TIME-LAG AUTO-ENCODERS FOR CHEMISTRY DIMENSIONALITY REDUCTION

L. Castellanos*, R. S. M. Freitas, A. Parente**, F. Contino***
luisa.castellanos@uclouvain.be
*IMMC-TFL, Universite Catholique de Louvain, Louvain La Neuve, Belgium
**Aerothermo-mechanics Department, Universite Libre de Bruxelles, Brussels, Belgium

Abstract.
Chemical kinetics modeling is a challenge. One of the challenges is related to the curse of dimensionality. Many reduction techniques have been applied, searching for alternatives to accelerate reacting flow simulations. Lately, Machine Learning techniques have been applied to this task. The present study assesses the applicability of Time-lag Auto-Encoders for chemistry reduction.

1 Introduction

Many industrial applications include reacting flows as a key issue to model and foresee. This means that efficient modeling techniques are a must; nowadays, the most common tools are Computational Fluid Dynamics (CFD) simulations, which unfortunately, hold a high computational overhead in proportion to the chemical mechanism considered and the number of physical phenomena to be modeled [9]. Since the principal issue regarding chemical kinetics is the curse of dimensionality, many dimensionality reduction techniques have been applied. For this study, there is a particular interest in Machine Learning (ML) techniques, from which, the application of PCA [6], and auto-encoders (AE) [12] are worth noticing. However, both come with advantages and disadvantages; in the case of PCA, even if it is an easy-to-apply technique, it just allows the representation of linear phenomena, failing to capture many of the chemistry kinetics non-linearities. On the other hand, AE offers a better representation of these non-linearities, however, as in the case of PCA, the reduced components are not directly related to chemical variables, making it difficult to develop closure models. In the present work, a time shift will be added to an AE network architecture, which gives place to a time-lag auto-encoder (TAE) [11]. The application of such a network adds a temporal characterization to the reduced components, which becomes a dynamical characterization of the chemistry. The aim of the present study is to assess the feasibility of such an approach for chemistry reduction.

2 Theoretical Background

A TAE follows the same concept and architecture as an Auto-encoders (AE), which is a type of neural network that is used for finding reduced representations of a given input vector [5]. The main difference is that a time shift is applied between the network's inputs and outputs, which means modeling a dynamical system as a time series starting from a thermochemical initial state. Moreover, the final goal of TAE

is to find an encoding and decoding which minimizes the time-lagged reconstruction loss (L_{TAE}); for having a physically constrained model, a physically aware loss function is proposed, which reflects the mass conservation principle:

$$L_{TAE} = \frac{1}{M}\sum_{t=0}^{t=t_{nt}} |X_{t+\Delta t} - \tilde{X}_{t+\Delta t}|^2 + \frac{\beta}{M}\sum_{t=0}^{t=t_{nt}} |Y_{t+\Delta t} - \tilde{Y}_{t+\Delta t}|^2 \quad (1)$$

In which variables with (˜) refer to the TAE's output, while X stands for the correct output state vector, and Y stands for the species mass fractions, which should sum up to a constant value. β is the regularization constant and M stands for the number of time samples in all the ignition cases. However, a reduction technique is evaluated in accordance with their respective reconstruction error, therefore, it is important to understand the reconstruction benefits of TAE regarding typical reconstruction techniques, such as PCA. Therefore, the question to answer is how much the reconstruction inaccuracies are 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 about the thermochemical state evolution, while PCA techniques just provide information about a present state. Thus, it is intended to analyze both techniques under equality of conditions, a one-layer auto-encoder is proposed to mimic the PCA technique, called a PCA-like auto-encoder.

3 Experimental Layout

In the present work, we test the capability of TAE as a feature extraction technique for hydrogen combustion. The dataset for training the machine learning models is obtained with isobaric batch reactors simulations, developed with the Cantera software [4]. The reduced version of the University of San Diego chemistry mechanism for hydrogen combustion is used [8], which consists of 9 species and 21 reactions. For the different ignition conditions, different values of equivalence ratio (ϕ) and initial temperature are considered. A Latin Hypercube Sampling was implemented [1] with a uniform distribution; 100 samples are obtained in a region delimited by $\phi \in [0.9, 1.2]$ and $T \in [1100, 1200]$ Kelvin. 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 ignition points. 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, so it is possible to discuss the number of components for accurate reconstruction. For the chemical carrier's identification, a correlation analysis is performed using Kendall's Tau B correlation index, due to its boundedness and resistance to outliers [2].

4 Results

The first item to assess is the quality of reconstruction achieved by the TAE networks while feeding with the training data. Random ignition cases are used, and the quality of reconstruction is assessed via R^2 scores [3] 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. The ignition cases ($\phi = 0.92$, $T_0 = 1140.35K$), and ($\phi = 1.09$, $T_0 = 1120K$) were sampled; both cases proportion a minimum R^2 score of 0.95, result compatible with a satisfactory quality of reconstruction. Once the quality of reconstruction is assessed in training data, it is possible to extrapolate to unseen conditions, which means ignition points that were not used during training. Three interpolation ignition cases will be studied; interpolation means an ignition point not considered in the training data but contained inside the LHS sampling limits. These ignition points are available in Table 1.

Table 1. Interpolation ignition cases

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

4.1 TAE 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 $[t_0, t_{nt-1}]$, where nt 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 $[t_{0+\Delta t}, t_{nt}]$. The R^2 score will be used again to assess the quality of reconstruction of the interpolation cases described in Table 1. The R^2 score values are available in Figure 1.

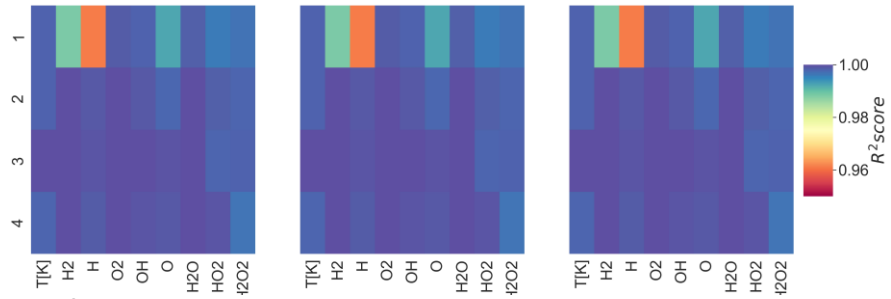


Figure 1. R^2 scores for interpolation cases, y-axe describes the number of latent variables considered by the model, at left: ignition case ($\phi = 0.93$, $T_0 = 1160K$), center: ignition case ($\phi = 1.0$, $T_0 = 1200K$), right: ignition case ($\phi = 1.10$, $T_0 =$

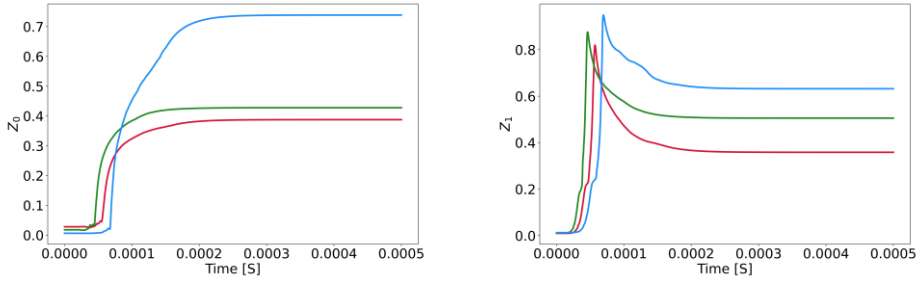
1130K). The numbers in the y-axis refer to the latent space dimension.

After observation, it is visible that the models retain a good representation of the dynamic behavior, returning R^2 scores larger than 0.96, while the lower scores belong to the latent space z = 1. Such output could be associated with combustion's high non-linearities, meaning that larger representations are required for an accurate description of the thermochemical states. In the previous section, the application of different bottleneck sizes is mentioned, which will be used to observe the chemical carriers that might appear in each case, such association is obtained using Kendall Tau B correlation index. The resulting associated species are available in Table 2.

Table 2. Chemical carriers' identification

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

It is important to mention that the same chemical carriers repeat for all the ignition cases, accordingly to the bottleneck size. The temperature can be considered a key variable since it appears in all the models. The temperature repetitiveness in the manifold z = 4 suggests that four variables lead to overparametrization of the states. Figure 2 shows the latent space behaviour of z = 2 for the three ignition cases under study. It is to notice the same behaviour of the curves in all cases, with magnitudes differences associated to the ignition case, further, the time shift from the Ignition Delay Time is observed. The chemical carrier's repetitiveness suggests that a chemical mechanism variance can be described by key thermochemical variables.



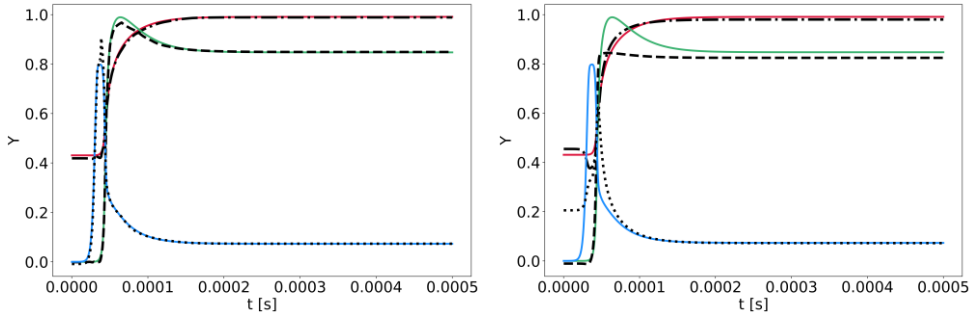
(a) (b)

Figure 2. Latent space visualization for manifold size z = 2, (a) shows the first latent variable, and (b) the second latent variable. The red curve stands for the ignition case (ϕ = 0.93, T_0 = 1160K), green curve for ignition case (ϕ = 1.0, T_0 =

1200K), and blue curve for ignition case (ϕ = 1.10, T_0 = 1130K)

4.2 PCA comparison

Here, the comparison will be performed based on the reconstruction error. For avoiding circumstances that could benefit one model over another, the PCA Autoencoder is trained with the same normalization and datasets used for the TAE training, and evaluations consider the same ignition cases. The scoring technique will be changed by the Mean Absolute Error (MAE) [7], so it is possible to estimate the average deviation between the predicted values and the real ones. The comparison highlights a significant difference between the numerical inaccuracies since the PCA reconstruction error (range [0, 0.04]) is easily four times higher than the one presented in TAE reconstruction (range [0, 0.01]). This suggests that TAE presents better reconstruction capabilities. However, the representation of fast phenomena must be assessed too. However, PCA does not manage to reproduce such events, issue that does not happen with TAE. The normalized reconstruction curves for T, OH, and HO2 are presented in Figure 3 for a latent space z = 2. Such selection of species searches to portray the behaviour in major, intermediate, and minor thermochemical variables.



(a) (b)

Figure 3. Normalized chemical species reconstruction for a latent space size z = 2, the red curve shows the simulated behavior or temperatures, the green curve shows OH behavior, and the blue curve shows the behavior of HO2 for an ignition case ϕ = 1.0, T0 = 1200K (a) shows the TAE reconstruction (b) shows the PCA reconstruction.

5 Conclusion

In this study, an exploration of TAE capabilities for chemistry reduction was presented. It is assessed how the resultant latent variables are a more efficient representation of the thermochemical variables, allowing at the same time, a temporal characterization. It is also important to remark that TAE's manifolds and

their associated species (chemical carriers) could be used for the development of surrogate models that promise greater interpolation and extrapolation characteristics in the low data limit. Future work should be aimed at the development of better strategies for the description of manifold dynamics, as well as testing possible applications for the identified chemical carriers.

References

[1] Cavazzuti, M., *Design of Experiments*, Springer Berlin Heidelberg, Berlin, Heidelberg, 2013, pp. 13-42.

[2] Croux, C., Dehon, C., "Influence functions of the spearman and kendall correlation measures", Statistical Methods & Applications 19:497-515 (2010).

[3] Darlington, R.B., Hayes, A.F., *Regression analysis and linear models*, New York,NY: Guilford, 2017, pp. 603–611.

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

[5] Hinton, G., Salakhutdinov, R., "Reducing the dimensionality of data with neural networks", *Science* (New York, N.Y.) **313**:504–507 (2006).

[6] Parente, A., Sutherland, J., Tognotti, L., Smith, P., "Identification of low-dimensional manifolds in turbulent flames", *Proceedings of the Combustion Institute* **32**:1579–1586 (2009).

[7] Sammut, C., Webb, G.I. (eds.), *Mean Absolute Error*, Springer US, Boston, MA, 2010, pp. 652–652.

[8] Saxena, P., Williams, F.A., "Testing a small detailed chemical-kinetic mechanism for the combustion of hydrogen and carbon monoxide", *Combustion and Flame* **145**:316–323 (2006).

[9] Veynante, D., Vervisch, L., "Turbulent combustion modelling", *Progress in Energy and Combustion Science* **28**:193–266 (03 2002).

[10] Wartha, E.M., Bosenhofer, M., Harasek, M., "Characteristic chemical time scales for reactive flow modeling", *Combustion Science and Technology* **193**(16), 2807– 2832 (2021).

[11] Wehmeyer, C., Noé F., "Time-lagged autoencoders: Deep learning of slow collective variables for molecular kinetics", *The Journal of Chemical Physics* **148** (Oct 2017).

[12] Zhang, P., Sankaran, R., "Autoencoder neural network for chemically reacting systems", *Journal of Machine Learning for Modeling and Computing* **3**:1–28 (2022).