Dimension Reduction in Chemical Reactions: Combustion Chemical Kinetics Analysis Based On Machine Learning

Zirui Liu
Princeton University
ziruil@princeton.edu

Dake Li
Princeton University
dakel@princeton.du

Abstract

Detailed chemical kinetics of the combustion of fuels contain a large number of species and elemental chemical reactions, which makes it computationally expensive in combustion simulation. To overcome this problem, people try to develop reduced mechanisms, which only keep important reactions and species to simplify combustion simulation. In this work, we applied different regression models to predict ignition delay time, and then selected important features as the reduced mechanism. The temperature profile and species concentration profiles are compared between the reduced and the detailed mechanism to show the accuracy of our reduction methods. The results show that Ridge regression and layer selection regression, which is developed by us in this work, are effective in mechanism reduction. Combustion reaction network is also analyzed using clustering and label propagation algorithm. Groups and communities are detected and enlighten us in the combustion chemistry studies.

1 Introduction

Combustion phenomena often involve complicate chemical kinetics and numerous elementary reactions. For example, hydrogen flame, which starts with reactants H_2 and O_2 and ends with H_2O, undergoes 19 elementary reactions, and 5 intermediate species H, O, OH, HO_2 and H_2O_2 appear in this process[1]. The information of all elementary reactions and species are gathered in the chemical mechanism of the fuel. For chemical mechanisms of larger fuels like C_4H_{10}, 230 species and 2457 elementary reactions are included. Practical fuels, such as gasoline or jet fuel, have even more complicate mechanisms. In the numerical simulation of combustion, it is computationally expensive to calculate all these elementary reactions, so people tried to reduce the size of the chemical mechanisms by removing some unimportant reactions and species without changing the results. Then the question is: how can we figure out which reactions are not important? Several methods are proposed, such as sensitivity analysis, Computational Singular Perturbation[2] and Direct Relation Graph[3]. However, two major disadvantages of these methods are (1) the reduced mechanisms highly depend on the initial condition (2) these methods require very detailed information of the chemical process.

In this work, we developed a data-driven mechanism reduction method using regularized regression models, which works in a wide range of initial conditions and does not need any information of chemistry during combustion. Besides this, we also construct a chemical reaction network of element flux between different species, and use unsupervised learning methods, such as clustering and label propagation algorithm to find groups and detect communities in the reaction network. This method provide us some insight in analyzing chemical process.

2 Related Work

There are several works on applying machine learning methods to simplify combustion chemical system or on using machine learning models to predict outcomes, such as flame temperature, ignition delay time, flame speed and etc. For example, Malik et al.[4] used Principal Component Analysis and nonlinear regression to parameterize a chemical transport system and reduce the large kinetic

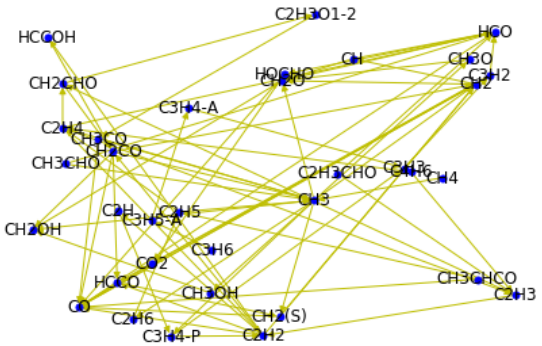


Figure 1: Reaction Path Diagram for the Flux of Carbon Element

models. Besides combustion chemistry, similar ideas have been explored in biochemical systems. Klimovskaia et al.[5] developed a sparse regression method to predict key bio-reactions and reaction network topology from a snapshot of certain time-dependent data. In our work, we pick reactions based on their prediction performance to simplify mechanism and also cluster species in light of reaction network.

3 Data

In this work, we use a extremely detailed butane (C_4H_{10}) mechanism from National University of Ireland, Galway[6], which contains 2457 reactions and 230 species as our target mechanism supposed to be simplified. We then use $Cantera$, which is an open-source solver for problems involving chemical kinetics, thermodynamics, and transport processes[7], to generate the data on all the reactions and the species in the complete ignition process. Data generated from $Cantera$ have been validated by chemical experiments and widely used in engineering literature, so it is reasonable to assume that our generated data are accurate enough and can be treated almost the same as the true data from doing chemical experiments.

An important outcome of the system is ignition delay time, which basically refers to the time for a fuel under a high temperature environment before it is ignited and depends on the property of the fuel and all the reactions involved. In general, the ignition delay time equals the time when the concentration of OH species or the gradient of temperature reaches maximum. In our work, we try to predict ignition delay time using all the 2457 reaction rates.

In addition, we use $Cantera$ together with $networkx$ package to generate the carbon element flux network in butane combustion. As shown in Figure 1, nodes stand for species, and if there is an directed edge between any two nodes, there is a flux of carbon element from one species to the other. For example, the edge from CH_3 species to CH_2 species indicates that CH_3 can react with other species and produce CH_2, and the carbon element is transferring from CH_3 to CH_2. Furthermore, the adjacent matrix of this network is generated by $networkx$ and used for clustering analysis later.

4 Machine Learning Methods

4.1 Data Pre-editing

In our dataset about ignition delay time and reaction rates, we split the data into the train set (2880 samples) and the test set (600 samples). Each of the 2457 reaction rates is considered as a single feature, whereas ignition delay time is the dependent variable.

Because the feature values vary from 1 to 10^8, the regression results can be robustly estimated only if we take logarithm for all the feature values. Also, the dependent variable, i.e. ignition delay time, is obviously a non-negative variable, and changes from 0.01 ms to 500 ms. We need to take into consideration that the dependent variable has a truncated distribution near 0, making the Gaussian assumption fails in the regression model. Moreover, the regression model will typically put equal weight on a 0.1 ms deviation when the true ignition delay time is 0.01 ms and on the same deviation

when the true value is 500 ms, which makes no sense in practice. Hence, we also take logarithm for the dependent variable to deal with these two issues.

Besides, we also delete the features which barely vary in the data, because their effects can not be identified in this data and they are not helping with prediction of ignition delay time. In this way, we drop around 150 features.

4.2 Prediction Task: Regularized Regression

In this project, we first try to predict ignition delay time using all the 2457 features we have. This is a supervised learning problem and can be solved by linear regression. Due to high dimensionality, the simple linear regression will perform extremely bad and over-fit the sample. Some regularized regression models are going to play an important role in this prediction task.

Among all the regularized regression models, we use Ridge, LASSO, forward stepwise regression, and our new method "layer selection regression", to explore the sparse structure in the regression model. Ridge regression penalizes the l_2 norm of the coefficients to achieve the sparsity, while LASSO penalizes the l_1 norm. The advantage of LASSO is that only a few coefficients will be non-zero. Since there are hyper-parameters for shrinkage, i.e. α in LASSO and Ridge regression, we choose regression models with built-in cross-validation functions(*LassoCV* and *RidgeCV* in the SciKitLearn Python libraries[8], to help us tune the hyper-parameters when doing shrinkage regression.

Forward stepwise regression (FSR) is another way of selecting important features in regression. We starts with no features in the model and gradually adds the features into the model one by one based on the correlation of residuals and the remaining features. After adding in all the features, we compares all these model by trading off between goodness of fit and model complexity, to finally choose the parsimonious model that can fit the data well. In this way, forward stepwise method avoids over-fitting and pick out a few important features in terms of prediction performance.

However, the sparse coefficients in LASSO and forward stepwise regression may still be difficult to be used as reduced mechanism in practice. A few crucial reactions are picked out due to statistical importance in these models, but we actually need reactions from each of the seven layers in the complete chemical process to ensure ignition and constitute a meaningful reduced mechanism. This requires that we need to pick out features based on not only the statistical importance, but also the chemical meaning. Thus, we come up with our own new method "layer selection regression" (LSR), to use forward stepwise regression on each layer to optimally pick out features on that layer, with some restrictions on the minimum number of selected features on that layer. In this way, we can make sure there are feature selected from each of the layers to make ignition really happen.

4.3 Clustering and Label Propagation Model

To analyze the reaction network, first we generate the adjacent matrix using $networkx$, where each row of this matrix stands for the flux connectivity with other species. K-Means clustering method is used to group all these vectors into several clusters. In each cluster, all species will have a similar connectivity with other species.

Besides, we further move on to community detection using LPA (Label Propagation Algorithm). The idea of LPA is letting the label propagate through the edge and affect other nodes. Initially, we assign a random label to each nodes. Then, based on the neighbour nodes' label and the weights of the edges in between, we change its label to the most probable one. After several iterations, when the label of every nodes does not change anymore, the nodes that share the same label are in one community. By iterating this algorithm, nodes which have more connections with neighbors are more likely to share the label with their neighbors

4.4 Evaluation Metrics

For each regression model, we predict the ignition delay time and calculate the mean squared error (MSE) when predicting in the train set and in the test set as

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2 \tag{1}$$

where Y_i is the true value, while $\hat{Y}_i$ is the predicted value. The smaller the MSE is, the better the model predicts. We compare MSE across linear regression, Ridge, LASSO, FSR and LSR.

Moreover, to evaluate the results from the chemical perspective, we compare key outcomes such as temperature profile and species profile in the $Cantera$ ODE solution, using both detailed chemical mechanism and our reduced chemical mechanism. The difference can be treated as the accuracy of our reduced mechanism in practice.

5 Spotlight Method: Forward Stepwise / Layer Selection Regression

In this section, we explain two methods we used in detail: forward stepwise regression and layer selection regression.

There are different methods of regularized regression, including shrinkage method such as LASSO, and information criteria method such as forward stepwise regression. Below, we describe forward stepwise regression in detail.

Forward stepwise regression is a regression model where we select the number of features based on Bayesian Information Criteria. Bayesian Information Criteria offers a score to trade off between the fit of the model and the complexity of the model. Now we have a prediction problem with $K = 2457$ features. This method begins with a empty model, and we then run K different linear regression models by gradually adding the features to the model one by one. Every time we want to add a feature, we follow the greedy algorithm to choose the feature which is the most correlated to the current residuals. Then, we evaluate the scores for different linear models based on their Bayesian Information Criteria (BIC) as

$$BIC = -2\log(\hat{L}) + k\log(n) = n\log(\hat{\sigma^2}) + k\log(n) \tag{2}$$

where $\hat{L}$ is the likelihood, n is the sample size, k is the number of features, and $\hat{\sigma^2}$ is the variance of residuals. The smaller the BIC score is, the better the model is.

The BIC score, on the one hand, favors the model with small residuals, but on the other hand, prefers the model with a small number of features. Choosing the optimal linear model with the minimal BIC score can effectively avoid the over-fitting issue. Thus we choose the optimal linear model among all the K models as our FSR model, to be used to predict the outcomes in the test set.

However, after we select a few features as key predictors, chemical researchers typically want to put these selected reactions into the $Cantera$ ODE solver to solve for the simulated experiment profile under this reduced mechanism. For a meaningful reduced chemical mechanism, we would need selected reactions from each of the seven layers in the complete chemical process. In general, our forward stepwise regression may not guarantee that there are selected features from each layer. Hence, we revise the classical forward stepwise regression, to run model selection layer by layer. In particular, we first run FSR in the first layer and pick out a few features based on minimized BIC. Then given the the features in the first features in the model, we jump to the second layer to run FSR to optimally add some other features from the second layer into the model. The process then gets iterated until we select from the last layer. In this way, we can make sure we add in some features from each layer and we are optimally controlling the number of features we are adding in from each layer. We also set some restriction on the minimum number of selected features on each layer to make sure each layer has a handful of features selected.

6 Results in Prediction Task

We run five different regression models to predict ignition delay time. We then try to compare their performance and interpret the estimated coefficients.

6.1 MSE in Predicted Outcome

The goodness of fit in the five regression models, including simple linear regression, Ridge, LASSO, forward stepwise regression and layer selection regression, vary dramatically. In Table 1, we compare the prediction errors using these regression models, both in the train set and in the test set.

The naive linear regression uses all the 2457 features to fit the data and thus has a serious over-fitting issue, which leads to huge prediction errors in the test set. After adding shrinkage over regression coefficients, Ridge performs really well both in sample and out of sample. LASSO selects 11 features and also does well in prediction. FSR optimally picks features by comparing BIC of various models, and thus avoids over-fitting, which makes the out-of-sample prediction extremely accurate. Similarly, our own method LSR optimally picks features from each layer of this chemical process,

Model	Train Set	Test Set
Linear	0.58358	9.66×10^{13}
Ridge	0.00758	0.00362
LASSO	0.02013	0.01555
FSR	0.00675	0.00320
LSR	0.00246	0.00313

Table 1: Mean Squared Error in Prediction

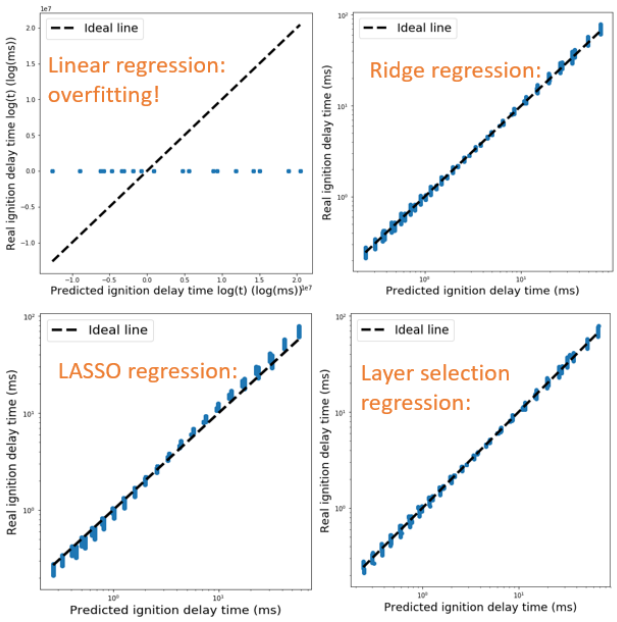


Figure 2: True Values v.s. Predicted Values in Test Set

and has a quite small MSE in out-of-sample prediction, indicating the chemical prior knowledge is helpful in choosing important features.

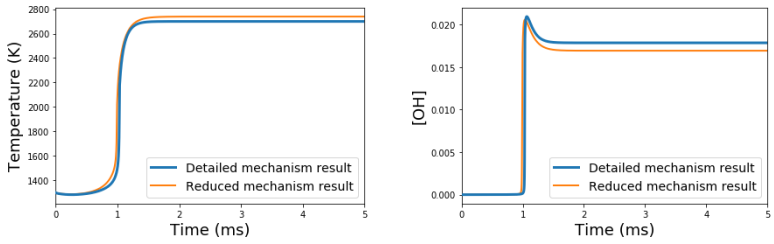
We also compare the true value versus the predicted values in the test set, as shown in Figure 2. Simple linear regression has serious over-fitting issue, and thus the scatter plots deviate the 45 degree line remarkably. All the other models predict the test set really well.

6.2 BIC in Model Selection

For forward stepwise regression, we gradually add features into the model one by one, and compare different models based on BIC, which is presented in Table 2.

# features	model BIC	# features	model BIC
1	-8020.6	11	-14051.8
2	-9492.8	12	-14172.8
3	-9519.9	13	-14230.1
4	-9663.5	14	-14234.6
5	-11394.2	15	-14226.9
6	-13416.9	16	-14244.6
7	-13567.0	17	-14237.5
8	-13957.4	***18***	***-14250.4***
9	-13987.6	19	-14242.5
10	-14000.2	20	-14234.8

Table 2: Model BIC in Forward Stepwise Regression



(a) Temperature profile of reduced and detailed mechanisms

(b) OH concentration profile of reduced and detailed mechanisms

Figure 3: Mechanism reduction result using Ridge regression (1281 reactions and 211 species)

The regression model with 18 features has the minimal BIC value, and thus FSR optimally choose 18 important features to predict ignition delay time. BIC values represent the trade-off between goodness of fit and model complexity, so the optimal model with 18 features is both a parsimonious model and a good fit for the data.

Besides the classical FSR, we also run our own LSR, trying to optimally pick features from each layer of the chemical process. In addition, we require the algorithm to at least pick a certain number of features from each layer based on our chemical prior knowledge, so that ignition is guaranteed to happen and the reduced mechanism makes sense in practice. Our method LSR finally picks 649 features in total.

6.3 Interpret Estimated Results

The coefficients in all the regression models can be interpreted as the percentage change of ignition delay time given 1 % increase in a certain reaction rate. The top 10 important reactions selected by LASSO are listed in Table 3

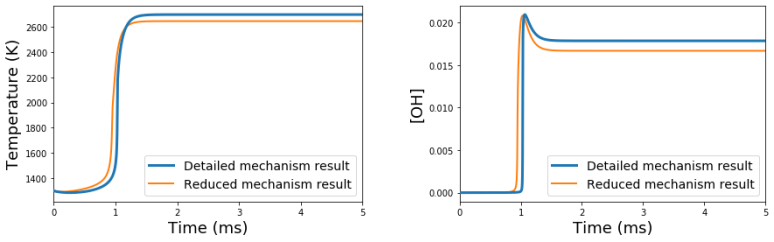
No.	Reaction	No.	Reaction
1	CH3OH (+M) <=> CH2OH + H (+M)_b	6	H + TC4H9 => IC4H10_f
2	HCO + M => CO + H + M_f	7	CH + H2 => CH2 + H_f
3	C2H3 + CH3 (+M) <=> C3H6 (+M)_f	8	CH3 + CO2 + M => CH3CO2 + M_f
4	CH3CO2 + M => CH3 + CO2 + M_f	9	OCHO + M => CO2 + H + M_f
5	OCH2OCHO + OH => HO2CH2OCHO_f	10	C2H5OH (+M) <=> C2H4 + H2O (+M)_b

Table 3: Top reactions selected by LASSO regression models

The reactions selected by our model are consistent with sensitivity analysis in chemistry research, which proves that our model is effective. However, since the reaction systems are nonlinear, and using linear to approximate will lose some of the information. So besides the reactions selected by the model, it is also necessary to add some other reactions to make the ignition happens. So it is a trade off between statistical regression and nonlinear chemical dynamics. Finally, using the reactions selected by our model and other reactions using prior chemical knowledge, we reduce the mechanism and input it into $Cantera$ solver and compare it with the original detailed mechanism. The results of ridge regression and LSR are shown in Figure 3 and Figure 4 . Ridge regression could reduce the mechanism to half size and keep very good accuracy, while our LSR model could reduce the mechanism to around a quarter with relatively good accuracy.

7 Results in Network Analysis

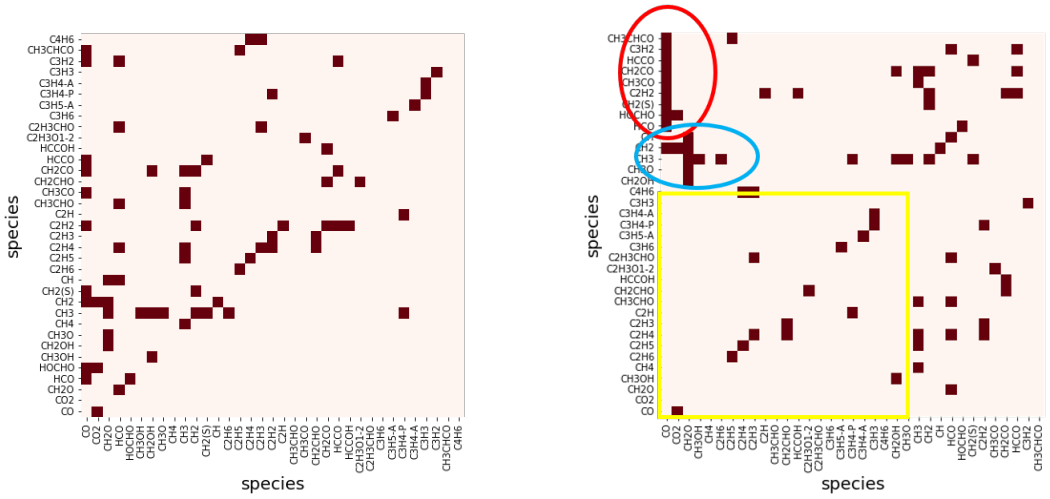
As we mentioned earlier, the chemical reaction system can be expressed using a large network, with the nodes being all intermediate species and edges being fluxes of elements. By using unsupervised learning methods, such as clustering, and community detection methods, we are able to find patterns from the network and study the chemical system from a new perspective.



(a) Temperature profile of reduced and detailed mechanisms

(b) OH concentration profile of reduced and detailed mechanisms

Figure 4: Mechanism reduction result using LSR (649 reactions and 174 species)



(a) Adjacent matrix of the reaction network

(b) Clustering results in adjacent matrix

Figure 5: Adjacent matrix and the clustering results

7.1 Estimated Latent Variables and Parameters

The adjacent matrix is generated and presented using $networkx$ and $seaborn$ packages, as shown in Figure 5a. Since the reaction network is a directed network, the matrix is not symmetric. For example, at (C_2H_4, HCO), the value is 1, indicating that there is a carbon element flux from C_2H_4 to HCO.

Inputting this matrix into Kmeans clustering method, we identify 3 different groups of species. The new adjacent matrix after changing the order of the rows to put species in the same cluster together is shown in Figure 5b. We observe patterns in the reordered adjacent matrix. The first group contains species from CH_3CHCO to HCO, and it turns out to be the major source of CO formation, and the second group contributes the most to CH_2O formation. Knowing this could assist us in the study of how to suppress the emission of carbon dioxide, which is very dangerous to human being, in the chemical reactions.

7.2 Interpret Community Assignments

To use LPA (label propagation algorithm) for the community detection, first we need to transform the network to a undirected one by assuming that the label could not only transfer in the flux direction, but also in the opposite direction of the flux. This is a valid assumption if we are looking for species communities within which the interaction is strong and outside which the interaction is very weak.

The results of LPA is shown in Figure 6. Three communities are identified by this algorithm. The first one contains C_3H_6, $C_3H_5 - A$ and $C_3H_4 - A$, which is obvious to us because these three species are almost "outside" of this network with only one edge connected $C_3H_4 - A$ to C_3H_3.

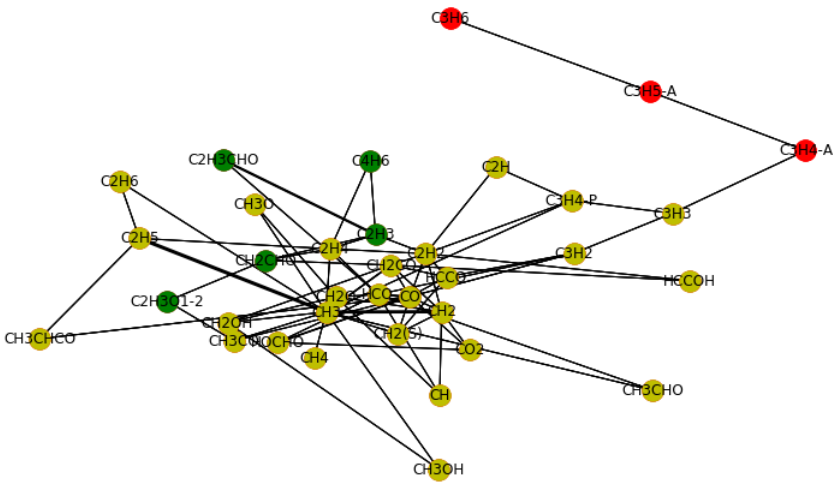


Figure 6: Community detection by Label Propagation Algorithm

So in the chemistry studies, if there's a concentration change of any species in this community, it is highly probable that we could find the corresponding response in the concentration of the other two species. Another community contains C_4H_6, C_2H_3, CH_2CHO, C_2H_3CHO and $C_2H_3O_1 - 2$, which also indicate that these five species have a stronger connection. We could interpret the label propagation as the propagation of the perturbation of the system. For example, in this reaction system, if there is a change of the concentration of certain species, other species in the same community are more sensitive to this change than the species outside the community. Thus, this method would be instructive in studying the dynamics of chemical systems.

8 Conclusion and Possible Extension

This work explores machine learning methods of using regression models to reduce the chemical kinetics and using clustering and community detection to find patterns in reaction networks. Ridge regression and LSR are found to be effective in model reduction. Patterns are discovered in the network using clustering and community detection.

Next step, it would be interesting if we include more outcomes in our regression model, such as the production rate of water and the flame speed. Also nonlinear regression could be applied here to approximate this system more accurate. In terms of network analysis, the stochastic block model could be used to reduce the dimension of the network and predict networks of a different fuel. It is always amazing to let machine learning help us find interesting things in science and engineering.

References

[1] C. K. Law, *Combustion physics*. Cambridge university press, 2010.

[2] S. Lam and D. Goussis, "The csp method for simplifying kinetics," *International Journal of Chemical Kinetics*, vol. 26, no. 4, pp. 461–486, 1994.

[3] T. Lu and C. K. Law, "A directed relation graph method for mechanism reduction," *Proceedings of the Combustion Institute*, vol. 30, no. 1, pp. 1333–1341, 2005.

[4] M. R. Malik, B. J. Isaac, A. Coussement, P. J. Smith, and A. Parente, "Principal component analysis coupled with nonlinear regression for chemistry reduction," *Combustion and flame*, vol. 187, pp. 30–41, 2018.

[5] A. Klimovskaia, S. Ganscha, and M. Claassen, "Sparse regression based structure learning of stochastic reaction networks from single cell snapshot time series," *PLoS computational biology*, vol. 12, no. 12, p. e1005234, 2016.

[6] D. Healy, N. Donato, C. Aul, E. Petersen, C. Zinner, G. Bourque, and H. Curran, "n-butane: Ignition delay measurements at high pressure and detailed chemical kinetic simulations," *Combustion and Flame*, vol. 157, no. 8, pp. 1526–1539, 2010.

[7] D. G. Goodwin, R. L. Speth, H. K. Moffat, and B. W. Weber, "Cantera: An object-oriented software toolkit for chemical kinetics, thermodynamics, and transport processes." `https://www.cantera.org`, 2018. Version 2.4.0.

[8] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, *et al.*, "Scikit-learn: Machine learning in python," *Journal of machine learning research*, vol. 12, no. Oct, pp. 2825–2830, 2011.