Unsupervised data analysis of Direct Numerical Simulation of a Turbulent Flame via Local Principal Component Analysis and Procrustes Analysis

Giuseppe D'Alessio[a,b,*], Antonio Attili[c], Alberto Cuoci[b], Heinz Pitsch[c], Alessandro Parente[a,d]

[a] *Université Libre de Bruxelles, Aero-Thermo-Mechanics Laboratory, Bruxelles, Belgium*
[b] *CRECK Modeling Lab, Department of Chemistry, Materials and Chemical Engineering, Politecnico di Milano, Piazza Leonardo da Vinci 32, 20131 Milano, Italy*
[c] *Institute for Combustion Technology, RWTH Aachen University, 52056 Aachen, Germany*
[d] *Université Libre de Bruxelles and Vrije Universiteit Brussel, Combustion and Robust Optimization Group (BURN), Brussels, Belgium*

Abstract

Direct Numerical Simulations (DNS) of reacting flows provide high-fidelity data for combustion model reduction and validation, although their interpretation is not always straightforward because of the massive amount of information and the data high-dimensionality. Several data mining and machine learning techniques have been tested by the combustion community to make the analysis easier, but the feature physical interpretation and the need to set the algorithms' hyperparameters often limit their potential and extendibility, as they are entrusted to the sole user experience. In this work, a completely unsupervised algorithm for data analysis is investigated on a data-set obtained from a temporally-evolving DNS simulation of a reacting n-heptane jet in air. The proposed algorithm combines the Local Principal Component Analysis (LPCA) clustering algorithm with a variables selection algorithm via dimensionality reduction and Procrustes Analysis. Unlike other data-analysis algorithms, it requires null or limited user expertise as all of its steps are unsupervised and solely entrusted to mathematical objective functions, without any hyperparameter tuning step required.

Keywords:
Data Analysis, Local variables selection, Principal Component Analysis, Direct Numerical Simulation, Turbulent flame

*Corresponding author:
Email address: giuseppe.dalessio@ulb.ac.be (Giuseppe D'Alessio)

1. Introduction

Combustion data obtained from high-fidelity numerical simulations such as Direct Numerical Simulations (DNS) are routinely used for model development and validation, as well as for the understanding of chemical and physical processes. In any case, the first step is always the analysis of the massive amount of information that large-scale simulations produce, as they are usually characterized by a large number of statistical observations and several variables. Many data-driven approaches are available in literature and have been tested on combustion data, such as linear and non-linear dimensionality reduction techniques, i.e. Principal Component Analysis (PCA), Autoencoders (AE), Kernel Principal Component Analysis (KPCA), Isomap and Dynamic Mode Decomposition (DMD) [1–4, 6, 8], as well as techniques for high-dimensionality space exploration and visualization, such as Self Organizing Maps (SOMs) and t-SNE [9–12]. Although the effectiveness of these techniques is not questioned, as they have all proved to be effective in extracting information from data, their common limitation is related to the physical interpretation of the features, which is, in all the mentioned cases, not driven by objective criteria, but entrusted to the sole user experience, constituting a limitation to the algorithm analysis potential and extendibility. For some of these algorithms, such as Autoencoders and t-SNE, good performances can be obtained only after an accurate tuning of the hyperparameters, for which a thorough sensitivity analysis or a significant user expertise are required. For other algorithms, the applicability to combustion is limited because of their intrinsic linearity (PCA and DMD), or because of their CPU-intensive nature (KPCA and Isomap).

In this work, a local unsupervised algorithm for data analysis, which combines the effectiveness of the Local Principal Component Analysis (LPCA) algorithm input-space partitioning [13–15] and an automatic variables selection criterion via dimensionality reduction and Procrustes Analysis [1, 16], is tested on a data-set representing a 2D slice of a 3D temporally-evolving DNS simulation of a reacting n-heptane jet [22–25]. The algorithm's performances were assessed comparing the selected main local principal variables (LPVs) with the ones obtained by means of another well-known method used for data analysis and feature extraction, which exploits the rotation of the local principal compo-

nents with the Varimax criterion [32]. The main advantage of the proposed local analysis is that both the partitioning and the LPV selection steps are accomplished according to mathematical criteria, not requiring any hyperparameter tuning, and no dependence from user expertise.

2. Theory

2.1. Variable selection via Principal Component Analysis and Procrustes Analysis

The PCA is a dimensionality reduction technique based on the eigenvalue-decomposition of a covariance matrix [1]. Given a matrix $\mathbf{X} \in \mathbb{R}^p$, consisting of n statistical observations of p variables, it is possible to compute the associated covariance matrix according to Equation 1, which can be then decomposed by means of Equation 2:

$$\mathbf{S} = \frac{1}{n-1}\mathbf{X}^T\mathbf{X}, \tag{1}$$

$$\mathbf{S} = \mathbf{A}\mathbf{L}\mathbf{A}^T. \tag{2}$$

The matrix $\mathbf{A}$ is an orthonormal basis of eigenvectors, the Principal Components (PCs), while $\mathbf{L}$ is a diagonal matrix of eigenvalues. The PCs are a linear combination of the original variables, and the dimensionality reduction is possible considering a subset of q eigenvectors, with $q < p$, associated to the q most powerful eigenvalues, such that the information loss is minimized.

In many applications, rather than reducing the dimensionality considering a new set of coordinates which are linear combination of the original ones, the main interest is to achieve a dimensionality reduction selecting a subset of m variables from the original set of p variables. One of the possible ways to accomplish this task is to couple the PCA dimensionality reduction with a Procrustes Analysis [1, 16]. To do that, PCA is firstly applied to the full data matrix $\mathbf{X} \in \mathbb{R}^p$, and a score matrix $\mathbf{Z} \in \mathbb{R}^q$ is obtained projecting the matrix $\mathbf{X}$ on the q-dimensional manifold spanned by the retained PCs:

$$\mathbf{Z} = \mathbf{X}\mathbf{A}. \tag{3}$$

After that, a subset consisting of m variables, with $q < m < p$, can be selected from the original matrix, thus obtaining the reduced matrix $\tilde{\mathbf{X}} \in \mathbb{R}^m$. At this point, PCA is

applied to $\tilde{\mathbf{X}}$, and a scores matrix $\tilde{\mathbf{Z}} \in \mathbb{R}^q$ is obtained also in this case. If the choice of the m variables is done correctly, the discrepancies between the two scores matrices $\mathbf{Z}$ and $\tilde{\mathbf{Z}}$ are minimal, while there are significant differences otherwise [16]. A Procrustes Analysis is thus carried out in order to quantitatively measure the similarity between the two matrices, calculating the sum of the squared differences between the points of $\mathbf{Z}$ and $\tilde{\mathbf{Z}}$. It consists in the computation of the M^2 coefficient:

$$M^2 = Tr(\mathbf{Z}\mathbf{Z}' + \tilde{\mathbf{Z}}\tilde{\mathbf{Z}}' - 2\mathbf{\Sigma}), \tag{4}$$

where $\mathbf{\Sigma}$ is the matrix of the singular values obtained from the decomposition of the square matrix $\tilde{\mathbf{Z}}'\mathbf{Z}$:

$$\tilde{\mathbf{Z}}'\mathbf{Z} = \mathbf{U}\mathbf{\Sigma}\mathbf{V}'. \tag{5}$$

By means of the minimization of M^2 as objective function, it is possible to build an iterative algorithm to select, in a totally unsupervised fashion, the best subset of m variables from the original set of p variables, as described in [16]:

1. The dimensionality of m is initially set equal to p.
2. Each variable is deleted from the matrix $\mathbf{X}$, obtaining p $\tilde{\mathbf{X}}$ matrices. The corresponding scores matrices $\tilde{\mathbf{Z}}$ are computed by means of PCA. For each of them, a Procrustes Analysis is performed as in Equation 4 with respect to the scores of the original matrix $\mathbf{X}$, and the corresponding M^2 coefficient is computed.
3. The variable which, once excluded, leads to the smallest M^2 coefficient is deleted from the $\tilde{\mathbf{X}}$ matrix.
4. Steps 2 and 3 are repeated until m variables are left, thus obtaining the reduced $\tilde{\mathbf{X}}$ $\in \mathbb{R}^m$ matrix.

2.2. Unsupervised data analysis via local principal variables

The coupling between PCA and the Procrustes Analysis, proposed by Krzanowski to select the main variables to preserve the multivariate data structure [16], can be easily extended to a local version by means of the LPCA clustering. The latter is an unsupervised algorithm to partition statistical observations in a high-dimensional space in clusters ($\mathbf{C}_i$, with $i \in [1, ..., k]$) via vector quantization (VQ), and after that the dimensionality reduction task is locally accomplished. This method has already been successfully applied

in combustion for clustering purposes [21] as well as for model reduction [14, 15]. The objective function for the unsupervised space partitioning is the PCA reconstruction error (ϵ_r), which is defined as:

$$\epsilon_r = ||\mathbf{x} - \tilde{\mathbf{x}}||, \tag{6}$$

where the vectors $\mathbf{x}$ and $\tilde{\mathbf{x}}$ in Equation 6 represent the original and the reconstructed (from the reduced manifold) vectors, respectively. If data are partitioned in k clusters, and in each of them PCA is performed, it is possible to find k reduced basis of eigenvectors (LPCs) $\mathbf{A}^{(j)} \in \mathbb{R}^q$, with $j \in [1, ..., k]$. Thus, for each observation $\mathbf{x}$ of the data-set $\mathbf{X} \in \mathbb{R}^p$ it is possible to iteratively compute k reconstruction errors and assign it to a cluster $\bar{k}$, such that:

$$\bar{k} \mid \epsilon_{r,\bar{k}} = \min_{j=1,...k} \epsilon_{r,j}, \tag{7}$$

until the error variation for the reconstruction of the full data matrix $\mathbf{X}$ is below a fixed threshold. Considering k local sets of PCs ($\mathbf{A}^{(j)} \in \mathbb{R}^q$, with $j \in [1, ..., k]$), the errors arising from the dimensionality reduction are lowered with respect to the global PCA. The local method is piecewise-linear and not globally linear, thus being effective also for non-linear applications. Moreover, the possibility to select *locally* relevant variables can be more attractive from both data analysis and model development perspective. Locally optimized combustion reduced models have already proved to have several advantages with respect to global reduced models [21], as subsets of variables which are locally more coherent with the physics can be extracted from each group.

The algorithm has the following steps:

1. *Partitioning of the input space in clusters:* the thermochemical space is partitioned in k clusters via minimization of the reconstruction error.
2. *LPCs and local scores computation:* in each cluster $\mathbf{C}_i$ ($i \in [1, ..., k]$) found in the partitioning step, a local set of LPCs $\mathbf{A}^{(i)} \in \mathbb{R}^q$ is computed, and the corresponding local scores matrices $\mathbf{Z}_i$ are computed by projection of the clusters' points on the local reduced manifold.
3. *Local variables selection:* The variables needed to preserve the local multivariate structure are retained by means of the Krzanowski algorithm [16].

3. Case description

The data chosen to test the proposed algorithm were obtained from a 2D slice of a 3D temporally evolving DNS simulation of a n-heptane jet [22–25]. The fuel jet is nitrogen-diluted (85% in volume) at 400K, arranged in a coflow configuration with the oxidizer stream (air) at 800K. The turbulent jet is initialized with a Reynolds' number equal to 15,000, and a layer at stoichiometric composition is inserted in the region of smooth transition between the fuel and the oxidizer. Both the gas phase hydrodynamics and combustion were modeled using a reactive unsteady Navier-Stokes equation formulation within the low Mach number limit [26]. For the resolution of the gas velocity field, as well as for the reactive scalar fields, a finite-differences scheme was chosen [27], while the advection-reaction equations for soot moments were solved by means of a Lagrangian particle method [28, 29]. Open boundary conditions were prescribed in the normal direction to the flame sheet in order to have a mass outflow for the combustion products, while periodic boundary conditions were imposed in the other two directions. The adopted kinetic mechanism for the n-heptane oxidation was reduced to 47 species and 290 reactions [31] from the detailed one developed by Blanquart et al. [30] .

The 2D slice of the simulation considered for the analysis consisted of 1,048,576 grid points, each of them characterized by a thermochemical vector $\boldsymbol{\phi}$ of temperature and 47 chemical species mass fractions. The data were organized as a matrix whose dimensions were $n \times p$, accounting for 1,048,576 observations of the 48 thermochemical variables.

4. Results

The local principal variables algorithm was applied to the data described in Section 3, with $k = 16$. In each cluster, the variables which were able to preserve the local multivariate structure (LPVs) were chosen according to a Procrustes Analysis applied to the local scores matrices. In Figure 1, the results obtained from the n-heptane jet clustering via LPCA are shown.

In each cluster, a range from 4 to 9 chemical species was retained, according to the local manifold dimensionality, within the original 47 species implemented in the chemical mechanism. By means of this variable selection process, it was possible to easily interpret, from a physical point of view, the results obtained from the clustering process, as many

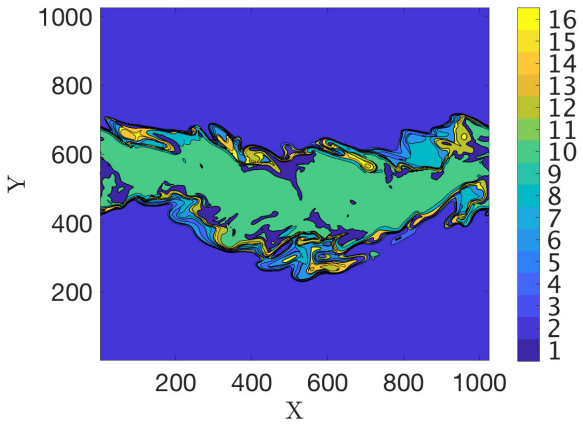

Figure 1: LPCA unsupervised partitioning of the selected 2D slice of the 3D DNS simulation with 16 clusters.

subset of variables resulted to be chemically coherent. For example, a subset containing 4 variables (oxygen radical, hydroxyl radical, hydrogen radical, hydroperoxy radical) was identified in cluster number 2. All of these variables are involved in the oxygen branching reactions and are H-atom abstractors, a key step in the PAH formation. In several clusters, the selected LPVs were the ones involved in the soot formation as they consisted of mainly PAHs, such as in cluster number 5, 6 and 9. In Table 1, the LPVs selected by the algorithm in each cluster, according to the Procrustes Analysis, are reported.

A first, qualitative, assessment of the data analysis algorithm performances can be done comparing the maps of the local principal variables with the cluster shapes. In Figure 2, the phenyl radical ($A1^-$) concentration map is compared with the shape of cluster number 9 (colored in yellow), where this species results to be a LPV. The maximum phenyl radical concentration values and gradients are placed in correspondence of the considered cluster, meaning that a correct variable was identified by the algorithm. Since a qualitative comparison by means of the contours shapes alone cannot be considered to be robust enough to evaluate the algorithm's performances in terms of data analysis, a quantitative assessment was carried out.

An assessment of the data analysis algorithm was done carrying out a comparison between the extracted LPVs and the variables which were considered important by another data analysis algorithm. The LPVs were compared with the variables having the highest weights on the LPCs, when rotated with the Varimax criterion. When PCA or

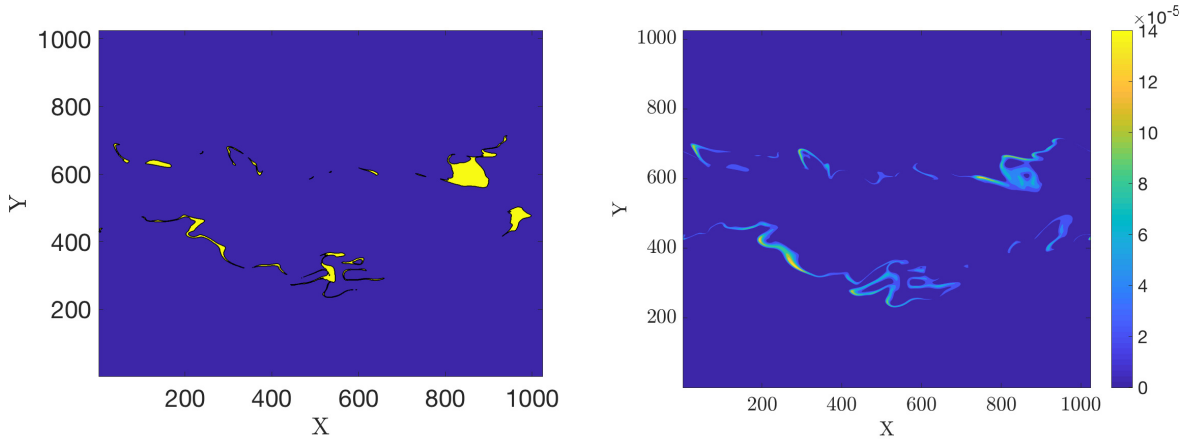

Figure 2: (a). Cluster number 9 (in yellow) identified by means of the LPCA unsupervised partitioning algorithm applied to the DNS data, with $k = 16$; (b). phenyl radical ($A1^-$) map of concentration for the selected 2D slice of the 3D DNS simulation.

LPCA are performed for data analysis tasks, the weights on the PCs must be visually inspected and interpreted, but it can easily happen that large weights are distributed on the eigenvectors over several variables, thus making impossible to associate the PC to a particular variable, nor a physical or chemical process. By means of rotation, instead, the PCs tend to align with only one or few variables, making their physical interpretation easier, as observed in [32]. A coefficient of participation ψ can be defined to represent the fraction of the LPVs having also the largest weight on the rotated LPCs, thus defined as the ratio between the number of LPVs found with largest weight on a rotated LPC in the considered cluster, and the total number of LPVs in that cluster:

$$\psi = \frac{N_{LPVs \in LPCs}}{N_{LPVs,tot}} \tag{8}$$

This coefficient can take values between zero and one, being equal to zero if the variables extracted by the two algorithms are completely different, and equal to one otherwise. Analyzing the ψ coefficients reported in Table 1 it is clear that, except for clusters number 8 and 10, all the PVs were found on the rotated LPCs. In particular, in clusters number 2, 4 and 12, all the selected LPVs were found to be important also by means of the rotation of the LPCs. Obtaining similar results by means of the two data analysis techniques is particularly relevant, as the analysis with the proposed local principal variables algorithm was achieved in an unsupervised fashion, without any visual inspection of the weights to

be required. This is a considerable strength of the proposed algorithm, as it is possible to analyze massive data also using many clusters, a task which would result to be unfeasible if the visual inspection of the first q PCs in each cluster would be required.

5. Conclusions

In this work, an algorithm for local unsupervised data analysis was proposed and tested on a massive dataset obtained from a DNS simulation of a n-heptane reacting jet. The method consists of two steps. The first one is the data-set partitioning in different clusters, accomplished via the LPCA algorithm. After that, in each cluster the main variables are selected by means of an iterative algorithm for variables selection employing a Procrustes Analysis.

A quantitative assessment of the algorithm' performances was carried out comparing the variables selected by means of the proposed algorithm with the ones selected by the rotation of the local principal components, and a satisfactory agreement was observed in all the clusters between the variables selected by the two algorithms. This result is particularly relevant, as it paves the way to the possibility to use a completely unsupervised tool to analyze the data, without any visual inspection nor interpretation of the weights.

The proposed algorithm for the local data analysis can constitute a functional tool aiding for the development and the validation of local reduced order models from DNS data. In fact, the formulation of local reduced order models has already shown to have several advantages over the global one, for example in the context of adaptive-chemistry simulations and the development of digital twins.

Acknowledgments

The first author acknowledges the support of the Fonds National de la Recherche Scientifique (FRS-FNRS) through a FRIA fellowship. A.A. and H.P. acknowledge funding from the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation program under grant agreement No 695747. A.P. acknowledges funding from the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation program, grant agreement No 714605.

References

[1] I. Jolliffe, Principal component analysis, Springer, 2011.

[2] J. C. Sutherland, A. Parente, Proceedings of the Combustion Institute 32 (2009) 1563–1570.

[3] M. Sakurada, T. Yairi, in: Proceedings of the MLSDA 2014 2nd Workshop on Machine Learning for Sensory Data Analysis, ACM, p. 4.

[4] H. Mirgolbabaei, T. Echekki, N. Smaoui, international journal of hydrogen energy 39 (2014) 4622–4633.

[5] H. Mirgolbabaei, T. Echekki, Combustion and Flame 161 (2014) 118–126.

[6] G. Bansal, A. A. Mascarenhas, J. H. Chen, Identification of Intrinsic Low Dimensional Manifolds in Turbulent Combustion using an Isomap based technique., Technical Report, Sandia National Lab.(SNL-CA), Livermore, CA (United States), 2011.

[7] J. H. Chen, H. Kolla, A. A. Mascarenhas, H. Yu, V. Pascucci, V. Krishnamoorthy, S. Liu, A. Gyulassy, K.-L. Ma, A. Tikhonova, et al., Data analysis and visualization of petascale combustion science simulation data., Technical Report, Sandia National Lab.(SNL-CA), Livermore, CA (United States), 2011.

[8] T. Grenga, J. F. MacArt, M. E. Mueller, Combustion Theory and Modelling 22 (2018) 795–811.

[9] M. Liukkonen, T. Hiltunen, E. Hälikkä, Y. Hiltunen, Environmental Modelling & Software 26 (2011) 605–614.

[10] J. Blasco, N. Fueyo, C. Dopazo, J. Chen, Combustion Theory and Modelling 4 (2000) 61–76.

[11] E. Fooladgar, C. Duwig, in: Direct and Large-Eddy Simulation XI, Springer, 2019, pp. 245–251.

[12] E. Fooladgar, C. Duwig, Combustion and Flame 191 (2018) 226–238.

[13] N. Kambhatla, T. K. Leen, Neural computation 9 (1997) 1493–1516.

[14] A. Parente, J. Sutherland, B. Dally, L. Tognotti, P. Smith, Proceedings of the Combustion Institute 33 (2011) 3333–3341.

[15] A. Parente, J. C. Sutherland, L. Tognotti, P. J. Smith, Proceedings of the Combustion Institute 32 (2009) 1579–1586.

[16] W. J. Krzanowski, Journal of the Royal Statistical Society: Series C (Applied Statistics) 36 (1987) 22–33.

[17] A. Coussement, O. Gicquel, A. Parente, Proceedings of the Combustion Institute 34 (2013) 1117–1123.

[18] A. Coussement, B. J. Isaac, O. Gicquel, A. Parente, Combustion and flame 168 (2016) 83–97.

[19] B. J. Isaac, A. Coussement, O. Gicquel, P. J. Smith, A. Parente, Combustion and flame 161 (2014) 2785–2800.

[20] B. J. Isaac, A. Parente, C. Galletti, J. N. Thornock, P. J. Smith, L. Tognotti, Energy & fuels 27 (2013) 2255–2265.

[21] G. D?Alessio, A. Parente, A. Stagni, A. Cuoci, Combustion and Flame 211 (2020) 68–82.

[22] A. Attili, F. Bisetti, M. E. Mueller, H. Pitsch, Combustion and flame 161 (2014) 1849–1865.

[23] A. Attili, F. Bisetti, M. E. Mueller, H. Pitsch, Combustion and Flame 166 (2016) 192–202.

[24] A. Attili, F. Bisetti, M. E. Mueller, H. Pitsch, Proceedings of the Combustion Institute 35 (2015) 1215–1223.

[25] A. Attili, F. Bisetti, Computers & Fluids 84 (2013) 164–175.

[26] A. Tomboulides, J. Lee, S. Orszag, Journal of Scientific Computing 12 (1997) 139–167.

[27] O. Desjardins, G. Blanquart, G. Balarac, H. Pitsch, Journal of Computational Physics 227 (2008) 7125–7159.

[28] G.-H. Cottet, P. D. Koumoutsakos, P. D. Koumoutsakos, et al., Vortex methods: theory and practice, Cambridge university press, 2000.

[29] P. Koumoutsakos, Annu. Rev. Fluid Mech. 37 (2005) 457–487.

[30] G. Blanquart, P. Pepiot-Desjardins, H. Pitsch, Combustion and Flame 156 (2009) 588–607.

[31] F. Bisetti, G. Blanquart, M. E. Mueller, H. Pitsch, Combustion and Flame 159 (2012) 317–335.

[32] A. Bellemans, G. Aversano, A. Coussement, A. Parente, Computers & chemical engineering 115 (2018) 504–514.

Table 1: Number of the cluster with the corresponding selected LPVs and coefficient of participation (ψ).

k	$LPVs$	ψ
1	CH_2O, CH_4, C_3H_6, C_4H_8, C_5H_6, A_1CHO	0.67
2	O, H, OH, HO_2	1
3	CH_2, HCO, C_2H_3, C_2H, HCCO, A_1^-, A_1CH_2	0.85
4	CH_2, O, CH, HCO	1
5	CH_3, A_2, A_1CH_2, $A_1C_2H^*$, A_2^-, A_1C_2H	0.67
6	CH_2, C_2H, A_1^-, $A_1C_2H^*$	0.75
7	A-C_3H_4, A_1, C_5H_6, A_2, $A_1C_2H_2$, A_2^-, A_1C_2H	0.71
8	CH_2O, C_2H_5, C_4H_8, C_5H_{11}, A_1CHO, C_7H_{15}	0.50
9	A_1^-, A_2^-, A_2, A_1CH_2, A_1C_2H	0.8
10	C_2H_6, C_4H_8, C_5H_{10}, A_1CHO	0.5
11	HO_2, HCO, CH_2O, CH_3, n-C_3H_7, C_7H_{15}	0.67
12	CH, HCO, C_2H, HCCO	1
13	CH_4, A-C_3H_5, C_4H_8, C_5H_6, C_5H_{11}, $A_1C_2H_2$, A_1CHO, C_7H_{15}	0.62
14	CH_2O, CH_3, C_2H_3, A-C_3H_5, n-C_3H_7, $A_1C_2H_2$, A_1CH_2	0.85
15	CH_2O, A-C_3H_5, n-C_3H_7, C_5H_{11}, $A_1C_2H_2$, C_7H_{15}	0.67
16	CH_2O, C_2H_3, A-C_3H_5, n-C_3H_7, A-C_3H_4, C_5H_6, $A_1C_2H_2$, A_1C_2H	0.75