

将问题定义为一个时间序列问题

# Transformers for modeling physical systems

Nicholas Geneva, Nicholas Zabaras\*



Scientific Computing and Artificial Intelligence (SCAI) Laboratory, University of Notre Dame, 311 Cushing Hall, Notre Dame, IN 46556, USA

## ARTICLE INFO

## Article history:

Received 18 January 2021

Received in revised form 8 November 2021

Accepted 22 November 2021

Available online 1 December 2021 预测代表物理现象  
的动力系统

## Keywords:

Transformers

使用Koopman的embeddings可以将任务

Deep learning

动态系统映射到矢量表示中，在使用

Self-attention

Transformer进行预测

Physics

Koopman

Surrogate modeling

## ABSTRACT

Transformers are widely used in natural language processing due to their ability to model longer-term dependencies in text. Although these models achieve state-of-the-art performance for many language related tasks, their applicability outside of the natural language processing field has been minimal. In this work, we propose the use of transformer models for the prediction of dynamical systems representative of physical phenomena. The use of Koopman based embeddings provides a unique and powerful method for projecting any dynamical system into a vector representation which can then be predicted by a transformer. The proposed model is able to accurately predict various dynamical systems and outperform classical methods that are commonly used in the scientific machine learning literature.<sup>1</sup>

© 2021 Elsevier Ltd. All rights reserved.

## 1. Introduction

The transformer model (Vaswani, et al., 2017), built on self-attention, has largely become the state-of-the-art approach for a large set of natural language processing (NLP) tasks including language modeling, text classification, question answering, etc. Although more recent transformer work is focused on unsupervised pre-training of extremely large models (Dai, Yang, Yang, Carbonell, Le, & Salakhutdinov, 2019; Devlin, Chang, Lee, & Toutanova, 2019; Liu, et al., 2019; Radford, Wu, Child, Luan, Amodei, & Sutskever, 2019), the original transformer model garnered attention due to its ability to out-perform other state-of-the-art methods by learning longer-term dependencies without recurrent connections. Given that the transformer model was originally developed for NLP, nearly all related work has been rightfully confined within this field with only a few exceptions. Here, we focus on the development of transformers to model dynamical systems that can replace otherwise expensive numerical solvers. In other words, we are interested in using transformers to learn the language of physics.

The surrogate modeling of physical systems is a research field that has existed for several decades and is a large ongoing effort in scientific machine learning. Past literature has explored multiple surrogate approaches including Gaussian processes (Atkinson & Zabaras, 2019; Bilionis & Zabaras, 2016; Bilionis, Zabaras, Konomi, 高斯回归

& Lin, 2013), polynomial chaos expansions (Xiu & Karniadakis, 2002), reduced-order models (Chakraborty & Zabaras, 2018; Gao, Wang, & Zahr, 2020), reservoir computing (Tanaka, et al., 2019) and deep neural networks (Geneva & Zabaras, 2020a; Tripathy & Bilionis, 2018; Zhu & Zabaras, 2018). A surrogate model is defined as a computationally inexpensive approximate model of a physical phenomenon that is designed to replace an expensive computational solver that would otherwise be needed to resolve the system of interest. An important characteristic of surrogate models is their ability to model a distribution of initial or boundary conditions rather than learning just one solution. This is arguably essential for the justification of training a deep learning model versus using a standard numerical solver, particularly in the context of utilizing deep learning methods which tend to have expensive training procedures. The most tangible applications of surrogates are for optimization, design and inverse problems where many repeated simulations are typically needed.

Standard deep neural network architectures such as auto-regressive (Geneva & Zabaras, 2020a; Mo, Zhu, Zabaras, Shi, & Wu, 2019), residual/Euler (González-García, Rico-Martínez, & Kevrekidis, 1998; Sanchez-Gonzalez, Godwin, Pfaff, Ying, Leskovec, & Battaglia, 2020), recurrent and LSTM based models (Geneva & Zabaras, 2020b; Maulik, Egele, Lusch, & Balaprakash, 2020; Mo et al., 2019; Tang, Liu, & Durlofsky, 2020) have been largely demonstrated to be effective at modeling various physical dynamics. Such models generally rely on the most recent time-steps to provide complete information on the current and past state of the system's evolution. Approaches that meld numerical time-integration methods with neural networks have also proven to be fairly successful, e.g. Wang and Lin (1998), Wessels, Weissenfels, and Wriggers (2020) and Zhu, Chang, and Fu (2018), but 将数值时间积分法与神经网络相结合的方法也取得了不错的结果

目的是代替昂贵的  
计算求解器来解决  
相关的问题

\* Corresponding author.

E-mail addresses: [n\\_geneva@nd.edu](mailto:n_geneva@nd.edu) (N. Geneva), [nzabaras@gmail.com](mailto:nzabaras@gmail.com) (N. Zabaras).URL: <https://zabaras.com/> (N. Zabaras).1 Code available at: <https://github.com/zabaras/transformer-phsys>.

固定的時間窗口

have a **fixed temporal window** from which information is provided. Present machine learning models lack generalizable time cognizant capabilities to predict **multi-time-scale phenomena** present in systems including turbulent fluid flow, multi-scale materials modeling, molecular dynamics, chemical processes, etc. Much work is needed to scale such deep learning models to complex physical systems that are of scientific and industrial interest. This work deviates from this pre-existing literature by investigating the use of transformers for the prediction of physical systems, relying entirely on self-attention to surrogate model dynamics. In the recent work of Shalova and Oseledets (2020), such self-attention models were tested to learn single solutions of several low-dimensional ordinary differential equations.

本文完全依赖  
自注意力机制

The novel contributions of this paper are as follows: (a) The application of self-attention transformer models for modeling physical dynamics; (b) The use of Koopman dynamics for developing physics inspired embeddings of high-dimensional systems with connections to embedding methods seen in NLP; (c) Discussion of the relations between self-attention with traditional numerical time-integration; (d) Demonstration of our model on high-dimensional partial differential equation problems that include chaotic dynamics, fluid flows and reaction-diffusion systems. To the authors best knowledge, this is the first work to explore transformer NLP architectures for the surrogate modeling of physical systems. The remainder of this paper is as follows: In Section 2, the machine learning methodology is discussed including the transformer decoder model in Section 2.1 and the Koopman embedding model in Section 2.2. Following in Section 3, the proposed model is implemented for a set of numerical examples of different dynamical nature. This includes classical chaotic dynamics in Section 3.1, periodic fluid dynamics in Section 3.2 and three-dimensional reaction-diffusion dynamics in Section 3.3. Lastly, concluding discussion and future directions are given in Section 4.

使用Transformer  
建立物理动态  
模型  
利用Koopman  
动力学设计  
物理学启发的  
高维系统嵌入  
讨论自注意力  
与传统数值  
时间积分之间  
的关系

## 2. Methods

We are interested in systems that can be described through a dynamical ordinary or partial differential equation of the form:

$$\phi_t = F(\mathbf{x}, \phi(t, \mathbf{x}, \eta), \nabla_{\mathbf{x}}\phi, \nabla_{\mathbf{x}}^2\phi, \phi \cdot \nabla_{\mathbf{x}}\phi, \dots), \quad (1)$$

$$t \in \mathcal{T} \subset \mathbb{R}^+, \quad \mathbf{x} \in \Omega \subset \mathbb{R}^m,$$

in which  $\phi \in \mathbb{R}^n$  is the solution of this differential equation of  $n$  state variables with parameters  $\eta$ , in the time interval  $\mathcal{T}$  and spatial domain  $\Omega$  with a boundary  $\Gamma \subset \Omega$ . This general form can embody a vast spectrum of physical phenomena including fluid flow and transport processes, mechanics and materials physics, and molecular dynamics. In this work, we are interested in learning the set of solutions for a distribution of initial conditions  $\phi_0 \sim p(\phi_0)$ , boundary conditions  $B(\phi) \sim p(B) \forall \mathbf{x} \in \Gamma$  or equation parameters  $\eta \sim p(\eta)$ . This accounts for modeling initial value, boundary value and stochastic problems. 初值、边界值和随机问题建模

To make the modeling of such dynamical systems applicable to the use of modern machine learning architectures, the continuous solution is discretized in both the spatial and temporal domains such that the solution of the differential equation is  $\Phi = (\phi_0, \phi_1, \dots, \phi_T); \phi_i \in \mathbb{R}^{n \times d}$ , for which  $\phi_i$  has been discretized by  $d$  points in  $\Omega$ . We assume an initial state  $\phi_0$  and that the time interval  $\mathcal{T}$  is discretized by  $T$  time-steps with a time-step size  $\Delta t$ . Hence, we pose the modeling a dynamical system as a **time-series problem**. The proposed machine learning methodology has two core components: the transformer for modeling dynamics and the embedding network for projecting physical states into a vector representation. Similar to NLP, the embedding model is trained prior to the transformer. This embedding model is then frozen and the entire data-set is converted to the embedded space in which the transformer is then trained as illustrated in Fig. 1.

Transformer用于动态建模，embedding网络将物理状态映射到矢量表示中。

During testing, the embedding decoder is used to reconstruct the physical states from the transformer's predictions.

### 2.1. Transformer

The transformer model was originally designed with NLP as the sole application with word vector embeddings of a passage of text being the primary input (Vaswani, et al., 2017). However, recent works have explored using attention mechanisms for different machine learning tasks (Fu, et al., 2019; Veličković, Cucurull, Casanova, Romero, Liò, & Bengio, 2018; Zhang, Goodfellow, Metaxas, & Odena, 2019) and a few investigate the use of transformers for applications outside of the NLP field (Chen, et al., 2020). This suggests that self-attention and in particular transformer models may work well for any problem that can be posed as a sequence of vectors.

#### 2.1.1. Transformer decoder

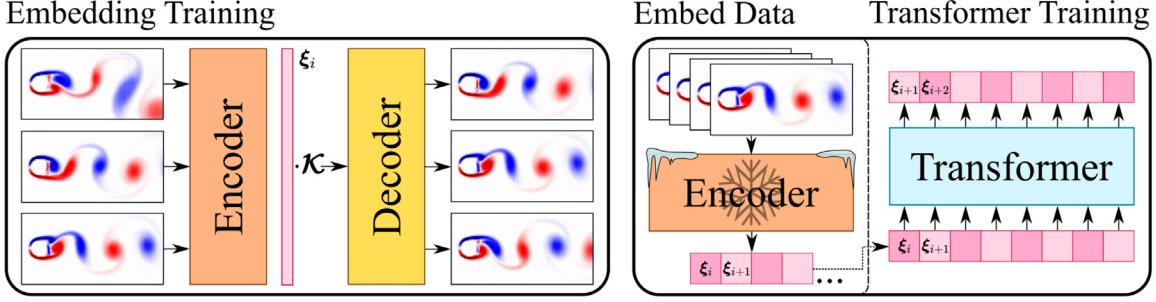
In this work, the primary input to the transformer will be an embedded dynamical system,  $\Xi = (\xi_0, \xi_1, \dots, \xi_T)$ , where the embedded state at time-step  $i$  is denoted as  $\xi_i \in \mathbb{R}^e$ . Given that we are interested in the prediction of a physical time series, this motivates the usage of a language modeling architecture that is designed for the sequential prediction of words in a body of text. We select the transformer decoder architecture used in the Generative Pre-trained Transformer (GPT) models (Radford, Narasimhan, Salimans, & Sutskever, 2018; Radford et al., 2019). Our model follows the GPT-2 architecture based on the implementation in the Hugging Face transformer repository (Wolf, et al., 2019), but is significantly smaller in size than these modern NLP transformers. This model consists of a stack of transformer decoder layers that use masked attention, as depicted in Fig. 2. The input to the transformer is the embedded representation of the physical system from the embedding model and a sinusoidal positional encoding proposed in the original transformer (Vaswani, et al., 2017). Since the transformer does not contain any recurrent or convolutional operation, information regarding the relative position of the embedded input sequence must be provided. The positional embedding is defined by the following sine and cosine functions:

$$PE_{pos,2j} = \sin(pos/10000^{2j/e}), \quad PE_{pos,2j+1} = \cos(pos/10000^{2j/e}), \quad (2)$$

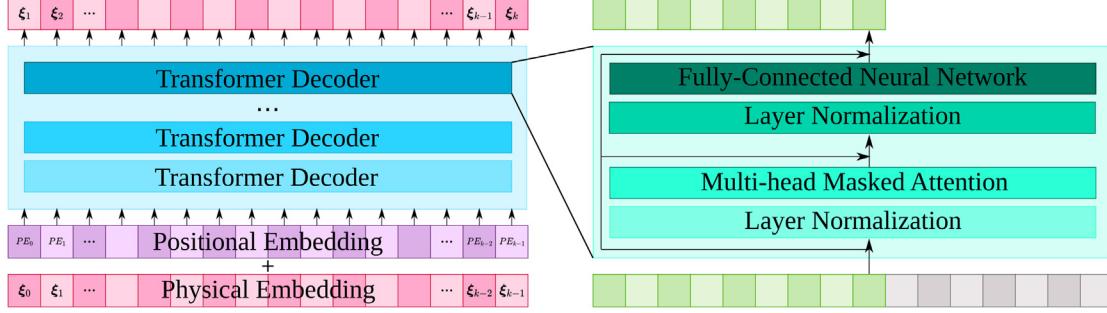
for the  $2j$  and  $2j + 1$  elements in the embedded vector,  $\xi_i$ .  $pos$  is the embedded vector's relative or global position in the input time series. To train the model, consider a data set of  $D$  embedded i.i.d. time-series  $\mathcal{D} = \{\Xi^i\}_{i=1}^D$  for which we can use the standard time-series Markov model (language modeling) log-likelihood:

$$L_{\mathcal{D}} = \sum_i^D \sum_j^T -\log p(\xi_j^i | \xi_{j-k}^i, \dots, \xi_{j-1}^i, \theta), \quad (3)$$

where  $\theta$  are the model's parameters and  $k$  is the transformer's context window. Contrary to the standard NLP approach which poses the likelihood as a softmax over a dictionary of tokens, the likelihood here is taken as a standard Gaussian between the transformer's prediction and the target embedded value resulting in a  $L_2$  loss. This is due to the fact that the solution to most physical systems cannot be condensed to a discrete finite set. Thus tokenization into a finite dictionary is not possible and a softmax approach is not applicable. Training is the standard auto-regressive method used in GPT (Radford et al., 2018), as opposed to the word masking (Devlin et al., 2019), constrained to the embedded space. The physical states,  $\phi_i$ , have the potential to be very high-dimensional and training the transformer in the lower-dimensional embedded space can significantly lower training costs.



**Fig. 1.** The two training stages for modeling physical dynamics using transformers. (Left to right) The embedding model is first trained using Koopman based dynamics. The embedding model is then frozen (fixed), all training data is embedded and the transformer is trained in the embedded space.



**Fig. 2.** The transformer decoder model used for the prediction of physical dynamics.

### 2.1.2. Self-attention

Self-attention is the main mechanism that allows the transformer to learn temporal dependencies. Prior to the seminal transformer paper, several works had already proposed a handful of different attention models typically integrated into recurrent models for language processing (Bahdanau, Cho, & Bengio, 2015; Graves, Wayne, & Danihelka, 2014; Luong, Pham, & Manning, 2015). With the popularization of the transformer model (Vaswani, et al., 2017), the most commonly used attention model is the scaled-dot product attention:

$$\mathbf{k}_i = \mathcal{F}_k(\mathbf{x}_i), \quad \mathbf{q}_i = \mathcal{F}_q(\mathbf{x}_i), \quad \mathbf{v}_i = \mathcal{F}_v(\mathbf{x}_i) \quad (4)$$

$$\mathbf{c}_n = \sum_{i=1}^k \alpha_{n,i} \mathbf{v}_i, \quad \alpha_{n,i} = \frac{\exp(\mathbf{q}_n^T \mathbf{k}_i / \sqrt{d_k})}{\sum_{j=1}^k \exp(\mathbf{q}_n^T \mathbf{k}_j / \sqrt{d_k})}, \quad (5)$$

in which we use  $\mathbf{x}_i \in \mathbb{R}^d$  and  $\mathbf{c}_n \in \mathbb{R}^{d_v}$  to denote an arbitrary input and context output, respectively.  $\mathbf{k} \in \mathbb{R}^{d_k}$ ,  $\mathbf{q} \in \mathbb{R}^{d_k}$  and  $\mathbf{v} \in \mathbb{R}^{d_v}$  are referred to as the key, query, and value vectors, respectively, calculated using neural networks  $\mathcal{F}_k$ ,  $\mathcal{F}_q$  and  $\mathcal{F}_v$ . The attention score,  $\alpha_{n,i}$ , is calculated by the soft-max of the dot product between the query and key vectors scaled by the dimension  $d_k$ . Due to the soft-max calculation note that the attention scores always sum to one,  $\sum_{i=1}^k \alpha_{n,i} = 1$ , for every input.

The attention calculation can be condensed for the entire context length of the model by a matrix representation  $\mathbf{C} = \text{softmax}(\mathbf{Q}\mathbf{K}^T / \sqrt{d_k}) \mathbf{V}$ , where  $\mathbf{Q} \in \mathbb{R}^{k \times d_k}$ ,  $\mathbf{K} \in \mathbb{R}^{k \times d_k}$  and  $\mathbf{V} \in \mathbb{R}^{k \times d_v}$ . As illustrated in Fig. 2, self-attention is typically implemented with a residual connection with multiple independent attention calculations referred to as attention heads (Vaswani, et al., 2017). While computationally inexpensive to evaluate, the memory requirement of scaled-dot product attention can become increasingly cumbersome as the context length,  $k$ , increases. Hence, several methods for approximating this calculation have been proposed which include the use of kernels and random projections in an attempt to lower the dimensionality of the self-attention calculation without loss of predictive accuracy (Kitaev,

Kaiser, & Levskaya, 2020; Sukhbaatar, Grave, Bojanowski, & Joulin, 2019; Sukhbaatar, Grave, Lample, Jegou, & Joulin, 2019).

While designed for natural language processing, in the context of dynamical systems, self-attention bears a very similar form to numerical time-integration methods. The use of time-integration methods such as Runge–Kutta schemes or Euler methods can be found in Neural ODEs (Chen, Rubanova, Bettencourt, & Duvenaud, 2018; Dupont, Doucet, & Teh, 2019) and Res-Net based models (González-García et al., 1998; Sanchez-Gonzalez et al., 2020). Self-attention offers a much larger learning capacity than using these traditional numerical methods for learning a unique latent time-integration method. The function space of a self-attention layer with a residual connection contains the space of explicit linear time-integration methods within an arbitrarily small non-zero amount of error.

**Theorem 1.** Consider a dynamical system of the form,  $d\phi/dt = f(t, \phi)$ ,  $\phi(t) \in \mathbb{R}^d$ , where  $f : \mathbb{R} \times \mathbb{R}^d \rightarrow \mathbb{R}^d$  is Lipschitz continuous with respect to  $\phi$  and continuous in  $t$ . Let the function  $\mathcal{A}_\theta(t, \phi_{t-k\Delta t:t}) = \hat{\phi}_{t+\Delta t}$  be a self-attention layer with a residual connection, of context length  $k$  and containing learnable parameters  $\theta \in \mathbb{R}^{d_\theta}$ . Suppose  $A := \{\mathcal{A}_\theta(t, \phi) \mid \forall \theta \in \mathbb{R}^{d_\theta}\}$  be the set of all possible self-attention calculations. Let  $\mathcal{M}_i(t, \phi_{t-(i-1)\Delta t:t}) = \phi_{t+\Delta t}$  be the  $i$ th order explicit Adams method time-integrator. The set of functions up to  $k$ th order  $M := \{\mathcal{M}_i \mid 1 \leq i \leq k\}$  is a subset of  $A$  such that  $\exists \mathcal{A}_{\theta_i} \in A$  s.t.  $\|\mathcal{M}_i - \mathcal{A}_{\theta_i}\|_\infty < O(\epsilon)$  for any  $\mathcal{M}_i \in M$  and  $\epsilon > 0$ .

**Proof.** The state variables are discretized w.r.t. time such that  $\phi_i = \phi(i\Delta t)$ ;  $t_i = i\Delta t$ . The general definition for linear multi-step methods follows:

$$\begin{aligned} \phi_{n+s} + a_{s-1} \cdot \phi_{n+s-1} + \cdots + a_0 \cdot \phi_n \\ = \Delta t \cdot (b_s \cdot f(t_{n+s}, \phi_{n+s}) + b_{s-1} \cdot f(t_{n+s-1}, \phi_{n+s-1}) \\ + \cdots + b_0 \cdot f(t_n, \phi_n)), \end{aligned} \quad (6)$$

in which  $a_i$  and  $b_i$  are coefficients determined by the integration method. For explicit Adams methods, the  $\phi_{n+s}$  is directly computed with  $b_s = 0$ ,  $a_{s-1} = -1$  and  $a_{0:s-2} = 0$ . The generalized formula for  $s$ -step explicit Adams methods can be represented as the following linear combination:

$$\mathcal{M}_s = \phi_{n+s} = \phi_{n+s-1} + \Delta t \sum_{j=0}^{s-1} b_j f(t_{n+j}, \phi_{n+j}), \quad (7)$$

which encapsulates up to and including  $s$ th order time integration (Stoer & Bulirsch, 2013). Consider a residual scaled dot-product self-attention calculation with context length,  $s$ , for output prediction  $\hat{\phi}_{n+s}$ , input states  $\phi_{n:n+s-1}$  and time  $t$ :

$$\begin{aligned} \mathcal{A}_{\theta_i} = \hat{\phi}_{n+s} &= \phi_{n+s-1} + \sum_{i=0}^{s-1} \alpha_{n+s-1,i} \mathbf{v}_i, \\ \alpha_{n+s-1,i} &= \frac{\exp(\mathbf{q}_{n+s-1}^T \mathbf{k}_i / \sqrt{d_k})}{\sum_{j=0}^{s-1} \exp(\mathbf{q}_{n+s-1}^T \mathbf{k}_j / \sqrt{d_k})}, \end{aligned} \quad (8)$$

for which  $\mathbf{k}_i = \mathcal{F}_k(t_{n+i}, \phi_{n+i})$ ,  $\mathbf{q}_i = \mathcal{F}_q(t_{n+i}, \phi_{n+i})$  and  $\mathbf{v}_i = \mathcal{F}_v(t_{n+i}, \phi_{n+i})$  vectors are outputs of differentiable functions parameterized by neural networks:

$$\mathcal{F}_k : \mathbb{R} \times \mathbb{R}^d \rightarrow \mathbb{R}^{d_k}, \quad \mathcal{F}_q : \mathbb{R} \times \mathbb{R}^d \rightarrow \mathbb{R}^{d_k}, \quad \mathcal{F}_v : \mathbb{R} \times \mathbb{R}^d \rightarrow \mathbb{R}^{d_v}. \quad (9)$$

It suffices to show that there exists a form of Eq. (8) that is equal to Eq. (7) within an error order  $O(\epsilon)$  such that

$$\|\phi_{n+s} - \hat{\phi}_{n+s}\|_\infty < O(\epsilon). \quad (10)$$

To this end, although  $\mathcal{F}_q$  and  $\mathcal{F}_k$  are fully-trainable neural networks, herein we consider a fixed single-entry query and key vectors resulting in the following simplified self-attention scores:

$$\begin{aligned} \mathbf{q}_i &= \mathcal{F}_q(t_{n+i}, \phi_{n+i}) = \sqrt{d_k} \mathbf{e}_m, \quad \mathbf{k}_i = \mathcal{F}_k(t_{n+i}, \phi_{n+i}) = \log(b_i) \mathbf{e}_m, \\ \alpha_{n+s-1,i} &= \frac{\exp(\log(b_i))}{\sum_{j=0}^{s-1} \exp(\log(b_j))} = \frac{b_i}{\sum_{j=0}^{s-1} b_j}, \end{aligned} \quad (11)$$

where  $\mathbf{e}_m \in \mathbb{R}^{d_k}$  is a unit vector with all elements set to zero except the  $m$ th element that is set to 1. By the universal approximation theorem (Cybenko, 1989; Hornik, 1991; Hornik, Stinchcombe, & White, 1989), the neural network,  $\mathcal{F}_v$ , is assumed to be of sufficient capacity such that it can approximate the r.h.s. of the governing equation:

$$\mathbf{v}_i = \mathcal{F}_v(t_{n+i}, \phi_{n+i}) = c f(t_{n+i}, \phi_{n+i}) + O(\epsilon), \quad \epsilon > 0, \quad (12)$$

where the constant  $c$  is taken as  $c = \Delta t \sum_{j=0}^{s-1} b_j$ . Eqs. (11) and (12) can then be combined leading to the following:

$$\begin{aligned} \phi_{n+s-1} + \sum_{i=0}^{s-1} \alpha_{n+s-1,i} \mathbf{v}_i &= \phi_{n+s-1} + \sum_{i=0}^{s-1} \frac{b_i}{\sum_{j=0}^{s-1} b_j} \\ &\quad c f(t_{n+i}, \phi_{n+i}) + O(\epsilon), \\ &= \phi_{n+s-1} + \Delta t \sum_{i=0}^{s-1} b_i f(t_{n+i}, \phi_{n+i}) + O(\epsilon). \end{aligned} \quad (13)$$

For a large capacity neural network, we can assume that the error can be made arbitrarily small ( $\epsilon \rightarrow 0$ ) and Eq. (10) is proved. Therefore, the form of the explicit Adams multi-step methods of order  $\leq s$  can be captured by a residual self-attention transformer layer.  $\square$

A single transformer layer is significantly more expressive than any numerical time integration scheme, enabling it to learn more complex temporal dependencies. However, the linear combination of weighted features from past time-steps is a commonality between self-attention and numerical time-integration methods. This mathematical similarity indicates self-attention models may be better suited for learning dynamical systems than alternative deep learning approaches. In particular traditional Euler time integration present in various models for modeling dynamics (González-García et al., 1998; Lu, Zhong, Li, & Dong, 2018; Sanchez-Gonzalez et al., 2020) is a specific case of this single layer attention calculation.

**Remark.** The inclusion of time,  $t$ , as an input is required for non-autonomous systems with explicit time-dependent terms. This has surprising connections to the implementation of the transformer which uses a positional embedding. Although the positional embedding is included for the sake of denoting the order of the input to the transformer, using current time of the given input state can accomplish the same goal.

## 2.2. Embedding model

The second major component of the machine learning methodology is the embedding model responsible for projecting the discretized physical state space into a 1D vector representation. In NLP, the standard approach is to tokenize then embed a finite vocabulary of words, syllables or characters using methods such as n-gram models, Byte Pair Encoding (Gage, 1994), Word2Vec (Mikolov, Chen, Corrado, & Dean, 2013; Mikolov, Sutskever, Chen, Corrado, & Dean, 2013), GloVe (Pennington, Socher, & Manning, 2014), etc. These methods allow language to be represented by a series of 1D vectors that serve as the input to the transformer. Clearly a finite tokenization and such NLP embeddings are not directly applicable to physics, thus we propose our own embedding method designed specifically for dynamical systems. Consider learning the generalized mapping between the system's state space and embedded space:  $\mathcal{F} : \mathbb{R}^{n \times d} \rightarrow \mathbb{R}^e$  and  $\mathcal{G} : \mathbb{R}^e \rightarrow \mathbb{R}^{n \times d}$ . Naturally, multiple approaches can be used especially if the dimensionality of the embedded space is less than that of the state-space but this is not always the case.

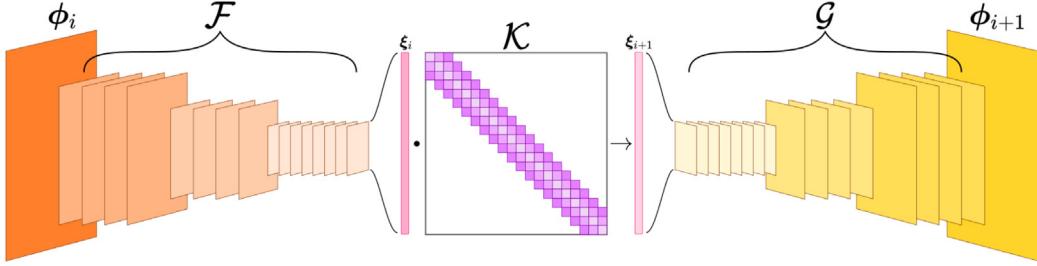
The primary approach that we will propose is a Koopman observable embedding which is a technique that can be applied universally to all dynamical systems. Considering the discrete time form of the dynamical system in Eq. (1), the evolution of the state variables can be abstracted by  $\phi_{i+1} = \mathbb{F}(\phi_i)$  for which  $\mathbb{F}$  is the dynamic map from one time-step to the next. The foundation of Koopman theory states any dynamical system can be represented in terms of an infinite dimensional linear operator acting on an infinite set of state observable functions,  $g(\phi_i)$ , such that:

$$\mathcal{K}g(\phi_i) \triangleq g \circ \mathbb{F}(\phi_i), \quad (14)$$

where  $\mathcal{K}$  is the infinite-dimensional linear operator referred to as the Koopman operator (Koopman, 1931). This implies that the system of observables can be evolved in time through repeated application of the Koopman operator:

$$\begin{aligned} g(\phi_{i+1}) &= \mathcal{K}g(\phi_i), \quad g(\phi_{i+2}) = \mathcal{K}^2g(\phi_i), \\ g(\phi_{i+3}) &= \mathcal{K}^3g(\phi_i), \dots \end{aligned} \quad (15)$$

in which  $\mathcal{K}^n$  denotes a  $n$ -fold composition, e.g.  $\mathcal{K}^3(g) = \mathcal{K}(\mathcal{K}(\mathcal{K}(g)))$ . Modeling the dynamics of a system through the linear Koopman space can be attractive due to its simplification of the dynamics but also the potential physical insights it brings along with it.



**Fig. 3.** Example of a Koopman embedding for a two-dimensional system using a convolution encoder-decoder model. The encoder model,  $\mathcal{F}$ , projects the physical states into the approximate Koopman observable embedding. The decoder model,  $\mathcal{G}$  recovers the physical states from the embedding.

Spectral analysis of the Koopman operator can reveal fundamental dynamical modes that drive the system's evolution in time.

Koopman theory can be viewed as a trade off between lifting the state space into observable space with more complex states but simpler dynamics. In practice, the Koopman operator must be finitely approximated. This finite approximation requires the identification of the essential measurement functions that govern the system's dynamics and the respective approximate Koopman operator. Data-driven machine learning has proven to be an effective approach for learning key Koopman observables for modeling, control and dynamical mode analysis of many physical systems (Korda & Mezić, 2018; Korda, Putinar, & Mezić, 2020; Li, Dietrich, Bollt, & Kevrekidis, 2017; Mezic, 2020). In recent years, the use of deep neural networks for learning Koopman dynamics has proven to be successful (Brunton, Budišić, Kaiser, & Kutz, 2021; Lusch, Kutz, & Brunton, 2018; Otto & Rowley, 2019; Takeishi, Kawahara, & Yairi, 2017). While deep learning methods have enabled greater success with discovering Koopman observables and operators such approaches have yet to be demonstrated for the long time prediction of high-dimensional systems. This is likely due to the approximation of the finite-dimensional Koopman observables, limited data and complete dependence on the discovered Koopman operator  $\mathcal{K}$  to model the dynamics. Suggesting the prediction of a system through a single linear transform clearly has significant limitations and is fundamentally a naive approach from a machine learning perspective.

In this work, we propose using approximate Koopman dynamics as a methodology to develop embeddings for the transformer model such that  $\mathcal{F}(\phi_i) \triangleq g(\phi_i)$ . As seen in Fig. 3, the embedding model follows a standard encoder-decoder model with the middle latent variables being the Koopman observables. In this model, the Koopman operator assumes the form of a learnable banded matrix that is optimized with the auto-encoder. Imposing some level of inductive bias on the form of the Koopman matrix is fairly common in similar deep Koopman works (Li, He, Wu, Kataib, & Torralba, 2020; Lusch et al., 2018; Otto & Rowley, 2019). We found that this reduction of learnable parameters helped encourage the model to discover better dynamical modes, preventing the model from overfitting to high-frequency fluctuations. Additionally, this form requires significantly less memory to store allowing models with embedding vectors of higher-dimensionality to be trained. This learned Koopman operator is disposed of once training of the embedding model is complete. Given the data set of physical state time-series,  $\mathcal{D}_\phi = \{\phi^i\}_i^D$ , the Koopman embedding model is trained with the following loss:

$$\begin{aligned} \mathcal{L}_{\mathcal{D}_\phi} = & \sum_{i=1}^D \sum_{j=0}^T \lambda_0 \underbrace{\text{MSE}(\phi_j^i, \mathcal{G} \circ \mathcal{F}(\phi_j^i))}_{\text{Reconstruction}} \\ & + \lambda_1 \underbrace{\text{MSE}(\phi_j^i, \mathcal{G} \circ \mathcal{K}^j \mathcal{F}(\phi_0^i))}_{\text{Dynamics}} + \lambda_2 \underbrace{\|\mathcal{K}\|_2^2}_{\text{Decay}}. \end{aligned} \quad (16)$$

This loss function consists of three components: the first is a reconstruction loss which ensures a consistent mapping to and from the embedded representation. The second is the Koopman dynamics loss which pushes  $\xi_j$  to follow linear dynamics. The last term decays the Koopman operator's parameters to help force the model to discover meaningful dynamical modes and further prevent overfitting.

In reference to traditional NLP embeddings, we believe our Koopman observable embedding has a motivation similar to Word2Vec (Mikolov, Chen, et al., 2013) as well as more recent embedding methods such as context2vec (Melamud, Goldberger, & Dagan, 2016), ELMo (Peters, Neumann, Iyyer, Gardner, Clark, Lee, & Zettlemoyer, 2018), etc. These methods are based on word context and association to develop a map where words that are related or synonymous to each other have similar embedded vectors. The Koopman embedding model has a similar objective encouraging physical realizations containing similar dynamical modes to also have similar embeddings. This is because the Koopman operator is time-invariant which means the embedded states must share the same basis functions that govern their evolution. As a result, the loss function of the embedding model rewards time-steps that are near each other in time or have the same underlying dynamics to have similar embeddings. Hence, our goal with the embedding model is to not find the true Koopman observables or operator, but rather leverage Koopman to enforce physical context and association using the learned dynamical modes.

### 3. Experiments and results

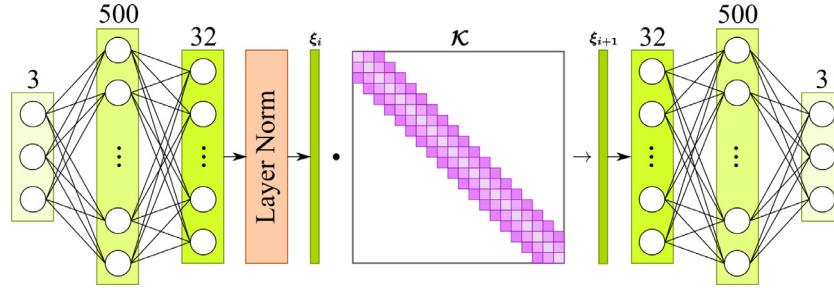
The proposed transformer model is implemented for three dynamical systems. The classical Lorenz system is used as a baseline test case in Section 3.1 to compare the proposed model to alternative machine learning approaches due to the Lorenz system's numerical sensitivity. Following in Section 3.2, we consider the modeling of two-dimensional Navier-Stokes fluid flow and compare different embedding methods for the transformer. Lastly, to demonstrate the models scalability, we demonstrate in Section 3.3 the transformer for the prediction of a three-dimensional reaction-diffusion system. These examples encompass surrogate modeling physical systems with stochastic initial conditions and parameters.

#### 3.1. Chaotic dynamics

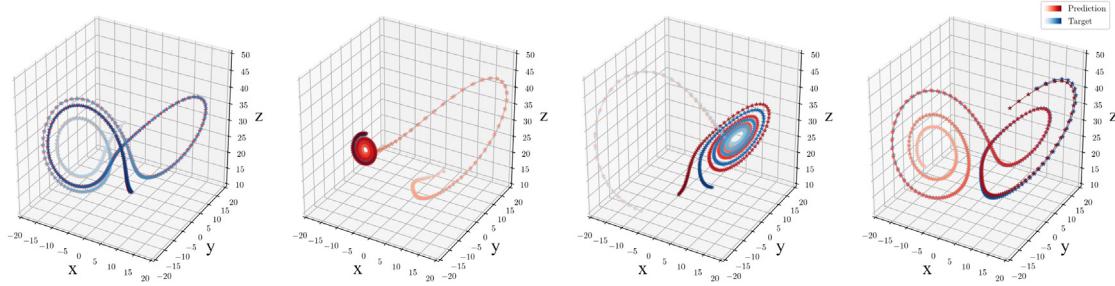
As a foundational numerical example to rigorously compare the proposed model to other classical machine learning techniques, we will first look at surrogate modeling of the Lorenz system governed by:

$$\frac{dx}{dt} = \sigma(y - x), \quad \frac{dy}{dt} = x(\rho - z) - y, \quad \frac{dz}{dt} = xy - \beta z. \quad (17)$$

We use the classical parameters of  $\rho = 28$ ,  $\sigma = 10$ ,  $\beta = 8/3$ . For this numerical example, we wish to develop a surrogate model



**Fig. 4.** Fully-connected embedding network with ReLU activation functions for the Lorenz system.



**Fig. 5.** Four test case predictions using the transformer model for 320 time-steps.

for predicting the Lorenz system given a random initial state  $x_0 \sim \mathcal{U}(-20, 20)$ ,  $y_0 \sim \mathcal{U}(-20, 20)$  and  $z_0 \sim \mathcal{U}(10, 40)$ . In other words, we wish to surrogate model various initial value problems for this system of ODEs. The Lorenz system is used because of its well known chaotic dynamics which make it extremely sensitive to numerical perturbations and thus an excellent benchmark for assessing a machine learning model's accuracy.

A total of four alternative machine learning models are implemented: a fully-connected auto-regressive model, a fully-connected LSTM model, a deep neural network Koopman model and lastly an echo-state model. All of these types of models have been proposed in past literature for predicting various physical systems (e.g. auto-regressive Geneva & Zabaras, 2020a, fully-connected LSTMs Zhao, Deng, Cai, & Chen, 2019, deep Koopman Lusch et al., 2018; Otto & Rowley, 2019 and echo-state models Chattopadhyay, Hassanzadeh, Subramanian, & Paley, 2020; Lukoševičius, 2012). Each is provided the same training, validation and testing data sets containing 2048, 64 and 256 time-series at a time-step size of  $\Delta t = 0.01$  solved using a Runge-Kutta numerical solver, respectively. The training data set contains time-series of 256 time-steps while the validation and testing data sets have 1024 time-steps. Each model is allowed to train for 500 epochs if applicable. The proposed transformer and embedding model are trained for 200 and 300 epochs, respectively, with an embedding dimension of 32. The embedding model is a simple fully-connected encoder-decoder model,  $\mathcal{F} : \mathbb{R}^3 \rightarrow \mathbb{R}^{32}; \mathcal{G} : \mathbb{R}^{32} \rightarrow \mathbb{R}^3$ , illustrated in Fig. 4. The transformer is trained with a context length of 64 with 4 transformer decoder layers. Both the training and validation data sets were chunked into a set of 64 steps for the alternative models, when applicable, to train on the same context length.

We plot four separate test cases in Fig. 5 for which only the initial state is provided and the transformer model predicts 320 time-steps. Several test predictions for alternative models are provided in Appendix A. In general, we can see that the transformer model is able to yield extremely accurate predictions even beyond its context length. Additionally, we plot the Lorenz solution for 25k time-steps from a numerical solver and predicted from the transformer model in Fig. 6. Note that both have the

same structure, which qualitatively indicates that the transformer indeed maintains physical dynamics.

The proposed transformer and alternative models' relative mean squared errors for the test set are plotted in Fig. 7(a) as a function of time and listed in Table 1 segmented into several intervals based on the transformer's context length. In general, we can see all deep learning models perform well in the time-series length for which they were trained with the deep Koopman model performing the best followed by the transformer. As we extrapolate our predictions past the trained context range, the benefits of the transformer become apparent with it achieving the best accuracy for later times. To quantify accuracy of the chaotic dynamics of each model, the Lorenz map is plotted in Fig. 7(b) which is a well-defined relation between successive  $z$  local maxima despite the Lorenz's chaotic nature. Calculated using 25k time-step predictions from each model, again we can see that the transformer model agrees the best with the numerical solver indicating that it has learned the best physical dynamics of all the tested models.

Additionally, we test the transformer's sensitivity to contaminated data by adding white noise to the training observations scaled by the magnitude of the state variables. Each model tested with clean data is retrained with data perturbed by 1% and 5% noise. The effects of these two different noise levels are qualitatively illustrated in Appendix A. The errors are listed in Table 2. In general, we can indeed see that the transformer can still perform adequately with noisy data by being the best performing model for 1% noise and still being competitive, particularly at later time-steps, with 5% noise.

The self-attention vectors,  $\alpha_i \in \mathbb{R}^{64}$ , for a single time-series prediction of 512 steps are plotted in Fig. 8. For time-steps over  $i \geq 64$ , the full context length of the transformer is used with the attention weight  $\alpha_{i,64}$  corresponding to the most recent time-step. Note that the transformer has learned multi-scale temporal dependencies not achievable with other machine learning architectures. Additionally, this verifies that each attention head is learning different dependencies suggesting that each head may be a particular component of the transformer's latent dynamics. This aligns with what is observed in NLP when using multi-head attention which increases the transformer's predictive accuracy (Vaswani, et al., 2017).

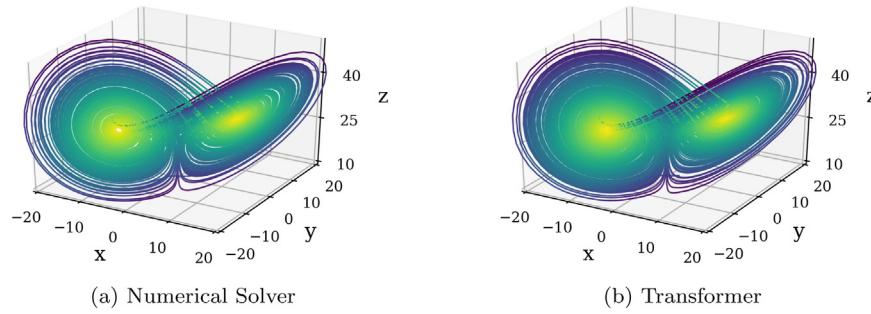
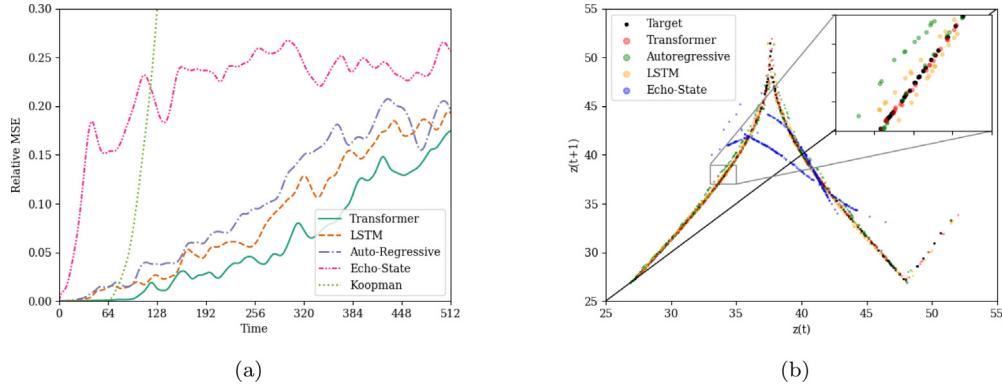
Fig. 6. Lorenz solution of 25k time-steps with  $\Delta t = 0.01$ .

Fig. 7. (a) The test relative mean-squared-error (MSE) with respect to time. (b) The Lorenz map produced by each model.

**Table 1**

Test set relative mean-squared-error (MSE) for surrogate modeling the Lorenz system at several time-step intervals.

Model	Parameters	Relative MSE		
		[0 – 64)	[64 – 128)	[128 – 192)
Transformer	36k/54k <sup>a</sup>	0.0003	<b>0.0060</b>	<b>0.0221</b>
LSTM	103k	0.0041	0.0175	0.0369
Autoregressive	92k	0.0057	0.0253	0.0485
Echo state	7.5k/6.3 m <sup>b</sup>	0.1026	0.1917	0.2209
Koopman	108k	<b>0.0001</b>	0.0962	2.0315

<sup>a</sup>Learnable parameters for the embedding/transformer model.

<sup>b</sup>Learnable output parameters/fixed input and reservoir parameters.

**Table 2**

Test set relative mean-squared-error (MSE) for surrogate modeling the Lorenz system at several time-step intervals with noisy data.

Model	Relative MSE 1% Noise			Relative MSE 5% Noise		
	[0 – 64)	[64 – 128)	[128 – 192)	[0 – 64)	[64 – 128)	[128 – 192)
Transformer	<b>0.0021</b>	<b>0.0216</b>	<b>0.0429</b>	0.0210	0.0759	<b>0.1292</b>
LSTM	0.0045	0.0218	0.0437	0.0212	<b>0.0758</b>	0.1324
Autoregressive	0.0114	0.0417	0.0901	0.0760	0.2060	0.2065
Echo state	0.0859	0.1686	0.2102	0.1000	0.1581	0.2051
Koopman	0.0047	0.1192	0.1597	<b>0.0200</b>	0.0787	0.1906

### 3.2. 2D fluid dynamics

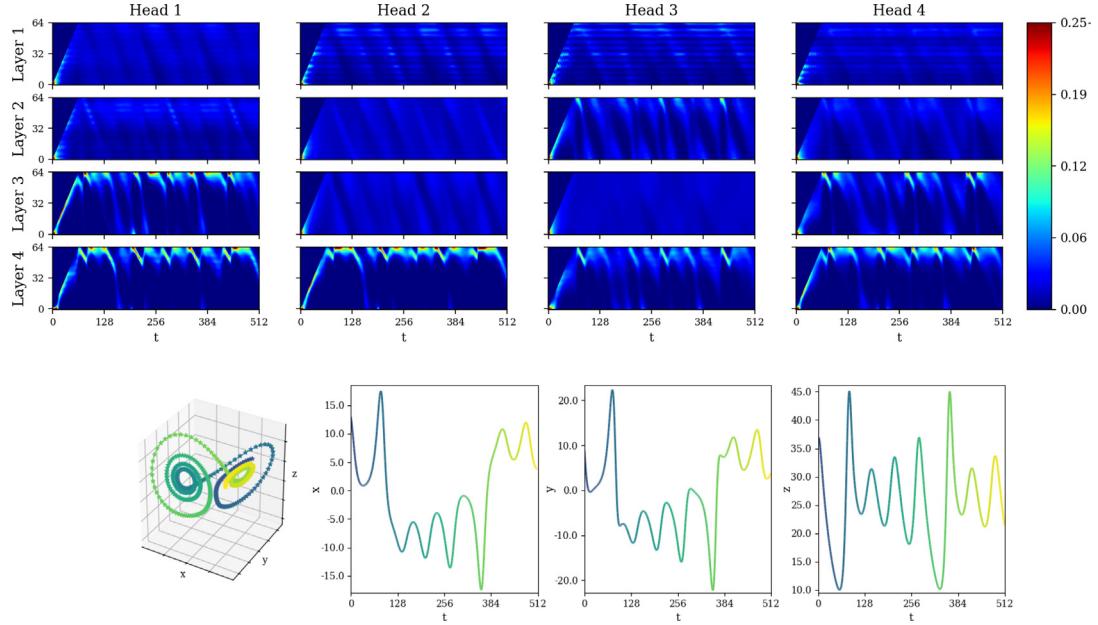
The next dynamical system we will test is transient 2D fluid flow governed by the Navier–Stokes equations:

$$\frac{\partial u_i}{\partial t} + u_j \frac{\partial u_i}{\partial x_j} = -\frac{1}{\rho} \frac{\partial p}{\partial x_i} + \nu \frac{\partial^2 u_i}{\partial x_j \partial x_j}, \quad (18)$$

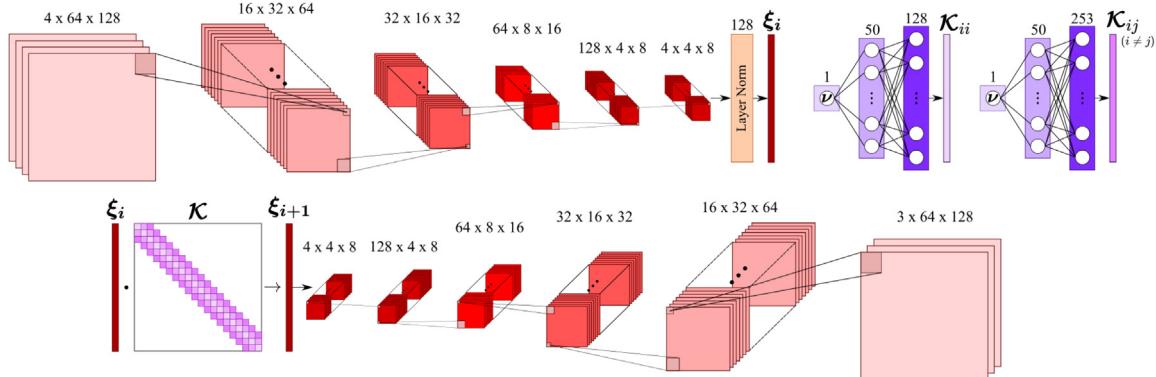
in which  $u_i$  and  $p$  are the velocity and pressure, respectively.  $\nu$  is the viscosity of the fluid. We consider modeling the classical problem of flow around a cylinder at various Reynolds number defined by  $Re = u_{in}d/\nu$  in which  $u_{in} = 1$  and  $d = 2$  are the inlet velocity and cylinder diameter, respectively. In this work, we choose to develop a surrogate model to predict the solution at any

Reynolds number between  $Re \sim \mathcal{U}(100, 750)$ . This problem is a fairly classical flow to investigate in scientific machine learning with various levels of difficulty (Geneva & Zabaras, 2020b; Han, Wang, Zhang, & Chen, 2019; Lee & You, 2017; Lusch et al., 2018; Morton, Jameson, Kochenderfer, & Witherden, 2018; Xu & Duraisamy, 2020). Here we choose one of the more difficult forms: model the flow starting at a steady state flow field at  $t = 0$ . Meaning the model is provided zero information on the structure of the cylinder wake during testing other than the viscosity.

Training, validation and testing data is obtained using the OpenFOAM simulator (Jasak, Jemcov, Tukovic, et al., 2007), from which a rectangular structured sub-domain is sampled centered around the cylinder's wake. Given that the flow is two-



**Fig. 8.** Attention vectors of the transformer model for the prediction of a single test case of the Lorenz system. The top contours illustrate the attention weights at each time-step; the bottom shows the respective state variables.



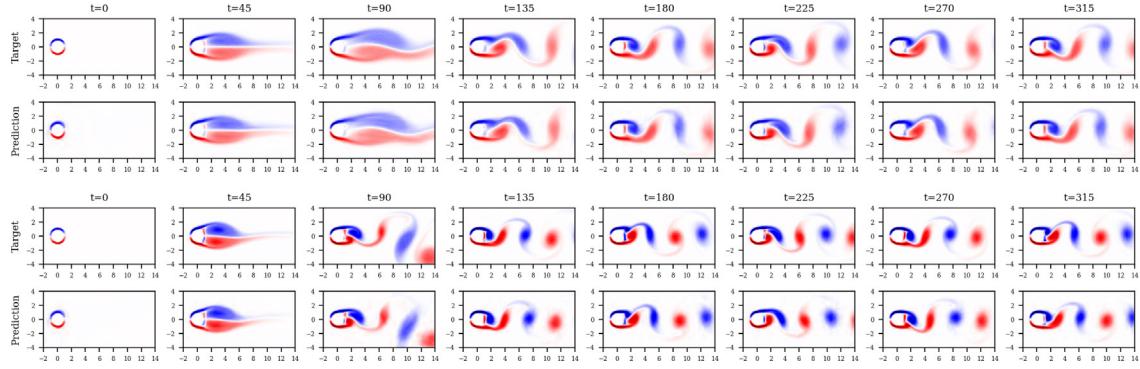
**Fig. 9.** 2D convolutional embedding network with ReLU activation functions for the flow around a cylinder system consisting of 5 convolutional encoding/decoding layers. Each convolutional operator has a kernel size of (3, 3). In the decoder, the feature maps are up-sampled before applying a standard convolution. Additionally, two auxiliary fully-connected networks are used to predict the diagonal and off-diagonal elements of the Koopman operator for each viscosity  $\nu$ .

dimensional, our model will predict the  $x$ -velocity,  $y$ -velocity and pressure fields,  $(\mathbf{u}_x, \mathbf{u}_y, p) \in \mathbb{R}^{3 \times 64 \times 128}$ . The training, validation and test data sets consist of 27, 6 and 7 fluid flow simulations, respectively with 400 time-steps each at a physical time-step size of  $\Delta t = 0.5$ . As a base line model, we train a convolutional encoder-decoder model with a stack of three convolutional LSTMs (Shi, Chen, Wang, Yeung, Wong, & Woo, 2015) in the center. The input for this model consists of the velocity fields, pressure field and viscosity of the fluid. We note that convolutional LSTMs have been used extensively in recent scientific machine learning literature for modeling various physical systems including fluid dynamics (Geneva & Zabaras, 2020b; Han et al., 2019; Maulik, Lusch, & Balaprakash, 2021; Tang et al., 2020; Wiewel, Becher, & Thuerey, 2019), thus can be considered a state-of-the-art approach. The convolutional LSTM model is trained for 500 epochs.

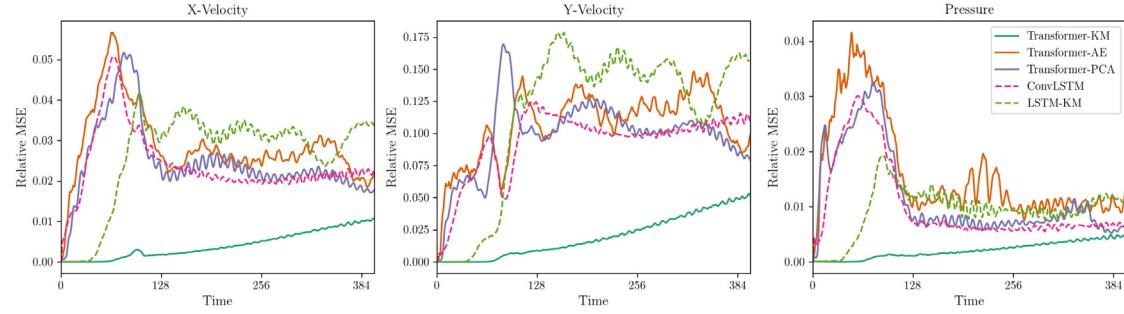
Additionally, three different embedding methods are implemented: the first is the proposed Koopman observable embedding using a convolutional auto-encoder illustrated in Fig. 9. This model encodes the fluid  $x$ -velocity,  $y$ -velocity, pressure and viscosity fields to an embedded dimension of 128,  $\mathcal{F} : \mathbb{R}^{4 \times 64 \times 128} \rightarrow$

$\mathbb{R}^{128}; \mathcal{G} : \mathbb{R}^{128} \rightarrow \mathbb{R}^{3 \times 64 \times 128}$ . The second embedding method is using the same convolutional auto-encoder model but without the enforcement of Koopman dynamics on the embedded variables. The third embedding method tested was principal component analysis (PCA) as a classical baseline. For each embedding method an identical transformer model with a context length of 128 and 4 transformer decoder layers is trained. Similar to the previous example, the embedding models are trained for 300 epochs when applicable and the transformer is trained for 200. To further isolate the impact of the transformer model from the embedding model, we also train a fully-connected model with 4 LSTM cells using the Koopman embedding as an input for 200 epochs.

Each trained model is tested on the test set by providing the initial laminar state at  $t = 0$  with the fluid viscosity and allowing the model to predict 400 time-steps into the future. Two test predictions using the proposed transformer model with Koopman embeddings are plotted in Fig. 10 in which the predicted vorticity fields are in good agreement with the true solution. The test set relative mean square error for each output field for each model is plotted in Fig. 11. The errors of each field over the entire time-series are listed in Table 3. Additional results are provided in Appendix B.



**Fig. 10.** Vorticity,  $\omega = \nabla_x \mathbf{u}_y - \nabla_y \mathbf{u}_x$ , of two test case predictions using the proposed transformer with Koopman embeddings at Reynolds numbers 233 (top) and 633 (bottom).



**Fig. 11.** Test set relative mean-squared-error (MSE) of the transformer with Koopman (KM), auto-encoder (AE) and PCA embedding methods, the convolutional LSTM model (ConvLSTM) and fully-connected LSTM with Koopman embeddings (LSTM-KM).

**Table 3**

Test set relative mean-squared-error (MSE) of each output field for surrogate modeling 2D fluid flow past a cylinder. Models listed include the transformer with Koopman (KM), auto-encoder (AE) and PCA embedding methods, the convolutional LSTM model (ConvLSTM) and fully-connected LSTM with Koopman embeddings (LSTM-KM).

Model	Parameters	Relative MSE [0 – 400]		
		$\mathbf{u}_x$	$\mathbf{u}_y$	$p$
Transformer-KM	224k/628k <sup>a</sup>	<b>0.0042</b>	<b>0.0198</b>	<b>0.0021</b>
Transformer-AE	224k/628k <sup>a</sup>	0.0288	0.1068	0.0161
Transformer-PCA	3.1m/628k <sup>a</sup>	0.0247	0.0984	0.0117
ConvLSTM	934k	0.0240	0.0938	0.0103
LSTM-KM	224k/759k <sup>b</sup>	0.0266	0.1155	0.0094

<sup>a</sup>Learnable parameters for the embedding/ transformer model.

<sup>b</sup>Learnable parameters for the embedding/ LSTM model.

For all alternative models, a rapid error increase can be seen between  $t = [0, 100]$  which is due to the transition from the laminar flow into vortex shedding. This error then plateaus since each model is able to produce stable vortex shedding, as illustrated in Figs. B.18 & B.19 in Appendix B. The proposed transformer with Koopman embeddings is the only model that can accurately match the instantaneous states from the numerical solution. These results indicate that the performance of the transformer is highly dependent on the embedding method used, which should be expected. However, the transformer with self-attention is equally important to the model's success as indicated by the performance of the Koopman embedding LSTM model. Compared to the widely used ConvLSTM model, the proposed transformer offers more reliable predictions for this fluid flow with less learnable parameters.

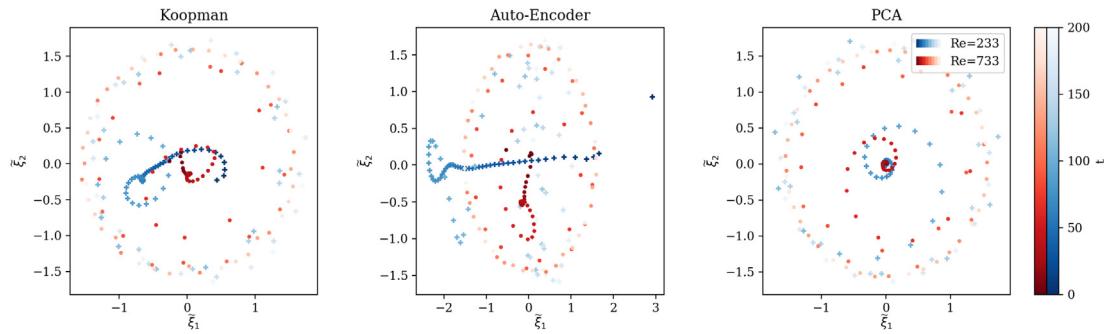
To gain a greater understanding of why the Koopman embedding performs better than the alternatives, we perform linear dimensionality reduction of the embedded states using PCA into two principle components,  $\xi_1, \xi_2$ , in Fig. 12. For all embedding methods, a circular structure is present reflecting the periodic dynamics of the vortex shedding. Compared to the auto-encoder, we

note that the Koopman principle subspace has a more consistent structure for later time-steps indicating that the Koopman loss term does indeed encourage the discovery of common dynamical modes. Yet for early time-steps, the Koopman model has a unique trajectory between Reynolds numbers. This lack of uniqueness with the PCA embedding at early time-steps results in the transformer not being able to differentiate between Reynolds numbers, lowering predictive accuracy with this embedding method. The Koopman approach strikes a balance between structure but uniqueness between flows for the transformer to learn with.

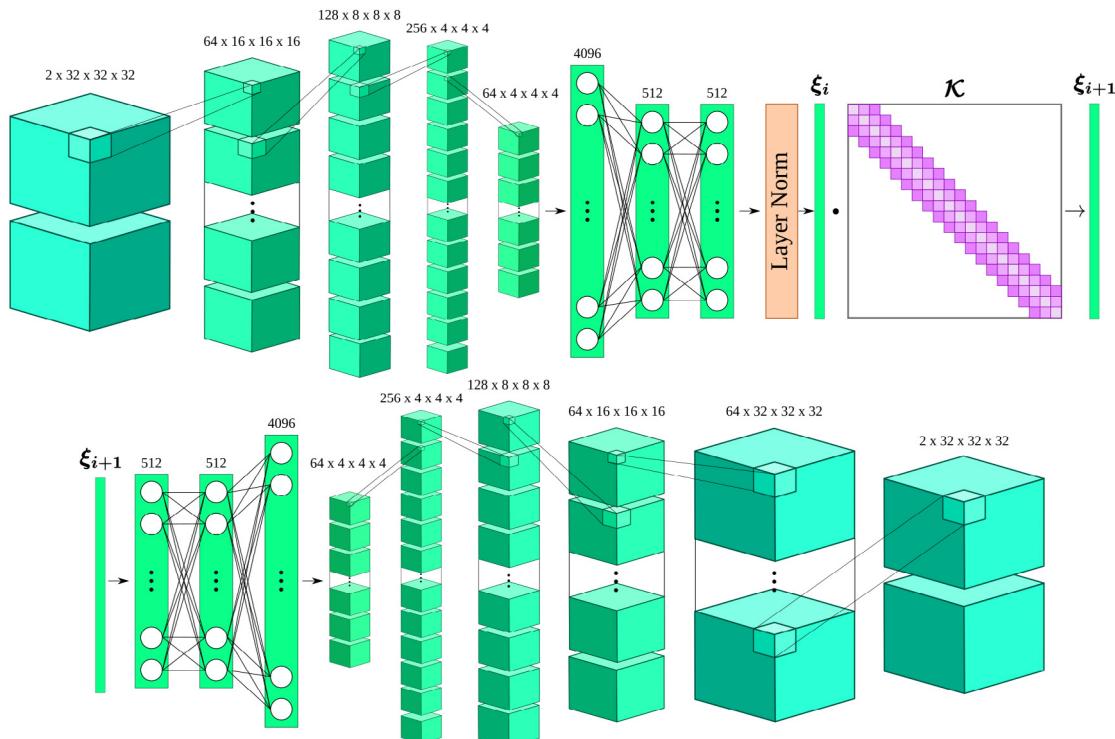
### 3.3. 3D Reaction-diffusion dynamics

The final numerical example to demonstrate the proposed transformer model is a 3D reaction-diffusion system governed by the Gray-Scott model:

$$\frac{\partial u}{\partial t} = r_u \frac{\partial^2 u}{\partial x_i^2} - uv^2 + f(1-u), \quad \frac{\partial v}{\partial t} = r_v \frac{\partial^2 v}{\partial x_i^2} + vu^2 - (f+k)v, \quad (19)$$



**Fig. 12.** The two-dimensional principle subspace of the embedded vectors,  $\xi_i$ , from each tested embedding model for two different Reynolds numbers.



**Fig. 13.** 3D convolutional embedding network with leaky ReLU activation functions for the Gray–Scott system. Batch-normalization used between each of the convolutional layers. In the decoder, the feature maps are up-sampled before applying a standard 3D convolution.

in which  $u$  and  $v$  are the concentration of two species,  $r_u$  and  $r_v$  are their respective diffusion rates,  $k$  is the kill rate and  $f$  is the feed rate. This is a classical system of particular application to chemical processes as it models the following reaction:  $U + 2V \rightarrow 3V; V \rightarrow P$ . For a set region of feed and kill rates, this seemingly simple system can yield a wide range of complex dynamics (Lee, McCormick, Ouyang, & Swinney, 1993; Pearson, 1993). Hence, under the right settings, this system is an excellent case study to push the proposed methodology to its predictive limits. In this work, we will use the parameters:  $r_u = 0.2$ ,  $r_v = 0.1$ ,  $k = 0.055$  and  $f = 0.025$  which results in a complex dynamical reaction. Akin to the first numerical example, the initial condition of this system is stochastic such that the system is seeded with 3 randomly placed perturbations within the periodic domain.

Training, validation and testing data are obtained from a Runge–Kutta finite difference simulation on a structured grid,  $(\mathbf{u}, \mathbf{v}) \in \mathbb{R}^{2 \times 32 \times 32 \times 32}$ . The training, validation and test data sets consist of 512, 16 and 56 time-series, respectively, with 200 time-steps each at a physical time-step size of  $\Delta t = 5$ . A 3D

**Table 4**

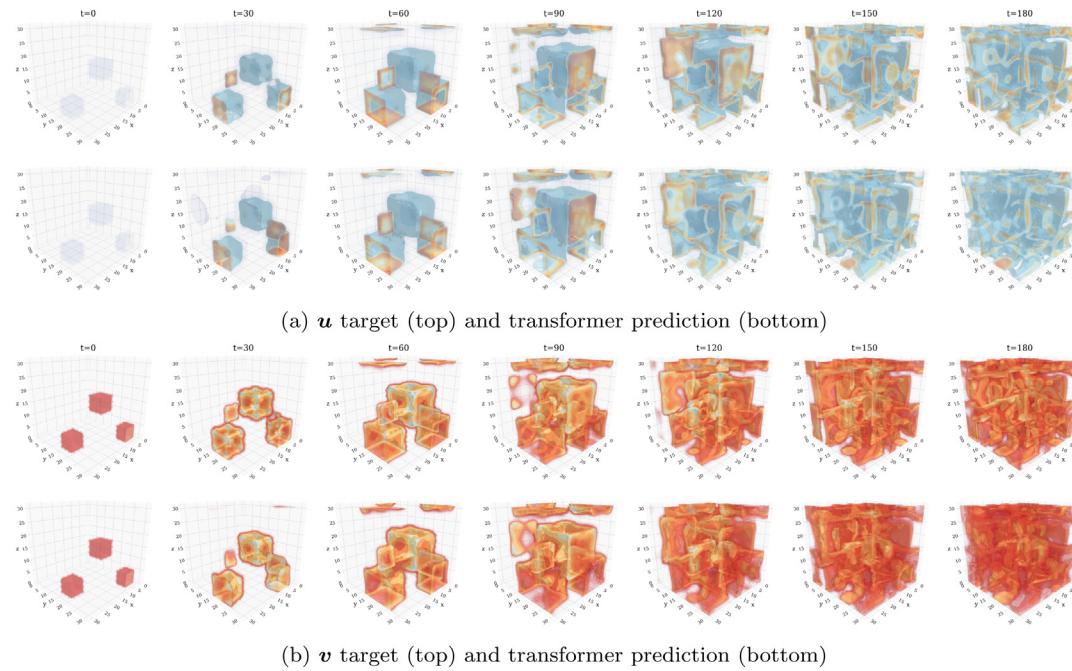
Test set relative mean-squared-error (MSE) for surrogate modeling 3D Gray–Scott system.

Model	Layers	Parameters	Relative MSE [0 – 200]	
			$u$	$v$
Transformer	2	6.2m/6.6m <sup>a</sup>	0.0159	0.0120
Transformer	4	6.2m/12.9m <sup>a</sup>	0.0154	0.0130
Transformer	8	6.2m/25.5m <sup>a</sup>	<b>0.0125</b>	<b>0.0101</b>

<sup>a</sup>Learnable parameters for the embedding/ transformer model.

convolutional encoder–decoder is used to embed the two species into a 512 embedding dimension,  $\mathcal{F} : \mathbb{R}^{2 \times 32 \times 32 \times 32} \rightarrow \mathbb{R}^{512}; \mathcal{G} : \mathbb{R}^{512} \rightarrow \mathbb{R}^{2 \times 32 \times 32 \times 32}$ , illustrated in Fig. 13.

Transformer models with varying depth are trained all with a context length of 128. All other model and training parameters for the transformer models are consistent. A test prediction using the transformer model is shown in Fig. 14 and the errors for



**Fig. 14.** Test case volume plots for the Gray–Scott system. Isosurfaces displayed span the range  $u, v = [0.3, 0.5]$  to show the inner structure.

each trained transformer are listed in [Table 4](#). Despite this system having complex dynamics in 3D space, the transformer is able to produce acceptable predictions with very similar structures as the numerical solver. To increase the model's predictive accuracy we believe the limitation here is not in the transformer, but rather the number of training data and the inaccuracies of the embedding model due to the dimensionality reduction needed. This is supported by the fact that increasing the transformer's depth does not yield considerable improvements for the test errors in [Table 4](#). Additional results are provided in [Appendix C](#).

#### 4. Conclusion

While transformers and self-attention models have been established as a powerful framework for NLP tasks, the adoption of such methodologies has yet to fully permeate other fields. In this work, we have demonstrated the potential transformers have for modeling dynamics of physical phenomena. The transformer architecture allows the model to learn longer and more complex temporal dependencies than alternative machine learning methods. Such models can be particularly beneficial for physical systems that exhibit dynamics that evolve on multiple time scales or consist of multiple phases such as turbulent fluid flow, chemical reactions or molecular dynamics. This can be attributed to the transformer's ability to draw information from many past time-steps directly with self-attention, learning accurate time integration without computationally expensive recurrent connections.

The key challenge of using the transformer model is identifying appropriate embeddings to represent the physical state of the system, for which we propose the use of Koopman dynamics to enforce dynamical context. Using the proposed methods, our transformer surrogate can outperform alternative models widely seen in recent scientific machine learning literature. The investigation of unsupervised pre-training of such transformer models as well as gaining a better understanding of what attention mechanisms imply for physical systems will be the subject of works in the future.

#### Declaration of competing interest

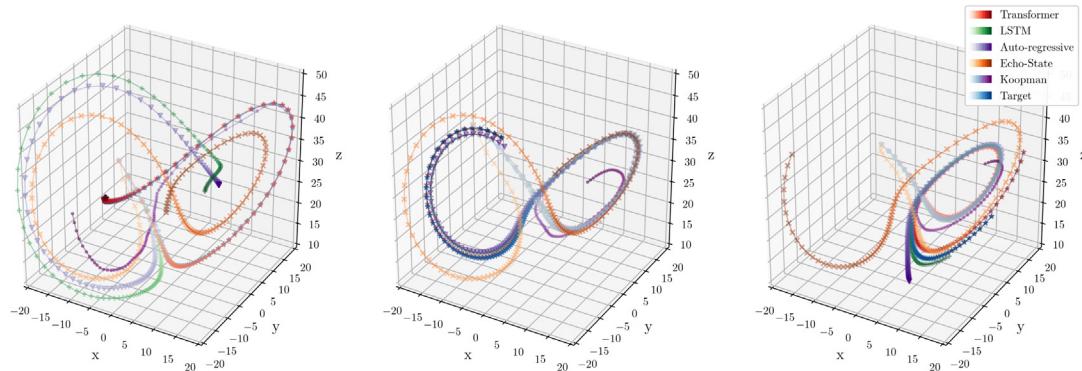
The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

#### Acknowledgments

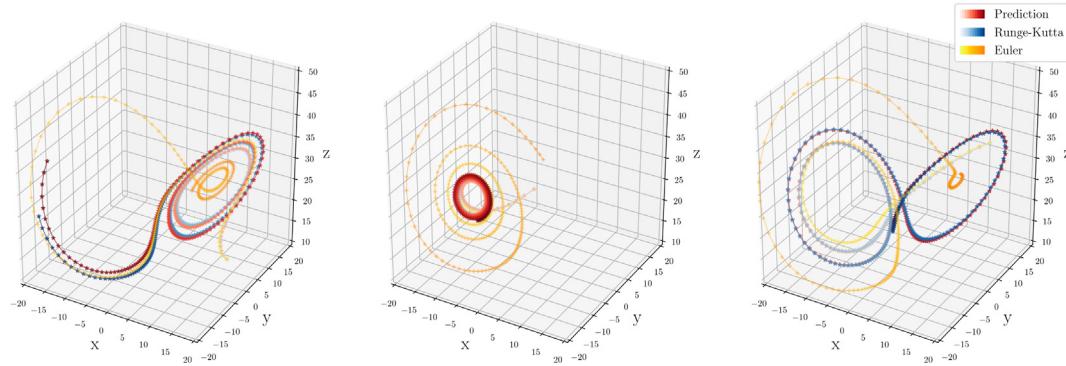
The anonymous reviewers are thanked for their significant effort in improving and clarifying this manuscript. The work reported here was initiated from the Defense Advanced Research Projects Agency (DARPA) under the Physics of Artificial Intelligence (PAI) program (contract HR00111890034). The authors acknowledge computing resources provided by the AFOSR Office of Scientific Research through the DURIP program and by the University of Notre Dame's Center for Research Computing (CRC). The work of NG was supported by the National Science Foundation (NSF), USA Graduate Research Fellowship Program grant No. DGE-1313583.

#### Appendix A. Lorenz supplementary results

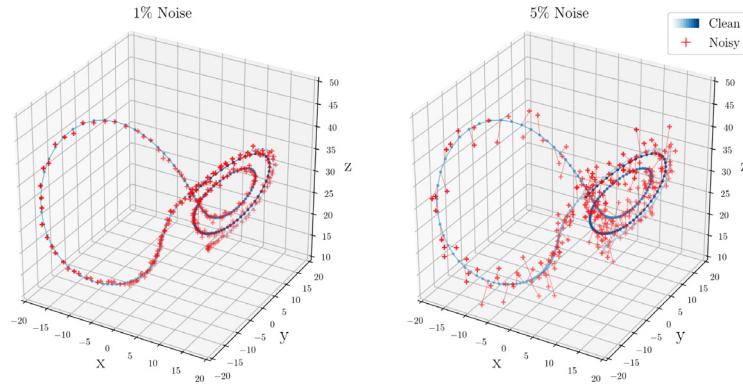
Predictions of the tested models for three test cases are plotted in [Fig. A.15](#). All machine learning methods are accurate for the first several time-steps, but begin to quickly deviate from the numerical solution in subsequent time-steps. The transformer model is consistently accurate within the plotted time frame, typically outperforming alternative methods at later time-steps. The predictions of the transformer model are also compared to the solution from a less accurate Euler time-integration method in [Fig. A.16](#). Due to the numerical differences between the Runge–Kutta ground truth and the Euler methods, the solutions are quick to deviate and the transformer clearly outperforms the Euler numerical simulator. Lastly, in [Fig. A.17](#) we illustrate the two different noise levels used to contaminate the training data for the results listed in [Table 2](#).



**Fig. A.15.** Three Lorenz test case predictions using each tested model for 128 time-steps.



**Fig. A.16.** Three Lorenz test solutions solved using Runge–Kutta and Euler time-integration methods with the prediction of the transformer for 256 time-steps.



**Fig. A.17.** Comparison between the clean and contaminated (noisy) training data.

## Appendix B. Cylinder supplementary results

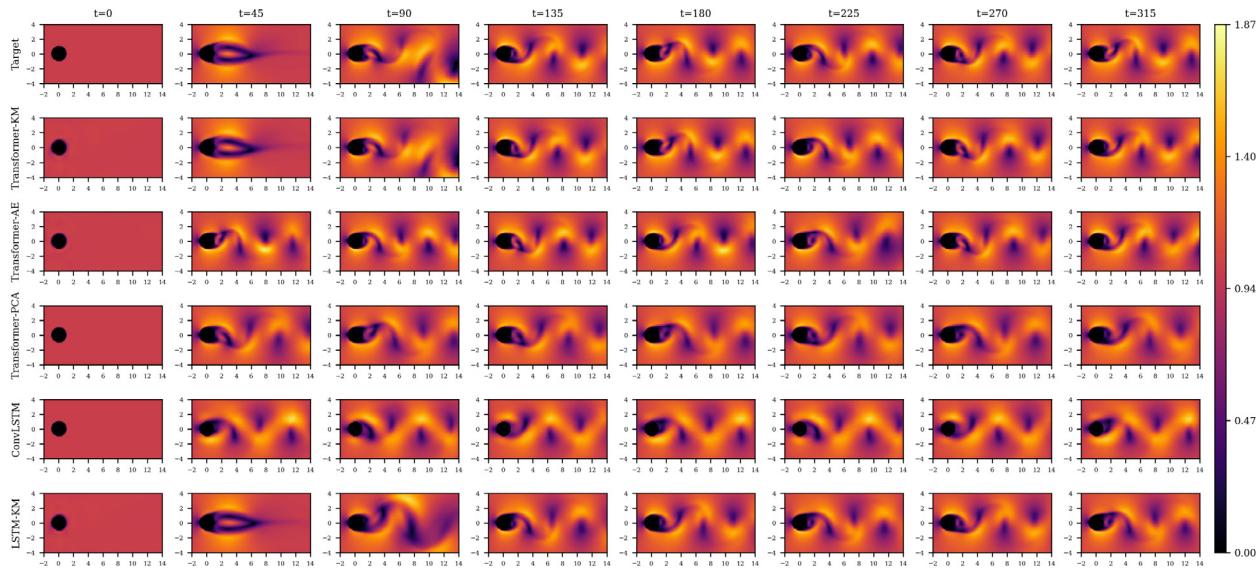
Prediction fields of all the tested models for a single test case are plotted in Figs. B.18 and B.19. While all models are able to perform adequately with qualitatively good results, the transformer model with Koopman embeddings (Transformer-KM) outperforms alternatives. Additionally, the training and validation errors during training are plotted in Fig. B.20 where all models have similar training and validation error indicating minimal overfitting.

The evolution of the flow field projected onto the dominant eigenvectors of the learned Koopman operator for two Reynolds numbers is plotted in Fig. B.21. This reflects the dynamical modes that were learned by the embedding model to impose physical “context” on the embeddings. Given that the eigenvectors are complex, we plot both the magnitude,  $|\psi|$ , and angle,  $\angle\psi$ . For both Reynolds numbers, it is easy to see the initial transition

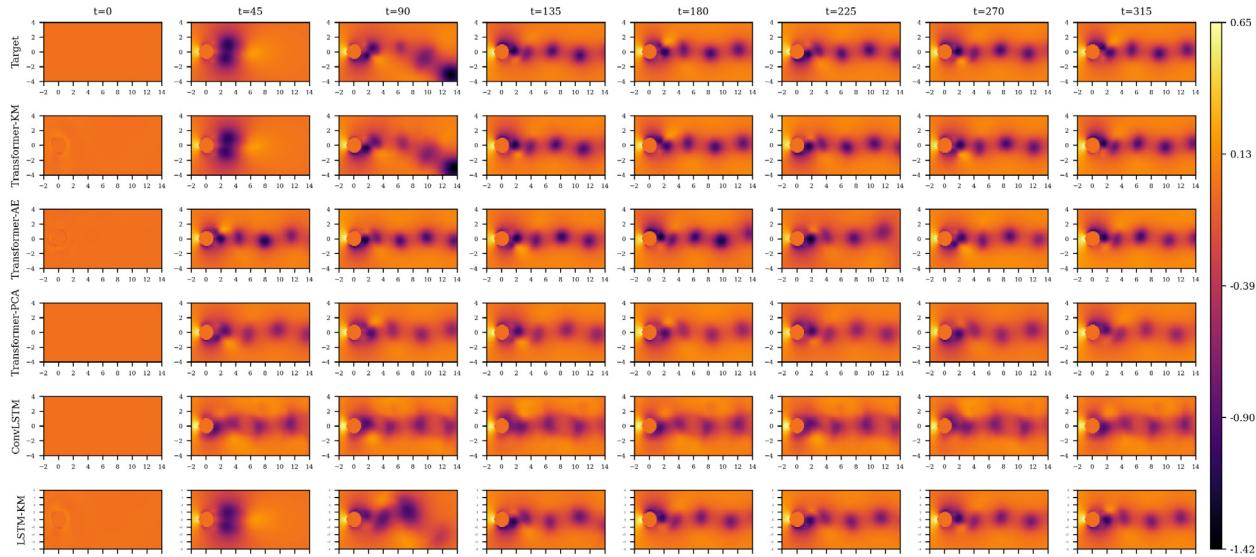
region,  $t < 100$ , before the system enters the periodic vortex shedding. Once in the periodic region, we can see that the higher Reynolds number has a higher frequency which reflects the increased vortex shedding speed.

## Appendix C. Gray–Scott supplementary results

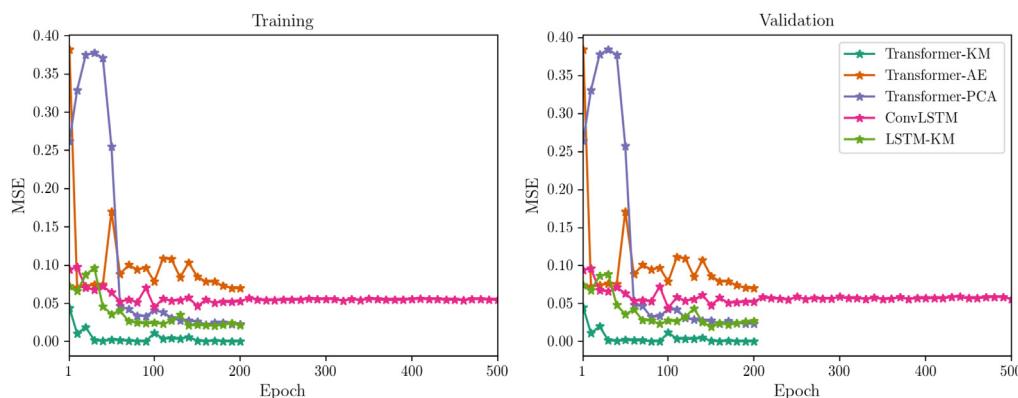
Volume plots of two test cases for both species are provided in Figs. C.22 & C.23. In addition to better visualize the accuracy of the trained transformer model, contour plots for three test cases are also provided in Fig. C.24. The transformer model is able to reliably predict the earlier time-steps with reasonable accuracy. As the system continues to react, the reaction fronts begin to interact resulting in the well-known complex structures of the Gray–Scott system (Lee et al., 1993; Pearson, 1993). During these later times, the transformer begins to degrade in accuracy yet does maintain notable consistency with the true solution.



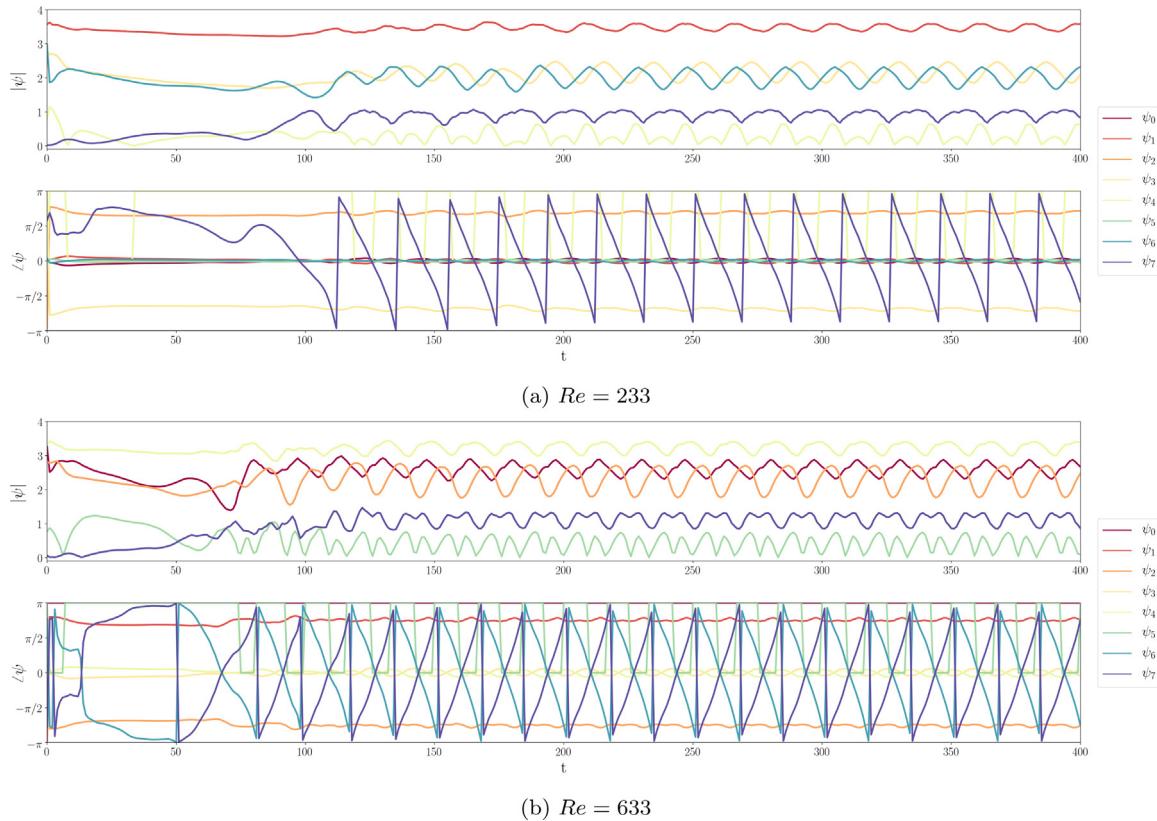
**Fig. B.18.** Velocity magnitude predictions of a test case at  $Re = 633$  using the transformer with Koopman (KM), auto-encoder (AE) and PCA embedding methods, the convolutional LSTM model (ConvLSTM) and fully-connected LSTM with Koopman embeddings (LSTM-KM).



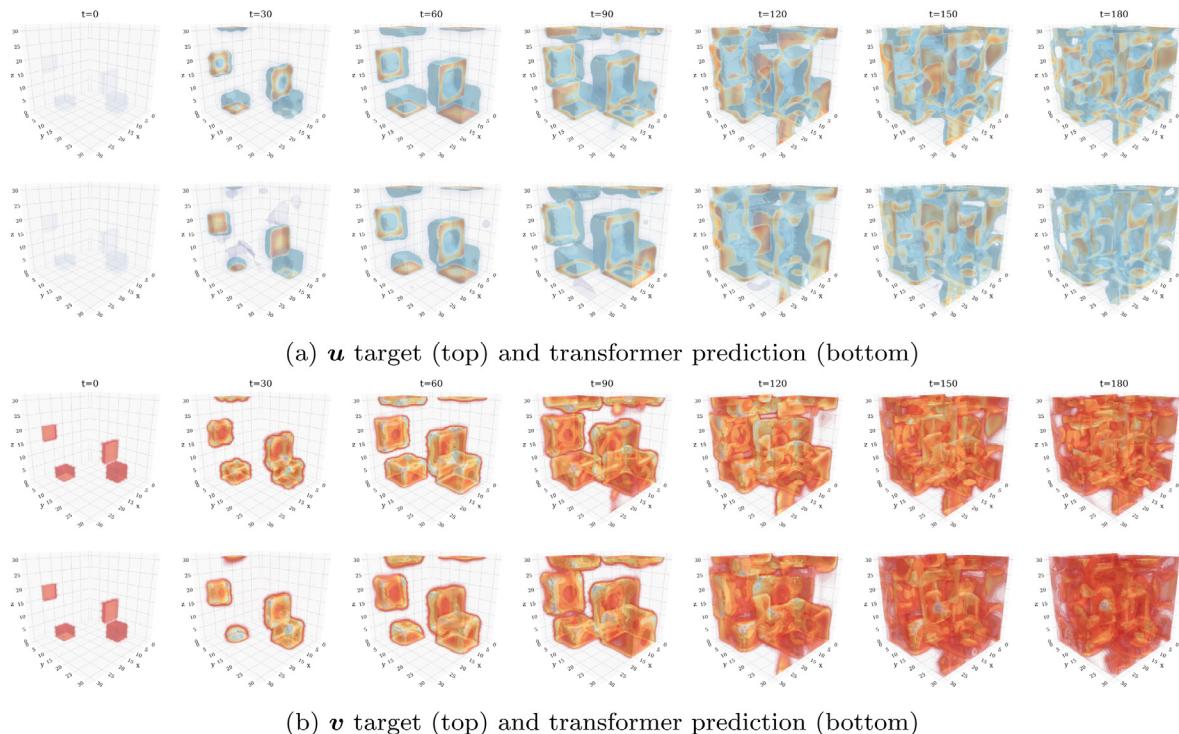
**Fig. B.19.** Pressure predictions of a test case at  $Re = 633$  using the transformer with Koopman (KM), auto-encoder (AE) and PCA embedding methods, the convolutional LSTM model (ConvLSTM) and fully-connected LSTM with Koopman embeddings (LSTM-KM).



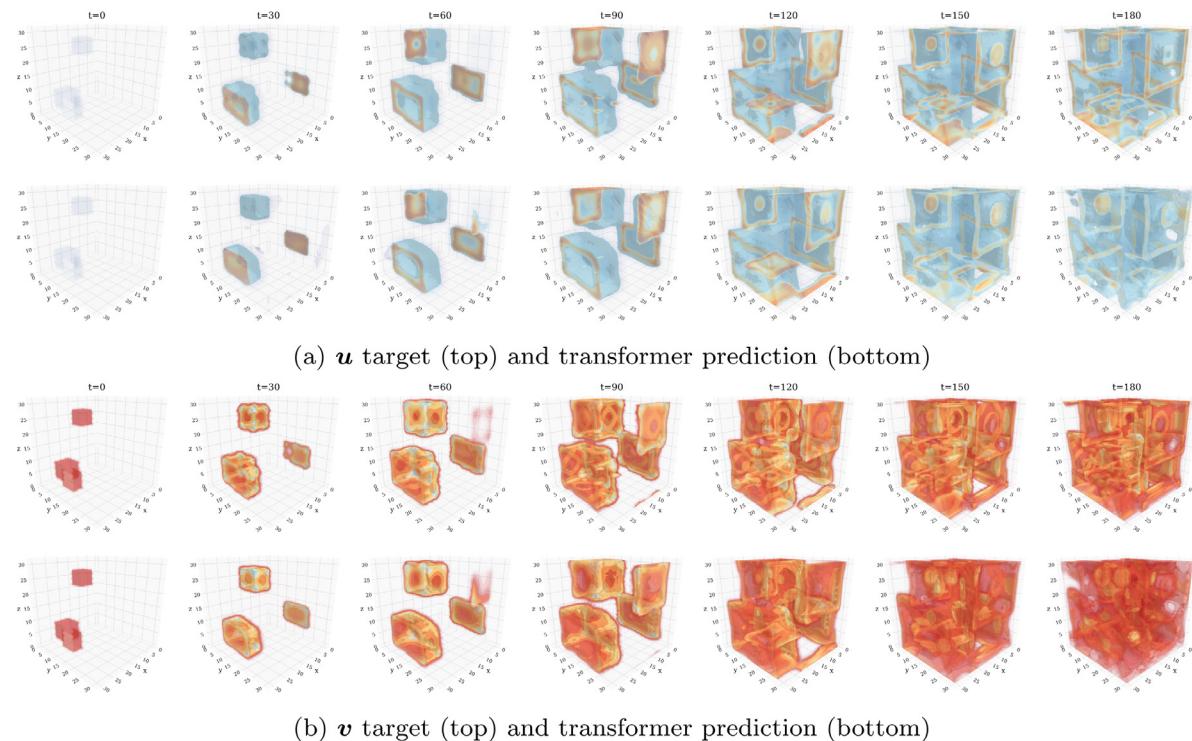
**Fig. B.20.** The training and validation mean square error (MSE) of the transformer with Koopman (KM), auto-encoder (AE) and PCA embedding methods, the convolutional LSTM model (ConvLSTM) and fully-connected LSTM with Koopman embeddings (LSTM-KM) during training.



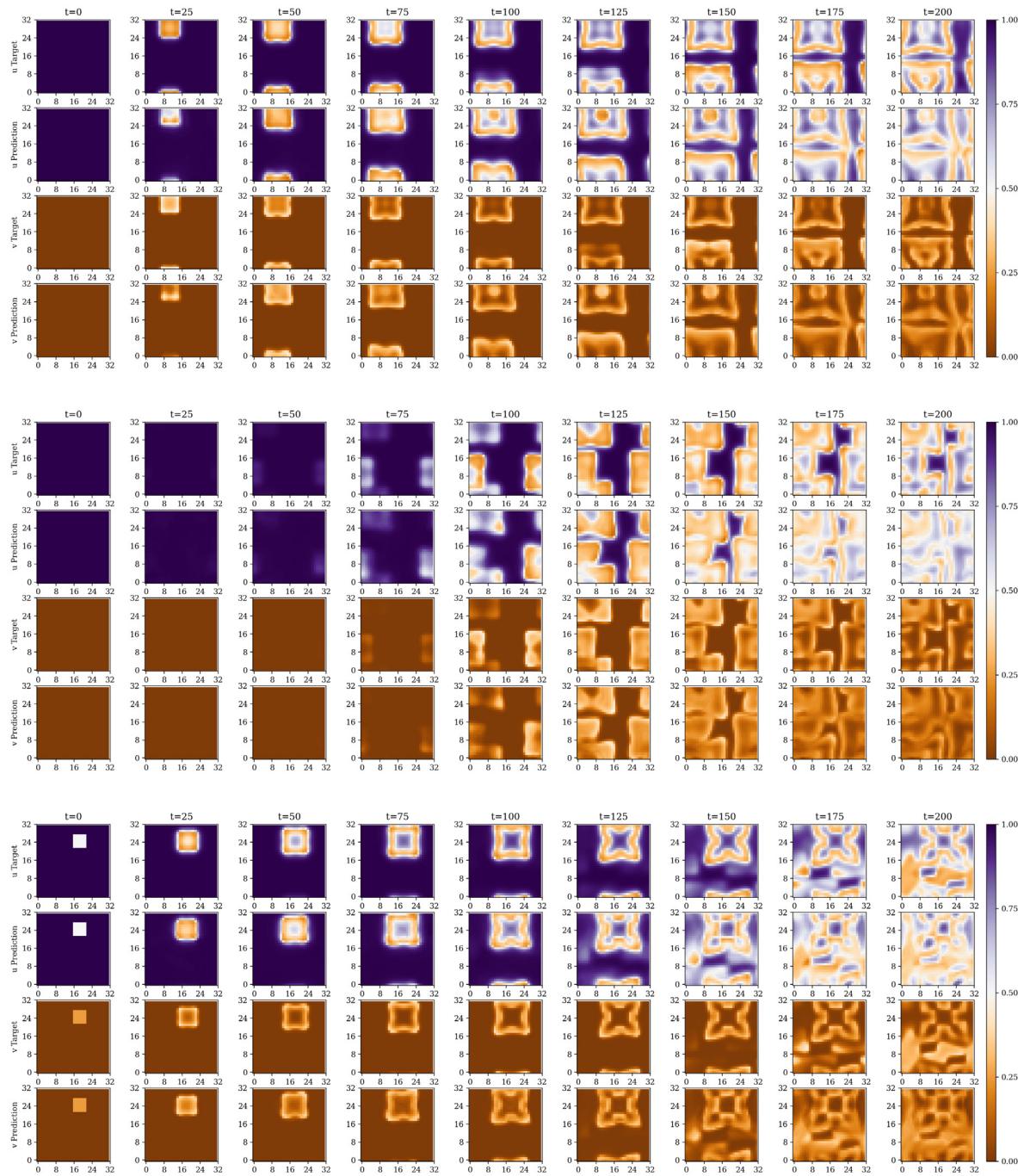
**Fig. B.21.** The dynamics of the fluid flow around a cylinder projected onto the 8 most dominant eigenvectors of the learned Koopman operator,  $\mathcal{K}$ , in the embedding model.



**Fig. C.22.** Test case volume plots for the Gray-Scott system. Isosurfaces displayed span the range  $u, v = [0.3, 0.5]$  to show the inner structure.



**Fig. C.23.** Test case volume plots for the Gray–Scott system. Isosurfaces displayed span the range  $\mathbf{u}, \mathbf{v} = [0.3, 0.5]$  to show the inner structure.



**Fig. C.24.**  $x$ - $y$  plane contour plots of three Gray-Scott test cases sliced at  $z = 16$ .

## References

- Atkinson, S., & Zabaras, N. (2019). Structured Bayesian Gaussian process latent variable model: Applications to data-driven dimensionality reduction and high-dimensional inversion. *Journal of Computational Physics*, 383, 166–195. <http://dx.doi.org/10.1016/j.jcp.2018.12.037>, URL <http://www.sciencedirect.com/science/article/pii/S0021999119300397>.
- Bahdanau, D., Cho, K., & Bengio, Y. (2015). Neural machine translation by jointly learning to align and translate. In *3rd International conference on learning representations*, arXiv:1409.0473.
- Bilionis, I., & Zabaras, N. (2016). Bayesian uncertainty propagation using Gaussian processes. In R. Ghanem, D. Higdon, & H. Owhadi (Eds.), *Handbook of uncertainty quantification* (pp. 1–45). Springer International Publishing, [http://dx.doi.org/10.1007/978-3-319-11259-6\\_16-1](http://dx.doi.org/10.1007/978-3-319-11259-6_16-1), URL [https://doi.org/10.1007/978-3-319-11259-6\\_16-1](https://doi.org/10.1007/978-3-319-11259-6_16-1).

Bilionis, I., Zabaras, N., Konomi, B. A., & Lin, G. (2013). Multi-output separable Gaussian process: Towards an efficient, fully Bayesian paradigm for uncertainty quantification. *Journal of Computational Physics*, 241, 212–239. <http://dx.doi.org/10.1016/j.jcp.2013.01.011>, URL <http://www.sciencedirect.com/science/article/pii/S0021999113000417>.

Brunton, S. L., Budišić, M., Kaiser, E., & Kutz, J. N. (2021). Modern Koopman theory for dynamical systems. arXiv preprint arXiv:2102.12086.

Chakraborty, S., & Zabaras, N. (2018). Efficient data-driven reduced-order models for high-dimensional multiscale dynamical systems. *Computer Physics Communications*, 230, 70–88. <http://dx.doi.org/10.1016/j.cpc.2018.04.007>, URL <http://www.sciencedirect.com/science/article/pii/S0010465518301176>.

Chattopadhyay, A., Hassanzadeh, P., Subramanian, D., & Palem, K. (2020). Data-driven prediction of a multi-scale Lorenz 96 chaotic system using a hierarchy of deep learning methods: Reservoir computing, ANN, and RNN-LSTM. *Nonlinear Processes in Geophysics*, 27, 373–389. <http://dx.doi.org/10.5194/npg-27-373-2020>, URL <https://doi.org/10.5194/npg-27-373-2020>.

- Chen, M., Radford, A., Child, R., Wu, J., Jun, H., Luan, D., et al. (2020). Generative pretraining from pixels. In H. D. III, A. Singh (Eds.), *Proceedings of machine learning research: 119, Proceedings of the 37th international conference on machine learning* (pp. 1691–1703). PMLR, URL <http://proceedings.mlr.press/v119/chen20s.html>.
- Chen, R. T. Q., Rubanova, Y., Bettencourt, J., & Duvenaud, D. K. (2018). Neural ordinary differential equations. In S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, & R. Garnett (Eds.), *31, Advances in neural information processing systems* (pp. 6571–6583). Curran Associates, Inc., URL <https://proceedings.neurips.cc/paper/2018/file/69386f6bb1dfed68692a24c8686939b-Paper.pdf>.
- Cybenko, G. (1989). Approximation by superpositions of a sigmoidal function. *Mathematics of Control, Signals, and Systems*, 2(4), 303–314. <http://dx.doi.org/10.1007/BF02551274>.
- Dai, Z., Yang, Z., Yang, Y., Carbonell, J. G., Le, Q., & Salakhutdinov, R. (2019). Transformer-XL: Attentive language models beyond a fixed-length context. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics* (pp. 2978–2988). <http://dx.doi.org/10.18653/v1/P19-1285>, URL <https://aclanthology.org/P19-1285>.
- Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019). Bert: Pre-training of deep bidirectional transformers for language understanding. In *NAACL-HLT* (pp. 4171–4186). Association for Computational Linguistics, <http://dx.doi.org/10.18653/v1/N19-1423>, URL <https://aclanthology.org/N19-1423>.
- Dupont, E., Doucet, A., & Teh, Y. W. (2019). Augmented neural ODEs. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d' Alché-Buc, E. Fox, & R. Garnett (Eds.), *32, Advances in neural information processing systems* (pp. 3140–3150). Curran Associates, Inc., URL <https://proceedings.neurips.cc/paper/2019/file/21be9a4bd4f81549a9d1d241981cec3c-Paper.pdf>.
- Fu, J., Liu, J., Tian, H., Li, Y., Bao, Y., Fang, Z., et al. (2019). Dual attention network for scene segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (pp. 3146–3154). URL [https://openaccess.thecvf.com/content\\_CVPR\\_2019/papers/Fu\\_Dual\\_Attention\\_Network\\_for\\_Scene\\_Segmentation\\_CVPR\\_2019\\_paper.pdf](https://openaccess.thecvf.com/content_CVPR_2019/papers/Fu_Dual_Attention_Network_for_Scene_Segmentation_CVPR_2019_paper.pdf).
- Gage, P. (1994). A new algorithm for data compression. 12, (2), (pp. 23–38). McPherson, KS: R & D Publications, c1987–1994, URL [https://www.derczynski.com/papers/archive/BPE\\_Gage.pdf](https://www.derczynski.com/papers/archive/BPE_Gage.pdf).
- Gao, H., Wang, J.-X., & Zahr, M. J. (2020). Non-intrusive model reduction of large-scale, nonlinear dynamical systems using deep learning. *Physica D: Nonlinear Phenomena*, 412, Article 132614. <http://dx.doi.org/10.1016/j.physd.2020.132614>, URL <http://www.sciencedirect.com/science/article/pii/S0167278919305573>.
- Geneva, N., & Zabaras, N. (2020a). Modeling the dynamics of PDE systems with physics-constrained deep auto-regressive networks. *Journal of Computational Physics*, 403, Article 109056. <http://dx.doi.org/10.1016/j.jcp.2019.109056>, URL <https://www.sciencedirect.com/science/article/pii/S0021999119307612>.
- Geneva, N., & Zabaras, N. (2020b). Multi-fidelity generative deep learning turbulent flows. *Foundations of Data Science*, 2, 391. <http://dx.doi.org/10.3934/fods.2020019>, URL <http://aimsces.org/article/id/3a9f3d14-3421-4947-a45f-a9cc74edd097>.
- González-García, R., Rico-Martínez, R., & Kevrekidis, I. (1998). Identification of distributed parameter systems: A neural net based approach. *Computers & Chemical Engineering*, 22, S965 – S968. [http://dx.doi.org/10.1016/S0098-1354\(98\)00191-4](http://dx.doi.org/10.1016/S0098-1354(98)00191-4), European Symposium on Computer Aided Process Engineering-8, URL <http://www.sciencedirect.com/science/article/pii/S0098135498001914>.
- Graves, A., Wayne, G., & Danihelka, I. (2014). Neural turing machines. arXiv preprint [arXiv:1410.5401](https://arxiv.org/abs/1410.5401).
- Han, R., Wang, Y., Zhang, Y., & Chen, G. (2019). A novel spatial-temporal prediction method for unsteady wake flows based on hybrid deep neural network. *Physics of Fluids*, 31(12), Article 127101. <http://dx.doi.org/10.1063/1.5127247>.
- Hornik, K. (1991). Approximation capabilities of multilayer feedforward networks. *Neural Networks*, 4(2), 251–257. [http://dx.doi.org/10.1016/0893-6080\(91\)90009-T](http://dx.doi.org/10.1016/0893-6080(91)90009-T), URL <https://www.sciencedirect.com/science/article/pii/089360809190009T>.
- Hornik, K., Stinchcombe, M., & White, H. (1989). Multilayer feedforward networks are universal approximators. *Neural Networks*, 2(5), 359–366. [http://dx.doi.org/10.1016/0893-6080\(89\)90020-8](http://dx.doi.org/10.1016/0893-6080(89)90020-8), URL <https://www.sciencedirect.com/science/article/pii/0893608089900208>.
- Jasak, H., Jemcov, A., Tukovic, Z., et al. (2007). OpenFOAM: A C++ Library for complex physics simulations. 1000, In *International workshop on coupled methods in numerical dynamics* (pp. 1–20). IUC Dubrovnik Croatia.
- Kitaev, N., Kaiser, L., & Levskaya, A. (2020). Reformer: The efficient transformer. In *International Conference on Learning Representations*. URL <https://openreview.net/forum?id=rkgNKKhTvB>.
- Koopman, B. O. (1931). Hamiltonian systems and transformation in Hilbert space. *Proceedings of the National Academy of Sciences of the United States of America*, 17(5), 315. <http://dx.doi.org/10.1073/pnas.17.5.315>.
- Korda, M., & Mezić, I. (2018). Linear predictors for nonlinear dynamical systems: Koopman operator meets model predictive control. *Automatica*, 93, 149–160. <http://dx.doi.org/10.1016/j.automatica.2018.03.046>, URL <http://www.sciencedirect.com/science/article/pii/S000510981830133X>.
- Korda, M., Putinar, M., & Mezić, I. (2020). Data-driven spectral analysis of the Koopman operator. *Applied and Computational Harmonic Analysis*, 48(2), 599–629. <http://dx.doi.org/10.1016/j.acha.2018.08.002>, URL <http://www.sciencedirect.com/science/article/pii/S1063520318300988>.
- Lee, K. J., McCormick, W. D., Ouyang, Q., & Swinney, H. L. (1993). Pattern formation by interacting chemical fronts. *Science*, 261(5118), 192–194. <http://dx.doi.org/10.1126/science.261.5118.192>, URL <https://science.sciencemag.org/content/261/5118/192>.
- Lee, S., & You, D. (2017). Prediction of laminar vortex shedding over a cylinder using deep learning. arXiv preprint [arXiv:1712.07854](https://arxiv.org/abs/1712.07854).
- Li, Q., Dietrich, F., Boltt, E. M., & Kevrekidis, I. G. (2017). Extended dynamic mode decomposition with dictionary learning: A data-driven adaptive spectral decomposition of the Koopman operator. *Chaos. An Interdisciplinary Journal of Nonlinear Science*, 27(10), Article 103111. <http://dx.doi.org/10.1063/1.4993854>.
- Li, Y., He, H., Wu, J., Kataib, D., & Torralba, A. (2020). Learning compositional Koopman operators for model-based control. In *International Conference on Learning Representations*. URL <https://openreview.net/forum?id=H1ldzA4tPr>.
- Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., et al. (2019). Roberta: A robustly optimized bert pretraining approach. arXiv preprint [arXiv:1907.11692](https://arxiv.org/abs/1907.11692).
- Lu, Y., Zhong, A., Li, Q., & Dong, B. (2018). Beyond finite layer neural networks: Bridging deep architectures and numerical differential equations. In *International conference on machine learning* (pp. 3276–3285). PMLR.
- Lukoševičius, M. (2012). A practical guide to applying echo state networks. In *Neural Networks: Tricks of the Trade* (pp. 659–686). Springer, [http://dx.doi.org/10.1007/978-3-642-35289-8\\_36](http://dx.doi.org/10.1007/978-3-642-35289-8_36).
- Luong, T., Pham, H., & Manning, C. D. (2015). Effective approaches to attention-based neural machine translation. In *EMNLP* (pp. 1412–1421). [arXiv:508.04025](https://arxiv.org/abs/1508.04025), URL <http://aclweb.org/anthology/D/D15/D15-1166.pdf>.
- Lusch, B., Kutz, J. N., & Brunton, S. L. (2018). Deep learning for universal linear embeddings of nonlinear dynamics. *Nature Communications*, 9(1), 1–10. <http://dx.doi.org/10.1038/s41467-018-07210-0>.
- Maulik, R., Egele, R., Lusch, B., & Balaprakash, P. (2020). Recurrent neural network architecture search for geophysical emulation. In *SC20: International conference for high performance computing, networking, storage and analysis* (pp. 1–14). <http://dx.doi.org/10.1109/SC41405.2020.00012>, URL <https://doi.org/10.1109/SC41405.2020.00012>.
- Maulik, R., Lusch, B., & Balaprakash, P. (2021). Reduced-order modeling of advection-dominated systems with recurrent neural networks and convolutional autoencoders. *Physics of Fluids*, 33(3), Article 037106. <http://dx.doi.org/10.1063/5.0039986>.
- Melamud, O., Goldberger, J., & Dagan, I. (2016). Context2vec: Learning generic context embedding with bidirectional LSTM. In *Proceedings of the 20th SIGNLL Conference on Computational Natural Language Learning* (pp. 51–61). URL <https://aclanthology.org/K16-1006.pdf>.
- Mezic, I. (2020). On numerical approximations of the koopman operator. arXiv preprint [arXiv:2009.05883](https://arxiv.org/abs/2009.05883).
- Mikolov, T., Chen, K., Corrado, G., & Dean, J. (2013). Efficient estimation of word representations in vector space. In *Workshop proceedings international conference on learning representations*.
- Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S., & Dean, J. (2013). Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems* (pp. 3111–3119). URL <https://proceedings.neurips.cc/paper/2013/file/9aa42b1882ec039965f3c4923ce901b-Paper.pdf>.
- Mo, S., Zhu, Y., Zabaras, N., Shi, X., & Wu, J. (2019). Deep convolutional encoder-decoder networks for uncertainty quantification of dynamic multiphase flow in heterogeneous media. *Water Resources Research*, 55(1), 703–728. <http://dx.doi.org/10.1029/2018WR023528>.
- Morton, J., Jameson, A., Kochenderfer, M. J., & Witherden, F. (2018). Deep dynamical modeling and control of unsteady fluid flows. In S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, & R. Garnett (Eds.), *31, Advances in neural information processing systems*. Curran Associates, Inc., URL <https://proceedings.neurips.cc/paper/2018/file/2b0aa0d9e30ea3a55fc271ced8364536-Paper.pdf>.
- Otto, S. E., & Rowley, C. W. (2019). Linearly recurrent autoencoder networks for learning dynamics. *SIAM Journal on Applied Dynamical Systems*, 18(1), 558–593. <http://dx.doi.org/10.1137/18M1177846>.
- Pearson, J. E. (1993). Complex patterns in a simple system. *Science*, 261(5118), 189–192. <http://dx.doi.org/10.1126/science.261.5118.189>, URL <https://science.sciencemag.org/content/261/5118/189>.
- Pennington, J., Socher, R., & Manning, C. D. (2014). Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)* (pp. 1532–1543).

- Peters, M. E., Neumann, M., Iyyer, M., Gardner, M., Clark, C., Lee, K., et al. (2018). Deep contextualized word representations. In *Proceedings of the 2018 conference of the North American chapter of the association for computational linguistics: Human language technologies, volume 1 (long papers)* (pp. 2227–2237). Association for Computational Linguistics, <http://dx.doi.org/10.18653/v1/N18-1202>, URL <https://aclanthology.org/N18-1202>.
- Radford, A., Narasimhan, K., Salimans, T., & Sutskever, I. (2018). Improving language understanding by generative pre-training. *OpenAI Blog*, URL [https://openai-assets.s3.amazonaws.com/research-covers/language-unsupervised/language\\_understanding\\_paper.pdf](https://openai-assets.s3.amazonaws.com/research-covers/language-unsupervised/language_understanding_paper.pdf).
- Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., & Sutskever, I. (2019). Language models are unsupervised multitask learners. *OpenAI Blog*, URL [https://cdn.openai.com/better-language-models/language\\_models\\_are\\_unsupervised\\_multitask\\_learners.pdf](https://cdn.openai.com/better-language-models/language_models_are_unsupervised_multitask_learners.pdf).
- Sanchez-Gonzalez, A., Godwin, J., Pfaff, T., Ying, R., Leskovec, J., & Battaglia, P. (2020). Learning to simulate complex physics with graph networks. In H. D. III, & A. Singh (Eds.), *Proceedings of machine learning research: 119, Proceedings of the 37th international conference on machine learning* (pp. 8459–8468). PMLR, URL <http://proceedings.mlr.press/v119/sanchez-gonzalez20a.html>.
- Shalova, A., & Oseledets, I. (2020). Tensorized transformer for dynamical systems modeling. arXiv preprint [arXiv:2006.03445](https://arxiv.org/abs/2006.03445).
- Shi, X., Chen, Z., Wang, H., Yeung, D.-Y., Wong, W.-k., & WOO, W.-c. (2015). Convolutional LSTM network: A machine learning approach for precipitation nowcasting. In C. Cortes, N. Lawrence, D. Lee, M. Sugiyama, & R. Garnett (Eds.), *28, Advances in neural information processing systems*. Curran Associates, Inc., URL <https://proceedings.neurips.cc/paper/2015/file/07563a3fe3bbe7e3ba84431ad9d055af-Paper.pdf>.
- Stoer, J., & Bulirsch, R. (2013). vol. 12, *Introduction to numerical analysis*. Springer, <http://dx.doi.org/10.1007/978-1-4757-2272-7>, URL <https://doi.org/10.1007/978-1-4757-2272-7>.
- Sukhbaatar, S., Grave, E., Bojanowski, P., & Joulin, A. (2019). Adaptive attention span in transformers. In *Proceedings of the 57th annual meeting of the association for computational linguistics* (pp. 331–335). Florence, Italy: Association for Computational Linguistics, <http://dx.doi.org/10.18653/v1/P19-1032>, URL <https://aclanthology.org/P19-1032>.
- Sukhbaatar, S., Grave, E., Lample, G., Jegou, H., & Joulin, A. (2019). Augmenting self-attention with persistent memory. arXiv preprint [arXiv:1907.01470](https://arxiv.org/abs/1907.01470).
- Takeishi, N., Kawahara, Y., & Yairi, T. (2017). Learning Koopman invariant subspaces for dynamic mode decomposition. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, & R. Garnett (Eds.), *30, Advances in neural information processing systems*. Curran Associates, Inc., URL <https://proceedings.neurips.cc/paper/2017/file/3a835d3215755c435ef4fe9965a3f2a0-Paper.pdf>.
- Tanaka, G., Yamane, T., Héroux, J. B., Nakane, R., Kanazawa, N., Takeda, S., et al. (2019). Recent advances in physical reservoir computing: A review. *Neural Networks*, *115*, 100–123. <http://dx.doi.org/10.1016/j.neunet.2019.03.005>, URL <http://www.sciencedirect.com/science/article/pii/S0893608019300784>.
- Tang, M., Liu, Y., & Durlofsky, L. J. (2020). A deep-learning-based surrogate model for data assimilation in dynamic subsurface flow problems. *Journal of Computational Physics*, *413*, Article 109456. <http://dx.doi.org/10.1016/j.jcp.2020.109456>, URL <https://www.sciencedirect.com/science/article/pii/S0021999120302308>.
- Tripathy, R. K., & Bilionis, I. (2018). Deep UQ: Learning deep neural network surrogate models for high dimensional uncertainty quantification. *Journal of Computational Physics*, *375*, 565–588. <http://dx.doi.org/10.1016/j.jcp.2018.08.036>, URL <http://www.sciencedirect.com/science/article/pii/S0021999118305655>.
- Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., et al. (2017). Attention is all you need. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, R. Garnett (Eds.), *30, Advances in neural information processing systems*. Curran Associates, Inc., URL <https://proceedings.neurips.cc/paper/2017/file/3f5ee243547dee91fb053c1c4a845aa-Paper.pdf>.
- Veličković, P., Cucurull, G., Casanova, A., Romero, A., Liò, P., & Bengio, Y. (2018). Graph attention networks. In *International conference on learning representations*, URL <https://openreview.net/forum?id=ijXMpikCZ>.
- Wang, Y.-J., & Lin, C.-T. (1998). Runge-kutta neural network for identification of dynamical systems in high accuracy. *IEEE Transactions on Neural Networks*, *9*(2), 294–307. <http://dx.doi.org/10.1109/72.661124>.
- Wessels, H., Weißfels, C., & Wriggers, P. (2020). The neural particle method – an updated Lagrangian physics informed neural network for computational fluid dynamics. *Computer Methods in Applied Mechanics and Engineering*, *368*, Article 113127. <http://dx.doi.org/10.1016/j.cma.2020.113127>, URL <https://www.sciencedirect.com/science/article/pii/S0045782520303121>.
- Wiewel, S., Becher, M., & Thürey, N. (2019). Latent space physics: Towards learning the temporal evolution of fluid flow. *Computer Graphics Forum*, *38*(2), 71–82. <http://dx.doi.org/10.1111/cgf.13620>, URL <https://onlinelibrary.wiley.com/doi/abs/10.1111/cgf.13620>.
- Wolf, T., Debut, L., Sanh, V., Chaumond, J., Delangue, C., Moi, A., et al. (2019). HuggingFace's Transformers: State-of-the-art natural language processing. arXiv preprint [arXiv:1910.03771](https://arxiv.org/abs/1910.03771).
- Xiu, D., & Karniadakis, G. E. (2002). The Wiener-Askey polynomial chaos for stochastic differential equations. *SIAM Journal on Scientific Computing*, *24*(2), 619–644. <http://dx.doi.org/10.1137/S1064827501387826>.
- Xu, J., & Duraisamy, K. (2020). Multi-level convolutional autoencoder networks for parametric prediction of spatio-temporal dynamics. *Computer Methods in Applied Mechanics and Engineering*, *372*, Article 113379. <http://dx.doi.org/10.1016/j.cma.2020.113379>, URL <https://www.sciencedirect.com/science/article/pii/S0045782520305648>.
- Zhang, H., Goodfellow, I., Metaxas, D., & Odena, A. (2019). Self-attention generative adversarial networks. In K. Chaudhuri, & R. Salakhutdinov (Eds.), *Proceedings of machine learning research: 97, Proceedings of the 36th international conference on machine learning* (pp. 7354–7363). PMLR, URL <http://proceedings.mlr.press/v97/zhang19d.html>.
- Zhao, J., Deng, F., Cai, Y., & Chen, J. (2019). Long short-term memory – fully connected (LSTM-FC) neural network for PM2.5 concentration prediction. *Chemosphere*, *220*, 486–492. <http://dx.doi.org/10.1016/j.chemosphere.2018.12.128>, URL <https://www.sciencedirect.com/science/article/pii/S0045653518324639>.
- Zhu, M., Chang, B., & Fu, C. (2018). Convolutional neural networks combined with Runge-Kutta methods. arXiv preprint [arXiv:1802.08831](https://arxiv.org/abs/1802.08831).
- Zhu, Y., & Zabaras, N. (2018). Bayesian deep convolutional encoder-decoder networks for surrogate modeling and uncertainty quantification. *Journal of Computational Physics*, *366*, 415–447. <http://dx.doi.org/10.1016/j.jcp.2018.04.018>, URL <http://www.sciencedirect.com/science/article/pii/S0021999118302341>.