Paper # 070RK-0277 **Topic: Reaction Kinetics**

8th U. S. National Combustion Meeting
Organized by the Western States Section of the Combustion Institute
and hosted by the University of Utah
May 19-22, 2013

Kinetics of the Reactions of H and CH_3 Radicals with *n*Butane: An Experimental Design Study using Reaction Network Analysis

David A. Sheen and Jeffrey A. Manion

Chemical and Biochemical Reference Data Division, National Institute of Standards and Technology, Gaithersburg, MD 20899

Abstract

The oxidation of hydrocarbon fuels proceeds through the attack of small radicals such as H and CH_3 on large molecules. These radicals abstract H atoms from the large molecules, which then usually proceed by β-scission to form C_2H_4 and C_3H_6. Quantifying these rates is critical to the development of chemical models for the oxidation of hydrocarbons. Study of this reaction system is confounded by the rapid dissociation of the intermediate radicals, which produces both additional H and additional CH_3, making it difficult to separate the behavior of the two radical species under many conditions.

In this work, we propose an experimental design algorithm that will be applied to measuring H and CH_3 attack rates on *n*-butane using a single-pulse shock tube. This design algorithm is based on the Method of Uncertainty Minimization using Polynomial Chaos Expansions (Sheen & Wang, Combust Flame 158, pp 2258-2374, 2011). We generate a set of proposed measurements covering a wide range of initial reactant concentrations, temperatures, and species concentration measurements. Hexamethylethane or *t*-butylperoxide, in mole fractions of less than 50 μL/L, generates H or CH_3 in the presence of *n*-butane, in mole fractions ranging from 100 μL/L to 100 000 μL/L. For some conditions, toluene was used as a radical-chain inhibitor in mole fractions of 20 000 μL/L. Temperatures ranged from 900 to 1100 K and pressures were assumed to be 2 atm. There are 16 different initial reactant concentrations and five temperatures for each concentration. For experiments using *t*-butylperoxide as a radical source, we consider measuring the absolute concentration of C_2H_4, C_3H_6 or measuring only the ratio between them; for hexamethylethane, we measure only the ratio, because the absolute concentrations depend on the highly-uncertain heating time. This gives a total of 160 proposed measurements.

To simulate the proposed experiments, we propose a candidate model to simulate these experiments, in this case the Jet Surrogate Fuel model. We then use a machine-learning algorithm to identify the best subset of experiments to perform. Of the original 160 conditions proposed, the algorithm selects seven as the best set. To test the machine algorithm, we compare its performance against an expert-recommended set of experimental measurements. The machine-generated experimental set performs better than the expert-generated experimental set. Therefore, the machine learning algorithm is therefore a suitable surrogate for an expert's evaluation of a set of experiments, and can be applied to many other database analysis and constraint problems.

Key Words: shock tube, kinetics, methyl radicals, H atoms, butane, uncertainty analysis, reaction networks

1 Introduction

The oxidation of hydrocarbon fuels proceeds by means of the attack of small radicals such as H and CH_3 on large molecules. These radicals abstract hydrogen atoms from the large molecules, which then usually proceed by β-scission to form C_2H_4 and C_3H_6. A quantitative understanding of the radical attack process is critical to the development of chemical models for the oxidation of hydrocarbons. Furthermore, it is necessary to determine how the uncertainty in measurements used to generate the model affects the model's final parameter estimates.

Butane is the simplest hydrocarbon that can react by H-abstraction and β-scission to form C_2H_4 and C_3H_6. It is therefore a useful subject for investigating the behavior of larger, more complex fuels. The radical (R) can attack H atoms bonded to either the primary or secondary carbon,

$$n\text{-}C_4H_{10} + R \leftrightarrow RH + C{\bullet}H_2C_3H_7 \qquad \text{(primary)}$$
$$n\text{-}C_4H_{10} + R \leftrightarrow RH + CH_3C{\bullet}HC_2H_5 \qquad \text{(secondary)}$$

The β-bond for $C{\bullet}H_2C_3H_7$ (p-C_4H_9) is the 2-3 bond, which upon β-scission produces ethylene and ethyl (which itself produces ethylene by H elimination). Conversely, the β-bond for $CH_3C{\bullet}HC_2H_5$ (s-C_4H_9) is the 3-4 bond, which upon β-scission produces propene and methyl. In principle, then, the branching ratio for the two processes could be determined by measuring the ratio of the ethylene and propylene produced in a single-pulse shock tube.

Typically, a single-pulse shock tube study begins by the selection of a set of experimental conditions that isolates the reactions under consideration. In the case of butane, the experimental condition would be a few parts per million of a radical precursor, either hexamethyl ethane (HME) to produce H atoms or di-tert-butyl peroxide (tBPO) to produce CH_3 radicals, in about 2% n-C_4H_{10} and 2% toluene with the balance argon. The purpose of the toluene is to act as an inhibitor for H atoms, which are eliminated from the C_2H_5 radicals and would confound measurements of CH_3 attack. Likewise, if the H atom pool builds too quickly in a study of H atom attack, the H atoms will begin to attack molecules other than the n-C_4H_{10}, producing ethylene by other pathways and thereby also confounding the measurements. Problematically, one process by which toluene inhibits the action of H atoms is by displacement of the CH_3 group to form benzene and a CH_3 radical, and so if the purpose of the experiment is to measure H atom attack on the parent fuel, some products will be produced by CH_3 radical attack. The measurements are further complicated by the self-combination of methyl radicals, which can produce ethylene and more H atoms.

Clearly, radical attack on saturated hydrocarbons is not sufficiently clean a system for traditional measurements. Unclean systems of this type are, however, amenable to analysis by large dataset analysis and optimization techniques such as DataCollaboration (Frenklach, Packard et al. 2004; Seiler, Frenklach et al. 2006; Russi, Packard et al. 2008) or the method of uncertainty minimization using polynomial chaos expansions (MUM-PCE) (Sheen and Wang 2011), as well as similar work by Turanyi and co-workers (Turányi, Nagy et al. 2012; Zsély, Varga et al. 2012). These techniques allow a complex model such as a chemical kinetic model to be constrained against a large number of experiments, characterizing the uncertainty in the model's parameters as a function of the uncertainty in the experimental measurements. At their core, these methods operate using Bayes' theorem. In the context of chemical kinetic modeling, it is usually common to talk about a candidate model $\mathcal{M}$ which is the collection of chemical and thermodynamic parameters along with their uncertainty. This model is then conditioned against a dataset $\mathcal{D}$ and an improved model is output; Bayes' theorem can then be expressed for this system as

$$p(\mathcal{M}^*) = p(\mathcal{M}|\mathcal{D}) = \frac{p(\mathcal{D}|\mathcal{M})p\left(\mathcal{M}^{(0)}\right)}{p(\mathcal{D})}$$

where $\mathcal{M}^{(0)}$ is the prior model and $\mathcal{M}^* = \mathcal{M}|\mathcal{D}$ is the improved posterior model, which has been improved by the measurements in the dataset $\mathcal{D}$.

Many chemical kinetic model optimization studies have been performed (Frenklach 1984; Yuan, Wang et al. 1991; Frenklach, Wang et al. 1992; Smith, Golden et al. 2000; Russi, Packard et al. 2008; Sheen, You et al. 2009; Sheen and Wang 2011; Sheen and Wang 2011; You, Russi et al. 2011; Turányi, Nagy et al. 2012; Zsély, Varga et al. 2012). Until very recently, however, there has not been any consideration of what comprises the best data set. When two measurements are taken independently by different researchers, it is assumed that those measurements are, in fact, statistically independent. However, the properties that are being measured are described by the same physical model, and indeed usually depend on a similar set of uncertain parameters in the model. As such, within the context of the physics that describes them, two seemingly independent measurements (say, of a laminar flame speed and of an ignition delay time) are not, in fact, independent, but are connected through the physics of the problem.

The objective of this paper is to present a machine learning algorithm that, if given a large database of experiments and a model for simulating them, is able to screen the database for redundancies and propose a new, reduced experimental database. This algorithm works by finding those experiments that have the greatest influence on a set of targeted applications. We apply the algorithm to selection of measurements of C_2H_4 and C_3H_6 in single-pulse shock tubes using mixtures of *n*-C_4H_{10}, toluene, HME and *t*BPO; the target applications are in this case the rate coefficients for the reactions of H abstraction by H and CH_3 from *n*-C_4H_{10}. The experimental dataset selected by this algorithm is compared with an expert-selected dataset and found to constrain the application rate coefficients generally better than the choices of the expert.

2 Methodology

2.1 Prior Model

The prior model, denoted $\mathcal{M}^{(0)}$, is the submodel for the oxidation of H_2, CO, and C_1-C_4 hydrocarbons from JetSurF 2, augmented as described in (Sheen, Rosado-Reyes et al. 2013). Rate constants are specified using a modified Arrhenius expression $k_n = A_n T^{b_n} \exp(E_n/T)$.

2.2 Application Dataset

Since the objective of the measurements is to determine the rate constants for radical attack on *n*-C_4H_{10}, the applications are the Arrhenius prefactors and activation energies for the reactions,

$n\text{-}C_4H_{10} + H \leftrightarrow H_2 + p\text{-}C_4H_9$	R633
$n\text{-}C_4H_{10} + H \leftrightarrow H_2 + s\text{-}C_4H_9$	R634
$n\text{-}C_4H_{10} + CH_3 \leftrightarrow CH_4 + p\text{-}C_4H_9$	R643
$n\text{-}C_4H_{10} + CH_3 \leftrightarrow CH_4 + s\text{-}C_4H_9$	R644

The application set $\mathcal{A}$ then consists of the Arrhenius prefactors A_{633}, A_{634}, A_{643} and A_{644}, the activation energies E_{633}, E_{644}, E_{643}, and E_{644}, as well as the ratios A_{633}/A_{634} and A_{643}/A_{644}, and the differences $E_{633} - E_{644}$ and $E_{643} - E_{644}$.

2.3 Experimental Dataset

The experimental measurements consist of measurements of C_2H_4 and C_3H_6 in mixtures of varying concentrations of *n*-C_4H_{10} and toluene, using HME or *t*BPO as a source of H or CH_3 radicals. The complete list of experiments is given in Table 1. In order to validate the experimental selection method, we consulted an outside expert to generate a dataset, denoted $\mathcal{D}^{expert}$. The expert suggested [W. Tsang, personal communication] that the experiments should be conducted with large excesses of toluene and butane, so this set consists of Experiments 1, 5, 6, 10, 125, and 130.

Mathematically, we characterize each measurement as a dataset element $\mathcal{D}_r$ so that the dataset $\mathcal{D} = \bigcup_r \mathcal{D}_r$, with $\bigcup_r$ denoting the union operator. Each $\mathcal{D}_r$ is described by a measurement value η_r^{obs} and a measurement uncertainty σ_r^{obs}, in addition to metadata such as composition, geometry, etc. Since the experiments have not yet been done, we take η_r^{obs} to be the same as the prior model's prediction and we assume a reasonable value for σ_r^{obs}, in this case 0.05 in the logarithm, equivalent to a 2σ uncertainty of 10%. Active parameters are determined in the same manner as (Sheen, Rosado-Reyes et al. 2013); briefly, For each experiment $\mathcal{D}_r$ and reaction rate parameter θ_i (either an Arrhenius prefactor or activation energy), the uncertainty-weighted sensitivity coefficient $\mathcal{S}_{i,k}$ was computed,

$$\mathcal{S}_{r,i} = \frac{d\eta_r}{d\theta_i}\frac{\theta_i}{\eta_r}\ln f_i \tag{1}$$

where η_r is the simulation prediction, θ_i is a generalized rate parameter and f_i is its uncertainty factor. The active rate parameters are those for which $\mathcal{S}_{r,i}/\mathcal{S}_{r,\max} > 0.02$. Uncertainty factors in the Arrhenius prefactors are taken from JetSurF 2 (Wang, Dames et al. 2010). Uncertainty factors in activation energies where estimated using $f_i = (E_k + T_c \ln F_k)/E_k$, where E_k is the activation energy of Rk and F_k is the uncertainty factor of Rk, with T_c = 1000 K. This formulation ensures that the activation energy contributes the same uncertainty to the rate constant as the Arrhenius prefactor at 1000 K. In simulations, the shock tube was treated as a homogeneous adiabatic reactor. Species concentrations following the shock were determined used the VODE solver (Brown, Byrne et al. 1989) to integrate the chemical rate equations supplied by Sandia CHEMKIN (Kee, Rupley et al. 1989) over a period of 500 μs. There are 25 active reactions and 44 active parameters, which are given in Table 2.

2.4 Model Constraint

Model constraint uses the method of uncertainty analysis using polynomial chaos expansions (MUM-PCE) (Sheen and Wang 2011). This is a method for finding the best set of rate parameters for a given set of experimental measurements. A prior model $\mathcal{M}^{(0)}$ is defined, and then conditioned against a set of experimental data $\mathcal{D}$, thus generating a posterior model $\mathcal{M}^{*}$, or, in probabilistic terms, $\mathcal{M}^{*} = \mathcal{M}|\left(\mathcal{D}, \mathcal{M}^{(0)}\right)$. It is assumed that the uncertain parameters in the model can be expressed as a vector $\mathbf{X} = \mathbf{x}^{(0)} + \mathbf{x}^{(1)}\boldsymbol{\xi}$, where $\mathbf{x}^{(0)}$ is the factorial variable vector whose elements are

$$x_i^{(0)} = \frac{\ln \theta_i/\theta_{i,0}}{\ln f_i} \tag{2}$$

where f_i is the uncertainty factor of the i^{th} generalized active parameter θ_i. $\boldsymbol{\xi}$ is a vector of independent, identically distributed normal random variables with mean 0 and variance 1, and $\mathbf{x}^{(1)}$ is a transformation matrix, so that $\mathbf{X}$ follows a multivariate normal distribution with mean $\mathbf{x}^{(0)}$ and covariance matrix $\boldsymbol{\Sigma} = \mathbf{x}^{(1)}\mathbf{x}^{(1)^T}$. $\mathcal{M}^{(0)}$ assumes $\mathbf{x}^{(0)} = 0$ and $\mathbf{x}^{(1)}$ equal to one-half times the identity matrix. This is equivalent to each rate coefficient being lognormally distributed about its nominal value with a 2σ uncertainty equal to its uncertainty factor. The rate parameters for the posterior model can then be estimated using Bayes' Theorem, which yields the following expression for the probability density function (PDF) of the rate parameters,

$$\ln P_{\mathbf{X}}(\mathbf{x}) \sim -\left[\sum_{r=1}^{N_e}\left(\frac{\eta_r(\mathbf{x}) - \eta_r^{\text{obs}}}{\sigma_r^{\text{obs}}}\right)^2 + \sum_{i=1}^{N_r} 4x_i^2\right] \tag{3}$$

where $\eta_r(\mathbf{x})$ is the model prediction of $\mathcal{D}_r$ as a function of the factorial variables $\mathbf{x}$, N_e is the number of experiments and N_r the number of active variables. This PDF is then approximated by a multivariate normal distribution with some $\mathbf{x}^{(0)*}$ and some $\boldsymbol{\Sigma}^{*}$. $\mathbf{x}^{(0)*}$ is found by finding the mode of the PDF in Eq. 3, equivalent to the least-squares optimization problem

$$\mathbf{x}^{(0)*} = \underset{\mathbf{x}}{\operatorname{argmax}} \ln P_{\mathbf{X}}(\mathbf{x}) \tag{4}$$

which is solved using the LMDIF solver in the MINPACK library (More, Garbow et al. 1999). $\boldsymbol{\Sigma}^{*}$ is found by linearizing the model predictions in the vicinity of $\mathbf{x}^{(0)*}$, which yields

$$\boldsymbol{\Sigma}^* = \left(\sum_{r=1}^{N_e} \frac{\mathbf{J}_r \mathbf{J}_r^T}{\left(\sigma_r^{\text{obs}}\right)^2} + 4\mathbf{I} \right)^{-1} \tag{5}$$

where $\mathbf{J}_r$ is the gradient of $\eta_r\left(\mathbf{x}^{(0)*}\right)$. Solution mapping (Frenklach 1984; Frenklach, Wang et al. 1992) is used to estimate the model predictions, which assumes that $\eta_r(\mathbf{X}) = \mathbf{X}^T \mathbf{b}_r \mathbf{X} + \mathbf{a}_r^T \mathbf{X} + \eta_0$, where $\mathbf{a}_r$ and $\mathbf{b}_r$ are expansion coefficients, calculated using the sensitivity-analysis-based method (SAB) (Davis, Mhadeshwar et al. 2004). Then $\mathbf{J}_r$ in Eq. 5 is $\mathbf{J}_r = 2\mathbf{b}_r \mathbf{x}^{(0)*} + \mathbf{a}_r$.

2.5 Experimental Discrimination

The form of Eq. 2 assumes that $p(\mathcal{D}|\mathcal{M}) = \prod_r p(\mathcal{D}_r|\mathcal{M})$, equivalent to saying that all of the experimental measurements are independent. This is not guaranteed or in fact very likely at all. A method such as DataCollaboration, because it assumes that the probability distributions are what is called an interval distribution, does not have this concern. However, interval distributions do not use a probabilistic interpretation, which some researchers have found to be problematic. Statistical methods such as MUM-PCE (Sheen and Wang 2011) and those employed by Turanyi and co-workers (Turányi, Nagy et al. 2012), on the other hand, are inherently probabilistic, but are therefore highly susceptible to over-constraining the data set by including many non-independent measurements.

If two measurements are not independent, we must figure out how to address this non-independence. The work of Turanyi and co-workers (Turányi, Nagy et al. 2012) uses a “size of the dataset” measure to normalize O(20,000) individual measurements of a species time history in the shock-tube oxidation of H_2 in O_2. This assumes that all of the measurements within a particular subset are equally correlated (and are independent of measurements outside that data subset), which might not be true. The algorithm proposed here addresses this question by finding the “most independent” set of experiments.

The amount of information provided by a particular experiment about a simulation can be estimated by the sensitivity S_H of the simulation’s uncertainty to the measurement uncertainty, which can be expressed as

$$S_{H,ij} = \frac{d\sigma_j^*}{d\sigma_i^{obs}} \frac{\sigma_i^{obs}}{\sigma_j^*} \tag{6}$$

where σ_i^{obs} is the uncertainty in the i^{th} experimental measurement and σ_j^* is the posterior uncertainty in the j^{th} simulation prediction. $S_{H,ij}$ is essentially the derivative of the entropy H of the simulation to that of the experimental measurement and it provides an estimate of how an incremental improvement in the measurement precision affects the simulation uncertainty, which gives an estimate of how important that experiment is with respect to the simulation under consideration. If $\mathbf{S}_H$ is cast as a matrix, then examining column i gives a measure of how much information is coming from experiment i into the simulations, while examining row j gives a measure of how much information is coming from the experiments into simulation j. A net information flux Φ_r can be defined as

$$\Phi_r = (\mathbf{S}_H^T \mathbf{S}_H - \mathbf{S}_H \mathbf{S}_H^T)_{rr} = \sum_{j=1}^{M} \left(S_{H,rj}^2 - S_{H,jr}^2 \right) \tag{7}$$

which is an estimate of the aggregate information coming into or out of an experiment. $\mathcal{D}^*$ is determined by the following method. Φ_r is calculated for each experiment, with uncertainties calculated by means of the method of uncertainty minimization using polynomial chaos expansions (MUM-PCE) (Sheen and Wang 2011). If any $\mathcal{D}_r$ has $\Phi_r < 0$, the one with the smallest value is removed from the experimental list and new values of Φ_r are calculated. This procedure is iterated until all remaining $\mathcal{D}_r$ have $\Phi_r > 0$. It should be noted that removed targets are not considered as applications; they are removed from consideration entirely. Additionally, it can be shown that, as the experimental uncertainty of a particular target, σ_i^{obs} gets smaller, its self-entropy derivative $S_{H,ii}$ approaches 1; this makes sense because the model uncertainty is now locked to the experimental uncertainty. However, at the same time, the other entropy derivatives $S_{H,ij}$ approach 0; in this case, simulations of other experiments depend on some chemistry that is not constrained by this experiment.

Examination of Φ_r reveals that $\sum_r \Phi_r = 0$, so that with each iteration approximately half of the members of $\mathcal{D} \cup \mathcal{A}$ will have $\Phi_r < 0$ and thus be eligible for removal. The presence of the application set $\mathcal{A}$, whose members must have $\Phi_r < 0$ because there is no measurement information about them, serves to increase Φ_r for the members of $\mathcal{D}$. This is a rigorous statement of the critical need for an experimental study to be performed with an application in mind; it does not make sense to simply do experiments in a vacuum. It is intuitively obvious that an application is necessary, but that it is an emergent property serves as a validation of the information flux selection criterion.

3 Results and Discussion

As can be seen in Table 2, the number of active reactions is much greater than four. For instance, the C_2H_4 concentration is strongly affected by R108 and R252,

$$2CH_3 \leftrightarrow H + C_2H_5 \qquad \text{R108}$$

$$C_2H_4 + H(+M) \leftrightarrow C_2H_5(+M) \qquad \text{R252}$$

The rates of H atom attack on toluene also affect the C_2H_4 concentration, toluene's purpose being to remove the H atoms before they can attack the *n*-C_4H_{10}. As expected, the system is not very clean, and would not be expected to be easily scrutinized by traditional methods, hence our desire to find which conditions give the most knowledge about this highly unclean system.

The entropy sensitivity matrix, $\mathbf{S}_H$, is presented in Fig. 1 for $\mathcal{D}^{(0)}$. Many of the measurement conditions are strongly correlated, as indicated by the large off-diagonal values of $\mathbf{S}_H$ and the comparatively small diagonal values. Furthermore, the correlations among the experiments under different conditions, for instance measurements of H attack and CH_3 attack, underscore how unclean the system is. However, they also indicate the possibility of finding a set of conditions that will effectively span the reaction rate space of interest.

Entropy sensitivity can be thought of as a kind of information flow, indicating how much a particular experiment tells about a particular simulation. The information flux Φ_r is then essentially a measure of the integrated information flux for a particular experiment. If this number is positive, then the r^{th} experiment provides more information about simulating other experiments than the other experiments provide about simulating it, and conversely if Φ_r is negative then other experiments provide more information about simulating the r^{th} experiment. The presence of application experiments is critical because two targets will always be coupled, that is, each will provide some information about simulating the others. In the limiting case of two targets and no applications, the information flux criterion will always eliminate one. Hence the application is critical to informing which measurements are the important ones.

In order to determine which experiments should be removed from consideration, the information flux Φ_r is calculated and presented in Fig. 2. It averages to 0 and so about half of the experiments are eligible for removal. The target with the lowest Φ_r is removed and the information flux recalculated until every target has a positive Φ_r. After the last iteration, six experiments remain, and their experimental conditions are listed in Table 3. $\mathbf{S}_H$ for the final iteration is presented in Fig. 3. Ideally, $\mathbf{S}_H$ would be diagonal, which would mean that there was very little cross-coupling among the members of $\mathcal{D}^*$; in reality it deviates from this ideal state, but it can be seen that the diagonals are much larger in the case of $\mathcal{D}^*$ than $\mathcal{D}^{(0)}$.

Once the final experimental list is determined, it could be asked whether this is actually the best set of experiments. To address this question, the uncertainties of the applications are presented in Table 4. The uncertainty of the reaction A factors is typically about 0.6 (compared to 1 for the prior model). The uncertainty of the activation energies is about 0.9, which is expected since the temperature range of the experiments is relatively small (900 K – 1100 K).

The information flux criterion selected seven experiments from the full set of 160. One question, then, is why these particular experiments were picked, given that many of the experiments are so similar. Another distressing result can be seen in Fig. 4, which shows the posterior uncertainties for all of the experiments. Many experiments have posterior uncertainties larger than 10%, which suggests that we would learn more about the model if we included these experiments, e.g. experiments 91-95. What is it, then, about these seven experiments that makes them the best?

We begin by compiling the expert-suggested dataset, $\mathcal{D}^{\text{expert}}$, which consists of six measurements, which are 10% n-butane, 2% toluene, and 50 parts per million HME or *t*BPO, at 900 and 1100 K. The uncertainties in $\mathcal{A}$ are tabulated in Table 4. The uncertainties for A_{643} and A_{644} are similar to (or slightly less than) those calculated using $\mathcal{D}^*$, but the uncertainties in A_{633} and A_{634} are much less in the case of $\mathcal{D}^*$ than for $\mathcal{D}^{\text{expert}}$, so the information flux algorithm can outperform the expert.

Some properties of $\mathcal{D}^*$ are obvious. For instance, in the experiments involving CH_3 attack on *n*-C_4H_{10}, (mixtures of *n*-C_4H_{10} and *t*BPO), measurements of $[C_2H_4]/[C_3H_6]$ and of absolute $[C_3H_6]$ are suggested, while measurements of absolute [C_2H_4] are contra-indicated. Obviously, these three values are related, and only two of them need be independently specified. It is not obvious why [C_2H_4] measurements are not included. To demonstrate this, we show in Fig. 5 the joint density functions of A_{643} and A_{644} for two cases, one where the reaction rates are constrained against the absolute [C_2H_4] and [C_3H_6] and one where the reaction rates are constrained against [C_2H_4]/[C_3H_6] and [C_3H_6]. It is impossible to see any difference in the figure; the joint density is slightly smaller in the latter case.

The information flux algorithm chose targets that predominantly have a low concentration of *n*-C_4H_{10}, contrary to the expert-selected data set. To explain why, we first examine how the uncertainty of the rate parameters for R643 and R644 changes depending on which measurements we choose to constrain the system. The marginal joint density function of A_{643} and A_{644} is shown in Fig. 6 as a function of the initial *n*-C_4H_{10} mole fraction when the system is constrained against measurements in mixtures of *n*-C_4H_{10} and *t*BPO. At an initial mole fraction of 0.01%, the lowest considered, the posterior uncertainty in A_{643} is the same as the prior value; there is no information about the rate of R643. When the initial mole fraction is increased to 0.1%, the uncertainty in A_{643} is reduced to 60% of its prior value. As the initial *n*-C_4H_{10} mole fraction is increased further, there is little change in the joint density function of A_{643} and A_{644}.

To address why there is so little information about A_{643} at low mole fractions of *n*-C_4H_{10}, we show in Fig 7. the marginal joint densities of A_{643} and A_{108} as a function of initial *n*-C_4H_{10} mole fraction. At 10%, there is no information about A_{108}, while A_{643} is strongly constrained. As the initial *n*-C_4H_{10} mole fraction is decreased, the uncertainty in the product $A_{643}A_{108}$ (as evidenced by the PDF's extend in the $y = x$ direction)is reduced at the expense of more uncertainty in the value of A_{643} (proportional to the ellipse's extent along the *x*-axis). Somewhere between a 0.1% and 0.01% initial *n*-C_4H_{10} mole fraction, there is a transition and A108 becomes strongly constrained, with very little information about A_{643}, similar to what was shown in Fig. 3. In these experiments, the thermal decomposition of *t*BPO very rapidly forms a large number of CH_3 radicals, and if [*n*-C_4H_{10}] is comparable to [CH_3], they are more likely to recombine with each other via R108 than to attack *n*-C_4H_{10} via R643 and R644. Most of the measured ethylene is formed through R108 rather than through R644, so measurements at low *n*-C_4H_{10} mole fractions are really measuring the rate of R108.

However, as the initial *n*-C_4H_{10} mole fraction is increased, more H atoms are formed through R644. When the initial *n*-C_4H_{10} mole fraction is 10%, H atoms are formed in sufficient quantity that a substantial amount of C_2H_4 and C_3H_6 comes from H attack rather than CH_3 attack. The joint density function of A633 and A634 is shown in Fig.8. At 0.1%, there is no information about these parameters, whereas at 10%, the uncertainty in A_{633}/A_{634} (the extent of the ellipse in the $y = -x$ direction) is constrained fairly strongly.

Figures 6 and 8 indicate that measuring [C_2H_4] and [C_3H_6] in a mixture of 10% *n*-C_4H_{10} and *t*BPO could provide a simultaneous measurement of all the title reactions, R633, R634, R643, and R644. This would seem to indicate that we can measure the CH_3 and H atom attack process simply with this one set of experiments. The information flux algorithm did not pick these experiments to measure A633/A634, however. In Fig. 9, we show the joint density function for A_{633} and A_{634}, as constrained by the [C_2H_4]/[C_3H_6] measurements in mixtures of *n*-C_4H_{10} and HME. For mixtures of 10% *n*-C_4H_{10} and HME, A_{633}/A_{634} is constrained about as well as it is by mixtures of 10% *n*-C_4H_{10} and *t*BPO. As the initial mole fraction is decreased, the uncertainty in A_{633}/A_{634} (the extent of the ellipse in the $y = -x$ direction) becomes smaller until 0.1% *n*-C_4H_{10}, after which it stops decreasing.

Toluene is used as an H-atom inhibitor. In experiments using *t*BPO, its purpose is to convert H atoms into CH_3 radicals via R674, thus reducing the confounding effect from H attack, which is almost 1000 times faster. In experiments using HME, the purpose is to reduce the H atom attack rate on butane by reducing the size of the H pool; the pool of CH_3 radicals never grows very large, so the confounding effect from CH3 attack is small, especially at low initial *n*-C_4H_{10} mole fractions. The question, then, what effect the toluene has on the measurements of the title rates. In Fig. 10, we present the joint density function for A_{633} and A_{634}, as in Fig. 6, except using toluene as a radical inhibitor. For fixed initial *n*-C_4H_{10} mole fraction, the uncertainty in A_{633}/A_{634} is slightly greater when toluene is used, although the effect is not easy to see in the figure.

The selection of 10% and 1% *n*-C_4H_{10} and *t*BPO represents a compromise. On the one hand, there is the desire to minimize the overlap between the H attack and CH_3 attack measurements, because two experiments measuring the same thing is bad; this requires low initial *n*-C_4H_{10}. On the other hand, for low initial *n*-C_4H_{10}, so much ethylene comes from R108 that A_{643} cannot be measured. Making measurements at an initial mole fraction of 1% *n*-C_4H_{10} and *t*BPO minimizes both of these effects. Conversely, experiments with 0.1% and 0.01% *n*-C_4H_{10} and HME are selected because there is no loss of information at low butane concentrations, but at high butane concentrations there is significant CH_3 production and therefore product formation through the CH_3 attack pathways.

Some selections are somewhat random. Experiments using HME and toluene are selected even though Fig. 10 suggests that these are not good experiments to choose. To see how important these experiments really are, we remove experiments with both HME and toluene and generate a new dataset, $\mathcal{D}^{*\dagger}$. The uncertainties in the reaction rate coefficients are shown in Table 4, and the uncertainties in $\mathcal{D}^{*\dagger}$ are essentially the same as those in $\mathcal{D}^{*}$. This means that the "objective function" being minimized by the algorithm, namely the uncertainty in the targeted rate parameters, is relatively flat, so that some fairly large motions within the experimental condition space are possible without changing the final rate parameter uncertainty. Whatever $\mathcal{D}^{*}$ the algorithm generates is therefore not unique; rather, the algorithm is presenting guidelines for how to select experimental measurements based on the initial dataset $\mathcal{D}^{(0)}$ that it was given as input.

4 Conclusion

A machine-learning algorithm was developed to determine a minimal set of experiments for constraining a model based on minimizing the uncertainty in the model's predictions of a set of applications. This algorithm was applied to measuring the rates of H-atom and methyl radical attack on normal butane. A candidate data set of 160 experimental conditions was compiled, and the algorithm chose seven conditions from this set. When the experimental set chosen by the algorithm was compared with an expert-selected set, it was found that the set chosen by the algorithm differed substantially from the expert-selected data set. In particular, the expert-selected set prefers large amounts of butane (several percent), whereas the algorithm generally preferred smaller amounts (0.1%). When the uncertainty in the applications resulting from using the algorithm-generated as opposed to the expert-selected data sets, it was found that the algorithm compared reasonably well at constraining the uncertainties in the rate constants of methyl radical attack, and better constrained the rate constants of H-atom attack than the expert-selected set. The machine-learning algorithm proposed here is, therefore, a reasonable surrogate for expert database analysis and evaluation and, although demonstrated here in the context of chemical kinetics, has potentially wide-reaching applications.

References

Brown, P. N., G. D. Byrne, et al. (1989). "VODE: A Variable-Coefficient ODE Solver." Journal on Scientific and Statistical Computing **10**(5): 1038-1051.

Davis, S. G., A. B. Mhadeshwar, et al. (2004). "A new approach to response surface development for detailed gas-phase and surface reaction kinetic model optimization." International Journal of Chemical Kinetics **36**(2): 94-106.

Frenklach, M. (1984). "Systematic Optimization of a Detailed Kinetic-Model Using a Methane Ignition Example." Combustion and Flame **58**(1): 69-72.

Frenklach, M., A. Packard, et al. (2004). "Collaborative data processing in developing predictive models of complex reaction systems." International Journal of Chemical Kinetics **36**(1): 57-66.

Frenklach, M., H. Wang, et al. (1992). "Optimization and Analysis of Large Chemical Kinetic Mechanisms Using the Solution Mapping Method - Combustion of Methane." Progress in Energy and Combustion Science **18**(1): 47-73.

Kee, R. J., F. M. Rupley, et al. (1989). CHEMKIN-II: A FORTRAN Chemical Kinetics Package for the Analysis of Gas-Phase Chemical Kinetics. Albequerque, NM, Sandia National Laboratories.

More, J., B. Garbow, et al. (1999). "Minpack." Retrieved March 6, 2012, from http://netlib.org/minpack/index.html.

Russi, T., A. Packard, et al. (2008). "Sensitivity Analysis of Uncertainty in Model Prediction." Journal of Physical Chemistry A.

Seiler, P., M. Frenklach, et al. (2006). "Numerical approaches for collaborative data processing." Optimization and Engineering **7**(4): 459-478.

Sheen, D. A., C. M. Rosado-Reyes, et al. (2013). "Kinetics of H atom attack on unsaturated hydrocarbons using spectral uncertainty propagation and minimization techniques." Proceedings of the Combustion Institute **34**(1): 527-536.

Sheen, D. A. and H. Wang (2011). "Combustion kinetic modeling using multispecies time histories in shock-tube oxidation of heptane." Combustion and Flame **158**(4): 645-656.

Sheen, D. A. and H. Wang (2011). "The method of uncertainty quantification and minimization using polynomial chaos expansions." Combustion and Flame **158**(12): 2358-2374.

Sheen, D. A., X. You, et al. (2009). "Spectral uncertainty quantification, propagation and optimization of a detailed kinetic model for ethylene combustion." Proceedings of the Combustion Institute **32**(1): 535-542.

Smith, G. P., D. M. Golden, et al. (2000). "GRI-Mech 3.0." Retrieved March 25, 2008, from http://www.me.berkeley.edu/gri_mech/.

Turányi, T., T. Nagy, et al. (2012). "Determination of rate parameters based on both direct and indirect measurements." International Journal of Chemical Kinetics **44**(5): 284-302.

Wang, H., E. Dames, et al. (2010). "A high-temperature chemical kinetic model of *n*-alkane (up to *n*-dodecane), cyclohexane, and methyl-, ethyl-, *n*-propyl and *n*-butyl-cyclohexane oxidation at high temperatures, JetSurF version 2.0." from http://melchior.usc.edu/JetSurF/JetSurF2.0.

You, X., T. Russi, et al. (2011). "Optimization of combustion kinetic models on a feasible set." Proceedings of the Combustion Institute **33**(1): 509-516.

Yuan, T., C. Wang, et al. (1991). "Determination of the Rate Coefficient for the Reaction H+O2-]Oh+O by a Shock-Tube Laser-Absorption Detailed Modeling Study." Journal of Physical Chemistry **95**(3): 1258-1265.

Zsély, I. G., T. Varga, et al. (2012). "Determination of rate parameters of cyclohexane and 1-hexene decomposition reactions." Energy **43**(1): 85-93.

Tables

Table 1. Experimental conditions in the initial dataset $\mathcal{D}^{(0)}$.

Composition (mole fraction)				
C_4H_{10}	Toluene	Precursor	Precursor	r [d]
10^{-1}	0.02	$5\ 10^{-5}$	tBPO	1-5 [a] ; 6-10 [b] ; 11-15 [c]
10^{-1}	0.02	$2.5\ 10^{-5}$	tBPO	16-20 ; 21-25 ; 26-30
10^{-2}	0.02	$5\ 10^{-5}$	tBPO	31-35 ; 36-40 ; 41-45
10^{-2}	0.02	$2.5\ 10^{-5}$	tBPO	46-50 ; 51-55 ; 56-60
10^{-3}	0.02	$5\ 10^{-5}$	tBPO	61-65 ; 66-70 ; 71-75
10^{-3}	0.02	$2.5\ 10^{-5}$	tBPO	76-80 ; 81-85 ; 86-90
10^{-4}	0.02	$5\ 10^{-5}$	tBPO	91-95 ; 96-100 ; 101-105
10^{-4}	0.02	$2.5\ 10^{-5}$	tBPO	106-110 ; 111-115 ; 116-120
10^{-1}	0	$5\ 10^{-5}$	HME	121-125 [c]
10^{-1}	0.02	$5\ 10^{-5}$	HME	126-130
10^{-2}	0	$5\ 10^{-5}$	HME	131-135
10^{-2}	0.02	$5\ 10^{-5}$	HME	136-140
10^{-3}	0	$5\ 10^{-5}$	HME	141-145
10^{-3}	0.02	$5\ 10^{-5}$	HME	146-150
10^{-4}	0	$5\ 10^{-5}$	HME	151-155
10^{-4}	0.02	$5\ 10^{-5}$	HME	156-160

a. Measuring [C_2H_4] ; b. Measuring [C_3H_6] ; c. Measuring [C_2H_4]/[C_3H_6] ; d. Five measurements at 50 K intervals from T_5 = 900 K to 1100 K.

Table 2. List of active rate coefficients and uncertainty factors. The title reactions are in bold.

		A		*E*	
n		*i*	*f*	*i*	*f*
	Title Reactions				
633	$\boldsymbol{C_4H_{10}+H \leftrightarrow p\text{-}C_4H_9+H_2}$	18	***3***	38	***1.2***
634	$\boldsymbol{C_4H_{10}+H \leftrightarrow s\text{-}C_4H_9+H_2}$	19	***3***	39	***1.2***
643	$\boldsymbol{C_4H_{10}+CH3 \leftrightarrow p\text{-}C_4H_9+CH_4}$	20	***3***	40	***1.2***
644	$\boldsymbol{C_4H_{10}+CH3 \leftrightarrow s\text{-}C_4H_9+CH_4}$	21	***3***	41	***1.2***
107	$2CH_3(+M) \leftrightarrow C_2H_6(+M)$	1	2	26	1.2
108	$2CH_3 \leftrightarrow H+C_2H_5$	2	5	27	1.2
131	$CH_4+CH_2{}^* \leftrightarrow 2CH_3$	3	5		-
252	$C_2H_4+H(+M) \leftrightarrow C_2H_5(+M)$	4	3	28	1.2
279	$C_2H_5+CH_3(+M) \leftrightarrow C_3H_8(+M)$	5	3		-
286	$C_2H_6+CH_3 \leftrightarrow C_2H_5+CH_4$	6	1.5	29	1.08
362	$C_3H_6+H \leftrightarrow C_2H_4+CH_3$	7	2	30	1.12
554	$1\text{-}C_4H_8+H(+M) \leftrightarrow sC_4H_9(+M)$	8	3	31	1.2
555	$1\text{-}C_4H_8+H \leftrightarrow C_2H_4+C_2H_5$	9	3	32	1.2
556	$1\text{-}C_4H_8+H \leftrightarrow C_3H_6+CH_3$	10	5	33	1.2
565	$2\text{-}C_4H_8+H(+M) \leftrightarrow sC_4H_9(+M)$	11	3	34	1.2
573	$i\text{-}C_4H_8+H \leftrightarrow i\text{-}C_4H_7+H_2$	12	3		-
574	$i\text{-}C_4H_8+H \leftrightarrow C_3H_6+CH_3$	13	3	35	1.2
582	$C_2H_4+C_2H_5 \leftrightarrow pC_4H_9$	14	3	36	1.2
592	$C_3H_6+CH_3(+M) \leftrightarrow sC_4H_9(+M)$	15	2	37	1.19
631	$nC_3H_7+CH_3(+M) \leftrightarrow C_4H_{10}(+M)$	16	2		-
632	$2C_2H_5(+M) \leftrightarrow C_4H_{10}(+M)$	17	2		-
673	$C_6H_5CH_3+H \leftrightarrow C_6H_5CH_2+H_2$	22	2	42	1.16
674	$C_6H_5CH_3+H \leftrightarrow C_6H_6+CH_3$	23	2	43	1.2
676	$C_6H_5CH_3+CH_3 \leftrightarrow C_6H_5CH_2+CH_4$	24	2	44	1.14
805	$C_6H_5CH_2+CH_3 \leftrightarrow C_6H_5C_2H_5$	25	2		-

Table 3. Experimental conditions in the final datasets $\mathcal{D}^*$ and $\mathcal{D}^{*\dagger}$.

Final dataset $\mathcal{D}^*$

		Composition (Mole fraction)					
i^*	i	C_4H_{10}	Toluene	Precursor	Precursor	T_5	Measurand
1	6	10^{-1}	0.02	0.00005	tBPO	900	C_3H_6
2	86	10^{-3}	0.02	0.000025	tBPO	900	C_2H_4/C_3H_6
3	89	10^{-3}	0.02	0.000025	tBPO	1000	C_2H_4/C_3H_6
4	96	10^{-4}	0.02	0.00005	tBPO	900	C_3H_6
5	145	10^{-3}	0	0.00005	HME	1100	C_2H_4/C_3H_6
6	150	10^{-3}	0.02	0.00005	HME	1100	C_2H_4/C_3H_6
7	156	10^{-4}	0.02	0.00005	HME	900	C_2H_4/C_3H_6

Final dataset, not considering experiments with toluene and HME, $\mathcal{D}^{*\dagger}$

$i^{*\dagger}$	i	C_4H_{10}	Toluene	Precursor	Precursor	T_5	Measurand
1	6	10^{-1}	0.02	0.00005	tBPO	900	C_3H_6
2	40	10^{-2}	0.02	0.00005	tBPO	1100	C_3H_6
3	41	10^{-2}	0.02	0.00005	tBPO	900	C_2H_4/C_3H_6
4	96	10^{-4}	0.02	0.00005	tBPO	900	C_3H_6
5	100	10^{-4}	0.02	0.00005	tBPO	1100	C_3H_6
6	145	10^{-3}	0	0.00005	HME	1100	C_2H_4/C_3H_6
7	151	10^{-4}	0	0.00005	HME	900	C_2H_4/C_3H_6

Table 4. Uncertainty in the applications $\mathcal{A}$ (A factors, activation energies E, and the parameter ratio) expressed as $\sigma^{(0)}/\sigma^*$ for the final datasets, $\mathcal{D}^*$ and $\mathcal{D}^{*\dagger}$, and the expert dataset $\mathcal{D}^{expert}$. If the uncertainty in an application is less than the corresponding value in $\mathcal{D}^*$, the entry is shown in bold italics.

	R633		R634		R643		R644		A_{633}/A_{634}	$E_{633}-E_{634}$	A_{643}/A_{644}	$E_{643}-E_{644}$
	A	E	A	E	A	E	A	E				
$\mathcal{D}^*$	0.66	0.86	0.62	0.92	0.64	0.84	0.52	0.86	0.41	0.79	0.51	0.81
$\mathcal{D}^{*\dagger}$	0.66	0.82	0.62	0.92	0.60	0.84	0.52	0.84	0.42	0.79	0.51	0.83
$\mathcal{D}^{expert}$	0.80	0.94	0.72	0.96	***0.60***	0.84	0.54	0.88	0.51	0.88	0.54	0.83

Figures

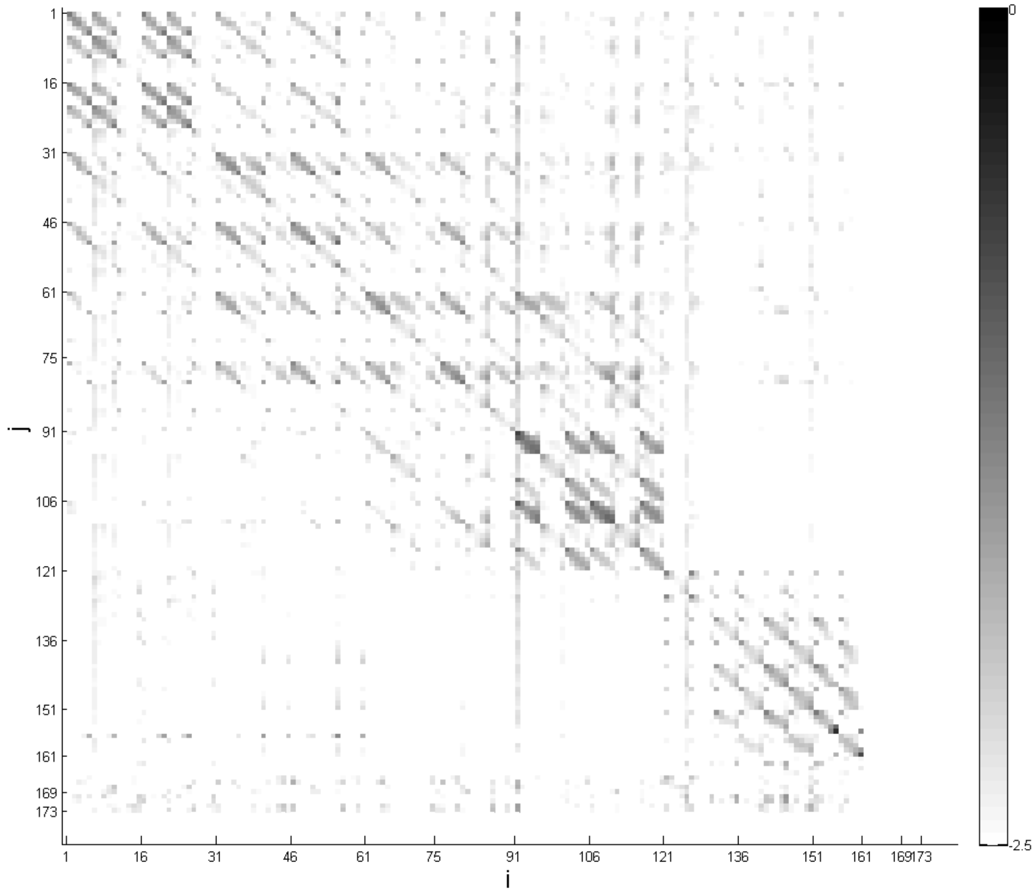

Figure 1. Entropy derivative matrix S_H for the experimental dataset $\mathcal{D}^{(0)}$ and application dataset $\mathcal{A}$. Values of *i* and *j* between 1 and 160 refer to experiments in $\mathcal{D}^{(0)}$; see Table 1 for index numbers. Values of *i* and *j* greater than 160 refer to the A-factors and activation energies in $\mathcal{A}$. Color indicates the value of $S_{H,ij}$

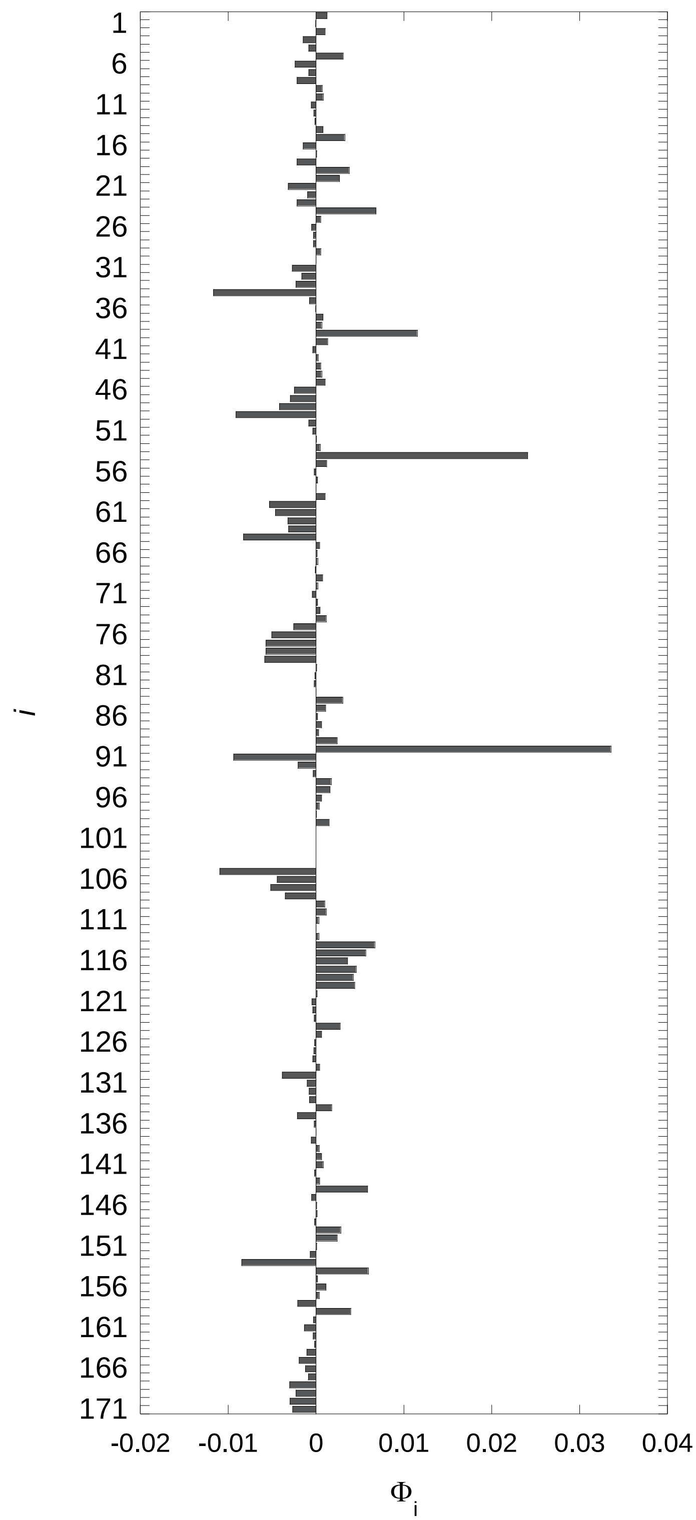

Figure 2. Information flux Φ_i for the experiments in $\mathcal{D}^{(0)}$. Values of *i* and *j* between 1 and 160 refer to experiments in $\mathcal{D}^{(0)}$; see Table 1 for index numbers. Values of *i* and *j* greater than 160 refer to the A-factors and activation energies in $\mathcal{A}$. See Table 1 for index numbers.

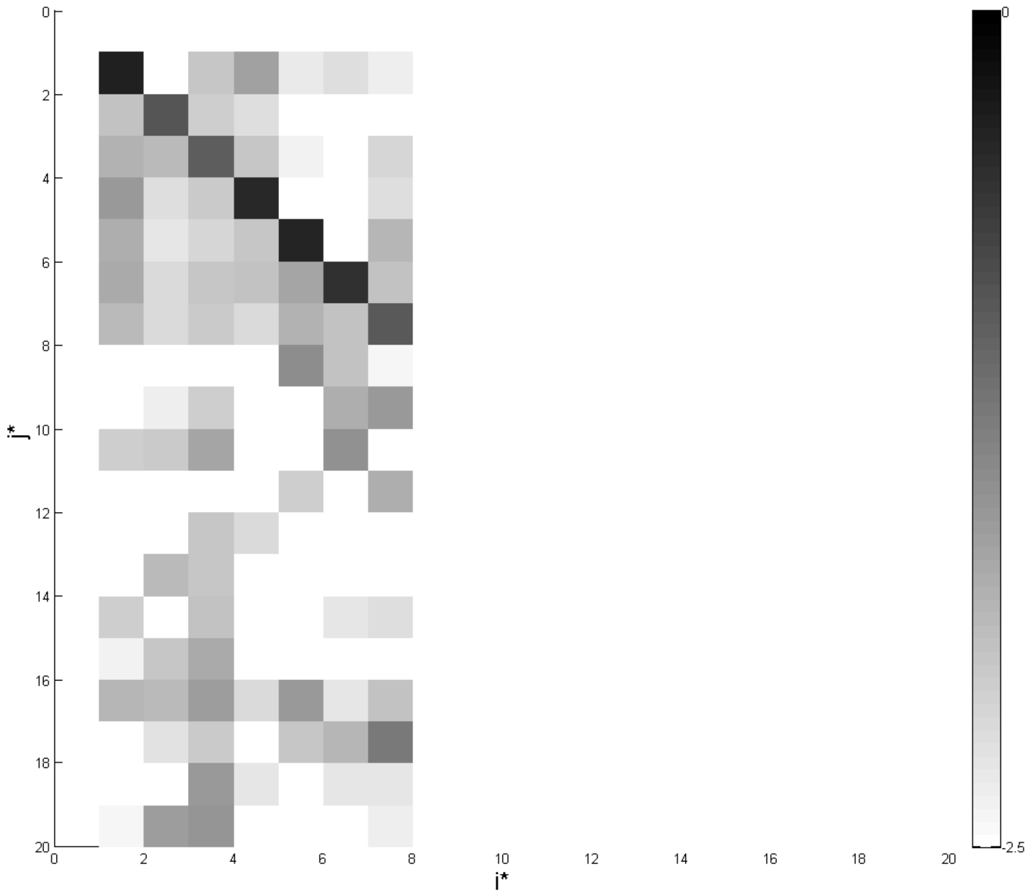

Figure 3. Entropy derivative matrix S_H for the experimental dataset $\mathcal{D}^*$ and application dataset $\mathcal{A}$. Values of i* and j* between 1 and 7 refer to experiments in $\mathcal{D}^*$; see Table 3 for index numbers. Values of i* and j* greater than 7 refer to the A-factors and activation energies in $\mathcal{A}$. Color indicates the value of $S_{H,ij}$

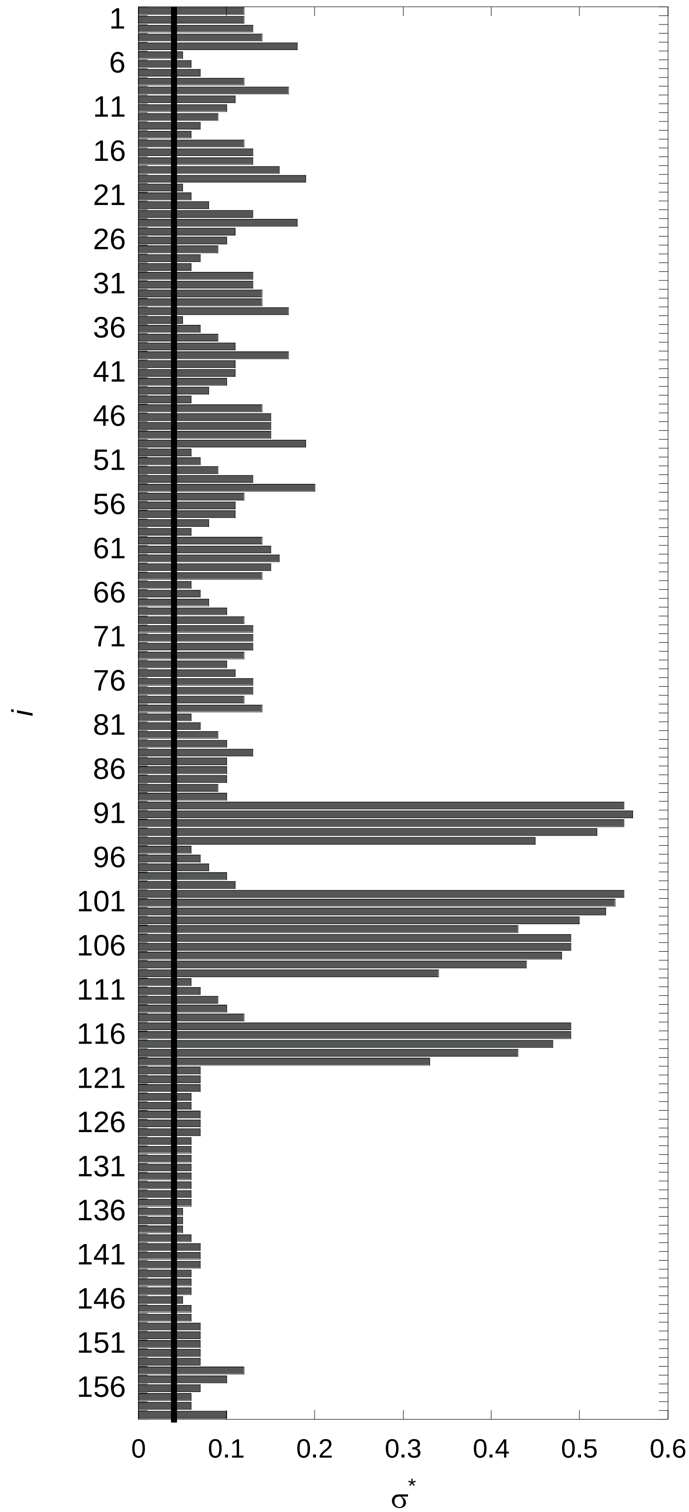

Figure 4. Posterior uncertainty for $\mathcal{M}^*|\mathcal{D}^*$. The experimental uncertainty of 0.05 is marked with the heavy black line. See Table 1 for index numbers.

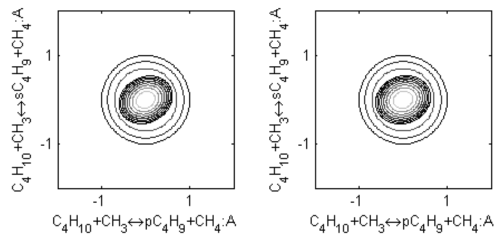

Figure 5. Joint probability density functions of the factorial variables corresponding to A_{643} and A_{644}. Concentric circles are the PDFs of the prior model. Concentric ellipses are the PDFs of the posterior models (left) considering measurements of absolute [C_2H_4] and [C_3H_6] and (right) considering measurements of absolute [C_3H_6] and [C_2H_4]/[C_3H_6]. The initial [*n*-C_4H_{10}] is 10% and the initial precursor is *t*-BPO.

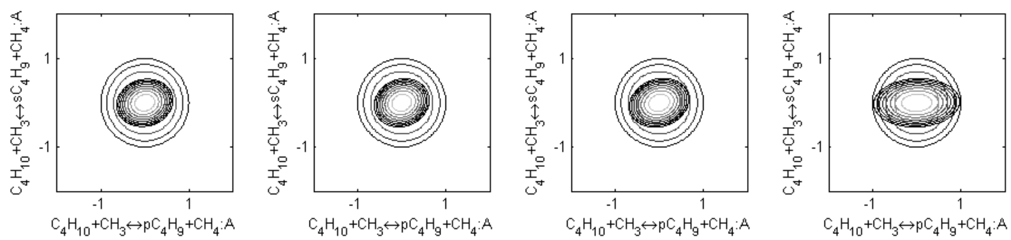

Figure 6. Joint probability density functions of the factorial variables corresponding to A_{643} and A_{644}. Concentric circles are the PDFs of the prior model. Concentric ellipses are the PDFs of the posterior models considering measurements of absolute [C_2H_4] and [C_3H_6]. The initial [*n*-C_4H_{10}] varies from 10% (far left) to 0.01% (far right), and the initial precursor is *t*-BPO.

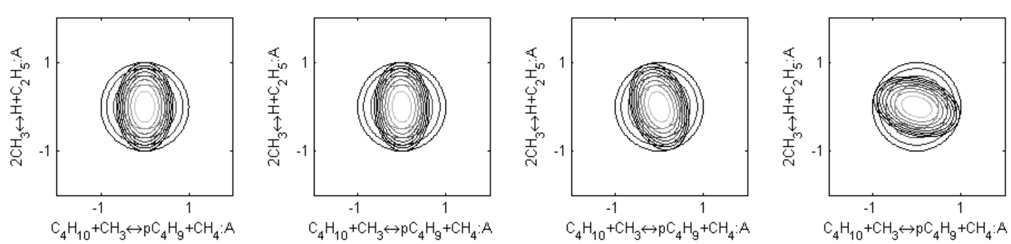

Figure 7. Joint probability density functions of the factorial variables corresponding to A_{643} and A_{108}. Concentric circles are the PDFs of the prior model. Concentric ellipses are the PDFs of the posterior models considering measurements of absolute [C_2H_4] and [C_3H_6]. The initial [*n*-C_4H_{10}] varies from 10% (far left) to 0.01% (far right) , and the initial precursor is *t*-BPO.

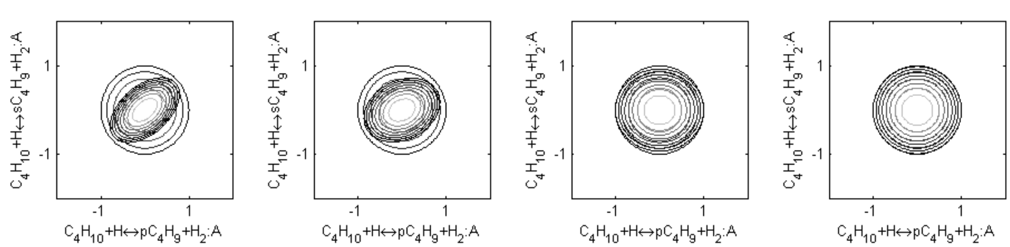

Figure 8. Joint probability density functions of the factorial variables corresponding to A_{633} and A_{634}. Concentric circles are the PDFs of the prior model. Concentric ellipses are the PDFs of the posterior models considering measurements of absolute [C_2H_4] and [C_3H_6]. The initial [*n*-C_4H_{10}] varies from 10% (far left) to 0.01% (far right) , and the initial precursor is *t*-BPO.

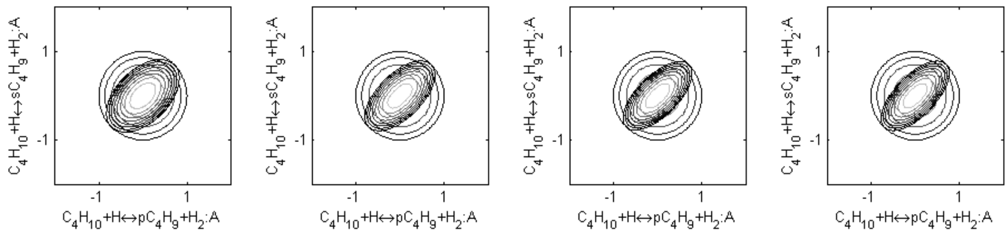

Figure 9. Joint probability density functions of the factorial variables corresponding to A_{633} and A_{634}. Concentric circles are the PDFs of the prior model. Concentric ellipses are the PDFs of the posterior models considering measurements of [C_2H_4]/[C_3H_6]. The initial [*n*-C_4H_{10}] varies from 10% (far left) to 0.01% (far right) , and the initial precursor is HME; no toluene is used.

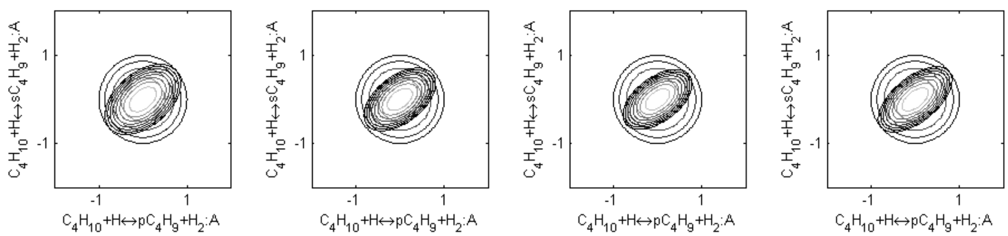

Figure 10. Joint probability density functions of the factorial variables corresponding to A_{633} and A_{634}. Concentric circles are the PDFs of the prior model. Concentric ellipses are the PDFs of the posterior models considering measurements of [C_2H_4]/[C_3H_6]. The initial [*n*-C_4H_{10}] varies from 10% (far left) to 0.01% (far right) , and the initial precursor is HME; 2% toluene is used as a radical inhibitor.