

Sparse Surrogate Modeling for Combustion Chemistry via Time-Lag Autoencoders and Gradient-Based Clustering

Luisa Castellanos^{a,b,c}, Rodolfo S. M. Freitas^d, Riccardo Malpica Galassi^e, Xi
Jiang^d, Alessandro Parente^{b,c,f}

^a*Institute of Mechanics, Materials, and Civil Engineering, Université Catholique de
Louvain, Place du Levant 2, Ottignies-Louvain-la-Neuve, 1348, Belgium*

^b*Aero-Thermo-Mechanics Department, Université Libre de Bruxelles, Avenue F. D. Roosevelt
50, Brussels, 1050, Belgium*

^c*Brussels Institute for Thermal-Fluid Systems and Clean Energy (BRITE), Université Libre de
Bruxelles, Avenue F. D. Roosevelt 50, Brussels, 1050, Belgium*

^d*School of Engineering and Materials Science, Queen Mary University London, Mile End
Road, London, E1 4NS, United Kingdom*

^e*Dipartimento di Ingegneria Meccanica e Aerospaziale, Università di Roma La Sapienza, Via
Eudossiana 18, Rome, 00184, Italy*

^f*WEL Research Institute, Avenue Pasteur 6, Wavre, 1300, Belgium*

Abstract

This work presents a data-driven framework for efficient and interpretable surrogate modeling of combustion chemistry, combining time-lag autoencoders (TAEs) with gradient-based clustering. The methodology targets applications where high-fidelity simulations are computationally prohibitive, enabling accurate prediction of thermochemical states with minimal input variables. Ignition trajectories are first encoded into low-dimensional, temporally informed latent spaces, from which physically meaningful chemical carriers are identified via statistical correlation. Cases are then grouped by progress-variable gradient similarity, and cluster-specific regression models reconstruct the full thermochemical state from the carriers and equivalence ratio. The proposed framework is demonstrated on methane/air combustion with GRI-Mech 2.11 over a broad range of equivalence ratios. The approach achieves a coefficient of determination greater than 0.9 in most cases and maintains robustness in interpolation and sparse-data scenarios. By integrating dynamical feature learning with gradient-informed clustering, the approach produces generalizable, interpretable, and computationally efficient surrogates for complex reaction systems. This facilitates their incorporation into

computational fluid dynamics (CFD) solvers to enable accelerated reactor simulations and optimize chemical engineering designs.

Keywords:

Chemical Kinetics, AI for science, Reduced manifold, Machine learning explainability, Data-driven modeling

Nomenclature

Acronyms

AE	Auto-encoder
CFD	Computational Fluid Dynamics
CNN	Convolutional Neural Network
DeepONet	Deep Operator Network
Gappy-AE	Gappy-Autoencoder
Gappy-POD	Gappy-Proper Orthogonal Decomposition
GPR	Gaussian Process Regressor
IDT	Ignition Delay Time
LPCA	Local-Principal Component Analysis
MAE	Mean Absolute Error
ML	Machine Learning
MSE	Mean Squared Error
NN	Neural Network
ODE	Ordinary Differential Equation
PCA	Principal Component Analysis
PFI	Permutation Feature Importance
POD	Proper Orthogonal Decomposition
PV	Progress Variable
RL	Reinforcement Learning
ROM	Reduced Order Model
TAE	Time-lag Auto-encoder

Parameters and variables

$\Delta\phi$	Equivalence ratio sampling variation
Δt	Time step
$\dot{\omega}_i$	Chemical source term of i-th species
λ	Regression regularization constant
$\mathcal{D}$	Decoder
$\mathcal{D}_i$	Arbitrary function that relates the set of chemical carriers Φ_t with the thermochemical state $X_{t+\Delta t}$
$\mathcal{E}$	Encoder
$\mathcal{F}$	Function describing Y_i behavior
$\mathcal{M}$	Mapping that relates a thermochemical state to a progress variable value
$\mathcal{M}^{-1}$	Inverse mapping that relates the progress variable value to a thermochemical state
$\mathcal{P}$	Function that relates the thermochemical state at the following time step and the previous chemical carriers values
$\mathcal{R}$	Batch reactor
Φ	Thermochemical state expressed using chemical carriers

ϕ	Equivalence ratio
ϕ_0	Equivalence ratio of mixture
$\Phi_{(t,n)}$	n-th ignition trajectory set of chemical carriers values at time t
τ_{ch}	Characteristic time scale
Θ	Temperature based progress variable
$\tilde{X}_{t+\Delta t}$	Thermochemical state at time $t+\Delta t$ predicted by TAE network
$\tilde{Y}_i$	TAE predicted mass fraction for the i-th chemical species
i_j	Permutation importance index for j-th input feature
K_i	i-th cluster
K_{i,n_k}	i-th cluster in n_k clustering strategy
m	$\mathcal{P}$ hypothesis polynomial degree
n	Total assemble of carriers by each cluster
n_0	First time step of a series
N_r	Number of reaction
n_t	Last time step of a series
n_z	Latent space size
n_k	Number of clusters
n_{traj}	Number of trajectories in the i-th cluster
p	Pressure
Q	Number of times the j-th species will be scrambled
R^2	Coefficient of determination
s	R^2 scores for the unscrambled hypothesis
$s_{q,j}$	R^2 scores for the j-th scrambled hypothesis, in q scrambling repetitions
T	Temperature at present time
T_0	Initial temperature
T_b	Temperature of fully burned mixture
T_u	Temperature of unburned mixture
X_t	Thermochemical state at time t
$X_{t+\Delta t}$	Thermochemical state at time $t+\Delta t$ predicted in the batch reactor $\mathcal{R}$
Y_i	i-th species mass fraction
Z	Latent space
Z_n	n-th variable in reduced manifold
X	Thermochemical state
M	Number of training samples
t	Time
Y	Fuel normalized mass fraction

1. Introduction

Modeling chemically reacting flows continues to represent a fundamental challenge in the field of computational combustion owing to the significant coupling between fluid dynamics and stiff, nonlinear chemical kinetics. To accurately describe complex combustion processes, the chemical mechanisms increase in size and complexity. Consequently, the computational cost of solving the associated ordinary differential equations (ODEs) rises substantially, often making numerical simulations impractical for real-world engineering applications [1, 2]. This computational barrier limits their direct application in practical reactor design, optimization, and scale-up, where prompt and accurate predictions are essential.

To address this, simplified modelling has been implemented, however, it comes with its own challenges [3, 4]; among such simplified models, reduced-order modeling (ROM) techniques have been developed to approximate the evolution of the thermochemical state using a lower-dimensional representation, which is a topic of interest for combustion modelling [5, 6].

A common approach involves projecting the entire chemical state onto a progress variable (PV), a scalar measure that indicates the extent of the chemical reaction processes [1]. While efficacious in numerous applications, PV-based models frequently require empirical definitions [5] and closure relations [7], which consequently restrict their generalizability and interpretability across diverse chemical regimes [8].

To overcome the inherent limitations of traditional PV methodologies while maintaining their advantageous aspects, recent efforts have progressively transi-

tioned towards the utilization of machine learning (ML) [9]. Indeed, ML has created new opportunities for data-driven reduction of combustion models without relying on empirically defined progress variables. In particular, ML models have been employed to develop ROMs that either interpolate chemical source terms or reconstruct complete thermochemical fields [10, 11]. Techniques such as neural networks (NN) [12, 13, 14, 15], autoencoders (AEs) [16, 17], and Gaussian process regression (GPR) [18] have facilitated the discovery of low-dimensional manifolds, improved predictive accuracy, and enhanced computational efficiency. Among methods for dimensionality reduction, principal component analysis (PCA) has played a crucial role by identifying low-dimensional manifolds that preserve the dominant variance in thermochemical data [19]. Although PCA provides interpretable linear projections, it frequently necessitates closure models [20, 21] and may face challenges when dealing with nonlinearities. To overcome these limitations, more advanced methodologies, including Local PCA (LPCA) [22, 23] and convolutional neural networks (CNNs) [24, 25], have been explored, enabling better performance in capturing complex combustion behavior.

Nevertheless, many of these models face significant limitations: they require extensive, densely sampled datasets; their latent representations are often physically incomprehensible; and they struggle to extrapolate to unseen conditions due to a lack of dynamical information. To address these challenges, this work introduces a new surrogate modeling framework that merges time-lag autoencoders (TAEs) with gradient-informed clustering of ignition trajectories. In the TAE method [26], the input and output of the autoencoder (AE) are thermochem-

1 ical state vectors at successive time steps, enabling the network to learn temporal
2 dynamics directly. The fundamental concept behind this work involves deriving
3 interpretable, low-dimensional chemical representations—referred to as chemical
4 carriers—from the TAE latent space, to be utilized as input features in sparse sur-
5rogate models. At the same time, ignition cases are grouped according to the
6 temporal gradient of the progress variable, which encodes essential dynamical
7 information crucial to the temporal evolution of combustion. Gradient-informed
8 clustering entails categorizing ignition cases based on the behavior of the tempo-
9ral derivatives of thermochemical variables. This approach is inspired by recent
10 research that incorporates function gradient information into model training or
11 clustering methodologies [27].

12 By integrating physically grounded carriers of TAE with cluster-specific re-
13 gression models, we develop models that are interpretable, accurate, and com-
14 putationally efficient. Additionally, we investigate a variant of the autoencoder
15 framework known as the Gappy-AE, wherein distinct TAE models are trained
16 for representative ignition conditions. Furthermore, this hybrid approach accom-
17 plishes three primary objectives:

- 18 1. Dynamic fidelity – TAEs incorporate temporal evolution into the latent space,
19 allowing the surrogate to model kinetic dynamics instead of static correla-
20 tions.
- 21 2. Interpretability – chemical carriers represent identifiable species that control
22 main pathways, enabling physical understanding and model validation.
- 23 3. Data efficiency – gradient-based clustering and sparse inputs allow accurate

1 surrogates to be trained on limited ignition data, reflecting the data con-
2 straints typical in both simulations and experiments.

3 This study concentrates on premixed combustion, where fuel and oxidizer
4 are homogeneously mixed before ignition. The framework is validated using
5 methane/air ignition in an isobaric batch reactor, which is chosen as a repre-
6 sentative test case, providing a controlled environment to evaluate the method-
7 ology and confirm the identification of chemical carriers. The proposed frame-
8 work shows high predictive accuracy ($R^2 \geq 0.9$) in both interpolation and ex-
9 trapolation scenarios, even with a significantly limited input space. Although the
10 results shown here are limited to premixed conditions, the framework is adapt-
11 able beyond this regime. In non-premixed flames, it can be combined with mix-
12 ture-fraction-based methods, enhancing traditional flamelet models by accurately
13 capturing local combustion states and providing better insight into chemical ki-
14 netics in stratified or mixing-limited environments [28, 29]. Furthermore, the pro-
15 posed framework is especially suitable for energy applications with limited data,
16 providing a pathway to real-time, reduced-order models for chemical engineering
17 uses such as combustion system optimization, digital twins, and energy process
18 monitoring.

19 The remainder of this paper is structured as follows. Section 2 presents the the-
20 oretical background, including gradient-based modeling and the TAE framework
21 for chemical kinetics. Section 3 describes the computational setup and numer-
22 ical experiments. Results are discussed in Section 4, and final conclusions are
23 provided in section 5.

1 **2. Theoretical background**

2 *2.1. Gradients in chemical kinetics*

3 Chemical kinetics is a multidimensional, stiff, and highly nonlinear problem.
4 In the absence of effects like convection or diffusion, the temporal evolution of the
5 mass fraction of a chemical species Y_i is governed by a source term that depends
6 on the thermochemical state:

$$\frac{dY_i}{dt} = \mathcal{F}(Y_k, T, p) = \dot{\omega}_i(Y_k, T, p). \quad (1)$$

7 Here, Y_k denotes the mass fractions of all the N_s chemical species, T is the tem-
8 perature, p is pressure, and t is time. The chemical source term $\dot{\omega}_i$ represents the
9 net production rate of species i due to all N_r chemical reactions. This approach is
10 widely used in splitting methods for reactive CFD [30].

11 This equation is written for each species in the mechanism, and together,
12 they describe the evolution of the full thermochemical state, from the initial pre-
13 ignition phase through the ignition and combustion stages, and eventually toward
14 a steady state. In general, the evolution within a batch reactor $\mathcal{R}$ can be interpreted
15 as a trajectory in composition space, connecting unreacted and fully reacted con-
16 ditions.

17 Since their source terms govern species evolution, the gradients (i.e., time
18 derivatives) of the state variables provide insights into the structure of this tra-
19 jectory. Typically, the gradient behavior is characterized by: near-zero values
20 during the induction phase (before ignition), rapid growth during ignition, a peak

1 associated with intense chemical activity, and decay toward zero as the system
2 approaches equilibrium. This trajectory often contains an inflection point, that is,
3 a time instant where the second derivative of a state variable (e.g., temperature)
4 changes sign. The location and characteristics of this inflection point mirror the
5 ignition dynamics and can assist in identifying common features across various
6 ignition cases. Significantly, the gradient behavior is influenced by the role of
7 each species. For instance, reactants typically exhibit a monotonic decrease in
8 mass fraction, while products increase. These consistent trends make the gradi-
9 ent profile a valuable foundation for clustering ignition trajectories and creating
10 gradient-informed reduced models.

11 In addition to analyzing species individually, it is feasible to establish global
12 progress indicators, such as the combustion PV, which encapsulates the overall
13 extent of reaction. Examining the gradient of the progress variable provides a
14 concise method for describing chemical dynamics. The following section offers
15 more details on the definition and role of the progress variable in this context.

16 2.2. *Progress Variables*

17 The progress variable is a widely used concept in combustion modeling. It pro-
18 vides a low-dimensional, non-dimensional representation of combustion progress,
19 simplifying the description of complex chemical kinetics. Many definitions of the
20 PV exist, including both non-dimensional temperature and linear combinations of
21 species mass fractions [31]. Here, as a proof of concept, we consider a basic *PV*
22 definition and its gradient characterization.

1 In general, the progress variable can be viewed as the result of a mapping $\mathcal{M}$
 2 from the thermochemical state $X = \{Y_i, T, p\}$ to a scalar value:

$$\mathcal{M}(X) \rightarrow PV \quad (2)$$

3 That is, any thermochemical state X can be associated with a corresponding value
 4 of the PV . An inverse mapping $\mathcal{M}^{-1}$ may also be defined, projecting a given
 5 value of the PV back to a consistent thermochemical state, provided that sufficient
 6 boundary conditions are known; indeed, the definition of these mappings is a wide
 7 area of research [32, 31, 6]. Although the mapping is not generally one-to-one,
 8 variations in the gradient of the PV often reflect variations in the underlying ther-
 9 mochemical variables. This motivates the use of PV gradient behavior to extract
 10 meaningful features of the combustion process.

11 A common approach uses two non-dimensional variables: a scaled fuel mass
 12 fraction Y (ranging from 1 to 0), and a non-dimensional temperature Θ (ranging
 13 from 0 to 1) [1, 33]. These are typically assumed to be inversely related, such that

$$Y + \Theta = 1 \quad (3)$$

14 This assumption implies that as fuel is consumed, temperature increases, reaching
 15 a maximum when all fuel is depleted.

16 In practice, many studies reduce the progress variable to the non-dimensional
 17 temperature alone [1], defined as:

$$\Theta = \frac{T - T_u}{T_b - T_u} \quad (4)$$

where T is the instantaneous temperature, T_u is the unburned gas temperature, and T_b is the temperature of fully burned products. While this definition is simple, it provides useful information about the system’s chemical state and energy release, and is often sufficient to characterize flame structure and evolution [24].

To compare progress variable trajectories across ignition cases, we employ the ignition delay time (IDT). This value is commonly defined as the time corresponding to the maximum temperature gradient. This point typically coincides with the inflection point of the PV curve and serves as a meaningful geometric descriptor. Figure 1 (a) shows several methane ignition trajectories computed in a 0-D isobaric batch reactor, with varying equivalence ratio ϕ . Panel (b) displays the corresponding progress variable (Θ), revealing substantial similarity across cases: each trajectory resembles a time-shifted version of a shared curve. In panel (c), the gradients of Θ are plotted, highlighting consistent patterns on both sides of the inflection point.

Moreover, IDT values can be predicted accurately using characteristic time scales (τ_{ch}) as input features for ML models [34]. Such a characteristic time scale can be defined as the slowest explosive time scale τ_e in the Jacobian eigenvalues decomposition [35, 36, 37]. This further supports the use of PV gradients as meaningful, low-dimensional indicators of combustion dynamics.

These observations motivate the use of progress variable derivatives for clus-

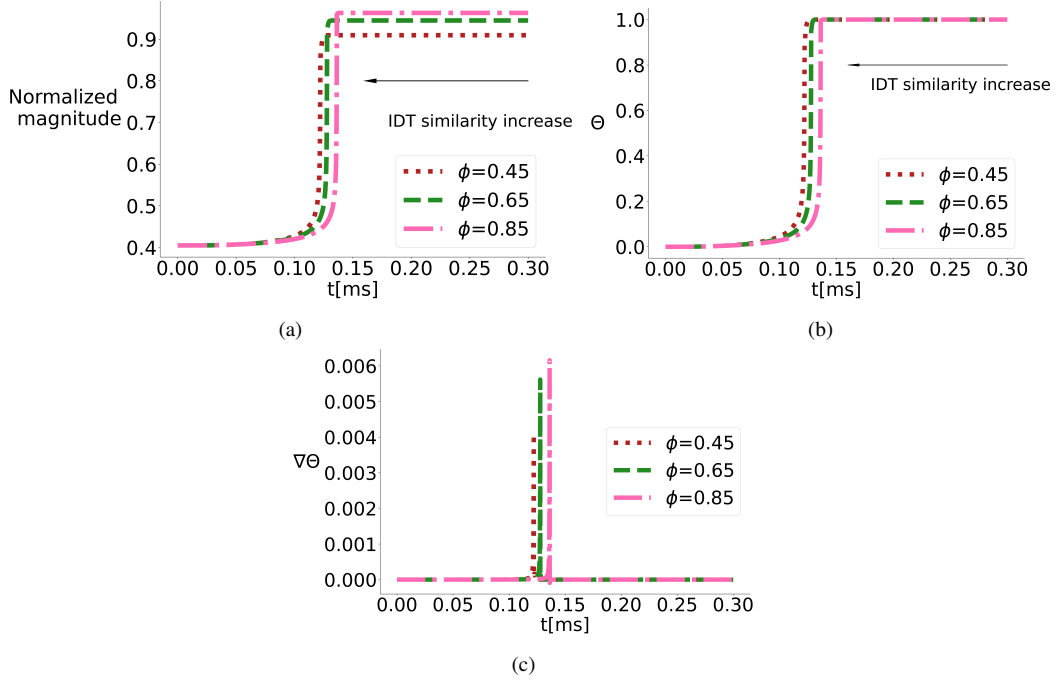


Figure 1: Ignition cases for methane combustion, with initial temperature $T_0 = 1418K$, $p = 4.33atm$, and a range of equivalence ratios. Temperature with a normalized magnitude (division by global maximum temperature value) are presented in panel (a); in panel (b) the respective progress variable curves are presented, resembling a time-shifted version of a shared curve; panel (c) shows the progress variable gradients, exhibiting well-defined trends at both sides of the inflection point.

1 tering ignition cases with similar kinetics. Hence, in this work, we employ IDT
2 values to guide a clustering algorithm that groups ignition trajectories. More
3 specifically, we will identify n_k clusters, each one containing n_{traj} ignition tra-
4 jectories.

5 The following section introduces Time-lag Autoencoders, which are used to
6 identify chemical species that preserve this gradient structure and enable sparse
7 representations of the full thermochemical state.

1 2.3. Time-lag Autoencoder

2 One of the major challenges of modeling chemical kinetics is the large num-
3 ber of chemical species involved, a situation that often calls for the identification
4 of key variables, in order to keep the models manageable. However, the same
5 fact of chemical kinetics multidimensionality complicates the identification of the
6 dominant variables controlling the chemical dynamics. Among the dimension-
7 ality reduction techniques present in Machine Learning (ML) literature, neural
8 networks are of interest for dimensionality reduction, in particular Autoencoders
9 (AE) [38], and specifically, Time-lag Autoencoders (TAE) [26]. Indeed, being
10 non-linear, they allow to account for non-linearities in chemical kinetics. How-
11 ever, Autoencoders can lead to interpretability challenges for what concerns the
12 identified manifold [39]. Conversely, Time-lag Autoencoders are of interest, since
13 they provide physically interpretable dimensionality reduction, shading light on
14 the relevance of specific chemical pathways. Indeed, a Time-lag Autoencoder can
15 be used to identify a small subset of species, termed chemical carriers, that control
16 the dominant kinetic behavior [26]. These carriers are identified using statistical
17 correlations between the latent variables and species mass fractions, exploiting the
18 temporal nature of the TAE since a TAE is mostly an autoencoder network with
19 the difference of having time shifted state vectors as inputs and outputs.

20 An example of a TAE network is shown in Fig. 2. The TAE processes the
21 thermochemical state vector at time t , denoted X_t , and encodes it into a latent
22 space representation Z via an encoder network ($\mathcal{E}$):

$$Z = \mathcal{E}(X_t) \quad (5)$$

1 This reduced representation is then passed to a decoder ($\mathcal{D}$) to reconstruct the
2 thermochemical state vector at the next time step $X_{t+\Delta t}$:

$$\begin{aligned} \tilde{X}_{t+\Delta t} &= \mathcal{D}(Z) \\ \tilde{X}_{t+\Delta t} &= \mathcal{D}(\mathcal{E}(X_t)) \end{aligned} \quad (6)$$

3 The network is trained using a loss function that minimizes the difference
4 between the network’s output $\tilde{X}_{t+\Delta t}$ and the true thermochemical state $X_{t+\Delta t}$
5 from the reactor simulations. A standard Mean Squared Error (MSE) loss function
6 can be used:

$$\mathcal{L}_{MSE} = \frac{1}{M} \sum_0^M \|X_{t+\Delta t} - \tilde{X}_{t+\Delta t}\|^2, \quad (7)$$

7 where M is the number of training samples. Chemical carriers Φ are extracted
8 from the bottleneck layer and correspond to the species most strongly correlated
9 with the learned latent variables Z . Such correlation is measured using correla-
10 tion indexes [26], such as the Kendall Tau B approach [40]. This compares the
11 behavior of the resulting latent variables and the thermochemical variables and
12 provides the best-matching thermochemical variable, allowing the identification
13 of chemical carriers under different conditions. These carriers are supposed to be
14 the species that efficiently parameterize the kinetic behavior of the mechanism,

1 enabling significant dimensionality reduction (e.g., reducing a 31-species mecha-
 2 nism to just 3 chemical carriers, each representative of distinct reactive pathways).

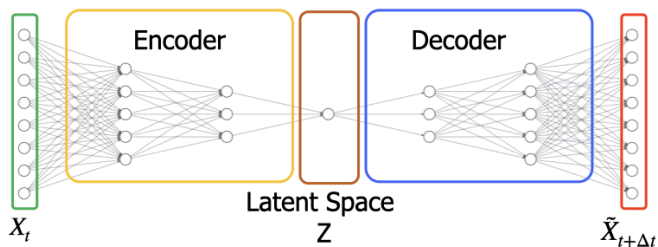


Figure 2: Illustration of a TAE. X_t represents the thermochemical state vector at time t , while $X_{t+\Delta t}$ represents the same thermochemical state vector at time $t + \Delta t$. The latent space layer, i.e., the bottleneck layer, enables the identification of latent variables and, consequently, the chemical carriers. The number of considered chemical carriers is represented by the number of neurons in Z .

3 It is essential to highlight that the latent variables obtained through the TAE
 4 differ fundamentally from those learned by standard autoencoders. While conven-
 5 tional autoencoders aim to compress and reconstruct instantaneous thermochem-
 6 ical states—essentially capturing the main modes of variance in the data—TAEs
 7 are specifically trained to reproduce the temporal evolution of these states. As
 8 such, the latent space of a TAE is optimized to represent the underlying dynamics
 9 of the system, rather than its static configuration [26]. This makes the TAE par-
 10 ticularly well-suited for identifying chemical carriers: the latent variables it learns
 11 serve as a compact representation of the system’s dynamics, encoding features
 12 that govern how the thermochemical state evolves. Consequently, these latent
 13 variables are expected to be more physically meaningful in a dynamical context
 14 than those derived from purely state-based dimensionality reduction techniques.

15 Despite their advantages, it is still true that autoencoder-based models require

1 substantial datasets covering a wide range of conditions to generalize effectively
 2 and provide effective dimensionality reduction. This is especially demanding
 3 for systems with high-dimensional state vectors. To address this, we propose a
 4 strategy inspired by Gappy-POD, where a POD (Proper Orthogonal Decomposi-
 5 tion) or PCA is performed on separate snapshots from different operating condi-
 6 tions, and intermediate states are interpolated from these localized components
 7 or orthonormal bases [41]. By analogy, we introduce the concept of “Gappy-
 8 Autoencoders”. Here, separate TAE networks are trained on a sparse set of ig-
 9 nition cases representative of distinct chemical regimes. Therefore, rather than
 10 a single model trained on a vast parameter space, multiple localized TAEs are
 11 used. Each network produces a tailored set of chemical carriers for its specific
 12 ignition case and that represents a set of chemical reactions, enabling sparse and
 13 interpretable representations across the sampled space.

14 In this setup, the TAE loss function for each ignition trajectory becomes:

$$\mathcal{L}_{TAE} = \frac{1}{n_t - n_0} \sum_{n_0}^{n_t-1} \|X_{t+\Delta t} - \tilde{X}_{t+\Delta t}\|^2, \quad (8)$$

15 where n_0 and n_t denote the initial and final time steps of one trajectory.

16 Each trajectory thus yields a different set of chemical carriers, $\Phi_{t,n}$. When
 17 these are considered in combination, it is possible to construct an explicit decoder
 18 that approximates the full thermochemical state in the entirety of the considered
 19 composition space as a combination of outputs from the different sets of carriers:

$$\tilde{X}_{t+\Delta t} = \mathcal{D}_1(\Phi_{t,1}) + \mathcal{D}_2(\Phi_{t,2}) + \dots + \mathcal{D}_{n_{traj}}(\Phi_{t,n_{traj}}) \quad (9)$$

1 where $\mathcal{D}_i$ are reconstruction/decoding operators associated with each trajectory
 2 (e.g., each trajectory belonging to a cluster).

3 In a more general form, the reduced model that outputs the full thermochem-
 4 ical state $X_{t+\Delta t}$ as a function of the sole chemical carriers Φ_t at time t , anywhere
 5 in the training composition space (e.g. the space visited by one cluster), can be
 6 expressed as:

$$X_{t+\Delta t} = \mathcal{P}(\Phi_t) \quad (10)$$

7 where the function $\mathcal{P}$ is an abstraction of the Gappy-AE formulation.

8 Such a function can be either a decoder network or any simpler function with
 9 sufficient degrees of freedom. Although neural decoders could potentially offer
 10 higher accuracy, they would reduce interpretability, increase the risk of overfit-
 11 ting, and make it more challenging to evaluate the relevance of chemical carriers.
 12 Being one focus of this work on interpretability, as testified by the use of chemi-
 13 cal carriers instead of the raw latent variables, we propose a polynomial decoder
 14 function, which is an intuitive first guess for representing non-linearities while
 15 keeping the number of tunable parameters small, which helps prevent overfitting
 16 in sparse-data regimes [42]. Specifically, the polynomial function predicting the
 17 mass fraction $\tilde{Y}_i$ species i at time $t + \Delta t$ can be expressed as:

$$\tilde{Y}_i(t+\Delta t) = \sum_{j=1}^n a_j \Phi_j + \sum_{j=1}^n \sum_{k=j+1}^n b_{jk} \Phi_j \Phi_k + \sum_{j=1}^n \sum_{k=j+1}^n c_{jk} \Phi_j^2 \Phi_k + \dots + \sum_{k=j+1}^n d_{jk} \Phi_j^m \Phi_k + e \phi_0, \quad (11)$$

1 where Φ_j denotes the concentration of the j -th chemical carrier belonging to
 2 a given cluster at time t (each TAE provides n_Z carriers, therefore $n = n_Z \cdot$
 3 $n_{traj}(K_i)$), a_j are the coefficients for the linear contribution of each chemical car-
 4 rier, b_{jk} are the second-order interaction coefficients accounting for pairwise car-
 5 rier interaction $\Phi_j \Phi_k$, c_{jk} are the third-order interaction coefficients, representing
 6 nonlinear interactions involving squared terms $\Phi_j^2 \Phi_k$, and ϕ_0 is the equivalence
 7 ratio of the mixture, which improves generalization, with e being its correspond-
 8 ing coefficient. Therefore, one function is calibrated for each chemical species $\tilde{Y}_i$
 9 for each cluster, obtaining the coefficients via regression on training data. This
 10 function holds for any time t . Additionally, it is crucial to note that separating the
 11 nonlinear encoding (via the TAE) from a simpler, regularised linear-polynomial
 12 decoding facilitates the identification of error sources and their assignment either
 13 to the latent-space representation or to the regression model.

14 Moreover, because TAEs produce physically interpretable chemical carriers,
 15 interpolation between ignition trajectories becomes feasible and meaningful. These
 16 carriers reflect key aspects of the underlying reaction kinetics, and their combina-
 17 tion enables the accurate reconstruction of thermochemical states for intermediate
 18 conditions — especially in sparse data regimes.

1 2.4. *Methodology Summary*

2 The proposed methodology involves the following steps: First, chemical car-
3 riers are identified for a set of ignition trajectories spanning different chemical
4 regimes. These trajectories are then clustered according to their IDT values,
5 grouping cases with similar temporal characteristics. For each cluster, a regres-
6 sion model is trained using the corresponding chemical carriers as input features,
7 with the target being the thermochemical state at the subsequent time step. This
8 enables each model to capture regime-specific dynamics. A visual summary of
9 the methodology is presented in Figure 3

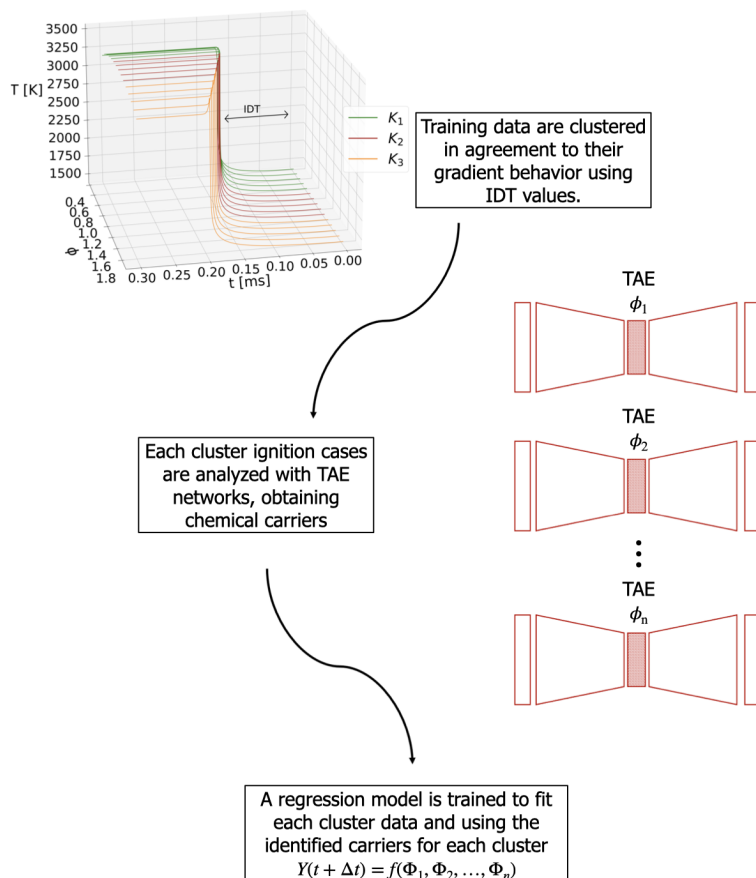


Figure 3: Overview of the proposed methodology for representation learning.

3. Computational setup

The proposed computational experiment is designed to assess the effectiveness of gradient-informed clustering for reduced-order modeling of chemical kinetics. Specifically, we test whether grouping ignition trajectories based on the behavior of their temporal gradients leads to better predictive models for thermochemical evolution.

To this end, we construct a dataset of ignition trajectories spanning a range

1 of ignition delay times, which correspond to different chemical regimes. The
2 selected cases are designed to: 1) span a broad range of IDT values, and 2) reflect
3 diverse dominant chemical pathways.

4 Since the goal is to cluster trajectories based on the evolution of a PV, and
5 because PV values are typically bounded in $[0, 1]$ regardless of the initial condi-
6 tions, we fix the initial temperature across all cases and vary only the equivalence
7 ratio ϕ . This approach yields ignition scenarios across lean and rich conditions,
8 facilitating the diversity of kinetic regimes. Methane combustion is selected as a
9 representative case, offering greater chemical complexity compared to hydrogen,
10 thereby enabling a broader assessment of the method’s applicability.

11 *3.1. Computational simulations*

12 We simulate homogeneous methane/air ignition trajectories for isobaric and
13 adiabatic batch reactors using the GRI-mech 2.11 mechanism [43], which includes
14 30 reactive species and 279 reactions. The simulations are performed with the
15 Cantera software [44], which implements a CVODE solver [45]; the pressure and
16 initial temperature are set to $p = 4.33$ atm and $T_0 = 1418$ K respectively. The mix-
17 ture composition is varied across equivalence ratios in the range $\phi \in [0.45, 1.85]$
18 in increments of $\Delta\phi = 0.1$, resulting in 15 distinct ignition trajectories. This sam-
19 pling strategy ensures coverage of a sufficient range of chemical regimes while
20 enabling sensitivity analysis with respect to the number and nature of the result-
21 ing clusters. Notably, some ignition trajectories may exhibit similar IDT values
22 despite originating from different equivalence ratios — a scenario of particular in-

1 terest for assessing the method’s ability to identify underlying kinetic similarities.

2 Simulations are carried out with a time step of $\Delta t = 2.5$ ns, which provides
3 sufficient temporal resolution to resolve the dynamics of fast-reacting species such
4 as CH_3 .

5 3.2. *Clustering setup*

6 The ignition cases are clustered according to their IDT values using the K-
7 means algorithm implemented via the Scikit-learn library [46]. To explore the sen-
8 sitivity to the number of clusters, values of $n_k = 2$, $n_k = 3$, and $n_k = 4$ are tested.
9 Larger values of n_k are not considered, as starting from $n_k = 5$, some ignition
10 cases are assigned to individual clusters, reducing the representativeness and gen-
11 erality of the clustering. Cluster quality assessment is done through the Davies-
12 Bouldin [47] and Calinski-Harabasz [48] scoring; with the Davies-Bouldin score
13 measuring the ratio between in-cluster distances and the separation between clus-
14 ters’ centroids, while the Calinski-Harabasz score measures the ratio between the
15 cluster’s centroid dispersion and each cluster’s compactness. Therefore, lower
16 Davies-Bouldin scores and higher Calinski-Harabasz scores both indicate better
17 clustering.

18 3.3. *Time-lag Autoencoder training parameters*

19 A separate TAE is trained for each ignition trajectory using the Python Ten-
20 sorflow package [49]. Each network receives as input the thermochemical state
21 vector at time t , consisting of temperature and species mass fractions, and predicts

1 the state at $t + \Delta t$. The training objective is to minimize a Mean Squared Error
2 (MSE) loss function between predicted and true states, as defined in Equation 8.

3 All input variables are magnitude-normalized by dividing each component by
4 a reference value across the dataset. This normalization preserves the exponential
5 trends characteristic of chemical species and ensures numerical stability during
6 training, as demonstrated previously [26].

7 The network architecture is selected using a combination of grid and random
8 search techniques, implemented via the Keras Tuner package [50], with hyperpa-
9 rameters tuned on an intermediate condition ($\phi = 1.25$). Cross-validation is then
10 performed for all other cases using a K-fold strategy, with an 80%-20% train-
11 validation split implemented via `Scikit-Learn` [46].

12 Each TAE uses Leaky ReLU activations with a negative slope of 1×10^{-3} ,
13 except in the bottleneck layer, which applies a sigmoid activation to constrain the
14 latent variables to the interval $[0, 1]$. The latent space dimension is fixed to $n_Z =$
15 3 across all cases to facilitate visual comparison of the reduced representations.
16 Training is performed using the Adam optimizer [51] with learning rate 1×10^{-3} ,
17 $\beta_1 = 0.9$, and $\beta_2 = 0.999$. Early stopping is triggered when validation loss does
18 not improve by at least 10^{-4} over five epochs, with a maximum of 2000 epochs
19 allowed. The trained networks are provided as supplementary material.

20 From each model, the chemical carriers are extracted based on correlation in-
21 dexes between latent variables and species mass fractions, using the Kendall Tau B
22 approach due to its robustness [40]. These chemical carriers are interpreted as key
23 reaction descriptors that preserve the statistical variance of the whole mechanism

1 while enabling sparse parameterizations of the chemical state.

2 3.4. Chemical carriers evaluation via permutation feature importance

3 To assess how well the identified chemical carriers parameterize the chemi-
4 cal mechanism, we apply a Permutation Feature Importance (PFI) analysis. This
5 technique evaluates the relative importance of input variables by quantifying the
6 degradation in model performance when the values of one input variable are ran-
7 domly permuted.

8 In this context, a linear regression model is trained to predict the mass fractions
9 of all non-carrier species using the identified chemical carriers as inputs. Although
10 the underlying relationships may be nonlinear, the use of a linear model provides
11 a first-order approximation to assess the explanatory power of the carriers. This
12 allows us to determine how much of the behavior of other species can be captured
13 by the identified subset, either through direct interactions (via chemical reactions)
14 or indirectly through coupled pathways.

15 Once the model is trained, the PFI procedure consists of independently per-
16 muting each chemical carrier’s values over the test set and measuring the reduction
17 in the model’s R^2 score. The importance index i_j for input feature j is defined as:

$$i_j = s - \frac{1}{Q} \sum_{q=1}^Q s_{q,j}, \quad (12)$$

18 where s is the model’s R^2 score without permutation, $s_{q,j}$ is the score after the
19 q -th permutation of feature j , and Q is the number of permutations performed.

20 The method is implemented using the scikit-learn library [46]. This analy-

sis provides a quantitative measure of how informative each chemical carrier is in predicting the evolution of other species, thereby validating its selection as a reduced-order parameter of the combustion system.

3.5. Regression model and error metrics

To reconstruct the temporal evolution of each chemical species within a given ignition cluster, we calibrate the coefficients of the surrogate decoder model in Equation (11) using *ridge regression*. Ridge regression provides a clear mapping from chemical carriers to species mass fractions, enabling direct examination of coefficient magnitudes and interactions essential for interpretability and explainability in combustion modeling. This approach minimizes the squared error between the model predictions and the reference data while applying L2 regularization to prevent overfitting; even if other regression techniques are possible, the model must allow for term diversification, that is, it should extract information from the carriers and perform the expansion process depicted by decoders, reason why shrinkage methods are not considered in first instance. The regression is performed independently for each cluster of ignition trajectories using the Ridge implementation from the Scikit-learn library [46], with a regularization parameter set to $\lambda = 0.3$. The model is trained on the chemical carrier trajectories identified by the TAE, along with the initial equivalence ratio.

To assess the accuracy of the reconstruction, we compare the predicted time series to the original simulation data using the coefficient of determination (R^2), the mean absolute error (MAE), Maximum Error, and visual inspection of selected

1 trajectories to evaluate temporal alignment and qualitative trends.

2 **4. Results**

3 The results are obtained through the codes available in the repository: [https :
4 //github.com/lmcastellanos/Combustion_Chemistry_Sparse_
5 Surrogate_Modeling_-_via_TAE_and_Gradient-Based_Clustering](https://github.com/lmcastellanos/Combustion_Chemistry_Sparse_Surrogate_Modeling_-_via_TAE_and_Gradient-Based_Clustering).

6 The proposed methodology is evaluated from two perspectives: (i) the consistency
7 and interpretability of the chemical carriers identified across a range of ignition
8 conditions, and (ii) the accuracy of thermochemical state reconstruction based on
9 clustering informed by progress variable dynamics. Accordingly, the results are
10 presented in the following two sections aligned with these objectives.

11 *4.1. Chemical Carriers Selection Analysis*

12 The chemical carriers are identified by analyzing the latent space of each
13 trained TAE model using correlation metrics between the latent variables and
14 species mass fractions. The species with the strongest correlations to the latent
15 variables are selected as chemical carriers; an example of how such correlation
16 indexes were analyzed is available in Figure 4, in which panel (a) shows the high-
17 est correlation indexes between the latent variable Z_3 and the thermochemical
18 variables, and panel (b) shows the behavior of the latent variable Z_3 and the se-
19 lected specie CH_3 . The results across all ignition cases are summarized in Table 1.
20 It is noted that only state variables (or chemical species) with positive correlation
21 coefficients are allowed, thereby avoiding divergence issues that might arise from

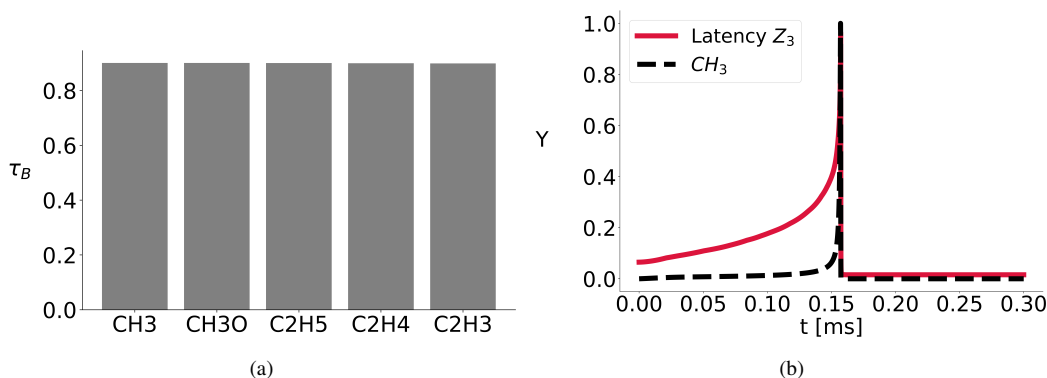


Figure 4: Correlation indexes analysis for the reduced manifold $\phi = 1.25$; results are shown for the specific latent variable Z_3 , with panel (a) showing the Kendall Tau B (τ_B) correlation indexes, and panel (b) showing the behavior comparison between the latent variable and the respective chemical carrier.

1 zero values when such carriers are used as model inputs.

2 For lean mixtures, particularly near the flammability limit (e.g., $\phi = 0.45$),
3 oxygen (O_2) emerges as a dominant carrier, which is expected given the oxidizer-
4 rich conditions. Other frequently selected species in lean regimes include H, OH,
5 and CH_2OH , all of which participate in early dehydrogenation reactions and hy-
6 drogen abstraction from methane. Intermediate species such as CO, HCCO, and
7 CO_2 appear more prominently as equivalence ratios increase, reflecting the tran-
8 sition to oxidation and CO_2 formation pathways.

9 In fuel-rich conditions, the carriers shift toward species related to incomplete
10 combustion and chain-branching limits, such as CH_3OH , CH_2OH , and C_2H_5 .
11 These species play a central role in the breakdown of fuel and modulation of
12 the radical pool. For instance, CH_3 is both a product of methane dehydrogenation
13 and a precursor in CH_4 reformation pathways (e.g. reactions R162 and R163 in
14 Table 2). These reactions also explain the appearance of CH_2OH and CH_3OH as

Table 1: Identified chemical carriers for each ignition trajectory

Equivalence Ratio ϕ	Identified Carriers
0.45	O ₂
0.55	HO ₂ , CO ₂ , H
0.65	HO ₂ , OH, CO
0.75	CH ₄ , HCCO, O
0.85	H, CH ₂ OH, CH ₄
0.95	OH, H, H ₂
1.05	H, CH ₂ OH, <i>H</i>
1.15	C, H, HCO
1.25	OH, CH ₃ OH, CH ₃
1.35	H, CH ₄ , <i>H</i>
1.45	CH ₃ O
1.55	H ₂ , C ₂ H ₅
1.65	H, H, CH ₄
1.75	CH ₄ , H, CH ₄
1.85	C ₂ H ₅ , H

chemical carriers. Similarly, radicals such as H, OH, and HO₂ are essential for sustaining oxidation processes and are thus often selected as carriers [52].

Overall, the carriers identified by the TAE, such as OH, CH₂OH, and CH₃, correspond to well-known kinetic controllers in methane combustion [53]. OH is a key chain-branching radical that accelerates oxidation and affects ignition delay times [54], whereas CH₂OH and CH₃ serve as essential intermediates in fuel decomposition and reformation pathways [55]. Their consistent selection across equivalence ratios confirms that the model captures the main reaction pathways that govern ignition and flame propagation. For engineers, this is significant because it reduces the dimensionality of the reaction mechanism without compromising the variables essential to heat release and pollutant formation. In practice, using these carriers as inputs for surrogate models allows for reliable prediction of ignition behavior and major species evolution with fewer equations, thereby

1 reducing computational cost in CFD simulations or reactor design tasks while
 2 preserving mechanistic interpretability.

Table 2: Examples of CH_3 and CH_3OH related reactions for CH_4 production in GRI-mech2.11

Reaction #	CH_4 reformation pathway
162	$\text{CH}_3 + \text{CH}_3\text{OH} \leftrightarrow \text{CH}_2\text{OH} + \text{CH}_4$
163	$\text{CH}_3 + \text{CH}_3\text{OH} \leftrightarrow \text{CH}_3\text{O} + \text{CH}_4$

3 To gain insights into how well the *chemical carriers* parameterize the chemical
 4 mechanism, we use the PFI test. Figure 5 presents the results of the PFI test
 5 analysis for the ignition case with initial conditions $T_0 = 1418\text{K}$, $\phi = 1.25$, where
 6 the identified chemical carriers are OH , CH_3OH , and CH_3 . Each point in the 3D
 7 scatter plot represents a chemical species, and its coordinates correspond to the
 8 permutation importance indices with respect to each chemical carrier. The Z-
 9 axis magnitude is also color-coded for clarity, and a 2D projection is included for
 10 reference.

11 The PFI result shows that nearly all species exhibit sensitivity to at least one
 12 of the chemical carriers, confirming the explanatory power of this approach. The
 13 scatter distribution forms three distinguishable branches, suggesting that a single
 14 carrier primarily influences different subsets of species. This separation supports
 15 the notion that the identified carriers effectively partition the chemical mechanism
 16 into interpretable subsystems, each governed by one dominant species.

17 These results validate the chemical relevance of the selected carriers and demon-
 18 strate their utility in parameterizing the thermochemical space under the consid-
 19 ered conditions. In the next subsection, we further analyze the clustering of igni-

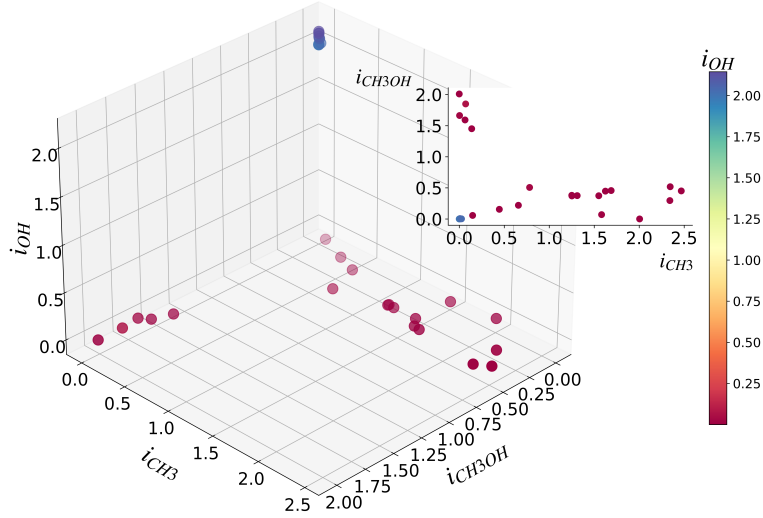


Figure 5: Permutation importance test results for GRI-Mech 2.11 under ignition conditions $T_0 = 1418\text{ K}$ and $\phi = 1.25$, with latent space dimension $n_Z = 3$. Each point represents a chemical species, and its position along the three axes corresponds to its permutation importance index with respect to the three identified chemical carriers: OH, CH₃OH, and CH₃. The Z-axis is also color-mapped to aid interpretation. The 2D inset provides a top view for additional clarity. The species cluster into three main groups, each predominantly sensitive to a different chemical carrier, illustrating the carriers' explanatory capacity in capturing distinct chemical subsystems.

tion cases and the associated selection of chemical carriers across different chemical regimes.

4.2. Clustering and surrogate model results

The clustering results are summarized in Figure 6, which shows the groups formed based on IDT values with 3 distinct numbers of clusters $n_k = 2, 3, 4$. Additionally, the clustering quality is assessed with the Davies-Bouldin [47] and Calinski-Harabasz [48] scores, as shown in Table 3. From the clustering quality analysis, it is observed that the Davies-Bouldin score shows low variance for all cases, which is likely related to the low sparsity between the clusters; that is,

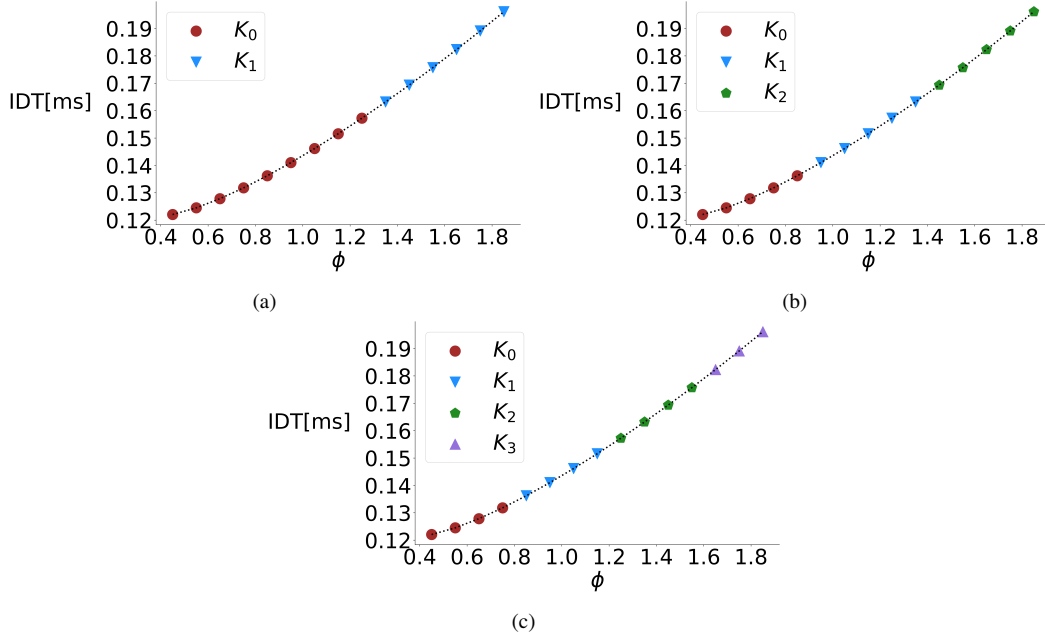


Figure 6: Clustering results for the ignition cases based on progress variable gradient similarity. Panels (a), (b), and (c) show the cluster distributions for $n_k = 2$, $n_k = 3$, and $n_k = 4$, respectively. Higher values of n_k are not considered, as they lead to over-segmentation, with some clusters containing only a single ignition trajectory.

1 the IDT values in the whole range are not widely separated from each other. The
2 Calinski-Harabasz score, on the other hand, varies proportionally to the number of
3 clusters, meaning that a higher number of clusters guarantees more compactness
4 and, therefore, higher similarity.

Table 3: Clustering Qualities Scoring

Clustering Strategy	Davies Bouldin Score	Calinski-Harabasz Score
K=2	0.48011	41.2873
K=3	0.48000	49.8137
K=4	0.48932	60.5845

5 For each resulting cluster, a regression model $\mathcal{P}(\Phi)$ (as in equation 11) is
6 trained. This calibration is performed using a Ridge regressor with a regulariza-

tion constant $\lambda = 0.3$ and a polynomial degree $m = 2$. To evaluate model robustness, we focus on ignition conditions near chemical regime transitions, where the separation of clustering may seem less intuitive. Specifically, the interpolation performance is analyzed for ignition cases at $\phi = 0.90$ and $\phi = 1.10$. The quality of the regression is assessed using R^2 scores and comparisons of reconstructed curves, as well as assessing the average Mean Absolute Error (MAE) and maximum error in the reconstructions of each ignition trajectory. Additional testing is done for the ignition cases $\phi = 0.5$ and $\phi = 1.8$, since they represent cases near the extremes of the sampled domain and show the model’s behavior when considering other clusters/chemical kinetics.

The chemical carriers assignments for $n_k = 2$ clusters are provided in Table 4; the higher amount of chemical carriers in cluster K_0 is proportional to the number of ignition trajectories in the cluster. After fitting the corresponding regression models, all training cases achieve $R^2 > 0.9$. The same performance is observed in the interpolation cases at $\phi = 0.90$ and $\phi = 1.10$. Representative reconstruction curves for cluster K_0 are shown in Figure 7. The reconstructions for $\phi = 0.5$ and $\phi = 1.80$ are available in Figure 8. It is to observe that a small peak deviation is presented in the minor species, specifically in Figure 8 panel (d), which can be associated with the small dimension of the input vector space and a requirement for further non-linearities expression.

The chemical carriers set for $n_k = 3$ are presented in Table 5. The regression models continue to show strong performance, with R^2 values exceeding 0.9 for all training and interpolation datasets. Reconstruction examples for the interpolation

Table 4: Chemical carriers distribution for $n_k = 2$ clusters

Cluster	Chemical Carriers
K_0	OH, CH ₃ OH, CH ₃ , H, CH ₂ OH, CH ₄ , HO ₂ , CO, O ₂ , CO ₂ , H ₂ , C, HCO, HCCO
K_1	H, CH ₄ , H ₂ , CH ₃ O, C ₂ H ₅

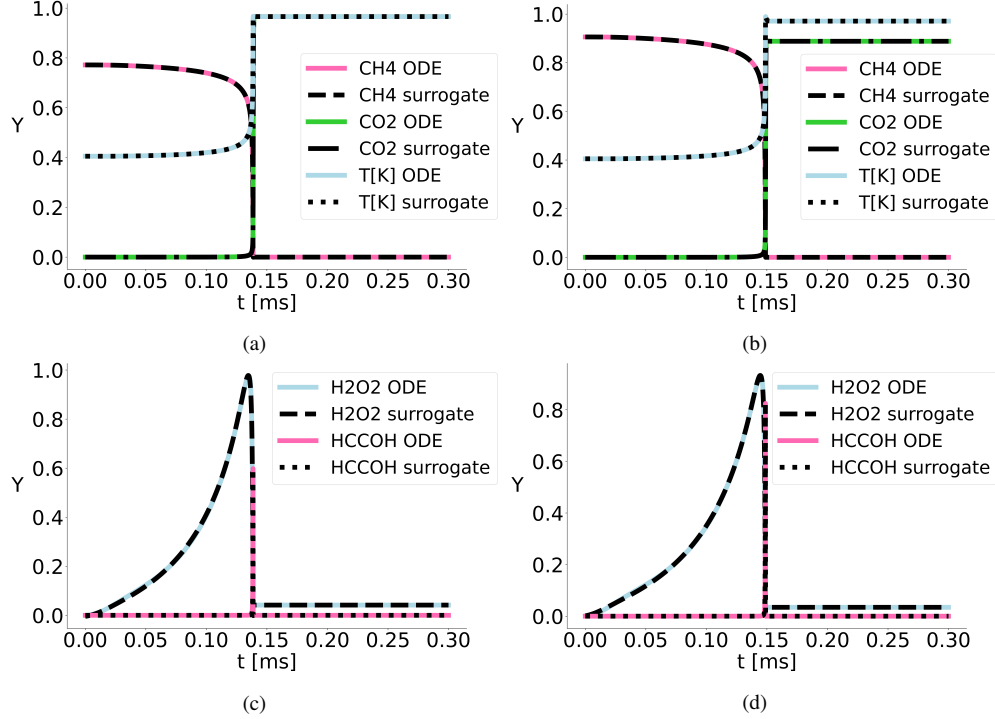


Figure 7: Interpolation results for cluster K_0 in the $n_k = 2$ clustering case; Y stands for the normalized magnitude. Panels (a) and (c) show the reconstructed ignition trajectory for $\phi = 0.90$, presenting an average MAE of 0.11×10^{-3} and maximum error of 3.68×10^{-3} ; instead panels (b) and (d) display the reconstruction for $\phi = 1.10$ with an average MAE of 0.10×10^{-3} and maximum error of 3.03×10^{-3} . These cases evaluate the model's interpolation performance near the transition between chemical regimes.

cases are illustrated in Figure 9. The reconstructions for the ignition trajectories $\phi = 0.5$ and $\phi = 1.80$ are displayed in Figure 10.

The chemical carrier distributions for $n_k = 4$ are shown in Table 6. While most clusters still achieve $R^2 > 0.9$, performance degrades slightly for cluster K_3 ,

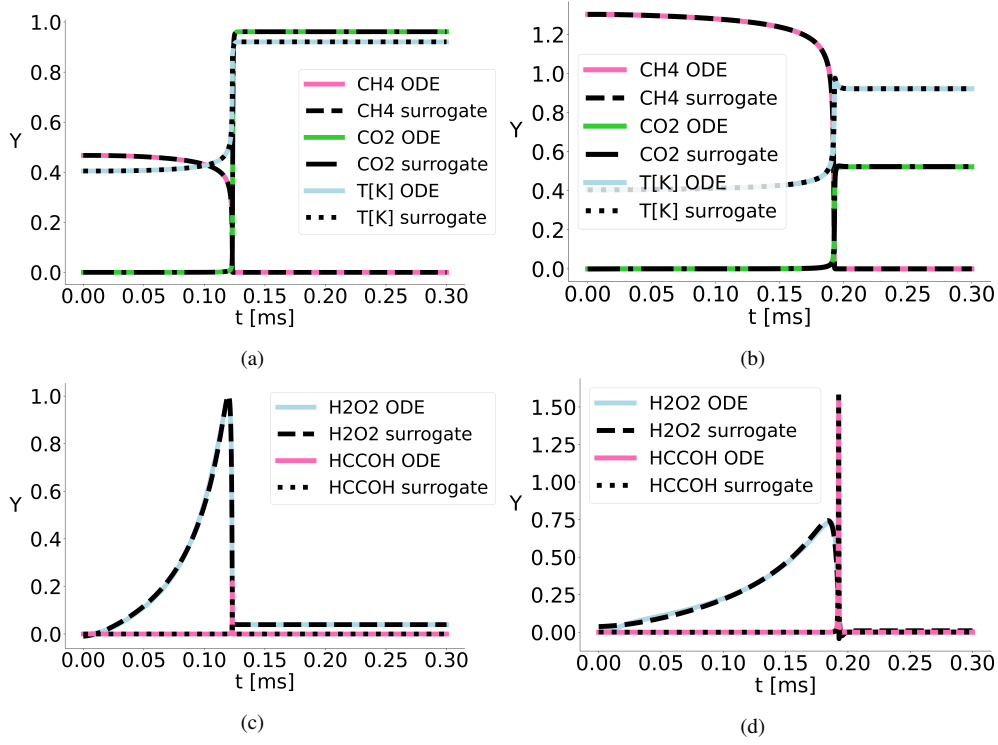


Figure 8: Interpolation results for clusters K_0 and K_1 in the $n_k = 2$ clustering case; Y stands for the normalized magnitude. Panels (a) and (c) show the reconstructed ignition trajectory for $\phi = 0.50$, presenting an average MAE of 0.19×10^{-3} and maximum error of 4.20×10^{-3} ; instead panels (b) and (d) display the reconstruction for $\phi = 1.80$ with an average MAE of 1.09×10^{-3} and maximum error of 40.10×10^{-3} , such maximum error might represent the small peak deviation in minor species; note the differences in errors magnitudes and the difference in number of considered carriers are inversely proportional.

1 where a few species show R^2 values between 0.7 and 0.8. The reduced accuracy
2 is observed in minor species, likely due to a sparser input vector (only 3 carriers
3 used instead of the full 31-species state) and the linear model's inability to capture
4 stronger nonlinear dynamics.

5 Figure 11 highlights reconstruction performance for key interpolation cases ,
6 namely $\phi = 0.90$ and $\phi = 1.10$; the ignition trajectory $\phi = 0.5$ is available in

Table 5: Chemical carriers distribution for $n_k = 3$ clusters

Cluster	Chemical Carriers
K_0	OH, H, CH ₂ OH, CH ₄ , HO ₂ , CO, O ₂ , CO ₂ , O, HCCO
K_1	OH, CH ₃ OH, CH ₃ , H, CH ₂ OH, CH ₄ , H ₂ , C, HCO
K_2	CH ₃ O, H, C ₂ H ₅ , CH ₄ , H ₂

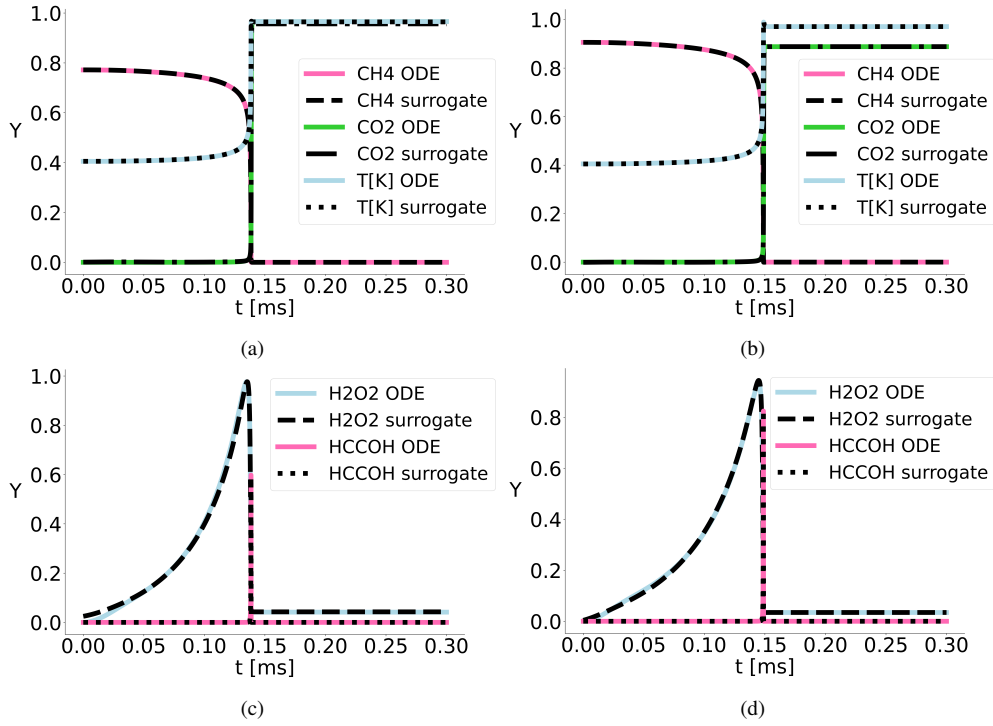


Figure 9: Interpolation results for cluster K_1 in the $n_k = 3$ clustering case; Y stands for the normalized magnitude. Panels (a) and (c) show the reconstructed ignition trajectory for $\phi = 0.90$, presenting an average MAE of 0.90×10^{-3} and maximum error 18.06×10^{-3} ; instead panels (b) and (d) display the reconstruction for $\phi = 1.10$, with MAE 0.25×10^{-3} and maximum error 8.70×10^{-3} . These cases evaluate the model's interpolation performance near the transition between chemical regimes.

1 Figure 12, as a representation of performance in cluster K_0 . For cluster K_3 , the
2 trends are generally preserved, though some concentration profiles—particularly
3 for the carbon atom (C) species—are not well captured at $\phi = 1.80$, as seen in

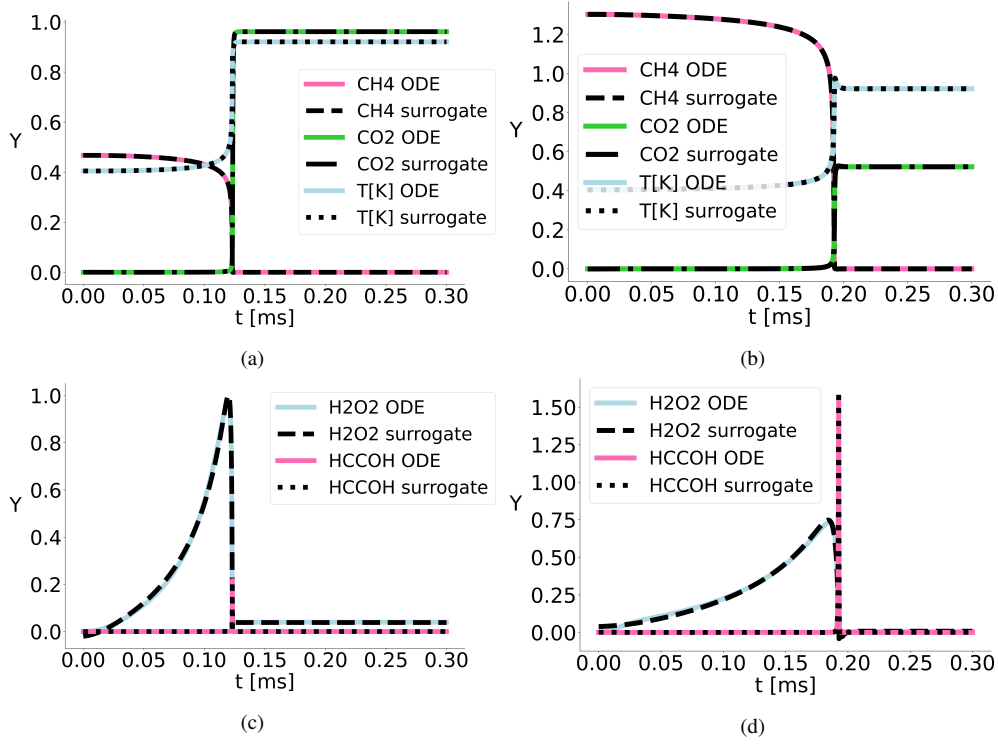


Figure 10: Interpolation results for cluster K_0 and K_2 in the $n_k = 3$ clustering strategy; Y stands for the normalized magnitude. Panels (a) and (c) show the reconstructed ignition trajectory for $\phi = 0.50$, presenting an average MAE of 0.36×10^{-3} and maximum error 12.76×10^{-3} ; instead panels (b) and (d) display the reconstruction for $\phi = 1.80$, with MAE 1×10^{-3} and maximum error 40.07×10^{-3} .

- 1 Figure 13. A histogram of the R^2 scores for these cases is presented in Figure 14,
- 2 indicating good interpolation performance for most species, despite the minimal
- 3 input set.

Table 6: Chemical carrier distribution for $n_k = 4$ clusters

Cluster	Chemical Carriers
K_0	OH, H, CH ₄ , HO ₂ , CO, O ₂ , CO ₂ , O, HCCO
K_1	OH, H, CH ₂ OH, CH ₄ , H ₂ , C, HCO
K_2	OH, CH ₃ OH, CH ₃ , H, CH ₄ , CH ₃ O, H ₂ , C ₂ H ₅
K_3	H, CH ₄ , C ₂ H ₅

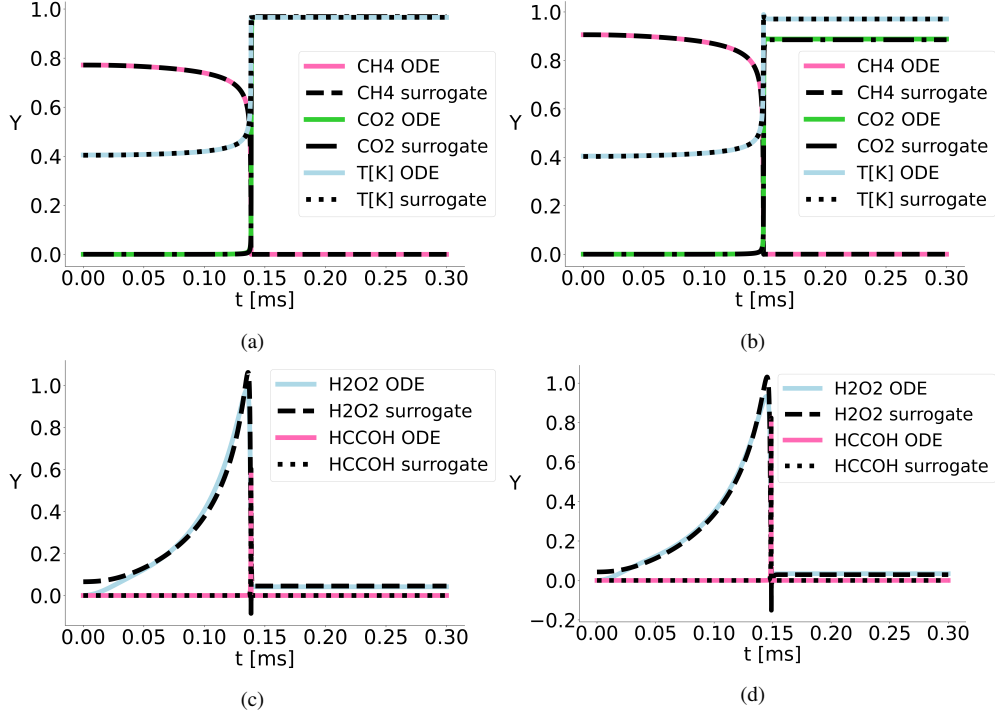


Figure 11: Interpolation results for cluster K_1 in the $n_k = 4$ clustering case; Y stands for the normalized magnitude. Panels (a) and (c) show the reconstructed ignition trajectory for $\phi = 0.90$, with a MAE of 1.47×10^{-3} and maximum error of 50.24×10^{-3} while (b) and (d) display the reconstruction for $\phi = 1.10$, with a MAE of 1.28×10^{-3} and maximum error of 28.43×10^{-3} . These cases evaluate the model's interpolation performance near the transition between chemical regimes.

4.3. Impact of cluster number

To assess the effect of varying the number of clusters on reconstruction performance, we analyze two key error metrics: the Mean Absolute Error (MAE), which evaluates overall model accuracy, and the Maximum Error, which highlights localized prediction failures that may be smoothed out in MAE averages. These metrics are reported in Table 7 and Table 8.

The results reveal that while MAE remains relatively consistent and low for

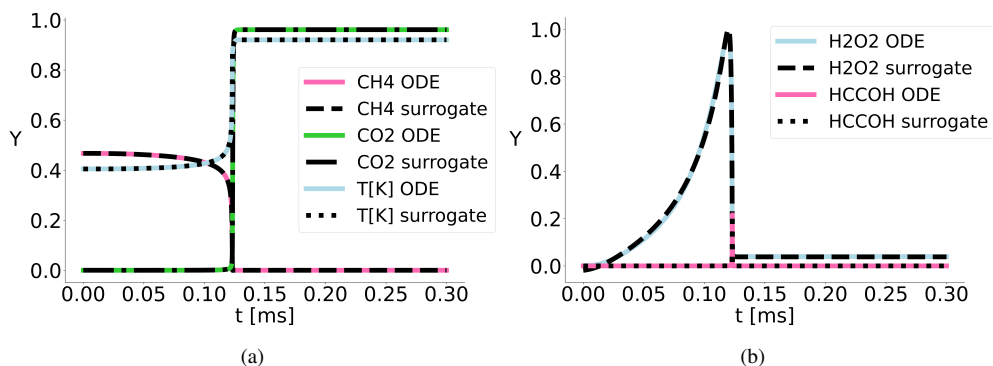


Figure 12: Interpolation results for the ignition trajectory $\phi = 0.5$ in cluster K_1 and the $n_k = 4$ clustering case; Y stands for the normalized magnitude. Panels (a) show the major species reconstructions, while (b) shows the minor species results; the MAE value stands in 0.37×10^{-3} while the maximum error amounts to 13.17×10^{-3} .

Table 7: MAE Error Metrics for Different Ignition Trajectories in Interpolation. All values escaled by $\times 10^{-3}$

Clustering Strategy	MAE [$\times 10^{-3}$]			
	$\phi = 0.5$	$\phi = 0.9$	$\phi = 1.1$	$\phi = 1.8$
$n_k = 2$	0.19	0.116	0.101	1.09
$n_k = 3$	0.36	0.904	0.254	1.00
$n_k = 4$	0.37	1.47	1.25	7.58

1 most of the ignition trajectories—indicating good general model performance—the
 2 Maximum Error exposes localized discrepancies in specific time instants or species.
 3 These spikes are particularly evident in the $n_k = 4$ clustering, as shown in Fig-
 4 ure 11. Such errors typically stem from inadequate parameterization of chemical
 5 kinetics within a cluster due to limited input variables. It must be highlighted
 6 that such cluster parametrizes the chemical kinetics using four variables, there-
 7 fore, almost 90% of variables have been eliminated form the input vector space.
 8 However, the appearance of such peak errors can be problematic due to their non-
 9 physical meaning (appearance of negative concentrations); these errors can be

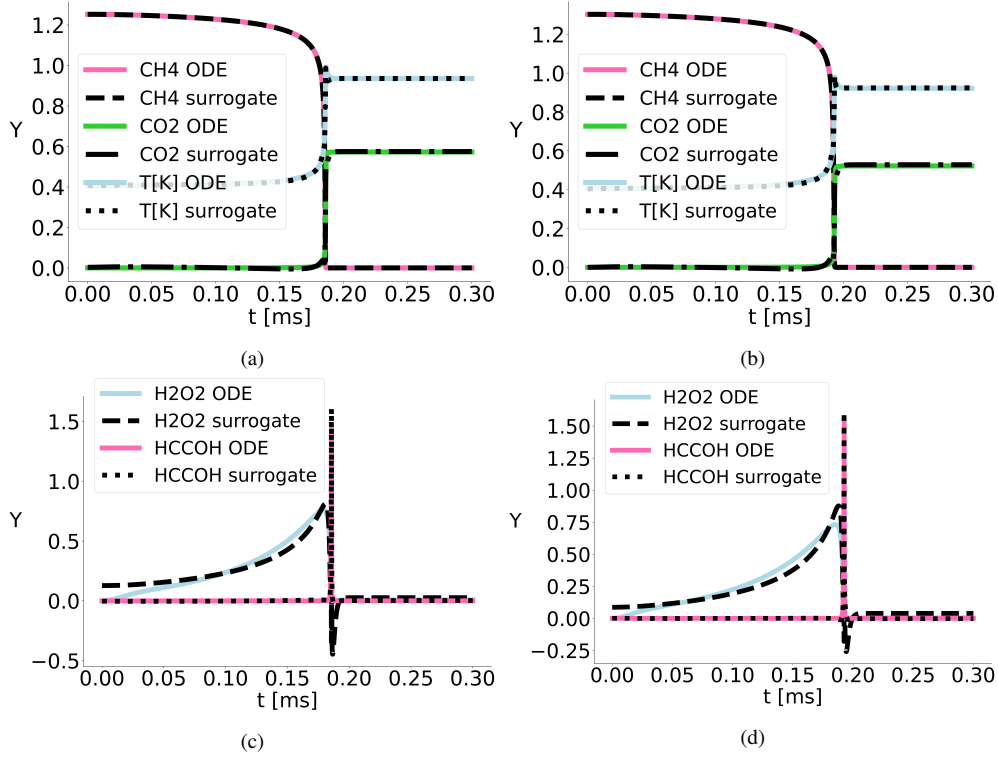


Figure 13: Interpolation curves for cluster K_3 in the $n_k = 4$ clustering case; Y stands for the normalized magnitude. Panels (a) and (c) show the reconstructed ignition trajectory for $\phi = 1.70$, with MAE 7.41×10^{-35} and maximum error of 215×10^{-3} , while (b) and (d) present the results for $\phi = 1.80$, with MAE 7.58×10^{-3} and maximum error of 195×10^{-3} ; the higher error metrics are explained due to the peak error values presented in specific time instances. These cases demonstrate the model’s ability to reconstruct trajectories within a sparse and chemically distinct cluster.

1 addressed making use of truncation methods, as the constrained Gaussian Pro-
2 cess Regressions (GPR) [56]. It should be mentioned that, due to the nature of
3 the regression model implemented, it is more challenging to implement physical
4 constraints, since every model holds many inputs to a single output structure.

5 To further investigate this limitation, we conduct an experiment in which all 18
6 chemical carriers identified across the entire dataset are used together as inputs to

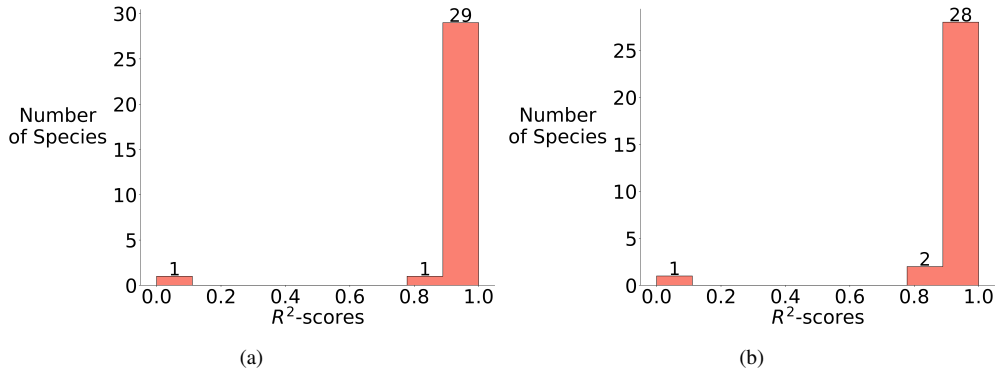


Figure 14: R^2 scores for interpolation cases in cluster K_3 of the $n_k = 4$ clustering configuration. Panel (a) corresponds to the ignition condition $\phi = 1.70$, and (b) to $\phi = 1.80$. In both cases, the reconstruction achieves high accuracy across most species, with notable loss of accuracy only for the carbon atom (C) species.

Table 8: Max Error Metrics for Different Ignition Trajectories in Interpolation. All values escaled by $\times 10^{-3}$

Clustering Strategy	Max Error [$\times 10^{-3}$]			
	$\phi = 0.5$	$\phi = 0.9$	$\phi = 1.1$	$\phi = 1.8$
$n_k = 2$	4.20	3.68	3.03	40.10
$n_k = 3$	12.76	18.06	8.70	40.07
$n_k = 4$	13.17	50.24	28.43	195

1 regression models. This test is conducted within the $n_k = 4$ clustering setup, i.e.,
2 we fit new regression models, and the new error metrics are presented in Table 9
3 and 10. Reconstruction curves for this configuration are shown in Figure 15 and
4 16.

5 The comparison demonstrates that increasing the number of chemical carriers
6 substantially reduces both MAE and Maximum Error, often by orders of magni-
7 tude. This is particularly noticeable when comparing Figures 11 and 15. However,
8 increasing the dimensionality of the latent space to extract more carriers requires
9 training additional ignition trajectories and modifying the autoencoder architec-

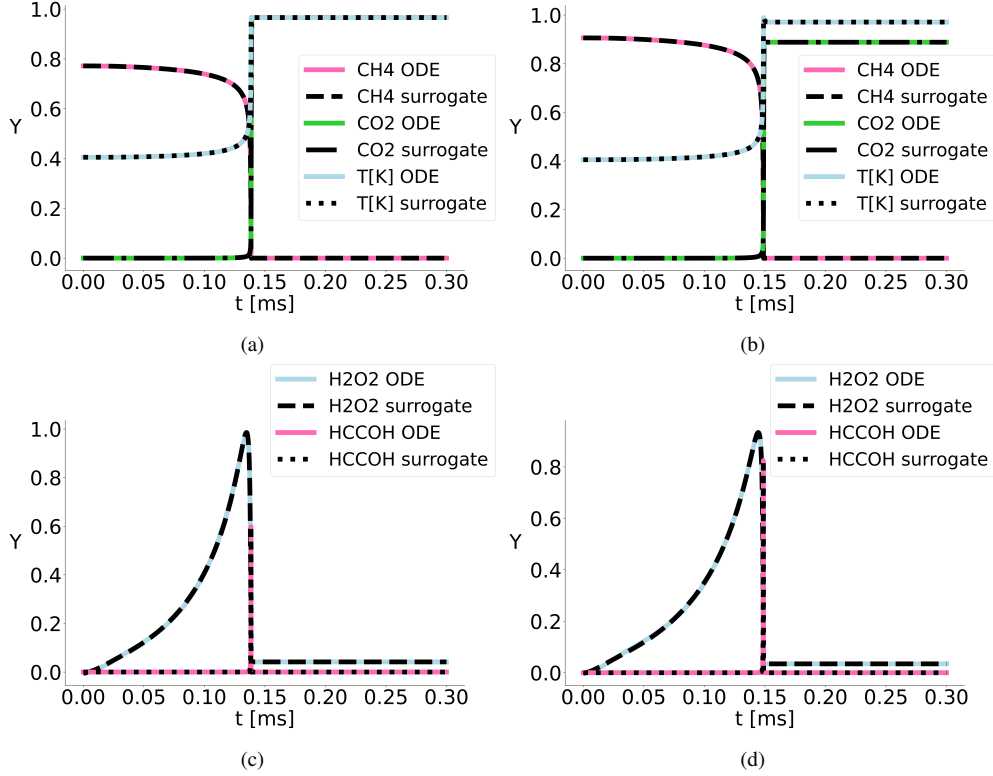


Figure 15: Interpolation results for cluster K_1 in the four-cluster ($n_k = 4$) configuration, using the full set of identified chemical carriers; Y stands for the normalized magnitude. Panels (a) and (c) correspond to the ignition case at $\phi = 0.90$, presenting a MAE of 0.11×10^{-3} and a maximum error of 2.63×10^{-3} while (b) and (d) show results for $\phi = 1.10$, with MAE 0.093×10^{-3} and maximum error 2.28×10^{-3} .

1 ture accordingly [26]; however, the final number of inputs must be chosen accord-
2 ing to the model requirements; that is, a model with an excessive number of inputs
3 could present higher variance or adaptability, and can improve error results, but it
4 can also produce overfitting due to multicollinearity [57], therefore, a balance must
5 be achieved between bias and variance. Additionally, the regression model used
6 here is simple and may not fully capture the nonlinearities learned by the neural
7 network's activation functions.

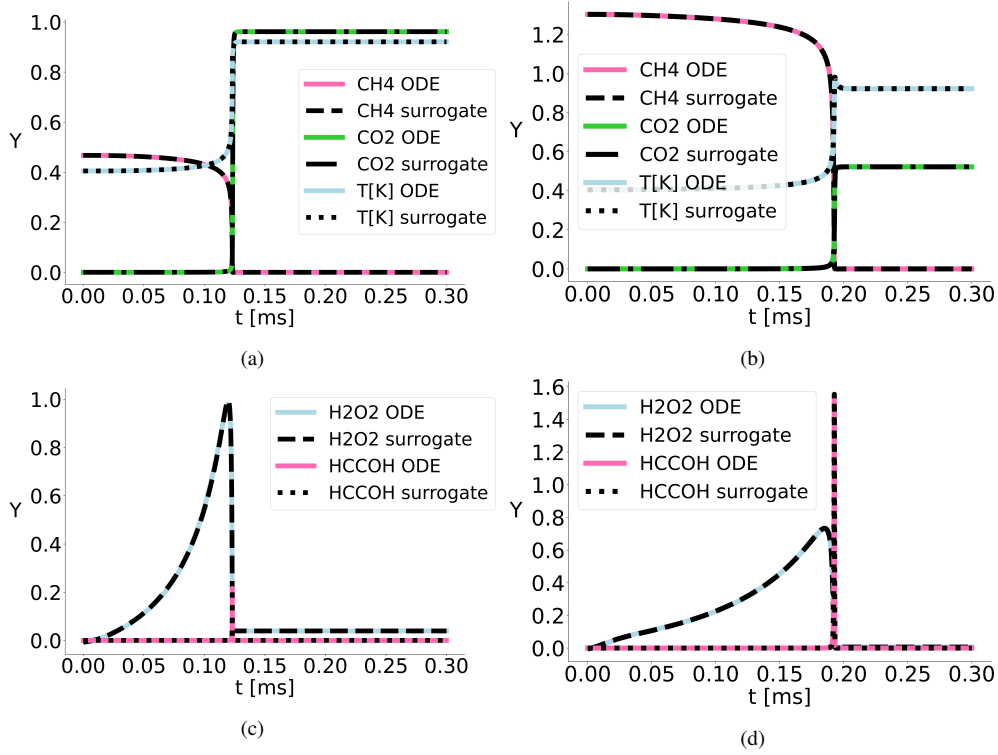


Figure 16: Interpolation results for cluster K_0 and K_3 in the $n_k = 4$ clustering strategy using all the chemical carriers; Y stands for the normalized magnitude. Panels (a) and (c) show the reconstructed ignition trajectory for $\phi = 0.50$, presenting an average MAE of 0.15×10^{-3} and maximum error 3.05×10^{-3} ; instead panels (b) and (d) display the reconstruction for $\phi = 1.80$, with MAE 0.13×10^{-3} and maximum error 2.28×10^{-3} .

1 Nevertheless, even when using all available carriers (18 variables), the reduced
2 representation still employs 42% fewer parameters than the full thermochemical
3 state vector (31 variables), indicating a significant dimensionality reduction. Fu-
4 ture work will explore optimal carrier selection strategies and Gappy-Autoencoder
5 sampling approaches to improve potential speed-ups; this is key since the im-
6 plementation of unnecessary inputs will traduce into unnecessary operations and
7 memory requirements.

1 To isolate the influence of trajectory similarity (especially in terms of progress
2 variable gradients), we compare specific clusters across configurations. Notably,
3 cluster $K_{0,2}$ in the $n_k = 2$ configuration (with 9 trajectories) is compared with
4 cluster $K_{1,4}$ in the $n_k = 4$ configuration (with 4 trajectories). Both are analyzed
5 using the full set of chemical carriers, allowing us to isolate the effect of cluster
6 diversity on model performance.

7 Error metrics for both clusters are reported in Table 11 (MAE) and Table 12
8 (Maximum Error), where mean values and 95% confidence intervals are provided.

9 While the mean errors are similar in both clusters, the confidence intervals are
10 narrower for $K_{1,4}$, indicating lower variability in model performance. This sug-
11 gests that although $K_{0,2}$ benefits from a larger set of input features, the increased
12 diversity within the cluster leads to higher regression variability. Conversely, $K_{1,4}$
13 shows more consistent performance due to its internal homogeneity.

Table 9: MAE Error Metrics for Different Ignition Trajectories in Interpolation with Differences in the Number of Carriers. All values escaled by $\times 10^{-3}$

Clustering Strategy	MAE [$\times 10^{-3}$]			
	$\phi = 0.5$	$\phi = 0.9$	$\phi = 1.1$	$\phi = 1.8$
$n_k=4$ (Cluster-wise Carriers)	0.37	1.47	1.28	7.58
$n_k=4$ (All Carriers)	0.15	0.115	0.0934	0.13

Table 10: Max Error Metrics for Different Ignition Trajectories in Interpolation with Differences in the Number of Carriers. All values escaled by $\times 10^{-3}$

Clustering Strategy	Max Error [$\times 10^{-3}$]			
	$\phi = 0.5$	$\phi = 0.9$	$\phi = 1.1$	$\phi = 1.8$
$n_k=4$ (Cluster-wise Carriers)	13.17	50.24	28.43	195
$n_k=4$ (All Carriers)	3.05	2.63	2.28	2.28

Table 11: MAE error metrics (mean and 95% confidence intervals) for all-carriers models comparing $n_k = 2$ and $n_k = 4$. All values are scaled by 10^{-3} .

Clusters case	$\phi = 0.9$		$\phi = 1.1$	
	Mean	Confidence	Mean	Confidence
$n_k=2$	0.103	$[-0.0078, 0.214]$	0.0883	$[-0.0049, 0.181]$
$n_k=4$	0.115	$[0.0131, 0.218]$	0.0934	$[0.0059, 0.181]$

Table 12: Maximum error metrics (mean and 95% confidence intervals) for all-carriers models comparing $n_k = 2$ and $n_k = 4$. All values are scaled by 10^{-3} .

Clusters case	$\phi = 0.9$		$\phi = 1.1$	
	Mean	Confidence	Mean	Confidence
$n_k=2$	2.80	$[1.33, 4.26]$	2.24	$[1.45, 3.04]$
$n_k=4$	2.63	$[2.04, 3.22]$	2.28	$[1.52, 3.05]$

1 Nonetheless, this homogeneity is a cause of drawbacks when predicting in
2 extrapolation conditions, that is, ignition trajectories outside the cluster’s region.
3 Indeed, homogeneity produces reliable and accurate behavior for similar cases,
4 while presenting higher errors in extrapolation. At the same time, when higher
5 variability is considered in the training process (i.e., higher variety of training
6 data), the model becomes less accurate in interpolation conditions while offering
7 higher performance in extrapolation conditions. Error metrics for cases $K_{0,2}$ and
8 $K_{1,4}$ tested in the extrapolation condition of $\phi=1.65$ are shown in Table 13, while
9 the reconstruction curves for $K_{0,2}$ are shown in Figure 17.

10 The extrapolation performance does not achieve the high accuracy required for
11 chemical kinetics applications since errors also appear in major species. There-
12 fore, best practices involve sufficient sampling for Chemical carriers while using
13 a diversity of clusters, generating many accurate and reliable models that describe

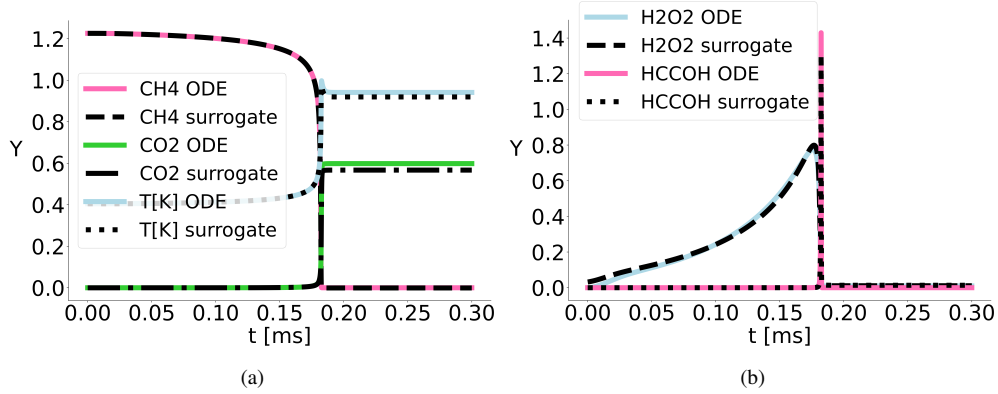


Figure 17: Extrapolation results for cluster K_0 in the $n_k = 2$ clustering case with all the carriers; Y stands for the normalized magnitude. Panels show the ignition case $\phi = 1.65$, a condition far away from the clusters limits.

the chemical kinetics. Nonetheless, the results presented for the cluster $K_{0,2}$ are of interest for generalization purposes; that is, adequate gradient-based clustering could increase model generalization even in the presence of a significant number of inputs.

Table 13: Max Error Metrics for all-carriers models comparing $n_k = 2$ and $n_k = 4$ in the extrapolation condition $\phi = 1.65$.

Clusters case	MAE [$\times 10^{-3}$]		Max Error [$\times 10^{-3}$]	
	Mean	Confidence	Mean	Confidence
$n_k=2$	8.95	[5.88, 12.02]	80.73	[41.32, 120.14]
$n_k=4$	27.97	[17.23, 38.71]	123.44	[78.78, 120.14]

Finally, it is worth noting that not all chemical carriers contribute equally to regression accuracy. Many features exhibit negligible influence, as shown by regression coefficients of order 10^{-5} or 10^{-6} . These could potentially be pruned in future implementations, offering further model simplification without compromising performance; therefore, the implementation of tailored shrinkage methods

1 will be analyzed in future work for each chemical species case , affecting speed-
2 ups by the number of operations. Increasing the Ridge regression regularization
3 constant did not yield better results.

4 In summary, beyond demonstrating predictive accuracy in ignition problems,
5 the proposed surrogate framework can be directly integrated into various areas
6 of chemical engineering practice. By replacing the direct use of stiff chemical
7 kinetics with sparse, interpretable surrogates significantly lowers computational
8 costs, allowing high-fidelity combustion chemistry to be integrated into large-
9 scale process simulations or multiphysics CFD solvers where detailed kinetics
10 would otherwise be prohibitively expensive. For reactor design, the framework
11 enables quick assessment of ignition and stability features across broad operating
12 ranges, aiding in optimization of fuel flexibility, pollutant reduction, and energy
13 efficiency. In optimization and control contexts, the low-dimensional chemical
14 carriers offer compact yet chemically meaningful state variables that are ideal
15 for integration with model predictive control or real-time digital twin platforms.
16 Therefore, the framework provides a practical route for speeding up engineering
17 design and decision-making in energy and process systems.

18 **5. Conclusions**

19 This work presents a novel framework for developing accurate surrogate mod-
20 els of combustion chemical kinetics by leveraging similarities in progress variable
21 gradient profiles, time-lagged autoencoders, and chemical carrier regression mod-
22 els. The approach enables a priori clustering of ignition trajectories based on

1 their gradient similarity, allowing the development of gradient-informed models.
2 These models operate using reduced input vectors consisting of selected chemical
3 species—the chemical carriers—identified through a Gappy-Autoencoder strat-
4 egy. The proposed framework enhances interpretable, data-efficient surrogate
5 modelling, validated by its high accuracy ($R^2 > 0.9$) and adaptability to sparse
6 data regimes. In summary, different aspects can be concluded considering the
7 clustering strategy, the Gappy-AE methodology, as well as some remarks for fu-
8 ture research and modeling approaches:

9 • **Gradient-based clustering:** The results demonstrate that clustering tra-
10 jectories by PV gradient similarity enhances both model convergence and
11 reconstruction accuracy, even when the input space is significantly reduced.
12 This suggests that PV -based gradient profiling retains critical dynamical
13 information and could serve as a promising basis for generalization, even
14 when considering different chemical regimes. In particular, this property
15 may be leveraged for transfer learning applications in more complex sce-
16 narios [58], where models trained in one regime can be efficiently adapted
17 to similar dynamic conditions through minimal fine-tuning. Moreover, the
18 clustering strategy may also support localized modeling in flame fronts [59]
19 or extrapolation to unseen regimes by exploiting shared structural properties
20 across chemical regimes.

21 • **Gappy-Autoencoder framework:** In terms of chemical carrier identifica-
22 tion, the accuracy of the surrogate models serves as a practical validation

1 of their effectiveness in parameterizing the reaction space. The Gappy-
2 Autoencoder methodology allows one to decompose a complex modeling
3 task into simpler subproblems while reducing the amount of training data
4 required. The number of carriers—or reduced dimensions—is determined
5 iteratively, akin to approaches such as PCA or Proper Orthogonal Decom-
6 position (POD) [41]. These carriers can be enriched by including additional
7 ignition trajectories or refining the clustering resolution. Crucially, since the
8 carriers encode the system’s chemical kinetics, they are potentially trans-
9 ferable to other regions of the thermochemical space sharing similar reac-
10 tion dynamics. The methodology is particularly well-suited for mechanisms
11 with large numbers of species. It offers a promising pathway for parame-
12 terizing combustion chemistry in scenarios with limited data availability—a
13 common challenge in both simulations and experiments. Future work will
14 explore the extension of this method to multidimensional data settings, as-
15 sess its integration into high-fidelity combustion models, and evaluate its
16 applicability for mechanism reduction.

- 17 • **Future research:** The structure of the proposed framework bears resem-
18 blance to the Deep Operator Network (DeepONet) architecture [60], wherein
19 a system of ODEs is decomposed into spatial (chemical carriers) and tem-
20 poral (progress variable) components. This analogy suggests that the pre-
21 sented approach may form a foundation for more general operator learn-
22 ing strategies in combustion modeling. Beyond the current offline formu-
23 lation, the framework could be expanded to include online adaptive surro-

gate models, which are capable of updating chemical carriers or regression coefficients as new data become available. Such adaptability would support integration into real-time applications, such as combustion monitoring or digital twin platforms. Additionally, combining the method with reinforcement learning (RL) could enable closed-loop optimization and control, where the surrogate model provides rapid predictions to direct actuators or operational conditions in response to evolving performance targets. However, further research is necessary to evaluate the method’s robustness under autoregressive deployment, particularly regarding error accumulation and long-term stability in predictive tasks. Moreover, although the present study focused on premixed flames, the methodology is broadly applicable. It can be extended to non-premixed flames, where it would supplement mixture–fraction–based approaches by accurately resolving local combustion states and enhancing kinetic fidelity.

Acknowledgments

The authors would like to acknowledge the insightful discussions regarding dynamical systems with Dr. Jean-Charles Delvenne.

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.

This study is supported by the UK Physical Sciences Research Council under Grant No. EP/X019551/1.

Computational resources have been provided by the Consortium des Équipements

1 de Calcul Intensif (CÉCI), funded by the Fonds de la Recherche Scientifique de
2 Belgique (F.R.S.-FNRS) under Grant No. 2.5020.11 and by the Walloon Region.
3 Additionally, this work has received funding from the European Union's Horizon
4 2020 research and innovation programme under Marie Skłodowska-Curie grant
5 agreement No. 801505.
6 RMG acknowledges the financial support from the Italian Ministry of University
7 and Research in the framework of the project "Italian Ministry of University and
8 Research" funded by EU – NextGenerationEU

9 **References**

- 10 [1] T. Poinso, D. Veynante, Theoretical and numerical combustion, RT Ed-
11 wards, Inc., 2005.
- 12 [2] D. Veynante, L. Vervisch, Turbulent Combustion Modelling, Progress in En-
13 ergy and Combustion Science 28 (2002) 193–266.
- 14 [3] M. Alvear, M. E. Fortunato, V. Russo, T. Salmi, M. D. Serio, Modelling
15 of transient kinetics in trickle bed reactors: Ethylene oxide production
16 via hydrogen peroxide, Chemical Engineering Science 248 (2022) 117156.
17 doi:<https://doi.org/10.1016/j.ces.2021.117156>.
18 URL [https://www.sciencedirect.com/science/article/](https://www.sciencedirect.com/science/article/pii/S0009250921007211)
19 [pii/S0009250921007211](https://www.sciencedirect.com/science/article/pii/S0009250921007211)
- 20 [4] L. Lu, X. Gao, J.-F. Dietiker, M. Shahnam, W. A. Rogers, Mfix based multi-
21 scale cfd simulations of biomass fast pyrolysis: A review, Chemical En-

- 1 gineering Science 248 (2022) 117131. doi:[https://doi.org/10.](https://doi.org/10.1016/j.ces.2021.117131)
2 1016/j.ces.2021.117131.
3 URL [https://www.sciencedirect.com/science/article/](https://www.sciencedirect.com/science/article/pii/S0009250921006965)
4 pii/S0009250921006965
- 5 [5] M. Ihme, L. Shunn, J. Zhang, Regularization of reaction progress variable
6 for application to flamelet-based combustion models, Journal of Computa-
7 tional Physics 231 (23) (2012) 7715–7721. doi:[https://doi.org/](https://doi.org/10.1016/j.jcp.2012.06.029)
8 10.1016/j.jcp.2012.06.029.
- 9 [6] K. Zdybał, J. C. Sutherland, A. Parente, Optimizing progress variables for
10 ammonia/hydrogen combustion using encoding–decoding networks, Com-
11 bustion and Flame 276 (2025) 114152. doi:[https://doi.org/10.](https://doi.org/10.1016/j.combustflame.2025.114152)
12 1016/j.combustflame.2025.114152.
- 13 [7] H. Pitsch, M. Ihme, An unsteady/flamelet progress variable method for
14 les of nonpremixed turbulent combustion, 43rd AIAA Aerospace Sciences
15 Meeting and Exhibit - Meeting Papers (01 2005). doi:10.2514/6.
16 2005–557.
- 17 [8] K. Bray, P. Domingo, L. Vervisch, Role of the progress variable in mod-
18 els for partially premixed turbulent combustion, Combustion and Flame
19 141 (4) (2005) 431–437. doi:[https://doi.org/10.1016/j.](https://doi.org/10.1016/j.combustflame.2005.01.017)
20 combustflame.2005.01.017.
- 21 [9] L. Zhou, Y. Song, W. Ji, H. Wei, Machine learning for combustion, En-

- 1 ergy and AI 7 (2022) 100128. doi:<https://doi.org/10.1016/j.egyai.2021.100128>.
- 2 egyai.2021.100128.
- 3 [10] O. Owoyele, P. Pal, ChemNODE: A neural ordinary differential equa-
4 tions framework for efficient chemical kinetic solvers, Energy and AI
5 7 (2022) 100118. doi:<https://doi.org/10.1016/j.egyai.2021.100118>.
- 6 2021.100118.
- 7 [11] M. Ihme, W. T. Chung, A. A. Mishra, Combustion machine learning: Prin-
8 ciples, progress and prospects, Progress in Energy and Combustion Science
9 91 (2022) 101010.
- 10 [12] T. Zhang, Y. Yi, Y. Xu, Z. X. Chen, Y. Zhang, W. E, Z.-Q. J. Xu, A
11 multi-scale sampling method for accurate and robust deep neural network
12 to predict combustion chemical kinetics, Combustion and Flame 245 (2022)
13 112319.
- 14 [13] R. Malpica Galassi, P. P. Ciottoli, M. Valorani, H. G. Im, An adaptive
15 time-integration scheme for stiff chemistry based on computational singu-
16 lar perturbation and artificial neural networks, Journal of Computational
17 Physics 451 (2022) 110875. doi:<https://doi.org/10.1016/j.jcp.2021.110875>.
- 18 jcp.2021.110875.
- 19 [14] V. Vijayarangan, H. A. Uranakara, S. Barwey, R. M. Galassi, M. R. Ma-
20 lik, M. Valorani, V. Raman, H. G. Im, A data-driven reduced-order model
21 for stiff chemical kinetics using dynamics-informed training, Energy and AI

- 1 15 (2024) 100325. doi:[https://doi.org/10.1016/j.egyai.](https://doi.org/10.1016/j.egyai.2023.100325)
2 2023.100325.
- 3 [15] Z. Wang, Z. Wu, Towards foundation model for chemical reactor modeling:
4 Meta-learning with physics-informed adaptation, Chemical Engineering Re-
5 search and Design 218 (2025) 839–853. doi:[https://doi.org/10.](https://doi.org/10.1016/j.cherd.2025.05.015)
6 1016/j.cherd.2025.05.015.
7 URL [https://www.sciencedirect.com/science/article/](https://www.sciencedirect.com/science/article/pii/S0263876225002473)
8 pii/S0263876225002473
- 9 [16] P. Zhang, R. Sankaran, Autoencoder neural network for chemically reacting
10 systems, Journal of Machine Learning for Modeling and Computing 3 (4)
11 (2022) 1–28.
- 12 [17] K. Zdybał, A. Parente, J. C. Sutherland, Improving reduced-order models
13 through nonlinear decoding of projection-dependent outputs, Patterns 4 (11)
14 (2023) 100859. doi:[https://doi.org/10.1016/j.patter.](https://doi.org/10.1016/j.patter.2023.100859)
15 2023.100859.
- 16 [18] M. R. Malik, P. Obando Vega, A. Coussement, A. Parente, Combustion mod-
17 eling using principal component analysis: A posteriori validation on sandia
18 flames d, e and f, Proceedings of the Combustion Institute 38 (2) (2021)
19 2635–2643. doi:[https://doi.org/10.1016/j.proci.2020.](https://doi.org/10.1016/j.proci.2020.07.014)
20 07.014.
- 21 [19] A. Parente, J. Sutherland, L. Tognotti, P. Smith, Identification of low-

- 1 dimensional manifolds in turbulent flames, Proceedings of the Combustion
2 tion Institute 32 (1) (2009) 1579–1586. doi:[https://doi.org/10.](https://doi.org/10.1016/j.proci.2008.06.177)
3 [1016/j.proci.2008.06.177](https://doi.org/10.1016/j.proci.2008.06.177).
- 4 [20] R. Ranade, T. Echekki, A framework for data-based turbulent combustion
5 closure: A priori validation, Combustion and Flame 206 (2019)
6 490–505. doi:[https://doi.org/10.1016/j.combustflame.](https://doi.org/10.1016/j.combustflame.2019.05.028)
7 [2019.05.028](https://doi.org/10.1016/j.combustflame.2019.05.028).
- 8 [21] R. Ranade, T. Echekki, A framework for data-based turbulent combustion
9 closure: A posteriori validation, Combustion and Flame 210 (2019)
10 279–291. doi:[https://doi.org/10.1016/j.combustflame.](https://doi.org/10.1016/j.combustflame.2019.08.039)
11 [2019.08.039](https://doi.org/10.1016/j.combustflame.2019.08.039).
- 12 [22] G. D’Alessio, A. Parente, A. Stagni, A. Cuoci, Adaptive chemistry via pre-
13 partitioning of composition space and mechanism reduction, Combustion
14 and Flame 211 (2020) 68–82. doi:[https://doi.org/10.1016/j.](https://doi.org/10.1016/j.combustflame.2019.09.010)
15 [combustflame.2019.09.010](https://doi.org/10.1016/j.combustflame.2019.09.010).
- 16 [23] R. Amaduzzi, G. D’Alessio, P. Pagani, A. Cuoci, R. Malpica Galassi, A. Par-
17 ente, Automated adaptive chemistry for large eddy simulations of turbulent
18 reacting flows, Combustion and Flame 259 (2024) 113136. doi:[https://doi.org/10.1016/j.](https://doi.org/10.1016/j.combustflame.2023.113136)
19 [combustflame.2023.113136](https://doi.org/10.1016/j.combustflame.2023.113136).
- 20 [24] C. J. Lapeyre, A. Misdariis, N. Cazard, D. Veynante, T. Poinso, Training
21 convolutional neural networks to estimate turbulent sub-grid scale reaction

- 1 rates, Combustion and Flame 203 (2019) 255–264. doi:[https://doi.](https://doi.org/10.1016/j.combustflame.2019.02.019)
2 [org/10.1016/j.combustflame.2019.02.019](https://doi.org/10.1016/j.combustflame.2019.02.019).
- 3 [25] R. Roncancio, A. El Gamal, J. P. Gore, Turbulent flame image classifica-
4 tion using convolutional neural networks, Energy and AI 10 (2022) 100193.
5 doi:<https://doi.org/10.1016/j.egyai.2022.100193>.
- 6 [26] L. Castellanos, R. S. M. Freitas, A. Parente, F. Contino, Deep learning dy-
7 namical latencies for the analysis and reduction of combustion chemistry
8 kinetics, Physics of Fluids 35 (10) (2023) 107143.
- 9 [27] F. Aggietti, F. Della Santa, A. Piano, V. Aglietti, Gradinn: Gradient in-
10 formed neural network, Arxiv (Sep 2024).
- 11 [28] M. Ihme, H. Pitsch, LES of a Non-Premixed Flame Using an Extended
12 Flamelet/Progress Variable Model. arXiv:[https://arc.aiaa.org/](https://arc.aiaa.org/doi/pdf/10.2514/6.2005-558)
13 [doi/pdf/10.2514/6.2005-558](https://arc.aiaa.org/doi/pdf/10.2514/6.2005-558), doi:10.2514/6.2005-558.
- 14 [29] Y. Hu, R. Kurose, Nonpremixed and premixed flamelets les of partially pre-
15 mixed spray flames using a two-phase transport equation of progress vari-
16 able, Combustion and Flame 188 (2018) 227–242. doi:[https://doi.](https://doi.org/10.1016/j.combustflame.2017.10.004)
17 [org/10.1016/j.combustflame.2017.10.004](https://doi.org/10.1016/j.combustflame.2017.10.004).
- 18 [30] Z. Ren, C. Xu, T. Lu, M. A. Singer, Dynamic adaptive chemistry with op-
19 erator splitting schemes for reactive flow simulations, Journal of Computa-
20 tional Physics 263 (2014) 19–36. doi:[https://doi.org/10.1016/](https://doi.org/10.1016/j.jcp.2014.01.016)
21 [j.jcp.2014.01.016](https://doi.org/10.1016/j.jcp.2014.01.016).

- 1 [31] A. Vasavan, P. de Goey, J. van Oijen, A novel method to automate fgm
2 progress variable with application to igniting combustion systems, *Com-*
3 *bustion Theory and Modelling* 24 (2) (2020) 221–244. doi:10.1080/
4 13647830.2019.1673902.
- 5 [32] Y.-S. Niu, L. Vervisch, P. D. Tao, An optimization-based approach to de-
6 tailed chemistry tabulation: Automated progress variable definition, *Com-*
7 *bustion and Flame* 160 (4) (2013) 776–785. doi:https://doi.org/
8 10.1016/j.combustflame.2012.11.015.
- 9 [33] K.-K. Kuo, R. Acharya, *Fundamentals of Turbulent and Multiphase Com-*
10 *bustion*, 2012. doi:10.1002/9781118107683.
- 11 [34] M. I. Radaideh, S. Rigopoulos, D. A. Goussis, Characteristic time scale as
12 optimal input in machine learning algorithms: Homogeneous autoignition,
13 *Energy and AI* 14 (2023) 100273. doi:https://doi.org/10.1016/
14 j.egyai.2023.100273.
- 15 [35] S. Lam, D. Goussis, Understanding complex chemical kinetics with com-
16 putational singular perturbation, *Symposium (International) on Combus-*
17 *tion* 22 (1) (1989) 931–941. doi:https://doi.org/10.1016/
18 S0082-0784(89)80102-X.
- 19 [36] A. Kazakov, M. Chaos, Z. Zhao, F. L. Dryer, Computational singular per-
20 turbation analysis of two-stage ignition of large hydrocarbons, *The Journal*

- 1 of Physical Chemistry A 110 (21) (2006) 7003–7009. doi:10.1021/
2 jp057224u.
- 3 [37] M. Valorani, S. Paolucci, E. Martelli, T. Grenga, P. P. Ciottoli, Dynamical
4 system analysis of ignition phenomena using the tangential stretching rate
5 concept, *Combustion and Flame* 162 (8) (2015) 2963–2990. doi:https:
6 //doi.org/10.1016/j.combustflame.2015.05.015.
- 7 [38] G. E. Hinton, R. R. Salakhutdinov, Reducing the dimensionality of data with
8 neural networks, *Science* 313 (5786) (2006) 504–507. arXiv:https:
9 //www.science.org/doi/pdf/10.1126/science.1127647,
10 doi:10.1126/science.1127647.
- 11 URL [https://www.science.org/doi/abs/10.1126/](https://www.science.org/doi/abs/10.1126/science.1127647)
12 [science.1127647](https://www.science.org/doi/abs/10.1126/science.1127647)
- 13 [39] T. Grassi, F. Nauman, J. P. Ramsey, S. Bovino, G. Picogna, B. Ercolano,
14 Reducing the complexity of chemical networks via interpretable autoen-
15 coders, *Astronomy and Astrophysics* 668 (2022) A139. doi:10.1051/
16 0004-6361/202039956.
- 17 URL <http://dx.doi.org/10.1051/0004-6361/202039956>
- 18 [40] C. Croux, C. Dehon, Influence Functions of the Spearman and Kendall Cor-
19 relation Measures, Tilburg University, Center for Economic Research, Dis-
20 cussion Paper 19 (Jan. 2010).
- 21 [41] T. Bui-Thanh, M. Damodarany, Proper orthogonal decomposition extensions

- 1 for parametric applications in transonic aerodynamics 2003 (06 2003). doi :
2 10.2514/6.2003-4213.
- 3 [42] S. L. Brunton, J. N. Kutz, Data-Driven Science and Engineering: Machine
4 Learning, Dynamical Systems, and Control, 2nd Edition, Cambridge Uni-
5 versity Press, 2022.
- 6 [43] C. Bowman, S. University, G. R. Institute, GRI-Mech 2.11, an Optimized
7 Detailed Chemical Reaction Mechanism for Methane Combustion and No
8 Formation and Reburning: Topical Report (February 1, 1997), Gas Research
9 Institute, 1997.
- 10 [44] D. G. Goodwin, R. L. Speth, H. K. Moffat, B. W. Weber, Cantera: An
11 Object-oriented Software Toolkit for Chemical Kinetics, Thermodynamics,
12 and Transport Processes (2021).
- 13 [45] S. D. Cohen, A. C. Hindmarsh, P. F. Dubois, Cvode, a stiff/nonstiff
14 ode solver in c, Computer in Physics 10 (2) (1996) 138–143. arXiv:
15 [https://pubs.aip.org/aip/cip/article-pdf/10/2/138/](https://pubs.aip.org/aip/cip/article-pdf/10/2/138/11945852/138_1_online.pdf)
16 11945852/138_1_online.pdf, doi:10.1063/1.4822377.
17 URL <https://doi.org/10.1063/1.4822377>
- 18 [46] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel,
19 M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos,
20 D. Cournapeau, M. Brucher, M. Perrot, E. Duchesnay, Scikit-learn: Machine

- 1 learning in Python, Journal of Machine Learning Research 12 (2011) 2825–
2 2830.
- 3 [47] D. L. Davies, D. W. Bouldin, A cluster separation measure, IEEE Trans-
4 actions on Pattern Analysis and Machine Intelligence PAMI-1 (2) (1979)
5 224–227. doi:10.1109/TPAMI.1979.4766909.
- 6 [48] T. Caliński, J. Harabasz, A dendrite method for cluster analysis, Com-
7 munications in Statistics 3 (1) (1974) 1–27. arXiv:<https://www.tandfonline.com/doi/pdf/10.1080/03610927408827101>,
8 doi:10.1080/03610927408827101.
- 9 [49] M. Abadi, A. Agarwal, P. Barham, E. Brevdo, Z. Chen, C. Citro, G. S. Cor-
10 rado, A. Davis, J. Dean, M. Devin, S. Ghemawat, I. Goodfellow, A. Harp,
11 G. Irving, M. Isard, Y. Jia, R. Jozefowicz, L. Kaiser, M. Kudlur, J. Lev-
12 enberg, D. Mané, R. Monga, S. Moore, D. Murray, C. Olah, M. Schuster,
13 J. Shlens, B. Steiner, I. Sutskever, K. Talwar, P. Tucker, V. Vanhoucke, V. Va-
14 sudevan, F. Viégas, O. Vinyals, P. Warden, M. Wattenberg, M. Wicke, Y. Yu,
15 X. Zheng, Tensorflow: Large-scale machine learning on heterogeneous dis-
16 tributed systems (2016). arXiv:1603.04467.
- 17 [50] T. O’Malley, E. Bursztein, J. Long, F. Chollet, H. Jin, L. Invernizzi, others,
18 KerasTuner (2019).
- 19 [51] D. Kingma, J. Ba, Adam: A Method for Stochastic Optimization, Interna-
20 tional Conference on Learning Representations (Dec. 2014).
- 21

- 1 [52] T. Zhang, L. Shou, L. Chen, H. Wang, Y. Long, Z. Wang, J. Chen, Detailed
2 and reduced chemical kinetic model of ch4/air mixtures combustion based
3 on high-precision first-principles molecular dynamics simulation, Chemical
4 Engineering Science 281 (2023) 119159. doi:[https://doi.org/10.](https://doi.org/10.1016/j.ces.2023.119159)
5 1016/j.ces.2023.119159.
- 6 [53] C. K. WESTBROOK, F. L. DRYER, Simplified reaction mechanisms for the
7 oxidation of hydrocarbon fuels in flames, Combustion Science and Technol-
8 ogy 27 (1-2) (1981) 31–43. doi:10.1080/00102208108946970.
- 9 [54] R. Atkinson, D. L. Baulch, R. A. Cox, J. N. Crowley, R. F. Hampson, R. G.
10 Hynes, M. E. Jenkin, M. J. Rossi, J. Troe, Evaluated kinetic and photochem-
11 ical data for atmospheric chemistry: Volume i - gas phase reactions of o_x,
12 ho_x, no_x and so_x species, Atmospheric Chemistry and Physics 4 (6) (2004)
13 1461–1738. doi:10.5194/acp-4-1461-2004.
- 14 [55] C. T. Bowman, Kinetics of pollutant formation and destruction in com-
15 bustion, Progress in Energy and Combustion Science 1 (1) (1975) 33–45.
16 doi:[https://doi.org/10.1016/0360-1285\(75\)90005-2](https://doi.org/10.1016/0360-1285(75)90005-2).
- 17 [56] M. A. Hafeez, A. Procacci, A. Coussement, A. Parente, Constrained
18 reduced-order modeling using bounded gaussian processes for physically
19 consistent reacting flow predictions, Energy and AI 21 (2025) 100554.
20 doi:<https://doi.org/10.1016/j.egyai.2025.100554>.
21 URL [https://www.sciencedirect.com/science/article/](https://www.sciencedirect.com/science/article/pii/S2666546825000862)
22 [pii/S2666546825000862](https://www.sciencedirect.com/science/article/pii/S2666546825000862)

- 1 [57] T. Hastie, R. Tibshirani, J. Friedman, The elements of statistical learning:
2 data mining, inference and prediction, 2nd Edition, Springer, 2009.
- 3 [58] J. Wang, Y. Chen, Introduction to Transfer Learning : Algorithms and Prac-
4 tice, 1st Edition, Machine Learning: Foundations, Methodologies, and Ap-
5 plications, Springer Nature Singapore, Singapore, 2023, publication Title:
6 Introduction to Transfer Learning : Algorithms and Practice.
- 7 [59] A. Coussement, O. Gicquel, A. Parente, Kernel density weighted princi-
8 pal component analysis of combustion processes, Combustion and Flame
9 159 (9) (2012) 2844–2855. doi:[https://doi.org/10.1016/j.](https://doi.org/10.1016/j.combustflame.2012.04.004)
10 [combustflame.2012.04.004](https://doi.org/10.1016/j.combustflame.2012.04.004).
- 11 [60] A. Kumar, T. Echehki, Combustion chemistry acceleration with deep-
12 onets, Fuel 365 (2024) 131212. doi:[https://doi.org/10.1016/](https://doi.org/10.1016/j.fuel.2024.131212)
13 [j.fuel.2024.131212](https://doi.org/10.1016/j.fuel.2024.131212).