Ramalli, E., Dinelli, T., Nobili, A., Stagni, A., Pernici, B. and Faravelli, T., 2023. Automatic validation and analysis of predictive models by means of big data and data science. Chemical Engineering Journal, 454, p.140149.

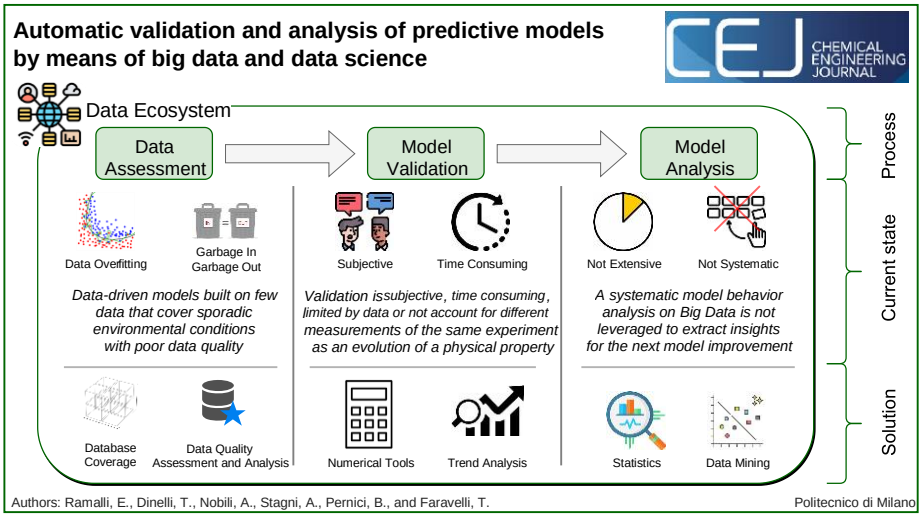
Accepted paper version, published with the following DOI: https://doi.org/10.1016/j.cej.2022.140149

Graphical Abstract

Automatic validation and analysis of predictive models by means of big data and data science

E. Ramalli, T. Dinelli, A. Nobili, A. Stagni,

B. Pernici, T. Faravelli



Automatic validation and analysis of predictive models by means of big data and data science

E. Ramalli[a,1], T. Dinelli[a,2], A. Nobili[a,2], A. Stagni[a,2],
B. Pernici[a,1], T. Faravelli [a,2,*]

[a]*Politecnico di Milano, Piazza Leonardo da Vinci 32, Milan, Italy*

Abstract

Validation is an essential procedure in the development of a predictive model in several engineering fields. In addition, recent data analysis techniques and the increasing availability of data have the potential to provide a deeper understanding of experimental data and simulation models. This work proposes a systematic, objective, and automatic methodology to validate and analyze experiments and models from a high-level perspective. The proposed methodology exploits the opportunities offered by the 'data ecosystem' concept, combining data and model evaluation and providing an integrated set of techniques to produce synthetic but comprehensive insights about the experiment and the predictive model. The methodology focuses on data assessment of the experiments used in the process, the use of a trend similarity comparison index to measure the model performance, and data science techniques to systematically extract models' behavior insight by analyzing a large number of validation results and linking them to the experiment characteristics.

*Corresponding author

Email addresses: edoardo.ramalli@polimi.it (E. Ramalli), timoteo.dinelli@polimi.it (T. Dinelli), andrea.nobili@polimi.it (A. Nobili), alessandro.stagni@polimi.it (A. Stagni), barbara.pernici@polimi.it (B. Pernici), tiziano.faravelli@polimi.it (T. Faravelli)

[1]Dept. of Electronics, Information, and Bioengineering, Politecnico di Milano

[2]Dept. of Chemistry, Materials, and Chemical Engineering "G. Natta", Politecnico di Milano

Preprint submitted to Chemical Engineering Journal October 26, 2022

The automated proposed approach follows the generality principle and can be extended to different application domains in which predictive models are validated against big data in the chemical engineering domain. As a case study, the proposed methodology is applied with hundreds of experimental datasets to evaluate a kinetic model that describes the pyrolysis and combustion of hydrocarbons.

Keywords: model validation, model analysis, data science, data mining, big data, knowledge extraction

---

1. Introduction

Over the past few decades, the progress in computing power, the availability of more and more data, and the tendency to share information boosted the development of many research and industrial areas [1]. The availability of predictive models to forecast a system state brought new insights into the comprehension of the phenomena, industrial applications, and social benefits, in many different sectors, from engineering to social science. In particular, in chemical engineering, increasingly complex physical-chemical models are capable of predicting at different scales diverse states of a domain [2, 3].

The generation procedure of complex models, due to a large amount of available data, is changing from an approach solely based on first principles to data-driven methodologies [4, 5]. In fact, the data collection phase is an increasingly fundamental step in a research plan to develop a predictive model [6]. In addition to data collection, it is also necessary to support a data preparation phase, in which the data collected through experimental campaigns and other available sources are validated and integrated for subsequent uses.

As these phases require a considerable effort, and due to the complexity and the many possibilities to model a domain, the "many-data many-models" problem originated [7]: many models are available to predict the same subject (i.e., the quantity or property of interest), but they differ in the number and form of

mathematical equations representing the phenomena or in the selection of parameters [7, 8]. These degrees of freedom and the "many-data" led to the development of many models of various complexity from different research groups concerning the same subject but based on a different subset of experiments. The result was the generation of inconsistent and not general models [9]. In addition, a manual evaluation of model quality through comparison with experimental data, and a univocal, quantitative ranking of the results, are not straightforward operations [10].

Therefore, there is a need to organize the available information, conceptualize the problem in terms of big data, and automate the model validation and analysis procedures. This approach can extract knowledge from the data to speed up the development process while reducing error-prone tasks [10], defining in practice what can be discovered [36, 37]. The validation procedure (or assessment) links the development of a complex model to the experimental data computing the predictive model performances by comparing the model predictions with the corresponding experimental measurements. Such comparison is traditionally performed manually via a graphical approach: experimental and simulated data are plotted together in the same figure, and the researcher evaluates whether the predictions are good enough to consider the model acceptable. Even if model validation is a "poorly posed problem" [7], this approach has two strong limitations: (i) it lacks objectivity since the same comparison can be good for someone and admissible, or even worse, for others. (ii) the validation is not extensive since the availability of human resources highly limits this time-consuming procedure.

As a consequence, one can easily lose control of the model development since, at each modification, it should be re-validated against a large number of experimental data to verify that the changes have not negatively impacted other model areas. This problem is known as the "short blanket" dilemma. As a side result, manual validation can not extract systematic features about the model

behavior based on large quantities of data, which could be helpful in providing suggestions for the next model improvement.

A numerical procedure to execute the model validation accounts for the first problem since it provides an unbiased model performance assessment. Some of these methodologies take only into account the distance between measurements and predictions, with metrics such as mean square error [27-29], R2 [30, 31], or customization of them based on the application, referred to as objective error function [32]; others also consider the dissimilarities and similarities among the shapes of the experimental and simulated data curves [33, 34]. However, there is the need to develop a more comprehensive approach to compare and analyze different numerical model validations on very large quantities of data, in addition to new ways of analyzing in-depth critical cases when identified, exploiting all available data. Therefore, an information system is essential to manage analyses of big data automatically. Over time, there were several initiatives aimed at collecting experimental data in frameworks or data ecosystems. Their typical challenges are the involvement of the scientific community in data sharing, providing services to users, and the standardization of data representation in agreed formats [11].

In the combustion domain, since the first example of the PrIMe (Process Informatics Model) system [12], these frameworks do not exist only as scientific data repositories but offer other domain-related services. PrIMe, in particular, also had the purpose of collecting predictive models and generating them based on specific user requests (e.g., operating conditions), providing services to control the consistency of the experimental data [13, 14], and validating the models [15]. The Bound-to-Bound Data Collaboration (B2BDC) methodology is a part of the PrIMe framework. It is rooted around the concept of consistency, and it is the first methodology that uses data to define constraints to bound a feasible space of variables [16]. The B2BDC casts the problem of model validation in an optimization setting inside the feasible space [13]. A model, in this space, can be generated, validated [15, 17], optimized, and the model uncertainty quantified

and analyzed [18-21]. As an evolution of PrIMe, CloudFlame was proposed [22, 23]. It offers cloud simulation computing capabilities, a data repository, and a model generation feature. Another framework is ReSpecTh which contains reaction kinetics, high-resolution molecular spectroscopy, and thermochemistry data [24]. It offers different functionalities, such as starting and running multiple simulations, visualizing data, and automatically validating models [24].

By leveraging a newly-conceived data ecosystem, this work proposes a systematic and automatic end-to-end methodology that first assesses the data used for validation, objectively validates the model and then analyzes the validation results with a data science approach to derive model behavior insights. The outcomes of such methodology are essential to guide the automatic and intelligent generation of new complex predictive models. Therefore, the proposed methodology aims at identifying a common direction, driving the scientific community toward an objective and uniquely consistent approach to define model quality, then fostering their further improvement. The methodology outcomes are synthetic analysis results that provide suggestions about ”which”, ”where”, ”why”, and ”how much” the model predictions are not satisfactory from a high-level perspective, together with a synthetic performance index of the model. The procedure does not rely upon or connect the prediction results to the application-specific component of a mechanism, thus keeping its generality. Instead, in the analyses, it links the model performances to the experiment characteristic (or metadata) to derive explainable model insights. A new data mining technique called interval analysis has also been implemented to estimate precisely ”how much” the model deviates from the experiments. Model component-level tools and analyses can be used as other sources of information to enrich the analysis accordingly to the applicative domain. For each step of the proposed approach, the current pitfalls and their mitigations are explained, together with the opportunities in terms of trustworthiness, comprehension, and development process improvement of the model that data science methodologies can bring to the field.

With the spread of frameworks for model generation, validation, and experiment collection, the growing number of increasingly complex models has involved development of several initiatives such as CaRMeN (Catalytic Reaction Mechanism Network) [10], those proposed by West et al. [25] or by Killingsworth et al. [26], which offered tools to check the physical consistency of predictive models, identify errors, and compare their performance.

Other successful applications in chemical engineering have shown the advantage of using data science [38, 29] or, more in general, computer science, such as PCA or Knowledge Graph approaches to extract new knowledge from data [39, 40] or inside an optimization procedure of existing models [32]. Furthermore, since many machine learning applications are spreading in this research area [41-46], it is important to apply and adapt to the chemical engineering domain the existing expertise in the computer science community to avoid well-known issues. One of the most important is related to the overfitting of the data during the model generation, which leads to a biased model [47].

Finally, as a case study, while this manuscript and the proposed methodology do not focus on a specific chemical engineering domain, the paper provided examples in the domain of chemical kinetics to illustrate the end-to-end methodology for data assessment, model validation, and analysis. In particular, the methodology is applied on a detailed kinetic mechanism describing the pyrolysis and combustion of conventional as well as next-generation fuels, like hydrogen, methane, and their blends, which can be derived from bio-feedstocks, using SciExpeM as data ecosystem [48].

2. Materials and methods

This section presents the proposed end-to-end methodology with the related techniques and tools, to validate and analyze a predictive model. The goal of the approach is to provide an *objective* and *systematic* procedure as illustrated in the following. An *objective* procedure implies that the model performance

assessment, i.e., the similarity measurements between the experimental and the corresponding simulated data, must be based on numerical approaches. There is a need to compare different numerical approaches and to select the more adequate ones in the different phases of the analysis, considering also computational constraints. In addition, the validation should include a large amount of experimental data with a certain quality and diversity; this would avoid the bias in relying on a predictive model that is tested only against a few experiments that cover sporadic environmental conditions with poor data quality. To this purpose, techniques for appropriate data selection and data quality assessment are needed. Finally, being objective suggests that the validation and the analysis procedures have to be defined as a detailed and replicable sequence of steps that will make the results reusable and the predictive models comparable while enhancing the trustworthiness of a new model release. On the other hand, a *systematic* procedure leverages the model validation results to detect recurring characteristics and patterns of the model predictive capability. Such methodology needs the largest available quantities of data; thus, the overall procedure should be automated.

This methodology compares the model's outcomes with experiments to provide knowledge on model behavior under a wide range of conditions. Other sources of information should enhance the analysis to improve the predictive capability of a model. For example, in the combustion fields, the use of well-known instruments, such as parameter sensitivity analysis and flux analysis, are a valid source of model component-level data [49, 27]

Therefore, this section presents (i) an approach oriented toward an automatic, standardized, and unbiased validation of the predictive model and (ii) data science-based techniques to analyze the model behavior providing suggestions for its later improvement, laying down the first stones for a machine learning-oriented predictive model development.

Section 2.1 introduces the characteristics that a data ecosystem should include to support this approach in terms of collecting, storing, and analyzing

experiments, simulations, and complex models. Section 2.2 presents a general overview of the methodology and the data science techniques used to assess a model's predictive performance and extract knowledge about the model behavior systematically. Finally, the techniques and tools proposed and developed to support the methodology are described in detail. Section 2.3 focuses on data assessment, while Section 2.4 and Section 2.5 describe the proposed validation and analysis tools respectively.

2.1. Data Management System

A data ecosystem (DE) has various characteristics and services, but the most important functionality is the capability to manage a large amount of data considering different aspects. This part of the system is also known as *data management system*. A DE for the development of a predictive model speeds up two important aspects. First, it stores experimental and simulated data together with models in the same place. As a result, it optimizes the reuse of resources, saves time in the search for experiments, and encourages data sharing among different researchers according to the Findable, Accessible, Interoperable, and Reusable (FAIR) principles [50]. Second, it is able to automatically manage large quantities of data and apply data science techniques, such as data mining, machine learning, and statistical analysis. This kind of study allows a deeper examination of the model and the collection of systematic insights from a broader set of data that could not be manually observable.

The automatic analysis of a larger amount of data is a game-changer in the predictive model development process, not only in the chemical engineering field: the researchers need to have a synthetic yet exhaustive overview of the model performances in many different conditions without the risk of overfitting the model only on the few data that they can manually handle. At the same time, such analysis can provide model developers with suggestions about *where*, *why*, and *how much* a model requires some improvements. Another advantage of a DE is in the "design of experiment" phase. Since the experiments are all stored and

categorized in a database, it is possible to know which domain area(s) lack experiments or suggest in which there are few, and the predictive model performances are not good enough; thus, other data are needed to comprehend better the phenomena.

A DE with these purposes must manage four data types: experiments, simulations, models, and analysis results. Each data type is linked to the other, such that each of them can be used to validate the other. For instance, to validate an experimental observation, it is possible to use both the chemistry theory data and the simulations, but also vice-versa. Therefore, a DE is a data-centralized structure that has the advantage of sharing and managing the knowledge between all the data sources. The drawback of this approach is a fast propagation of an error, but if proper data quality rules are set (see Section 2.3.1), this hazard can be managed. Moreover, having all the data in the same place incentivize more users to use the DE, enhancing the trustworthiness of the data and of the system itself. The more the system is trustworthy the more users are likely to use it; a positive vicious cycle is started.

The term experiment (or experimental data) is used in this work to refer to both a set of experimental observations about a specific target or measured property as well as data from chemical theory, given specific environmental conditions (or experimental setting). Similarly, a simulation (or simulated dataset, or simulated data) is the collection of data points obtained using a model to forecast the output of a system given a set of environmental conditions.

2.2. Model Evaluation Methodology

The model evaluation methodology includes three phases (see Figure 1) that follow the principles of *objectivity* and *systematicity* explained previously, solving the limitations presented in the introduction.

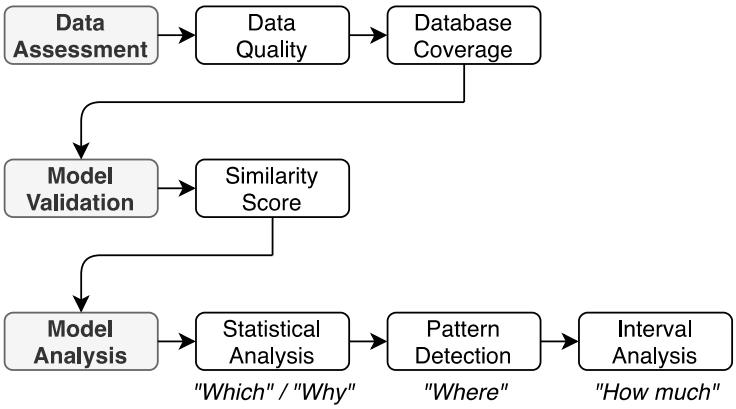


Figure 1: The proposed model evaluation process.

The first phase is the data assessment, illustrated in Section 2.2.1, which evaluates different characteristics of the data used in this process since they directly impact the quality of the outcomes. Then, during model validation (Section 2.2.2), the similarity (score) between the experimental data and the model prediction (or simulated data) is properly quantified. The simulation's similarity score (or index) is the similitude, computed according to a score function, between the experimental data and the corresponding simulated data. The score range is [0,1], where 1 is the perfect similarity. In general, several possible scoring methods can be used and compared during the analysis, as discussed in Section 2.2. Finally, the model analysis described in Section 2.2.3 uses the validation results and data science to derive insights about *which*, *where*, *why*, and *how much* a model performs inaccurately.

A proper data ecosystem is required to implement such an automatic procedure to elaborate *big data* and therefore standardize and significantly reduce both the time spent and the human-related errors during the steps of the overall model development process.

The model evaluation procedure can be applied both in the case of evaluating a single model performances or comparing those of two different ones at the

		Validation		Analysis
		Qualitative	*Quantitative*	*Quantitative*
Grain	*Coarse*		Trend Score	Pattern Detection Statistical
	Fine	Visualization	Point-Wise Score*	Interval

Table 1: Techniques used to validate and analyze a model. * denote a tool that is present in the literature, but it is not used in our approach. Each tool is quantitative or qualitative and provides detailed (Fine grain) or general information (Coarse grain).

same time. In the first case, the similarity score is used in absolute terms. In the latter, the similarity score is used in terms of relative variations, therefore evaluating improvements or worsening in the model. In both cases, the overall evaluation procedure is not different.

Table 1 reports the available techniques for the model validation, distinguishing between quantitative and qualitative approaches. Next to the validation tools, the analysis technologies are listed, to provide insights about the model behavior systematically. Both validation and analysis techniques are divided into coarse and fine grain, denoting respectively methodologies capable of summarizing multiple aspects and large quantities of data or deeper and punctual study. The evaluation procedure presented in this paper considers visualization an optional step, useful only to check a few particular cases further.

2.2.1. Data Assessment

This first phase quantifies the quality and diversity of the validation set of experimental data selected for the following model validation and analysis. The experimental data collection should be the most extensive, diverse, and high quality possible to avoid the hazard of overfitting the model on the selected data or providing wrong information during the following phases. This phase can in part be performed during the model evaluation procedure and in part be

supported by the automatic data quality control procedures defined within the management of the data ecosystem, independently of the development of a single model. As explained in more detail in Section 2.3.1, data quality controls employ a combination of rule and threshold approaches to establish if an experiment is reliable to be used in the data ecosystem. The rule-based approach is independent of the quantity of available data, whereas the threshold guarantees more reliable results with more data. However, predictive models from different sources can be used to make up for the missing data since it is reasonable to think that they are not perfect but reliable enough in most cases.

2.2.2. Model Validation

This phase entails the quantification of the predictive model performances. To be *objective*, the procedure employs a quantitative approach that, by measuring the similarity between the trend (Trend score) of the experimental data points as a whole against the corresponding simulated data, provides a synthetic index of the model's performance. It is fair to point out that deciding which objective metric to use is not subjective. In fact, based on the specific application domain, some metrics are proved to be better than others. In chemical engineering, most models are evaluated not only on the prediction capabilities to forecast a single point, but also on the ability to apprehend the trend of point series as in a parametric analysis. For this reason, generally speaking, comparing the trend is a safe choice since it also includes a point-wise comparison. Moreover, the similarity score must account for the experimental uncertainty when comparing the experiment to the simulations.

The similarity score is automatically computed for each experimental and simulated data pair. Once this operation is concluded, the model validation is completed: a general and synthetic performance overview of the model is now available as an average of all similarity indexes. Instead, when comparing two models, the percentage variation between them is evaluated.

2.2.3. Model Analysis

The third and last macro phase of the evaluation procedure consists of analyzing the similarity indexes computed during the model validation. The model analysis leverages data science techniques to collect knowledge about the predictive model's behavior systematically. Notwithstanding that the proper interpretation of analysis results is dependent on the type of application.

As in many computer science applications, also in this case there is a need to address what is known as the "curse of dimensionality" [51]. In fact, for the model validation, for each pair of experimental and simulated data is computed a similarity score, and an average of all of them is a fair indicator of the general model performance. However, such an average is not able to provide detailed information about the behavior of the predictive model since it depends on many variables. The application of data science techniques allows managing the many dimensions that define a domain and the thousands of similarity indexes computed during validation to extract insights automatically.

The analysis phase is characterized by the following three steps that can be used to study the model simulation results more and more in depth: statistical analysis, pattern detection, and interval analysis.

Statistical Analysis. Experimental data are provided with additional information (also referred to as characteristics, metadata, or properties) that can be leveraged to statistically analyze the model performances in a complex and multidimensional domain. First, it is possible to group the similarity indexes based on common characteristics of the experiment to know *which* combination of them indicates the worst model performance. Second, using correlation on the experiment metadata, it is possible to investigate *why* the model does not perform well enough, i.e., outside the experimental uncertainties, or in other words, the contributing causes.

First of all, a collection of experiments is filtered based on their similarity score, whether it is below the first quartile (25^{th} percentile or 1^{st} quartile).

The percentile is computed with respect to the global distribution of similarity

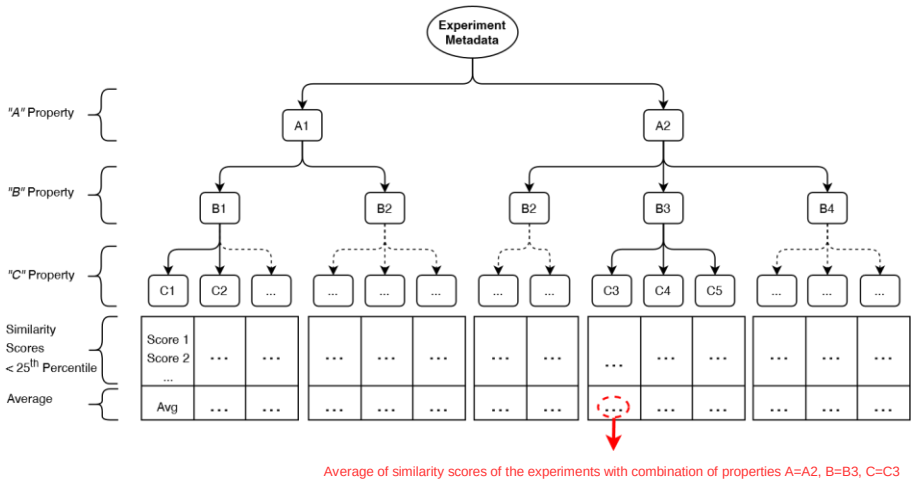


Figure 2: Subdivision and average of the similarity scores of the experiments with the same combination of categorical attributes (A, B, C).

scores. In such a way, the focus is immediately shifted to the experiments with the associated worst predictive model performance.

However, it is necessary to find out *which* combinations of the properties of the experiments correspond to such behavior. Without losing generality, let us assume that each experiment is characterized by the values assumed in correspondence of three categorical properties *A,B,C*. Since each property is categorical, only a precise set of values can be assumed, and not all combinations of property values are possible in a domain. Let us assume, for instance, that the values for each category are defined as follows:

$A = \{A1,A2\}$ $B = \{B1,B2,B3,B4\}$, $C = \{C1,...,C5\}$.

Following the idea pictured in Figure 2, it is possible to compute the average (or other statistical measures) of the similarity scores of the experiment that have a precise combination of experimental properties. Then the combinations of properties that are statistically relevant are observed. A combination is statistically relevant whether it has a considerable number of cases in the quartile

and a high percentage of them with respect to the total number of existing ones with that combination of properties. In a second moment, a correlation analysis between all the experiments metadata, such as type of experiment and environmental conditions, together with the similarity score, suggests, for example, that the model performance is due the use of particular conditions (like equipment) when a specific variable (species for example) is measured. To this purpose, both clustering and classification techniques can be adopted to analyze the results on a large scale.

Pattern Detection. Pattern detection algorithms, such as clustering, applied to the similarity index, together with (numerical and continuous) physical properties associated with an experiment such as temperature and pressure, can automatically distinguish the portions of the domain *where* the model does show larger mispredictions. In fact, clustering algorithms group similar experiments in the same cluster: taking the most representative cluster(s) with the lowest variation of the associated performance scores, it is possible to know which combination of physical property range is responsible for the worst performance.

Interval Analysis. Once the analysis has been identified for *which* combination(s) of metadata and *where* the model is more deficient, with the developed *ad-hoc* analysis of the intervals, it is possible to quantify (*how much*) the average deviation of the experimental curve from the simulated one.

In the following sections, the approaches proposed in this paper for data assessment and to validate and analyze the model through the techniques mentioned above are described in detail.

2.3. Data assessment techniques

Data assessment is related to the evaluation of the quality of the dataset used in the subsequent validation and analysis phase, but also in all the phases of the model development process. The dataset is considered in terms of both the quality of the data and the representativeness (coverage) with respect to a domain.

2.3.1. Data Quality

Data have an increasingly central role in all data-driven applications, and their quality is critical since it directly influences the reliability of all the downstream uses [52]. If the Data Quality rules are properly set, they mitigate the typical "garbage-in garbage-out" hazard of all data-driven applications and fast spread of wrong information in data centralized architecture. In the last decades, research on Data Quality has defined analysis dimensions and metrics to define and assess the quality of data. Data Quality identifies different characteristics of a dataset and presents quantitative measures of the corresponding quality dimensions. In the end, Data Quality quantifies and highlights the strengths and criticalities of a dataset. Over time, hundreds of different data quality dimensions were defined, each quantifying a different quality aspect of the data [53]. A DE that hosts thousands of experimental and simulated data points must automatically ensure a certain quality of the repository by assessing the proper dimensions of data quality. Following the *fitness for use* concept [53], a DE for the development of data-driven models based on experimental data needs to consider completeness, consistency, and accuracy as data quality dimensions since they are the most widely used across different domains and provide a good assessment of the quality of data products.

Completeness. Completeness measures how much mandatory information is missing in a database. To ensure this data quality aspect, it is sufficient to define the mandatory database fields as a set of rules. An example is the measurement unit of a quantity.

Consistency. Consistency, through the definition of a list of rules, quantifies whether the information stored in different parts of the database, but semantically connected to each other and regarding the same experiment, is congruent. Given the type of a measured property and the unit of measurement stored in two different database fields regarding the same experiment, an example of a consistency rule is the plausibility of the unit of measurement regarding the

reported property. In other words, if the type of the measured property is "pressure", possible units of measurement are "atm", "Pa" (Pascal), "bar", etc., but not "K" (Kelvin), for instance.

Accuracy. Accuracy is related to the precision of the data in representing real world values. Given a ground truth, accuracy measures the discrepancy between the value reported in the database with respect to the real one. Following the previous example, a bunch of valid units is plausible for "pressure", but only one value is correct for a measured value given the unit of measurement. However, accuracy is also strictly connected to experimental uncertainty (that, unfortunately, is not always provided together with experimental data [54]). Measuring accuracy is not an easy task since a ground truth is needed for its evaluation. For numerical values, the accuracy is determined using difference data sources and thresholds. An exhaustive discussion on how this dimension could be quantified is out of the scope of this work; as an example, Section 3.1 discusses more in detail how the accuracy is handled for a case study. In our scenario, the concept of accuracy is also related to the Data Quality dimension of consistency (or agreement) of different experiments concerning the same (or similar) experimental observation. Therefore, evaluating the consistency between different experiments regarding the same condition can be reduced to their accuracy evaluation.

2.3.2. Database Coverage

The reliability of the predictive model validation and analysis does not depend only on the quantity and quality of the experimental data. It is essential to be aware of the diversity of the data involved in terms of coverage of the domain that the model aims to represent. This discipline is known as database coverage (or diversity) [47]. For instance, having many experiments to be used for validation, all representing the same portion of the domain, may not be enough to establish whether the result of the model validation is sufficiently reliable. In fact, if a model is not validated in many different environmental conditions, its performance may

worsen unexpectedly when used to predict an unexplored (untested) but physically relevant portion of the domain. Generally speaking, not only in chemical engineering, a lack of adequate coverage in the dataset, thus a non extensive testing of the generality capabilities of a model, can result in a biased reliability of the model validation [55]. Therefore, it is important to have a collection of experiments, the largest and most diverse as possible against which the model can be validated.

In the literature [55-57], the most common techniques for quantifying the coverage of a database concerning a domain share a similar set-up phase and then differ in the method of calculating such coverage. The first step of the set-up phase entails the identification of the dimensions that define a domain. Subsequently, for each dimension (or axis), the possible values that can be assumed are specified, discretizing them in the case of continuous numerical values, or dividing them into categories in the case of literal (or categorical) values. After that, having defined the dimensions, a corresponding multidimensional matrix M is constructed [56]. This solution allows gathering different levels of granularity about the database coverage as needed in the analysis. Leveraging *bucketization*, the matrix is populated with the number of experiments available in a given region of the domain that corresponds to one or more cells (or boxes) of the multidimensional matrix. Bucketization consists in having all similar values in the same bucket. For example, if a dimension has been split into buckets with values 0,5,10,15, a data with the value for that dimension of 2 is associated with bucket 0, 3 with 5, 8 and 12 with bucket 10. It is worth mentioning that there is no constraint on how the buckets are defined, for example, whether they should follow a linear division in the case of a numerical axis. The buckets definition is domain-dependent. Then, each entry of the database can be associated to a cell of the matrix M, i.e., a sub-portion of the domain. Finally, the coverage of the database can be measured, given a threshold t as the ratio between the number of different cells that have at least t associated experiments or data $|cells(t)|$, over the total

amount of cells in the matrix M, as in Equation (1). Therefore, the database coverage definition is resilient to multiple experiments associated with the same

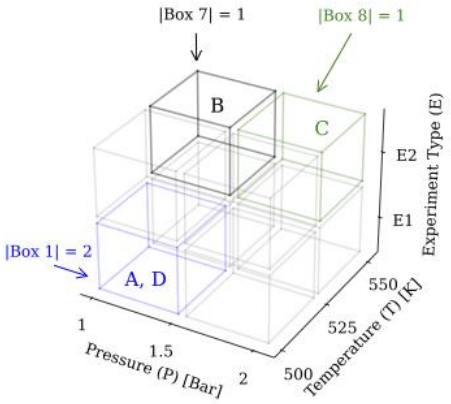


Figure 3: An example of the coverage of a database is computed: first the domain is divided in blocks according to dimensions that are defined by the metadata (properties) of the experiments, then, using bucketization, each experiment is associated to a block. The coverage $C(k)$ is the percentage of blocks that have at least k associated experiments.

bucket that could not bring any additional information in terms of the extensiveness of the model testing.

$$C(t) = \frac{|cells(t)|}{|M|} \quad (1)$$

One or multiple multidimensional matrices can be used to represent the diversity of a database according to different situations.

Experiment	T [K]	P [bar]	E	Box
A	508	1.2	E1	1
B	540	1.2	E2	7
C	537	1.1	E2	8
D	520	1.3	E1	1

Table 2: The experimental data set used for the running example in Figure 3.

Box	[T, P, E]	Cardinality
1	[(500, 525), (1.0, 1.5), E1]	2
2	[(500, 525), (1.5, 2.0), E1]	0
3	[(500, 525), (1.0, 1.5), E2]	0
4	[(500, 525), (1.5, 2.0), E2]	0
5	[(525, 550), (1.0, 1.5), E1]	0
6	[(525, 550), (1.5, 2.0), E1]	0
7	[(525, 550), (1.0, 1.5), E2]	1
8	[(525, 550), (1.5, 2.0), E2]	1

Table 3: The results of bucketization for the running example in Figure 3 using as data set Table 2.

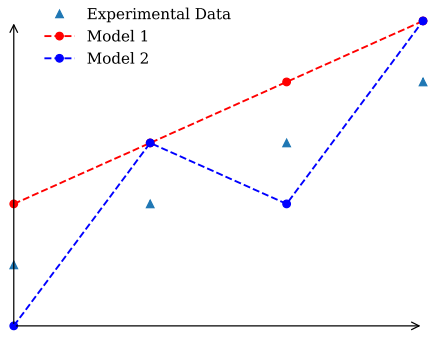
In Figure 3, a "small world" example is represented, where the dimensions are temperature (T), pressure (P), and experiment type (E) that mark the diversity of the dataset. In this scenario, each experiment in the dataset is characterized by a value property (dimension) that will place it in a specific position of the domain. These dimensions range respectively from 500K to 550K, 1 bar to 2 bar, and the possible experiment types are 'E1' and 'E2'. In this case, each numerical dimension is divided equally into two buckets, but this is not mandatory in the general case. Instead, the categorical properties (as 'experimental type' in the example) define themselves the number of buckets. As we can see from Figure 3, this configuration determines eight cells of the matrix (or boxes) that partition the "small word" domain. Table 2 reports the experimental data set used, containing four experiments, each with its own features in terms of temperature, pressure, and experiment type. In a second moment, using bucketization, each experiment is associated with a box (Table 3). Finally, the number of boxes with at least 1 associated experiment is 3, therefore the coverage index in this case $C(1) = 3/8 \sim 38\%$, and if the threshold is 2 the coverage index will therefore be $C(2) = 1/8 \sim 13\%$.

2.4. Model validation techniques

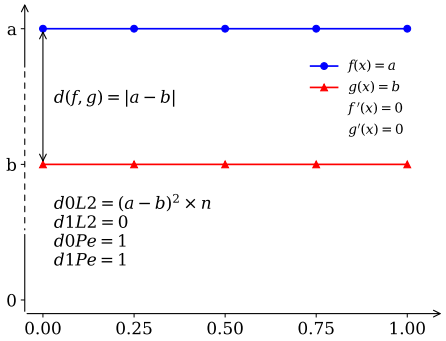
Qualitative and quantitative are the two macro families of model validation techniques used in the literature to compare experiments to simulations. Some of these techniques rely on *Cubic spline interpolation* to derive a continuous function from a discrete data set like in parametric experimental measurements. Cubic spline interpolation defines piecewise function using third-order polynomials, which pass through the given set of data points [58].

Visualization is a subjective, and thus qualitative, comparison of the experiments against the simulated data. The users evaluate, based on their expertise, the predictive model performance without quantifying the prediction quality. Moreover, different experts could have dissimilar opinions on the same experimental and simulated pair comparison.

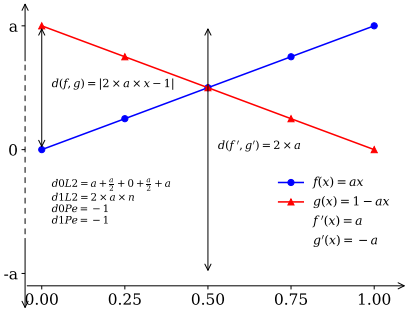
Point-wise approaches define a set of score functions to measure quantitatively the similarity between the experimental and simulated data-set evaluating the error point by point. This approach requires that both the datasets are defined over the same points on the x-axis. To overcome this assumption, in the general case, usually a spline is computed for the simulated dataset, and the error is computed over a set of x-values defined in the experimental dataset. These approaches are fast to compute, but they miss the fact that the points are not stand-alone but belong to a set, a chemical-physical trend of measurements. So, even if the score is quite high, the trend between the two datasets could be quite different; thus, the approach is objective, automatable, but misleading. In Figure 4a an example of this pitfall: even if the trend of the simulated data point of *Model 1* is quite different from *Model 2*, the point-wise error of the models when computed against the experimental data is the same. In such a family of scores, one of the most frequently used is the following function.



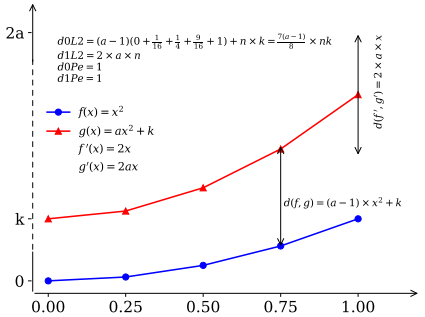
(a) The pitfall of using point-wise approaches to compute the similarity be- tween the models and the experimental data. This kind of similarity functions could report the same score for the two models even if the trend is different.



(b) Curve matching of f and g functions that differ for a vertical translation |a−b|. This setting can be understood a posteriori by looking at the curve matching indexes since the $d^0_{L2} \neq 0$ and $d^1_{L2} = 0$.



(c) Curve matching of two symmetric monotone functions. Except for the sign, they share the same first derivative, in fact $d^0_{Pe} = -1$ and $d^1_{Pe} = -1$.



(d) Curve matching of two functions that differ only for a multiplicative factor. In fact d^0_{Pe}=1 and d^1_{Pe}=1.

Figure 4. Pitfall of the point-wise approaches (Figure 4a), and same explanatory examples of curve matching between two functions (Figures 4b to 4d).

Definition 2.1 (**Sum Squared Error (SSE)).** SSE computes the sum of the squared difference between the experimental f and the simulated g data-points.

$$SSE = \sum (f(x_i) - g(x_i))^2 \tag{2}$$

Similar definitions are provided for other score functions such as Mean Squared Error (MSE), Mean Absolute Error (MAE), and Root-Mean-Square Error (RMSE).

Curve Matching (CM) is a quantitative *trend approach*, that overcomes the limitations of the point-wise approaches, accounting also for the fact that each data point is a part of the trend. CM measures the similarity of two functions f and g with a score ∈ [0,1], where 1 is the perfect similarity, after normalization.

A detailed description of the CM definition is available in the work by Pelucchi et al. [33].

Given the following definitions:

- F and G, i.e., the continuous curves generated using the cubic spline interpolation, representing experiment and model data points, respectively, and F' and G' their derivatives
- D, i.e. the intersection of the domains of F and G
- $\|h\|$, i.e. the norm of a generic curve h in the L^2 space:

$$||h|| = \sqrt{\int_D h(x)^2\, dx} \qquad (3)$$

It is possible to define the following dissimilarity measurements.

Definition 2.2 (d^0_{L2})**.** It is a generalization of the SSE to the continuous case.

$$d^0_{L_2}(F,G) = \frac{1}{1+\frac{||F-G||}{D}} \in (0,1) \qquad (4)$$

Definition 2.3 (d^1_{L2})**.** It is the generalization of the SSE to the continuous case of the first derivative. If $F(x) = G(x)+k$, where $k \in \mathbb{R}$, then $d^1_{L2} = 0$. So d^1_{L2} is invariant to vertical translations, but quantify if the two functions have similar slope.

$$d^1_{L_2}(F,G) = \frac{1}{1+\frac{||F'-G'||}{D}} \in (0,1) \qquad (5)$$

Definition 2.4 (d^0_{Pe})**.** It is the Pearson correlation index that measures whether the trend of a function is in agreement or disagreement with the other. In other words, if $F(x) = G(x)*k+a$, where $k,a \in \mathbb{R}$, then $d^0_{Pe} = 1$. The Pearson index,

in fact, is invariant to translation and dilatation.

$$d_{Pe}^{0}(F,G) = 1 - \frac{1}{2}\left|\left|\frac{F}{||F||} - \frac{G}{||G||}\right|\right| \in (0,1) \quad (6)$$

Definition 2.5 ($d^{1}{}_{Pe}$)**.** It is similar to Eq. 6, but on the first derivative of the functions.

$$d_{Pe}^{1}(F,G) = 1 - \frac{1}{2}\left|\left|\frac{F'}{||F'||} - \frac{G'}{||G'||}\right|\right| \in (0,1) \quad (7)$$

Definition 2.6 (S)**.** The shift S measures the dissimilarities in terms of horizontal shift between the two functions, as follows.

$$S = \max\left(1 - \frac{\delta}{D}, 0\right) \in (0,1) \quad (8)$$

Where δ is the horizontal shift between the two curves, obtained maximizing the sum of Eqs. (4) to (7):

$$\delta = \text{argmax}_{\delta}\left(d_{L_2}^{0} + d_{L_2}^{1} + d_{Pe}^{0} + d_{Pe}^{1}\right) \quad (9)$$

From a modeling point of view, using these indices has different advantages. The Pearson indexes and the SSE computed on both the function and the first derivative capture whether the model trend agrees or disagrees with the experimental data, while the SSE on the function still quantifies the difference point-to-point. The shift instead measures if the two functions are horizontally translated.

Therefore, CM is defined as the arithmetic average of five indexes, $d^{0}{}_{L2}$, $d^{1}{}_{L2}$, $d^{0}{}_{Pe}$, $d^{1}{}_{Pe}$, S, where S is weighted twice since it accounts both for the left and right horizontal shift.

Definition 2.7 (**Curve Matching (CM)).**

$$CM(f,g) = \frac{d_{L_2}^{0} + d_{L_2}^{1} + d_{Pe}^{0} + d_{Pe}^{1} + 2S}{6} \quad (10)$$

Figure 4b to 4d show examples curves' comparison using the same indexes used by CM without normalizing neither the values of the indexes or that of the curves. In all the examples, for simplicity, but without losing generality, the axes are adimensional and the x-values range from 0 to 1.

Curve Matching also accounts for experimental uncertainty, using a bootstrapping procedure [33]. If the uncertainty is not provided, Curve Matching uses a default uncertainty as suggested in the work of Olm et al. [28].

2.5. Model Analysis techniques

This section presents more details of the techniques used for the model analysis phase presented in Section 2.2.

The arithmetic mean, median, and standard deviation are mainly used as statistical indexes for this work. In addition, the Pearson correlation [59], the point-biserial [60] and the logistic regression [61] are used when it is needed to correlate two variables that could be continuous or categorical. All correlation indexes range from -1 to 1, where 1 indicates two closely and positively correlated variables.

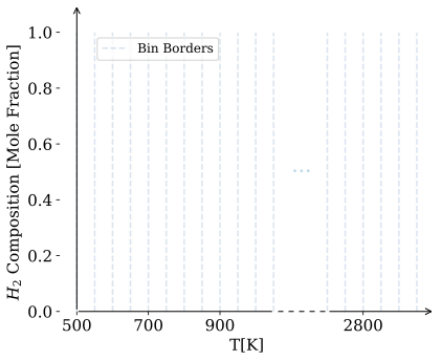
Data mining is a field of data science that applies a series of techniques to extract hidden features from large quantities of data. In particular, pattern detection or recognition is the process of discovering patterns and regularities in the data. Clustering is a typical unsupervised machine learning algorithm that allows examining a collection of data and, given a measure of distance, groups them into clusters based on their similarity. Once the data are organized in clusters, it is possible to analyze their common features and understand the pattern, the discriminant that has brought the data together. The clustering algorithms can be divided into two large classes. The main difference is that the first class of algorithms, known as *hierarchical*, start from the definition of a cluster for each point and then gradually merge the clusters until a stop criterion is reached. On the other hand, the second class of algorithms start from a predetermined number of clusters, and assign the other points to them based on

their similarity. In this work, Affinity Propagation [62] is used as a hierarchical clustering algorithm. Affinity Propagation selects a number of samples from the dataset as representatives of all the others. The algorithm exchanges a message between pairs of samples to determine which one is suitable to represent the other one. Representatives are continuously selected until convergence, at which point the final clusters are given. Affinity Propagation, by definition, establishes the number of clusters based on the data provided. However, two parameters need to be set: the preference, which controls how many exemplars are used, and the damping factor, which controls the message flow, damping some of them to avoid numerical oscillations.

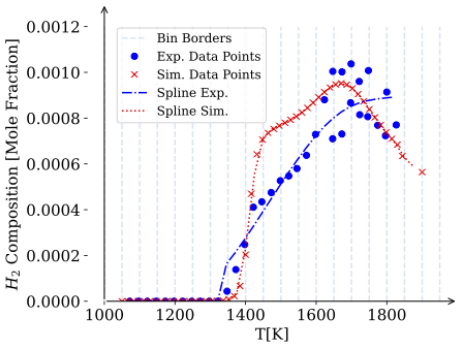
CM, working with a large number of data, provides a synthetic score about how good a model is. However, this synthetic result hides the detailed behavior of a predictive model. Instead, interval analysis, given a set of experimental and simulated data, computes the error of the model in predicting specific targets, in terms of quantitative overestimation and underestimation, in different ranges of a physical property (e.g., temperature). The basic idea of dividing the physical domain into intervals for different purposes has been used several times in the literature. However, either they use a point-wise similarity score to assess the model performance in an interval [27], or, leveraging the concept of data consistency and constraint definition [63, 64], they identify a region in the domain called ”feasible set” in which a model can be generated and optimized [65, 66]. Interval analysis, instead, uses a trend similarity score and measures the model performance in each interval for model validation purposes. In other words, curve matching summarizes the similarity between two curves, while the interval analysis maintains the axial dimension and quantifies the overestimation or underestimation of one curve with respect to the other. The disadvantage of maintaining a physical dimension comes with the curse of dimensionality. However, in the procedure proposed in this paper, this algorithm is used as the last step. The previous analyses have identified the single physical dimension and

group of experiments (in terms of common features) that significantly impact the model performance.

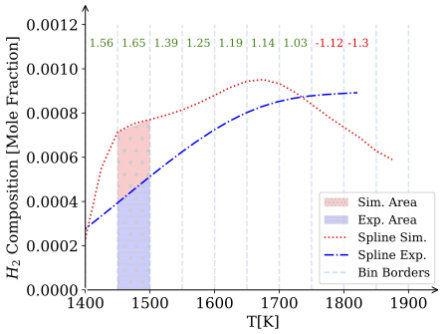
Given as input a set of experimental and simulated data pairs having the same variable on the abscissa and ordinate axes as input, the interval analysis algorithm is divided into four phases. Firstly (Figure 5a), given the independent variable operative domain, it is divided into *n* parts. The division could be equally distributed or not. For example, if the independent variable is the temperature and has an operating domain from 500K to 2500K, this dimension can be divided into 200 sectors or *bins*, each of 10K, such as (500K, 510K), (510K, 520K), and so on until the last one (2490K, 2500K). Secondly (Figure 5b), for each pair of experimental and simulated data, their corresponding splines are generated. Subsequently (Figure 5c), for each bin in which the experimental and simulated splines are defined, the area underlying the sub-portion of the domain delimited by the bin's ends is calculated. Then the ratio of the two areas is calculated and stored, providing a precise quantification of overestimation or underestimation of the simulated data concerning the experimental data. Finally (Figure 5d), once each pair is analyzed, following the previously described procedure, the model behavior can be summarized by averaging the ratios for each bin, distinguishing for each case whether it is an underestimation or an overestimation. The result of such analysis provides punctual information about the model behavior as the value on the x-axis changes.



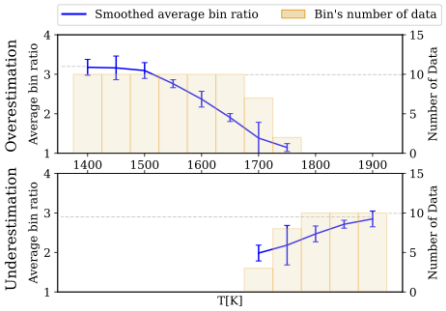
(a) Step 1. Division of the x-axis dimension (Temperature, in this case) into bins, while the y-axis measures a given property (H_2 composition, in this case)



(b) Step 2. Starting from the experimental and simulated data points, the experimental and simulated splines are generated.



(c) Step 3. Zoom-in on a portion of the x- axis of the example in Step 2. The ratio between the experimental and simulated area under the curve for each bin is computed. The simulation overestimates (in green) the experimental data if the ratio is bigger than one. Otherwise, it is underestimating (in red). Steps 2 and 3 are repeated for every available pair (in this case, T vs. H_2) in the database.



(d) Step 4. All the results of Step 3 are aggregated in order to know, on average, the amount of overestimation or underestimation in each bin. Error bars represent the standard deviation from the mean value. Bar plot represents the number of available pairs in a bin. In this case, the model tends to overestimate by a factor 3 in the low temperature, while at higher temperature underestimates.

Figure 5. The four steps of the interval analysis procedure.

3. Results and discussion

The methodology presented in Section 2.2 was applied to analyze and validate a detailed kinetic model for combustion applications. Generally speaking, a detailed kinetic model is an ensemble of thousands of reactions that describe the overall chemical conversion of a specific fuel into the final products. In particular,

665 the model describes the formation and consumption of each species represented by means of mathematical equations involving Arrhenius rate parameters. Moreover, since combustion experiments can be carried out in a wide range of conditions (i.e., operating conditions, geometry of the equipment, pressure and temperature ranges spanned, and so on), the model must be able to predict the
670 combustion evolution in such a variety of conditions.

The overall development procedure of such kind of models follows two main principles: *hierarchy* and *modularity*. *Hierarchy* means that the simplest subsystems must be included in all the more complex ones. *Modularity* leverages the already defined basic elementary steps to define more complex fuels [67].

675 The design of a reaction mechanism can vary a lot depending on the elementary steps taken into account by different research groups [68-71]. This aspect, coupled with the ever increasing number of experimental data, led to the scope of this work applied to combustion kinetics.

Therefore, this case study focuses on the critical aspects of the kinetic
680 modeling activity for combustion applications validating and analyzing the latest model (hereafter referred to as CRECK 2100 release) developed by CRECK modeling group [72]. All the numerical simulations were carried out using the OpenSMOKE++ framework as a numerical solver [73]. The CRECK 2100 release hereinafter presented consists of 365 species and 11,887 reactions which
685 describe through different merged sub-models the chemical evolution of different fuels, from H_2, with no C-atoms, to complex Polycyclic Aromatic Hydrocarbons (PAHs), with up to 20 C-atoms.

3.1. Data assesment case study

This case study used SciExpeM[1], a freely available data ecosystem with a
690 micro-service structure to manage scientific and simulated data together with predictive models and analysis results. The purpose of SciExpeM is to offer

[1] https://sciexpem.polimi.it

different data collection, management, and analysis services through a REST Application Program Interface (API) [1] that makes all these functionalities programming language-independent and versatile for many different uses, combining or integrating them in other systems, according to the user preferences.

The combustion modeling activity involves many quantities of interest expressed in different conditions. This domain is intrinsically complex since it is composed of many experimental structures. More generally, experimental data can be categorized based on the type of experiment that states which phenomena are of interest during the experimental measurement. Furthermore, only a precise set of reactors, i.e., the experimental facilities in which the measurements was carried out, are possible for each experiment type. Appendix C reports the reactor-experiment type association adopted within SciExpeM, which reflects the one proposed by Varga et al. [24]. Other categorical properties that characterize an experiment are the fuel and the target.

Ignition Delay Time measurements (IDTM) involve all the experiments types where the Ignition Delay Time, which is the time interval between the end of compression for Rapid Compression Machine (RCM) or the arrival of the reflected shock wave in Shock Tubes (ST), and the beginning of combustion, which is determined by pressure measurements or peak of CH and OH emission.

Speciation Measurement (SM) is the set of experimental activities in which inside ideal reactors such as shock Tube, jet stirred and plug flow reactors, are used to measure, at a specific experimental condition, the final concentration or the time evolution of mole fraction of a species.

Laminar Burning Velocity Measurement (LBVM) reports the laminar burning velocity (also referred as “laminar flame speed” or “speed”) of an experimental setting which is the velocity of a steady one-dimensional adiabatic free flame propagating in the doubly infinite domain. Usually, the laminar flame speed is

[1] https://pypi.org/project/SciExpeM-API/

720 studied against the equivalence ratio of the reacting mixture which is defined as the ratio of the fuel-to-oxidizer ratio to the stoichiometric fuel-to-oxidizer ratio, denoted as ϕ. It is important to highlight that the Laminar Burning Velocity is never measured directly, but it is derived from other measurements such as flame speed or inlet gas velocity.

725 For this work, a subset of the experimental data collected available in Sci-ExpeM was used. Specifically, 438 experimental datasets containing more than 10,000 data points. This collection of experiments involves ignition delay times, outlet concentration measurements, concentration time profile measurements and laminar burning velocity measurements. Appendix A shows a complete and 730 detailed overview of the experiments used, including type of experiment, type of reactor, and starting fuel.

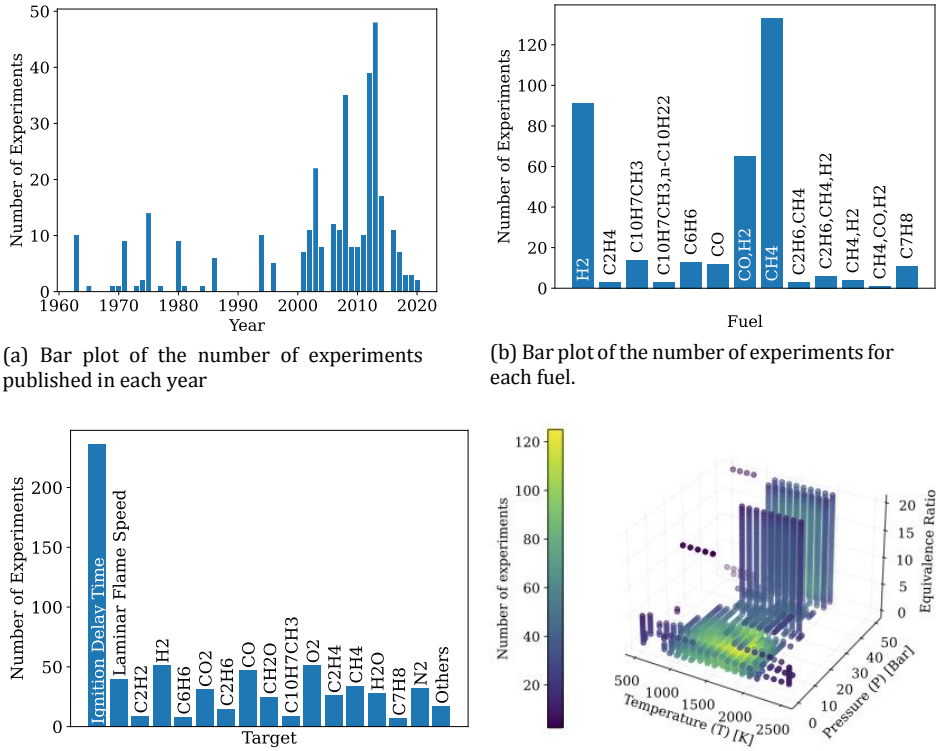
The data quality dimensions as presented in Section 2.3 are ensured by using SciExpeM not accepting data that does not meet the rules of completeness, consistency, and accuracy. In particular, regarding experimental uncertainty, if 735 available, it is accounted with bootstrapping procedures [74, 33]. In addition, SciExpeM implements two automatic strategies. First, its data management system can verify if multiple experiments in the same conditions are present in the database; thus, the uncertainty can be estimated as the standard deviation of the reported measurements [75, 76], and verify the consistency between different 740 experimental observations. Second, to detect evident outliers, but also in the case of no duplicated observations or to disambiguate inconsistent experiments, it leverages the idea that experimental data are used to validate a predictive model, but also a predictive model can be used to validate the experimental data [77]. A significant discrepancy, i.e., a similarity index below the first quartile of similarity 745 indexes computed with the model, between the experimental and simulated data suggests an unreliable experiment or model. Multiple models developed by different research groups can be used to repeat the same procedure and disambiguate whether the experiment is unreliable or not.

In combustion kinetics the physical properties that characterise an experiment are Temperature, Pressure and Equivalence ratio [78], therefore, they define the dimensions of the coverage index. The coverage of the collection of experiment used during this case study is $C(1) \sim 40\%$, $C(2) \sim 33\%$ as shown in Figure 6d (for more details see Appendix D). The coverage index is not exceptionally high, but some considerations about this data assessment procedure have to be made. During the computation of the database coverage, in a real-world domain, not all possible combinations of properties are meaningful or experimental data are available. Moreover, the coverage index depends on the setting of the algorithm, such as the number and distribution of the buckets. However, providing all this information properly creates awareness and trustworthiness for the model end-user. In addition, if the settings are standardized, then all the model release performances will be easily comparable.

Figure 6 shows some characteristics of the experimental data used. The majority of the experiments are recent (Figure 6a) and mainly involve $H_2, CH_4, H_2 + CO$ as fuels (Figure 6b). The main measured quantities (i.e., subjects or targets) are ignition delay time, laminar flame speed, and mole fraction of H_2, O_2, H_2O, CO (Figure 6c).

3.2. Model Validation case study

Model validation, as presented in Section 2.2, was carried out by comparing the experimental and simulated data through the calculated Curve Matching (CM) indexes as similarity score. In the present case study, the CM indexes of 997 experimental-simulated pairs belonging to 438 experiments were computed. Table 4 reports the five CM indices of the predictions of the CRECK 2100 release along with the *Score* (also referred with "general score"), which represents the global performance index computed following Definition 2.7.



(a) Bar plot of the number of experiments published in each year

(b) Bar plot of the number of experiments for each fuel.

(c) Bar plot of the number of experiments for each target. Each specie name at the top of the bins is intended to be representative of the mole fraction of the specie.

(d) Visualization of the database coverage of the employed experiments collection.

Figure 6

In general, the results show a good agreement between the experimental and simulated data, with an overall score higher than 0.8. In particular, the d^0_{L2} and d^1_{L2} represent the adjacency of the model simulation profiles and their first derivative with respect to the corresponding experimental ones. In kinetic models, discrepancies of such curves representing, for instance, concentration profiles, suggest inaccurate or wrong activation energy of reactions governing the chemical evolution of related species. On the other hand, the low value on d^0_{Pe} and d^1_{Pe} suggests a mismatch in the estimation of intermediate reaction products.

Finally, the *Shift* index estimates the horizontal misalignments with the experimental data suggesting a revision of the reaction rate or a not accounted experimental uncertainty among the x-axis.

	Score	$d^0_{L_2}$	$d^1_{L_2}$	d^0_{Pe}	d^1_{Pe}	Shift
Average	0.84	0.88	0.83	0.88	0.76	0.82
Median	0.87	0.98	0.97	0.92	0.79	0.88
Min	0.34	0.07	0.00	0.01	0.02	0.00
Max	0.97	0.99	0.99	0.98	0.99	0.99
St. Dev	0.11	0.17	0.22	0.16	0.13	0.17
Variance	0.01	0.03	0.04	0.02	0.01	0.03
P_{25}	0.79	0.80	0.79	0.82	0.78	0.78
P_{75}	0.92	0.98	0.98	0.97	0.96	0.96

Table 4: Curve Matching indixes and global score for the 993 experimental-simulated pair for the CRECK model 2100 release. In the table are illustrated their minimum (min), maximum (max), average, median (P_{50}), standard deviation (St. Dev.), variance, and the 25^{th},50^{th},75^{th} percentile (P_{25},P_{50},P_{75}, respectively).

3.3. Model analysis case study

3.3.1. Statistical analysis

As stated previously, a typical experiment of combustion study is characterized by the following properties: experiment type, reactor, fuel, and target. Curve Matching general scores range from 0.34 to 0.97, with the first quartile P_{25} = 0.79, i.e., the scores of all the worst performing model simulations is < 0.79, as shown in Table 4. Therefore, the 997 CM scores are filtered accordingly to the P_{25}. With such collection of pairs (and related experiments), it is possible to identify the combinations among the properties of the available experiments that determine the worst model performance. In the present case study, all the ~ 20,000 possible combinations are evaluated, while in Appendix A only those with

at least three cases (of experimental-simulated pairs, i.e., corresponding to CM scores) are reported. As the table suggests, the most numerous cases regard *IDTM* as experiment type, and fuels H_2, $H_2 + CO$, or CH_4. Moreover, the percentage of such cases in the P_{25} percentile with respect with the total number of existing ones ($\%\frac{\#P_{25}}{\#Total}$) is particularly high: 45%, 37%, and 62% respectively. At the same time, although there is no statistically significant combination of the whole experiment properties in Appendix A, there are several cases with fuel C_6H_6 whose model performances are not satisfactory, also confirmed by the tables in Appendix A. These evidences, therefore, highlight the areas of intervention of the model to improve the predictive capabilities.

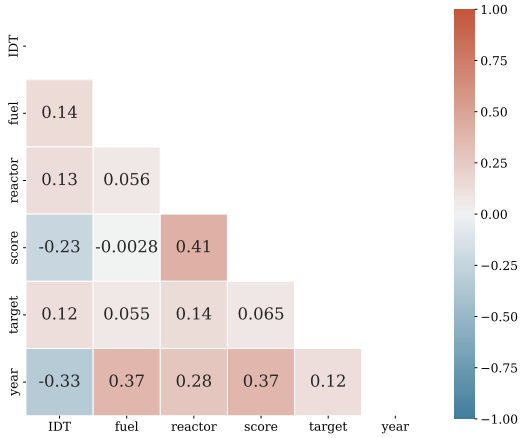


Figure 7: Correlation matrix between numerical and categorical property of the experiments. The correlation matrix is computed within the similarity indexes within the first quartile.

Correlating the properties of the experiments and the global performance index of Curve Matching allows identifying contributing factors to better understand the model behavior. Since from the previous analysis the ignition delay measurements are an improvable aspect of the model, in this phase, the experiment metadata were enriched with a new property to establish which species was used to detect the ignition. Figure 7 shows the correlation matrix between the experiment metadata such as reactor, experiment type, fuel, target,

experiment publication year, type of IDT, and the CM score of each experimental-simulated data pair. It is immediate to notice the interdependence between the Ignition Delay Time types, the year, and the similarity score. The moderate correlation between the score and the year suggests that the "oldest" experimental data are the ones that show a larger discrepancy to the model, probably due to the lower accuracy of the measured value with older instruments. Moreover, the moderate correlation between the reactor and score suggests that the prediction of some reactor is worse than for others. This fact is supported by the link between the IDT type and the score that implies that there are species to detect the ignition on which the model does not perform well. A further investigation revealed that inside the older experiments the ignition delay time was defined and therefore computed as the minimum baseline intercept of CO_2, this probably led to more inaccurate measurements.

Cluster	# Exp.	T[K]	P[Bar]	Eq. Ratio [-]	CM Score
A	10	1466-1700	2-15	1-2	0.77
B	10	1165-1383	3-31	1-1	0.72
C	7	1774-1956	0-20	0-2	0.497
D	9	1786-1995	1-15	0-6	0.593
E	6	1756-1800	7-11	1-4	0.582

Table 5: Results of the clustering algorithm. In the table the five most statistically representative clusters are reported.

3.3.2. Pattern Detection

In this scenario, the clustering algorithm is applied to the experimental-simulated pairs belonging to the first quartile. The algorithm clusters the data based on the similarity score and the physical properties that collocate an experiment in the domain. In combustion kinetics, the physical properties that characterize and experiment are temperature, pressure, and equivalence ratio. This analysis considers only the most statistically relevant cluster, i.e., the cluster

with at least six associated experiments with low variance in terms of their similarity scores. Consequently, it is ensured that a cluster contains the pairs in which the model performs similarly, and the experiments are physically located closely in the domain.

Clustering results are summarized in Table 5. The first two columns report respectively the five most statistically relevant clusters and the number of experiments belonging to a specific cluster. Moreover, the table shows the average ranges of temperature, pressure, and equivalence ratio of the experiments that belong to a cluster and the arithmetic average of the global performance index computed for each experimental-simulated pair belonging to the same cluster. By looking at the Table 5, it is clear that generally, the experiments at a high temperature (approximately from 1,700 K to 2,000 K, i.e., Cluster C, D, E) are predicted less accurately. Looking at the values for the pressure, no particular trend is observed since several values for the pressure are covered by the experiments with very different performances. Therefore, a variation in the pressure is not a discriminant factor for the model performance. From what concerns the equivalence ratio instead, it is possible to observe a pattern with the temperature. The performances at a low equivalence ratio and high temperature (Cluster A) are worse than at a similar equivalence ratio but at lower temperatures (Cluster B); thus, the temperature has a critical impact on such a range of equivalence ratio. In summary, inside the first quartile, the worst performances are associated with the experiments at elevated temperatures. However, if we consider the entire set of experiments on which the analysis is carried on, it is not true. These results would require further investigation with appropriate, domain-specific tools (Rate Of Production Analysis, Sensitivity Analysis, Elemental Flux Analysis) to understand the real source of deviation in model components.

3.3.3. *Interval Analysis*

Finally, the Interval Analysis was applied to the collection of experiments regarding the ignition delay times pre-selected by the first quartile previously analyzed when the fuels is CH_4 (see Appendix A). Figure 8 reports a possible result of interval analysis. Inside the plot, two different curves are reported: the one at the top accounting for the overestimation, the other for the underestimation of the model with respect to the experimental data. Moreover, in the background, the histogram accounts for the number of experimental datasets inside each interval. It can be clearly observed an underestimation by a factor of up to 3 at elevated temperatures ($T > 1650K$), while at lower temperatures the model performs accurately, i.e., within the typical experimental uncertainty of 10% [28]. This result is congruent with the results of the previous analyses.

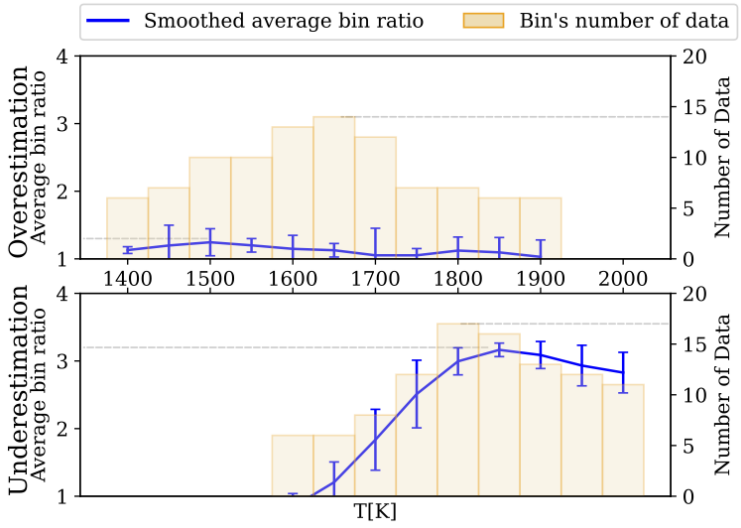


Figure 8: Results of the Interval Analysis when applied to a set of IDT experiments selected by the first quartile when the fuel is methane.

4. Concluding remarks

This work proposes an end-to-end validation and analysis methodology supported by an integrated set of tools to assess and comprehend the model performance. It leverages the increasing amount of data and (hidden) information to improve the predictive model further. Therefore, this paper proposes a model evaluation technique that combines methodologies and technologies from big data, data science, and data management fields. The result is a three-phased systematic, objective, and automatic procedure: (i) Following the data ecosystem concept, it is necessary to develop a data management system that facilitates the sharing of scientific data between researchers, the reuse of resources, while ensuring a certain data quality level and diversity of the repository to mitigate as much as possible the risk of delivering a model based on unreliable data or overfitted; (ii) Using numerical methods, abandoning the typical visual, subjective, and error-prone data comparison, the model validation is conducted, and the results are presented as a synthetic performance overview to the model developer; (iii) Applying data science and developed ad-hoc techniques such as interval analysis, it is possible to leverage the model validation results to analyze the model behavior in many different conditions, providing suggestions in terms of where, why, how much, and in which situations of the domain the model needs to be improved. Finally, the methodology is applied to a combustion kinetic model release to demonstrate the opportunities and effectiveness of such an approach.

Future work is related to leverage such information to build a predictive model starting from the domain variables intrinsic relationships present in the data laying down the foundations for a model development process based on artificial intelligence. Moreover, thanks to the use of such a data ecosystem as a repository of information, future work will regard the estimation of experimental uncertainty, leveraging multiple model predictions as a “ground truth” and

experiments in similar conditions present in the database to evaluate their accuracy using both statistical analysis and machine learning.

Appendix A. Curve Matching Result

Exp. Type	*Reactor*	*Fuel*	*Target*	$\#\mathbf{P_{25}}$	*Average*	*Min*	*Max*	*Median*	*St.Dev*	*Var.*	$\%\frac{\#P_{25}}{\#Total}$
Ignition Delay Time Measurement	*Shock Tube*	$H_2 + CO$	*Ignition Delay Time*	21	0.74	0.59	0.79	0.74	0.05	0.00	37
		C_2H_4		3	0.68	0.59	0.74	0.72	0.08	0.01	100
		CH_4		62	0.61	0.35	0.79	0.66	0.14	0.02	62
		$C_{10}H_7CH_3$		4	0.77	0.74	0.78	0.77	0.02	0.00	80
		C_5H_6		3	0.56	0.52	0.62	0.53	0.05	0.00	100
		C_6H_6		4	0.65	0.55	0.73	0.65	0.08	0.01	40
		H_2		25	0.72	0.59	0.79	0.74	0.06	0.00	45
Speciation Measurement		C_6H_6	C_6H_6	5	0.56	0.34	0.70	0.61	0.14	0.02	100
	Plug Flow	C_6H_6	C_6H_6	3	0.72	0.68	0.76	0.71	0.04	0.00	100
			C_5H_6	3	0.66	0.62	0.73	0.64	0.06	0.00	100
			C_2H_2	3	0.73	0.69	0.76	0.75	0.04	0.00	100
			C_2H_4	3	0.73	0.69	0.76	0.75	0.04	0.00	100
	Jet Stirred	CH_4	C_2H_4	4	0.57	0.43	0.67	0.58	0.10	0.01	56
		$C_{10}H_7CH_3$	NC_6H_{12}	3	0.71	0.65	0.74	0.74	0.05	0.00	100
			$NC_{10}H_{22}$	5	0.76	0.75	0.77	0.76	0.01	0.00	71
		C_6H_6	C_3H_6	3	0.67	0.48	0.79	0.73	0.17	0.03	75
	Plug Flow	CH_4	O_2	3	0.76	0.72	0.78	0.78	0.04	0.00	23
	Shock Tube	C_6H_6	C_4H_2	3	0.67	0.60	0.78	0.64	0.10	0.01	100
LBVM	*Flame*	C_7H_8	*Speed*	6	0.64	0.51	0.70	0.68	0.08	0.01	100

Table A.6: The combination of properties of the experiments in terms of experiment type (Exp. Type), reactor, fuel, and target, whose CM scores (below P25) are at least 3 cases. The table shows the average, min, max, median, standard deviation (St. Dev.), and Variance (Var.) of the associated scores. Moreover, $\%\frac{\#P_{25}}{\#Total}$ accounts for the proportion of pairs with a particular combination of properties below P25 respect with the number of existing ones.

Appendix B. P_{25} Analysis

Fuel	*Average*	*Min*	*Max*	*Median*	*St.Dev*	*Var*	$\#\mathbf{P_{25}}$	$\%\ \mathbf{P_{25}}$	*#Total*	*% Total*	$\%\frac{\#P_{25}}{\#Total}$
$C_{10}H_7CH_3$	0.69	0.39	0.79	0.73	0.1	0.01	33	13	204	20	16
C_2H_4	0.68	0.59	0.74	0.72	0.08	0.01	3	1	3	0	100
C_5H_6	0.56	0.52	0.62	0.53	0.05	0.00	3	1	3	0	100
C_6H_6	0.7	0.34	0.79	0.73	0.09	0.01	58	24	165	17	35
C_7H_8	0.63	0.46	0.79	0.67	0.12	0.01	20	8	125	13	16
CH_4	0.62	0.35	0.79	0.66	0.14	0.02	78	32	207	21	38
CO	0.74	0.7	0.78	0.74	0.06	0.00	2	1	12	1	17
$H_2 + CO$	0.74	0.59	0.79	0.74	0.05	0.00	21	9	60	6	35
H_2	0.72	0.59	0.79	0.74	0.06	0.00	25	10	102	10	25

Table B.7: A detailed first quartile analysis for 8 fuels (out of the existing 16) with at least 2 cases. The table reports the average, min, max, median, standard deviation (St.Dev.), and Variance(Var.) of the CM scores with a particular fuel. Moreover, $\#P_{25}$ is the number of experimental-simulated pairs (or cases). $\%\ P_{25}$ is the corresponding percentage value. *#Total* accounts for how many pairs with a given fuel are existing in total (therefore, also with CM score above the P_{25}). *% Total* represents the corresponding percentage. $\%\frac{\#P_{25}}{\#Total}$ measures the proportion of pairs below P_{25} respect with the number of existing ones.

Target	*Average*	*Min*	*Max*	*Median*	*St.Dev*	*Var*	$\#\mathbf{P_{25}}$	$\%\ \mathbf{P_{25}}$	*#Total*	*% Total*	$\%\frac{\#P_{25}}{\#Total}$
C_2H_2	0.71	0.63	0.76	0.73	0.05	0.00	7	3	33	3	21
C_2H_4	0.65	0.43	0.76	0.67	0.11	0.01	9	4	52	5	17
C_2H_6	0.63	0.44	0.78	0.66	0.14	0.02	4	2	25	3	16
C_3H_6	0.67	0.48	0.79	0.73	0.17	0.03	3	1	11	1	27
C_4H_2	0.71	0.6	0.78	0.75	0.09	0.01	5	2	6	1	83
C_4H_4	0.77	0.75	0.78	0.78	0.02	0.00	3	1	14	1	21
C_4H_6	0.76	0.74	0.78	0.76	0.01	0.00	6	2	17	2	35
C_5H_6	0.66	0.62	0.73	0.64	0.06	0.00	3	1	14	1	21
C_6H_5OH	0.71	0.67	0.78	0.67	0.06	0.00	3	1	10	1	30
C_6H_6	0.64	0.34	0.79	0.67	0.13	0.02	10	4	30	3	33
CH_4	0.76	0.67	0.79	0.78	0.06	0.00	4	2	60	6	7
CO_2	0.71	0.64	0.77	0.71	0.07	0.01	4	2	58	6	7
CO	0.74	0.7	0.78	0.74	0.03	0.00	8	3	77	8	10
INDENE	0.72	0.68	0.77	0.72	0.04	0.00	3	1	8	1	38
$NC_{10}H_{22}$	0.76	0.75	0.77	0.76	0.01	0.00	5	2	7	1	71
NC_6H_{12}	0.71	0.65	0.74	0.74	0.05	0.00	3	1	3	0	100
O_2	0.76	0.72	0.78	0.78	0.03	0.00	5	2	44	4	11
Speed	0.64	0.51	0.7	0.68	0.08	0.01	6	2	42	4	14
IDT	0.66	0.35	0.79	0.7	0.13	0.02	126	51	248	25	51

Table B.8: A detailed first quartile analysis for 19 targets (out of the existing 48) with at least 3 cases. The table reports the average, min, max, median, standard deviation (St.Dev.), and Variance(Var.) of the CM scores with a particular target. Moreover, $\#P_{25}$ is the number of experimental-simulated pairs (or cases). $\%\ P_{25}$ is the corresponding percentage value. *#Total* accounts for how many pairs with a given target are existing in total (therefore, also with CM score above the P_{25}). *% Total* represents the corresponding percentage. $\%\frac{\#P_{25}}{\#Total}$ measures the proportion of pairs below P_{25} respect with the number of existing ones.

Appendix C. Experiment types and reactors

Exp.Type / Reactor	*Ignition DelayTime*	Speciation	Laminar Burning Velocity
Shock Tube	✓	✓	
Rapid Compression Machine	✓	✓	
Jet Stirred		✓	
Plug Flow		✓	
Premixed Laminar Flame		✓	✓

Table C.9: Possible reactors for each type of experiment according to Varga et al. [24].

Appendix D. Database Coverage

The database coverage was computed using the following buckets values.

- Temperature (K) T = {500,750,1000,1250,1500,2000,2250,2500}
- Pressure (bar) P = {0,1,2,5,10,25,50,75}
- Equivalence ratio ϕ = {0,0.25,0.75,1,2,5,10,50}

The selection of the buckets for each dimension has a direct impact on the coverage index. Since the bucket definition depends on each applicative domain, providing a general rule to define the buckets is challenging. For example, some combinations of properties value, and therefore, the corresponding "boxes" could not be physically admissible in a domain, and they will be empty, affecting the coverage index. A proper selection of the buckets' values mitigates this problem. Therefore, each scientific community in each chemical engineering sector should reach an agreement on the buckets definition. Doing so will make the predictive model performances immediately comparable. In the meanwhile, providing supplementary materials about the buckets definition with the model validation

results, it will make the user model aware of the extensiveness of the validation test, thus the reliability of the predictive model.

Acknowledgements

The authors thank Alberto Cuoci, Matteo Pelucchi, and Luna Pratali Maffei for the discussions during this work.

Fundings

The work of ER is supported by the interdisciplinarity Ph.D. project of Politecnico di Milano.

References

[1] C. Tenopir, E. D. Dalton, S. Allard, M. Frame, I. Pjesivac, B. Birch, D. Pollock, K. Dorsett, Changes in data sharing and data reuse practices and perceptions among scientists worldwide, PloS ONE 10 (8) (2015) e0134826.

[2] S. Raimondeau, D. Vlachos, Recent developments on multiscale, hierarchical modeling of chemical reactors, Chemical Engineering Journal 90 (1) (2002) 3–23, Catalytic Reaction and Reactor Engineering EuropaCat V Limerick, Sept 2-7 2001. doi:https://doi.org/10.1016/ S1385-8947(02)00065-7.

[3] S. Madanikashani, L. A. Vandewalle, S. De Meester, J. De Wilde, K. M. Van Geem, Multi-scale modeling of plastic waste gasification: Opportunities and challenges, Materials 15 (12). doi:10.3390/ma15124215.

[4] S. H. Rudy, S. L. Brunton, J. L. Proctor, J. N. Kutz, Data-driven discovery of partial differential equations, Science Advances 3 (4) (2017) e1602614.

[5] S. L. Brunton, J. L. Proctor, J. N. Kutz, Discovering governing equations from data by sparse identification of nonlinear dynamical systems, Proceedings of the National Academy of Sciences 113 (15) (2016) 3932–3937.

[6] J. Farrell, N. Cernansky, F. Dryer, D. G. Friend, C. Hergart, C. Law, R. Mc-David, C. Mueller, A. Patel, H. Pitsch, Development of an experimental database and kinetic models for surrogate diesel fuels, SAE technical paper No. 2007-01-0201 (2007).

[7] H. Wang, D. A. Sheen, Combustion kinetic model uncertainty quantification, propagation and minimization, Progress in Energy and Combustion Science 47 (2015) 1–31.

[8] J. N. Kutz, Data-driven modeling & scientific computation: methods for complex systems & big data, Oxford University Press, 2013.

[9] F. Farazi, J. Akroyd, S. Mosbach, P. Buerger, D. Nurkowski, M. Salamanca, M. Kraft, Ontokin: An ontology for chemical kinetic reaction mechanisms, Journal of Chemical Information and Modeling 60 (1) (2019) 108–120.

[10] H. Gossler, L. Maier, S. Angeli, S. Tischer, O. Deutschmann, Carmen: an improved computer-aided method for developing catalytic reaction mechanisms, Catalysts 9 (3) (2019) 227.

[11] C. Cappiello, A. Gal, M. Jarke, J. Rehof, Data ecosystems: Sovereign data exchange among organizations (dagstuhl seminar 19391), in: Dagstuhl Reports, Vol. 9:9, Schloss Dagstuhl-Leibniz-Zentrum fuer Informatik, 2020.

[12] M. Frenklach, Transforming data into knowledge—process informatics for combustion chemistry, Proceedings of the Combustion Institute 31 (1) (2007) 125–140.

[13] Hegde, Arun and Li, Wenyu and Oreluk, James and Packard, Andrew and Frenklach, Michael, Consistency analysis for massively inconsistent datasets in bound-to-bound data collaboration, SIAM/ASA Journal on Uncertainty Quantification 6 (2) (2018) 429–456.

[14] Feeley, Ryan and Seiler, Pete and Packard, Andrew and Frenklach, Michael, Consistency of a reaction dataset, The Journal of Physical Chemistry A 108 (44) (2004) 9573–9583.

[15] You, Xiaoqing and Packard, Andrew and Frenklach, Michael, Process informatics tools for predictive modeling: Hydrogen combustion, International Journal of Chemical Kinetics 44 (2) (2012) 101–116.

[16] Frenklach, Michael and Packard, Andrew and Seiler, Pete, Prediction uncertainty from models and data, in: Proceedings of the 2002 American Control Conference (IEEE Cat. No. CH37301), Vol. 5, IEEE, 2002, pp. 4135–4140.

[17] Li, Wenyu and Hegde, Arun and Oreluk, James and Packard, Andrew and Frenklach, Michael, Representing model discrepancy in bound-to-bound data collaboration, SIAM/ASA Journal on Uncertainty Quantification 9 (1) (2021) 231–259.

[18] Yeates, Devin R and Li, Wenjun and Westmoreland, Phillip R and Speight, William and Russi, Trent and Packard, Andrew and Frenklach, Michael, Integrated data-model analysis facilitated by an instrumental model, Proceedings of the Combustion Institute 35 (1) (2015) 597–*.

[19] Oreluk, James and Liu, Zhenyuan and Hegde, Arun and Li, Wenyu and Packard, Andrew and Frenklach, Michael and Zubarev, Dmitry, Diagnostics of data-driven models: uncertainty quantification of PM7 semi-empirical quantum chemical method, Scientific reports 8 (1) (2018) 1–12.

[20] Russi, Trent and Packard, Andy and Frenklach, Michael, Uncertainty quantification: Making predictions of complex reaction systems reliable, Chemical Physics Letters 499 (1-3) (2010) 1–8.

[21] Frenklach, Michael and Packard, Andrew and Garcia-Donato, Gonzalo and Paulo, Rui and Sacks, Jerome, Comparison of statistical and deterministic

frameworks of uncertainty quantification, SIAM/ASA Journal on Uncertainty Quantification 4 (1) (2016) 875–901.

[22] G. L. Goteng, N. Nettyam, S. M. Sarathy, Cloudflame: Cyberinfrastructure for combustion research, in: 2013 International Conference on Information Science and Cloud Computing Companion, IEEE, 2013, pp. 294–299.

[23] Reyno-Chiasson, Z., et al. CloudFlame and PrIMe: accelerating combustion research in the cloud. 9th In-ternational Conference on Chemical Kinetics, Ghent, Belgium.

[24] T. Varga, T. Turányi, E. Czinki, T. Furtenbacher, A. Császár, Respecth: a joint reaction kinetics, spectroscopy, and thermochemistry information system, in: Proceedings of the 7th European Combustion Meeting, Vol. 30, Citeseer, 2015, pp. 1–5.

[25] V. R. Lambert, R. H. West, Identification, correction, and comparison of detailed kinetic models, in: 9th US Natl Combust Meeting, Cincinnati, OH, 2015, pp. 1–8.

[26] N. J. Killingsworth, M. J. McNenly, R. A. Whitesides, S. W. Wagnon, Cloud based tool for analysis of chemical kinetic mechanisms, Combustion and Flame 221 (2020) 170–179.

[27] P. Zhang, I. G. Zsély, M. Papp, T. Nagy, T. Turányi, Comparison of methane combustion mechanisms using laminar burning velocity measurements, Combustion and Flame 238 (2022) 111867.

[28] Olm, C., Zsély, I. G., Pálvölgyi, R., Varga, T., Nagy, T., Curran, H. J., & Turányi, T. Comparison of the performance of several recent hydrogen combustion mechanisms, Combustion and Flame 161 (9) (2014) 2219–2234.

[29] D. Q. Gbadago, J. Moon, M. Kim, S. Hwang, A unified framework for the mathematical modelling, predictive analysis, and optimization of reaction systems using computational fluid dynamics, deep neural network and

genetic algorithm: A case of butadiene synthesis, Chemical Engineering Journal 409 (2021) 128163.

[30] J.-P. Simonin, On the comparison of pseudo-first order and pseudo-second order rate laws in the modeling of adsorption kinetics, Chemical Engineering Journal 300 (2016) 254–263.

[31] J. Fermoso, M. V. Gil, C. Pevida, J. Pis, F. Rubiera, Kinetic models comparison for non-isothermal steam gasification of coal–biomass blend chars, Chemical Engineering Journal 161 (1-2) (2010) 276–284.

[32] Kelly, Mark and Dooley, Stephen and Bourque, Gilles, Toward machine learned highly reduced kinetic models for methane/air combustion, in: Turbo Expo: Power for Land, Sea, and Air, Vol. 84942, American Society of Mechanical Engineers, 2021.

[33] M. Pelucchi, A. Stagni, T. Faravelli, Addressing the complexity of combustion kinetics: Data management and automatic model validation, in: Computer Aided Chemical Engineering, Vol. 45, Elsevier, 2019, pp. 763– 798.

[34] M. S. Bernardi, M. Pelucchi, A. Stagni, L. M. Sangalli, A. Cuoci, A. Frassoldati, P. Secchi, T. Faravelli, Curve matching, a generalized framework for models/experiments comparison: An application to n-heptane combustion kinetic mechanisms, Combustion and Flame 168 (2016) 186–203.

[35] T. Varga, C. Olm, A. Busai, I. G. Zsély, Respecth kinetics data format specification v2. 0 (2017).

[36] Frenklach, Michael and Packard, Andrew and Seiler, Pete and Feeley, Ryan, Collaborative data processing in developing predictive models of complex reaction systems, International journal of chemical kinetics 36 (1) (2004) 57–66.

[37] C. Allan, J.-M. Burel, J. Moore, C. Blackburn, M. Linkert, S. Loynton, D. MacDonald, W. J. Moore, C. Neves, A. Patterson, et al., Omero: flexible, model-

driven data management for experimental biology, Nature Methods 9 (3) (2012) 245–253.

[38] D. A. Beck, J. M. Carothers, V. R. Subramanian, J. Pfaendtner, Data science: Accelerating innovation and discovery in chemical engineering, AIChE Journal 62 (5) (2016) 1402–1416.

[39] F. Farazi, M. Salamanca, S. Mosbach, J. Akroyd, A. Eibeck, L. K. Aditya, A. Chadzynski, K. Pan, X. Zhou, S. Zhang, et al., Knowledge graph approach to combustion chemistry and interoperability, ACS Omega 5 (29) (2020) 18342–18348.

[40] N. Liu, J. Wang, S. Sun, C. Li, W. Tian, Optimized principal component analysis and multi-state bayesian network integrated method for chemical process monitoring and variable state prediction, Chemical Engineering Journal 430 (2022) 132617.

[41] S. Mittal, S. Pathak, H. Dhawan, S. Upadhyayula, A machine learning approach to improve ignition properties of high-ash indian coals by solvent extraction and coal blending, Chemical Engineering Journal 413 (2021) 127385.

[42] P. P. Plehiers, I. Lengyel, D. H. West, G. B. Marin, C. V. Stevens, K. M. Van Geem, Fast estimation of standard enthalpy of formation with chemical accuracy by artificial neural network correction of low-level-of-theory ab initio calculations, Chemical Engineering Journal 426 (2021) 131304.

[43] Y. Ouyang, L. A. Vandewalle, L. Chen, P. P. Plehiers, M. R. Dobbelaere, G. J. Heynderickx, G. B. Marin, K. M. Van Geem, Speeding up turbulent reactive flow simulation via a deep artificial neural network: A methodology study, Chemical Engineering Journal 429 (2022) 132442.

[44] F. H. Vermeire, W. H. Green, Transfer learning for solvation free energies: From quantum chemistry to experiments, Chemical Engineering Journal 418 (2021) 129307.

[45] X. Chen, L. G. Wang, F. Meng, Z.-H. Luo, Physics-informed deep learning for modelling particle aggregation and breakage processes, Chemical Engineering Journal 426 (2021) 131220.

[46] A. Shokry, S. Medina-Gonzalez, P. Baraldi, E. Zio, E. Moulines, A. Espunã, A machine learning-based methodology for multi-parametric solution of chemical processes operation optimization under uncertainty, Chemical Engineering Journal 425 (2021) 131632.

[47] M. Drosou, H. Jagadish, E. Pitoura, J. Stoyanovich, Diversity in big data: A review, Big Data 5 (2) (2017) 73–84.

[48] E. Ramalli, G. Scalia, B. Pernici, A. Stagni, A. Cuoci, T. Faravelli, Data ecosystems for scientific experiments: managing combustion experiments and simulation analyses in chemical engineering, Frontiers in Big Data 4 (2021) 1–19. doi:10.3389/fdata.2021.663410.

[49] Tomlin, Alison S and Turányi, Tamás, Investigation and improvement of reaction mechanisms using sensitivity analysis and optimization, in: Cleaner Combustion, Springer, 2013, pp. 411–445.

[50] M. D. Wilkinson, M. Dumontier, I. J. Aalbersberg, G. Appleton, M. Axton, A. Baak, N. Blomberg, J.-W. Boiten, L. B. da Silva Santos, P. E. Bourne, et al., The FAIR guiding principles for scientific data management and stewardship, Scientific Data 3 (1) (2016) 1–9.

[51] J. Leskovec, A. Rajaraman, J. D. Ullman, Mining of massive data sets, Cambridge University Press, 2020.

[52] S. Garc´ıa, J. Luengo, F. Herrera, Data preprocessing in data mining, Vol. 72, Springer, 2015.

[53] R. Y. Wang, D. M. Strong, Beyond accuracy: What data quality means to data consumers, Journal of Management Information Systems 12 (4)

(1996) 5–33.

[54] W. Dai, S. Cremaschi, H. J. Subramani, H. Gao, Estimation of data uncertainty in the absence of replicate experiments, Chemical Engineering Research and Design 147 (2019) 187–199.

[55] A. Asudeh, Z. Jin, H. Jagadish, Assessing and remedying coverage for a given dataset, in: 2019 IEEE 35th International Conference on Data Engineering (ICDE), IEEE, 2019, pp. 554–565.

[56] E. Ramalli, B. Pernici, Know your experiments: interpreting categories of experimental data and their coverage, in: SeaData at VLDB 2021, CEUR Workshop Proceedings, 2021, pp. 27–33.

[57] Y. Lin, Y. Guan, A. Asudeh, H. Jagadish, Identifying insufficient data coverage in databases with multiple relations, Proceedings of the VLDB Endowment 13 (11) (2020) 2229–2242.

[58] McKinley, Sky and Levine, Megan, Cubic spline interpolation, College of the Redwoods 45 (1) (1998) 1049–1060.

[59] J. Benesty, J. Chen, Y. Huang, I. Cohen, Pearson correlation coefficient, in: Noise reduction in speech processing, Springer, 2009, pp. 1–4.

[60] R. F. Tate, Correlation between a discrete and a continuous variable. Point biserial correlation, The Annals of Mathematical Statistics 25 (3) (1954) 603–607.

[61] D. G. Kleinbaum, K. Dietz, M. Gail, M. Klein, M. Klein, Logistic regression, Springer, 2002.

[62] D. Dueck, Affinity propagation: clustering data by passing messages, Citeseer, 2009.

[63] Seiler, Pete, Frenklach, Michael, Packard, Andrew, Feeley, Ryan, Numerical approaches for collaborative data processing, Optimization and

Engineering 7 (4) (2006) 459–478.

[64] Edwards, David E, Zubarev, Dmitry Yu, Packard, Andrew, Lester Jr, William A, Frenklach, Michael, Interval prediction of molecular properties in parametrized quantum chemistry, Physical review letters 112 (25) (2014) 253003.

[65] Frenklach, Michael, Packard, Andrew, Feeley, Ryan, Optimization of reaction models with solution mapping, Comprehensive Chemical Kinetics 42 (2007) 243–291.

[66] You, Xiaoqing, Russi, Trent, Packard, Andrew, Frenklach, Michael, Optimization of combustion kinetic models on a feasible set, Proceedings of the Combustion Institute 33 (1) (2011) 509–516.

[67] E. Ranzi, T. Faravelli, P. Gaffuri, A. Sogaro, Low-temperature combustion: automatic generation of primary oxidation reactions and lumping procedures, Combustion and Flame 102 (1-2) (1995) 179–192.

[68] KAUST: Combustion kinetic mechanisms, King Abdullah University of Science and Technology., https://cloudflame.kaust.edu.sa/mechanisms.

[69] C3 NUIG: Combustion kinetic mechanisms, National University of Ireland Galway., https://c3.nuigalway.ie/combustionchemistrycentre/ mechanismdownloads/

[70] Lawrence Livermore National Laboratory, Combustion mechanisms, https://combustion.llnl.gov/mechanisms

[71] UC San Diego, The San Diego Mechanism - Chemical-kinetic mechanisms for combustion applications, https://web.eng.ucsd.edu/mae/groups/ combustion/mechanism.html

[72] T. Faravelli, E. Ranzi, A. Frassoldati, A. Cuoci, M. Mehl, M. Pelucchi, A. Stagni, P. Debiagi, L. P. Maffei, A. Bertolino, et al., The CRECK Modeling Group, http://creckmodeling.chem.polimi.it/.

[73] A. Cuoci, A. Frassoldati, T. Faravelli, E. Ranzi, OpenSMOKE++: An object-oriented framework for the numerical modeling of reactive systems with detailed kinetic mechanisms, Computer Physics Communications 192 (2015) 237–264.

[74] J. U. Hjorth, Computer intensive statistical methods: Validation model selection and bootstrap, Chapman and Hall/CRC, 2017.

[75] Moffat, Robert J, Using uncertainty analysis in the planning of an experiment, Journal of Fluids Engineering 107 (2) (1985) 173–178.

[76] Peters, Catherine A, Statistics for analysis of experimental data, Environmental engineering processes laboratory manual (2001) 1–25.

[77] M. L. Lavadera, A. A. Konnov, Data consistency of the burning velocity measurements using the heat flux method: syngas flames, Energy & Fuels 34 (3) (2020) 3725–3742.

[78] A. Bertolino, M. Fürst, A. Stagni, A. Frassoldati, M. Pelucchi, C. Cavallotti, T. Faravelli, A. Parente, An evolutionary, data-driven approach for mechanism optimization: theory and application to ammonia combustion, Combustion and Flame 229 (2021) 111366.