An Intelligent System for Reaction Kinetic Modeling and Catalyst Design

Santhoji Katare,† James M. Caruthers, W. Nicholas Delgass, and Venkat Venkatasubramanian*

Center for Catalyst Design, School of Chemical Engineering, Purdue University, West Lafayette, Indiana 47907

The continuing development of high throughput experiments (HTEs) in catalysis has dramatically increased the amount of data that can be collected in relatively short periods of time. Even when HTEs can afford "Edisonian" discovery, how can the increasing amounts of data be converted to knowledge to guide the next search in the vast design space of catalytic materials? To address this question, we recently proposed a catalyst design architecture that uses detailed kinetic models. In this paper, we describe Reaction Modeling Suite—a rational, automated, and intelligent environment, based on systems, artificial intelligence, and optimization techniques that aid the development of kinetic models. We demonstrate its utility in developing a kinetic model for propane aromatization on zeolite. We also show the proof-of-concept of how a genetic algorithm-based search strategy can be used to search for kinetic parameters that correspond to an improved catalyst.

1. Introduction

The design of new materials possessing desired macroscopic properties or performance characteristics is an important, although difficult, problem. Materials design finds applications in the development of diverse materials such as polymers, polymeric composites, blends, paints and varnishes, refrigerants, solvents, drugs, pesticides, and so on. The traditional approach requires the designer to hypothesize a molecule or material, synthesize it, and experimentally evaluate it to see if it meets the desired properties or performance criteria and to reformulate the design if the targets are not met. This Edisonian guess-and-test method is time-consuming, expensive, cumbersome, and complicated—time-consuming and expensive because of the nature of the experiments and cumbersome and complicated because of the underlying huge, nonlinear search space.

In the area of catalyst design, experimentally tuning catalyst structure to improve performance is well known.[1] Advances in surface science techniques[2] that enable manipulation of individual atoms on the catalyst surface in real time have[3] immensely contributed to improved understanding of the catalysts. With the advent of high throughput and combinatorial methods, experimental guidance techniques such as hierarchical screening,[4] evolutionary ideas,[5] and those based on statistics[6] have become relevant. Despite these efforts, the nonlinearity and the size of the underlying search space still pose a strong challenge to systematic design. Also, design techniques that are mainly driven by experiments will only enable in the collection of information, and unless there is a method to convert that information into knowledge and insight, a general catalyst design methodology would remain an unsolved problem.

Theory and model based catalyst design strategies are well known in the literature. These include the idea of using qualitative reasoning and knowledge-based systems,[7,8] efforts toward using computational models and calculations to guide the search for new materials,[9,10] and those that use detailed microkinetic models to study catalytic systems.[11] A more comprehensive review of catalyst design techniques is available elsewhere.[12]

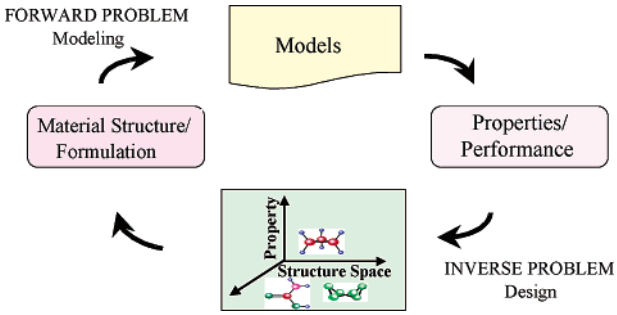


Figure 1. Schematic of the forward and inverse problems in computer-aided materials design.

Computer-aided materials design[13] offers an attractive alternative to the above approaches, whereby the design problem involves the use of computer-based procedures to systematically identify appropriate molecular structures that satisfy a set of desired properties. In general, the overall task requires the solution of two subproblems as shown in Figure 1: the forward problem, which involves the computation of performance measures or physical, chemical, and/or biological properties from the product structure and formulation/composition; and the inverse problem, which entails the identification of the appropriate molecular structure or composition given the desired macroscopic properties. To solve the inverse problem, which is the true design problem, a robust forward model is essential. This forward model development is complicated because the underlying system is often complex. The main challenges include identification of the key descriptors that characterize the system under study and development of a methodology to link the material descriptors to performance. We recently proposed a methodology[14] for building forward models for designing catalysts. This

* To whom correspondence should be addressed. E-mail: venkat@ecn.purdue.edu. Phone: 765 494 0734. Fax: 765 494 0805.

† Current address: Department of Chemical Engineering, University of Houston, Houston, TX 77204.

10.1021/ie034067h CCC: $27.50
Published on Web 02/11/2004

involves a systematic, rational, and iterative technique for knowledge extraction (KE) from high throughput experimentation data. The KE procedure facilitates convergence to a quality predictive model from an initial approximate model by systematically incorporating any new information about the system as and when it is available. In this paper, we describe the Reaction Modeling Suite (RMS)—a collection of systems, optimization, and artificial intelligence based tools that enable KE by aiding the expert in building robust kinetic models.

The main objective of this work is the design and development of systems tools for integrating large amounts of diverse sets of data with the hypothesis generation and screening process, at a pace commensurate with the rate of data production. In particular, emphasis would be on developing user-driven systems tools to offer a systematic, less error-prone, automated environment for an expert to postulate, evaluate/optimize, and refine reaction mechanisms. The tools would allow rigorous analysis of multiple reaction mechanisms in the light of data, with minimum human intervention. This would make the whole process user friendly and quick. The rest of this paper is organized as follows. RMS will be described in the next section along with a brief review about the requirements of an automated kinetic model building system and the state-of-the-art in this area. In section 3, the various capabilities of the RMS tools in hypothesis screening, reaction network analysis, model refinement, model discrimination, and experimental formulation will be demonstrated by developing a kinetic model for propane aromatization on H-ZSM-5 zeolite catalyst. This section will also include a proof-of-concept of the inverse problem that involves the search for an improved catalyst for paraffin aromatization. The main contributions will be summarized and general conclusions will be drawn in the final section.

2. Reaction Modeling Suite

The key requirement of any model building procedure is the rapid screening of an expert-postulated reaction mechanistic hypothesis that could explain the data. This process should be fast enough to keep pace with the rate of data generation from combinatorial and high throughput experiments. Another challenge is to develop user-driven tools that naturally relate to the domain expert. Toward this end, we have developed the RMS that enables rational, automated reaction kinetic modeling and thus facilitates knowledge archiving and retrieval. The software in RMS allows easy encoding of reaction chemistry knowledge in terms of pseudo-English language rules and enables automated and fast hypothesis testing by screening through multiple hypothesis in a systematic manner.

Traditionally, the reaction-modeling problem has been an art tackled by chemists and chemical engineers or other domain experts. On the basis of their experience or knowledge about the system at hand, the experts first formulate a set of heuristics or rules that appear to govern the process. These rules directly translate to a reaction mechanism, and a mathematical model is then constructed from it. Depending on the discrepancies in the predictions of the model and experimental results, the experts go back to the initial stage of rule formulation and consider alternative or additional rules. When the reaction network consists of a large number of reactions and chemical species, the development of the mathematical model becomes cumbersome. The overall process as shown in Figure 2 is therefore "Edisonian" and is often protracted, cumbersome, and expensive. It is protracted because even a slight change in one of the reactions in the mechanism leads to multiple changes in the mathematical equations that represent these reactions. Since building a feasible mechanism starting from a plausible set of steps is often iterative, the whole process becomes time-consuming and highly prone to errors. Any efforts to automate the same using computer-assisted methods can lead to considerable savings in time and money.

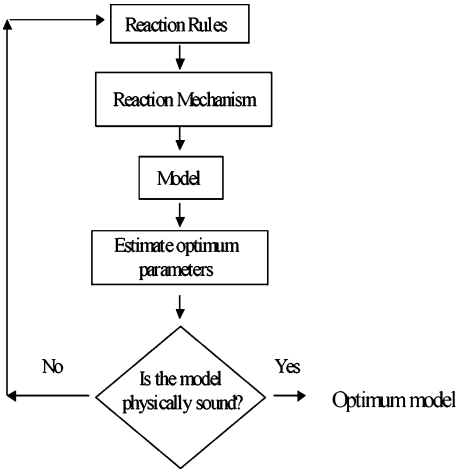


Figure 2. Traditional model building process.

The importance of developing robust kinetic models for understanding the underlying chemical system is well-known in the literature. For example, concepts such as stoichiometric network analysis,[15] mathematically controlled comparison and canonical representation of differential equation models,[16] development of large scale reaction networks based on elementary reactions,[11,17] deduction of reaction mechanisms given a set of elementary steps,[18] and reverse engineering of reaction mechanisms[19,20] have attracted widespread attention. Software tools for automating the process of reaction model building have also been available. Table 1 shows a list of software tools available in the literature that aid the process of modeling chemical reaction networks. The tools have been categorized based on their ability to (1) formulate a reaction network from high level chemistry rules, (2) visualize the reaction network, (3) parse the reaction network to get a mathematical model, (4) solve the model and optimize the parameters, and (5) analyze the results statistically. For detailed reviews of software tools for reaction kinetic modeling, the reader is referred to Arkin[21] and Katare.[12]

Drawing ideas from the traditional modeling methodologies, we propose a set of tools with a systems viewpoint for effectively and efficiently implementing the ideas of chemical reaction modeling within the perspective of computer-aided materials design. Specifically, the current work deals with a framework for developing forward prediction models for surface reactions from a catalyst design (Figure 1) perspective. The key steps involved in the process of model building are as follows: generation of the simplest plausible reaction mechanism; translation of the mechanism to a computationally tractable mathematical model; solving the model to estimate the parameters in light of high throughput and/or insufficient experimental data; refin-

Table 1. List of Software Tools That Aid in the Process of Modeling Chemical Reaction Networks[a]

no.	software	descriptors	reference
1	Reaction Description Language	F	Pricket and Mavrovouniotis, 1997[30]
2	DBsolve	VPOS	Goryanin et al., 1999[99]
3	E-cell	VPO[b]	Tomita et al., 1999[100]
4	Gepasi	POS	Mendes, 1993[101]
5	CRNT	POS	ftp://ftp.che.rochester.edu/pub/feinberg/
6	Dynetica	VPO	You et al., 2003[102]
7	XMG	FPOS	Green et al., 2001[39]
8	NetGen	FPOS	Broadbelt et al., 1994[24]
9	IBM CKS	VPOS	www.almaden.ibm.com/st/msim/ckspage.html
10	MKM	POS	http://www.aue.auc.dk/~stoltze/mkm/main.html
11	Mitsubishi	FPOS	Hostrup and Balakrishna, 2001[75]
12	Chemkin	POS	Kee et al., 1989[103]
13	KINAL A	POS	Turanyi, 1990[38]

[a] The descriptors show the ability of the tool to formulate a reaction network from higher level rules (F), visualize the network (V), parse the network into mathematical model (P), solve the model and optimize the parameters (O) and analyze the results statistically (S). [b] Solves but does not optimize the parameters.

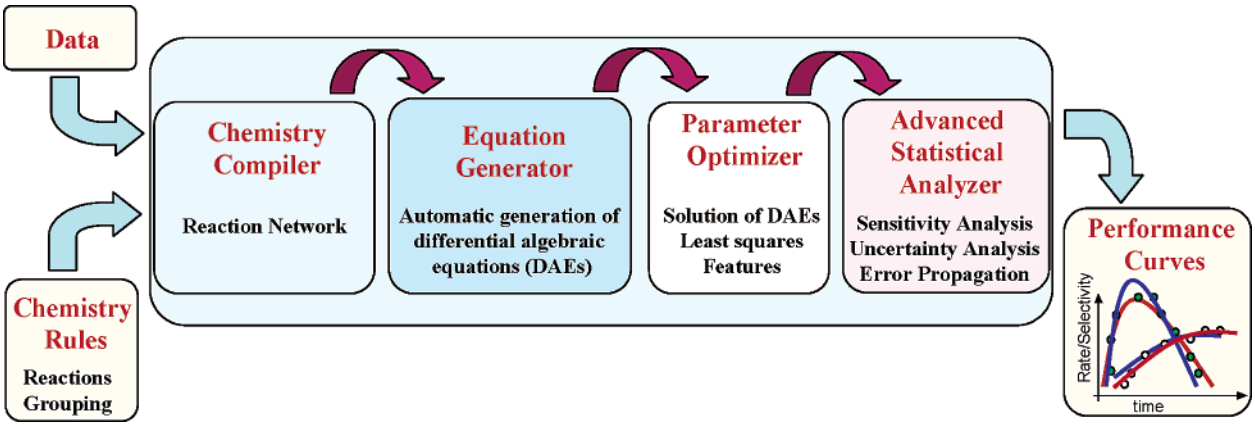


Figure 3. Reaction Modeling Suite.

ing the model to better fit the data by altering the mechanism; suggesting new experiments that could help discriminate among multiple models.

The main challenges involved are as follows:

Mechanism Generation from Reaction Rules.

1. Unambiguous representation of the large number of reactions and species.

2. A compiler that understands the generic reaction rules and a network generator that applies these rules recursively to all possible reactant species.

3. Pruning the reaction mechanism to get the simplest possible model that can explain the data consistently.

4. Assimilating thermokinetic data and experimental information involving the various reactions/species to minimize the number of thermodynamic and/or kinetic parameters to be optimized and to aid the process of parameter estimation.

Parameter Estimation.

5. Robust solvers that handle the large number of differential-algebraic equations and parameter estimation techniques that will evaluate the validity of the proposed mechanism to model data.

6. Evaluation of the multiple solutions for the parameters that explain the data equally well.

Feature Extraction.

7. Feature extraction techniques to identify the discrepancies between the key features of the model predictions and that of the data so that they can be used for MR and experimental formulation.

8. Mapping the feature discrepancies to the mechanistic rules that generated the reaction mechanism.

Statistical Analysis.

9. Estimation of the robustness of the model.

We now describe RMS (Figure 3), the tools that provide solutions to most of the above challenges. Specifically, in this section, we present our implementation of an English language rules-to-reaction network compiler that translates pseudo-English language rules into a chemical reaction network. Then we describe a hybrid algorithm for parameter estimation that affords a thorough and efficient search of the nonlinear parameter space. This is then followed by a feature extraction procedure that enables a natural way for evaluating the validity of a model in light of the data. Finally, we explain the statistical analysis tools that have been developed to analyze the robustness of a kinetic model.

2.1. English Language Rules-to-Reaction Network Compiler. Building a kinetic model is initiated by an expert who proposes an initial set of reaction rules that is most likely to explain the experimentally observed product distribution. For example, for a solid acid catalyst system, adsorption, desorption, protolysis, beta-scission, oligomerization, dehydrogenation, aromatization, etc., form a plausible rule set which gives rise to a large number of elementary reactions. The first step of postulating a hypothesis as rules is the most critical one as it drives the subsequent process of model building and evaluation. Moreover, the task of model refinement based on the model−data mismatch is typically aimed at altering one or more of these basic reaction rules rather than independently changing the elementary reaction steps. This is because changing a single rule affects several chemically similar elementary steps. Thus, the process of hypothesis screening is largely dependent on how well the expert is able to postulate and iteratively manipulate these reaction rules. There-

Table 2. Attributes of a Molecule, Atom, Bond, and a Fragment in RDL++

	attributes
molecule	atom, bond, fragment, reactant/product, network to which it belongs
atom	neighboring atoms, list of bonds, charge, element type, molecule or fragment to which it belongs
bond	list of atoms, list of neighboring atoms, order, molecule or fragment to which it belongs
fragment	type, list of atoms and bonds, molecule to which it belongs

fore, any effort to automate the model building process should aid the expert as much as possible in his/her natural working language.

2.1.1. Representation of Molecules and Reactions. Software that enables hypothesis screening and model building should provide as much flexibility as possible to the expert in postulating and manipulating the various reaction rules. It should be possible to include new reaction rules readily. The molecules and reaction rules should be encoded in the natural language that is used by an expert while formulating them. Previous works on automated model generation have used several representations including extended SMILES notation[22] and bond-electron (BE) matrices[23] to represent molecules and BE matrices,[24] functional group vectors,[25−27] and pseudo-English language rules[28] to represent reaction rules. SMILES and BE matrices can become cumbersome[29] as the complexity of the reactive intermediates increases. Extending the functional group vectors representation to non-hydrocarbon chemistry may not be straightforward.

Prickett and Mavrovouniotis[30] have tried to overcome the above shortcomings by introducing a pseudo-English language to describe reaction rules along with a compiler to translate these instructions into a reaction network. Their Reaction Description Language (RDL) uses a sequence of commands to locate the reaction site, to manipulate the reactant to form the product, and to prune the reaction network with a syntax that mimics the way reactions are typically described by the chemists.

RMS has been designed to facilitate hypothesis generation and screening, and one of its key requirements is that it should enable the initiation of this process using a natural language interface. RDL[31] satisfies this condition, and so we have designed and developed the Reaction Description Language Plus Plus (RDL++) as a system that extends RDL. RDL++ can be used to model reaction mechanisms on solid acid-based catalysts, like zeolites, and is designed to be more user-driven and extendible. Also RDL++ has been integrated with other tools that are geared toward building robust reaction kinetic models.

2.1.2. The Reaction Description Language Plus Plus (RDL++). RDL++ is a compiler that translates chemistry rules in pseudo-English language to a reaction network. Molecules and reaction networks in RDL++ are represented in an object oriented fashion along the lines of RDL.[28] Molecules are graphs whose nodes represent the atoms and the edges denote the connectivity between the atoms. The various attributes of a molecule are shown in Table 2. The molecule is characterized by its atoms, bonds, fragments (rings and chains), and its role as a reactant or a product in a particular reaction. An atom's attributes include its neighboring atoms, the list of bonds, its charge, element type, and the molecule or the fragment to which it belongs. A bond has its list of atoms and the list of its neighboring atoms, order, and the identity of the molecule or fragment to which it belongs. Finally, the fragment is represented by its type (ring or chain), a list of atoms and bonds, and the molecule to which it belongs.

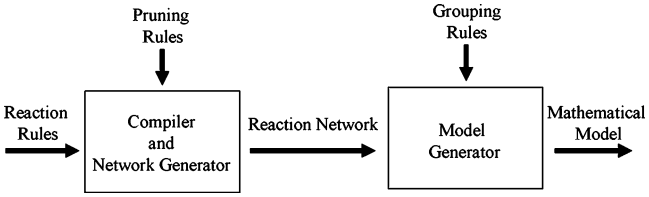


Figure 4. Schematic of the Reaction Description Language Plus Plus.

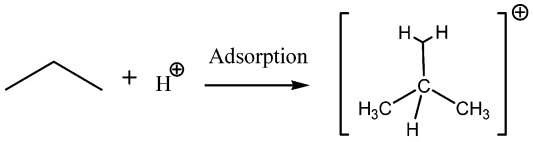
Table 3. A Typical Rule in RDL++: Adsorption of a Paraffin To Give a Carbonium Ion

```
{
// description of the rule
(reaction-name "adsorption of paraffin")

// rate constant definition
(rate-constant kaa)

// reaction site identifier
(label-site m1 reactant)
(label-site c1 (find neutral-carbon))

// local pruning statements
(require (paraffin m1))
(forbid (cyclic m1))
(forbid (less-than (size-of m1) 2))

// transformation statements
(add-charge c1)
(connect c1 neutral-hydrogen)
}
```

A schematic of the RDL++ is shown in Figure 4. RDL++ consists of (1) a compiler and a network generator that transforms reaction rules and global pruning rules to a reaction network and (2) a model generator that generates a model from the reaction network and a set of grouping rules. The compiler translates the English language rules to intermediate code or patterns which are then recursively applied by the network generator to all the species in the reaction mixture to generate the reaction network.

Reaction rules in RDL++ are similar to that of RDL[28] and consist of three important blocks of statements with specific roles: (1) identification of the reaction site(s) among the reactive species, (2) transformation of the reactive sites to products, and (3) local pruning of the reactions based on the reactive sites or on the products formed. The pruning rules restrict the type of reactants that can enter a reaction and the products that can be formed. A typical reaction rule that describes adsorption of a paraffin to form a carbonium ion (eq R.1) is shown in Table 3. Every statement is in the form of a production rule and is enclosed in parantheses. Comments are preceded by a pair of forward slashes. The statement at the beginning of every rule describes the name of the reaction. This is followed by the identification of the rate constant of the reactions generated by this rule. The rule then identifies the reaction site—a neutral carbon in the reactant. The local pruning rules constrain the various possibilities, and in the current version of paraffin adsorption (Table 3), it is required that the reactant be a paraffin and that any cyclic species be forbidden. Also,

the size or the number of carbon atoms in the reactant has been constrained to be less than 8. So, the pruning rules enforce the condition that paraffin adsorption can only take place on an acyclic paraffin up to C_7. Note that any of these pruning rules can be relaxed as per the requirement of the system under consideration.[12]



(R. 1)

2.1.3. Network Generator. The reaction rules are converted by the compiler to an intermediate code that contains the information about the generic reactions. These patterns are now applied to all the species in the reaction mixture to create the actual reaction network. For example, if there are four rules and two initial reactants, after the application of the four rules to the two initial species, the rules are again applied to the products formed from in the first pass. This process is repeated until each of the rules is applied to all the species in the reaction mixture.

2.1.4. Model Reduction—Pruning of Reaction Networks. Since the analysis of complex reaction networks typically requires more data than are often available, the mathematical model of the reaction network formed from the reaction rules is often pruned. The area of simplification of mathematical models that describe reaction mechanisms has been reviewed by Tomlin and co-workers[32] and Mavrovouniotis.[33] The methods of model reduction can be widely divided into two parts—reduction based on the time scale analysis and the techniques that are not based on the time evolution of the species. Pseudo-steady-state analysis that converts the differential equations into algebraic equations, computational singular perturbation based on reaction rates, and the inertial low dimensional manifold technique of Mass and Pope[34] that use the species rate trajectory to distinguish between the slow and fast rates are examples of the former class. Sensitivity analysis, which studies the importance of reactions and species to identify the redundancy in the reaction network, lumping of a group of species such as isomers or chemically similar groups, forms the basis of time-independent techniques for model reduction.

In summary, the reaction network is often pruned using a variety of techniques,[32,33,35] including sensitivity analysis,[36−38] math-programming methods,[39−42] and manifold techniques.[34,43] However, these methods depend on the elimination of species and/or kinetic steps that are not important for a particular data set, where there is no assurance that this species and/or reaction mechanism will not become important for other reaction conditions. In contrast, Mavrouvouniotis and Prickett[31] have suggested model reduction methods where known reactivity relationships between different species are used to eliminate unimportant reactions; alternatively, reaction rate based techniques have been used to control the size of the network.[44]

RDL++ consists of two types of pruning rules—local pruning rules that are restricted to a particular reaction rule and global pruning rules that are applied to all the reaction rules. The local pruning rules include (1) forbidding a particular reactant from undergoing a reaction or a product from being formed, (2) limiting the number of carbons in the reactants and/or the products, (3) requiring or forbidding a particular pattern in the reactant and/or product. For example, adsorption of paraffin is restricted; paraffin with fewer than two carbons—methane—will not adsorb (Table 3). The global pruning rules apply to all the reaction rules and hence can restrict the formation of certain types of products by any reaction rule. As shown in Table 4, the global pruning rules, for example, forbid the formation of species with two adjacent double bonds, triple bonds among the products—defined as a "trifin" product, species that have a double bond and a positive charge and species with charges on two different atoms. Another powerful global pruning rule is forbidding the formation of any isomers. This reduces the size of the network to a great extent and is particularly useful when building models with data that cannot distinguish between different isomers. Although the word "pruning" implies that it happens after the actual transformation takes place, pruning rules defined in terms of the reactants forbid the concerned reaction from being executed for unqualified reactants.

Table 4. Global Pruning Rules

```
{
(forbid (adjacent double-bond))
(forbid (trifin product))
(forbid positive-carbon attached-to double-bond)
(forbid (double charge))
(forbid (isomer product))
}
```

Table 5. Carbonium Ion Desorption To Give a Paraffin

```
{
(reaction-name "desorption of carbonium")
(rate-constant kad)
(label-site c1+ (find positive-carbonium))
(label-site h1 (find neutral-hydrogen attached-to c1+))
(disconnect c1+ h1)
(subtract-charge c1+)
}
```

Table 6. Dehydrogenation of a Carbonium Ion Gives Rise to a Carbenium Ion and H_2

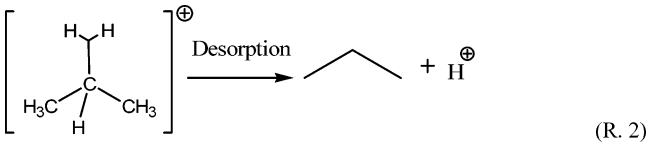
```
{
(reaction-name "dehydrogenation of carboniums")
(rate-constant kcd)
(label-site c1+ (find positive-carbonium))
(forbid (quaternary c1+))
(label-site h1 (find neutral-hydrogen attached-to c1+))
(label-site h2 (find neutral-hydrogen attached-to c1+))
(disconnect h1 c1+)
(disconnect h2 c1+)
(connect h1 h2)
}
```

2.1.5. Examples of Network Generation with RDL++. We illustrate the utility and versatility of the RDL++ chemistry compiler in translating the pseudo-English language rules into a reaction network using an example of a set of paraffin reactions on a zeolite catalyst. This reaction mechanism consisting of paraffin adsorption, desorption, dehydrogenation, and protolysis is a critical subset of the reactions leading to paraffin aromatization, a commercially important reaction for the transformation of paraffin to gasoline. The RDL++ rules corresponding to these reactions are shown in Table 3 and Tables 5−7 and their representative reactions in R.1 through R.4, respectively. Specifically, reaction set S_1 consists of (1) adsorption of a paraffin to form a carbonium ion (Table 3), (2) desorption of the carbonium ion to give back the paraffin (Table 5), (3) carbonium ion dehydrogenation to give a carbenium ion

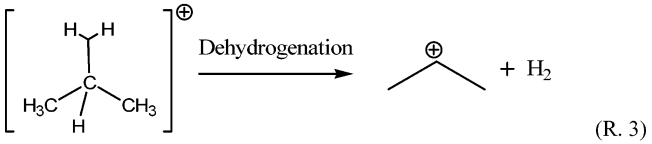
Table 7. Protolysis of a Carbonium Ion To Form a Carbenium Ion and a Paraffin

```
{
(reaction-name "protolysis of carbonium ions")
(rate-constant kp)
(label-site c1+ (find positive-carbonium))
(label-site c2 (find neutral-carbon attached-to c1+))
(label-site h1 (find neutral-hydrogen attached-to c1+))
(disconnect c1+ c2)
(disconnect c1+ h1)
(connect c2 h1)
}
```

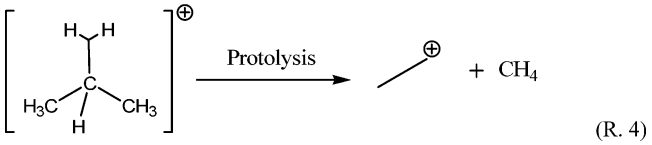
and H_2 (Table 6), and (4) protolysis of carbonium ion to give a paraffin and a carbenium ion (Table 7).



(R. 2)



(R. 3)



(R. 4)

These reactions describe the adsorption of paraffin to form carbonium ions which subsequently protolyze and dehydrogenate to give carbenium ions or desorb to give back the paraffin. To eliminate infeasible products, the first four global pruning rules as shown in Table 4, which forbid adjacent double bonds and triple bonds, a carbon with a charge attached to a double bond, and the same molecule with two charges, are used in generating the reaction network.

To study the effect of the reaction rules on the input species, different reactants—propane (C_3), butane (C_4), isobutane (2m−C_3), pentane (C_5), 2-methylbutane (2m−C_4), and 2,2-dimethylpropane (2,2m−C_3)—have been used to generate the reaction networks. The choice of the input species was based on the fact that we wanted to examine the effect of the size of the reaction network and the type of species formed depending on the various isomers of small alkanes as input reactants. Also, the effect of the symmetric nature of the reactants on the resultant reaction network will be studied. The number of paraffin, carbonium, and carbenium ions and the total number of species including H_2 in the product mixture for reaction mechanism S_1 with different input species are tabulated in Table 8.

Propane adsorbs to give a three-carbon carbonium ion which can then protolyze to give methane and ethyl carbenium ion, or ethane and methyl carbenium ion. Ethane subsequently can adsorb to form a two-carbon carbonium ion, which can then protolyze to give a methyl carbenium ion and methane. In the final reaction mixture methane, ethane, and propane along with their adsorbed carbonium species are present. The two- and three-carbon carbenium ions are mainly formed by the dehydrogenation of the respective carbonium ions and a methyl carbenium ion is formed when ethyl

Table 8. Product Distribution as a Result of Reaction Set S_1 with Different Paraffins as Input[a]

input	paraffin	carbonium ions	carbenium ions	total no. of species including H_2
C_3	3	3	4	11
C_4	4	5	6	16
2m-C_3	4	5	6	16
C_5	5	8	9	23
2m-C_4	6	11	12	30
2,2m-C_3	5	7	7	20

[a] All isomers of all the species are generated.

Table 9. Number of Reactions as a Result of Reaction Set S_1[a]

input	reaction type I	II	III	IV	total no. of reactions
C_3	3	3	3	3	12
C_4	5	5	5	6	21
2m-C_3	5	5	5	5	20
C_5	8	8	8	10	34
2m-C_4	11	11	11	14	47
2,2m-C_3	7	7	6	7	27

[a] All isomers of all the species are generated.

carbonium ion undergoes protolysis. The total number of elementary reactions that result from each of the four rules is as shown in Table 9. Due to the restriction in the paraffin adsorption rule (Table 3), species with fewer than two carbons cannot undergo adsorption; hence, methane does not adsorb to give a single carbon carbonium ion. The increase in the number of isomers with increasing carbon numbers in the reactant molecule, and hence the increase in the number of possible valid reaction sites, is responsible for the increase in the number of reactions in the various reaction networks.

The symmetry of the molecules also affects the number of isomers and hence affects the number of reactions due to each of the reaction rules and the total of species that are present in the reaction network. For example, 2-methylbutane, which has the most number of isomers as compared to that of the other input species, gives rise to the maximum number of ions. Also the number of reactions due to protolysis, which involves breaking of a bond between two carbons of a carbonium ion, increases for reactants that are asymmetric since asymmetric species have more isomers. Similarly, since 2,2-dimethylpropane is the most symmetric species among all the five carbon reactants considered in this study, it gives rise to the lowest number of reactions and total number of species among all the five carbon reactants. All the above computations took only a few seconds on an Intel Xeon dual processor machine with 1 GHz processors, 512K cache, and 2 GB RAM running under the RedHat Linux 7.3 operating system. The computer program approximately consists of 14K lines of C++ code.

Reaction set S_1 primarily consisted of paraffin activation reactions in which the paraffin was adsorbed and the resulting carbonium ion was transformed to paraffin and carbenium ions. Besides the chemistry of carbonium ions, the process of paraffin aromatization also involves the transformation of carbenium ions. The carbenium ions formed due to the dehydrogenation and protolysis of carbonium ions or by the adsorption of an olefin (R.5) can desorb to give olefins (R.6), break into smaller carbenium ions (R.7), combine with an olefin to form a larger carbenium ion (R.8), swap its positive charge with

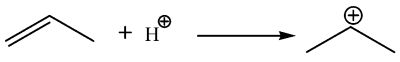
Table 10. Adsorption of Olefin To Form a Carbenium Ion

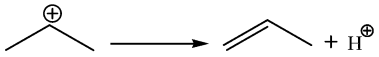
```
{
(reaction-name "adsorption of olefin")
(rate-constant koa)
(label-site b1 (find double-bond))
(label-site c1 (find neutral-carbon attached-to b1))
(label-site c2 (find neutral-carbon attached-to b1))
(forbid (diene m1))
(decrease-order-of b1)
(add-charge c1)
(connect c2 neutral-hydrogen)
}
```

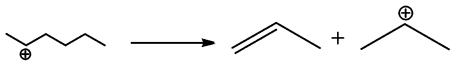
Table 11. Desorption of the Carbenium Ion To Give Back the Olefin

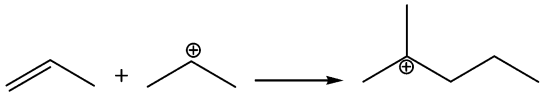
```
{
(reaction-name "desorption of adsorbed olefins")
(rate-constant kod)
(label-site c1+ (find positive-carbon))
(label-site c2 (find neutral-carbon attached-to c1+))
(label-site b1 (find single-bond connecting c1+ c2))
(label-site h1 (find neutral-hydrogen attached-to c2))
(disconnect c2 h1)
(increase-order-of b1)
(subtract-charge c1+)
}
```

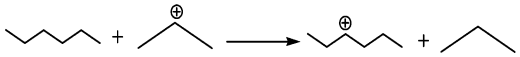
a paraffin/monoene/diene (R.9), or combine with a paraffin to give a larger paraffin (R.10).

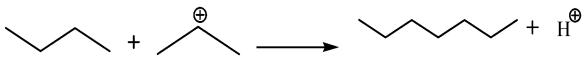

(R. 5)


(R. 6)


(R. 7)


(R. 8)


(R. 9)


(R. 10)

The above reactions are only representative of the various reaction rules. These reactions can be easily manipulated by changing only a few words in the rules.[12] In summary, reaction set S_2 consists of the following paraffin and olefin reactions: (1) adsorption of paraffin to form a carbonium ion (Table 3); (2) desorption of the carbonium ion to give back the paraffin (Table 5); (3) carbonium ion dehydrogenation that results in a carbenium ion and H_2 (Table 6); (4) protolysis of carbonium ion to give a paraffin and a carbenium ion (Table 7); (5) adsorption of olefin to form a carbenium ion (Table 10); (6) Desorption of the carbenium ion to give back the olefin (Table 11); (7) β-scission of a carbenium ion to a smaller carbenium ion and an olefin (Table 12); (8) oligomerization of a carbenium ion and an olefin to form a larger carbenium ion (Table 13); (9) hydride transfer between a carbenium ion and a paraffin/monoene/diene (Table 14); (10) alkylation of a carbenium ion with a paraffin to give rise to a larger paraffin (Table 15). It is important to recall that the network generator operates recursively on all the

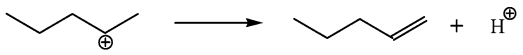
Table 12. β-Scission of a Carbenium Ion in to a Smaller Carbenium Ion and an Olefin

```
{
(reaction-name "Beta-scission")
(rate-constant kb)
(label-site m1 reactant)
(label-site c1+ (find positive-carbon))
(label-site c2 (find neutral-carbon attached-to c1+))
(label-site c3 (find neutral-carbon attached-to c2))
(label-site c4 (find neutral-carbon attached-to c3))
(label-site b1 (find single-bond connecting c1+ c2))
(label-site b2 (find single-bond connecting c3 c2))
(label-site b3 (find single-bond connecting c3 c4))
(forbid (less-than (size-of m1) 4))
(disconnect c2 c3)
(add-charge c3)
(subtract-charge c1+)
(increase-order-of b1)
}
```

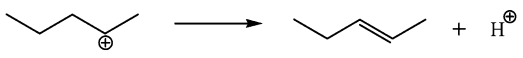
Table 13. Oligomerization of a Carbenium Ion and an Olefin To Form a Larger Carbenium Ion

```
{
(reaction-name "Oligomerization")
(rate-constant kolig)
(label-site m1 reactant)
(label-site b1 (find double-bond))
(label-site c1 (find neutral-carbon attached-to b1))
(label-site c2 (find neutral-carbon attached-to b1))
(forbid (diene m1))
(search-network-for
   (label-site m2 reactant)
   (label-site c3+ (find positive-carbon))
)
(require (less-than (plus (size-of m1) (size-of m2)) 8))
(decrease-order-of b1)
(connect c1 c3+)
(add-charge c2)
(subtract-charge c3+)
}
```

new species formed during the course of generation of the network starting from the initial set of reactants specified by the user. When the generation of isomers is forbidden by the global rule (Table 4), all isomers are considered to be the same. For example, when a five-carbon carbenium ion desorbs to give an olefin, the double bond could be placed on either side of the carbon atom that originally had the charge. Specifically


(R. 11)

or


(R. 12)

However, when the formation of isomers is forbidden, 2-pentene is considered to be the same species as 1-pentene, and so the second reaction (R.12) does not become part of the reaction network. Consequently, the number of species and the total number of reactions in the reaction network reduce to a great extent when the global rule that forbids the generation of isomers is used. This could be very useful in generating compact reaction networks especially when the kineticist does not have access to analytical data that can distinguish among the various isomers. Although the current implementation of RDL++ retains the first isomer formed and rejects the subsequent ones, one could use thermokinetic information to make this procedure more chemically consistent.

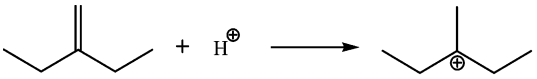
Table 14. Hydride Transfer That Transfers a Charge from a Carbenium Ion to a Paraffin

```
{
(reaction-name "Hydride transfer")
(rate-constant kh)
(label-site m1 reactant)
(label-site c1+ (find positive-carbon))
(forbid (allylic m1))
(search-network-for
   (label-site m2 reactant)
   (label-site c2 (find neutral-carbon))
   (label-site h1 (find neutral-hydrogen attached-to c2))
   (require (less-than (plus (size-of m1) (size-of m2)) 8))
   (require (or (and (paraffin m2) (at-least (size-of m2) 2))))
   (forbid (allylic m2))
)
(disconnect c2 h1)
(add-charge c2)
(connect c1+ h1)
(subtract-charge c1+)
}
```

Table 15. Alkylation of a Carbenium Ion with a Paraffin To Give Rise to a Larger Paraffin

```
{
(reaction-name "Alkylation")
(rate-constant kalk)
(label-site m1 reactant)
(label-site c1+ (find positive-carbon))
(search-network-for
   (label-site m2 reactant)
   (label-site c2 (find neutral-carbon))
   (label-site h1 (find neutral-hydrogen attached-to c2))
   (require (paraffin m2))
)
(require (less-than (plus (size-of m1) (size-of m2)) 8))
(connect c1+ c2)
(disconnect h1 c2)
(subtract-charge c1+)
}
```

To demonstrate the effect of forbidding the generation of isomers on the size of the reaction network, we will analyze the reaction networks that result from reaction set S_2 with and without the global rule that restricts isomer generation. The product distribution as a result of the reaction network from reaction set S_2 when isomers are not generated is shown in Tables 16−19. The number of reactions as a result of the various reaction rules is shown in Table 20. The notations in the tables follow the IUPAC convention. The abbreviations m and e stand for methyl and ethyl, respectively. Also the superscript = denotes the double bond and the prefix to the carbon (C) denotes the location of the double bond or that of the positive charge as the case may be. The total number of paraffin, olefin, carbonium, and carbenium ions is shown in Table 16. It is interesting to note that the number of all the species generated remains the same irrespective of the input reactants except for the number of olefins when 2,2-dimethylpropane (2,2m−C_3) is the input. However, as shown in Tables 17−19, the paraffins, olefins, and carbonium ions formed from each of these species as inputs are quite different from each other. For example, the five-, six-, and seven-carbon paraffins generated when pentane is the input reactant are linear as against the branched products obtained when 2-methylbutane is the input. All the paraffins except methane, as shown in Table 17, adsorb to give the corresponding carbonium ions shown in Table 19. This is because the paraffin adsorption rule (Table 3) forbids adsorption of methane that leads to the formation of the highly unstable single-carbon carbonium ion. Similar to the paraffins and the carbonium ions, the distribution of carbenium ions (not shown as they are identical to that of carbonium ions shown in Table 19) is similar to the distribution of olefins (Table 18). This is intuitive because olefins adsorb to give the corresponding carbenium ions. For example, for the case when propane is the input species, 2-ethyl-1-butene (2e-1$C_4^=$) adsorbs to give rise to 3m-3C_5^+ as follows:



(R. 13)

Table 20 shows the number of reactions as a result of each of the reaction rules and the total number of reactions in the reaction network. The alkylation reaction (Table 15) that involves the fusion of a carbenium ion with an alkane to give rise to a larger alkane is the least restricted rule and so accounts for the most number of reactions. This is because any of the paraffin and carebenium ion species in the reaction network can undergo this reaction provided that the resultant paraffin species has fewer than eight carbon atoms. The β-scission reaction involves the fragmentation of a carbenium ion to a smaller carbenium ion and an olefin. Also this rule requires the presence of a three-carbon linear chain attached to positive carbon. When propane (C_3), 2-methylpropane (2m-C_3), or 2,2-dimethylpropane (2,2m-C_3) are the input reactants, carbenium ions that satisfy this constraint are not created, and hence β-scission reactions do not occur.

The statistics of the reaction network that results from the seaction set S_2 when the generation of all isomers of all species is allowed is shown in Table 21 and Table 22. From Table 23, it is evident that the number of species and the number of reactions in the reaction network increase when isomers are generated. The reaction network generated is the smallest when the most symmetric molecule—2,2-dimethyl propane (2,2m-C_3)—is used as the input reactant. This is because this molecule results in products that are highly symmetric and hence have fewer isomers. The vast change in the size and type of the reaction networks that result because of just one change in the reaction rules (forbid-

Table 16. Product Distribution as a Result of Reaction Set S_2 with Different Paraffins as Input[a]

input	no. of paraffin	no. of olefin	total no. of gas-phase species	no. of carbonium ions	no. of carbenium ions	total no. of ions
C_3	7	6	14	6	7	13
C_4	7	6	14	6	7	13
2m-C_3	7	6	14	6	7	13
C_5	7	6	14	6	7	13
2m-C_4	7	6	14	6	7	13
2,2m-C_3	7	5	13	6	7	13

[a] Isomers of different species are ignored. Total number of gas phase species includes the H_2 molecule.

Table 17. Various Paraffin Species Formed as a Result of Reaction Set S_2[a]

C_3	C_4	2m-C_3	C_5	2m-C_4	2,2m-C_3
C_1	C_1	C_1	C_1	C_1	C_1
C_2	C_2	C_2	C_2	C_2	C_2
C_3	C_3	C_3	C_3	C_3	C_3
C_4	C_4	2m-C_3	C_4	C_4	2m-C_3
C_5	C_5	2m-C_4	C_5	2m-C_4	2,2m-C_3
3m-C_5	C_6	2m-C_5	C_6	3m-C_5	2,3m-C_4
3e-C_5	3m-C_6	2,3m-C_5	C_7	3m-C_6	2,2,3m-C_4

[a] Isomer information is ignored.

Table 18. Olefins Formed as a Result of Reaction Set S_2[a]

C_3	C_4	2m-C_3	C_5	2m-C_4	2,2m-C_3
$C_2^=$	$C_2^=$	$C_2^=$	$C_2^=$	$C_2^=$	$C_2^=$
$1C_3^=$	$1C_3^=$	$1C_3^=$	$1C_3^=$	$1C_3^=$	$1C_3^=$
$1C_4^=$	$1C_4^=$	2m-$1C_3^=$	$1C_4^=$	$2C_4^=$	2m-$1C_3^=$
$2C_5^=$	2m-$1C_4^=$	3m-$1C_4^=$	$1C_5^=$	2m-$1C_4^=$	3,3m-$1C_4^=$
2e-$1C_4^=$	$3C_6^=$	1m-$2C_5^=$	$1C_6^=$	3m-$1C_5^=$	2,3,3m-$1C_4^=$
3e-$2C_5^=$	2e-$1C_5^=$	2(1m-C_2)$1C_4^=$	$1C_7^=$	4m-$1C_6^=$	–

[a] Isomer information is ignored.

Table 19. Carbonium Ions Formed as a Result of Reaction Set S_2[a]

C_3	C_4	2m-C_3	C_5	2m-C_4	2,2m-C_3
C_2^+	C_2^+	C_2^+	C_2^+	C_2^+	C_2^+
$1C_3^+$	$1C_3^+$	$2C_3^+$	$1C_3^+$	$1C_3^+$	$2C_3^+$
$2C_4^+$	$1C_4^+$	2m-$1C_3^+$	$1C_4^+$	$2C_4^+$	2m-$2C_3^+$
$2C_5^+$	$2C_5^+$	3m-$2C_4^+$	$1C_5^+$	2m-$1C_4^+$	2,2m-$1C_3^+$
3m-$3C_5^+$	$2C_6^+$	4m-$2C_5^+$	$2C_6^+$	3m-$2C_5^+$	2,3m-$2C_4^+$
3e-$3C_5^+$	3m-$3C_6^+$	2,3m-$3C_5^+$	$2C_7^+$	4m-$2C_6^=$	2,3,3m-$2C_4^+$

[a] Isomer information is ignored.

Table 20. Number of Reactions as a Result of Reaction Set S_2. Isomers of Different Species Are Ignored

	reaction type										total no. of
input	I	II	III	IV	V	VI	VII	VIII	IX	X	reactions
C_3	6	6	6	9	6	6	0	15	15	21	90
C_4	6	6	6	10	6	6	4	15	15	21	95
2m-C_3	6	6	6	10	6	6	0	15	15	21	91
C_5	6	6	6	8	6	6	4	15	15	21	93
2m-C_4	6	6	6	9	6	6	3	15	15	21	93
2,2m-C_3	6	6	6	8	5	5	0	13	15	21	85

ding isomers) demonstrates the power and utility of the RDL++ compiler for translating English language rules to elementary reactions.

The computation time for the various runs of reaction set S_2 both with and without taking into consideration the various isomers is given in Table 23. This table also shows the number of gas phase (paraffin + olefin + H_2) and surface species (carbonium and carbenium ions) reactions and the total number of reactions for the various input reactants. Clearly, the time taken to generate a reaction network increases with its size. Accounting for isomers takes longer, 38−42 s, as compared to under 1 s when the isomer information is excluded. All the computations were performed on the Intel Xeon dual processor machine with 1 GHz processors, 512K cache, and 2 GB RAM. The effectiveness of the compiler is evident as it can be used to generate multiple reaction networks by minimal change in highly intuitive rules rapidly.

2.1.6. Automatic Reaction-Network-to-Model Generator. The first phase of the RDL++ compiler generates the reaction network that is consistent with the chemistry rules as specified by the user. However, to test the validity of these reaction networks against experimental data, a mathematical model has to be generated. Typically in reaction engineering examples, this corresponds to formulating an ordinary differential equation model to explain transient data or an algebraic equation model to fit steady-state data. Law of mass action kinetics is used to translate a reaction network to such a mathematical model. This process could be very tedious and error prone when dealing with reaction networks with more than 10−20 elementary steps. The automatic reaction-network-to-model generator (Figure 4) automates this step by transforming the reaction network into a set of differential or algebraic equations depending on the data available. Each of the elementary steps generated by RDL++ is scanned for every species and the law-of-mass-action terms contributing to the consumption, and/or production of these species is constructed. As an example, consider the following reactions:

$$\mathrm{A} + \mathrm{B} \xrightarrow{k_1} \mathrm{C}$$

$$\mathrm{C} + 2\mathrm{B} \xrightarrow{k_2} \mathrm{D} \qquad \text{(R.14)}$$

Reaction terms that affect the concentration of species B (C_B) are $-k_1C_AC_B$ and $k_2C_CC_B^2$. Corresponding terms for the species A, B, and C are then used to construct the mathematical model:

$$dC_A/dt = -k_1C_AC_B$$

$$dC_B/dt = -k_1C_AC_B - k_2C_CC_B^2$$

$$dC_C/dt = -k_1C_AC_B - k_2C_CC_B^2$$

The reactions in the network have parameters that can be grouped according to the concept of "similar species undergo similar reactions under similar rates". Empirical grouping rules such as Polanyi relations to correlate activation energies to adsorption energies, reactivity relationships, variation of activation energies with carbon numbers etc. based on experiments, computations, and theory can be used to group the various

Table 21. Product Distribution as a Result of Reaction Set S_2 with Different Paraffins as Input Taking into Account the Various Isomers of Different Species[a]

input	no. of paraffin	no. of olefin	total no. of gas-phase species	no. of carbonium ions	no. of carbenium ions	total no. of ions
C_3	24	55	80	76	79	155
C_4	23	56	80	82	81	163
2m-C_3	23	53	77	83	78	161
C_5	22	51	74	77	76	153
2m-C_4	22	56	79	78	79	157
2,2m-C_3	22	51	74	76	73	149

[a] Total number of gas phase species includes the H_2 molecule.

Table 22. Number of Reactions as a Result of Reaction Set S_2 Taking into Account the Various Isomers of Different Species

input	reaction type I	II	III	IV	V	VI	VII	VIII	IX	X	total no. of reactions
C_3	89	76	71	114	96	96	54	96	81	114	887
C_4	80	80	75	119	98	98	56	96	82	116	900
2m-C_3	88	83	78	123	98	98	57	98	84	118	925
C_5	77	77	72	115	96	97	55	96	81	115	881
2m-C_4	78	78	73	117	96	96	54	96	81	114	883
2,2m-C_3	76	76	71	114	98	96	54	96	81	114	876

reactions so that the number of model parameters can be reduced.[45] These grouping rules also impose constraints to make sure that the model parameters are correlated such that they do not vary independently leading to nonphysical values. For example, the rate parameters of a reversible reaction cannot vary independent of the equilibrium constant of that reaction.

2.2. Genetic Algorithm Based Hybrid Pseudoglobal Parameter Estimator. The development of predictive models is a time-consuming, knowledge intensive, iterative process where an approximate model is proposed to explain experimental data, the model parameters that best fit the data are determined, and the model is subsequently refined to improve its predictive capabilities. Ascertaining the validity of the proposed model is based upon how thoroughly the parameter search has been conducted in the allowable range. The determination of the optimal model parameters is complicated by the complexity/nonlinearity of the model, potentially large number of equations and parameters, poor quality of the data, and lack of tight bounds for the parameter ranges. Thorough search of the parameters is necessary to obviate the wrong conclusions about the effectiveness of a proposed mechanism.

Recently, we critically evaluated a hybrid search procedure[12,46] that employs a genetic algorithm for identifying promising regions of the solution space followed by the use of an optimizer to search locally in the identified regions. We also reported that this algorithm is capable of finding global minima for test case problems[47] as determined by a deterministic global optimizer,[48] but with significant savings in time. The performance of this hybrid method in the presence of noise was found to be satisfactory. Also, this hybrid technique has been able to locate multiple solutions that are nearly as good with respect to the "sum of squares" error criterion but imply significantly different physical situations. In this section, we will compare this methodology with another stochastic technique—adaptive random search.[49] In section 3, we will propose a 13-parameter model that results in 60 differential algebraic equations for propane aromatization on a zeolite catalyst as a more challenging test case to validate this algorithm.

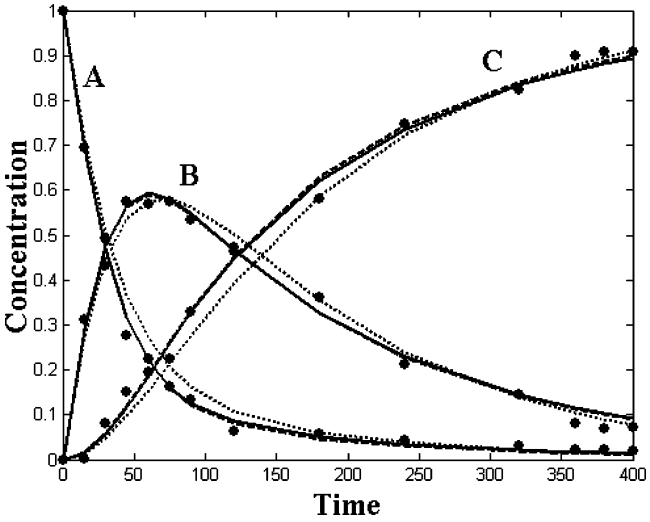


Figure 5. Concentration−time curves for species A, B, and C in the toluene hydrogenation model.[51] Solid lines correspond to the parameters reported by Belohlav and co-workers,[51] dashed lines represent the best solution obtained by the hybrid procedure, and the dotted lines show the predictions for the parameter set whose objective function value is at most five times that of the best solution of the hybrid procedure.

We will now compare the hybrid procedure with a popular parameter estimation method available in the literature—the direct search optimization technique based on use of randomly chosen sample points and adaptive reduction of the search space.[50] Belohlav and co-workers[51] have used this method for estimating the parameters of a model for toluene dehydrogenation based on the following reaction scheme

$$A \rightleftarrows B \rightarrow C$$

where A, B, and C represent toluene, methylcyclohexene, and methylcyclohexane, respectively. The model[51] consists of a set of three ordinary differential equations to describe the time evolution of the concentration of species A, B, and C and 14 data points for each of the species is used for estimating the five parameters (k_1−k_5) in the model. To minimize the correlation among the estimated parameters, the authors have used the determinant of the multiresponse data as the criterion for estimating the parameters as suggested by Box and Draper.[52] The first two rows of Table 24 show the best set of parameters as reported by Belohlav and co-workers[51] and those obtained by our hybrid search procedure, respectively. The corresponding predictions are shown by the solid and dashed curves, respectively, in Figure 5. It is interesting to note that although the predictions are nearly indistinguishable, the parameters are slightly different and the objective function value obtained by the hybrid procedure is marginally lower

Table 23. Comparison of Reaction Networks Generated by Reaction Set S_2 with and without Isomer Generation

input	S_2 without isomers: no. of surface species	no. of gas-phase species	total no. of reactions	time (s)	S_2 with isomers: no. of surface species	no. of gas-phase species	total no. of reactions	time (s)
C_3	13	14	90	0.86	155	80	887	39
C_4	13	14	95	0.91	163	80	900	40
2m-C_3	13	14	91	0.86	161	77	925	42
C_5	13	14	93	0.86	153	74	881	38
2m-C_4	13	14	93	0.92	157	79	883	38
2,2m-C_3	13	14	85	0.83	149	74	876	37

Table 24. The Performance of the Hybrid Procedure on the Toluene Hydrogenation Model [51] Where k_1 to k_5 Are the Parameters[a]

model	k_1	k_2	k_3	k_4	k_5	objective
Belohlav et al.	0.023	0.005	0.011	1.9	1.8	3.088×10^{-8}
hybrid, best	0.0234	0.0039	0.0106	1.6958	1.6953	2.883×10^{-8}
hybrid, worst	0.0226	0.0034	0.0071	1.2139	0.8438	1.424×10^{-7}

[a] The best solution from the hybrid method and the solution with at most five times the objective function value of this best solution are reported. The corresponding concentration−time curves are as in Figure 5.

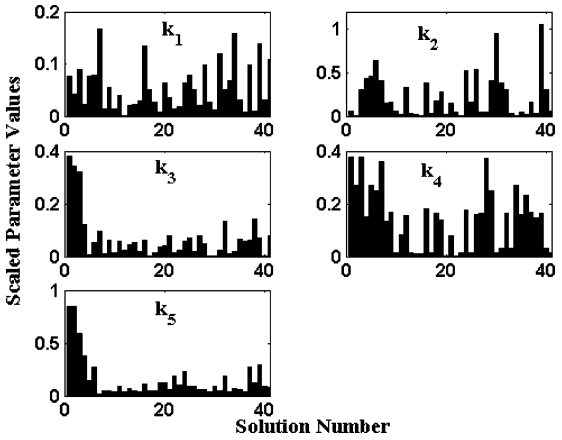


Figure 6. Multiple solutions for the toluene hydrogenation problem[51] whose objective function value is up to five times that of the best solution but whose parameter values are widely different. The scaled parameter values have been calculated as the absolute difference between the actual value and the average value and then scaled by the average value.

than that reported by Belohlav et al.[51] This may be because of the errors in the integration routines used.

A crucial point commonly associated with most nonlinear parameter estimation problems is the multiplicity of the solutions. This becomes especially important when there is error associated with the data available for estimating the parameters. To further investigate this issue, we examined all the solutions of the hybrid procedure with at most five times the objective function value (SSE) corresponding to that of the best solution. The dotted lines in Figure 5 show the predictions of the worst solution of this set. The parameter values corresponding to this worst solution are shown in Table 24. With the assumption that a typical kinetic experiment has 15−20% error in data, the predictions from this worst solution cannot be distinguished from the predictions corresponding to the best solution as shown in the dashed lines in Figure 5. Figure 6 shows the relative variation of the 41 solutions whose objective function value is at most five times that of the best solution. The relative variation in a parameter is determined as the absolute value of the difference between the parameter and the average value of the parameter scaled by the average value; specifically, $k_i^{\text{scale}} = |k_i - k_i^{\text{avg}}|/k_i^{\text{avg}}$. It is interesting to note that parameters k_2 and k_5 can vary up to almost 100% of their average value. This means that these parameters could be twice as much as their average values among all the solutions. Multiple solutions and large parameter variation among them could be an indication that we have insufficient data to effectively estimate the parameters or that the proposed mechanism does not explain the data completely. These solutions could be of potential interest in planning further experiments for discriminating among competitive models for this problem.

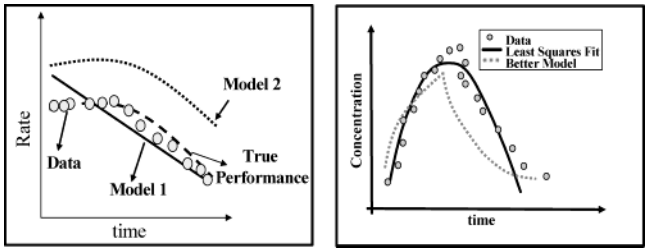


Figure 7. Need for feature extraction: (a) rate vs time; (b) concentration vs time.

As discussed in the previous paragraph, the hybrid search procedure, using GA for identifying promising initial guess values followed by the application of a traditional local optimizer, is able to find the global optimum. It is important to note that local optimizers alone can be very successful for relatively small refined problems with well-defined and small parameter bounds; however, for initial screening of large amounts of data, the above procedure would be natural choice. Also, this method is useful from a design perspective when the expert is interested in multiple solutions. In section 3, we will examine how this hybrid method works for much larger problems that are of particular interest in determining optimal reaction networks for real systems.

2.3. Feature Extractor. The main aim of developing an automated, user-driven tool kit such as Reaction Modeling Suite is to aid an expert in building large-scale kinetic models. One of the key postulates in this effort is that any system to aid an expert should follow the thought process of the expert. In our opinion, the human expert does not primarily think in terms of the detailed mathematical formulation of a model; rather he or she thinks in terms of the "rules" that lead to that model and the features that result from the model. Accordingly the RDL++ compiler acts as an information gathering tool from the user through which the expert can key in the rules and it also translates the input rules into a mathematical model automatically. The model parameters are then robustly estimated using the GA-based hybrid parameter estimation technique as explained in section 2.2. The expert is now interested in analyzing the predictions that resulted from the model that was based on the rules.

The analysis of predictions vs data, at least during the early stages of model development, is not primarily via a least-squares fit but rather through a comparison of the "features" of the data vs those of the model. As shown in Figure 7a, the expert would vote for model 2 that captures the features of the true performance (dotted line) even though model 1 has better quantitative fit to the data. Clearly, the expert does not think in terms of the squared errors at individual data points or in terms of the sum of the squared errors or in other statistical lack-of-fit measures that quantitatively address the difference between the model predictions and the data. For example, in the simple catalytic reaction A goes to B, the "rule" that A must be adsorbed reversibly leads to the "feature" that the rate of production of B will show a maximum with increasing temperature. Similarly, the features could be the initial slope of the rate curve, the kink at the top of the curve (Figure 7b), or the time at which the selectivity curve saturates—essentially the key landmarks that the expert is interested in explaining through the model.

Information about the mismatch between the features of the data and the model predictions is used to

postulate new rules or to modify the existing rules to improve the model—an iterative process that we call model refinement. Evidently, the ability to automatically extract the features facilitates model refinement, and in this section, we explain the process of automatically and robustly extracting the features from a curve which could be either experimental or based on the model predictions.

To avoid the scenarios shown in Figure 7 and to aid the process of model refinement, we propose a feature-based model evaluation to screen and compare models. These models could be different mathematical realizations of a process that are derived based on different assumptions or could be because of the multiple parameter values for the same underlying model.

The objective is to come up with a criterion for estimating the goodness-of-fit of multiple models in their ability to explain data using an objective function based upon the critical features of the model predictions and the data. For example, consider a model with two parameters. If n_f is the number of features identified to be critical in the data, $M_{f_i}(x_1,x_2)$ corresponds to the ith feature as predicted by the model, and $D_{f_i}(x_1,x_2)$ corresponds to the corresponding feature in the data, then the following objective function would suit our purposes

$$\min_{x_1x_2} \sum_{i=1}^{n_f} w_i||M_{f_i}(x_1,x_2) - D_{f_i}(x_1,x_2)|| \qquad (1)$$

where w_i is the weighing factor for the feature i that depends on the importance of the feature and $\sum_{i=1}^{n_f} w_i = 1$. For the purpose of illustration, let us choose two slopes s_1 and s_2 as the critical features and let m_1 and m_2 be the corresponding model predicted slopes. Then the above objective function would reduce to

$$\min_{x_1,x_2} \sum_{i=1}^{2} w_i||(m_i - s_i) \qquad (2)$$

An automatic feature extraction procedure would facilitate the evaluation of the above criterion simple. This becomes more important especially when the data are available at a higher rate and accuracy and automated model discriminations strategies are required to build robust kinetic models.

In the chemical engineering literature, feature extraction techniques have been used for the purposes of trend analysis of process data in order to exploit the temporal information and to reason about the process state. The main activities involved are (i) identification of qualitative trends and (ii) mapping from trends to operational conditions. To deal effectively with a multitude of process data and extract the underlying important trends and events in the process, Janusz and Venkatasubramanian[53] proposed a framework for the automatic generation of such qualitative process trend descriptions directly from sensor data. A trend is represented as a sequence of seven primitives that are piecewise unimodal or quadratic segments. These primitives form the alphabets of their trend description language. This qualitative filter provided a meaningful compaction of large amounts of numerical data without losing the essential information about the trends.

Other examples of qualitative process trend analysis include use of an expandable "composite" shape library to approximate a noisy process signal by a polynomial,[54] a knowledge-based interpretation of sensor patterns,[55] a technique for data compression and trending called piecewise linear online trending that adapts to process variability and noisy data,[56] pattern matching between the observed fault trends and the ones in a knowledge base,[57] application of trend based temporal techniques for medical diagnosis,[58] a B-Spline based technique for data compression and automatic trend extraction,[59] combination of the primitives based trend description language[53] with a fuzzy-logic-based multivariate inference framework for temporal-reasoning,[60] and automated identification of process trends based on an interval-halving procedure.[61] A more recent review of the qualitative methods used in the process trend analysis and fault diagnosis is available elsewhere.[62,63]

There are significant differences between the sensor network process data and kinetic data that are the focus of this paper. Unlike process data from plants, kinetic data are typically available in small amounts. Although the volume of data from high throughput experiments has been increasingly available at a higher rate and accuracy in the recent years, this is still not a match to the historical process data available from chemical plants. Kinetic data generally do not show abnormal deviations or unexplainable trends. If there are irregularities in curves from good experimental setups, they will mostly be systematic and repeatable. The data from experiments is noisy, but the noise is far less as compared to that in process data.

Qualitative process trend extraction algorithms do not use a priori information about the processes as this is typically not available in sensor network data; however, in the case of kinetic data, one typically knows the curve signatures for faulty experiments such as malfunction in reactor setup, catalyst deactivation, temperature spike, etc., and this information can be used to reject certain features in the curves that are not interesting. Also, the expert typically has some information about which features are important and which are not. This information can be used so that the feature extraction algorithm does not have to look for the unimportant features. Considering the above differences between the kinetic data and the data from process plants, we cannot use the automatic trend extraction algorithms such as the interval-halving procedures[61] that have been developed mainly to deal with large volumes of noisy data without user intervention and with minimal a priori information.

Any feature extraction algorithm devised for characterizing kinetic data should be able to overcome a different set of challenges. First, it is highly likely that in the context of kinetic data, some features that occur only for very short periods of time, and hence treated as noise by automatic noise rejection algorithms, could actually be the most important features. So, a progressive, detailed-to-coarse feature identification with updates from the user is required. Second, not all the features are equally important for a kineticist trying to model data. For example, the initial lag in a rate curve may be more important than the relatively large deviations at saturation at later times. Also, features such as the slope of the curve, offset, etc., at lower space times in a plug flow reactor may be more important as they nonlinearly affect the features at the end of the reactor. Similarly the concentrations of species with fewer carbon numbers may be more important than that of the long-chain hydrocarbons in a polymerization reactor,

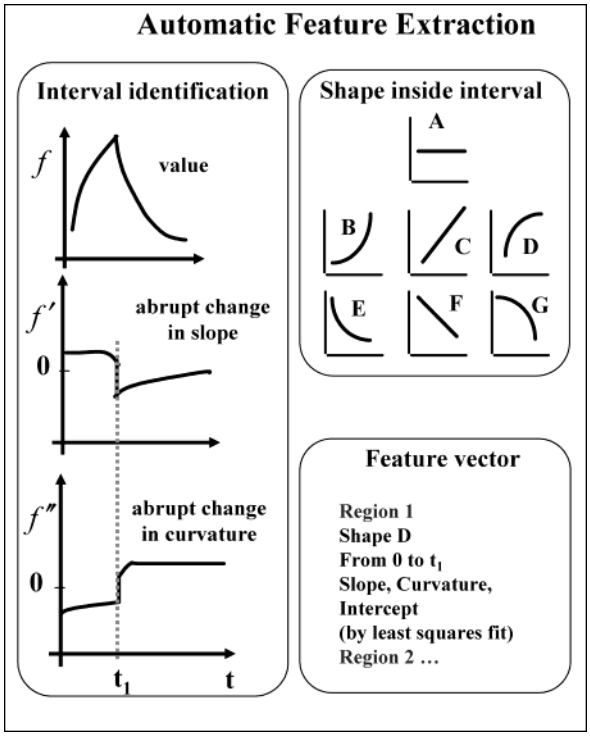


Figure 8. Methodology for automated feature extraction.

as the smaller species act as the seeds for the longer ones. So, a mechanistic rank ordering of features that are important for the problem at hand is required so as to facilitate the understanding of which features would need to be fixed first and which are relatively unimportant in the process of MR. Third, in the case of kinetic data we are interested in not only the qualitative trends in terms of the slopes of the curves but also the absolute values of at least some of the features. Finally, data may be sparse in certain time ranges. Unlike process data that has almost equal density over large periods of time, kinetic data may not be available over the entire design parameter space. So any feature extraction algorithm for kinetic data should not rely on the density of data to extract meaningful features.

We propose the following feature extraction algorithm that systematically interprets the kinetic data curves by identifying key features of the curve, e.g., increasing rate, decreasing rate, change of slope, inflection point, etc. realized through the generation of a "feature vector".

(1) With the help of a human expert, draw a smooth curve passing through the experimental data. This ensures that the features generated are not affected by the noise and irregularities in the data.

(2) As shown in Figure 8, identify the critical points where there are abrupt changes in the value, first or second derivatives of the curves. Report any unimodal or quadratic primitives that match this section of the curve. Populate the feature vector with the primitive, slope, curvature, intercept at end points, and the range of the independent variable (typically time) that corresponds to this section of the curve.

(3) Interacting with the user, rank order features both in terms of the important species and time regimes according to their importance. If the user flags certain features to be unimportant according to the user, redo step 2 by merging time ranges of the identified features with the adjacent ones. If the user specifies a particular time range to be of greater importance, calculate the feature vector in that range.

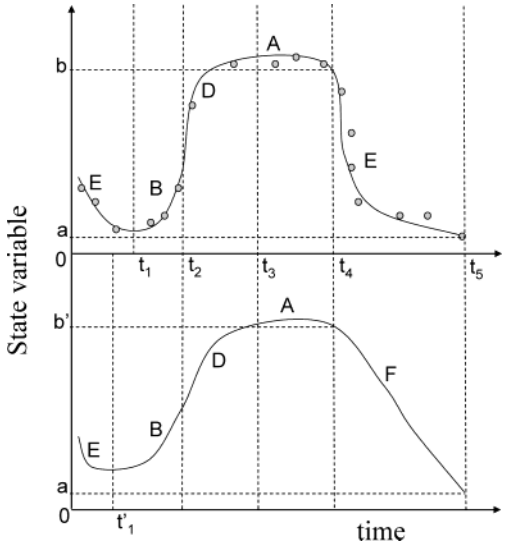


Figure 9. Example to illustrate the feature extraction algorithm to compute the similarity between two curves. Data and the expert postulated curve through data are shown in the top figure (a) and the hypothetical model curve is in the bottom figure (b).

Table 25. Critical Features of the Data and the Model Curves Shown in Figure 9

	data		model	
time interval	primitive	feature	primitive	feature
t_1-0	E		E	
range		t_1		t_1'
t_2-t_1	B		B	
slope at t_2		s		s
t_3-t_2	D		D	
–		–		–
t_4-t_3	A		A	
ordinate at t_4		b		b′
t_5-t_4	E		F	
ordinate at t_5		a		a

(4) Repeat step 2 for the curves generated by the model simulations and calculate the sum of the squared deviations between the model and the data features. Populate the feature vector of the model curve with this metric. Optimize on the model parameters using this objective function criterion with any parameter estimation routine such as the GA-based hybrid algorithm explained in section 2.2.

The example shown in Figure 9 demonstrates the feature extraction technique to compute the similarity between two curves. The expert identifies the significant features in the data (Figure 9a) by analyzing the abrupt changes in the value, slope, and the curvature. The data curve is partitioned into five intervals within which the curve is characterized by one feature. For example, in the first time interval, the expert is interested in the lag period and in the second interval between t_1 and t_2, the slope of the curve (at t_2) is considered to be important. The points at which the curve saturates at t_4 and at t_5 are considered to be the critical features in the last two intervals. The primitives that closely match the curves in each of the time intervals is extracted by computing the slope and curvature and is as shown in Table 25.

Now consider a model that results in the curve as shown in Figure 9b. This curve is analyzed for the critical points, and the various time intervals are calculated. The primitives and the important features of this curve are also shown in Table 25. It is interesting to note that the first time point at which significant

change in slope occurs is different in the data (t_1) and the model (t_1') curves. Similarly, the saturation point at time t_4 is different between the two curves, b and b'. Also, the basic shape as characterized by the primitive between t_4 and t_5 in the data curve is E whereas that in the model curve is F. The objective function in eq 1 is used to account for the differences in the two curves as

$$\min w_1(t_1 - t_1')^2 + w_2(b - b')^2 + w_3(E - F)^2 \quad (3)$$

and $w_1 + w_2 + w_3 = 1$. To quantify the differences in the qualitative aspects of the primitives E and F, the fuzzy similarity matching indices[60] to quantify the differences between the various primitives. For example, primitives A and C are not completely different and so they are assigned a similarity index of 0.25. This means that primitive A is 25% similar to primitive C. Similarly, primitives E and F are closer to each other by 75%.

The above algorithm for feature extraction is simple and extensively uses domain knowledge about the system available from the user regarding the features and their relative importance. Unlike interval-halving-based fully automated algorithms,[60] it does not rely on the density of the data. This algorithm also allows for iterative correction of features and so if it misses an important feature, it can go back and locate it with the help of the user. The user-defined features on the experimental data are used to guide the automatic extraction of the features from the model curves using the critical points. The feature-based objective function (eq 2) can thus be computed for estimating the parameters and screening through multiple models.

2.4. Statistical Analyzer. Statistical analysis of the models is necessary in order to screen, compare, and improve them. The most common statistical method used for analyzing the quality of the model is based on the sensitivity of the model output with respect to the model parameters defined as a partial derivative introduced via a Taylor series expansion

$$c_i(t,\underline{k} + \Delta\underline{k}) = c_i(t,\underline{k}) + \sum_{j=1}^{m} \frac{\partial c_i}{\partial k_j} + ... \quad (4)$$

where the partial derivatives $\partial c_i/\partial k_j$ known as the first-order local concentration sensitivity coefficients are evaluated by the parameters kj, one at a time at time t and the effect on concentrations measured at time $t + \Delta t$. Kinetic models generally involve ordinary differential equations such as

$$\mathrm{d}\underline{c}/\mathrm{d}t = \underline{f}(\underline{c},\underline{k}),\underline{c}(0) = \underline{c}^0 \quad (5)$$

where $\underline{c}$ is a dimensional concentration vector. The above ODEs can be differentiated with respect to the model parameters kj to give

$$\frac{\mathrm{d}\underline{c}}{\mathrm{d}t}\frac{\partial \underline{c}}{\partial k_j} = \underline{J}(t)\frac{\partial \underline{c}}{\partial k_j} + \frac{\partial \underline{f}(t)}{\partial k_j} \quad (6)$$

$$j = 1, ..., m$$

where $\underline{J}(t) = \partial \underline{f}/\partial \underline{c}$ and the initial condition for $\partial \underline{c}/\partial k_j$ is a zero vector. Equations 5 and 4.6 are coupled through the matrices $\partial \underline{f}/\partial \underline{c}$ and $\partial \underline{f}/\partial \underline{k}$; hence eq 6 can only be solved if the concentration values calculated in eq 6 are available at times where these matrices are calculated during the numerical solution of eq 6. This is achieved by solving the $(m + 1)n$ equations in eqs 5 and 6 simultaneously. A more efficient algorithm for the solution of the sensitivity differential equations is the decoupled direct method[64] which uses the fact that eqs 5 and 6 have the same Jacobian. The result of the above solution procedure is a set of sensitivity coefficients $\partial c_i/\partial k_j$. Since the parameters and the various output quantities of the model may have different units especially when the reactions are of different reaction orders, normalized sensitivity matrix defined as the fractional change in concentration c_i caused by a fractional change of parameter k_j

$$\underline{S} = \frac{k_j}{c_i}\frac{\partial c_i}{\partial k_j} = \frac{\partial \ln c_i}{\partial \ln k_j} \quad (7)$$

is typically used for further analysis. The variance on the parameter estimates can be computed using

$$\mathrm{Var}(\underline{k}) = J^{\mathrm{T}}J^{-1}\,\mathrm{s}^2 \quad (8)$$

where s^2 is an unbiased estimate of the model prediction error expressed as the difference between the model predictions (c_i) and the experimental data ($\hat{c}_i$) as

$$s^2 = \sum_{i=1}^{n} (c_i - \hat{c}_i)^2 n - p \quad (9)$$

for a model with p parameters and in the case where there are n experimental data points.

The local sensitivity coefficients and the other metrics defined above have been used for identifying redundant species and redundant reactions[35,42] and hence to reduce a kinetic model. Other techniques such as concentration sensitivity analysis,[32] reaction rate analysis,[36] principal component analysis,[65] and lumping analysis[66] have been used for the investigation and reduction of reaction mechanisms especially for the description of combustion reactions. For a more comprehensive review of the statistical methods for the analysis of reaction mechanisms, refer to Tomlin et al.[32] Computer software packages such as SENKIN[67] and KINALC[38] implement one or more of these methods. An alternative to these local methods would be the global sensitivity analysis procedure—study of the effect of the parameters on the model output without the assumption of any individual solution. For example, methods[68] such as the Fourier amplitude sensitivity test simultaneously perturb all rate parameters by sine functions with different frequencies and analyzes its effect on the concentrations. These methods are computationally expensive especially for models with a large number of parameters.

To address the above concerns, the Statistical Analyzer in RMS uses techniques from both local and global sensitivity analysis for ascertaining the robustness of a kinetic model. A kinetic model is defined to be robust if it is accurate in explaining the data even when the model parameters have not been estimated with sufficient accuracy. Typically kinetic model parameters such as rate and equilibrium constants cannot be estimated accurately because of the errors in the model and the data and the errors in the estimation of the parameters. It is useful to see how these errors are propagated through to the model predictions. We postulate that the various errors are localized as errors in

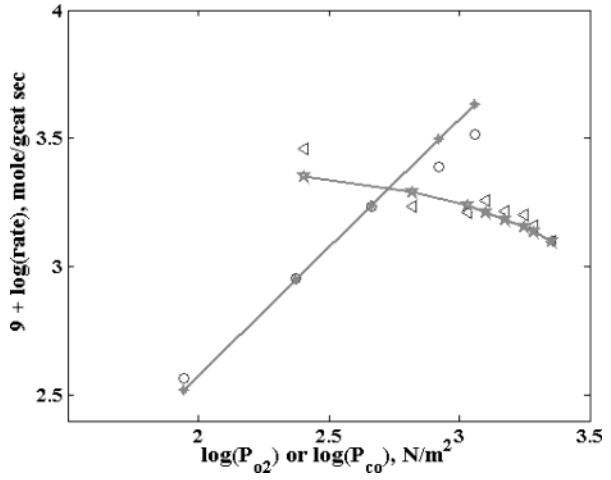


Figure 10. Predictions of CO oxidation model where carbon-monoxide adsorption is quasi-equilibrated and adsorption of oxygen is irreversible. The parameters are as in Table 26.

Table 26. Values of the Rate Constants, Standard Deviations, 80% Confidence Interval, and the Correlation Matrix for the CO Oxidation Model Where Carbon Monoxide Adsorption Is Quasi-Equilibrated and Adsorption of Oxygen Is Irreversible

			80% confidence interval		correlation matrix		
parameter	mean	std dev					
log K_1	−3.75	0.24	−4.09	−3.42	1.000	0.967	−0.753
log k_2	−3.87	0.11	−4.03	−3.72		1.000	−0.849
log k_3	3.57	168.53	−222.18	229.30			1.000

the estimated parameters, and we use these errors to compute the error in the model predictions.

Local sensitivity analysis deals with sensitivity coefficients computed by local perturbations around the best minimum and is expressed as the standard deviations, confidence bounds, and correlation matrix[69,70] of the parameters. To illustrate the concepts of the local sensitivity analysis, we consider a model for CO oxidation on a supported metal catalyst as described in Appendix A. Specifically, we show as to how the error in the model, in the data, and in the parameter estimation procedure can be propagated to the model predictions. This model consists of three steps: molecular adsorption of CO in a quasi-equilibrated manner where the forward step of adsorption and the reverse step of desorption are of almost the same rate; dissociative irreversible adsorption of oxygen; the surface reaction between the two adsorbed species to give CO_2. The three parameters that describe this process are the equilibrium constant of the CO adsorption process (K_1), the rate constant for oxygen adsorption (k_2), and the rate constant for the surface reaction (k_3). Table 26 shows the values of the estimated parameters, and the corresponding predictions of the rate of CO_2 production with respect to the variation in the partial pressures of CO and O_2 are as in Figure 10. The large standard deviations and the confidence intervals in parameter k_3 show that for a reasonably good prediction of the data, the value of k_3 has not been estimated accurately.

Figure 11 pictorially represents the inference regions[70] of the parameters for 90 and 95% confidence limits. The large variation in the ordinate of the first plot again shows that the parameter k_3 has not been estimated accurately. Also, the plots show how the parameters are correlated to each other. Parameters k_2 and k_3 are correlated negatively—when k_2 increases k_3 decreases—and parameters K_1 and k_2 are positively

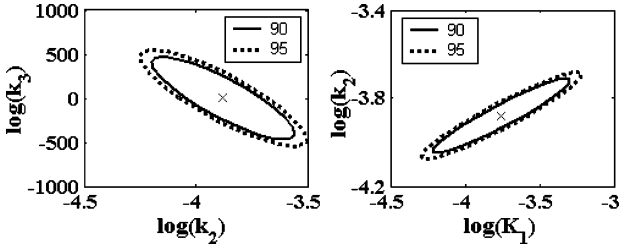


Figure 11. Inference regions based on the confidence intervals of the parameters for the CO oxidation model where carbon monoxide adsorption is quasi-equilibrated and adsorption of oxygen is irreversible.

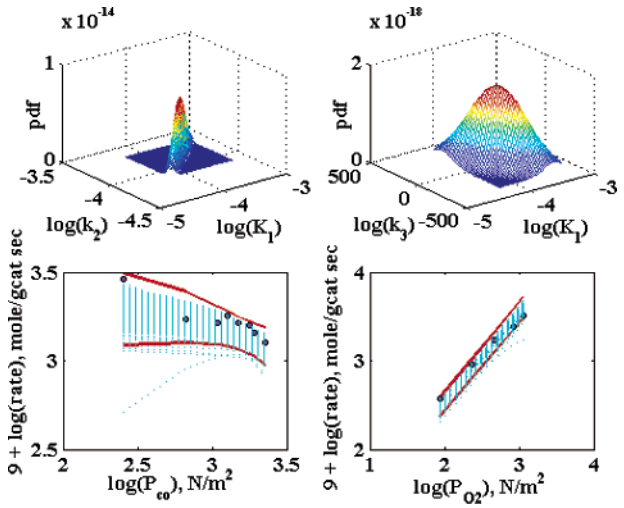


Figure 12. Parameter correlations assuming a Gaussian distribution and the $\mu \pm 3\sigma$ error bars on the model predictions for the CO oxidation model where carbon monoxide adsorption is quasi-equilibrated and adsorption of oxygen is irreversible. The dots in the bottom plots show the data from Cant, Hicks, and Lennon.[71]

correlated. This information is also available from the correlation matrix shown in Table 26. We claim that errors in the model, in the data, and in the estimation procedure have been cast as the errors in the model parameters, and we now propagate this error to the model predictions in the following manner. Assuming that the parameters follow a Gaussian distribution with mean as the best estimates and standard deviation given by the errors in the estimates (Figure 12), we randomly sample 1000 parameter sets from this distribution and simulate the model with these parameter sets. The resulting predictions are shown in the bottom two plots in Figure 12. The thick lines show the 3σ deviation of the model predictions and the dots show the experimental data.[71] The error bars on the model predictions show clearly that even though the parameter k_3 has not been accurately estimated, the model is robust—the predictions are accurate. Also, we can see that the accuracy of parameter k_3 affects the predictions of the variation of CO_2 rate with the partial pressure of CO more than that of the predictions of the variation of CO_2 rate with the partial pressure of O_2. In the case of a large reaction network, inferences of this kind can be used to ascertain as to which part of the reaction network is sensitive to which parameters.

Information available by locally perturbing the parameters around the best solution may not be sufficient to analyze the model especially when a large number of parameter sets explain the data equally well. Typically the ranges in which parameters of large reaction networks lie are not known accurately. Also the models

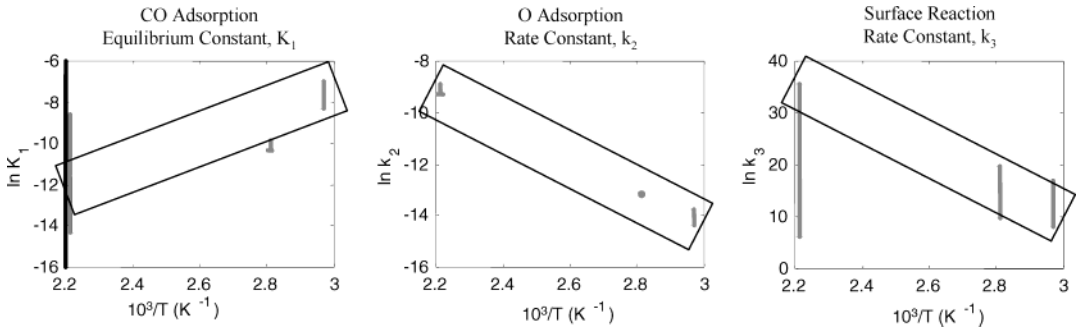


Figure 13. Multiple values of rate constants for the CO oxidation model where carbon monoxide adsorption is quasi-equilibrated and adsorption of oxygen is irreversible.

that describe these networks may not have all the required components either stoichiometrically or in terms of constraints among the model parameters at least in the beginning of the model-building procedure. In such situations, the data can be predicted by a large number of parameter sets equally well. It is important to estimate these multiple minima as they may correspond to physically different situations of the underlying system. This information is especially useful when designing a catalyst. For example a set of parameters could correspond to the high coverage of CO on the catalyst surface and yet another set of parameters could correspond to the high coverage of O_2. It is likely that for the range of partial pressures of CO (higher than that of O_2) used to collect the data, high coverage of O_2 is physically infeasible. Hence, it is important to identify and analyze the multiple minima.

To address this concern, the Statistical Analyzer in Reaction Modeling Suite uses techniques to identify multiple minima and analyzes them. The GA-based hybrid parameter estimation technique discussed earlier in this section was one such technique. The effectiveness of this method in identifying multiple minima was shown in section 2.2 and Figure 6. Another method known as the trough-walking algorithm that tracks the minima in the local neighborhood of any given minima is described in Appendix B. This method starts with multiple initial guess values, and once a local optimum is found, the local neighborhood is analyzed to find any other close minima. This ensures that we find most of the local minima around any given minima. The number of local minima that we are able to find depends on the value of the parameter that controls the size of the local neighborhood searched at any step (EDBOUND = 0.05) and also on the ruggedness of the fitness landscape.

The vertical bars in Figure 13 show the parameter value ranges of all the minima found at the various temperatures for the CO oxidation model described in Appendix A. Each of these minima predicts the data within a sum of squared errors between the model predictions and the data of 0.1. It is interesting to note that the O_2 adsorption rate constant (k_2) has been found with great accuracy unlike the other parameters (K_1 and k_3) that show large variations. This information is also available from the standard deviations of the parameters from the local sensitivity analysis (Table 26). More importantly, the information about the multiple minima helps us in validating the model predictions. For example in Figure 13, the equilibrium constant K_1 for the exothermic reaction of CO adsorption has a positive slope and the rate constants have negative slopes. This is in accordance to the physical realities of this system.

2.5. Discussion—Reaction Modeling Suite. In this section, we describe a user-driven, automated set of tools—Reaction Modeling Suite (RMS) that aids the expert in constructing robust kinetic models. Specifically, RMS is designed to allow the expert to initiate the kinetic modeling sequence in a simple reaction chemistry language, converts the reaction network into a mathematical model, optimizes the model parameters using a hybrid algorithm, extracts the features of the data and model prediction curves, and statistically analyzes the robustness of the model.

The RDL++ compiler that translates the English-language rules to a reaction network is generic, extendable, and intuitive, thereby affording an easy-to-use interface for a practitioner. The English-language rules input is more user friendly as compared to that of the bond order—bond electron matrices[24] and the structure oriented lumping vectors[25] as the rules are in the natural language used by a chemist to describe the reactions. The human expert can readily create multiple hypotheses and change the size of the reaction networks from a few species and reactions to several hundreds of species and reactions by manipulating a few steps in the reaction rules. Any new rule can be easily added, and the existing rules can be changed with little effort. The capability of RDL++ to track down all the isomers and generate all reaction steps that involve all the isomers of any species is very useful for describing reaction networks whose characteristics change with the three-dimensional structure of the species involved. Also, during the initial stages of modeling a network, the expert can simply turn off the isomer generation global rule and, with the limited amount of analytical data, try to explain the reaction. The expert can also manipulate the size of the reaction network by changing the number of carbon atoms present in a reactant or a product in any of the reaction rules. The use of global rules to prevent the formation of chemically infeasible species—allylics, species with a positive charge and a double bond, trifins, species with triple bonds, species with more than two double bonds, species with a positive carbon attached to a double bond, etc.—enables the expert to keep the reaction network feasible.

RDL++ has been designed and implemented along the lines of Reaction Description Language,[31] but RDL++ forms a part of RMS which handles all the operations of building a kinetic model starting from the formulation of chemistry rules to the analysis of the performance of a kinetic model. With this in mind, RDL++ has been designed to be more extendable, user-driven, and efficient than RDL. Specifically, new rules such as desorption, cyclization, and hydride transfer have been

developed based on the language for solid acid chemistry and reactions on catalytic surfaces. New keywords and syntax for carbonium ions, trifin (species with triple bonds), allylic (species with a double bond and a positive charge), and monoene, diene, and triene (species with three double bonds), have been included in RDL++ so as to enrich the palette for the user. New model pruning concepts to reduce the size of the reaction network have been introduced. The concept of global rules prevents duplication of pruning steps in individual reaction rules. Also the user now has the powerful ability to forbid the formation of isomers. The size of the resulting model is controlled by the size of the carbon chain in the hydrocarbon reactants or products rather than the less-intuitive "generation count"[72] that is based on the depth of the reaction network. We have also shown as to how a tool such as RDL++ can be integrated with other tools for automated hypothesis generation and testing in order to build robust kinetic models. Finally, RDL++ has been developed in a C++ environment which is more structured and user friendly compared to that of LISP.

Possible improvements to RDL++ include an XML (http://www.w3c.org) based interface to interactively define new keywords and to extend existing keywords. An intelligent backtracking mechanism that enables causal reasoning of the individual terms in the mathematical model to the elementary reaction steps of the network and/or directly to the section of the rules would make the whole compiling process more transparent and could have potential implications in model refinement. With this added feature to perform qualitative sensitivity analysis, the user will be able to selectively modify a set of rules to manipulate the terms in the model which would result in different features in the performance curves.

The Feature Extractor module in RMS is used to aid the expert in extracting the features in the data curves and then use this information to develop an objective function to compare the features in the model and the data. An expert-system-like framework that can be continually updated with user-supplied information about the different features and their relative importance can make this process more efficient. Also, the primitives that have been currently used in RMS are based on the description of the first- and second-order derivatives of the curves. New definitions of primitives that use the domain knowledge would be more attractive. For example, simple kinetic reaction mechanisms give rise to standard rate laws[73] which in turn give rise to specific features in the performance curves. For example, a second-order surface reaction gives rise to a square in the denominator of the Langmuir−Hinshelwood rate expression and a saturation curve. Similar trends can be encoded as primitives and more complex rate expressions can be derived by superposition of these primitives. A list of common features and the corresponding rate laws can be used to enhance the set of currently available primitives. This would help in transparent and intuitive model refinement.

The Statistical Analyzer affords inference and analysis of performance curves obtained from the user-postulated hypothesis and acts as a feedback mechanism for the user to refine the model. We have used information from local sensitivity analysis to propagate the error in the data, model, and the parameter estimation procedures to model predictions and thereby evaluate the robustness of the models. We also analyzed the multiple solutions of the parameters that explain the data equally well in order to understand the variations in parameters physically.

The current work on RMS aims at the design and development of new tools and modification and integration of existing qualitative and quantitative concepts and tools to aid an expert in all the steps involved in building robust kinetic models. This framework affords a usable and practical methodology to systematize reaction modeling for materials design. Generally RMS can be used to study any kinetic system typically modeled as a set of elementary reactions[11] leading to models based on ordinary differential or algebraic equations. The current implementation of RDL++ is geared toward carbenium/carbonium-based chemistry, and the overall design is such that it can be extended to other systems, for example, reactions on transition metal catalysts, metabolic reaction networks, etc. All other tools in the RMS require little or no modification to apply them to other systems. We have used an ideal plug flow reactor model to explain the details of RMS. However, there may be situations where the basic assumptions of the flow through the reactor may warrant more complicated multimode reactor models.[74] Also, RMS provides a good tool box to rapidly screen through multiple kinetic mechanisms especially in the light of high throughput kinetic data; however, for detailed analysis of catalytic processes, one would require much richer understanding of the concepts such as aging, poisoning, coverage-dependent surface energetics, etc.

Similar work in developing an integrated framework includes the efforts by Hostrup and Balakrishna[75] who use reaction modeling for process design and by Stoltze[76] (http://www.aue.auc.dk/~stoltze/mkm/main.html) for a reactor design based on the Langmuir−Hinshelwood reaction mechanism. On the basis of the descriptors such as (1) the ability of the tool to formulate a reaction network from higher level rules, (2) visualize the network, (3) parse the network into mathematical model, (4) solve the model and optimize the parameters, and (5) perform statistical analysis of the results to evaluate software systems that facilitate kinetic model building as discussed at the beginning of this section (Table 1), RMS would be an effective option.

3. Case Study—Propane Aromatization on Zeolites

The effectiveness of the various components of the RMS has been demonstrated on simpler problems; however, to truly test its hypothesis screening abilities, we will now apply it to develop a kinetic model of a complex and industrially relevant reaction—propane aromatization on HZSM-5 zeolite catalyst. A number of kinetic models have been proposed for aromatization of alkanes over ZSM-5;[45,77,78] however, a model with predictive capabilities remains a challenge. Our kinetic model is based on a reaction scheme involving adsorption, desorption, protolysis, dehydrogenation, hydride transfer, β-scission, oligomerization, and aromatization reactions. The rules are encoded in a similar manner as described in section 2.1.

The proposed set of reaction "rules" generates a very large number of individual reactions. Statistics-based model reduction techniques[40,42] do lead to the reduction in the number of parameters; however, they may result in ad hoc elimination of reactions leading to chemically infeasible reaction networks. To reduce the number of

Table 27. Model Characteristics for the Propane Aromatization on a Zeolite[a]

model components	reaction families
231 reactions	protolysis of carbonium ions
31 gas-phase species	carbonium ion desorption
28 surface species	carbenium ion desorption
1 vacant site	β-scission
31 ODEs	aromatization
29 algebraic equations	alkane adsorption
28 surface species balances	hydride transfer
1 site balance	olefin adsorption
species up to C9 have been considered	oligomerization
	carbonium ion dehydrogenation

model parameters	bounds
protolysis of carbonium ions	$10^2 \le k_p \le 10^7$
carbenium ion desorption	$10^4 \le k_{od} \le 10^{10}$
increase in adsorption enthalpy for alkenes with carbon number	$6 \le \Delta q_{od} \le 14$
carbonium ion desorption	$10^3 \le k_{ad} \le 10^9$
increase in activation energy for alkanes desorption with carbon number	$6 \le \Delta q_{od} \le 12$
β-scission	$10^3 \le k_b \le 10^8$
aromatization	$10^7 \le k_a \le 10^{13}$
alkane adsorption	$10^{-3} \le k_{aa} \le 10^2$
hydride transfer reactions	$10^{-3} \le k_h \le 10^2$
olefin adsorption	$10^{-1} \le k_{oa} \le 10^4$
carbonium ion dehydrogenation	$10^2 \le k_{cd} \le 10^8$
increase in the activation energy for carbonium ion dehydrogenation with carbon number	$2 \le \Delta q_{od} \le 6$
entropy factor for determining the equilibrium between β-scission and oligomerization	$18 \le S \le 25$

[a] First-order rate constants (k_p, k_{od}, k_{ad}, k_b, k_a, k_{cd}) are in terms of mol/(g/h); second-order rate constants (k_{aa}, k_h, k_{oa}) in m^3/(g/h); energy terms (Δq_{od}, Δq_{ad}, Δq_{cd}) are in kJ/mol; and the entropy term, ΔS has been normalized by the universal gas constant (J/(mol/K)).

parameters involved, we invoke the principle of similar species undergo similar reactions at similar rates. Thus we categorize the reactions into various families, and all reactions in a particular family were assumed to have the same rate constant or a set of rate constants that are a specific function of the carbon number of the species. The model consists of 31 gas-phase species, 29 surface species, and 271 reaction steps, which have been categorized into 33 different families. Each reaction family is parametrized in terms of either a rate constant or an equilibrium constant and the carbon number dependence within a family is considered in terms of the Polanyi relation. Transition state theory has been used to estimate bounds on the preexponential factors, and literature values have been used to bound the activation energies[79−83] and provide interrelationships between reaction families[84,85] to reduce the number of parameters to 13. The proposed model assumes that the reactions of neutral surface alkoxy species[81,86,87] take place through carbenium/carbonium ion transition states. The details of the model and the parametrization methods will be communicated in a future publication. Table 27 shows a summary of the reaction families, the various model parameters, and their allowable bounds.

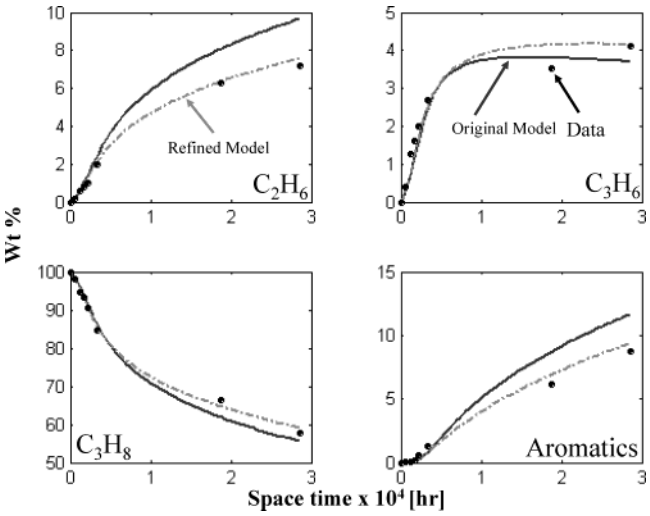


Figure 14. Improvement in performance curves for propane aromatization on HZSM-5. Dots correspond to experimental data from Lukyanov, Gnep, and Guisnet,[45] solid lines indicate the original model predictions, and the dashed line indicates the refined model predictions. The *x*-axis is in terms of the space-time $\times$ 10^4 (hours), and the *y*-axis is the weight percentage of the various species.

3.1. Parameter Estimation and Statistical Analysis. Clearly the above kinetic model for propane aromatization, with 29 algebraic equations and 31 ODEs, is much more complicated as compared to that of the test case considered in section 2.2 for demonstrating the GA-based hybrid pseudoglobal parameter estimator. The parameter bounds as shown in Table 27 are so large that solving the system is not possible using local optimization algorithms with multiple initial guesses that are randomly or uniformly spaced. Experimental data for propane aromatization at 500 °C and 1 bar as reported by Lukyanov and co-workers[45] has been used to fit the model. The search space is complicated, and intuitive initial guesses for all the parameters are difficult. The GA-based hybrid search procedure[46] (section 2.2) with 50 generations and 100 members in each generation was able to identify a pseudoglobal solution with a normalized sum-of-squared error (SSE) of 0.19, where SSE was calculated by the ratio of the sum of the squared differences between the model predictions and the experimental data, as scaled by the experimental data and normalized by the number of data points. The model predictions corresponding to this minimum are shown with the experimental data in Figure 14 for the various species. The performance of this hybrid procedure as compared to randomly generated points, its effectiveness in finding multiple solutions, global statistical analysis,[46] and local statistical analysis of the results[14] are discussed elsewhere.[12]

3.2. Model Refinement. The predictions from the proposed model for paraffin aromatization on HZSM-5 (Figure 14) is reasonable and to the best of our knowledge is better than that available in the literature.[45] However, under close observation it is clear that at lower space times, the slope of the C_2 model prediction curve is higher than that of the data and the slope of the model curve for aromatics concentration is lower than that of the corresponding data. This means that we overpredict C_2 concentration and underpredict the concentration of the aromatics. It is important to account for such small discrepancies at lower space times that correspond to the inlet of the plug flow reactor. This is because the concentrations at the lower space times can nonlinearly affect those at higher space times. So, although the predictions at higher space times are far worse than those that at lower space times, we choose to concentrate on the discrepancies at the lower space times. The overprediction of C_2 and the underprediction of aromatics suggest that we might be missing a reaction step that transforms the light paraffin to aromatics.

To address this concern, we tried a variety of alternate rules and the most effective single rule addition was alkylation of alkoxy species with light alkanes.[88,89] Intuitively, this step that creates larger alkanes from smaller alkanes should be able to drain the smaller alkanes and produce more aromatics. This is because the larger alkane formed by alkylation can further adsorb to form carbonium ions which can then undergo protolysis to give carbenium ions which are longer than those that were present before. These carbenium ions can in turn cyclize and produce aromatics. The addition of this new rule results in the creation of 27 additional elementary reactions and one more model parameter—the rate constant of alkylation. Thus the new model consists of 298 steps and 14 parameters. However, with the help of RMS, the new model can be formulated and evaluated very efficiently. The dashed lines in Figure 14 show the predictions of this refined model. Clearly the predictions at lower space times for C_2 and for aromatics are much better than the predictions corresponding to our earlier model. More importantly, the predictions at the higher space times have also improved substantially for C_2 and aromatics. Since this reaction network is highly coupled, the addition of the alkylation rule has also improved the predictions of other species such as ethane, propane, ethylene, and propylene.

In summary, the human expert made a very reasonable, but incomplete, initial hypothesis to initiate the process of kinetic model building, then with the help of the various tools of RMS, the expert determined an improved rule set with associated kinetic parameters; i.e., knowledge extraction has been demonstrated. The example outlined above regarding improving the quality of the model by the addition of a rule prompted by the feature mismatch in the model and the data should clearly show the general principle behind model refinement. Model refinement is an iterative model, experiment, and expert guided process of adding, deleting, or modifying the rules until a model that satisfactorily explains the data is obtained and, in general, is a difficult task. This is a difficult inverse mapping and search problem, where the expert looks for a new set of rules or modifications to the existing rule set from a large and combinatorial rule hyperspace.

To address this concern, we reformulate MR as a search problem for the true rule set among a combinatorially large rule space. The objective is to down select a set of reaction rules that define a kinetic model for olefin chemistry that plays a critical role in paraffin aromatization as explained in section 2.1. The possible reaction rules are (1) olefin adsorption to produce a carbenium ion, (2) desorption of a carbenium ion to produce an olefin, (3) alkylation of a smaller paraffin by a nonallylic carbenium ion to produce a larger paraffin, (4) β-scission of a carbenium ion to produce an olefin and a smaller carbenium ion, (5) oligomerization of a carbenium ion and an olefin to give rise to a larger carbenium ion, and (6) hydride transfer between a carbenium ion and an olefin/paraffin/diene to yield an alkane and a carbenium ion. Each of these rules can assume several different variations as shown in Table 28. For example, there are three different rule variations for olefin adsorption: no reaction; only 2° and 3° carbenium ion with up to seven or eight or nine carbon atoms as the reactant; and any carbenium ion with up to seven or eight or nine carbon atoms as the reactant. Olefin desorption can only assume two variations and the bimolecular reaction; hydride transfer can take place in any of the 34 different forms depending on the first and second reactants. These variations typically form the palette from which a modeler chooses for explaining the olefin chemistry in the context of paraffin adsorption. A rule set is constructed by picking one variation of each rule. Considering all the different possibilities in Table 28, a total of 333 200 different rule sets are possible. Every such rule set is equivalent to a kinetic model, and the set of all rule sets corresponds to the rule space. RDL++ translates the rule set into the corresponding kinetic model by using propane and propylene as the starting reactants. All reaction networks are restricted to species with up to 12 carbons and are characterized by six rate constants, one for each of the reaction class.

The objective of this study is to find a rule set from the given rule space (Table 28) that best corresponds to the data through successive model refinement. For demonstration purposes, we choose the rule set as shown in Table 29 as the target. This rule set leads to a kinetic model with 5 gas-phase species (propane, propylene, hexane, hexene, and hex-1,2-ene), 3 surface species (carbenium ion formed by the adsorption of propene, hexene, and hex-1,2-ene), and 15 reactions. Table 29 also shows the bounds on the various model parameters and their values for the target model. Product distribution data corresponding to propane, propene, and lumps of all other paraffin and all other monoenes is simulated using this model. This will be used as the target experimental data in the model refinement case study.

A knowledge-based, guided stochastic search based on genetic algorithms[90,91] (GA) is used to search for the model that best corresponds to the simulated data. Each of the solutions is represented as a string of numbers that represents each of the rules. For example, the string corresponding to the target rule is 626 333. This represents the fact that this rule set involves the sixth variation (Table 28) of the rule for olefin adsorption: All carbeniums up to C_8 are formed, the second variation of carbenium desorption, and so on. Fitness of each of the strings is calculated by translating the rule string into the corresponding kinetic model using RDL++ and subsequently fitting this model to the target data using the hybrid algorithm based on GA as discussed in

Table 28. Possible Variations among the Different Reaction Rules of Olefin Chemistry that Constitute the Rule Space Used for Automated Model Refinement.

rule variations	no. of possibilities
olefin adsorption	
no reaction	1
no 1° carbenium formed, up to C_7, C_8, or C_9	3
all carbenium formed, up to C_7, C_8, or C_9	3
total	7
carbenium desorption	
no reaction	1
all carbenium desorb	1
total	2
alkylation	
no reaction	1
no 1° carbenium as reactant, up to C_7, C_8, or C_9	3
ny carbeninum can react, up to C_7, C_8, or C_9	3
total	7
β-scission	
no reaction	1
all carbenium except allylics react, up to C_7, C_8 or C_9	3
all carbenium ($\geq C_5$) except allylics react, up to C_7, C_8, or C_9	3
no 1° carbenium product formed, up to C_7, C_8, or C_9	3
total	10
oligomerization	
no reaction	1
all carbenium except allylics react, up to C_7, C_8, or C_9	3
all carbenium except 1° and allylics react, up to C_7, C_8, or C_9	3
no 1° carbenium product formed, up to C_7, C_8, or C_9	3
total	10
hydride transfer	
no reaction	1
first reactant	
all carbenium except allylics react, up to C_7, C_8, or C_9	
second reactant	
paraffin or monoene	3
monoene	3
diene	3
paraffin	3
monoene or diene	3
first reactant	
any carbenium can react, up to C_7, C_8, or C_9	
second reactant	
paraffin	3
monoene	3
diene	3
monoene or diene	3
paraffin, monoene, or diene	3
first reactant	
all carbenium except allylics react, up to C_7, C_8 or C_9	
no 1° carbenium formed	
second reactant	
paraffin, monoene, or diene	3
total	34

[a] Allylic stands for a species with a double bond and a positive charge and carbenium means a positively charged tricoordinated carbenium ion.

section 2.2. The resultant SSE is the fitness value of this model. The GA is used to find the model with the lowest SSE which is the one that best explains the target data.

A hybrid fitness proportionate and random selection criterion is used in every generation to identify the individuals to be manipulated by the GA operators. Uniform crossover and single-point mutation with probabilities 0.85 and 0.5, respectively, have been used, and the top 10% of the solutions in every generation are preserved in the next generation. A custom operator known as the complement-carbon-number that modifies the size of the reactants has been defined. For example, this operator can change the fifth variation of the olefin adsorption rule which corresponds to the formation of carbenium ions up to C_7 to the sixth variation of the rule that corresponds to the formation of carbenium ions up to C_8. Thus, this operator modifies the rule sets such that the size of the resultant reaction network changes.

All the three genetic operators—uniform crossover, single point mutation and complement-carbon-number—maintain the feasibility of the rule set. The maximum size of the reactants in all the reaction rules in any rule set should be the same in order to prevent any inconsistencies in the size of the species involved in the different reactions. Adsorption, desorption, and alkylation are assumed to be always present in every rule set, and the rule variations for β-scission and oligomeriza-

Table 29. The Target Rule Set Used in the Automated Model Refinement Case Study[a]

rules	no. of reactions	bounds on the rate constants	rate constants for the target model
olefin adsorption: any carbenium up to C_8 can form	3	$1 \le k_{oa} \le 10^3$	10
carbenium desorption: all carbenium desorb	3	$100 \le k_{od} \le 10^5$	$100k_{oa}$
alkylation: all carbenium up to C_8 react	1	$0.1 \le k_{alk} \le 100$	80
β-scission: all carbenium up to C_8 except allylics react	1	$10^5 \le k_b \le 10^7$	10^6
oligomerization: all carbenium up to C_8 except allylics reacts	1	$10^4 \le k_{olig} \le 10^6$	$0.1k_b$
hydride transfer: first reactant, all carbenium up to C_8 except allylic; second reactant, paraffin or monoene	6	$1 \le k_h \le 10^4$	10^3

[a] The rule set has been down selected from the rule space defined in Table 28. Number of reactions generated by RDL++ corresponding to each rule, bounds on the six rate constants and their values used for generating the data for the target model are also given. Allylic stands for a species with a double bond and a positive charge and carbenium means a positively charged tricoordinated carbenium ion.

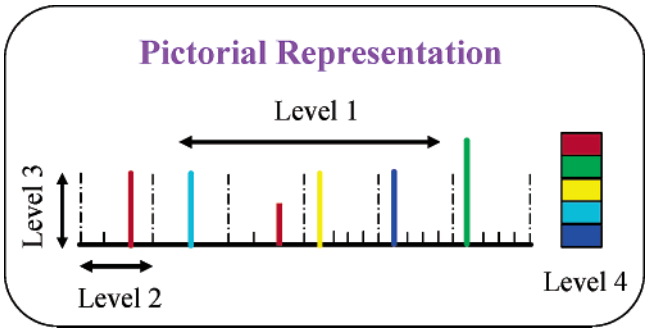


Figure 15. Pictorial representation of the rule sets in the automated model refinement case study.

tion should be similar as they are reverse of each other. These restrictions reduce the search space from 333 200 to approximately 13 000. To demonstrate the power of the GA as a search procedure, we sample only four generations with six rule sets in each. This corresponds to a total of 24 solutions, which is less than 0.2% of the total of 13 000 solutions. This ensures that we do not exhaustively search through the space of possible solutions.

We now introduce a pictorial representation of the rule sets as shown in Figure 15 in order to explain the search results. There are six bins corresponding to each of the rules in the rule set. The finer divisions within a bin represent the second major variation within each rule. For example, as shown in Figure 15, the vertical bar at the second location in the first bin corresponds to the rule variation that all possible carbenium ions are formed during adsorption of an olefin. The no-reaction variation is not considered for pictorial representation of the rule of adsorption as it is an infeasible alternative. Similarly, the fourth, fifth, and the sixth bins have four divisions in them corresponding to the four different variations. The next level of rule variation corresponding to the number of carbons specified in the reaction rules is represented by the height of the vertical bars.

Rules that have restrictions on the carbon numbers up to C_7 are represented by the smallest height followed by those with the C_8 restriction and then by the case with the rules that allow for reactants up to C_9. If there are any more variations, then this is depicted by the color of the bars. For example, the hydride transfer rule has variations based on its second reactant. This is represented by the color of the bar. The target rule 626 333 is shown pictorially in Figure 16. The bar of intermediate size in the second division of the first bin represents olefin adsorption that allows the formation of all carbenium ions up to C_8. It is important to note

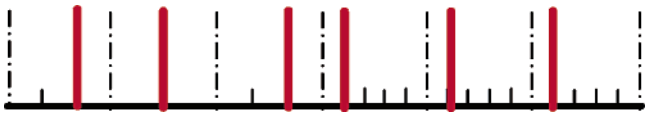


Figure 16. Pictorial representation of the target rule 626333. The rule variations are as shown in Table 29.

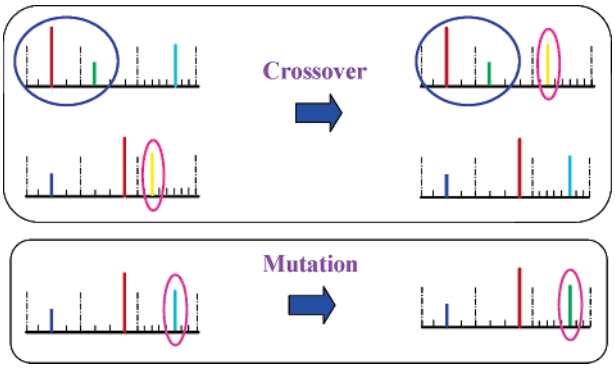


Figure 17. Pictorial representation of the genetic operators of uniform crossover and single point mutation.

that all the bars have the same height. This is because, for any chemically consistent rule, the maximum number of carbons in any reactant in all the reaction rules should be the same. The red bar in the first subdivision of the last bin corresponding to hydride transfer represents that the second reactant could be a paraffin or a monoene (Figure 16) and that all carbenium ions except allylics up to C_8 can react. Figure 17 pictorially show the crossover and mutation operators employed in the GA search. It is interesting to note that crossover makes large jumps in the search space and mutation leads to only small changes in the rules. The complement-carbon-number operator will change the height of the vertical bar in any of the bins; however, to maintain chemical consistency, the heights of the other bars are also modified to be the same.

The GA search procedure for the rule set that closely corresponds to the target model shown in Table 29 is seeded with a set of chemically consistent random rule sets. These random rules are shown pictorially in Figure 18. Also shown in this figure is the true model. The data generated by the true target model are shown as circles, and the solid lines represent the predictions from the randomly generated models. Clearly the predictions due to the randomly generated models are not good. After four generations of hybrid selection and chemically feasible genetic operations, the predictions from the new rule sets have improved as shown in Figure 19. The figure also shows the true model, the two best models from the first generation of GA and the three best models from the GA search in the increasing order of

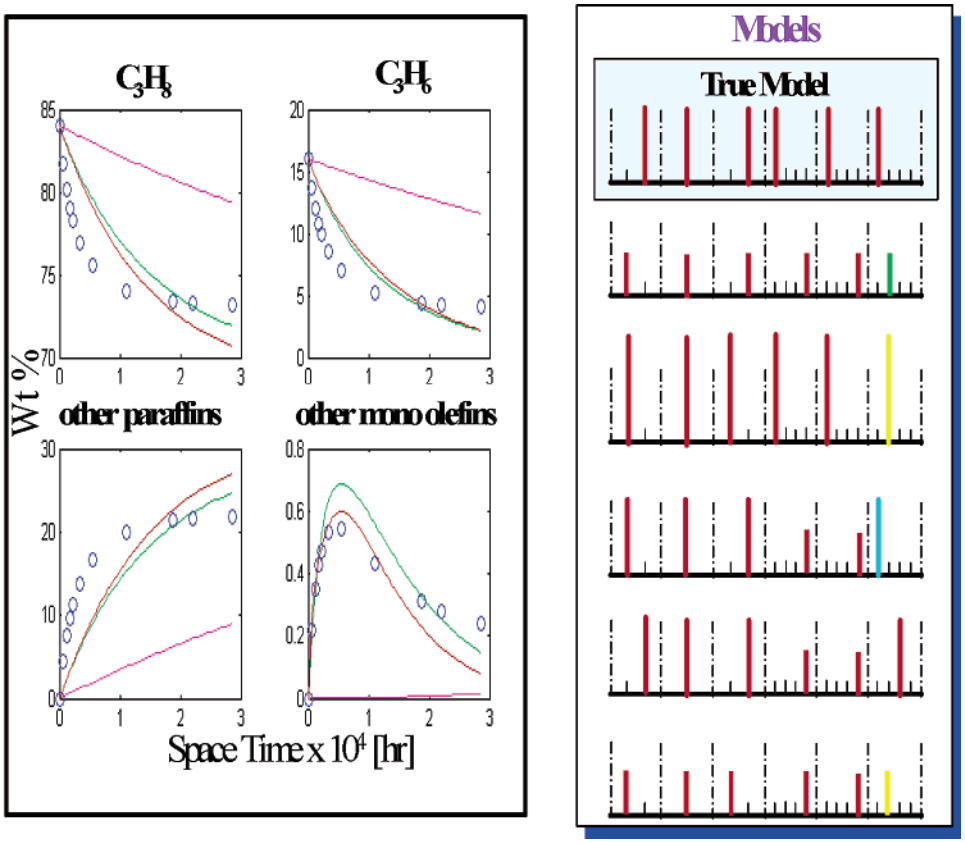


Figure 18. Pictorial representation of the randomly generated initial set of rules to seed the genetic algorithm-based model refinement search. Also shown is the true model corresponding to the rule variation shown in Table 29. The solid lines on the plots show the predictions corresponding each of the random set of models and the circles correspond to that of the data generated by the true model using the parameters in Table 29.

SSE. The best solution of the GA search clearly resembles the true model much more than any of the solutions in the initial solution. The height of the bars of this best solution that represent the maximum size of the reactant is same as that of the true model. However, it is clear that the locations of the bars of this best model and those corresponding to the true model are quite different. It is also interesting to note that many more models now predict the data well as compared to the predictions of the initial set of models shown in Figure 18.

3.3. Experimental Formulation. During the iterative procedure of hypothesis generation and testing, it is typical that the suggested hypothesis does not explain the data or more than one hypothesis explains the data equally well. In the former case, we need to identify the cause for the discrepancy by exercising model refinement (MR). We did this through a GA-based search procedure in section 3.2. In the latter case, when more than one model explains the data equally well, we need to discriminate among the multiple models. This can be achieved by evaluating the equally good models against new discriminatory data—analytical measurements of a new set of species or measurements of species with different feeds or at different temperatures, etc. Models that could explain the data from one part of the network may not be able to explain the data from a different part of the network. Thus the task of discriminating among models leads to suggestions about new sets of experiments and we call this step as the formulation of experiments.

At the end of the GA search as discussed in section 3.2, we found improved models whose predictions were better than the initial set of randomly generated models; however, now there are several models that explain the data equally well. The current data set corresponds to the two input species (propane and propylene) and lumped forms of other downstream species (all paraffins other than propane and all monoenes other than propylene). These kinds of lumped analytical measurements are useful during the initial stages of model building; however, now we are unable to discriminate among the different models with these data. If further details about the species are available, we may be able to distinguish among the various models. So, we choose to measure at least one downstream species, say, hexane, and one species which has two double bonds, say, hex-1,2-ene, which also appears downstream of the initial set of reactants. Figure 20 shows the predictions of the top three models at the end of the GA search against this new data set. The predictions for propane and propylene remain the same as in Figure 19; however, there is only one model that predicts the time evolution of hex-1,2-ene reasonably well. The other two models form very little of this species and so are not as good in explaining the data. Clearly using the additional resolution in the data, we have achieved model discrimination. However, we still do not have the right model that can explain the data. Hence, we choose to perform another iteration of the GA-based MR, now with the new data set.

Figure 21 pictorially shows the three best models at the end of the second iteration of MR with measurements from new species suggested as part of experimental formulation. This figure also shows the target model and the two best models from the previous

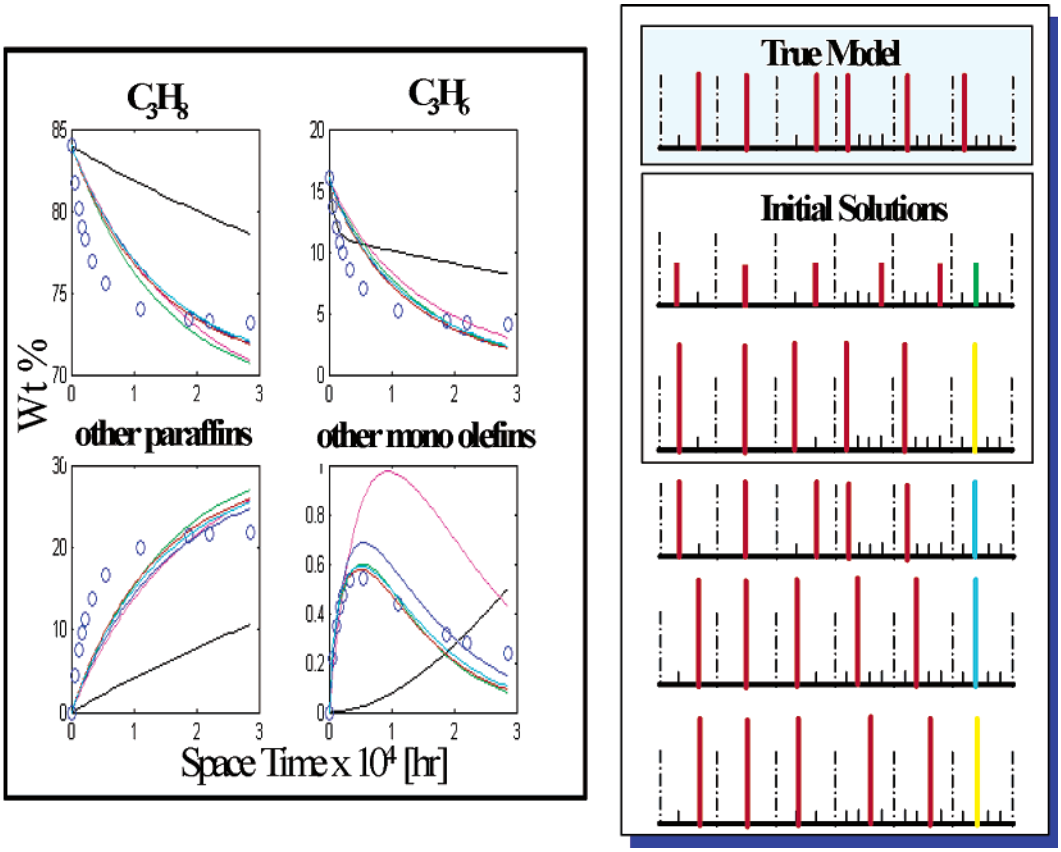


Figure 19. Pictorial representation of the three best models after four generations of the genetic algorithm search. The true model and two of the initial solutions are also shown. The solid lines represent the predictions corresponding to the solutions after the genetic algorithm search and the circles represent the predictions of the true model.

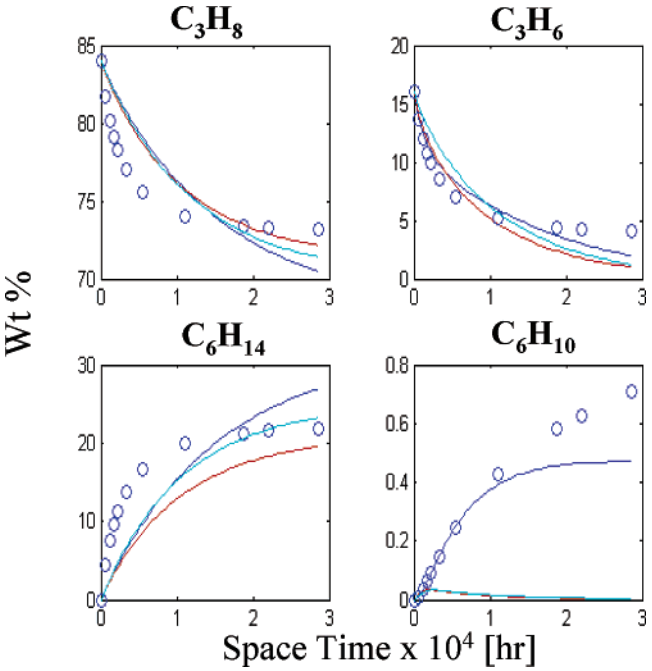


Figure 20. Predictions of the models at the end of the first iteration of model refinement for a new set of species. Measurements of downstream species such as hexane and hex-1,2-diene instead of lumped paraffin and monoenes lead to discrimination among models.

iteration of MR in the increasing order of SSE. The predictions of these models have substantially improved as compared to the models from the previous iteration of MR as shown in Figure 20. The current best model explains the data very well. The pictorial representation of the current best model resembles that of the true model except for the height of the vertical bars. This means that these two models have exactly the same set of rules except that the maximum size of the reactant in these rules is different—up to C_8 in the true model and up to C_7 in the current best model. The predictions do not seem to be affected because of this variation. With a few more iterations of MR, we are able to find a rule set that is identical to the target rule set; however, after one iteration of experimental formulation and two iterations of MR, we could refine an initial set of random models, discriminate among equally good models, and find a model whose predictions are almost identical to that of the target. Thus, we have demonstrated the iterative procedure of MR and experimental formulation.

Typically, the process of refining models is approached in the form of deleting and adding an elementary reaction.[20,92] Our modeling step starts from the chemistry rules of the domain expert rather than the individual reactions. The process of changing rules rather than individual reactions is chemically more intuitive because the experts think in terms of rules and also a single rule change can affect a large number of reactions that may not be in the local vicinity of one another. Also, the allowable search space in a typical reaction network with 300 elementary steps, for a single instance of the rule set, is of the order of 3^{300}, the three possibilities signifying the absence of an individual reaction and its presence in the forward or reverse direction. On the contrary, the rule space is typically of the order of 15 with around five variations in each of

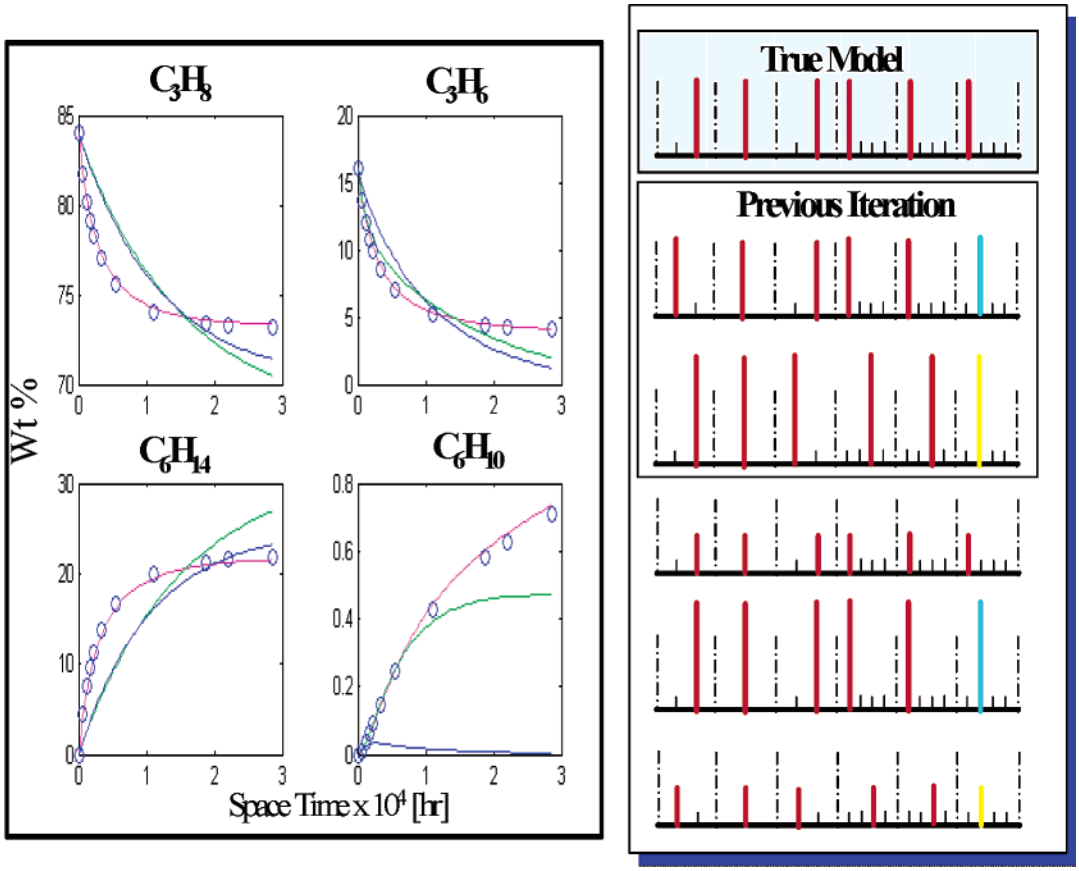


Figure 21. Pictorial representation of models after the second iteration of model refinement with measurements from new species suggested as part of experimental formulation. The best model from the genetic algorithm search closely resembles that of the true model. Predictions of the best model after two iterations of model refinement explain data better than compared to that of the initial set of models shown in Figure 20.

these, thereby giving rise to a search space of 10^{10}, which is much smaller than 3^{300}.

The current automated MR strategy is based on knowledge-guided stochastic search of the rules. This is based on a custom genetic algorithm and is driven by the causal models postulated by the expert about how the rules affect the resultant model predictions. Provisions for probabilistic acceptance of the initial causal models, learning novel causal patterns that arise in the evolution process, and guidance using the local concentration or rate sensitivity coefficients[93] will enhance the capability of this procedure. RMS has been used to faithfully translate the knowledge from the expert into quantitative models and evaluate them; however, using the above model refinement procedure, it is possible to discover combinations of rules resulting in new pathways that may not have been considered by the expert because of the sheer size of the number of possibilities.

3.4. Search for Novel Catalyst Formulations Using Genetic Algorithms. The main objective of developing robust forward models is to use them in the context of computer-aided materials design (Figure 1) in order to design catalysts that meet a set of desired performance criteria.[13,94] The most important challenge in designing a zeolite catalyst for paraffin aromatization is in developing a good forward model that would predict the product distribution (paraffin, olefin, diene, cyclic olefin, aromatics, etc.) with the catalyst structure and process conditions such as the contact time on the catalyst, temperature, pressure, etc. The kinetic model presented in section 3.1 that predicts the product distribution with the contact time on the catalyst is a valuable starting point towards this.

We now demonstrate how the kinetic model developed in section 3.1 can be used to design a new catalyst with improved aromatics yield. The catalyst will be characterized by the set of kinetic model parameters as in Table 27. The objective here is to search for a set of parameters that would give the maximum aromatics yield, given that the catalyst will behave according to the kinetic model that we have developed. Specifically, we want to find a catalyst with the maximum aromatics yield using our model as it can explain the product distribution of the current catalyst with 7% aromatics yield.[45] The hybrid GA-based search procedure used for parameter estimation as discussed in section 2.2 is used for this search; however, we now minimize the reciprocal of the aromatics yield instead of the sum of the squared error evaluated as the difference between the model and the data. The aromatics yield is defined as the weight percentage of the aromatics in the product mixture.

Figure 22 shows the set of parameters corresponding to this new catalyst that gives 58% aromatics yield as the third vertical bar. This figure also compares the values of these parameters to that corresponding to the catalyst that was used for validating our model (second vertical bar) and the lower (first bar) and upper (fourth bar) bounds of the parameters (Table 27). It is interesting to note that the rate constant for carbonium ion dehydrogenation (parameter 11) has been increased to reach its upper bound. It is well-known that the presence of metal additives such as Ga as the extraframe-

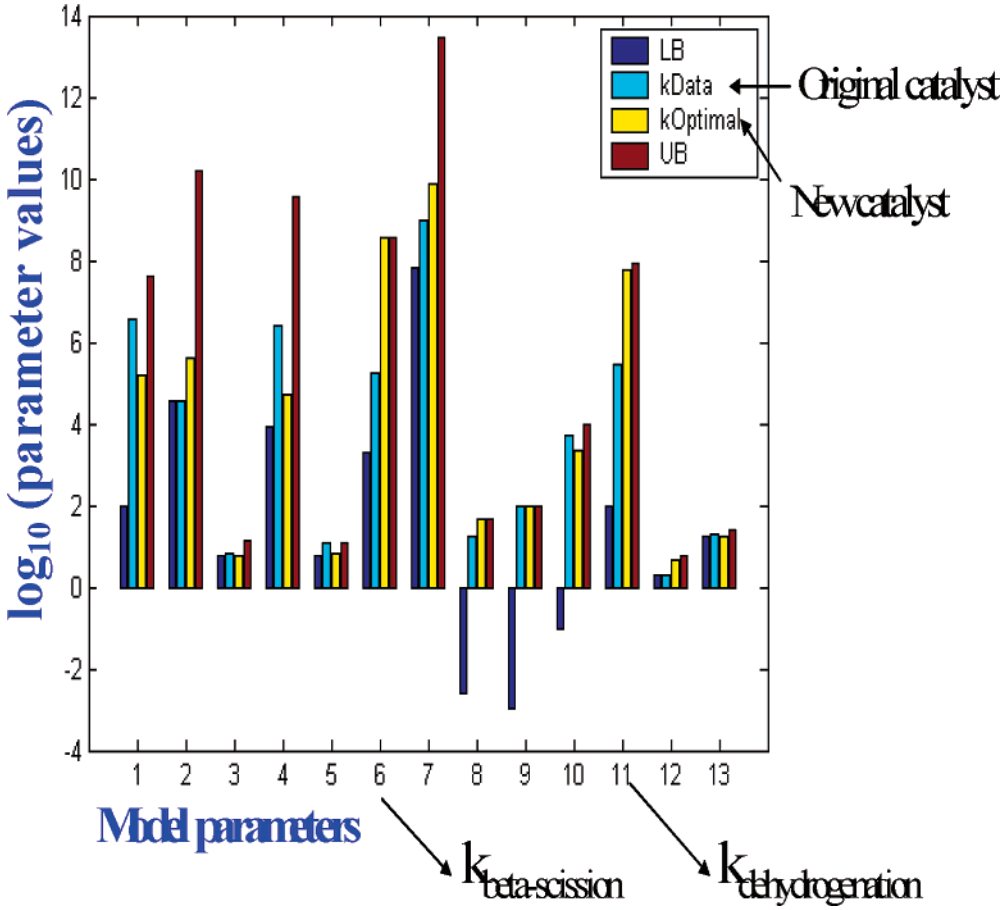


Figure 22. Parameters corresponding to an improved catalyst with 58% aromatics yield found by the inverse search procedure by using the kinetic model for paraffin aromatization on zeolites. LB and UB correspond to the lower and upper bounds on the parameter values. kData corresponds to the value of the rate constants corresponding to the catalyst with 7% aromatics yield[45] and kOptimal represents the new catalyst with 58% aromatics yield.

work metal ions increase the aromatics yield of zeolite catalysts.[78] Also it is intuitive to see that the rate constant for the aromatization reaction (parameter 7) is increased. Another interesting change is the increase in the value of the rate constant for β-scission (parameter 6). The rate constant for oligomerization has been defined as the product of the rate constant of β-scission and an equilibrium constant for the reversible reaction of β-scission/oligomerization. This equilibrium constant, in turn, is characterized by the entropy as in parameter 13. Hence, the marginal decrease in the entropy and substantial increase in the rate constant for β-scission signifies the increase in the rate of oligomerization. This is intuitive as an increase in the rate of oligomerization would increase the rate at which larger carbenium ions are formed and hence increases the rate of formation of cyclic olefins and aromatics. Thus the change in the values of the model parameters is consistent with the objective function of increasing the aromatics yield.

The above procedure has been able to find a set of model parameters that correspond to a catalyst with an improved aromatics yield. However, the aromatics yield predicted by this new catalyst is the upper bound that can ever be achieved using this chemistry. Any additional restrictions on the model parameters in the form of either relationships between the catalyst structure and the kinetics or interrelationships among the existing parameters will only reduce the number of degrees of freedom among the model parameters. This may lead to a decrease in the aromatic yield reported here. This process of catalyst design and improvement can be enhanced with a model that maps the effect of the catalyst structural and electronic descriptors to the reaction kinetics.[14] This model will help to relate the new parameters as found in the current study to the actual structure of the catalyst that the expert could make in the laboratory.

The upper bound on the aromatics yield as found by the GA-based search can be used as follows. If for economical reasons or otherwise one cannot afford to retrofit an existing reactor setup in the plant unless a catalyst with 60% yield is available, searching in the design space of a zeolite catalyst with carbonium/carbenium chemistry may be futile as it can only yield a maximum of 58% yield. Either a completely different catalyst or a modified zeolite catalyst that follows new chemistry has to be evaluated for this purpose. This kind of guidance for eliminating possibilities can potentially save a lot of time and effort in the process of catalyst development.

4. Summary and Discussion

In this paper, we have demonstrated how ideas from process systems, artificial intelligence, and machine learning can be used to design, develop, customize, and integrate a set of tools that aid an expert in building robust kinetic models to be used for catalyst design (Figure 1). As shown in Figure 23, the model-building procedure starts with the expert down-selecting a specific rule set from a large rule space, formulating a reaction network from the selected rule set, translating

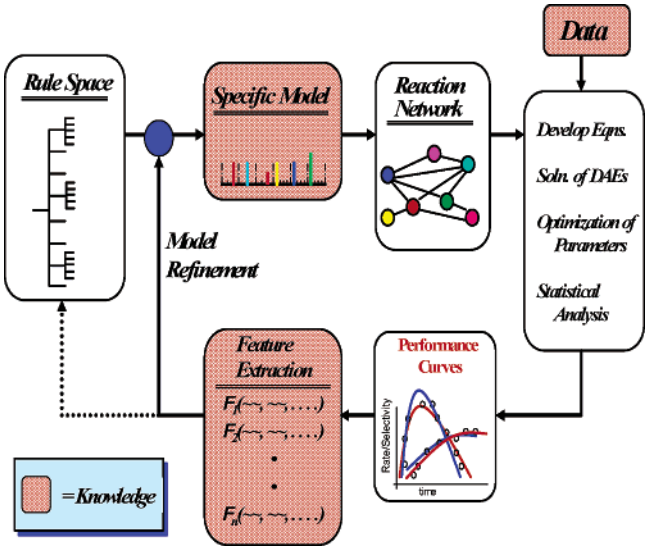


Figure 23. Schematic of the model building procedure designed, developed, and implemented in this paper.

Table 30. Different Models for CO Oxidation within Each Model Type

model no.	CO reaction	O_2 reaction
1	quasi-equilibrated	irreversible
2	reversible	irreversible
3	irreversible	irreversible
4	quasi-equilibrated	reversible
5	reversible	reversible
6	irreversible	reversible
7	quasi-equilibrated	quasi-equilibrated
8	reversible	quasi-equilibrated
9	irreversible	quasi-equilibrated

the reaction network into a mathematical model, solving the equations, estimating the model parameters, statistically analyzing the results, extracting the features of the model predictions, and using the model−data feature mismatch to initiate the process of model refinement. The pieces of knowledge that the user can directly relate to are the model, data, and the model−data feature mismatch.

Reaction Modeling Suite (Figure 3) facilitates rapid formulation of hypothesis, thorough screening of models and their analysis. The overall idea is based on the postulate that any computer tool should mimic the thought process of the human expert. Toward this end, we have developed the various tools in the RMS. For example, the Reaction Description Language Plus Plus is a compiler that enables an expert to initiate kinetic model building in the natural language of the chemist. The hybrid parameter estimation procedure affords a thorough search of the vast nonlinear space of parameters in a computationally efficient manner. The Feature Extractor ensures that the output of the model building exercise is in a language that is highly intuitive to the expert. The Statistical Analyzer evaluates the robustness of the models. The concept of model refinement that allows us to start with an approximate model and iteratively converge to a better predictive model and the idea of experimental formulation for discriminating among similar models has been demonstrated. All the above ideas have been used to develop a kinetic model for propane aromatization on zeolites. This kinetic model has been subsequently used to search for a set of kinetic parameters that correspond to a catalyst with improved aromatics yield.

Acknowledgment

The authors wouuld like to thank the Indiana 21st Century Research and Technology fund for their support of this research. We also gratefully acknowledge the support of the U.S. Department of Energy, Office of Basic Energy Sciences, through the Catalysis Science grant number DE-FG02-03ER15466.

Appendix A: CO Oxidation on Supported Metal Catalysts

CO oxidation involves the adsorption of CO and O_2 on the surface of a catalyst and their reaction to give CO_2. This problem in its own right is important as it finds application in the automobile catalytic converter. Due to its relative simplicity, both the experimentalists and the modelers study this reaction. Models available in the literature assume surface homogeneity—CO and O_2 adsorb randomly on any part of the surface with equal probability and react to give CO_2. Razon and Schmitz[95] have reviewed a large part of this modeling literature until 1986, and recent work by Lund et al.[96] support this. As noted by Mukesh and co-workers,[97] it is somewhat surprising that the elementary step models predict experimental data so well since the true situation on the catalyst surface even for relatively simple reactions is complicated.[98]

Models for CO oxidation in the literature fall under four major categories depending on the major assumptions: Langmuir−Hinshelwood models with one type of adsorption site for both CO and O_2 and a mean field approximation about the availability of the adsorbed species for reaction; two-site models which consider two different sites for the adsorption of CO and O_2; models that assume adsorption of CO to be on two different sites; the perimeter models that allow reaction between CO and O_2 only at the perimeter of the CO islands. Within these different types of models,

Table 31. Typical Reaction Model Representation for CO Oxidation[a]

model tag	reactions	model equations
LH1	$CO + S \leftrightarrow COS$ (K_1)	$x - K_1P_{CO}(1 - x - y) = 0$
	$O_2 + 2S \rightarrow 2OS$ (k_2)	$k_2P_{O_2}(1 - x - y)^2 - k_3xy = 0$
	$COS + OS \rightarrow CO_2$ (k_3)	$r = k_3xy$
TS5	$CO + S_1 \leftrightarrow COS_1$ (k_{11}, k_{12})	$k_{11}P_{CO}(1 - x) - k_{12}x - k_3xy = 0$
	$O_2 + 2S_2 \leftrightarrow 2OS_2$ (k_{21}, k_{22})	$k_{21}P_{O_2}(1 - y)^2 - k_{22}y^2 - k_3xy =$
	$COS_1 + OS_2 \rightarrow CO_2 + S_1 + S_2$ (k_3)	$r = k_3xy$
COTS9	$CO + 2S \rightarrow SCOS$ (k_1)	$K_1PC_O(1 - x - y)^2 - k_3xy = 0$
	$O_2 + S \leftrightarrow 2OS$ (K_2)	$y_2 - K_2P_{O_2}(1 - x - y)^2 = 0$
	$SCOS + OS \rightarrow CO_2 + 3S$ (k_3)	$r = k_3xy$
PM1	$CO + S \leftrightarrow COS$ (K_1)	$x - K_1P_{CO}(1 - x - y) = 0$
	$O_2 + 2S \rightarrow 2\ OS$ (k_2)	$k_2P_{O_2}(1 - x - y)^2 - k_3x^{1/2}y = 0$
	$COS + OS \rightarrow CO_2$ (k_3)	$r = k_3x^{1/2}y$

[a] The symbol $\leftrightarrow$ represents an equilibrium reaction.

Table 32. Pseudoalgorithm of the Trough-Walking Algorithm

estimate the parameter initial guesses from the parameter bounds
for all parameter initial guesses
  obtain a local minimum, K_{opt}, using the parameter estimator
  evaluate the Jacobian matrix, **J** at K_{opt}
  evaluate the inverse of the maximum eigen values, *s*, and the eigen vector (ev) corresponding to the minimum eigen value of $\mathbf{J}^T\mathbf{J}$
  calculate $k_{init-new} = k_{opt} + (ev\ s)$
  start local minima search from $k_{init\text{-}new}$
  accept the new minima K_{curr} if Euclidean distance between k_{curr} and k_{opt} is less than EDBOUND
end

depending on the reversibility, irreversibility, or the equilibrated nature of the CO and O_2 adsorption steps, we can formulate nine different models (Table 30). This leads to a total of 36 different models. The reactions and the equations of a representative set of four models are shown in Table 31. The model denotedby LH1 allows for a molecular CO adsorption that is quasi-equilibrated and a dissociative and irreversible O_2 adsorption. This model has been used in section 2.4 to demonstrate the Statistical Analyzer of the RMS.

Appendix B: Trough-Walking Algorithm for Locating Multiple Minima

Parameter estimation is a nonlinear optimization problem for which local optimization algorithms have been traditionally used. Although deterministic global optimization algorithms[48] have been recently developed for this purpose, modelers still rely on the local optimization algorithms due to their ready availability in commercial packages. One of the major drawbacks of the local optimization algorithms is that they are strongly affected by the initial guess values and can get trapped into local minima. In this study, we use a modified Levenberg−Marquardt local optimization algorithm to search the local neighborhood of the minima found by starting with multiple initial guess values. Thus this trough-walking algorithm is an attempt to identify as many local minima as possible.

Any local minimum obtained by starting from an initial guess is searched for other minima in its neighborhood. The neighborhood is defined in terms of a Euclidean distance. When a local optimum value is located, the eigen vector corresponding to the minimum eigen value of the $\mathbf{J}^T\mathbf{J}$ matrix (**J** is the Jacobian matrix consisting of the derivatives of the objective function with respect to the parameters) at the optimum identifies the direction in which the objective function changes the minimum.[65] A new initial guess value is obtained as

$$k_{init-new} = k_{opt} + (ev\ s) \qquad \text{(B1)}$$

where k_{opt} is the minimum around which the search is being carried out, ev is the eigen vector corresponding to the minimum eigen value, and *s* is a step size (taken to be the inverse of the maximum eigen value of the $\mathbf{J}^T\mathbf{J}$ matrix). The minimum obtained by the local search from this initial guess value is accepted if it satisfies the SSE criterion and if it is sufficiently (defined by another user defined parameter, EDBOUND=0.05) far from all minima already found. The distance between any two minima is measured in terms of the Euclidean distance between them. This is similar to locating all the optima along a shallow trough. The pseudocode of this trough-walking algorithm is given in Table 32. The overall idea for this intensive search is to locate all possible optimum values that satisfy the experimental data mathematically.

Literature Cited

(1) Jones, C. W.; Tsuji, C.; Davis, M. E. Organic-Functionalized Molecular Sieves as Shape-Selective Catalysts. *Nature* **1998**, *393*, 52−54.

(2) Somorjai, G. A. *Introduction to Surface Chemistry and Catalysis*; John Wiley and Sons: New York, 1994; p 688.

(3) Hahn, J. R.; Ho, W., *Phys. Rev. Lett.* **2001**, *87*, 166102.

(4) Boussie, T. R.; Diamond, G. M.; Goh, C.; Hall, K. A.; LaPointe, A. M.; Leclerc, M.; Lund, C.; Murphy, V.; Shoemaker, J. A. W.; Tracht, U.; Turner, H.; Zhang, J.; Uno, T.; Rosen, R. K.; Stevens, J. C. Fully Integrated High-Throughput Screening Methodology for the Discovery of New Polyolefin Catalysts: Discovery of a New Class of High-Temperature Single-Site Group (Iv) Copolymerization Catalysts. *J. Am. Chem. Soc.* **2003**, *125* (14), 4306−4317.

(5) Buyevskaya, O. V.; A., B.; Kondratenko, E. V.; Wolf, D.; Baerns, M., Fundamental and Combinatorial Approaches in the Search for and Optimization of Catalytic Materials for the Oxidative Dehydrogenation of Propane to Propene. *Catal. Today* **2001**, *67*, 369−378.

(6) Cawse, J. N. *Experimental Design for Combinatorial and High Throughput Materials Development*; John Wiley & Sons: New York, 2002; p 336.

(7) Banares-Alcantara, R.; Westerberg, A. W.; Ko, E. I.; Rychener, M. D. Decade—a Hybrid Expert System for Catalyst Selection-I. Expert System Consideration. *Comput. Chem. Eng.* **1987**, *11* (3), 265−277.

(8) Banares-Alcantara, R.; Ko, E. I.; Westerberg, A. W.; Rychener, M. D., Decade—a Hybrid Expert System for Catalyst Selection-Ii. Final Architecture and Results. *Comput. Chem. Eng.* **1988**, *12* (9−10), 923−938.

(9) van Santen, R. A. Theory, Spectroscopy and Kinetics of Zeolite Catalyzed Reactions. *Catal. Today* **1999**, *50*, 511−515.

(10) Jacobsen, J. H. C.; Dahl, S.; Clausen, B. S.; Bahn, S.; Logadottir, A.; Norskov, J. K. Catalyst Design by Interpolation in the Periodic Table: Bimetallic Ammonia Synthesis Catalysts. *J. Am. Chem. Soc.* **2001**, *123*, 8404−8405.

(11) Dumesic, J. A.; Rudd, D. F.; Aparicio, L. M.; Rekoske, J. E.; Trevino, A. A. *The Microkinetics of Heterogeneous Catalysis*; American Chemical Society: Washington, DC, 1993; p 316.

(12) Katare, S. A Rational Automated Knowledge Framework for Reaction Kinetic Modeling and Catalyst Design. Ph.D. Thesis, Purdue University, West Lafayette, IN, 2003.

(13) Sundaram, A.; Ghosh, P.; Caruthers, J. M.; Venkatasubramanian, V. Design of Fuel Additives Using Neural Networks and Evolutionary Algorithms. *AIChE J.* **2001**, *47* (6), 1387−1406.

(14) Caruthers, J. M.; Lauterbach, J. A.; Thomson, K. T.; Venkatasubramanian, V.; Snively, C. M.; Bhan, A.; Katare, S.; Oskarsdottir, G. Catalyst Design: Knowledge Extraction from High-Throughput Experimentation. *J. Catal.* **2003**, *216* (1−2), 98−109.

(15) Clarke, B. L. Stoichiometric Network Analysis of the Oxalate-Persulfate-Silver Oscillator. *J. Chem. Phys.* **1992**, *97* (4), 2459−2472.

(16) *Canonical Nonlinear Modeling*: *S-System Approach to Understanding Complexity*; Voit, E. O., Ed.; Van Nostrand Reinhold: New York, 1991; p 384.

(17) Lukyanov, D. B. Development of Kinetic Models for Reactions of Light Hydrocarbons over Zsm-5 Catalysts. Experimental

Studies and Kinetic Modelling of Ethene Transformation and Deactivation of Hzsm-5 Catalyst. *React. Kinet. Dev. Catal. Processes* **1999**, *122*, 299−306.

(18) Happel, J.; Sellers, P. H.; Otarod, M. Mechanistic Study of Chemical Reaction Systems. *Ind. Eng. Chem. Res.* **1990**, *29*, 1057−1064.

(19) Tsuchiya, T.; Ross, J. Application of Genetic Algorithm to Chemical Kinetics: Systematic Determination of Reaction Mechanism and Rate Coefficients for a Complex Reaction Network. *J. Phys. Chem* **2001**, *105*, 4052−4058.

(20) Koza, J. R.; Mydlowec, W.; Lanza, G.; Yu, J.; Keane, M. A. *Reverse Engineering and Automatic Synthesis of Metabolic Pathways from Observed Data Using Genetic Programming*; SMI-2000-0851; Stanford University, 2000.

(21) Arkin, A. P. Synthetic Cell Biology. *Curr. Opin. Biotechnol.* **2001**, *12*, 638−644.

(22) Prickett, S. E. Generation and Enumeration of Conjugation Chains in Acyclic Compounds. M.S. Thesis, University of Maryland at College Park, 1992.

(23) Ugi, I.; Bauer, J.; Brandt, J.; Freidrich, J.; Gasteiger, J.; Jochum, C.; Schubert, W. New Applications of Computers in Chemistry. *Angew. Chem., Int. Ed. Engl.* **1979**, *18*, 111−123.

(24) Broadbelt, L. J.; Stark, S. M.; Klein, M. T. Computer Generated Pyrolysis Modeling: On-the-Fly Generation of Species, Reactions and Rates. *Ind. Eng. Chem. Res.* **1994**, *33*, 790−799.

(25) Quann, R. J.; Jaffe, S. B. Structure-Oriented Lumping: Describing the Chemistry of Complex Hydrocarbon Mixtures. *Ind. Eng. Chem. Res.* **1992**, *31* (11), 2483−2497.

(26) Quann, R. J.; Jaffe, S. B. Building Useful Models of Complex Reaction Systems in Petroleum Refining. *Chem. Eng. Sci.* **1996**, *51* (10), 1615.

(27) Quann, R. J. Modeling the Chemistry of Complex Petroleum Mixtures. *Environ. Health Perspect.* **1998**, *106*, 1441−1448.

(28) Prickett, S. E. Object-Oriented Generation of Complex Reaction Systems for Chemical Processes. Ph.D. Thesis, University of Maryland at College Park, 1995.

(29) Klinke, D. J., II; Broadbelt, L. J. Construction of a Mechanistic Model of Fischer−Tropsch Synthesis on Ni (111) and Co (0001) Surfaces. *Chem. Eng. Sci.* **1999**, *54*, 3379−3389.

(30) Prickett, S. E.; Mavrovouniotis, M. L. Construction of Complex Reaction Systems—I. Reaction Description Language. *Comput. Chem. Eng.* **1997**, *21* (11), 1219−1235.

(31) Mavrovouniotis, M. L.; Prickett, S. E. Generating Complex Systems in the Domain of Chemical Reactions. *Knowl.-Based Syst.* **1998**, *10*, 199−211.

(32) Tomlin, A. S.; Turanyi, T.; Pilling, M. J. Mathematical Tools for the Construction, Investigation and Reduction of Combustion Mechanisms. In *Low-Temperature Combustion and Autoignition*; Pilling, M. J., Ed.; Elsevier: Amsterdam, 1997; Vol. 35, pp 293−437.

(33) Okino, M. S.; Mavrovouniotis, M. L. Simplification of Mathematical Models of Chemical Reaction Systems. *Chem. Rev.* **1998**, *98* (2), 391−408.

(34) Maas, U.; Pope, S. B. Simplifying Chemical Kinetics: Intrinsic Low-Dimensional Manifolds in Composition Space. *Combust. Flame* **1992**, *88*, 239−264.

(35) Turanyi, T. Reduction of Large Reaction-Mechanisms. *New J. Chem.* **1990**, *14* (11), 795−803.

(36) Turanyi, T.; Berces, T.; Vajda, S. Reaction-Rate Analysis of Complex Kinetic Systems. *Int. J. Chem. Kinet.* **1989**, *21* (2), 83−99.

(37) Turanyi, T. Sensitivity Analysis of Complex Kinetic Systems—Tools and Applications. *J. Math. Chem.* **1990**, *5* (3), 203−248.

(38) Turanyi, T., Kinal—a Program Package for Kinetic-Analysis of Reaction- Mechanisms. *Comput. Chem.* **1990**, *14* (3), 253−254.

(39) Green, W. H.; Barton, P. I.; Bhattacharjee, B.; Matheu, D. M.; Schwer, D. A.; Song, J.; Sumathi, R.; Carstensen, H.-H.; Dean, A. M.; Grenda, J. M. Computer Construction of Detailed Chemical Kinetic Models for Gas-Phase Reactors. *Ind. Eng. Chem. Res.* **2001**, *40*, 5362−5370.

(40) Androulakis, I. P. Kinetic Mechanism Reduction Based on an Integer Programming Approach. *AIChE J.* **2000**, *46* (2).

(41) Edwards, K.; Edgar, T. F.; Manousiouthakis, V. I. Kinetic Model Reduction Using Genetic Algorithms. *Comput. Chem. Eng.* **1998**, *22* (1−2), 239−246.

(42) Petzold, L.; Zhu, W. J. Model Reduction for Chemical Kinetics: An Optimization Approach. *AIChE J.* **1999**, *45* (4), 869−886.

(43) Skodje, R. T.; Davis, M. J. Geometrical Simplification of Complex Kinetic Systems. *J. Phys. Chem.* **2001**, *105*, 10356−10365.

(44) Susnow, R. G.; Dean, A. M.; Green, W. H.; Peczak, P.; Broadbelt, L. J. Rate-Based Construction of Kinetic Mocdels for Complex Systems. *J. Phys.l Chem.* **1997**, *101*, 3731−3740.

(45) Lukyanov, D. B.; Gnep, N. S.; Guisnet, M. R. Kinetic Modeling of Propane Aromatization Reaction over Hzsm-5 and Gahzsm-5. *Ind. Eng. Chem. Res.* **1995**, *34* (2), 516−523.

(46) Katare, S.; Bhan, A.; Caruthers, J. M.; Delgass, W. N.; Venkatasubramanian, V. A Hybrid Genetic Algorithm for Efficient Parameter Estimation of Large Kinetic Models. Submitted for publication in *Comput. Chem. Eng.*

(47) Floudas, C. A.; Pardalos, P. M.; Adjiman, C. S.; Esposito, W. R.; Gumus, Z.; Harding, S. T.; Klepeis, J. L.; Meyer, C. A.; Schweiger, C. A. *Handbook of Test Problems for Local and Global Optimization*; Kluwer Academic Publishers: Dordrecht, 1999.

(48) Esposito, W. R.; Floudas, C. A. Global Optimization for the Parameter Estimation of Differential-Algebraic Systems. *Ind. Eng. Chem. Res.* **2000**, *39*, 1291−1310.

(49) Luus, R. Direct Search Luus-Jaakola Optimization Procedure. In *Encyclopedia of Optimization*; Floudas, C. A.; Pardalos, P. M., Eds.; Kluwer Academic Publishers: Dortrecht, 2001; Vol. 1, pp 440−444.

(50) Luus, R.; Jaakola, T. H. I. Optimization by Direct Search and Systematic Reduction of the Size of Search Region. *AIChE J.* **1973**, *19*, 760−766.

(51) Belohlav, Z.; Zamostny, P.; Kluson, P.; Volf, J. Application of Random-Search Algorithm for Regression Analysis of Catalytic Hydrogenations. *Can. J. Chem. Eng.* **1997**, *75*, 735−742.

(52) Box, G. E. P.; Draper, N. R. The Bayesian Estimation of Common Parameters from Several Responses. *Biometrika* **1965**, *52*, 355−365.

(53) Janusz, M.; Venkatasubramanian, V. Automatic Generation of Qualitative Description of Process Trends for Fault Detection and Diagnosis. *Eng. Appl. Artif. Intell.* **1991**, *4* (5), 329−339.

(54) Konstantinov, K. B.; Yoshida, T. Real-Time Qualitative Analysis of the Temporal Shapes of Bioprocess Variables. *AIChE J.* **1992**, *38* (11), 1703−1715.

(55) Whiteley, J. R.; Davis, J. F. Knowledge-Based Interpretation of Sensor Patterns. *Comput. Chem. Eng.* **1992**, *16* (4), 329−346.

(56) Mah, R. S. H.; Tamhane, A. C.; Tung, S. H.; Patel, A. N Process Trending with Piecewise Linear Smoothing. *Comput. Chem. Eng.* **1995**, *19* (2), 129−137.

(57) Oh, Y. S.; Moo, K. J.; Yoon, E. S.; Yoon, J. H. Fault Diagnosis Based on Weighted Symptom Tree and Pattern Matching. *Ind. Eng. Chem. Res.* **1997**, *36*, 2672−2678.

(58) Haimowitz, I. J.; Le, P. P.; Kohane, I. S. Clinical Monitoring Using Regression Based Templates. *Artif. Intell. Med.* **1995**, *7*, 473−496.

(59) Vedam, H. Op-Aide: An Intelligent Operator Decision Support System for Diagnosis and Assessment of Abnormal Situations in Process Plants. Ph.D. Thesis, Purdue University, West Lafayette, 1999.

(60) Dash, S.; Maurya, M. R.; Rengaswamy, R.; Venkatasubramanian, V. A Novel Interval-Halving Framework for Automated Identification of Process Trends. *AIChE J.* **2003**.

(61) Dash, S. Data-Driven Qualitative and Model-Based Quantitative Approaches to Fault Diagnosis. Ph.D. Thesis, Purdue University, School of Chemical Engineering, 2001.

(62) Venkatasubramanian, V.; Rengaswamy, R.; Kavuri, S. N.; Yin, K., Review of Process Fault Diagnosis—Part Iii: Process History Based Methods. *Comput. Chem. Eng.* **2003**, *27* (3), 327−346.

(63) Venkatasubramanian, V.; Rengaswamy, R.; Kavuri, S. N. Review of Process Fault Diagnosis—Part Ii: Qualitative Models and Search Strategies. *Comput. Chem. Eng.* **2003**, *27* (3), 313−326.

(64) Dunker, A. M., The Decoupled Direct Dmethod for Calculating Sensitivity Coefficients in Chemical Kinetics. *J. Chem. Phys.* **1984**, *81*, 2385−2393.

(65) Vajda, S.; Valkó, P.; Turányi, T., Principal Component Analysis of Kinetic Models. *Int. J. Chem. Kinet.* **1985**, *17*, 55−81.

(66) Li, G.; Rabitz, H.; Toth, J. A General Analysis of Exact Nonlinear Lumping in Chemical Kinetics. *Chem. Eng. Sci.* **1994**, *49*, 343−361.

(67) Lutz, A. E.; Kee, R. J.; Miller, J. A. *Senkin*: *A Fortran Program for Predicting Homogeneous Gas-Phase Chemical Kinetics with Sensitivity Analysis*; SAND87-8248; Sandia National Laboratories Report, 1988.

(68) Cukier, R. I.; Levine, H. B.; Shuler, K. E. Nonlinear Sensitivity Analysis of Multiparameter Model Systems. *J. Comput. Phys.* **1978**, *26*, 1−42.

(69) Bard, Y. *Nonlinear Parameter Estimation*; Academic Press: New York, 1974.

(70) Bates, D. M.; Watts, D. G. *Nonlinear Regression Analysis and Its Applications*; Wiley: New York, 1988; p 384.

(71) Cant, N. W.; Hicks, P. C.; Lennon, B. S. Steady-State Oxidation of Carbon Monoxide over Supported Nobel Metals with Particular Reference to Platinum. *J. Catal.* **1978**, *54*, 372−383.

(72) Prickett, S. E.; Mavrovouniotis, M. L. Construction of Complex Reaction Systems—Iii. An Example: Alkylation of Olefins. *Comput. Chem. Eng.* **1997**, *21* (12), 1325−1337.

(73) Boudart, M.; Djega-Mariadassou, G. *Kinetics of Heterogeneous Catalytic Reactions*; Princeton University Press: Princeton, NJ, 1984; p 243.

(74) Balakotaiah, V.; Chakraborty, S. Low-Dimensional Models for Describing Mixing Effects in Laminar Flow Tubular Reactors. *Chem. Eng. Sci.* **2002**, *57*, 2545−2564.

(75) Hostrup, M.; Balakrishna, S. Systematic Methodologies for Chemical Reaction Analysis. *Comput. Aided Chem. Eng.* **2001**, 401−406.

(76) Stolze, P. Microkinetic Simulation of Catalytic Reactions. *Prog. Surf. Sci.* **2000**, *65*, 65−150.

(77) Bandiera, J.; Taarit, Y. B. Ethane Conversion: Kinetic Evidnece for the Competition of Consecutive Steps for the Same Active Centre. *Appl. Catal., A* **1997**, *152*, 43−51.

(78) Lukyanov, D. B.; Gnep, N. S.; Guisnet, M. S. Kinetic Modelling of Ethene and Propene Aromatization over Hzsm5 and Gahzsm5. *Ind. Eng. Chem. Res.* **1994**, *33*, 223−234.

(79) Narbeshuber, T. F.; Brait, A.; Seshan, K.; Lercher, J. A. Dehydrogenation of Light Alkanes over Zeolites. *J. Catal.* **1997**, *172*, 127−136.

(80) Krannila, H.; Haag, W. O.; Gates, B. C. Monomolecular and Bimolecular Mechanisms of Paraffin Cracking: N−Butane Cracking Catalyzed by Hzsm5. *J. Catal.* **1992**, *135*, 115−124.

(81) Kazansky, V. B. Adsorbed Carbocations as Transition States in Heterogeneous Acid-Catalyzed Transformations of Hydrocarbons. *Catal. Today* **1999**, *51*, 419−434.

(82) Guisnet, M. S.; Gnep, N. S. Mechanism of Short-Chain Alkane Transformation over Protonic Zeolites, Alkylation, Dispropotionation and Aromatization. *Appl. Catal., A* **1996**, *146*, 33−64.

(83) Narbeshuber, T. F.; Vinek, H.; Lercher, J. A. Monomolecular Conversion of Light Alkanes over H-Zsm-5. *J. Catal.* **1995**, *157*, 388−395.

(84) Kazansky, V. B.; Frash, M. V.; van Santen, R. A. A Quantum-Chemical Study of Hydride Transfer in Catalytic Transformations of Paraffins on Zeolites. Pathways through Adsorbed Nonclassical Carbonium Ions. *Catal. Lett.* **1997**, *48*, 61−67.

(85) Buchanan, J. S.; Santiesteban, J. S.; Haag, W. O. Mechanistic Considerations in Acid-Catalyzed Cracking of Olefins. *J. Catal.* **1996**, *158*, 279−287.

(86) Aronson, M. T.; Gorte, R. J.; Farneth, W. E.; White, D. ^{13}C NMR Identification of Intermediates Formed by 2-Methyl-2-Propanol Adsorption in H-Zsm-5. *J. Am. Chem. Soc.* **1989**, *111*, 840−846.

(87) Kazansky, V. B. The Catalytic Site from a Chemical Point of View. *Stud. Surf. Sci. Catal.* **1994**, *85*, 251−272.

(88) Boronat, M.; Viruela, P.; Corma, A., A Theoretical Study of the Mechanism of the Hydride Transfer Reaction between Alkanes and Alkenes Catalyzed by an Acidic Zeolite. *J. Phys. Chem. A* **1998**, *102*, 9863−9868.

(89) Boronat, M.; Viruela, P.; Corma, A. Ab Initio and Density-Functional Theory Study of Zeolite-Catalyzed Hydrocarbon Reactions: Hydride Transfer, Alkylation and Disproportionation. *Phys. Chem. Chem. Phys.* **2000**, *2* (14), 3327−3333.

(90) Holland, J. H. *Adaptation in Natural and Artificial Systems*; University of Michigan: Ann Arbor, MI, 1975.

(91) Goldberg, D. E. *Genetic Algorithms in Search, Optimization, and Machine Learning*; Addison-Wesley: Reading, MA, 1989.

(92) Bay, S. D.; Shrager, J.; Pohorille, A.; Langley, P. Revising Regulatory Networks: From Expression Data to Linear Causal Models. *J. Biomed. Informatics*.

(93) Ni, T. C.; Savageau, M. A. Model Assessment and Refinement Using Strategies from Biochemical Systems Theory: Application to Metabolism in Human Red Blood Cells. *J. Theoret. Biol.* **1996**, *179*, 329−368.

(94) Ghosh, P. A Systematic Framework for Computer-Aided Design of Engineering Rubber Formulations. Ph.D. Thesis, Purdue University, West Lafayette, IN, 2002.

(95) Razon, L. F.; Schmitz, R. A. Intrinsically Unstable Behavior During the Oxidation of Carbon Monoxide on Platinum. *Catal. Rev.-Sci. Eng.* **1986**, *28* (1), 89−164.

(96) Lund, C. D.; Surko, C. M.; Maple, M. B.; Yamamoto, S. Y. Model Discrimination in Oscillatory Co Oxidation on Platinum Catalysts at Atmospheric Pressure. *Surf. Sci.* **2000**, *459*, 413−425.

(97) Mukesh, D.; Morton, W.; Kenney, C. N.; Cutlip, M. B. Island Models and the Catalytic Oxidation of Carbon Monoxide−Olefin Mixtures. *Surf. Sci.* **1984**, *138*, 237−257.

(98) Zambelli, T.; Wintterlin, J.; Trost, J.; Ertl, G. Identification of the "Active Sites" of a Surface-Catalyzed Reaction. *Science* **1996**, *273*, 1688−1690.

(99) Goryanin, I.; Hodgman, T. C.; Selkov, E. Mathematical Simulation and Analysis of Cellular Metabolism and Regulation. *Bioinformatics* **1999**, *15* (9), 749−758.

(100) Tomita, M.; Hasimoto, K. E.-Cell: Software Environment for Whole-Cell Simulation. *Bioinformatics* **1999**, *15* (1), 72−84.

(101) Mendes, P. Gepasi: A Software Package for Modleing the Dynamics, Steady States and Control of Biochemical and Other Systems. *Comput. Appl. Biosci.* **1993**, *9*, 563−571.

(102) You, L.; Hoonlor, A.; Yin, J. Modeling Biological Systems Using Dynetica—a Simulator of Dynamic Networks. *Bioinformatics* **2003**, *19*, 435−436.

(103) Kee, R. J.; Rupley, F. M.; Miller, J. A. *Chemkin*; SAND89-8003; Sandia National Laboratories Report, 1989.

Received for review August 14, 2003
Revised manuscript received November 24, 2003
Accepted December 1, 2003

IE034067H