Alma Mater Studiorum - Università di Bologna

DOTTORATO DI RICERCA IN

AUTOMOTIVE PER UNA MOBILITÀ INTELLIGENTE

Ciclo 34

Settore Concorsuale: 09/C1 - MACCHINE E SISTEMI PER L'ENERGIA E L'AMBIENTE

Settore Scientifico Disciplinare: ING-IND/08 - MACCHINE A FLUIDO

DEVELOPMENT OF ADVANCED METHODS FOR THE SIMULATION OF THE REACTING MIXTURE FORMATION IN INTERNAL COMBUSTION ENGINES WITH THE USE OF MACHINE LEARNING ALGORITHMS

Presentata da: Leonardo Pulga

Coordinatore Dottorato

Nicolò Cavina

Supervisore

Gian Marco Bianchi

Esame finale anno 2022

Acknowledgements

My Ph.D. journey has been defined from the first days by the support and assistance of many invaluable people to whom I am most grateful.
First of all I would like to thank my supervisor Professor Gian Marco Bianchi, for the wonderful opportunity to pursue this path and for his constant support and expertise. The same level of deep gratitude is for Dr. Claudio Forte, my industry reference and guidance, for his constant support and belief in me, whose insightful feedback have allowed me to constantly improve.

I would like to thank my parents, they have always been an example of ethics on work as well as in life for me, and my entire wonderful family, for their constant presence and help.

A special thanks goes to the love of my life, Silvia, for her continuous support and cheering in the darker days and for remaining stuck at home with me for the long weekends spent writing research papers and this dissertation without complaining.

Last but not least, I would like to thank Dr. Valerio Mariani, a dear friend, for his constant presence and fruitful discussion during the master and Ph.D. years and for the most theoretical discussions in front of a slice of cake or pizza. Together with Dr. Gabriele Losi, my oldest friend whom I am grateful to, they have been a clear reference during all these years.

“Truth is much too complicated to allow anything but approximations”

John von Neumann

Abstract

Besides increasing the share of electric and hybrid vehicles, in order to comply with more stringent environmental protection limitations, in the mid-term the auto industry must improve the efficiency of the internal combustion engine and the well to wheel efficiency of the employed fuel. To achieve this target, a deeper knowledge of the phenomena that influence the mixture formation and the chemical reactions involving new synthetic fuel components is mandatory, but complex and time intensive to perform purely by experimentation. Therefore, numerical simulations play an important role in this development process, but their use can be effective only if they can be considered accurate enough to capture these variations.

The most relevant models necessary for the simulation of the reacting mixture formation and successive chemical reactions have been investigated in the present work, with a critical approach, in order to provide instruments to define the most suitable approaches also in the industrial context, which is limited by time constraints and budget evaluations. To overcome these limitations, new methodologies have been developed to conjugate detailed and simplified modelling techniques for the phenomena involving chemical reactions and mixture formation in non-traditional conditions (e.g. water injection, biofuels etc.). Thanks to the large use of machine learning and deep learning algorithms, several applications have been revised or implemented, with the target of reducing the computing time of some traditional tasks by orders of magnitude. Finally, a complete workflow leveraging these new models has been defined and used for evaluating the effects of different surrogate formulations of the same experimental fuel on a proof-of-concept GDI engine model.

Nomenclature

Considering the large number of equations present in the following chapters, if a variable appears with a different meaning from the one defined here, it will be explicitly stated, while apices/pedices will always be introduced.

Abbreviations

AEOI: After End Of Injection
AI: Artificial Intelligence
AL-PIONA: Alcohol – n-Paraffins i-Paraffins Olefins Naphthenes Aromatics
AMD: Average Mean Diameter
ANN: Artificial Neural Network
DNN: Deep Neural Network
BDC: Bottom Dead Center
BMEP: Brake Mean Effective Pressure
C: Carbon Atom
CAD: Crank Angle Degrees
CFD: Computational Fluid Dynamics
CO: carbon monoxide
CO2: carbon dioxide
EGR: Exhaust Gas Recirculation
EOI: End Of Injection
ETRF: Ethanol Toluene Reference Fuel
GDI: Gasoline Direct Injection
GP: Gaussian Process
HC: unburned hydrocarbons
HCCI: Homogeneous Charge Compression Ignition
ICE: Internal Combustion Engine
IDT: Ignition Delay Time
IVC: Intake Valve Closing
IVO: Intake Valve Opening
LFS: Laminar Flame Speed
LFT: Laminar Flame Thickness
LGI: Liquid-Gas Interface
MARE: Mean Absolute Relative Error
MAPO: Max Absolute Pressure Oscillation
MFB50: Angle of 50% of Mass Fuel Burnt
MON: Motor Octane Number
NBP: Normal Boiling Point
NOx: nitrogen oxides
ON: Octane Number

Nomenclature

PDF: Probability Density Function
PFI: Port Fuel Injection
PISO: Pressure-Implicit with Splitting Operators
PM: Particulate Matter
PN: Particulate Number
PRF: Primary Reference Fuel
RANS: Reynolds Average Navier-Stokes
RON: Research Octane Number
RSME: Root Squared Mean Error
RVP: Reid Vapor Pressure
SA: Spark Advance
SACI: Spark Assisted Compression Ignition
SI: Spark Ignition
SKR: Soave-Redlich-Kwong
SOI: Start Of Injection
aSOI: After Start of Injection
TDC: Top Dead Center
TKE: Turbulent Kinetic Energy
TRF: Toluene Reference Fuel
TRFE: Toluene Reference Fuel with Ethanol addition
TX: Distillation temperature at X% volume evaporated
UCB: Upper Confidence Bound
VLE: Vapor-Liquid Equilibrium

1 Introduction

1.1 Motivation and Objectives

Starting from 1992, the European car manufacturers have been required to satisfy more and more stringent regulations regarding the green-house gases and particulate matter emissions of their fleet in order to register new cars [1]. Considering that the internal combustion engine is arguably the most relevant source of these substances from the vehicle, such regulations have always pushed the development of new strategies and technologies for both SI and CI engines. The latest regulations, in particular, starting from Euro6a have introduced limitations to all major environmental pollutants: CO, HC, NOx, Particulate Matter mass and parcel number, as well as a CO2 target to be reached on predefined driving cycles. The expectations for the new regulations Euro7, are for a tightening of the limitations [1], as well as a testing procedure on more extreme testing conditions, for which electrification of the powertrain system is expected to play an essential role. On the other hand, several studies are currently being carried out by governments and private research facilities ([2] [3] [4] [5]) in order to assess the potential in efficiency and emission reduction of synthetic fuels, employed as dense energy carrier, or by introducing more relevant bio-derived fractions inside the final blend, to reduce the CO2 generated by the whole process. The main idea behind these studies is that in the mid-term, electrification solutions will not be available worldwide, and the most immediate response to climate change would come from a reduction of the emissions generated by the current technology. To enforce this perspective, the European average passenger cars fleet age is reported, underlying the fraction of vehicles older than 10 years [Acea 2021] in Figure 1-1. It clearly shows that it cannot be reasonably expected to reach a renewal of the entire fleet towards hybrid or full electric vehicles before the next 10 to 15 years in most countries, that could benefit, however, from upgrades in the fuel technology that they would operate on.

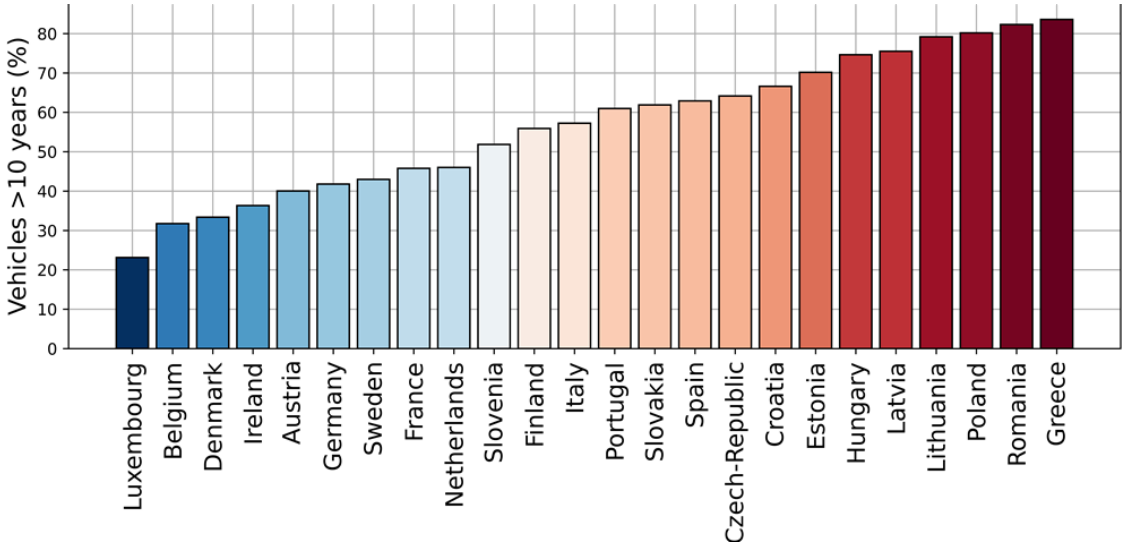

Figure 1-1: Fraction of vehicles older than 10 years registered in Europe in 2020

Motivation and Objectives

The gasoline fuel that is typically found at the pump across Europe is required by law to respect a minimum set of properties [6] . In particular, limitations about its composition are provided, such as a maximum fraction of metals (Manganese, Lead), a maximum fraction of benzene, olefins and aromatics molecules (due to their higher tendency to produce harmful emissions) and a controlled fraction of alcohols (ethanol and methanol). Other properties that are required to satisfy precise limitations are the density at ambient conditions, the distillation residues at the final boiling point, the evaporated volume fraction at 100 °C and 150 °C and the vapor pressure, in order to set the standard evaporative behaviour of the liquid fuel. Finally, the RON and MON of the gasoline fuel must be verified following the ISO 5164 and ISO 5163 testing methods, through a variable compression ratio carburetted engine named CFR (Cooperative Fuel Research), for a minimum value of RON≥95 and MON≥85. Besides minimum law requirements, few information are usually available for the complete characterization of the pump gasoline, whereas many research activities have been carried out to obtain a more complete description of some specific fuel formulations.

To support the development of new combustion engines with innovative strategies or the integration with renewable and synthetic molecules through computational methods, it is essential to perform CFD simulations sensitive to variations in fuel composition. Considering a standard CFD simulation of the internal combustion engine following the RANS approach, the model that are most sensitive to the fuel definitions are:

1) Liquid spray model, which includes the definition of the liquid phase properties (surface tension, mass and energy transfer coefficients), atomization phase inside the injector holes, break-up phase due to instabilities and aerodynamics forces, spray wall interaction, for which the boiling and critical temperature of the fuel play a crucial role in the formation of liquid film.
2) Ignition and combustion models, that rely on the chemical reactivity properties of the fuel for the prediction of the energy release rate of the mixture and flame front propagation.
3) Knock model, which is used for evaluating the auto-ignition tendency of the mixture as a function of local thermodynamics conditions.
4) Emission models, which are used to predict the main gaseous and solid pollutant formation in the burnt region, and based on the level of detail of the chosen model they may depend on the average composition of specific fractions of some components.

The main objective of the present work is to provide a deeper understanding of the effects that new fuel properties might have on the performance of existing SI engine models from the point of view of chemical kinetics, by leveraging innovative techniques for the definition of fuel surrogates for computational analysis. At the same time, a general fuel surrogate formulation is provided as a reference for further numerical studies regarding the European market.

1.2 Structure of the thesis

The works conducted through this thesis is organized as follows:

- Chapter 2 describes the fundamentals of chemical kinetics and chemical kinetics mechanisms available for modelling in detail the processes of flame propagation and self-ignition of hydrocarbons and alcohols and their relevance in engine modelling. Particular focus will be given to the availability of reactions describing the chemical pathways of the most common molecules used for the definition of gasoline fuel surrogates. Then, a state of the art in artificial intelligence algorithms and workflows for regression problems is carried out, to lay out the theoretical ground for their application in the following chapters.

- Chapter 3 analyses the auto-ignition phenomenon for its relevance in internal combustion engines modelling and how it is coupled with chemical kinetics simulations. A validation study on the chosen chemical kinetics mechanism is performed on the palette of molecules that are used for the fuel surrogate formulation and on a selection of reference mixtures. Then, a dataset of ignition delay time values is defined for the training and validation of machine learning techniques to be employed for the prediction of ignition delay times on newly defined surrogates, without the need to perform new simulations.

- Chapter 4 introduces to the modelling of laminar flame speed, the numerical methodology employed for its computation and several factors affecting its computation complexity. Then, an introduction to automatic mechanism reduction methods is reported, as well as several computational strategies involving data-driven models to increase the speed of computation of the laminar flame speed for new mixtures. Thereafter, a focus is given on new strategies to reduce the computing time required to generate laminar flame speed and thickness look-up tables in case of the inclusion of water vapour as an independent variable in the simulations, together with pressure, temperature, equivalence ratio and EGR mass fraction that are usually employed.

- Chapter 5 is dedicated to the definition of an optimization algorithm for the characterization of gasoline fuel surrogates and applied to 2 different targets. The theoretical ground and numerical strategies for the computation of each property of the target fuels are presented, as well as an in-depth analysis of the strength and limitations of any fuel surrogate definition. In particular, the availability of experimental properties is correlated with the unicity of the surrogate formulation with a custom merit function. Three families of surrogates are therefore formulated, depending on the level of information available and compared, to identify a minimum set of data that would be required for a robust surrogate formulation.

Structure of the thesis

- Chapter 6 proposes an application of the defined workflows to the simulation of a proof of concept GDI engine, with particular focus on the knock prediction at peak power conditions. A general description of the models chosen for the CFD simulation is proposed, with particular focus on the impact that the fuel surrogate formulation plays in their predictive accuracy. Besides, the newly formulated temperature wall function is described and applied for a more accurate prediction of the knock risk related to the gas temperature in the peripherical regions.

2 Chemical kinetics and AI fundamentals

The second chapter provides an introduction to the fundamental concepts of chemical kinetics and machine learning that will be adopted during the thesis applied to the simulation of fuel combustion simulations. It includes:

1) An introduction to the chemical kinetics simulations of hydrocarbons and the selection of the numerical schemes used for the activities.

2) An introduction to the fundamental regression algorithms and data processing strategies that have been applied to the following problems.

2.1 Chemical kinetics fundamentals for CFD

Chemical phenomena play a crucial role in the internal combustion engine functioning, from combustion, to knock, to pollutant formation, but their understanding and modelling constitutes a challenging problem. In fact, hydrocarbons oxidation presents peculiarities with respect to other processes, for example the cool flame and negative temperature coefficient during a two-stage ignition at intermediate temperature, as well as the requirement to predict correctly both low temperature decomposition intro intermediate species that consistently affect the high temperature reactions. These aspect make the number of intermediate species and reactions required for an accurate description of the phenomena occurring in an internal combustion engine extremely high. The fundamental reaction path diagram for an exemplary linear hydrocarbon oxidation can be represented as in Figure 2-1, where the main steps are highlighted:

1) The main alkyl radicals thermally decompose, after reaching a sufficiently high temperature and H abstraction:
    a. Into smaller alkyl radicals and olefin species following a high temperature path which makes it undergo β scission mainly due to the $\dot{H} + O_2 = \dot{O} + \dot{OH}$ reaction
    b. Into ketohydroperoxide species due to the reaction with O_2 at lower temperature.
2) $Q\dot{O}OH$ with temperature increase start to form cyclic ether species, conjugate olefins and β-decomposition products that compete with the formation of the ketohydroperoxide species. This phenomenon is particularly relevant, due to the overall reduction in the reactivity of the system at intermediate temperatures, which is commonly referred to as NTC behavior when analyzing the ignition delay time of hydrocarbons.
3) At the same time, $Q\dot{O}OH$ can react with molecular O_2 leading to $O_2Q\dot{O}OH$ species (peroxyalkylhydroperoxide) which can isomerize and form ketohydroperoxide and $\dot{OH}$ radicals, through an internal H abstraction.

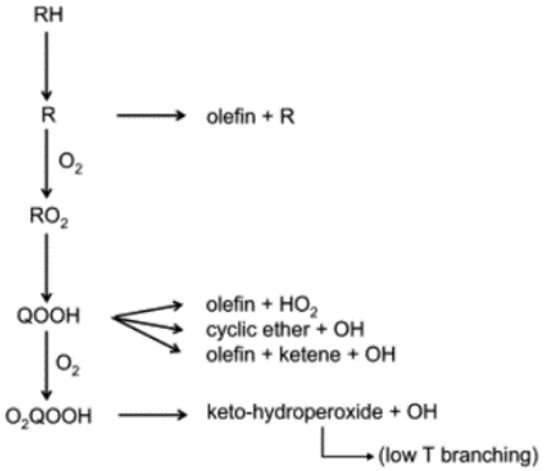

Figure 2-1: simplified reaction path for generic linear N-Alkane [7]

Considering the importance that the availability of O_2 molecules play in the hydrocarbon decomposition, it is evident that the equivalence ratio plays a crucial role in the evolution of the reactions, as well as the mixture pressure due to the pressure dependency of the addition of molecular oxygen to alkyl radicals [8].

The detailed description of all the reaction pathways involved in the oxidation of large hydrocarbon molecules is still an open research topic, currently investigated by several major international research facilities [9]. The state of the art chemical kinetics mechanisms are constructed upon reference experimental values and molecular dynamics computations and can accurately describe the oxidation and mutual interactions of several large molecules used to represent gasoline fuels.

Several chemical mechanisms are available in literature for hydrocarbons oxidation, but thanks to their extensive validation and precise nomenclature, three have been taken into consideration for this work: NUI2021 [10], LLNL2021 [11] and CRECK2019 [12] in their most complete forms. The number of species and reactions present in each scheme is collected in Table 1-1, together with the potential gasoline fuel components available.

Table 2-1: Chemical kinetics schemes available for hydrocarbon combustion simulation

	NUI2021	LLNL2021	CRECK2019
#SPECIES	2746	1956	582
#REACTIONS	11279	10371	21174
ISO-OCTANE	Y	Y	Y
N-HEPTANE	Y	Y	Y
TOLUENE	N	Y	Y
ETHANOL	Y	Y	Y
METHYLCYCLOHEXANE	Y	Y	Y
CYCLOHEXANE	Y	Y	Y
CYCLOPENTANE	Y	Y	N
1-HEXENE	Y	Y	Y

1-PENTENE	Y	Y	Y
ISO-PENTANE	Y	Y	N
124-TRIMETHYLBENZENE	N	Y	Y
N-DECANE	N	Y	Y

Due to hardware limitations, acknowledged by the same authors of the mechanism, the only feasible calculation of laminar flame speed with the general mechanism has been performed with the CRECK2019 mechanism, which has also been chosen for the research activity.

The choice of the molecular palette required has been defined after extensive literature review, to account with at least one molecule for all the main components that are identified in the pump gasoline, as well as for more research-oriented applications, as reported in Figure 2-2 where the typical composition of a European standard gasoline fuel was collected with gas-chromatography by [13]. The W and S letters in the molar fraction of cyclic paraffins indicate the winter and summer grades expected variations, with respect to the mean value.

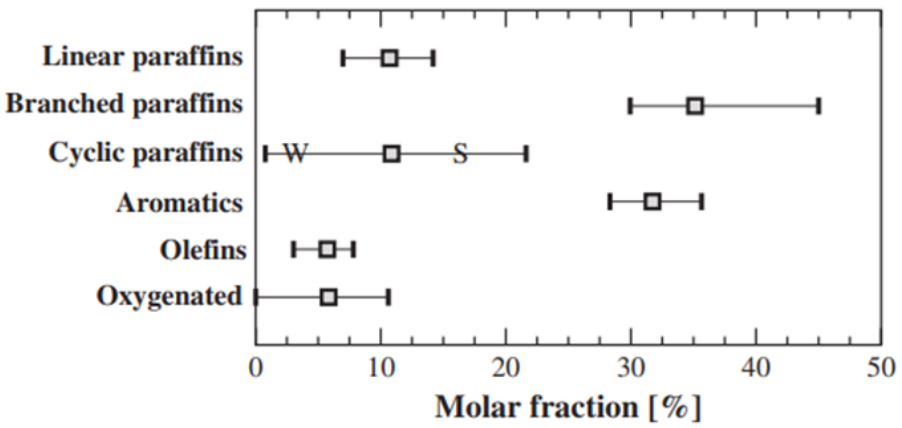

Figure 2-2: Generic composition of ULG95 European standard gasoline fuel [13]

The chosen molecules, together with their most relevant properties for the chemical kinetics analysis are collected in Table 1-2, where LHV and HHV correspond to the Lower and Higher heating value of the stoichiometric combustion, AFS is the mass ratio between oxidizer and fuel for the complete combustion and RON and MON are anti-knock indices that will be defined in more detail in Chapter 3.

Table 2-2: Main properties of the molecules constituting the reference palette

Component	Formula	Class	RON/MON	LHV-HHV	AFS
Iso-octane	iC_8H_{18}	I-Alkanes	100/100	44.611 - 48.079	15.028
n-pentane	nC_5H_{12}	N-Paraffins	62/62	45.338 - 48.998	15.227
n-heptane	nC_7H_{16}	N-Paraffins	0/0	44.531 – 47.668	14.686
Ethanol	C_2H_5OH	Alcohols	109/90	27.728 – 30.594	8.934
1-hexene	$1\text{-}C_6H_{12}$	Olefins	76/63.4	44.792 – 47.929	14.686
cyclohexane	$C6H_{12}$	Naphthenes	82.5/77.2	43.818 – 46.955	14.686
methylcyclohexane	C_7H_{14}	Naphthenes	74.1/71	43.723 – 46.861	14.686

toluene	$C_6H_5CH_3$	Aromatics	118/103.5	40.93 – 42.84	13.414
124-trimethylbenzene	$C_6H_5C_6H_5$	Aromatics	107.4/97.9	41.315 – 43.387	13.582
1-pentene	$1\text{-}C_5H_{10}$	Olefins	90/77.1	44.815 – 47.952	14.686
n-decane	$nC_{10}H_{22}$	N-Paraffins	-15/0	45.338 – 48.998	15.227

The use of detailed chemical kinetics mechanism for hydrocarbon oxidation has been traditionally addressed by means of simplified problems, in order to focus more on the effects of the chemical pathways and reaction constants rather than on the species transport and potential discretization errors of more complex domains . Therefore, as will be described in further details in the next chapters, the reference problems that employ the direct integration of the chemical reactions are the constant volume, adiabatic, perfectly stirred reactor for 0D computations and the 1D steady, adiabatic, planar flame tube. The software library used to simulate these simplified domains is the Python binding to the open source Cantera package [14], which is capable of easily represent these reference problems as well as to deal with a set of relevant reaction families:

1) Elementary reactions, with pressure-independent rate coefficients, with a forward rate described by the classical Arrhenius form, where A is the pre-exponential factor, E_a the activation energy, b the temperature exponent and T and R the temperature and gas constant:

$$B + C \leftrightarrow D + E$$

$$R_f = [B][C] \cdot AT^b e^{-\frac{E_a}{RT}}$$

2) Three-body reactions, in the form $A + B + M \leftrightarrow AB + M$ where a generic molecule M acts as a stabilizer for the AB formation, or catalyst for the AB scission, with a given efficiency.
3) Falloff reactions, usually displayed by polyatomic dissociation reactions, for example a three-body reaction, whose reaction rate is linear to the molecular concentration of a generic molecule [M] at low pressure, but reduces this proportionality as pressure increases. The same pressure dependency can be displayed by other reactions, where the collisional stabilization of the reaction intermediates increases with pressure, thus reducing the overall reactivity.

2.2 Machine learning general definitions

Machine learning is a branch of artificial intelligence concerning a series of computer algorithms developed to leverage available information in order to perform tasks for which they were not explicitly programmed. The operations associated with machine learning are traditionally divided into two different tasks [15]: supervised and unsupervised learning. Supervised learning refers to the availability for the training of the model of examples for which the true output is available, while in the unsupervised problems, the output is not known a-priori. The tasks associated with supervised learning can be further divided into 2 groups, represented graphically for a simplified problem on the left side of Figure 2-3:

- Regression, when the target output is a continuous variable, for the example in figure with 1 input feature (x) and the known target output (y).
- Classification, when the target output is a category (therefore with 2 or more possible outcomes), for example on the right side of Figure 2-3 with 2 features (x,y) describing the points and the known category they belong to.

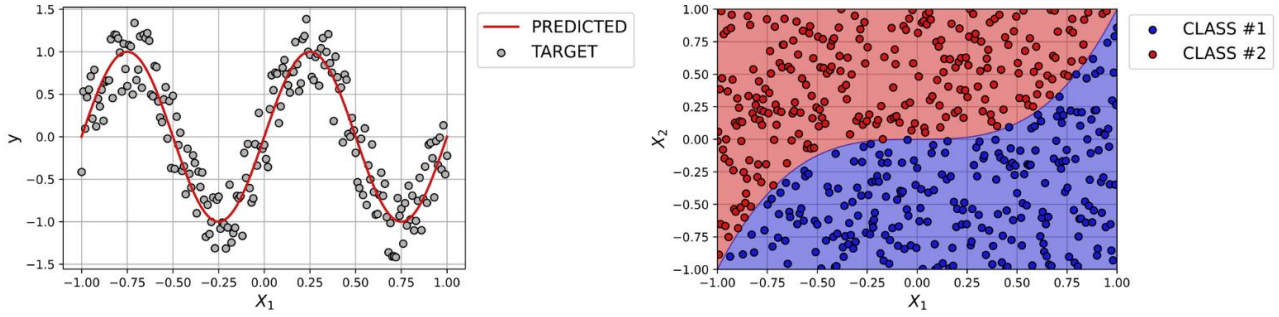

Figure 2-3: simplified regression and classification problem descriptions

As far as unsupervised learning is concerned, 2 main tasks are usually concerned:

- Clustering, when the objective is to perform an efficient and representative grouping of the available data, without the need to use a target variable or class known a-priori.
- Dimensionality reduction, when the objective is to find a set of independent variables that best represent the system, reducing the number of necessary features.

A third kind of machine learning task is called reinforcement learning, and it is concerned with the development of smart 'agents' that are trained to identify the best actions in a given environment by performing a trial and error training. As long as the system, or the environment, is defined in a way to provide feedback (positive or negative) to the agent, this group of algorithms are trained to become more and more efficient in choosing the best set of operations to perform a given task.

Deep learning, on the other hand, can be defined as a sub-field of machine learning, with which it shares all the previous definitions, but the models are all in the form of artificial neural networks. This architecture is more versatile, and allows, with the appropriate transformations, to operate with inputs of different format, not only tabular data (continuous or categorical data), but also with n-D arrays as a whole. For example, a time series (1-D array) should be divided into N features (with N equals the length of the array) for a traditional machine learning algorithm, while it can be regarded as a single feature input to some kinds

of neural network (for example 1D convolutional neural network) that will extract the most relevant features from the signal during training.

The task that is more closely associated with the engineering practice, and that will be more thoroughly employed in the present thesis is regression. The most relevant algorithms employed for a regression task will be briefly discussed in the following sections, while those employed for each experiments will be further outlined it their respective chapter.

2.2.1 Regression algorithms

As previously defined, a regression model is aimed at predicting a continuous target variable (y) from a set of continuous or categorical features X, for which examples are available. This task is performed by training the model at minimizing a given objective function (MAE, MSE, RMSE, MAPE, MSLE, PHLOSS) for which, given N the number of elements tested, y_i the i-th target variable and $\tilde{y}_i$ the i-th predicted variable, the definitions are as follows:

Mean Absolute Error

$$MAE = \frac{1}{N}\sum_{i=1,N} |y_i - \tilde{y}_i|$$

Mean Squared Error

$$MSE = \frac{1}{N}\sum_{i=1,N} (y_i - \tilde{y}_i)^2$$

Root Mean Squared Error

$$RMSE = \sqrt{\frac{1}{N}\sum_{i=1,N} (y_i - \tilde{y}_i)^2}$$

Mean Absolute Percentage Error

$$MAPE = \frac{1}{N}\sum_{i=1,N} \left|\frac{y_i - \tilde{y}_i}{y_i}\right| \cdot 100$$

Mean Squared Logarithmic Error

$$MSLE = \frac{1}{N}\sum_{i=1,N} (\log(1 + y_i) - \log(1 + \tilde{y}_i))^2$$

Pseudo-Huber Loss

$$PHLOSS = \delta^2 \sum_{i=1,N} \left(\sqrt{1 + \left(\frac{y_i - \tilde{y}_i}{\delta}\right)^2} - 1\right)$$

The specificity of the Pseudo-Huber Loss function [16] is that it can be regarded as a combination of MAE and MSE (also called L1 and L2 loss functions), blended with the term δ. Its advantages are that it is convex near the target, making it more stable at converging than L1 loss, but at the same time it is more robust to outliers than the L2 loss (depending on the chosen value of δ). These differences can be seen in Figure 2-4 where the loss function are represented as a function of the normalized error committed.

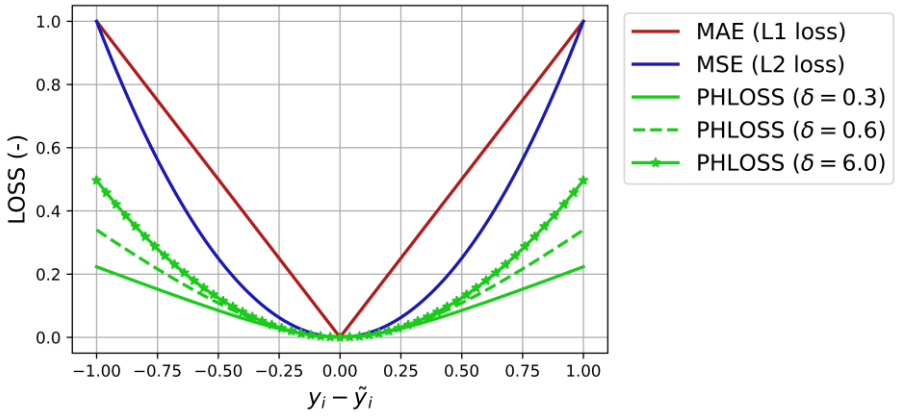

Figure 2-4: Effect of different loss function definition on the sensitivity to higher errors

2.2.2 Machine learning regression task anatomy

Before going further into the details of each model, particular attention should be given to the steps that they all have in common in order to improve the efficiency of the algorithm. The workflow typically followed for a regression task that leverages machine learning algorithms can be summarized as follows:

1) Data collection, backed by expert's knowledge, this step is of paramount importance, in order to provide a set of relevant input features, usually indicated with matrix **X** and targets, usually identified with matrix **y**, that can be useful for the solution of engineering (or any other kind of) problems. The number of entries in the dataset should be much greater than the number of features, in order to allow the models to perform a sufficient amount of optimization steps and identify all possible relations.

2) Data pre-processing, which can be further divided into 5 steps:

    a. Missing data handling, in case some features or targets are unavailable for some entries, these must be identified and corrected, if possible, by experts, or removed from the dataset. Typical solutions to this problem, beside removing either the entry or the feature, are to assign to these missing data the mean value of the feature in the whole dataset or of a subset (eventually identified by clustering techniques).

    b. Multicollinearity identification, in order to make the prediction algorithm more robust, it is considered a good practice to reduce the number of features in case some of them are highly correlated, considering that only a minimal loss of information is expected to occur, and the system is expected to be more robust against random errors in the targets. The possibility of 2 target points being randomly different with the same input variable would, in fact, be considered twice in case there was a highly correlated feature, thus producing a higher deviation from the optimal solution. The correlation between features can be calculated by using different metrics, the most common being Spearman's rank coefficient and Pearson's calculated as $R_P = 1 - \frac{6\sum_i d_i^2}{n(n^2-1)}$ where **x** and **y** transformed into coupled ranks, for which d_i is

their difference. The Spearman's coefficient is considered more general, since it is not based on the assumption of linear dependence between the 2 features considered, but it only requires the function to be monotonic.

c. Feature engineering and outlier removal, in order to provide more useful features through transformations of single features or combinations based on automatic exploration or expert evaluation on the physical meaning of the variables and removing data points that lay outside confidence intervals due to possible collection errors. In addition, in case some input variables are of categorical type, a label encoding could be performed, in order to avoid the possibility that the regression model could infer fallacious ordinary properties of the labels.

d. Dataset split between train and validation set, in order to verify the performance of the whole prediction pipeline (scaling and prediction) on a hold-out portion of the available data (usually between 10 and 20% depending on the number of samples available).

e. Feature scaling, aimed at reduce the difference between the dimensionality if different features that would, otherwise, induce unjustified higher sensitivity to some inputs. Classical scaling strategies are the standard scaling, which generates a variable distribution with mean value at zero and unitary standard deviation, range scaling (for example minmax scaling), which generates a distribution limited between 0 and 1, logarithmic scaling, used for features with different orders of magnitude and robust scaling, which is similar to standard scaling, but more robust to outliers, since it employs the median value and the percentiles instead of mean and standard deviation. The scaler's parameters must be defined on the train set, and only applied on new points to be evaluated.

Standard Scaling	$\widehat{\boldsymbol{x}_\iota} = \frac{(\boldsymbol{x_i} - \bar{x}_\iota)}{\sigma_{\boldsymbol{x_i}}}$
MinMax Scaling	$\widehat{\boldsymbol{x}_\iota} = \frac{\left(\boldsymbol{x_i} - min(\boldsymbol{x_i})\right)}{(max(\boldsymbol{x_i}) - min(\boldsymbol{x_i}))}$
Log Scaling	$\widehat{\boldsymbol{x}_\iota} = log(1 + \boldsymbol{x_i})$
Robust Scaling	$\widehat{\boldsymbol{x}_\iota} = \frac{(\boldsymbol{x_i} - median(\boldsymbol{x_i}))}{p_{75}(\boldsymbol{x_i}) - p_{25}(\boldsymbol{x_i})}$

3) Model definition and optimization which can be performed through heuristics or by leveraging optimization algorithms for deciding the best hyperparameters or architecture of the regression model. Depending on the number of features available and on the size of the dataset, it can be decided to remove from the training set a portion of the data to be considered as test set for evaluating the performance of the model during its optimization. Alternatively a k-fold cross validation can be employed, which requires the split of the available dataset into a number k of sub-groups, and

retrain the model on k-1 sub-groups while using the left-out group as test set for k-1 times. The final score of the model is the average of the performance of all evaluations, which makes the evaluation on the performance of the model less sensitive to the choice of the train test split, especially useful when the number of available points is small. The graphical example for a cross-validation with 5 folds is reported in Figure 2-5.

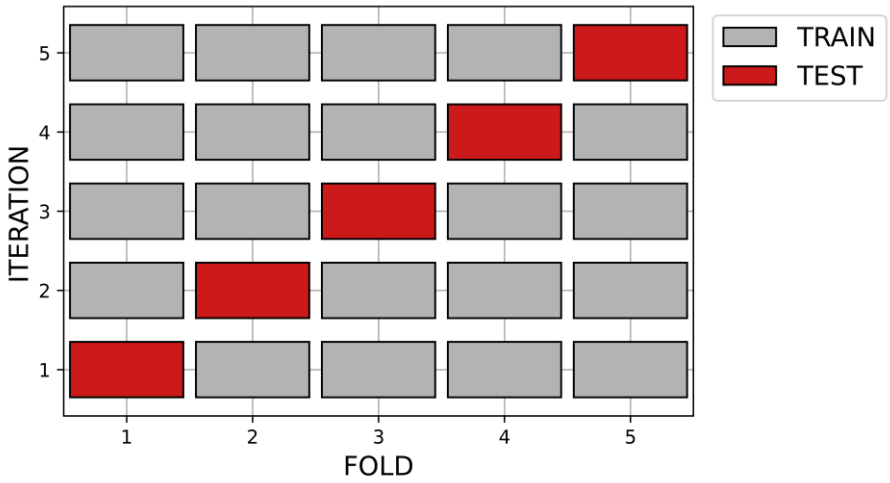

Figure 2-5: Simplified k-fold cross validation scheme

In order to optimize the prediction algorithms, several hyperparameters (i.e. characteristics of the machine learning model that are not optimized during training, but influence its behaviour) can be leveraged. Besides expert's hypothesis, hyperparameters' choice is often performed with 2 different approaches [17]:

a) Grid search optimization, which performs training and evaluation of the performance of a model for all the combinations of parameters considered relevant. This approach can be extremely time-consuming, but it ensures that all possible configurations will be evaluated.

b) Global optimization, which can be implemented with different algorithms (for example Bayesian optimization [18]) and it is aimed at minimizing an unknown function, in this case the performance of the regression model on the test set (or cross validation performance) given a set of hyperparameters.

4) Pipeline validation on the hold-out portion of the dataset defined at step 2d. In order to assess as accurately as possible the ability of the defined models to perform predictions on unseen data, this process should be used only to once, and not repeated to optimize the model. In case the performance was not satisfactory, the entire workflow should be re-considered, and not only the model, otherwise it would risk to overfit the validation dataset instead of validating its robustness against new points.

A short introduction to the most relevant algorithms for regression problems will be presented, in order to provide the foundational knowledge to understand the specificities of each of them and the reasons behind the choice of one over the others that have been taken during the following work.

2.2.3 Regression models

Linear Regression

Linear regression is possibly the most classical example of data-driven model, widespread due to the intrinsic simplicity in calculating its coefficients without the need of high computing power. The model calculates the output target as a linear combination of the input features, fitting weights and biases to minimize the L2 loss function, in the form:

$$\min_{\boldsymbol{w},\boldsymbol{b}} (\boldsymbol{X} \cdot \boldsymbol{w} + \boldsymbol{b} - \boldsymbol{y})^2$$

Where the linear coefficients **w** and intercepts (or biases) **b** have the same number of elements as the number of features in matrix **X**. Particular care must be given in the use of linear regression models with multiple features, to avoid collinearity between the inputs, which would make it less robust to errors in the target variables.

Elastic-Net regression

Elastic-net regression is a particular kind of linear model trained not only at minimizing the distance between predicted and real values, but also a linear combination of L1 and L2 regularization losses of the coefficients **w** [19]. The general form, with α as regularization coefficient and ρ as L1 to L2 ratio is:

$$\min_{\boldsymbol{w},\boldsymbol{b}} \left(\frac{\mathbf{1}}{\mathbf{2N}} (\boldsymbol{X} \cdot \boldsymbol{w} + \boldsymbol{b} - \boldsymbol{y})^{\mathbf{2}} + \alpha\rho \sum |\boldsymbol{w_i}| + \frac{\alpha(1-\rho)}{2} \sum \boldsymbol{w_i^2} \right)$$

The main advantages of Elastic-Net over traditional linear models is its ability to represent the solution with the least number of relevant weights, thus reducing the risk of overfitting and increasing the explainability of the model.

Polynomial regression

Polynomial regression is an extension of the linear models based on the assumption that, creating an improved set of features, containing also the exponentials of the original input, a linear model can be optimized for predicting also non-linear dependencies. For example, if the classical linear regression model can be described as:

$$\tilde{\boldsymbol{y}}(\boldsymbol{w}, \boldsymbol{x}) = \boldsymbol{w_2 x_2} + \boldsymbol{w_1 x_1} + \boldsymbol{w_0}$$

If we define a new set of variables from $\boldsymbol{x_1}$ and $\boldsymbol{x_2}$, as $\boldsymbol{u} = [\boldsymbol{x_1}, \boldsymbol{x_2}, \boldsymbol{x_1 x_2}, \boldsymbol{x_1^2}, \boldsymbol{x_2^2}]$ the linear regression model can be described as:

$$\tilde{\boldsymbol{y}}(\boldsymbol{w}, \boldsymbol{u}) = \boldsymbol{w_5 u_5} + \boldsymbol{w_4 u_4} + \boldsymbol{w_3 u_3} + \boldsymbol{w_2 u_2} + \boldsymbol{w_1 u_1} + \boldsymbol{w_0}$$

Which can also be seen as:

$$\tilde{\boldsymbol{y}}(\boldsymbol{w}, \boldsymbol{x}) = \boldsymbol{w_5 x_2^2} + \boldsymbol{w_4 x_1^2} + \boldsymbol{w_3 x_2 x_1} + \boldsymbol{w_2 x_2} + \boldsymbol{w_1 x_1} + \boldsymbol{w_0}$$

Support Vector Machine regression

The support vector algorithm was initially developed for classification tasks, but it can also be employed in regression problems with some adaptation in the optimization steps.

Differently from linear regression, a support vector machine identifies a linear correlations between features and targets while, at the same time, identifying a margin within which the prediction can be considered acceptable [20] .
The most relevant hyperparameter that the model relies on is the choice of a suitable kernel (linear, polynomial or radial basis function) to map the data into a higher dimensional space in which it is possible to perform a linear separation.

Gaussian Process regression

The Gaussian process regression algorithm is a specific model capable of predicting the target value but also the uncertainty in the prediction, by predicting the probability distribution over all the possible functions that fit the training data (linearly or based on a kernel). Thanks to the availability of the probability distribution, the mean value can be computed, to perform the prediction, and the standard deviation can be used to describe the confidence of the prediction [21].
The regression model of the Gaussian Process is given by:

$$P(\boldsymbol{f}|\boldsymbol{X}) = \mathcal{N}(\boldsymbol{f}|\boldsymbol{\mu}, \boldsymbol{K})$$

Where $\boldsymbol{X} = [\boldsymbol{x_1}, \boldsymbol{x_2}, \dots, \boldsymbol{x_n}]$ are the input features, $\boldsymbol{f} = [f(\boldsymbol{x_1}), f(\boldsymbol{x_2}), \dots, f(\boldsymbol{x_n})]$ are the function outputs of the model, $\boldsymbol{\mu} = [m(\boldsymbol{x_1}), m(\boldsymbol{x_2}), \dots, m(\boldsymbol{x_n})]$ are the men values and $\boldsymbol{K_{ij}} = \boldsymbol{k}(\boldsymbol{x_i}, \boldsymbol{x_j})$ where *k* represents a positive definite kernel function.
A classical implementation of the GPR algorithm is as follows:

$$L = Cholesky(K + \sigma_n^2 I)$$
$$\boldsymbol{\alpha} = L^T \backslash (L \backslash y)$$
$$\overline{\boldsymbol{f}_*} = \boldsymbol{k}_*^{\boldsymbol{T}} \boldsymbol{\alpha}$$
$$V = L \backslash \boldsymbol{k}_*$$
$$Var(\boldsymbol{f}_*) = k(\boldsymbol{x}_*, \boldsymbol{x}_*) - \boldsymbol{v}^{\boldsymbol{T}} \boldsymbol{v}$$
$$Log\ p(\boldsymbol{y}|\boldsymbol{X}) = -\frac{1}{2} y^T (K + \sigma_n^2 I)^{-1} \boldsymbol{y} - \frac{1}{2} log\ det(K + \sigma_n^2 I) - \frac{n}{2} \log(2\pi)$$

With X as inputs, y as targets, k the covariance function, σ_n^2 the noise level, $\boldsymbol{x}_*$ is the test input, whereas the outputs are: $\overline{\boldsymbol{f}_*}$ that represent the mean of the outputs, $Var(\boldsymbol{f}_*)$ its variance and $Log\ p(y|X)$ is the log marginal likelihood of the solution. In Figure 2-6, 3 different predictions of the confidence interval are reported without noise in the data, representing the output of the pure prior distribution, and with 1 and 6 points, for which the kernel is capable of perfectly predict the target as long as the prediction point is within the training interval. In case of a noisy distribution of the targets, Figure 2-6 d, the GPR does not perfectly fit the sample points, because an additional term has been added to the kernel matrix, which can be identified as the variance of additional gaussian measurement noise.

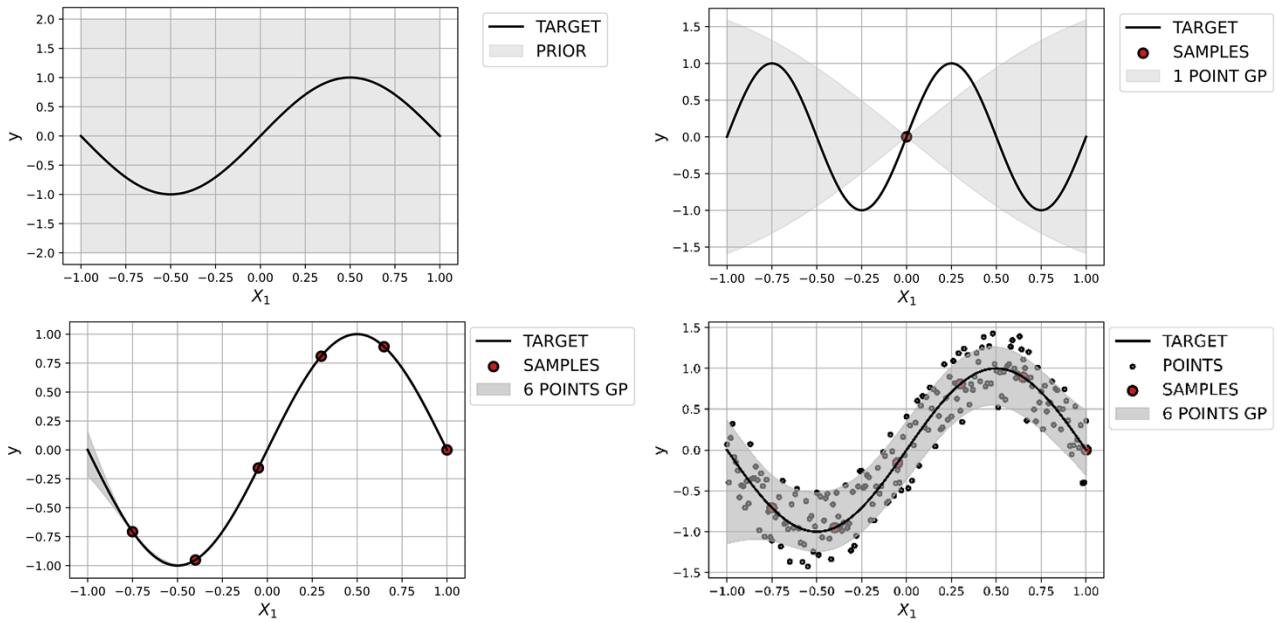

Figure 2-6: Improvement of the predictive capability of GP regressor prior with additional training points

Decision Tree regression

A decision tree regression algorithm is a non-parametric model used to predict the target variable by using a piecewise constant approximation, whose conditions are learned by the training data [22]. The decision tree regressors' hyperparameters are concerned with the definition of the splits and values of each leaf.

The algorithm states that, for example, if we have N_m samples at each node *m,* there is a candidate split couple $\theta = (j, t_m)$ where *j* is a given feature and t_m its threshold value, that can divide the data at the node in left and right subsets. The choice of the split based on the couple $\theta = (j, t_m)$ at node m is defined by a loss function L (for example RMSE) that must be minimized and must satisfy the condition:

$$G(Q_m, \theta) = Frac_{left} \cdot L(Q_{left}(\theta)) + Frac_{right} \cdot L(Q_{right}(\theta))$$

Where Q_m is the total set of points at node m, Q_{left} and Q_{right} its subsets based on split threshold θ, that must be found as:

$$\theta^* = argmin_\theta \, G(Q_m, \theta)$$

The hyperparameters that can further describe the regression tree algorithm are related to controls over the possibility to compute new nodes, by limiting the depth that a nested condition can reach, or by defining a minimum number of examples that must be found in a subset.

Ensemble Methods

Classical machine learning algorithms are best suited for specific kinds of problems, but might become less robust when the task is not perfectly defined, therefore a common strategy to improve generalizability of the predictions and reduce the variance that a single estimator might display is to combine the predictions of several models. The most straightforward approach, i.e. to perform an average of the predictions of the chosen set of algorithms (called base learners in the context of ensemble) is called 'bagging' [23]. A bagging strategy can be developed to combine different base algorithms, but also the same

algorithm trained on different portions of the dataset, in order to induce a variation in the optimized parameters. On the other hand, a different approach is called 'boosting', and it describes the sequential training of base estimators on modified versions of the dataset, aimed at compensating the errors of the previous models. Both techniques have been found to be extremely useful for several kinds of problems, therefore they some of the most frequent algorithms will be described more in detail.

Random forest

The random forest algorithm is a bagging method consisting of a combination of several regression trees trained on a subset of the whole training dataset and using, for each tree, all input features at each node [24]. These 2 choices are both sources of randomness in the training process, that are aimed at reducing the risk of overfitting that traditional regression trees usually display. The combination, by averaging the outputs, of all the base learners is expected to compensate the increased bias with a significantly reduced variance.

Gradient boosting

The gradient boosting method is a general name for a family of ensemble algorithms where the base learners are recursively trained on the residuals of the previous models, in order to 'correct' the predictions of the previous models [25]. The loss function that must be minimized by the regression algorithm is expected to be differentiable, in order to optimize the weights of the models by searching for the minimum of the overall error function of the ensemble model. The gradient descent algorithm followed is the following:

Define the initial set of parameters $\theta = \theta_0$

For every iteration it (until convergence or maximum number of steps allowed), calculate the gradient of the loss function as:

$$\nabla L_\theta(\theta^{it}) = \left[\frac{\partial L\big(y, f(x,\theta)\big)}{\partial \theta}\right]_{\theta=\theta^{it}}$$

Update the weights with the gradient of the loss function multiplied by a learning rate value ρ (coefficient to stabilize the optimization updates:

$$\theta^{it+1} = \theta^{it} - \rho \nabla L_\theta(\theta^{it})$$

The final model will, therefore use the final updated version of the weights θ^N, and the same prediction algorithm and base learners used for the computation of the loss function.

A particularly efficient implementation of the gradient boosted algorithm has gained particular attention recently, especially do to its capacity to handle larger datasets more efficiently than the standard gradient boosting implementations, named XGBoost [26].

Adaptive boosting

The AdaBoost model is considered a very effective boosting technique that draws on any base learner [27], given the hypothesis that their prediction is expected to be, at least, slightly more accurate than random guessing. If the hypothesis holds true, then the following algorithm, represented graphically in Figure 2-7 also known as AdaBoost.R2 (considering that the original AdaBoost implementation was developed for classification tasks) will recursively increment the performance of the whole model:

Define an initial value for each sample's weight, with N the number of sample points.

$$w_i = \frac{1}{N}, \quad i = 1,2,..,N$$

For a number of iterations (it) equals to the number (m) of base learners chosen (identified by f(x)), repeat:

$$l^{it}(i) = |f(x_i) - y_i|$$

$$\overline{L^{it}} = \sum_{i=1,N} l^{it} \cdot w_i$$

$$\beta^{it} = \frac{\overline{L^{it}}}{1 - \overline{L^{it}}}$$

$$w_i^{it+1} = \frac{w_i^{it}\, \beta^{1-l^{it}(i)}}{\sum_{i=1,N} w_i^{it}\, \beta^{1-l^{it}(i)}}, \quad i = 1,2,..,N$$

After the algorithm has been repeated for the number of iterations chosen, the output value is calculated.

$$y = inf\left[y \in Y: \sum_{it:f_t(x) \le y} log\left(\frac{1}{\beta^{it}}\right) \ge \frac{1}{2} \sum_{it=1,m} log\left(\frac{1}{\beta^{it}}\right) \right]$$

In this case, the Loss function is considered to be a mean absolute error, but it could eventually be modified. As it results from the definition, $\overline{L^{it}}$ is the average error committed by the weak learner at iteration *it*, following a distribution of weights which is initialized as uniform. β^{it} is a term representing the confidence of the prediction at the iteration *it* and it is used to update the weights that will be used during the next model training step in order to provide exponentially more importance to the dataset points that were less captured by the current model. The lower the value of β^{it}, the higher the confidence of the current model, and therefore a smaller weight update will be performed. The output result is a weighted median calculation of the predictions of all the possible solutions provided by the weak learners, using the logarithm of the inverse of their average confidence as weight.

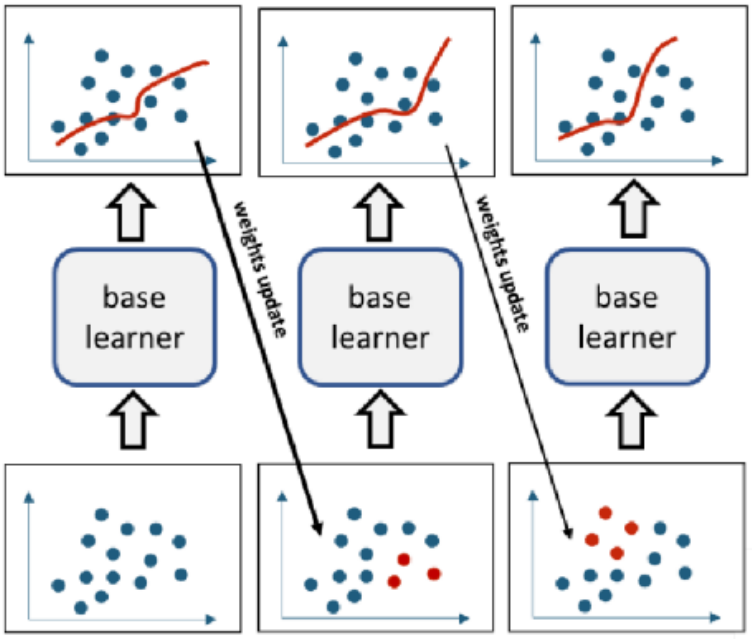

Figure 2-7: AdaBoost model training workflow

Neural networks

Neural networks are a set of algorithms inspired by the biological neurons as far as the terminology is considered, whose structure is reported in Figure 2-8.

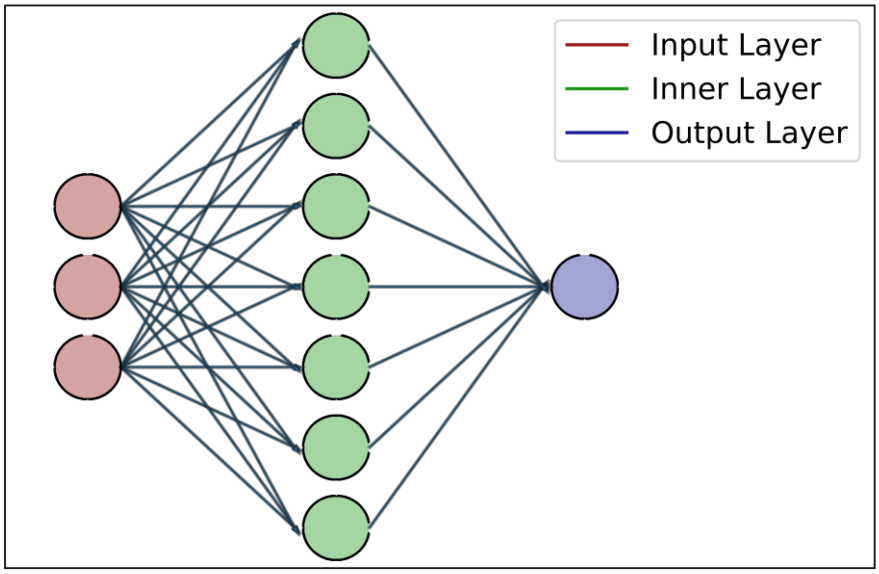

Figure 2-8: Generic representation of a 1-hidden layer neural network

They consist of a set of neurons, single cells, defined by a set of inputs x_i and one output variable y_i calculated as:

$$y_i = f\left(\sum_{j=1}^{n} w_j x_j + b\right)$$

Where *f* is called activation function, and can take the form of a hyperbolic tangent, or linear, or sigmoid functions for example, whose input is a linear combination of the inputs received from the neuron, scaled by an array of weights and a bias term. For a regression task, the output layer consists of a number of neurons equal to the number of target variables to predict simultaneously, and the activation function is expected to be a linear function, in order to obtain a scaled real value. Considering that the number of neurons and layers can be arbitrarily large, neural networks are considered general approximators of a target distribution. The optimization of weight and biases for each neuron is performed using different implementations of the gradient descent algorithm, based on the problem definition in the form of a minimization problem for which the derivative is known. The capability to calculate the derivative of the loss function with respect to each weight is due to the back propagation algorithm [28], based on the chain rule of partial derivatives, which is particularly efficient to compute when the activation functions are convenient to differentiate, which is the main reason why the most popular choices are the ones represented in Figure 2-9 where also their derivatives are represented. Besides the linear function, which is mostly used for the output neuron in regression tasks, the other activation function are all non-linear and differentiable.

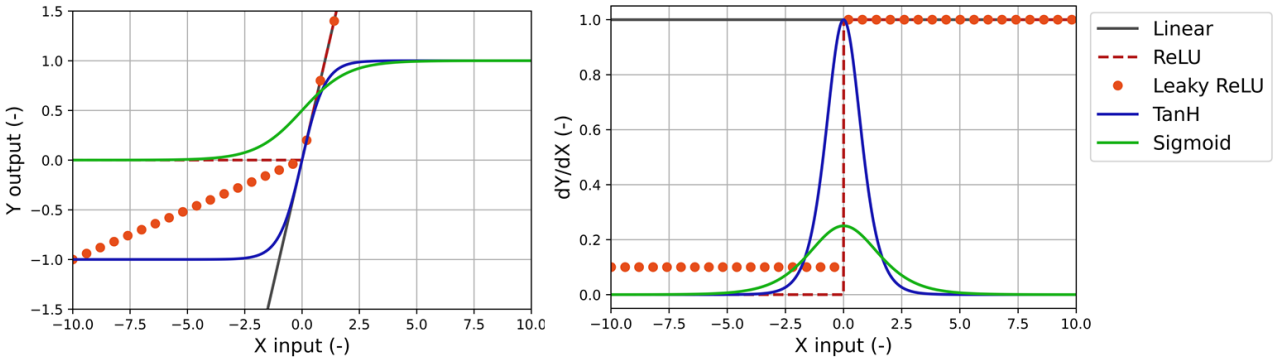

Figure 2-9: Typical activation functions for neural network neurons and their derivative

Defining as E the error performed by the neural network with output y with respect to target t, and for each neuron j defining the output as o_j calculated as: $o_j = \phi\left(\sum_{k=1,n} w_{kj} o_k + b_j\right)$ where ϕ is the activation function applied to the weighted sum of all the inputs of the neuron (outputs from the previous layer). The partial derivative of the error with respect to the weights can be computed following the chain rule as:

$$\frac{\partial E}{\partial w_{ij}} = \frac{\partial E}{\partial o_j}\frac{\partial o_j}{\partial w_{ij}} = \frac{\partial E}{\partial o_j}\frac{\partial o_j}{\partial net_j}\frac{\partial net_j}{\partial w_{ij}}$$

Where $net_j = \sum_{k=1,n} w_{kj} o_k + b_j$, which leads to the simplified partial derivative $\frac{\partial net_j}{\partial w_{ij}} = o_i$ that corresponds to x_i for the first layer. At the same time, a recursive formulation can be derived for the term $\frac{\partial E}{\partial o_j} = \sum_{l \in L}\left(\frac{\partial E}{\partial o_l}\frac{\partial o_l}{\partial net_l} w_{jl}\right)$ with L the set of all the neurons receiving input from neuron *j.* By following these observations, the partial derivative of the error with respect to each weight can be computed as:

$$\frac{\partial E}{\partial w_{ij}} = o_i \delta_j$$

With $\delta_j = \frac{\partial E}{\partial o_j}\frac{\partial o_j}{\partial net_j} = \frac{\partial \phi(net_j)}{\partial net_j} \cdot \sum_{l \in L}\left(w_{jl}\delta_l\right)$ for all inner neurons, and $\delta_j = \frac{\partial L(o_j,t)}{\partial o_j}\frac{\partial \phi(net_j)}{\partial net_j}$ for the output neurons, where *t* is the target and $L(o_j, t)$ is the loss function to be minimized. The gradient descent algorithm is then applied to identify the update on weights that will reduce the error *E*, proportionally to the error committed, with a proportionality factor named learning rate, which is one of the most relevant hyperparameters, connected with stability and performance of the final solution.

3 Knock modelling with chemical kinetics

Knock in SI engines is a phenomenon limiting performance and efficiency especially at high load conditions [29], thus it is extremely important to prevent its occurrence. It is defined as the spontaneous ignition of the reacting mixture due to the thermodynamics conditions reached inside the combustion chamber during combustion, with consequent nearly instantaneous localized heat release.

Historically, the auto-ignition tendency has been investigated by means of rapid compression machines, or shock tube experiments [30], where it is possible to control the thermodynamical and compositional properties of the analysed mixtures. More recently, the same experiments have been conceptually converted into simplified models geometries (single cell, homogeneous reactors) resolved by means of detailed chemical kinetics simulations successfully [31]. Despite the time to analyse the impact of different mixture formulations on the ignition delay times has been drastically reduced, it is still a limiting factor when used for near real-time applications, such as when coupled to a fluid dynamics solver , or to the acquisition function of a fuel surrogate optimizer. Considering that the latter is one of the potential applications that is going to be investigated in the current work, in the present chapter two main issues will be addressed:

1) The generation of ignition delay time values for a generic mixture, and the definition of a dataset for the exploration of the effect of different compositions of the fuel surrogate using molecules from the palette introduced in the previous chapter.
2) The definition of a workflow for the prediction of ignition delay time values for any combination of thermodynamics conditions and fuel surrogate composition and the optimization of the regression model.

3.1 Constant volume reactor model

Whereas the insurgence of mixture auto-ignition is usually a consequence of regions of the boundary walls not sufficiently cooled (so called hot-spots, especially near the exhaust valves) or carbon deposits, therefore at a localized level, the fuel octane number, intake air temperature and the spark timing have the highest influence in the possible knock occurrence from a design and control perspective.
The simulation by means of quasi dimensional multi-zone or CFD3D models can be a valuable tool to address these dependencies and their correlation with other control and design parameters, such as the chamber shape, injection strategy or cooling circuit requirements. On the other hand, they can also be used to tailor specific minimum requirements a fuel must present to comply with specific safety levels during its usage. Obviously, this is not the classical task that engine manufacturers face, considering that they must provide robustness only against potential variations in the pump fuels, but the current research targets a broader audience of research and development engineers whose task could also be to co-optimize fuel and engine to gain the most profit.
The chemical kinetics phenomenon that leads to engine knock is the mixture self-ignition that can be evaluated both experimentally and numerically in a very detailed way. Two kinds

of experimental procedures are usually employed [32] to identify the ignition delay time of a given mixture in a more controlled environment:

1) High pressure shock tube, where the mixture is constrained inside a constant volume tube subjected to reflected shock waves that increase the gas pressure and temperature to the prescribed temperature and pressure.
2) Rapid compression machine, which is a machine emulating the compression stroke of a single engine cycle in order to rapidly heat the gas to the desired temperature and pressure values, by keeping the heat losses to a minimum. This is the main reason why ignition delay times higher than 100 ms may present inconsistencies, due to the increased importance of heat losses through the walls.

The use of reflected pressure waves in shock tubes, and of the varying volume of the rapid compression machine are limiting factors in comparing precisely with a simplified numerical model the conditions that the mixture is undertaking before ignition occurs. Therefore, a simplified approach is usually adopted [33] to validate reaction rates and the behavior of different fuels, by considering a constant volume perfectly stirred ideal gas reactor [34]. The mixture is initialized at the reference pressure temperature and composition of the test and numerically integrated until the temperature inside the reactor reaches a threshold value, often indicated as 400 K above the initial temperature, to avoid considering first stage ignitions.
The equations that describe the 0D ideal gas constant volume reactors are the conservation of species and mass, and the energy conservation equation for a closed system:

$$c_V \frac{dT}{dt} = -\sum_k u_k \frac{dY_k}{dt} \tag{3-1}$$

It can be noticed that no heat transfer to the walls is considered, as well as mechanical work, highly simplifying the solution of the system of equations related to the temperature dependence of the heat release of the species. Considering that the system is closed, the species conservation can be simply written for each species as:

$$m \frac{dY_k}{dt} = \dot{m}_{k,gen} \tag{3-2}$$

The simplified methodology for the computation of ignition delay times consists in the integration of the system of ODEs defined by the chemical kinetics reactions and the energy and species conservation equations, starting from the initial reference state.
The identification of the ignition delay time depends on the definition of the main combustion event, which is relevant especially for molecules exhibiting a negative temperature coefficient (NTC) region, which may lead to first stage ignitions, particularly at higher pressure. This phenomenon is justified by the competing reactions of RO_2 isomerization, together with QOOH decomposition and β scission reaction of the alkyl-radicals competing with the specific low temperature chemistry chain reactions. An example of NTC behaviour is reported in Figure 3-1 where the temperature, fuel and O2 profiles are reported normalized

for n-heptane at 100 bar and 800K (dashed lines), with initial temperature increase and successive reduction and 1000 K as can be seen from Figure 3-1.

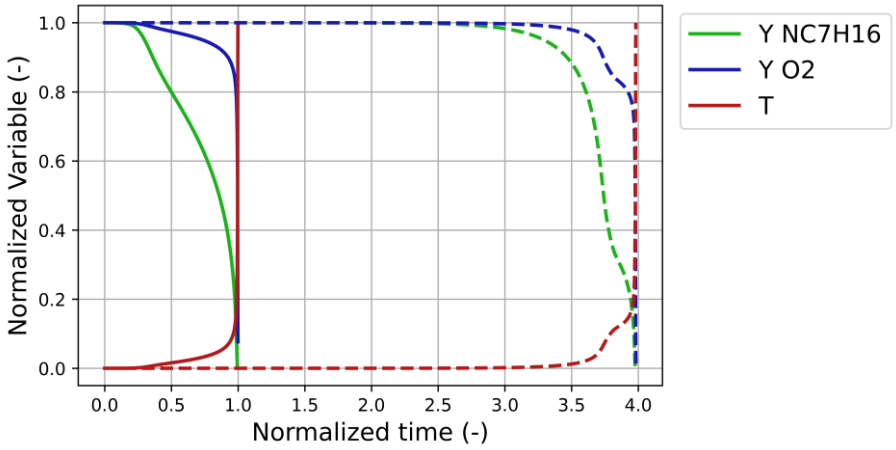

Figure 3-1: Dimensionless profile of selected variables evolution inside a constant volume ignition delay time calculation for starting temperature if 800K (dashed lines) and 1000 K (solid lines) and P=100 bar

Therefore, following [33], it was chosen to define a simplified approach to the definition of the ignition delay time, as the time required to raise the mixture temperature at constant volume by more than 400 K, considering the lack of information regarding each experimental facility used for the validation data. The validation of the chemical kinetics scheme adopted, for each of the single molecules of the palette, where this methodology was applied, will be reported in the next section with sufficiently accurate results.

3.2 Validation results

A set of experimental data found from different sources in literature has been compared with the results of detailed chemical kinetics simulations performed with the CRECK-2019 mechanism. Overall, the accuracy of the mechanism can be considered acceptable for the intended use for most basic molecules, for which a basic validation has been performed (Figure 3-2 to 3-13), considering the wide availability of validation studies for this mechanism. It can be noticed, however a general underestimation of the reaction speed at low temperature for the aromatics components toluene and 124-trimethylbenzene as well as for the low temperature chemistry of ethanol. The application on experimental fuel surrogates with varying ethanol fraction [35], in the formulation reported in Table 3-1 however, demonstrates with a good agreement of the results, the predictivity of the scheme also for ethanol containing fuels up to E40 (Figure 3-13).

Table 3-1: Definition of the fuel surrogates for the chemical kinetics scheme validation and their RON

Gasoline surrogate	Liquid volume fraction (%)				RON
	Ethanol	Iso-octane	N-heptane	Toluene	
Surrogate-A (E40)	40.0	37.8	10.2	12.0	98.75
Surrogate-B (E20)	20.0	62.0	18.0	-	92.0
Surrogate-C (E0)	-	69	17.0	14.0	87.0

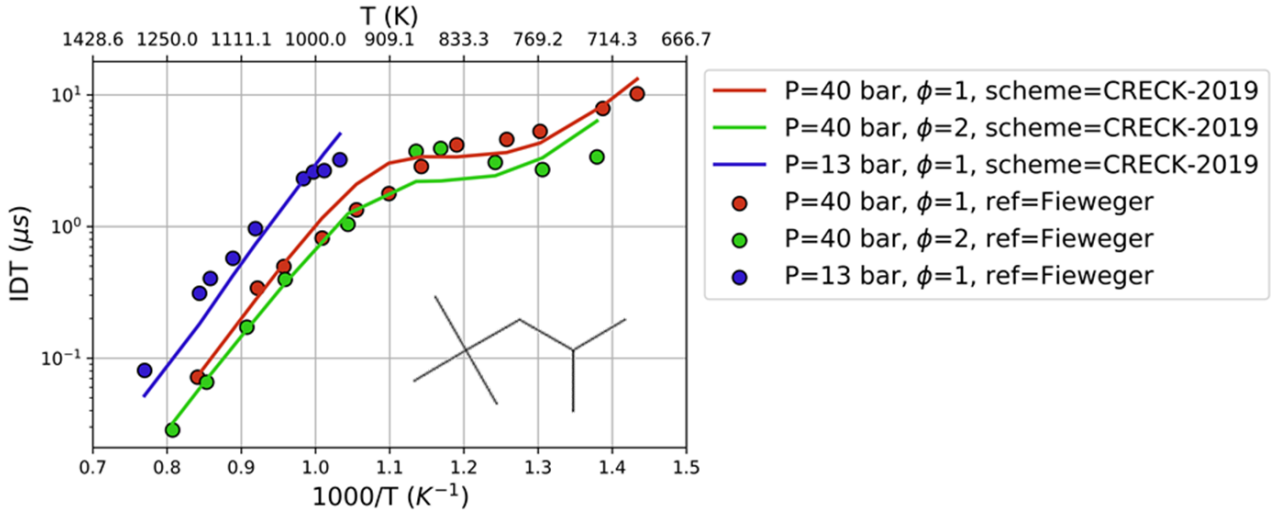

Figure 3-2: Validation data for the ignition delay time of iso-octane in air, Fieweger [36]

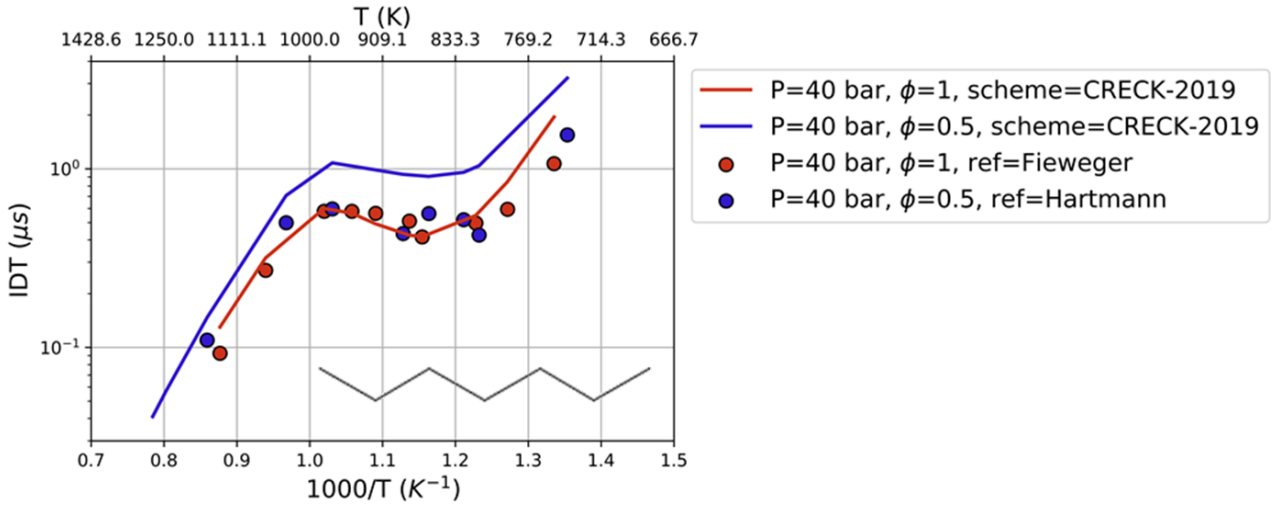

Figure 3-3: Validation data for the ignition delay time of n-heptane in air, Fieweger [36], Hartmann [37]

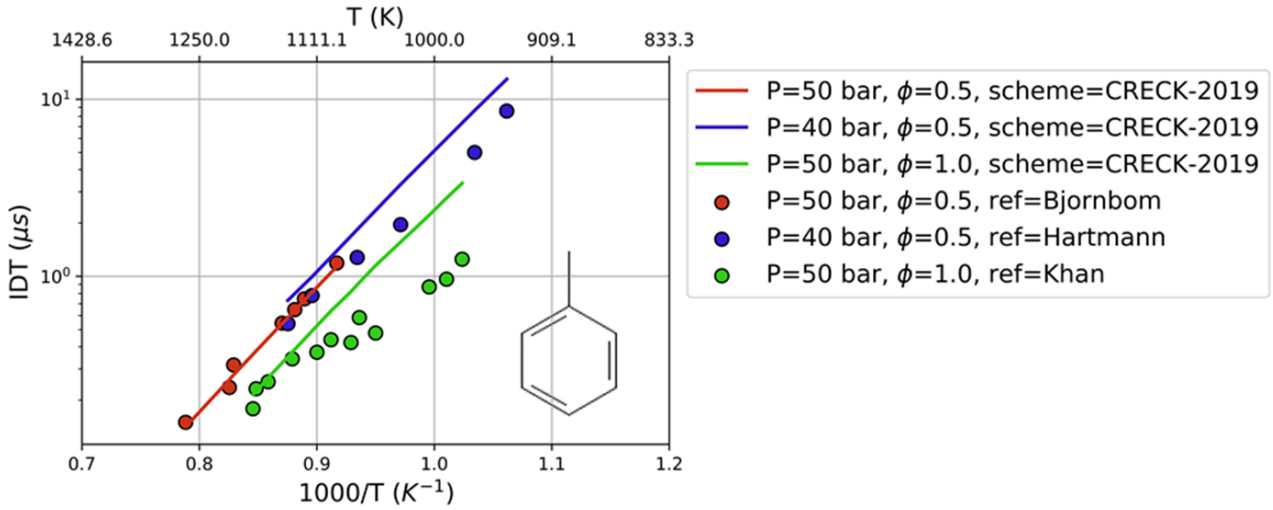

Figure 3-4: Validation data for the ignition delay time of toluene in air, Bjornbom [38], Hartmann [37], Khan [39]

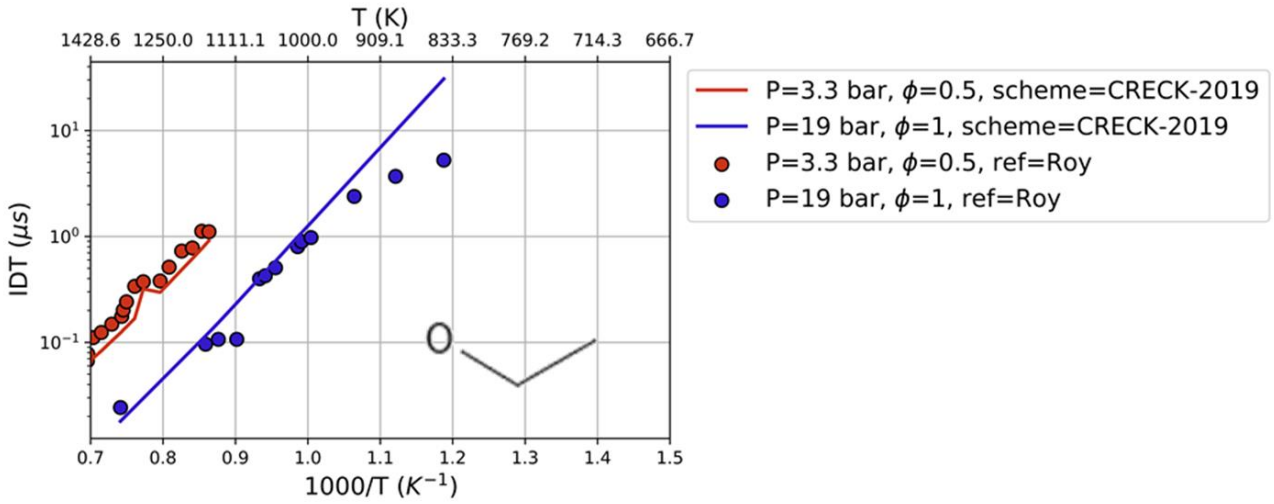

Figure 3-5: Validation data for the ignition delay time of ethanol in air, Roy [40]

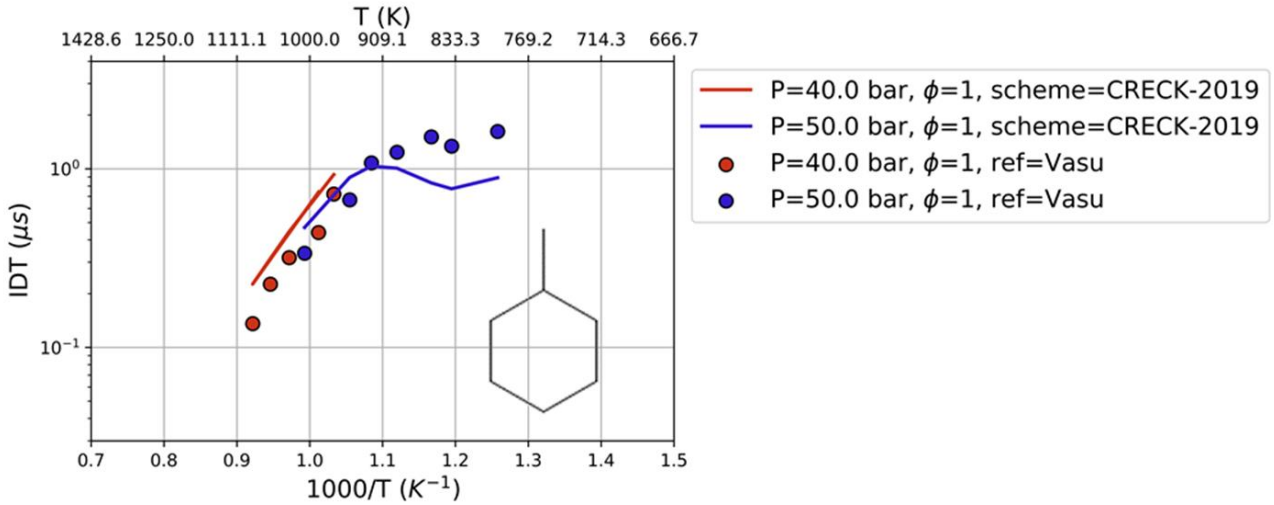

Figure 3-6: Validation data for the ignition delay time of methylcyclohexane in air, Vasu [41]

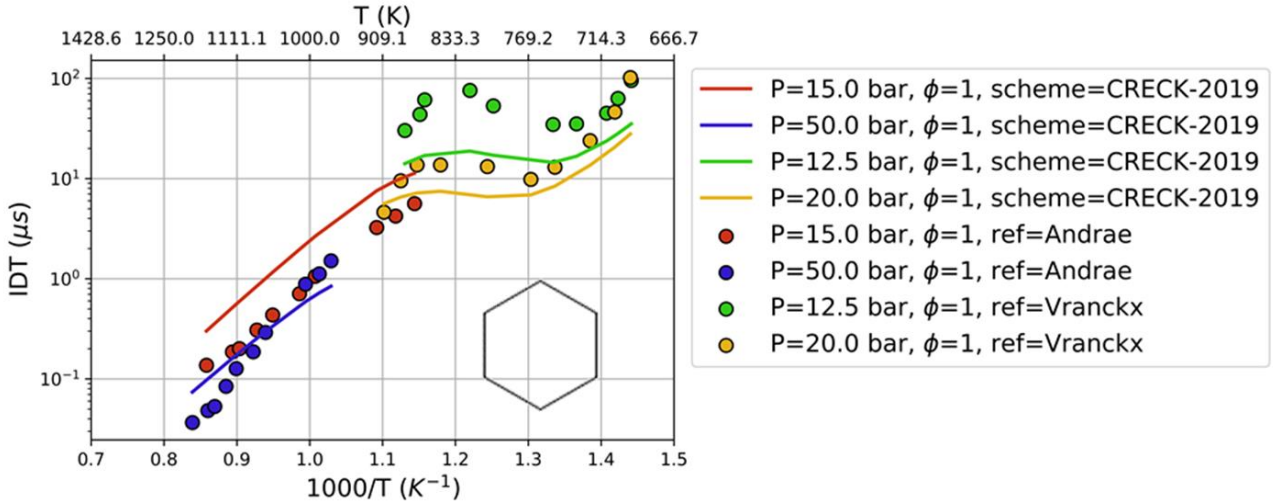

Figure 3-7: Validation data for the ignition delay time of cyclohexane in air, Andrae [42], Vranckx [43]

Validation results

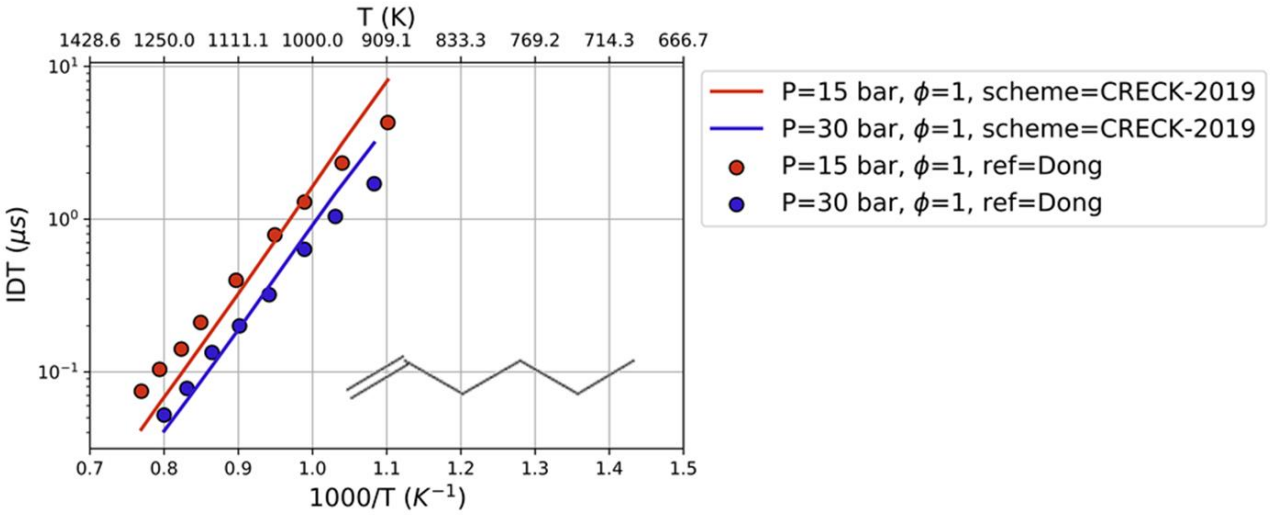

Figure 3-8: Validation data for the ignition delay time of 1-hexene in air, Dong [44]

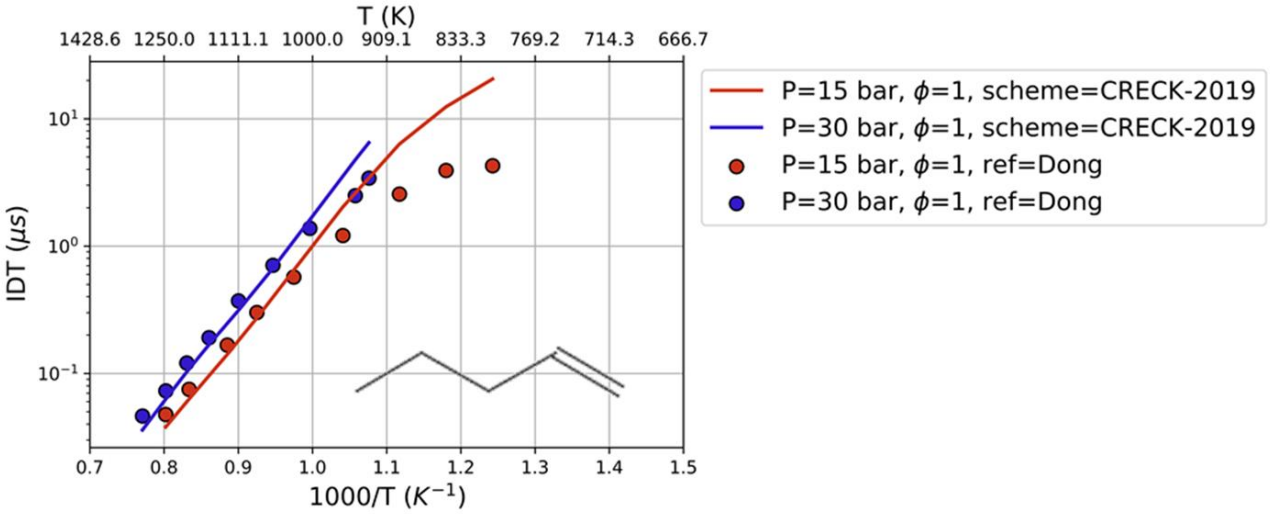

Figure 3-9: Validation data for the ignition delay time of 1-pentene in air, Dong [44]

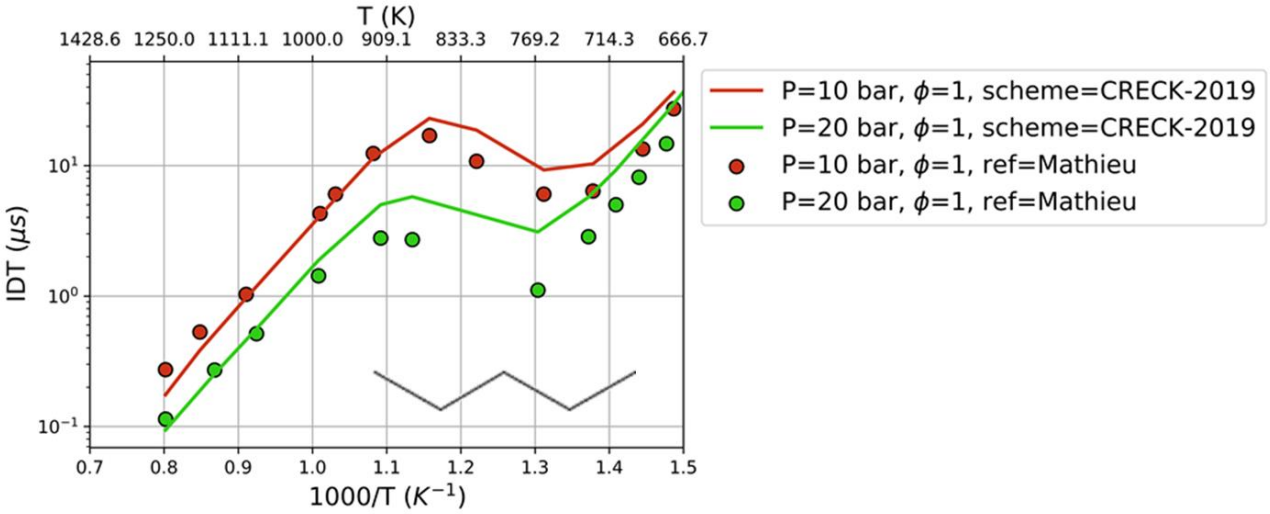

Figure 3-10: Validation data for the ignition delay time of n-pentane in air, Mathieu [45]

Validation results

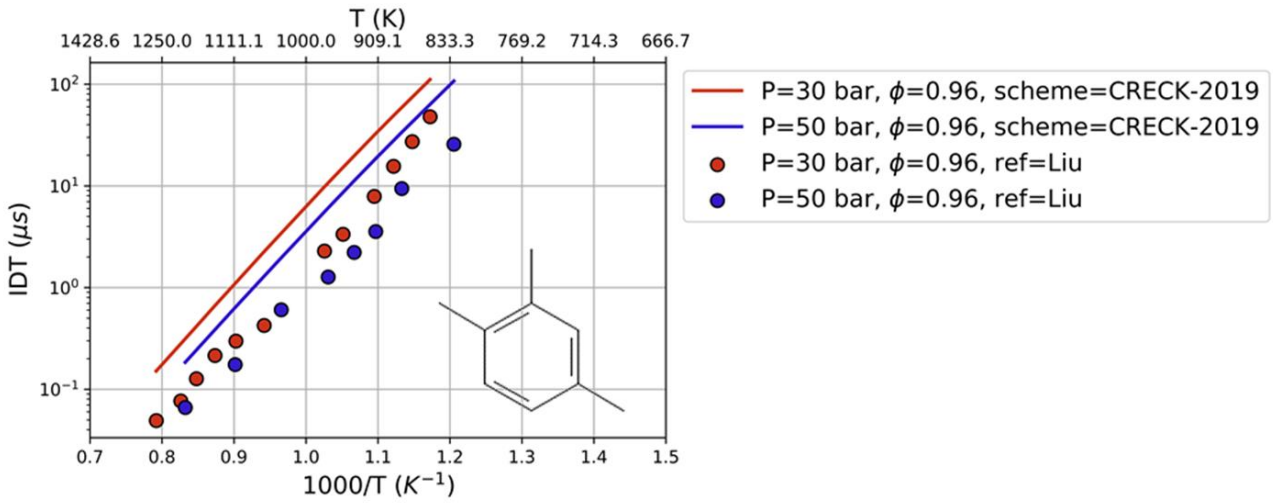

Figure 3-11: Validation data for the ignition delay time of 124-trimethylbenzene in air, Liu [46]

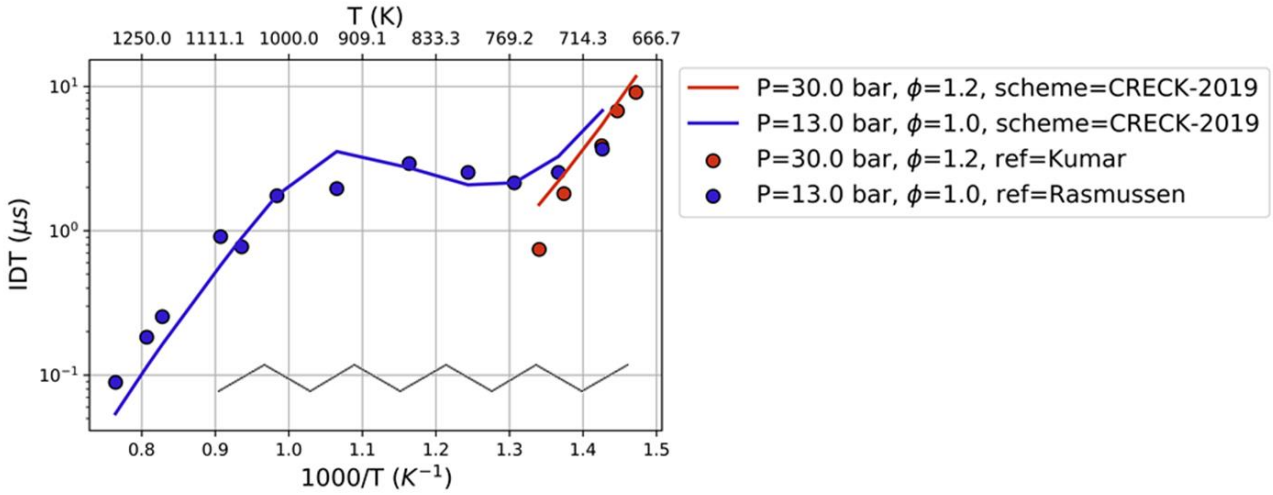

Figure 3-12: Validation data for the ignition delay time of n-decane in air, Kumar [47], Ramussen [48]

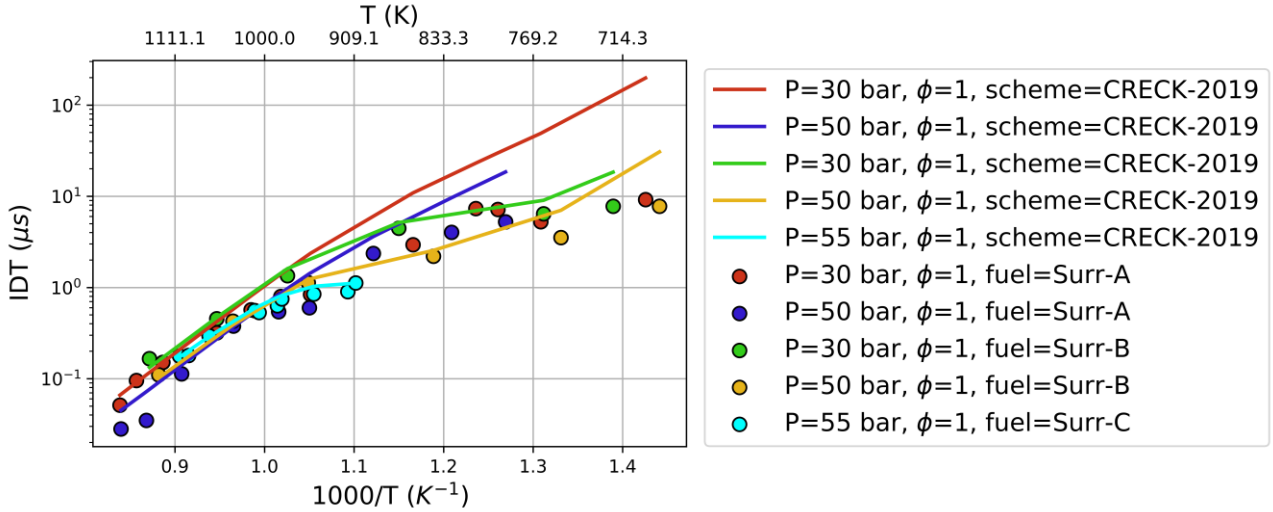

Figure 3-13: Multi-component surrogates [35]

3.3 Dataset generation

The availability of well distributed data and of a sufficiently wide training domain is considered one of the key assets for the successful implementation of machine learning strategies, therefore the definition of the dataset of simulation results has been carried out carefully. In particular, the choice of the thermodynamics conditions has been performed by identifying any possible scenario that could occurred inside the cylinder, comprising hot spots on the walls even at low pressure and very lean or rich local regions, with a sufficiently detailed resolution, leading to the values of Table 3-2.

Table 3-2: Operating conditions analyzed

Variable	# points	Values
Pressure (bar)	5	10,30,60,100,140
Temperature (K)	9	700,750,800,850,900,950,1000,1050,1100,1150,1200
Equivalence ratio (-)	7	0.4,0.8,1.0,1.2,1.5,2.0,3.0

As far as the fuel composition is concerned, the dataset has been defined by following a 3 step approach:

1) For each molecule, a limit has been assigned, depending on its class and RON (from 0 to 60% for iso-paraffins, from 0 to 20% for n-paraffins, from 0 to 30% for ethanol, from 0 to 20% for olefins, from 0 to 40% for aromatics and from 0 to 20% for naphthene), following loosely the limitations imposed on transportation fuels in Europe.
2) A random normal distribution has been defined for each molecule allowing the limitation to be enforced at least for 2 standard deviations, with the mean value set equal to the arithmetic mean of the boundaries.
3) For each sample point, the sum of the fractions is normalized to unitary sum, and the values lower than 0.04 are removed, leading to a re-normalization step.

The distribution of each single molecule inside the surrogates formulated in the dataset is collected in Figure 3-14 together with the pairwise relationship between each molecule, which underlines qualitatively the resolution of the dataset.
In order to reduce the number of simulations required, for each surrogate, only 60% of the thermodynamics conditions, selected randomly, was actually simulated, leading to the definition of about 100’000 points for 520 combinations.

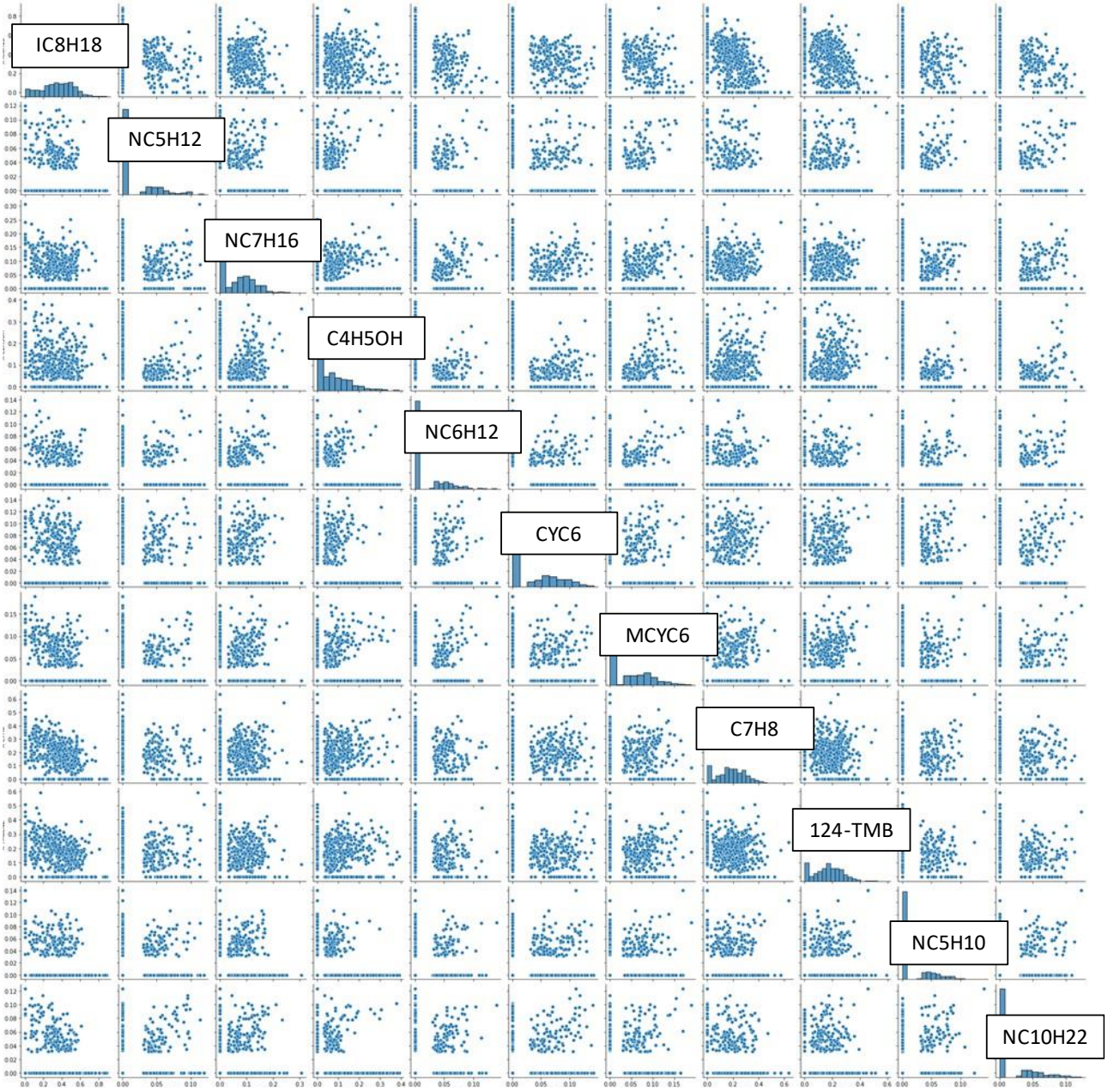

Figure 3-14: Pairwise relationship plot describing the distribution of the single molecules in the available dataset and their pair-wise relative distribution

3.4 ML Experiments

A series of Machine learning algorithms has been applied to the presented dataset, in order to predict the ignition delay time for any combination of thermodynamics properties and fuel surrogate composition. The nature of the problem is a regression task with tabular data, therefore classical ML algorithms have the potential to perform accurately, especially if combined through boosting or bagging. In addition an optimized deep neural network has been tested, as well as a novel algorithm based on a combination of deep neural networks. For all the models, the same workflow introduced in Chapter 1 has been applied, with preprocessing of the dataset, algorithm optimization via cross validation and Bayesian

optimization of the hyperparameters and a final validation against an hold-out validation set. All the steps will be outlined in the following sections.

1) Preprocessing

   One of the advantages of tree based algorithms is the fact that they do not require a-priori scaling of the input features, since they are not sensitive to variance between the data. However, to perform a more fair comparison among all the models, all the features have been scaled in order to have their minimum value set to 0 and their maximum value set to 1 [49]. Considering that the target value spans several orders of magnitude, an adequate scaling is required before training the model. In particular, it is straightforward to perform a logarithmic scaling on the ignition delay time, which leads to the distribution change of Figure 3-15

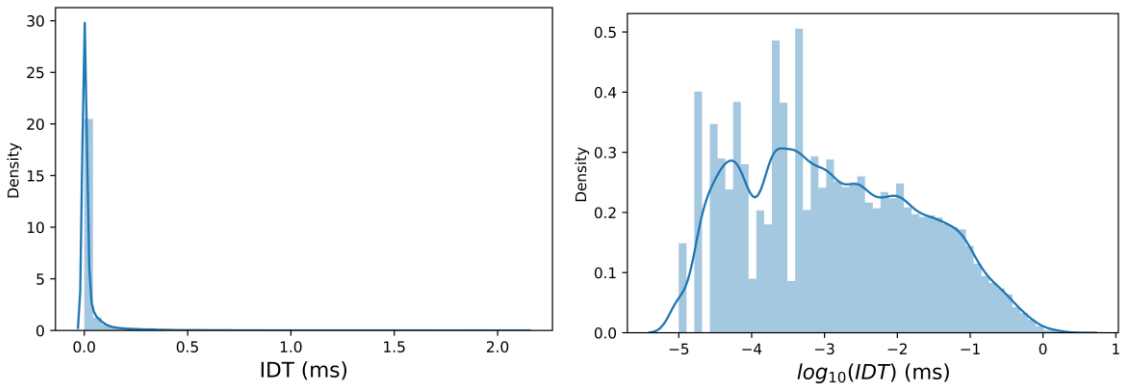

Figure 3-15: Target ignition delay time distribution before and after the logarithmic transformation

2) Algorithm optimization

   Before optimizing the hyperparameters of the algorithm, a portion equal to 25% of the entire dataset was kept as reference test set and removed from the training set. The portion was composed of different surrogates and not randomly distributed, in order to assess the predictivity of the models towards totally unseen data.

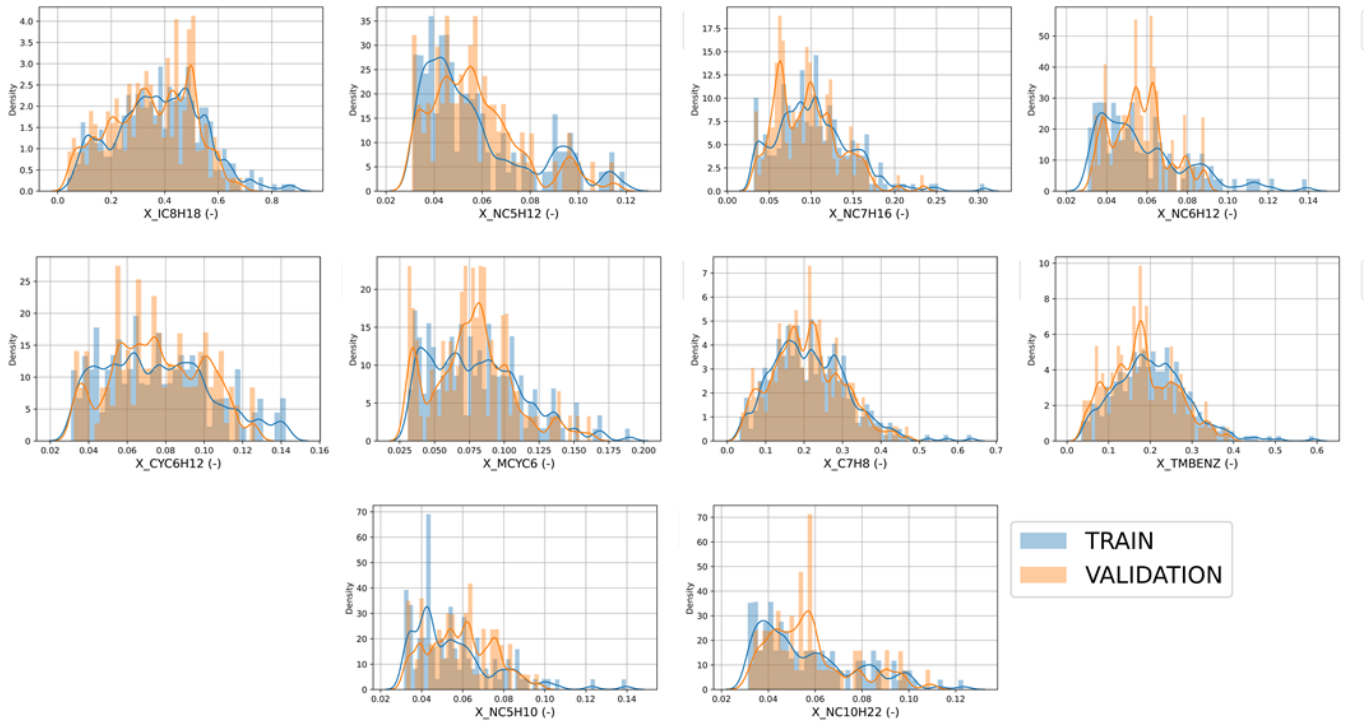

Figure 3-16: Distribution of the molecular fractions inside the training set and the hold-out validation set

3.4.1 Optimization and Validation set results for XGBoost

The first machine learning experiment performed was based on the gradient boosting tree method, considering the tabular nature of the problem, and the expected non-linearity of the features. In addition, the size of the dataset does not require the model to extrapolate, therefore a boosted tree algorithm is expected to work properly. In order to reduce the variance of the prediction, the model has been trained on 5 random fractions of the training set composed of 80% of the total dataset. The prediction of all the 5 models on the validation dataset is then averaged in order to obtain the final value, thus combining the boosting method with the bagging ensemble technique, as described in chapter 1.

The optimized hyperparameters for the gradient boosting regressor, in its implementation in XGBoost are: 150 base estimators (which are regression tress with a maximum depth set to 20), trained successively with a learning rate of the boosting algorithm set to 0.05 and a subsample rate for the training of each new model set to 80% of the overall dataset, and the performance metrics for the transformed target on the validation set are reported inTable 3-3. An additional metric, specifically defined for the ignition delay time prediction has been added, named ERR_G, and calculated as in Eq. 3-3 after the Livengood-Wu integral method [50].

$$ERR_G = \frac{\sum_{i=1,N}\left(\frac{1}{y_i} - \frac{1}{\tilde{y}_i}\right)}{\sum_{i=1,N}\frac{1}{y_i}} \tag{3-3}$$

Table 3-3: Performance metrics of the XGBoost algorithm

MAE	0.034
RMSE	0.045
R2	0.998
ERR_G	4.56

3.4.2 Validation set results for other regression models

Considering the general simplicity in the optimization of other classical machine learning algorithms, their description is left to the introduction chapter, while in the present section, only their performance on the hold-out validation set is reported.

Table 3-4: Performance metrics of the classical machine learning algorithms

Model	MAE	RMSE	R2	ERR_G
Linear regression	0.102	0.158	0.859	12.5
Elastic net regression	0.084	0.088	0.924	9.8
Polynomial regression with 2nd order features	0.093	0.113	0.812	11.8
Polynomial regression with 3rd order features	0.075	0.103	0.835	10.9
Support vector machine regression with RBF kernel	0.072	0.094	0.877	10.6
Gaussian process regressor with RBF kernel	0.065	0.085	0.896	8.9
Decision tree regressor	0.132	0.372	0.723	19.2
Random forest regressor	0.046	0.055	0.991	6.1
Adaboost regressor	0.041	0.049	0.993	5.7

From the classic linear and polynomial regression performance, which improves with the addition of higher order terms, it is clear that non-linearities are required to be modelled in order to perform an adequate prediction. On the other hand, ensemble algorithms, such as random forest or adaboost, which are considered prone to overfitting, clearly outperform the simpler methods, also on the validation dataset, underlying the importance of the accurate dataset generation step in representing all possible combinations.

3.4.3 Optimization and Validation set results for Deep Neural Network

Considering the tabular nature of the regression task presented, the application of deep neural networks has been generally regarded as an over-complication of a problem that can be successfully addressed with tree-based algorithm [51]. However, this statement is correct for general machine learning problems, with dozens of features, and some of them are classes. For the present task, however, the continuous nature of the input, as well as the need for a continuous approximation of the target makes the use of deep neural networks advantageous. By following the standard approach introduced in the previous section, the

architecture has been optimized, up to the structure with 65’283 trainable weights presented in Figure 3-17.

The final structure has been obtained by applying a Bayesian optimization algorithm on a simplified block-structure made of dense+normalization layers with size of integer multiples of the input size of 14. The use of batch normalization layer is usually applied as a method for improving training stability by normalizing (zero mean and unitary standard deviation) the output of the intermediate layers of the neural network, with the additional effect of increasing the accuracy and generalization of the neural network trained with larger learning rate [52]. The optimization of the weights of the neural network has been therefore performed with the Adam optimizer and a learning rate set to 0.01, thanks to the presence of batch normalization layers, for an infinite number of iterations, stopped after 30 steps without reduction in the MAE loss function of a random fraction of 10% of the dataset used for testing, technique called early stopping. The final performance, reported in Table 3-5.

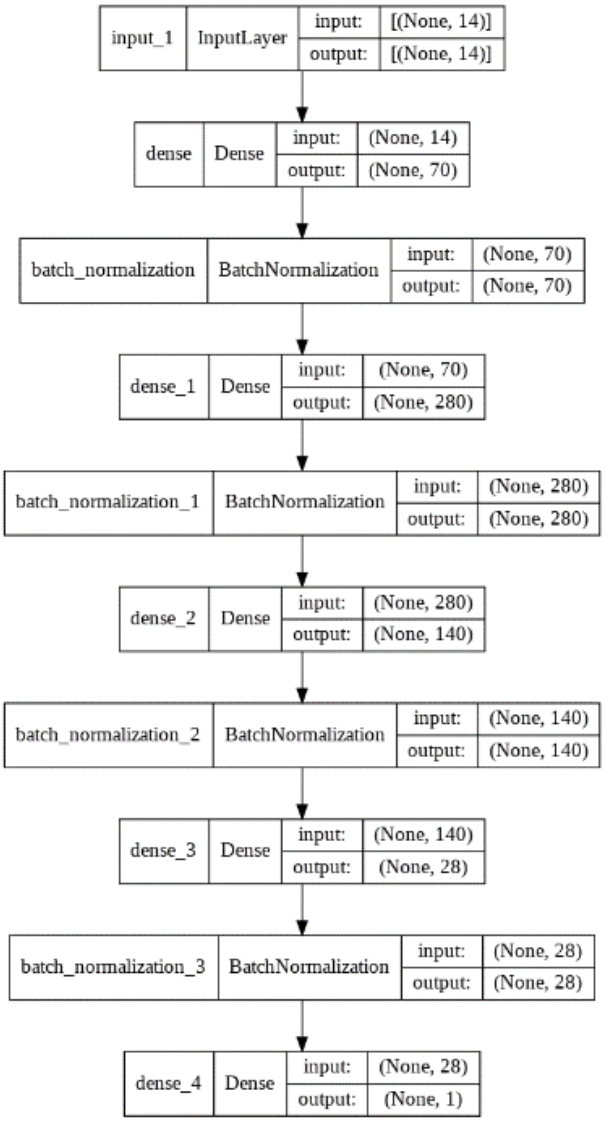

Figure 3-17: Architecture of the optimized neural network

Table 3-5: Performance metrics of the DNN

MAE	0.046
RMSE	0.057
R2	0.989
ERR_G	6.12

3.4.4 Optimization and Validation set results for Meta-learner

A significant drawback of the use of the neural network is its suffering from the distribution of the targets, which is skewed towards the value of -3, values around which the model is optimized. In fact, the simplicity of the MAE loss function makes the training of the regression model leaning towards the mean value (e.g. if the model is unable to find any correlation between the features and the output, the minimization of the MAE loss function produces a model that would always provide as outcome the mean value of the training dataset). In order to compensate this modelling issue, a meta-learning methodology has been developed, where multiple NNs are stacked in two consecutive levels. The workflow is based on the common approach of developing a prediction model for the ignition delay time by combining multiple correlations fitted with reference values from different temperatures intervals [53]. At the first level, during the training phase, the temperature range of the dataset (700-1200 K) is divided as follows:

a) the low-temperature Arrhenius region (700-950 K);
b) the expected NTC region (900-1100 K);
c) the high-temperature Arrhenius region (1050-1200 K).

These regions were created with a slight temperature overlap between each other, in a manner to account for a similar number of points per region. The points in each region are provided to train three different and independent NNs, whose predictions will result in three different IDT values.

In the second level, a larger NN is trained on the full database, but the input layer receives three additional features, namely the IDT values predicted by the three independent NNs at the first level, independently from their temperature. The second level FFNN is defined to receive as input features the physical and mixture properties and the 3 outputs from level 1. For sake of illustration, Fig. 3-17 shows a schematic view of the meta-learning workflow and architecture.

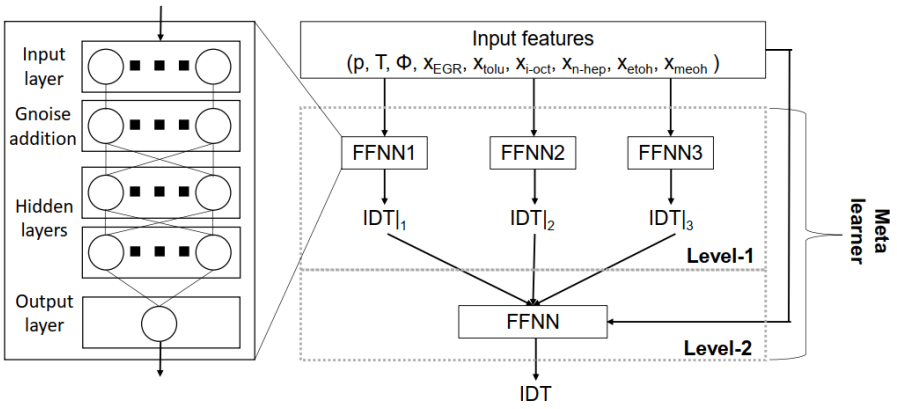

Figure 3-17: General structure of the Meta-learner algorithm

The performance of each single network significantly decreases significantly outside of its training region, however the second level network is trained at recognizing the validity of each input and therefore the overall performance of the 2-step model is almost as good as each specialist network, and reduces the percentage relative error with respect to a single network by 30% on average, and up to 50% at lower temperatures, without any correlation between the predictive error and the temperature of the mixture.

Table 3-6: Performance metrics of the meta-learner algorithm

MAE	0.016
RMSE	0.021
R2	0.999
ERR_G	2.4

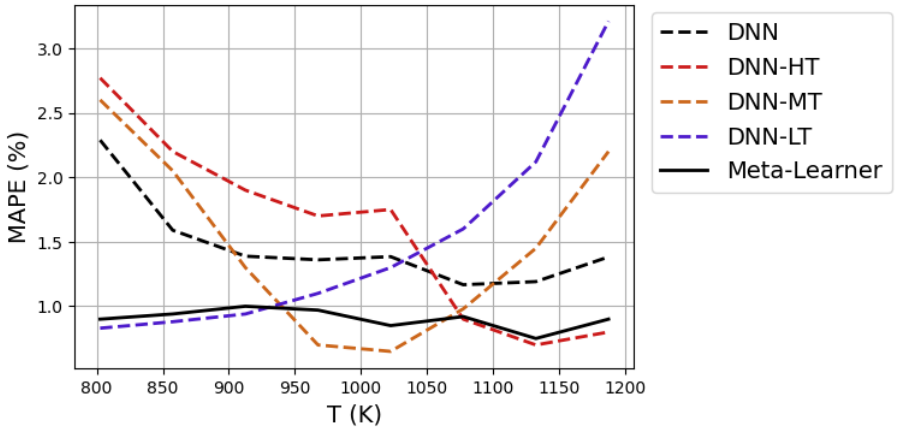

Figure 3-18: Performance of the meta-learner algorithm compared to a single optimized DNN and the specialized subnetworks

4 Laminar flame propagation with chemical kinetics

The flame propagation inside a turbulent flow field is a complex phenomenon highly influenced by both chemical reactions and flame turbulence interaction. The energy released by the fuel oxidation and the density variation between fresh and burnt mixture are, in fact, a source of induced turbulence that, conversely, increases the molecular diffusivity inside the flame front allowing for a faster burning rate [54]. Considering the chemical time (~O(1e-8 s)) and length (~O(1e-7 m)) scales in play, it is unfeasible to directly resolve the entire flame structure during a combustion simulation with sufficient accuracy for industrial applications. Most frequently, the results of detailed chemical kinetics simulations in simplified domains, or experimental data under controlled environments are employed as the basic blocks for simplified models. The requirement to understand in-depth the flame structure, as well as the possibility to predict the flame front propagation speed without limitations in temperature and pressure values has made the use of detailed chemical kinetics simulation a common practice in recent years, accompanied by a higher level of detail reached by chemical kinetics schemes, as underlined in Figure 4-1 On the other hand, the increase in size of the chemical reactions and intermediate species considered, has made the solution of simplified 1D flame propagation models a resource intensive activity [55].

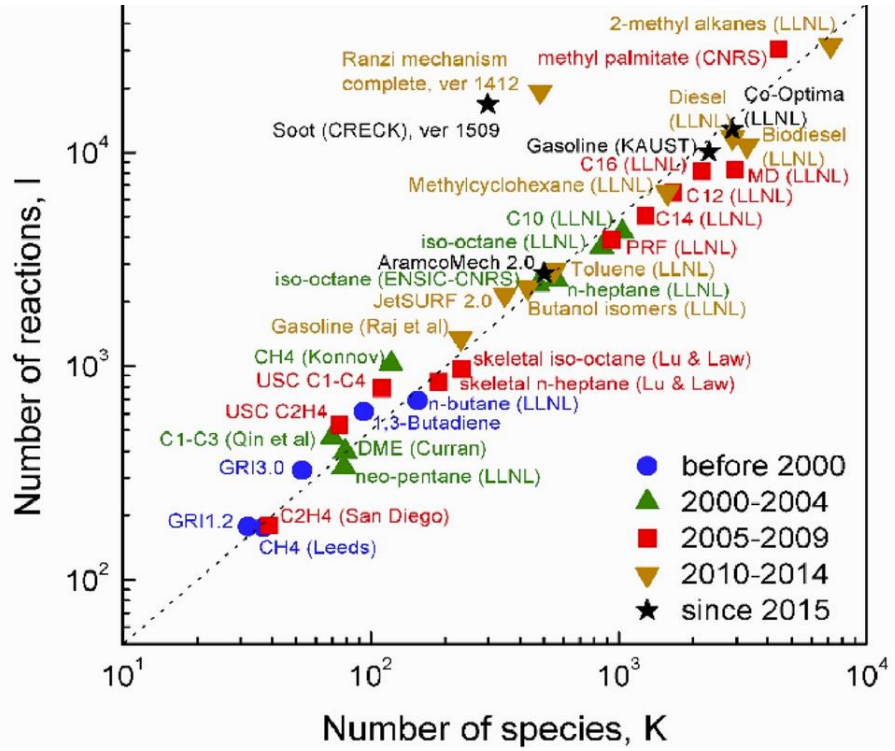

Figure 4-1: Evolution of the dimensions of chemical kinetics schemes for combustion simulation [56]

To reduce the time requirements for the generation of laminar flame speed databases given new fuel surrogate formulations, in the present chapter a set of techniques is defined, implemented and tested starting from the standard flame definition with the following targets:

1) Introduce the calculation strategy for laminar flame speed and laminar flame thickness values from simplified 1D chemical kinetics simulations.
2) Describe the traditional chemical kinetics reduction strategies that can be employed for the automatic derivation of simplified models.
3) Define and validate a data-driven methodology for the generation of laminar flame speed and thickness look-up tables in the minimum possible time, while controlling the accepted error from the reference values [57].

In the second part of the chapter, attention will be given to the simulation of a flame front propagating into a mixture containing additional water vapour, essential for the CFD simulation of internal combustion engines employing a water injection knock mitigation strategy. The application of hybrid machine learning methods to the generation of a laminar flame speed database considering the additional water fraction variable will be shown to be an extremely valuable tool for the reduction of the overall computing time [58].

4.1 Laminar flame speed simulations

Considering the central role that laminar flame speed plays in the combustion modelling, it is of paramount importance to make a precise computation of its value for all conditions that might be encountered during the engine combustion simulation, in order to provide in runtime, by means of correlations or look-up tables, an efficient way to retrieve an accurate estimate of its value.
Historically, several authors have proposed correlations of laminar flame speed with temperature, pressure, and equivalence ratio for different fuels, with similar forms.

$$s_L = {s_L}^0 \cdot \left(\frac{T}{T_0}\right)^{\alpha} \cdot \left(\frac{P}{P_0}\right)^{\beta} \cdot (1 - k \cdot {X_{EGR}}^{\gamma}) \tag{4-1}$$

$${s_L}^0 = B_m - B_f\,(\phi - \phi_m)^2 \tag{4-2}$$

$$\alpha = \alpha_0 - \alpha_1(\phi - 1) \tag{4-3}$$

$$\beta = \beta_0 + \beta_1(\phi - 1) \tag{4-4}$$

Where T_0 is the reference unburnt gas temperature and P_0 the reference pressure, usually taken at ambient condition, subscript 1 indicates the unburnt (fresh) mixture, and 2 indicates the burnt phase, $k \cdot {X_{EGR}}^{\gamma}$ is a correction factor introduced to account for the presence of EGR as inert (k has values found in literature between 1.7 and 2.3). The coefficients in the previous equations need to be modified as a function of the chosen correlation (the most used are those of Metghalchi and Keck [59], Heywood [60] or Gülder [61]), and are presented in literature for different surrogates and PRFs, obtained through experimental campaigns, while ϕ_m indicates the equivalence ratio at which maximum laminar flame speed is reached at reference conditions. Despite their diffusion, these correlations, based on experimental data, lack the ability to accurately describe the effect of different fuel compositions, and their validity is confirmed only under thermodynamics conditions close to

ambient, that are far from the higher pressure and temperature ranges reached during engine operation. Therefore, in recent years, it has become common practice to perform a-priori simplified laminar flame speed calculations with detailed chemical kinetics solvers, in order to obtain an appropriate description of the sensitivity of the chemical reactions occurring inside the flame front to the conditions actually reached inside the combustion chamber. The results of these simulations, in terms of both laminar flame speed and thickness, then, can be employed in the form of look-up tables to be interpolated in run-time, or through new and more precise correlations.

Starting from the definition of laminar premixed flames, these can be regarded in their simplest form as a one dimensional laminar and adiabatic propagation of the flame front into a homogeneous air-fuel mixture. The equations that describe the mass, species and energy conservation for this type of problem can be defined as follows [62]:

$$\frac{\partial \rho}{\partial t} + \frac{\partial \rho u}{\partial x} = 0 \tag{4-5}$$

$$\frac{\partial \rho Y_k}{\partial t} + \frac{\partial}{\partial x}(\rho(u + V_k)Y_k) = \dot{\omega}_k \quad , \quad for\ k = 1, N-1 \tag{4-6}$$

$$\rho c_p \left(\frac{\partial T}{\partial t} + u\frac{\partial T}{\partial x}\right) = -\sum_{k=1,N} h_k \dot{\omega}_k = +\frac{\partial}{\partial x}\left(\lambda \frac{\partial T}{\partial x}\right) - \rho \frac{\partial T}{\partial x}\left(\sum_{k=1,N} c_{p,k} Y_k V_k\right) \tag{4-7}$$

These equations describe a transient wave propagating from the burnt to the fresh zone of the 1D domain, where N is the number of molecules that appear in the mixture, Y_k their mass fraction, V_k their diffusion velocity and $\dot{\omega}_k$ their reaction rate.

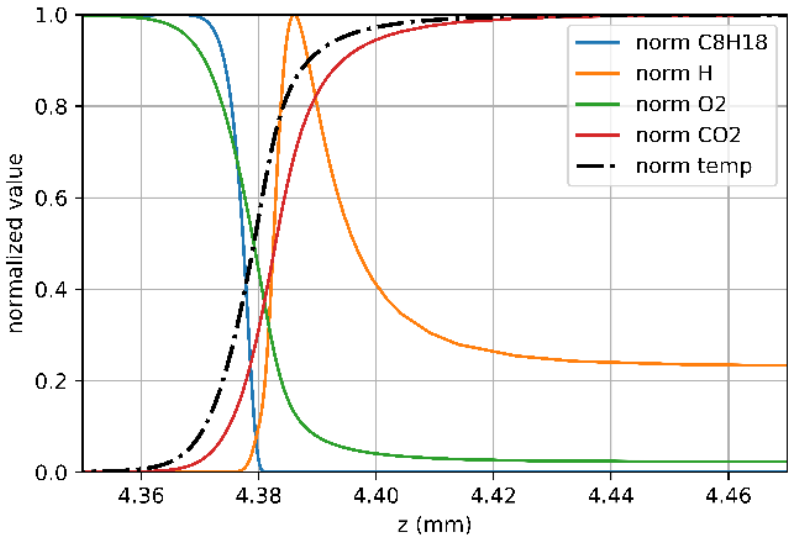

Figure 4-2: Non-dimensional profile of selected variables near the flame front section

The second essential property of the laminar flames is the front thickness, which can be computed as the ratio between thermal diffusivity of the fresh gas and the laminar flame speed. A more useful definition of laminar flame thickness is given by [63] using the temperature profile, reported in Eq. 4-8, where subscript 1 indicates the unburnt (fresh) mixture, and 2 indicates the burnt phase:

$$\delta_L{}^0 = \frac{T_2 - T_1}{max\left(\left|\frac{dT}{dx}\right|\right)} \tag{4-8}$$

The definition and the evaluation of the laminar front thickness is notable since it affects the wrinkling efficiency of turbulence eddies. The chosen definition of laminar flame thickness in Eq. 4-8 is described as thermal thickness, and it is reported as the most representative for laminar flames, but the application in the CFD simulations of ICEs is more complex, because of the several hypotheses made in the combustion and turbulence models.
A correlation based on scaling laws was introduced to overcome the lack of experimental data regarding the LFT, as proposed in Eq. 4-9 calculated with the properties of the fresh mixture, where λ is the thermal conductivity of the gas, ρ is its density, C_P the specific heat at constant pressure and s_L is the laminar flame speed in that condition.

$$\delta_L{}^0 = \frac{\lambda}{\rho C_P s_L} \tag{4-9}$$

Blint [64] corrected the previous correlation by introducing a correction factor based on the burnt gas temperature, leading to Eq. 4-10, where subscript 1 indicates that the property refers to the fresh mixture, and subscript 2 is the condition of the burnt zone.

$$\delta_L{}^{Blint} = \delta_L{}^0 \frac{(\lambda/C_P)_2}{(\lambda/C_P)_1} \tag{4-10}$$

Since the value of the LFT in turbulent combustion modelling is extremely important, the accuracy of the Blint's correlation has been evaluated under engine relevant conditions with respect to values obtained by means of detailed chemical simulations. It was chosen to focus only on the data available for a generic gasoline surrogate fuel and the results reported in Table 4-1 show an overall good agreement between the two values.

Table 4-1: Performance metrics of the Blint correlation with respect to detailed chemical kinetics

Slope	R^2	MAE (m)	RMSE (m)
0.626	9.91E-1	2.11E-5	5.07E-5

In particular, the slope can be interpreted as a proportionality coefficient between the calculated thermal thickness and that obtained from the temperature profile, which can be used as a constant of proportionality for simpler CFD models. Besides the requirement for

scaling factor. After the application of such correction, the MAE and RMSE are reduced by over 40% and the final relative error distribution is presented in Figure 4-3.

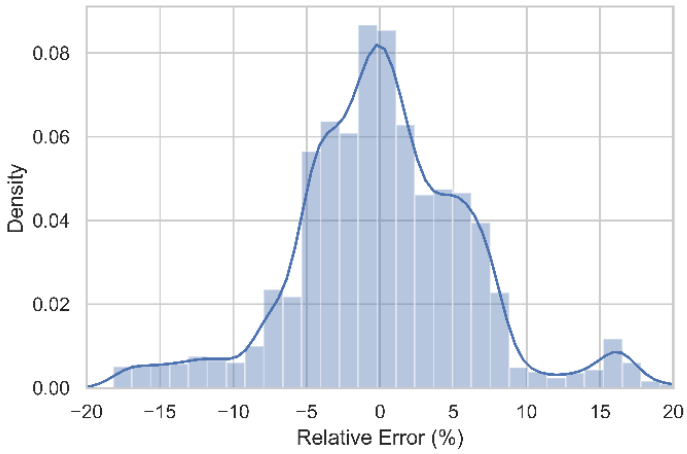

Figure 4-3: Relative error distribution of the laminar flame thickness prediction methods

The solution procedure followed by the Cantera implementation is based on the finite difference solution of the equations on a adaptively refined grid, using a hybrid Newton / time stepping algorithm, where the flame is considered to be steady in the inner region of the domain, with the inlet velocity boundary that equals the consumption speed of the flame. After reaching the steady solution of the problem, the temperature profile along the centerline can be used to identify the position of the flame front and its width can be calculated [65]. Arguably the most significant drawback of this approach is the requirement to impose a fixed width of the domain from the first iteration. In particular, it has been directly observed that the solution of the problem might not reach the steady state for a too short domain, since the computational method is not capable of identifying a suitable flame front stabilization point, thus leading to the requirement to increase the domain width and run a new simulation. At the same time, the choice of a domain width too large allows for the development of cool flame phenomena before the flame front, up to the point of fresh mixture autoignition, which is identified by the solver as an extremely high laminar flame speed (since it is not expected to handle deflagration to detonation transitions). For sake of clarity, in Figure 4-4 the laminar flame speed of n-heptane, which is a molecule usually referred to for its NTC behaviour, is reported as a function of the chosen domain width for a sufficiently high initial pressure, and rich mixture, in order to increase the cool flame effect (in this case, the chosen operating point was P=120 bar, T=850K, equivalence ratio=1.2 and 0% EGR). Besides the possibility of an a-posteriori analysis of the flame temperature profile, which would allow to identify the insurgence of cool flames before the planar flame front [66], in order to reduce the computational time required, a precise domain width definition strategy has been defined:

1) The first step is the identification of a reasonable laminar flame speed value for the operating point, obtained with the classical Heywood correlation for Isooctane. Even if the final laminar flame speed is not expected to correspond precisely to the value predicted by the correlation, this initial guess is necessary to perform further computations.

2) The first stage ignition delay time of the mixture in the desired conditions is easily computed by integrating a 0D constant volume domain under the same conditions as for autoignition simulations. To be more cautious, the identification of the insurgence of cool flame phenomena corresponds to the first instant with a positive temperature gradient.
3) An equivalent distance is computed by multiplying the guessed laminar flame speed by the first stage ignition delay time, and its value is considered a upper-bound limit for the laminar flame front stabilization.
4) The lower-bound limit is identified in order to reduce the computing time required and avoid the need to increase the domain width in order to allow the numerical method to converge to steady state. Its value is identified by scaling an initial value of 0.1 m by the laminar flame thickness of the mixture in the simulated conditions (obtained by applying the Blint equation to the guessed laminar flame speed).

The domain width is then chosen as the average of the lower and upper -bound limits, as long as the correlation does not provide values that invert their order, in which case, the requirement to avoid cool-flame insurgence is enforced.

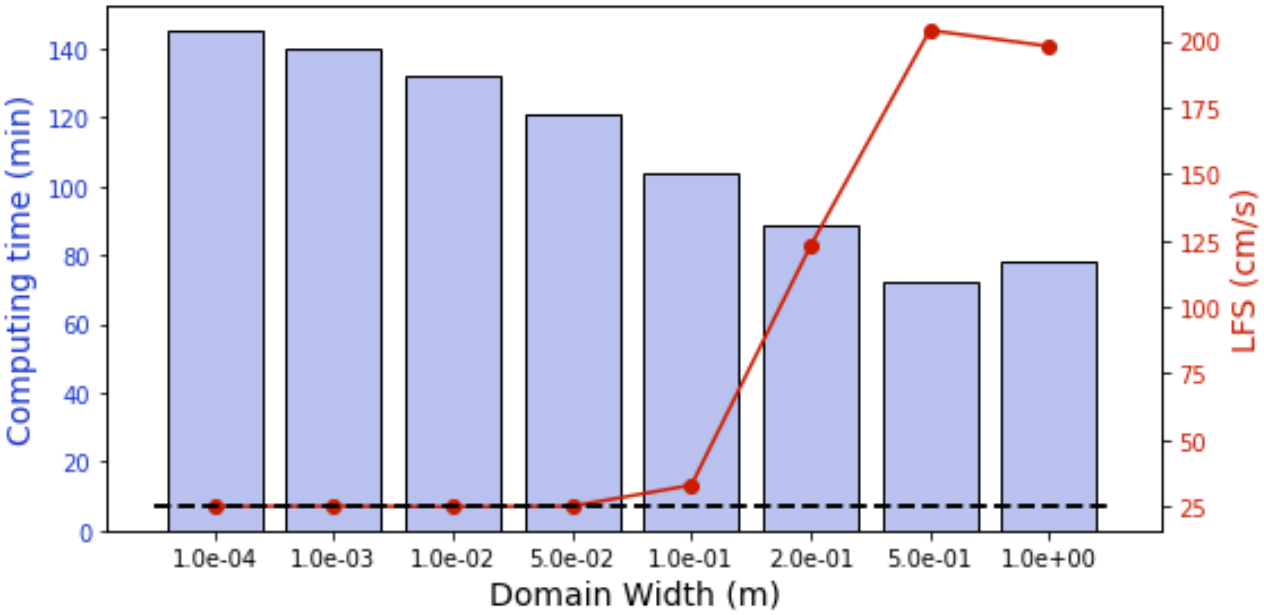

Figure 4-4: Influence of domain width on the computing time and cool flame phenomenon insurgence

The computing time of the problems, independently from the solution procedure adopted, is expected to scale linearly with the number of reactions, and quadratically with the number of species that the chemical kinetics mechanism accounts for.

4.2 Brief introduction to DRGEP-SA

Considering the high computing time required by the computation of a single laminar flame speed value, the generation of new look-up tables for evaluating sensitivity analysis to fuel

formulations, as well as optimizing fuel composition, might become unfeasible. A possible approach to reduce this requirement is based on the use of an optimized version of the chemical kinetics scheme [67], obtained by removing the less relevant reactions for the problem and conditions to be simulated. To this aim, several approaches have been proposed in literature, but the use of automatic reduction algorithms is generally considered an efficient approach. In particular, the DRGEP (Directed Relation Graph with Error Propagation) combined with a subsequent Sensitivity Analysis, proposed by Niemeyer [68] [69] [70] [71] are usually referred as efficient ways to reduce the size of a chemical kinetics scheme while controlling the solution accuracy in 0D reactor simulations at reference conditions. The method is based on a three-step algorithm:

1) Direct interaction coefficients between each species pair A, B is computed as:

$$r_{AB} = \left| \frac{\sum_{i=1,N_{reactions}} \nu_{A,i}\omega_i \delta_{B,i}}{\max(P_A, C_A)} \right|$$

Where $P_A = \sum_{i=1,N_{reactions}} \max(0, \nu_{A,i}\omega_i)$ and $C_A = \sum_{i=1,N_{reactions}} \max(0, -\nu_{A,i}\omega_i)$ and $\delta_{B,i}$ is a Kroenhecker delta equal to 1 if the ith reaction involves B.

2) Differently from the DRG method, a modified Dijkstra's algorithm [72] has been applied to calculate the overall interaction coefficient efficiently between each target T and all other species B as the maximum of the interaction coefficients along all paths:

$$R_{TB} = \max\left(\prod_{j=1}^{n-1} r_{S_j S_{j+1}} \right)$$

3) After the computation of each overall interaction coefficients, each species B is removed when $R_{TB} < \epsilon$, set to 1e-4 for this application.

After the application of the reduction procedure, the idea of Niemeyer et al. was to apply a sensitivity analysis reduction step on a subset of the reactions defined by their interaction coefficients. The sensitivity analysis is, in fact a very computationally intensive technique, which is based on the repeated simulation at reference conditions of modified versions of the base mechanism. The change in the target results (for example ignition delay times) with respect to the change in the rate of each reaction is then compared with a threshold value and the reaction is eventually removed.

Since real engine CFD simulations are concerned, and due to the availability of experimental data [73], the selected fuel surrogate is the *TAE7000* (RON 98.1, composition: 13.7%v n-heptane C_7H_{16}, 42.9%v iso-octane C_8H_{18}, 43.4%v toluene C_7H_8), which was found to provide accurate laminar flame speed predictions to that of a TOTAL commercial gasoline.

From the chemical kinetics schemes presented in chapter 1, considering that toluene is not included in the NUI2021 mechanism, only the LLNL2021 and CRECK2019 are considered as reference mechanisms. Each of them has also been reduced with the DRGEP-SA approach, admitting a reduction error of up to 5% and up to 10% on ignition delay time simulations performed on a set of 10 operating points with the TAE7000 surrogate. The chosen conditions are reported in Table 4-2

Table 4-2: Operating points considered for the autoignition simulations employed by the mechanism reduction algorithm

OP #	P (bar)	T (K)	Equivalence ratio (-)
1	10	900	1.0
2	50	800	1.2
3	100	1000	1.0
4	150	1400	0.8
5	10	950	1.0
6	50	850	1.3
7	100	1100	0.7
8	150	1200	1.5
9	100	1000	1.0
10	100	1000	0.5

For each chemical kinetics scheme (referred to as 'complete'), two reductions are therefore available, and their size is reported in table, together with the average time required to perform a laminar flame speed computation with the same numerical setup on 5 operating points on a machine having the following hardware and software configuration: Intel Xeon Platinum 3.0 GHz, 36 cores,144 GB ram, Cantera version 2.3.0 on Python 3.7.2 [74, p. 3].

Table 4-3: Number of species and reactions and computing time required for the solution of a laminar flame speed simulation for complete and reduced chemical kinetics mechanisms

	LLNL2021	L_RED1	L_RED2	CRECK2019	C_RED1	C_RED2
#SPECIES	1956	1004	891	582	279	226
#REACTIONS	10371	7853	6543	21174	11980	8703
Time/sim (CPUhr)	N/A	14.1	10.3	6.8	2.5	1.9

The simulations for the laminar flame speed with the LLNL2021 scheme could not be performed due to hardware limitation, given the large size of the mechanism and the memory requirements of the employed method.
The differences between the complete and reduced schemes in term of species and reactions considered generate a great unbalancing in terms of the CPU time required for

simulating a complete database of reference points. Due to its time cost requirements, the remaining analysis has been performed only on the CRECK2019 model and its reductions. Besides a reduction in the computing time, a reduced accuracy of the result can be identified when comparing the laminar flame speed values of simple gasoline fuel surrogates with the detailed and reduced versions of the chemical kinetics mechanism, as presented in Figure 4-5 for the *TAE7000* gasoline fuel surrogate. Further analysis on the error induced by the reduced schemes on a wider range of conditions will be reported in the next sections.

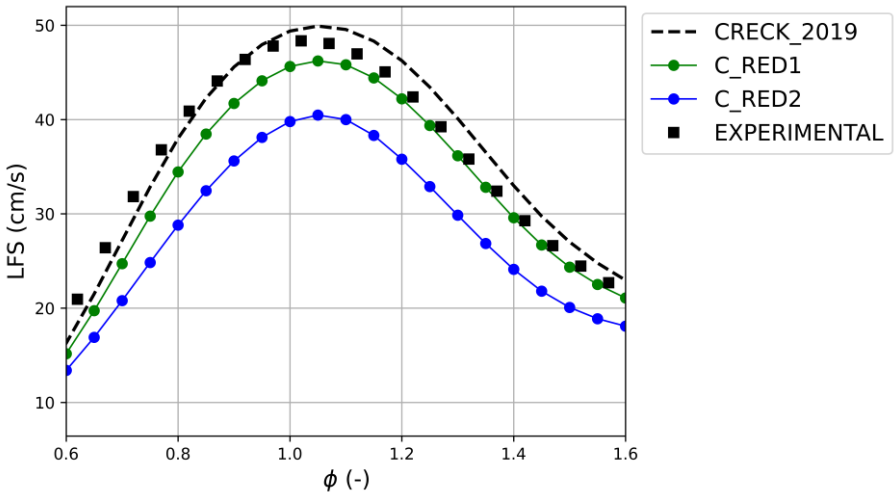

Figure 4-5: Laminar flame speed of TAE7000 with different methods for different values of Equivalence ratio, P=50 bar, T=750 K, EGR=5%

4.3 Look-up table generation

To compute a lookup table of laminar flame speed and thickness, to be substituted to the classical correlation approach, the detailed 1D chemical kinetics simulation must be performed for a range of potential thermodynamics and mixture conditions, capable of representing all possible engine operation scenarios. Considering the idea of populating a full grid of combinations of the independent variable that describe the problem after the selection of the fuel is performed, a total number of 4 independent variable is required, and their operating limits are reported in Table 4-4.

Table 4-4: Lower and upper bounds for the intervals of the properties defining an operating condition

Variable	Min Vlaue	Max Value
Pressure (bar)	5.0	160.0
Unburnt Temperature (K)	450.0	950.0
Equivalence Ratio (-)	0.3	3.0
EGR (%)	0.0	30.0

In order to define the number of points for each variable, considering their influence on the overall grid of computations required, a sensitivity study has been performed assuming that the on-line interpolation of the look-up table during CFD simulation will be based on a multi-

linear approach. An exemplary effect of the effect of the number of points for a simplified domain is reported in Figure 4-6, where the impact of choosing an increasing number of points for the discretization of the dependency of the laminar flame speed on the equivalence ratio is illustrated.

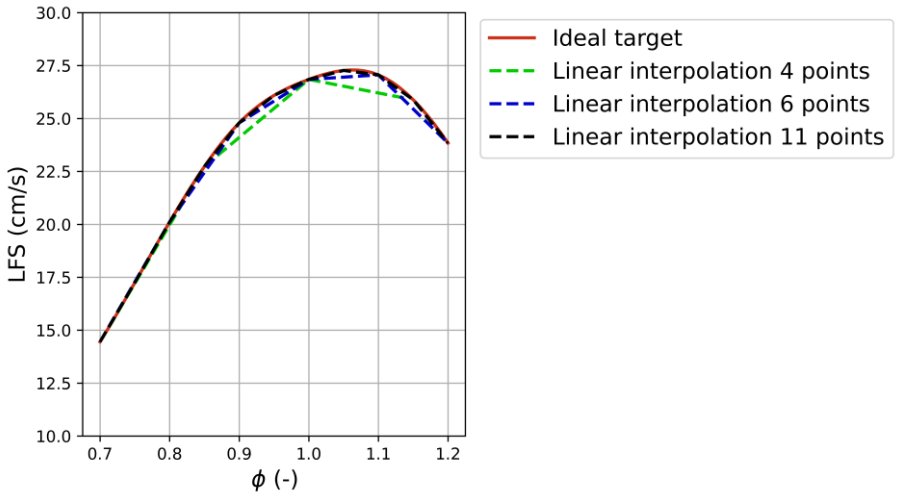

Figure 4-6: Laminar flame speed of TAE7000 for different values of Equivalence ratio, P=1bar, T=358 K, EGR=0% with varying number of discretization points

A deeper analysis has been performed on a set of available simulation data for a sufficiently refined grid (whose step size is at least double than the smallest step size considered for the analysis). The results reported in Figure 4-7, clearly show that after a threshold limit for each variable a sufficiently accurate result can be obtained, therefore the step sizes have been chosen accordingly (Equivalence ratio step size=0.1, Temperature step size = 50 K, Pressure step size=20 bar, EGR step size=5%).

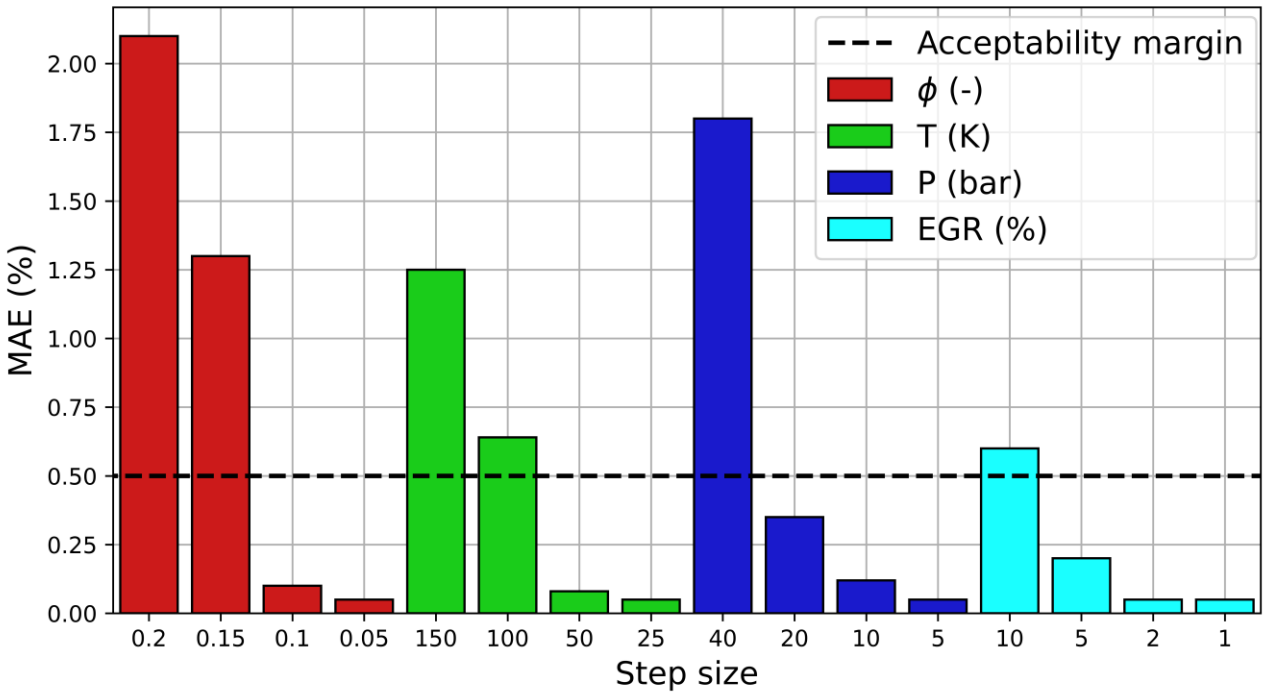

Figure 4-7: MAE committed during multi-linear interpolation function of variable step size

The total number of simulations to be performed is given by the full combinations of the reference data, i.e. 19’404 points.

4.4 ML Experiments

The possible approaches to integrate a machine learning model with detailed chemistry simulations to generate new tabulation points, previously unseen, can be classified with respect to the level of autonomy that is given to the model into two different approaches:

- Employ a traditional regression model to infer the values of the entire map after training on a limited number of observations generated by simulating a section of the database.
- Define a hybrid regression strategy based on the use of reduced chemical kinetics mechanisms and machine learning algorithms in order to leverage the increased computational speed of the reduced mechanism compensating its reduced accuracy with a machine learning model, following Eq. 4-11 where the term η_{ML} is a proportionality coefficient inferred by the trained model, $(s_L)_{1D\_red}$ is the result of the reduced mechanism and $\widetilde{s_L}$ the final output.

$$\widetilde{s_L} = (s_L)_{1D\_red} \cdot (1 + \eta_{ML}) \tag{4-11}$$

The use of the correction term in the second method is expected to compensate most part of the error committed by the reduced scheme, but in a fraction of the time that a simulation with the complete scheme would require, leading to an optimal trade-off between computation time and accuracy. A further advantage of the second approach is the intrinsic robustness of the methodology, considering that the underlying solution is still based on chemical kinetics computations that are expected not to display unphysical results, even when using reduced models.

In order to compute an estimate of the accuracy correction term (η_{ML}) for a given combination of input parameters, the ML model must be trained on the available dataset of both reduced and full chemical simulations at the reference points.

Before addressing the proposed solution, the differences in the results obtained with the complete and the reduced schemes in terms of laminar flame speed and thickness must be analyzed.

By comparing the values obtained on a subset of reference conditions of the full grid, computed for both CRECK2019 and C_RED2 mechanisms, the relative difference between the results of the simulations displays a non-linear behaviour with the physical and chemical properties of the simulated points. From a preliminary analysis on a dataset of about 1000 points, the relative error committed by using the reduced mechanism (with respect to the solution provided by the reference complete mechanism using *TAE7000* as representative fuel) shows different behaviours depending on the equivalence ratio, with an almost exponential profile for richer mixtures. On the opposite, such an error is parabolic with respect to mixture composition, as reported in Figure 4-9 and increases exponentially for lower pressure values, as reported in Figure 4-8 for different initial temperatures. Whereas

the main source of error between the complete and the reduced mechanisms is the reduced number of intermediate species leading to the more exothermic reaction pathways, which might lead to expect an overall laminar flame speed value reduction, the equilibrium between $\dot{QOOH}$ decomposition into cyclic ether species, conjugate olefins and β-decomposition products and the formation of the ketohydroperoxide species is not expected to respond linearly with temperature pressure and mixture composition, especially for more complex thermodynamics conditions.

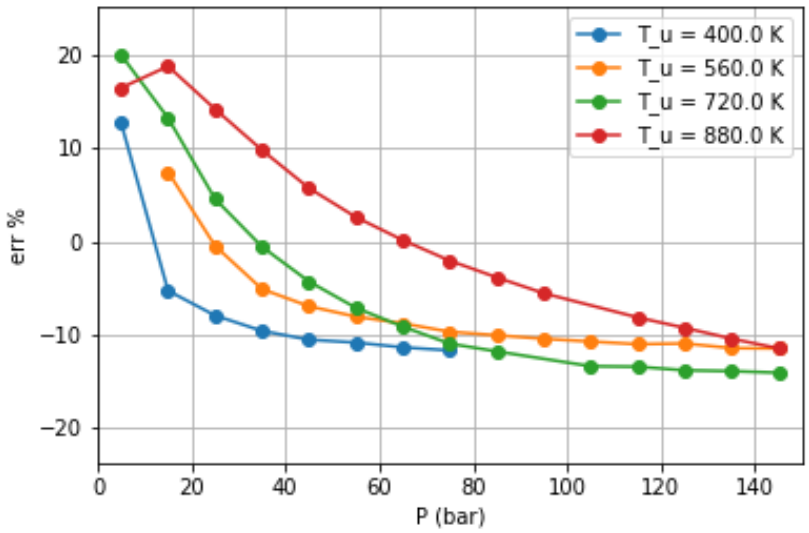

Figure 4-8: Relative difference between laminar flame speed obtained with complete and reduced chemical kinetics scheme for different values of P and T, at ϕ=0.6, EGR=0 %.

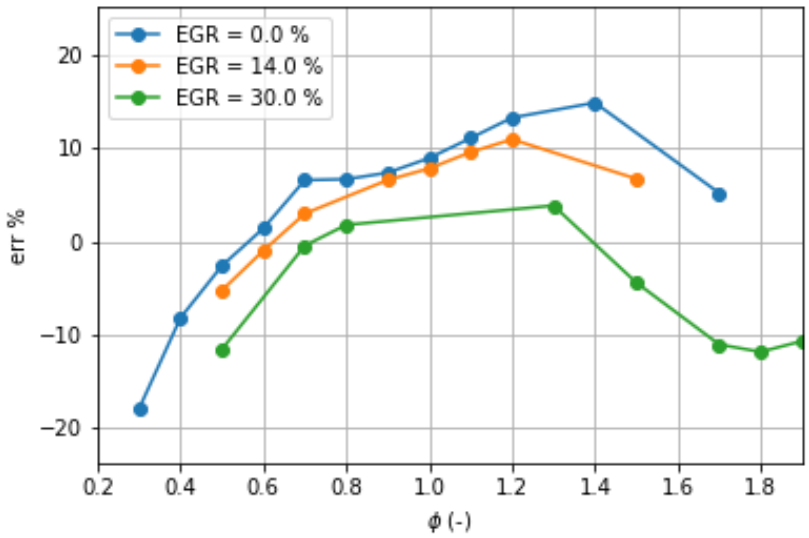

Figure 4-9: Relative difference between laminar flame speed obtained with complete and reduced chemical kinetics scheme for different values of EGR and ϕ, at P=105 bar, T=840 K.

4.4.1 Optimization and Validation set results for the first methodology

Considering the recognized applicability of the Blint correlation for laminar flame thickness, particular focus has been given on the prediction of the laminar flame speed values with machine learning techniques. In order to perform a fair comparison between all the methods and algorithms, a validation dataset composed of 1000 normally distributed random points (approximately 5% of the overall dataset) has been removed from the dataset available for training and analysis of the regression algorithms. The first test is therefore based on the pure tabular regression of the laminar flame speed values, for which classic ensemble models are known to even outperform deep neural networks [51].
Considering the requirement to avoid unphysical results when predicting values that must be used during a simulation, a new metric J has been introduced in order to perform a better comparison between the models, computed on the validation dataset.

$$J = 0.9 \cdot E_{L1} + 0.1 \cdot E_{L\infty} \tag{4-12}$$

A further aspect to be considered is related to the dependency of the performance with respect to the training set size, therefore all the models have been compared on 4 normally distributed random subsets of decreasing size (80%, 50%, 20%, 10% of the entire training dataset) and a further set of approximately 5% of the full database, hand-picked and composed by the most engine-relevant conditions.
The algorithm optimization workflow is the same applied in Chapter 3 and introduced in Chapter 2, with the Bayesian optimization of the main hyperparameters of the model over the k-fold cross validation performance. The XGBoost algorithm [26] has been chosen as initial testing model for the identification of the minimum acceptable dataset size. The results, reported in Figure 4-10 with respect to the error committed by the reduced mechanism C_RED2 clearly highlight the reduced accuracy of the predictions in unknown regions, which lead to maximum relative error of almost 30% even when 80% of the training dataset is employed. On the other hand, the average performance of the regression model is even closer to the target values with train set fractions of at least 20% of the overall dataset (approximately 3000 points).

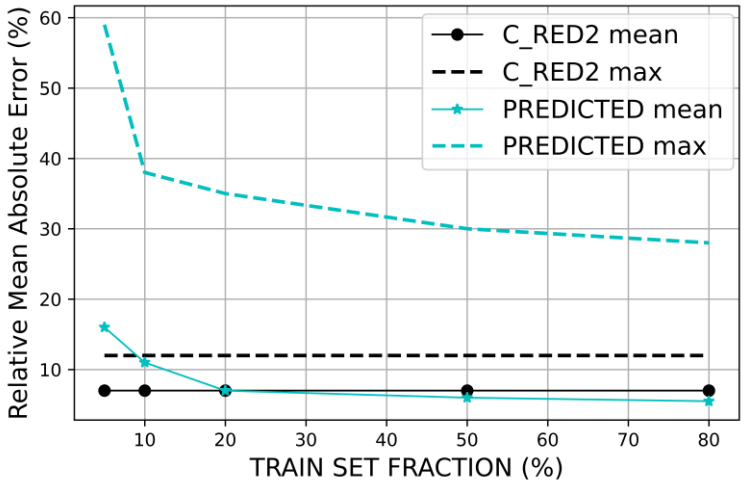

Figure 4-10: Relative difference (mean and max) for the C_RED2 mechanism and regression model

ML Experiments

Considering the dependency of the error a function of the methodology rather than the single algorithm, the train set fraction of 20% is selected to compare the other regression algorithms, with optimized performance reported in Table 4-5.

Table 4-5: Performance metrics on the classical machine learning regression models

Model	MAE (m/s)	RMSE (m/s)	R2	J (m/s)
Elastic net regression	0.177	0.294	0.76	0.27
Polynomial regression with 2nd order features	0.116	0.198	0.87	0.21
Polynomial regression with 3rd order features	0.143	0.268	0.79	0.24
Support vector machine regression with RBF kernel	0.096	0.142	0.89	0.18
Gaussian process regressor with RBF kernel	0.083	0.128	0.91	0.15
Decision tree regressor	0.115	0.194	0.88	0.19
Random forest regressor	0.032	0.067	0.94	0.044
Adaboost regressor	0.014	0.052	0.97	0.021

The same conclusion that was derived in Chapter 3, regarding the non-linearity nature of the target, can be drawn by observing the increased accuracy of the polynomial regression by increasing the order of the features. Also in this case, the capability of ensemble methods to combine several simpler models and optimize different areas of the domain highlights their applicability to this type of problems.

In order to capture all nonlinearities, a different approach, based on deep neural networks is applicable, thanks to the availability of a sufficiently large dataset. Also in this case, the optimization methodology is based on the workflow presented in Chapter 2, and the resulting neural network is re-trained on 20% of the entire training dataset. The final structure is composed of a simplified block-structure made of dense+normalization layers with size of integer multiples of the input size of 4, as reported in Figure 4-11. The optimization of the weights of the neural network has been therefore performed with the Adam optimizer [75] and a learning rate set to 0.05, thanks to the presence of batch normalization layers, for an infinite number of iterations, stopped by the early stopping techniques after 100 iteration without improvement on a random 5% validation sub-set.

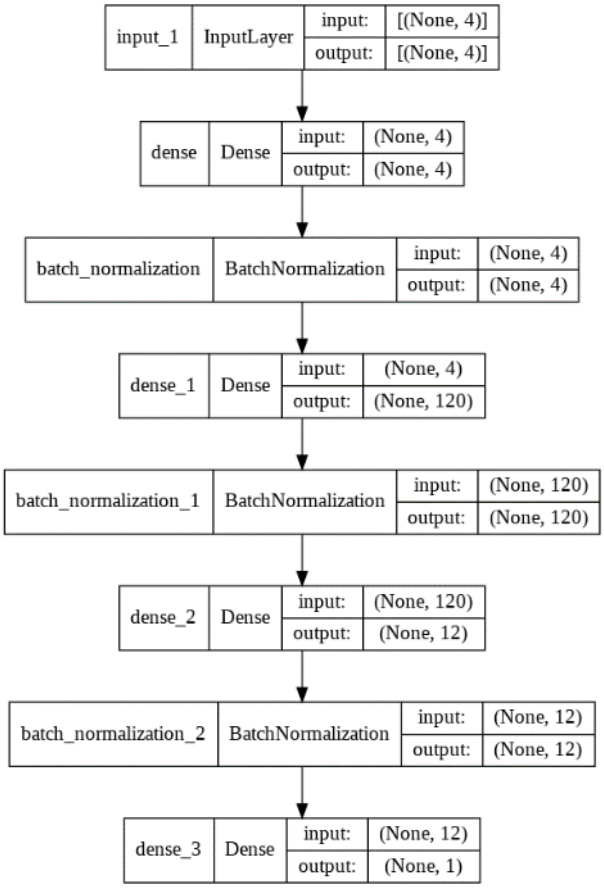

Figure 4-11: Architecture of the optimized DNN model

To visually understand the meaning of the employed metrics, the results of laminar flame speed for the 3 mechanisms and the values predicted by the XGBoost algorithm are shown as a function of the equivalence ratio for an unseen temperature and pressure condition. From the visual comparison, the predicted values present a reasonable profile, as well as values closer to the target with respect to the results obtained with the most reduced mechanism. However, the relative difference changes direction passing from lean to rich mixtures, which is an important aspect to be considered, since it might lead to erroneous interpretation of the combustion results.

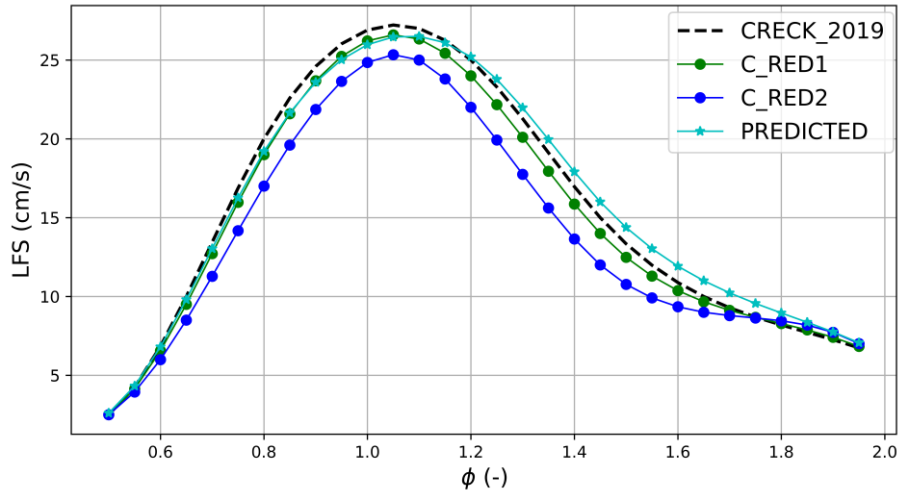

Figure 4-12: Laminar flame speed of TAE7000 with different methods for different values of Equivalence ratio, P=25 bar, T=600 K, EGR=5%

4.4.2 Optimization and Validation set results for the accuracy recovery methodology

The workflow adopted for the definition of the machine learning regression problem with accuracy recovery is based on the assumption that the relative error between the laminar flame speed obtained with the complete mechanism and that obtained with a reduced version (in this case CRECK2019 and C_RED2) differ by a proportionality coefficient that behaves non linearly with the thermodynamics properties and thus needs to be inferred by non-linear regression models. The first step for the application of this approach is, therefore, the definition of the new target variable, defined as the relative difference between detailed and reduced results:

$$\eta_{ML} = \frac{(s_L)_{1D_{detailed}} - (s_L)_{1D_{red}}}{(s_L)_{1D_{red}}} \quad (4\text{-}13)$$

From this equation, it is straightforward that the recovery of the approximate laminar flame speed solution will be derived from Eq. 4-11.

The results of the same algorithms employed for the previous approach are directly applied on the final laminar flame speed value, and not to the intermediate target variable. Also in this case, an initial optimization of the required dataset size has been performed with the optimized XGBoost algorithm, leading to the identification of the minimum dataset-size required to outperform the reduced mechanism in terms of both mean and maximum relative error. The first conclusion that can be drawn is related to the overall better performance of the accuracy recovery method with respect to the standard regression, with a drop from about 7% to less than 2% average error with as little as 5% of the training set size. In order to control also the maximum error committed, however, the chosen training set size has been defined as 10% of the available dataset.

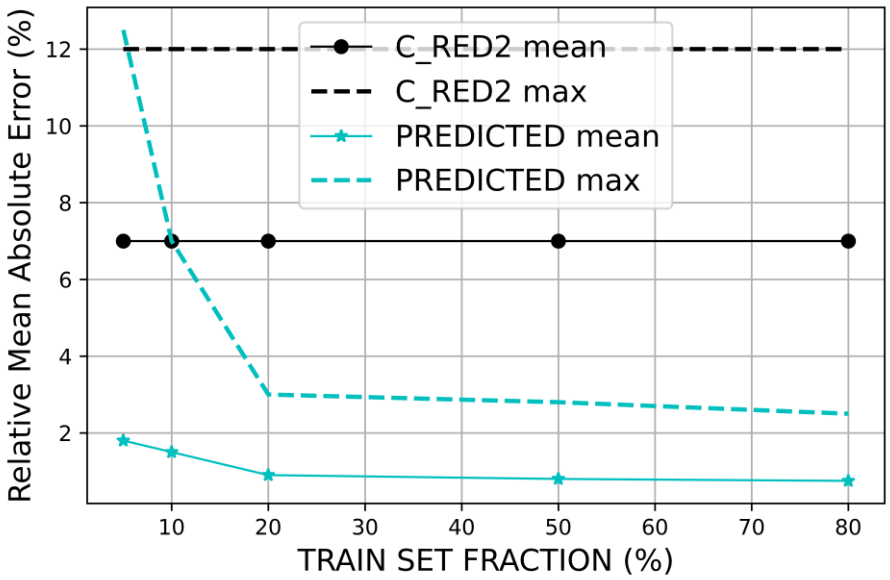

Figure 4-13: Relative difference (mean and max) for the C_RED2 mechanism and accuracy recovery

The results of all the regression algorithms considered are reported in Table 4-6 together with the optimized neural network architecture, confirming the capability of the boosted tree model to outperform other regression techniques when dealing with tabular data.

Table 4-6: Performance metrics on the classical machine learning regression models

Model	MAE (m/s)	RMSE (m/s)	R2	J (m/s)
Elastic net regression	0.041	0.049	0.956	0.076
Polynomial regression with 2nd order features	0.038	0.042	0.974	0.064
Polynomial regression with 3rd order features	0.031	0.037	0.985	0.067
Support vector machine regression with RBF kernel	0.028	0.033	0.992	0.053
Gaussian process regressor with RBF kernel	0.026	0.031	0.995	0.049
Decision tree regressor	0.032	0.043	0.985	0.061
Random forest regressor	0.019	0.023	0.998	0.041
Adaboost regressor	0.015	0.019	0.999	0.032
XGBoost	0.011	0.014	0.999	0.027
Optimized DNN	0.012	0.017	0.999	0.029

The results in terms of visual comparison of the same operating condition reported for the simple regression problem are reported in Figure 4-14 for the accuracy recovery method, underlying a very accurate result with respect to the reference values.

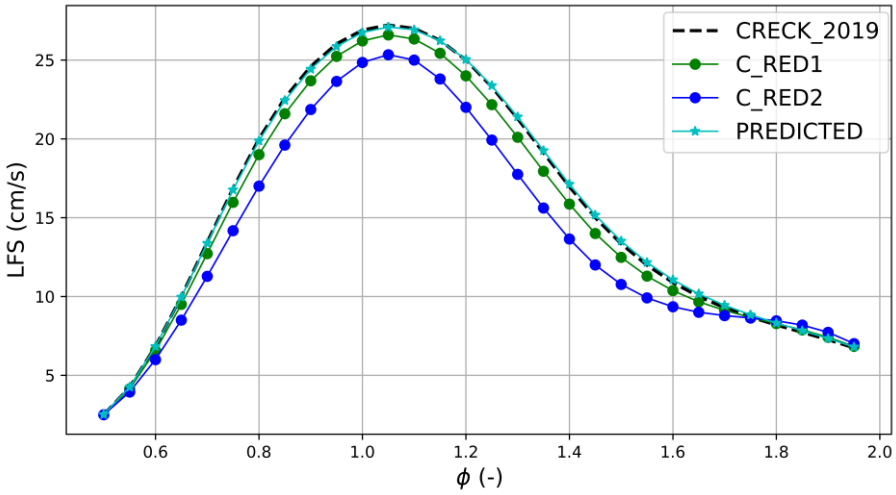

Figure 4-14: Laminar flame speed of TAE7000 with different methods for different values of Equivalence ratio, P=25 bar, T=600 K, EGR=5%

The relative error of the different methodologies is reported in Figure 4-15, where it can be noticed that the accuracy recovery approach displays a superior performance both in terms of mean error and error distribution, which is almost normally distributed around 0 and limited between -11% and 9%.

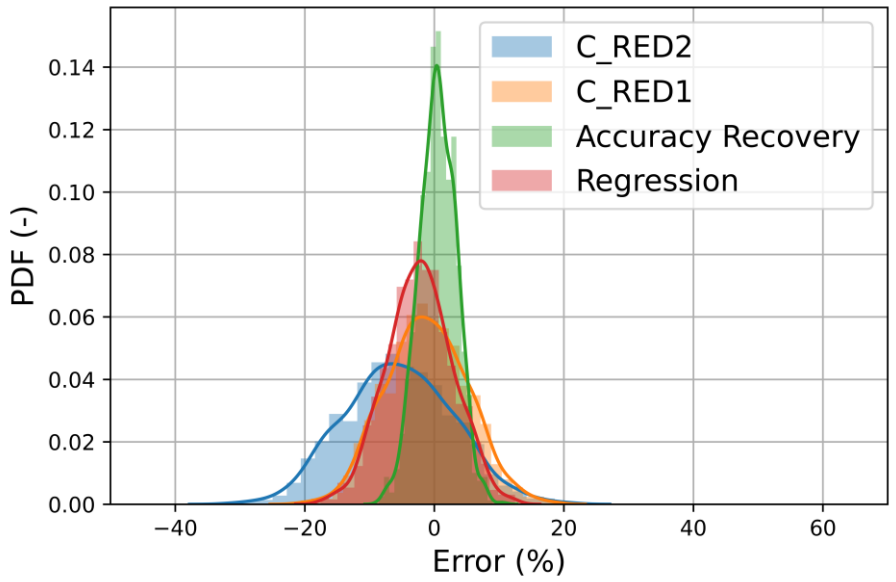

Figure 4-15: Relative error distribution on the whole dataset for the different approaches proposed

On the other hand, this approach carries a higher computational cost compared with the pure regression model, as reported in Table 4-7 which is only partly alleviated by the requirement for a smaller training dataset. The training and inference time of the machine learning algorithms overhead is not considered, given the difference by order of magnitudes with chemical kinetics simulation times.

Table 4-7: Computing time required for the generation of a complete look-up table of laminar flame speed values

GENERATION of 19000 points	Simulation time with CRECK2019 (CPUhr)	Simulation time with C_RED1 or 2 (CPUhr)	TOTAL TIME (CPUhr)
CRECK2019	129’200	-	129’200
C_RED1	-	47’500	47’500
C_RED2	-	36’100	36’100
Regression	25’840	-	25’840
Accuracy Recovery	12’920	36’100	49’020

4.5 Water Injection in internal combustion engines

The water injection strategy has become of increasing interest for the development of SI engines, even if application examples have been provided for several decades for increasing the thrust at low speed of jet engines, or mitigate the knock risk of gasoline engines before the introduction of the intercooler technology [76] [77]. In recent years, interest towards this technology has grown, since it is considered a viable alternative to reach the targets of reduction of CO2, NOx and fuel consumption that are required by the current regulations, in particular as enhancer for the lambda=1 strategy that is expected to be enforced by future regulations on the whole engine map [78] [79] [80].

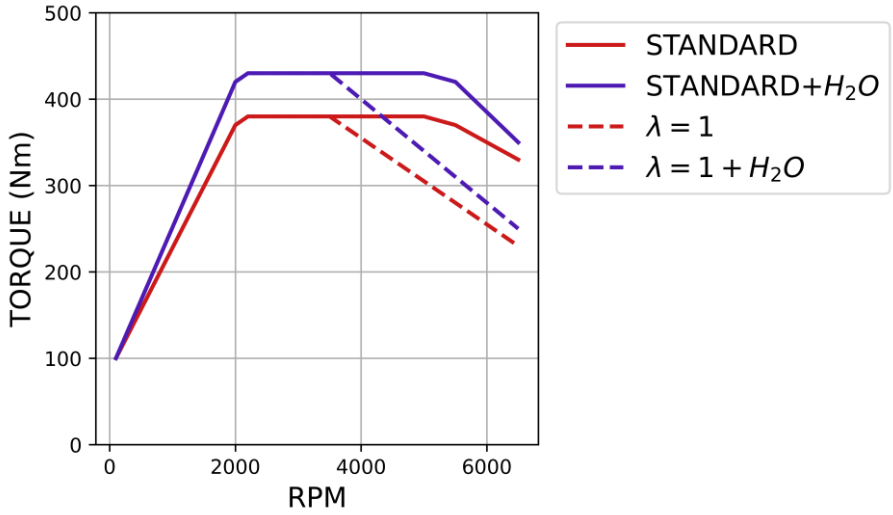

Figure 4-16: Performance map enhancement at lambda=1 thanks to water injection strategy adoption

Recently, a new correlation was proposed for keeping into account also the effect of water vapour as a diluent, based on detailed chemical simulations performed on relevant conditions that are reached during engine operation [81]. The results of the correlation, in terms of sensitivity to the diluent effect of water addition at relevant conditions, were adequately in agreement with the values of the detailed chemical simulations. On the other hand, it resulted that the use of a literature-standard power-law function for higher P and T

values led to the prediction of unphysical values of LFS for ϕ greater than stoichiometric. The effect of the addition of water and EGR in reducing the LFS was, however, captured well by using a linear correlation, in the form of Eq. 4-14:

$$\frac{s_L}{s_{L_0}} = (1 - k_{EGR} \cdot X_{EGR}) \cdot (1 - k_{wat} \cdot X_{H2O}) \quad (4\text{-}14)$$

where s_L is the actual LFS, s_{L_0} indicates LFS at the same P, Tu and ϕ but with $X_{EGR} = 0$ (EGR mass fraction) and $X_{H2O} = 0$ (water vapour mass fraction), and k_{EGR} and k_{wat} are the correlation coefficients.

In order to overcome these limitations, and following the ideal target of generating a laminar flame speed look-up table for online interpolation during the CFD analysis, the water vapour mass fraction value must be considered as an additional independent variable. The upper bound for the water mass fraction inside the mixture is fixed at 6%, considering that the evaporation times of the liquid droplets would not allow to practically employ fractions higher than 4%, which is assumed to locally reach values not higher than 6%. The analysis reported in this section refers to the simulations performed on the TAE7000 surrogate but can be qualitatively extended to the other fuels. It can be noticed that the effect of the addition of water vapour to the unburnt mixture definition leads to a decrease in terms of LFS. This effect, as reported for two combinations of P and T_u in Figure 4-17 is higher for near stoichiometric reacting mixtures, underlying a nonlinear behaviour that a traditional correlation would have omitted.

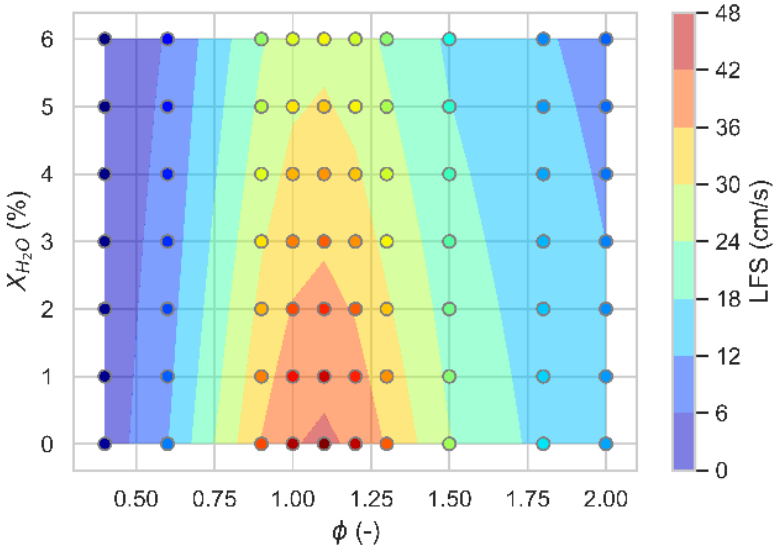

Figure 4-17: Laminar flame speed of the TAE7000 gasoline fuel surrogate as a function of Equivalence ratio and water vapour mass fraction at P=50 bar, T=600 K and EGR=1%

From Eq. 4-14, the effective value of a linear correlation coefficient k_{wat} can be calculated from the simulated values of laminar flame speed, where s_L represents the actual LFS with

a mass fraction of water vapour equal to X_{H2O} and s_{L_0} the LFS at the same conditions of P, Tu, ϕ and X_{EGR} without water addition:

$$k_{wat} = \left(1 - \frac{s_L}{s_{L_0}}\right) / X_{H2O} \tag{4-15}$$

The distribution of the relative error committed in calculating the LFS on a subset of a full-grid database, using a single value of k_{wat} with respect to the output of the simulations is reported in Figure 4-18. It can be noticed that the peak of the distribution is near 0, which derived from the choice of the best fitting value of k_{wat}, but in more than 50% of the simulated points the effect of water addition is overestimated.

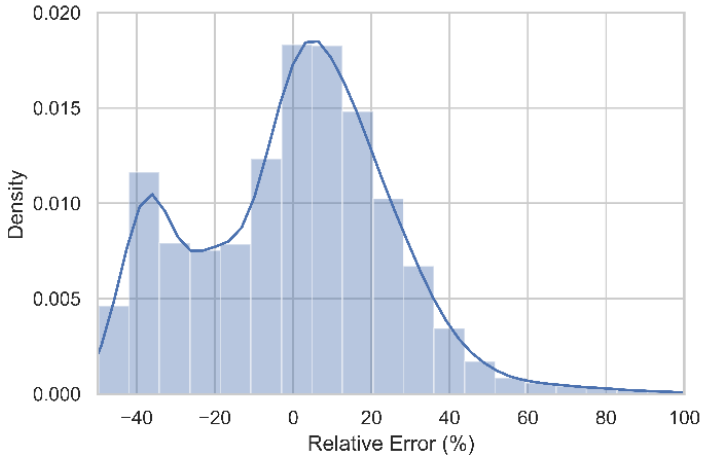

Figure 4-18: Relative error distribution between the optimized constant k_{wat} and its real value

4.5.1 Machine learning strategy for laminar flame speed look-up tables accounting for water vapour

Starting from the overall acceptable performance of a simple constant linear coefficient for the inclusion of water vapour into the definition of laminar flame speed, the hybrid machine learning methodology employed consists in the definition of non-linear, variable proportionality coefficient, whose value needs to be inferred by a machine learning algorithm, as reported in Eq. 4-16.

$$\widetilde{s_L} = s_{L_0} \cdot (1 - k_{ML}(P, T_u, \phi, X_{EGR}, X_{H2O}) \cdot X_{H2O}) \tag{4-16}$$

The additional hypothesis of a linear correlation of LFS with X_{H2O}, even if performed with a variable proportionality coefficient, enforces the equivalence of the predicted laminar flame speed with the reference value, which is a desirable property. It is however necessary to include the water vapour mass fraction inside the input variable for the regression algorithm, considering that it is expected to learn the mutual relations between the other variables and

the water mass fraction. Also for this application, the use of the minimum number of simulated point is a requirement for the applicability of the method, therefore an analysis on the performance on a hold-out validation dataset has been performed as a function of the size of the training set. The results are reported in Table 4-8 and highlight the fact that the application of the hybrid approach requires a limited training dataset in terms of new simulated points for extending a pre-defined laminar flame speed look-up table of reference size to include the effect of water vapour inside the domain. The comparison was carried out only with the optimized XGBoost algorithm, considering the optimal results obtained for previous similar tasks.

Table 4-8: Performance metrics of XGBoost as a function of the size of the additional dataset

Additional points	MAE (m/s)	RMSE (m/s)	R2 (-)
1%	0.008	0.008	0.98
2%	0.005	0.003	0.99
5%	0.003	0.002	0.99
10%	0.003	0.001	0.99
20%	0.002	0.001	0.99

5 Algorithm for fuel surrogate definition and analysis

While the methods for simulating the behaviour of hydrocarbon fuels are essentially well established within the scientific community, a unique definition of the fuel composition is still an open issue. Focusing purely on gasoline fuel and bio-derived blends, which are currently the most widely adopted energy sources for light duty transport, the main reasons behind this issue can be divided into 2 groups:

1) Variability in pump gasoline composition due to different origins, seasonality and experimental uncertainty.
2) Lack of detailed molecular modelling for the potentially thousands of single molecules constituting the target blend.

While the first concern can only be contained by providing information related to the experimental accuracy, but not entirely solved, the second point has been addressed by the research community through the development of simplified mixtures, defined to match some relevant properties of the real target. These simplified models are generally referred to as fuel surrogates and have been the subject of several studies over the years.
In the present chapter, after a review of the evolution of gasoline fuel surrogate formulations, attention will be given to:

1) the identification of potential target fuels, for which a wide experimental characterization is available,
2) the definition of mixing laws for the rapid identification of mixture properties after the fuel surrogate formulation,
3) the development of a multi-objective optimization algorithm for the efficient identification of suitable surrogates.

The methodology will then be applied to 2 target fuels, identifying a correlation between the type and number of experimental data available, and the expected unicity of the results.

5.1 Classical gasoline fuel surrogate formulations

Many approaches have been proposed in literature to define surrogates matching some known properties of the real fuel [82] [83] [84], from the common PRFs (mixtures of isooctane and n-heptane that can easily match RON values) [85] to TRF-E (PRF + toluene and ethanol, in order to better represent the composition of the pump gasoline and allow for different sensitivity values [86] [87] [88]) to more complex mixtures. For example, in the seminal work by Mehl e al. [89] the authors define a 4-components fuel surrogate mimicking RON/MON and information regarding the composition of the target fuel by leveraging their expertise and using validated correlations to define the properties of the mixture. Puduppakkam et al. [90] have described the process for formulating a complex 5-component fuel surrogate emulating the known properties of a target gasoline validated under HCCI conditions, underlying how the capability of a more complete surrogate could outperform the predictions of simpler mixtures. Despite the frequent availability of some information regarding both the liquid and vapor phases of the fuel, it is frequently found in literature a

differentiation between surrogates employed for the liquid phase (necessary for spray and mixing simulations) and for the gas phase (used for the simulation of the combustion process). In the present chapter the focus is on developing surrogates for the gas phase prevalently, however, physical properties of the liquid phase are considered as well, due to their larger availability, with the aim of providing the optimizer with more information that could correlate with reactivity data. Considering the fact that some values, like laminar flame speed or ignition delay times, are rarely identified during the characterization of target fuels due to the requirement of more expensive [91], longer or more complicated experimental tests, the aim of the next sections is to identify the minimum set of properties that are required to define a representative gasoline surrogate for combustion CFD simulations within a given confidence interval.

Furthermore, the possibility to quantify the predictive consistency of simulations performed with different surrogates of the same fuel, as a function of the number of optimized variables, is expected to quantify the attended level of uncertainty that the research community must consider from the definition of a gasoline fuel surrogate with a limited number of known properties. The definition of the minimum set of necessary properties and the attended confidence level is expected to increase the speed of development of fuel mixtures with new bio-derived components, using CFD simulations.

5.2 Target fuels

The developed approach will be validated against two well described fuels available in literature with properties in line with pump gasoline, and the conclusions regarding the representability of the final surrogates will be drawn from these two applications. Therefore, it is necessary to underline the fact that they can be considered relevant especially for the simulation of gasoline fuel combustion, which is the main problem addressed, but they may not hold true for different groups of combustibles.

The first target is a fuel mixture developed by Shell Global Solutions defined ‘Shell-D’ and its properties (composition, density, molecular weight, rvp, distillation curve, ignition delay time and laminar flame speed) were experimentally characterized by Xu et al. [92]. It is considered representative of an E5 pump gasoline due to its main composition, octane number and LHV. The second target fuel is a pump RON96 E5 gasoline analyzed by Esposito et al. [93] for which the main properties are available but lack experimental data regarding laminar flame speed and ignition delay time. To complement these data, experimental values were taken as representative of the same fuel from where a RON96 E5 gasoline was experimentally tested to identify the chemical kinetics behavior [94].

The detailed properties of both targets are reported in Table 5-1

Table 5-1: Properties collected for the target fuels

Properties	SHELL-D	MARKET GASOLINE
RON (-)	96.2	96.5
MON (-)	82	86
LHV (MJ/kg)	41	42
H/C (-)	1.9	1.8

O/C (-)	0.03	0.02
AFS (-)	14.6	14.3
AL-P-I-O-N-A (%mass)	10-10-30-8-6-36	5-50-NA-12-NA-33
IDT (ms)	Available	Available
LFS (cm/s)	Available	Available
Liquid Density (kg/m3)	742	729.9
T10-T50-T90 (°C)	60-111-188	40-80-160
Mw (g/mol)	97.4	95.5
RVP (psi)	7	9

Table 5-2: Properties chosen for the different surrogate levels

Properties	Level 1	Level 2	Level 3
RON	Yes	Yes	Yes
MON	Yes	Yes	Yes
LHV	Yes	Yes	Yes
H/C	Yes	Yes	Yes
O/C	Yes	Yes	Yes
AFS	Yes	Yes	Yes
AL-P-I-O-N-A	No	Yes	Yes
IDT	No	No	Yes
LFS	No	No	Yes
Additional properties			
Mw	Yes	Yes	Yes
Liquid Density	Yes	Yes	Yes
T10-T50-T90	No-No-No	No-No-No	No-Yes-No
RVP	No	No	No

The availability of experimental data regarding the properties of a given fuel usually requires a period in the order of several days and, depending on the availability of appropriate experimental instrumentation, a longer amount of time is expected to be required for the identification of the reactivity of the mixture. Considering the data available in literature from the extensive literature review on fuel surrogate formulations [95], an attempt has been made internally at identifying 3 sets of properties with increasing level of resources required for their identification, reported in Table 5-2.

Level 1 surrogates are the most found in literature, matching properties that can be obtained from liquid phase characterizations and requirements in terms of RON MON and LHV. For reaching level 2 accuracy, it is required to include a detailed molecular analysis in the characterization of the fuel, which is reported for more detailed surrogates. Level 3, on the other hand, is often not reached in literature because the values of laminar flame speed or ignition delay times are rarely publicly available. The addition of the temperature at which 50% of the distilled volume is recovered adds an additional information that could help identify a unique surrogate, whereas the T10 and T90 values have not been added as

optimization targets due to the expected difficulty in reaching reasonable values while targeting primarily the gas phase [96].

5.3 Optimization algorithm

5.3.1 Bayesian Optimization

The optimization of a non-linear function ***f(x)*** in a domain **D**, is a typical task addressed in scientific literature. Many approaches have been proposed, for example:
Grid search, a classical parameter sweep where every combination of the discretized domain **D** is tested. This approach guarantees that the best result is achieved in case of problems with discrete inputs, but it relies on the spatial discretization of the domain in case of continuous inputs. Despite the straightforward parallelization of the algorithm, its main drawback is the time required to evaluate all the combinations, especially for higher dimensional problems, or for tasks where the acquisition function is computationally expensive.
Random search is a derivation of the grid search algorithm, where the combinations are chosen following a probability distribution, in order to reduce the sampling time at cost of a decreased accuracy.
Gradient based optimization is a family of algorithms that exploit the knowledge (where applicable) not only of the point value at the sampled inputs combination, but also of its derivatives, in order to better evaluate the next optimal evaluation points. The main drawbacks of this method are the requirement to calculate the derivative of the acquisition function, which is not always possible, and the reduced parallelization exploitability.
Genetic algorithms are a family of global optimization algorithms that mimic some aspects of the biological evolutionary theory, in search of the best performing set of inputs as ranked by a fitness function. They are based on a repetition of a set of operations aimed at continuously increase the performance, while, at the same time explore new combinations of inputs. The basic process adopted by these algorithms is the following:

- initialization of the population (definite number of initial points in domain **D** for which ***f(x)*** needs to be evaluated);
- evaluation of each set of points concurrently evaluated in the current generation (based on the objective function to be maximized), step that can eventually be parallelized;
- selection of the fittest candidate points in the generation, whose features are used (following appropriate strategies, eventually tuned) to perform crossovers (recombination of the parameters in the fittest elements) and mutations (changing some parameters in order to continue exploring a larger search domain), in order to define the following generation;
- after a prescribed number of generations, or when reaching a satisfying result, the process can be stopped.

Depending on the application fields, many variations have been proposed to this simple procedure, but they are mainly devoted to the identification of specialized strategies to

perform the crossover and mutation steps in more effective ways, leaving the underlying method unchanged.
Despite its applicability to any kind of complex optimization problems, the use of evolutionary algorithms greatly relies on their implementation with respect to the problem and heuristics, to adapt the mutation and crossover strategies to the task. In addition, there is no formal indication that the genetic algorithm will converge to a global optimal solution in a prescribed number of iterations.
A second kind of population based algorithms, named Bayesian Optimization, has gained much attention in the last decade, due to its ability to reach optimal results in a limited number of iterations, when compared to other methods. Its main drawback is the complexity of the method, which is $\mathcal{O}(n^3)$, where n is the number of evaluations available for fitting the surrogate model, which is, for most application, a gaussian process regressor. The high computational costs makes this method difficult to apply for more than a few hundred points, but, on the other hand, it has been empirically showed on standard optimization benchmark problems, that the reached performance would be similar to that reached by other algorithms in a larger number of evaluations. Therefore, the use of the Bayesian optimization algorithm is being widely adopted especially for the optimization of functions that require long evaluation times, in order to compensate the computational overhead induced by the algorithm [Time Efficiency in Optimization with a Bayesian-Evolutionary Algorithm], for example during the hyperparameter tuning of a machine learning algorithms. The Bayesian optimization algorithm follows a series of steps tat can be summarized as:

- Define a gaussian process model (see ch. 1) with a prior describing the function ***f(x)*** and evaluate with the acquisition function a number of initial points according to a design of experiments methodology.
- For a prescribed number of iterations:
    - update the posterior probability of the gaussian process model using all the available data sampled,
    - use an acquisition strategy, to define the next sampling point based on the outputs of the GP model on the entire domain (both mean values and variance).
    - Evaluate function ***f(x)*** at the point that maximizes the acquisition function and add the result to the available data for the update of the GP model.

The Bayesian optimization algorithm greatly relies on the definition of the acquisition strategy for its efficiency, which is commonly defined as an Upper Confidence Bound method. This approach formalizes the so called ‘exploration vs exploitation tradeoff’ [97], since it consists in maximizing a combination of both mean value and variance.

$$UCB(\boldsymbol{x}, \beta) = \mu(\boldsymbol{x}) - \beta\sigma(\boldsymbol{x}) \tag{5-1}$$

The constant coefficient β is expected to have positive value, and it clearly represents the chosen exploration term of the function, whereas the maximization of the mean value of the predicted GP regressor represents the exploitation term. The acquisition function consists in the choice of the input variables $\boldsymbol{x}$ that maximize the acquisition function. The most

commonly used acquisition function is called Expected Improvement and it leverages to a higher extent the posterior description provided by the GP model. In particular, the expected improvement function is set to 0 for previously acquired points, or points with null standard deviation from the GP regressor, otherwise it is defined as:

$$EI(\boldsymbol{x}) = (\mu(\boldsymbol{x}) - f(\boldsymbol{x}^*) - \xi)\boldsymbol{\Phi}(Z) + \sigma(\boldsymbol{x})\phi(Z) \quad (5\text{-}2)$$

With $\boldsymbol{x}^*$ defined as the best sample so far, ξ defined as the parameter that quantifies the ratio between exploitation and exploration during the optimization approach, $\mu(\boldsymbol{x})$ and $\sigma(\boldsymbol{x})$ the mean and standard deviation predicted by the GP posterior, while $\boldsymbol{\Phi}(Z)$ and $\phi(Z)$ are the CFD and PDF of variable Z defined as:

$$Z = \frac{\mu(\boldsymbol{x}) - f(\boldsymbol{x}^+) - \xi}{\sigma(\boldsymbol{x})} \quad (5\text{-}3)$$

5.3.2 Multi-objective optimization problem definition

Considering the non-linearity of several properties of the target fuel and the size of the palette, a multi-objective optimization algorithm should be defined. To perform a sufficiently refined grid search in the whole domain, the number of combinations would be prohibitive, therefore a more advanced algorithm should be employed. A Bayesian optimization algorithm was implemented with a custom acquisition function necessary considering the large number of target properties that describe the multi-objective optimization problem.
Considering the strong importance that the research and industrial community has always attributed to the RON value for differentiating between fuels, it has been considered as a constraint, in order to reduce the number of evaluations that could not be considered acceptable. Considering the values of RON of the single molecules and the fact that the calculation method is expected to be accurate within a 3% error, the optimization algorithm is based on 10 variables, of which 9 are the mole fractions of the single molecules except for the aromatics, and the 10th variable is the ratio between toluene and 124TMB. The first step is the calculation of the RON value of the normalized surrogate without aromatics (RON_surr0), and of the mixture of the 2 aromatics molecules (RON_A), from which the fraction of aromatics in the final surrogate is calculated following Eq. 5-4.

$$X_A = \frac{RON_{target} - RON_{surr0}}{RON_A - RON_{surr0}} \quad (5\text{-}4)$$

After this initial step, all the other properties and errors are calculated with the methods described in the next sections and the final target function is computed to account for the multi-objective nature of the optimization problem.
The error function J to be minimized is defined as in Eq. 5-5 as the normalized weighted Euclidean distance of the absolute relative errors for each target plus a penalty term introduced to reduce the complexity of the final surrogate.

Optimization algorithm

$$J = \left(\frac{1}{\sum_{i=1}^{N_{targets}} w_i} \cdot \sum_{i=1}^{N_{targets}} w_i \cdot abs\left(\frac{property_i - target_i}{target_i} \right)^2 \right)^{0.5} + w_{complexity} \cdot \frac{\sum_{i=1}^{Npalette} 1 \; if \; X_i > 0 \; else \; 0}{\#palette} \quad (5\text{-}5)$$

The weights used are the same for each level analyzed and have been obtained by heuristics, and their value is reported in Table 5-3. In particular, larger importance has been given to values such as AFS or LHV that would directly impact the simulation results, and the importance of the laminar flame speed and ignition delay time targets have been recomputed based on the number of available points, in order to make the system independent of the size of the target dataset.

Table 5-3: Weights associated with each single property during the multi-objective optimization

Properties	weight
RON	1
MON	2
LHV	3
H/C	1
O/C	1
AFS	4
AL-P-I-O-N-A	2-1-1-1-1-1
IDT	5/#points
LFS	5/#points
Additional properties	
Liquid Density	1
YSI	0
T10-T50-T90	0-2-0
RVP	0

The Bayesian optimization algorithm employed has been thoroughly described in [96] which is based on a 2-step approach:

- an initial number of evaluations is performed by randomly searching inside the exploration domain, in order to gain information regarding the distribution of the target function and provide an initial training set for the surrogate model. In this case, the regression algorithm chosen is the gaussian process regressor (with a radial basis function kernel and a value of gamma = 0.5), this regression algorithm was chosen since it allows to perfectly fit the training data (even if the value of gamma has been adequately tuned to avoid excessive overfitting) and to predict the accuracy of the estimate on unknown points.
- for a prescribed number of optimization steps, a search algorithm (in this case the Upper Confidence Bound algorithm) is used to identify the next acquisition point to be evaluated with the aim of maximizing the target function and increase the accuracy of the posterior distribution approximated by the Gaussian Process algorithm.

After internal testing on the effect of the size of the initial training set, an empirical value of 400 evaluations was identified as a threshold number for leading to a consistent result during the optimization step, independently from the number of optimization steps (at least 30).

5.4 Computation of mixture properties

5.4.1 Average molecular properties

Several properties of a hydrocarbon mixture can be assumed as linear combination of the properties of the single molecules weighted with their molar fractions. This hypothesis is inherently true for HC ratio, OC ratio, stoichiometric Air Fuel ratio, density and ALPIONA composition, synthetized in Eq. 5-6 to 5-11.

$$M_w = \sum_{i=1}^{N_{PIONA}} x_i \, M_{w_i} \tag{5-6}$$

$$H/C = \frac{\sum_{i=1}^{N_{PIONA}} x_i \, H_i}{\sum_{i=1}^{N_{PIONA}} x_i \, C_i} \tag{5-7}$$

$$O/C = \frac{\sum_{i=1}^{N_{PIONA}} x_i \, O_i}{\sum_{i=1}^{N_{PIONA}} x_i \, C_i} \tag{5-8}$$

$$AFS = \sum_{i=1}^{N_{PIONA}} x_i \, (C_i + \frac{H_i}{4} - \frac{O_i}{2}) \tag{5-9}$$

$$AL - P - I - O - N - A = \sum_{i=1}^{N_{PIONA}} x_i \tag{5-10}$$

$$\rho_l = \sum_{i=1}^{N_{PIONA}} x_i \, \rho_{l_i} \tag{5-11}$$

Since no experimental data are available regarding the sooting tendency of the target fuels, its value, calculated for each surrogate is employed only to measure the relative distance between two formulations. The sooting tendencies of the surrogates, identified by means of the Yield Soot Index, are predicted following [93] by a linear mixing rule based on the mole fraction of the components. The values for each molecule were experimentally evaluated in [98] except for 1-pentene, for which the YSI value was estimated based on its molecular structure.

As far as the lower heating value is concerned, a simple weighted mean is often proposed [99], whereas, in order to improve the accuracy of the prediction, in the present work a different approach has been developed involving chemical kinetics equilibrium computations. After defining the surrogate, the Lower heating value is obtained by performing a stoichiometric reaction at ambient temperature and pressure and then computing the difference in enthalpy between products and reactants divided by the mass fraction of fuel in the reactants, without considering the condensation of water (which would lead to the HHV value instead). These values have been validated against the single molecules with respect to experimental data from [100] with a relative error below 1%.

5.4.2 RON and MON

The octane number of a fuel mixture, both in terms of RON and MON is experimentally is classically computed with the mole weighted mean of the values of the single molecules numerically. It has been demonstrated [101] [102], however, that this holds accurate for simple TRF mixtures, while the addition of oxygenates to the blend, chosen as target in the present work, requires a non-linear treatment. The chosen model is the one presented by Anderson et al. [102] implemented as a two-step approach following Eq. and Eq. . First the RON/MON value for the mixture without alcohol is computed as a weighted mean, to which a second term is added computed as the alcohol mole fraction with a quadratic dependence, multiplied by a coefficient K_g with values 0.45 for RON and 0.94 for MON according to the guidelines provided by the authors.

$$ON_g = \sum_{i=1}^{N_{PIONA}} x_i \, ON_i \tag{5-12}$$

$$ON = \sum_{i=1}^{N} x_i \, ON_i + K_g \, x_{AL} \, (1 - x_{AL}) \left(ON_{AL} - ON_g \right) \tag{5-13}$$

5.4.3 Ignition delay time

For the ignition delay time prediction, the model developed in Chapter 3 has been directly applied to the newly defined surrogates at the same operating conditions available from the experimental characterizations. The attended accuracy is expected to be below 2% with respect to detailed chemical kinetics simulations, after the thorough validation that has been performed.

5.4.4 Laminar flame speed

The consumption speed of a reacting mixture, necessary for the flamelet combustion models, can be obtained by performing detailed chemical kinetics simulations of stationary freely propagating adiabatic flames, in which the displacement speed at steady state equals the rate of consumption on the flame front. In order to reduce the computational time, a one-dimensional domain is usually employed (acceptable given the conditions that induce a planar flame front propagation), and the algorithm defined in chapter 4 was followed to perform a more robust and accurate prediction. Considering the time required to perform a single simulation with the reduced mechanism (approximately 2 hours on a single processor even with the reduced version of the chemical kinetics mechanism), it is unfeasible to perform complete simulations for each surrogate evaluated, therefore a mixing law has been implemented.

A series of approaches has been tested to compensate this limitation:

- the classical Le Chatelier's mixing rule which assumes that the laminar flame speed of a fuel surrogate is equal to the molar weighted sum of its components [103], following Eq. 5-14 where Z_i indicates the molar fraction of the i-th component in the fuel surrogate definition

Computation of mixture properties

$$LFS_{mix} = \frac{1}{\sum_{i=1,N} \frac{Z_i}{LFS_i}} \quad (5\text{-}14)$$

- A modified Le Chatelier's rule based on the energy fraction of the single component, calculated as the product of the molar fraction of the component and its lower heating value LHV (Eq. 5-15) introduced by Sileghem [104].

$$LFS_{mix} = \sum_{i=1,N} LFS_i \cdot \left(\mathrm{LHV_i} \cdot \frac{X_i}{LHV_{mix}}\right) \quad (5\text{-}15)$$

- The Hirasawa mixing rule [105], based on a logarithmic correlation with the adiabatic flame temperature of each component as the weighting term as in Eq. 5-16 where n represents the number of moles

$$LFS_{mix} = \exp\left(\sum_{i=1,N} \ln(LFS_i) \cdot \left(\frac{Z_i n T_{ad,i}}{n_{mix} T_{ad,mix}}\right)\right) \quad (5\text{-}16)$$

  In order to estimate the adiabatic flame temperature for all configurations, considering its relevance in the estimation of the laminar flame speed of the mixture, a numeric procedure can be applied. It is required to identify the composition and temperature of the exhaust gas (considering the variation in H/C and O/C ratio and the different equivalence ratio that must be accounted for) by reaching the equilibrium state in a constant enthalpy and constant pressure 0D reactor, which can be computed in a few milliseconds.

- The Yumlu mixing rule [106] which is a simplification of the correlation proposed by Semenov [107] where the additivity of the laminar flame speed of the single components to the mixture is accounted for by the term ω_i which is calculated as the normalized mass fraction of fuel component I and the corresponding mass fraction of oxidant, leading to Eq. 5-17.

$$LFS_{mix}^2 = \sum_{i=1,N} LFS_i^2 \cdot \omega_i \quad (5\text{-}17)$$

- a data driven approach, following the same workflow presented in the previous section, i.e. an optimized machine learning algorithm with normalized mole fractions and test conditions as input and the laminar flame speed value as output. The final model selected after repeated k-fold cross validation on the dataset is a neural network composed of the input layer with 15 units, 3 hidden layers (respectively of 60,180,30 units and ReLu activation function) and 1 neuron in the output layer with

linear activation function, due to its better generalization with respect to tree-based regression algorithms.

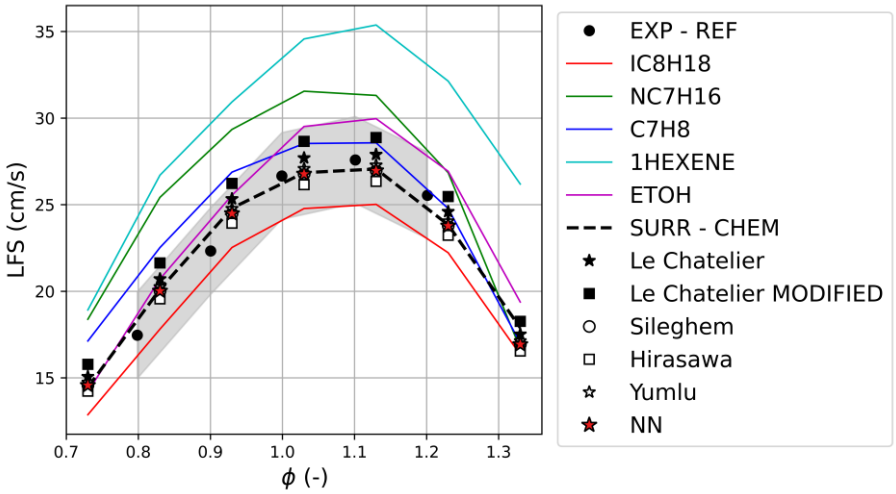

Figure 5-2: Laminar flame speed at reference conditions (P=10 bar, T=373 K) for a generic gasoline fuel

A dataset of laminar flame speed for generic mixtures of components from the defined palette has been generated through chemical kinetics simulations. Since the operating conditions for which experimental data are known, the dataset is limited to variations in fuel mixture and equivalence ratio, while pressure and temperature are kept constant. The values of the single components, as well as that of the whole dataset are reported in Figure 5-2 compared to the experimental data for a generic gasoline fuel under the same thermodynamics conditions [108]. The same dataset of mixture fractions used for the ignition delay time computation was employed for the laminar flame speed calculation and the validation set of about 90 mixtures was used for validation and comparison with the other mixing rules in terms of R2 and RMSE.

Table 5-4: Performance metrics of the selected methods for the laminar flame speed of a mixture

	RMSE (m/s)	R2
Le Chatelier	0.04	0.981
Modified Le Chatelier	0.03	0.992
Hirasawa	0.018	0.998
Yumlu	0.02	0.999
DNN	0.015	0.999

Considering the dependency of the data-driven approach to the training dataset, in the end the choice of using the Hirasawa mixing rule has been taken, which requires the computation of the laminar flame speed values for the single molecules before the optimization process starts, but only for the experimental reference conditions, which are expected to be limited. For the presented examples, only 6 to 8 operating conditions are available, making the a-priori computation of the laminar flame speed for each single molecule feasible.

5.4.5 Liquid-Vapour equilibrium properties

The properties of the liquid-vapor equilibrium have been computed by an in-house developed algorithm described in [96], that performs the iterative solution of the system of equations Eq. 5-18, which accounts for:

i) the liquid-vapor equilibrium for non-ideal fluids (1st Eq. 5-18), which is an implicit in the temperature and non-linear equation;
ii) the cubic Soave-Redlich-Kwong (SKR) equation of state (2nd-3rd Eq. 5-18);
iii) the solution of the Rayleigh differential equation describing the liquid consumption law in the no-reflux simple batch distillation (4th Eq. 5-18);
iv) conservation of the residual liquid mass (5th Eq. 5-18).

Two SKR equations of state are solved for the liquid phase and the vapor phase, respectively. The Rayleigh's solutions and the equilibrium equations are solved for each of the n pure compounds comprised in the mixture at the J distillation step. The system of equations is solved for a constant pressure set at the atmospheric value, the solution array is composed of the liquid and the vapor compressibility factors (Z), the composition of the distilled vapor (y) and of the residual liquid (x), the equilibrium temperature at ambient pressure measured in the liquid phase (TK). k is the equilibrium ratio, A, B are the parameters of the SKR equation, r is the ratio between the residual liquid at the J-th step and the residual at the previous step.

$$\begin{cases} y_i - k_i(Z, T, x_i, y_i) \cdot x_{i,J} = 0 \ \forall i = 1, \dots, n \\ Z_L^3 - Z_L^2 + (A_L - B_L - B_L^2) \cdot Z_L - A_L \cdot B_L = 0 \\ Z_V^3 - Z_V^2 + (A_V - B_V - B_V^2) \cdot Z_V - A_V \cdot B_V = 0 \\ x_{i,J} - x_{i,0} \cdot \prod_{j=1}^{J} r_j^{k_i - 1} = 0 \ \forall i = 1, \dots, n \\ \sum_{i=1}^{n} x_{i,J} - 1 = 0 \end{cases} \quad (5\text{-}18)$$

In computing the RVP of the mixture, the system of equations in Eq. 5-18 is solved once for the initial mixture composition. In computing the distillation curve of the mixture, Eq. 5-18 is solved for each distillation step J. Considering the iterative nature of the problem, in order to increase the accuracy of the initial solution attempt at each distillation step, the distillation range was divided in incremental steps of 1% evaporated volume, resulting in 100 computations to cover the whole distillation range. As a consequence it is possible to assume that the equilibrium temperature of each distillation step is the one calculated for the previous step increased by 0.5 K. Eq. 5-18 was solved by means of a modified Powell's method for the multivariate root-finding problem implemented in the SciPy-Optimize open library by setting custom values for tolerance (1x10-6) and number of iterations (5000) [109]. Once the equilibrium temperature in the liquid phase is returned by the afore-mentioned model, the equilibrium temperature at the top of the distillation apparatus is calculated by means of the energy balance (Eq. 5-19) between the distilled fuel vapor, the air and the solid wall.

$$T_{H,J} = \left[\left(m_{air} c_{p,air} + m_{wall} c_{p,wall} \right) \cdot T_{H,J-1} + m_{fuel(V)} c_{p,fuel(V)} \cdot T_{K,J} \right] / \left(m_{air} c_{p,air} + m_{wall} c_{p,wall} + m_{fuel(V)} c_{p,fuel(V)} \right) \quad (5\text{-}19)$$

Fig.5-3 shows the validation of the described model by means of the comparison between the experimental and the modelled distillation curve and the corresponding temperature difference (model – experimental) at each distilled volume for two mixtures i.e., a 50%-50% pentane-heptane mixture (light mix), representative of the lighter composition of commercial gasolines (C4-C7 n-paraffins); a 80%-20% dodecane-124TMB mixture (heavy mix), representative of the heavier composition of commercial gasolines (heavy n-paraffins and aromatics). From Fig.5-3 it is visible that the temperature difference around the end-tails of the curve is not significant regardless of the volatility of the components. It must be underlined that the bell-shaped temperature difference is characterized by the maximum value around half of the distilled volume likely due to the transition from 2-component mixture to almost pure component. Furthermore, it is observed that the temperature difference is higher for the light mixture with respect to the heavy mixture.

Reid vapor pressure (RVP) i.e., the value of the absolute vapor pressure at 100°F (310 K), is a synthetic index representative of the volatility of fuel mixtures. Considering the typical injection temperature under early injection conditions in GDI engines (40-45°C), it is clear that RVP can describe the early evaporative behavior of real gasolines at the engine cold states, which are the most critical for the tailpipe emission under both WLTP and real driving conditions. As seen for the distillation temperatures, even RVP is not additive and needs a dedicated modelling. In this research project RVP is predicted by using the same pure thermodynamic model adopted for T0 adjusted to solve the classical bubble point pressure problem for non-ideal mixtures. Therefore, the same objective function shown in Eq. 5-18 has been numerically minimized, however the temperature has been set to the constant value of 310 K (100°F) instead of being the equilibrium unknown, whilst the pressure has been set as the implicit solution instead of being the environmental pressure. The initial solution guess is the same as that of the first distillation temperature except for the equilibrium unknown which is initially set as the moles weighted average among the saturation pressure values at 310 K of the pure components. It is underlined that as seen for T0, the initial values of the equilibrium ratios (k) have been estimated by means of the Wilson correlation [110], which is still reliable due to the typical range of gasoline RVP values

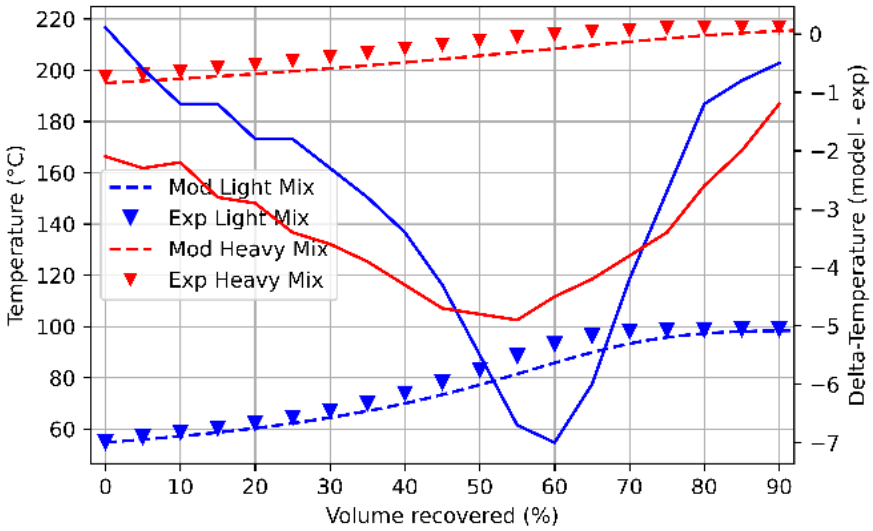

Figure 5-3: Validation of the distillation curve computation methodology

5.5 Results of the optimization process

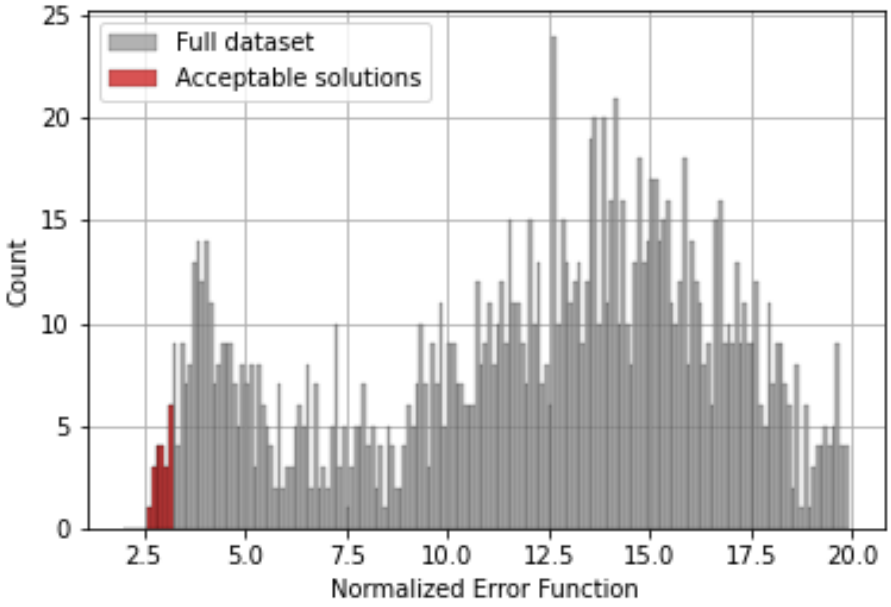

Figure 5-4: Distribution of the error function and identification of the acceptable solutions region

The optimization algorithm has been applied 3 times, for each fuel, targeting the properties collected in Table 5-2 with the weights of the target function described in the previous sections. The second use of the dataset was to identify not only the best optimized surrogate, but a family of equally acceptable alternatives, in order to identify the differences among them. To this end, it was chosen to define acceptable all surrogates with a target function within 10% of the standard deviation from the optimal value. Within this domain, all the surrogates have been compared with the values obtained with the optimization process and the most distant surrogate has been identified by means of a Similarity parameter S defined in Eq. 5-20.

$$S = \left(\frac{1}{\sum_{i=1}^{N_{properties}} w_i} \cdot \sum_{i=1}^{N_{properties}} w_i \cdot abs\left(\frac{property_i - property_{OPT}}{property_{OPT}} \right)^2 \right)^{0.5} \quad (5\text{-}20)$$

The properties used for the computation are all those reported in Table 5-2 and not considered during the optimization step, with unitary weight for mean properties, a value calculated for ignition delay times and laminar flame speed as the reciprocal of the number of points available in the dataset and a value of 0.3 for the mass fraction of the molecular groups. In the next sections, for each target fuel, 6 surrogates will be defined, in couples for each level of characterization. The surrogates defined with (*) in the images are the furthest acceptable solutions obtained by maximizing the similarity function of Eq. 5-20.

5.5.1 Shell-D surrogates

In the current section, the main differences between all the surrogates defined in the optimization and verification steps are compared in terms of elements that would impact the CFD combustion simulations. From the point of view of the composition, the levels of target for which the ALPIONA is available can satisfactorily represent the target, while the fraction of aromatics in the level 1 optimized surrogates is overestimated. In general, despite the availability of multiple molecules for each group, the complexity penalty has led most of the optimized solutions to be composed by at most 7 components, in line with the number of components usually found in literature for the most detailed fuel surrogates [91].

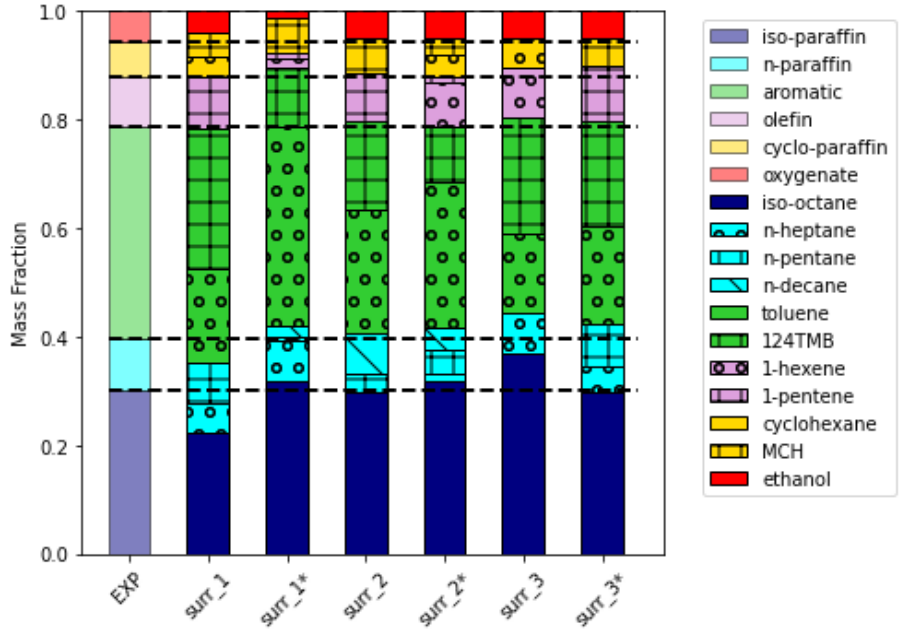

Figure 5-5: Shell-D surrogate compositions

Results of the optimization process

The ignition delay time at reference condition of φ=0.9, p=40 bar is optimized only for level 3 surrogates, whereas a very good results is also obtained for surrogates from other levels. On the other hand, the difference between two surrogates from the same level is significantly reduced by including more targets, up to an almost unique result for level 3 (where the values of ignition delay time are one of the targets) as presented in Figure 5-6. Regarding the laminar flame speed, there are variations between 1% and 3% for all surrogate levels, Figure 5-7, indicating an overall insensitivity of the burning speed with respect to the fuel surrogate formulation within the Shell-D properties range.

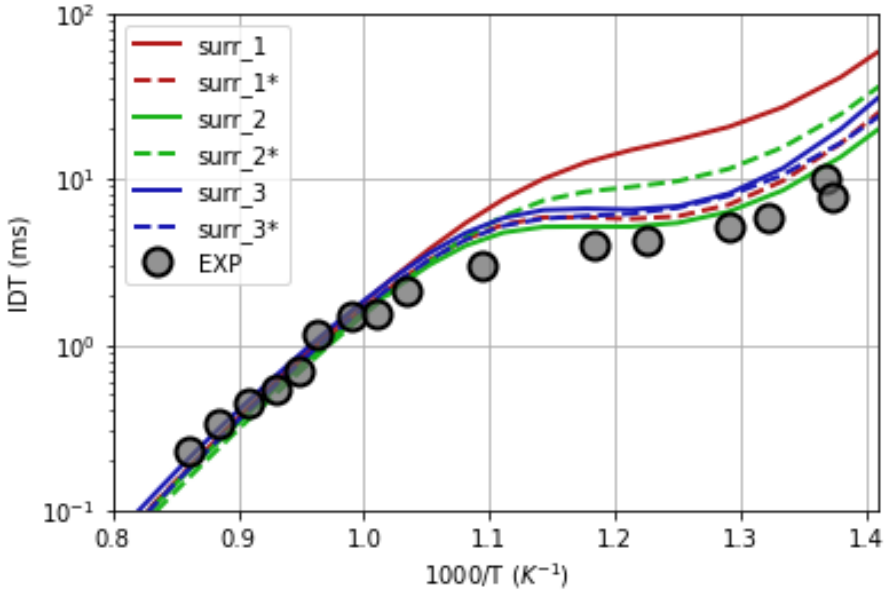

Figure 5-6: Shell-D surrogate ignition delay time at reference conditions

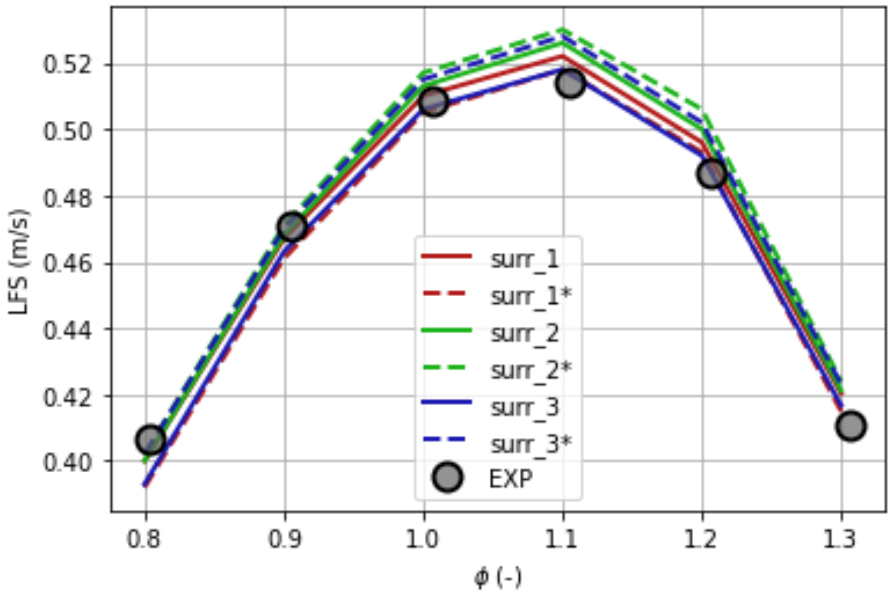

Figure 5-7: Shell-D surrogates laminar flame speed at reference conditions

5.5.2 RON96E5 surrogates

Differently from the optimization of Shell-D fuel, the optimization of the RON95E5 gasoline surrogate has not led to a significantly close definition of the composition. In particular, even for level 3, where the composition is known, the fraction of aromatics is overestimated, due to the enforcement of the other requirements.

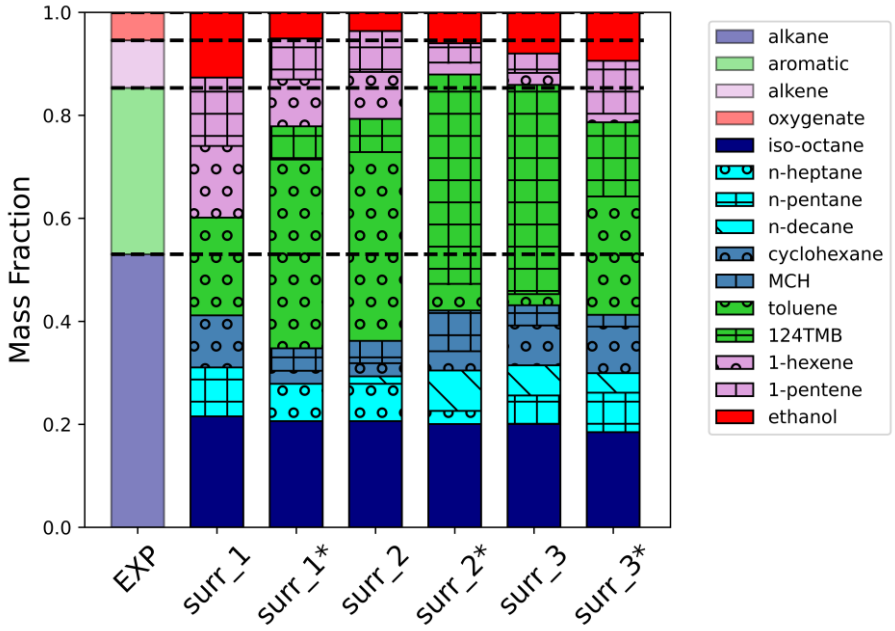

Figure 5-8: RON96-E5 surrogate compositions

The surrogates obtained from level 1 targets display significant differences in terms of ignition delay time at low and high temperature, due to the overestimation of the ethanol content which increase the ignition delay time. As far as the laminar flame speed is concerned, differently from the surrogates of Shell-D gasoline, the first level of optimization produces differences up to 10%, while the dependence on the equivalence ratio is consistent for all surrogates with the experimental data, thanks to the close stoichiometric air fuel ratio as can be seen from Figure 5-9. Level 2 and 3 surrogates present similar compositions, however with higher values of the aromatics mass fraction with respect to the experimental data. Regarding ignition delay time, reported in Figure 5-10, from level 2 there are only minor differences, while the laminar flame speed is consistent only in the case that it is used as a target.

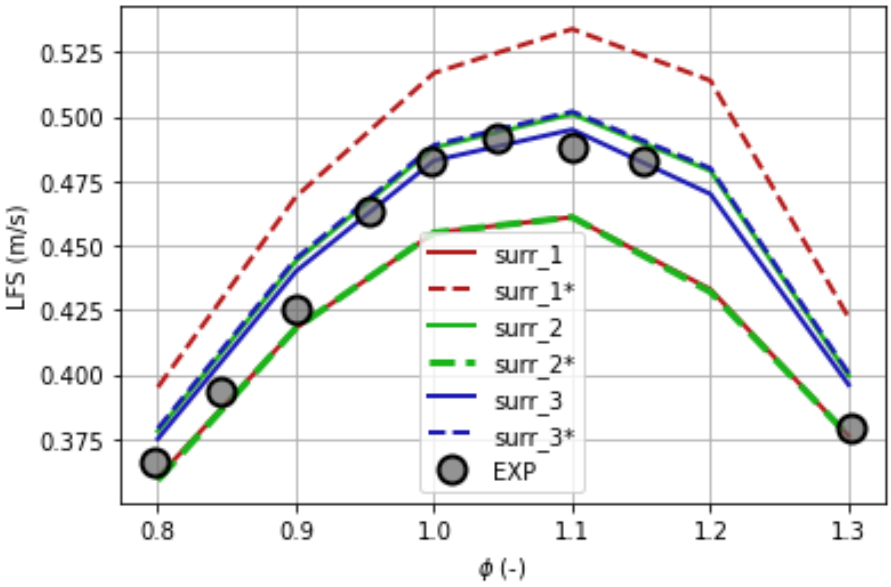

Figure 5-9: RON96-E5 surrogate laminar flame speed at reference conditions

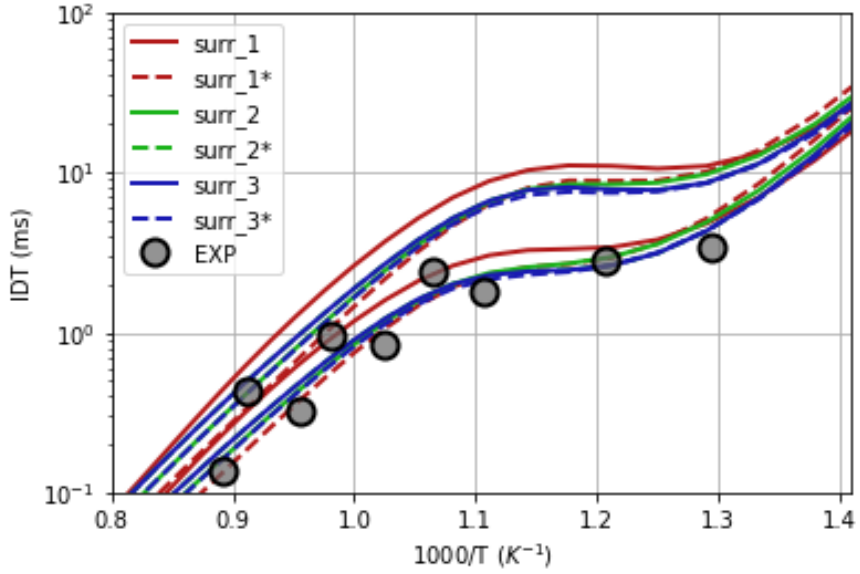

Figure 5-10: RON96-E5 surrogate ignition delay times at reference conditions

All the evaluated combinations have been stored to perform an analysis of the properties of the search domain of all the combinations with RON within 95 and 98, corresponding to acceptable values of the pump gasoline. The Pearson's correlation coefficient [111] is a measure of the ranks of each property with the others, where an absolute value of 1 represents a monotonically perfect correlation (differently from Pearson's correlation which requires linear dependence). The relationships with RON are not assessed in this report, because no significant variations in RON are collected. From the correlation matrix presented in Figure 5-11, several conclusions can be inferred:

- the aromatics mass fraction of the fuel is strongly and directly correlated to the YSI, density and the temperature of the distillation profile, with greater impact on the first half of the curve,

- the paraffinic content of the fuel is directly correlated to LHV, AFS and inversely proportional to YSI and density with high absolute values of the correlation coefficient,
- alkene, oxygenates and naphtene components are weakly correlated to several properties, which is attributed to the reduced fraction that these groups represent in the region of compatible combinations.

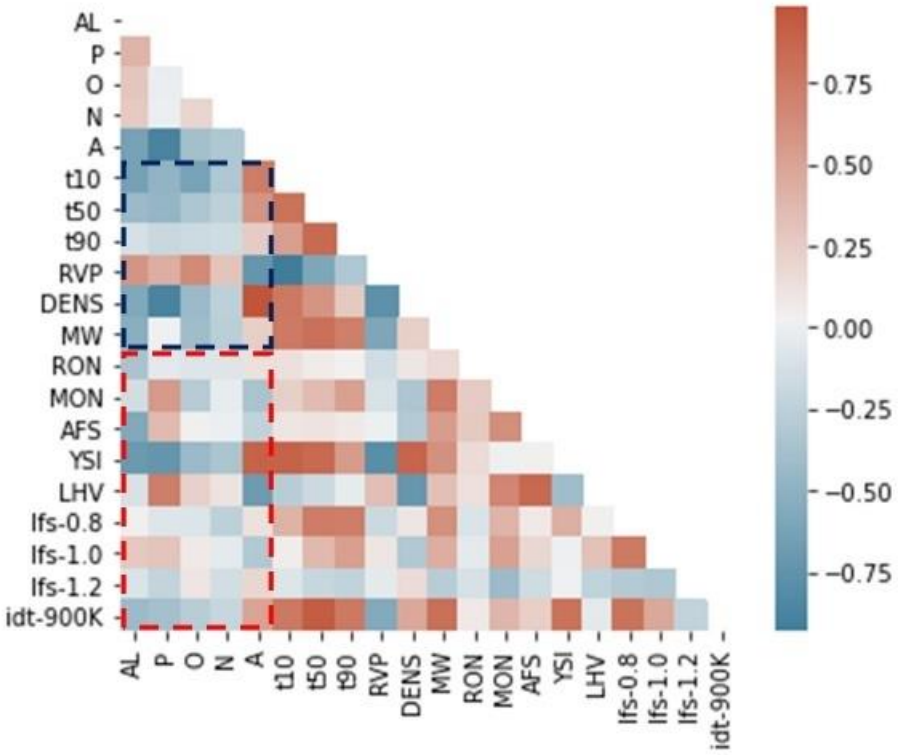

Figure 5-11: Pairwise correlation coefficients between the different properties inside the search domain

Another important conclusion that can be derived from the correlation matrix, is the inter-correlation between targets. In particular, it can be noticed that the ignition delay time at low temperature (900 K) is strongly correlated with the values of the distillation curve, molecular weight, YSI and laminar flame speed at lean to stoichiometric conditions while LHV is strongly correlated with AFS.

These correlations explain the reasons behind the similitudes of level 2 and 3 optimizations, that have the composition among the target properties.

6 CFD sensitivity analysis

The objective of the last chapter before the conclusion of the thesis, is to collect and implement all the methods proposed and developed previously with two targets:

1) Test the applicability of each methodology to a complete CFD simulation use case, starting from the fuel and engine specifications, to the fuel surrogate definition, to the accelerated generation of look-up tables for laminar flame speed and ignition delay time.
2) Assess the potential variability in the results of CFD combustion simulations performed with different surrogates of the same fuels as a function of the properties available for its definition, beyond the theoretical evaluations performed in the previous chapter.

To this aim, a specific engine test case has been designed and an accurate RANS set-up has been adopted for reducing potential sources of variability in the computation. Particular care has been given to the spray modelling, considering its relevance in the mixture distribution prediction, as well as to the gas-wall heat transfer, which is an important topic when modelling auto-ignition, considering its role in the calculation of the gas temperature near the walls. These models will be outlined in further detail in the following sections, together with the fundamental aspects of the combustion model that are affected by the fuel surrogate definition. Finally, the results of the compression phase will be reported to understand the potential areas of interest for the knock onset, which will be assessed through a sweep of spark timing for each of 4 fuel surrogates, obtained with different target properties imposed to the optimizer.

6.1 Engine specifications and 0D/1D model

The studied engine is a proof of concept digital model of a state of the art turbo charged 4 cylinder GDI engine developed in the past years at the University of Bologna for the conceptualization of novel combustion and knock control techniques [112] [113] [114]. The specification of the full engine, as well as the single cylinder simulated by means of the STAR-CD CFD solver are reported in Table 6-1. The use of a commercial software in place of open source alternatives derives from the previous background of the research group that has developed several in-house models over the years.

Table 6-1: Main properties of the GDI engine modelled

Compression Ratio (-)	9.5
Unit displacement (cm^3)	471.05
Bore (mm)	84
Stroke (mm)	85
Conrod length (mm)	165.6
Number of Valves (-)	4
S/D (-)	1.01
Intake D_v/D (-)	0.36
Exhaust D_v/D (-)	0.33

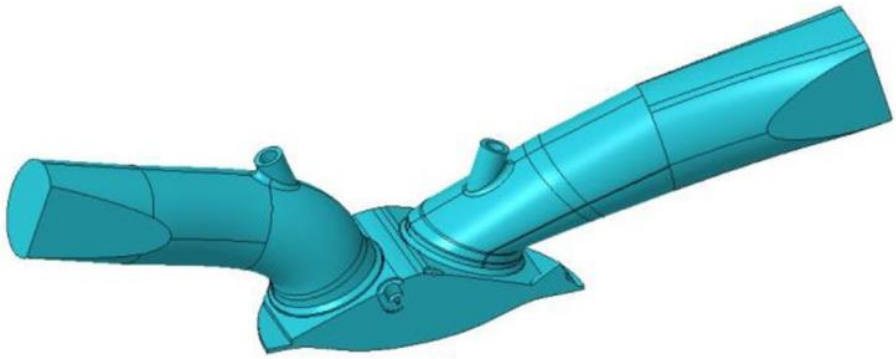

Figure 6-1: Geometry of the GDI engine model

The boundary conditions for the simulations were generated using a 0D/1D fluid dynamics model defined in OpenWAM [115] of the full engine (4 cylinders, detailed air paths and manifolds, as depicted in figure). OpenWAM is a 0D/1D Opensource code developed at CMT Motores implementing a TVD algorithm for the solution of the 1D Navier-Stokes equations to capture sharp profiles (accounting for example for shock waves inside intake/exhaust pipes). The numerical setup employed is quite standard, with application of the ideal gas hypothesis, courant number limited to 0.4 for stability, convergence reached after 20 steady cycles and heat release rate defined a-priority based on literature research.

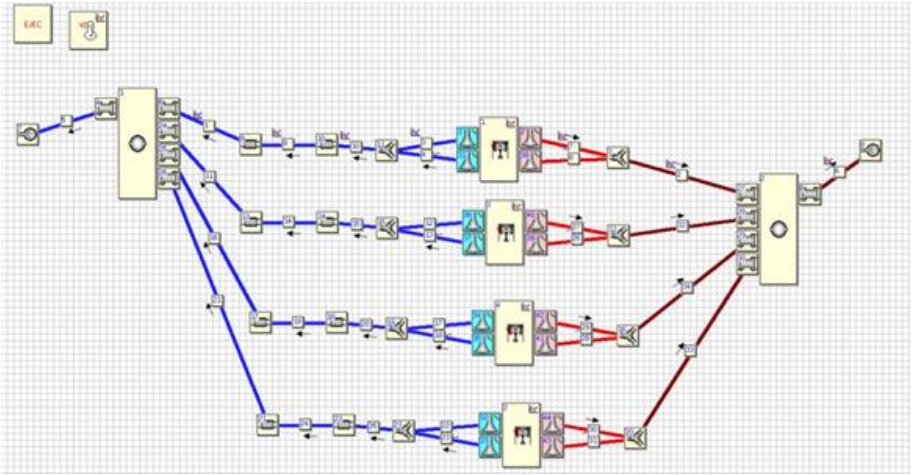

Figure 6-2: Architecture of the system simulation in OpenWAM

The best practice of the CFD software indicates that, for stability reasons, the intake pressure should be defined as a total quantity, taken at node 5 in Figure 6-2, while the exhaust pressure is imposed static taken at point 12 in figure. The temperature of the walls was set as uniformly distributed, due to the non-availability of the full engine model for a complete conjugate heat transfer loop, differentiating each surface with the values reported in the following table.

Table 6-2: Boundary conditions for the peak power operating condition

BOUNDARY NAME	BND #	T (K)
Dome	1	505
Piston	2	555

Liner	3	455
Spark Plug	4	550
Intake Valve Plate	5	400
Intake Valve Stem	6	383
Intake Manifold	7	373
Exhaust Valve Plate	8	855
Exhaust Valve Stem	9	844
Exhaust Manifold	10	423

The valve strategy was designed in order to allow the introduction of the target mass without exceeding the limits of the compressor (maximum intake pressure set to 2.7 bar), while reducing to a minimum the scavenging to avoid superposition of the open periods, leading to the choice of an inlet vale opening period between 362 and 598 CAD aTDC and an exhaust valve opening period between 136 and 376 CAD aTDC.

6.2 3-D CFD MODELLING

6.2.3 Mesh and Numerical setup

The meshing process was performed following a robust internal standard practice, backed by several previous validations. In particular a mesh base size of 0.8 was adopted for the mesh template definition, resulting in the template of Figure 6-3. The approach used by the software is extrude the base template along the cylinder axis, and perform adaptation steps on the boundary regions, where a cell size of 0.3 mm is imposed. To handle the mesh motion, the code is expected to stretch the cells up to a threshold value, after which the bottom layer of extruded cells will be removed. In order to better capture the flame kernel release, as well as to resolve in a more detailed way the fluid motion around the electrodes, a successive mesh refinement has been applied on the spark region, to reduce by 8 the volume of each computational cell, as detailed in Figure 6-3. In order to reduce the number of cells, given the symmetric nature of the combustion chamber designed, only half of the cylinder has been actually simulated, leaving to a symmetry plane the definition of the motion at the interface.

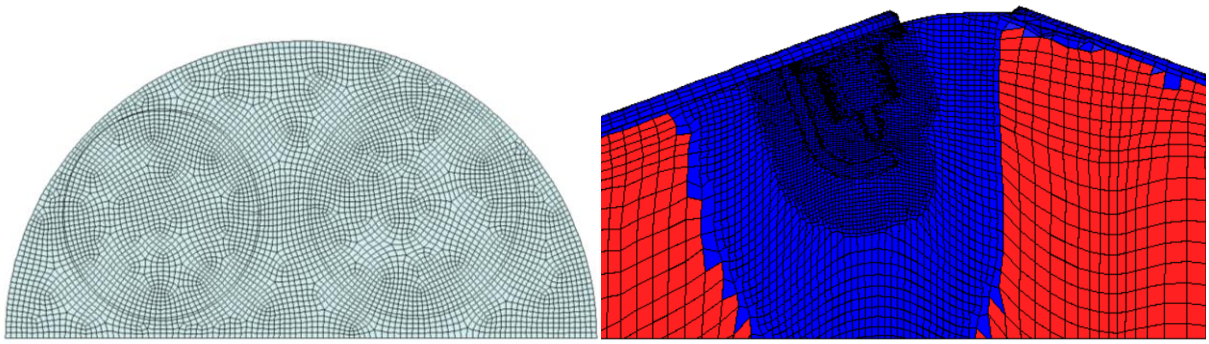

Figure 6-3: Basic mesh template and detail of the refinement levels near the spark plug

The resulting mesh is composed of 1.2 million cells at TDC, and 1.8 million at BDC, requiring on average 0.3 CPUh to update the solution by 1 CAD on a standard workstation with Intel Xeon Platinum 8000 @3 GHz.
Following the internal best practice, the turbulence model chosen was the k-epsilon RNG [116], while the chosen differencing scheme for mass, momentum and energy transport equations are all second order accurate. The transient solution is resolved with the standard PISO algorithm, adopting a convergence threshold of 1e-4 for all residual errors and using a constant time step set to 1e-6 seconds, in order to assure a limited Courant number. Considering the RANS nature of the simulation, several models have been adopted for the simulations of relevant phenomena, that will be described in further details in the next sections.

6.2.4 Multi-scale models

As outlined in the introductory chapter, the solution of the complex interconnected phenomena that occur inside the cylinder during a full engine cycle requires the coupling of the compressible flow solver with numerical models. In particular, multi-scale and multi-phase phenomena must be addressed carefully. A brief review of the models employed for the simulation where the fuel surrogate definition plays a crucial role will be therefore reported, to underline the causes of potential variations in the results. Another aspect that is essential for the correct prediction of mixture auto-ignition in internal combustion engine simulations is the gas temperature, especially near the walls, which is mostly influenced by the temperature wall function employed, which will be described in more detail.

6.2.5 The ECFM-3Z combustion model

The ECFM-3Z [117] is a turbulent combustion model based on the flamelet approach, applied to both premixed and diffusive combustion simulations, as well as knock and emissions formation. A schematic of the cell subgrid model is represented in Figure 6-4 where three zones are clearly visible:

1) Unmixed fuel zone (both in the burnt and fresh states)
2) Unmixed air zone composed of air and eventually diluents from previous cycles
3) Mixed zone, where all three main components (fuel, air and EGR) are present, and where combustion phenomena actually take place (combustion, auto-ignition and pollutant formation).

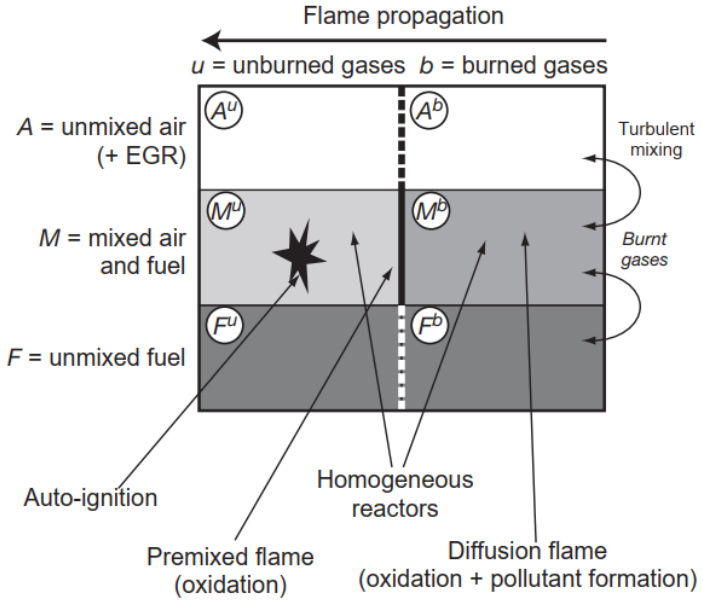

Figure 6-4: Standard visualization of the ECFM-3Z subgrid model [117]

When premixed combustion takes place, a progress variable relative to each cell indicates the fraction of mixture in the burnt and unburnt states and its evolution represents the flame front propagation. The model is completed with a set of transport equations that control the transition between the unmixed and mixed state of air and fuel as a function of turbulence, as well as their presence in the fresh or burnt phase. The standard ECFM model is sensitive to the fuel composition in 2 ways:

1) The number of carbon, hydrogen and oxygen atoms of the fuel molecule is essential to the definition of the equilibrium products and therefore the composition inside the zones.
2) The flame front propagation and interaction with turbulence are governed by chemical reactions that depend directly on the fuel composition.

Of these 2 aspects, the former is not expected to be influenced by the fuel surrogate formulation, since its definition is expected not to affect the H/C and O/C ratio of the fuel. On the other hand, different surrogates of the same fuel, as shown in the previous chapter, can display different laminar flame speed and thickness, leading to different flame propagation predictions. The flame propagation speed can be simulated with a set of hypothesis that constitute the flamelet models, which assume that the turbulent flame front can be represented as an aggregate of infinitesimal laminar flamelets. Based on this assumption, several models have been proposed and successfully validated but two of them have been mostly investigated in literature, namely the G-equation model and the ECFM-3Z model. The G-equation model [118] describes the flame front propagation inside a turbulent flow field through the solution of the transport equation of the flame front tracking scalar 'G', with a turbulent displacement speed defined s_T for which experimental as well as theoretical correlations have been proposed, all in the form of Eq. 6-1

$$\frac{s_T}{s_L} = 1 + \alpha \left(\frac{u'}{s_L}\right)^n \qquad (6\text{-}1)$$

Where α and n are model constants, u' is the turbulent velocity (RMS) and s_L the laminar flame speed function of the local thermodynamics and mixture conditions. The transport of the scalar G is then modelled as:

$$\bar{\rho}\frac{\partial G}{\partial t} + \bar{\rho}\widetilde{u_i}\frac{\partial G}{\partial x_i} = \rho_0 s_T |\nabla G| \quad (6\text{-}2)$$

Where ρ_0 is the density of the fresh mixture associated to a given level of the flame front displacement.

A second important and frequent model, valid under the flamelet assumption, is based on the description of the flame surface density (FSD [119]), that describes the available flame surface area per unit volume as a linear scaling factor for the fuel consumption rate, finally computed as in Eq. 6-3.

$$\dot{\omega}_f = \rho_u s_L \Sigma \quad (6\text{-}3)$$

The quantity FSD allows for dividing the consumption rate from the displacement affected by the turbulence interaction but requires the solution of a transport equation that has been proposed in several forms. One of the simplest forms is reported as:

$$\frac{\partial \Sigma}{\partial t} + \frac{\partial \widetilde{u_i}\Sigma}{\partial x_i} = \frac{\partial}{\partial x_i}\left(\frac{\nu_t}{\sigma_\Sigma}\frac{\partial \Sigma}{\partial x_i}\right) + \kappa_m \Sigma + \kappa_t \Sigma - D \quad (6\text{-}4)$$

Where D is the flame surface consumption term and κ_m and κ_t are the strain rate terms given by the mean flow (κ_m) and turbulence fluctuations (κ_t) for which several closure models have been proposed based on the interaction between turbulence and flame length scales. What these approaches have in common is that the chemical kinetics effects are not resolved, but are kept into account indirectly in two forms:

1) The laminar flame speed is considered not only as the base displacement term, enhanced by turbulence interaction, but also as the reference consumption rate for the fuel oxidation.
2) The laminar flame thickness is used as a factor of ‘stiffness’ of the flame elements to turbulence effects in most closure models.

In order to consider the different fuel surrogate formulation in the engine CFD simulation, a look-up table of laminar flame speed and thickness value has been generated for any definition. The table is composed by 4 independent variables (Pressure, Unburnt temperature, equivalence ratio and EGR mass fraction) and 2 target variables (laminar flame speed and thickness), which requires, for a complete multi-linear interpolation, the retrieval of 16 nodal points and relative distance to the nodes and their weighted sum.

6.2.6 The knock model

As presented in the previous chapters, mixture auto-ignition can be simulated efficiently by means of detailed chemical kinetics simulations in simplified 0D reactors. Similarly, the same computation could be performed to update the composition and eventually the enthalpy of each computational cell in the engine domain. However, a simplified approach in the framework of the flamelet combustion models can be implemented [120], to reduce the computing time with respect to the a full coupling of chemical and fluid solvers. The model is based on a 2-step procedure:

- The first part is based on the Livengood-Wu integral approach, introduced in Chapter 3, with the transported scalar Y_{IG}function defined as in Eq. 6-5:

$$\frac{d\rho Y_{IG}}{dt} + \nabla \cdot (\rho \vec{u} Y_{IG}) = \nabla \cdot (\Gamma \nabla Y_{IG}) + \frac{\rho}{\tau} \quad (6\text{-}5)$$

  Where $\frac{\rho}{\tau}$ represents the source term that increases the knock precursor variable Y_{IG} up to the point at which it reaches the threshold value chosen to start releasing the energy associated with the auto-ignition phenomenon. τ indicates the ignition delay time that can be obtained from chemical kinetics simulations, stored in efficient look-up tables and interpolated in runtime during the simulation, highly reducing the computational time required for the computation.
- The second part is related to the energy release rate to define after the threshold value is reached: in this case, a further tabulation is required, to store an indicator of the consumption speed of the reactions, which depend on both thermodynamics conditions and knock progress, as presented in Figure 6-5 for a generic gasoline fuel surrogate. The rate of reaction (indicated with AIrr) is then applied in Eq. 6-6 to compute the additional heat release rate of the standard ECFM-3Z model.

$$\dot{\omega} = LHV \cdot \rho_u \cdot Y_{Fu} \cdot s_L \cdot \Sigma + LHV \cdot \rho_u \cdot Y_{Fu} \cdot AIrr \quad (6\text{-}6)$$

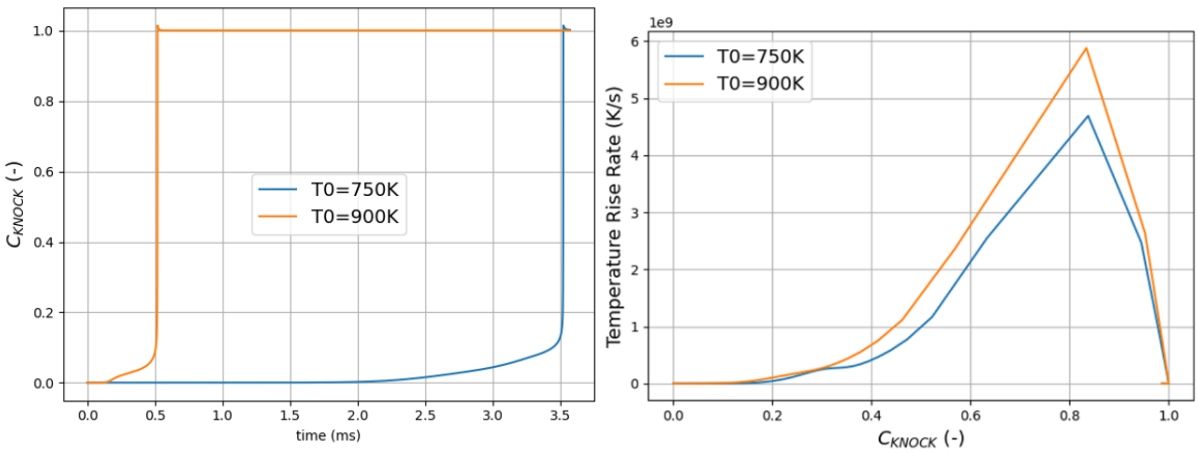

Figure 6-5: Dimensionless progress variable evolution with time and temperature rise rate profile as a function of the progress variable for 2 different temperature levels

The most relevant choices that concern this modelling approach lie in the definition of the progress variable and its threshold value for the heat release rate. While for the first definition is generally considered arbitrary between the oxygen consumption and temperature rise (in this case the temperature based progress variable was chosen), the second choice must be considered carefully, in order to avoid 2 possible drawbacks:

1) The choice of a threshold value too low, for example 0, implies that the model is always active on the domain, which would undermine the standard flame propagation model, which is expected to work properly in standard conditions. In addition, many numerical aspects might affect the output of this kind of simulations, due to the simplified nature of the basic assumptions.
2) The choice of a threshold value too high would reduce the impact of the heat release rate due to autoignition of the mixture, especially within the cool flame regime, where the temperature rise would promote faster reactions.

By analyzing several profiles, the choice of a threshold value was set to 2%, in order to account also for the cool flame regime, while maintaining a good control over the general combustion model. Considering the sensitivity of the ignition delay time, as well as the heat release rate to the fuel surrogate formulation, 2 look-up tables have been generated for each definition. The first look-up table stores the values of τ for each thermodynamics state of the mixture, and it is always interpolated in runtime, while the second table is composed of a 5th independent variable, which is the knock precursor progress variable, and stores the AIrr values (in this case computed as the temperature gradient during the ignition event), which are computed only after the threshold value is reached in a computational cell.

6.2.7 Temperature wall function

Considering the great importance that temperature distribution in the gas near the boundary regions play in the prediction of knock occurrence, particular focus has been given to the gas-wall heat exchange. In the framework of RANS simulations, the heat transfer between gas and walls is traditionally not directly resolved, which would require a refinement of the computational mesh to capture the high temperature gradients, but it is modelled by assuming a pre-defined profile (named wall function) of the temperature with respect to the non-dimensional distance from the wall, as presented in Figure 6-6.

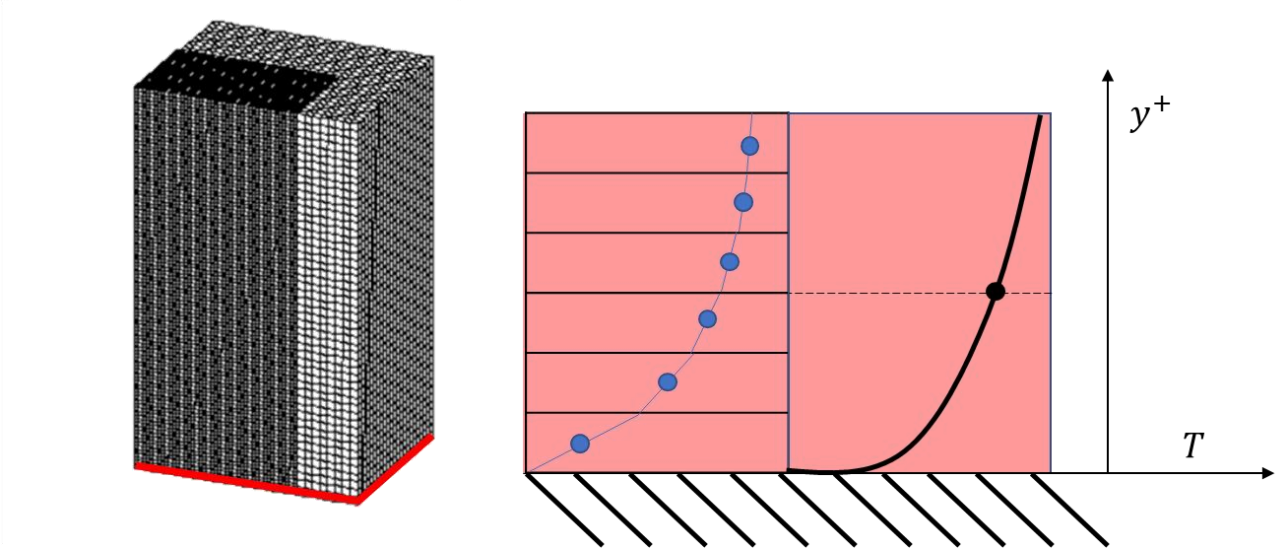

Figure 6-6: Graphical description of the difference between a low-Reynolds approach and a wall function boundary layer treatment

The black line corresponds to the assumed temperature profile with respect to the dimensionless normal distance from the wall, introduced to normalize the profile with respect to fluid dynamics conditions, as proposed by Angelberger and Poinsot et al. [63]:

$$y_w{}^+ = \frac{u_\tau}{\nu_w} y \tag{6-7}$$

Where u_τ is the friction velocity, which can be computed from turbulence properties and ν_w is the kinematic viscosity in the boundary region. The blue profile, on the other hand, would be obtained by a finer discretization of the boundary region, which should, however hold valid during all engine phases, in order to guarantee a sufficiently accurate result (values of $y_w{}^+$ below 3, which would require a geometric distance of about 5micronfor the combustion phase, when values of $u_\tau > 300$ are usually reached).

The development of a wall function for engine in-cylinder applications usually relies on several simplified assumptions, coming from the boundary layer theory of turbulent flows:

1. Derivatives parallel to the wall are neglected. The flow is mainly parallel to the wall
2. Pressure gradients is assumed to be zero in the near-wall cell, i.e., *p=p(t)* only. In the absence of spontaneous combustion, this can be true
3. The Mach number is low, so no energy associated with small-scale eddy motion is converted into thermal energy
4. Fully turbulent boundary layer
5. The flow in the boundary layer is non-reactive
6. The gas inside the cylinder is ideal
7. Radiation is neglected

In accordance with the above assumptions, the one-dimensional energy equation inside the boundary layer can be written in the following form:

$$\frac{\partial \rho T}{\partial t} = \frac{\partial}{\partial y}\left[(\sigma_{eff})\frac{\partial T}{\partial y}\right] + \frac{\partial P}{\partial t} - Q_{comb} \quad (6\text{-}8)$$

Even though the flow inside internal combustion engines is intrinsically unsteady, the dynamic of the heat transfer inside the boundary layer is described by the additional simplifying assumption of quasi-steady flow condition. As a result, both the unsteady and pressure terms are removed from the balance equation. Another assumption comes from the flame-wall interaction physics. The flame does not touch the walls since it quenches a few micrometres before because of the heat transfer. Thus, the heat source term due to combustion can be ignored, leading to the simplification of Eq. 6-8 into:

$$\frac{\partial}{\partial y}\left[(\sigma_{eff})\frac{\partial T}{\partial y}\right] = 0 \quad (6\text{-}9)$$

where the effective thermal diffusivity is the sum of a laminar and turbulent component:

$$\sigma_{eff} = \sigma_l + \sigma_t \quad (6\text{-}10)$$

where the eddy thermal diffusivity σ_t is introduced for including the effect of the additional fluctuating (turbulent) quantities. The integration of Eq. 6-8 from the wall (y=0) to the first node gives:

$$(\lambda + \lambda_t)\frac{\partial T}{\partial y} = -q_w \quad (6\text{-}11)$$

and through the definitions of laminar and turbulent Prandtl numbers along with the relation between dynamic and kinematic viscosity, the Eq. 6-11 becomes:

$$\frac{c_p}{\rho}\left(\frac{\nu}{Pr} + \frac{\nu_t}{Pr_t}\right)\frac{\partial T}{\partial y} = -q_w \quad (6\text{-}12)$$

From Eq. 6-12 after some rearrangements, the following Eq. 6-13 can be derived, showing the relationship between temperature and the dimensionless distance parameter:

$$\frac{1}{\frac{1}{Pr}+\frac{\nu^+}{Pr_t}} dy^+ = -\frac{\rho c_p u_\tau}{q_w} dT \quad (6\text{-}13)$$

Isolating the left side of Eq. 6-13 identified as the derivative of T^+ between the boundary wall and the cell centre, and considering that the density inside the boundary layer is not constant and its profile is not known a priori, an assumption of a power law, consistent for several mixture in a wide temperature range (500 K to 1500 K), is introduced as:

$$\frac{\rho_w}{\rho_c} \simeq \left(\frac{T_w}{T_c}\right)^m \quad (6\text{-}14)$$

$$m = \frac{log(\rho_w/\rho_c)}{log(T_w/T_c)} \quad (6\text{-}15)$$

Because of the assumption of ideal gas and uniform pressure inside the control volume, the equivalence: $P = \rho_c R T_c = \rho_w R T_w$ brings to write *m* as follows:

$$m = \frac{log(T_c/T_w)}{log(T_w/T_c)} = -1 \quad (6\text{-}16)$$

Integrating the right-hand side of the Eq. from the wall to the first computational cell centroid yields:

$$T^+ = -\int_{T_w}^{T_c} \frac{\rho c_p u_\tau}{q_w} dT = -\frac{\rho_w c_p u_\tau T_w}{q_w} \ln\left(\frac{T_c}{T_w}\right) \quad (6\text{-}17)$$

where T^+ is positive, because the sign convention is that, for a cold wall surrounded by hot gases, q_w is negative.

The integration of the left-hand side of Eq. 6-17 to identify the dimensionless temperature profile requires a relation for the variation of v^+/Pr_t with y^+, which is the basic differentiating point for the different wall function formulations. In fact, the definition of $T^+(y^+)$ is essential for the computation of the heat transfer to the walls, but it cannot be defined analytically. Most correlations are therefore obtained by leveraging an experimental correlation obtained by Han and Reitz [121] through a set of incompressible measurements, which has led to one of the most widely adopted wall function for internal combustion engine modelling, the one proposed by Angelberger [63], defined as:

$$T^+ = \begin{cases} Pr \cdot y^+, & y^+ \leq 13.2 \\ 2.075\, ln(y^+) + 3.9, & y^+ > 13.2 \end{cases} \quad (6\text{-}18)$$

Other notable functions were proposed by Launder and Spalding [122], Kays and Crawford [123] and Han and Reitz [121], all based on the same assumption of a linear behaviour in the viscous sub-layer (the region closer to the wall) and a logarithmic profile in the so called 'logarithmic' layer, after a threshold value of y^+.

An extensive review of the existing heat transfer models was performed by Rakopoulos et al. [124] who finally proposed a new wall heat transfer model, validated against several engine conditions, which led to a series of conclusions:

1) the isothermal wall functions are not capable of describing both high and low load conditions, as well as all the phases of the engine cycle;
2) the wall functions which included the pressure work term performed better in capturing the timing of the heat flux peak and the following reverse energy exchange in the expansion phase;
3) the model proposed by Han and Reitz proved to be the most reasonable trade-off between simplicity and accuracy, despite the unsteady pressure term is not accounted for.

Starting from the analysis of Rakoupolos, in order to deal with a more robust and accurate methodology for the assessment of the engine in-cylinder wall heat transfer, over the very large range of loads typical of both SI and CI engines currently under production, a more comprehensive fully non-isothermal wall function has been developed. In order to consider the effects of flow density variations in the near wall region, Keum et al. [125] proposed to modify the incompressible part of Han and Reitz formulation, by adding a non-dimensional temperature parameter, that includes the non-isothermal effects. A similar, more detailed approach has been used here, where no correction term is added to the general formulation, but the empirical relationship between dimensionless temperature and distance is directly addressed. In fact, the incompressible nature of the experimental measurements is in

contrast with the compressibility hypothesis of the boundary layer, leading to a misleading lack of sensitivity in the core model, leading to Eq. 6-19.

$$\frac{\nu^+}{Pr_t} = \begin{cases} \Pi \cdot (a + by^+ + c(y^+)^2), & y^+ \leq y_0^+ \\ \Pi \cdot (my^+), & y^+ > y_0^+ \end{cases} \quad (6\text{-}19)$$

where *a*, *b* and *c* are model constants as reported in [121] and the term $\Pi = \left(\frac{T_w}{T_c}\right)^{\alpha}$ is used as a correction coefficient which corresponds to 1 when the temperature value at the wall is the same of that in the bulk flow (i.e the hypothesis of incompressible boundary layer is verified, which might be considered sufficiently accurate for particularly high pressure values) and *α* is a coefficient that accounts for the boundary layer compressibility effect on the temperature profile, set to a constant value of -1.

It is now possible to integrate the left-had side of Eq. 6-17 with a fully non-isothermal approach, evaluating the temperature wall function:

$$T^+ = \begin{cases} PrA_1, & y^+ \leq y_0^+ \\ PrA_2, & y^+ > y_0^+ \end{cases} \quad (6\text{-}20)$$

where the model parameters A_1 and A_2, that summarize the integrals found in the definition of T^+ are reported in Eq. 6-21 and 6-22.

$$A1 = \left(-2\, arctan\left(\frac{0.025\sqrt{\Pi\, Pr}}{\sqrt{0.048+0.004175\Pi\, Pr}}\right) + 2\, arctan\left(\frac{(0.025+0.024\, y^+)\sqrt{\Pi\, Pr}}{\sqrt{0.048+0.004175\Pi\, Pr}}\right)\right) \quad (6\text{-}21)$$

$$A2 = \frac{1}{\Pi Pr}\left[\sqrt{\{\Pi Pr\}}\left(-2\, arctan\left(\frac{0.025\sqrt{\Pi Pr}}{\sqrt{0.048+0.004175\Pi Pr}}\right) + 2\, arctan\left(\frac{0.9226\sqrt{\Pi Pr}}{\sqrt{0.048+0.004175\Pi Pr}}\right)\right) \cdot \frac{}{\sqrt{0.048+} \quad (6\text{-}22)$$

$$-2.09776\, log(1 + 17.8286\Pi Pr) + 2.09776\, log(1 + 0.4767\Pi\, Pr\, y^+)$$

Interestingly, as can be seen from Figure 6-7, the profile of the newly defined wall function is almost coincident with the standard Angelberger definition, as long as the incompressibility hypothesis is considered satisfied $\left(\frac{T_w}{T_c} = 1\right)$. When the temperature ratio increases, thus indicating that the gradient between the boundary and gas temperature is steeper, the value of T^+ increases faster with y^+ than the classical formulation, reducing the predicted heat flux, while for the viscous sublayer only small variations are obtained.

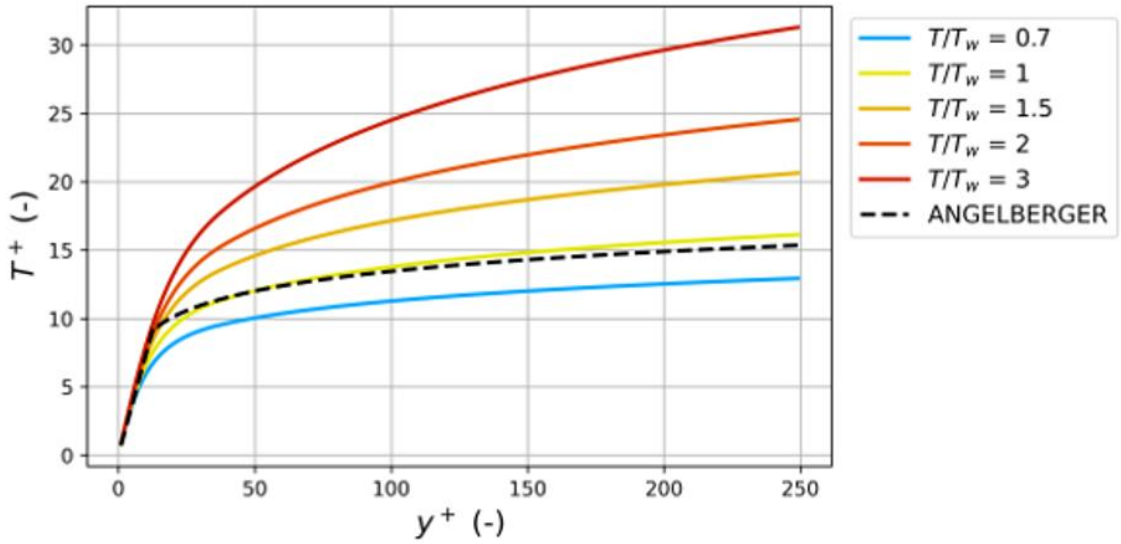

Figure 6-7: Dimensionless temperature profile as a function of the dimensionless distance from the wall

The in-cylinder heat flux across the wall can be finally calculated as:

$$-q_w = \begin{cases} \dfrac{u_\tau c_p \rho_w T_w \, ln\left(\frac{T_c}{T_w}\right)}{Pr \cdot A1}, & y^+ \leq y_0^+ \\ \dfrac{u_\tau c_p \rho_w T_w \, ln\left(\frac{T_c}{T_w}\right)}{Pr \cdot A2}, & y^+ > y_0^+ \end{cases} \tag{6-23}$$

Despite the lack of experimental data for the validation of the proposed wall function on high power density engine configurations, the trend to predict a reduced heat transfer with respect to the reference Angelberger formulation is in line with the most recent observations in academia and industry [126] which results in a higher confidence in the new method. Additionally, the dependency of the wall function formulation on the temperature ratio between cell and wall is expected to allow a higher robustness against different boundary mesh dimensions, which has as a direct consequence the variation in the average cell temperature calculation. To verify the second observation a simplified test case, usually referenced in literature has been implemented and tested: the GM pancake engine. It must be reminded that this case is of interest because of the availability of heat flux measurements on several thermocouples placed on the head of the engine, which are mandatory for the validation of the wall-functions. The main limitation, however, is that the operating condition of the engine is representative of a low load point for modern engines. The engine has been discretized with a hexahedral mesh with base size = 0.8 mm, but the boundary layer region on the head has been defined in order to generate 3 different configurations, reported in Table 6-3.

Table 6-3: Mesh size on the boundary regions for the GM-Pancake test case

	Base size (mm)	Boundary dome (mm)	Boundary piston (mm)	Boundary liner (mm)
Mesh 1	0.8	0.3	0.3	0.3
Mesh 2	0.8	0.2	0.3	0.3
Mesh 3	0.8	0.1	0.3	0.3

The simulations cover the range from 117 CA BTDC to 30 CA ATDC, and the general modelling setup can be found in [127], together with the initial and boundary conditions imposed. The combustion pressure trace is aligned with experimental data from the measurements available in Alkidas [128], as can be seen from Figure 6-8 for the reference case (Mesh 1 and Angelberger wall function).

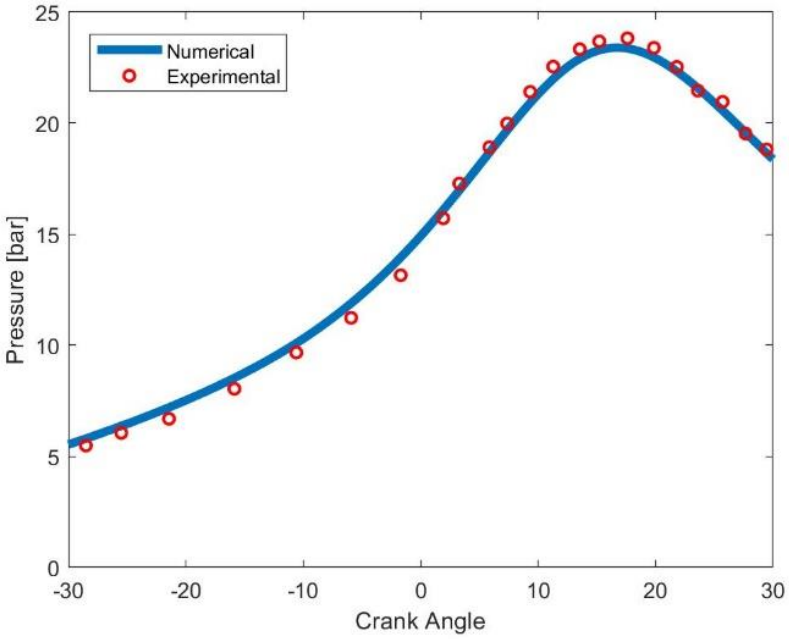

Figure 6-8: Experimental and numerical pressure trace for the reference GM-Pancake combustion phase

The results, in term of comparison with the experimental data of the Angelberger wall function are in agreement with the experimental values, whereas there is a general shift of the peak in the heat flux recorded by the thermocouple closest to the spark plug, whose punctual value is reported in Figure 6-9. Additionally, the integral heat transfer values of the Angelberger wall function case depend on the mesh size leading to variations up to 15% changing from mesh 1 to mesh 3. The proposed wall function, on the other hand, displays no significant variation in the heat flux profile, and only up to 5% variation in the average heat flux prediction.

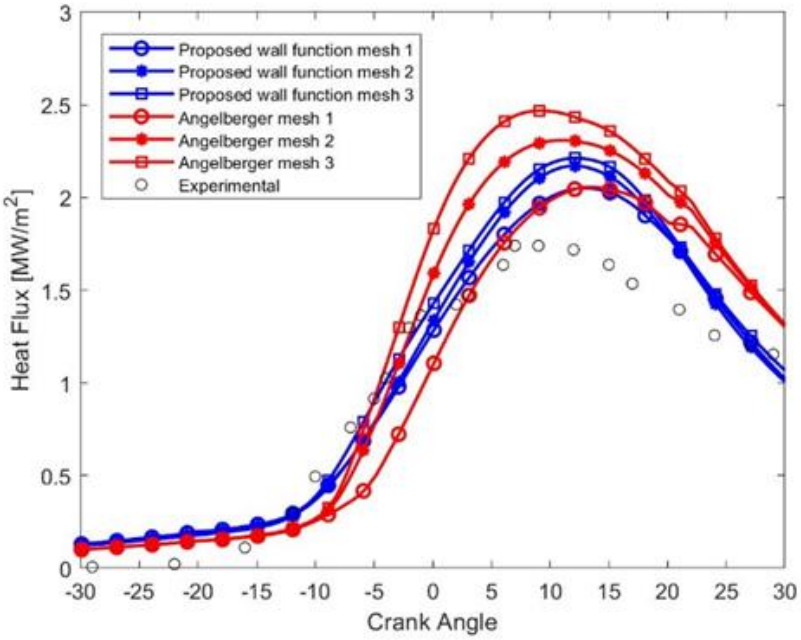

Figure 6-9: Heat flux on thermocouple 1 with different mesh sizes for 2 wall funcitons

6.2.8 Spray modelling

The solution of the coupling between liquid spray and the external gas is usually addressed by following a lagrangian approach in RANS simulations [129]. Specifically, the liquid spray injected into the computational domain (gaseous) is treated as an aggregate of droplets under the assumption that their volume is minimal with respect to the computational cell. The initial velocity and direction of each computational parcel is defined by randomly selecting an axial angle within the spray cone angle, which is an experimental property of the injector usually between 20 and 30 degrees. The velocity of the liquid spray needs to be defined a priori, considering that the flow inside the nozzle is not resolved directly. In particular, the fluid velocity can be computed with a classical Bernoulli equation, scaled by a velocity reduction constant C_v, which summarizes the friction losses inside the nozzle:

$$v_0 = C_v\sqrt{\frac{2\Delta P}{\rho}} \tag{6-24}$$

The velocity reduction coefficient for GDI injector simulations is usually considered in the range 0.8-0.95, and for the present application a value of 0.84 has been chosen, after detailed in-nozzle multiphase flow simulations performed within the research group. The second main property of the initialized droplets is the evolution of their diameter, within the injector nozzle, and immediately outside. For this application, a slightly modified version of the model proposed by Bianchi et al. has been employed, which assumes that the droplets are introduced inside the domain as a column of mono-dispersed droplets, whose diameter corresponds to the effective diameter of each hole and can be obtained by multiplying the nominal orifice diameter by a flow contraction coefficient C_C which can be obtained from the overall discharge coefficient and the velocity reduction hypothesis:

$$C_c = \frac{4Q_{static}}{\pi d_{nom}^2 C_v \sqrt{2\Delta P \rho}} \quad (6\text{-}25)$$

Where Q_{static} is the experimental effective flow rate, and d_{nom} the nominal hole diameter. After the introduction of the droplets inside the domain (which in this phase are called blobs), their size is rapidly eroded to account for turbulent and cavitating phenomena not considered in the computation of the contraction coefficient, up to the effective atomization of each blob into new droplets which represent the effective spray atomization phase, guided by the blob relative velocity, and an assumed probability distribution of the diameters of the droplets.

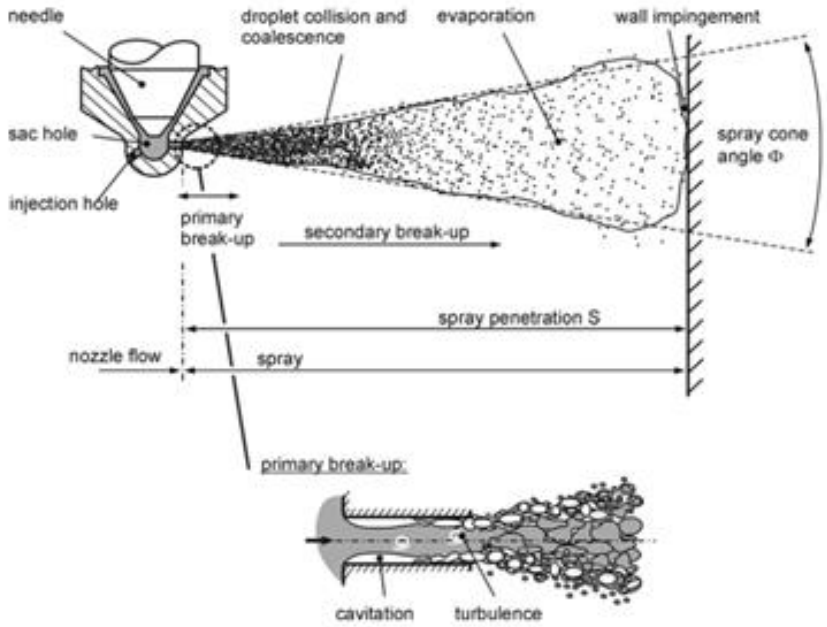

Figure 6-10: Classical representation of the liquid spray evolution inside the surrounding medium

The interaction between aerodynamics forces and the fluid surface is usually modelled by means of break-up models for the droplets, instead of directly resolved, due to the characteristic length scale of the problem. One of the most renowned models, based on the non-dimensional Weber number, is the Pilch-Erdmann model used in the current work [130]. This method is based on the calculation of a characteristic break-up time obtained with an experimental campaign by the authors and a stable diameter the droplets tend to, all functions of the Weber number, defined as the ratio between aerodynamic drag force and surface cohesion forces.

$$\frac{dD}{dt} = -\frac{(D - D_{stable})}{\tau} \quad (6\text{-}26)$$

The break-up model implemented is a customized version of the Pilch and Hermann model, which considers 5 different break-up regimes as a function of the droplet break-up number. The basic idea behind the algorithm is that, based on the Weber number, the droplet will behave differently, and its break-up will occur with different time scales, registered after a set of experimental validation studies.

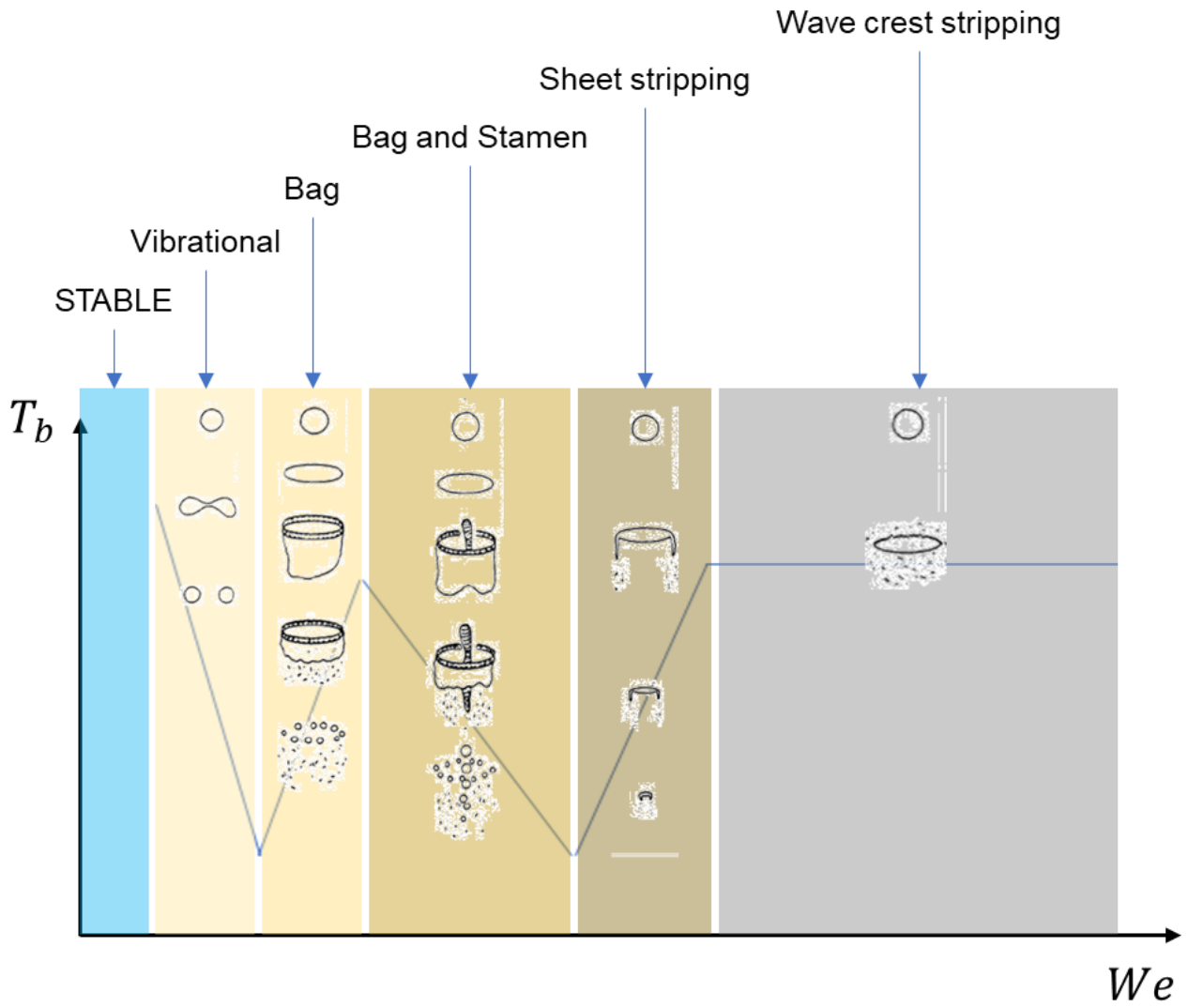

Figure 6-11: Visualization of the different break-up regime identified as a function of the Weber number

Based on the density ratio between fluid and gas, the characteristic time scale is computed from the dimensionless break-up time as:

$$\tau = T\frac{D}{V\left(\frac{\rho}{\rho_d}\right)^{0.5}} \quad (6\text{-}27)$$

Consequently, from the We_cinformation and definition to a reference value of 12, the stable diameter at which the droplet will tend is defined as:

$$D_{stable} = C \cdot We_c \frac{\sigma}{pV^2} \quad (6\text{-}28)$$

From these 2 data (stable diameter and characteristic time scale, the STAR-CD solver will define the droplet diameter at the next time step.

In order to validate the numerical setup for the spray modelling, it is frequent in literature to compare the numerical predictions with experimental data from a constant volume vessel test. To this aim, the choice of the injector for the proof of concept GDI engine was directed

towards a model with experimental characterization available in literature [131]. The injector is a prototype 5-hole with a stepped-hole design, and 125microndiameter holes. The spray pattern was not available, but it was retrieved from the available spray images for the validation study, and adapted heuristically in order to avoid phenomena of spark plug fouling while maximizing the spatial distribution of the spray for the engine simulation case.
To perform an accurate validation of the spray model and setup, a simplified constant volume domain is selected, with size 50x100x100 mm and discretized with hexahedral cells of reference size=0.8 mm, in line with the average size of the engine computational grid, except for the most external region which is composed of 1.6 mm cells to reduce the size of the problem, from 700’000 to about 500’000 cells, as in Figure 6-12.

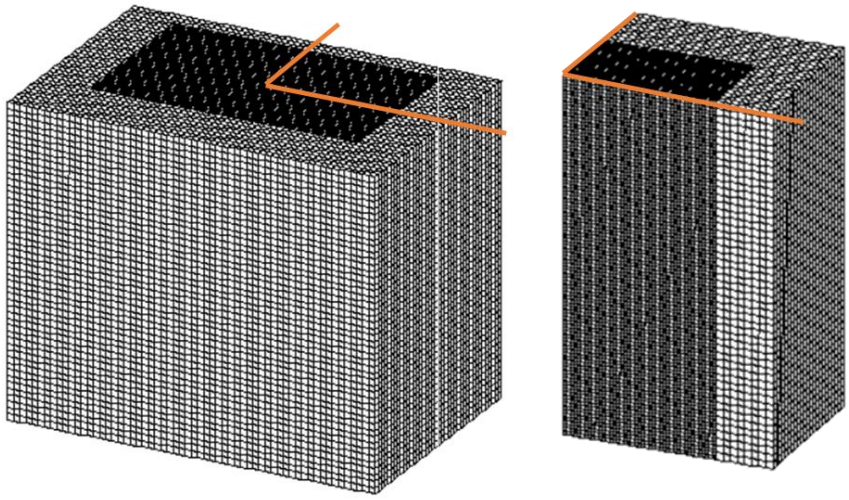

Figure 6-12: Computational domain employed for the simulation of the reference injection test case

The boundary conditions are set to adiabatic walls, which are expected to not influence the overall result in terms of spray modelling, and at the same time, provide sufficient computing stability. The time step for the transient solution is set to 1e-6 s, in line with the expected time step size employed during the engine simulation, and the solution algorithm is the classical PISO solver implemented in the commercial CFD software Star-CD, with pressure under relaxation term set to 0.1, and residual tolerances for the main turbulence, velocity, pressure and temperature fields set to 1e-5 and the discretization is second order accurate in line with the engine case setup.
The results have been compared in terms of granulometry on a 30 mm distance plane and liquid tip penetration, as well as spray plumes morphology, reported in Figure 6-13.

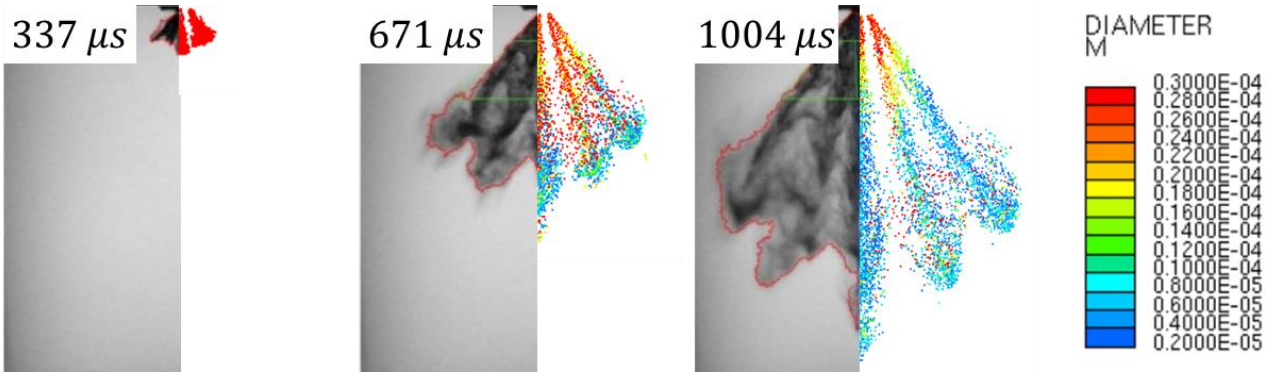

Figure 6-13: Visual validation of the spray morphology with respect to the experimental data

The liquid spray tip penetration is compared by considering the furthermost point from the injector tip reached by 99% of the total injected mass, while the granulometry is compared in terms of d10 and SMD, which are calculated as in Eq. 6-29 and 6-30 and their values are compared in Table 6-4:

$$d_{10} = \frac{1}{N}\sum_{i=1,N} d \tag{6-29}$$

$$SMD = \frac{\sum_{i=1,N} d^3}{\sum_{i=1,N} d^2} \tag{6-30}$$

Table 6-4: Comparison of experimental and numerical results of the spray simulation

	EXP	CFD
D10 (mum)	7	6.8
D32 (mum)	12	12.1
z@0.2 ms aSOI (mm)	14	14.2
z@0.4 ms aSOI (mm)	31	30.8
z@0.6 ms aSOI (mm)	42	41.5
z@0.8 ms aSOI (mm)	53	51.9

6.3 Peak Power simulation results

6.3.1 Cold Flow Analysis

Considering that the comparison between the different fuel surrogates is based on the combustion phase of the engine cycle, only a short analysis of the cold flow phase will be reported, to understand the mixture properties before the spark ignition event. A single liquid phase gasoline fuel surrogate has been employed following a multi-component formulation in order to isolate the chemical kinetics related phenomena, choosing the surrogate 3 of the Shell-D gasoline as reference.

Table 6-5: Molar composition of the different fuel surrogates for Shell-D gasoline

SHELL-D SURROGATES (Mol%)						
	Level-1	**Level-1***	**Level-2**	**Level-2***	**Level-3**	**Level-3***
Iso-octane	1.8344e-01	2.7977e-1	2.5058e-01	2.6452e-01	**3.1850e-01**	2.4129e-01
n-heptane	5.0414e-02	7.3584e-2	0.0000e+00	1.1121e-02	**7.3414e-02**	4.6099e-02
n-pentane	9.4730e-02	1.0933e-2	4.4991e-02	6.0857e-02	**0.0000e+00**	9.8087e-02
toluene	1.7762e-01	4.0259e-1	2.3716e-01	2.7552e-01	**1.5411e-01**	1.8391e-01
124TMB	1.9982e-01	8.9177e-2	1.3174e-01	8.2400e-02	**1.7658e-01**	1.4904e-01
1-hexene	0.0000e+00	2.2738e-2	0.0000e+00	9.1562e-02	**1.0788e-01**	0.0000e+00
1-pentene	1.3025e-01	1.4989e-2	1.2076e-01	1.3684e-02	**0.0000e+00**	1.3487e-01
cyclohexane	4.0385e-02	0.0000e+00	0.0000e+00	4.3356e-02	**6.1644e-02**	0.0000e+00
MCH	4.0291e-02	6.4441e-2	6.2522e-02	3.2129e-02	**0.0000e+00**	4.8649e-02
Ethanol	8.3046e-02	2.7305e-2	1.0265e-01	9.8840e-02	**1.0788e-01**	9.8052e-02
n-decane	0.0000e+00	1.4478e-2	4.9609e-02	2.6012e-02	**0.0000e+00**	0.0000e+00

The overall mixture distribution 10 degrees before top dead center, shows a sufficiently high homogenization, due to the high tumble (maximum tumble ratio = 0.95), with a leaner region on the squish plane on the exhaust side, which could indicate a potential hotter region (due to a reduced charge cooling).

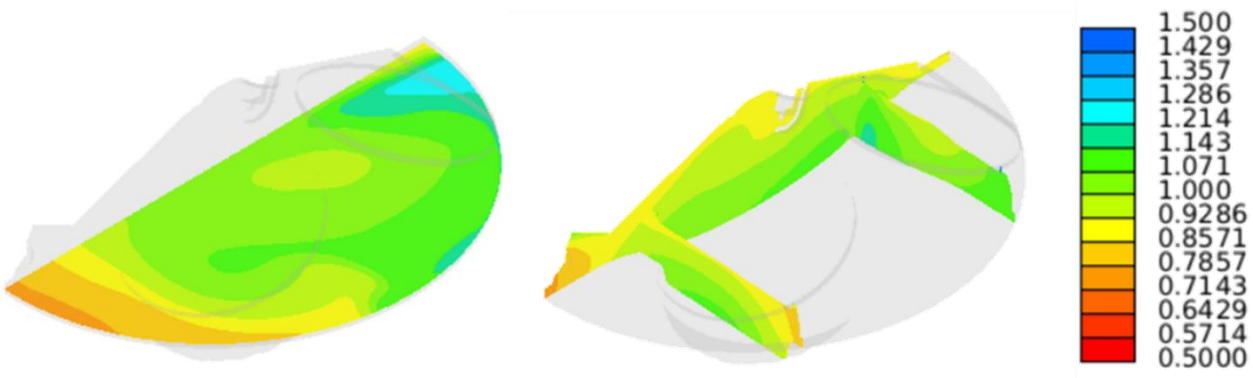

Figure 6-14: Mixture fraction distribution (λ) inside the cylinder @710 CAD

The temperature distribution inside the cylinder, presented in Figure 6-15, due to different charge cooling effects, presents a temperature variation of up to 100 K, with minimum values

near the richer region on the intake side, and a hotter region under the exhaust valves due to higher heat transfer from the valve faces and lower cooling for the liquid fuel evaporation.

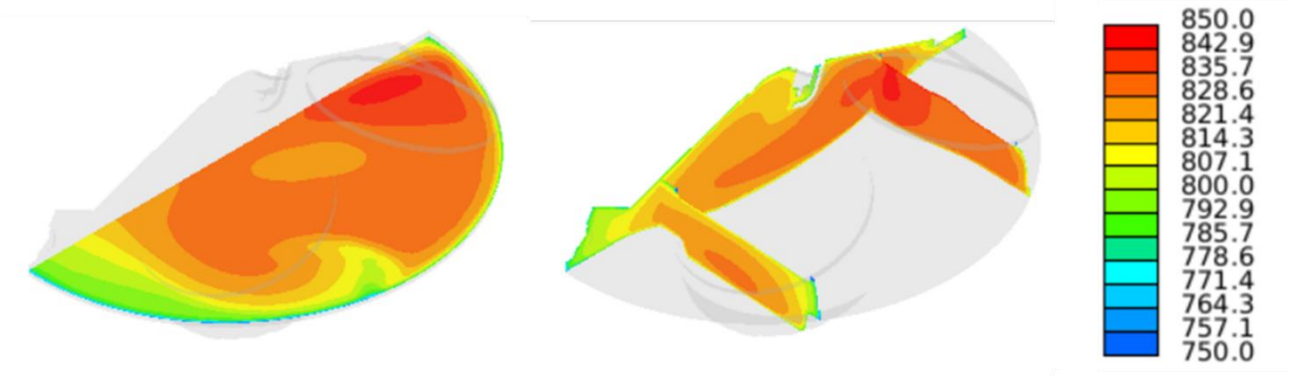

Figure 6-15: Mixture temeprature distribution inside the cylinder @710 CAD

The turbulent kinetic energy inside the cylinder is extremely dependent on the piston bowl and head shapes, that lead to the formation of a non-symmetric higher turbulence region at TDC on the intake side, up to the spark plug. The flame front propagation is expected to be driven by the turbulence distribution, therefore leading to an increase in the time available for the knock precursor to develop on the exhaust side.

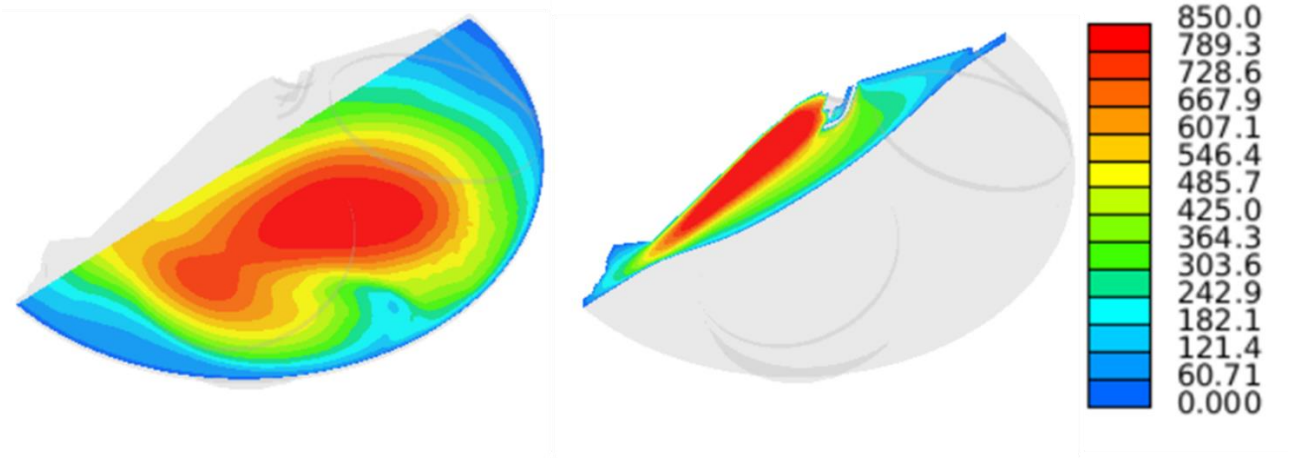

Figure 6-16: Turbulent kinetic energy distribution inside the cylinder @710 CAD

6.3.2 Sa sweep analysis

As expected from the preliminary analysis, the combustion progress follows the turbulence distribution propagating towards the intake side before expanding towards the exhaust region. In Figure 6-17 the combustion progress variable is presented for the reference cycle, at 736 CAD, which corresponds to 50% of the total mass fraction burnt for the cases with SA=721 CAD.

Peak Power simulation results

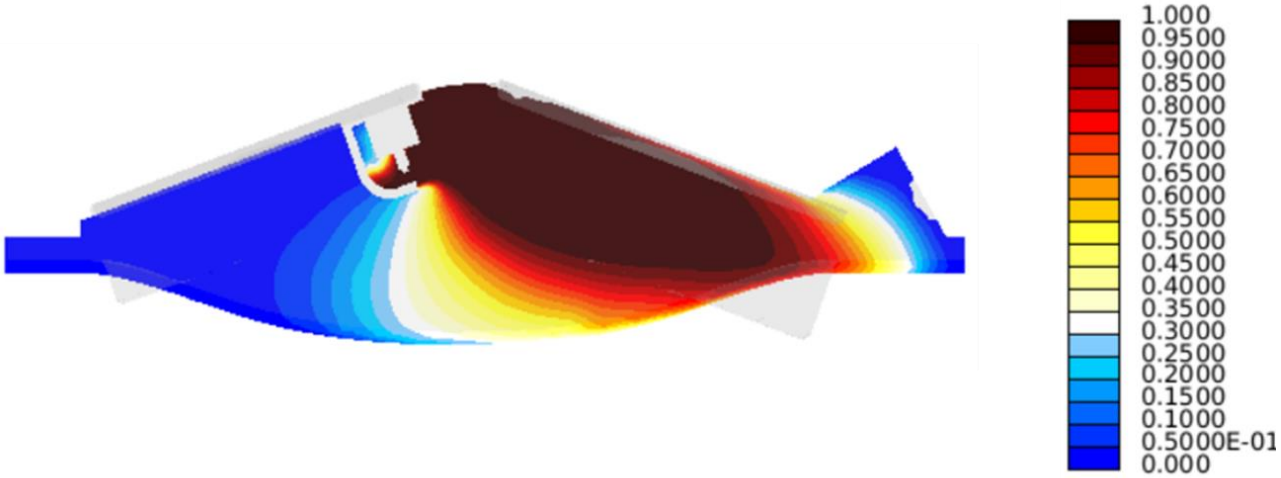

Figure 6-17: Combustion progress variable for the reference cycle (not-knocking) @736 CAD

In order to provide a pressure trace comparable with experimental results, leading to the possible definition of knock indicators, such as the MAPO index, calculated as in Eq. 6-31, the requirement to store punctual values is obtained by placing virtual sensors in several positions of the engine, near the boundaries, as reported in Figure 6-18.

$$MAPO = \max(|P|) \tag{6-31}$$

Where P is the filtered pressure trace, usually obtained with a high-pass filter limited at 6KHz.The presented results were obtained from the same initial field at 700 CAD, changing only the laminar flame speed and ignition delay time look-up tables interpolated during the simulation. Considering the time requirements for the generation of entire databases, even applying the presented methodologies, it was chosen to verify the robustness of the surrogate definitions only with respect to the most detailed optimized surrogate (SURR-3).

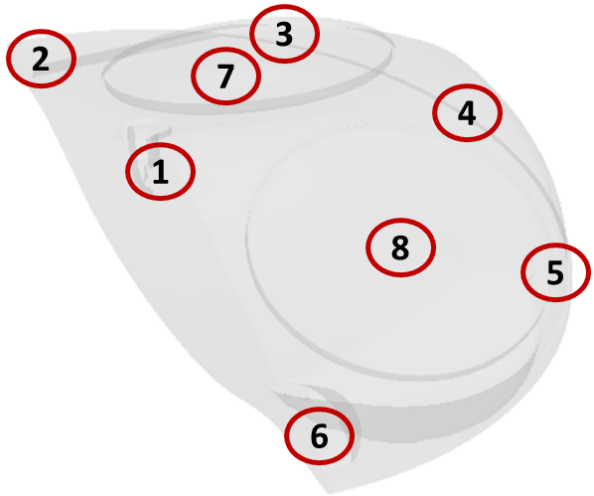

Figure 6-18: Distributions of the punctual virtual sensors placed on the engine

The results of a variation in the spark timing, with consequent modification in the combustion phasing of the engine are reported in term of punctual pressure traces obtained at sensor

#2, which is, from the previous observations, the one which could be more subject to knocking phenomena.

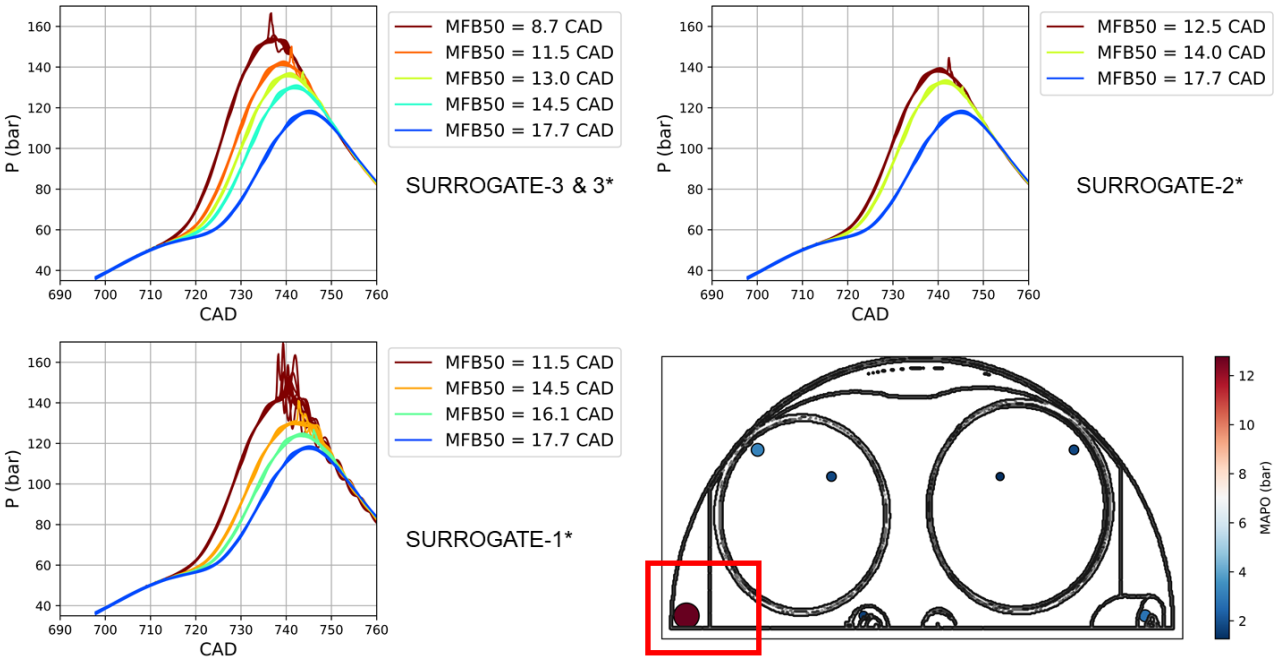

Figure 6-19: Pressure trace of the SA sweeps performed with the different surrogates and identification of the most relevant location for knock insurgence

The effect of the fuel surrogate definition does not display any significant variation in KLSA identification when the full experimental validation data are employed (SURR_3 and SURR-3*) but it clearly shows diminishing robustness when reducing the number of common properties up to the minimal set of information of SURR_1*, which underestimate the potential optimal MFB50 value by over 3 CAD, by considering the BOSCH limit MAPO rule (limit=RPM/1000) [132]. On the other hand, the results obtained with SURR-2* highlight the potential of the use of fuel surrogates even without the presence of experimental kinetics data, as long as a sufficient molecular characterization is available.

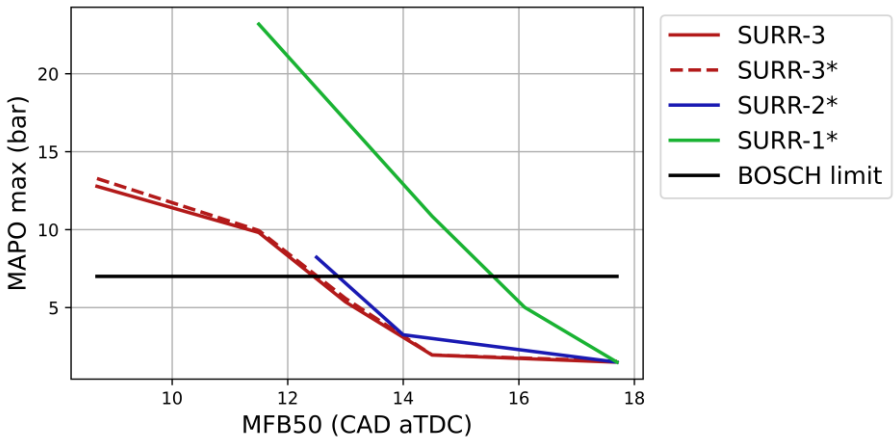

Figure 6-20: Identification of the Knock-limited SA by means of the BOSCH limit method for the different gasoline fuel surrogates

Conclusions and improvements

The automotive industry, in the last decades, has seen an ever-increasing level of complexity to meet customer demand as well as facing regulatory limitations mainly concerned with the internal combustion engine functioning. The use of numerical simulation tools, in particular fluid dynamics solvers have played a central role in the understanding and optimization of the latest generation of powertrains, from several points of view, from duct shape optimization to injection pattern definition, to combustion and knock prediction. Even with the increasing electrification share, the use of numerical methods is regarded as the fastest approach to the combined optimization of engines and fuels, which could help reaching the emission targets set by the international community. The use of synthetic fuels or the increased fraction of bio-derived components inside the more traditional pump gasoline can, indeed represent a viable alternative in the mid-term, when most light-duty transportation vehicles are still powered by a combustion engine.
The level of detail that industry-standard simulation models reach needs, however, to be sensitive enough to potential variations in the fuel definitions, with an unprecedented level of accuracy and adaptability. In this context, the present work is intended to contribute to the development of new optimized fuels and engines by including fuel-dependent chemical kinetics properties in the simulation of industry-grade engine combustion simulations. To this aim, the use of machine learning techniques has played a crucial role, thanks to its excellent approximation properties. The role of artificial intelligence algorithm is, on the other hand, gaining more and more attention from the scientific community, highlighting the potentials of the introduction of data-driven algorithms into more traditional scientific domains. The main outcomes of the present work are:

- An analysis on the potential methods to reduce the time required for the generation of laminar flame speed databases for new fuels. These include the reduction of the chemical kinetics scheme, as well as a set of data-driven methodologies to overcome the accuracy reduction induced by the reduced mechanism.
- A proposed method for the rapid inclusion of water vapour into the laminar flame speed databases as an independent variable, by increasing the simulation requirements by less than 5% with a minimum error.
- A method for the direct prediction of ignition delay time for any engine relevant thermodynamics condition and gasoline fuel surrogate definition from a palette of 11 validated molecules.
- An optimized workflow for the definition of gasoline surrogates tested against 2 target reference fuels demonstrating the potential of the technique, as well as its direct relation with the number of target properties experimentally available. This aspect is of particular interest for the applicability of the entire simulation workflow, providing confidence levels for the predicted performance metrics with new fuels, as a function of the quantity of experimental data available.
- The same conclusions have been drawn from the simulation of a proof-of-concept GDI engine at peak power conditions, to which a set of advanced industry-standard modelling approaches, expected to be sensitive to the fuel surrogate formulation has been applied. The conclusions drawn from the combustion and knock analysis

employing different surrogates representing the same target fuel have confirmed that the availability of data regarding the fuel main molecular composition (AL-P-I-O-N-A) together with the more classical average properties (RON/MON, AFS, LHV) can substitute the need for more specialized experimental data about the reactivity of the fuel, which are expensive and time-consuming properties to be identified for new fuel formulations.

A natural continuation of the current work is related to the inclusion in the analysis of the effect of the mixture definition in terms of pollutant formation emissions, and the validation of the results against a set of experimental engine data running with different fuels, for which the surrogate should be defined. Besides traditional SI engines, moreover, the effect of fuel surrogate definition on different combustion modes, such as HCCI should be investigated, considering the high relevance that fuel reactivity properties play in the performance prediction of such concepts.

References

[1] V. Macián, J. Monsalve-Serrano, D. Villalta, and Á. Fogué-Robles, 'Extending the potential of the dual-mode dual-fuel combustion towards the prospective EURO VII emissions limits using gasoline and OMEx', *Energy Convers. Manag.*, vol. 233, p. 113927, Apr. 2021, doi: 10.1016/j.enconman.2021.113927.

[2] T. Hübner and S. von Roon, 'Synthetic fuels in the German industry sector depending on climate protection level', *Smart Energy*, vol. 3, p. 100042, Aug. 2021, doi: 10.1016/j.segy.2021.100042.

[3] A. C. Kulzer *et al.*, 'Sustainable Mobility Using Fuels with Pathways to Low Emissions', Apr. 2020, pp. 2020-01–0345. doi: 10.4271/2020-01-0345.

[4] J. Villforth *et al.*, 'Methods to Investigate the Importance of eFuel Properties for Enhanced Emission and Mixture Formation', Sep. 2021, pp. 2021-24–0017. doi: 10.4271/2021-24-0017.

[5] E. Rossi *et al.*, 'Experimental and Numerical Investigation for Improved Mixture Formation of an eFuel Compared to Standard Gasoline', Sep. 2021, pp. 2021-24–0019. doi: 10.4271/2021-24-0019.

[6] J. P. Szybist *et al.*, 'What fuel properties enable higher thermal efficiency in spark-ignited engines?', *Prog. Energy Combust. Sci.*, vol. 82, p. 100876, Jan. 2021, doi: 10.1016/j.pecs.2020.100876.

[7] C. K. Westbrook and W. J. Pitz, 'A Comprehensive Chemical Kinetic Reaction Mechanism for Oxidation and Pyrolysis of Propane and Propene', *Combust. Sci. Technol.*, vol. 37, no. 3–4, pp. 117–152, May 1984, doi: 10.1080/00102208408923750.

[8] H. J. Curran, P. Gaffuri, W. J. Pitz, and C. K. Westbrook, 'A Comprehensive Modeling Study of n-Heptane Oxidation', *Combust. Flame*, vol. 114, no. 1–2, pp. 149–177, Jul. 1998, doi: 10.1016/S0010-2180(97)00282-4.

[9] S. Dong *et al.*, 'A new detailed kinetic model for surrogate fuels: C3MechV3.3', *Appl. Energy Combust. Sci.*, vol. 9, p. 100043, Mar. 2022, doi: 10.1016/j.jaecs.2021.100043.

[10] S. S. Nagaraja *et al.*, 'A hierarchical single-pulse shock tube pyrolysis study of C2–C6 1-alkenes', *Combust. Flame*, vol. 219, pp. 456–466, Sep. 2020, doi: 10.1016/j.combustflame.2020.06.021.

[11] S. Cheng *et al.*, 'Autoignition and preliminary heat release of gasoline surrogates and their blends with ethanol at engine-relevant conditions: Experiments and comprehensive kinetic modeling', *Combust. Flame*, vol. 228, pp. 57–77, Jun. 2021, doi: 10.1016/j.combustflame.2021.01.033.

[12] E. Ranzi, A. Frassoldati, A. Stagni, M. Pelucchi, A. Cuoci, and T. Faravelli, 'Reduced Kinetic Schemes of Complex Reaction Systems: Fossil and Biomass-Derived Transportation Fuels: REDUCED KINETIC SCHEMES OF COMPLEX REACTION SYSTEMS', *Int. J. Chem. Kinet.*, vol. 46, no. 9, pp. 512–542, Sep. 2014, doi: 10.1002/kin.20867.

[13] C. Pera and V. Knop, 'Methodology to define gasoline surrogates dedicated to auto-ignition in engines', *Fuel*, vol. 96, pp. 59–69, Jun. 2012, doi: 10.1016/j.fuel.2012.01.008.

[14] 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*. Zenodo, 2018. doi: 10.5281/ZENODO.1174508.

[15] F. Maleki, K. Ovens, K. Najafian, B. Forghani, C. Reinhold, and R. Forghani, 'Overview of Machine Learning Part 1', *Neuroimaging Clin. N. Am.*, vol. 30, no. 4, pp. e17–e32, Nov. 2020, doi: 10.1016/j.nic.2020.08.007.

[16] S. Ghafari, M. Ghobadi Tarnik, and H. Sadoghi Yazdi, 'Robustness of convolutional neural network models in hyperspectral noisy datasets with loss functions', *Comput. Electr. Eng.*, vol. 90, p. 107009, Mar. 2021, doi: 10.1016/j.compeleceng.2021.107009.

[17] M. Shahhosseini, G. Hu, and H. Pham, 'Optimizing ensemble weights and hyperparameters of machine learning models for regression problems', *Mach. Learn. Appl.*, p. 100251, Jan. 2022, doi: 10.1016/j.mlwa.2022.100251.

[18] J. Snoek, H. Larochelle, and R. P. Adams, 'Practical Bayesian Optimization of Machine Learning Algorithms', in *Advances in Neural Information Processing Systems*, 2012, vol. 25. Accessed: Jan. 31, 2022. [Online]. Available: https://papers.nips.cc/paper/2012/hash/05311655a15b75fab86956663e1819cd-Abstract.html

[19] S.-J. Kim, K. Koh, M. Lustig, S. Boyd, and D. Gorinevsky, 'An Interior-Point Method for Large-Scale -Regularized Least Squares', *IEEE J. Sel. Top. Signal Process.*, vol. 1, no. 4, pp. 606–617, Dec. 2007, doi: 10.1109/JSTSP.2007.910971.

References

[20] J. C. Platt, 'Probabilistic Outputs for Support Vector Machines and Comparisons to Regularized Likelihood Methods', in *Advances in Large Margin Classifiers*, 1999, pp. 61–74.

[21] B. Wang, L. Yan, X. Duan, T. Yu, and H. Zhang, 'An integrated surrogate model constructing method: Annealing combinable Gaussian process', *Inf. Sci.*, vol. 591, pp. 176–194, Apr. 2022, doi: 10.1016/j.ins.2022.01.021.

[22] L. Breiman, J. H. Friedman, R. A. Olshen, and C. J. Stone, *Classification And Regression Trees*, 1st ed. Routledge, 2017. doi: 10.1201/9781315139470.

[23] L. Breiman, 'Bagging predictors', *Mach. Learn.*, vol. 24, no. 2, pp. 123–140, Aug. 1996, doi: 10.1007/BF00058655.

[24] P. Geurts, D. Ernst, and L. Wehenkel, 'Extremely randomized trees', *Mach. Learn.*, vol. 63, no. 1, pp. 3–42, Apr. 2006, doi: 10.1007/s10994-006-6226-1.

[25] G. Ridgeway, 'Generalized Boosted Models: A Guide to the GBM Package', *Compute*, vol. 1, pp. 1–12, Jan. 2005.

[26] T. Chen and C. Guestrin, 'XGBoost: A Scalable Tree Boosting System', in *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, San Francisco California USA, Aug. 2016, pp. 785–794. doi: 10.1145/2939672.2939785.

[27] Y. Freund and R. E. Schapire, 'A Decision-Theoretic Generalization of On-Line Learning and an Application to Boosting', *J. Comput. Syst. Sci.*, vol. 55, no. 1, pp. 119–139, Aug. 1997, doi: 10.1006/jcss.1997.1504.

[28] Y. A. LeCun, L. Bottou, G. B. Orr, and K.-R. Müller, 'Efficient BackProp', in *Neural Networks: Tricks of the Trade*, vol. 7700, G. Montavon, G. B. Orr, and K.-R. Müller, Eds. Berlin, Heidelberg: Springer Berlin Heidelberg, 2012, pp. 9–48. doi: 10.1007/978-3-642-35289-8_3.

[29] D. Vuilleumier and M. Sjöberg, 'Significance of RON, MON, and LTHR for Knock Limits of Compositionally Dissimilar Gasoline Fuels in a DISI Engine', *SAE Int. J. Engines*, vol. 10, no. 3, pp. 938–950, Mar. 2017, doi: 10.4271/2017-01-0662.

[30] J. Cho and H. H. Song, 'Understanding the Effect of Inhomogeneous Mixing on Knocking Characteristics of Iso-Octane by Using Rapid Compression Machine', *SAE Int. J. Engines*, vol. 11, no. 6, pp. 769–781, Apr. 2018, doi: 10.4271/2018-01-0212.

[31] R. A. Messerly *et al.*, 'Understanding how chemical structure affects ignition-delay-time φ -sensitivity', *Combust. Flame*, vol. 225, pp. 377–387, Mar. 2021, doi: 10.1016/j.combustflame.2020.11.004.

[32] E. L. Petersen *et al.*, 'Discrepancies between shock tube and rapid compression machine ignition at low temperatures and high pressures', in *Shock Waves*, K. Hannemann and F. Seiler, Eds. Berlin, Heidelberg: Springer Berlin Heidelberg, 2009, pp. 739–744. doi: 10.1007/978-3-540-85168-4_119.

[33] R. Fang *et al.*, *Autoignition experiments and kinetic modeling of selected highly-branched C8–C16 iso-alkanes for surrogate fuel applications*. 2019.

[34] J. Melguizo-Gavilanes and L. Bauwens, 'On the Validity of the Constant Volume Assumption in Shock Tube Experiments', in *28th International Symposium on Shock Waves*, K. Kontis, Ed. Berlin, Heidelberg: Springer Berlin Heidelberg, 2012, pp. 179–184. doi: 10.1007/978-3-642-25688-2_27.

[35] L. R. Cancino *et al.*, 'A six-compound, high performance gasoline surrogate for internal combustion engines: Experimental and numerical study of autoignition using high-pressure shock tubes', *Fuel*, vol. 261, p. 116439, Feb. 2020, doi: 10.1016/j.fuel.2019.116439.

[36] K. Fieweger, R. Blumenthal, and G. Adomeit, 'Self-ignition of S.I. engine model fuels: A shock tube investigation at high pressure', *Combust. Flame*, vol. 109, no. 4, pp. 599–619, Jun. 1997, doi: 10.1016/S0010-2180(97)00049-7.

[37] M. Hartmann, I. Gushterova, M. Fikri, C. Schulz, R. Schießl, and U. Maas, 'Auto-ignition of toluene-doped n-heptane and iso-octane/air mixtures: High-pressure shock-tube experiments and kinetics modeling', *Combust. Flame*, vol. 158, no. 1, pp. 172–178, Jan. 2011, doi: 10.1016/j.combustflame.2010.08.005.

[38] J. C. G. Andrae, P. Björnbom, R. F. Cracknell, and G. T. Kalghatgi, 'Autoignition of toluene reference fuels at high pressures modeled with detailed chemical kinetics', *Combust. Flame*, vol. 149, no. 1–2, pp. 2–24, Apr. 2007, doi: 10.1016/j.combustflame.2006.12.014.

[39] A. F. Khan, P. J. Roberts, and A. A. Burluka, 'Modelling of Self-Ignition in Spark-Ignition Engine Using Reduced Chemical Kinetics for Gasoline Surrogates', *Fluids*, vol. 4, no. 3, p. 157, Aug. 2019, doi: 10.3390/fluids4030157.

References

[40] S. Roy and O. Askari, 'A New Detailed Ethanol Kinetic Mechanism at Engine-Relevant Conditions', *Energy Fuels*, vol. 34, no. 3, pp. 3691–3708, Mar. 2020, doi: 10.1021/acs.energyfuels.9b03314.

[41] S. S. Vasu, D. F. Davidson, Z. Hong, and R. K. Hanson, 'Shock Tube Study of Methylcyclohexane Ignition over a Wide Range of Pressure and Temperature', *Energy Fuels*, vol. 23, no. 1, pp. 175–185, Jan. 2009, doi: 10.1021/ef800694g.

[42] J. C. G. Andrae, 'Kinetic Modeling of the Influence of Cyclohexane on the Homogeneous Ignition of a Gasoline Surrogate Fuel', *Energy Fuels*, vol. 32, no. 3, pp. 3975–3984, Mar. 2018, doi: 10.1021/acs.energyfuels.7b04023.

[43] S. Vranckx, C. Lee, H. K. Chakravarty, and R. X. Fernandes, 'A rapid compression machine study of the low temperature combustion of cyclohexane at elevated pressures', *Proc. Combust. Inst.*, vol. 34, no. 1, pp. 377–384, Jan. 2013, doi: 10.1016/j.proci.2012.06.071.

[44] S. Dong *et al.*, 'A comprehensive experimental and kinetic modeling study of 1- and 2-pentene', *Combust. Flame*, vol. 223, pp. 166–180, Jan. 2021, doi: 10.1016/j.combustflame.2020.09.012.

[45] J. Bugler *et al.*, 'An ignition delay time and chemical kinetic modeling study of the pentane isomers', *Combust. Flame*, vol. 163, pp. 138–156, Jan. 2016, doi: 10.1016/j.combustflame.2015.09.014.

[46] Y.-X. Liu and Z.-Y. Tian, 'Oxidation chemistry of four C9H12 isomeric transportation fuels: Experimental and modeling studies', *Combust. Flame*, vol. 205, pp. 165–179, Jul. 2019, doi: 10.1016/j.combustflame.2019.04.006.

[47] K. Kumar, G. Mittal, and C.-J. Sung, 'Autoignition of n-decane under elevated pressure and low-to-intermediate temperature conditions', *Combust. Flame*, vol. 156, no. 6, pp. 1278–1288, Jun. 2009, doi: 10.1016/j.combustflame.2009.01.009.

[48] M. Skjøth-Rasmussen, M. Braun-Unkhoff, C. Naumann, and P. Frank, *Experimental and Numerical Study of n-Decane*. 2003.

[49] P. Juszczak, D. M. J. Tax, and R. P. W. Duin, 'Feature scaling in support vector data description'.

[50] Z. Hu, B. L. M. T. Somers, R. F. Cracknell, and D. Bradley, 'Investigation of the Livengood–Wu integral for modelling autoignition in a high-pressure bomb', *Combust. Theory Model.*, vol. 20, no. 1, pp. 77–98, Jan. 2016, doi: 10.1080/13647830.2015.1118159.

[51] S. O. Arik and T. Pfister, 'TabNet: Attentive Interpretable Tabular Learning', *ArXiv190807442 Cs Stat*, Dec. 2020, Accessed: Jan. 31, 2022. [Online]. Available: http://arxiv.org/abs/1908.07442

[52] J. Davis and L. Frank, 'Revisiting Batch Normalization', *ArXiv211013989 Cs*, Oct. 2021, Accessed: Jan. 31, 2022. [Online]. Available: http://arxiv.org/abs/2110.13989

[53] J. Pan, P. Zhao, C. K. Law, and H. Wei, 'A predictive Livengood–Wu correlation for two-stage ignition', *Int. J. Engine Res.*, vol. 17, no. 8, pp. 825–835, Oct. 2016, doi: 10.1177/1468087415619516.

[54] A. A. Verbeek, P. A. Willems, G. G. M. Stoffels, B. J. Geurts, and T. H. van der Meer, 'Enhancement of turbulent flame speed of V-shaped flames in fractal-grid-generated turbulence', *Combust. Flame*, vol. 167, pp. 97–112, May 2016, doi: 10.1016/j.combustflame.2016.02.022.

[55] A. D'Adamo, M. Del Pecchia, S. Breda, F. Berni, S. Fontanesi, and J. Prager, 'Chemistry-Based Laminar Flame Speed Correlations for a Wide Range of Engine Conditions for Iso-Octane, n-Heptane, Toluene and Gasoline Surrogate Fuels', Oct. 2017, pp. 2017-01–2190. doi: 10.4271/2017-01-2190.

[56] H. J. Curran, 'Developing detailed chemical kinetic mechanisms for fuel combustion', *Proc. Combust. Inst.*, vol. 37, no. 1, pp. 57–81, 2019, doi: 10.1016/j.proci.2018.06.054.

[57] L. Pulga, G. M. Bianchi, S. Falfari, and C. Forte, 'A machine learning methodology for improving the accuracy of laminar flame simulations with reduced chemical kinetics mechanisms', *Combust. Flame*, vol. 216, pp. 72–81, Jun. 2020, doi: 10.1016/j.combustflame.2020.02.021.

[58] L. Pulga, G. M. Bianchi, M. Ricci, G. Cazzoli, and C. Forte, 'Development of a Novel Machine Learning Methodology for the Generation of a Gasoline Surrogate Laminar Flame Speed Database under Water Injection Engine Conditions', *SAE Int. J. Fuels Lubr.*, vol. 13, no. 1, pp. 04-13-01–0001, Nov. 2019, doi: 10.4271/04-13-01-0001.

[59] M. Metghalchi and J. C. Keck, 'Burning velocities of mixtures of air with methanol, isooctane, and indolene at high pressure and temperature', *Combust. Flame*, vol. 48, pp. 191–210, Jan. 1982, doi: 10.1016/0010-2180(82)90127-4.

References

[60] V. De Bellis, E. Malfi, L. Teodosio, P. Giannattasio, and F. Di Lenarda, 'A Novel Laminar Flame Speed Correlation for the Refinement of the Flame Front Description in a Phenomenological Combustion Model for Spark-Ignition Engines', *SAE Int. J. Engines*, vol. 12, no. 3, pp. 03-12-03–0018, Apr. 2019, doi: 10.4271/03-12-03-0018.

[61] Ö. L. Gülder, 'Laminar burning velocities of methanol, ethanol and isooctane-air mixtures', *Symp. Int. Combust.*, vol. 19, no. 1, pp. 275–281, Jan. 1982, doi: 10.1016/S0082-0784(82)80198-7.

[62] 'Chemically Reacting Flow: Theory, Modeling, and Simulation, 2nd Edition | Wiley', *Wiley.com*. https://www.wiley.com/en-ie/Chemically+Reacting+Flow%3A+Theory%2C+Modeling%2C+and+Simulation%2C+2nd+Edition-p-9781119184874 (accessed Jan. 31, 2022).

[63] T. Poinsot and D. Veynante, 'Theoretical and Numerical Combustion', *Prog Energy Combust Sci*, vol. 28, Jan. 2005.

[64] R. J. Blint, 'The Relationship of the Laminar Flame Width to Flame Speed', *Combust. Sci. Technol.*, vol. 49, no. 1–2, pp. 79–92, Sep. 1986, doi: 10.1080/00102208608923903.

[65] S. Hann, M. Grill, and M. Bargende, 'Reaction Kinetics Calculations and Modeling of the Laminar Flame Speeds of Gasoline Fuels', Apr. 2018, pp. 2018-01–0857. doi: 10.4271/2018-01-0857.

[66] H. Lin, P. Zhao, and H. Ge, 'A Computational Study on Laminar Flame Propagation in Mixtures with Non-Zero Reaction Progress', Apr. 2019, pp. 2019-01–0946. doi: 10.4271/2019-01-0946.

[67] K. E. Niemeyer and C.-J. Sung, 'Reduced Chemistry for a Gasoline Surrogate Valid at Engine-Relevant Conditions', *Energy Fuels*, vol. 29, no. 2, pp. 1172–1185, Feb. 2015, doi: 10.1021/ef5022126.

[68] T. Lu and C. K. Law, 'A directed relation graph method for mechanism reduction', *Proc. Combust. Inst.*, vol. 30, no. 1, pp. 1333–1341, Jan. 2005, doi: 10.1016/j.proci.2004.08.145.

[69] T. Lu and C. K. Law, 'Linear time reduction of large kinetic mechanisms with directed relation graph: n-Heptane and iso-octane', *Combust. Flame*, vol. 144, no. 1–2, pp. 24–36, Jan. 2006, doi: 10.1016/j.combustflame.2005.02.015.

[70] T. Lu and C. K. Law, 'On the applicability of directed relation graphs to the reduction of reaction mechanisms', *Combust. Flame*, vol. 146, no. 3, pp. 472–483, Aug. 2006, doi: 10.1016/j.combustflame.2006.04.017.

[71] P. Pepiotdesjardins and H. Pitsch, 'An efficient error-propagation-based reduction method for large chemical kinetic mechanisms', *Combust. Flame*, vol. 154, no. 1–2, pp. 67–81, Jul. 2008, doi: 10.1016/j.combustflame.2007.10.020.

[72] E. W. Dijkstra, 'A note on two problems in connexion with graphs', *Numer. Math.*, vol. 1, no. 1, pp. 269–271, Dec. 1959, doi: 10.1007/BF01386390.

[73] P. Dirrenberger *et al.*, 'Laminar burning velocity of gasolines with addition of ethanol', *Fuel*, vol. 115, pp. 162–169, Jan. 2014, doi: 10.1016/j.fuel.2013.07.015.

[74] G. van Rossum and F. L. Drake, *The Python language reference*, Release 3.0.1 [Repr.]. Hampton, NH: Python Software Foundation, 2010.

[75] D. P. Kingma and J. Ba, 'Adam: A Method for Stochastic Optimization', *ArXiv14126980 Cs*, Jan. 2017, Accessed: Jan. 31, 2022. [Online]. Available: http://arxiv.org/abs/1412.6980

[76] S. Falfari, G. M. Bianchi, G. Cazzoli, C. Forte, and S. Negro, 'Basics on Water Injection Process for Gasoline Engines', *Energy Procedia*, vol. 148, pp. 50–57, Aug. 2018, doi: 10.1016/j.egypro.2018.08.018.

[77] N. Cavina, N. Rojo, A. Businaro, A. Brusa, E. Corti, and M. De Cesare, 'Investigation of Water Injection Effects on Combustion Characteristics of a GDI TC Engine', *SAE Int. J. Engines*, vol. 10, no. 4, pp. 2209–2218, Sep. 2017, doi: 10.4271/2017-24-0052.

[78] M. S. Gern, A. Vacca, and M. Bargende, 'Experimental Analysis of the Influence of Water Injection Strategies on DISI Engine Particle Emissions', Sep. 2019, pp. 2019-24–0101. doi: 10.4271/2019-24-0101.

[79] J. Khatri, I. Denbratt, P. Dahlander, and L. Koopmans, 'Water Injection Benefits in a 3-Cylinder Downsized SI-Engine', Jan. 2019, pp. 2019-01–0034. doi: 10.4271/2019-01-0034.

[80] F. Hoppe, M. Thewes, J. Seibel, A. Balazs, and J. Scharf, 'Evaluation of the Potential of Water Injection for Gasoline Engines', *SAE Int. J. Engines*, vol. 10, no. 5, pp. 2500–2512, Sep. 2017, doi: 10.4271/2017-24-0149.

References

[81] G. Cazzoli, S. Falfari, G. M. Bianchi, and C. Forte, 'Development of a chemical-kinetic database for the laminar flame speed under GDI and water injection engine conditions', *Energy Procedia*, vol. 148, pp. 154–161, Aug. 2018, doi: 10.1016/j.egypro.2018.08.043.

[82] N. Al-Esawi and M. Al Qubeissi, 'A new approach to formulation of complex fuel surrogates', *Fuel*, vol. 283, p. 118923, Jan. 2021, doi: 10.1016/j.fuel.2020.118923.

[83] S. Gail *et al.*, 'Evaluating a novel gasoline surrogate containing isopentane using a rapid compression machine and an engine', *Proc. Combust. Inst.*, vol. 38, no. 4, pp. 5643–5653, 2021, doi: 10.1016/j.proci.2020.07.103.

[84] S. Gail *et al.*, 'THIP: A new TPRF-like fuel surrogate development approach to better match real fuel properties', *Fuel*, vol. 286, p. 119395, Feb. 2021, doi: 10.1016/j.fuel.2020.119395.

[85] V. S. Bhavani Shankar *et al.*, 'Primary Reference Fuels (PRFs) as Surrogates for Low Sensitivity Gasoline Fuels', Apr. 2016, pp. 2016-01–0748. doi: 10.4271/2016-01-0748.

[86] G. Kalghatgi, H. Babiker, and J. Badra, 'A Simple Method to Predict Knock Using Toluene, N-Heptane and Iso-Octane Blends (TPRF) as Gasoline Surrogates', *SAE Int. J. Engines*, vol. 8, no. 2, pp. 505–519, Apr. 2015, doi: 10.4271/2015-01-0757.

[87] N. Morgan, A. Smallbone, A. Bhave, M. Kraft, R. Cracknell, and G. Kalghatgi, 'Mapping surrogate gasoline compositions into RON/MON space', *Combust. Flame*, vol. 157, no. 6, pp. 1122–1131, Jun. 2010, doi: 10.1016/j.combustflame.2010.02.003.

[88] M. Di Lorenzo, P. Brequigny, F. Foucher, and C. Mounaïm-Rousselle, 'Validation of TRF-E as gasoline surrogate through an experimental laminar burning speed investigation', *Fuel*, vol. 253, pp. 1578–1588, Oct. 2019, doi: 10.1016/j.fuel.2019.05.081.

[89] M. Mehl, J. Y. Chen, W. J. Pitz, S. M. Sarathy, and C. K. Westbrook, 'An Approach for Formulating Surrogates for Gasoline with Application toward a Reduced Surrogate Mechanism for CFD Engine Modeling', *Energy Fuels*, vol. 25, no. 11, pp. 5215–5223, Nov. 2011, doi: 10.1021/ef201099y.

[90] K. V. Puduppakkam, L. Liang, C. V. Naik, E. Meeks, S. L. Kokjohn, and R. D. Reitz, 'Use of Detailed Kinetics and Advanced Chemistry-Solution Techniques in CFD to Investigate Dual-Fuel Engine Concepts', *SAE Int. J. Engines*, vol. 4, no. 1, pp. 1127–1149, Apr. 2011, doi: 10.4271/2011-01-0895.

[91] M. Del Pecchia and S. Fontanesi, 'A methodology to formulate multicomponent fuel surrogates to model flame propagation and ignition delay', *Fuel*, vol. 279, p. 118337, Nov. 2020, doi: 10.1016/j.fuel.2020.118337.

[92] R. Xu *et al.*, 'A physics-based approach to modeling real-fuel combustion chemistry – VI. Predictive kinetic models of gasoline fuels', *Combust. Flame*, vol. 220, pp. 475–487, Oct. 2020, doi: 10.1016/j.combustflame.2020.07.020.

[93] S. Esposito, L. Cai, M. Günther, H. Pitsch, and S. Pischinger, 'Experimental comparison of combustion and emission characteristics between a market gasoline and its surrogate', *Combust. Flame*, vol. 214, pp. 306–322, Apr. 2020, doi: 10.1016/j.combustflame.2019.12.025.

[94] M. Del Pecchia, V. Pessina, F. Berni, A. d'Adamo, and S. Fontanesi, 'Gasoline-ethanol blend formulation to mimic laminar flame speed and auto-ignition quality in automotive engines', *Fuel*, vol. 264, p. 116741, Mar. 2020, doi: 10.1016/j.fuel.2019.116741.

[95] H. Chu, L. Xiang, X. Nie, Y. Ya, M. Gu, and J. E, 'Laminar burning velocity and pollutant emissions of the gasoline components and its surrogate fuels: A review', *Fuel*, vol. 269, p. 117451, Jun. 2020, doi: 10.1016/j.fuel.2020.117451.

[96] V. Mariani, L. Pulga, G. M. Bianchi, S. Falfari, and C. Forte, 'Machine Learning-Based Identification Strategy of Fuel Surrogates for the CFD Simulation of Stratified Operations in Low Temperature Combustion Modes', *Energies*, vol. 14, no. 15, p. 4623, Jul. 2021, doi: 10.3390/en14154623.

[97] '[1808.03030] Policy Optimization as Wasserstein Gradient Flows'. https://arxiv.org/abs/1808.03030 (accessed Jan. 31, 2022).

[98] D. D. Das, P. C. St. John, C. S. McEnally, S. Kim, and L. D. Pfefferle, 'Measuring and predicting sooting tendencies of oxygenates, alkanes, alkenes, cycloalkanes, and aromatics on a unified scale', *Combust. Flame*, vol. 190, pp. 349–364, Apr. 2018, doi: 10.1016/j.combustflame.2017.12.005.

References

[99] X. Su, Y. Ra, and R. D. Reitz, 'A Surrogate Fuel Formulation Approach for Real Transportation Fuels with Application to Multi-Dimensional Engine Simulations', *SAE Int. J. Fuels Lubr.*, vol. 7, no. 1, pp. 236–249, Apr. 2014, doi: 10.4271/2014-01-1464.

[100] G. Richard. Handrick, 'Heats of Combustion of Organic Compounds', *Ind. Eng. Chem.*, vol. 48, no. 8, pp. 1366–1374, Aug. 1956, doi: 10.1021/ie50560a039.

[101] A. S. AlRamadan, S. M. Sarathy, M. Khurshid, and J. Badra, 'A blending rule for octane numbers of PRFs and TPRFs with ethanol', *Fuel*, vol. 180, pp. 175–186, Sep. 2016, doi: 10.1016/j.fuel.2016.04.032.

[102] J. E. Anderson *et al.*, 'Octane Numbers of Ethanol-Gasoline Blends: Measurements and Novel Estimation Method from Molar Composition', Apr. 2012, pp. 2012-01–1274. doi: 10.4271/2012-01-1274.

[103] A. Bertolino, A. Stagni, A. Cuoci, T. Faravelli, A. Parente, and A. Frassoldati, 'Prediction of flammable range for pure fuels and mixtures using detailed kinetics', *Combust. Flame*, vol. 207, pp. 120–133, Sep. 2019, doi: 10.1016/j.combustflame.2019.05.036.

[104] L. Sileghem, J. Vancoillie, J. Demuynck, J. Galle, and S. Verhelst, 'Alternative Fuels for Spark-Ignition Engines: Mixing Rules for the Laminar Burning Velocity of Gasoline–Alcohol Blends', *Energy Fuels*, vol. 26, no. 8, pp. 4721–4727, Aug. 2012, doi: 10.1021/ef300393h.

[105] T. Hirasawa, C. J. Sung, A. Joshi, Z. Yang, H. Wang, and C. K. Law, 'Determination of laminar flame speeds using digital particle image velocimetry: Binary Fuel blends of ethylene, n-Butane, and toluene', *Proc. Combust. Inst.*, vol. 29, no. 2, pp. 1427–1434, Jan. 2002, doi: 10.1016/S1540-7489(02)80175-4.

[106] V. S. Yumlu, 'Prediction of burning velocities of saturated carbon monoxide-air flames by application of mixing rules', *Combust. Flame*, vol. 11, no. 5, pp. 389–396, Oct. 1967, doi: 10.1016/0010-2180(67)90059-4.

[107] G. A. Lavoie, 'Correlations of Combustion Data for S. I. Engine Calculations – Laminar Flame Speed, Quench Distance and Global Reaction Rates', *SAE Trans.*, vol. 87, pp. 1015–1033, 1978.

[108] S. Jerzembeck, N. Peters, P. Pepiotdesjardins, and H. Pitsch, 'Laminar burning velocities at high pressure for primary reference fuels and gasoline: Experimental and numerical investigation', *Combust. Flame*, vol. 156, no. 2, pp. 292–301, Feb. 2009, doi: 10.1016/j.combustflame.2008.11.009.

[109] P. Virtanen *et al.*, 'SciPy 1.0: fundamental algorithms for scientific computing in Python', *Nat. Methods*, vol. 17, no. 3, pp. 261–272, Mar. 2020, doi: 10.1038/s41592-019-0686-2.

[110] D. J. Gaspar *et al.*, 'Measuring and predicting the vapor pressure of gasoline containing oxygenates', *Fuel*, vol. 243, pp. 630–644, May 2019, doi: 10.1016/j.fuel.2019.01.137.

[111] W. Kirch, Ed., 'Pearson's Correlation Coefficient', in *Encyclopedia of Public Health*, Dordrecht: Springer Netherlands, 2008, pp. 1090–1091. doi: 10.1007/978-1-4020-5614-7_2569.

[112] L. Pulga, S. Falfari, G. M. Bianchi, M. Ricci, and C. Forte, 'Advanced Combustion Modelling of High BMEP Engines under Water Injection Conditions with Chemical Correlations Generated with Detailed Kinetics and Machine Learning Algorithms', *SAE Int. J. Adv. Curr. Pract. Mobil.*, vol. 3, no. 1, pp. 77–94, Sep. 2020, doi: 10.4271/2020-01-2008.

[113] S. Falfari, G. M. Bianchi, L. Pulga, and C. Forte, 'PWI and DWI Systems in Modern GDI Engines: Optimization and Comparison Part II: Reacting Flow Analysis', Apr. 2021, pp. 2021-01–0454. doi: 10.4271/2021-01-0454.

[114] S. Falfari, G. Cazzoli, M. Ricci, and C. Forte, 'PWI and DWI Systems in Modern GDI Engines: Optimization and Comparison Part I: Non-Reacting Flow Analysis', Apr. 2021, pp. 2021-01–0461. doi: 10.4271/2021-01-0461.

[115] J. R. Serrano, F. J. Arnau, P. Piqueras, and O. García-Afonso, 'Application of the two-step Lax and Wendroff FCT and the CE-SE method to flow transport in wall-flow monoliths', *Int. J. Comput. Math.*, vol. 91, no. 1, pp. 71–84, Jan. 2014, doi: 10.1080/00207160.2013.783206.

[116] O. S. A *et al.*, 'Renormalization Group Modeling and Turbulence Simulations.', *-Wall Turbul. Flows*, pp. 1031–1046, 1993.

[117] O. Colin and A. Benkenida, 'The 3-Zones Extended Coherent Flame Model (Ecfm3z) for Computing Premixed/Diffusion Combustion', *Oil Gas Sci. Technol.*, vol. 59, no. 6, pp. 593–609, Nov. 2004, doi: 10.2516/ogst:2004043.

[118] N. Peters and M. Dekena, 'Combustion Modeling with the G-Equation', *Oil Gas Sci. Technol.*, vol. 54, no. 2, pp. 265–270, Mar. 1999, doi: 10.2516/ogst:1999024.

References

[119] A. Trouvé and T. Poinsot, 'The evolution equation for the flame surface density in turbulent premixed combustion', *J. Fluid Mech.*, vol. 278, pp. 1–31, Nov. 1994, doi: 10.1017/S0022112094003599.

[120] S. Chevillard, O. Colin, J. Bohbot, M. Wang, E. Pomraning, and P. K. Senecal, 'Advanced Methodology to Investigate Knock for Downsized Gasoline Direct Injection Engine Using 3D RANS Simulations', Mar. 2017, pp. 2017-01–0579. doi: 10.4271/2017-01-0579.

[121] Z. Han and R. D. Reitz, 'A temperature wall function formulation for variable-density turbulent flows with application to engine convective heat transfer modeling', *Int. J. Heat Mass Transf.*, vol. 40, no. 3, pp. 613–625, Feb. 1997, doi: 10.1016/0017-9310(96)00117-2.

[122] B. E. Launder and D. B. Spalding, 'The numerical computation of turbulent flows', *Comput. Methods Appl. Mech. Eng.*, vol. 3, no. 2, pp. 269–289, Mar. 1974, doi: 10.1016/0045-7825(74)90029-2.

[123] *Convective Heat and Mass Transfer*. Accessed: Jan. 31, 2022. [Online]. Available: https://books.google.com/books/about/Convective_Heat_and_Mass_Transfer.html?id=PpkeAQAAIAAJ

[124] C. D. Rakopoulos, G. M. Kosmadakis, and E. G. Pariotis, 'Critical evaluation of current heat transfer models used in CFD in-cylinder engine simulations and establishment of a comprehensive wall-function formulation', *Appl. Energy*, vol. 87, no. 5, pp. 1612–1630, May 2010, doi: 10.1016/j.apenergy.2009.09.029.

[125] S. Keum, H. Park, A. Babajimopoulos, D. N. Assanis, and D. Jung, 'Modelling of heat transfer in internal combustion engines with variable density effect', *Int. J. Engine Res.*, vol. 12, no. 6, pp. 513–526, Dec. 2011, doi: 10.1177/1468087411410015.

[126] F. Berni, S. Fontanesi, G. Cicalese, and A. D'Adamo, 'Critical Aspects on the Use of Thermal Wall Functions in CFD In-Cylinder Simulations of Spark-Ignition Engines', *SAE Int. J. Commer. Veh.*, vol. 10, no. 2, pp. 547–561, Mar. 2017, doi: 10.4271/2017-01-0569.

[127] M. Ricci, L. Pulga, G. M. Bianchi, S. Falfari, and C. Forte, 'Numerical Aspects Affecting Heat Transfer in ICE Applications and Definition of a Temperature Wall Function Accounting for the Boundary Layer Compressibility', *SAE Int. J. Engines*, vol. 12, no. 5, pp. 03-12-05–0034, Aug. 2019, doi: 10.4271/03-12-05-0034.

[128] A. C. Alkidas, 'Heat Transfer Characteristics of a Spark-Ignition Engine', *J. Heat Transf.*, vol. 102, no. 2, pp. 189–193, May 1980, doi: 10.1115/1.3244258.

[129] F. Brusiani, G. M. Bianchi, and A. Tiberi, 'Primary Breakup Model for Turbulent Liquid Jet Based on Ligament Evolution', Apr. 2012, pp. 2012-01–0460. doi: 10.4271/2012-01-0460.

[130] M. Pilch and C. A. Erdman, 'Use of breakup time data and velocity history data to predict the maximum size of stable fragments for acceleration-induced breakup of a liquid drop', *Int. J. Multiph. Flow*, vol. 13, no. 6, pp. 741–757, Nov. 1987, doi: 10.1016/0301-9322(87)90063-2.

[131] L. Postrioti, A. Cavicchi, G. Brizi, F. Berni, and S. Fontanesi, 'Experimental and Numerical Analysis of Spray Evolution, Hydraulics and Atomization for a 60 MPa Injection Pressure GDI System', Apr. 2018, pp. 2018-01–0271. doi: 10.4271/2018-01-0271.

[132] K. Steurs, C. K. Blomberg, and K. Boulouchos, 'Knock in an Ethanol Fueled Spark Ignition Engine: Detection Methods with Cycle-Statistical Analysis and Predictions Using Different Auto-Ignition Models', *SAE Int. J. Engines*, vol. 7, no. 2, pp. 568–583, Apr. 2014, doi: 10.4271/2014-01-1215.