Principal component analysis acceleration of rovibrational coarse-grain models for internal energy excitation and dissociation

Aurélie Bellemans* and Alessandro Parente†
Service d'Aéro-Thermo-Mécanique, Université libre de Bruxelles.
50 avenue F.D. Roosevelt, 1050 Bruxelles, Belgium.

Thierry Magin‡
Aeronautics and Aerospace Department, von Karman Institute for Fluid Dynamics
72 Chaussée de Waterloo, 1640 Rhode-Saint-Genèse, Belgium.
(Dated: December 18, 2017)

The present work introduces a novel approach for obtaining reduced chemistry representations of large kinetic mechanisms in strong non-equilibrium conditions. The need for accurate reduced-order models arises from the need to compress large *ab initio* quantum chemistry databases for their use in fluid codes. The method presented in this paper builds on the known physics-based strategies and proposes a new approach based on the combination of a simple coarse grain model with Principal Component Analysis (PCA). The internal energy levels of the chemical species are regrouped in distinct energy groups with a uniform lumping technique. Following the philosophy of machine learning, PCA is applied on the training data provided by the coarse grain model to find an optimally reduced representation of the full kinetic mechanism. Compared to recently published complex lumping strategies, no expert judgment is required before the application of PCA. In this work we will demonstrate the benefits of the combined approach, stressing its simplicity, reliability and accuracy. The technique is demonstrated by reducing the complex quantum $N_2(^1\Sigma_g^+)$-$N(^4S_u)$ database for studying molecular dissociation and excitation in strong non-equilibrium. Starting from detailed kinetics, an accurate reduced model is developed and used to study non-equilibrium properties of the $N_2(^1\Sigma_g^+)$-$N(^4S_u)$ system in shock relaxation simulations.

I. INTRODUCTION

Detailed kinetic models are a prerequisite to conduct accurate predictive simulations of non-equilibrium plasma in various high-tech applications [1]. Some examples are the prediction of the material recession of ablative heat shields during atmospheric re-entry [2, 3], the study of non-equilibrium plasma to model the ignition time in plasma assisted combustion applications [4–6], or the design of ion thrusters for electric propulsion [7]. All these applications unite the chemistry and engineering communities to couple computational chemistry to computational fluid dynamics (CFD) in order to conduct high-fidelity predictions.

Two main categories of models have been developed to describe the physico-chemical state of non-equilibrium plasma: collisional and multi-temperature models. Multi-temperature models (MT) are based on experimental data and distribute the energy levels of a species according to their most probable distribution at equilibrium, the Maxwell-Boltzmann distribution. All the populations follow an equilibrium distribution at their own temperature, which can be the translational, vibrational, rotational or electronic temperature depending on the excited degrees of freedom of the species. Although these models are simple and computationally efficient, they are only valid for the equilibrium conditions determined by the experimental setup. For atmospheric entry flows, MT models have been investigated in detail by Park for Earth and the Martian atmosphere [13, 14]. He showed their applicability is limited as they are only correct for Local Thermodynamic Equilibrium (LTE) assumptions where the free-stream velocity remains low and the pressure high [15]. In order to extend the validity of the physico-chemical models and improve their accuracy, *ab initio* collisional or State-To-State (STS) models have been developed. Collisional models require a deep knowledge of the internal structure of the system. Accurate kinetic data must be collected for each excited level under the form of ab initio cross-sections or rate coefficients for all elementary processes which are obtained after complex quantum calculations [10–12]. In STS models, all excited levels of an atom or molecule are included and solved as pseudo-species during the calculation [8, 9]. Non-equilibrium effects, i.e. deviations of the inner energy levels from the equilibrium Maxwell-Boltzmann distribution, are automatically accounted for as the level of detail is conserved. As the complexity of the model increases for a higher accuracy, STS models require significantly more computational power than MT models. This may lead to prohibitively expensive calculations in the case of 3D flow simulations [16].

Different strategies can be considered to reduce the complexity of detailed kinetic mechanisms while conserving a high level of information and precision. One possible approach consists in projecting the system

* aurelie.bellemans@ulb.ac.be; Also at von Karman Institute.
† alessandro.parente@ulb.ac.be
‡ thierry.magin@vki.ac.be

on a reduced base of variables according to a time-scale based classification. In the Rate-Controlled Constrained Equilibrium (RCCE) theory introduced by Keck [17, 18], the fast kinetics is projected on the slow varying base of the system providing a reduced representation of the dynamics. This reduced base is obtained by setting constraints, where the fast reactions are assumed to be at equilibrium [19–21]. A possible alternative to the aforementioned methods is provided by the lumping or binning techniques. In combustion, those lumping techniques propose to regroup species with similar compositions and functionality. These new recombined species are solved as one identity throughout the simulation [22, 23]. In other cases where the internal energy levels of a species are excited, the mechanism reduction is realized using a coarse grain model which lumps the inner levels into bins (where a distribution for the levels is prescribed). As an example, coarse grain models have been developed to reduce the electronic excitation and ionization mechanism in air (ABBA model [24–26]). Le et al. [39] have focused on the reduction of atomic systems and were able to capture the main features of macroscopic ionization of hydrogen using only 2 energy bins. Guy et al. [40] have reduced the 68 vibrational levels of $N_2(^1\Sigma_g^+)$ into 3 bins to study non-equilibrium phenomena in nitrogen nozzle flows. The vibrational kinetics of CO_2 has been reduced by using an adaptive binning scheme by Sahai et al.[41] maximizing the entropy in each bin. As a follow-up on this method, they reduced the NASA database for $N_2(^1\Sigma_g^+)$-$N(^4S_u)$ combining a spectral clustering algorithm with the maximum entropy principle reducing the 9391 species in the mechanism to 15-20 bins [42]. Multiple research has been published on the $N_2(^1\Sigma_g^+)$-$N(^4S_u)$ database developed by the Computational Quantum Chemistry Group at NASA Ames Research Center [31–33] to gain insight on molecular dissociation and internal energy excitation in hypersonic flows [34–36]. The first coarse grain models for the NASA database have been proposed by Magin, reducing the rovibrational levels into distinct energy bins [27]. Using the energy binning approach, the results obtained for flows across normal shock waves, within nozzles and along the stagnation-line of blunt bodies have shown that the proposed coarse-grained models can correctly reproduce the dynamics of $N_2(^1\Sigma_g^+)$ dissociation by using only 10-20 energy bins [27, 37, 38]. Alternative methods have been proposed by Yen [28, 29] and Zhu [30]. A long lasting literature has been published on the development and improvement of coarse grain models for reducing the large $N_2(^1\Sigma_g^+)$-$N(^4S_u)$ mechanism, using every time more complex algorithms to cluster the inner levels. In this paper we will present a novel technique inspired from machine learning, combining a simple coarse grain model with Principal Component Analysis.

Principal Component Analysis (PCA) is a statistical approach for projecting a system on a reduced reference system identified by the so-called principal components [43]. The PCA method learns from data as in the machine learning approaches [44–46]. These principal components correspond to the directions with the largest variance in the system. The reduced base is obtained after solving an eigenvalue problem on the covariance of the full thermo-chemical state. Over the years, the method has demonstrated its ability to reduce large combustion mechanisms as shown by Sutherland and Parente [47, 48]. Global and local Manifold-Generated PCA have been derived from the method as shown by Isaac [49] and Coussement [50]. As a first attempt of applying the method to plasma flows, Peerenboom [51] has coupled PCA with non-linear regression to reduce the vibrational levels of CO_2. More recently, PCA has successfully been applied on a collisional-radiative model for argon plasma to study non-equilibrium phenomena in shock relaxation calculations reducing the dimensionality of the system by 90 % [52, 53]. PCA can be used as a tool to analyze the dynamics of a reacting system and to retrieve its main variables which can thereafter be used in a reduced model. Moreover, PCA has demonstrated it could conserve the level of accuracy of the original model which is not the case using a coarse grain model. The method is simple, and doesn't require detailed knowledge about the energy levels as is required using complex lumping strategies. The objective of the present work is to combine PCA with physical methods such as the binning techniques providing an optimal reduced representation of large kinetic mechanisms. We will demonstrate that the PCA-based reduced model is more accurate, and offers a valuable alternative to existing reduction methods based on coarse grain models.

The present paper describes how to combine a simple coarse grain model with principal component analysis to optimally reduce large kinetic mechanisms. To demonstrate the method, the full physical model for the $N_2(^1\Sigma_g^+)$-$N(^4S_u)$ NASA ARC mechanism will be reduced in shock relaxation calculations using re-entry free-stream settings. The database for the $N_2(^1\Sigma_g^+)$-$N(^4S_u)$ mixture is presented in Section II together with its coarse grain model based on a simple uniform binning strategy. Principal component analysis is described in Section III and the results are presented in Section IV. The conclusions are summarized in Section V.

II. PHYSICAL MODELING

A. NASA Ames database for the $N_2(^1\Sigma_g^+)$-$N(^4S_u)$ mechanism

The NASA Ames Research Center (ARC) database gives accurate state-to-state data for the thermodynamics and kinetics of rovibrational excitation, dissociation and predissociation of molecular N_2 with N [33, 54, 55]. Besides the N_2-N interaction, the database also provides

detailed information for N_2-N_2 collisions [56, 57]. In total, 9390 rovibrational energy levels $N_2(\nu,\ J)$ are provided for the electronic ground state of N. The index ν stands for the vibrational quantum number. As 61 vibrational levels can be considered when there is no rotation, this index can vary from 0 to 60. The rotational quantum number is given by the index J. A total number of 279 rotational levels are considered for $\nu = 0$.

Three main reactions have been computed with the $N_2(^1\Sigma_g^+)$-$N(^4S_u)$ database: collisional dissociation, pre-dissociation and excitation between all states. In this work, the pre-dissociation processes have been neglected only using the following processes,

1. Collisional dissociation of bound states and predissociated states:

$$N_2(\mathrm{i}) + N \underset{k_b^{diss}}{\overset{k_f^{diss}}{\rightleftharpoons}} N + N + N, \tag{1}$$

2. Inelastic collisional excitation between all states:

$$N_2(\mathrm{i}) + N \underset{k_b^{in}}{\overset{k_f^{in}}{\rightleftharpoons}} N_2(\mathrm{j}) + N, \tag{2}$$

using the convention $i = i(\nu, J)$ ordered by increasing energy, with $i < j$ and $i, j \in \{1, ..., 9390\}$.

State-to-state collision cross-sections have been computed using quasi-classical trajectory (QCT) quantum calculations to provide rate coefficients for the three dissociation and inelastic collisional processes. The temperature range that was used for these cross-section computations ranged between 7,500 and 50,000 K. When considering all exchanges between energy levels, 44 million processes can be computed. However, only 19 million non-zero excitation rate coefficients were retained from the QCT computation. The backward rate coefficients are computed based on micro-reversibility. The equilibrium constants K for dissociation is given by the ratio of the forward rate coefficient k_f and backward rate coefficient k_b,

$$K^{diss} = \frac{k_f^{diss}}{k_b^{diss}} = \frac{[a_N Q_N^t(T)]^2}{a_{N_2(i)} Q_{N_2}^t(T)} \exp \frac{-(2E_N - E_{N_2(i)})}{k_B T}, \tag{3}$$

and for inelastic collisional excitation,

$$K^{in} = \frac{k_f^{in}}{k_b^{in}} = \frac{a_{N_2(j)}}{a_{N_2(i)}} \exp \frac{-(E_{N_2(j)} - E_{N_2(i)})}{k_B T}, \tag{4}$$

with $i < j$ and $i, j \in \{1, ..., 9390\}$, a the degeneracy, Q^t the translational partition function, E the energy, T the temperature and k_B the Boltzmann constant. The degeneracy of the energy levels of N_2 depends on the rotational quantum number $J(i)$ for each level,

$$a_{N_2(i)} = (2J(i) + 1)a^{NS}, \tag{5}$$

where the nuclear spin degeneracy a^{NS} equals 6 for even and 3 for odd $J(i)$ respectively. The degeneracy for single nitrogen a_N equals 12 summing the nuclear and electronic spin contributions. The translational partition functions Q_N^t and $Q_{N_2}^t$ are given as a function of the translational temperature T,

$$Q_j^t(T) = \left(\frac{2\pi k_B m_j T}{h_P^2} \right)^{3/2}, \tag{6}$$

with $j \in \{N, N_2\}$, h_P the Planck constant and m_k the mass of species j.

B. Rovibrational collisional model

The NASA ARC database has been used to develop a detailed rovibrational collisional model [34] to study molecular dissociation and excitation in nitrogen shocks. The governing equations are written as Euler equations in the shock frame and conserve the species continuity, for both N and all the rovibrational levels of N_2. The total energy of the mixture is conserved with respect to the translational temperature of the species, which is the only temperature considered in this model,

$$\frac{\partial(\rho y_N u)}{\partial x} = \omega_N, \tag{7}$$

$$\frac{\partial(\rho y_{N2(i)} u)}{\partial x} = \omega_{N2(i)}, \tag{8}$$

$$\frac{\partial(\rho u^2 + p)}{\partial x} = 0, \tag{9}$$

$$\frac{\partial(\rho u H)}{\partial x} = 0, \tag{10}$$

with $i \in \{1, ..., 9390\}$.

The thermodynamic properties are obtained by summing the contributions for all N_2 levels together with the contribution for N, with the number density of the gas:

$$n = n_N + n_{N2}, \tag{11}$$

and the pressure:

$$p = n_N k_B T + n_{N2} k_B T. \tag{12}$$

The thermal energy density is also composed out of a translational part and the formation contributions of both N and N_2,

$$\rho e(T) = \frac{3}{2} n k_B T + n_N E_N + \sum_{i=1}^{9390} n_i E_i, \tag{13}$$

with $n_{N2} = \sum_{i=1}^{9390} n_{N2(i)}$.

Additional temperatures can be retrieved after post-processing the data to obtain the internal, vibrational

and rotational temperatures of the N_2 levels. Our temperature of interest is the internal temperature of the bins, T_{int}, is obtained by solving [58],

$$\sum_{i=1}^{9390} n_{N_2(i)} E_{N_2(i)} - n_{N2} E_{N2}^{int}(T_{int}) = 0. \tag{14}$$

C. Coarse grain models

Previous studies have shown that the full RVC model can be reduced to a simplified model using coarse-grain models. The objective is to lump the excited levels of the N_2 molecule into several energy bins as demonstrated in the works of Magin [27], Munafò [37], Torres [59]. Nowadays, more sophisticated binning strategies exist based on maximum entropy principles instead of energy separation as shown is the work of Sahai [41]. However, we will focus on the simple binning methods as they are simple and easy to implement. In the early lumping history, two models have been developed, which cluster the levels into simple energy bins : the Boltzmann and the uniform rovibrational collisional model, denoted by BRVC and URVC respectively. Their difference lies in the method used to average the properties of the energy bins. In the BRVC model, the properties of each bin are averaged through the use of an equilibrium Boltzmann distribution function, similar to the MT models. The population distribution of each bin has been averaged using this Boltzmann average at the translational temperature T,

$$\frac{n_{N_2(i)}}{n_k} = \frac{a_{N_2(i)}}{Q_k(T)} \exp\left(-\frac{\Delta E_k(i)}{k_B T}\right), \tag{15}$$

with n_k the population distribution of bin k, and E_k the energy distribution of the bin:

$$n_k = \sum_{N_2(i)\in B^k} n_{N_2(i)}, \tag{16}$$

$$E_k = E(1) + \Delta E_k. \tag{17}$$

The energy E_k depends on the energy of the first level in energy bin k, and the ΔE_k which represents the energy difference between the level k and this first energy level. The partition function of the energy bin k, $Q_k(T)$, can be considered as the bin's degeneracy and is given by:

$$Q_k(T) = \sum_{N_2(i)\in B^k} a_{N_2(i)} \exp\left(-\frac{\Delta E_k(i)}{k_B T}\right). \tag{18}$$

Thermodynamic properties for the BRVC model contain an extra term because of the temperature dependence of the Boltzmann distribution in the definition of the specific heats and energies:

$$\rho e(T) = \frac{3}{2} n k_B T + n_N E_N + \sum_{N_2(i)\in B^k} n_{N_2(i)} E_{N_2(i)} + \sum_{N_2(i)\in B^k} n_{N_2(i)} k_B T^2 \frac{\partial \ln Q_k(T)}{\partial T}, \tag{19}$$

with the specific heats,

$$c_{V,N}(T) = \frac{3}{2}\frac{k_B}{m_N}, \tag{20}$$

$$c_{V,k}(T) = \frac{3}{2}\frac{k_B}{m_{N_2}} + \frac{\partial}{\partial T}\left[\frac{k_B T^2}{m_{N_2}} \frac{\partial \ln Q_k(T)}{\partial T}\right]. \tag{21}$$

The difficulty in these lumping methods lies in the choice of the correct number of bins to represent the detailed NASA ARC database in the full RVC model. After a bin sensitivity analysis, Munafò [60] defined the optimal number of Boltzmann bins to use to represent the detailed RVC model. He showed that both the translational and internal temperature are bin dependent. The more bins are used, the higher the translational temperature becomes. Starting from 10 bins, both temperatures converge to the exact equilibrium value. Starting from 100 bins, the temperature follows its converged solution. A good approximation of the N mole fraction can already be obtained using 10 bins. To conclude on the BRVC bins, there can be stated a minimum of 10 bins is necessary to be able to well represent the dissociation rate of nitrogen. Ideally, 100 bins are needed to reconstruct a correct post-shock temperature field which converges to the exact equilibrium solution. Unfortunately, using 100 bins is still a high amount of variables to use within a 2D code for solving a re-entry problem.

When considering the uniform formulation for the URVC bins, the energy of a bin, E_k is averaged according to the degeneracy of each corresponding excited state i in the bin Bk,

$$E_k = \frac{1}{a_k} \sum_{N_2(i)\in B^k} a_{N_2(i)} E_{N_2(i)}. \tag{22}$$

The same kind of relation using the degeneracy exist when expressing the population of every bin,

$$\frac{n_{N_2(i)}}{n_k} = \frac{a_{N_2(i)}}{a_k}, \tag{23}$$

with the degeneracy, energy and population of bin k expressed respectively as,

$$a_k = \sum_{N_2(i)\in B^k} a_{N_2(i)}, \tag{24}$$

$$E_k = \sum_{N_2(i)\in B^k} E_{N_2(i)}. \tag{25}$$

$$n_k = \sum_{N_2(i)\in B^k} n_{N_2(i)}. \tag{26}$$

Calculating the thermodynamic properties for the URVC model is straight forward as each bin can be treated as a separate species with its own characteristics; each bin has its own energy level and degeneracy

according to the previously described relations. The contributions of each bin are summed together to evaluate the total thermodynamic properties of molecular nitrogen the mixture. The energy of N_2 can be represented as the sum of the energies over the uniform bins k in the expression for the gas thermal energy density:

$$\rho e(T) = \frac{3}{2} n k_B T + n_N E_N + \sum_{N_2(i) \in B^k} n_{N_2(i)} E_{N_2(i)}. \quad (27)$$

When considering Boltzmann bins, an additional temperature-dependent term should be added because of the expression for the partition function. For URVC bins, $Q_k = a_k$, which simplifies the determination of the thermodynamic properties significantly. The specific heats are also free from a temperature-dependent term and are given by the relations:

$$c_{V,N}(T) = \frac{3}{2} \frac{k_B}{m_N}, \quad (28)$$

$$c_{V,k}(T) = \frac{3}{2} \frac{k_B}{m_{N_2}}. \quad (29)$$

with $k \in B^k$.

A bin sensitivity study has also been performed for the URVC bins by Munafò [60]. In this investigation, the minimal number of bins has been determined to represent the full RVC solution by comparing shock tube simulations. For a describing the temperature behavior within 1% accuracy, 100 bins are needed. For determining an accurate dissociation rate, allowing an error on the temperatures, the number of bins can be decreased to 10. The translational temperature is not bin-dependent when using the URVC bins. However, this is the case when using the BRVC model.

Unfortunately, using the URVC model as it is leads to wrong equilibrium calculations. By averaging the populations uniformly according to their degeneracy, important non-equilibrium properties of the flow are being neglected during the simulation. As the thermodynamics has been altered, a correct equilibrium state cannot be retrieved using the URVC model as it is. To solve this problem, a correct equilibrium can be forced by imposing the forward and backward reaction rate by micro-reversibility.

As mentioned previously, an internal temperature for the energy bins can be extracted after post-processing the solution for the flow field. By expressing the ratio of the sum of the population in each bin over the total population of N_2, the following expression for the internal temperature T_{int}, can be retrieved,

$$\frac{\sum n_{N_2(i)} E_{N_2(i)}}{n_{N_2}} = \frac{\sum E_k a_k \exp(-E_k/k_B T_{int})}{\sum a_k \exp(-E_k/k_B T_{int})} \quad (30)$$

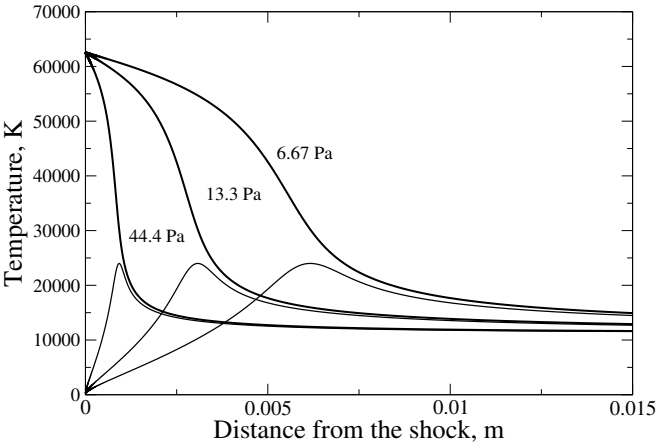


FIG. 1: Temperature profiles for the URVC model using 100 bins after a normal shock wave. Free-stream velocity: 10 km/s. Thick lines translational temperature T, thin lines internal temperature T_{int}.

TABLE I: Free-stream conditions for the RVC model shock relaxation simulations.

T [K]	300
p [Pa]	6.67, 13.3, 44.4
p [torr]	0.05, 0.1, 0.33
v [km/s]	10

Figure 1 represents the temperature profiles when using the URVC model with 100 bins. The thick lines stand for the translational temperature of the species, and the thin lines the internal one. The free-stream pressure equals 10 km/s and the pressure ranges from 6.67 to 44.4 Pa as shown in Table I. When increasing the pressure, the number of collisions between the species increases, which results in a shorter relaxation time. All internal temperatures reach a peak before melting with the translational one and converging to thermal equilibrium. This peak in internal temperature reaches values up to 25,000 K for a free-stream velocity of 10 km/s. The initial mass fraction for single nitrogen has been set to 2.8 % to enhance the dissociation in the beginning of the simulation. The 100 bins case will be used as a benchmark for providing the training sample in the PCA reduction.

III. PRINCIPAL COMPONENT ANALYSIS FOR CHEMISTRY REDUCTION

Principal component analysis offers a way to reduce the dimensionality of the reacting manifold by projecting the system on a truncated base made up out of its principal components. These principal components

are uncorrelated and retain most of the variance of the system. The computational cost decreases considerably as only a smaller number of variables, the principal components, are taken into account to solve the set of governing equations. One can choose to work in the space given by the eigenvectors which correspond to the principal components: these are called scores and relate to the PCA-score method. The CFD code has to be modified accordingly to allow a change of variables from the conserved ones, such as mass fractions and temperatures, to the scores. Another method consists in relating the chosen principal components to variables expressed in the original space of mass fractions. This technique has lead to development of MG-PCA, which has already been discussed and applied to argon plasma in previous work by the authors [52].

This section of the paper describes the PCA-score technique in more detail and shows how it can be coupled to a rotation method, such as the VARIMAX method, for retrieving a more stable formulation of the source terms and increasing the robustness of the code.

A. PCA-scores

PCA starts with a training data set containing the value of all conserved variables for every observation. These conserved variables correspond to mass fractions, temperatures and the velocity. However, previous work[61] has shown the reduction works best when using only mass fractions when carrying out PCA. In this particular case, the mass fractions are retrieved at different distances from the shock front after a shock relaxation simulation. The sample data is collected in matrix $\mathbf{Y}$, which has the size $[n \times Q]$ with n the number of observations or points in space and Q the number of variables. For the N-N_2 system, Q=9,391 as this number of species is considered within the model,

$$\mathbf{Y} = \begin{bmatrix} y_{11} & \cdots & y_{1Q} \\ \vdots & \ddots & \vdots \\ y_{n1} & \cdots & y_{nQ} \end{bmatrix}. \tag{31}$$

Some pre-processing techniques are applied to prepare the data for PCA. After centering the variables, they are scaled by dividing them by a scaling factor which is determined by a suitable scaling technique. Choosing a good scaling method is essential as it can affect the size and the accuracy of the reconstruction of the manifold after PCA reduction. An overview of different scaling techniques (variable stability, pareto, max, ...) are given in the work of Parente and Sutherland[62]. In previous work on the reduction of collisional-radiative chemistry, Pareto scaling appeared to be the most convenient method.

To correlate information of this system in terms of variance, an eigenvalue problem is solved on the covariance matrix, given by $\mathbf{S}$, to obtain the eigenvalues, $\mathbf{L}$, and eigenvectors, $\mathbf{A}$ (Eq. 32).

$$\mathbf{S} = \frac{1}{n-1}\mathbf{Y}^T\mathbf{Y} = \mathbf{A}\mathbf{L}\mathbf{A}^T \tag{32}$$

As our main interest lies in the eigenvectors containing most of the variance of the system, we only select those with the highest eigenvalue. The matrix of eigenvectors can be truncated to a matrix $\mathbf{A}_q$ containing only the $q < N_s$ eigenvectors with the highest variance.

$$\mathbf{Z}_q = \mathbf{Y}\mathbf{A}_q \tag{33}$$

$$\tilde{\mathbf{Y}}_q = \mathbf{Z}_q\mathbf{A}_q^T \tag{34}$$

When projecting the original data set on this truncated matrix, one obtains the principal components or scores which correspond to the most important directions of the reduced system. The scores are a linear combination of the original species as they contain each variable weighted by a PCA-defined loading. When inverting relation 33 one can find back the original sample as shown in Eq.34.

The scores do not relate to the original space given by the conserved variables which are in our case the mass fractions, but do relate to the eigenvectors of the variables. The governing equations must be solved in this space and should accordingly be rewritten in the CFD code. More generally, if a set of transport equations exists under the following conservative form:

$$\frac{\partial}{\partial t}\rho\mathbf{y} + \nabla \cdot (\rho\mathbf{u} \otimes \mathbf{y}) = \omega_y \tag{35}$$

than it can be rewritten in the score space as follows:

$$\frac{\partial}{\partial t}\rho\mathbf{z} + \nabla \cdot (\rho\mathbf{u} \otimes \mathbf{z}) = \omega_z \tag{36}$$

In these relations, $\mathbf{y}$ is a mass fraction for a single species and $\mathbf{z}$ a single principal component, which are each individual realizations of the vectors $\mathbf{Y}$ and $\mathbf{Z}$ respectively. The species source terms should also be transformed to the space of principal components by using the truncated matrix of eigenvectors $\mathbf{A}_q$:

$$\omega_Z = \omega_Y\mathbf{A_q} \tag{37}$$

IV. RESULTS

The one-dimensional SHOCKING code, developed by Munafò[60] has been used to simulate the relaxation of ionizing shocks in the N-N_2 mixture with the Score-PCA based reduced model. Applying PCA on the full mechanism containing almost 10000 species remains expensive. The PCA model of the full mechanism retained around 100 scores, which is still too expensive for 2D or 3Dl calculations. The idea is to start from a coarse grain

model given by the BRVC or URVC formulation and apply Score-PCA on the model to further reduce the cost of the simulation. The data from the coarse grain models using 100 bins for the N_2 states has been used as a starting point for Score-PCA. Using 100 bins ensures reproducing detailed information, such as post-shock temperatures, dissociation rates and populations, with high accuracy.

First the BRVC model will reduced. Next, Score-PCA will be applied on the URVC binning model. The goal is to reduce the 101 variables (100 bins of N_2 + N) to the smallest number possible while conserving detailed chemistry features. To assess if a reduced model is accurate, we will compare the post-shock temperature, dissociation rate and populations. The principal components will thereafter be analyzed to investigate how much of the original system is expressed within the reduced model.

A. Score-PCA on the BRVC model

The BRVC model already provides a large reduction of the original RVC mechanism as shown in the work of Munafó [60]. As a reminder of his work, the 9391 species were reduced to a model using 100 bins for a correct temperature estimation. When allowing an error on the translational temperature, 10 bins are sufficient to predict the molar mass fraction of dissociated nitrogen. Ideally, we would like to find a reduced model which reproduces a correct temperature and nitrogen dissociation rate with less than 100 variables.

The solution for 100 bins has been used as a training data set for the Score-PCA method, in order to start from a model which was closer to the full RVC solution. The data for the mass fractions of the bins has been centered and scaled according to the PARETO scaling method [62].

Figure 2b shows the dissociation rate of N and the post-shock temperature respectively for a reduced BRVC model based on 6 scores against a URVC reduction with 5 scores. This was the best BRVC reduction that could be obtained as the solution diverges when using only 5 scores. An important model reduction has been obtained as 6 variables have been retained out of the 100 bins. Expressing this result as a global reduction, we can state the NASA ARC database has been reduced by more than 99 %.

Our interest lies in using the cheapest binning method to reduce the total computational cost of the simulation. The BRVC lumping is a little more expensive and complicated as its uniform counterpart as the properties should be averaged using an equilibrium distribution in each bin. We have therefore decided to investigate the URVC model in detail for the different free-stream parameters.

TABLE II: Testing conditions for assessing the reduced model. Free stream conditions are given by the index 1. Calculated post-shock conditions by the index 2.

T_1 [K]	v_1 [km/s]	p_1 [Pa]	T_2 [K]	v_2 [km/s]	p_2 [Pa]
300	10	44.4	62,547	2.5	36,861
300	10	13.3	62,547	2.5	11,041
300	10	6.67	62,547	2.5	5,537

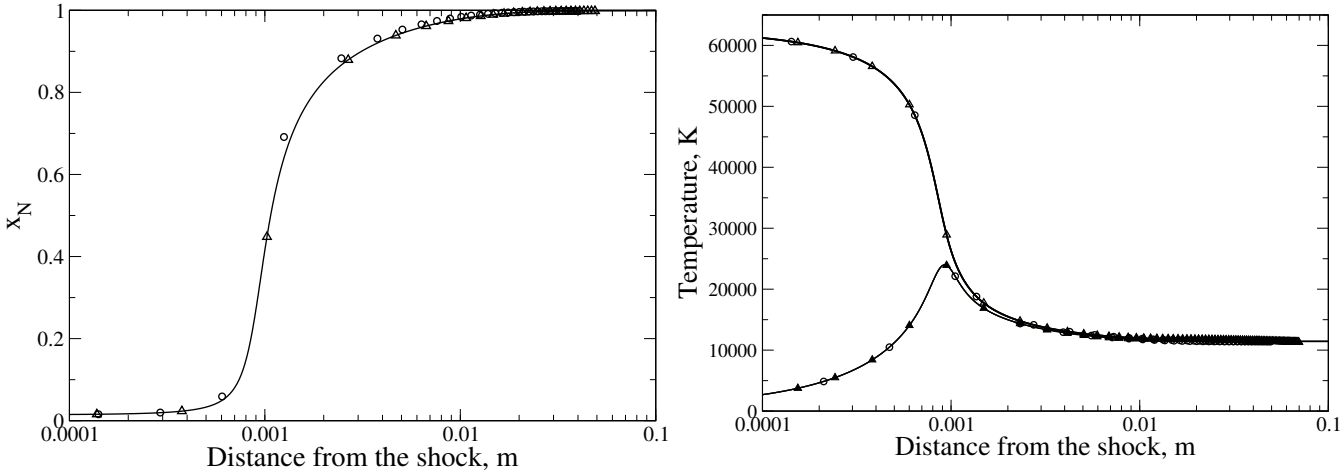
The URVC model using the non-uniform energy grid of Torres [59] has shown to provide the best results with respect to coarse grain models. In the next paragraph, we will show how this coarse grain binning model can be combined with Score-PCA to further reduce the model.

B. Score-PCA on the URVC model

Also in this case, the scores have been retrieved starting from a full data field containing the mass fractions for 100 bins. Centering and PARETO scaling have been applied to the training data set. More severe free-stream conditions, reaching lower pressures have been tested with the URVC bins as shown in Table II. These low pressure and high speed cases are representative of re-entry missions. Re-entry velocities of 10 km/s are typical for lunar returns.

The results for the reduced model against the original RVC solution can be visualized in Figure 2b for the 44.4 Pa case. The dissociation rate of N and the post-shock temperatures have been plotted for the reduced models based on 5 scores against the full solution. Small discrepancies of the order of some percent can be observed for the 5 scores model in the representation for the N mole fraction. Figure 3b shows similar results for the 13.3 Pa case. Also for this intermediate pressure case, the model can be represented by 5 scores allowing 1 % error on the mass fraction of single nitrogen. The results for the third test case with a pressure of 6.67 Pa, can be visualize in Figure 4b. The model can be represented by 6 sores allowing a little error on the both the temperatures and the dissociation rate for N.

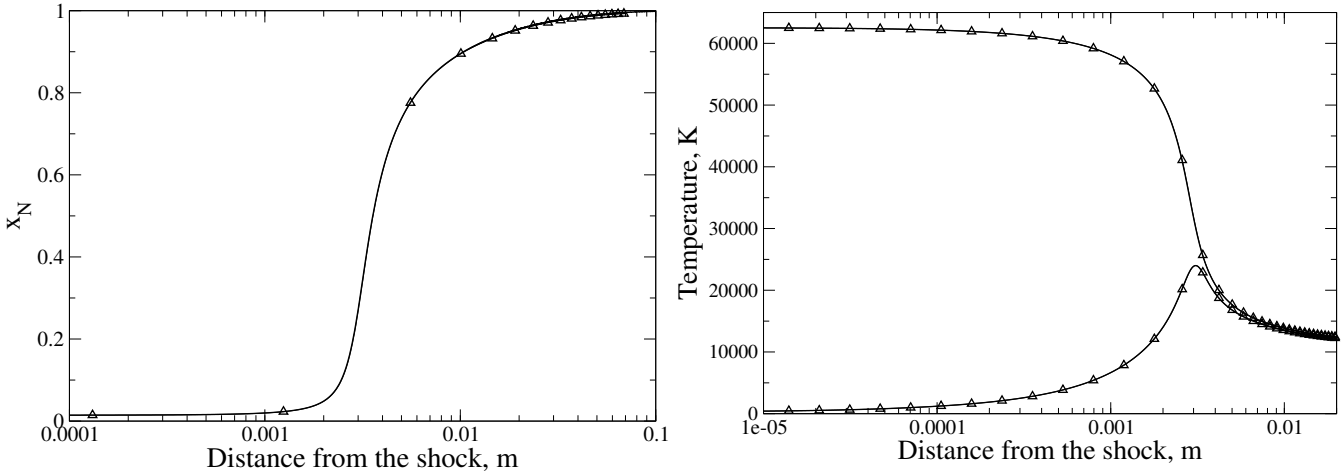
To evaluate the model, the relative error has been calculated between the full RVC model and its reduced representation with 6 scores in the re-entry conditions of 6.67 Pa and 10 km/s. As a reference case, the error of a URVC(20) model has been compared with PCA and the full case. Figures 6a and 6b represent the error on the translational and internal temperature respectively, which lies around 3.5 % for the score model. For the URVC(20) model the error on the internal temperature T_{int} lies significantly higher : 43 %. The error on the molar fraction of single nitrogen is shown in Figure 6c and has a value of 10 % for the score model and 17% for the URVC(20) case. The relative error of the PCA



(a) N mole fraction.

(b) Temperatures.

FIG. 2: Comparison between the RVC solution and PC-scores. Pre-shock conditions: u_1= 10 km/s and p_1= 44.4 Pa. Full lines RVC, lines with circles 6 scores BRVC, lines with triangles 5 scores URVC.



(a) N mole fraction.

(b) Temperatures.

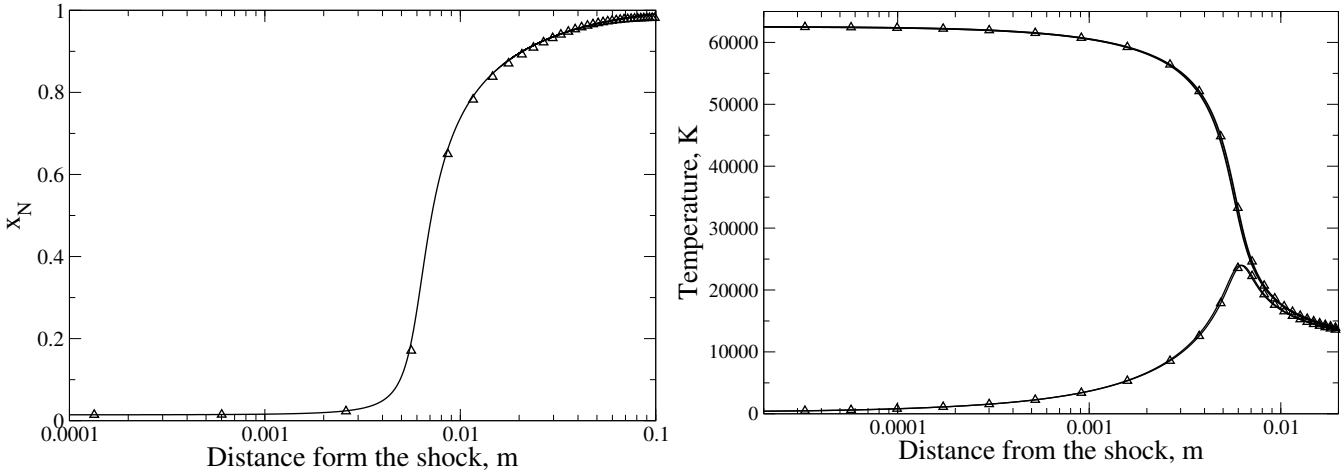
FIG. 3: Comparison between the RVC solution and PC-scores. Pre-shock conditions: u_1= 10 km/s and p_1= 13.33 Pa. Full lines RVC, lines with triangles 5 scores URVC.

scores model is significantly lower than the one obtained using the URVC(20) representation. Comparing these results with the error of advanced binning techniques as presented in the works of Sahai et al. [42], we can conclude PCA performs better in terms of accuracy for the test case investigated here.

To finalize the evaluation of the reduced model, the populations are visualized for the most severe case using the lowest pressure in a Boltzmann plot in Figure 5a. The 6-scores model has been used to retrieve the populations of the 100 URVC bins. This is possible as the scores or principal components are a linear combination of the original variables. We are able to retrieve the same detail on the distribution function as the 100 bins at a reduced cost with respect to alternative methods such as the spectral methods [42] as can be seen in Figure 5b where the populations have been plotted for the URVC(10) model. This unique property of PCA stresses its suitability for reducing large plasma mechanisms.

V. CONCLUSION

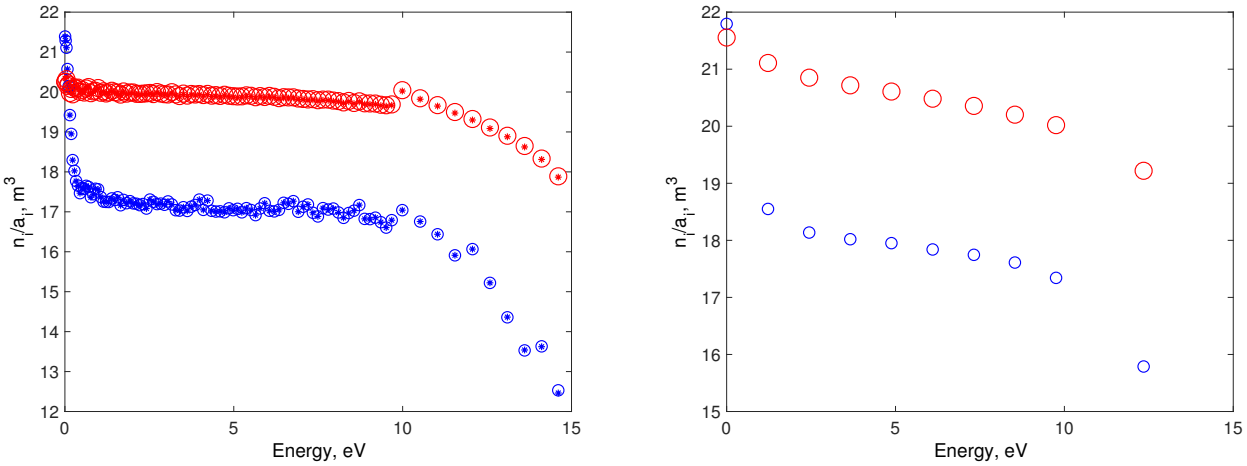
The detailed rovibrational collisional model for the large $N_2(^1\Sigma_g^+)$-$N(^4S_u)$ NASA ARC database has been reduced using a novel technique combining two reduction methods. A simple and cheap lumping technique has been applied condensing the 9390 rovibrational states of N_2 into 100 energy bins, conserving detailed information.



(a) N mole fraction.

(b) Temperatures.

FIG. 4: Comparison between the RVC solution and PC-scores. Pre-shock conditions: u_1= 10 km/s and p_1= 6.67 Pa. Full lines RVC, lines with triangles 6 scores URVC.



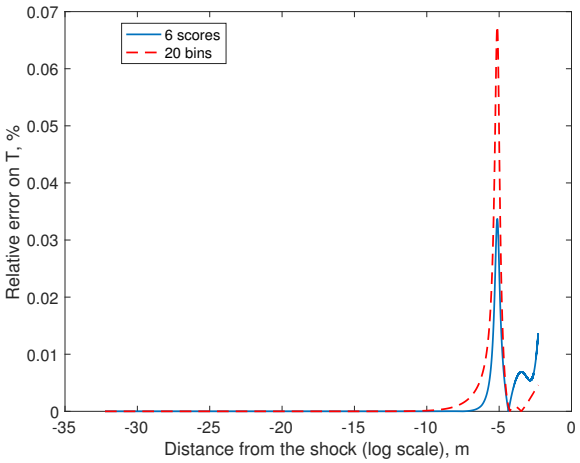
(a) URVC(100) (circles) vs. PC-score(6) (stars).

(b) URVC(10).

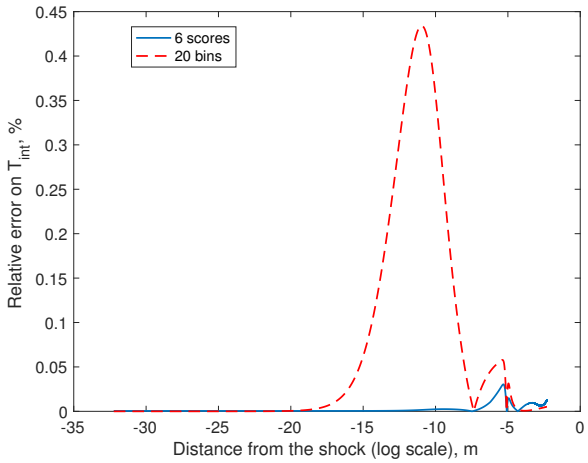
FIG. 5: Population distribution against their energy in a Boltzmann plot for 6.67 Pa and 10 km/s. Small blue dots: x= 4×10^{-4} m , big red dots: x = 5×10^{-2} m. Full RVC solution in Figure (a).

These 100 energy bins have been used in a shock relaxation code to provide a training sample for carrying out PCA. Score-PCA has thereafter been applied reducing these 100 bins to 6 scores for the BRVC model. Applying the method on the URVC model lead to a reduction using 5 scores for the 13.3 and 44.4 Pa case. For the low pressure case of 6.67 Pa, 6 scores were retained from the 100 variables. Score-PCA reduces the coarse grain model with 95%, decreasing the complexity of the mechanism significantly. Globally, the 9391 species in the $N_2(^1\Sigma_g^+)$-$N(^4S_u)$ mechanism have been reduced to 6 new variables leading to an impressive compression of the mechanism. The Score-PCA technique allows to retrieve detailed chemistry features accurately, such as post-shock temperatures, dissociation rates and the populations of the energy bins. Retrieving the populations of the energy bins with such a high precision is a unique property of the method compared to the coarse grain models. The relative error between the full model and the PCA-based reduction shows higher accuracy for the results at a lower computation cost than recently published work using advanced lumping techniques [42].

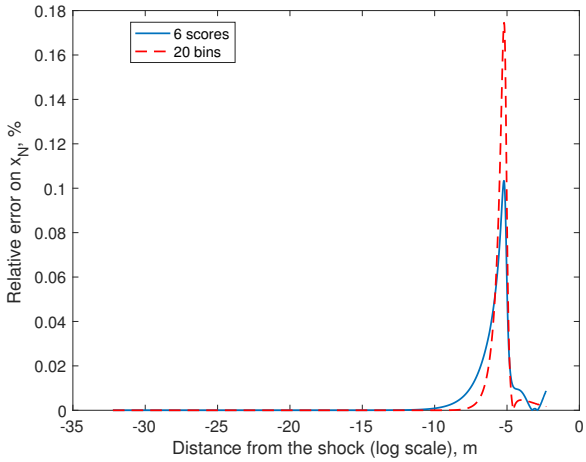
The present work has shown how lumping techniques and principal component analysis can be combined to reduce large and complex chemistry models. The technique is user-friendly as it is automatic and simple to implement. No expert judgment is required before the application of PCA and no complicated fine-tuning is necessary to optimize the method. The method follows



(a) Relative error on the temperature, T.



(b) Relative error on the internal temperature, T_{int}.



(c) Relative error on the mole fraction of N, x_N.

FIG. 6: Relative error on the temperatures and nitrogen mole fraction between the 6 Score-PCA and URVC(20) reduced model against the full RVC solution for 6.67 Pa and 10 km/s.

the philosophy of machine learning where complex systems can be studied and solved by analyzing and developing algorithms for simple training cases. Following this line of thought, the developed one-dimensional reduction will be applied to solve multi-dimensional cases in future work.

ACKNOWLEDGMENTS

The research of A. Bellemans has been sponsored by a FRIA fellowship of the Belgian research fund F.R.S.-FNRS. The research of A. Parente is sponsored by the European Research Council, Starting Grant No. 714605. The authors would like to acknowledge the help of Dr. Erik Torres for providing the URVC model with non-uniform energy grid and Dr. Alessandro Munafò for discussions about the shock relaxation code.

[1] Yen Liu, Marco Panesi, Amal Sahai, and Marcel Vinokur, "General multi-group macroscopic modeling for thermo-chemical non-equilibrium gas mixtures," The Journal of chemical physics **142**, 134109 (2015).

[2] Chul Park, *Nonequilibrium hypersonic aerothermodynamics* (John Wiley and Sons, New York, 1989).

[3] Bernd Helber, Olivier Chazot, Annick Hubin, and Thierry E Magin, "Microstructure and gas-surface in-

teraction studies of a low-density carbon-bonded carbon fiber composite in atmospheric entry plasmas," Composites Part A: Applied Science and Manufacturing **72**, 96–107 (2015).

[4] Andrey Starikovskiy and Nickolay Aleksandrov, "Plasma-assisted ignition and combustion," Progress in Energy and Combustion Science **39**, 61–110 (2013).

[5] G Vanhove, M-A Boumehdi, S Shcherbanev, Y Fenard, P Desgroux, and SM Starikovskaia, "A comparative experimental kinetic study of spontaneous and plasma-assisted cool flames in a rapid compression machine," Proceedings of the Combustion Institute **36**, 4137–4143 (2017).

[6] Sergey V Pancheshnyi, Deanna A Lacoste, Anne Bourdon, and Christophe O Laux, "Ignition of propane–air mixtures by a repetitively pulsed nanosecond discharge," IEEE Transactions on Plasma Science **34**, 2478–2487 (2006).

[7] Burak Korkut, Deborah A Levin, and Ozgur Tumuklu, "Simulations of ion thruster plumes in ground facilities using adaptive mesh refinement," Journal of Propulsion and Power (2017).

[8] M. G. Kapper and J.-L. Cambier, "Ionizing shocks in argon. part 1: collisional-radiative model and steady-state structure," J. Appl. Phys. **109**, 113308 (2011).

[9] A. Bogaerts, R. Gijbels, and J. Vlcek, "Collisional-radiative model for an argon glow discharge," Journal of Applied Physics **84**, 121 (1998).

[10] Wei Lin, Rubén Meana-Pañeda, Zoltan Varga, and Donald G Truhlar, "A quasiclassical trajectory study of the n2 (x 1σ)+ o (3 p) no (x 2π)+ n (4 s) reaction," The Journal of chemical physics **144**, 234314 (2016).

[11] Daniil A Andrienko and Iain D Boyd, "Rovibrational energy transfer and dissociation in o2–o collisions," The Journal of chemical physics **144**, 104301 (2016).

[12] Inga S Ulusoy, Daniil A Andrienko, Iain D Boyd, and Rigoberto Hernandez, "Quantum and quasi-classical collisional dynamics of o2–ar at high temperatures," The Journal of chemical physics **144**, 234311 (2016).

[13] C. Park, "Review of chemical-kinetic problems of future NASA missions, I: Earth entries," J. Thermophys. Heat Transfer **7**, 385–398 (1993).

[14] C. Park, J. T. Howe, R. L. Jaffe, and G. V. Candler, "Review of chemical-kinetic problems of future NASA missions, II: Mars entries," J. Thermophys. Heat Transfer **8**, 9–23 (1994).

[15] Chul Park, "The limits of two-temperature model," AIAA Paper **911**, 2010 (2010).

[16] M. Capitelli, C. M. Ferreira, B. F. Gordiets, and A. I. Osipov, *Plasma Kinetics in Atmospheric Gases* (Springer, 2000).

[17] James C Keck and David Gillespie, "Rate-controlled partial-equilibrium method for treating reacting gas mixtures," Combustion and Flame **17**, 237–241 (1971).

[18] J. C. Keck, "Rate-controlled constrained equilibrium theory of chemical reactions in complex systems," Prog. Energy Combust. Sci. **16**, 125–154 (1990).

[19] S. B. Pope, "The computation of constrained and unconstrained equilibrium compositions of ideal gas mixtures using gibbs function continuation," FDA **03-02** (2003).

[20] Z. Ren, S. B. Pope, A. Vladimirsky, and J. M. Guckenheimer, "The invariant constrained equilibrium edge preimage curve method for the dimension reduction of chemical kinetics," The Journal of Chemical Physics **124**, 114111 (2006).

[21] WP Jones and Stelios Rigopoulos, "Rate-controlled constrained equilibrium: Formulation and application to nonpremixed laminar flames," Combustion and Flame **142**, 223–234 (2005).

[22] Perrine Pepiot-Desjardins and Heinz Pitsch, "An automatic chemical lumping method for the reduction of large chemical kinetic mechanisms," Combustion Theory and Modelling **12**, 1089–1108 (2008).

[23] Vikram Reddy Ardham and Frdric Leroy, "Communication: Is a coarse-grained model for water sufficient to compute kapitza conductance on non-polar surfaces?" The Journal of Chemical Physics **147**, 151102 (2017).

[24] A. Bultel, B. G. Chéron, A. Bourdon, O. Motapon, and I. F. Schneider, "Collisional-radiative model in air for earth re-entry problems," Phys. Plasmas **13**, 043502 (2006).

[25] M. Panesi, T. E. Magin, A. Bourdon, A. Bultel, and O. Chazot, "Electronic excitation of atoms and molecules for the FIRE II flight experiment," J. Thermophys. Heat Transfer **25**, 361–374 (2011).

[26] Gianpiero Colonna, Iole Armenise, Domenico Bruno, and Mario Capitelli, "Reduction of state-to-state kinetics to macroscopic models in hypersonic flows," Journal of thermophysics and heat transfer **20**, 477–486 (2006).

[27] T.E. Magin, M. Panesi, A. Bourdon, R.L. Jaffe, and D. W. Schwenke, "Coarse-graining model for internal energy excitation and dissociation of molecular nitrogen," Chem. Phys. **398**, 90 – 95 (2012).

[28] Y. Liu, M. Vinokur, M. Panesi, and T. E. Magin, *A multi-group maximum entropy model for thermochemical nonequilibrium*, AIAA Paper 2010–4332 (10th AIAA/ASME Joint Thermophysics and Heat Transfer Conference, Chicago, IL, 2010).

[29] Y. Liu, M. Panesi, M. Vinokur, and P. Clarke, *Microscopic simulation and macroscopic modeling of thermal and chemical non-equilibrium gases*, AIAA Paper 2013–3146 (44th AIAA Thermophysics Conference, San Diego, CA, 2013).

[30] Tong Zhu, Zheng Li, and Deborah A Levin, "Development of a two-dimensional binning model for n2–n relaxation in hypersonic shock conditions," The Journal of Chemical Physics **145**, 064302 (2016).

[31] G. Chaban, R. L. Jaffe, D. W. Schwenke, and W. Huo, *Dissociation cross-sections and rate coefficients for nitrogen from acccurate theoretical calculations*, AIAA Paper 2008–1209 (46th AIAA Aerospace Sciences Meeting and Exhibit, Reno, NV, 2008).

[32] D. W. Schwenke, "Dissociation cross-sections and rates for nitrogen," in *Non-Equilibrium Gas Dynamics - From Physical Models to Hypersonic Flights*, Lecture Series (von Karman Institute for Fluid Dynamics, 2008).

[33] R. L. Jaffe, D. W. Schwenke, and G. Chaban, *Theoretical analysis of* N_2 *collisional dissociation and rotation-vibration energy transfer*, AIAA Paper 2009–1569 (47th AIAA Aerospace Sciences Meeting and Exhibit, Orlando, FL, 2009).

[34] M. Panesi, R. L. Jaffe, D. W. Schwenke, and T. E. Magin, "Rovibrational internal energy transfer and dissociation of $N(^4S_u) + N_2(^1\Sigma_g^+)$ sytem in hypersonic flows," J. Chem. Phys **138**, 044312 (2013).

[35] Jae Gang Kim and Iain D Boyd, "Monte carlo simulation of nitrogen dissociation based on state-resolved cross sections," Physics of Fluids **26**, 012006 (2014).

[36] Jae Gang Kim and Iain D Boyd, "State-resolved master equation analysis of thermochemical nonequilibrium of nitrogen," Chemical Physics **415**, 237–246 (2013).

[37] A. Munafò, M. Panesi, and T. E. Magin, "Boltzmann rovibrational collisional coarse-grained model for internal energy excitation and dissociation in hypersonic flows," Phys. Rev. E **89**, 023001 (2014).

[38] A. Munafò and T. E. Magin, "Modeling of stagnation-line nonequilibrium flows by means of quantum based collisional models," Phys. Fluids **26**, 097102 (2014).

[39] H. P. Le, A. R. Karagozian, and J. L. Cambier, "Complexity reduction of collisional-radiative kinetics for atomic plasma," Phys. Plasmas **20**, 123304 (2013).

[40] A. Guy, A. Bourdon, and M.-Y. Perrin, "Consistent multi-internal-temperatures models for nonequilibrium nozzle flows," Chem. Phys. **420**, 15–24 (2013).

[41] Amal Sahai, Bruno E Lopez, Christopher O Johnston, and Marco Panesi, "Novel approach for co2 state-to-state modeling and application to multidimensional entry flows," in *55th AIAA Aerospace Sciences Meeting* (2017) p. 0213.

[42] A Sahai, B Lopez, CO Johnston, and M Panesi, "Adaptive coarse graining method for energy transfer and dissociation kinetics of polyatomic species," The Journal of Chemical Physics **147**, 054107 (2017).

[43] Jonathon Shlens, "A tutorial on principal component analysis," arXiv preprint arXiv:1404.1100 (2014).

[44] Thomas H-J Uhlemann, Christoph Schock, Christian Lehmann, Stefan Freiberger, and Rolf Steinhilper, "The digital twin: Demonstrating the potential of real time data acquisition in production systems," Procedia Manufacturing **9**, 113–120 (2017).

[45] Benjamin Schleich, Nabil Anwer, Luc Mathieu, and Sandro Wartzack, "Shaping the digital twin for design and production engineering," CIRP Annals-Manufacturing Technology (2017).

[46] Gianmarco Aversano, Alessandro Parente, Olivier Gicquel, and Axel Coussement, "Application of reduced-order models based on the combination of pca & kriging on 1d flames," Fuel (2017).

[47] James C Sutherland and Alessandro Parente, "Combustion modeling using principal component analysis," Proceedings of the Combustion Institute **32**, 1563–1570 (2009).

[48] A. Parente, J. C. Sutherland, L. Tognotti, and P. J. Smith, "Identification of low-dimensional manifolds in turbulent flames," Proc. Combust. Inst. **32** (2009).

[49] B. Isaac, A. Coussement, O. Gicquel, P. J. Smith, and A. Parente, "Reduced-order pca models for chemical reacting flows," Combustion and Flame **161**, 2785 – 2800 (2014).

[50] Axel Coussement, Benjamin J Isaac, Olivier Gicquel, and Alessandro Parente, "Assessment of different chemistry reduction methods based on principal component analysis: Comparison of the mg-pca and score-pca approaches," Combustion and Flame **168**, 83–97 (2016).

[51] Kim Peerenboom, Alessandro Parente, Tomas Kozak, Annemie Bogaerts, and Gerard Degrez, "Dimension reduction of non-equilibrium plasma kinetic models using principal component analysis," Plasma Sources Science and Technology **24**, 025004 (2014).

[52] Aurélie Bellemans, TE Magin, A Coussement, G Degrez, and A Parente, "Mg-local-pca method for the reduction of a collisional-radiative argon plasma mechanism," in *45th AIAA Thermophysics Conference* (2015) p. 3105.

[53] Aurélie Bellemans, Thierry Magin, Axel Coussement, and Alessandro Parente, "Reduced-order kinetic plasma models using principal component analysis: Model formulation and manifold sensitivity," Physical Review Fluids **2**, 073201 (2017).

[54] R. L. Jaffe, D. W. Schwenke, G. Chaban, and W. Huo, *Vibrational and rotational excitation and relaxation of nitrogen from accurate theoretical calculations*, AIAA Paper 2008–1208 (46th AIAA Aerospace Sciences Meeting and Exhibit, Reno, NV, 2008).

[55] KL Heritier, RL Jaffe, V Laporta, and M Panesi, "Energy transfer models in nitrogen plasmas: Analysis of n 2 (x σ g+ 1)–n (4 s u)–e- interaction," The Journal of chemical physics **141**, 184302 (2014).

[56] Jason D Bender, Sriram Doraiswamy, Donald G Truhlar, and Graham V Candler, "Potential energy surface fitting by a statistically localized, permutationally invariant, local interpolating moving least squares method for the many-body potential: Method and application to n4," The Journal of chemical physics **140**, 054302 (2014).

[57] Jason D Bender, Paolo Valentini, Ioannis Nompelis, Yuliya Paukku, Zoltan Varga, Donald G Truhlar, Thomas Schwartzentruber, and Graham V Candler, "An improved potential energy surface and multi-temperature quasiclassical trajectory calculations of n2+ n2 dissociation reactions," The Journal of chemical physics **143**, 054304 (2015).

[58] M. Panesi, A. Munafò, T. E. Magin, and R. L. Jaffe, "Study of the non-equilibrium shock heated nitrogen flows using a rovibrational state-to-state method," Phys. Rev. E **90**, 013009 (2014).

[59] E Torres, Ye A Bondar, TE Magin, Andrew Ketsdever, and Henning Struchtrup, "Uniform rovibrational collisional n2 bin model for dsmc, with application to atmospheric entry flows," in *AIP Conference Proceedings*, Vol. 1786 (AIP Publishing, 2016) p. 050010.

[60] A. Munafò, *Multi-Scale Models and Computational Methods for Aerothermodynamics*, Ph.D. thesis, Ecole Centrale Paris, Châtenay-Malabry, France (2014).

[61] B. Isaac, *Reduced-order modelling for reacting flows based on principal component analysis*, Ph.D. thesis, University of Utah (2014s).

[62] A. Parente and J. Sutherland, "Prinicpal component analysis of turbulent combustion data: Data preprocessing and manifold sensitivity," Combustion and Flame **160**, 340–350 (2013).