Combustion and Flame 259 (2024) 113192

Contents lists available at ScienceDirect

Combustion and Flame

journal homepage: www.sciencedirect.com/journal/combustion-and-flame

A co-kurtosis PCA based dimensionality reduction with nonlinear reconstruction using neural networks

Dibyajyoti Nayak [a], Anirudh Jonnalagadda [a], Uma Balakrishnan [b], Hemanth Kolla [b], Konduri Aditya [a,*]

[a] Department of Computational and Data Sciences, Indian Institute of Science, Bangalore, India
[b] Sandia National Laboratories, Livermore, CA, USA

ARTICLE INFO

Keywords:
Dimensionality reduction
Principal component analysis
Co-kurtosis tensor
Deep neural networks
Reconstruction

ABSTRACT

For turbulent reacting flow systems, identification of low-dimensional representations of the thermo-chemical state space is vitally important, primarily to significantly reduce the computational cost of device-scale simulations. Principal component analysis (PCA), and its variants, are a widely employed class of methods. Recently, an alternative technique that focuses on higher-order statistical interactions, co-kurtosis PCA (CoK-PCA), has been shown to effectively provide a low-dimensional representation by capturing the stiff chemical dynamics associated with spatiotemporally localized reaction zones. While its effectiveness has only been demonstrated based on *a priori* analyses with linear reconstruction, in this work, we employ nonlinear techniques to reconstruct the full thermo-chemical state and evaluate the efficacy of CoK-PCA compared to PCA. Specifically, we combine a CoK-PCA-/PCA-based dimensionality reduction (encoding) with an artificial neural network (ANN) based reconstruction (decoding) and examine, *a priori*, the reconstruction errors of the thermo-chemical state. In addition, we evaluate the errors in species production rates and heat release rates, which are nonlinear functions of the reconstructed state, as a measure of the overall accuracy of the dimensionality reduction technique. We employ four datasets to assess CoK-PCA/PCA coupled with ANN-based reconstruction: zero-dimensional (homogeneous) reactor for autoignition of an ethylene/air mixture that has conventional single-stage ignition kinetics, a dimethyl ether (DME)/air mixture which has two-stage (low and high temperature) ignition kinetics, a one-dimensional freely propagating premixed ethylene/air laminar flame, and a two-dimensional dataset representing turbulent autoignition of ethanol in a homogeneous charge compression ignition (HCCI) engine. Results from the analyses demonstrate the robustness of the CoK-PCA based low-dimensional manifold with ANN reconstruction in accurately capturing the data, specifically from the reaction zones.

Novelty and significance

The co-kurtosis PCA (CoK-PCA) based encoding method has been shown to provide a more accurate reduced manifold relative to PCA in capturing stiff dynamics relevant to combustion datasets. However, contrary to PCA, the efficacy of the co-kurtosis PCA technique has not been explored in conjunction with widely used nonlinear reconstruction methods (decoding). This paper provides a rigorous performance analysis of CoK-PCA with an artificial neural network (ANN) based reconstruction for diverse combustion datasets. Further, the accuracy of both PCA and CoK-PCA in capturing the complex kinetics of two-stage autoignition is evaluated. Such *a priori* analyses performed in this study are foundational towards establishing the accuracy of dimensionality reduction techniques and their implementation into reacting flow solvers for significantly reducing computational costs.

1. Introduction

The multi-scale, multi-physics nature of turbulent reacting flows necessitates the use of high-fidelity simulations to accurately model chemical kinetics and turbulence–chemistry interactions. However, when representing chemical kinetics using first principles, e.g., direct numerical simulations with detailed kinetics, the governing system of equations has large dimensionality due to tens of chemical species participating in hundreds of chemical reactions [1–5]. As a result, the computational costs become prohibitively expensive for simulations of practical device-scale problems. Indeed, as the chemistry calculations associated with even the simplest of reaction mechanisms present themselves as the main driver of the large computational cost [6], reduced order modeling techniques become invaluable.

* Corresponding author.
E-mail address: konduriadi@iisc.ac.in (K. Aditya).

https://doi.org/10.1016/j.combustflame.2023.113192
Received 14 July 2023; Received in revised form 3 November 2023; Accepted 5 November 2023
Available online 9 November 2023

With the advent of data-driven techniques, low-dimensional manifold (LDM) representations of the thermo-chemical state space, identified from relevant training data, can effectively model the species dynamics of an otherwise large chemical system. Among the various available strategies to obtain these LDMs, principal component analysis (PCA) and its many flavors have been most widely employed [7–12]. However, the principal components obtained by PCA are optimized with respect to second-order joint statistical moment, i.e., covariance, of the training data and may not be sensitive to the presence of extreme-valued samples characteristic of localized spatiotemporal events such as the formation of ignition kernels [13]. In contrast, the statistical signature of such events is shown to be favorably captured by principal components of higher-order joint statistical moments, specifically the fourth-order co-kurtosis tensor [13]. Building upon this observation, a dimensionality reduction procedure that constructs LDMs represented by principal components of the co-kurtosis tensor, namely the co-kurtosis PCA (CoK-PCA) method, was proposed [14]. Additionally, analogous to PCA, a recently proposed online low-rank approximation algorithm known as dynamically bi-orthogonal decomposition (DBO), which is based on time-dependent low-dimensional subspaces, has been shown to effectively characterize strongly transient events in turbulent compressible reacting flows [15].

It is noteworthy that, while the CoK-PCA method was shown to represent the thermo-chemical state as well as nonlinear functions of the thermo-chemical state, such as species production rates (PRs) and heat release rates (HRRs), better than PCA in the localized spatiotemporal regions corresponding to strong chemical activity, the transformation from the principal components of the LDM to the full thermo-chemical state was performed through linear operators [14]. However, due to the inherent nonlinear nature of the combustion phenomenon, the use of linear reconstruction has long been known not to be sufficiently accurate. Thus, the main objective of the present study is to address these concerns by studying the CoK-PCA method with nonlinear reconstruction techniques and comparing the accuracy relative to both PCA and a simple linear reconstruction. While performing an *a priori* validation is one way of assessing the efficacy of CoK-PCA relative to PCA, recent studies have proposed various other techniques that focus on analyzing the topology and uniqueness of the LDMs to examine their quality [16–18].

For PCA-based LDMs, several studies have explored nonlinear reconstruction techniques such as artificial neural networks (ANNs), kernel methods, Gaussian process regression (GPR), and their hybrid approaches [19–23]. Nonlinear reconstruction using ANN models provides flexibility to capture complex relationships, scalability for large datasets, meaningful representation learning, robustness to noise and irregularities, and the ability to generalize well to unseen data [24]. Therefore, within the confines of this paper, our primary emphasis is directed towards nonlinear reconstruction utilizing ANNs. In this study, we compare the reconstruction performance of ANNs with linear methods [14] and subsequently aim to evaluate the efficacy and superiority of nonlinear approaches in accurately capturing and predicting important combustion variables. Following Jonnalagadda et al. [14], the quality of the CoK-PCA-based/PCA-based encoder and ANN-based decoder models, hereafter called the CoK-PCA-ANN and PCA-ANN models, respectively, are compared via the conventionally considered reconstruction errors of the thermo-chemical scalars as well as more sensitive PRs and HRRs for four combustion datasets namely premixed ethylene-air in a homogeneous reactor, two-stage autoignition of dimethyl ether (DME)-air, a one-dimensional freely-propagating planar laminar flame of premixed ethylene-air, and a homogeneous charge compression ignition of ethanol-air mixture. To reiterate, the primary objective of this work is to establish, in an *a priori* setting, the effectiveness of the CoK-PCA-ANN method in capturing the thermo-chemical signature of data from regions exhibiting strong chemical activity, such as reaction zones/ignition kernels. Much like the work presented by Abdelwahid et al. [25] and Kumar et al. [26], the critical application of this technique would be to incorporate the CoK-PCA transformation to solve a reduced set of the CoK-PCA principal components transport equations, which not only accelerates computations but is also expected to provide better representations of the reaction zones.

The remainder of this paper is organized as follows. In Section 2, we briefly illustrate the dimensionality reduction procedure and outline the PCA and the CoK-PCA methods to obtain the low-dimensional manifolds (LDMs). Section 3 describes the artificial neural network (ANN) based nonlinear reconstruction procedure to predict the thermo-chemical scalars from the principal components of the LDMs. The results from the *a priori* analyses to evaluate the performance of the two LDMs based on ANN reconstruction are presented in Section 4. Finally, we summarize the paper and provide future directions in Section 5.

2. Dimensionality reduction

Following convention, we arrange the scaled training data as a matrix $\mathbf{X} \in \mathbb{R}^{(n_g \times n_v)}$ with n_g observations (e.g., spatial locations, temporal checkpoints) each having n_v real-valued variables or features (e.g., species concentrations, temperature). With respect to the feature space, $\mathbf{X}$ can be represented in terms of column vectors as $\mathbf{X} = \left\{ x_i \in \mathbb{R}^{(n_g \times 1)} \; \forall \; i \in \{1, \dots, n_v\} \right\}$. The purpose of dimensionality reduction, within the context of combustion, is to find a column subspace of dimension $n_q < n_v$, representing an LDM of the feature space by some measure of optimality. Note that dimensionality reduction could also denote techniques that seek an optimal row subspace, which reduces the size of n_g, but our interest here is strictly on reducing n_v.

2.1. Principal component analysis (PCA) based low-dimensional manifold

For PCA, the principal vectors align in the directions of maximal variance as captured by the second order data covariance matrix, $\mathbf{C} \in \mathbb{R}^{(n_v \times n_v)}$, represented using index notation as:

$$(\mathbf{C})_{ij} \equiv C_{ij} = \mathbb{E}(x_i x_j), \quad i, j \in \{1, \dots, n_v\}, \tag{1}$$

where $\mathbb{E}$ is the expectation operator. The required principal vectors ($\mathbf{A}$) are the eigenvectors of the covariance matrix obtained through an eigenvalue decomposition, $\mathbf{C} = \mathbf{A}\mathbf{L}\mathbf{A}^T$. It should be noted that the data used in the definition of joint moments is assumed to be centered around the mean.

2.2. Co-kurtosis tensor based low-dimensional manifold

Similarly, with the higher order moment of interest, i.e., the fourth-order co-kurtosis tensor, the principal vectors represent the directions of maximal kurtosis in the data. The co-kurtosis tensor is defined as:

$$\mathcal{T}_{ijkl} = \mathbb{E}(x_i x_j x_k x_l), \quad i, j, k, l \in \{1, \dots, n_v\} \tag{2}$$

By drawing an analogy to independent component analysis (ICA) [13], for a non-Gaussian data distribution, the fourth-order cumulant tensor, i.e., co-kurtosis $\mathbf{K}$ is computed by subtracting the excess variance given as:

$$K_{ijkl} = \mathcal{T}_{ijkl} - C_{ij}C_{kl} - C_{ik}C_{jl} - C_{il}C_{jk} \tag{3}$$

Again note that as the data is centered around the mean, only the second moment terms appear in the evaluation of the cumulant tensor.

The next step involves a suitable decomposition of the co-kurtosis tensor $\mathbf{K}$ to obtain the required principal components. Directly computing the higher-order joint moment tensors is expensive due to the curse of dimensionality, i.e., in our case for the co-kurtosis tensor, computational complexity would be n_v^4 where n_v is the number of features. The symmetric nature of the co-kurtosis tensor can be leveraged to result in roughly half of n_v^4 computations. However, the existing well-defined matrix decomposition techniques cannot be directly extended to higher-order tensors. Therefore, alternate tensor decomposition methods, such

as symmetric canonical polyadic (CP), higher order singular value decomposition (HOSVD), etc., should be explored to obtain the principal kurtosis vectors and values. Following [27,28], Aditya et al. [13] showed that the cumulant tensor $\mathbf{K}$ could be *reshaped* into a $n_v \times n_v^3$ matrix $\mathbf{T}$ following which the principal vectors $\mathbf{U}$ are determined from the SVD of $\mathbf{T} = \mathbf{USV}^T$.

After obtaining the principal components, we can reduce the dimensionality of the original data by projecting it onto a low-dimensional manifold. This is typically performed by selecting the most informative subset of principal vectors to project $\mathbf{X} \in \mathbb{R}^{(n_g \times n_v)}$ onto the reduced space represented as $\mathbf{Z}_q \in \mathbb{R}^{(n_g \times n_q)}$, where $n_q (< n_v)$ corresponds to the number of principal vectors retained. The conventional forward projection procedure in PCA employs a simple matrix transformation,

$$\mathbf{Z}_q = \mathbf{X}\mathbf{A}_q, \tag{4}$$

where $\mathbf{A}_q \in \mathbb{R}^{(n_v \times n_q)}$ represents the truncated subset of principal vectors (eigenvectors of the covariance matrix). For CoK-PCA, we obtain $\mathbf{A}_q$ as the n_q leading left singular vectors of $\mathbf{U}$ as described above. The contrast between PCA and CoK-PCA has been illustrated using a synthetic bivariate dataset with a few extreme-valued samples collectively representing anomalous events [13,14]. It was observed that while the first PCA principal vector aligned in the direction of maximal variance, the first CoK-PCA principal vector aligned itself in the direction of the anomalous cluster, supporting the hypothesis that CoK-PCA is more sensitive to extreme-valued samples than PCA.

3. Reconstruction methodology

To assess the quality of the reduced manifold, we need to evaluate the reconstruction accuracy of the original state space from the low-dimensional subspace. Note that errors in the reconstructed variables are incurred at two stages: while projecting data into the low-dimensional space and during the reconstruction.

3.1. Linear reconstruction

The standard procedure of obtaining the original thermo-chemical state is a linear reconstruction through a matrix inversion, given as:

$$\mathbf{X}_q = \mathbf{Z}_q \mathbf{A}_q^T, \tag{5}$$

where $\mathbf{X}_q$ denotes the reconstructed data in the original feature space. Now, a comparison between $\mathbf{X}_q$ and $\mathbf{X}$ would provide a quantitative measure of the quality of the two reduced manifolds obtained by CoK-PCA and PCA, respectively. Jonnalagadda et al. [14] analyzed the maximum and average values of the absolute reconstruction error $\left(\varepsilon = |\mathbf{X} - \mathbf{X}_q|\right)$, $\varepsilon_m = \max(\varepsilon)$ and $\varepsilon_a = \text{mean}(\varepsilon)$, respectively to quantify the accuracy in each reconstructed variable. Specifically, they examined the error ratio,

$$r_i = \ln\left\{\frac{\varepsilon_i^{\text{PCA}}}{\varepsilon_i^{\text{CoK-PCA}}}\right\}, \tag{6}$$

to analyze the performance of CoK-PCA relative to PCA; the subscript i can represent either the maximum (r_m) or average (r_a) errors.

3.2. Nonlinear reconstruction through ANNs

It is clear that while CoK-PCA exhibits improved accuracy in capturing stiff dynamics compared to PCA [14], both methods incur significant errors while employing a linear reconstruction of the original thermo-chemical state from the reduced manifold, particularly for an aggressive truncation (low n_q). Therefore, to fully establish the efficacy of CoK-PCA relative to PCA in capturing stiff dynamics, it is imperative to investigate its efficacy coupled with a nonlinear reconstruction approach. In this paper, we employ fully-connected deep neural networks to accomplish the required nonlinear reconstruction task. Since strong dependencies or relationships exist between different thermo-chemical scalars, it is appropriate to consider a fully-connected network where every subsequent layer is fully connected with the previous layer, ensuring the flow of information (of dependencies) across the network. In this regard, we also hypothesize that the use of a skip connection, i.e., introducing a sort of regularization in deeper networks by skipping some of the layer outputs during backpropagation, would not be suitable. However, it should be noted that using artificial neural networks (ANNs) is an intuitive choice. Alternate nonlinear regression methods, such as Gaussian process regression (GPR), polynomial regression, least squares, etc., exist and can be incorporated in a similar manner as described in this study.

With significant advancements in deep learning in recent times, ANNs have proven their potential to model highly complex nonlinear relationships between any set of inputs and outputs. The goal of an ANN or, specifically, a deep feedforward neural network is to approximate some underlying function f^*. For example, for a classifier, $\mathbf{y} = f^*(\mathbf{x})$ maps an input $\mathbf{x}$ to a category $\mathbf{y}$, but more generally in case of regression problems $\mathbf{x}$ is a vector of real numbers and $\mathbf{y}$ output of a vector-valued function. A feedforward network defines a mapping $\mathbf{y} = f(\mathbf{x};\boldsymbol{\theta})$ and learns the value of the parameters $\boldsymbol{\theta}$ that result in the best function approximation. The nonlinear reconstruction step in a dimensionality reduction algorithm can be viewed as a nonlinear mapping from the reduced manifold (or input PCs) to the original feature space (or output features). We leverage the property of ANNs being *universal function approximators* [24] to achieve this task.

Consider a reduced data representation of the original state space $\mathbf{X}$ given by the *score* matrix, $\mathbf{Z}_q = \mathbf{X}\mathbf{A}_q$, where $\mathbf{A}_q \in \mathbb{R}^{(n_v \times n_q)}$ comprises the chosen subset of principal vectors (kurtosis or variance). Now, the objective is to use an ANN to predict (or reconstruct) $\mathbf{X}_q$ from $\mathbf{Z}_q$ where $\mathbf{X}_q$ represents the reconstructed data in the original feature space, which is as close to $\mathbf{X}$ as possible. This is a supervised learning problem where for every kth feature vector from (kth row of) the design matrix $\mathbf{Z}_q$, $z_{k*} \in \mathbb{R}^{n_q}$, the network should accurately predict the target vector (kth row of $\mathbf{X}$) $x_{k*} \in \mathbb{R}^{n_v}$, i.e., the ANN should provide the mapping $z_{k*} \mapsto x_{k*}, \quad \forall k \in \{1, 2, \ldots, n_g\}$. In other words, the goal of training a neural network is to drive its prediction $\mathbf{X}_q$ to match $\mathbf{X}$. Since it is a regression problem, we evaluate the performance or accuracy of the model by using a mean squared error (MSE) loss defined as:

$$\mathcal{L}_{MSE} = \frac{1}{m}\sum_{k=1}^{m}(\hat{x}_{k*} - x_{k*})^2 \tag{7}$$

where $\hat{x}_{k*}$, x_{k*}, and m are the model prediction, ground truth, and the number of samples, respectively. Note that m can differ from n_g depending on how the entire dataset is split into training and test sets.

A simple feedforward neural network or ANN computes the output of a neuron by a linear combination of all weights and biases associated with it followed by a nonlinear activation function. From our numerous experiments, we found that employing Tanh for the hidden layer activations provides better stability, robustness, and smoother training of the network than ReLU, and exhibits minimal sensitivity to different random seeds. Additionally, Tanh can map inputs to spaces with both positive and negative values, unlike ReLU and sigmoid. Further, we use a custom sigmoid-based activation function at the output layer to ensure the model predictions are within the same limits as the original state. To obtain an optimal solution for the non-convex loss function, we employ the widely used Adam optimization algorithm, which is a variant of stochastic gradient descent (SGD). The network is implemented in the *TensorFlow* framework along with the use of techniques such as stochastic weight averaging, model averaging, and ensemble averaging in the network training phase to ensure consistency in model predictions.

3.3. Error metrics

Once trained, the network is used to predict the thermo-chemical scalars, which include species mass fractions and temperature. The

species production rates and heat release rate are also computed based on these reconstructed thermo-chemical scalars. The motivation behind calculating the species production rates and heat release rate is their nonlinear dependence on the species mass fractions and temperature, which provides a more stringent metric for assessing the reconstruction accuracy of the full thermo-chemical state and the overall dimensionality reduction strategy. Further, apart from having a tangible physical meaning, the reconstruction error associated with the heat release rate also provides an overall assessment of the quality of the reduced manifold since the heat release rate represents an aggregate effect of all the quantities of interest. A key point to note is that the network predictions correspond to a scaled version of the original state since the network is trained with scaled input feature vectors. Hence, we suitably unscale the network outputs before calculating the errors in the reconstruction of thermo-chemical scalars. Analogous to the error metrics in [14], we examine the following error ratios,

$$r_i = \log_{10}\left\{\frac{\varepsilon_i^{\text{CoK-PCA}}}{\varepsilon_i^{\text{CoK-PCA-ANN}}}\right\}, \tag{8}$$

$$r_i = \log_{10}\left\{\frac{\varepsilon_i^{\text{PCA}}}{\varepsilon_i^{\text{PCA-ANN}}}\right\}, \tag{9}$$

$$r_i = \log_{10}\left\{\frac{\varepsilon_i^{\text{PCA-ANN}}}{\varepsilon_i^{\text{CoK-PCA-ANN}}}\right\}, \tag{10}$$

to compare the relative performance of different methods such as CoK-PCA, PCA, CoK-PCA-ANN, and PCA-ANN considered in our study. Again, the subscript i can represent either the maximum (m) or average (a) errors. The value of r_i will be positive if the ratio inside the logarithm is greater than unity (the error in the denominator is lower), indicating that the technique represented by the denominator is more accurate than that represented by the numerator. In the results to be shown, following [14], we will denote positive r_i by blue and negative by brown colored bars.

4. Results

To investigate the accuracy of the proposed reconstruction methodology for combustion datasets, we consider four test cases representative of various physical and chemical phenomena (e.g., autoignition, flame propagation) ubiquitous in such scenarios:

- autoignition of a premixed ethylene/air mixture in a homogeneous reactor,
- autoignition, with two-stage ignition kinetics, of a dimethyl ether (DME)/air mixture in a homogeneous reactor,
- one-dimensional freely propagating planar laminar premixed flame of an ethylene/air mixture,
- two-dimensional turbulent autoignition of an ethanol/air mixture at homogeneous charge compression ignition (HCCI) conditions.

The datasets represent an increasing order of complexity of chemical kinetics and flow-chemistry interactions. The first two cases represent homogeneous (spatially zero-dimensional) autoignition, albeit ethylene/air with conventional ignition kinetics, while DME/air has more complex low and high temperature ignition kinetics. The third case incorporates spatial variation, including convection and diffusion effects in the canonical planar laminar premixed flame configuration. The fourth case represents complex turbulence–chemistry interactions in a spatially two-dimensional configuration under conditions relevant to practical devices.

4.1. Premixed ethylene-air in a homogeneous reactor

In this section, we consider the dataset that characterizes spontaneous ignition in a simple homogeneous (zero-dimensional) reactor. For dataset generation, we simulate a constant pressure reactor with

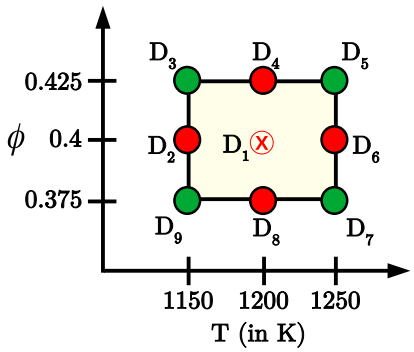

Fig. 1. Illustration of train–test split in ensemble training. Training states: D_1, D_2, D_4, D_6, D_8 and testing states: D_3, D_5, D_7, D_9. To generate the LDMs, PCs are computed based on the reference state D_1.

a premixed ethylene-air mixture at a pressure $P = 1.72\,\text{atm}$ for a suite of nine different thermo-chemical states, i.e., $D_i\ \forall i \in \{1, 2, \ldots, 9\}$, each with a different initial temperature (T) and equivalence ratio (ϕ) as illustrated in Fig. 1. Specifically, we perturb the initial conditions (T, ϕ) from a reference state of $D_1 \equiv (T = 1200\,\text{K},\ \phi = 0.4)$ by $\Delta T = \pm 50\,\text{K}$ and $\Delta\phi = \pm 0.25$. Thus, each state is parameterized by a combination of initial (T, ϕ) where $T \in \{1150\,\text{K}, 1200\,\text{K}, 1250\,\text{K}\}$ and $\phi \in \{0.375, 0.4, 0.425\}$. The chemistry is represented by a 32-species, 206-reactions mechanism [29]. The homogeneous reactor simulations are performed with *Cantera* [30], and each state is computed for different duration to ensure that the profiles remain nearly similar. For the reference state, the reactor is evolved for 2.5 ms with a time step of 1 μs to yield 2501 data samples. Hence, in this case, the original design matrix $\mathbf{D}$ consists of $n_g = 2501$ points and $n_v = 33$ variables, comprising 32 species mass fractions and temperature. The next step involves a data preprocessing stage where the design matrix for each state is zero-centered by subtracting with the mean feature vector and normalized with the absolute maximum feature vector to obtain the scaled data matrix, $\mathbf{X}$. This ensures an unbiased data representation with equal weight to all the features. This scaling procedure is followed for the other remaining test cases as well. To generate the low-dimensional manifolds, i.e., using PCA and CoK-PCA, we compute the principal vectors and values based on the scaled reference state (X_1), which eventually forms the basis for constructing the training/validation data. Next, we perform an aggressive truncation of the reduced manifolds by retaining $n_q = 5$ dominant principal vectors out of the $n_v = 33$ vectors that capture approximately 99% of the variance and 98% of the kurtosis in the dataset, respectively. Using the principal vectors computed on the scaled reference state (X_1), we obtain the LDM representation (score matrices) $\mathbf{Z}_q^4$ and $\mathbf{Z}_q^2$ through the dimensionality reduction procedure discussed in Section 2 for the CoK-PCA and PCA reduced manifolds, respectively. It should be noted that this projection is a linear operation.

After obtaining the LDMs with PCA and CoK-PCA, the next step in the *a priori* analysis is to evaluate the reduced manifolds in conjunction with the nonlinear reconstruction of the original thermo-chemical state through ANNs. For the ANN training phase, the input feature vectors are the rows of the score matrices ($\mathbf{Z}_q^4, \mathbf{Z}_q^2$) and the output vectors are the corresponding rows of the scaled original thermo-chemical state matrix $\mathbf{X}$; these matrices are arranged based on the different states (D_js) using train–test split shown in Fig. 1 (states D_1, D_2, D_4, D_6, and D_8 are used for ANN training only). The split between training and validation data is kept at 60/40. Through hyperparameter tuning, the best network architecture is ascertained with four hidden layers of widths of 40, 64, 40, and 32 neurons, respectively. In addition, the widths of input and output layers correspond to $n_q = 5$ and $n_v = 33$ neurons, respectively. Further, hyperbolic tangent activation in the hidden layers, custom sigmoid-based activation at the output layer, and Adam

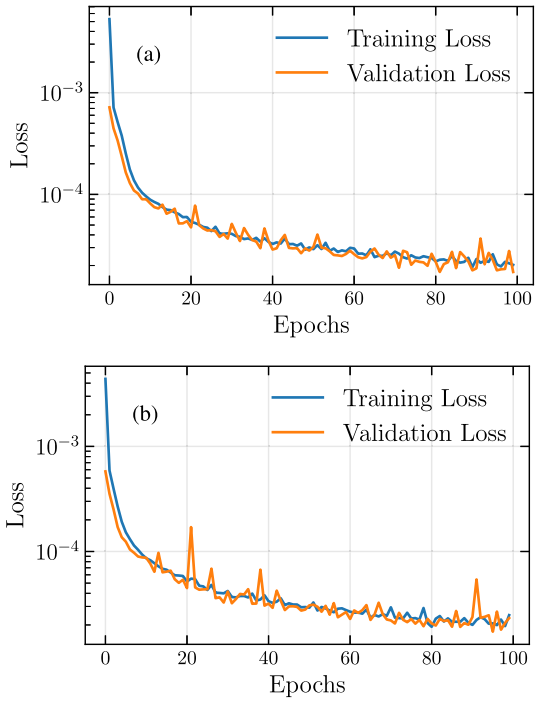

Fig. 2. Training and validation loss curves for (a) CoK-PCA-ANN and (b) PCA-ANN, respectively, for the premixed ethylene-air homogeneous reactor dataset.

optimizer (learning rate $= 1 \times 10^{-3}$) are used. Fig. 2 depicts the loss curves obtained for CoK-PCA-ANN and PCA-ANN, where convergence is achieved at around 100 epochs with a validation loss of about 2×10^{-5}. Notably, the close alignment of the validation loss with the training loss signifies a well-performing model capable of effectively generalizing to unseen states.

Having trained on a subset of the states, we use the neural network to predict (or reconstruct) the scaled species mass fractions and temperature for the test states, i.e., $D_j \ \forall j \in \{3, 5, 7, 9\}$. To ensure that the reconstructed thermo-chemical state results in a unit sum of species mass fractions, as is the standard practice, all reconstructed species mass fractions which yield negative values (that are slightly smaller than zero) are taken to be zero, after which any deviation from the sum equaling unity is adjusted for in the non-participating or bath species. Using the reconstructed thermo-chemical scalars, $\mathbf{D_q}$, we proceed to compute the species production rates and heat release rates. The reconstructed quantities are compared against the original thermo-chemical state, $\mathbf{D}$, and their derived quantities (species production rates, heat release rates) using the error metrics, r_a and r_m.

In Fig. 3, we compare error ratios of linear and ANN reconstruction (based on Eqs. (8) and (9)) of thermo-chemical scalars for both the dimensionality reduction methods. N_2 being an inert species has not been included here. For most variables, ANN reconstruction performs better than linear reconstruction (demonstrated by blue bars) with respect to the average (r_a) and maximum (r_m) error metrics. An exception is temperature, where linear reconstruction performs marginally better in terms of r_m (demonstrated by brown bars). This observation is consistent for both methods, i.e., PCA and CoK-PCA. In general, as n_q increases, the accuracy improvements obtained with ANN in comparison to linear reconstruction decrease as the reduced manifold becomes an increasingly better linear approximation of the original state; in the limit of $n_q = n_v$ linear reconstruction is exact, which is a scenario with no reduction in dimensionality. This is evident from Fig. 4, which presents the variation of the maximum error ratio (r_m) in the reconstruction of heat release rate with different values of n_q for CoK-PCA and CoK-PCA-ANN methods. The figure shows that for $n_q \leq 8$, positive values of r_m are obtained, which indicates a better accuracy of ANN reconstruction than linear reconstruction. However, for $9 \leq n_q \leq 22$, there is a steep decrease in the magnitude of r_m; it becomes largely negative, demonstrating the improved reconstruction accuracy of linear methods over ANN reconstruction. For $n_q \geq 23$, we observe an increase in the value of r_m towards the positive side, indicating the gain in reconstruction performance of ANNs and eventually matching that of linear methods. As dimensionality needs to be reduced as aggressively as possible, one can conclude that ANN is better suited for reconstructing data from aggressively-truncated low-dimensional manifolds.

Table 1
Cumulative errors in the normalized heat release rates (up to a progress variable of 0.99) from the two dimensionality reduction methods for the ethylene-air homogeneous reactor case. The error is computed individually for each of the four test states.

Method	Cumulative ϵ_{HRR}			
	D_3	D_5	D_7	D_9
PCA-ANN	29.626	1.867	19.966	13.767
CoK-PCA-ANN	9.414	1.666	14.710	9.166

Next, we compare the two dimensionality reduction techniques against each other, both with ANN reconstruction. Fig. 5 shows the error ratios comparing PCA-ANN and CoK-PCA-ANN (see Eq. (10)) in reconstructing the thermo-chemical scalars (parts (a) and (b)), and species production rates and heat release rates (parts (c) and (d)). For the scalars, it can be clearly seen that CoK-PCA-ANN performs better than PCA-ANN in predictions of 19 and 24 (out of 32) variables for r_a and r_m metrics, respectively. The trend becomes more prominent in the case of species production rates and heat release rates where CoK-PCA-ANN predicts production rates more accurately for 22 out of the 31 species with the r_a metric and 23 out of the 31 species with the r_m metric. Notably, CoK-PCA-ANN captures heat release rate better than PCA-ANN in terms of both error metrics.

While r_a and r_m are global error metrics, it is instructive to examine the temporal distribution of reconstruction errors and determine whether the errors are low/high in the unburnt, igniting, or fully burnt portions of the flame. Fig. 6 presents the normalized reconstruction error of heat release rate plotted against time for the four test states: D_3, D_5, D_7, and D_9. The normalized error in the reconstructed heat release rate, (ϵ_{HRR}), is defined as:

$$\epsilon_{\text{HRR}} = \frac{|\text{HRR}_q - \text{HRR}_{original}|}{\max(|\text{HRR}_{original}|)}, \tag{11}$$

For reference, the progress variable, which is computed based on the temperature, is plotted on the right y-axis of each figure. Both methods incur significant error in the reaction zones, with the peak at intermediate values of the progress variable, which occurs at 0.8 ms, 0.4 ms, 1 ms, and 1.9 ms for D_3, D_5, D_7, and D_9, respectively. As expected, the error is much lower on the unburnt and the fully burnt portions. Further, for D_3 and D_9, CoK-PCA-ANN incurs a significantly lower peak reconstruction error than PCA-ANN (demonstrated by the blue peaks smaller in magnitude than the red peaks), which is reflected in the r_m error presented in Fig. 5(d). However, the peak error for D_7 is higher for CoK-PCA-ANN. For D_5, both the methods incur essentially the same magnitude of errors and perform at par with each other. Nonetheless, across the four test states, CoK-PCA-ANN yields an overall smaller average reconstruction error than PCA-ANN, as reflected in the r_a error presented in Fig. 5(c). An alternative measure of the accuracy is the cumulative error in normalized heat release rate (ϵ_{HRR}) summed over each individual state. Table 1 shows this cumulative error for each of the states, and these comparisons provide further evidence that the proposed CoK-PCA-ANN method predicts the overall chemical kinetics in the reaction zone better than PCA-ANN.

It has been well established in the literature that the reduced manifolds identified by PCA are highly sensitive to the data subsampling

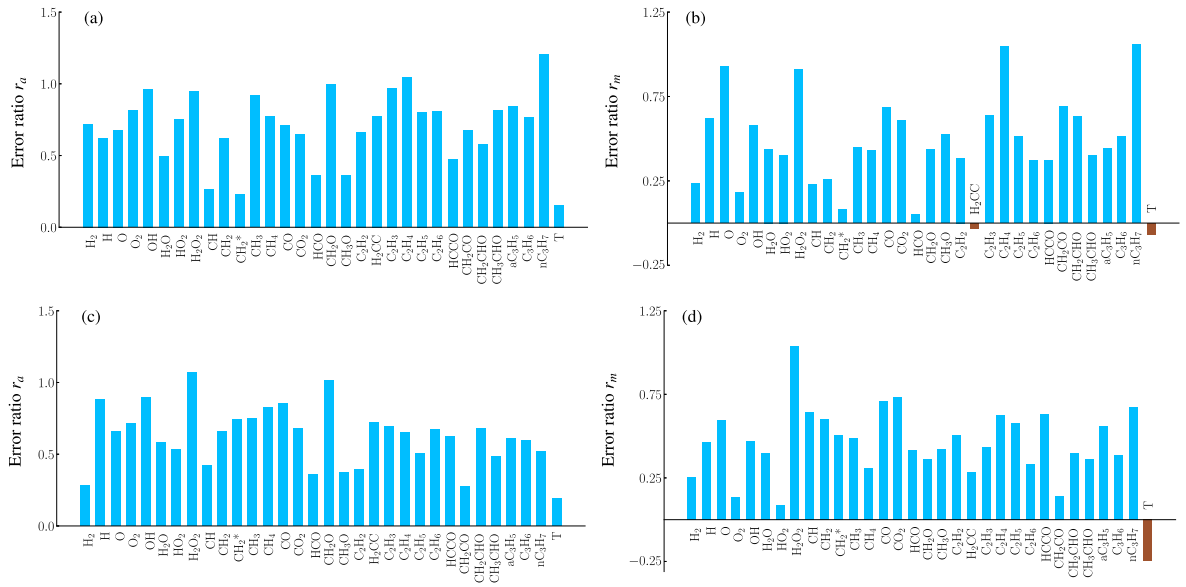

Fig. 3. Comparison of the reconstruction errors in thermo-chemical scalars between CoK-PCA and CoK-PCA-ANN (parts (a), (b)) as well as between PCA and PCA-ANN (parts (c), (d)) for the premixed ethylene-air homogeneous reactor dataset.

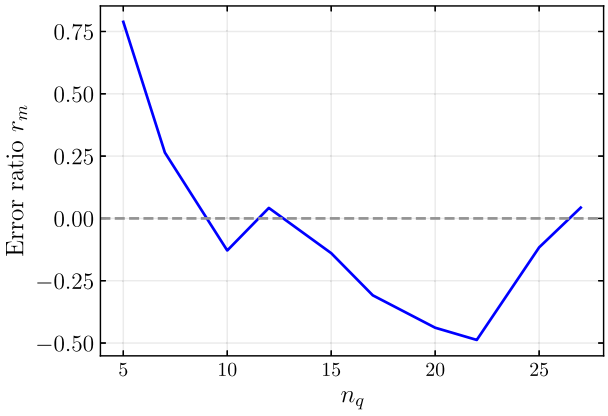

Fig. 4. Variation of maximum error ratio (r_m) in the reconstruction of heat release rate between CoK-PCA and CoK-PCA-ANN for different values of n_q.

strategy adopted, particularly in large DNS datasets where there is an absence of an equitable distribution of samples in the igniting and non-igniting regions [31]. This leads to the biasing of average errors (r_a) because of a significantly high number of data samples representing zero reaction rates. Among several studies to address this issue, Coussement et al. [23] performed a kernel density-based sampling approach for PCA. However, the optimal sampling strategy in combustion is highly dataset-dependent and, therefore, remains an open research problem [31]. In this study, we adopt a simple sampling strategy as performed in [14], which considers a uniform sampling of data points from the zones of unburnt reactants, chemical reactions, and burnt products. To segregate the temporal domain into different regions of varying chemical activity, we assume that the start/end of reactions occurs when the instantaneous heat release rate increases/decreases below 1% of the peak heat release rate. The number of samples considered in the non-igniting regions is one-third of those in the reacting region. Thereafter, we obtain the score matrix ($\mathbf{Z}_q^2$) for PCA using a new set of principal vectors computed based on this subsampled data. In contrast, the CoK-PCA score matrix ($\mathbf{Z}_q^4$), obtained based on the entire dataset, remains unchanged. The next step involves reconstructing the original thermo-chemical state using the full score matrices for PCA and CoK-PCA as computed above. Fig. 7 presents the comparison of reconstruction errors in the species production rates and heat release rates between PCA-ANN performed on the subsampled dataset and CoK-PCA-ANN performed on the entire dataset. It can be observed that CoK-PCA-ANN performs better than PCA-ANN in predicting production rates accurately for 22 out of 31 species in terms of r_a error metric. However, in the case of r_m error metric, both methods perform at par with each other. Moreover, CoK-PCA-ANN performs a more accurate reconstruction of the overall heat release rate in terms of both the error metrics.

4.2. Two-stage autoignition of dimethyl ether-air mixture

In contrast to ethylene, which has conventional single-stage ignition chemistry, a class of hydrocarbon fuels characterized by more complex two-stage ignition (a low-temperature and a high-temperature) chemistry are increasingly considered suitable for novel combustion concepts such as homogeneous charge compression ignition (HCCI) [32]. HCCI relies on volumetric autoignition of a (nearly) homogeneous fuel charge and realizes the benefits of low emissions due to fuel-lean combustion while also achieving high efficiencies. However, controlling the ignition timing is the biggest challenge since the charge ignites spontaneously due to compression heating. Consequently, modeling the ignition processes of two-stage ignition fuels under engine-relevant conditions is an open challenge. Dimethyl ether (DME) is a prominent example, and its ignition behavior resulting from turbulence–chemistry interactions at engine-relevant conditions has been widely studied using DNS [32–34]. From a dimensionality reduction perspective, DME ignition presents distinct challenges from that of ethylene; the chemical pathways and the participating chemical species for the low-temperature ignition chemistry are different from high-temperature chemistry. This motivates us to test the capability of CoK-PCA-ANN in reconstructing the original state space from the reduced manifold for the two-stage ignition of DME.

We consider a constant pressure zero-dimensional homogeneous reactor of a stoichiometric mixture of hydrogen-enriched DME fuel and air. The ratio of hydrogen to DME is 3:2 in the fuel mixture, similar to that in [33]. The initial pressure is 1 atm while the initial temperature is varied from 600 K to 800 K in increments of 25 K, for a total of nine flames. This range of initial temperatures is such that the flames contain both two-stage as well as single-stage ignition behavior. Finite rate chemistry is specified using the 39-species, 175-reactions

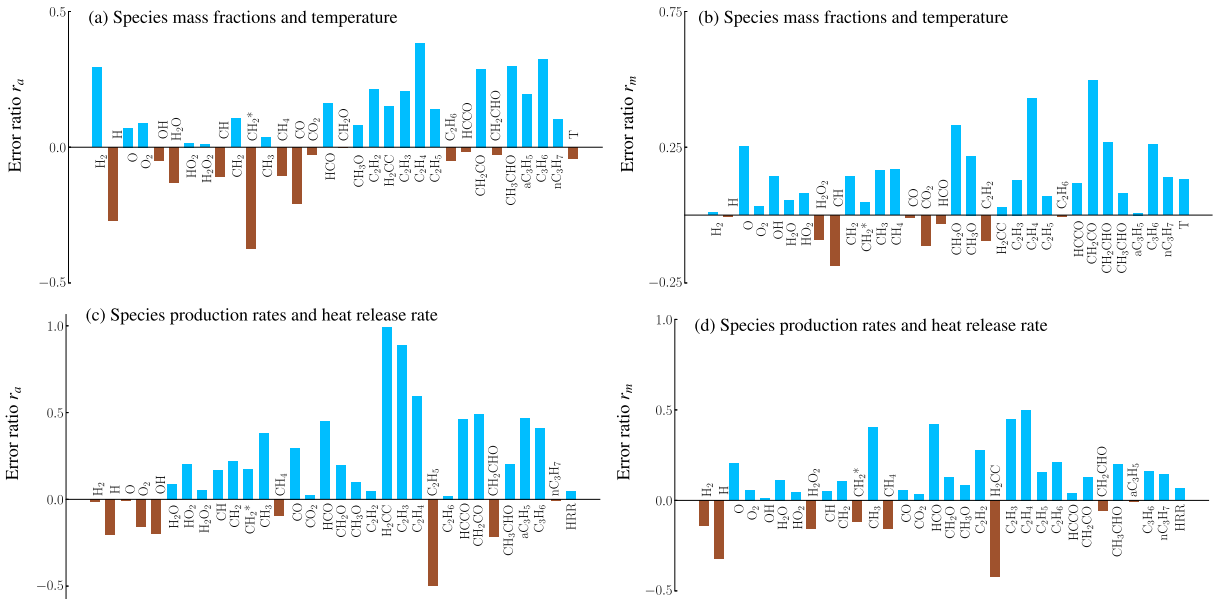

Fig. 5. Comparison of the reconstruction errors in thermo-chemical scalars (parts (a), (b)), species production rates and heat release rate (parts (c), (d)) between PCA-ANN and CoK-PCA-ANN for the premixed ethylene-air homogeneous reactor dataset.

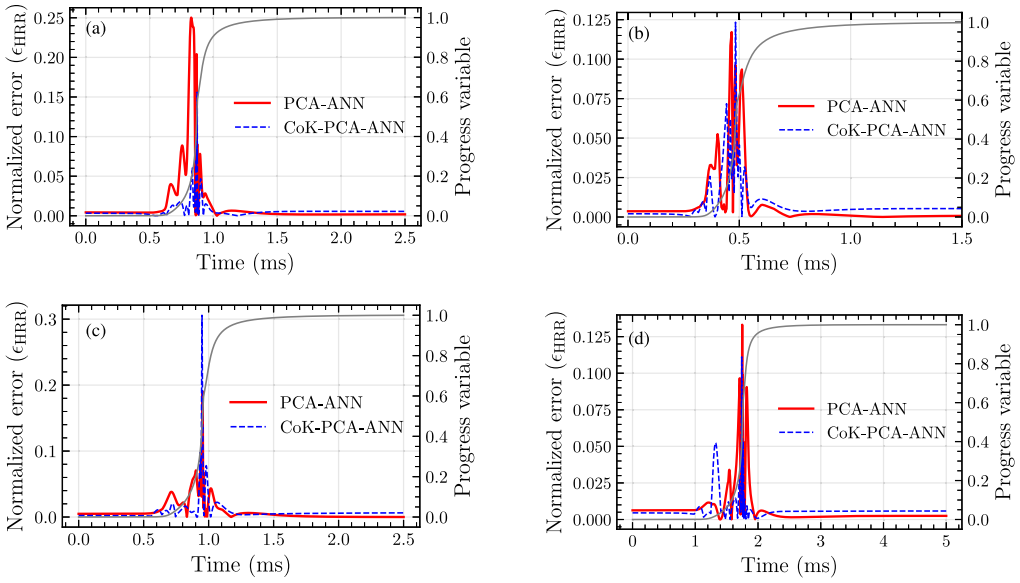

Fig. 6. Temporal evolution of normalized errors (see Eq. (11)) in reconstructed heat release rate for the test states (a) D_3, (b) D_5, (c) D_7, and (d) D_9 for the premixed ethylene-air homogeneous reactor dataset. The progress variable is plotted in solid gray for reference.

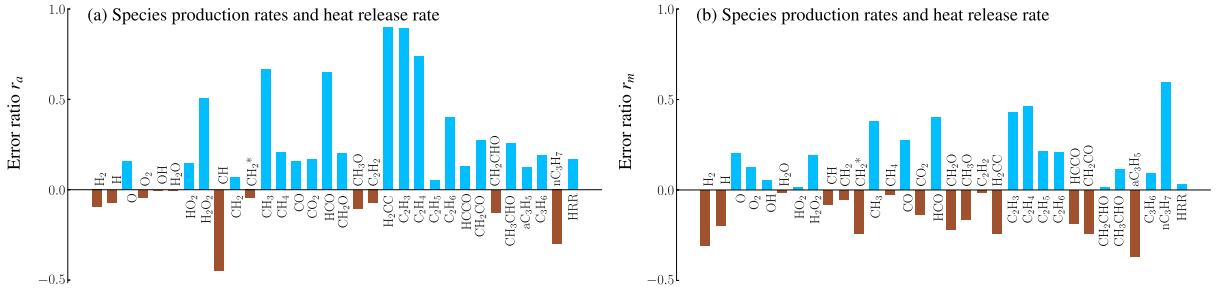

Fig. 7. Comparison of the average (part (a)) and maximum (part (b)) error ratios in the reconstruction of species production rates and heat release rate for the premixed ethylene-air homogeneous reactor dataset using PCA-ANN on sub-sampled data and CoK-PCA-ANN on the entire dataset.

skeletal mechanism developed in [33], and the flames are simulated with *Cantera* [30] for a duration of 1 s with a fixed time step of 0.1 ms. In this case, the original design matrix **D** consists of $n_g = 10001$ points and $n_v = 40$ variables, comprising 39 species and temperature.

Traditional dimensionality reduction techniques, such as PCA, may not effectively capture the nonlinear interactions present in the data. The data associated with the two-stage ignition of DME is high-dimensional and contains intricate patterns. This includes time-dependent or transient behavior, multiple ignition modes, and variations under different operating conditions. This complexity makes it difficult to find a low-dimensional representation that captures the essential information while discarding irrelevant or redundant features. The reconstruction of two-stage ignition dataset using ANNs is expected to improve the overall accuracy of the dimensionality reduction techniques used, which will be demonstrated next.

CoK-PCA and PCA are performed using the data of all nine flames, and dimensionality is reduced to $n_q = 5$. To train the ANNs for reconstructing the full thermo-chemical state from the reduced state, the data is split into training and testing sets, with five flames (initial temperatures of 600 K, 650 K, 700 K, 750 K, 800 K) comprising the former, and the rest, the latter. We randomly shuffle the training dataset and set aside 20% for the validation process. Hyperparameter tuning for the network width and depth and learning rate optimization were performed prior to defining the network architecture. The optimal network architecture is ascertained with two hidden layers of widths 10 and 20 neurons, respectively, with a learning rate of 1×10^{-3}. The input and output layers have a width of $n_q = 5$ and $n_v = 40$ neurons, respectively. A hyperbolic tangent activation function for the hidden layers, a custom sigmoid-based activation function for the output layer, and the Adam optimizer are used as before. The intricate nature of the two-stage DME dataset manifests complicated relationships, making it prone to overfitting when NN with greater width and depth are employed. However, a streamlined network architecture, comprising solely of two hidden layers, demonstrates exceptional efficacy and efficiency in accurately predicting outcomes for unseen instances within this complex dataset. While the best NN architecture for this case might appear simpler than that of the previous ethylene-air dataset, it was arrived at after a careful tuning of the hyperparameters, and it reflects the tradeoff between the data complexity, NN generalizability, and avoiding overfitting.

Fig. 8 shows the training and validation loss for PCA-ANN and CoK-PCA-ANN. It is evident that the validation loss remains consistently only slightly higher than the training loss ($\sim 2.5 \times 10^{-4}$) for a significant number of epochs (200–500), and the model has converged. We employ early stopping to achieve this convergence, thereby saving computational resources and preventing overfitting. This indicates that the model is generalizing well to unseen data. Despite a minor difference in loss, the model demonstrates robustness and reliability in its predictions. This suggests that the model has learned complex patterns present in the two-stage ignition dataset and features from the training data that allow it to make accurate predictions on unseen data, resulting in a reliable and effective model.

Fig. 9 illustrates the relative error ratios between PCA-ANN and CoK-PCA-ANN for thermo-chemical scalars (parts (a) and (b)), species production rates, and heat release rates (parts (c) and (d)). The errors in reconstructed thermo-chemical scalars show mixed trends, unlike the ethylene-air dataset for which CoK-PCA-ANN was consistently more accurate than PCA-ANN. However, the accuracy of species production rates and, more importantly, the heat release rate for CoK-PCA-ANN is better than PCA-ANN. This result reinforces the notion that error metrics based only on thermo-chemical state reconstruction may not be sufficient measures of accuracy. Going beyond the error ratios, and similar to the ethylene-air case, we plot the normalized errors in reconstructed heat release rates (Eq. (11)) for one of the DME-air flames from the test set with an initial temperature of 625 K as shown in Fig. 10. Since this mixture has two-stage ignition, the heat release rate

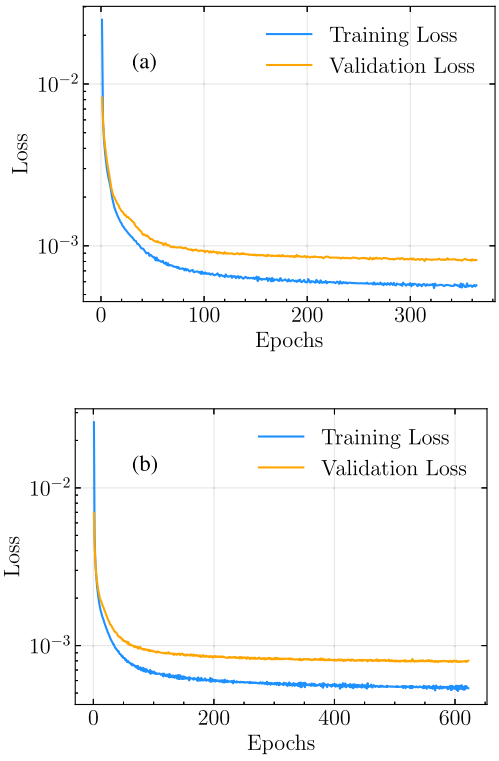

Fig. 8. Training and validation loss curves for (a) PCA-ANN and (b) CoK-PCA-ANN, respectively, for the DME two-stage autoignition dataset.

for the second stage (at ~ 0.25 ms) is orders of magnitude larger than the first stage (at ~ 0.047 ms). To make the comparison clearer, insets in Fig. 10 show the regions zoomed on the two stages. It is evident that the normalized errors in reconstructed heat release rates are greater by up to an order of magnitude with linear reconstruction (part (a)) compared with ANN-based reconstruction (part (b)). Moreover, while the errors for the first stage are comparable between PCA-ANN and CoK-PCA-ANN, for the second stage, CoK-PCA-ANN is more accurate.

4.3. Premixed ethylene-air laminar flame

The third case we consider is a one-dimensional freely-propagating planar laminar premixed flame of the ethylene-air mixture. In addition to the chemical reactions that govern the evolution of homogeneous reactors of the previous two cases, this case has effects of convection and diffusion that influence the thermo-chemical evolution. The chemistry is represented by the same 32-chemical species, 206-reactions mechanism [29], resulting in $n_v = 33$ variables. The freely-propagating flame is simulated in a one-dimensional domain of 0.02 m discretized with a grid of around 550 points. The pressure is kept at 1 atm, and a parametric variation is considered for the unburnt mixture conditions. Analogous to the ensemble training performed in Section 4.1, to construct the required training and testing data, we perturb the unburnt mixture temperature and equivalence ratio, (T, ϕ) by $\Delta T = \pm 50$ K and $\Delta\phi = \pm 0.25$ from the reference state, i.e., $D_1 \equiv (T = 300\,\text{K}, \phi = 0.6)$. This effectively results in nine configurations, $D_i\ \forall i \in \{1, 2, \ldots, 9\}$, one for each combination of (T, ϕ) where $T \in \{250\,\text{K}, 300\,\text{K}, 350\,\text{K}\}$ and $\phi \in \{0.575, 0.6, 0.625\}$. Again, to generate the CoK-PCA and PCA reduced manifolds, the principal components are computed with respect to the scaled reference state, X_1, by selecting $n_q = 5$ leading principal vectors out of the $n_v = 33$ vectors that capture approximately 99% of the variance and 98% of the kurtosis in the dataset, respectively. Following the dimensionality reduction procedure in Section 2, we compute the

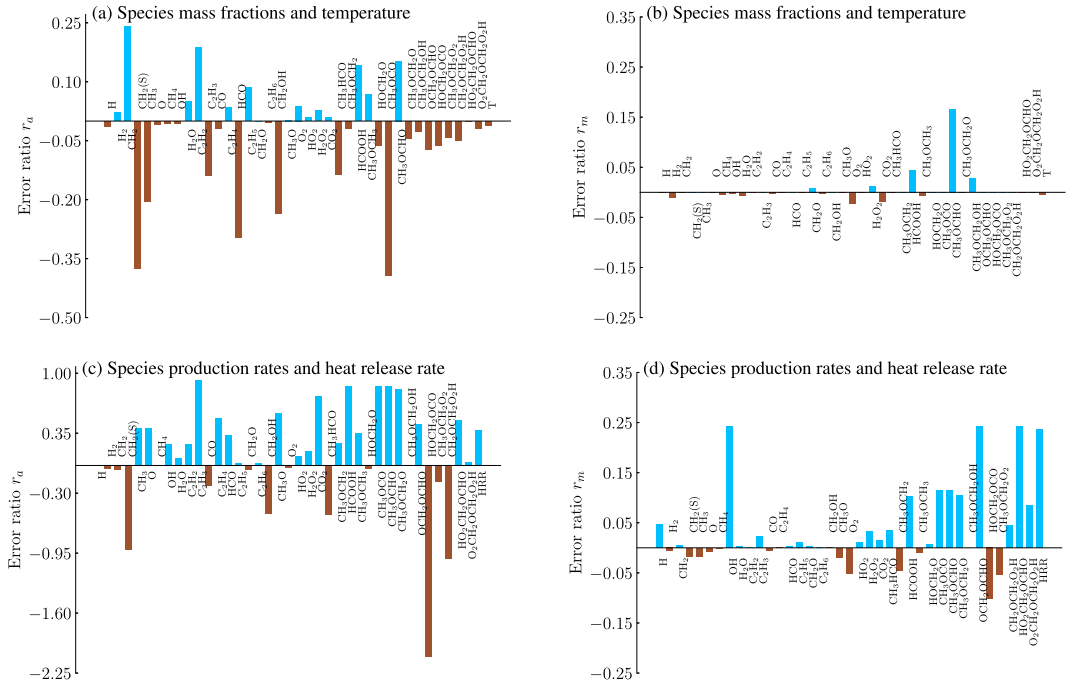

Fig. 9. Comparison of the reconstruction errors in thermo-chemical scalars (parts (a), (b)), species production rates and heat release rate (parts (c), (d)) between PCA-ANN and CoK-PCA-ANN for the DME two-stage autoignition dataset.

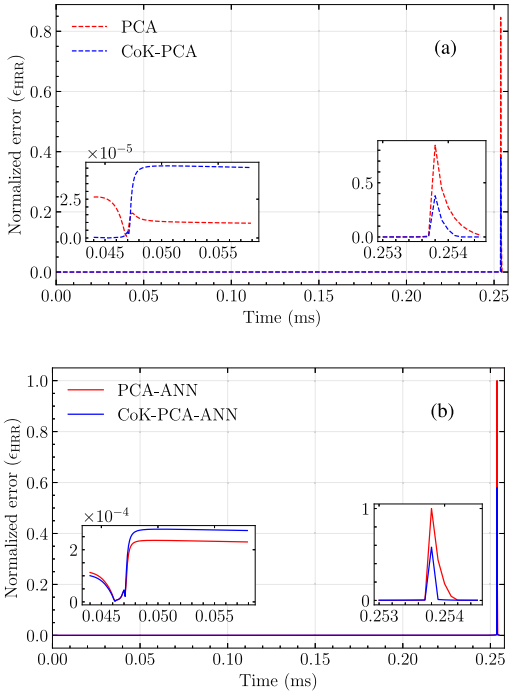

Fig. 10. Temporal evolution of normalized errors (see Eq. (11)) in the reconstruction of heat release rate for the DME two-stage autoignition flame with an initial temperature of 625 K.

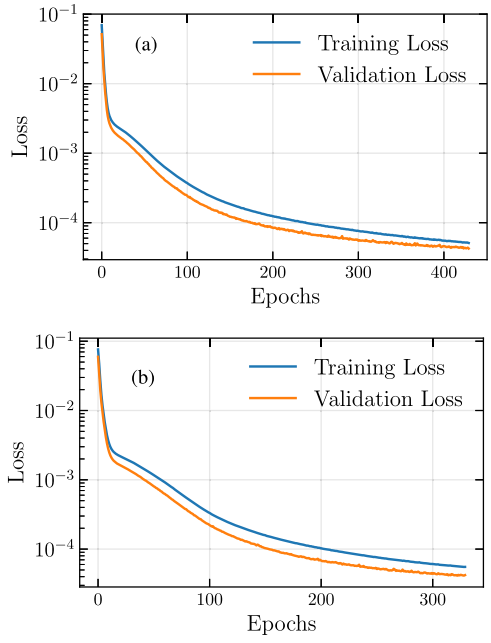

Fig. 11. Training and validation loss curves for (a) CoK-PCA-ANN and (b) PCA-ANN, respectively, for the one-dimensional planar laminar premixed ethylene-air flame dataset.

score matrices, $\mathbf{Z}_q^4$ and $\mathbf{Z}_q^2$ for the CoK-PCA and PCA low-dimensional manifolds, respectively.

For the ANN training, a similar split of the data into training and testing sets, as described in Section 4.1, is performed here; D_1, D_2, D_4, D_6, and D_8 are used for training and the rest for testing. Again, the training–validation split of data is kept at 60/40. Accordingly, we construct the input feature vectors and ground truths to train a neural network with four hidden layers of widths 48, 48, 48, and 56 neurons. The widths of input and output layers are $n_q = 5$ and $n_v = 33$ neurons, respectively. The layer activation functions remain the same as before with the use of Adam optimizer (learning rate = 1×10^{-4}) for training. Figs. 11(a) and (b) depict the loss curves obtained for CoK-PCA-ANN and PCA-ANN, respectively. Again, it should be emphasized that the training and validation losses are comparable to each other, indicating that the model has effectively learned the underlying training data

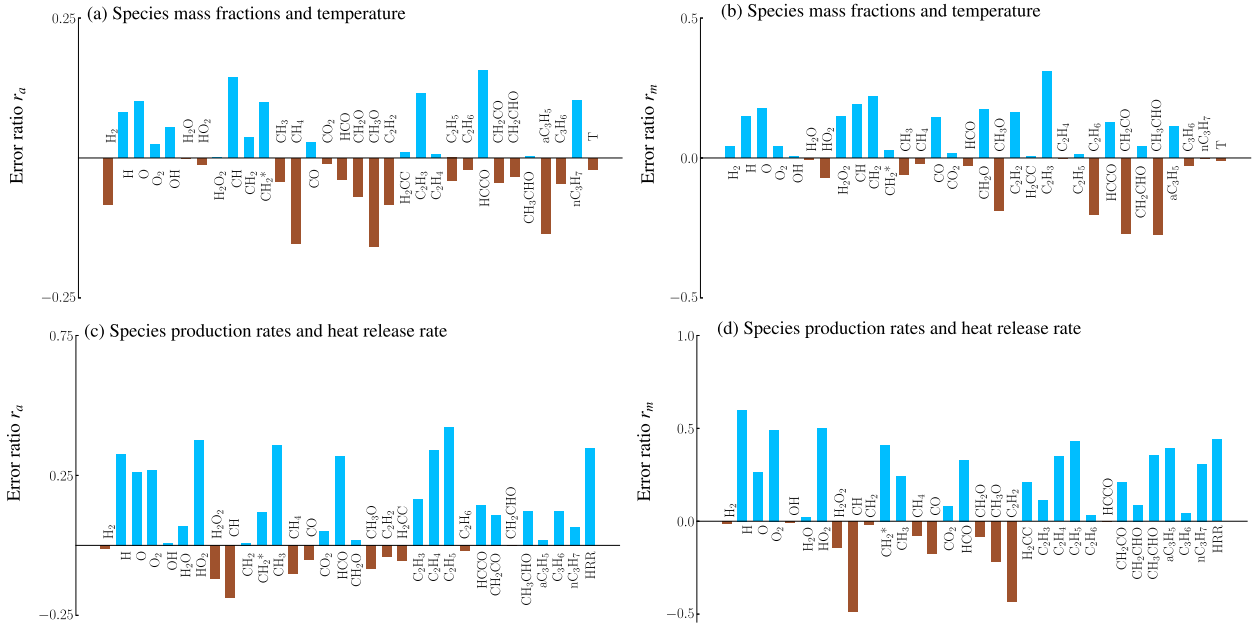

Fig. 12. Comparison of the reconstruction errors in thermo-chemical scalars (parts (a), (b)), species production rates and heat release rate (parts (c), (d)) between PCA-ANN and CoK-PCA-ANN for the one-dimensional planar laminar premixed ethylene-air flame dataset.

distribution and exhibits low generalization error in making predictions on unseen states.

Following Section 4.1, we assess the reconstruction accuracy of the trained models on the test states, i.e., $D_j \ \forall j \in \{3, 5, 7, 9\}$. Similar to the trends observed in previous cases, ANN reconstruction performs better than linear reconstruction for all the quantities of interest, the plots of which are not presented here for brevity. With reconstruction based on ANNs, we next focus on the performance of CoK-PCA-ANN against PCA-ANN in terms of the error ratios (r_a, r_m), which are presented in Fig. 12. For the accuracy of thermo-chemical scalars, we observe a different trend in this case, with PCA-ANN being more accurate than CoK-PCA-ANN for 17 out of the 32 variables for r_a. However, CoK-PCA-ANN performs better than PCA-ANN in terms of the r_m metric in accurate predictions of 19 out of the 32 variables. Further, while comparing errors in the reconstruction of species production rates and heat release rates, CoK-PCA-ANN dominates over PCA-ANN in both error ratios. In particular, CoK-PCA-ANN significantly improves upon PCA-ANN by predicting production rates for 20 out of 31 species in terms of the r_m error and 21 out of 31 species in terms of the r_a error. More importantly, it incurs lower errors in reconstructing the heat release rate in both metrics, which is an overall measure of the fidelity of the chemical system. This case clearly illustrates the fact that errors in reconstructing the thermo-chemical state alone might not be a sufficient measure of accuracy for a given dimensionality reduction technique, and a broader set of metrics might be prudent.

The profile of normalized errors in reconstructed heat release rates (Eq. (11)) obtained for both the methods, CoK-PCA-ANN (dashed blue) and PCA-ANN (solid red), is shown in Fig. 13 for the four test states, D_3, D_5, D_7, and D_9. We observe that CoK-PCA-ANN performs better than PCA-ANN in accurately predicting the steady-state flame location for all the test states, thereby characterizing flame propagation better. This behavior is consistent with the r_m errors presented in Fig. 12(d). Further, both techniques capture the non-reacting regions reasonably well in all the test states. However, in these regions, CoK-PCA-ANN performs marginally better than PCA-ANN by predicting nearly zero heat release rates for the test flames, D_5, D_7, and D_9 (Figs. 13 (b)–(d)). It should be noted that reconstruction errors incurred by the methods in these regions (i.e., predicting non-zero heat release in the non-reacting zones) can be attributed to statistical inconsistencies or stochasticity of the ANN training process. Consequently, this is reflected in the r_a metric (average error), which is lesser in the case of CoK-PCA-ANN than PCA-ANN (demonstrated by blue bars) in Fig. 12(c). Moreover, as with the homogeneous reactor case, a comparison of the cumulative normalized heat release rate error for each of the test flames of this data set, shown in Table 2, quantifies the superior accuracy of CoK-PCA-ANN over PCA-ANN.

Table 2
Cumulative errors in the normalized heat release rates (up to a progress variable of 0.99) for the test states obtained from different reduced manifolds for the one-dimensional freely propagating premixed ethylene-air laminar flame case.

Method	Cumulative ϵ_{HRR}			
	D_3	D_5	D_7	D_9
PCA-ANN	16.601	14.822	21.153	21.089
CoK-PCA-ANN	9.260	4.717	8.926	9.840

To gain insights into the reconstruction accuracy of selected thermo-chemical scalars, we choose two chemical species, namely, C_2H_3 and CH_3CHO, which yield maximum and minimum values for r_m, respectively, and comparatively assess their reconstructed profiles with the original scalar profile. Fig. 14 presents the original scalar (solid green) along with PCA-ANN (dashed red) and CoK-PCA-ANN (dashed blue) reconstructed scalar profiles of the aforementioned species for flamelet D_3. It is evident that the reconstruction of the scalar profiles for species C_2H_3 and CH_3CHO performed by both the methods, PCA-ANN and CoK-PCA-ANN, has a strong agreement with their original profiles and demonstrates the efficacy of the dimensionality reduction procedures.

4.4. Homogeneous charge compression ignition

In this section, we examine a dataset that encompasses the influence of spatial transport involving convection and diffusion in turbulent flows. The dataset is from a simulation of high-pressure, high-temperature autoignition of a turbulent mixture composed of premixed ethanol-air and combustion products, emulating the process of "exhaust gas recirculation" (EGR), representative of homogeneous charge compression ignition (HCCI) engine conditions [35]. The simulation was performed using S3D solver [36], which solves the reacting compressible flow governing equations. A doubly periodic domain with a two-dimensional spatial grid of 672 × 672 points was used. The initial

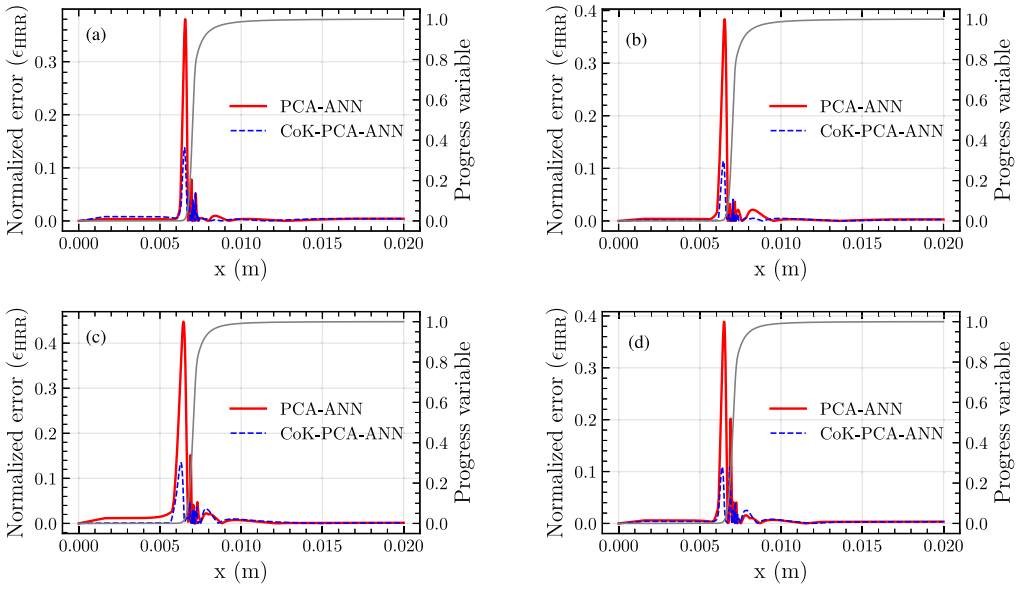

Fig. 13. Spatial variation of normalized errors (see Eq. (11)) in reconstructed heat release rates for the test states — (a) D_3, (b) D_5, (c) D_7, and (d) D_9 for the one-dimensional planar laminar premixed ethylene-air flame dataset. The progress variable is plotted in solid gray for reference.

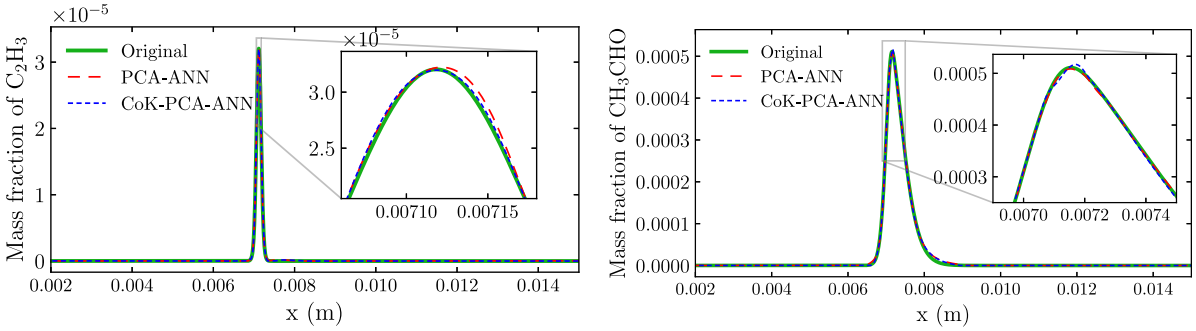

Fig. 14. Original scalar profile (solid green) along with PCA-ANN (dashed red) and CoK-PCA-ANN (dashed blue) reconstructed scalar profiles of the species C_2H_3 (left) and CH_3CHO (right) for the premixed ethylene-air laminar flame dataset (D_3). (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

conditions include a nominal pressure of 45 atm and a mean temperature of 924 K. The reactants are set to an equivalence ratio of 0.4. To account for the uneven mixing caused by EGR, a spatial temperature fluctuation and a separately computed divergence-free turbulent velocity field are superimposed onto the system. Furthermore, the simulation also considers the effects of compression heating resulting from the motion of the piston. The chemistry is represented by a 28-species reaction mechanism. Thus, at each simulation snapshot, the design matrix, $\mathbf{D}$, consists of $n_g = 672 \times 672$ data samples and $n_v = 29$ thermo-chemical scalars. For this study, we consider the temporal checkpoint at $t = 1.2$ ms [13], which corresponds to the propagation of the flame fronts in the bulk of the domain, as shown in the heat release rate contours in Fig. 15, which has been saturated to a peak heat release rate of 1×10^9 Jm^{-3}s^{-1} in order to demonstrate the growth in the size of the ignition kernels.

For the testing state, we consider the simulation snapshot at 1.19 ms. In other words, we are interested in investigating the efficacy of the proposed CoK-PCA-ANN method in predicting the thermo-chemistry at an unseen state ($t = 1.19$ ms) while being trained on a subsequent checkpoint at $t = 1.2$ ms. To obtain the score matrices $\mathbf{Z}_q^4$ and $\mathbf{Z}_q^2$, we use the principal vectors computed on the reference state, i.e., on $t = 1.2$ ms. The low-dimensional manifolds are constructed by retaining $n_q = 5$ out

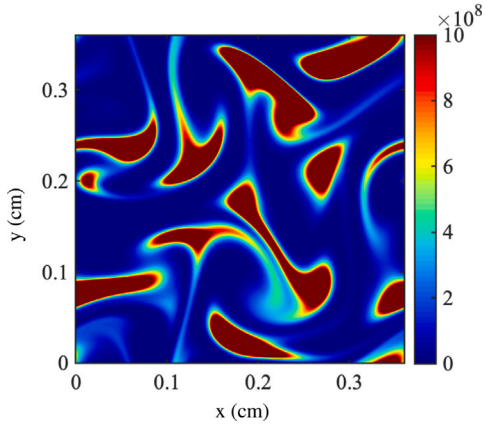

Fig. 15. Instantaneous contour plot of heat release rates (Jm^{-3}s^{-1}) from the two-dimensional HCCI dataset at t =1.2 ms.

of $n_v = 29$ principal vectors that correspond to approximately 99% of the variance and kurtosis in the PCA and CoK-PCA reduced manifolds,

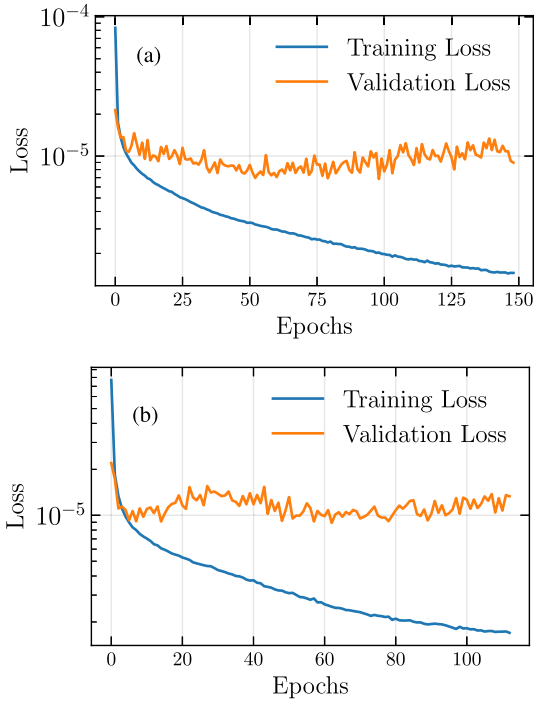

Fig. 16. Training and validation loss curves for (a) CoK-PCA-ANN and (b) PCA-ANN, respectively, for the two-dimensional HCCI dataset.

respectively. A neural network with three hidden layers of widths 8, 8, and 64 neurons is trained till convergence with an Adam optimizer (learning rate = 0.00274). In addition, early stopping is employed to ensure the network does not lead to overfitting on the training data. The corresponding loss curves obtained for both the manifolds are presented in Fig. 16. Although the validation loss is marginally higher than the training loss by a magnitude of $\sim 9 \times 10^{-6}$, the network has reached convergence at around 100–150 epochs, as shown in Fig. 16. It is worth noting that we are dealing with a two-dimensional temporally evolving dataset with complex turbulence–chemistry interactions. Despite this complexity, the NN successfully learns the underlying intrinsic patterns and complicated relationships present in the HCCI dataset at $t = 1.2\,\text{ms}$, enabling the creation of a robust and generalizable model capable of predicting outcomes on unseen data at a different time, $t = 1.19\,\text{ms}$. We use the trained network to predict the thermo-chemical scalars at $t = 1.19\,\text{ms}$ for both the CoK-PCA and PCA reduced manifolds. In a similar manner, using the reconstructed thermo-chemical scalars, species production rates and heat release rates are computed.

From the comparison in Fig. 17 (parts (a) and (b)), it is evident that CoK-PCA-ANN performs significantly better than PCA-ANN in the reconstruction of thermo-chemical scalars with more accurate predictions of 20 out of 29 species in both r_a and r_m errors. The superior accuracy of CoK-PCA-ANN is even more significant for reconstructed species production rates; it is more accurate than PCA-ANN for 25 and 26 of the 28 species in terms of r_a and r_m metrics, respectively (parts (c) and (d) in the figure). This results in lower reconstruction errors for heat release rates in both metrics (r_a, r_m) for CoK-PCA-ANN. Contrary to the observations in [14], where the CoK-PCA based LDM performed poorly in terms of the average errors (r_a) while considering the entire spatial domain, CoK-PCA, when coupled with ANN, overcomes this issue and better represents the stiff chemical dynamics in the average error as well. Note that the conclusions drawn from these error plots remain unchanged when the training is performed with $t = 1.19\,\text{ms}$ and testing is carried out on $t = 1.2\,\text{ms}$. Next, we plot the contours of the normalized errors in the reconstructed heat release rate (Eq. (11)) in Fig. 18 to compare the spatial error distribution for the two methods. Due to the inherent ability of excess kurtosis to suitably capture extreme-valued samples, CoK-PCA-ANN identifies the ignition zones better (e.g., as observed in the circled regions of the figure where CoK-PCA-ANN shows lower peak values) in the entire domain, which is in good agreement with the error metrics in Fig. 17 for heat release rate.

5. Conclusions and future work

In this paper, we have proposed an enhanced version of the co-kurtosis PCA (CoK-PCA) based dimensionality reduction method, namely CoK-PCA-ANN, which leverages the potential of artificial neural networks (ANNs) to model complex nonlinear relationships inherent between the aggressively truncated low-dimensional manifolds and the original thermo-chemical state. The rationale behind this work is (i) to assess the collective effectiveness and performance of the nonlinear reconstruction using ANNs (CoK-PCA-ANN and PCA-ANN) with linear reconstruction (CoK-PCA and PCA); (ii) to evaluate the overall efficacy of CoK-PCA-ANN in comparison with PCA-ANN and expand its applicability to chemically reacting systems presenting stiff dynamics. While other nonlinear reconstruction methods, such as Gaussian process regression (GPR), kernel density methods, autoencoders, etc., have been used in conjunction with PCA in previous studies, we have focused on ANNs in this study.

The framework of the proposed CoK-PCA-ANN dimensionality reduction method was presented with a discussion on the generation of the low-dimensional manifold using linear projection (encoding) with CoK-PCA followed by nonlinear reconstruction of the original thermo-chemical state space (decoding) using ANNs. Sufficient rigor was followed in the training of ANNs, specifically with regard to the appropriate selection of training and testing data, hyperparameter tuning, avoiding overfitting, and ensuring convergence. The performance of the CoK-PCA-ANN method was evaluated in comparison to the linear reconstruction (CoK-PCA) and PCA-ANN methods across four distinct combustion test cases that span conventional single-stage to complex two-stage ignition kinetics, different combustion regimes (autoignition, flame propagation), as well as a simple homogeneous reactor to a spatiotemporally evolving two-dimensional flow.

The training of a NN necessitates adapting to the intricacies inherent to the data, thereby leading to the potential variation of hyperparameters (such as learning rate, number of layers, number of neurons, and regularization strength) in accordance with the level of complexity involved. This process involves iterative experimentation and fine-tuning to identify the hyperparameter configuration that yields optimal performance and generalization for each dataset considered in this study. Overall, nonlinear reconstruction using ANNs demonstrated significantly high accuracies, as compared to linear reconstruction, for the CoK-PCA and PCA manifolds in terms of thermo-chemical scalars, species production rates, and heat release rates with aggressive truncation (low n_q). As expected, and akin to nonlinear reconstruction from PCA manifolds, CoK-PCA-ANN incurred lower reconstruction errors as compared to CoK-PCA with a simple linear reconstruction [14] in the average error metric (r_a). Additionally, CoK-PCA-ANN also presented a better representation of the non-igniting regions, i.e., unburnt reactants and burnt products in all the test cases. The quality of the PCA and CoK-PCA manifolds in conjunction with ANN reconstruction was further comparatively assessed using average (r_a) and maximum (r_m) error metrics for the reconstruction of the three aforementioned quantities. We find that CoK-PCA-ANN performs better than PCA-ANN in all the test cases in terms of both the error metrics for all the considered quantities. Most importantly, the smaller reconstruction errors associated with the heat release rates provide further evidence that the chemical kinetics prevalent in the reaction zones representative of stiff dynamics are captured more accurately by a CoK-PCA manifold than PCA. To summarize, the results from the above analyses suggest that CoK-PCA-ANN realizes the advantages of both CoK-PCA as well as ANNs to

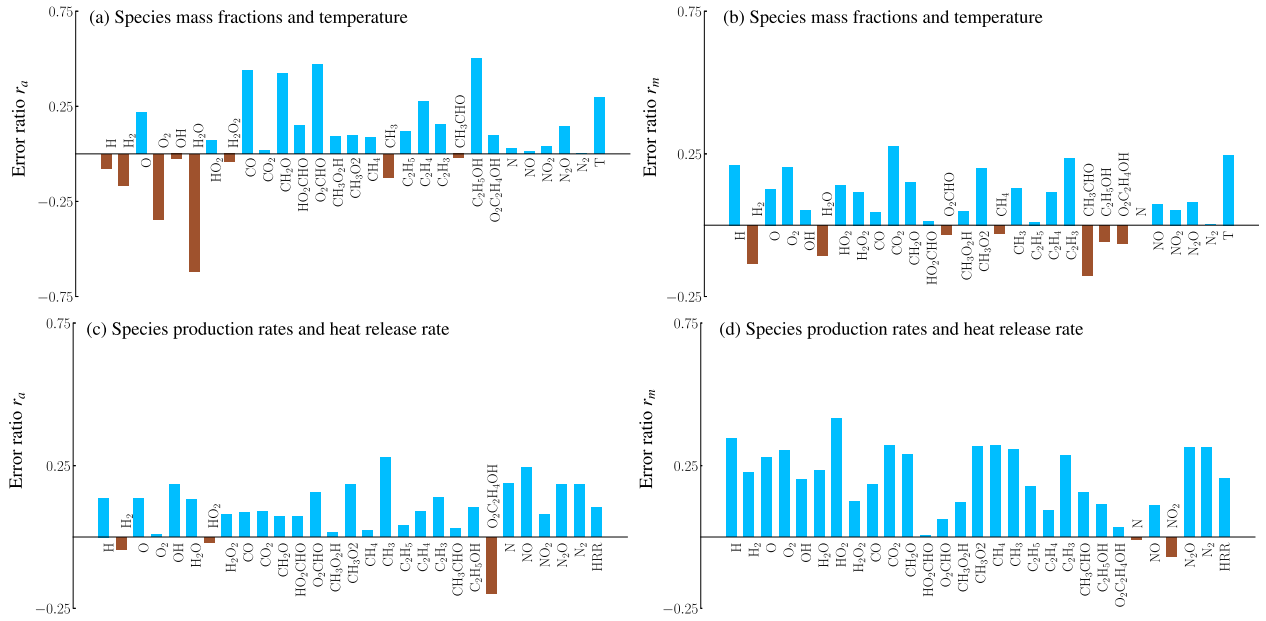

Fig. 17. Comparison of the reconstruction errors in thermo-chemical scalars (parts (a), (b)), species production rates and heat release rate (parts (c), (d)) between PCA-ANN and CoK-PCA-ANN for the two-dimensional HCCI dataset.

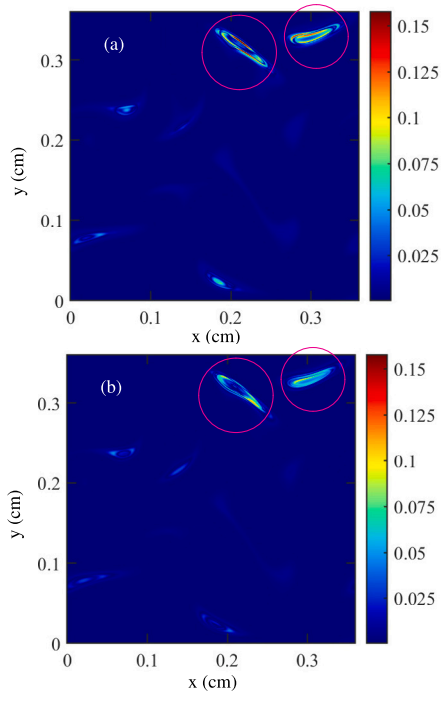

Fig. 18. Instantaneous contour plots of absolute errors in reconstructed heat release rates normalized by the peak heat release rate (2.55×10^{10} $Jm^{-3}s^{-1}$) for the two-dimensional HCCI dataset at t = 1.19 ms for (a) PCA-ANN and (b) CoK-PCA-ANN, respectively.

yield a reliable, robust, and generalizable low-dimensional manifold representation of complex combustion datasets.

However, it should be remarked that the investigation of the CoK-PCA-based nonlinear reconstruction using ANNs in this paper was carried out in an *a priori* setting. It is well known that these data-driven dimensionality reduction methods are capable of accelerating numerical simulations of reacting flows by solving a reduced set of principal component transport equations as opposed to solving a very high-dimensional system of species conservation equations. Such *a posteriori* validations, performed by other studies for PCA, remain to be explored for CoK-PCA and constitute future work.

CRediT authorship contribution statement

Dibyajyoti Nayak: Conceptualization, Implementation (datasets, neural networks), Analysis, Writing – original draft. **Anirudh Jonnalagadda:** Conceptualization, Implementation (chemistry), Analysis, Writing. **Uma Balakrishnan:** Conceptualization, Implementation (datasets, neural networks), Analysis, Writing – original draft. **Hemanth Kolla:** Conceptualization, Analysis, Writing, Supervision. **Konduri Aditya:** Conceptualization, Analysis, Writing, Supervision.

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 work at IISc was supported under a project from the National Supercomputing Mission, India. DN is a recipient of the Ansys M.Tech. (Research) Fellowship. AJ was funded by a project from Shell Technology Center, Bengaluru, India. KA is a recipient of the Arcot Ramachandran Young Investigator award, IISc. Work by HK and UB was part of the ExaLearn Co-design Center, supported by the Exascale Computing Project (17-SC-20-SC), a collaborative effort of the U.S. Department of Energy Office of Science and the National Nuclear Security Administration. Sandia National Laboratories is a multi-mission laboratory managed and operated by National Technology and Engineering Solutions of Sandia, LLC., a wholly owned subsidiary of Honeywell International, Inc., for the U.S. Department of Energy's National Nuclear Security Administration under contract DE-NA-0003525. The views expressed in the article do not necessarily represent the views of the U.S. Department of Energy or the United States Government.

References

[1] K. Aditya, A. Gruber, C. Xu, T. Lu, A. Krisman, M.R. Bothien, J.H. Chen, Direct numerical simulation of flame stabilization assisted by autoignition in a reheat gas turbine combustor, Proc. Combust. Inst. 37 (2019) 2635–2642.

[2] B. Savard, E.R. Hawkes, K. Aditya, H. Wang, J.H. Chen, Regimes of premixed turbulent spontaneous ignition and deflagration under gas-turbine reheat combustion conditions, Combust. Flame 208 (2019) 402–419.

[3] L. Berger, R. Hesse, K. Kleinheinz, M.J. Hegetschweiler, A. Attili, J. Beeckmann, G.T. Linteris, H. Pitsch, A DNS study of the impact of gravity on spherically expanding laminar premixed flames, Combust. Flame 216 (2020) 412–425.

[4] G. Nivarti, S. Cant, Direct numerical simulation of the bending effect in turbulent premixed flames, Proc. Combust. Inst. 36 (2017) 1903–1910.

[5] S. Desai, Y.J. Kim, W. Song, M.B. Luong, F.E.H. Pérez, R. Sankaran, H.G. Im, Direct numerical simulations of turbulent reacting flows with shock waves and stiff chemistry using many-core/GPU acceleration, Comput. & Fluids 215 (2021) 104787.

[6] H.A. Uranakara, S. Barwey, F.E.H. Pérez, V. Vijayarangan, V. Raman, H.G. Im, Accelerating turbulent reacting flow simulations on many-core/GPUs using matrix-based kinetics, Proc. Combust. Inst. (2022).

[7] J.C. Sutherland, A. Parente, Combustion modeling using principal component analysis, Proc. Combust. Inst. 32 (2009) 1563–1570.

[8] A. Biglari, J.C. Sutherland, A filter-independent model identification technique for turbulent combustion modeling, Combust. Flame 159 (2012) 1960–1970.

[9] Y. Yang, S.B. Pope, J.H. Chen, Empirical low-dimensional manifolds in composition space, Combust. Flame 160 (2013) 1967–1980.

[10] R. Ranade, T. Echekki, A framework for data-based turbulent combustion closure: A priori validation, Combust. Flame 206 (2019) 490–505.

[11] A. Parente, J. Sutherland, B.B. Dally, L. Tognotti, P. Smith, Investigation of the MILD combustion regime via principal component analysis, Proc. Combust. Inst. 33 (2011) 3333–3341.

[12] A. Parente, J.C. Sutherland, L. Tognotti, P.J. Smith, Identification of low-dimensional manifolds in turbulent flames, Proc. Combust. Inst. 32 (1) (2009) 1579–1586.

[13] K. Aditya, H. Kolla, W.P. Kegelmeyer, T.M. Shead, J. Ling, W.L. Davis, Anomaly detection in scientific data using joint statistical moments, J. Comput. Phys. 387 (2019) 522–538.

[14] A. Jonnalagadda, S. Kulkarni, A. Rodhiya, H. Kolla, K. Aditya, A co-kurtosis based dimensionality reduction method for combustion datasets, Combust. Flame 250 (2023) 112635.

[15] J. Chen, S. Desai, H. Babaee, S. Yamajala, Direct numerical simulation with time dependent subspaces for reduced-order modeling (ROM) of turbulent compressible reacting flows, Bull. Am. Phys. Soc. (2023).

[16] E. Armstrong, J.C. Sutherland, A technique for characterising feature size and quality of manifolds, Combust. Theory Model. 25 (4) (2021) 646–668.

[17] K. Zdybał, E. Armstrong, J.C. Sutherland, A. Parente, Cost function for low-dimensional manifold topology assessment, Sci. Rep. 12 (1) (2022) 14496.

[18] K. Zdybał, E. Armstrong, A. Parente, J.C. Sutherland, PCAfold: Python software to generate, analyze and improve PCA-derived low-dimensional manifolds, SoftwareX 12 (2020) 100630.

[19] H. Mirgolbabaei, T. Echekki, The reconstruction of thermo-chemical scalars in combustion from a reduced set of their principal components, Combust. Flame 162 (5) (2015) 1650–1652.

[20] T. Echekki, H. Mirgolbabaei, Principal component transport in turbulent combustion: A posteriori analysis, Combust. Flame 162 (5) (2015) 1919–1933.

[21] M.R. Malik, P.O. Vega, A. Coussement, A. Parente, Combustion modeling using Principal Component Analysis: A posteriori validation on Sandia flames D, E and F, Proc. Combust. Inst. 38 (2021) 2635–2643.

[22] A. Bellemans, M.R. Malik, F. Bisetti, A. Parente, A machine-learning framework for plasma-assisted combustion using principal component analysis and Gaussian process regression, in: Int. J. Uncertain. Quantif., Springer, 2020, pp. 379–392.

[23] A. Coussement, O. Gicquel, A. Parente, Kernel density weighted principal component analysis of combustion processes, Combust. Flame 159 (2012) 2844–2855.

[24] K. Hornik, M. Stinchcombe, H. White, Multilayer feedforward networks are universal approximators, Neural Netw. 2 (5) (1989) 359–366.

[25] S. Abdelwahid, M.R. Malik, H.A.A.K. Hammoud, F.E. Hernández-Pérez, B. Ghanem, H.G. Im, Large eddy simulations of ammonia-hydrogen jet flames at elevated pressure using principal component analysis and deep neural networks, Combust. Flame 253 (2023) 112781.

[26] A. Kumar, M. Rieth, O. Owoyele, J.H. Chen, T. Echekki, Acceleration of turbulent combustion DNS via principal component transport, Combust. Flame 255 (2023) 112903.

[27] L. De Lathauwer, B. De Moor, J. Vandewalle, Independent component analysis and (simultaneous) third-order tensor diagonalization, IEEE Trans. Signal Process. 49 (10) (2001) 2262–2271.

[28] A. Anandkumar, R. Ge, D. Hsu, S. Kakade, M. Telgarsky, Tensor decompositions for learning latent variable models, J. Mach. Learn. Res. 15 (2014) 2773–2832.

[29] Z. Luo, C.S. Yoo, E.S. Richardson, J.H. Chen, C.K. Law, T. Lu, Chemical explosive mode analysis for a turbulent lifted ethylene jet flame in highly-heated coflow, Combust. Flame 159 (2012) 265–274.

[30] D.G. Goodwin, H.K. Moffat, I. Schoegl, R.L. Speth, B.W. Weber, Cantera: An object-oriented software toolkit for chemical kinetics, thermodynamics, and transport processes, 2022, Version 2.6.0, https://www.cantera.org.

[31] K. Zdybał, J.C. Sutherland, A. Parente, Manifold-informed state vector subset for reduced-order modeling, Proc. Combust. Inst. 39 (4) (2023) 5145–5154.

[32] G. Bansal, A. Mascarenhas, J.H. Chen, Direct numerical simulations of autoignition in stratified dimethyl-ether (DME)/air turbulent mixtures, Combust. Flame. 162 (3) (2015) 688–702.

[33] A. Bhagatwala, Z. Luo, H. Shen, J.A. Sutton, T. Lu, J.H. Chen, Numerical and experimental investigation of turbulent DME jet flames, Proc. Combust. Inst. 35 (2) (2015) 1157–1166.

[34] A. Krisman, E.R. Hawkes, M. Talei, A. Bhagatwala, J.H. Chen, A direct numerical simulation of cool-flame affected autoignition in diesel engine-relevant conditions, Proc. Combust. Inst. 36 (3) (2017) 3567–3575.

[35] A. Bhagatwala, J.H. Chen, T. Lu, Direct numerical simulations of HCCI/SACI with ethanol, Combust. Flame 161 (2014) 1826–1841.

[36] J.H. Chen, A. Choudhary, B. De Supinski, M. DeVries, E.R. Hawkes, S. Klasky, W.-K. Liao, K.-L. Ma, J. Mellor-Crummey, N. Podhorszki, et al., Terascale direct numerical simulations of turbulent combustion using S3D, Comput. Sci. Discov. 2 (1) (2009) 015001.