

Transforming Data into Knowledge—Process Informatics for Combustion Chemistry

Michael Frenklach

*Department of Mechanical Engineering, University of California, and
Environmental Energy Technologies Division, Lawrence Berkeley National Laboratory,
Berkeley, CA 94720, USA*

Corresponding author: Professor Michael Frenklach
Department of Mechanical Engineering
University of California at Berkeley
Berkeley, CA 94720-1740, USA
Phone: (510) 643-1676
Fax: (510) 643-5599
E-mail: myf@me.berkeley.edu

Running Title: Process Informatics for Combustion Chemistry

Topical Review

31th International Symposium on Combustion

University of Heidelberg, Germany, August 6-11, 2006

(Last modified May 13, 2006; with typos corrected)

Abstract

The present frontier of combustion chemistry is the development of *predictive* reaction models, namely, chemical kinetics models capable of accurate numerical predictions with quantifiable uncertainties. While the usual factors like deficient knowledge of reaction pathways and insufficient accuracy of individual measurements and/or theoretical calculations impede progress, the key obstacle is the inconsistency of accumulating data and proliferating reaction mechanisms. Process Informatics introduces a new paradigm. It relies on three major components: proper organization of scientific data, availability of scientific tools for analysis and processing of these data, and engagement of the entire scientific community in the data collection and analysis. The proper infrastructure will enable a new form of scientific method by considering the entire content of information available, assessing and assuring mutual scientific consistency of the data, rigorously assessing data uncertainty, identifying problems with the available data, evaluating model predictability, suggesting new experimental and theoretical work with the highest possible impact, reaching community consensus, and merging the assembled data into new knowledge.

1. Introduction

1.1. Progress in Combustion Depends on Reliable Reaction Kinetics

Research is motivated by the human desire to explain phenomena surrounding us and the passion for new discoveries. Driven by individual satisfaction such activities aim at improving the quality of human experience. One can generalize the stimulus for research as developing the ability of *making predictions*. Indeed, predicting properties of a material before it is synthesized, predicting performance of a device before it is built, or predicting the magnitude of a natural disaster ahead of time all have obvious benefits to society.

In the area of combustion, we have all the above elements: the fascination with fire from ancient times, its broad application to numerous aspects of everyday life, and the definite need for reliable predictions (see Fig 1). In recent decades the combustion research was largely driven by the desired increase in energy efficiency, reduction in pollutant formation, and material synthesis. The need for tying the research to practical applications was further emphasized in the Hottel address of the 28th Symposium by Professor Glassman [1].

One of the key areas of present and future research in combustion, emphasized in essentially all Plenary Lectures of the 30th Symposium [2-5], is chemical kinetics. The combustion community has realized from very early on that dynamics of reaction systems plays a critical role in combustion phenomena. An important intellectual accomplishment was early recognition of following the road of what became known as detailed kinetics—composing large reaction models and then reducing their mathematical form, rather than a parsimonious approach of starting from the smallest possible, usually empirically-based model. The latter approach is rooted in the presumption that the system is very complex and one will never know all the details. Hence, one may start with an empirical form and fit it to the available data. Advances in

quantum and rate theories [3] make the latter argument go away, and the present state of modeling in combustion [5] clearly testifies to the success of the detailed-kinetics approach. The combustion kinetics community's foresight in following this course is hard to understate.

A detailed account of the developments and accomplishments of chemical kinetics modeling was given in the Plenary Lectures of the 30th Symposium. Essentially all reiterated that future progress in combustion science and engineering depends on availability of reliable kinetics. Formally, we would like to pose this as the quest for *predictive reaction models*, where “predictive” means that the model can reproduce a large set of well-defined experimental data and model predictions (for the known as well as unknown) are rigorously quantified by their respective uncertainty bounds (Fig. 1). The objective of this Review is to address how we get there, with the main thesis being that the road to building such predictive reaction models lies through organization of data and methods into a new community infrastructure.

1.2. “Embarrassment of Success”

One of the pioneers of modern combustion chemistry, Fred Kaufman, made the following comment, addressed to authors of several modeling papers presented at the 19th Symposium [6]:

“There seems to exist, suddenly, an embarrassment of success: A great variety of experimental results is well modeled so that one may be tempted to claim that most combustion problems are solved. Yet our knowledge of the ~100–200 reaction rate constants is far from satisfactory. My questions are: (1) Are the mechanisms much too large so that most of the more complicated steps could be omitted? (2) How should sensitivity analysis be used in these problems and how has it been used? (3) The state of affairs of rate constant collection, evaluation, and selection seems to be chaotic. Can the authors suggest improvements, regularized interactions with kinetics practitioners, etc.

so that they will provide the best, up-to-date information, and so that the kineticists will be made aware of priority ordering of rate problems?”

Kaufman’s comments celebrate the detailed-kinetics approach; his penetrating questions raise not a criticism but the challenge to practitioners, whose successful resolution is anticipated to fully substantiate the approach.

The questions on the mechanism size and sensitivity analysis have been extensively studied and consensus has been reached. Essentially everyone accepts building over-large sets of reactions and then reducing them in size. A large number of mathematical and numerical methods have been developed over the years, which can be classified as *pruning*, methods based on comparing individual reaction terms and removing those below a certain threshold [7-16], *lumping*, transformation of species concentrations into a few dynamically equivalent lumped variables [17-28], *parameterization*, capturing relationships between model responses and model variables by tabulation [29] or simple algebraic models [9,30-35] that replace the calls to differential-equation solvers, and combinations of such techniques [13,36-38].

Likewise, the analysis of sensitivity has become a standard tool in kinetics work, and a variety of techniques emerged [39]. Recent developments place further emphasis on global aspects [40-44] and extend sensitivity analysis to transport properties [45,46] and uncertainties [47-50].

However, more than two decades after Kaufman’s comment, the chaotic state of affairs for rate constant collection, evaluation, and selection is still alive and well, and the “embarrassment of success” is still with us. For instance, a recent paper of Tamás Turányi was titled “How can very different combustion models produce similar results?” [51,52]. To find solutions, we need to understand the problem.

2. The Root of the Problem

2.1. A Set of Individual Uncertainties Does Not Represent the True Compound Uncertainty

Detailed chemical kinetics models are composed of individual reactions steps. Each reaction step has a prescribed rate law, which is characterized by a set of parameters. If these parameters were known or could be known exactly, without any uncertainties, then the construction of a reaction model would entail selecting all pertinent reactions and evaluating their individual roles through such techniques as reaction-path and sensitivity analyses. The adequacy of such a model or discrimination among several of them would then be assessed by comparing model predictions to experimental observations. Unfortunately, whether the rate parameters are determined experimentally or computed theoretically they always have a measurable level of uncertainty, and the process of reaching conclusions on the adequacy of the reaction mechanism is thus coupled to parameter identification. It is pertinent to mention that as difficult as the problem is, the detailed-kinetics approach of constructing reaction models from elementary reactions makes it much simpler as it eliminates the uncertainty associated with the reaction rate laws and provides means to evaluate rate parameters from theory.

The first obvious conclusion is that we need to know the uncertainties, and this goal is now often voiced as a critical need. Yet documentation of uncertainties, while absolutely necessary, will not on its own resolve the problem. The heart of the difficulties we face is the inherent correlation among model parameters. To visualize this, consider that all rate parameters, $\mathbf{k} = \{k_1, k_2, \dots\}$, are known along with their respective error bars, $\Delta\mathbf{k} = \{\Delta k_1, \Delta k_2, \dots\}$. The latter specification carves a “hypercube”, $\mathbf{k} \pm \Delta\mathbf{k} = \{k_1 \pm \Delta k_1, k_2 \pm \Delta k_2, \dots\}$, in the parameter space, as illustrated in Fig. 2 [53]. Yet not all points in this hypercube can reproduce pertinent experimental data within their respective uncertainty bounds; and those that can, form a *feasible*

set, F . Typically, the feasible set does not have a simple geometry and becomes more complex with the increase in the dimensionality of the parameter space and with the increase in the number of experimental observations to be matched. A realistic three-dimensional example is given in Fig. 3. One must appreciate that the current reaction systems deal typically with 10^2 – 10^3 -dimensional parameter spaces!

2.2. *Artificial Controversies*

Many “controversies” seen in chemical kinetics arise from explicit or implicit disregard of the feasible set, by treating all parameters and their uncertainty ranges as essentially independent of one another. A typical situation goes like this. The researcher fixes several (or many) parameters at their “recommended values” and adjusts just a few, the most influential ones, to match some (usually the researcher’s own) experimental observations. At first glance, this may seem to be perfectly harmless, and actually to follow the sensitivity-analysis rationale. Yet, artificially restricting the motion in the parameter space by doing so, one quickly leaves the feasible region. Let us consider the two-dimensional example of Fig. 2. Adjusting, say, k_1 at a constant k_2 will most likely end up outside the shaded area in Fig. 2. Exactly the same, yet hard to visualize, situation occurs in multi-dimensional cases: a single-parameter change moves away from the feasible set, and several of such changes bring to the state of chaos noted by Kaufman. Further analysis of the origin of such “controversies” can be found in Refs. [40,54,55] and examples of feasible sets encountered in combustion kinetics in Refs. [47,54,56-58].

2.3. *An “Expert Choice” of Parameter Values is Like a Needle in a Haystack*

Many, not only Kaufman, recognized the “state of chaos” as one of the key problems plaguing the progress of combustion chemistry. For instance, David Golden, referring to conflicting assignment of rate coefficients in kinetics models and calling for doing it in a

consistent manner [59], used to say that “Any agreement between model and experiment is sheer luck”. While intended as a friendly attention caller, the “sheer luck” does correctly portray the reality. Consider that the feasible set is approximated by a hypercube with each of its sides equal to one half of the respective Δk (which is a great exaggeration as the usual feasible set is much smaller by volume). The fraction of such a feasible set of the entire uncertainly region is 2^{-n} , where n is the dimension of the hypercube. For $n = 100$, the size of the GRI-Mech 3.0 parameter space [60], this fraction is roughly 8×10^{-31} . This means that selecting a point at random within the uncertainty region $\mathbf{k} \pm \Delta \mathbf{k}$ one would chose a point within the feasible set with the probability of 8×10^{-31} . To get a measure how “lucky” this is, let us compare this probability to the following ones. The probability of randomly selecting your own name from a list of all the people on Earth is 2×10^{-10} and that of a given atom in a human body is $\sim 10^{-28}$, both much “luckier” outcomes than selecting a point in our exaggerated feasible set.

2.4. *Comprehensive Hierarchical Mechanisms*

One of the earlier concepts advanced for systematization of kinetic model building was *comprehensive hierarchical mechanism* of Westbrook and Dryer [61,62]. The model is assembled in a *hierarchical* manner, starting from the most primary subsystem, say, a set of reactions describing hydrogen oxidation, validating it against pertinent experimental data, adding the next set of reactions, those describing CO oxidation and validating the combined set against pertinent CO experimental targets, adding the next set, and so on. The *comprehensiveness* is understood as the ability of the (single) model to reproduce all available experimental data at once. The key feature of the hierarchical building is the presumption that once the smaller reaction subset is validated, it is “frozen” for the remainder of the process; in the above example, matching the CO targets is accomplished by tuning only the added reaction subset of the CO

subsystem, while the preceding reaction subsystem, hydrogen oxidation, is remained unchanged. Charlie Westbrook used to liken this model assembly process to building a house of cards, with an experimenter's report of an updated rate constant for one of the validated reactions having an effect equivalent to pulling a card from the house foundation. In terms of the parameter hypercube discussed above, the hierarchical parameter freezing creates artificial constraints with a consequent departure from the feasible set in the process.

2.5. Model Comparison

There is also a proposition for direct comparison of different models, selecting one which exhibits the best overall fit to data. Building on the established, centuries-tested standards of reasoning aimed at discrimination among competing models, this approach served well at the dawn of detailed kinetics models, when the mechanisms were composed of at most a dozen reactions and the key unknown was identification of the reaction steps themselves. The situation is much different now, when, for the most part, the community has reached general consensus on the principal reaction pathways of combustion (with the important exception, perhaps, of aromatics and soot formation). The focus has shifted from "guessing" what reactions are possible to asking where one should place the effort, as the modern tools of quantum and reaction-rate theory can produce the order-of-magnitude estimates (e.g., [3,63,64]). The kinetic models increase in size, and the number of models continues to rise. Even if the process is completely and cleverly automated, comparison of models on its own offers no systematic provision for efficient merging of models or data. From the perspective of the parameter-space hypercube, this approach is reminiscent of a random walk in the parameter space with unspecified convergence.

2.6. Optimized Models and Solution Mapping

Let us consider again the example in Fig. 2. We noted earlier that changing k_1 at a constant k_2 will most likely end up moving out of the feasible set. Yet, k_1 can be also adjusted by simultaneously changing k_2 , so as to always remain in the feasible set. This stipulates a properly constructed constrained optimization. A familiar example is the correlated variation of the preexponential factor and activation energy in development of an Arrhenius expression for rate coefficient values determined at a set of temperatures. One would not approach this task by calculating a set of expressions for different pairs of scattered experimental points and selecting the best among them, but resort to a least-squares analysis of the entire data set. There is no fundamental (or philosophical) difference between this simple example and the task we face of developing a predictive reaction model. We still have to fit the model (parameters) to match all the data available. The difficulty of a practical solution lies in the vast dimensionality of the parameter space, the large number and heterogeneity of the experimental observations, and the implicitly specified relations between model responses and parameters requiring solution of systems of differential equations. Hence, the “head-on” approach of employing an optimization code that calls directly shock-tube, flame, etc. solvers for evaluation of experimental targets is computationally inconceivable for the foreseeable future.

These practical difficulties can be finessed by the effect sparsity and Solution Mapping (SM) methodology. Under conditions of an individual experiment, the model responses that correspond to the experimental observations do not depend sensitively on all the parameters. In fact it has been noted by many that usually only a small fraction of the parameters show a significant effect on measured responses. This phenomenon has been termed *effect sparsity* and the influential parameters *active variables* [65,66].

The SM approach [40,55,56,67] decouples the optimization problem, first reducing the differential equation models to simple algebraic relations between model responses and model active variables, and then performing multi-response, multi-parameter optimization using these relations. The algebraic relations, referred to as statistical surrogates, are developed through statistical techniques of response surface design [66,68,69], by performing computer experiments with the “complete” reaction model. The key feature that needs to be emphasized in the context of the present discussion is the fact that the statistical surrogates are developed only in active variables; i.e., for each model response we develop a statistical surrogate in its own set of active variables. In this way, we combine the quest for comprehensiveness overall with the quest for locally smallest possible models. The analysis can include as many data sets as needed and has no dependence on the order of their inclusion, thus offering an alternative to the hierarchical model building discussed above. SM offers additional advantages that will be presented later.

2.7. GRI-Mech

The term “GRI-Mech” refers to several aspects of that project [60]: GRI-Mech *Organization*, GRI-Mech *Scientific Methodology*, GRI-Mech *Database*, GRI-Mech *Best Current Dataset*, etc. Much of the experience gained through this project paved the way to the main subject of this Review, Process Informatics (aka PrIME), and hence many details, common to both GRI-Mech and PrIME, will be covered in the next Section. The objective here is to capture some of the historical perspective in the context of the preceding discussion.

The GRI-Mech project was created in 1990 by the initiative of four of the contractors funded prior to that individually by the Gas Research Institute in the area of chemical kinetics: SRI (David Golden, Greg Smith, David Crosley), Stanford University (Tom Bowman and Ron Hanson), University of Texas (Bill Gardiner), and Pennsylvania State University (and later

University of California at Berkeley, Michael Frenklach). These researchers proposed to work jointly on a single, unified reaction model for natural gas combustion, instead of building four individual, competing models. The premise of this proposal was development of a systematic and robust methodology that will produce predictive reaction models, and that such models will be available to the user on a regular basis and in a convenient manner.

The scientific underpinning of the GRI-Mech approach was an *iterative* process, with a single cycle comprised of a new round of data (re)evaluation and model optimization. The completion of the optimization cycles resulted in new GRI-Mech reaction datasets released to the public, with a frequency of about once or twice a year. The dissemination of the GRI-Mech results was done using the emerging then internet technology, an open-access Web site [60].

Each cycle began by expanding the GRI-Mech database from the preceding release by bringing in additional reaction subsystems, additional experimental targets, and revising and updating the old data. The GRI-Mech protocol included assessment of thermochemical data, reaction rate data, experimental targets, and their respective uncertainties. A list of active parameters was identified for each experimental target by ranking impact factors, $|\text{sensitivity} \times \text{uncertainty}|$. The uncertainties in all active variables establish a hypercube in the parameter space over which statistical surrogates were developed and joint optimization of all selected experimental targets was performed. The closeness of model fit to the experimental values weighted against the uncertainties assessed for the experimental targets formed the basis for selection of the “best current data set”—the new GRI-Mech release. *The most problematic issues encountered in the process suggested the next experiments to perform.*

An important feature of the GRI-Mech process was the protocol for reaching consensus. In a usual panel evaluation, the decisions are made based upon the final reports of individual studies;

in other words, an evaluation panel is discerning among already formed opinions. In the GRI-Mech protocol, the decisions on “quality” were done preceding this final step, preempting the “controversies”. The team deliberations began with considering different reports and opinions for each parameter (agreed upon a priori to be in the list of tasks for the next release). If there was a disagreement, either among the literature values or among GRI-Mech researchers, a wider range of uncertainty for such a parameter was accepted at that stage, with a provision that later on, after the optimization, these assumptions and choices will be revisited. The same was done for the experimental targets, and optimization itself. The work was facilitated by the “right” tools. For instance, the optimization code was set on a single server, but the process of optimization and reporting of the results was automated. In this way, each member of the GRI-Mech team was able to perform optimization on his own, select a specific set of constraints and statistical weights, and form his own opinion. Also, the use of the Solution Mapping approach allowed the team to perform the optimization of the entire GRI-Mech set on the fly, thus enabling and supporting team deliberations.

The popularity of GRI-Mech releases extended beyond the project termination, and the accuracy of its predictions was demonstrated repeatedly in the literature (see, e.g., most recent numerical studies of combustion phenomena [70-72]). The success attests to the efficacy of the GRI-Mech process. Still, as the project progressed, the GRI-Mech researchers became painfully aware of the problems. They were of two kinds: technical and sociological. On the technical side,

- The increase in the amount of data needed to be processed started to cause errors in data recording, data copying, file misplacement, and the like, all pointing to a critical need for automation of data handling;

- The increase in the ability to handle large sets of heterogeneous data calls for automation of data analysis, discussed further in Section 4 under Data Collaboration;
- The amount of effort and the level of expertise that goes into expanding the data set to include additional subsystems (i.e., to increase the data set comprehensiveness) is beyond the ability of a single researcher, a single research group, or even a small team of such groups (such as the GRI-Mech team was), suggesting that the solution should come from involving more people, and ultimately, the entire community.

There is definitely a strong sociological side to the subject matter, and this section will be incomplete without mentioning that the GRI-Mech project drew some negative reactions, like “GRI-Mech does not predict my data”, “GRI-Mech does not correctly predict methanol oxidation” (while there was no methanol chemistry included in the GRI-Mech releases), and “one cannot do basic science through fitting”. Encouragingly, the subject of model optimization has become more “acceptable” in recent years, as evidenced by the rising number of publications of new optimization [73,74] and error-analysis [48,49,75] methods for extracting kinetic information from experimental observations. Recent efforts of Ruscic and co-workers on “Active Thermochemical Tables” [76] and of Frenkel and co-workers on “ThermoData Engine” [77] corroborate further the benefits of global multi-dataset optimization.

3. A New Paradigm: Process Informatics

3.1. There Must Be a System

We are now ready to tackle Kaufman’s question on how to improve the chaotic state of affairs of rate constant collection, evaluation, and selection. There is an increasing attention given to this very issue, identifying data organization as a critical need for progress in developing reliable chemical kinetics models in combustion. The motivation often cited is the impact made

to the thermal sciences by the JANAF [78] and NASA thermodynamics tables, to gas-phase reaction kinetics by the Leeds [79], NASA/JPL [80], and NIST [81] data evaluation and compilation efforts, and to the experimental data collection by the International Workshop on Measurement and Computation of Turbulent Nonpremixed Flames [82]. With the Internet coming of age, the Web sites reporting thermochemical, reaction, and experimental data sets begin to proliferate. However, while all such efforts are commendable, the problem we face will not be solved just by mere collection and Web publication of (usually conflicting) data and models.

Dmitri Mendeleev, the discoverer of the Periodic System of Elements, wrote [83]:

“The mere accumulation of facts, even an extremely extensive collection, ... does not constitute scientific method; it provides neither a direction for further discoveries nor does it even deserve the name of science in the higher sense of that word. The cathedral of science requires not only material, but a design, harmony ... a design ... for the harmonic composition of parts and to indicate the pathway, by which the most fruitful new material might be generated.”

This one-and-a-half-century old statement is echoed by modern ones, like that voiced in the summary of a recent NRC workshop on “Bioinformatics: Converting Data to Knowledge” [84],

“... how well data are turned into knowledge depends on how they gathered, organized, managed, and exhibited—and those tasks are increasingly arduous as the data increase. ... databases can be far more than repositories—they can serve as tools for creating new knowledge”,

that of experienced computer scientists, whose recent Physics Today article [85], titled “Computational Science Demands a New Paradigm”, begins with a declaration:

“The field has reached a threshold at which better organization becomes crucial”,
and that of Peter Cochrane, a modern-technology visionary [86],

“We have deployed technology that speeds up our communication and ability to enact decisions, but without simultaneously investing in models and aids to visualize and understand the outcome. In short, we have just not invested in the appropriate management tools and systems”.

All these quotes have a direct implication to our current topic. For the reasons discussed in Section 2, just assembling and comparing data, in old or modern form, does not offer means for synthesis and integration, and thus unlikely to resolve the “state of chaos”. In fact, it may actually divert from the real solution by creating a new level of “embarrassment of success”—imagine 5-10 years from now someone making a statement: “A great variety of reaction models is posted at numerous Web sites so that one may be tempted to claim that most combustion problems are solved. Yet our ability to reproduce the combustion in a diesel engine is far from satisfactory” (cf. Kaufman’s citation).

To succeed, we need to build an entirely new type of organization of data, methods, and community attitude—Process Informatics.

3.2. Scope of Process Informatics

Process Informatics (aka PrIME for Process Informatics Model) is a data-centric approach to developing predictive models for complex chemical reaction systems. It deals with all aspects of integration of pertinent data of complex systems (industrial processes and natural phenomena) whose complexity originates from *chemical reaction networks*. The primary goal of process informatics is information gathering, validation, and transformation into a useable form. The

latter includes development of *predictive* (numerical/computer) models with quantified degrees of reliability.

Chemical reaction models will never be complete. A problem is that the data on which models are based are scattered over different sources and are not properly evaluated. Most importantly, these data cannot be applied directly to practical problems—they have to be “transformed” into useful models. Such models, however, cannot be created by simple “compilation” of the data, as discussed above. The goal of Process Informatics is to convert such model building into science, automate the methodology, and make the results available in a prompt and convenient form for the user.

3.3. Origin of PrIME

As of this writing, PrIME is a “grass-roots” initiative. One of the defining events for creation of PrIME happened at the 77th International Bunsen Discussion Meeting held in Bad Herrenalb, Germany, in October 2001. The session and “corridor” discussions on building reaction chemistry a la GRI-Mech turned into an ad hoc session on Process Informatics [87]. The outcome of that was a “Position Paper” [88]; and the preceding text is an excerpt from that document.

In 2002, during the 29th Combustion Symposium in Sapporo, Japan, there was a follow-up meeting of the PrIME enthusiasts. The “Sapporo Report” [88] documents, among other things, postulated principles of open membership and democratic governance for PrIME.

The PrIME Initiative was launched at a meeting at the University of California at Berkeley on April 21-22, 2006 [88].

3.4. Data and Models

In pursuing the new roadmap we have to revisit some old concepts, introduce new ones, and adapt to new terminology. Here, we begin by the keystones of PrIME, data and models.

For small reaction systems, such as those dealt with at the dawn of detailed kinetic modeling, terms “reaction *mechanism*” and “reaction *model*” were used interchangeably. Unfortunately, this “equivalence” was propagated to larger reaction sets, with further additions of such terms as “rate coefficient *compilation*” and “Chemkin *file*”. It is imperative, however, to differentiate among these terms and use them appropriately.

As an example, there is broad consensus on the *mechanism* of methane oxidation, namely that CH₄ is converted to CH₃, which is oxidized to CH₂O, followed by conversion to CO, which is oxidized by OH to CO₂, etc. Yet, there are numerous detailed kinetic *models* available and being continuously produced for methane oxidation. Even if we all to agree on the perfect “comprehensive mechanism” of methane combustion, we would want to produce various reduced models for different applications. So, *mechanism* and *model* are not equivalent concepts. And “compilation of rate coefficients” is neither mechanism nor a model—it is a collection of *data*.

We thus should reserve term *mechanism* for what it really means—principal reaction pathways responsible for the phenomena in question (like coupling between soot and NO formation). For better known systems, like methane oxidation, we no longer question the mechanism but rather seek a model—to interpret experimental observations, to predict the levels of NO forming in a new engine design, or to be used in an automated air-pollution control device. In other words, in the context of the present discussion, a *model* is a set of mathematical equations describing the phenomena in question or its computer/numerical equivalent. Given a

set of elementary reactions (i.e., a detailed mechanism) specifies the corresponding set of differential equations, i.e., the mathematical model. With a (known) set of parameters, such a model generates prediction for model responses. A *predictive* model specifies the bounds for these predictions.

Experimental observations are classified as *data*. It is by matching the latter we identify model parameters or test model validity. However, not necessarily every experimental “point” needs to be taken for such analysis. Those experimental features chosen for this purpose are referred to as the *training data set* and the process of fitting the model parameters by matching the training set as *model training*.

While “experimental data” is a clearly identifiable concept, we have some ambiguity in classifying “thermochemical and rate data”. Say, when we compile a set of rate coefficients for reaction $\text{H} + \text{O}_2 \rightarrow \text{OH} + \text{H}$, we consider this set to be “data”. But when we use one of these expressions in a specific model, it becomes part of the model, a *model parameter*.

Finally, with above in mind, we should replace the term “comprehensive mechanisms” with “predictive models” and “comprehensive data base” and not mix the two. The “data base” can be considered to consist of both experimental and thermochemical-and-rate data, as the two are closely interrelated with each other, as will be discussed in Section 4. We can then ask how to convert the information content of the data into practical predictions, i.e., into predictive models.

3.5. PrIME Infrastructure

The Process Informatics infrastructure, referred to as *Process Informatics Model* (PrIME) and under construction at the time of this writing [88], will have two principal components: a *Data Warehouse* and a collection of *Tools*. The PrIME Data Warehouse will consist of *Data Depository* and *Data Library*. The Depository will represent the most currently complete set of

data available in the field of combustion chemistry (conceptually an “infinite data bank” [9]). It will contain experimental data, on both complex systems and on elementary reactions, molecular properties determined from quantum chemical calculations, reaction-rate coefficients, obtained from reaction-rate theories, and similar information. The Process Informatics *Tools* will be of two general kinds, those enabling the collection, transfer, organization, display, curation, and mining of the data—i.e., those maintaining the infrastructure of PrIme, and those enabling processing and analysis of the data along with assembly of the data into models—i.e., scientific and numerical tools.

The PrIme Data Depository will contain collections of bibliographic references, chemical elements, chemical species, chemical reactions, and experimental records. One of the distinctive features is that the Data Depository is organized not by the source or origin of the data, typical of most other data-collection activities, but by merging the different sources together according to the scientific meaning of the data (while tracking, of course, the origin and other metadata). The initial collection is being prepared by merging in this matter the NIST Kinetics Database [81], GRI-Mech 3.0 [60], and European-Union kinetics database [79]. It is envisioned that further, permanent-mode population of the Data Depository will be accomplished through participation of the entire research community.

Evaluation of these data will be performed by small groups of PrIme team members, open to all and created on the basis of matching the expertise. For instance, such evaluation groups are envisioned for H₂/O₂ subsystem, C₁-C₆ chemistry, PAHs, soot, etc. Collectively, the participants of these groups will constitute the PrIme Scientific Council. Its mission is “quality control” of the knowledge buildup of combustion chemistry. The results of their evaluation will constitute the “best current recommendations” that will be recorded in PrIme Data Library—one entry per

each property. For instance, while the Depository will store all suggested rate expressions for reaction $\text{H} + \text{O}_2 \rightarrow \text{OH} + \text{H}$, the Library will have a single rate expression, the best current recommendation, either selected among the set in the Depository or generated on the basis of this set by the evaluating group. In many ways the Scientific Council is similar to a Data Evaluation Panel, an established practice today for database quality control. The difference—and hence the novelty underlying the shift in the scientific approach—is that the Council activity will be based on the analysis of the *entire knowledge* available in the field. It is the goal of Process Informatics to develop infrastructure, tools, and protocols to enable such operation of the Scientific Council.

3.6. *PrIMe Vision*

The two principal customers of the Process Informatics System are the *data provider* and the *model user*. During the development stage, there will also be a *model builder*, whose role will eventually be automated.

A *Data Provider* (an experimental or theorist chemist; the “Chemist” in Fig. 1) submits new observations or new computational results, which are placed in Depository. The protocol assures completeness of the data submission. The deposited data are immediately analyzed for consistency with the Library and the results are reported both to the data provider and to the scientific council. Upon approval of the council, the Library is modified. In other words, the PrIMe Data Library will be fluid and will be continuously modified and these modifications will be documented.

A *Model User* (a design engineer, CFD researcher, atmospheric modeler, a policymaker, etc; the Policymaker and the “Engineer” in Fig. 1) requests a kinetic model (or a simulation with such a model) and specifies the conditions of interests, the desired level of accuracy, and the

mathematical form of the model (detailed, reduced, parameterized, etc). The system checks for the existence of such a model, if none is available one is generated.

Another user might be a project manager (a member of the Scientific Council, or a researcher) who wants to know whether a proposed experiment/calculation will improve the current knowledge, and by how much. Various scenarios with the envisioned experiment or calculation can be evaluated. A similar question can be posed differently: what needs to be done to improve the predictability of a given model? Repeat old experiments? Under what conditions? New experiments? Which ones? New calculations? What level of local error must be maintained to accomplish the stated goal?

Thus, the goal of PrIME is not merely a collection of data and tools, but a shift in the paradigm of the scientific process: building targeted knowledge by the entire community and providing the wealth of information in its entirety to every user. In a general sense, Process Informatics establishes a process for reaching community consensus on firm scientific grounds.

4. Data Collaboration

The notion of Process Informatics relies on the presumption that it is feasible to meaningfully study all the data taken together, and that such analysis provides definite benefits over other forms of data processing. While the suggested PrIME data organization can support various modes of data analysis and different types of community and individual activities, here we will focus on the approach called *Data Collaboration* [47,54,55,89-91], which can serve the symbiosis of the Process Informatics.

4.1. Data Collaboration Concepts

Dataset. The concept of a dataset lays down the foundation for the Data Collaboration methodology. We associate with an experiment, E, a *dataset unit*, which consists of the

measured value, d_e , reported uncertainties in the lower and upper bounds on the measurement, l_e and u_e , respectively, and a mathematical model, M_e . The model M_e is defined as the functional relation between the model active variables, \mathbf{k} , and the prediction for d_e . The feasible values of \mathbf{k} are those that satisfy $l_e \leq M_e - d_e \leq u_e$; this ties together data, model, and uncertainty. A *dataset*, D , is a collection of such dataset units $U_e = (d_e, u_e, l_e, M_e)$. In the present Data Collaboration methodology, the models are the statistical surrogates developed in computer experiments, namely $M_e = s_e(\mathbf{k})$, and so $U_e = (d_e, u_e, l_e, s_e)$ and $D = \{U_e\}$. These surrogate models are developed using the response surface technique, following the methodology of Solution Mapping described in Section 2.6. The concept *dataset* of Data Collaboration will take the explicit expression in building the PrIME Data Library.

The creation and organization of a dataset is guided by the system in question, like the formation of nitrogen oxides in combustion of natural gas or formation of soot in flames of automotive fuels. A single experiment cannot provide complete information on such a system, but rather probes its particular aspect. A collection of such individual “bits” of pertinent information (i.e., dataset units) forms a dataset. The more extensive and diverse the collection, the more complete is the understanding of the system. The unifying principle, the one that determines the “pertinence” of a given experiment to a given dataset, is a presumption that there exists a single chemical kinetics model, common to all dataset units, that is expected to predict d_e when exercised at the conditions of experiment E. In other words, it is presumed that broad consensus exists (at least tentatively) regarding the necessary reaction steps of the system and hence the mathematical structure of the unifying kinetic model is known, and that this mathematical model is sufficient, in principle (with the “right” choice of parameter values), to predict all experimental observations included in the dataset.

Initial Hypercube. We assume that there is prior information on the possible values of the model parameters. This prior information can be expressed as the confinement of possible values of the active variables to an n -dimensional hypercube, $H := \{\mathbf{k} \in \mathbb{R}^n : k_{i,\min} \leq k_i \leq k_{i,\max}\}$, where $k_{i,\min}$ and $k_{i,\max}$ are the lower and upper bounds, respectively, on k_i for $i = 1, 2, \dots, n$. Each edge of the hypercube H represents the presumed interval of “physically-allowed” values of the corresponding model parameter, either the estimated uncertainty or a range containing the differing values. The latter information will come from the PrIME Data Library, which will store the best current data including the ranges of uncertainty.

Feasible Region. As mentioned earlier, not every point \mathbf{k} in H predicts all experimental observations of the dataset within their specified uncertainties. The collection of parameter values that are both contained in the hypercube and satisfy $l_e \leq s_e - d_e \leq u_e$ for every dataset unit e in the dataset form the *feasible region*, F . A point \mathbf{k} that is not contained in F has been eliminated from consideration as a possible value for the dataset active variables by either the prior information, through the bounds of H , or by the experimental observations of the dataset, through intervals $(d_e + l_e, d_e + u_e)$. It is in this manner that experimental observations increase our knowledge of kinetic parameters: an experiment may eliminate portions of the hypercube H from consideration, thereby decreasing the uncertainty in the values of the kinetic parameters.

4.2. Data Collaboration Approach

A dataset constitutes a multitude of assertions about the feasible region within the parameter space. One would like to do reasoning on this collection of assertions. For instance:

- Is the feasible region empty? If so, then something is wrong about the dataset, invalidating at least one of the dataset units and/or the underlying reaction model which is common to all of the units.
- Which dataset units have the most impact on the overall dataset's (in)consistency? Answering this can signal dataset units that are possibly incorrect, though self-consistent.
- Which model assumptions have the most impact on the dataset's (in)consistency?
- What is the tightest range of predictions about an additional experiment, given that these predictions must be consistent with the dataset?
- Which experiment or/and parameter bound is mostly responsible for the given bound on the model prediction? In other words, which experiment/parameter should be improved to tighten the range of model uncertainty?
- What is the utility of a hypothetical experiment to further knowledge regarding the system? In this framework, "what-if" questions can be posed and addressed. For instance, *what if* additional shock tube experiments are performed at such and such conditions; what is the anticipated contribution that these experiments could make to further knowledge of this system? In other words, what is the likelihood that these new experiments will tighten the range of the model prediction?

All of these questions lead to set-intersection questions. Some of these questions can be answered with the methodology available today (e.g., by combining solution mapping, sum-of-squares programming, heuristic search methods, and branch-and-bound strategies [47,54,90,91]), and some require further development.

The Data Collaboration methodology, thus, puts models, theory, and data on the same footing. It does not change the way experimentation is done, but requires a different approach to

analyzing even one's own observations and, as a consequence, places new standards on data reporting. In this approach, measured data, its estimated uncertainty, and a model of the experimental system are treated as an assertion whose correctness depends on the suitability of the model and the reliability of the measured data. Taken together, the model and measurement constitute a (low dimensional) constraint in the global unknown parameter space. Specifically, only those parameters that are consistent with the model/measurement pair are possible values of the unknown parameters. By considering only these parameter values, one can harvest a majority of the information content of the data [54], determine realistic bounds on model predictions [89], test consistency of a dataset (see Fig. 4) [47], or discriminate among competing models [91]. This numerical methodology avoids unnecessary overconstraining of model parameters that plagues many other techniques due to inherent correlations among parameters, and allows one to explore more closely the true feasible region of the parameter space in a computationally efficient manner.

5. Data-Centric Predictive Models

Development of predictive models marks human civilization. Ancient models were philosophical in nature and relied entirely on logical deduction; one example of this can be the concept of corpuscular (i.e., atomic) nature of matter arrived at by the ancient Greeks two millennia ago. The invention of geometry and algebra in the Middle East was motivated by and applied to agricultural problems. Early scientific discoveries were expressed in terms of simple phenomenological laws, like Fick's Law of diffusion and Boyle's Law of gases. Development of calculus expressed these laws in terms of differential equations and formulated solutions to their linearizations. The advent of computers provided means to solve these differential equations in

their complete nonlinear formulation, such as the Navier Stokes equation of fluid mechanics or Schrodinger equation of quantum mechanics.

Thus, the concept of a model and the “physical” form it assumes have changed with time: from conceptual statements to simple algebraic relationships to differential equations to numerical algorithms to computer programs and files. The present problems call for integration of a variety of computer programs, from quantum chemistry to fluid mechanics. The primary difficulty, common to essentially all fields of science and engineering, is the fact that more authentic models introduce larger numbers of parameters. The general expectation is that advancement in science should provide means to establish these parameters with required accuracy. Experience both supports and disproves this expectation. Indeed, the unprecedented advance in scientific instrumentation (e.g., laser spectroscopy) and computer technology (the increase in speed and memory and the reduction in cost) provides powerful means to determine the parameter values on sound experimental and theoretical grounds. At the same time, this often (if not always) involves another model, either instrumental or/and theoretical, which in turn introduces additional parameters and additional uncertainties.

The current practice of model development/use is based on the following paradigm: (a) A model is postulated introducing a set of parameters; (b) A “unique” set of parameter values are determined from experiment or/and theory, and (ideally) supplied along with a corresponding set of individual uncertainties; and (c) the model is then applied to conditions of interest employing the unique set of parameter values. The natural uncertainties of the underlying experiment and theory must somehow be transferred into the final prediction uncertainty using the uncertainties that were assigned to the model parameters. As discussed in the present paper, this conventional paradigm does not lead to a desirable quality of prediction.

The methodology of Data Collaboration integrated with PrIMe lends itself toward data-centric approach to model building. In so doing the paradigm of a model is shifted from considering model parameters as “unique”, predetermined values with individual, uncorrelated uncertainties to including the actual experimental data, along with the physicochemical theoretical constraints if available, as the integral part of the model, with model parameters playing a role of internal variables. In other words, instead of the two-stage approach—i.e, estimation of model parameters from fitting experimental data followed by model predictions using the obtained estimates—the uncertainties of the “raw” data are transferred into model prediction directly.

The advantage of fully collaborative environments, in which models and data can be shared allowing global optimization-based tools to reason quantitatively with the community information, can be illustrated with an example from Ref. [54]. A series of “toy” numerical experiments were performed to gauge the information loss by doing “traditional” analysis without the benefit of Data Collaboration at the raw data level. A measure of information gained as a result of the data processing was defined as a relative decrease in the predicted range R of (an arbitrary) model, s_0 ,

$$I := 1 - \frac{\Delta R_{\text{posterior}}}{\Delta R_{\text{prior}}},$$

where ΔR is the length of an interval R . Prior to data processing, the variations of k 's are confined to the initial hypercube H , and

$$\Delta R_{\text{prior}} = \max_{\mathbf{k} \in H} s_0(\mathbf{k}) - \min_{\mathbf{k} \in H} s_0(\mathbf{k})$$

is the range of model prediction on H . $\Delta R_{\text{posterior}}$ denotes the range obtained after considering the dataset constraints; for instance, in the Data Collaboration mode, it becomes

$$\Delta R_{\text{posterior,Data Collaboration}} = \max_{\mathbf{k} \in F} s_0(\mathbf{k}) - \min_{\mathbf{k} \in F} s_0(\mathbf{k}),$$

where k 's are confined to the feasible region F . Recall that the feasible region F is a subset of H , $F \subseteq H$, and hence $\Delta R_{\text{posterior,Data Collaboration}} \leq \Delta R_{\text{prior}}$. If the analysis does not follow the Data Collaboration mode and engage just a subset of the dataset information, we obtain $\Delta R_{\text{posterior,Data Collaboration}} \leq \Delta R_{\text{posterior,Without Data Collaboration}} \leq \Delta R_{\text{prior}}$.

We can compare how I changes depending on the mode of data analysis. For this purpose, let us define the *Information Loss*, L , due to not using Data Collaboration as

$$L = \frac{I_{\text{Data Collaboration}} - I_{\text{without Data Collaboration}}}{I_{\text{Data Collaboration}}}.$$

A series of 100 computer runs were performed with a randomly assigned model to predict, s_0 , constrained to the GRI-Mech 3.0 dataset [54]. The results, reported in Fig. 5, show about 90 % loss—a significant price to pay for the lack of collaboration.

5. Concluding Remarks

The success of Process Informatics will depend on broad community involvement. In addition to scientific and technological challenges, there are also sociological ones. Perhaps the foremost among the latter is sharing data—not the usual “read my paper” or posting the published results on the Web, but essentially contributing one’s own “raw” measurements to the community-wide analysis. One should not underestimate the difficulty of this requirement [92], but the encouragement comes from the success of the GRI-Mech pilot. With proper technical

protocols, thoughtful (re)organization of scientific activities, and honest consideration of sociological elements, it is feasible to build an infrastructure that will support and encourage individual researchers to share the data. One of the PrIME objectives is to find such a way.

The pursuit of PrIME presents the combustion research community an opportunity to lead the science of complex reaction networks. Like many other fields, and most notably biology, combustion deals with complex chemical systems. Yet, unlike other fields, combustion chemistry is ahead of others: the “anti-parsimonious” approach of detailed mechanisms reduces the “mystery” of chemistry to sets of elementary reactions, modern quantum chemistry and reaction-rate theory provide a solid framework for estimation of reaction rates, sophisticated laser diagnostics allow collection of very specific data, present numerical algorithms and simulation codes are capable of reliable predictions for basic combustion apparatus, the demonstrated scope of model optimization reassures its practicality, and developed reduction strategies make it feasible to use realistic chemistry in fluid-dynamic simulations of practical combustion. Putting it all into a system such as PrIME will attain the stated goal of maintaining data consistency, identifying best experiments to perform, and developing predictive models, and in so doing will intimately bridge fundamental science from the atomic level to laboratory experiments, to design of practical reactors and combustors, to assessment of human impact on the environment.

Acknowledgments

The presented here ideas have been developed as part of several projects—including GRI-Mech [60,93,94], Data Collaboration [47,54,89-91], CMCS [95], and PrIME [88]—working closely with many colleagues, students, postdocs, and program managers, whose valuable contribution I would like to acknowledge, along with the financial support over the years by

various sources including NASA, AFOSR, GRI, and recently by the NSF Information Technology Research Program, Grant No. CTS-0113985, the NSF Chemistry Division initiative on Collaborative Research: Cyberinfrastructure and Research Facilities, Grant No. CHE-0535542, and in part by the Director, Office of Energy Research, Office of Basic Energy Sciences, Chemical Division of the U.S. Department of Energy, under contract No. DE-AC03-76SF00098, and as part of the CMCS project supported by the U.S. Department of Energy of Energy's Office of Mathematical, Information, and Computational Sciences, Grant No. DE-FG02-01ER25445. The manuscript benefited from helpful comments of David Golden, Andy Packard, Steve Pope, Tamás Turányi, Charlie Westbrook, and two anonymous reviewers.

Among many events on the journey, there were several defining moments and individuals. The late Bill Gardiner, PhD , was the first to appreciate what I was talking about and invited me to present these ideas in a Chapter [40]. David Golden, concerned with the consistency of the thermochemical and rate data used in combustion models, recognized the benefits of a joint data analysis (with Solution Mapping) and encouraged the teaming to attain this objective; Tom Bowman placed the team goals ahead of his personal "secure funding"; Robert Gemmer, acquainted with the statistical response-surface methods, and Jim Kezerle pushed the new concept of a unified, GRI-Mech consortium through the GRI bureaucracy; and Robert Serauskas, a fan of novel electronics, endorsed the use of the emerging Web technology for dissemination of the GRI-Mech results.

Christopher Nokleberg and Mikhail Goldenberg created the first menu-driven Web application of chemical kinetics, "GRI-Mech Calculator". Crossing roads with Andrew Packard put the collaborative data analysis on firmer mathematical grounds. Gregory Smith, a seasoned GRI-Mech-er, tested the new tools of Data Collaboration on an atmospheric chemistry problem.

A broader community outreach that led to what became known PrIME started with Michael Pilling expressing an interest in combining the kinetics efforts across the Atlantic Ocean in a la GRI-Mech approach. Wing Tsang called GRI-Mech a “gold standard”, Gregory Rosasco understood the benefits of community-led management of data fostered by PrIME, and Gregory McRae and William Green merged their vision of a roadmap with PrIME. Tom Allison converted NIST Kinetics Database into PrIME format and Zoran Djuriscic made the PrIME Data Depository a reality [96]. Adel Sarofim, Philip Smith, and their colleagues at the University of Utah became actively involved in PrIME, as well as Branko Ruscic and his colleagues at the Argonne National Laboratory. David Davidson, Ronald Hanson, Hai Wang, and Phillip Westmoreland are working on the next milestone in data management; and a large number of people across the Globe, who expressed support to PrIME, are ready to get engaged in the activities.

David Golden has been a devoted PrIME comrade and a personal confidant.

References

- [1] I. Glassman, *Proc. Combust. Inst.* 28 (2000) 1-10.
- [2] J. Buckmaster, P. Clavin, A. Liñán, M. Matalon, N. Peters, G. Sivashinsky, F.A. Williams, *Proc. Combust. Inst.* 30 (2005) 1-19.
- [3] J.A. Miller, M.J. Pilling, J. Troe, *Proc. Combust. Inst.* 30 (2005) 43-88.
- [4] K. Kohse-Höinghaus, R.S. Barlow, M. Aldén, J. Wolfrum, *Proc. Combust. Inst.* 30 (2005) 89-123.
- [5] C.K. Westbrook, Y. Miropuchi, T.J. Poinso, P.J. Smith, J. Warnatz, *Proc. Combust. Inst.* 30 (2005) 125-157.
- [6] C.K. Westbrook, *Proc. Combust. Inst.* 19 (1982) 127-141.
- [7] M. Frenklach, K. Kailasanath, E.S. Oran, in: J.R. Bowen, J.C. Leyer, R.I. Soloukhin (Eds.), *Dynamics of Reactive Systems Part II: Modeling and Heterogeneous Combustion*. Am. Inst. Aeronautics Astronautics, Washington, D.C., 1986, pp. 365-376.
- [8] T. Turányi, T. Bérces, S. Vajda, *Int. J. Chem. Kinet.* 21 (1989) 83-99.
- [9] M. Frenklach, in: E.S. Oran, J.P. Boris (Eds.), *Numerical Approaches to Combustion Modeling*. American Institute of Aeronautics and Astronautics, Washington, D.C., 1991, pp. 129-154.
- [10] H. Wang, M. Frenklach, *Combust. Flame* 87 (1992) 365-370.
- [11] L.R. Petzold, W. Zhu, *Am. Inst. Chem. Eng. J.* 45 (1999) 869-886.
- [12] T. Løvås, D. Nilsson, F. Mauss, *Proc. Combust. Inst.* 28 (2000) 1809-1815.
- [13] W.H. Green, P.I. Barton, B. Bhattacharjee, D.M. Matheu, D.A. Schwer, J. Song, R. Sumathi, H.H. Carstensen, A.M. Dean, J.M. Grenda, *Ind. Eng. Chem. Res.* 40 (2001) 5362-5370.

- [14] B. Bhattacharjee, D.A. Schwer, P.I. Barton, W.H. Green, *Combust. Flame* 135 (2003) 191-208.
- [15] T. Lu, C.K. Law, *Proc. Combust. Inst.* 30 (2005) 1333-1341.
- [16] S. Mosbach, H. Su, M. Kraft, *Proc. Combust. Inst.* 30 (2005) 1301-1308.
- [17] J.C. Keck, D. Gillespie, *Combust. Flame* 17 (1971) 237-241.
- [18] N. Peters, F.A. Williams, in: J. Warnatz, W. Jäger (Eds.), *Complex Chemical Reaction System, Mathematical Modelling and Simulation*. Springer-Verlag, Berlin, 1987, pp. 310-317.
- [19] S. Vajda, P. Valko, T. Turányi, *Int. J. Chem. Kinet.* 17 (1985) 55-81.
- [20] J.Y. Chen, *Combust. Sci. Technol.* 57 (1988) 89-94.
- [21] U. Maas, S.B. Pope, *Combust. Flame* 88 (1992) 239-264.
- [22] S.H. Lam, D.A. Goussis, *Int. J. Chem. Kinet.* 26 (1994) 461-486.
- [23] M. Frenklach, *Chem. Eng. Sci.* 40 (1985) 1843-1849.
- [24] M. Frenklach, *Chem. Eng. Sci.* 57 (2002) 2229-2239.
- [25] E. Ranzi, M. Dente, A. Goldaniga, G. Bozzano, T. Faravelli, *Prog. Energy Combust. Sci.* 27 (2001) 99-139.
- [26] T. Løvås, F. Mauss, C. Hasse, N. Peters, *Proc. Combust. Inst.* 29 (2002) 1403-1410.
- [27] W.P. Jones, S. Rigopoulos, *Proc. Combust. Inst.* 30 (2005) 1325-1331.
- [28] H. Huang, M. Fairweather, J.F. Griffiths, A.S. Tomlin, R.B. Brad, *Proc. Combust. Inst.* 30 (2005) 1309-1316.
- [29] K.N.C. Bray, N. Peters, in: P.A. Libby, F.A. Williams (Eds.), *Turbulent Reacting Flows*. Academic, San Diego, CA, 1994, pp. 63-113.

- [30] A.R. Marsden, M. Frenklach, D.D. Reible, *J. Air Pollution Control Assoc.* 37 (1987) 370-376.
- [31] T. Turányi, *Comput. Chem.* 18 (1994) 45-54.
- [32] T. Turányi, *Proc. Combust. Inst.* 25 (1995) 948-955.
- [33] S.B. Pope, *Combust. Theory Modeling* 1 (1997) 41-63.
- [34] B. Yang, S.B. Pope, *Combust. Flame* 112 (1997) 17-32.
- [35] S.R. Tonse, N.W. Moriarty, N.J. Brown, F. M., *Isr. J. Chem.* 39 (1999) 97-106.
- [36] B. Yang, S.B. Pope, *Combust. Flame* 112 (1998) 85-112.
- [37] M. Frenklach, in: K.J. Bathe (Ed.) *Computational Fluid and Solid Mechanics*. New York, Elsevier, 2001, pp. 1177-1179.
- [38] Z. Ren, S.B. Pope, A. Vladimirovsky, J.M. Guckenheimer, *J. Chem. Phys.* 124 (2006) 114111.
- [39] A. Saltelli, K. Chan, E.M. Scott (Eds.), *Sensitivity Analysis*. Wiley, Chichester, England, 2000.
- [40] M. Frenklach, in: W.C. Gardiner, Jr. (Ed.) *Combustion Chemistry*. Springer-Verlag, New York, 1984, pp. 423-453.
- [41] Z. Zhao, J. Li, A. Kazakov, F. Dryer, *Int. J. Chem. Kinet.* 37 (2005) 282-295.
- [42] J.J. Scire, Jr., F.L. Dryer, R.A. Yetter, *Int. J. Chem. Kinet.* 33 (2001) 784-802.
- [43] I.G. Zsély, J. Zádor, T. Turányi, *J. Phys. Chem A* 107 (2003) 2216-2238.
- [44] K. König, U. Maas, *Proc. Combust. Inst.* 30 (2005) 1317-1323.
- [45] N.J. Brown, K.L. Revzan, *Int. J. Chem. Kinet.* 37 (2005) 538-553.
- [46] Y. Dong, A.T. Holley, M.G. Andac, F.N. Egolfopoulos, S.G. Davis, P. Middha, H. Wang, *Combust. Flame* 142 (2005) 374-387.

- [47] R. Feeley, P. Seiler, A. Packard, M. Frenklach, *J. Phys. Chem. A* 108 (2004) 9573-9583.
- [48] M.T. Reagan, H.N. Najm, P.P. Pébay, O.M. Knio, R.G. Ghanem, *Int. J. Chem. Kinet.* 37 (2005) 368-382.
- [49] J. Zádor, I.G. Zsély, T. Turányi, M. Ratto, S. Tarantola, A. Saltelli, *J. Phys. Chem A* 109 (2005) 9795-9807.
- [50] I.G. Zsély, J. Zádor, T. Turányi, *Proc. Combust. Inst.* 30 (1994) 1273-1281.
- [51] I.G. Zsély, J. Zádor, T. Turányi, *30th International Symposium on Combustion*, Chicago, IL, 2004, pp. Poster 1F2-11.
- [52] I.G. Zsély, J. Zádor, T. Turányi, *Combust. Theory Modeling* 9 (2005) 721-738.
- [53] For clarity in presentation the notation and terminology are somewhat simplified, and several remarks are in order. The “rate parameter k ” could represent the value of k , if k is constant, or parameters of the rate coefficient expression, like the pre-exponential factor or activation energy. The uncertainties in k represented by $k \pm \Delta k$ can be asymmetric, with the explicit notation introduced later, in Section 4. The “hypercube” is actually a hyper-parallelogram in space \mathbf{k} , with uneven sides in different dimensions of \mathbf{k} ; yet, expressed in factorial representation, it is a hyper-cube, as each of its sides is scaled to the respective interval of variation in k .
- [54] M. Frenklach, A. Packard, P. Seiler, R. Feeley, *Int. J. Chem. Kinet.* 36 (2004) 57-66.
- [55] M. Frenklach, A. Packard, R. Feeley, in: R.W. Carr (Ed.) *Modeling Chemical Kinetics*. 2006, to be published, pp.
- [56] M. Frenklach, H. Wang, M.J. Rabinowitz, *Prog. Energy Combust. Sci.* 18 (1992) 47-73.
- [57] B. Eiteneer, C.-L. Yu, M. Goldenberg, M. Frenklach, *J. Phys. Chem. A* 102 (1998) 5196-5205.

- [58] D.M. Golden, G.P. Smith, A.B. McEwen, C.-L. Yu, B. Eiteneer, M. Frenklach, G.L. Vaghjiani, A.R. Ravishankara, F.P. Tully, *J. Phys. Chem. A* 102 (1998) 8598-8606.
- [59] D.M. Golden, J.A. Manion, *Advances in Chemical Kinetics and Dynamics* 1 (1992) 187-276.
- [60] G.P. Smith, D.M. Golden, M. Frenklach, N.W. Moriarty, B. Eiteneer, M. Goldenberg, C.T. Bowman, R. Hanson, S. Song, W.C. Gardiner, Jr., V. Lissianski, Z. Qin, http://www.me.berkeley.edu/gri_mech/.
- [61] C.K. Westbrook, F. Dryer, *Proc. Combust. Inst.* 18 (1981) 749-766.
- [62] C.K. Westbrook, F. Dryer, *Prog. Energy Combust. Sci.* 10 (1981) 1-57.
- [63] R. Sumathi, W.H. Green, *Theor. Chem. Acc.* 108 (2002) 187-213.
- [64] J.P. Senosiain, C.B. Musgrave, D.M. Golden, *J. Phys. Chem. A* 105 (2001) 1669-1675.
- [65] G.E.P. Box, R.D. Meyer, *J. Res. Nat. Bur. Stand.* 90 (1985) 494-496.
- [66] R.H. Myers, D.C. Montgomery, *Response Surface Methodology: Process and Product Optimization Using Designed Experiments*. Wiley, New York, 2002.
- [67] D. Miller, M. Frenklach, *Int. J. Chem. Kinet.* 15 (1983) 677-696.
- [68] G.E.P. Box, W.G. Hunter, J.S. Hunter, *Statistics for Experimenters. An Introduction to Design, Data Analysis, and Model Building*. Wiley, New York, 1978.
- [69] G.E.P. Box, N.R. Draper, *Empirical Model-Building and Response Surfaces*. Wiley, New York, 1987.
- [70] R.R. Cao, S.R. Pope, *Combust. Flame* 143 (2005) 450-470.
- [71] J.B. Bell, M.S. Day, I.G. Shepherd, M.R. Johnson, R.K. Cheng, J.F. Grcar, V.E. Beckner, M.J. Lijewski, *Proc. Nat. Acad. Sci.* 102 (2005) 10006-10011.
- [72] L. Wang, D.C. Haworth, S.R. Turns, M.F. Modest, *Combust. Flame* 141 (2005) 170-179.

- [73] W. Tang, L. Zhang, A.A. Linninger, R.S. Tranter, K. Brezinsky, *Ind. Eng. Chem. Res.* 44 (2005) 3626-3637.
- [74] A.B. Singer, J.W. Taylor, P.I. Barton, W.H. Green, *J. Phys. Chem. A* 110 (2006) 971-976.
- [75] B.D. Phenix, J.L. Dinaro, M.A. Tatang, J.W. Tester, J.B. Howard, G.J. McRae, *Combust. Flame* 112 (1998) 132-146.
- [76] B. Ruscic, R.E. Pinzon, M.L. Morton, G. von Laszewski, S.J. Bittner, S.G. Nijssure, K.A. Amin, M. Minkoff, A.F. Wagner, *J. Phys. Chem. A* 108 (2004) 9979-9997.
- [77] M. Frenkel, R.D. Chirico, V. Diky, X. Yan, Q. Dong, C. Muzny, *J. Chem. Inf. Model.* 45 (2005) 816-838.
- [78] D.R. Stull, H. Prophet, *JANAF Thermochemical Tables*. U.S. Nat. Bur. Stand., Washington, D.C., 1971.
- [79] D.L. Baulch, C.J. Cobos, R.A. Cox, P. Frank, G. Hayman, T. Just, J.A. Kerr, T. Murrells, M.J. Pilling, J. Troe, R.W. Walker, J. Warnatz, *J. Phys. Chem. Ref. Data* 23 (1994) 847-1033.
- [80] S.P. Sander, R.R. Friedl, D.M. Golden, M.J. Kurilo, R.E. Huie, V.L. Orkin, G.K. Moortgat, A.R. Ravishankara, C.E. Kolb, M.J. Molina, B.J. Finlayson-Pitts, <http://jpldataeval.jplnasa.gov>.
- [81] <http://webbook.nist.gov>.
- [82] R.S. Barlow, <http://www.ca.sandia.gov/TNF/>.
- [83] J. Tennenbaum, http://american_almanac.tripod.com/mendel94.htm.
- [84] R. Pool, J. Esnayra, *Bioinformatics: Converting Data to Knowledge*. National Research Council, National Academies, Washington, D.C., 2000.
- [85] D.E. Post, L.G. Votta, *Physics Today* 58 (2005) 35-41.

- [86] P. Cochrane, *Uncommon Sense*. Capstone, West Sussex, 2004.
- [87] H. Hippler, *Phys. Chem. Chem. Phys.* 4 (2002) vii-11.
- [88] <http://primekinetics.org>.
- [89] M. Frenklach, A. Packard, P. Seiler, *Proceedings of the American Control Conference*. 2002, pp. 4135-4140.
- [90] P. Seiler, M. Frenklach, A. Packard, R. Feeley, *Optim. Eng.* (2006) in press.
- [91] R. Feeley, M. Frenklach, M. Onsum, T. Russi, A. Arkin, A. Packard, *J. Phys. Chem. A* (2006) in press.
- [92] K. Seashore Louis, L.M. Jones, E.G. Campbell, *Am. Sci.* 90 (2002) 304-307.
- [93] M. Frenklach, H. Wang, M. Goldenberg, G.P. Smith, D.M. Golden, C.T. Bowman, R.K. Hanson, W.C. Gardiner, V. Lissianski, *GRI-Mech—An optimized detailed chemical reaction mechanism for methane combustion*, The Gas Research Institute GRI-95/0058, 1995.
- [94] C.T. Bowman, R.K. Hanson, W.C. Gardiner, V. Lissianski, M. Frenklach, M. Goldenberg, G.P. Smith, *GRI-Mech 2.11—An Optimized Detailed Chemical Reaction Mechanism for Methane Combustion and NO Formation and Reburning*, The Gas Research Institute GRI-97/0020, 1997.
- [95] <http://cmcs.org>.
- [96] Z.M. Djuricic, D. Amusin, T. Berekyei, T.C. Allison, M. Frenklach, Fall Meeting of the Western States Section of the Combustion Institute, Stanford, CA, 2005, Paper 05F-43.

Figure Captions

Figure 1. The *Prolog* from NSF proposal “Collaborative Research: Cyberinfrastructure and Research Facilities: Process Informatics for Chemical Reaction Systems” by M. Frenklach, A. Packard, C. T. Bowman, D. Golden, W. H. Green, and G. McRae, 2005.

Figure 2. An illustration of a two-dimensional rate-constant uncertainty region. The shaded area is the feasible set.

Figure 3. A feasible set obtained in a joint analysis of two GRI-Mech 3.0 experimental targets (E_{66} and E_{67}) with all rate constants but k_{44} , k_{45} , and k_{34} set to the respective literature values [54].

Figure 4. Pairwise test [47]: Threshold uncertainty levels u_{ef} are calculated for each $D_{ef} = \{U_e, U_f\}$ pair of the GRI-Mech dataset units. The highest peak is $u_{57,58}$, flagging U_{57} and U_{58} as possible outliers (see [47] for further details).

Figure 5. Information loss without Data Collaboration [54,55].

Sometime in the year 2008 ...

| | |
|-------------------|--|
| Chemist to PrIME: | I have an idea of how to measure the elusive reaction between $C_{14}H_7$ and C_3H_3 forming $C_{16}H_8$ and CH_2 . What impact would such a measurement have on present knowledge concerning the nucleation of interstellar dust? |
| PrIME to Chemist: | If the rate coefficient is established to within 3% accuracy, I will be able to discriminate between two competing hypothesis, A and B. |
| Chemist to PrIME: | I do not think my experiment can attain better than 10% accuracy. What is the next best thing can I do experimentally to advance knowledge of this subject? |
| PrIME to Chemist: | Measure the reaction between $C_{10}H_7$ and C_3H_2 ; I can then discriminate between hypotheses B and C. |

... in the year 2010 ...

| | |
|-----------------------------------|---|
| Engineer to PrIME: | What fueling rate produces peak output power while holding NO_x yields within the EPA prescribed limits in a HCCI engine running GTL prescribed fuel #22 with the following design and operating parameters: xx,yy, ... |
| PrIME to Engineer: | How well do you want to know this? |
| Engineer to PrIME: | I need 5% accuracy! |
| PrIME to Engineer: | This accuracy is not achievable. The uncertainty range on the fueling rate runs from 1.22 to 1.35 g/s. |
| Engineer to PrIME: | What is the dominant source of this uncertainty? |
| PrIME to Engineer: | 80% of the uncertainty in the predicted fuel consumption rate is caused by the uncertainty in the estimated rate constant for the reaction $(CH_3)_2CCHCH_2 + O_2 \rightarrow (CH_3)_2CCCH_2 + HO_2$. There are no literature data on this reaction, only indirect estimates. The needed data can be obtained via quantum chemistry calculations (time 2 days, cost \$\$) or by performing a series of experiments (time 2 years, cost \$\$\$\$); what is your choice? |
| Engineer to PrIME: | The quantum chemistry calculations right away and start experiments as well. |
| PrIME to Engineer (2 days later): | Optimal fueling rate needed is 1.31 g/s, based on computed rate constant of 5×10^9 cc/mol s. Prediction uncertainty range is 1.29 - 1.33 g/s. |
| PrIME to Chemists: | Perform measurements to determine the rate constant for the reaction $(CH_3)_2CCHCH_2 + O_2 \rightarrow (CH_3)_2CCCH_2 + HO_2$. |

... and in the year 2020 ...

| | |
|---|---|
| Policymaker to PrIME: | How much longer will there be an Antarctic ozone hole? |
| PrIME to Policymaker: | 50 to 150 years. |
| Policymaker to PrIME: | Can I get the answer more accurately, within at worst a 5 year interval, and what would it take? |
| PrIME to Policymaker: | My result is the best current estimate based on all available data and the scientific community consensus. To get to the requested level of certainty the heat of formation of Cl_2O_2 has to be known to within 1 kJ/mol and the rate coefficient for the reaction between two ClO radicals to within 10%. |
| Policymaker to Chemists (via funding agency): | Please (re)measure/(re)calculate the heat of formation of Cl_2O_2 and the rate coefficient for the reaction between two ClO radicals to the above accuracy. |

Fig. 1

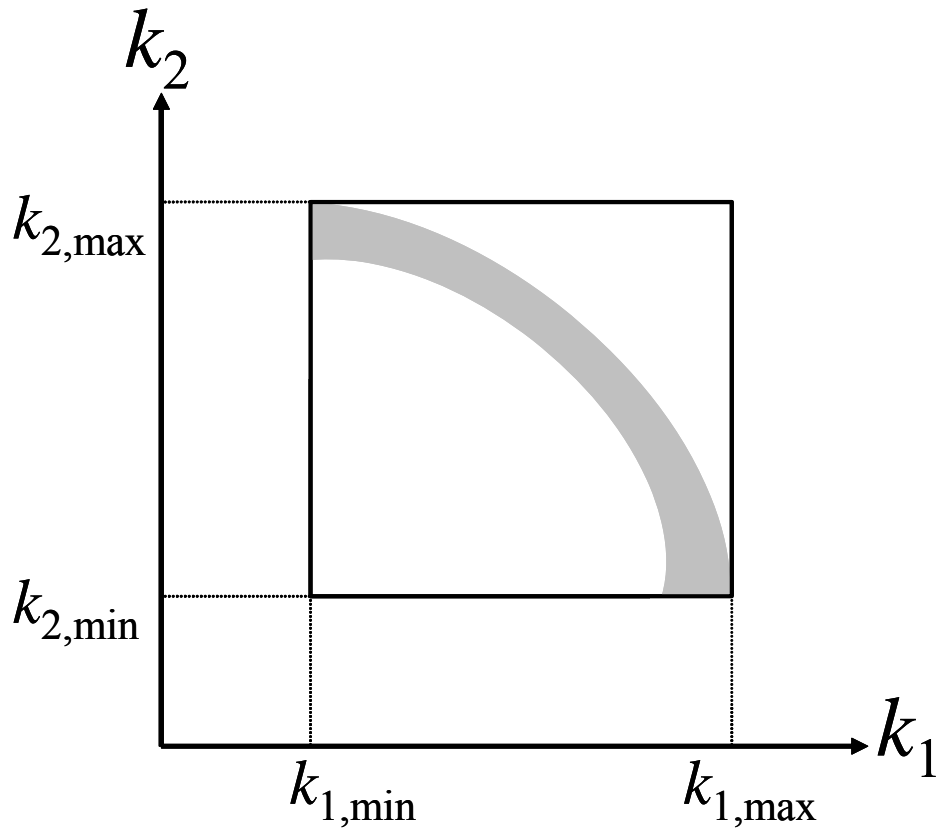


Fig. 2

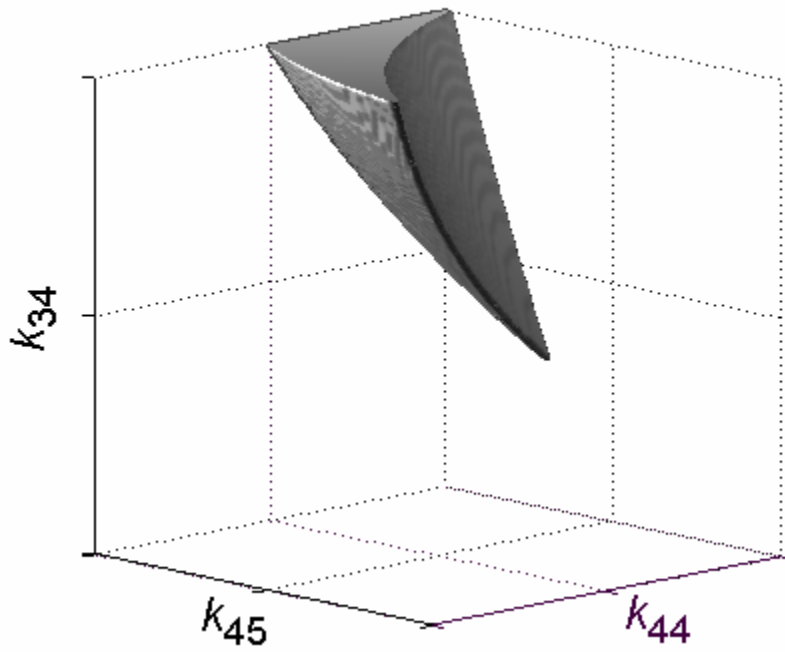


Fig. 3

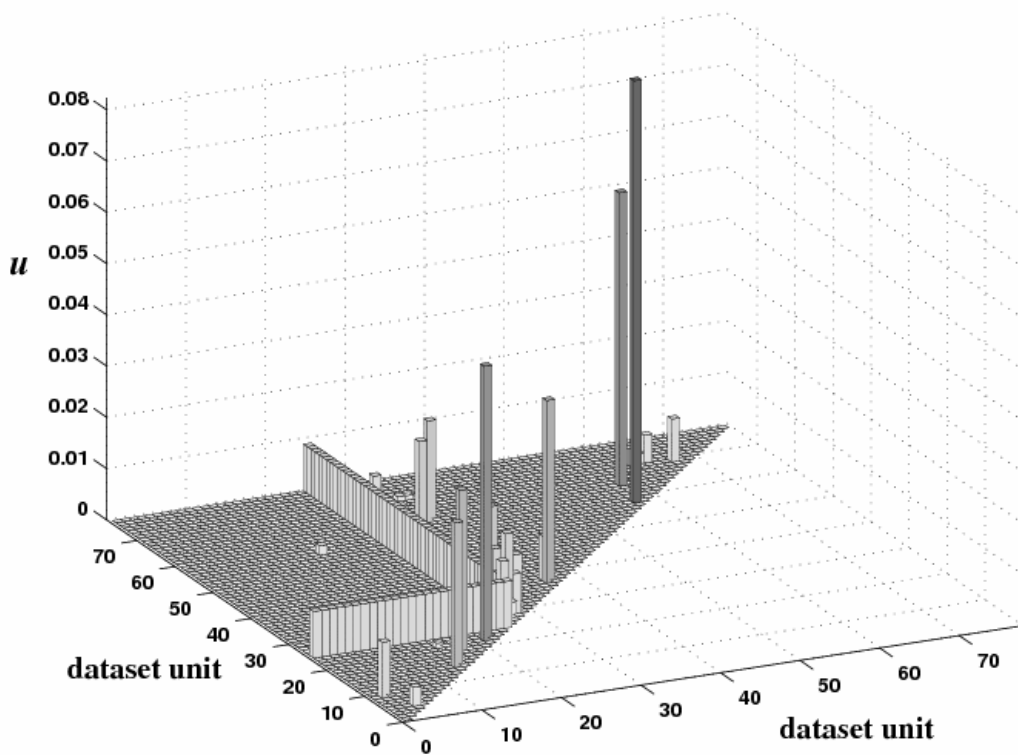


Fig. 4

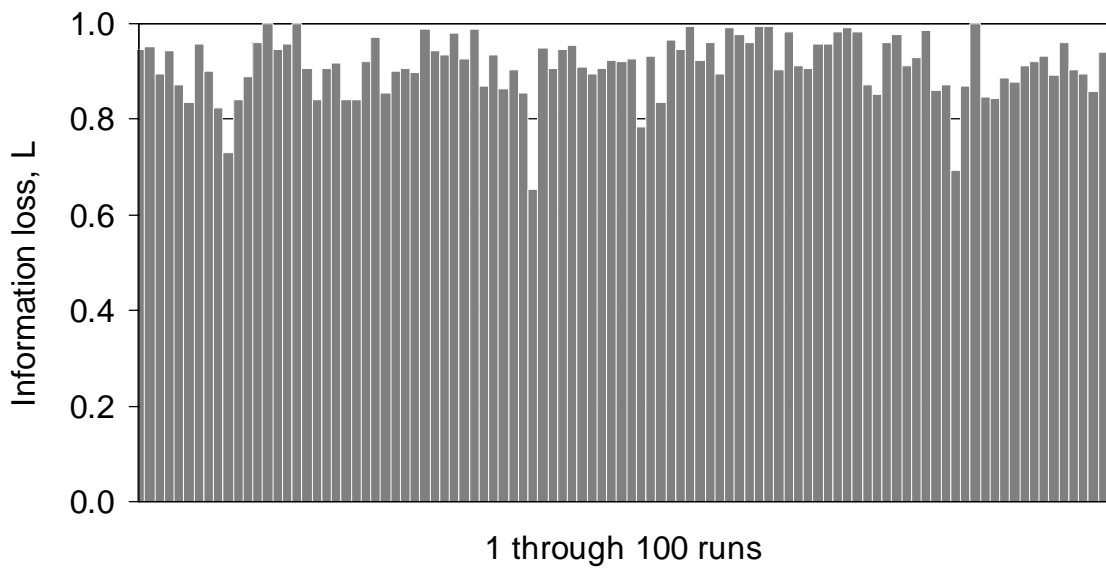


Fig. 5