ACCEPTED MANUSCRIPT
Physics of Fluids
AIP Publishing
This is the author's peer reviewed, accepted manuscript. However, the online version of record will be different from this version once it has been copyedited and typeset.
PLEASE CITE THIS ARTICLE AS DOI: 10.1063/5.0167110

Accepted to Phys. Fluids 10.1063/5.0167110

Preprint submitted to Physics of Fluids

manuscript.tex

Deep Learning Dynamical Latencies for the Analysis and Reduction of Combustion Chemistry Kinetics

Luisa Castellanos[1,2,*], Rodolfo S. M. Freitas[2], Alessandro Parente[2], Francesco Contino[1]

[1]*Université Catholique de Louvain, École polytechnique de Louvain,*
Institute of Mechanics, Materials and Civil Engineering,
Ottignies-Louvain-la-Neuve, Belgium and
[2]*Université Libre de Bruxelles, École polytechnique de Bruxelles,*
Aero-Thermo-Mechanics Laboratory, Brussels, Belgium

(*luisa.castellanos@uclouvain.be)

(Dated: September 8, 2023)

Abstract

The modeling of chemical kinetics holds many challenges, as well as a necessity for more efficient modeling techniques, together with dimensionality reduction techniques. This work studies the application of time-lag auto-encoders (TAEs) for the analysis of combustion chemistry kinetics. Such a technique allows a better reconstruction of the thermochemical temporal advancement in relation to traditional reduction techniques (Principal Component Analysis or PCA) while applying a potential denoising operation. Moreover, the reduced manifolds or latencies are provided with physical meaning, which further analysis, gives insight into key chemical reactions and interactions between chemical species, allowing for a deeper understanding of the chemical mechanism itself.

ACCEPTED MANUSCRIPT
Physics of Fluids
AIP Publishing
This is the author's peer reviewed, accepted manuscript. However, the online version of record will be different from this version once it has been copyedited and typeset.
PLEASE CITE THIS ARTICLE AS DOI: 10.1063/5.0167110

Accepted to Phys. Fluids 10.1063/5.0167110

NOMENCLATURE

Abbreviations

AE	Autoencoder
CFD	Computational Fluid Dynamics
CombML	Combustion Machine Learning
D	Decoder
DMD	Dynamic Mode Decomposition
E	Encoder
HODMD	High-Order Dynamic Mode Decomposition
ISAT	In Situ Adaptive Tabulation
LHS	Latin Hypercube Sampling
LPCA	Local Principal Component Analysis
MAE	Mean Absolute Error
NNs	Neural Networks
PCA	Principal Component Analysis
SPARC	Sample-Partitioning Adaptive Reduced Chemistry
TAE	Time-lag Autoencoder
TDAC	Tabulation of Dynamic Adaptive Chemistry

Greek letters

β	Regularization constant
β_1	Adam's optimizer first constant
β_2	Adam's optimizer second constant

ACCEPTED MANUSCRIPT
Physics of Fluids
AIP Publishing
This is the author's peer reviewed, accepted manuscript. However, the online version of record will be different from this version once it has been copyedited and typeset.
PLEASE CITE THIS ARTICLE AS DOI: 10.1063/5.0167110

ϕ	Equivalence ratio
ψ	Encoder Neural Network vectors
τ	Time interval
ξ	Decoder Neural Network vectors

Latin letters

$\mathbb{R}$	Dimensionality
$\mathscr{L}_{TAE}$	Time-lagged reconstruction loss
i_j	Coefficient of importance for the j-th variable
K	Number of time series
M	Number of time samples
N	Number of variables of high-dimensional space
n	Number of variables of low-dimensional space
Q	Number of permutation repetitions
R^2	Coefficient of determination
s	Linear regression coefficient of determination
$s_{q,j}$	Linear regression coefficient of determination for the j-th variable in q permutation repetition
T	Temperature
T_0	Initial temperature
t_0	First time step
t_{nt}	Last time step
X	Thermochemical sate vector

ACCEPTED MANUSCRIPT
Physics of Fluids
AIP Publishing
This is the author's peer reviewed, accepted manuscript. However, the online version of record will be different from this version once it has been copyedited and typeset.
PLEASE CITE THIS ARTICLE AS DOI: 10.1063/5.0167110

Accepted to Phys. Fluids 10.1063/5.0167110

X_0 Initial thermochemical state

Y_i Mass fraction of i-th species

Z Reduced manifold size

I. INTRODUCTION

Reacting flows are a common topic when it comes to industrial applications; this also means that efficient techniques that foresee their behavior are a necessity. Simulation techniques are a common technique used for the chemistry reactions' approximations, more precisely, CFD techniques have become part of common knowledge applications. Unfortunately, many of the current simulation techniques are unfeasible for their application in real industrial cases, mainly due to the computational overhead that they suppose. In brief, the number of equations increases linearly in the measure that more chemical species are added to the model, which becomes necessary to accurately describe the chemical evolution of the reacting system [1], at the same time, the computational cost for the solution increases exponentially in accordance to the number of chemical species. Additionally, there are many other physical phenomena to be considered, such as turbulence models, thermal irradiation models, acoustics, etc., from which, each one considers the addition of phenomenological closure models. For high-fidelity detailed chemistry mechanisms, indeed, the chemical reaction mechanism is frequently the most computationally demanding of a computational combustion model, as a reference, above 50 species, it takes about 90% of the computational effort [2].

As an answer to this issue, many efforts have been invested in the development of reduced efficient chemistry representation. Among these approaches, a very popular technique is the In Situ Adaptative Tabulation (ISAT) [3], which consists in storing thermochemical states through binary trees, however, the principal drawback of this technique is its high memory requirements, which makes the technique unviable for cases considering an elevated number of species, which are most of the real-cases scenarios [4].

Going further, many other techniques for reduced representations have been applied, such as the application of limited chemical steps [5] in which many reactions that lead to minor species are neglected; additionally, we can mention the development of skeleton mechanisms [6] that neglect many of the intermediate species while considering just major species and final products; it is to

ACCEPTED MANUSCRIPT
Physics of Fluids
AIP Publishing
This is the author's peer reviewed, accepted manuscript. However, the online version of record will be different from this version once it has been copyedited and typeset.
PLEASE CITE THIS ARTICLE AS DOI: 10.1063/5.0167110

mention the Tabulation of Dynamic Adaptive Chemistry (TDAC) methodology, in which there is a coupling of ISAT tabulation of previous results, plus a generation of skeleton mechanisms on the fly [7], however, the calculations of the reduced mechanisms can be demanding. Many other approaches can be mentioned, each one of them presenting its own advantages and disadvantages, however, none of the developed techniques represents a general solution to all the possible cases. With the current development of numerous Machine Learning techniques, it is possible to find a variety of works in which they are applied to combustion problems, giving place to the area of Combustion Machine Learning (CombML [8]). From these applications, it is important to mention the application of Principal Component Analysis (PCA) [9], a feature reduction technique, and the application of Neural Networks (NNs) as a substitute for the ODE time integrator [10], [11].

When it comes to PCA applications, its use as a feature reduction technique is limited since the resulting components are usually a linear combination of chemical species, which supposes some difficulties for the treatment of source terms ([12]). This technique has found a more efficient use as the application of Local PCA (LPCA) for clustering, as it is done in SPARC (Sample-Partitioning Adaptive Reduced Chemistry) chemistry models [13]. Another popular technique for feature reduction is the Dynamic Mode Decomposition (DMD), which has been proven useful for the identification of principal modes or frequency patterns in flows through matrix decompositions [14]. For the case of complex dynamics, High-Order Dynamic Mode Decomposition (HODMD) has proven effective [15] and introduces a windowing concept to apply DMD locally [16, 17].

On the other hand, NNs can be used as universal approximators, which means they are suitable for the representation of chemistry non-linearities, some works that present this kind of applications are such as [18, 19]; additionally, it is to mention hybrid applications, in which techniques such as PCA or HODMD decompositions are used for a dimensionality reduction of datasets, the same data that is used afterward for the training of neural networks (NNs), which integrate the variables in time [20]; the application of these methodologies reduces significantly the computational cost of the hyperparameters optimization. However, as it is observed in [21], a drawback of their implementation is the huge dimensions of the training datasets, which might require the application of generative models. Moreover, the application of neural networks and CombML techniques in general, address the problem mostly from a statistical perspective in most cases and give back very little understanding of principal species that play a role in the mechanism, as well as the principal reactions involved.

Nonetheless, this neural network function approximation can be used for dimensionality reduc-

ACCEPTED MANUSCRIPT
Physics of Fluids
AIP Publishing
This is the author's peer reviewed, accepted manuscript. However, the online version of record will be different from this version once it has been copyedited and typeset.
PLEASE CITE THIS ARTICLE AS DOI: 10.1063/5.0167110

tion, as it is pictured in autoencoder (AE) applications, such as [22] and [23]. As expected, even if an AE provides with a better representation of chemistry non-linearities, the resulting reduced representations are difficult to interpret, leading to model closure issues. In the present work, a time shift will be added to an AE network architecture, which gives place to a time-lag autoencoder (TAE) [24]. The application of such a network allows to assess an AE topology for the task of time integration itself, as well as it adds a temporal characterization to the reduced components, which becomes a dynamical characterization of the chemistry. Additionally, this kind of network allows for the development of reduced dynamic models [25], which are expressed in terms of the network's reduced components. Furthermore, it should be assessed if the reduced components enable the recognition of principal chemical features, yielding better insights about the chemistry mechanism behavior from a dynamic perspective.

In the next sections, a conceptualization of the TAE approach is provided with its applicability as a feature extractor of chemistry kinetics (Sec. II A). Additionally, some notions regarding PCA as a reduction technique are given, principally when used in the form of neural networks (subsection II B). A proof-of-concept study for hydrogen-air homogeneous autoignition combustion and a discussion of the capability of TAE's for chemistry estimation, such an experiment intends to provide some general guidelines that will help for the technique implementation in other reactive scenarios. The paper ends with a summary of our main findings and future perspectives to leverage the application of the proposed methodology in combustion science.

II. THEORETICAL BACKGROUND

A. Time-lagged Autoencoders

Auto-encoders (AE) are a type of neural network that is used for finding reduced representations of a given input vector [22]. In its original form, the network hyperparameters search to minimize the error of a loss function while reconstructing the same input parameters as the network's outputs, through the application of two main operations, which are named the encoder (E) and decoder (D) networks. These networks are separated by a bottleneck layer, from which it is possible to obtain this reduced representation. Generally speaking, the encoder network is the one that takes the input parameters and transforms them into the reduced representation, while the decoder network takes this reduced or latent representation, and reconstructs the original parameters.

ACCEPTED MANUSCRIPT
Physics of Fluids
AIP Publishing
This is the author's peer reviewed, accepted manuscript. However, the online version of record will be different from this version once it has been copyedited and typeset.
PLEASE CITE THIS ARTICLE AS DOI: 10.1063/5.0167110

As suggested by the technique's name, a TAE follows the same concept and architecture as an AE, with the difference that a time shift is applied between the network's inputs and outputs. In particular, the inputs and outputs to be applied are the respective thermochemical states $X \in \mathbb{R}^N$ of N state variables (i.e. species mass fractions and temperature), in the time interval $\tau \in t \subset \mathbb{R}^+$. More specifically, it is assumed that the time interval τ is discretized by M time steps with a time-step size Δt. Thus, the modeling of a dynamical system is posed as a time-series problem propagating from the initial state X_0.

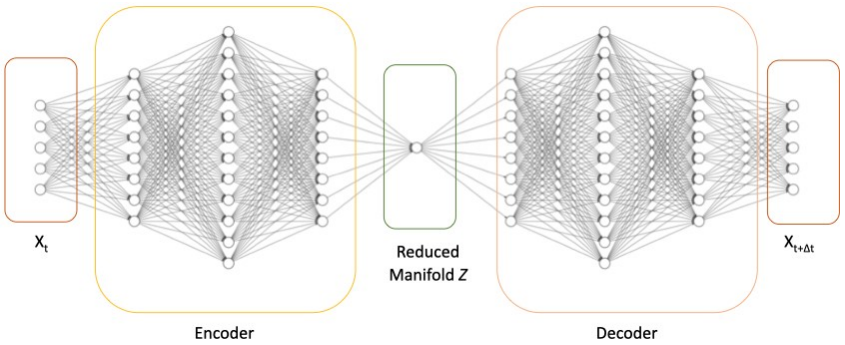


FIG. 1: Schematic that describes the composition of a TAE network, such network architecture is to be applied for a dynamical analysis of chemistry kinetics.

Mathematically, the encoder network can be understood as a sequence of nonlinear transformations that brings the vector X from dimensionality $\mathbb{R}^N \rightarrow \mathbb{R}^n$,

$$Z_t = E(X_t). \tag{1}$$

Here, Z represents the low-dimensional latent space $Z \in \mathbb{R}^n$ with $n << N$; for the case of the decoder, the transformation is such that enables the reconstruction of $\mathbb{R}^n \rightarrow \mathbb{R}^N$, but with a temporal advancement, therefore, it is possible to write:

$$\tilde{X}_{t+\Delta t} = D(Z_t) = D(E(X_t)). \tag{2}$$

ACCEPTED MANUSCRIPT
Physics of Fluids
AIP Publishing
This is the author's peer reviewed, accepted manuscript. However, the online version of record will be different from this version once it has been copyedited and typeset.
PLEASE CITE THIS ARTICLE AS DOI: 10.1063/5.0167110

In the present study, fully connected neural networks are employed to define the encoding and decoding models, as shown in Fig. 1. Thus, the final result is a deep learning model $\tilde{X}_{t+\Delta t} = D_\xi(E_\psi(X_t))$ parameterized by vectors ξ and ψ, like trained deep neural networks. Thus, the final goal is to find an encoding and decoding which minimizes the time-lagged reconstruction loss $\mathscr{L}_{TAE}$,

$$\mathscr{L}_{TAE} = \sum_{0}^{M-1} \|X_{t+\Delta t} - \tilde{X}_{t+\Delta t}\|^2, \tag{3}$$

for a family of functional mappings E and D. Additionally, neural networks are trained through the minimization of a loss function, which ideally, should reflect physical laws constraints. Following this aim, we proposed a physics-aware loss function to ensure mass conservation at each time-step:

$$\mathscr{L}_{TAE} = \frac{1}{M-1}\sum_{0}^{M-1} \|X_{t+\Delta t} - \tilde{X}_{t+\Delta t}\|^2 + \frac{\beta}{M-1}\sum_{0}^{M-1} \|\sum Y_{t+\Delta t} - \sum \tilde{Y}_{t+\Delta t}\|^2. \tag{4}$$

Variables with $(\tilde{\cdot})$ refer to the TAE's output, and Y stands for the species mass fractions, which should sum up to one. β is the regularization constant, that is to say, how much the violation of mass conservation will be penalized; lastly, M stands for the number of time samples.

B. Comparison with other typical reduction methods

Considering that TAE can be used as a feature extraction technique, it is important to understand the benefits that a TAE implementation holds in relation to traditional decompose techniques. Even though it is true that PCA gained so much popularity due to its low computational cost and the fact that almost any data matrix can be decomposed by the technique [26], [27]. Moreover, PCA is also a linear technique, which means that it assumes the existence of a linear map that can represent the dataset, which is not necessarily true in many real-world cases [28]. Regardless of this intrinsic characteristic, even if the method can be applied successfully for the feature extraction of the thermochemical state [29], the resulting variables might not be of good significance for the application of predictive methods for the chemistry temporal advancement, which can be rather complex.

Since autoencoders hold a non-linear character, therefore, the question to answer is how much the reconstruction cost is altered by the application of a TAE while keeping a constant number of features for both methods. Additionally, it should be emphasized that TAE gives direct information

ACCEPTED MANUSCRIPT
Physics of Fluids
AIP Publishing
This is the author's peer reviewed, accepted manuscript. However, the online version of record will be different from this version once it has been copyedited and typeset.
PLEASE CITE THIS ARTICLE AS DOI: 10.1063/5.0167110

about the thermochemical state evolution, while PCA techniques just provide information about a present state. Because it is intended to analyze both techniques under equality of conditions, a one-layer autoencoder is proposed to mimic the PCA technique, called a PCA-like autoencoder. Since PCA is a linear technique, the PCA-like autoencoder is constructed by applying a linear activation function in the hidden layer. Thus, the feature variables are obtained from a linear transformation of the original thermochemical state. A diagram of this type of autoencoder is visualized in Figure 2.

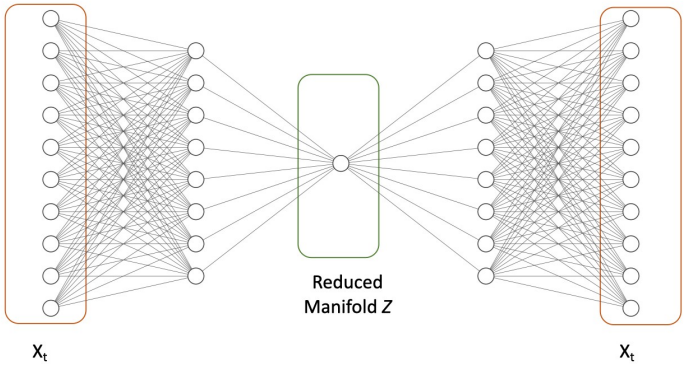


FIG. 2: Schematic of the parts of a PCA-like autoencoder, it is to notice the difference in complexity regarding the displayer TAE architecture.

III. COMPUTATIONAL EXPERIMENT.

In the present work, we test the capability of TAE as a feature extraction technique, while it performs the time integration of the thermochemical variables, this can be traduced into a time integration that eliminates unnecessary inputs; in this case, both capabilities will be applied to hydrogen combustion. The dataset for training the machine learning models is obtained with isobaric batch reactors simulations, developed with the Cantera software [30]. The simulations were performed using a constant time step, which should be small enough to ensure a sufficient number of time samples for different ignition conditions, additionally, it should provide a good description of the chemical species' mass fractions time evolution, this can be ensured if the time step is in agreement with the inverse reaction rate chemical time scale [31], which is in order of

ACCEPTED MANUSCRIPT
Physics of Fluids
AIP Publishing
This is the author's peer reviewed, accepted manuscript. However, the online version of record will be different from this version once it has been copyedited and typeset.
PLEASE CITE THIS ARTICLE AS DOI: 10.1063/5.0167110

10^{-4}, therefore, in order to capture all the chemical time scales with detail, the applied time step is $10^{-7}s$ for a time length of $0.5ms$. The reduced version of the University of San Diego chemistry mechanism for hydrogen combustion was used [32], which consists of 9 species and 21 reactions. Since non-reacting species are neglected, the resulting thermochemical state holds dimensionality $X \in \mathbb{R}^9$, which considers 8 reacting chemical species, plus temperature. The application of a small mechanism helps to understand the principles behind this application without any loss of generality.

Furthermore, correct modeling of chemical kinetics is characterized by multi-physics and multi-scale phenomena, in which the properties might vary by several ranges of magnitudes. As the main goal here is to build predictive models capable of producing reliable predictions for different conditions, the reconstruction model must be able to bear a wide range of scenarios defined by state variables, also called control variables, such as temperature, pressure, and equivalence ratio in various flame configurations. So, the training dataset is composed of K time-series of an isobaric zero-dimensional reactor using different ignition conditions.

For the different ignition conditions, different values of equivalence ratio (ϕ) and initial temperature are considered. A Latin Hypercube Sampling (LHS) was implemented [33] with the Python PyDOE library [34]. A uniform distribution is used to obtain 100 samples in a region delimited by $\phi \in [0.9, 1.2]$ and $T \in [1100, 1220]$. The number of samples was chosen through a try-error approach in which the number of ignition cases was progressively increased from 50 to 75, and finally to 100, where a satisfactory accuracy was achieved. Here, it is worth highlighting that for cases with higher dimensionality such as methane combustion the number of ignition cases needed to train the TAE model might be larger. In such cases, generative models can be used to build large training datasets [21]. The resulting sampling is visible in Figure 3.

The thermochemical state vectors are defined using mass fractions values and temperature values, the dataset normalization is done with a general maximum value per quantity, meaning that each quantity is divided by its maximum occurrence among all the datasets. This allows all the datasets to lie in the same manifold. The applied time shift equals a single time step.

At the same time, different sizes of bottleneck layers were explored; the latent spaces range is described by $Z \in [1, 4]$, therefore, four different networks are meant to be studied. The networks are built using the Python Tensorflow package[35], each of the networks' hyperparameters has been tunned using the optimization functions available in the package Keras Tuner [36], and this optimization process provides a set of different networks that differ between them just in the num-

ACCEPTED MANUSCRIPT
Physics of Fluids
AIP Publishing
This is the author's peer reviewed, accepted manuscript. However, the online version of record will be different from this version once it has been copyedited and typeset.
PLEASE CITE THIS ARTICLE AS DOI: 10.1063/5.0167110

Accepted to Phys. Fluids 10.1063/5.0167110

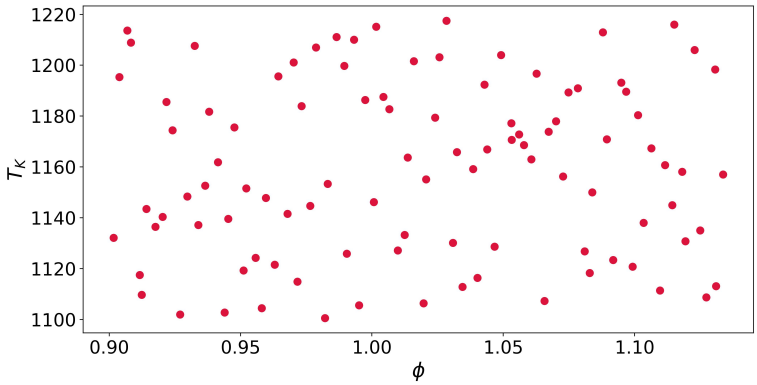


FIG. 3: LHS sampling of ignition conditions considered for training data

ber of neurons per layer, which can be observed in Table I, these values are specifically obtained making use of a Random Search algorithm, while the rest parameters are obtained making use of grid search.

The Leaky ReLu functions used, hold an activation constant of 1×10^{-3}, which allows a better representation of combustion non-linearities. Moreover, a sigmoid activation function is used in the latent layer to constrain the latent space in $Z \in [0,\ 1]$ [24]. The neural network parameters are initialized using a Glorot normal [37]. The physics-aware loss function available in Equation (4) is employed to train the TAE model. The Adam optimizer [38] is employed with a learning rate of 1×10^{-3}, $\beta_1 = 0.9$, and $\beta_2 = 0.999$, for a total of 2000 epochs. Early stopping is used with a minimum change of 1×10^{-4} for five epochs. Due to the computational overhead that implies using the whole dataset, the optimization process involves the utilization of 75% of the whole dataset, while posterior cross-validation is implemented using the whole dataset. In each phase, 80% of the data is used as a training dataset while the remaining 20% is used as validation data.

Moreover, the resulting latent space will be used for the determination of latent chemical species; with latent species, it is intended chemical species that hold a strong correlation with the latent variables. The aim is to assess whether these chemical carriers may bear the potential to act as input features for the development of physically explainable reduced-order models. The identification is carried out through a correlation analysis using Kendall's Tau B correlation index, due to its boundedness and resistance to outliers [39]. This feature identification is achieved

ACCEPTED MANUSCRIPT
Physics of Fluids
AIP Publishing
This is the author's peer reviewed, accepted manuscript. However, the online version of record will be different from this version once it has been copyedited and typeset.
PLEASE CITE THIS ARTICLE AS DOI: 10.1063/5.0167110

TABLE I: TAE Network's Architectures

Operation	Number Layer	Number of Neurons				Activation Function
		Z=1	Z=2	Z=3	Z=4	
Input Layer	1	9	9	9	9	Linear
Encoder	2	161	245	318	210	Leaky ReLu
	3	339	319	227	186	Leaky ReLu
	4	46	79	96	76	Leaky ReLu
	5	322	336	285	324	Leaky ReLu
Code Layer	6	1	2	3	4	Sigmoid
Decoder	7	322	336	285	324	Leaky ReLu
	8	46	79	96	76	Leaky ReLu
	9	339	319	227	186	Leaky ReLu
	10	161	245	318	210	Leaky ReLu
Output Layer	11	9	9	9	9	Linear

through the usage of the Scipy package [40].

IV. RESULTS.

As a first step, it is important to assess the quality of reconstruction that is achieved using the TAE networks in training data. For this, the training ignition cases will be reconstructed, and the quality of these reconstructions will be assessed with the analysis of the R^2 scores [41], which evaluates the quality of the model's output, scoring it within a constrained range of $[0, 1]$, where 0 is given to a poorly performing model, and 1 to a perfect prediction. A general R^2 is to be calculated for each ignition case, this means calculating the score for each one of the components of the thermochemical variables and averaging these results. This process provides a mean R^2 score of 0.991, with a standard deviation of $4.6x10^{-3}$, such results are an indicator of a satisfactory reconstruction. The histogram with the distribution of the respective R^2 scores is available in Figure 4, this for a TAE model $z = 1$, in which a lower accuracy is to be expected.

ACCEPTED MANUSCRIPT
Physics of Fluids
AIP Publishing
This is the author's peer reviewed, accepted manuscript. However, the online version of record will be different from this version once it has been copyedited and typeset.
PLEASE CITE THIS ARTICLE AS DOI: 10.1063/5.0167110

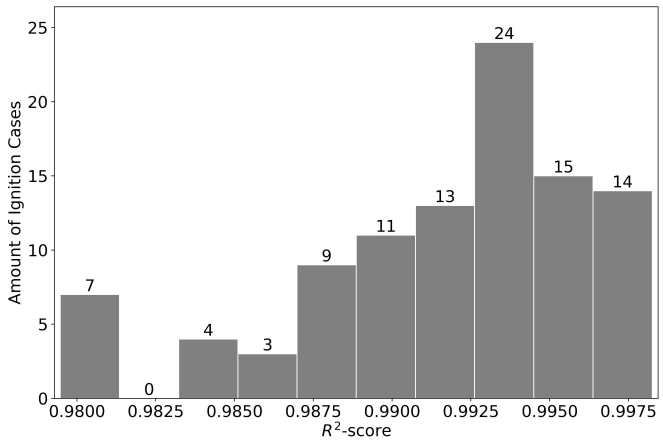


FIG. 4: Distribution of R2 scores for training ignition cases tested in TAE model $z = 1$

, the y-axis describes the count of ignition cases for each range of values, while the x-axis presents the ranges of R^2 scores.

Since the studied models give a good reconstruction in training data, the outcomes of this study will be organized as follows: first, an analysis of the TAE-reduced manifolds will be presented. For this test, ignition conditions that can be seen as interpolation cases are considered; interpolation makes reference to conditions contained inside the limits of training data. Lastly, a comparison of TAE as a reduction technique will be performed regarding PCA capabilities for a range of latent space dimensions or a number of principal components, adding in this case, some extrapolation conditions, that is to say, conditions outside of the training limits. The interpolation and extrapolation conditions considered are available in Tables II and III respectively.

TABLE II: Interpolation Ignition cases

Case	$T[K]$	ϕ
1	1160	0.93
2	1200	1.0
3	1130	1.10

ACCEPTED MANUSCRIPT
Physics of Fluids
AIP Publishing
This is the author's peer reviewed, accepted manuscript. However, the online version of record will be different from this version once it has been copyedited and typeset.
PLEASE CITE THIS ARTICLE AS DOI: 10.1063/5.0167110

Accepted to Phys. Fluids 10.1063/5.0167110

TABLE III: Extrapolation Ignition cases

Case	$T[K]$	ϕ
1	1160	0.85
2	1300	1.0

A. Time-lag Auto-encoder Reduced Manifolds.

For obtaining these manifolds, the TAE network is fed with K time series that describe the hydrogen-air homogeneous autoignition problem in different ignition scenarios at the time interval $[t_0, t_{n_t-1}]$, where n_t stands for the number of time steps. The expected output is the future values of the time series, it is to say, the state vectors for the time interval $[t_{0+\Delta t}, t_{n_t}]$.

Going further, we evaluate the ability of the TAE model to generalize to unseen ignition scenarios. Specifically, we assess the ability of the TAE integrator in interpolation and extrapolation scenarios. The interpolation scenario represents the homogeneous ignition cases described in Table II. The R^2 score will be used again to assess the quality of reconstruction; this assessment is the first step in the analysis since the validity of the latent manifold is dependant on the outputs reconstruction accuracy. The R2 score values are available in Figure 5.

After observation, it is visible that can see that the models retain a good representation of the dynamic behavior, returning R^2 scores larger than 0.96. Moreover, we note that in the latent space with a dimension equal to 1, the model accuracy decreases. That might be explained by the fact that combustion is a highly non-linear phenomenon, and latent spaces with larger dimensions are required to describe accurately the whole thermochemical states, mainly for more complex chemical mechanisms.

Once the accuracy of the reconstruction is assessed, the next objective is to observe the latent manifolds for different bottleneck sizes, in order to assess the minimum number of latent variables for a good reconstruction. Such manifolds will be associated with chemical species, in order to obtain what can be considered as *chemical carriers*. To perform such an association, a correlation analysis is performed using the Kendall Tau B correlation index due to its robustness and resistance to outliers [39]; in other words, there is a comparison of the temporal behavior between the thermochemical variables, and the reduced latencies from the same ignition case. In general, when there is a similarity between two objects, they are likely to hold a high positive correlation index

ACCEPTED MANUSCRIPT
Physics of Fluids
AIP Publishing
This is the author's peer reviewed, accepted manuscript. However, the online version of record will be different from this version once it has been copyedited and typeset.
PLEASE CITE THIS ARTICLE AS DOI: 10.1063/5.0167110

Accepted to Phys. Fluids 10.1063/5.0167110

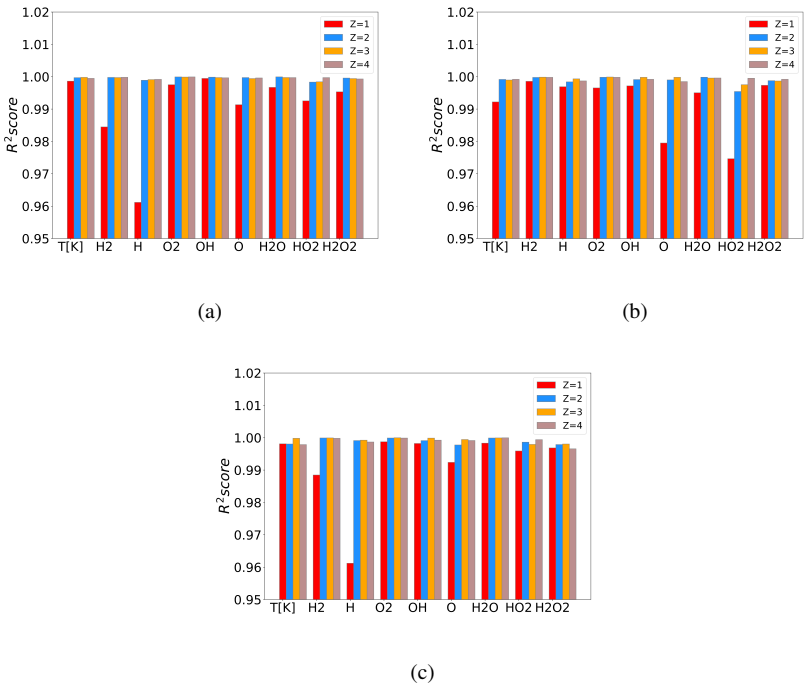


FIG. 5: Interpolation cases R2 scores for all reduced manifolds sizes (Z), y-axes describes the number of latent variables considered by the model, (a) ignition case ($\phi = 0.93, T_0 = 1160K$), (b) ignition case ($\phi = 1.0, T_0 = 1200K$), (c) ignition case ($\phi = 1.10, T_0 = 1130K$).

[42]. The resulting associated species are available in Table IV.

The chemical carriers presented in Table IV are identifiable due to the dataset preprocessing, since the easiness of pattern or key features recognition is related to the representation of datasets under study [42]. It is important to highlight the presence of just one set of chemical carriers, which is due to the repetitiveness of the same chemical carriers for all the interpolation ignition cases considered, for the cases in extrapolation, the majority of chemical carriers are retained, however, one chemical specie might change in each case, i.e. for extrapolation case 1 ($\phi = 0.85, T_0 = 1160K$) the combination T, O is favored in $z = 2$, while in extrapolation case 2 ($\phi = 1.0, T_0 = 1300K$) the combination T, H is preferred, nevertheless, this change in preference is due to a very small variation in correlation indexes, around 0.05 maximum, holding always a high similarity with the chemical

ACCEPTED MANUSCRIPT
Physics of Fluids
AIP Publishing
This is the author's peer reviewed, accepted manuscript. However, the online version of record will be different from this version once it has been copyedited and typeset.
PLEASE CITE THIS ARTICLE AS DOI: 10.1063/5.0167110

TABLE IV: Chemical Carriers Identification

Manifold Size (Z)	Chemical Carriers
1	T
2	T, O
3	H2O2, T, OH
4	O2, O, T, T

carriers obtained in interpolation; this can be interpreted as a potential change in the chemical carriers in agreement to the analyzed combustion regime when applied to a wider range of ignition cases. However, it is seen that temperature is a key variable appearing as a chemical carrier for all models. Moreover, in all cases we can see that for the latent space with dimensionality 4, there is a repetitiveness of the temperature. That may suggest that a latent space with dimension 3 is enough to describe accurately the whole thermochemical states, and for larger dimensions, the model can be over-parametrized; such over-parametrization is likely a product of high-dimensionality in the latencies manifold size, making it difficult to recognize patterns properly [43], producing a non-optimum identification; it is to keep in mind that the chemical carriers are obtained through an interpretation of the patterns given by the neural network's signals. From this analysis, it is also possible to relate the TAE decomposition with the DMD-like techniques, considering that both obtain a dynamical characterization, that is however, from different perspectives, while DMD-like techniques use a regression approach [16], while TAE uses a variational approach [24].

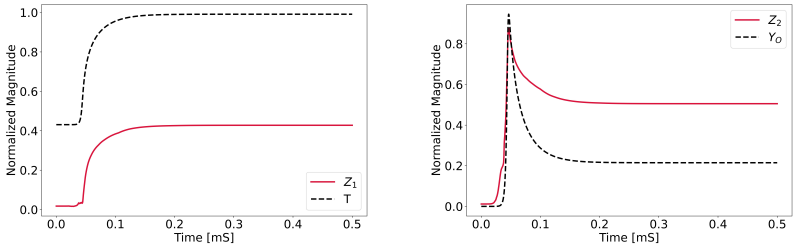
Aiming to show the degree of correlation between the *chemical carriers* in the interpolation cases, a visual comparison between the chemical carrier and the respective latent variable is displayed in Figure 6, while an example of the obtained correlation indexes is available in Figure 7. This is done for the ignition case ($\phi = 1.0, T_0 = 1200K$), for a latent space size $z = 2$.

On the other hand, Figure 8 shows the latent space behavior of $Z = 2$ for the three ignition cases under study. It is noted an identical behavior, which just differs in relation to the magnitude of the latent variables at each time step for the different ignition scenarios. Furthermore, there is a clear time shift for each ignition case due to the different ignition delay times, which are accurately represented.

The repetition of *chemical carriers* for different ignition conditions, suggests that there might

ACCEPTED MANUSCRIPT
Physics of Fluids
AIP Publishing
This is the author's peer reviewed, accepted manuscript. However, the online version of record will be different from this version once it has been copyedited and typeset.
PLEASE CITE THIS ARTICLE AS DOI: 10.1063/5.0167110

Accepted to Phys. Fluids 10.1063/5.0167110



(a) (b)

FIG. 6: Latent variables associations for the model manifold size $z = 2$, ignition case ($\phi = 1.0, T_0 = 1200K$), (a) presents the behavior of the first latent variables with its associated thermochemical variable, while (b) presents the same items for the second latent variables; it is to mention that the same association is repeated for all the tested ignition cases.

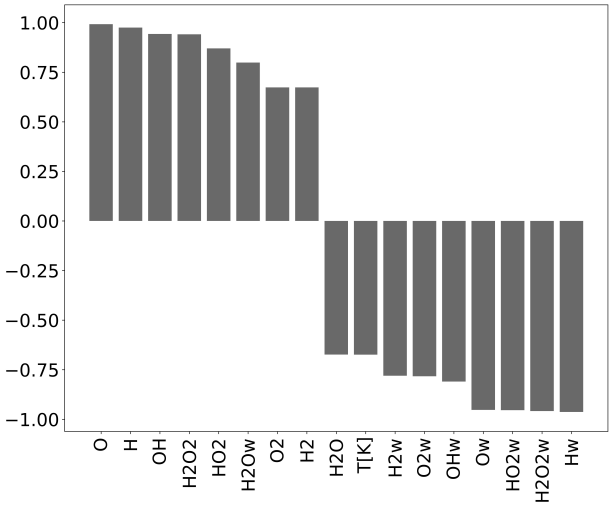
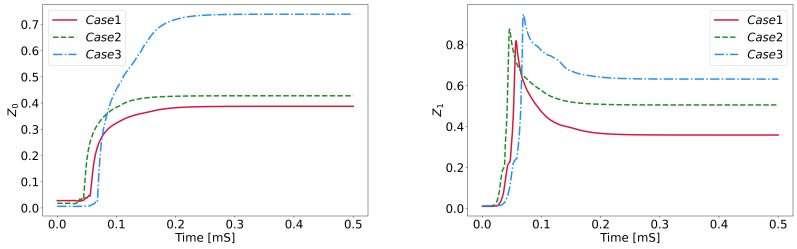


FIG. 7: Kendall Tau B correlation indexes for the second latent variable, manifold size $z = 2$, ignition case ($\phi = 1.0, T_0 = 1200K$), the tendency to extreme correlation indexes is repeated in all cases.

ACCEPTED MANUSCRIPT
Physics of Fluids
AIP Publishing
This is the author's peer reviewed, accepted manuscript. However, the online version of record will be different from this version once it has been copyedited and typeset.
PLEASE CITE THIS ARTICLE AS DOI: 10.1063/5.0167110

Accepted to Phys. Fluids 10.1063/5.0167110



(a) (b)

FIG. 8: Latent space visualization for manifold size $z = 2$, (a) shows the first latent variable, and (b) the second latent variable. The Case 1 curve stands for the ignition case ($\phi = 0.93, T_0 = 1160K$), Case 2 for ignition case ($\phi = 1.0, T_0 = 1200K$), and Case 3 for ignition case ($\phi = 1.10, T_0 = 1130K$)

.

be key chemical species that are capable of explaining the variance of a chemical reaction. Aiming to verify this fact, a permutation importance test is proposed [44]; conceptually, the test consists of the elaboration of a model that relates a set of inputs and output variables, once the model is established, each input variable will be permuted individually. If the permuted variable is meaningful for the description of the output, the error in predictions will increase, and therefore a high permutation weight will be assigned.

For testing purposes, a linear regression model will relate the latent chemical species with the future values of the thermochemical state vector, the intentionality of this is to relate the direct statistical relationships between the variables, which can be very key for finding chemical species interactions or dependencies. For the scoring process, an R^2 score is used again. Additionally, the Sci-Kit Learn built-in function for the score of determination, which means that it allows also negative values, a situation that reflects the model performing worst than a constant model with the mean value as an output; this consideration enables the R^2 score to be in the range of $[-\infty, 1]$.

Moreover, the Sci-Kit Learn permutation importance scoring follows the Equation:

$$i_j = s - \frac{1}{Q}\sum_{n=1}^{Q} s_{q,j}. \tag{5}$$

ACCEPTED MANUSCRIPT
Physics of Fluids
AIP Publishing
This is the author's peer reviewed, accepted manuscript. However, the online version of record will be different from this version once it has been copyedited and typeset.
PLEASE CITE THIS ARTICLE AS DOI: 10.1063/5.0167110

where i_j stands for the coefficient of importance for the j-th variable, s is the R^2 score from the linear regression without permuting the input variables, and $s_{q,j}$ is the resulting R^2 with a single scrambled input variable. Q refers to the number of repetitions in which the permutation is performed and the determination score is re-calculated [45].

However, Equation (5) also provides a warning, and it is that from case to case, the magnitude of importance weights can vary greatly, not only depending on the number of latent variables considered but also in relation to the scrambled variable. Therefore, for each output variable, the resulting importance weights will be added, and this total will be used to normalize the actual importance weights; such a procedure searches to allocate all the weights in the same numerical scale for better interpretation. For the case of the latent space $z = 1$, normalization is applied with the consideration of the maximum importance weight among all the output variables. Figure 9 shows the results of the permutation tests. Here, the importance of the chemical carriers for each thermochemical variable is displayed in proportion to the height of the bar.

It is observed that just in the case of latent space $z = 1$ the input variables do not hold influence in every output of the mechanism, this finding is also representative of the lower R^2 scores that come from the network's reconstruction in Figure 5. Additionally, from this permutation importance test it is to mention the Figure 9 (c), which the analysis is made for the manifold size $z = 3$. In general, the most important radicals for the ignition development in hydrogen combustion are OH and HO2 [46], however, in this case, the network has chosen as the third chemical carrier the H2O2 chemical species, this is sustained since the TAE network searches for statistical patterns, and therefore, HO2 is not considered since H2O2 holds part of its statistical variance, which is only surpassed by the influence of temperature, therefore, the network judges HO2 as temperature dependent, same that holds significative importance for other chemical species; lastly, in Figure 9 (a) it is observed that temperature does not hold a direct significance for H2O2, justifying why this variable is chosen. Another similar analysis is possible from Figure 9 (b), in which the strong relationship between OH, HO2, and Temperature is displayed, at the same time, the interactions between the H and O species is highlithed; the strong bond between T and OH does not appear in Figure 9 (c) since OH is considered as a chemical carrier. This analysis reinforces the fact that TAE highlights the chemical species that are likely to preserve the variance of the mechanism, moreover, such a result suggests that TAE's temporal integration is likely to eliminate unnecessary variables for the description of future thermochemical spaces.

ACCEPTED MANUSCRIPT
Physics of Fluids
AIP Publishing
This is the author's peer reviewed, accepted manuscript. However, the online version of record will be different from this version once it has been copyedited and typeset.
PLEASE CITE THIS ARTICLE AS DOI: 10.1063/5.0167110

Accepted to Phys. Fluids 10.1063/5.0167110

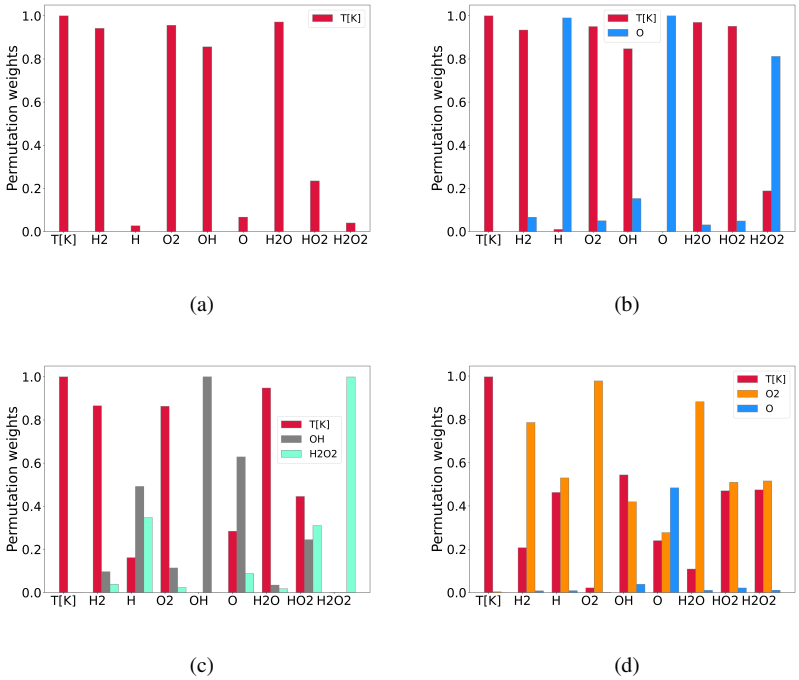


FIG. 9: Permutation importance test weights for chemical carriers for ignition case ($\phi = 1.10, T_0 = 1130K$); (a) latent space $z = 1$, (b) latent space $z = 2$, (c) latent space $z = 3$, and (d) latent space $z = 4$. In each picture, the bars' height is proportional to the importance of each chemical carrier for each of the considered thermochemical variables, which are displayed on the x-axis.

B. Comparison with Principal Component Analysis.

The most important point of comparison regarding other reduction techniques is the error they introduce at the moment of reconstruction. To avoid circumstances that could benefit one model over another, the PCA-Autoencoder is trained with the same normalization and datasets used for the TAE training, afterward, their performance is evaluated in interpolation cases.

For this case, the chosen scoring technique is the Mean Absolute Error (MAE) [47], which gives an estimation of the average deviation between the predicted values and the real ones, the inclusion

ACCEPTED MANUSCRIPT
Physics of Fluids
AIP Publishing
This is the author's peer reviewed, accepted manuscript. However, the online version of record will be different from this version once it has been copyedited and typeset.
PLEASE CITE THIS ARTICLE AS DOI: 10.1063/5.0167110

of this scoring technique ensures a more comprehensive evaluation of the model's performance. Additionally, since the chemical variables are constrained between values of zero and one, the MAE also gives information on the average percentual error in the predictions. The results for TAE and PCA reconstructions are available in Figure 10, it is to notice the significant difference in the reconstruction errors, which are around four times higher in the PCA technique than in TAE.

From Figure 10, it is evident that TAE presents better reconstruction capabilities when considering each chemical species separately; it is to highlight the order difference in the error that is visible in the Figures comparison; since the error is significantly lower for the TAE reconstructions, it is possible to state this technique as more accurate. Moreover, PCA is not able to reproduce the fast phenomena in the chemical reaction, fact that is visible in Figure 11, in which different reconstruction curves are presented for a latent space size $z = 2$, it is observable in the OH and HO2 curves that PCA fails to represent any behavior that takes place for small periods of time, in other words, it is unable to reproduce small chemical time scales.

As expected, the reconstruction error is higher in extrapolation cases, as it is possible to observe in Figure 12. However, it is to remark that reconstruction behaviors maintain a similar behavior to the ones in interpolation conditions, this indicates that models actually associate the extrapolation case with the closest known conditions. For the TAE case, the temperature deviations might be also a product of the MSE reconstruction that is applied, since the physically aware term from Equation 4 regularizes only chemical mass fractions.

In agreement with the provided results, although PCA provides a much faster method to reduce the dimensionality of combustion systems via matrix decomposition [48], it is observed that TAE reconstruction is more accurate than PCA. Going further, TAE provides interpretable latent spaces allowing an understanding of the underlying physics of the chemical kinetics that may be useful to construct reduced chemical mechanisms. However, when it comes to defining a minimum amount of reduced variables, PCA brings results more straightforwardly, considering the eigenvalues of the matrix decomposition, while TAE requires space exploration for the determination of optimums.

ACCEPTED MANUSCRIPT
Physics of Fluids
AIP Publishing
This is the author's peer reviewed, accepted manuscript. However, the online version of record will be different from this version once it has been copyedited and typeset.
PLEASE CITE THIS ARTICLE AS DOI: 10.1063/5.0167110

Accepted to Phys. Fluids 10.1063/5.0167110

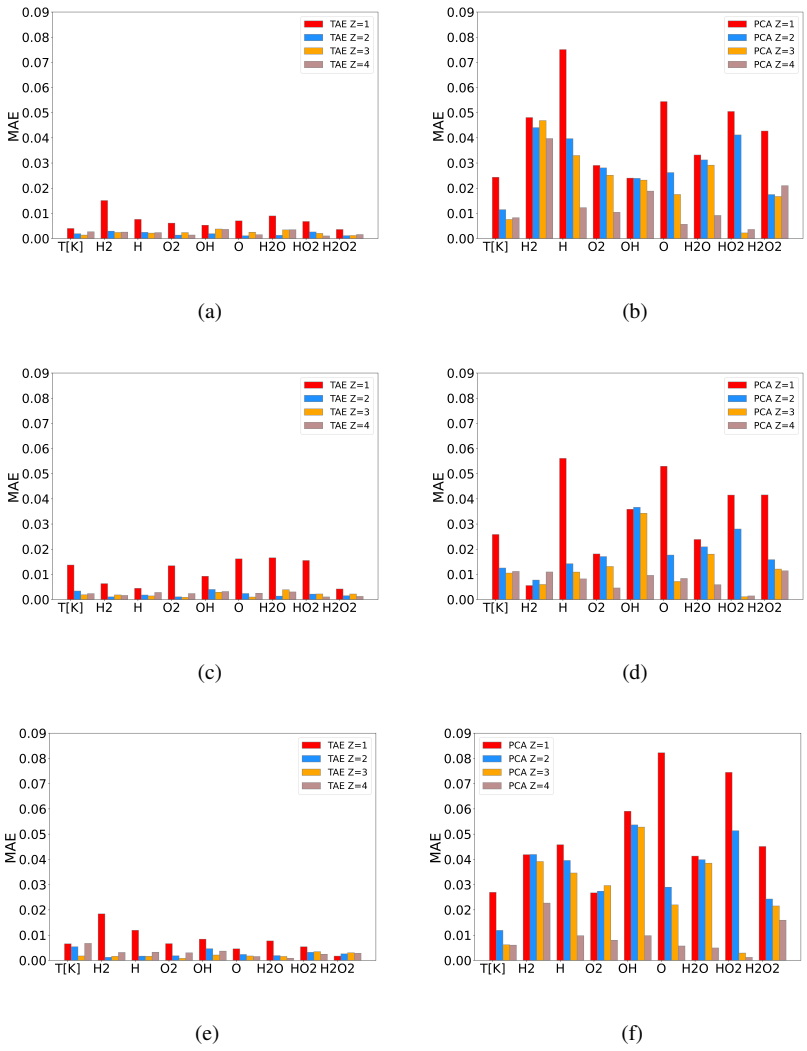
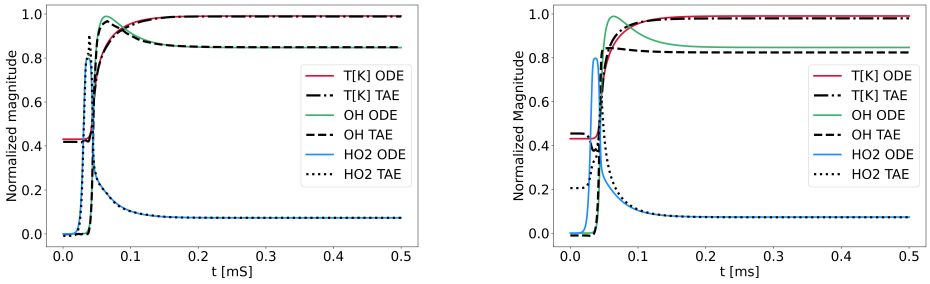


FIG. 10: TAE and PCA MAE scores for interpolation cases and all manifold sizes (Z), figures at the left present TAE's values, while figures at the right present PCA's ones; (a) and (b) present ignition case ($\phi = 0.93, T_0 = 1160K$), (c) and (d) the ignition case ($\phi = 1.0, T_0 = 1200K$), and, (e) and (f) ignition case ($\phi = 1.10, T_0 = 1130K$). All the graphs hold the same y-axis limits for a better comparison of magnitudes between techniques.

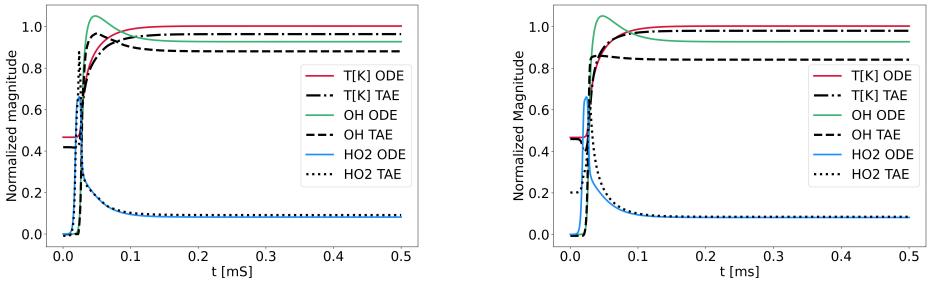
ACCEPTED MANUSCRIPT
Physics of Fluids
AIP Publishing
This is the author's peer reviewed, accepted manuscript. However, the online version of record will be different from this version once it has been copyedited and typeset.
PLEASE CITE THIS ARTICLE AS DOI: 10.1063/5.0167110

Accepted to Phys. Fluids 10.1063/5.0167110



(a) (b)

FIG. 11: Normalized chemical species reconstruction for a latent space size $z = 2$ in interpolation, at the left the TAE reconstructions are available, while at the right the PCA reconstructions are displayed for an ignition case $\phi = 1.0, T_0 = 1200K$.



(a) (b)

FIG. 12: Normalized chemical species reconstruction for a latent space size $z = 2$ in extrapolation, at the left the TAE reconstructions are available, while at the right the PCA reconstructions are displayed for an ignition case $\phi = 1.0, T_0 = 1300K$.

ACCEPTED MANUSCRIPT
Physics of Fluids
AIP Publishing
This is the author's peer reviewed, accepted manuscript. However, the online version of record will be different from this version once it has been copyedited and typeset.
PLEASE CITE THIS ARTICLE AS DOI: 10.1063/5.0167110

V. CONCLUSION.

The modeling of combustion chemistry kinetics is a challenging problem for which, no absolute solutions have been found. Therefore, exploration of new perspectives and methodologies is a must. Time-lag auto-encoders (TAE) are an interesting ML model, which provides a time integration scheme while allowing the identification of reduced representations. This paper centers on TAE applications for dimensionality reduction, which holds a dynamical characterization of chemistry kinetics. As a dimensionality reduction technique, TAE provides a higher quality of reconstruction in relation to standard techniques (principally PCA), while providing at the same time, a physical interpretation of the reduced manifolds (*chemical carriers* identification). Considering that the *chemical carriers* are the thermochemical variables that hold all the statistical variance of the mechanism, the set of network transformations can be easily interpreted as a denoising operation, in which the influence of statistically meaningless variables is canceled. Moreover, with the appropriate treatment, the *chemical carriers* could lead to the study and better interpretation of key chemical pathways, improving the knowledge of chemical mechanisms.

Further work must be developed for the assessment of TAE's application as a time integrator, as well as to assess its benefits regarding simpler neural network architectures. Additionally, better model training techniques should be tested for the improvement of autoregressive models, since most of the literature cases consider just the mass conservation principle leaving the temperature variable without any sort of regularization or constraint, allowing for the arising of outliers [49]. Lastly, possible applications for the identified latencies for the development of alternative surrogate models should be evaluated.

ACKNOWLEDGMENTS

This research is funded by the Fonds de la Recherche Scientifique de Belgique (F.R.S.-F.N.R.S) project METRIC, under grant No.T.0175.21.
Computational resources have been provided by the Consortium des Équipements de Calcul Intensif (CÉCI), funded by the Fonds de la Recherche Scientifique de Belgique (F.R.S.-FNRS) under Grant No. 2.5020.11 and by the Walloon Region. Also, this work has received funding from the European Union's Horizon 2020 research and innovation programme under Marie Skłodowska-Curie grant agreement No 801505.

ACCEPTED MANUSCRIPT
Physics of Fluids
AIP Publishing
This is the author's peer reviewed, accepted manuscript. However, the online version of record will be different from this version once it has been copyedited and typeset.
PLEASE CITE THIS ARTICLE AS DOI: 10.1063/5.0167110

Accepted to Phys. Fluids 10.1063/5.0167110

DATA AVAILABILITY STATEMENT

The data that support the findings of this study are openly available in Time-lag-AE-Chemistry-Analysis-Reduction at https://sandbox.zenodo.org/record/1238601 reference number [1238601].

[1] D. Veynante and L. Vervisch, Progress in Energy and Combustion Science **28**, 193 (2002).

[2] A. Avdić, G. Kuenne, F. di Mare, and J. Janicka, Combustion and Flame **175**, 201 (2017), special Issue in Honor of Norbert Peters.

[3] S. B. Pope, Combustion Theory and Modelling **1**, 41 (1997), publisher: Taylor & Francis _eprint: https://doi.org/10.1080/713665229.

[4] M. Singer, S. Pope, and H. Najm, Combustion and Flame - COMBUST FLAME **147**, 150 (2006).

[5] P. Boivin, C. Jiménez, A. L. Sánchez, and F. A. Williams, Proceedings of the Combustion Institute **33**, 517 (2011).

[6] P. Pepiot and H. Pitsch, 4th Jt Meet US Sect Combust Inst (2005).

[7] F. Contino, H. Jeanmart, T. Lucchini, and G. D'Errico, Proceedings of the Combustion Institute **33**, 3057 (2011).

[8] M. Ihme, W. T. Chung, and A. A. Mishra, Progress in Energy and Combustion Science **91**, 101010 (2022).

[9] A. Parente, J. Sutherland, L. Tognotti, and P. Smith, Proceedings of the Combustion Institute **32**, 1579 (2009).

[10] O. Owoyele and P. Pal, Energy and AI **7** (2021), 10.1016/j.egyai.2021.100118.

[11] R. Malpica Galassi, P. P. Ciottoli, M. Valorani, and H. G. Im, Journal of Computational Physics **451**, 110875 (2022).

[12] J. C. Sutherland and A. Parente, Proceedings of the Combustion Institute **32**, 1563 (2009).

[13] G. D'Alessio, A. Cuoci, G. Aversano, M. Bracconi, A. Stagni, and A. Parente, Energies **13**, 2567 (2020).

[14] P. J. Schmid, Annual Review of Fluid Mechanics **54**, 225 (2022), https://doi.org/10.1146/annurev-fluid-030121-015835.

[15] S. Le Clainche and J. M. Vega, Complexity **2018**, 6920783 (2018).

ACCEPTED MANUSCRIPT
Physics of Fluids
AIP Publishing
This is the author's peer reviewed, accepted manuscript. However, the online version of record will be different from this version once it has been copyedited and typeset.
PLEASE CITE THIS ARTICLE AS DOI: 10.1063/5.0167110

[16] A. Corrochano, G. D'Alessio, A. Parente, and S. L. Clainche, "Hierarchical higher-order dynamic mode decomposition for clustering and feature selection," (2023), arXiv:2301.07976 [physics.flu-dyn].

[17] A. Corrochano, G. D'Alessio, A. Parente, and S. Le Clainche, International Journal of Mechanical Sciences **249**, 108219 (2023).

[18] W. Ji, W. Qiu, Z. Shi, S. Pan, and S. Deng, The Journal of Physical Chemistry A **125**, 8098 (2021), pMID: 34463510, https://doi.org/10.1021/acs.jpca.1c05102.

[19] A. Sharma, R. Johnson, D. Kessler, and A. Moses (2020).

[20] A. Hetherington, A. Corrochano, R. Abadía-Heredia, E. Lazpita, E. Muñoz, P. Díaz, E. Moira, M. López-Martín, and S. L. Clainche, "Modelflows-app: data-driven post-processing and reduced order modelling tools," (2023), arXiv:2305.17150 [cs.CE].

[21] T. Zhang, Y. Yi, Y. Xu, Z. X. Chen, Y. Zhang, W. E, and Z.-Q. J. Xu, Combustion and Flame **245**, 112319 (2022).

[22] G. Hinton and R. Salakhutdinov, Science (New York, N.Y.) **313**, 504 (2006).

[23] P. Zhang and R. Sankaran, Journal of Machine Learning for Modeling and Computing **3**, 1 (2022).

[24] C. Wehmeyer and F. Noé, The Journal of Chemical Physics **148** (2017), 10.1063/1.5011399.

[25] K. Champion, B. Lusch, J. N. Kutz, and S. L. Brunton, Proceedings of the National Academy of Sciences **116**, 22445 (2019).

[26] K. P. F.R.S, The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science **2**, 559 (1901), publisher: Taylor & Francis _eprint: https://doi.org/10.1080/14786440109462720.

[27] S. Wold, K. Esbensen, and P. Geladi, Chemometrics and Intelligent Laboratory Systems **2**, 37 (1987).

[28] L. van der Maaten, E. Postma, and H. Herik, Journal of Machine Learning Research - JMLR **10** (2007).

[29] A. Parente and J. Sutherland, Combustion and Flame **160**, 340 (2013).

[30] D. G. Goodwin, R. L. Speth, H. K. Moffat, and B. W. Weber, "Cantera: An Object-oriented Software Toolkit for Chemical Kinetics, Thermodynamics, and Transport Processes," (2021).

[31] E.-M. Wartha, M. Bösenhofer, and M. Harasek, Combustion Science and Technology **193**, 2807 (2021), https://doi.org/10.1080/00102202.2020.1760257.

[32] P. Saxena and F. A. Williams, Combustion and Flame **145**, 316 (2006).

[33] M. Cavazzuti, "Design of experiments," in *Optimization Methods: From Theory to Design Scientific and Technological Aspects in Mechanics* (Springer Berlin Heidelberg, Berlin, Heidelberg, 2013) pp.

ACCEPTED MANUSCRIPT
Physics of Fluids
AIP Publishing
This is the author's peer reviewed, accepted manuscript. However, the online version of record will be different from this version once it has been copyedited and typeset.
PLEASE CITE THIS ARTICLE AS DOI: 10.1063/5.0167110

13–42.

[34] M. Baudin, (2015).

[35] Martín Abadi, Ashish Agarwal, Paul Barham, Eugene Brevdo, Zhifeng Chen, Craig Citro, Greg S. Corrado, Andy Davis, Jeffrey Dean, Matthieu Devin, Sanjay Ghemawat, Ian Goodfellow, Andrew Harp, Geoffrey Irving, Michael Isard, Y. Jia, Rafal Jozefowicz, Lukasz Kaiser, Manjunath Kudlur, Josh Levenberg, Dandelion Mané, Rajat Monga, Sherry Moore, Derek Murray, Chris Olah, Mike Schuster, Jonathon Shlens, Benoit Steiner, Ilya Sutskever, Kunal Talwar, Paul Tucker, Vincent Vanhoucke, Vijay Vasudevan, Fernanda Viégas, Oriol Vinyals, Pete Warden, Martin Wattenberg, Martin Wicke, Yuan Yu, and Xiaoqiang Zheng, "TensorFlow: Large-Scale Machine Learning on Heterogeneous Systems," (2015).

[36] T. O'Malley, E. Bursztein, J. Long, F. Chollet, H. Jin, L. Invernizzi, and others, "KerasTuner," (2019).

[37] X. Glorot and Y. Bengio, in *AISTATS* (2010).

[38] D. Kingma and J. Ba, International Conference on Learning Representations (2014).

[39] C. Croux and C. Dehon, Tilburg University, Center for Economic Research, Discussion Paper **19** (2010), 10.1007/s10260-010-0142-z.

[40] P. Virtanen, R. Gommers, T. E. Oliphant, M. Haberland, T. Reddy, D. Cournapeau, E. Burovski, P. Peterson, W. Weckesser, J. Bright, S. J. van der Walt, M. Brett, J. Wilson, K. J. Millman, N. Mayorov, A. R. J. Nelson, E. Jones, R. Kern, E. Larson, C. J. Carey, Polat, Y. Feng, E. W. Moore, J. VanderPlas, D. Laxalde, J. Perktold, R. Cimrman, I. Henriksen, E. A. Quintero, C. R. Harris, A. M. Archibald, A. H. Ribeiro, F. Pedregosa, P. van Mulbregt, and SciPy 1.0 Contributors, Nature Methods **17**, 261 (2020).

[41] R. B. Darlington and A. F. Hayes, New York, NY: Guilford , 603 (2017).

[42] R. P. W. Duin and E. Pękalska, "Object representation, sample size, and data set complexity," in *Data Complexity in Pattern Recognition*, edited by M. Basu and T. K. Ho (Springer London, London, 2006) pp. 25–58.

[43] C. M. Bishop, *Pattern Recognition and Machine Learning (Information Science and Statistics)* (Springer-Verlag, Berlin, Heidelberg, 2006).

[44] P. BIECEK, *Explanatory model analysis: Explore, explain, and examine predictive models* (CHAPMAN amp; HALL CRC, 2022).

[45] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and

ACCEPTED MANUSCRIPT
Physics of Fluids
AIP Publishing
This is the author's peer reviewed, accepted manuscript. However, the online version of record will be different from this version once it has been copyedited and typeset.
PLEASE CITE THIS ARTICLE AS DOI: 10.1063/5.0167110

E. Duchesnay, Journal of Machine Learning Research **12**, 2825 (2011).

[46] M. Blocquet, C. Schoemaecker, D. Amedro, O. Herbinet, F. Battin-Leclerc, and C. Fittschen, Proc Natl Acad Sci U S A **110**, 20014 (2013).

[47] C. Sammut and G. I. Webb, eds., "Mean absolute error," in *Encyclopedia of Machine Learning* (Springer US, Boston, MA, 2010) pp. 652–652.

[48] I. Jolliffe, *Principal component analysis* (Springer Verlag, New York, 2002).

[49] D. M. Hawkins, "Introduction," in *Identification of Outliers* (Springer Netherlands, Dordrecht, 1980) pp. 1–12.