PREDICTIVE KINETICS: A NEW APPROACH FOR THE 21ST CENTURY

William H. Green, Jr.[*]

Department of Chemical Engineering, Massachusetts Institute of Technology, Cambridge, MA 02139, USA

---

[*] Corresponding author. E-mail: balkwill@MIT.EDU

Advances in Chemical Engineering, vol. 32
ISSN 0065-2377
DOI 10.1016/S0065-2377(07)32001-2

Copyright © 2007 by Elsevier Inc.
All rights reserved.

Abstract

The capability to reliably predict the behavior of reactive chemical systems would allow rational *a priori* design of chemical reaction systems. Recent progress toward overcoming several technical obstacles to predictive kinetics for homogeneous gas-phase systems is reviewed. The focus is on (1) updates to the fundamental data model used in kinetic modeling, (2) methods for solving large kinetic simulations efficiently without relying on uncontrolled approximations, and (3) methods for determining whether or not the model predictions are consistent with experimental data. Appropriate handling and archiving of experimental data from different sources, and of the many uncertainties in the data embedded throughout the kinetic models, is a major challenge facing the kinetics community as kinetics becomes a predictive science.

I. Introduction

Chemical kinetic modeling has been one of the pillars of chemical engineering since the field was founded. In many respects, it is the aspect that distinguishes chemical engineering from other fields of engineering. However, despite a huge amount of work by thousands of researchers, we still have a very limited ability to predict what will happen in a newly proposed reactive system, even in well-characterized gas-phase systems with simple fluid mechanics. A number of technical problems significantly impeded progress in kinetics during the 20th century. Here we briefly review the history and the goals of chemical kinetic modeling, and highlight a few of the technical areas where significant advances have been made in the first years of the 21st century.

A. The Historical Applications of Kinetic Modeling

1. *Predictive vs. Postdictive Models*

Chemical kinetic models play multiple roles. In the 20th century, these models were most often used to assist in the interpretation of experimental data, and for the interpolations used in process optimization and control. Kinetic models also played an essential role in the process of generalizing new chemical knowledge from experimental results. In both of these applications, a large amount of experimental data were measured first and then kinetic models were used to rationalize or interpolate between these experimental data. It would not be too unfair to characterize most of this work as "postdictive", i.e. the modelers were not able to say anything interesting until *after* a considerable amount of experimental work had been completed.

Of course, chemical kinetic models would be even more useful if they could accurately ***predict*** the behavior of reacting systems under conditions significantly different from those that have already been measured. If these extrapolative predictions were accurate enough, chemical kinetic models could become valuable tools in process and product design, and by reducing the need to do so many experiments in order to gain a small amount of information, the models could accelerate the pace of innovation. Reliable predictive kinetic models would be particularly helpful in situations where it is impractical to do the experiments, e.g. in the public policy arena, where a failed experiment could be prohibitively costly, or in situations where the experiment is impossible (e.g. predicting what happens in very slow or very fast processes).

In recent years, it has become possible to extrapolate accurately using detailed chemical kinetic models to predict quantitatively the behavior of some rather complicated chemical systems. The most famous examples of this success are the detailed atmospheric chemistry models whose predictions underlie the Montreal Protocol on ozone-depleting chemicals. However, these atmospheric chemistry models were developed through a huge international effort over several decades, based heavily on a large number of laboratory experiments. Much more rapid and efficient methods of model development are required for detailed predictive chemical kinetics to become a practical everyday design tool for chemical engineering.

2. *Technical Hurdles in the 20th Century*

Several technical issues significantly hampered progress in chemical kinetic modeling in the 20th century; among these were serious difficulties solving kinetic simulations numerically, difficulties constructing and checking (and peer-reviewing) large simulations, problems with estimating reaction rates when thermochemistry and experimental data were missing, and problems checking whether model predictions were consistent with experimental data.

First, significant numerical problems plagued the kinetic models. Almost all chemical kinetic systems are intrinsically "stiff", with a large separation of timescales between the most reactive and least reactive species in a mixture. Although a number of clever approximations were used to circumvent this problem, the situation was highly unsatisfactory (e.g. it was very difficult to prove that the approximations were valid) until William Gear invented the first algorithm capable of numerically solving stiff systems of differential equations with error control in the 1970s (Gear, 1971). Also, most chemical kinetic systems of interest to engineers involve a relatively large number of species, which made them hard to handle until computers with significant core memory became widely available ~1980. The development of the CHEMKIN® software in the 1980s, which dealt with the most common data-handling issues and also incorporated a reliable stiff differential equation solver that evolved from Gear's algorithm, was a major step-forward in chemical kinetic modeling (Kee *et al.*, 1989). Toward the end of the last century, modelers benefited tremendously from the rapid improvements in computer hardware and from major advances in the capability to solve multidimensional, multi-species kinetic simulations. But there are still many important reactor types whose simulations are plagued by numerical issues and by hardware limitations; at present there are only a handful of well-tested widely distributed solvers with error control, and these solvers work only for specific idealized reactor types.

As the numerical solvers have become capable of solving rather complicated chemical kinetic simulations, modelers have had to accept that we do not really know the chemistry of most processes as well as one might hope, and in some cases the reacting mixtures are so complicated that it is doubtful anyone could write down all the reactions occurring, even if we did know all the chemistry. (For complicated reactor geometries, there is a similar problem associated with constructing a suitable mesh for solving the transport equations.) The problem of complex chemistry was recognized even before CHEMKIN became widely available, and has led to the development of dozens of automated mechanism-construction programs over the past two decades (Blurock, 2004a, b; Broadbelt *et al.*, 1994; Chevalier *et al.*, 1990; Chinnick *et al.*, 1988; Dente *et al.*, 1979; DiMaio and Lignola, 1992; Hillewaert *et al.*, 1988; Kojima, 1994; Matheu *et al.*, 2003a; Prickett and Mavrovouniotis, 1997a, b, c; Quann and Jaffe, 1992; Ranzi *et al.*, 1995; Tomlin *et al.*, 1997; Warth *et al.*, 2000; Zarth *et al.*, 2002). These programs have been successful in modeling some systems, but as discussed below there are many significant unresolved technical issues which have prevented these programs from having a broad impact on the field. A very important problem is that most of the model-construction software is not sufficiently flexible; in particular, it is rather hard to add additional chemistry and modify the rate parameters as knowledge improves. As a result, the software can easily become obsolete. Also, it is very difficult to check or peer-review large complicated simulations; this combined with the tendency for postdictive modeling makes many researchers skeptical about the usefulness of large simulations.

For most of the 20th century, the only solid basis for assigning a numerical value to a rate parameter was experimental data. However, in most experimental systems, the number of parameters determinable from the experimental rate data is much less than the number of rate parameters needed in the chemical kinetic model. Benson (1976) and others proposed methods to estimate some rate parameters, (primarily by analogy) to reactions whose rate constants had been measured, but there were still a large number of reactions that could not be estimated at all, and it was hard to know how confidently one could use the analogy-based estimates. Of necessity, researchers developing large kinetic models developed rules-of-thumb to estimate reaction rates for the reaction families of greatest importance in the systems they were simulating, but this was necessarily very *ad hoc* in the absence of any data with which to check the estimates. In the last decade of the 20th century, computation of reaction rates (and thermochemistry) using quantum chemistry became feasible, making it possible to check the accuracy of the analogy-based estimates. With the quantum chemistry techniques available at that time, this was most feasible for thermochemistry, and several researchers, notably Melius (Melius and Allendorf, 2000; Zachariah and Melius, 1998) and Bozzelli (Lay and Bozzelli, 1997a, b; Lay *et al.*, 1997) and their collaborators, used quantum chemical calculations to significantly improve the quality of thermochemical (and some rate) estimates used in combustion models. By the end of the 20th century, similar techniques were being widely used to improve barrier height and A-factor estimates, usually using transition state theory (Truhlar *et al.*, 1996). Sumathi and Green (2002) reviewed the state of the art in rate estimation for large-scale kinetic modeling around the turn of the century.

Finally, there are multiple hurdles that made it difficult to appropriately compare kinetic model predictions with experimental data in the 20th century, and which still make it very difficult today. As discussed in Section IV below, there are some numerical issues which make it hard to determine definitively whether or not a model is consistent with a data set. If the model is consistent, one should be able to refine the model parameters using the experimental data, but this is numerically challenging when one is working with large-scale non-linear kinetic simulations and large data sets. Response surface methods were used successfully for this parameter refinement step toward the end of the 20th century (Aghalayam *et al.*, 2000; Frenklach *et al.*, 1992).

A bigger problem than the numerical issues is the difficulty in coming up with the data and the error bars needed to make a meaningful comparison between the model and experiment. Very often one can no longer locate all the original experimental data in order to make a comprehensive comparison, and even the data one can find seldom comes with reliable error bars. Also, researchers rarely compute the error bars in a large chemical kinetic model's predictions. So, although thousands of kinetic model vs. data comparisons were published in the 20th century, in most cases these plots are insufficient to determine conclusively whether the model is really consistent with the data, and it would usually be difficult-to-impossible to determine the error bars on any of the model parameters based on the published comparison.

With this history of difficulties in mind, in the next section we reconsider what are the real goals of chemical kinetic modeling.

B. The Goal of Chemical Kinetic Modeling

The primary goal of chemical kinetic modeling is to make predictions: given our current understanding of chemistry, what do we expect to happen in a particular reacting mixture under specified reaction conditions? If desired, these

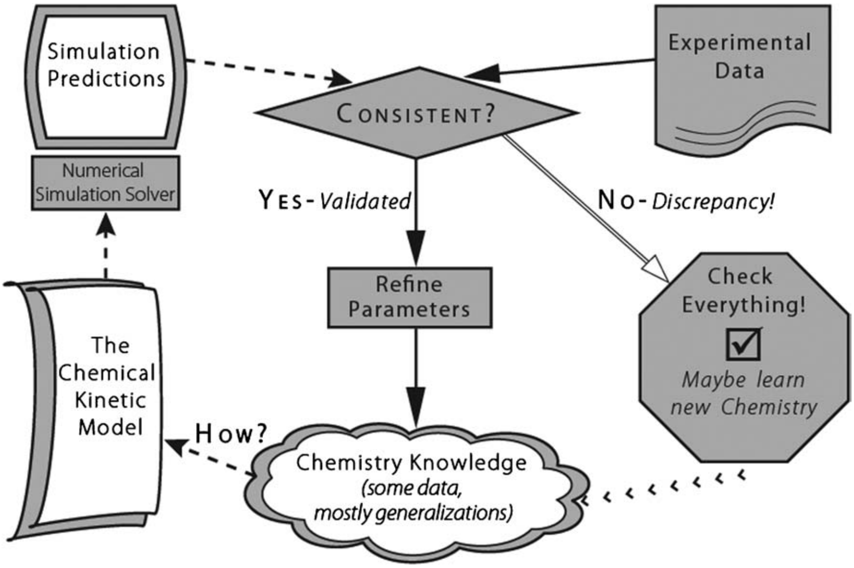


FIG. 1. The Model-Prediction-Data loop. This article is focused on the steps on the left side, from model-construction through consistency testing.

model predictions can then be quantitatively compared with experimental data, to determine if the two are consistent, Fig. 1. If the model and the data are inconsistent, at least one of them must be wrong, and we have identified an interesting opportunity to learn something new. If the model and the data are consistent within the error bars on each, then we say the model has been *validated* (never *proven*), and the experimental data can be used to refine the model parameters.

To the extent that we have confidence in the predictions of chemical kinetic models, we could then use these models to design new products, processes, and reactors to plan new experiments, to design model-based control systems and safety systems, and to inform critical business and public policy decisions. However, to develop the required level of confidence in the model predictions, the loop in Fig. 1 must be functioning effectively.

In the present paper we focus on a few of the steps involved in the left side of the loop shown in Fig. 1:

(1) Construction of predictive chemical kinetic models
(2) Efficient and accurate numerical solution of chemical kinetic simulations, and
(3) Quantitative tests whether the model and data are consistent.

Each of these three steps is addressed in a separate section below.

The right-hand half of the loop is, of course, equally important, and many of the difficult issues involved in moving from experimental kinetics data to chemical knowledge have been discussed in the literature by others (Feeley *et al.*, 2004; Feng and Rabitz, 2004; Milford *et al.*, 1992; Shenvi *et al.*, 2002; Tomlin *et al.*, 1997). The model should be tested for consistency against all relevant data, but there are significant difficulties assembling all the data in a form suitable for making this test, an issue being addressed by the PrIMe project (http://www.primekinetics.org). Combining data from different experiments can greatly increase the amount of information gained (Feeley *et al.*, 2004). For some nice recent examples of the practical value of combining different data from different experiments, observe how the Active Tables approach has revolutionized determinations of high-accuracy molecular thermochemistry (Ruscic *et al.*, 2004).

II. Construction of Predictive Chemical Kinetic Models

Constructing an accurate chemical kinetic model is quite a challenging task, both because “Chemistry Knowledge” is very large and somewhat amorphous, and because it often requires hundreds of differential equations and thousands of numerical parameters (rate constants, molecular thermochemistry, etc.) to accurately describe the details of the processes occurring in a reacting mixture. Constructing these models by hand is a very time-consuming and error-prone process, so over the past three decades many research groups have developed software to automate the process of constructing chemical kinetic simulations (Blurock, 2004a, b; Broadbelt *et al.*, 1994; Chevalier *et al.*, 1990; Dente *et al.*, 1979; Kojima, 1994; Matheu, 2003; Prickett and Mavrovouniotis, 1997a, b, c; Quann and Jaffe, 1992; Tomlin *et al.*, 1997; Warth *et al.*, 2000; Zarth *et al.*, 2002).

In brief, model-construction software works by applying reaction operators to a set of molecules, to generate new product molecules. The reaction operators break and make bonds following some pre-defined motif. Each time a new chemical reaction is constructed by this process, the corresponding rate parameters are estimated, drawing on some library of rate estimation parameters (e.g. Arrhenius parameters, or Evans–Polanyi parameters). The thermochemistry of each species is also estimated, drawing on another library of estimation parameters (e.g. Benson group values). Often there are some experimental data which can be used instead of estimates, and these can be drawn from another library. The reaction operators are then applied to the product species to generate byproducts and so on iteratively.

Computer memory limitations usually require that one terminate the iterative process before $\sim 10^6$ reactions have been generated. Several criteria have been used for terminating the model-construction process; this author prefers the rate-based species-selection method (Susnow *et al.*, 1997). It should be noted that different model-construction software will construct different models for the same physical situation, both because different reactions/species will be included in the models and also because different estimates are made for the various rate and thermochemical parameters. Also model-construction software will often build different models for the same reactant(s) reacting under different conditions, since very different reactions might dominate under different conditions. Rather than having some gigantic model which includes everything that could take place, real-world models are quite finite, and so have limited ranges where they are accurate or valid.

Many of the technical issues involved in computer-aided model-construction have previously been reviewed by Tomlin *et al.* (1997). Several researchers, most notably Bozzelli, have extended Benson's method for estimating molecular thermochemistry using quantum chemistry (Lay and Bozzelli, 1997a, b; Lay *et al.*, 1995). Sumathi and Green (2002) have discussed how quantum chemistry can supplement experiments in developing rate estimation. Matheu *et al.* have shown how to automate the computation of rates of chemically-activated (pressure-dependent) reactions (Matheu, 2003; Matheu *et al.*, 2003a, b). Here we focus on a few issues which have not been so thoroughly discussed in the literature:

(1) The challenge of documenting large chemical kinetic simulations
(2) A data model for chemical reactions and molecular properties that facilitates extensibility.

A. The Challenge of Documenting Large Simulations

Because the "Chemistry Knowledge" input, the procedure for developing the correct list of differential equations corresponding to this input, and the estimates of the numerical values of the rate and molecular parameters are all known imperfectly, it is quite important that each aspect be carefully documented to facilitate scientific progress. Ideally, all the assumptions and simplifications made along the way from "Chemistry Knowledge" through to quantitative predictions, Fig. 1, should be clearly and precisely documented, so each can be tested, and each number in both the inputs and the outputs should be assigned an uncertainty band. (Ideally, the correlations between all these uncertainties should also be documented, but this has seldom been done so far.) Clarity about the assumptions and simplifications would facilitate identification of the root causes why different models give different predictions, and why predictions differ from experimental results. Clear understanding and documentation of the origins of these discrepancies will dramatically improve the efficiency of the world-wide efforts to improve both chemical understanding and kinetic modeling methodology.

There is more than just scientific progress at stake. Predictions based on chemical kinetic models are increasingly used to inform major policy and business decisions, often with large impacts on society, so it is critical that the uncertainties and assumptions associated with these predictions be clearly enunciated and understood. Clarity about our current level of ignorance is essential both to avoid misleading decision makers and to facilitate future work to test the assumptions and reduce the uncertainties in the predictions.

B. Data Models for Chemical Kinetics

Clarity in expressing information (including information about assumptions and uncertainties) can be greatly facilitated by using a good "data model". A "data model" is a standard way of expressing information; it includes both the format and the relationships between various data objects.

Currently, the most popular way to document a chemical kinetic model is to list the reactions and corresponding rate parameters (typically in a CHEMKIN format (Kee *et al.*, 1989)), and separately to list the molecular parameters including enthalpies and heat capacities (typically in a NASA polynomial format).

This popular data model, Fig. 2, has many positive aspects: it unambiguously expresses all the information necessary to construct and solve the differential equations, and it is flexible enough that it can be used for a broad range of chemical systems.

However, the popular List-of-Reactions data model has several serious deficiencies. First of all, it is difficult in this data model to adequately document the origins of all the numerical parameters used in the simulation, and the associated (usually highly correlated) uncertainties in these parameters. The format is also not very user-friendly: it is almost impossible for a human to check that all the numbers and reactions in a large kinetic model expressed in this format are consistent and reasonable, much less correct. The conventional List format does not allow for graphical representation of chemical structures, leading to sometimes serious inconsistencies in naming conventions.

More fundamentally, the current data model does not provide any way to document why some chemical reactions were included in the simulation, while others were left out. Typically, a large number of assumptions about which reactions/species are likely to be important (under the reaction conditions of interest to the simulation's author) are made by the person (or computer program) who assembles the list of species and reactions, based on his or her (or its) chemistry knowledge. Because the current data model does not provide any convenient way to document these assumptions, most of these assumptions are never documented at all, and this contributes to the impression that this information is not worth recording. The fact that it is usually very difficult to uncover complete information about the assumptions behind a reaction list

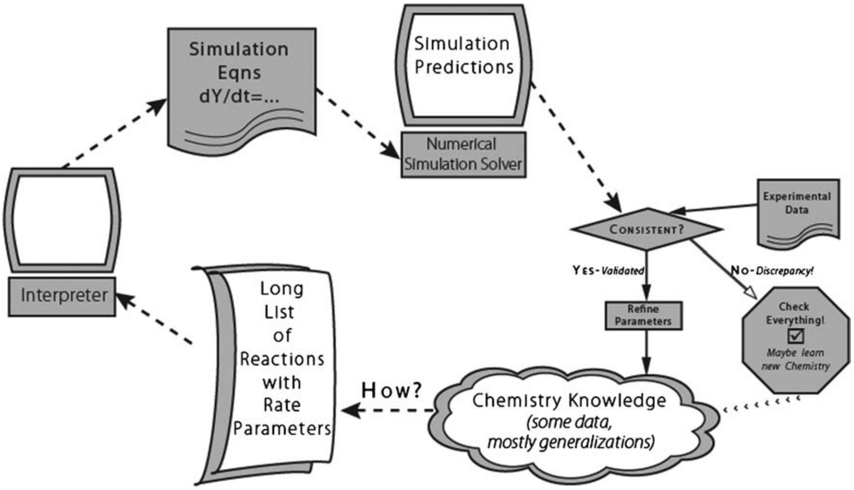


Fig. 2. Often the primary documentation of a kinetic model is the List-of-Reactions, and the steps from that List to the predictions are often well documented. However, the upstream steps that led to the List-of-Reactions are usually poorly documented.

discourages careful review of the critical steps "upstream" of the CHEMKIN-format file, Fig. 2.

It is important to realize that the CHEMKIN-format file really represents an intermediate step in the process of predicting chemical behavior. For large mechanisms, almost all of the parameter values in this data file are estimates that come from some simple (but not always documented, or consistently applied) rate estimation rule. Similarly, almost all of the numbers in the molecular property files are estimates and theoretical extrapolations. In large models, the thousands of NASA coefficients in the thermo file are typically derived from a much smaller set of Benson-type (Benson, 1976) group-additivity values. The molecular transport parameters are typically estimated even more approximately by rough analogy. So, although the typical CHEMKIN-type file contains thousands of numbers, the real information content is often much smaller: the numerical values in the simulation could usually be represented much more compactly in terms of group values and rate estimation parameters. Reducing the number of numerical parameters would greatly facilitate the important (but now rarely performed) task of comparing two competing kinetic models, to understand why they give different predictions.

Here we propose a fundamentally different data model, Fig. 3, where the crucial documentation lies upstream of the conventional List-of-Reactions. In this data model, the estimation rules for the rates and thermochemistry are the main data, supplemented by a set of numerical parameters for individual species and reactions (whose accurate values are known from experiments or

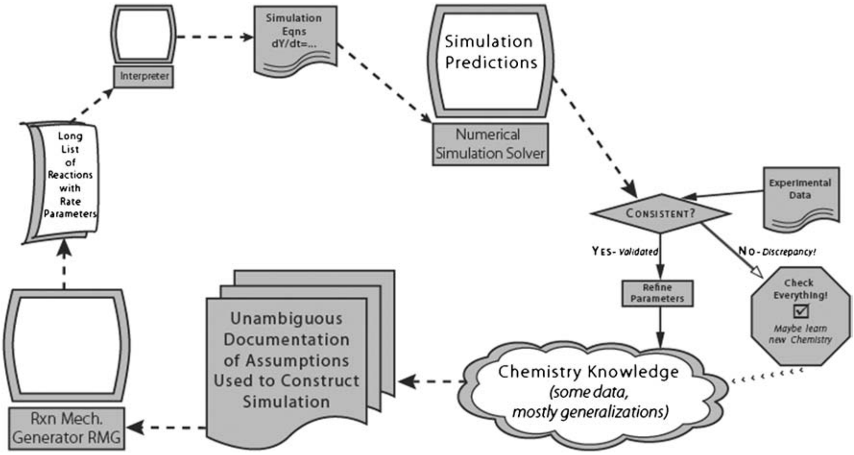


FIG. 3. In the proposed data model, the fundamental documentation is the information on how one estimates reaction rates (and molecular properties), and the other assumptions used to construct the simulation. All the subsequent steps in the process are automated, well-documented procedures.

high-quality quantum calculations). This new data model overcomes many of the objections to the List-of-Reactions data model it replaces. However, for this data model to be successful, it is necessary that unambiguous procedures, with clearly documented assumptions and tolerances, exist for converting the Estimation Rules into the corresponding List-of-Reactions, and thence (as is done currently) to Predictions. It is most convenient if these procedures can be performed automatically, using a computer.

C. AUTOMATED CONSTRUCTION OF LISTS-OF-REACTIONS

The awkwardness of the List-of-Reactions data model became apparent to many researchers modeling pyrolysis and combustion soon after CHEMKIN was developed, and since the 1980s many software packages have been developed to automate the process of constructing the long lists of reactions required, as reviewed by Tomlin *et al.* (1997). Many calculations have been published demonstrating that automated mechanism-construction can be successful. However, all of these software packages had significant flaws, many have been abandoned, and to our knowledge none of this software has yet been successfully distributed by its author to another research group.

To understand the problem, it is helpful to think about what the inputs are to these computer programs, i.e. what does one need to know in order to construct the appropriate List-of-Reactions chemistry model, and so correctly predict the behavior of a reacting system?

1. Fundamental Inputs for Chemical Kinetic Model-Construction

When one begins to construct a chemical kinetic model, there are several different types of required "input" information, Fig. 4. Obviously, one needs some specification of the initial concentrations of the reactants, and of the reaction conditions (e.g. T, P, timescale) of interest. Normally one wants to numerically solve the kinetic model to predict species and/or temperature profiles, so the inputs must also include some specification of numerical tolerances on these outputs, and options for the differential equation solver. The most complicated "input" information required to construct a kinetic model is the chemistry: what species, reactions, or reaction types will be considered? How will all the thermochemical and rate parameters be estimated?

A naïve answer is that one should just assemble the list of all the known species and all the known elementary-step reactions connecting them in the literature, and use all the literature values for the rate and thermochemical parameters. However, except in a very small number of very simple cases (e.g. H_2/O_2 combustion) this naïve approach is both inefficient and seriously inaccurate. First of all, unselectively constructing a kinetic model out of a very large set of reactions has the extremely undesirable effect of making it difficult or impossible for a human to understand/check/peer-review the model. Also, almost invariably, most of the reactions in any large compilation are unimportant under the specific reaction conditions of interest, so one is doing a lot more work both in checking and in solving the large model than is necessary. But the biggest problem is that in most branches of chemistry, only a very small fraction

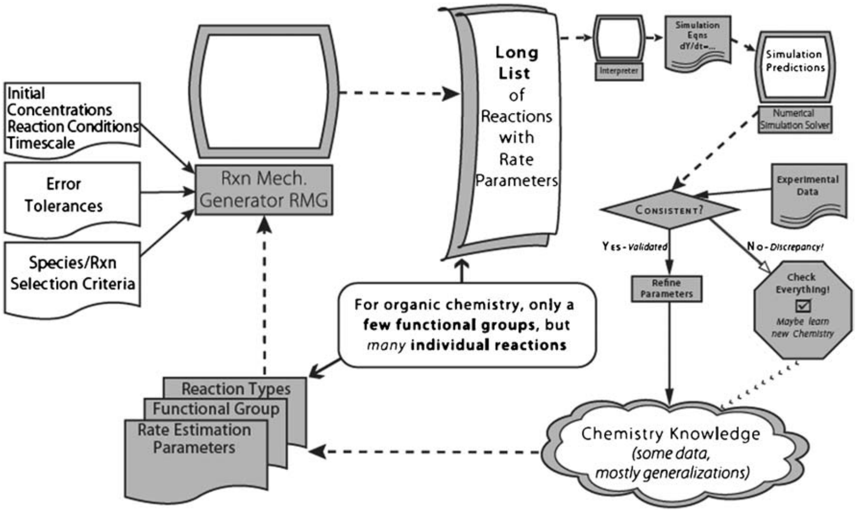


FIG. 4. The inputs needed to construct a chemical kinetic simulation. The Chemistry Knowledge underlying the model is embedded in the functional group database. This is the most important input to any large kinetic model, but also the most challenging to obtain and to document.

of the important elementary-step reactions have ever been studied at all. So models constructed by just assembling a list of previously studied reactions from the literature are almost always missing important reaction steps, and therefore give seriously erroneous predictions.

2. *Estimations and Generalizations*

One might ask: how can one make sensible predictions when no data on one or more of the important reaction steps can be found in the literature? The answer is that chemists have abstracted a tremendous number of generalizations from the limited number of experimental (and, recently, theoretical) studies that have been performed on particular reactions. So when it is said that a particular reaction is "well-understood", it is not meant that someone has actually measured that reaction's rate to high precision under every possible reaction condition; instead, it means that enough measurements and/or calculations have been done on that reaction or perhaps on some similar reaction so that one can reasonably generalize and confidently estimate any particular parameter that is needed. Using generalizations has a tremendous advantage compared to attempting to list every possible individual reaction: one generalization can be used to estimate the parameters of hundreds or thousands of individual reactions. In fact, most of the rate parameters in the existing large kinetic models are actually estimates from generalizations, not from individual reaction experiments or quantum calculations.

Most automated model-construction software uses only the conventional List-of-Reactions data model. Since there is no established data model for generalized rate estimates, nor for documenting the rules that determine which species and reactions are considered important, this information is typically not treated as "data" at all—instead, much of it is hard-coded into the software.

In our view, the main problem with existing software is that it is poorly designed to deal with the true complexity of the real chemistry and the corresponding level of detail needed to represent it accurately, nor was it designed to be easily modified to incorporate new or improved chemistry knowledge. Because the chemistry has many details, and is imperfectly known, it is imperative that the software be easily extensible, so that additional chemical detail can be added as desired (preferably by the wide community of chemists). In this way, the model predictions can be continuously improved, each time incorporating the latest Chemistry Knowledge. If a modeling package is not extensible in this way, it will soon become obsolete.

D. A New Data Model for Chemical Reactions and Properties

Here we propose replacing the conventional List-of-Reactions data model with a data model that accurately represents the chemistry knowledge that goes into constructing a large simulation: a relatively small number of data on individual reactions which have been extensively studied in the literature plus a relatively small number of generalized rate and thermo estimation procedures. These inputs are used to estimate most of the thousands of parameters in the kinetic differential equations. Despite early work along these lines by Blurock (1995), this approach of treating the reaction-construction procedures themselves as input data rather than as part of the software has not yet caught on with the kinetics community. However, very recently several papers have presented kinetics software based on this data model (Blurock, 2004a, b; Ratkiewicz and Truong, 2006). It appears that the reaction mechanism generator (RMG) program described below will be the first software package enabling use of the new data model to become widely disseminated throughout the kinetics community.

We have developed a convenient data format for storing the rate and thermochemistry estimation parameters, as well as a graphical user interface (GUI), which makes it much easier to improve the rate estimates, to add additional reaction types, and to handle complicated functional groups. As we have shown elsewhere (Song, 2004), one can devise transparent unambiguous procedures, with clearly documented assumptions, for converting these estimation rules into lists of reactions, and thence into differential equations and quantitative predictions as is done currently. The new database format, GUI, and algorithm for constructing the List-of-Reactions from our new tree-structured rate estimation database allows us to achieve the design shown in Fig. 2.

1. Hierarchical Tree Structure for Functional Group Parameters

The information needed to compute both the molecular thermochemistry and also the elementary-step rate parameters can typically be associated with functional groups. To use these estimation approaches during mechanism-construction, we need a reliable, efficient, and unambiguous method for rapidly identifying which functional group values should be used for any given molecule. To maximize scientific progress, we must display the functional groups and the numerical values we associate with each group in such a way that can be easily used, understood, scrutinized, criticized, amended, and extended by other researchers.

We have accomplished these goals by storing all the functional group definitions in a hierarchical tree database developed by R. Sumathi (Song, 2004). As chemistry knowledge improves, the functional groups can be specified more precisely, allowing distinctions between related functional groups.

For example, at the roughest level of approximation, all carbonyl carbons can be treated as identical: they all have similar thermochemistry and all can undergo certain types of reactions (e.g. they all can be formed by beta-scission reactions of the corresponding alkoxy radicals). However, if one looks at the situation more carefully, the carbonyl groups in ketones, ketenes, aldehydes,

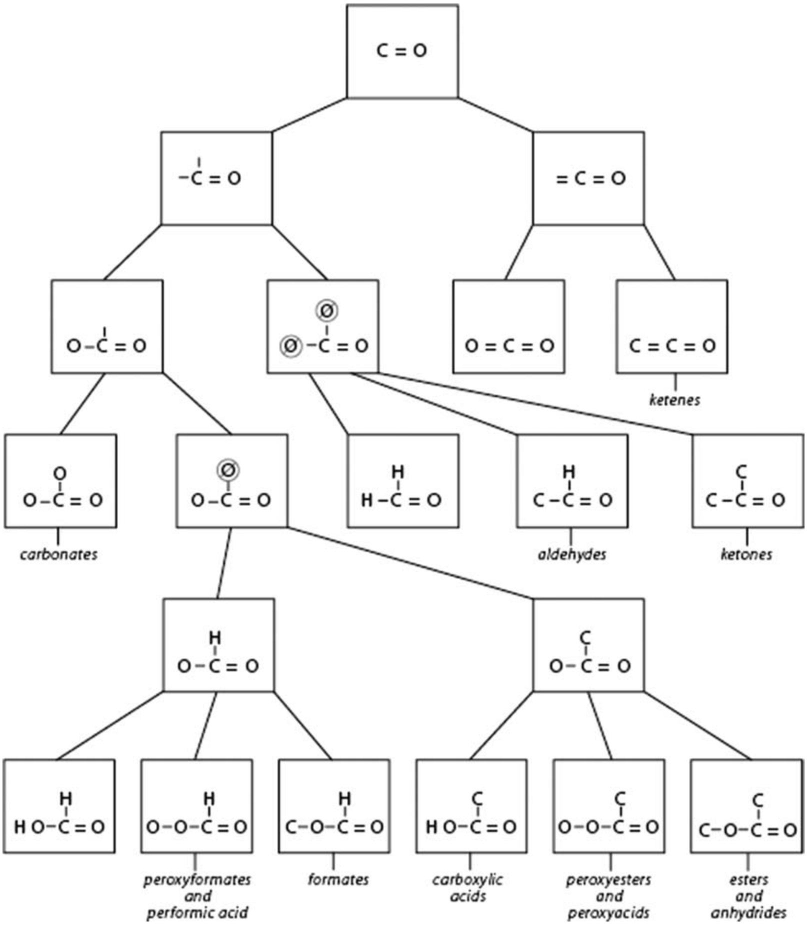
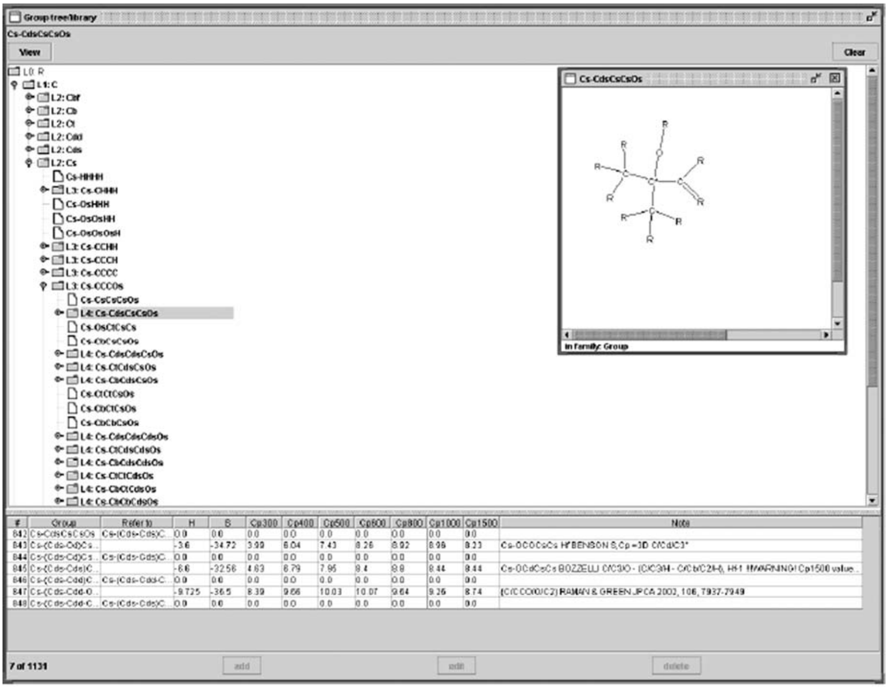


FIG. 5. Portion of a hierarchical functional group tree for classifying carbonyls. The slashed O means any atom except oxygen.

carboxylic acids, peroxyacids, and esters are all significantly different, Fig. 5. Still finer levels of distinction than are shown in Fig. 5 would certainly be justified chemically, e.g. unsaturated ketones certainly have different thermal properties than dialkyl ketones.

The functional group tree databases are read both by RMG and by a GUI developed by J. Robotham, Fig. 6. The RMG program performs substructure searches on each molecule in the model, to identify the most specific (i.e. most extended) functional group in the tree database that matches each portion of the molecule. The thermochemistry of each molecule and radical is estimated using Benson-type (Benson, 1976) group-additivity (one group value per non-terminal atom), but the functional group hierarchical tree allows one to define much



- L0: R
  - L1: C
    - L2: Cbf
    - L2: Cb
    - L2: Ct
    - L2: Cdd
    - L2: Cds
    - L2: Cs
      - Cs-HHHH
      - L3: Cs-CHHH
      - Cs-OsHHH
      - Cs-OsOsHH
      - Cs-OsOsOsH
      - L3: Cs-CCHH
      - L3: Cs-CCCH
      - L3: Cs-CCCC
      - L3: Cs-CCCOs
        - Cs-CsCsCsOs
        - L4: Cs-CdsCsCsOs
        - Cs-OsCtCsCs
        - Cs-CbCsCsOs
        - L4: Cs-CdsCdsCsOs
        - L4: Cs-CtCdsCsOs
        - L4: Cs-CbCdsCsOs
        - Cs-CtCtCsOs
        - Cs-CbCtCsOs
        - Cs-CbCbCsOs
        - L4: Cs-CdsCdsCdsOs
        - L4: Cs-CtCdsCdsOs
        - L4: Cs-CbCdsCdsOs
        - L4: Cs-CtCtCdsOs
        - L4: Cs-CbCtCdsOs
        - L4: Cs-CbCbCdsOs

#	Group	Refer to	H	S	Cp300	Cp400	Cp500	Cp600	Cp800	Cp1000	Cp1500	Note
842	Cs-CdsCsCsOs	Cs-(Cds-Cds)C...	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	
843	Cs-(Cds-Od)Cs...		-3.6	-34.72	3.99	6.04	7.43	8.26	8.92	8.96	8.23	Cs-OCOCsCs H'BENSON S,Cp =3D C/Cd/C3*
844	Cs-(Cds-Cd)Cs...	Cs-(Cds-Cds)C...	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	
845	Cs-(Cds-Cds)C...		-6.6	-32.56	4.63	6.79	7.95	8.4	8.8	8.44	8.44	Cs-OCdCsCs BOZZELLI C/C3/O - (C/C3/H - C/Cb/C2/H), Hf-1 !!WARNING! Cp1500 value...
846	Cs-(Cds-Cdd)C...	Cs-(Cds-Cdd-C...	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	
847	Cs-(Cds-Cdd-O...		-9.725	-36.5	8.39	9.66	10.03	10.07	9.64	9.26	8.74	{C/C CO/O/C2} RAMAN & GREEN JPCA 2002, 106, 7937-7949
848	Cs-(Cds-Cdd-C...	Cs-(Cds-Cds)C...	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	

FIG. 6. The graphical user interface (GUI) for displaying and modifying the functional group trees used to estimate thermochemistry.

more specific, complicated, and extended functional groups than are typically used, for example the extended groups for computing the thermochemistry of polycyclic aromatic hydrocarbons (Yu *et al.*, 2004). In each functional group the "central atom" of the group must be indicated; this is the non-terminal atom whose group thermochemistry is being estimated. In Fig. 5, the central atom is always the C double bonded to the oxygen.

At present, the functional group trees are certainly incomplete. For example, at present we only have reasonably comprehensive information on functional groups made of C, H, and O atoms. There is a limited amount of thermochemical (and detailed rate) information available in the literature on functional groups involving halogens, N, P, and S, but this is not yet reflected in the functional group trees. Very little generalized information is available for functional groups involving other elements. For certain classes of compounds, such as fused-ring compounds, there is some question about the correct method for generalizing from the limited available information to estimate the thermochemistry of never-studied species. It will be many decades before the functional group trees will be at all complete. Hence, it is essential to design a functional group tree structure which can easily incorporate new chemistry information.

Users can conveniently view and modify the thermochemical group values, or add new functional groups, using the GUI, Fig. 6. Most commonly, an expert will subdivide an existing functional group into several more detailed functional groups, each with distinct thermochemical parameters. The tree structure is intuitively obvious to chemists, since it groups similar chemical entities together. The tree structure of the database also facilitates efficient search for the best-matching functional group in each molecule encountered while the computer is building the reaction mechanism.

Following Bozzelli and Ritter (Lay *et al.*, 1995; Ritter and Bozzelli, 1991), the thermochemistry of free radicals is estimated by adding Hydrogen Bond Increments (HBI) to the energy of the corresponding stable molecule where an H has capped the radical site. The HBI groups are also stored in a functional group tree.

It should be noted that there are some types of molecules where it is not clear if the functional group tree approach is viable. For example, some transition metal complexes have fluxional structures, so there can be some debate about which atoms are bonded to each other. In biochemistry, non-bonded interactions often have very large cumulative effects on the thermochemistry. For multi-ring compounds and molecules with chiral centers, there are often several distinct three-dimensional structures corresponding to the same connectivity diagram, and these might each have discernibly different thermochemistry. However, the functional group approach has been found to be quite effective for a very large fraction of all molecules studied to date.

So far, we have compiled and tree-classified more than 700 functional groups with estimates of their thermochemical group values. However, only about 110 of these group values have been carefully derived; the rest are mostly estimates based on analogy. More technical details on the tree classification of functional groups and the thermochemistry estimation procedure used in the RMG software package are given at http://web.mit.edu/greengp/RMG and elsewhere (Song, 2004).

2. *Functional Group Trees for Reaction Rate Estimation*

a. Functional group tree rate estimation: How it works. As with thermochemistry, functional groups can be used to classify chemical reactions in a way that is easily extensible and easy for a human to understand. The functional group concept also corresponds well to physical reality: at a microscopic level, the barrier height for a chemical reaction is controlled primarily by the local structure close to the bonds that are being made and broken. For many reactions, the Arrhenius A-factor is also strongly affected by the nearby functional groups.

However, this is considerably more complicated than for thermochemistry, because most chemical reactions involve two functional groups, not just one, and some types of reactions have a rather complex dependence on the extended molecular structure. When representing functional groups for chemical reactions, one usually needs to specify several special atoms rather than just one central atom: one must specify all the atoms whose bonding is changing in the course of the reaction.

For each type of reaction, we have first specified the "Reaction Recipe", which specifies which bonds are made and broken and which gives the numbering for the special "central" atoms. Then we have constructed functional group trees for each of the functional groups involved, typically two functional groups for a bimolecular reaction, one on each reactant.

In order to find the correct rate estimation parameters for a bimolecular reaction, one must first find the best-match functional group A in one reactant, using a substructure search algorithm very similar to that used to find thermochemical group values. Then one must find the best-match functional group B in the second reactant in a similar way. Then one can look up the rate estimation parameters for this reaction type using A and B as the keys.

As a simple example, consider the addition of ethyl radical to the O end of the carbonyl group in allyl acetate ($CH_3C(O)OCH_2CHCH_2$). The Reaction Recipe for radical-addition-to-double-bond is shown in Fig. 7.

This reaction will occur whenever one reactant has a radical center, and the other reactant has a double bond. In the example, the first reactant, ethyl radical, has a radical site, designated as atom 1 in the Reaction Recipe. The second reactant, allyl acetate, has two double bonds. The program would first select one of the double bonds, and then select one end of the double bond to be atom 2 in the Reaction Recipe, Fig. 7. (After completing this reaction, the computer will successively try all the other permutations.) Suppose the program selected the $C{=}O$ double bond, and selected the O atom to be atom number 2. The carbonyl carbon would be designated atom number 3. Immediately, the computer would perform the reaction operation as shown in Fig. 7, generating a new product molecule, Reaction (1).

$$CH_3CH_2 + CH_3C(O)OCH_2CHCH_2 \rightarrow CH_3C(OEt)OCH_2CHCH_2 \qquad (1)$$

If this product molecule did not yet exist in the reaction mechanism, the computer would immediately classify all of its functional groups using the thermochemical group value tree described above, to compute its thermochemistry. (The product thermochemistry will certainly be needed to compute the heat of reaction and the reverse reaction rate, and it might be needed to compute the forward reaction rate also.)

Now the computer must identify the best rate estimation parameters for the forward reaction. First the radical site in the first reactant, ethyl radical, will be classified as "primary alkyl" using the functional group tree shown in Fig. 8.

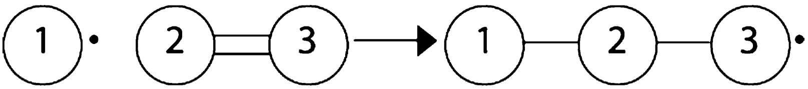


FIG. 7. Reaction Recipe for radical addition to a double bond.

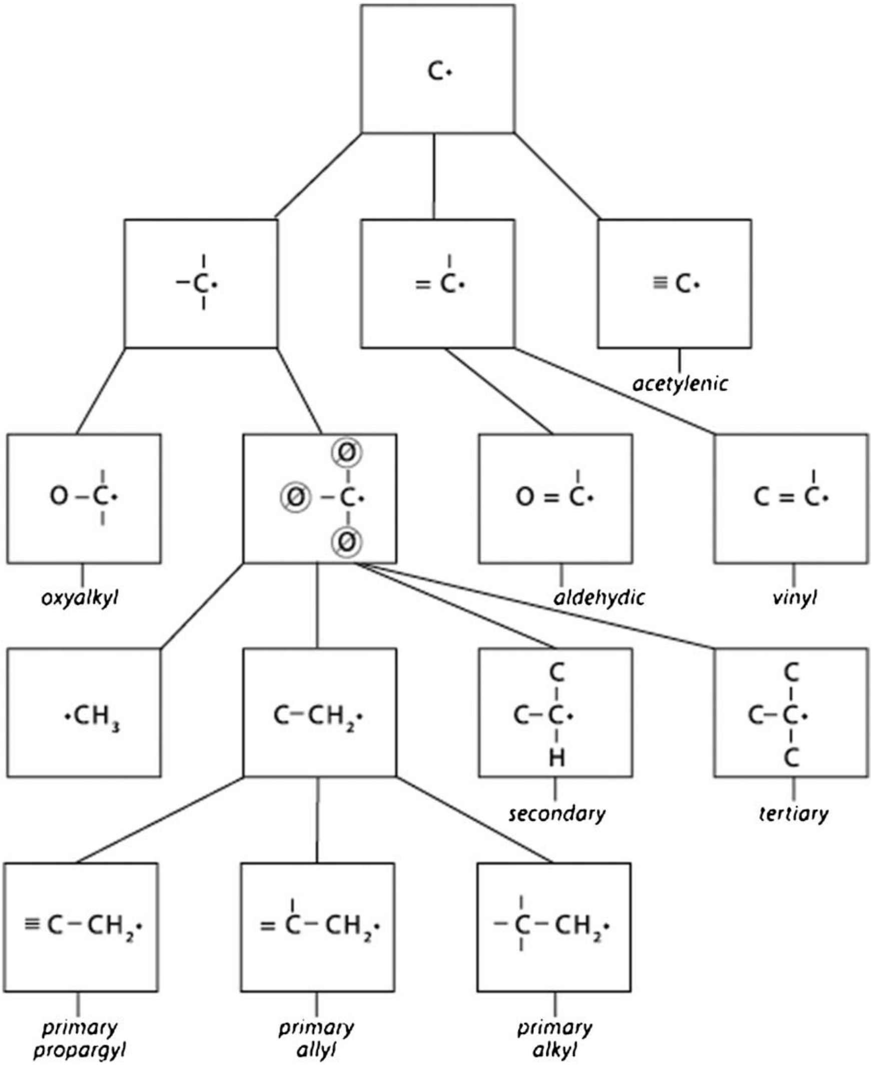


FIG. 8. Portion of one of the functional group trees for addition reactions: classification of site 1 (the radical center) in the Reaction Recipe shown in Fig. 7.

Then, from the tree in Fig. 9 the O-ended double bond in allyl acetate would be classified as an "ester O". Note the numbers 2 and 3 in this functional group tree. These numbers indicate which O and C atoms in the functional group participate directly in the reaction. (Of course the neighboring O and C atoms affect the rate to some extent, even though they do not change their valence bonding in this reaction.)

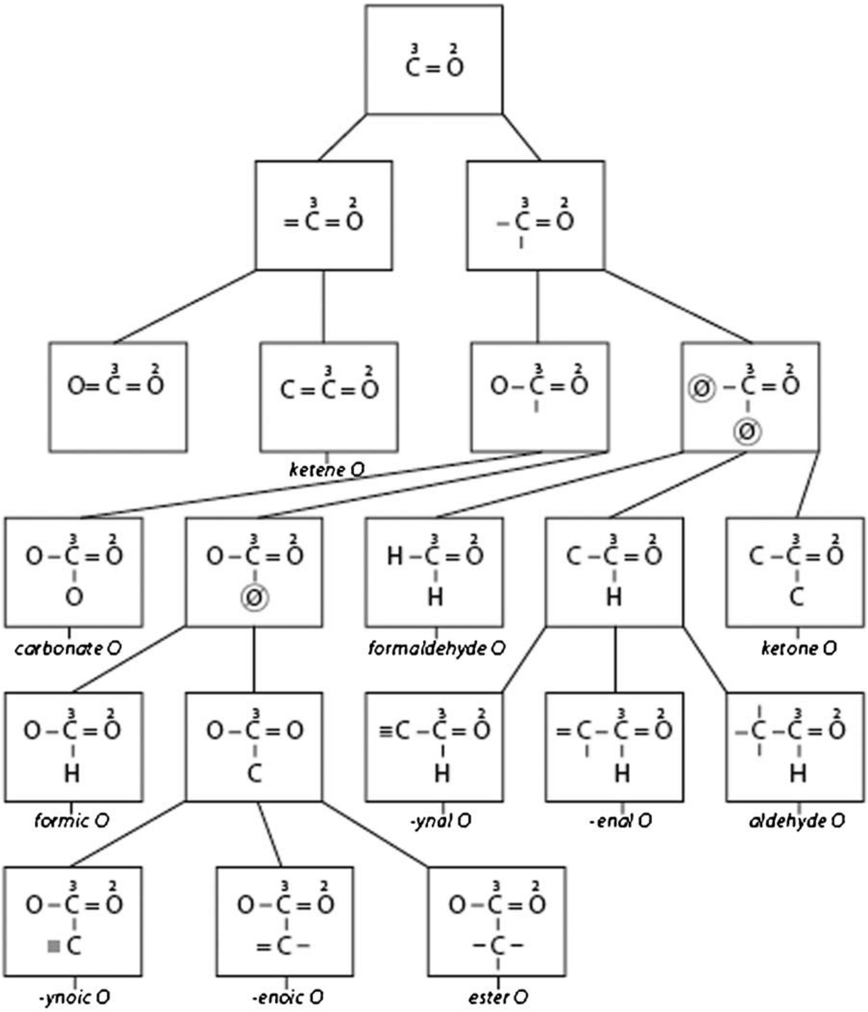


FIG. 9. Portion of the one of the functional group trees for addition reactions: classification of the 2–3 double bond in the Reaction Recipe shown in Fig. 7.

The computer would then look in the rate estimation library for Radical Addition reactions to find the rate parameters that correspond to "primary alkyl + ester O". There it would find a set of numerical parameters E_o, α, A, n with estimated uncertainties; one could then calculate the forward rate coefficient for the reaction of interest using this Evans–Polanyi modified Arrhenius form:

$$k_{\text{forward}} = \mathrm{A}T^{n}\exp\left(\frac{-(\mathrm{E_o} + \alpha\Delta H_{\text{reaction}})}{RT}\right) \tag{1}$$

As discussed below, the k_{forward} computed this way is only accurate in the high-pressure-limit. The reverse rate would be computed from k_{forward} using the thermochemistry, for example if all the species are ideal gases:

$$k_{\text{reverse}} = k_{\text{forward}} \left(\frac{RT}{1\text{atm}}\right)^{\Delta\text{moles}} \exp\left(\frac{\Delta G_{\text{reaction}}}{RT}\right) \tag{2}$$

where (Δmoles) = change in number of moles in the reaction. Note that both the forward and reverse rates depend exponentially on the enthalpy of the reaction, so it is critical that the thermochemistry be computed accurately.

The functional group tree classification of reaction rate parameters is conceptually related to an extension of the thermochemical group-additivity to transition states. In conventional transition state theory, for ideal gases and neglecting tunneling and recrossings, the reaction rate coefficient is given by:

$$k_{\text{forward}} = \left(\frac{RT}{1\text{atm}}\right)^{\text{Nreactants}-1} \left(\frac{k_B T}{h}\right) \exp\left(\frac{G_{\text{reactants}} - G_{\text{TS}}}{RT}\right) \tag{3}$$

In automated model-construction programs, $G_{\text{reactants}}$ is normally estimated using group-additivity (as discussed above), i.e. $G_{\text{reactants}}(T) = \Sigma G_{\text{group},n}$ where the program looks up the value of $G_{\text{group}}(T)$ for each of the thermochemical groups in the molecule. It is possible to estimate G_{TS} in a similar way (Sumathi & Green, 2002), if one first divides the transition state structure (i.e. the geometry at the saddle point) into a "reacting" functional group which includes all the bonds that are being made and broken, and "inert" functional groups:

$$G_{\text{TS}}(T) = G_{\text{reactive}}(T) + \Sigma G_{\text{group},n} \tag{4}$$

For the inert functional groups one can look up the same group values $G_{\text{group}}(T)$ as for any ordinary molecule; in Eq. (3) these will identically cancel the corresponding terms in $G_{\text{reactants}}(T)$. The values for the reacting functional groups are of course very different, but these group values can be computed using quantum chemistry and then stored in the functional group trees for reactions instead of the more conventional A, E_a, etc. In principle, it would be more accurate to store $G_{\text{reactive}}(T)$ rather than forcing all rates to conform to the modified Arrhenius form. However, with our current very imperfect knowledge of transition states and uncertainties in measured rates there aren't yet many cases where we can discriminate between the rate models in Eq. (1) and in Eqs. (3) and (4).

b. Difficulties with functional group rate estimates

1. Paucity of reliable data. A serious practical issue is that we do not yet have enough information to make accurate rate estimates for every possible combination of functional groups A + B. As noted above, we have identified about 700 individual functional groups for estimating thermochemistry of molecules $C_xH_yO_z$. We have identified about 30 distinct reaction types (with different reaction recipes) that are important in gas-phase thermal chemistry of C/H/O compounds; most of these involve more than one functional group. We would need $O(10^5)$ distinct sets of rate estimation parameters if we were to consider rate estimation at the same level of chemistry detail as we did thermochemistry (and surely if the thermochemistry varies with a change in the functional group, the reaction rate will be affected as well). But we have very few reliable data suitable for making these rate estimates. At present we are only able to use truncated functional group trees (or equivalently, use the same rate parameters for several similar functional groups) since we do not have estimates of the effects of subtler variations in the functional groups. Fortunately, the extensible tree structure makes it feasible to enlist the larger chemistry community in the task of converting "Chemistry Knowledge" into a more comprehensive listing of specific functional group-based rate estimation parameters that can be used in quantitative kinetic models.

One of the reasons for the paucity of data is that for many reaction families it is hard-to-impossible to set up an experiment that probes only the process we want to measure. For example, in the gas phase, it might be difficult to observe the reaction used as an example in Section II.D.2.a above, Reaction (1), for a couple of reasons. First, there are several competing reaction pathways (e.g. H-abstraction, addition to the $C=C$ double bond instead of the $C=O$ double bond) leading to different products. Second, as discussed further below, the product shown above may be so unstable that it cannot be detected, even if the reaction is actually running at a fast rate. Often researchers will report the rate for the sum of all the reactions, Reaction (2),

$$CH_3CH_2 + CH_3C(O)OCH_2CHCH_2 \rightarrow \text{Allproducts} \tag{2}$$

but it is difficult to use this information to derive rate estimation parameters for Reaction (1). To some extent this difficulty can be alleviated by quantum chemistry calculations, but if no experimental data are available to check quantum chemistry calculations on a new type of reaction, there is always some doubt about their reliability.

2. Chemical activation and fall-off. Gas-phase reactions that form an energized product cause particular difficulty in kinetic model-construction. For example, in Reaction (1) the unstable product will be formed with excess energy (due to the exothermicity of the bond-forming addition reaction), that it will rapidly dissociate to ethyl acetate + allyl radical, Reaction (3).

$$CH_3C(OEt)OCH_2CHCH_2 \rightarrow CH_3C(O)OCH_2CH_3 + CH_2CHCH_2 \tag{3}$$

At the low pressures used in many laboratory kinetics experiments, the rate of formation of allyl radicals by this two-step pathway will be much higher than one would compute if one assumed that the energized intermediate $CH_3C(OEt)OCH_2CHCH_2$ was in thermal equilibrium with the bath gas; the rate for the apparently direct fast Reaction (4)

$$CH_3CH_2 + CH_3C(O)OCH_2CHCH_2 \rightarrow CH_3C(O)OCH_2CH_3 + CH_2CHCH_2 \tag{4}$$

will strongly depend on the bath gas temperature, pressure, and composition. This effect, where a reaction appears to go directly through several elementary steps in a single step, is called "chemical activation". (If the energized product does not fragment to form new species, but has a significant probability of falling apart back to the reactants before being thermalized, the effect is called "fall-off".)

Chemical activation causes many problems. First of all, rates are now a function of P (as well as T), adding another dimension to the rate estimation functions. Also, chemically-activated rates depend on properties of the complete adduct molecule, not just some of its functional groups. And, in order to compute the chemically-activated rates, one must enumerate all the important product channels, i.e. identify a large number (Matheu *et al.*, 2003a, b) of different reactions and products, which could be of completely different types. As a result, each chemically-activated reaction has unique rate parameters, not amenable to easy generalization. It would not be practical to compute or to store the rate parameters for every conceivable chemically-activated reaction.

Our solution to these challenges is to have the computer compute the rate parameters for the chemically-activated reactions on-the-fly as needed during the model-construction process, using the algorithm of Matheu *et al.* (2003a). The density of states is estimated from the heat capacity using the three-frequency method devised by Bozzelli (Bozzelli *et al.*, 1997) complemented by the Stein–Rabinovitch method for convolving rotors with vibrations (Stein and Rabinovitch, 1974, 1973). The micro-canonical rates $k(E)$ are estimated using the inverse Laplace transform method (Forst, 1973) or (often equivalently) Quantum Rice-Rampsberger-Kassel (QRRK) (Dean, 1985). The master equations are solved approximately using the modified strong collision approximation as implemented in CHEMDIS (Chang *et al.*, 2000). For more details, see discussions by Matheu (2003).

In the database we only need to store high-pressure-limit (i.e. thermally-equilibrated intermediates) rate estimation parameters for elementary-step reactions; from this information, the molecular structure, and the thermochemical parameters (Section II.D.1) one can compute the pressure-dependent reaction rates. The elementary-step high-pressure-limit rate estimation parameters depend primarily on the local functional group structure, and so are very well-suited to functional group tree classification.

3. Intramolecular reactions. Intramolecular reactions, such as intramolecular addition reactions and intramolecular H-abstraction reactions, introduce serious complications. In these reactions, the reacting functional groups A and B are both contained in the same molecule, tied together by a third functional

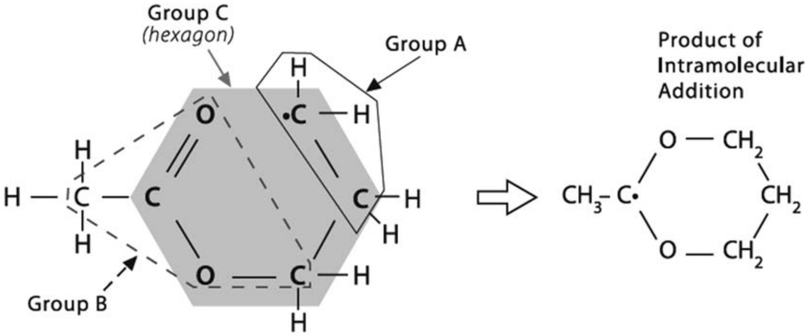


FIG. 10. A molecule undergoing an intramolecular addition (cyclization) reaction, indicating the definitions of functional groups A, B, and C.

group C, Fig. 10. In order to estimate the reaction rate, one must specify all three functional groups, each of which can be comparable in complexity to the functional groups used to describe thermochemical groups and bimolecular reactions. So one can imagine a very large number of distinct types of intramolecular reactions with unique rate constants specified by the functional groups involved (A, B, and C).

For the reaction illustrated in Fig. 10, groups A and B could be classified using the trees shown in Figs. 8 and 9, but the tether group C requires a new functional group tree, Fig. 11. In Fig. 11, the tether group is classified solely based on the bonding arrangement inside the ring; in the future this tree should be extended further to differentiate according to which atoms are at each node in the tether, and to account for substituent effects.

Unfortunately, only a very small fraction of the many conceivable distinct intramolecular reactions have been studied in the literature, due in part to technical problems. Experimentally, it is often difficult to prepare a run on high concentrations of thermalized molecules with a high-reactivity functional group A close to another functional group B that reacts with A. It is therefore very difficult to measure the (often rather high) rate at which this reaction occurs. To make the reaction slow enough to measure, one can run it in the endothermic direction—but then one needs a very efficient way to trap the product before it converts exothermically back to the reactants.

Also, in gas-phase experiments, many of these reactions are pressure-dependent, and special care must be taken in extrapolating the experimental measurements to the high-pressure-limit. Quantum chemical calculations are also rather difficult for most reactions of this type, since molecules that contain three distinct functional groups A, B, and C necessarily include a rather large number of heavy atoms. For this reason, so far relatively few high-level calculations have been performed on these sorts of reactions.

At present, these types of reactions are represented using three different extensible functional group trees, one for each of the groups involved. In principle,

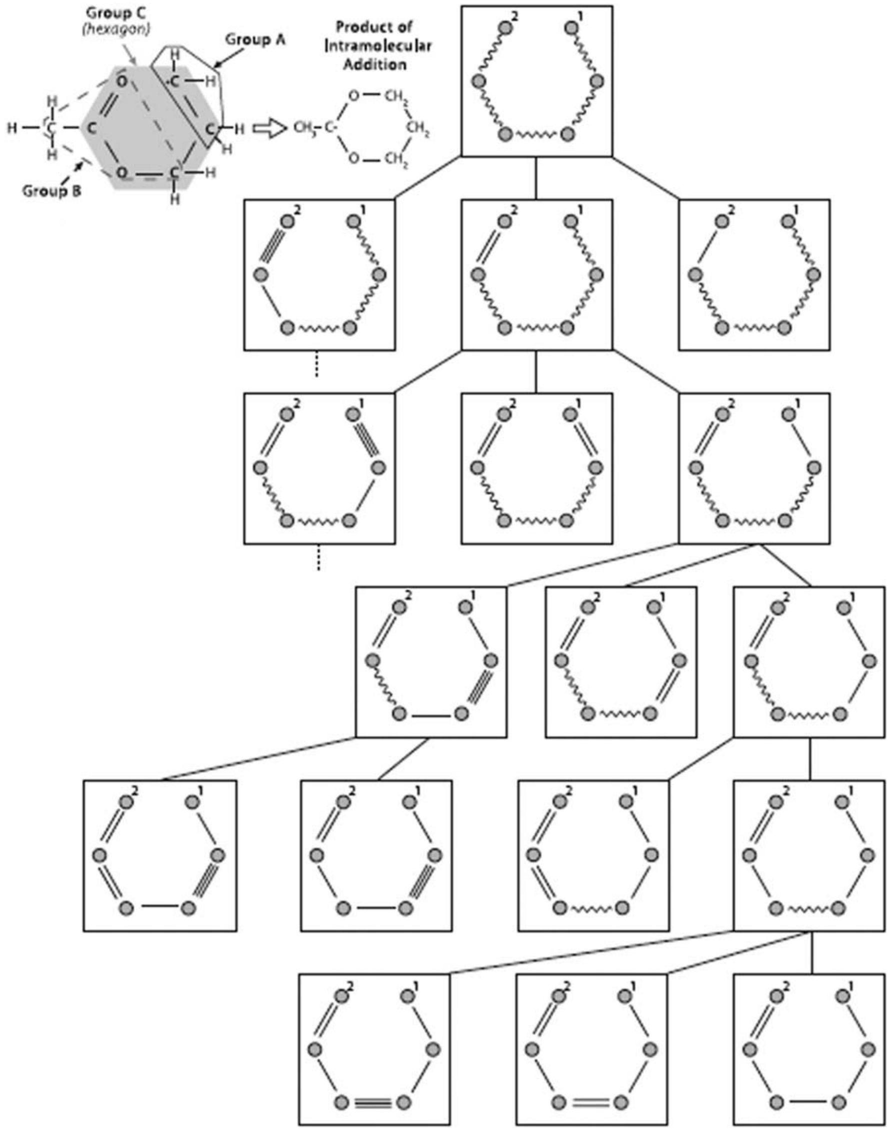


FIG. 11. A portion of a functional group tree for classifying tether functional groups, e.g. group C in Fig. 10.

this requires a much larger table of rate estimation parameters (keyed from A, B, and C). However, due to lack of information, we and many others working on these reactions often make untested separability assumptions, e.g. assuming that the high-pressure-limit rate parameters for an intramolecular reaction involving groups (A, B, and C) can be computed using Eq. (5):

$$k = k_{\text{bimolecular}}(\text{A} + \text{B}) \times D_{\text{C}} \exp\left(\frac{-E_{\text{C}}}{RT}\right) \tag{5}$$

where $k_{\text{bimolecular}}(\text{A}+\text{B})$ is the rate we estimate for the bimolecular reaction $\text{A}+\text{B}$ (i.e. if the tether C was cut somewhere in the middle), E_{C} is the estimated ring-strain involved in bringing reactive sites close to each other when they are connected by a tether C, and D_{C} is an estimate of the ratio between a bimolecular Arrhenius A factor and the corresponding intramolecular Arrhenius A factor; this depends significantly on the structure of the tether C (due to variations in the number loss of rotors in the transition state). By making the separability assumption in Eq. (5), one only needs to add a simple table that returns D_{C} and E_{C} given as a key a single tether group C. To move beyond this separability approximation and account for interactions between functional groups A, B, and C, one would need to extend to a much larger three-key table.

E. The Reaction Mechanism Generator (RMG)

The new data model for chemical kinetics will work best if there is a reproducible, deterministic, automated procedure for going from the fundamental inputs (a library of the small number of known rate constants, and the hierarchical tree of rate estimation parameters for all the other reactions) to the desired outputs (e.g. predicted yield/selectivity profiles). Ideally, the software that performs this automated procedure will not need to be altered if someone changes an initial condition, a rate estimate, or even the structure of the hierarchical tree—reducing the need to touch the source code of the software dramatically reduces both the programming burden on the kineticist and the likelihood of introducing a bug into the software. This design also has the practical advantage that exactly the same software could be used by many kineticists for many different problems, making it much easier to reproduce work by another, while also reducing the difficulty of maintaining the software.

We have developed a software package, RMG, for exactly this purpose. The software is designed to take the hierarchical database trees as its main input, and to return a kinetic model as its main output. This kinetic model can be in the conventional List-of-Reactions format suitable for use in CHEMKIN or other integrators, or, if desired, the RMG program package can perform the integration itself and return product yield/selectivity profiles.

RMG, which was implemented by Jing Song (2004), uses a version of the rate-based model-construction algorithm described by Susnow *et al.* (1997), but it treats non-thermalized chemically-activated reaction paths on an equal basis with ordinary thermal reactions, as done by Matheu *et al.* (Matheu, 2003; Matheu *et al.*, 2003a, b). Byproducts and activated reaction intermediates are considered important species if their estimated instantaneous formation rates ever exceed $\varepsilon\|\omega(t)\|$, where $\|\omega(t)\|$ is the norm of the major species flux array (Song *et al.*, 2002) and ε is the error tolerance. The concentrations needed to compute the rates of change $\omega(t)$ are estimated using the best mechanism available so far. The concentration estimates can be substantially in error early in the process when the mechanism may be missing important reactions. However, this procedure has been demonstrated to settle down and become quite stable and robust after a modest number of iterations even for rather complex systems (Matheu *et al.*, 2003a). At convergence, all of the neglected byproducts are formed at rates at least a factor of ε lower than the rates of the important reactions in the system.

Unlike most model-generation software described in the literature, RMG correctly handles pressure and temperature variations; it does this by using $k(T,P)$ computed for the chemically-activated reactions at discrete (T,P) to determine coefficients in a Chebyshev form (Venkatesh *et al.*, 1997) suitable for use in the differential equation solver.

At the heart of the RMG algorithm is the process of identifying all the possible reactions of some species of interest with itself and all the species included in the model so far, and estimating each reaction's rate. Unlike most previous mechanism generators, the RMG software handles this procedure in a very general way: all the specifics about functional groups and reaction types, as well as all the associated parameters, are external to the program, contained in the tree databases. This has the huge advantage that nothing inside the RMG program has to change when the chemistry information is updated: chemists only need to modify the tree databases using the GUI, they do not need to do any programming or re-compiling.

F. APPLICATIONS OF RMG

So far, RMG has been applied to several problems: the pyrolysis (steam-cracking) of *n*-hexane (Van Geem *et al.*, 2006), the laser-initiated oxidation of neo-pentane (Petway, 2005), the supercritical water oxidation of methane [http://web.mit.edu/cfgold/www/RMG/testcases.html], and the pyrolysis and oxidation of butane (Song *et al.*). In each case, the software constructed models involving $O(100)$ species, and the model predictions were about as accurate as one would expect given the factor of 2–3 uncertainties in the rate parameters and equilibrium constants arising from the estimation techniques used to generate the model parameters.

As a simple example, consider the predicted yield vs. conversion data for butane oxidation at 715 K, Figs. 12 and 13, plotted on top of experimental data points measured by Cernansky's group at Drexel (Wilk *et al.*, 1995). As is typical, most of the discrepancies between the model predictions and the experimental data can be safely attributed to the uncertainties in the rate

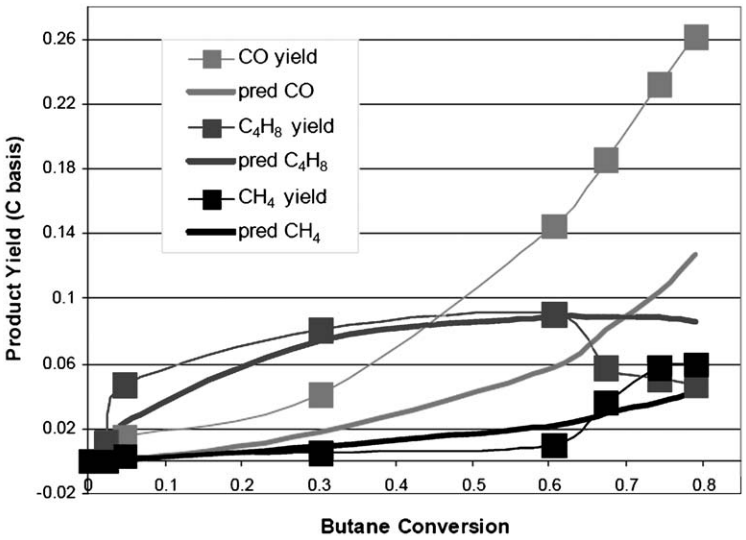


FIG. 12. Model predictions vs. experimental measurements of Wilk *et al.* (1995) from butane oxidation at 715 K. The predicted yields are generally within a factor of two of the experimental data, reflecting roughly factor of two uncertainties in rate constant estimates. As is typical, the discrepancies are largest at high conversions, both because the errors in the parameters cumulate, and because the model may be missing some reactions of the minor byproducts (Please see Color Plate Section in the back of this book).

constant estimates (and related uncertainties in the thermochemistry). However, the discrepancy between the predicted and observed CO_2 yields at very small conversion is likely due to the fact that the model does not accurately reproduce the true initial conditions (the model assumes instantaneous perfect mixing and a perfectly isothermal reactor).

Many detailed chemical kinetic models show discrepancies at high conversions similar to what is seen in Fig. 12, in large part due to the time-integrated effects of even relatively small errors in the estimated rate parameters, and in some cases also due to the omission of side reactions, which become important as products build up.

Several other mechanism-generation computer programs have been developed which have shown comparable ability to generate predictions in good accord with experimental data; for a couple of impressive recent examples see Matheu's models of methane (Matheu *et al.*, 2003a) and ethane/ethene pyrolysis (Matheu and Grenda, 2005a, b) and the Nancy group's models for oxidation of fuels in the gasoline range (Buda *et al.*, 2006; Glaude *et al.*, 2002). The case of methane pyrolysis is particularly interesting as an example where the mechanism-generation program immediately identified critical reaction steps that the human experts had overlooked for more than 20 years (Matheu *et al.*, 2003a).

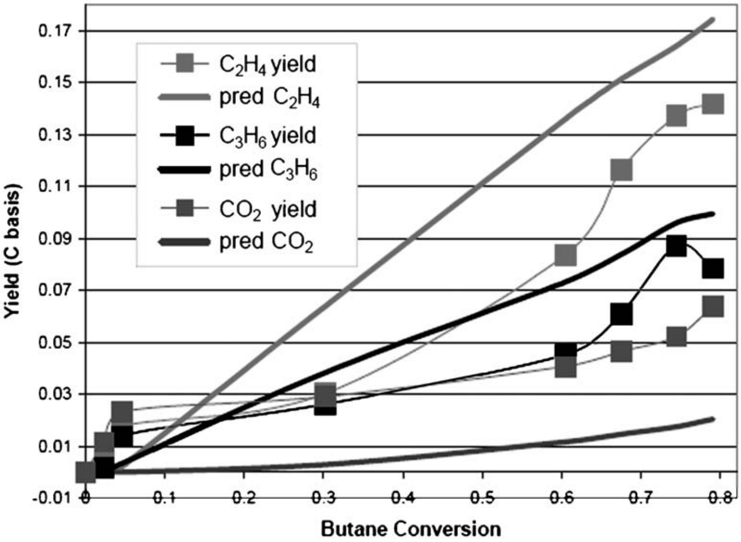


FIG. 13. Model predictions vs. experimental measurements of Wilk *et al.* (1995) for ethene, propene, and CO_2, same conditions as Fig. 12. The large discrepancy in the CO_2 predictions at the lowest conversions suggests that the model does not accurately represent the true boundary conditions at the inlet (Please see Color Plate Section in the back of this book).

The new RMG software combines high-accuracy chemistry estimation methods (including methods for automatically identifying chemically-activated reaction paths) with the extensible 21st century functional group tree data model described above, so it is now much easier for users to add, modify, and document the chemical assumptions that underlie the models than it was using the previous generation of model-construction software, and the software does not need to be modified or recompiled when new chemistry is added.

The RMG software, the rate and thermo estimation databases, and the GUI for reading and modifying the functional group tree databases are all available as open-source software at http://web.mit.edu/greengp/. It is hoped that experts in all branches of chemical kinetics will contribute to improving and extending the rate and molecular property estimates used by RMG.

III. Efficiently and Accurately Solving Large Kinetic Simulations

The methods described in Section II carry out the first step in Fig. 1, constructing a detailed chemical kinetic model from our current understanding of chemistry. However, this step is only useful if we can numerically solve the chemical kinetic simulation to obtain quantitative predictions. In many cases, solving the model is even more challenging than constructing it.

Most real reacting systems are accurately described by a large system of partial differential equations in (x, y, z, t); most commonly the spatial coordinates are discretized so that one is actually solving equations like this at each mesh point:

$$\frac{\mathrm{d}Y^p}{\mathrm{d}t} = \Theta^p[Y] + \omega^p(Y^p) \tag{6}$$

where Y is the long array of state variables, Θ are algebraic equations representing the transport between mesh point "p" and nearby mesh points, and ω^p the algebraic equations representing the rate of change of the state variables occurring at mesh point p due to chemical reactions, i.e. the chemical source terms. The differential equations Eq. (6) are usually supplemented by some algebraic equations (e.g. boundary conditions, equations of state). While a few of the state variables Y represent fluid dynamic variables such as momenta, for even relatively simple reacting chemical systems the great majority of the state variables are chemical species concentrations or mass fractions, and in most systems of interest to reaction engineers the (usually very stiff) chemistry terms consume most of the CPU time (Schwer *et al.*, 2002).

Detailed chemical kinetic models are usually very stiff in time due to the large separation of timescales between reactive intermediates and the more stable major species. This is a serious problem for explicit methods and stochastic methods, but stiff ordinary differential equation (ODE) systems describing up to O(10^4) variables can be handled routinely using modern implicit solvers like DASPK (Bernan *et al.*, 1989) and VODE (Brown *et al.*, 1989). However, spatially inhomogeneous reacting systems often lead to stiff systems with many more state variables, which cannot be efficiently solved with available off-the-shelf black-box software. Two-dimensional and three-dimensional simulations with large (> 100 species) chemical kinetic models are almost impossible to solve with existing software. This is unfortunate, since mechanism-generation programs such as RMG described in Section II above regularly produce models containing more than 100 chemical species.

There are many ways one can try to reduce the computational burden. Ideally, one would find numerical methods which are guaranteed to retain accuracy while speeding the calculations, and it would be best if the procedure were completely automatic: i.e. it did not rely on the user to provide any special information to the numerical routine. Unfortunately, often one is driven to make physical approximations in order to make it feasible to reach a solution. Common approximations of this type are the quasi-steady-state approximation (QSSA), the use of reduced chemical kinetic models, and interpolation between tabulated solutions of the differential equations (Chen, 1988; Peters and Rogg, 1993; Pope, 1997; Tonse *et al.*, 1999). All of these methods were used effectively in the 20th century for particular cases, but all of these approximated-chemistry methods share a serious problem: it is hard to know how much error is introduced by making these approximations (unless one is able to solve the simulation using the full chemistry model—in which case there may have been no need to make the approximation). Because there is no bound on the error introduced by reducing the chemistry this way, usually no one really knows how much accuracy one is sacrificing in order to make the calculation feasible.

Below we present 21st century numerical solution methods suitable for large chemical kinetic simulations. We present both a method that does not introduce any approximations, and an automated model-reduction method that allows rigorous error control in steady-state simulations.

A. Fast Solution of Large Systems of Chemistry Equations Using Sparsity

Before making approximations to speed up a calculation or reduce its memory requirements, it is worthwhile to consider if there are CPU time or RAM reducing alternatives that do not introduce any error. For large chemical kinetic simulations, one is typically solving rather large systems of differential-algebraic equations (DAEs), where the differential equation system is extremely stiff. The most popular solvers for these types of problems are DASSL (Petzold, 1982), DASAC (Caracotsios and Stewart, 1995), and DASSL's successor DASPK (Bernan *et al.*, 1989). By time-profiling these types of solvers, one discovers that most of the CPU time is consumed constructing the Jacobians needed to solve stiff systems of chemistry equations.

The structure of these Jacobians depends on which variables one uses when setting up the equations; for most obvious choices the chemistry Jacobians are dense (i.e. more non-zero elements than zeroes in the Jacobian matrix). One can always rewrite the equations, however, to give unstructured sparse Jacobians (i.e. most of the elements in the matrix are identically zero) (Schwer *et al.*, 2002). With standard Jacobian-construction and solvers, it makes little difference whether the Jacobian is dense or sparse. But if one uses software that takes advantage of the sparsity, one can reduce both the memory requirements and the CPU time required by a significant factor just by omitting all the entries known to be zero. Determining which Jacobian entries are always zero requires a lot of book-keeping in a large chemical kinetic system, but this book-keeping can be done automatically using programs like DAEPACK, which analyze FORTRAN programs symbolically (Schwer *et al.*, 2002; Tolsma and Barton, 2000). Taking advantage of sparsity in homogeneous charge compression ignition (HCCI) engine simulations with very large chemistry models can speed up the calculations by a factor of 60, and reduce the memory requirements enough that rather large calculations can be run on ordinary laptop computers (Yelvington *et al.*, 2004).

It is not hard to change the variables to make the Jacobian sparse, and it is simple to run a simulation using DAEPACK rather than DASPK or VODE, to see whether the sparsity-aware software is less demanding of computer resources than the conventional solvers. In our experience, if the chemistry model involves more than ~100 species, it is usually better to use the sparse solver approach. For more details on how to use DAEPACK to analyze a FORTRAN simulation and to handle the sparsity, see Schwer (Schwer *et al.*, 2002) and http://yoric.mit.edu/

B. Using Reduced Chemistry Models in Multidimensional Simulations without Introducing Error

The CPU time required to solve large chemical kinetic simulations scales approximately linearly with the number of chemical reactions and with the number of finite volumes in the mesh, and approximately quadratically with the number of chemical species. (The LU decomposition of the Jacobian performed in most stiff solvers, which scales as N^3, has a much smaller prefactor than the N^2 terms and usually takes negligible time compared to other steps in chemical kinetic simulations (Schwer *et al.*, 2002).) So for large chemical kinetic models (10^2~10^3 species, 10^3~10^4 reactions) there is a significant potential to speed up the calculation by replacing the large chemistry model with a smaller model (fewer reactions, maybe also fewer species). This is particularly important in multidimensional reacting-flow simulations, where it is not unusual for calculations to take weeks to converge even on very powerful computers. As mentioned above, there are many ways to approximate the chemistry: the challenge is how to avoid losing accuracy when making these approximations.

It is relatively straightforward to bound the error in the transport and chemistry terms on the right-hand side of Eq. (6) caused by certain types of approximations to the Θ and ω operators, since Θ and ω are relatively simple algebraic operators. Methods for bounding/controlling the errors associated with the size of the mesh in the transport operator Θ are very well-understood, and form the basis of adaptive meshing technology (Berger and Oliger, 1984; MacNeice *et al.*, 2000). Methods for bounding the error in ω made by using a reduced chemistry model are discussed below in Section III.C. For the moment, let us suppose we have constructed an approximation $\omega_{\mathrm{approx}}(Y)$ to the true $\omega(Y)$, and that we can set bounds on the deviation between these two chemistry source terms.

We would like to solve the approximate model, Eq. (7) rather than the full model, Eq. (6) above, but we need to have some control over the error introduced by this approximation.

$$\frac{\mathrm{d}Y^p_{\mathrm{approx}}}{\mathrm{d}t} = \Theta^p(Y_{\mathrm{approx}}) + \omega^p_{\mathrm{approx}}(Y^p_{\mathrm{approx}}) \tag{7}$$

The difficulty is that $Y(t)$ and $Y_{\mathrm{approx}}(t)$ can follow rather different trajectories; while we can bound the error $|\omega_{\mathrm{approx}}(Y) - \omega(Y)|$ at a single point Y, it is not so easy to bound the error $|\omega_{\mathrm{approx}}(Y_{\mathrm{approx}}(t)) - \omega(Y(t))|$ along the whole trajectory. (Despite the difficulty, very recently Singer and Barton have developed a rigorous method to bound $|Y(t) - Y_{\mathrm{approx}}(t)|$. So far this approach has only been applied to problems with only relatively small number of state variables; making it viable for large-scale chemical kinetic simulation is a challenge for the future.)

For cases where explicit integration methods are viable, one could adjust the time step to ensure that $Y_{\mathrm{approx}}(t)$ very closely tracks $Y(t)$ (roughly speaking, to ensure that $|\omega - \omega_{\mathrm{approx}}|^{*}\Delta t$ is small compared to other errors in the calculation). But for most large-scale chemical kinetic problems explicit integration methods are not viable due to the stiffness, so this idea is not very appealing.

There is, however, one important class of reacting-flow simulations where bounding the deviation between $\omega(Y)$ and $\omega_{\mathrm{approx}}(Y)$ immediately allows one to control the approximation error: steady-state simulations. For steady-states, the left-hand side of Eq. (6) and Eq. (7) vanish, and one wants to ensure that the solution Y_{approx} to the approximate equation

$$0 = \Theta(Y_{\mathrm{approx}}) + \omega_{\mathrm{approx}}(Y_{\mathrm{approx}}) \tag{8}$$

is very similar to the (unknown) solution Y to the original equation:

$$0 = \Theta(Y) + \omega(Y) \tag{9}$$

If one has already found a converged solution to Eq. (8) so that

$$-\mathrm{tol}_{\mathrm{approx}} < \Theta(Y_{\mathrm{approx}}) + \omega_{\mathrm{approx}}(Y_{\mathrm{approx}}) < \mathrm{tol}_{\mathrm{approx}} \tag{10}$$

one can check whether Y_{approx} is an acceptable solution to the original equation by a single function evaluation

$$-\mathrm{tol}_{\mathrm{total}} < \Theta(Y_{\mathrm{approx}}) + \omega(Y_{\mathrm{approx}}) < \mathrm{tol}_{\mathrm{total}} \tag{11}$$

where $\mathrm{tol}_{\mathrm{total}}$ is the user-specified convergence criterion for a solution to be considered acceptable. Note that Y_{approx} is probably not exactly the same solution as one would find if one had solved Eq. (9) directly, but that by the user's error specification Eq. (11), Y_{approx} is a perfectly acceptable solution.

This approach is great if one finds that the approximate solution satisfies Eq. (11). But one would prefer not to risk the possibility that after grinding a large-scale steady-state reacting-flow simulation to convergence, a process that could take months, one could discover that the solution obtained, Y_{approx}, does not satisfy the user-specified error tolerance Eq. (11); at that point it would not be clear how to proceed. Also, it would be helpful to know how to set the error tolerances for the reduced model ω_{approx} and the convergence criteria $\mathrm{tol}_{\mathrm{approx}}$ in order to be certain of finding an acceptable solution. One does not want to set the tolerances too tight, otherwise one will waste CPU time, by using reduced chemistry models that are larger than necessary, or by running the calculation determining Y_{approx} to higher accuracy than is necessary.

If one were able to guarantee that the reduced chemistry model reproduced the full-chemistry source term within a tolerance ε for all Y's in some range $\boldsymbol{Y}$

$$-\varepsilon < \omega(Y) - \omega_{\text{approx}}(Y) < \varepsilon \qquad \forall Y \in \boldsymbol{Y} \tag{12}$$

and if the Y_{approx} found by solving the reduced-model computational fluid dynamics (CFD) simulation in Eq. (8) numerically fell in the range, i.e. $Y_{\text{approx}} \in \boldsymbol{Y}$, then by adding Eqs. (10) and (12)

$$-\left(\text{tol}_{\text{approx}} + \varepsilon\right) < \theta\left(Y_{\text{approx}}\right) + \omega\left(Y_{\text{approx}}\right) < \text{tol}_{\text{approx}} + \varepsilon \tag{13}$$

Comparison of Eqs. (13) and (11) reveals that if one ran the CFD calculation to a computational to a convergence $\text{tol}_{\text{approx}} = \text{tol}_{\text{total}} - \varepsilon$, one could be certain that the resulting Y_{approx} would be a satisfactory solution of the full chemistry model, i.e. Eq. (11) would be satisfied. Methods for constructing reduced models guaranteed to satisfy Eq. (13) over a known range of conditions are discussed in detail in Section III.C below. The practical effectiveness of this overall error control approach, once one has an approximate source term ω_{approx} that satisfies Eq. (13) has recently been demonstrated for 1–d and 2–d steady laminar flame simulations.

The challenge is to construct reduced chemistry models which are fast to evaluate, yet which still satisfy the error tolerance Eq. (13). One effective approach to this is the Adaptive Chemistry method (Schwer *et al.*, 2003a, b), where different reduced chemistry models are used under different local reaction conditions. For example, in the 1−d steady premixed flame studied by Oluwole *et al.* (Oluwole *et al.*, 2006), six different reduced chemistry models were used, and the full chemistry model only had to be used at about 20% of the grid points, Fig. 14.

This approach makes it easier to satisfy Eq. (13) using small chemistry models, but it requires a method to easily obtain a set of reduced models appropriate for different conditions. In the next section, we present a method for constructing the various reduced models, and for identifying the range of reaction conditions where each should be used.

C. Constructing Reduced Chemistry Models Satisfying Error Bounds Over Ranges

The algebraic equations in the chemistry operator ω can be much more complicated than the equations usually used for the transport operator Θ, but it is still possible to bound the error introduced by approximating ω, by using interval analysis. Interval analysis is a branch of mathematics that considers how mathematical operations affect intervals (ranges) $[Y_{\text{low}}, Y_{\text{high}}]$ rather than single points Y. For error control, one wants to rigorously bound the error

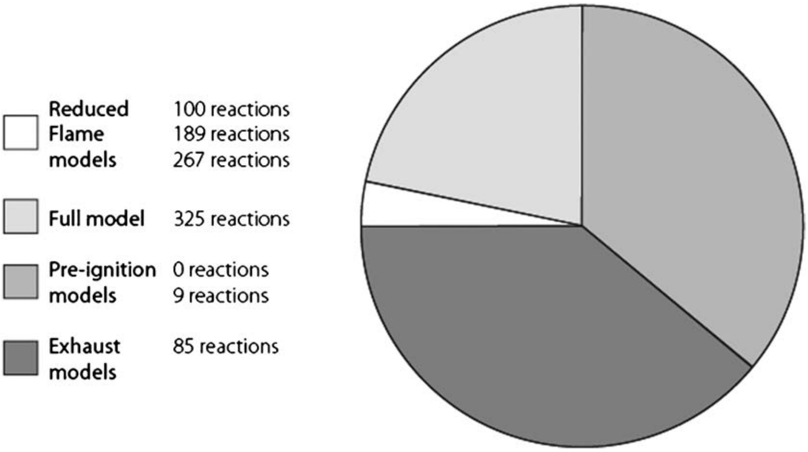


FIG. 14. Fraction of the grid points used by various models during a 1–d Adaptive Chemistry simulation of a premixed stoichiometric methane–air flame. The full chemistry model is GRI-Mech 3.0. Reduced models were accurate at about 80% of the grid points.

made by approximating ω not just at a single point Y, but over entire ranges of Y values that might be encountered by the solver in the process of solving Eq. (8). This is particularly important since ω is usually a highly nonconvex function of Y, i.e. just because an approximation to $\omega(Y)$ works well at certain points Y_1, Y_2, etc., does not guarantee that the approximation will be accurate at other points lying between Y_1, Y_2, etc. (Oluwole *et al.*, 2006; Sirdeshpande *et al.*, 2001). Hence understanding the behavior of ω and its approximations over intervals $[Y_{\text{low}}, Y_{\text{high}}]$ is critical to achieve effective error control.

Interval analysis is trickier than ordinary mathematics, since in interval analysis some arithmetic operations give outputs which are overestimates of the true range. In other words, if a function $f(x)$ takes on values between $f_{\min}$ and $f_{\max}$ for input values of x in the interval $[x_{\text{low}}, x_{\text{high}}]$, interval analysis may give you an output range $[f^*_{\min}, f^*_{\max}]$, where $f^*_{\max} > f_{\max}$ and/or $f^*_{\min} < f_{\min}$.

As a result, interval analysis tends to overestimate error bounds. However, there are clever ways to reduce this overestimation, for example by appropriately grouping terms. One of the best and most computationally efficient ways to minimize overestimation of the bounds on the approximation error is to replace terms $f(Y)$ in $\omega(Y)$ by their Taylor models, e.g. the first-order Taylor model is given by Eq. (14):

$$f(Y) = f(Y^{(o)}) + \sum_n \left(\frac{\partial f}{\partial Y_n}\right)_{Y^{(o)}} \left(Y_n - Y_n^{(o)}\right) + \frac{1}{2}\sum_{m,n} \left(Y_m - Y_m^{(o)}\right)\left(\frac{\partial^2 f}{\partial Y_m \partial Y_n}\right)_{\zeta} \left(Y_n - Y_n^{(o)}\right) \tag{14}$$

In ordinary arithmetic one normally neglects the final term, making this an approximation rather than a rigorous equality, since all one knows is the interval in which ζ lies:

$$\zeta \in [Y^{(o)}, Y] \tag{15a}$$

But with interval analysis, it is no problem to rigorously compute a range that is guaranteed to bound that final term in Eq. (14) over any specified range of Y. The first terms are simple polynomials which can be bounded easily and accurately using interval analysis. So, although we do not know the value of ζ, we can still rigorously bound any function $f(Y)$ over this range of Y as long as we know its value and its first derivative evaluated at $Y^{(o)}$, and the algebraic form of its second derivatives. The Taylor models themselves can be constructed automatically by the computer using automatic differentiation (AD) technology, for example using the DAEPACK software to do all the book-keeping and to compute all the derivatives (Tolsma and Barton, 2000). So with this approach one can easily construct a fairly tight bound on the error made by straightforward approximations to $\omega(Y)$ over any specified range $[Y_{\min}, Y_{\max}]$.

1. *Reaction Elimination via Optimization*

A particularly simple method of approximating the chemical source term $\omega(Y)$ is "reaction elimination": all one does is delete reactions which are numerically insignificant under the current reaction conditions. As shown by Bhattacharjee *et al.* (2003), at any set of single point reaction condition(s), one can rigorously identify the smallest possible set of reactions which reproduces the rate at which species are made or consumed chemically, and the rate at which heat is released due to chemical reactions, cumulatively $\omega(Y)$, to within a user-specified tolerance vector "tols", Eq. (16). Mathematically, the process of finding the smallest possible model is a constrained interval optimization

$$\min \Sigma z_m \tag{15b}$$

subject to the constraint

$$-\text{tols} < \omega_{\text{true}} - \omega_{\text{approx}}(\text{Z}) < \text{tols} \tag{16}$$

where Z is a vector of ones and zeros; if $z_m = 1$ reaction m is included in the approximate model. Because both the function to be optimized and the constraint depend linearly on the optimization variables z_m, it can be solved easily to global optimality.

A free computer service that automatically performs this type of chemistry model-reduction is available through the CMCS Web portal (http://www.cmcs.org). This web software can also provide an interval where the reduced model is guaranteed to replicate the full model to within user-specified tolerances, i.e. a range $[Y_{\text{low}}, Y_{\text{high}}]$ where the error constraint Eq. (16) is guaranteed to be satisfied (Oluwole *et al.*, 2006). However, sometimes the intervals returned by this software are inconvenient for use in reacting-flow simulations (e.g. the computed range where a reduced model is guaranteed to be accurate might be too narrow to be practically useful), and the range-computing software is limited to models containing about 100 species.

2. *Constructing Reduced Models with User-Specified Valid Ranges*

For CFD reacting-flow simulation applications, instead of having the computer it is often preferable to pre-specify a desired range of reaction conditions $[Y_{\mathrm{low}}, Y_{\mathrm{high}}]$, and then have the computer find a reduced model which is guaranteed to be accurate over this range. To accomplish this, we replace the error constraints Eq. (16) with Eqs. (17)

$$\begin{aligned} &\Sigma \max \mathrm{TaylorModel}_m([Y_{\mathrm{low}}, Y_{\mathrm{high}}]) \times (1 - z_m) < \mathrm{tols} \quad m = 1 \ldots N_{\mathrm{reactions}} \\ &\Sigma \min \mathrm{TaylorModel}_m([Y_{\mathrm{low}}, Y_{\mathrm{high}}]) \times (1 - z_m) > - \mathrm{tols} \end{aligned} \tag{17}$$

where $\mathrm{maxTaylorModel}_m$ is a guaranteed upper bound on $\omega_m(Y)$ over the range $[Y_{\mathrm{low}}, Y_{\mathrm{high}}]$ and $\mathrm{minTaylorModel}_m([Y_{\mathrm{low}}, Y_{\mathrm{high}}])$ is a guaranteed lower bound on $\omega_m(Y)$. The two expressions in Eq. (17) thus represent guaranteed upper and lower bounds, respectively, on each element of $\omega_{\mathrm{true}} - \omega_{\mathrm{approx}}(\mathrm{Z})$ over the range $[Y_{\mathrm{low}}, Y_{\mathrm{high}}]$. These bounds are computed automatically by interval analysis using DAEPACK (Tolsma and Barton, 2000).

The constraints Eqs. (17) are still linear in z_m, so one can still find the global optimum, i.e. the reduced model with the smallest number of reactions that satisfies error constraints Eqs. (17). However, because applying interval analysis to the Taylor Model of $\omega_{\mathrm{true}}(Y) - \omega_{\mathrm{approx}}(Y)$ overestimates the deviation between ω_{true} and ω_{approx}, Eqs. (17) are a tighter constraint than Eq. (16), and the reduced model obtained will generally contain more reactions than that obtained using the first formulation. However, despite the loss in speed up due to the overestimate, this second approach is often much more convenient in practice, and our Taylor Model software can handle very large reaction mechanisms easily. For more technical details, see Oluwole *et al.* (Oluwole *et al.*, 2007).

An example: The temperature field computed for a partially-premixed radially-symmetric methane/air flame is shown in Fig. 15. This is the same $\phi = 2.464$ laminar flame simulated by Bennett *et al.* (2000). We used the same 217 reaction full chemistry model used by Bennett *et al.* (2000) to compute the temperature field shown on the left-hand side of Fig. 15. On the left-hand side is shown the temperature field computed using the full chemistry model everywhere. On the right-hand side is shown the temperature field computed by the Adaptive Chemistry method using 13 different reduced models ranging in size from zero reactions to 156 reactions. As guaranteed by the error control

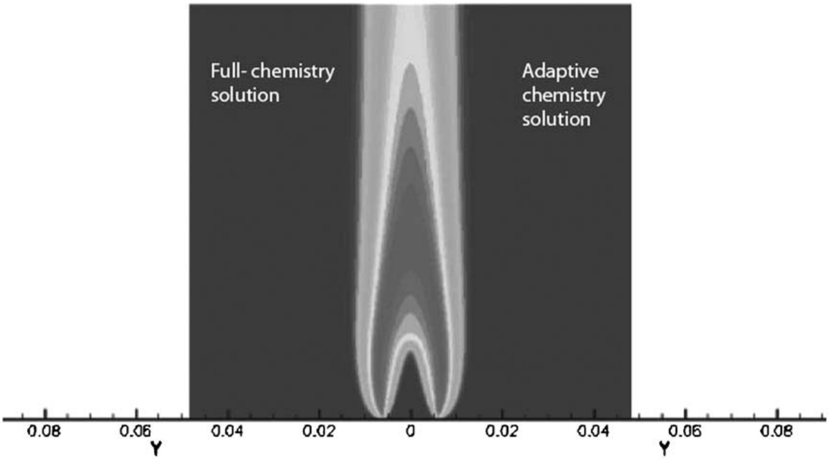


FIG. 15. Computed temperature field for a radially-symmetric partially-premixed methane–air laminar jet flame. The left-hand side shows the temperature field computed using the full chemistry simulation, and the right-hand side shows the temperature field computed using a set of reduced chemistry models that satisfy the error control constraints.

equations above, the two sets of model predictions agree to within the specified error tolerance.

IV. Are Model Predictions Consistent with Experimental Data?

Frequently, the next step after numerically solving a chemical kinetic simulation is to compare the model predictions with some experimental data, to check whether it is consistent with reality at least in one case. This is called "validating" the model. In the 20th century, it was a common practice to plot chemical kinetic model predictions with some experimental data, without any attempt to indicate the uncertainties in either. The reader then had to make his or her own judgment about whether the model and the data were close enough to be considered "consistent", or whether the data had disproved the model.

If the uncertainty information is available, then one can make a much more satisfactory quantitative comparison between the model predictions M_i (uncertainty range u_i) and the corresponding data D_i (uncertainty range σ_i), to determine whether the two are truly consistent. The extent of deviation between the model and the data scaled by the uncertainties is measured by χ^2, Eq. (18).

$$\chi^2 = \Sigma \left[\frac{(D_i - M_i)}{(\sigma_i + u_i)} \right]^2 \tag{18}$$

From statistics, assuming normally distributed errors, the probability that the data and the model would differ so much that $\chi^2 > \Omega$ is given by Eq. (19).

$$P(\Omega) = \int_{\Omega}^{\infty} \frac{t^{m-1} e^{-t/2}}{\Gamma(m) 2^m} \mathrm{d}t \tag{19}$$

where $m = (N_{\text{data}} - N_{\text{adjustable}})/2$. If we have a pure prediction (no parameters adjusted to improve the fit with the data) then $m = N_{\text{data}}/2$. If $P(\chi^2) < 0.1$ that means that there is less than 10% chance that the model and the data are consistent. Or in other words, we can say with 90% confidence that the model and the data are inconsistent, i.e. that the data disproved the model.

A. How to Prove Inconsistency Even if Some Parameters Are Highly Uncertain

The presumption of the approach above is that all the model parameters that significantly affect the model predictions M_i are known with reasonable accuracy, so it is not completely unreasonable to assume that the errors in the M_i are normally distributed (although a normal distribution is only reasonable if the model prediction could vary over the full range $-\infty$ to $+\infty$; for bounded quantities appropriately bounded distributions should be used). However, very often some of the important model parameters are not well established from prior work, so we do not know if their uncertainties are normally distributed. In fact, often it would be most reasonable to use the experimental data to try to determine these highly uncertain parameters $\mathbf{x}$.

When experimental data is unavailable the critical question becomes: is there any physically reasonable choice of these very uncertain parameters $\mathbf{x}$ that makes the model and the data consistent? If not, we can be confident that the model and the data are inconsistent, and if we trust the data and the estimated error bars we can reasonably conclude that the data have disproved the model. Mathematically, we are confident the model and the data are inconsistent if and only if

$$\max_{\mathbf{x}} P(\chi^2(\mathbf{x})) < 0.1 \tag{20}$$

where

$$\chi^2(\mathbf{x}) = \Sigma \frac{(D_i - M_i(\mathbf{x}))^2}{(\sigma_i + u_i(\mathbf{x}))^2} \tag{21}$$

$$M_i(\mathbf{x}) = M_i[Y(t; \mathbf{x})] \tag{22}$$

Where $Y(t;\, \mathbf{x})$ is not known explicitly, but only implicitly as the solution of a system of stiff differential or DAEs such as Eq. (7) which depend parametrically on $\mathbf{x}$, and $\mathbf{x}$ is confined to lie in its physically reasonable range. The functional M_i that relates the recorded experimental measurement M_i to the underlying species concentrations and other state variables could in principle be quite complicated, but often it is linearly dependent on just a few of the species concentrations at some time $t_{\text{measurement}}$. As discussed in a later section, our estimate of the error bar in the model prediction for a fixed $\mathbf{x}$ might depend on $\mathbf{x}$, that is why $u_i(\mathbf{x})$ is written in Eq. (21).

While this is conceptually simple, numerically it is very difficult to find the global maximum of P, because of the fact that the dependence of Y on $\mathbf{x}$ is only known implicitly through a complicated numerical procedure (solving the whole simulation).

Note that if at any point we find any physically acceptable $\mathbf{x}$ which causes $P(\chi^2(\mathbf{x})) > 0.1$, then we know immediately that the model could be consistent with the data, and there is then no need to continue to try to find the global maximum of P. Once one has found any reasonable choice of model parameters that makes the model consistent with the data, then one should move onto questions about what ranges of parameters give predictions consistent with the data, and how the experimental data can be used to help refine and tighten the uncertainty ranges on the parameter values. This parameter-refinement step is indicated in Fig. 1 and has been discussed at length in the literature. It is complicated for nonlinear models with many significantly uncertain correlated parameters, since the range of parameter values consistent with the experiment will generally have a rather complex, often nonconvex, shape in a high-dimensional space. For a recent discussion of how one can deal with the parameter correlations in practice, see Feeley *et al.* (2004).

In the present work, we focus on the opposite case: if we cannot find any choice of $\mathbf{x}$, which makes the model consistent with the data, how do we prove that the model and data are inconsistent? To prove this, we must establish that all physically acceptable $\mathbf{x}$'s give $P(\chi^2(\mathbf{x})) < 0.1$, and for this we need to find or at least set an upper bound on the global maximum of P over the entire range of $\mathbf{x}$. It is in general much more difficult to find guaranteed-globally-optimal maxima than it is to find local maxima. Many of the most efficient methods for finding global optima rely on knowing the explicit function you are trying to optimize. But here the function we are trying to optimize is an integral involving the solution of a complicated stiff system of nonlinear differential equations that depend on the value of $\mathbf{x}$.

In the 20th century there was no way known to solve this type of optimization problem. However, very recently Singer and Barton (2006) have developed a branch-and-bound algorithm that is guaranteed to find the global optimum of $P(\chi^2(\mathbf{x}))$ for the important case where $Y(t;\mathbf{x})$ is specified by a system of nonlinear ODEs (as an initial value problem).

$$\frac{\mathrm{d}Y}{\mathrm{d}t} = F(Y;\mathbf{x}) \tag{23a}$$

$$Y(t_0) = Y_0(\mathbf{x}) \tag{23b}$$

$$\mathrm{M} = \mathrm{M}[Y(t;\mathbf{x});\mathbf{x}] \tag{23c}$$

The numerical procedure used in Singer's global dynamic optimization code (GDOC) software involving establishing bounds on the algebraic equations on the right-hand sides of Eqs. 23a, 23b and 23c over the physically reasonable range of the parameters $\mathbf{x}$ using interval analysis techniques similar to those used in Section III above. Then one constructs linear systems of differential equations that bound Y and P. Finally, one can systematically remove regions of $\mathbf{x}$ space which probably cannot contain the maximum of P. As the $\mathbf{x}$ space is subdivided, all the bounds tighten, until the procedure converges to the true global minimum of $\chi^2(\mathbf{x})$ (which is the global maximum of P). The details with proofs are presented elsewhere. A more concise explanation aimed at kineticists, with some practical examples, has been published recently. Practical kineticists interested in using the GDOC software to convincingly disprove chemical kinetic models are welcome to download it from http://yoric.mit.edu/

At present, the GDOC software can only handle problems with <6 highly uncertain parameters $\mathbf{x}$. It is hoped and expected that in the near future, more efficient algorithms will be found for handling systems with many more uncertain parameters, so that one could more easily determine, even in messy real-world systems, whether the discrepancies one observes between model predictions and data really indicate that the model must be wrong, or just indicate that one has not yet found the best values for the uncertain parameters.

B. How Standard Operating Practice Must Change in 21st Century

As discussed above, numerical tools now exist that make it possible to rigorously test whether chemical kinetic models and data are quantitatively consistent for many important cases. There is good reason to expect that the numerical methods will continue to improve rapidly, so that these quantitative models vs. data tests will become possible for ever more complex experiments. As discussed in Sections II and III, our capability to construct and solve complex simulations is also improving rapidly.

However, it is far from certain that we will be able to access the experimental data we need for these consistency checks (and for subsequent refinement of the models and model parameters) in a useful format. Historically, the field of kinetics has not been consistent in requiring that complete sets of experimental data be archived in the open literature, and in many journals error bars are optional. Instead, researchers often report only fitting parameters that they extracted from their experimental data using a model. The raw data is often lost in a few years, when the researcher or student who measured it moves on to a different position, and the media on which it is stored becomes obsolete. This makes it very difficult to try to reconstruct what was actually observed; in the common case where the model originally used to interpret the data had some flaw this can make the literature report about the experimental measurement almost useless.

The PrIMe project (http://www.primekinetics.org) is addressing these serious problems by developing software tools to make it easier for experimentalists to archive their measurements in an electronic format, so that it will be easier for future researchers to use the data for consistency checks and to refine model parameters.

In many respects, the situation is much worse in kinetic modeling than in experimental work. One seldom sees any kind of uncertainty estimate on a chemical kinetic model prediction, and as discussed extensively in Section II, many models have been incompletely documented, in part due to their reliance on a 20th century data model which made it inconvenient to document uncertainties and information on the origin of rate/thermo/transport parameter estimates. Very commonly, estimated rate constants and thermochemical parameters have been copied from prior reports (without error bars) without any serious attempt to check their veracity. After this cycle repeats several times, rough estimates made 50 years ago eventually become ‘‘firmly established’’ because they have been repeated so often in the literature, not because the numbers were ever really determined precisely. As a consequence, it seems likely that most chemical kinetic models in the literature will be discarded and replaced by new models as efficient reaction-mechanism-generation software based on well-documented rate and thermochemical estimates become widely deployed over the next few years: why bother to track down all the poorly documented assumptions in prior modeling work, when one can construct a brand new, possibly more complete, model-based on the latest rate and thermochemical data and estimates in just a few hours? This paradigm shift in kinetics will be most useful if it is accompanied by much more consistent treatment of uncertainty and uncertainty propagation.

C. Estimating Error Bars on Model Predictions

In order to estimate the uncertainties in a model's predictions (the u_i in Eq. (21)) for consistency checking, one must have estimates of the uncertainties in the model's input parameters. In the 20th century, model input uncertainties were usually converted into model prediction uncertainties using normalized first-order sensitivity coefficients S_{ij}, Eq. (24a). For ODEs and DAEs, first-order sensitivities can be computed very efficiently at the same time the simulation is solved (Maly and Petzold, 1996).

$$S_{ij} = \frac{\partial(\ln M_i)}{\partial(\ln k_j)} \quad i = 1 \ldots N_{\text{predictions}}, \; j = 1 \ldots 2N_{\text{reactions}} \tag{24a}$$

Knowing which of the S_{ij} are large in magnitude can be very helpful to a modeler, but Eq. (24a) requires some care. First, in systems where T and P are not constant, one usually means the sensitivity with respect to the pre-exponential factor in the rate constant, i.e. a uniform scaling of $k_j(T,P)$ at all T,P (Kee *et al.*, 1989).

More importantly, as written, the partial derivative implies that a single rate constant is to be varied, holding all the others constant, and indeed this is the way it is implemented in many sensitivity analysis routines. The index "j" runs out to $2N_{\text{reactions}}$ because each reaction has two rate constants, one for the forward direction, and one for the reverse. However, in order for the model to remain consistent with the laws of thermodynamics, the rate constant for the reverse of reaction "j" must vary simultaneously with the forward reaction, since the two rate constants must maintain a detailed-balance ratio related to the $\Delta G_{\text{reaction}}$, Eq. (2). This can be assured by specifying that the partial derivative is taken only for the forward reaction, while holding the thermochemistry fixed. Note that this also cuts the number of partial derivatives to be computed in half. These sensitivities should then be supplemented with sensitivities to the individual species' thermochemistry as in Eq. (25); overall the number of partial derivatives to be computed per model prediction M_i is $(N_{\text{reactions}} + N_{\text{species}})$, not $2N_{\text{reactions}}$.

$$S_{ij} = \frac{\partial(\ln M_i)}{\partial(\ln k_j)} \quad i = 1 \ldots N_{\text{predictions}}, \; j = 1 \ldots N_{\text{reactions}} \text{ (forward directions only)} \tag{24b}$$

$$S_{im} = \frac{\partial(\ln M_i)}{\partial(\Delta H_f(\text{species } m))} \quad i = 1 \ldots N_{\text{predictions}}, \; m = 1 \ldots N_{\text{species}} \tag{25}$$

If one (optimistically) believes the boundary conditions, initial conditions, and other molecular properties (e.g. transport properties, heat capacities) have all been specified very precisely, and one thinks the uncertainties in the rate constants and the enthalpies are not highly correlated, one might then estimate the uncertainty in prediction M_i by Eq. (26).

$$\delta M_i = u_i \sim [\Sigma|S^*_{ij}(\delta \ln k_j)|^2 + \Sigma|S^*_{im}\delta\Delta H_f \text{ (species } m)|^2]^{1/2} \tag{26}$$

If one believes that the uncertainties in rate constants and thermo are highly correlated, a more conservative (pessimistic) estimate would be

$$\delta M_i = u_i \sim \Sigma|S_{ij}^*(\delta \ln k_j)| + \Sigma|S_{im}^*\delta\Delta H_f \text{ (species } m)| \tag{27}$$

1. *Moving Beyond First-Order Sensitivity Analysis*

Of course, Eq. (24b) and Eq. (25) are only the first derivatives, and Eqs. (26) and (27) are just low-order approximations to u_i; for cases where the first derivative S_{ij} is small or where the uncertainties $\delta\ln(k_j)$ and $\delta\Delta H_f$ are large, one must consider higher-order effects (Li *et al.*, 2002; Phenix *et al.*, 1998; Vuilleumier *et al.*, 1997). Many of the methods of including the higher-order effects, including Monte Carlo simulations, can be computationally expensive if the original kinetic simulation is hard to solve, but these calculations are certainly not impossible with modern computers, particularly if one only wants to consider the effects of large variations in a few key parameters. Certain quantities involving second-order sensitivities can be computed efficiently even for very large systems using adjoint methods (Oezyurt and Barton, 2005). But in the near future, for most cases involving large simulations, we expect that the routine first approximation to the error will be something like Eq. (26) or Eq. (27). In a limited number of important cases, e.g. when an apparent discrepancy has been identified, a more accurate method for uncertainty propagation, and a careful scrutiny of the most important estimated input uncertainties $\delta\ln(k_j)$ and $\delta\Delta H_f$(species m), will be required.

2. *Correlated Uncertainties in Model Input Parameters*

Often the uncertainties in the rate constants are highly correlated. For example, in large kinetic models, most of the rate constants are derived from a small number of functional group-based estimates, such as Evans–Polanyi parameters (Eq. (1)), some of which are significantly uncertain. In these cases, it would make sense to vary the Evans–Polanyi parameters, and so adjust the whole group of related rate constants, rather than allowing each rate constant to vary separately from all of the rest. For example, the high-pressure-limit rate for pent-2-yl radical addition to butadiene is not known very precisely, it could be uncertain by a factor of 10 in certain temperature ranges, but it almost certainly must be within a factor of two of the high-pressure-limit rate for but-2-yl + butadiene, because of the similarity between the functional groups involved. Similarly, many of the enthalpy values used in large kinetic models are very highly correlated, many of them coming from group-additivity estimates using the group values published more than 30 years ago in Benson's book (Benson, 1976). It would be very reasonable as a first approximation to vary the group values rather than varying the thermochemistry of all the individual species independently. However, to our knowledge, no one has yet produced software that makes it convenient to compute how these correlated uncertainties in the inputs to large kinetic models affect the model predictions.

3. Model Truncation Error

Model predictions can also be inaccurate due to the incompleteness of the chemical model, e.g. if some reactions or species were incorrectly omitted from the mechanism. If the missing species or reactions are completely missing from the database used by the model-construction software, there is no easy way to detect them (though perhaps a human expert might notice the omission in the tree databases described in Section II). There is certainly chemistry which is not well understood, even in the well-studied thermal gas-phase chemistry of small organic molecules; for example some of the important reactions of peroxyl radicals are still unclear (Taatjes, 2006), the true reaction path for $CH + N_2$ was only recently identified (Moskaleva and Lin, 2000), and recently some reactions that occur over ridges rather than saddle points have been identified (Townsend *et al.*, 2004). It will be some time before there is a community consensus on how to correctly generalize from some of these observations.

It is also possible that reactions or species might be omitted from the model even if the chemistry is well-understood and was correctly included in the databases, for example if the species-selection tolerances were not set tightly enough during model-construction. Song *et al.* (2002) have discussed methods for trying to detect and avoid this type of omission. It should be possible to improve this approach using interval analysis techniques (Oluwole *et al.*, 2007; Oluwole *et al.*, 2006); perhaps it is possible to rigorously bound the error in the model predictions due to the finite error tolerance in species-selection, at least for certain types of simulations.

4. Other Sources of Error in Model Predictions

Finally, the model predictions always include some error because the simulation never exactly models the physical situation in the experiment, and because of numerical errors in solving the simulations. Both types of errors can be very important, even dominant, though in many systems these errors are smaller than the uncertainties in the chemistry model. It is very difficult to deal with the distressingly common problem that the model does not match the physical experimental situation; at present only human experts can address this problem (e.g. by numerically tests which demonstrate that including more experimental details does not materially change the simulation predictions). On the other hand, most of the numerical errors that arise when solving the simulations can be rigorously addressed through correctly designed algorithms and software.

At present, black-box programs for solving initial value problems typically maintain fairly rigorous automatic control of the numerical errors. However, for multidimensional reacting-flow simulations (boundary value problems), error control is usually not automated, relying on the user performing the correct checks to try to ensure that the calculation has actually converged to something close to the true solution. Schwer *et al.* (2003b) show how a relatively straightforward calculation can converge to an unphysical "solution". Many algorithms use low-order upwinding, but this introduces "numerical diffusion" which can lead to significantly erroneous conclusions in some cases; this issue is often difficult to deal with because removing the upwinding approximation can make the simulations numerically unstable. A major effort is needed to upgrade the numerical solvers for multidimensional reacting-flows and to incorporate error control, to make it easier to ensure that numerical errors are not significantly perturbing the simulation results, and to increase our confidence that two different researchers will arrive at the same converged solution if they start with the same reacting-flow equations. After this is achieved, we would then be in a better state to rationally decide which algorithms are really the most efficient and robust for solving different types of reacting-flow simulations.

V. Summary and Outlook

In this article, we have presented an overview of predictive chemical kinetics, a very broad field which is rapidly advancing in many ways. This manuscript also discusses some recent technical progress addressing a few of the sub-problems that must be overcome before predictive chemical kinetics can completely fulfill its potential:

(1) making the automated chemical mechanism-construction more extensible
(2) speeding multidimensional reacting-flow simulations involving complex chemistry, while maintaining rigorous error control
(3) conclusively determining whether or not model predictions are consistent with experimental data, when both are uncertain.

While there has been very significant recent progress on all three sub-problems, as discussed in Sections II, III, and IV, clearly much remains to be done before the iterative predict-measure-learn loop shown in Fig. 1 can function smoothly for most systems of interest to chemical engineers. For example, Section II describes a good method for storing and updating chemical information, but these data trees will need to be significantly expanded before predictive chemical kinetics will be usefully accurate for most systems. Similarly, in Section III we present a promising method for faster, error-controlled reacting-flow simulations, but this method is limited to steady-state processes, and these computations are still extremely difficult. Section IV discusses a method for performing rigorous consistency checks, but this method presumes that one has access to a lot of data and error bars on both data and model predictions that are usually not available at present. The overall message is that predictive chemical kinetics looks very promising in the early 21st century, but very significant efforts, including changes in the day-to-day functioning of the kinetics community, will be required in order to achieve this promise.

This manuscript certainly does not address all of the challenges to the advance of predictive chemical kinetics; for example it focuses only on homogeneous systems, and it is not even a complete review of efforts on any of the three sub-problems discussed in detail. However, this overview will hopefully be helpful to researchers trying to comprehend this broad and somewhat fragmented area of chemical engineering research, a field that will grow rapidly in the 21st century, both in technical capability and in practical importance.

ACKNOWLEDGMENTS

The work reported here was made possible through financial support from the US Department of Energy, the US National Science Foundation, and Alstom Power. The RMG program was written by Jing Song, with contributions from David Matheu, Sarah Petway, Sandeep Sharma, Paul Yelvington, and Joanna Yu. The functional group trees were designed by Sumathy Raman, who also computed many of the parameters, and the GUI was written by John Robotham. The numerical methods presented in Sections III and IV were developed in Prof. Paul Barton's group at MIT by John Tolsma and Adam Singer, and were adapted to chemical kinetic problems by Binita Bhattacharjee, O. Oluwole, Douglas Schwer, and James Taylor. Assistance from Greg Beran and Susan Lanza in the preparation of this manuscript is gratefully acknowledged. The figures were drawn by Tammy Keithley.

NOMENCLATURE

χ^2	weighted sum of squares of deviations between model and experiment
α	Evans–Polanyi slope
n	temperature exponent
P	probability that one would measure data giving such a low χ^2 value if the model were true
Ω	cut-off in χ^2 for a model to acceptably fit the data
Z	integer array indicating which reactions are kept or deleted
T [=]Kelvin	temperature
P [=]atm	pressure
t [=]sec	time

E_o [=]kJ/mole	Evans–Polanyi barrier for hypothetical isothermic reaction
k [=]sec^{-1} or [=]liter/mole-second	Rate constant (unimolecular or bimolecular)
A [=](units of k)/(Kelvin)^n	modified-Arrhenius prefactor
E_a [=]kJ/mole	activation energy
R [=]J/mole-Kelvin	gas constant
k_B [=]J/Kelvin	Boltzmann's constant
h [=]Js	Planck's constant
G [=]kJ/mole	Gibb's free energy
H [=]kJ/mole	enthalpy
Θ [=]mole/s	rate of change in species due to transport
ω [=]mole/s	rate of change in species due to chemical reaction Θ, ω are generalized to include change in other state variables as well as species.
Y	the state variables in a CFD simulation
D	the array of measured data values
M	the array of model predictions for these data
σ	uncertainties in D
u	uncertainties in M
$\mathbf{x}$	adjustable parameters in a model
F	function returning rate of change of Y
M	functional for computing M from Y
S	normalized sensitivity of model predictions M

REFERENCES

Aghalayam, P., Park, Y. K., and Vlachos, D. G. *AIChE J.* **46**, 2017–2029 (2000).

Bennett, B. A. V., McEnally, C. S., Pfefferle, L. D., and Smooke, M. D. *Combust. Flame* **123**, 522–546 (2000).

Benson, S. W., "Thermochemical Kinetics". 2d ed. Wiley, New York (1976).

Berger, M. J., and Oliger, J. *J. Comput. Phys.* **53**, 484–512 (1984).

Bernan, K. E., Campbell, S. L., and Petzold, L. R., "Numerical Solution of Initial-Value Problems in Differential-Algebraic Equations". 2nd ed. Elsevier, New York (1989).

Bhattacharjee, B., Schwer, D. A., Barton, P. I., and Green, W. H. *Combust. Flame* **135**, 191–208 (2003).

Blurock, E. S. *J. Chem. Inf. Comput. Sci.* **35**, 67 (1995).

Blurock, E. S. *J. Chem. Inf. Comput. Sci.* **44**, 1336–1347 (2004a).

Blurock, E. S. *J. Chem. Inf. Comput. Sci.* **44**, 1348–1357 (2004b).

Bozzelli, J. W., Chang, A. Y., and Dean, A. M. *Int. J. Chem. Kinet.* **29**, 161–170 (1997).

Broadbelt, L. J., Stark, S. M., and Klein, M. T. *Ind. Eng. Chem. Res.* **33**, 790–799 (1994).

Brown, P. N., Byrne, G. D., and Hindmarsh, A. C. *SIAM J. Sci. Stat. Comput*. 1038–1064 (1989).

Buda, F., Heyberger, B., Fournet, R., Glaude, P.-A., Warth, V., and Battin-Leclerc, F. *Energy Fuels* **20**, 1450–1459 (2006).

Caracotsios, M., and Stewart, W. E. *Comput. Chem. Eng.* **19**, 1019–1030 (1995).

Chang, A. Y., Bozzelli, J. W., and Dean, A. M. *Zeitschrift fuer Physikalische Chemie (Muenchen)* **214**, 1533–1568 (2000).

Chen, J. Y. *Combust. Sci. Technol.* **57**, 89–94 (1988).

Chevalier, C., Warnatz, J., and Melenk, H. *Berichte der Bunsen-Gesellschaft* **94**, 1362–1367 (1990).

Chinnick, S. J., Baulch, D. L., and Ayscough, P. B. *Chemom. Intell. Lab. Syst.* **5**, 39 (1988).

Dean, A. M. *J. Phys. Chem.* **89**, 4600–4608 (1985).

Dente, M., Ranzi, E., and Goossens, A. G. *Comput. Chem. Eng.* **3**, 61–75 (1979).

DiMaio, F. P., and Lignola, P. G. *Chem. Eng. Sci.* **47**, 2713 (1992).

Feeley, R., Seiler, P., Packard, A., and Frenklach, M. *J. Phys. Chem. A* **108**, 9573–9583 (2004).

Feng, X.-J., and Rabitz, H. *Biophys. J.* **86**, 1270–1281 (2004).

Forst, W., "Theory of Unimolecular Reactions". Academic Press, New York (1973).

Frenklach, M., Wang, H., and Rabinowitz, M. J. *Prog. Energy Combust. Sci.* **18**, 47 (1992).

Gear, C. W. *Commun. ACM* **14**, 176–179 (1971).

Glaude, P. A., Conraud, V., Fournet, R., Battin-Leclerc, F., Come, G. M., Scacchi, G., Dagaut, P., and Cathonnet, M. *Energy Fuels* **16**, 1186–1195 (2002).

Hillewaert, L. P., Diericks, J. L., and Froment, G. F. *AIChE J.* **34**, 17 (1988).

Kee, R. J., Rupley, F. M., and Miller, J. A. Sandia National Laboratories, Livermore, CA, USA (1989).

Kojima, S. *Combust. Flame* **99**, 87–136 (1994).

Lay, T. H., and Bozzelli, J. W. *Chem. Phys. Lett.* **268**, 175–179 (1997a).

Lay, T. H., and Bozzelli, J. W. *J. Phys. Chem. A* **101**, 9505–9510 (1997b).

Lay, T. H., Bozzelli, J. W., Dean, A. M., and Ritter, E. R. *J. Phys. Chem.* **99**, 14514–14527 (1995).

Lay, T. H., Yamada, T., Tsai, P.-L., and Bozzelli, J. W. *J. Phys. Chem. A* **101**, 2471–2477 (1997).

Li, G., Wang, S.-W., Rabitz, H., Wang, S., and Jaffe, P. *Chem. Eng. Sci.* **57**, 4445–4460 (2002).

MacNeice, P., Olson, K. M., Mobarry, C., de Fainchtein, R., and Packer, C. *Comput. Phys. Commun.* **126**, 330–354 (2000).

Maly, T., and Petzold, L. R. *Appl. Numer. Math.* **20**, 57 (1996).

Matheu, D. M. Thesis, PhD. Dissertation Massachusetts Institute of Technology (2003).

Matheu, D. M., Dean, A. M., Grenda, J. M., and Green, W. H. *J. Phys. Chem. A* **107**, 8552–8565 (2003a).

Matheu, D. M., Green, W. H., and Grenda, J. M. *Int. J. Chem. Kinet.* **35**, 95–119 (2003b).

Matheu, D. M., and Grenda, J. M. *J. Phys. Chem. A* **109**, 5332–5342 (2005a).

Matheu, D. M., and Grenda, J. M. *J. Phys. Chem. A* **109**, 5343–5351 (2005b).

Melius, C. F., and Allendorf, M. D. *J. Phys. Chem. A* **104**, 2168–2177 (2000).

Milford, J. B., Gao, D., Russell, A. G., and McRae, G. J. *Environ. Sci. Technol.* **26**, 1179–1189 (1992).

Moskaleva, L. V., and Lin, M. C. *Proc. Combust. Inst.* **28**, 2393–2401 (2000).

Oezyurt, D. B., and Barton, P. I. *Ind. Eng. Chem. Res.* **44**, 1804–1811 (2005).

Oluwole, O. O., Bhattacharjee, B., Barton, P. I., and Green, W. H. *Combust. Flame* **146**, 348–365 (2006).

Oluwole, O. O., Bhattacharjee, B., Tolsma, J., Barton, P. I., and Green, W. H. *Combust. Theor. Model.* **11**(1), 127–146 (2007).

Peters, N., and Rogg, B., "Reduced Kinetic Mechanisms for Applications in Combustion Systems". Springer-Verlag, Berlin and New York (1993).

Petway, S. V. Application of Automatic Reaction Model Generation to Chlorine-Initiated Neopentane Oxidation, (Poster Presentation) 6th International Conference on Chemical Kinetics, NIST (Gaithersburg, MD) (2005).

Petzold, L. R. Report SAND82-8637. Sandia National Laboratories (1982).

Phenix, B. D., Dinaro, J. L., Tatang, M. A., Tester, J. W., Howard, J. B., and McRae, G. J. *Combust. Flame* **112**, 132–146 (1998).

Pope, S. B. *Combust. Theor. Model.* **1**, 41–63 (1997).

Prickett, S. E., and Mavrovouniotis, M. L. *Comput. Chem. Eng.* **21**, 1219–1235 (1997a).

Prickett, S. E., and Mavrovouniotis, M. L. *Comput. Chem. Eng.* **21**, 1237–1254 (1997b).

Prickett, S. E., and Mavrovouniotis, M. L. *Comput. Chem. Eng.* **21**, 1325–1337 (1997c).

Quann, R. J., and Jaffe, S. B. *Ind. Eng. Chem. Res.* **31**, 2483–2497 (1992).

Ranzi, E., Faravelli, T., Gaffuri, P., and Sogaro, A. *Combust. Flame* **102**, 179–192 (1995).

Ratkiewicz, A. T., and Truong, T. N. *Int. J. Quantum Chem.* **106**, 244–255 (2006).

Ritter, E. R., and Bozzelli, J. W. *Int. J. Chem. Kinet.* **23**, 767–778 (1991).

Ruscic, B., Pinzon, R. E., Morton, M. L., von Laszevski, G., Bittner, S. J., Nijsure, S. G., Amin, K. A., Minkoff, M., and Wagner, A. F. *J. Phys. Chem. A* **108**, 9979–9997 (2004).

Schwer, D. A., Lu, P., and Green, W. H. *Combust. Flame* **133**, 451–465 (2003a).

Schwer, D. A., Lu, P., Green, W. H.Jr., and Semiao, V. *Combust. Theor. Model* **7**, 383–399 (2003b).

Schwer, D. A., Tolsma, J. E., Green, W. H., and Barton, P. I. *Combust. Flame* **128**, 270–291 (2002).

Shenvi, N., Geremia, J. M., and Rabitz, H. *J. Phys. Chem. A* **106**, 12315–12323 (2002).

Singer, A. B., and Barton, P. I. *J. Global Optim.* **34**, 159–190 (2006).

Sirdeshpande, A. R., Ierapetritou, M. G., and Androulakis, I. P. *AIChE J.* **47**, 2461–2473 (2001).

Song, J. Thesis, PhD. Dissertation Massachusetts Institute of Technology (2004).

Song, J., Stephanopoulos, G., and Green, W. H. *Chem. Eng. Sci.* **57**, 4475–4491 (2002).

Song, J., Sumathi, R., Green, W. H., and Van Geem, K. *Int. J. Chem. Kinet.* (submitted, under revision).

Stein, S. E., and Rabinovitch, B. S. *J. Chem. Phys.* **58**, 2438–2445 (1973).

Stein, S. E., and Rabinovitch, B. S. *J. Chem. Phys.* **60**, 908–917 (1974).

Sumathi, R., and Green, W. H. Jr. Theor. Chem. Acc. **108**, 187–213 (2002).

Susnow, R. G., Dean, A. M., Green, W. H., Peczak, P., and Broadbelt, L. J. *J. Phys. Chem. A* **101**, 3731–3740 (1997).

Taatjes, C. A. *J. Phys. Chem. A* **110**, 4299–4312 (2006).

Tolsma, J., and Barton, P. I. *Ind. Eng. Chem. Res.* **39**, 1826–1839 (2000).

Tomlin, A. S., Turanyi, T., and Pilling, M. J. *Compr. Chem. Kinet.* **35**, 293–437 (1997).

Tonse, S. R., Moriarty, N. W., Brown, N. J., and Frenklach, M. *Isr. J. Chem.* **39**, 97–106 (1999).

Townsend, D., Lahankar, S. A., Lee, S. K., Chambreau, S. D., Suits, A. G., Zhang, X., Rheinecker, J., Harding, L. B., and Bowman, J. M. *Science* **306**, 1158–1161 (2004).

Truhlar, D. G., Garrett, B. C., and Klippenstein, S. J. *J. Phys. Chem.* **100**, 12771–12800 (1996).

Van Geem, K., Reyniersa, M. F., Marina, G., Song, J., Matheu, D. M., and Green, W. H. *AIChE J.* **52**(2), 718–730 (2006).

Venkatesh, P. K., Chang, A. Y., Dean, A. M., Cohen, M. H., and Carr, R. W. *AIChE J.* **43**, 1331–1340 (1997).

Vuilleumier, L., Harley, R. A., and Brown, N. J. *Environ. Sci. Technol.* **31**, 1206–1217 (1997).

Warth, V., Battin-Leclerc, F., Fournet, R., Glaude, P. A., Come, G. M., and Scacchi, G. *Comput. Chem.* **24**, 541–560 (2000).

Wilk, R. D., Cohen, R. S., and Cernansky, N. P. *Ind. Eng. Chem. Res.* **34**, 2285–2297 (1995).

Yelvington, P. E., Rallo, M. B. I., Liput, S., Tester, J. W., Green, W. H., and Yang, J. *Combust. Sci. Technol.* **176**, 1243–1282 (2004).

Yu, J., Sumathi, R., and Green, W. H. *J. Am. Chem. Soc.* **126**, 12685–12700 (2004).

Zachariah, M. R., and Melius, C. F., Bond-additivity correction of ab initio computations for accurate prediction of thermochemistry, *Comput. Thermochem.* **677**, 162–175 (1998).

Zarth, C., Tirtowidjojo, M., and West, D. Annual Meeting Archive – American Institute of Chemical Engineers, Indianapolis, IN, United States (November 3–8, 2002), pp. 2627–2646 (2002).

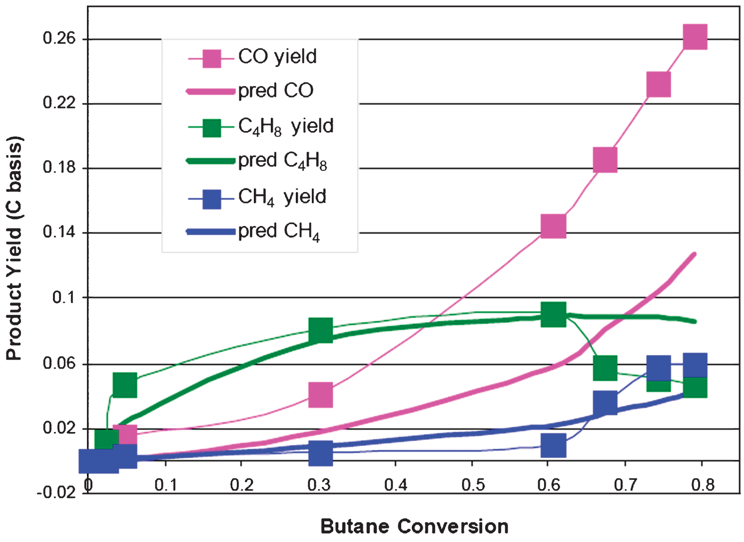


Plate 1. Model predictions vs. experimental measurements of Wilk *et al.* (1995) from butane oxidation at 715 K. The predicted yields are generally within a factor of two of the experimental data, reflecting roughly factor of two uncertainties in rate constant estimates. As is typical, the discrepancies are largest at high conversions, both because the errors in the parameters cumulate, and because the model may be missing some reactions of the minor byproducts (For Black and White version, see page 28).

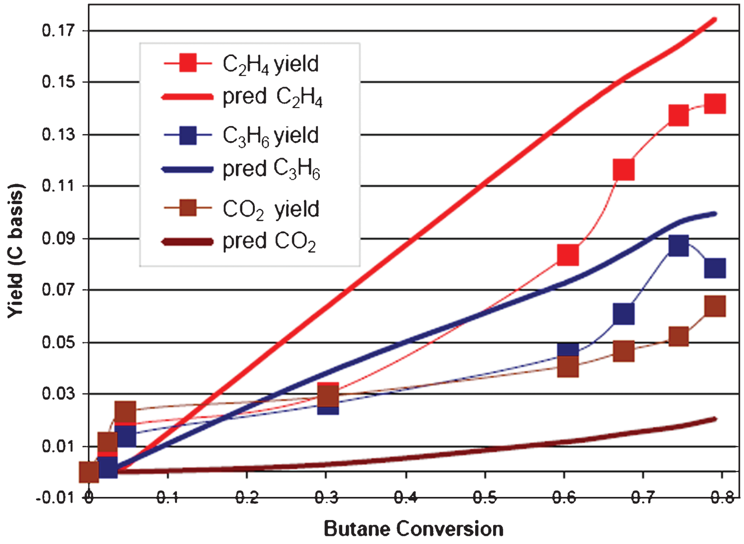


Plate 2. Model predictions vs. experimental measurements of Wilk *et al.* (1995) for ethene, propene, and CO_2, same conditions as Fig. 12. The large discrepancy in the CO_2 predictions at the lowest conversions suggests that the model does not accurately represent the true boundary conditions at the inlet (For Black and White version, see page 29).