. Data-driven operator inference for nonintrusive projection-based model reduction. *Comp. Meth. Appl. Mech. Eng.*, 306:196–215, 2016.
- [35] Joshua L Proctor, Steven L Brunton, Bingni W Brunton, and JN Kutz. Exploiting sparsity and equation-free architectures in complex systems. *Europ. Phy. J. Spec. Top.*, 223(13):2665–2684, 2014.
- [36] Maziar Raissi, Paris Perdikaris, and George E Karniadakis. Physics-informed neural networks: A deep learning framework for solving forward and inverse problems involving nonlinear partial differential equations. *J. Comput. Phys.*, 378:686–707, 2019.
- [37] Clarence W Rowley, Igor Mezić, Shervin Bagheri, Philipp Schlatter, and Dans Henningson. Spectral analysis of nonlinear flows. *J. Fluid Mech.*, 641(1):115–127, 2009.
- [38] Samuel H Rudy, Steven L Brunton, Joshua L Proctor, and J Nathan Kutz. Data-driven discovery of partial differential equations. *Sci. Adv.*, 3(4):e1602614, 2017.
- [39] Samuel H Rudy, J Nathan Kutz, and Steven L Brunton. Deep learning of dynamics and signal-noise decomposition with time-stepping constraints. *J. Comput. Phys.*, 396:483–506, 2019.
- [40] Hayden Schaeffer, Russel Caffisch, Cory D Hauck, and Stanley Osher. Sparse dynamics for partial differential equations. *Proc. Nat. Acad. Sci. U.S.A.*, 110(17):6634–6639, 2013.

- 
- [41] Peter J Schmid. Dynamic mode decomposition of numerical and experimental data. *J. Fluid Mech.*, 656:5–28, 2010.
- [42] Michael D Schmidt, Ravishankar R Vallabhajosyula, Jerry W Jenkins, Jonathan E Hood, Abhishek S Soni, John P Wikswo, and Hod Lipson. Automated refinement and inference of analytical models for metabolic networks. *Phy. Biology*, 8(5):055011, 2011.
- [43] Johan AK Suykens, Joos PL Vandewalle, and Bart L de Moor. *Artificial Neural Networks for Modelling and Control of Non-Linear Systems*. Springer, 1996.
- [44] Robert Tibshirani. Regression shrinkage and selection via the lasso. *J. Royal Stat. Soc.: Series B (Methodological)*, 58(1):267–288, 1996.
- [45] Peter Van Overschee and Bart de Moor. *Subspace Identification of Linear Systems: Theory, Implementation, Applications*. Kluwer Academic Publishers, 1996.
- [46] Wen-Xu Wang, Rui Yang, Ying-Cheng Lai, Vassilios Kovanis, and Celso Grebogi. Predicting catastrophes in nonlinear dynamical systems by compressive sensing. *Phy. Rev. Letters*, 106(15):154101, 2011.
- [47] Matthew O Williams, Ioannis G Kevrekidis, and Clarence W Rowley. A data-driven approximation of the Koopman operator: Extending dynamic mode decomposition. *J. Nonlinear Sci.*, 25(6):1307–1346, 2015.
- [48] Hao Ye, Richard J Beamish, Sarah M Glaser, Sue CH Grant, Chih-hao Hsieh, Laura J Richards, Jon T Schnute, and George Sugihara. Equation-free mechanistic ecosystem forecasting using empirical dynamic modeling. *Proc. Nat. Acad. Sci. U.S.A.*, 112(13):E1569–E1576, 2015.
- [49] Bin Zhen, Zhenhua Li, and Zigen Song. Influence of time delay in signal transmission on synchronization between two coupled FitzHugh-Nagumo neurons. *Applied Sciences*, 9(10):2159, 2019.